UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

ANÁLISIS DE SUPERVIVENCIA DEL USO DE UNA PLATAFORMA DIGITAL DE CONTRATACIONES "PER DIEM" POR PARTE DE PROFESIONALES EN ENFERMERÍA EN LOS ESTADOS UNIDOS DE AMÉRICA

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Profesional en Estadística

JAVIER PADILLA HUETE

Ciudad Universitaria Rodrigo Facio, Costa Rica

2026

DEDICATORIA Y AGRADECIMIENTOS

A mi familia y amistades, especialmente a mi mamá Yolanda y a mi papá José Carlos, por su amor incondicional, su ejemplo, sus enseñanzas, y su gran apoyo constante. Sin su acompañamiento, su paciencia y su confianza en mí, este logro académico no habría sido posible.

Extiendo también mi gratitud a todas las personas docentes que han acompañado mi formación desde la niñez hasta el presente. En particular, al profesor Gilbert Brenes y al profesor Guaner Rojas, por sus guías, sus enseñanzas y su disposición para acompañar este proceso; así como a Federico Castro, por su generosa ayuda y acompañamiento en el desarrollo de este trabajo.

Como señala Lev Vygotsky, "El aprendizaje humano presupone una naturaleza social específica y un proceso mediante el cual los niños acceden a la vida intelectual de aquellos que los rodean" y, en esa misma línea, "A través de otros llegamos a ser nosotros mismos".

Dado que el individuo es, en gran medida, creación de su entorno social, cultural y ambiental, este trabajo es fruto de todas las personas e instituciones que han formado parte del mío.

Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística.

ROY ALBERTO CAMPOS RETANA (FIRMA)

Digitally signed by ROY ALBERTO CAMPOS RETANA (FIRMA)
Date: 2026.06.03 18:33:39 -06'00'

Dr. Roy Campos Retana
Representante de la Decanatura
Sistema de Estudios de Posgrado

GILBERT BRENES CAMACHO (FIRMA)
PERSONA FISICA, CPF-01-0843-0678.
Fecha declarada: 04/06/2026 08:44:25 a. m.
Esta es una representación gráfica únicamente, verifique la validez de la firma.

Dr. Gilbert Brenes Camacho
Profesor Guía

Firmado por GUANER DAVID ROJAS ROJAS (FIRMA)
PERSONA FISICA, CPF-05-0328-0120.
Fecha declarada: 03/06/2026 12:41 PM
Razón: Aprobación
Lugar: Montes de Oca Contacto: guaner.rojas@ucr.ac.cr

Dr. Guaner Rojas Rojas
Lector

FEDERICO CASTRO MORA (FIRMA)
PERSONA FISICA, CPF-01-1326-0504.
Fecha declarada: 03/06/2026 11:32:06 PM
Esta es una representación gráfica únicamente, verifique la validez de la firma.

M.Sc. Federico Castro Mora
Lector

Dra. Alejandra Arias Salazar
Representante del Director del Programa de Posgrado en Estadística

JAVIER PADILLA HUETE (FIRMA)
PERSONA FISICA, CPF-01-1346-0629.
Fecha declarada: 2026/06/04/ 07:51:32 AM
Lugar: San José.

Javier Padilla Huete
Sustentante

TABLA DE CONTENIDO

RESUMEN

Este trabajo analiza la fuga de clientes (*churn*) en una plataforma digital de contratación "*per diem*" de personal de enfermería en Estados Unidos mediante modelos de análisis de supervivencia. El objetivo es caracterizar el tiempo de permanencia e identificar factores asociados al abandono de los usuarios entre diciembre de 2021 y octubre de 2024, de forma que además se identifiquen perfiles de usuarios más o menos vulnerables a dicho abandono del servicio.

Se trabajó con una base de datos de 48 437 usuarios. El análisis incluyó una etapa exploratoria, la estimación de curvas de supervivencia mediante el método de Kaplan-Meier y curvas de *hazard* no paramétricas, un modelo semi-paramétrico de Cox extendido con covariables dependientes del tiempo, y la estimación de seis modelos paramétricos. Las covariables incluidas fueron edad, región de trabajo, tipo de licencia de enfermería e ingresos generados en los primeros dos meses de uso de la plataforma. Se evaluaron los supuestos de los modelos y se comparó su desempeño mediante las métricas AIC, BIC, C-index y Brier Score Integrado con validación cruzada estratificada. Además, se llevó a cabo un análisis de sensibilidad de la definición de la fecha de censura. A partir de los modelos con mejor ajuste (Log-normal, Gamma y Cox extendido) se estimaron tiempos de supervivencia y probabilidades de permanencia, y se analizaron los efectos de las covariables para distintos perfiles de usuarios.

Los resultados indican que el riesgo de abandono es mayor en las primeras semanas tras el ingreso y disminuye con el tiempo; luego, alrededor de las 90 semanas (aproximadamente 21 meses), el riesgo de abandono vuelve a aumentar (se observa un repunte del riesgo instantáneo de abandono). La media de supervivencia para el usuario promedio de la plataforma (con ingresos en los primeros dos meses), según el modelo de Cox extendido es de 55.4±2.6 semanas (aproximadamente 12.8 meses), de acuerdo con el modelo Log-normal es de 48.8±1.4 semanas (aproximadamente 11.4 meses), y de acuerdo con el modelo Gamma es de 48.4±1.2 semanas (aproximadamente 11.3 meses). También se encontraron diferencias sustantivas en las probabilidades de supervivencia según edad, región, tipo de licencia e ingresos iniciales. En conjunto, los modelos muestran un desempeño predictivo aceptable y permiten identificar segmentos de alto riesgo, ofreciendo insumos concretos para diseñar estrategias de retención y optimizar la gestión de la plataforma.

ABSTRACT

This study examines *customer churn* in a digital *per diem* nursing staffing platform in the United States using survival analysis models. The objective is to characterize user retention time and identify factors associated with platform abandonment between December 2021 and October 2024, thereby identifying user profiles that are more or less vulnerable to *churn*.

A database of 48 437 users was analyzed. The study included an exploratory stage, estimation of survival curves using the Kaplan-Meier method and nonparametric hazard curves, a semi-parametric extended Cox model with time dependent coefficients, and the estimation of six parametric models. The covariates considered were age, work region, nursing license type, and earnings generated during the first two months of platform use. Model assumptions were assessed, and performance between models was compared using AIC, BIC, the C-index, and the Integrated Brier Score with stratified cross-validation. In addition, a sensitivity analysis was conducted regarding the definition of the censoring date. Based on the best-fitting models (Log-normal, Gamma, and extended Cox), survival times and retention probabilities were estimated, and covariate effects were examined across different user profiles.

The results indicate that churn risk is highest during the first weeks after onboarding and decreases over time; then, around 90 weeks (about 21 months), churn risk increases again (an uptick in the instantaneous hazard of churn is observed). For the average platform user (with earnings in the first two months), the mean survival time is 55.4±2.6 weeks (about 12.8 months) under the extended Cox model, 48.8±1.4 weeks (about 11.4 months) under the Log-normal model, and 48.4±1.2 weeks (about 11.3 months) under the Gamma model. Substantial differences in survival probabilities were also found by age, region, license type, and early earnings. Overall, the models show acceptable predictive performance and enable the identification of high-risk segments, providing concrete inputs to design retention strategies and optimize platform management.

LISTA DE CUADROS

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

ADN: *Associate Degree in Nursing* (grado de asociado en enfermería).

AED: Análisis exploratorio de los datos.

AFT: *Accelerated Failure Time* (tiempo de falla acelerado).

AIC: *Akaike Information Criterion* (criterio de información de Akaike).

ANOVA: *Analysis of Variance* (análisis de varianza).

AUC: Área bajo la curva ROC (*Receiver Operating Characteristic*).

BIC: *Bayesian Information Criterion* (criterio de información bayesiano).

BiLSTM: *Bidirectional Long Short-Term Memory* (memoria a largo y corto plazo bidireccional).

BS: *Brier Score*.

BSN: *Bachelor of Science in Nursing* (bachillerato en ciencias de la enfermería).

C-index: *Concordance Index* (índice de concordancia), también conocido como el estadístico C de Harrell.

CNA: *Certified Nursing Assistant* (asistente de enfermería certificado/a).

CNN: *Convolutional Neural Network* (red neuronal convolucional).

CRM: *Customer Relationship Management* (gestión de las relaciones con los clientes).

CSV: *Comma-Separated Values* (valores separados por comas).

CV: *Cross-Validation* (validación cruzada).

Churn: término utilizado para "fuga de clientes"; se refiere al abandono voluntario o involuntario de un cliente en un periodo determinado.

DNN: *Deep Neural Network* (red neuronal profunda).

EE: Error estándar (*Standard Error*).

FWER: *Family-Wise Error Rate* (Tasa de error familiar).

GNA: *Geriatric Nursing Assistant* (asistente de enfermería geriátrica).

GVIF: Factor de Inflación de la Varianza Generalizado.

HR: *Hazard Ratio* (razón de riesgos).

IBS: *Integrated Brier Score* (Brier Score integrado).

IC: Intervalo(s) de confianza.

IPCW: *Inverse Probability of Censoring Weighting* (pesos de probabilidad inversa de censura).

KM: Kaplan-Meier.

LPN: *Licensed Practical Nurse* (enfermera/o práctica/o licenciada/o).

LVN: *Licensed Vocational Nurse* (enfermera/o vocacional licenciada/o).

NA: *Not Available* (no disponible); se utiliza para denotar valores faltantes.

NCLEX: *National Council Licensure Examination* (examen de licencia del Consejo Nacional de Enfermería).

OECD: *Organisation for Economic Co-operation and Development* (Organización para la Cooperación y el Desarrollo Económicos).

PH: *Proportional Hazards* (riesgos proporcionales).

RN: *Registered Nurse* (enfermera/o registrada/o).

RNN: *Recurrent Neural Network* (red neuronal recurrente).

ROC: *Receiver Operating Characteristic* (curva característica operativa del receptor).

SSD: Servicio de Suscripción Digital (*Digital Subscription Service*).

UCR: Universidad de Costa Rica.

UI: *User Interface* (interfaz de usuario).

VC: Validación cruzada.

VE: Valores extremos.

VIF: *Variance Inflation Factor* (factor de inflación de la varianza).

CAPÍTULO I. INTRODUCCIÓN

1.1. CONTEXTO DEL ESTUDIO

En la era digital, la gestión de servicios mediante plataformas digitales ha experimentado un crecimiento exponencial, transformando diversos sectores, entre ellos, el de la salud. En particular, la contratación de personal de enfermería bajo el modelo "*per diem*" ha ganado relevancia, pues permite a los profesión ales aceptar turnos de trabajo de manera flexible y sin compromisos contractuales a largo plazo. No obstante, el éxito de estas plataformas está estrechamente vinculado a la retención de sus usuarios, dado que una alta tasa de deserción puede afectar la disponibilidad del personal y la estabilidad del sistema de atención sanitaria.

El fenómeno de "fuga de clientes" (*customer churn*) ha sido ampliamente estudiado en distintos sectores por su importante impacto en la sostenibilidad y rentabilidad de los negocios (Hung et al., 2006; Lai & Zeng, 2014; Larivière & Van den Poel, 2004; Masarifoglu & Buyuklu, 2019). En el contexto de la salud, la pérdida de usuarios en plataformas de contratación de personal puede comprometer la prestación de servicios, generando ineficiencias operativas y afectando la calidad de la atención ofrecida. Por ello, comprender los factores que influyen en la permanencia o abandono de los usuarios resulta fundamental para diseñar estrategias que optimicen la retención y garanticen la continuidad del servicio.

Estudios previos han demostrado que la "fuga de clientes" puede tener un impacto financiero significativo en las empresas, reduciendo la rentabilidad y aumentando los costos de adquisición de nuevos clientes (Keaveney & Parthasarathy, 2001; Reichheld & Sasser, 1990; Lai & Zeng, 2014; Hung et al., 2006; Khattak et al., 2023). De hecho, investigaciones como las de Reichheld y Sasser (1990) han evidenciado que la retención de clientes puede generar beneficios sustanciales a largo plazo. Además, de acuerdo con Lai y Zeng (2014), una gestión eficiente de la retención de clientes constituye una estrategia clave para optimizar el desempeño empresarial, especialmente en sectores donde la lealtad del usuario se traduce en estabilidad y crecimiento.

Dada la relevancia económica, operativa y asistencial de la fuga de clientes, por sus efectos sobre la rentabilidad de las plataformas, la calidad del servicio y la continuidad de la atención, mencionados anteriormente, este estudio se centra en analizar el fenómeno de "fuga de clientes" (*customer churn*) en una plataforma digital (página web y aplicación móvil) de contratación *per diem* de personal de enfermería en los Estados Unidos. En este contexto, la deserción de usuarios constituye un desafío significativo, ya que puede generar pérdidas económicas, afectar la calidad del servicio y comprometer la eficiencia del sistema de salud. Por ello, comprender los factores que inciden en la permanencia o abandono de los usuarios resulta clave para diseñar estrategias que optimicen la retención y fortalezcan la sostenibilidad de la plataforma.

Para abordar la problemática de la fuga de clientes dentro de la plataforma digital en estudio, se emplearon modelos de análisis de supervivencia, una metodología estadística que permite estudiar el tiempo hasta que ocurre un evento de interés, en este caso, la deserción de los usuarios de la plataforma. A través de modelos como el estimador de Kaplan-Meier, el modelo de riesgos proporcionales de Cox y diversos modelos paramétricos, se pretende identificar patrones de comportamiento y evaluar los factores que influyen en la retención de los profesionales de enfermería dentro de la plataforma.

A diferencia de los modelos tradicionales de clasificación, el análisis de supervivencia no solo permite determinar la probabilidad de que un usuario abandone la plataforma, sino también predecir cuándo es más probable que esto ocurra. Esta capacidad predictiva es crucial para la formulación de estrategias de retención más efectivas. Además, este tipo de análisis permite estudiar y cuantificar el impacto de diversas variables explicativas en la probabilidad de abandono de los usuarios.

En síntesis, este estudio aplica el análisis de supervivencia como una metodología robusta para investigar la "fuga de clientes" en plataformas de contratación de personal de enfermería en los Estados Unidos. Mediante el desarrollo de modelos estadísticos, se busca identificar patrones de abandono y evaluar los factores que influyen en la permanencia de los profesionales dentro de la plataforma, contribuyendo así tanto al conocimiento académico como a la optimización de la gestión en entornos digitales de salud.

1.2. ANTECEDENTES

El análisis de la deserción de clientes, conocido como *"churn"*, ha sido ampliamente estudiado en diversos sectores mediante técnicas de análisis de supervivencia. Estas metodologías permiten modelar el tiempo hasta que ocurre un evento específico, como la decisión de un cliente de abandonar un servicio. Sin embargo, la aplicación de estas técnicas en el ámbito de la salud, específicamente en plataformas digitales de contratación "*per diem*" para profesionales de enfermería, es sumamente limitada.

En el contexto costarricense, destaca el estudio reciente de Ortiz Robles (2022), quien analizó la probabilidad de retiro de los funcionarios de los Ministerios de Gobierno a corto, mediano y largo plazo mediante modelos de supervivencia. Para ello, empleó datos históricos de funcionarios públicos. Las técnicas de análisis aplicadas demostraron ser especialmente valiosas para el estudio de eventos temporales como el retiro laboral, ya que permiten examinar no solo la ocurrencia del evento, sino también su probabilidad y los factores asociados, superando así las limitaciones de los métodos tradicionales.

En dicho estudio de Ortiz Robles (2022) se utilizaron diversos modelos de supervivencia, entre ellos el modelo de Cox y los modelos paramétricos Exponencial, Normal, Weibull, Logístico, Log-logístico y Log-normal. Además, se estableció una metodología para determinar cuál de estos modelos proporcionaba estimaciones más precisas del comportamiento de retiro de los funcionarios. Para evaluar el rendimiento de los modelos, se utilizaron los siguientes estadísticos: Índice C de Harrell, el Área bajo la curva ROC con Sensibilidad al Incidente y Especificidad Dinámica (I/D), el *Brier Score* como medida de exactitud, y la Desviación estándar del estadístico C para evaluar la variabilidad en las predicciones. En ese estudio se encontró que el modelo con predicciones más confiables fue el modelo Logístico.

Entre los hallazgos más relevantes del estudio de Ortiz Robles (2022), se encontró que el comportamiento del retiro varía significativamente según las variables analizadas, particularmente el sexo, la categoría profesional y la edad inicial al ingresar al servicio público. En particular, se observó una mayor probabilidad de retiro en mujeres y en funcionarios no profesionales, especialmente durante los primeros cinco años de servicio. De

esta forma, este estudio, realizado en el ámbito laboral costarricense, demuestra que los modelos de supervivencia son herramientas sólidas para analizar eventos dinámicos y generar resultados valiosos para la gestión de recursos humanos y la formulación de políticas públicas.

Por su parte, el estudio de Sánchez Brenes (2021), analizó la reincidencia de personas egresadas del sistema penitenciario costarricense entre 2016 y 2018, con el objetivo de estimar el tiempo hasta una nueva privación de libertad y determinar los factores asociados a este evento. Para ello, se utilizaron modelos de análisis de supervivencia, como Kaplan-Meier, pruebas de Log-Rank y Wilcoxon, así como modelos paramétricos (Weibull y Log-logístico) y el modelo semi-paramétrico de Cox. Estos métodos permitieron comparar grupos y evaluar la influencia de variables como el sexo, la edad al egreso, el tipo de delito, y la duración de la condena.

Los resultados del estudio de Sánchez Brenes (2021) mostraron que la edad y el tipo de delito son factores significativos tanto en la probabilidad de reincidencia como en el tiempo hasta que esta ocurre. Por ejemplo, las personas que egresan del sistema penitenciario a una edad más avanzada presentan menores tasas de reincidencia. Este estudio demuestra la relevancia de los modelos de supervivencia en investigaciones sobre eventos que evolucionan a lo largo del tiempo, ya que permiten incorporar información censurada y analizar con mayor precisión los factores asociados a la ocurrencia y duración de eventos como la reincidencia, fortaleciendo así su valor en estudios criminológicos y de política pública.

En el contexto de estudiantes universitarios, el estudio de Gallardo Allen et al. (2016) analizó el tiempo necesario para que los estudiantes de la Universidad de Costa Rica (UCR) obtuvieran el grado de bachillerato universitario, considerando su relación con factores sociodemográficos y de aptitud académica. Los resultados indicaron que las mujeres y los estudiantes de sedes regionales tenían una mayor probabilidad de graduarse en menos tiempo. En contraste, llamó la atención que el nivel socioeconómico no resultó un factor determinante, posiblemente debido al impacto de las becas otorgadas. Asimismo, se encontró que aproximadamente un 38% de los estudiantes desertaban antes de completar sus estudios. Al excluir a este grupo, se estimó que el 65% de los alumnos que permanecen logran

graduarse en un plazo de ocho años, un tiempo considerablemente alto en comparación con otras universidades dentro y fuera de Costa Rica.

Para analizar el tiempo hasta la graduación, Gallardo Allen et al. (2016) emplearon técnicas de análisis de supervivencia, en particular el estimador de Kaplan-Meier y pruebas de Log-Rank, para comparar las curvas de graduación a lo largo del tiempo. Además, aplicaron modelos de regresión de Cox para evaluar la influencia de variables como el tipo de colegio de procedencia, el género, la puntuación en pruebas de aptitud académica y la recepción de becas socioeconómicas sobre el tiempo de graduación. Estos modelos permitieron identificar factores asociados a la permanencia y al éxito académico, demostrando su utilidad en estudios educativos y de retención estudiantil.

Por otro lado, a nivel internacional, uno de los sectores con mayor número de estudios relacionados con el *churn*, es el de las telecomunicaciones. Entre estos se encuentra el estudio de Masarifoglu y Buyuklu (2019), quienes emplearon análisis de supervivencia para estimar la probabilidad de abandono del servicio por parte de los clientes de una empresa de telecomunicaciones, considerando variables como campañas promocionales, tarifas y métodos de pago. Este estudio enfatiza la importancia de analizar el *churn* voluntario, es decir, la decisión del cliente de abandonar el servicio por factores como precio o calidad, para identificar oportunamente los perfiles de usuarios propensos a cancelar su suscripción. Para abordar esta problemática, los autores aplicaron técnicas de análisis de supervivencia, específicamente el modelo de riesgos proporcionales de Cox, utilizando datos reales de la empresa de telecomunicaciones.

El objetivo del estudio de Masarifoglu y Buyuklu (2019) fue evaluar cómo variables como la edad, la duración del contrato, el tipo de tarifa, la participación en campañas y el uso del método de pago automático afectan el tiempo de permanencia del cliente. Los resultados mostraron que estos factores tienen un impacto significativo en la probabilidad de cancelación del servicio, lo que permite a la empresa segmentar a sus clientes y diseñar estrategias proactivas de retención. En esta investigación, el análisis de supervivencia demostró ser una herramienta eficaz para anticipar el *churn* y orientar la toma de decisiones dentro del marco de la gestión de relaciones con el cliente.

De forma similar, el estudio de Monika et al. (2021), realizado en Indonesia, analizó el tiempo hasta el *churn* de clientes en una empresa de telecomunicaciones, aplicando modelos de análisis de supervivencia para comparar enfoques no paramétricos y semi-paramétricos. El objetivo principal fue determinar cuál de estos métodos proporciona estimaciones más precisas para modelar la duración de la permanencia de los clientes, un aspecto clave en la retención de usuarios en servicios competitivos como las telecomunicaciones. Se encontró que el modelo semi-paramétrico de Cox presenta un mejor ajuste en comparación con los métodos no paramétricos como Kaplan-Meier, permitiendo evaluar el impacto individual de cada covariable en el tiempo hasta la desvinculación. El desempeño de los modelos se comparó utilizando métricas como el error estándar y medidas de bondad de ajuste. Los resultados sugieren que los modelos semi-paramétricos permiten mayor precisión al incluir múltiples factores, lo cual los hace especialmente útiles en la planificación de estrategias de retención de clientes.

Otro estudio destacado sobre *churn* es el de Lai y Zeng (2014), quienes analizaron el comportamiento de abandono de usuarios en bibliotecas digitales, una problemática poco explorada a pesar de su impacto directo en la sostenibilidad y efectividad de estos servicios. Los autores reconocen que, al igual que en otras industrias de servicios, el *churn* de usuarios puede tener consecuencias negativas en los ingresos y en el retorno de inversión en captación de clientes. Por ello, comprender la duración de la relación entre los usuarios y las bibliotecas digitales se vuelve crucial para implementar estrategias efectivas de retención. En este estudio, el uso de modelos de supervivencia demostró ser fundamental para comprender mejor la duración de la relación cliente-servicio y anticipar comportamientos de abandono.

Para abordar esta problemática, Lai y Zeng (2014) aplicaron métodos no paramétricos de análisis de supervivencia sobre una muestra de 8 054 usuarios de una biblioteca digital en China. En particular, utilizaron el estimador de Kaplan-Meier y un modelo de riesgos proporcionales para analizar el comportamiento del riesgo de abandono. Los resultados revelaron una alta tasa de *churn*, especialmente durante los primeros tres meses posteriores al registro, así como diferencias marcadas en el tiempo de permanencia entre distintos grupos de usuarios. Los hallazgos subrayan la importancia de integrar el análisis de supervivencia

en la gestión de relaciones con los usuarios en bibliotecas digitales, ya que esta metodología permite no solo modelar la duración de la relación con cada usuario, sino también identificar momentos críticos de abandono y patrones de comportamiento específicos susceptibles de intervención mediante estrategias personalizadas.

En el ámbito financiero, el estudio de Mavri e Ioannou (2008) analizó el comportamiento de *churn* de clientes en el sector bancario griego mediante técnicas de análisis de supervivencia. El objetivo principal fue evaluar la duración de la relación entre los clientes y los bancos, e identificar los factores que influyen en la decisión de cambiar de entidad financiera. Para ello, se utilizaron el estimador de Kaplan-Meier, que permitió describir las curvas de supervivencia sin necesidad de asumir una distribución específica, y el modelo de riesgos proporcionales de Cox, que permitió estimar el impacto de variables sociodemográficas y de percepción del servicio sobre el tiempo de permanencia de los clientes.

Entre los hallazgos más relevantes del estudio de Mavri e Ioannou (2008), se identificó que los clientes más jóvenes, con mayor nivel educativo e ingresos más altos, presentaban una mayor propensión a cambiar de banco. Además, factores como la percepción de la calidad del servicio y el uso de canales electrónicos influyeron significativamente en la duración de la relación cliente-banco. Este estudio es relevante porque demuestra cómo los modelos de análisis de supervivencia permiten capturar con precisión el comportamiento temporal del abandono, lo que los convierte en herramientas particularmente útiles para la toma de decisiones estratégicas en contextos donde la retención de usuarios es prioritaria. Su aplicación en el sector bancario pone de manifiesto la utilidad y versatilidad de estas técnicas en una amplia variedad de sectores.

Por su parte, Larivière y Van den Poel (2004) llevaron a cabo un estudio en una importante entidad financiera belga, en el cual analizaron el papel de las características de los productos en la prevención del *churn* de clientes. El objetivo principal fue analizar cómo atributos específicos de los productos de ahorro e inversión, como su duración (plazo fijo vs. indefinido) y el riesgo asociado, influyen en la decisión y el momento en que los clientes abandonan la entidad. Con este propósito, los autores categorizaron los productos según estas dimensiones y emplearon técnicas de análisis de supervivencia. Inicialmente, utilizaron el

estimador de Kaplan-Meier para obtener una visión exploratoria de los tiempos de abandono en las diferentes categorías de productos. Posteriormente, aplicaron un modelo de *split-hazard* (una variante del análisis de supervivencia) para identificar los factores demográficos y de relación con el cliente (como la edad, la antigüedad y el número de productos contratados) que predicen la probabilidad y el momento del *churn*.

Entre los hallazgos más relevantes del estudio de Larivière y Van den Poel (2004), se evidenció que tanto los patrones como el momento del abandono varían significativamente según las características del producto, particularmente su duración y nivel de riesgo. Además, se identificó que las variables asociadas al cliente y a la profundidad de su relación con la entidad también son predictores clave de la propensión al abandono. Este estudio resulta especialmente relevante porque subraya la necesidad de no tratar el *churn* de manera homogénea, sino de incorporar las características específicas de los productos en el análisis. De esta forma, las instituciones financieras pueden comprender mejor qué clientes son más propensos a abandonar según el tipo de portafolio que poseen y, en consecuencia, diseñar estrategias de retención más específicas y efectivas.

Además de los modelos estadísticos de análisis de supervivencia, la problemática del *churn* también se ha estudiado en los últimos años mediante modelos de *Machine Learning* y *Deep Learning*, los cuales presentan ventajas y desventajas en comparación con los enfoques más clásicos. Por ejemplo, en Chile, Pérez Pizarro (2022) investigó la fuga de clientes en un servicio de suscripción digital, es decir, en plataformas donde los usuarios pagan periódicamente por acceder a contenidos como videos, noticias o software. El objetivo principal fue identificar los factores que explican el abandono del servicio y predecir qué clientes tienen mayor probabilidad de cancelarlo. Para ello, utilizó datos de 5 000 clientes, tanto activos como inactivos, considerando variables demográficas, de uso del servicio, tipo de dispositivo e interacción con el soporte técnico.

En este estudio, Pérez Pizarro (2022) aplicó diversas técnicas de análisis, entre ellas modelos de clasificación como árboles de decisión, regresión logística, *Random forest* y *XGBoost*, siendo este último el que presentó la mayor precisión, con un Área bajo la curva ROC (*Receiver Operating Characteristic*) (AUC, por sus siglas en inglés) superior al 90 %. Entre

los factores más relevantes asociados al abandono destacaron la baja frecuencia de uso, el consumo limitado de tipos de contenido, el uso exclusivo de dispositivos móviles y una alta interacción con el centro de soporte, lo cual puede reflejar insatisfacción. Este trabajo demuestra cómo el uso de herramientas analíticas permite mejorar la retención de clientes y optimizar decisiones estratégicas en servicios digitales. Asimismo, evidencia la utilidad del análisis cuantitativo para anticipar eventos de deserción en servicios de suscripción, lo que refuerza la pertinencia de este tipo de estudios para investigaciones como la presente.

Por su parte, Hung et al. (2006) aplicaron técnicas de minería de datos para gestionar la deserción de clientes en el sector de telecomunicaciones móviles de Taiwán, un mercado altamente competitivo tras su desregulación. El estudio se centró en identificar patrones predictivos del *churn*, con el objetivo de desarrollar un modelo que permitiera detectar a los clientes con mayor riesgo de abandonar la empresa, como parte de una estrategia de gestión de relaciones. Los autores destacan que retener clientes es más rentable que adquirir nuevos, lo que convierte al *churn* en un factor crítico para la sostenibilidad y rentabilidad del sector.

En el estudio de Hung et al. (2006) se emplearon técnicas de minería de datos utilizando información real de una empresa taiwanesa. Se compararon modelos de árboles de decisión (C5.0 y CART) y redes neuronales, evaluando su desempeño con variables independientes como información demográfica, historial de facturación, cambios contractuales y patrones de uso del servicio. La validación se realizó mediante métricas de precisión de clasificación y análisis de importancia de variables, demostrando que ambos métodos pueden predecir con alta precisión la cancelación de clientes.

Otro estudio que recientemente utilizó técnicas de *Deep Learning* (DNN) para predecir el *churn* es el de Khattak et al. (2023), desarrollado en el sector de telecomunicaciones. Los autores propusieron un modelo híbrido basado en redes neuronales bidireccionales de memoria a largo y corto plazo y redes neuronales convolucionales (BiLSTM-CNN), con el fin de clasificar clientes en riesgo de abandono a partir de información histórica. Para ello, emplearon variables independientes de tipo demográfico, contractual y de facturación, tales como género, condición de adulto mayor, pareja, dependientes, antigüedad del cliente, tipo de contrato, método de pago, cargos mensuales, cargos totales y distintas características del

servicio contratado. Los resultados muestran que el modelo propuesto alcanzó una exactitud de 81%, superando a modelos con otras arquitecturas profundas, específicamente redes neuronales convolucionales (CNN), redes neuronales recurrentes (RNN) y redes neuronales de memoria a largo y corto plazo bidireccionales (BiLSTM), encontrando que el modelo propuesto obtuvo el mejor desempeño. Además, reportaron para el modelo propuesto una precisión de 66%, un *recall* de 64% y un F-score de 65%. En síntesis, estos hallazgos resaltan un gran potencial de los enfoques híbridos de *Deep Learning* para capturar patrones complejos en problemas de predicción de abandono de clientes.

De manera complementaria, un estudio muy relevante es el de Equihua et al. (2023), quienes propusieron un marco secuencial de supervivencia basado en redes neuronales recurrentes (*Recurrent Neural Networks*, RNN) para modelar el abandono de clientes en el sector minorista. En este enfoque, los parámetros del modelo de supervivencia se aprenden directamente a partir de las secuencias de compras de cada cliente, sin necesidad de recurrir a variables demográficas ni a atributos construidos manualmente, lo que reduce de forma importante el esfuerzo de *feature engineering.* Además, el modelo está diseñado para contextos no contractuales, donde la fecha exacta de abandono no es observable y la información disponible se resume en secuencias de transacciones censuradas.

El marco propuesto en el estudio de Equihua et al. (2023) permite obtener, para cada cliente, una distribución de supervivencia individual sobre el tiempo hasta su próxima compra (o hasta el *churn*), capturando así la evolución dinámica del riesgo de abandono a lo largo del tiempo. En términos empíricos, Equihua et al. (2023) comparan su método con dos enfoques clásicos de análisis de supervivencia utilizados en *churn*: el modelo de riesgos proporcionales de Cox y el estimador de Kaplan-Meier, empleando tanto datos reales de transacciones de un minorista como un conjunto de datos simulado que replica este comportamiento. Los resultados muestran un desempeño predictivo comparable al de estos métodos establecidos, pero con ventajas adicionales: el modelo profundo trabaja únicamente con las secuencias de compras, evita la construcción manual de atributos y, sobre todo, proporciona estimaciones personalizadas de la probabilidad de *churn* para cada cliente en el tiempo. En síntesis, el estudio concluye que un marco de supervivencia profundo y secuencial constituye una

alternativa viable a los modelos tradicionales, especialmente cuando se busca explotar información transaccional de alta dimensión sin depender de una ingeniería de variables intensiva.

En conjunto, los estudios de Khattak et al. (2023) y de Equihua et al. (2023) coinciden en que el *Deep Learning* ofrece ventajas relevantes frente a varios enfoques tradicionales. Ya sea mediante BiLSTM o CNN para clasificación de *churn* o mediante marcos secuenciales de supervivencia basados en RNN, estos enfoques permiten modelar patrones complejos, escalar a grandes volúmenes de datos y reducir el sesgo asociado al diseño manual de variables sensibles. Asimismo, resaltan su potencial para optimizar las estrategias de retención al anticipar el abandono de manera proactiva, personalizar la estimación del riesgo a nivel de cliente y facilitar intervenciones oportunas en entornos altamente competitivos.

Sin embargo, los modelos tradicionales de análisis de supervivencia (como el modelo de Cox y los modelos paramétricos) pueden ser preferibles por su alta interpretabilidad, ya que proporcionan coeficientes claros que cuantifican el impacto de cada variable en el riesgo de abandono, lo cual resulta especialmente crítico en sectores regulados o en contextos donde se exige transparencia en la toma de decisiones. Además, estos enfoques suelen funcionar adecuadamente con tamaños muestrales pequeños y son menos propensos al sobreajuste. Al requerir menos datos, ofrecen una mayor eficiencia y una comprensión más directa de los resultados, a diferencia de los modelos de *Deep Learning*, que generalmente necesitan grandes volúmenes de información para alcanzar un buen desempeño.

Esta revisión de antecedentes refuerza la pertinencia del análisis de supervivencia como metodología robusta para modelar eventos temporales, destacando su versatilidad en disciplinas diversas. Su potencial para identificar patrones de abandono en entornos digitales, aún no explorados en profundidad, justifica su adopción en futuros estudios, incluido el presente trabajo, cuyo objetivo es abordar esta laguna mediante un enfoque cuantitativo adaptado a las particularidades del sector salud.

No obstante, pese a la amplia difusión de estas técnicas en otros ámbitos, aún persiste un vacío importante en su aplicación al ámbito de la salud digital, específicamente en plataformas de contratación *per diem* para profesionales de enfermería. La falta de

investigaciones centradas en este contexto limita el desarrollo de estrategias basadas en evidencia para mitigar la deserción de usuarios, un factor crítico que afecta la sostenibilidad y la calidad de los servicios de salud; vacío que el presente estudio busca comenzar a subsanar.

1.3. OBJETIVOS

Objetivo general

- Analizar el comportamiento de "abandono" (*churn*) de los usuarios de una plataforma digital de contratación "*per diem"* de personal de enfermería en los Estados Unidos de América, durante el periodo entre diciembre de 2021 y octubre de 2024, mediante la aplicación de modelos de análisis de supervivencia no paramétricos, semi-paramétricos y paramétricos que permitan identificar los factores asociados al tiempo hasta el abandono y generar estimaciones y predicciones del *churn* a corto, mediano y largo plazo.

Objetivos específicos

1. Caracterizar descriptivamente la población usuaria de la plataforma y el patrón observado de tiempos de supervivencia y censura durante el periodo de estudio.
2. Ajustar distintos modelos de análisis de supervivencia, incluyendo un modelo no paramétrico (Kaplan-Meier), modelos semi-paramétricos (modelo de riesgos proporcionales de Cox, en su versión estándar y extendida), y varios modelos paramétricos (Exponencial, Weibull, Log-normal, Log-Logístico, Gamma, y Gompertz), para analizar y predecir la fuga de clientes (*churn*) en la plataforma digital de contratación *"per diem"* de personal de enfermería.
3. Estimar y comparar las funciones de supervivencia y de riesgo, tanto globales como estratificadas por región, tipo de licencia, nivel de ingresos y grupos de edad, utilizando pruebas estadísticas apropiadas y el análisis de los coeficientes de los

modelos, con el fin de identificar diferencias en el comportamiento de *churn* entre subgrupos de usuarios.

4. Generar predicciones de tiempos de supervivencia y probabilidades de permanencia a corto, mediano y largo plazo para perfiles de usuarios de interés de la plataforma digital.

1.4. JUSTIFICACIÓN

Como se evidenció en la sección de Antecedentes, el estudio del *churn* es de suma importancia para empresas, gobiernos y otras instituciones. Una metodología que ha demostrado ser eficaz para abordar este fenómeno es el análisis estadístico de supervivencia. En la siguiente sección de "Marco Teórico" se profundizará en el concepto de *churn* y en cómo los modelos de supervivencia ayudan a estudiarlo y comprenderlo y, por tanto, pueden contribuir al diseño de estrategias para mitigarlo, en caso de que así se desee.

Keaveney y Parthasarathy (2001) señalan que la "fuga de clientes" (*churn*) reduce significativamente los ingresos y la rentabilidad de las empresas, ya que implica el desperdicio de la inversión inicial realizada para atraer clientes (por ejemplo, costos de consultoría, mercadeo y publicidad) y, además, genera costos adicionales asociados con la captación continua de nuevos clientes. En la misma línea, Reichheld y Sasser (1990) sostienen que la fuga de clientes tiene un impacto más fuerte sobre los ingresos empresariales que muchos otros factores tradicionalmente asociados con las ventajas competitivas.

Como se menciona en el estudio realizado por Lai y Zeng (2014), la retención de clientes, que cabe ser entendida como el opuesto a la "fuga de clientes", puede convertirse en la mejor estrategia de marketing en el futuro. Por ello, el análisis de la fuga de clientes ha cobrado creciente relevancia, particularmente en industrias de ventas y servicios, como el sector financiero (Mavri & Ioannou, 2008) y en las telecomunicaciones (Datta et al., 2000; Khattak et al., 2023). Además, en las últimas décadas se ha evidenciado una relevancia particular en el contexto digital y de servicios por suscripción, donde la retención de clientes se traduce

directamente en sostenibilidad y crecimiento de las organizaciones (Lai & Zeng, 2014; Monika et al., 2021).

En este contexto, el análisis de supervivencia se ha consolidado como una herramienta estadística avanzada para estudiar este tipo de fenómenos, proporcionando modelos rigurosos para la comprensión y predicción del abandono de clientes (Masarifoglu & Buyuklu, 2019; Khattak et al., 2023). La aplicación de modelos no paramétricos como el de Kaplan-Meier; semi-paramétricos, como el modelo de riesgos proporcionales de Cox; y diversos modelos paramétricos ha permitido identificar patrones de abandono y, con ello, diseñar estrategias de retención más efectivas. Dichas estrategias pueden traducirse en importantes beneficios económicos para empresas basadas en suscripción u otras en las que la permanencia de los usuarios resulta clave.

En el sector de la salud, particularmente en el ámbito de la contratación *per diem* de personal de enfermería a través de servicios web, la retención de usuarios adquiere una relevancia adicional. La prestación de servicios de salud mediante plataformas digitales enfrenta el desafío de mantener una base de usuarios activa y comprometida, aspecto fundamental para asegurar una oferta adecuada de personal de enfermería y, por ende, una atención de calidad a los pacientes. La fuga de usuarios en este contexto no solo implica una pérdida económica directa para las plataformas que facilitan la contratación de personal de enfermería, sino que también puede repercutir negativamente en la calidad y continuidad de la atención sanitaria.

La aplicación del análisis de supervivencia en plataformas digitales de contratación *per diem* de personal de salud representa una contribución relevante desde la perspectiva académica, profesional, y social. Este estudio responde a un vacío en la literatura, al implementar técnicas estadísticas avanzadas para describir, modelar y predecir el comportamiento de fuga de clientes en plataformas digitales del sector salud.

Desde un enfoque práctico, los resultados de este análisis permiten a los administradores de estas plataformas tomar decisiones basadas en evidencia, optimizando la experiencia del usuario, reduciendo la tasa de abandono y mejorando la eficiencia operativa en la contratación de personal de enfermería bajo la modalidad *per diem*. Además, la aplicación

específica del análisis de supervivencia aporta herramientas cuantitativas innovadoras para abordar problemas reales en la gestión del talento humano en entornos digitales.

Por otra parte, desde el ámbito académico, esta investigación amplía el conocimiento sobre la aplicabilidad del análisis de supervivencia en contextos digitales poco explorados, proporcionando así un marco metodológico sólido que puede ser aprovechado en futuras investigaciones afines.

En resumen, este trabajo contribuye significativamente al fortalecimiento de estrategias efectivas para la retención de usuarios, apoyando la optimización del recurso humano en el sector salud y favoreciendo, en última instancia, la calidad y continuidad de la atención sanitaria a través de entornos digitales.

CAPÍTULO II. MARCO TEÓRICO

2.1. FUNDAMENTOS CONCEPTUALES DEL FENÓMENO EN ESTUDIO

La presente investigación se enmarca en el uso del análisis estadístico de supervivencia como herramienta metodológica para estudiar el comportamiento de usuarios en servicios digitales, específicamente en una plataforma digital (aplicación móvil y página web) dedicada a la contratación *per diem* de personal de enfermería en los Estados Unidos. Este enfoque analítico ha demostrado ser eficaz para comprender fenómenos relacionados con el tiempo hasta la ocurrencia de un evento, siendo ampliamente aplicado en sectores como los servicios financieros (Larivière & Van den Poel, 2004), la educación superior (Gallardo Allen et al., 2016), las bibliotecas digitales (Lai & Zeng, 2014), entre muchos otros.

En el contexto de esta investigación, el evento de interés es la "fuga de clientes", también conocida por su denominación en inglés como *customer churn* o simplemente *churn*. Berson et al. (2000) definen la fuga de clientes como la rotación de clientes en industrias de servicios, destacando el cambio de los usuarios desde un proveedor hacia otro. Por su parte, Boote (1998) resalta que este comportamiento refleja la decisión consciente de un cliente de dejar de adquirir un servicio o de interrumpir su vínculo con una empresa prestadora de servicios. Aunque este concepto surgió en la industria de las telecomunicaciones, su uso se ha extendido ampliamente a otros sectores donde la retención de clientes resulta fundamental.

De forma más general, la fuga de clientes se entiende como el fenómeno donde los usuarios de una empresa o servicio dejan de hacer negocios con dicha empresa, ya sea cancelando su suscripción, dejando de comprar productos o servicios, o simplemente dejando de interactuar con la empresa. Para este estudio, la fuga de clientes consiste en la interrupción del uso del servicio digital de contratación *per diem* de personal de enfermería por parte de los usuarios del mismo, y es el comportamiento que se desea estudiar en este trabajo.

Comprender el fenómeno de *churn* resulta crucial para las empresas, ya que representa no solo una posible pérdida de ingresos, sino también una oportunidad para optimizar la experiencia del usuario, identificar factores de abandono y desarrollar estrategias de fidelización más efectivas. En este sentido, el análisis de supervivencia ofrece un marco

riguroso para modelar y predecir el momento en que los usuarios dejarán de utilizar un servicio, proporcionando información valiosa para la toma de decisiones estratégicas.

Se ha descubierto que el comportamiento de fuga de clientes reduce las ganancias y los beneficios de la empresa (Keaveney & Parthasarathy, 2001), ya que resulta en el desperdicio de la inversión inicial en los clientes (por ejemplo, costo de consultoría o publicidad) y la necesidad adicional de costos para atraer nuevos clientes. Además, Reichheld y Sasser (1990) consideran que el comportamiento de fuga de clientes (*churn*) tiene un efecto más fuerte en los ingresos de las empresas que muchos otros factores que generalmente se asocian con una ventaja competitiva para la empresa.

En este estudio, el análisis se centra en plataformas digitales orientadas a la contratación de personal de enfermería bajo la modalidad *per diem*. Esta forma de contratación implica la asignación de turnos laborales por día o por tarea específica, sin generar un vínculo laboral a largo plazo. Los trabajadores *per diem* son remunerados por cada jornada trabajada (usualmente de 8 o 12 horas), sin acceso a beneficios tradicionales como seguros médicos, vacaciones o licencias pagadas. Esta modalidad representa una solución flexible tanto para empleadores (que pueden ajustar sus necesidades de personal según la demanda) como para trabajadores que buscan autonomía y mayor control sobre su disponibilidad.

Por su parte, las plataformas digitales, entendidas como sitios web o aplicaciones móviles, han revolucionado la forma en que se gestionan estas contrataciones temporales. Su estructura tecnológica permite conectar, de manera rápida y eficiente, a las instituciones de salud con profesionales disponibles para cubrir turnos específicos. Estas herramientas han cobrado gran relevancia en los últimos años, impulsadas por la digitalización de procesos laborales, el crecimiento del trabajo remoto y los cambios en el mercado ocasionados por eventos globales como la pandemia de COVID-19.

El auge de estas plataformas ha sido facilitado por el uso masivo de teléfonos inteligentes, los cuales permiten a los usuarios acceder a la oferta laboral de manera inmediata y desde casi cualquier lugar. Esta disponibilidad permanente mejora la experiencia tanto de los empleadores como de los trabajadores, contribuyendo al dinamismo del mercado laboral y a la creciente popularidad del modelo *per diem*.

En las últimas décadas, las plataformas digitales han experimentado un crecimiento exponencial, transformando significativamente diversos sectores económicos y sociales. Según la Organización para la Cooperación y el Desarrollo Económicos (OECD, 2020a), la digitalización ha alterado profundamente la manera en que se realiza la ciencia, la tecnología y la innovación, facilitando nuevas formas de colaboración y difusión de la información. Este fenómeno de la digitalización ha impulsado la creación de nuevas oportunidades de negocio en múltiples industrias, fomentando la innovación continua y convirtiéndose en motores clave de crecimiento económico y transformación social (OECD, 2020b; Adigital & Boston Consulting Group, 2020).

Además, las plataformas digitales han redefinido la organización del trabajo, permitiendo modalidades más flexibles y eficientes. Este cambio ha facilitado la inclusión laboral de diversos grupos, como mujeres, jóvenes y personas con discapacidad, al ofrecer nuevas formas de empleo y acceso a mercados laborales previamente inaccesibles (Organización Internacional del Trabajo, 2021). Asimismo, la digitalización ha permeado todos los aspectos de la sociedad, transformando la manera en que trabajamos, nos comunicamos y realizamos nuestras actividades diarias, como se evidencia en estudios recientes sobre la influencia de estas plataformas en entornos laborales contemporáneos (Araujo Delgado, 2024).

Para analizar los datos se usarán métodos estadísticos de análisis de supervivencia, también conocidos como análisis de tiempo de vida, los cuales son una rama de la estadística que se centra en el estudio de la duración de tiempo hasta que ocurra un evento particular (O'Quigley, 2021; Hosmer et al., 2008). Este tipo de análisis tiene en cuenta tanto el tiempo de observación como la posibilidad de datos censurados, donde el evento de interés no se ha observado durante el período de estudio. Es de gran utilidad, ya que permite estimar funciones de supervivencia, tasas de riesgo y compararlas entre diferentes grupos o tratamientos. Estos métodos son ampliamente descritos en los libros relacionados con este tema (O'Quigley, 2021; Hosmer et al., 2008; Kleinbaum & Klein, 2005). Además, estos métodos han sido utilizados múltiples veces para entender fenómenos similares al de "fuga de clientes" (Ortiz Robles, 2022; Sánchez Brenes, 2021; Masarifoglu & Hakan Buyuklu, 2019; Gallardo Allen et al., 2016; Lai & Zeng, 2014). Una explicación más amplia y específica de estos modelos estadísticos se realiza más adelante en las siguientes secciones.

2.2. FUNCIONAMIENTO DE LA PLATAFORMA DIGITAL

La plataforma digital en estudio está especializada en la contratación de personal de enfermería bajo la modalidad *per diem,* conectando a profesionales de la salud con centros médicos que requieren cubrir turnos de manera flexible y en tiempo real. Esta plataforma opera a nivel nacional en Estados Unidos y se presenta como una alternativa moderna a las agencias tradicionales de personal, facilitando la interacción directa entre clínicas, hospitales y profesionales de enfermería sin intermediarios ni contratos a largo plazo.

Para utilizar este servicio (la aplicación móvil o la página web) los profesionales de enfermería deben empezar por registrarse y crear su perfil en la plataforma. Durante este proceso, los usuarios pueden brindar información como nombre, fecha de nacimiento, dirección, entre otros datos (algunos obligatorios y otros no). Además, se recomienda que registren (carguen) en la plataforma sus credenciales verificadas, como licencias y certificaciones de enfermería. La plataforma presenta un riguroso sistema de verificación de la identidad de los enfermeros así como de sus licencias y credenciales.

Una vez verificados, los y las enfermeras pueden explorar y solicitar turnos (trabajos *per diem)* disponibles en su área en tiempo real y de forma gratuita, eligiendo aquellos que se ajusten a su disponibilidad y preferencias. Los profesionales tienen la libertad de seleccionar cuántos turnos desean tomar, sin compromisos a largo plazo, lo que les permite equilibrar su vida laboral y personal según sus necesidades. Sin embargo, una vez realizada la solicitud de trabajo, deben esperar a ser o no aceptados por el establecimiento donde desean trabajar.

En cuanto a la contraparte, las instalaciones médicas, estas también pueden registrarse gratuitamente en la plataforma, y crear un perfil con información sobre su establecimiento. Luego, pueden publicar los turnos que necesitan cubrir, especificando detalles como horarios, tarifas y requisitos del puesto (licencias y credenciales, entre otros detalles menores).

Al recibir solicitudes de profesionales interesados, las instalaciones pueden revisar sus perfiles y credenciales para seleccionar al candidato que consideran más adecuado. La

plataforma permite una comunicación directa entre las instalaciones y los profesionales, facilitando la coordinación y resolución de cualquier duda o requerimiento adicional.

La plataforma en estudio brinda un servicio al cliente sumamente eficiente y dedicado a resolver las necesidades tanto de las instalaciones como de los profesionales en enfermería, ya sean problemas técnicos, preguntas, u otras necesidades.

Una vez que los profesionales de enfermería completan sus turnos y presentan el informe correspondiente a través de la plataforma, esta procesa el pago por los servicios prestados, calculado según las horas efectivamente trabajadas.

Posteriormente, la plataforma factura a las instalaciones médicas el monto correspondiente al pago del profesional, al cual se añade una comisión por el uso del servicio. Este modelo de precios transparente permite a las instalaciones aprobar el costo total de cada turno antes de publicarlo, asegurando que no haya cargos ocultos y facilitando una gestión financiera eficiente.

Por lo tanto, este servicio a través de la plataforma digital, permite a las instalaciones cubrir turnos de manera rápida y eficiente, reduciendo el tiempo y los recursos dedicados a la contratación tradicional. Al eliminar intermediarios y presentar tarifas competitivas, se reducen los costos asociados a las agencias de personal. Sumado a lo anterior, esta plataforma presenta la ventaja de que, tanto los profesionales como las instalaciones tienen acceso a información detallada y actualizada sobre el trabajo, lo que facilita una toma de decisiones informada y personalizada. Asimismo, la plataforma se adapta relativamente fácil a las necesidades cambiantes del entorno de atención médica, ofreciendo soluciones flexibles para cubrir demandas variables de personal.

2.3. EXPLICACIÓN DE LAS VARIABLES

En esta sección se definen y describen las variables utilizadas en el presente estudio de supervivencia.

Las variables independientes consideradas en este análisis son: *Censura*, *Edad*, *Licencia*, *Región* e *Ingresos*. Estas serán descritas a continuación y analizadas en profundidad en

apartados posteriores. Por su parte, la variable dependiente del estudio es el Tiempo de Supervivencia, que también será explicada a continuación.

Censura

La censura, en los análisis de supervivencia, se refiere a la situación en la que no se tiene información completa sobre el tiempo hasta que ocurre un evento de interés (como la falla, o abandono de un servicio) para algunos individuos. Esto significa que, para esos individuos, el tiempo real del evento no puede ser completamente observado, ya sea por diversos motivos, entre los que se encuentra que el evento no haya ocurrido durante el período de observación (censura por la derecha).

En el contexto de un análisis de supervivencia, la censura es de gran ayuda, porque permite incluir datos de participantes que no completaron el evento durante el periodo de estudio, evitando sesgos y maximizando el uso de la información disponible.

En este estudio, la censura ocurre únicamente por la derecha y se presenta cuando, al finalizar la recopilación de la información (1 de octubre de 2024), los usuarios aún no han incurrido en la falla; es decir, continúan activos y no se conoce cuándo dejarán de serlo.

Este estudio tiene la particularidad, y dificultad, de que el servicio brindado por la plataforma digital no es de suscripción pagada, por lo que los usuarios no necesitan cancelar su suscripción cuando ya no la van a utilizar más. Sumado a lo anterior, no se puede suponer que un usuario haya abandonado el servicio solo porque no estuvo activo en la última semana analizada, ya que es común que los usuarios no estén activos durante una o más semanas, y luego vuelvan a utilizar el servicio.

En consecuencia, en este estudio, un usuario se considera como censurado cuando haya realizado una solicitud de trabajo entre la fecha de finalización de la recopilación de la información (1 de octubre de 2024) y 60 días antes de dicha fecha (1 de agosto de 2024), es decir, si el usuario tiene al menos una solicitud de trabajo después del día 1 de agosto de 2024 este se clasifica como censurado (no se conoce con exactitud su tiempo de vida utilizando la plataforma). Adicionalmente, se realizará un análisis de sensibilidad para

evaluar si este margen temporal de 60 días es adecuado, el cual se presenta en la sección 4.5. de los Resultados.

Tiempo de Supervivencia (variable dependiente)

En este estudio, la variable *Tiempo de Supervivencia* representa la duración del uso de la plataforma digital por parte de los usuarios, o lo que sería lo mismo, el tiempo hasta el abandono del servicio (*churn*). Para medir la actividad de los usuarios en la plataforma digital, se registran las solicitudes de trabajo que los mismos realizan en el servicio, de forma que, si el usuario se encuentra constantemente realizando solicitudes de trabajo, se tiene la garantía de que se mantiene activo.

Para este análisis, se consideraron exclusivamente aquellos usuarios que realizaron al menos una solicitud de empleo a través de la plataforma.

Esta variable se calcula inicialmente como el tiempo, en días, entre la fecha de creación de la cuenta del usuario en la plataforma y la fecha de su última solicitud de empleo registrada en el sistema. Posteriormente, para efectos prácticos del estudio, este tiempo se convierte a semanas dividiendo el valor original entre 7.

En el caso de los usuarios que fueron denominados como Censurados, su tiempo de vida se calcula como el tiempo transcurrido entre la fecha de creación de la cuenta del usuario en la plataforma y la fecha de finalización de la recopilación de la información (1 de octubre de 2024).

Edad

La variable de edad, una característica demográfica de los individuos evaluados, representa la cantidad de años de vida de cada persona. Esta se mide en años completos y se incluye como una covariable en el análisis para explorar su relación con el tiempo de supervivencia.

En este estudio, la edad se calculó como la diferencia en años entre la fecha de nacimiento del usuario (reportada en la plataforma) y el 1 de enero de 2024. Los valores resultantes se

redondearon al número entero más cercano, eliminando cualquier decimal, con el objetivo de garantizar uniformidad y claridad en la interpretación de los datos.

Un 1% de los usuarios de la plataforma (307 usuarios) no reportan su fecha de nacimiento. Para abordar esta limitación, se creó una segunda variable independiente de naturaleza dicotómica. Esta variable asigna un valor de cero a los usuarios que han registrado su fecha de nacimiento y un valor de uno a aquellos que no la han proporcionado. A esta variable se le denominó "Sin_Edad".

Licencias

Esta variable se refiere a la licencia de enfermería que poseen los usuarios de la plataforma, considerando que estas licencias son específicas del sistema sanitario de los Estados Unidos. En dicho país, el personal de enfermería puede obtener diferentes tipos de licencias que delimitan su nivel de responsabilidad y las funciones que pueden desempeñar en el ámbito de la atención sanitaria (U.S. Bureau of Labor Statistics, 2024).

La regulación de las licencias de enfermería en Estados Unidos es una tarea compartida entre entidades estatales y organizaciones nacionales (U.S. Bureau of Labor Statistics, 2024). Las Juntas Estatales de Enfermería (Boards of Nursing - BONs) en cada estado, distrito y territorio son responsables de establecer estándares para la práctica segura de la enfermería, evaluar solicitudes, emitir y renovar licencias, así como tomar acciones disciplinarias cuando sea necesario. Por su parte, el Consejo Nacional de Juntas Estatales de Enfermería (National Council of State Boards of Nursing - NCSBN) es una organización independiente que reúne a todas las juntas estatales y territoriales. El NCSBN facilita la colaboración entre estas juntas en asuntos de interés común y administra el Examen Nacional de Licenciamiento para Enfermeras (NCLEX).

El NCLEX es una prueba estandarizada que evalúa las competencias y conocimientos fundamentales de los graduados en enfermería. Es un requisito obligatorio para obtener la licencia de práctica profesional como Enfermera Registrada (RN) o Enfermera Práctica

Licenciada (LPN), asegurando que los profesionales cumplan con los estándares mínimos de seguridad y calidad en el cuidado de la salud.

A continuación, se presentan las principales licencias de enfermería otorgadas en Estados Unidos:

CNA (Certified Nursing Assistant): Proporcionan atención básica a los pacientes bajo la supervisión de enfermeras registradas (RN) o enfermeras prácticas con licencia (LPN). Sus tareas incluyen ayudar con actividades diarias como bañarse, vestirse y alimentarse, además de tomar signos vitales y asistir en la movilidad de los pacientes. Dentro de los requisitos para obtener esta licencia se encuentra completar un programa de formación aprobado por el estado, que generalmente dura entre 4 a 12 semanas, y aprobar un examen de competencia para obtener la certificación (U.S. Bureau of Labor Statistics, 2024).

GNA (Geriatric Nursing Assistant): Similar al CNA, pero especializado en el cuidado de pacientes geriátricos. Los GNAs se enfocan en atender las necesidades específicas de los adultos mayores en diversos entornos, como hogares de ancianos y centros de atención a largo plazo. Además de la certificación CNA, se requiere formación adicional enfocada en geriatría y, en algunos estados, una certificación específica como GNA (U.S. Bureau of Labor Statistics, 2024).

RN (Registered Nurse): Las y los enfermeros RN tienen una formación más avanzada y son responsables de evaluar las necesidades de los pacientes, desarrollar planes de cuidado, administrar medicamentos y coordinar con otros profesionales de la salud. Pueden especializarse en diversas áreas de la medicina. Dentro de los requisitos para obtener esta licencia se encuentra obtener un título de asociado en enfermería (ADN) o una licenciatura en ciencias de la enfermería (BSN) y aprobar el examen NCLEX-RN para obtener la licencia (U.S. Bureau of Labor Statistics, 2024).

LPN (Licensed Practical Nurse): Las y los enfermeros LPNs brindan atención médica básica, como administrar medicamentos, supervisar a CNAs y monitorear signos vitales, bajo la supervisión de una RN o un médico. En algunos estados, también se les conoce como LVN (Licensed Vocational Nurse). Dentro de los requisitos para obtener esta licencia se encuentra

completar un programa de formación en enfermería práctica, que suele durar alrededor de un año, y aprobar el examen NCLEX-PN para obtener la licencia (U.S. Bureau of Labor Statistics, 2024).

Otras Licencias: En este estudio, se creó la categoría denominada "Otras" licencias para agrupar todas aquellas licencias que no se incluyen en las categorías principales de esta variable. Algunas de las licencias incluidas en esta categoría son: QMAP (Qualified Medication Administration Person), MA-C (Medication Aide Certified), RT (Respiratory Therapist), CG (Caregiver), CMA (Certified Medical Assistant), y CRMA (Certified Residential Medication Aide).

Sin licencia: Sumado a las categorías anteriores, se creó esta categoría para clasificar a los usuarios que no poseen ninguna licencia registrada en la plataforma.

Regiones

Esta otra variable independiente corresponde a la clasificación geográfica de los usuarios de la plataforma según las principales regiones de los Estados Unidos. Su propósito en el estudio es analizar cómo la ubicación geográfica puede influir en los patrones de uso de la plataforma, el tiempo de supervivencia y las características asociadas al personal de enfermería.

En este estudio, se utilizaron siete regiones específicas que agrupan diversos estados con características socioeconómicas y demográficas similares. Estas regiones son las siguientes:

1. Utah & Arizona: Comprende a los usuarios residentes en los estados de Utah y Arizona, una región conocida por su crecimiento poblacional acelerado y su mezcla de áreas urbanas y rurales.
2. Rockies & Red River: Agrupa a los estados de las Montañas Rocosas y el área del Río Rojo, caracterizada por su diversidad geográfica y económica.

3. California & Nevada: Incluye a los estados de California y Nevada, una región destacada por su alta densidad poblacional, su dinamismo económico y su diversidad cultural.
4. Northeast: Representa los estados de la región noreste del país, conocida por su historia, alto nivel de urbanización y liderazgo en sectores como la educación y la salud.
5. Pacific Northwest: Engloba a los estados del noroeste del Pacífico, reconocidos por su crecimiento tecnológico, sostenibilidad ambiental y calidad de vida.
6. Southeast: Incluye a los estados del sureste, una región diversa en términos culturales y económicos, con una importante presencia rural.
7. Midwest: Representa a los estados del medio oeste, caracterizados por su economía agrícola, industria manufacturera y comunidades cohesionadas.

El análisis basado en esta variable permitirá identificar posibles diferencias en el comportamiento de los usuarios según su región de residencia.

En la sección A de los Anexos se presenta el Cuadro A.1. que contiene los estados que componen cada una de las regiones mencionadas.

Ingresos

La variable independiente *Ingresos* representa un aspecto clave del análisis, relacionado con el desempeño económico de los usuarios en la plataforma. Esta variable es dicotómica y clasifica a los usuarios en dos categorías según si han obtenido o no ingresos económicos utilizando la plataforma para sus actividades profesionales.

En concreto:

- **Valor 0**: Indica que el usuario no generó ingresos a través de la plataforma durante sus dos primeros meses luego de haber hecho su primera solicitud de trabajo.

- **Valor 1**: Señala que el usuario sí generó ingresos utilizando la plataforma en los dos primeros meses luego de su primera solicitud de trabajo, reflejando su participación activa y éxito en el uso de esta como medio para obtener oportunidades laborales remuneradas.

Esta variable permite analizar las diferencias en el comportamiento y las características de los usuarios según su nivel de participación económica en la plataforma. Al incluirla en el modelo, se busca evaluar su asociación con el tiempo de supervivencia y explorar cómo el hecho de generar ingresos influye en las demás variables estudiadas, como la edad, la región y la licencia de enfermería.

2.4. MODELOS ESTADÍSTICOS DE SUPERVIVENCIA

El análisis estadístico de supervivencia es una rama de la estadística que se centra en el estudio del tiempo que transcurre hasta que ocurre un evento específico de interés. Este evento puede ser, por ejemplo, la muerte de un paciente, la deserción de un cliente, el fallo de una máquina o el abandono de un servicio. A diferencia de otros tipos de análisis, el de supervivencia tiene en cuenta no solo si el evento ocurrió, sino también cuándo ocurrió, lo que lo convierte en una herramienta especialmente útil para estudiar procesos dinámicos y temporales. Además, permite manejar adecuadamente la presencia de datos censurados, es decir, aquellos casos en los que no se ha observado aún el evento, pero de los que se conoce el tiempo de observación hasta cierto punto.

Uno de los componentes centrales del análisis de supervivencia es la función de supervivencia, que estima la probabilidad de que un individuo "sobreviva" (es decir, no experimente el evento) más allá de un determinado tiempo (Hosmer et al., 2008). También se utiliza la función de riesgo o *hazard*, que mide el riesgo instantáneo de que ocurra el evento en un momento específico, dado que el individuo ha sobrevivido hasta ese instante (Hosmer et al., 2008). Para estimar estas funciones, se pueden emplear diversos métodos estadísticos como el estimador de Kaplan-Meier, el análisis de regresión de Cox (modelo de riesgos proporcionales) y diversos modelos paramétricos, los cuales permiten explorar la influencia de distintas variables sobre el tiempo hasta el evento.

Existen tres enfoques principales en los modelos estadísticos de supervivencia: los modelos no paramétricos, los semi-paramétricos y los paramétricos. Cada uno de estos enfoques será descrito en detalle en secciones posteriores. Aunque en los últimos años se han desarrollado modelos más complejos basados en técnicas de aprendizaje automático y redes neuronales profundas (*Machine Learning* y *Deep Learning*), estos no son abordados en la presente investigación, ya que el enfoque se mantiene en los métodos clásicos ampliamente validados en la literatura estadística.

En contextos contemporáneos, como los servicios digitales y las plataformas tecnológicas, el análisis de supervivencia se ha adoptado con notable eficacia para estudiar fenómenos como la retención de usuarios, la fuga de clientes o el abandono del servicio (*churn*). Al permitir modelar el comportamiento de los usuarios a lo largo del tiempo y anticipar cuándo podrían abandonar una plataforma, esta técnica se convierte en una herramienta valiosa para la toma de decisiones estratégicas. Su uso permite implementar acciones preventivas, mejorar la personalización de la experiencia del cliente y aumentar la fidelización, contribuyendo así al crecimiento y sostenibilidad de las organizaciones (Hung et al., 2006; Larivière & Van den Poel, 2004).

2.4.1. Modelos no paramétricos

En el marco de la inferencia estadística y el modelado matemático, un modelo no paramétrico se define fundamentalmente por su ausencia de asunciones sobre una forma funcional específica predefinida para la distribución de probabilidad subyacente a los datos, o para la relación funcional entre variables. A diferencia de los modelos paramétricos, que postulan que los datos provienen de una familia de distribuciones indexada por un vector de parámetros de dimensión finita y fija (por ejemplo, la distribución Normal, $N(\mu, \sigma)$, definida por su media μ y su desviación estándar σ), los modelos no paramétricos operan bajo supuestos considerablemente más débiles.

La característica distintiva reside en que el espacio de parámetros efectivo en un modelo no paramétrico es, conceptualmente, de dimensión infinita. En lugar de estimar un número fijo

de parámetros escalares, el objetivo suele ser estimar directamente una función completa (como una función de densidad de probabilidad, una función de regresión o una función de supervivencia) o características específicas de la distribución (como cuantiles o la mediana) sin restringirla a una forma paramétrica particular.

Esta flexibilidad, entendida como la ausencia de una forma funcional rígida impuesta a priori, permite a los modelos no paramétricos adaptarse a una amplia gama de estructuras de datos, incluyendo relaciones complejas, no lineales o distribuciones con formas inusuales, que serían mal aproximadas por modelos paramétricos estándar. La estructura del modelo no se impone *a priori*, sino que emerge o se determina en gran medida a partir de los propios datos observados.

El estimador no paramétrico por excelencia para la función de supervivencia es el estimador de Kaplan-Meier, también conocido como estimador producto-límite. Este estimador proporciona una estimación empírica, escalonada y decreciente de la función de supervivencia $S(t)$ basada únicamente en los tiempos de evento observados y la información de censura, sin asumir ninguna forma distributiva para el tiempo T.

De manera similar, el estimador de Nelson-Aalen proporciona una estimación no paramétrica de la función de riesgo acumulado $H(t)$. Se construye a partir de los mismos tiempos de evento ordenados $t_{(1)} < t_{(2)} < \cdots$, registrando en cada uno de ellos el número de eventos observados d_j y el número de individuos en riesgo n_j justo antes de ese tiempo. En cada instante con eventos, el incremento del riesgo acumulado se aproxima por la fracción d_j/n_j; luego, la estimación de $H(t)$ se obtiene sumando todos estos incrementos para los tiempos de evento menores o iguales que t. El resultado es una función escalonada, creciente, que aproxima la integral en el tiempo del riesgo instantáneo y que, bajo supuestos regulares, es consistente para la verdadera función de riesgo acumulado. Además, mediante la relación aproximada $S(t) \approx \exp\{-H(t)\}$, el estimador de Nelson-Aalen también puede utilizarse para obtener una alternativa no paramétrica a la función de supervivencia.

Además del estimador de Nelson-Aalen, que proporciona una estimación escalonada de la función de riesgo acumulado $H(t)$, existen otros métodos no paramétricos que permiten

estimar directamente la función de riesgo instantáneo $h(t)$. En este trabajo, se utilizó el paquete "*muhaz*" (Hess & Gentleman, 2021) que implementa el método de suavizado propuesto por Müller y Wang (1994), el cual genera una curva continua y estable de riesgo, ajustada a los datos observados mediante técnicas *kernel*. Esta aproximación resulta más informativa para el análisis gráfico de la evolución del riesgo en el tiempo.

La ventaja principal de estos métodos no paramétricos en supervivencia radica en su robustez frente a la especificación incorrecta del modelo distributivo. Permiten una exploración inicial de los datos y la obtención de estimaciones de $S(t)$ o $H(t)$ que no están sesgadas por asunciones paramétricas potencialmente erróneas. Sin embargo, esta flexibilidad puede venir a costa de una menor eficiencia (mayor varianza en las estimaciones) en comparación con un modelo paramétrico correctamente especificado, especialmente con tamaños muestrales pequeños. Además, la extrapolación más allá del rango de los datos observados es intrínsecamente más problemática y menos directa que en los modelos paramétricos.

Suponiendo que se tiene un conjunto de *n* observaciones, donde cada una representa el tiempo hasta el evento o el tiempo censurado para un individuo.

Sea:

- $t_1 < t_2 < \cdots < t_k$: los tiempos distintos en los que ocurren eventos.
- d_j: el número de eventos que ocurren exactamente en el tiempo t_j.
- n_j: el número de individuos en riesgo justo antes del tiempo t_j.

La estimación de la función de supervivencia de Kaplan-Meier se define como se muestra en la Ecuación E.01, y de manera equivalente puede expresarse en la forma recursiva mostrada en la Ecuación E.02.

$$\hat{S}(t) = \prod_{j:\, t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \qquad \text{(Ecuación E.01)}$$

$$\hat{S}(t_j) = \hat{S}(t_{j-1})\left(1 - \frac{d_j}{n_j}\right), \qquad j = 1, \ldots, k, \quad \hat{S}(t_0) = 1 \qquad \text{(Ecuación E.02)}$$

Esta fórmula indica que la probabilidad estimada de sobrevivir más allá del tiempo t se obtiene como el producto acumulado de las probabilidades condicionales de sobrevivir en cada tiempo de evento t_j, dado que se ha sobrevivido hasta justo antes de ese instante (Hosmer et al., 2008; O'Quigley, 2021).

Dentro de las grandes ventajas de estos modelos, es que manejan adecuadamente las censuras. Si un individuo es censurado (es decir, se pierde el seguimiento sin que haya ocurrido el evento), se cuenta en el número en riesgo n_j hasta justo antes del momento en que se censura, pero no en el número de eventos d_j.

Aunque el modelo de Kaplan-Meier no se basa explícitamente en un enfoque de máxima verosimilitud para su estimación, el estimador de Kaplan-Meier puede verse como el resultado de maximizar la función de verosimilitud en un modelo no paramétrico donde los tiempos de falla son discretos y con censura no informativa (Hosmer et al., 2008; O'Quigley, 2021).

La función de verosimilitud en ese caso, asumiendo que las fallas ocurren en tiempos distintos y bajo censura no informativa, puede expresarse como se muestra en la Ecuación E.03.

$$\mathrm{L} = \prod_{\mathrm{j}=1}^{\mathrm{k}} \left(\frac{\mathrm{d_j}}{\mathrm{n_j}}\right)^{\mathrm{d_j}} \left(1 - \frac{\mathrm{d_j}}{\mathrm{n_j}}\right)^{\mathrm{n_j - d_j}} \qquad \text{(Ecuación E.03)}$$

donde, d_j es el número de eventos (muertes) en el tiempo t_j, n_j es el número de individuos en riesgo justo antes de t_j, y k es el número de tiempos distintos de falla.

La varianza de Kaplan-Meier puede estimarse con el estimador de Greenwood (Hosmer et al., 2008; O'Quigley, 2021), presentado en la Ecuación E.04.

$$\widehat{\mathrm{Var}}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_j \le t} \frac{d_j}{n_j\left(n_j - d_j\right)} \qquad \text{(Ecuación E.04)}$$

Esto permite construir intervalos de confianza (IC), por ejemplo, usando el método log-log o log-menos-log (O'Quigley, 2021), como se muestra en la Ecuación E.05.

$$\mathrm{IC} \;=\; \log\left[-\log\left(\hat{S}(t)\right)\right] \pm z_{\alpha/2} \cdot \frac{1}{\log\left(\hat{S}(t)\right)} \cdot \sqrt{\mathrm{Var}\left[\hat{S}(t)\right]} \qquad \text{(Ecuación E.05)}$$

El modelo no paramétrico de Kaplan-Meier constituye una de las herramientas más versátiles y robustas en el análisis de datos de tiempo hasta evento (Kleinbaum & Klein, 2005). Su principal fortaleza radica en la capacidad de estimar la función de supervivencia sin necesidad de asumir una forma funcional específica para la distribución subyacente de los tiempos de fallo o abandono. Esta propiedad lo hace particularmente adecuado en contextos donde no se cuenta con una función paramétrica que se ajuste completamente a los datos. Además, el estimador de Kaplan-Meier ofrece una representación intuitiva y continua de la probabilidad de supervivencia a lo largo del tiempo, permitiendo comparaciones claras entre diferentes grupos de estudio. Esta construcción secuencial otorga al modelo gran precisión, incluso en situaciones con datos heterogéneos.

En términos computacionales, la estimación no paramétrica de la función de supervivencia mediante el estimador de Kaplan-Meier es altamente accesible mediante el lenguaje de programación R (R Core Team, 2023), específicamente a través de la función *survfit*() contenida en la biblioteca *survival* (Therneau, 2024). Esta herramienta automatiza el cálculo de las estimaciones de supervivencia y, junto con paquetes complementarios como *survminer* (Kassambara et al., 2024), facilita la generación de gráficos interpretables y la comparación estadística entre curvas. De este modo, R se posiciona como un entorno robusto, eficiente y reproducible para el análisis no paramétrico de supervivencia, adoptado en una amplia gama de disciplinas del conocimiento, tanto en la investigación teórica como en contextos aplicados.

2.4.2. Modelos semi-paramétricos

El modelo semi-paramétrico de riesgos proporcionales de Cox, propuesto originalmente por David R. Cox en 1972 (Cox, 1972), es una de las herramientas más utilizadas en el análisis de datos de supervivencia. Este modelo permite estudiar la relación entre el tiempo hasta la ocurrencia de un evento (por ejemplo, muerte, abandono o falla) y un conjunto de variables explicativas o covariables. Su principal fortaleza radica en que no requiere especificar la forma funcional de la distribución del tiempo de supervivencia, diferenciándose así de los modelos paramétricos tradicionales. El modelo de Cox incorpora explícitamente el efecto de las covariables sobre el riesgo (*hazard*) de ocurrencia del evento, y se considera semi-paramétrico porque combina una parte no paramétrica (la función de riesgo basal) con una parte paramétrica (el efecto exponencial de las covariables).

La función de riesgo $h(t)$ es una medida del potencial instantáneo por unidad de tiempo de que ocurra el evento dado que un sujeto ha sobrevivido hasta el tiempo (Hosmer et al., 2008). El principal supuesto del modelo es que las funciones de riesgo para los grupos comparados son proporcionales a lo largo del tiempo, siendo el riesgo base $h_0(t)$ una función no especificada. Matemáticamente, este modelo se expresa como se muestra en la Ecuación E.06.

$$\mathrm{h}(t|X) = h_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right) \qquad \text{(Ecuación E.06)}$$

donde $h(t|X)$ es la función de riesgo condicional en el tiempo t, $h_0(t)$ es la función de riesgo base (no especificada), y $\beta_1, \ldots, \beta_p$ son los coeficientes asociados a las covariables $X_1, \ldots, X_p$. La estructura exponencial de este modelo permite interpretar los coeficientes como logaritmos de razones de riesgo, lo cual facilita una interpretación directa del impacto de cada covariable sobre el riesgo instantáneo del evento, manteniendo constantes las demás variables. Debido a su flexibilidad, interpretabilidad y solidez estadística, este modelo ha sido ampliamente adoptado en contextos que requieren modelar la duración hasta un evento bajo la influencia de múltiples factores.

En cuanto a la razón de riesgo (*Hazard Ratio, HR por sus siglas en inglés*), esta se define como el riesgo instantáneo de un individuo dividido por el riesgo instantáneo de otro individuo de referencia. Los individuos comparados pueden diferenciarse por los valores de un conjunto específico de predictores o covariables (Kleinbaum & Klein, 2005). El supuesto de riesgos proporcionales implica que la razón de riesgos (HR) es constante a lo largo del tiempo o, en términos equivalentes, que el riesgo para un individuo es proporcional al riesgo para cualquier otro individuo, siendo la constante de proporcionalidad independiente del tiempo. Matemáticamente, la HR para dos individuos se expresa como se muestra en la Ecuación E.07.

$$\widehat{HR} = \frac{\hat{h}(t, X^*)}{\hat{h}(t, X)} = \frac{\hat{h_0}(t) exp\left[\sum_{i=1}^{p} \beta_i X_i^*\right]}{\hat{h_0}(t) exp\left[\sum_{i=1}^{p} \beta_i X_i\right]} = exp\left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right] \qquad \text{(Ecuación E.07)}$$

donde X es el vector de valores de referencia para las covariables, y X^* es el vector con los valores específicos para otro individuo o grupo con el que se desea realizar la comparación.

Como se ha mencionado, una razón clave para la popularidad del modelo de Cox es que, aunque el riesgo basal no está especificado, se pueden obtener estimaciones razonablemente buenas de los coeficientes de regresión, las razones de riesgo de interés y las curvas de supervivencia ajustadas para una amplia variedad de situaciones (Kleinbaum & Klein, 2005). Otra forma de decir esto es que el modelo de riesgos proporcionales de Cox es un modelo "robusto", de modo que los resultados obtenidos con él se aproximarán estrechamente a los resultados obtenidos por modelos paramétricos correctamente especificados.

La estimación de los coeficientes de las covariables, β, del modelo de riesgos proporcionales de Cox se realiza mediante la verosimilitud parcial. Esta verosimilitud se basa en el orden observado de los eventos en lugar de la distribución conjunta de los tiempos en que ocurren dichos eventos. Por esta razón, se le denomina una "verosimilitud parcial" (Kleinbaum & Klein, 2005).

Una propiedad clave de esta verosimilitud parcial es que la función de riesgo basal (es decir, la parte del modelo que representa el riesgo común a todos los individuos en el tiempo) se

cancela en cada uno de los términos de la verosimilitud. Esto implica que no es necesario especificar la forma funcional de la función de riesgo basal al ajustar un modelo de Cox, ya que esta no influye en la estimación de los parámetros de regresión (Kleinbaum & Klein, 2005).

Sea $\mathrm{t_1} < \mathrm{t_2} < \cdots < \mathrm{t_D}$ los tiempos de falla observados (eventos no censurados) en una muestra de n individuos, y sea $R(t_j)$ el conjunto de riesgo en el tiempo t_j, es decir, el conjunto de individuos que están en riesgo justo antes del tiempo t_j.

La verosimilitud parcial para el modelo de Cox se muestra en la siguiente Ecuación E.08.

$$L(\beta) = \prod_{\mathrm{j=1}}^{\mathrm{D}} \frac{\exp\left(\beta^T x_{(j)}\right)}{\sum_{\mathrm{k \in R(t_j)}} \exp(\beta^T x_k)} \qquad \text{(Ecuación E.08)}$$

La forma logarítmica (log-verosimilitud parcial, l), que se puede maximizar numéricamente, se muestra en la siguiente Ecuación E.09.

$$l(\beta) = \sum_{\mathrm{j=1}}^{\mathrm{D}} \left[\beta^T x_{(j)} - \log\left(\sum_{k \in R(t_j)} \exp(\beta^T x_k) \right) \right] \qquad \text{(Ecuación E.09)}$$

donde, $x_{(j)}$ es el vector de covariables del individuo que falla en el tiempo t_j. Mientras que, x_k es el vector de covariables del individuo l en el conjunto de riesgo.

Este método resulta especialmente ingenioso, ya que evita estimar directamente la función de riesgo basal.

No obstante, este modelo semiparamétrico de riesgos proporcionales de Cox descansa sobre ciertos supuestos críticos que es importante verificar para asegurar la validez y robustez de las inferencias realizadas a partir de él. Como ya se ha mencionado, el supuesto fundamental de estos modelos, del cual deriva precisamente su nombre, es el de proporcionalidad de los riesgos. Este supuesto implica que la razón de riesgos (HR) entre dos individuos o grupos, definida por las covariables incluidas en el modelo, es constante a lo largo del tiempo (Kleinbaum & Klein, 2005; Hosmer et al., 2008; O'Quigley, 2021). En términos prácticos,

esto significa que el impacto relativo de cada covariable sobre el riesgo instantáneo del evento es el mismo independientemente del momento temporal considerado. Cuando este supuesto se cumple, los riesgos relativos pueden interpretarse fácilmente y tienen una única estimación global que es independiente del tiempo, facilitando la interpretación de los resultados.

Sin embargo, el supuesto de proporcionalidad puede violarse en diversas situaciones prácticas. Una violación ocurre cuando el efecto de una covariable cambia a lo largo del tiempo; por ejemplo, cuando una variable que inicialmente tiene una influencia fuerte sobre la supervivencia pierde importancia a medida que transcurre el tiempo o viceversa. En estos casos, el HR no permanece constante, sino que varía, comprometiendo la validez de los resultados derivados del modelo estándar de Cox.

Para evaluar este supuesto de proporcionalidad de los riesgos, se pueden emplear diferentes estrategias, por ejemplo, se pueden utilizar pruebas gráficas y estadísticas. Gráficamente, se puede analizar la proporcionalidad mediante gráficos del logaritmo negativo acumulado de la supervivencia, $\log(-\log[S(t)])$, contra el tiempo (Kleinbaum & Klein, 2005); si las curvas son aproximadamente paralelas, se asume proporcionalidad. Además, es posible graficar los residuos de Schoenfeld frente al tiempo y evaluar visualmente si el supuesto se cumple. Si los residuos de Schoenfeld no presentan ningún patrón definido en función del tiempo, esto indica que el supuesto de riesgos proporcionales probablemente se cumple. En cambio, si estos residuos presentan una tendencia clara o patrón evidente en función del tiempo, se sugiere una violación del supuesto (Grambsch & Therneau, 1994).

Estadísticamente, se puede utilizar la prueba de bondad de ajuste utilizando residuos escalados de Schoenfeld, desarrollada por Grambsch y Therneau (1994), para probar y diagnosticar formalmente el supuesto de riesgos proporcionales en el modelo de Cox utilizando residuos ponderados. Estos métodos permiten detectar desviaciones respecto a la proporcionalidad visualizando gráficamente cómo cambia el efecto de las covariables en el tiempo mediante suavizaciones gráficas, facilitando pruebas estadísticas que contrastan la hipótesis de proporcionalidad mediante regresiones ponderadas de los residuos escalados de Schoenfeld contra funciones del tiempo predefinidas.

Además del supuesto clave mencionado, el modelo de Cox también asume que las covariables tienen una relación lineal y aditiva con el logaritmo del riesgo (Kleinbaum & Klein, 2005; Hosmer et al., 2008). Esto implica que las covariables entran en la ecuación del modelo de forma lineal (aunque la transformación exponencial introduce no linealidad en la relación con el riesgo mismo), y que los efectos combinados de múltiples covariables sobre el logaritmo del riesgo son aditivos. La violación de esta linealidad puede llevar a sesgos importantes en las estimaciones obtenidas. Para evaluar la validez de este supuesto se utilizan frecuentemente métodos gráficos y técnicas estadísticas avanzadas, tales como gráficos de residuos Martingala frente a los valores predichos de las covariables continuas, que permiten explorar visualmente posibles desviaciones de la linealidad.

Finalmente, el modelo de Cox también asume la existencia de no colinealidad relevante entre las covariables, un supuesto que es común a muchos modelos de regresión. La presencia de alta colinealidad o multicolinealidad, es decir, la existencia de fuertes correlaciones entre dos o más covariables, puede generar inestabilidad en la estimación de los coeficientes de regresión, provocando errores estándar exagerados y reduciendo significativamente la precisión estadística del modelo (Hosmer et al., 2008; O'Quigley, 2021). Para diagnosticar problemas de colinealidad en el contexto del modelo de Cox, se pueden emplear técnicas estadísticas como el análisis del factor de inflación de la varianza (VIF, por sus siglas en inglés) (Belsley et al., 1980) o análisis de matrices de correlación entre covariables.

La generación, ajuste y análisis de modelos de riesgos proporcionales de Cox puede realizarse eficientemente utilizando el software estadístico R (R Core Team, 2023), ampliamente utilizado en investigación y análisis estadísticos avanzados. En particular, se cuenta con las librerías especializadas "*survival*" (Therneau, 2024) y "*survminer*" (Kassambara et al., 2024), que ofrecen diversas funciones robustas y flexibles para la estimación, diagnóstico, evaluación, y visualizaciones gráficas de estos modelos.

Residuos de los Modelos de Supervivencia:

En el modelo de riesgos proporcionales de Cox, los residuos de martingala (M_i) miden la discrepancia entre los eventos observados y los esperados según el modelo para cada individuo. Matemáticamente, para un sujeto i con tiempo de seguimiento T_i y covariables X_i, el residuo de martingala se define como se muestra en la Ecuación E.10.

$$M_i = \delta_i - \widehat{H_0}(T_i)\exp\left(\boldsymbol{X_i}^{\top}\hat{\beta}\right) \qquad \text{(Ecuación E.10)}$$

donde δ_i es el indicador de evento (1 si el individuo presentó el evento, 0 si su dato está censurado) y $\widehat{H_0}(T_i)$ es la función de riesgo (*hazard*) acumulada basal estimada hasta el tiempo T_i, $\hat{\beta}$ es el vector estimado de coeficientes. Por ello, en esencia, M_i es el número de eventos observado (δ_i) menos el número de eventos esperado según el modelo de Cox para ese individuo (Hosmer et al., 2008).

Un valor de M_i cercano a 1 indica que el individuo tuvo el evento antes de lo previsto por el modelo (o "vivió menos de lo esperado"), mientras que valores negativos indican que el evento ocurrió después de lo esperado por el modelo ("sobrevivió más de lo que el modelo esperaba").

A diferencia de los residuos utilizados en modelos lineales, los residuos de martingala presentan una distribución altamente asimétrica, ya que no se distribuyen alrededor de cero. Estos residuos están acotados en el intervalo $(-\infty, 1]$, siendo que los individuos que presentan el evento tienen residuos siempre menores o iguales a 1.

Debido a esta distribución sesgada, a menudo se transforman en *residuos de deviancia* para obtener una escala más simétrica alrededor de 0, facilitando la detección de valores atípicos. No obstante, los residuos de martingala siguen siendo muy útiles para evaluar la adecuación del modelo: en especial, para diagnosticar si la relación entre una covariable continua y el riesgo es correctamente especificada (lineal en el log-riesgo).

Por su parte, los residuos de deviancia (D_i) son una transformación de los residuos de martingala diseñada para mejorar su interpretación. Estos se definen como se muestra en la siguiente Ecuación E.11.

$$D_i = sign(M_i)\sqrt{-2[M_i + \delta_i \log(\delta_i - M_i)]} \qquad \text{(Ecuación E.11)}$$

donde, $sign(M_i)$ es la función signo aplicada al residuo de martingala.

Estos residuos son utilizados principalmente en el modelo de riesgos proporcionales de Cox con el objetivo de identificar observaciones atípicas o individuos con un posible mal ajuste del modelo.

Mientras que los residuos de martingala son útiles, presentan una distribución altamente asimétrica, lo que puede dificultar la interpretación directa. Los residuos de deviancia solucionan este problema mediante una transformación no lineal que los hace más simétricos alrededor de cero, favoreciendo así un análisis gráfico más claro y facilitando la identificación de valores atípicos o sujetos que contribuyen de forma inusual a la función de verosimilitud.

2.4.3. Modelo de Cox extendido para coeficientes que varían con el tiempo

Como se mencionó anteriormente, existe una gran cantidad de situaciones en las cuales el supuesto de riesgos proporcionales no se cumple para una o más de las covariables del modelo. Una de las formas de solucionar o contrarrestar este problema es utilizar el Modelo de Cox extendido (*Extended Cox Model*) (Kleinbaum & Klein, 2005; Therneau & Grambsch, 2000; Zhang et al., 2018), que permite covariables que varían con el tiempo. En este modelo, la función de riesgo $h(t)$ puede expresarse matemáticamente como se muestra en la siguiente Ecuación E.12.

$$\mathrm{h}\big(t, X(t)\big) = h_0(t)\exp\left(\sum_{i=1}^{p_1}\beta_i X_i + \sum_{j=1}^{p_2}\delta_j X_j(t)\right) \qquad \text{(Ecuación E.12)}$$

En esta ecuación, las variables X_i son covariables independientes del tiempo, y $X_j(t)$ representan covariables dependientes del tiempo, permitiendo una estructura más flexible que no requiere que se cumpla el supuesto de riesgos proporcionales para todas las variables durante todo el periodo de estudio (Kleinbaum & Klein, 2005; Therneau & Grambsch, 2000). Además, β_{i} son los coeficientes fijos asociados a cada covariable, y δ_{j} son los coeficientes que multiplican las interacciones temporales.

Por su parte, la fórmula general para la razón de riesgo en el modelo de Cox extendido, comparando dos individuos o grupos en un tiempo t, es la presentada en la Ecuación E.13

$$\widehat{LHR}(\mathrm{t}) = \frac{\hat{\mathrm{h}}\big(\mathrm{t}, \mathrm{X}^*(\mathrm{t})\big)}{\hat{\mathrm{h}}\big(\mathrm{t}, \mathrm{X}(\mathrm{t})\big)} = \exp\left[\sum_{\mathrm{i}=1}^{\mathrm{p}_1}\beta_{\mathrm{i}}(\mathrm{X}_{\mathrm{i}}^* - \mathrm{X}_{\mathrm{i}}) + \sum_{\mathrm{j}=1}^{\mathrm{p}_2}\delta_{\mathrm{j}}\left(\mathrm{X}_{\mathrm{j}}^*(\mathrm{t}) - \mathrm{X}_{\mathrm{j}}(\mathrm{t})\right)\right] \qquad \text{(Ecuación E.13)}$$

Aquí, $X^*(t)$ y $X(t)$ indican los valores de las covariables para dos individuos o grupos en el tiempo t.

La verosimilitud parcial del modelo de Cox extendido para la estimación de los coeficientes se presenta en la Ecuación E.14.

$$\mathrm{L} = \prod_{\mathrm{j}=1}^{\mathrm{D}} \frac{\exp\left(\beta^{\mathrm{T}}\mathrm{X}_{(\mathrm{j})}(\mathrm{t}_{\mathrm{j}})\right)}{\sum_{\mathrm{l}\in\mathrm{R}(\mathrm{t}_{\mathrm{j}})}\exp\left(\beta^T X_l(t_j)\right)} \qquad \text{(Ecuación E.14)}$$

donde $X_j(t_j)$ es el vector de covariables, incluyendo las dependientes del tiempo, del individuo que experimenta el evento en el tiempo t_j, y $R(t_j)$ representa el conjunto de individuos en riesgo en ese mismo tiempo t_j. La diferencia fundamental respecto al modelo Cox estándar radica en que, en este caso, las covariables pueden tomar diferentes valores para diferentes tiempos (crecer o decrecer con el tiempo), brindando mayor flexibilidad para

abordar situaciones donde no se cumple el supuesto de riesgos proporcionales. Ambas fórmulas consideran el orden observado de eventos, pero la versión extendida incorpora explícitamente variaciones temporales en las covariables (Kleinbaum & Klein, 2005).

Utilizando el software estadístico R (R Core Team, 2023) y la librería especializada "*survival"* (Therneau, 2024), es posible ajustar modelos de Cox que permiten coeficientes que varían con el tiempo (modelos de Cox extendidos). La función *tt()* dentro de *coxph()* facilita esta implementación al transformar una covariable en función del tiempo, permitiendo modelar efectos no proporcionales. Por ejemplo, se puede especificar que el efecto de la covariable cambie logarítmicamente con el tiempo. Esta herramienta es muy práctica, versátil y eficiente, ya que facilita y amplía notablemente la capacidad de análisis del modelo de Cox, permitiendo capturar dinámicas temporales complejas con una sintaxis sencilla y altamente personalizable.

2.4.4. Modelos paramétricos

En estadística, un modelo paramétrico es aquel que asume una forma funcional específica para la distribución de los datos, partiendo del supuesto de que los valores observados en la muestra o población provienen de una distribución de probabilidad conocida, caracterizada por un conjunto finito de parámetros (Kleinbaum & Klein, 2005; Hosmer et al., 2008; O'Quigley, 2021).

En el contexto del análisis de supervivencia, los modelos paramétricos suponen que los tiempos hasta la ocurrencia del evento de interés (como el *churn* en este estudio) siguen una distribución estadística conocida, dentro de las más utilizadas en este contexto se encuentra la distribución Exponencial, Weibull, Gamma, Gompertz, Log-normal o Log-logística, entre otras. Esta suposición permite describir completamente la función de supervivencia y la función de riesgo mediante fórmulas matemáticas explícitas que dependen de los parámetros del modelo. A diferencia de los modelos semi-paramétricos, como el de Cox, los modelos paramétricos permiten realizar extrapolaciones más allá del período observado y ofrecen estimaciones más precisas cuando la distribución asumida es apropiada para los datos (los

datos se ajustan adecuadamente a la distribución seleccionada). No obstante, su validez depende críticamente de que la elección de la distribución sea adecuada para el fenómeno en estudio, es decir, que el comportamiento de los datos refleje de forma coherente las características fundamentales de la distribución elegida. En caso contrario, incluso modelos bien especificados desde el punto de vista técnico pueden producir estimaciones sesgadas o poco representativas de la realidad.

Kleinbaum y Klein (2005) indican que, de conocerse la distribución paramétrica correcta para los datos analizados, lo más apropiado sería utilizar el modelo correspondiente del cual dichos datos provienen (sobre otro tipos de modelos como los semi-paramétricos o los no-paramétricos). Sin embargo, en muchos casos surge la dificultad de que, aunque existen varios métodos para evaluar la bondad de ajuste, es muy difícil tener plena certeza de que una distribución dada sea realmente la adecuada para el fenómeno en estudio.

El principal beneficio de los modelos paramétricos es que permiten una interpretación clara del efecto de las covariables sobre el tiempo hasta el evento a través de la función de densidad y de riesgo. Además, estos modelos pueden incorporar el efecto de las covariables de forma multiplicativa, lo que implica que el efecto combinado de las covariables sobre el tiempo hasta el evento es una multiplicación de sus efectos individuales. Esto contrasta con modelos no paramétricos como el modelo de Cox, donde el riesgo base no se especifica y sólo se modela el efecto proporcional de las covariables sobre el riesgo.

Estas distribuciones paramétricas, pueden ser representadas por medio de la función de densidad, $\phi(t)$, la cual se relaciona con la función de supervivencia, $S(t)$, y la función de riesgo, $h(t)$, como se muestra en las Ecuaciones E.15, E.16, E.17, y E.18.

$$S(t) = P(T > t) = \int_t^{\infty} \phi(u)\, du \qquad \text{(Ecuación E.15)}$$

$$h(t) = \frac{-d\,[S(t)]/dt}{S(t)} \qquad \text{(Ecuación E.16)}$$

$$S(t) = exp\left(-\int_0^t h(u)\,du\right) \quad \text{(Ecuación E.17)}$$

$$\phi(t) = h(t) \cdot S(t) \quad \text{(Ecuación E.18)}$$

A continuación, en el Cuadro 1, se presentan las ecuaciones de las funciones de supervivencia $S(t)$ y de riesgo $h(t)$ correspondientes a seis modelos paramétricos utilizados en esta investigación: el modelo Exponencial (Ecuaciones E.19 y E.20), el modelo Weibull (Ecuaciones E.21 y E.22), el modelo Log-logístico (Ecuaciones E.23 y E.24), el modelo Log-normal (Ecuaciones E.25 y E.26), el modelo Gamma (Ecuaciones E.27 y E.28) y el modelo Gompertz (Ecuaciones E.29 y E.30). Cabe destacar que existen muchas otras distribuciones que pueden explorarse según la naturaleza de los datos.

Cuadro 1.

Funciones de supervivencia S(t) y de Riesgo h(t) para los Modelos Paramétricos utilizados

Distribución	$S(t)$	$h(t)$
Exponencial	$e^{-\lambda t}$ (E. E. 19)	λ (E. E. 20)
Weibull	$e^{-\lambda t^p}$ (E. E. 21)	$\lambda p t^{p-1}$ (E. E. 22)
Log-logística	$\frac{1}{1+\lambda t^p}$ (E. E. 23)	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$ (E. E. 24)
Log-normal	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$ (E. E. 25)	$\frac{\frac{1}{t\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log t-\mu)^2}{2\sigma^2}\right)}{1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)}$ (E. E. 26)
Gamma	$1-\mathrm{F}[t;\ T\sim\Gamma(\alpha)]$ (E. E. 27)	$\frac{\lambda\,(\lambda t)^{\alpha-1}\,e^{-\lambda t}}{\Gamma(\alpha)\{1-F[t;T\sim\Gamma(\alpha)]\}}$ (E. E. 28)
Gompertz	$\exp\left(\frac{h_0}{a}(1-e^{at})\right)$ (E. E. 29)	$h_0 e^{at}$ (E. E. 30)

E. : Ecuación.

En las expresiones anteriores, cada parámetro tiene una interpretación particular que permite adaptar la forma de la función de supervivencia y de riesgo a distintos contextos. En el modelo exponencial, el parámetro $\lambda > 0$ representa la tasa de riesgo constante en el tiempo, y esto se mantiene para el modelo Weibull. Mientras que $p > 0$ es un parámetro de forma que permite modelar riesgos crecientes ($p > 1$) o decrecientes ($p < 1$) a lo largo del tiempo t (Kleinbaum & Klein, 2005).

De forma similar, en la distribución Log-logística, los parámetros λ y p ajustan respectivamente la escala y la forma de la distribución. Es importante destacar que, en el ajuste de modelos de supervivencia paramétricos, es común que el parámetro de escala λ sea reparametrizado en función de variables predictoras (a través de regresores), $\lambda_i = \exp\left(x_i^{\top\beta}\right)$, mientras que el parámetro p, conocido también como parámetro de forma, suele mantenerse constante a lo largo del proceso de modelado (Kleinbaum & Klein, 2005).

En el caso del modelo Log-normal, los parámetros μ y $\sigma > 0$ corresponden a la media y desviación estándar del logaritmo del tiempo hasta el evento. Y, la función $\Phi(\cdot)$ denota la función de distribución acumulada de la normal estándar (Liu, 2012). En este modelo Log-normal, la media μ del logaritmo del tiempo se reparametriza usando regresores: $\mu_i = x_i^{\top\beta}$.

En el caso del modelo Gamma, $\alpha > 0$ es el parámetro de forma y $\lambda > 0$ representa el parámetro de escala; además, $\Gamma(\alpha)$ representa la función Gamma (Ecuación E.31), mientras que, $F[t;\, T \sim \Gamma(\alpha)]$ representa la función de distribución acumulada, también llamada, función Gamma incompleta (Ecuación E.32) (Liu, 2012).

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t}\, dt \qquad \text{(Ecuación E.31)}$$

$$F[t;\, T \sim \Gamma(\alpha)] = \int_0^{t} \frac{t^{\alpha-1} e^{-u}}{\Gamma(\alpha)}\, du, \quad t > 0 \qquad \text{(Ecuación E.32)}$$

Por último, en el modelo Gompertz (Ecuaciones E.29 y E.30), $a > 0$ controla la aceleración del riesgo en el tiempo, es decir, es la tasa de crecimiento de la función de riesgo a lo largo

del tiempo, y $h_0 > 0$ es el nivel inicial de la tasa de riesgo (*hazard*), es decir, el valor de la función de riesgo al tiempo cero (Liu, 2012). Estos parámetros permiten capturar distintos patrones de riesgo temporal, lo cual hace que estos modelos sean especialmente útiles en contextos donde el riesgo no es constante.

Modelos de riesgos proporcionales (PH) y Modelos acelerados de falla (AFT)

Una de las distinciones fundamentales dentro del análisis de supervivencia se encuentra en el tipo de modelo que se utiliza para describir la relación entre las covariables y el tiempo hasta la ocurrencia del evento de interés. En particular, destacan dos enfoques ampliamente reconocidos: los modelos de riesgos proporcionales (PH, por sus siglas en inglés) y los modelos acelerados de falla (AFT, por sus siglas en inglés). Aunque ambos permiten analizar cómo las covariables influyen en el proceso de supervivencia, lo hacen desde perspectivas conceptualmente distintas y con diferentes implicaciones interpretativas.

Los modelos de riesgos proporcionales (PH), entre los cuales el modelo de Cox, exponencial, y Weibull son los más representativos, se basan en la idea de que el efecto de las covariables actúa multiplicando la función de riesgo base a lo largo del tiempo. Es decir, se asume que el riesgo relativo entre dos individuos con diferentes valores de covariables permanece constante a lo largo del tiempo. Esta característica da origen al nombre "proporcionales". Matemáticamente, el modelo PH se expresa como se presenta en la siguiente Ecuación E.33.

$$h(t \mid X) = h_0(t) \cdot \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right) \qquad \text{(Ecuación E.33)}$$

donde $h(t \mid X)$ es la función de riesgo para un individuo con vector de covariables X; $h_0(t)$ es la función de riesgo base (no especificada en el modelo de Cox, pero sí es conocida y específica en los modelos paramétricos de tipo PH), y los coeficientes β_j cuantifican el efecto de cada covariable X_j en términos de razones de riesgo. Este enfoque permite modelar de forma flexible y robusta situaciones donde el riesgo cambia en el tiempo, pero la relación relativa entre grupos se mantiene constante (Kleinbaum & Klein, 2005).

Por otro lado, los modelos AFT adoptan una visión distinta: modelan directamente el tiempo hasta el evento, bajo el supuesto de que las covariables afectan el tiempo de supervivencia mediante un factor de aceleración o desaceleración. En lugar de afectar el riesgo, las covariables modifican la velocidad con la que ocurre el evento. La forma funcional general de un modelo AFT es como se presenta en la Ecuación E.34.

$$\log(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \qquad \text{(Ecuación E.34)}$$

donde T es el tiempo hasta el evento, y ε es un término de error que sigue una distribución específica (normal, logística, Gamma, etc.). En este caso, los coeficientes β_j tienen una interpretación directa sobre el tiempo de vida esperado: un coeficiente negativo implica un acortamiento del tiempo hasta el evento (aceleración del fallo), mientras que uno positivo sugiere una extensión del tiempo de supervivencia (ralentización del fallo) (Kleinbaum & Klein, 2005; Liu, 2012).

Una diferencia importante entre ambos modelos es que los modelos PH se enfocan en comparar tasas de ocurrencia del evento (riesgo instantáneo), mientras que los AFT modelan directamente el tiempo como resultado. Además, las distribuciones utilizadas en cada caso no son equivalentes: por ejemplo, el modelo Weibull puede ser formulado tanto bajo un enfoque PH como AFT, mientras que los modelos Log-normal, log-logístico, y Gamma pertenecen exclusivamente al marco AFT, y los modelos Cox y Gompertz son únicamente modelos de tipo PH (Kleinbaum & Klein, 2005; Liu, 2012; Jackson, 2016). Esto se muestra de forma ilustrativa en el siguiente Cuadro 2.

Cuadro 2.

Clasificación de los modelos de supervivencia según los marcos PH y AFT

Modelo	¿PH?	¿AFT?	¿Ambos?
Exponencial	Sí	Sí	Sí
Weibull	Sí	Sí	Sí
Gamma	No	Sí	No
Log-normal	No	Sí	No
Log-logístico	No	Sí	No
Gompertz	Sí	No	No
Cox	Sí	No	No

Modelo Weibull

El modelo de supervivencia Weibull parte de la suposición fundamental de que los tiempos hasta el evento siguen una distribución Weibull, lo que implica que la función de riesgo (*hazard*) tiene una forma funcional determinada, la cual puede ser creciente, constante o decreciente en el tiempo, dependiendo del parámetro de forma p. Si $p > 1$, la tasa de riesgo aumenta con el tiempo; si $p < 1$, disminuye; y si $p = 1$, se reduce al modelo exponencial, con riesgo constante. Además de esta especificación funcional, se asume que las observaciones son independientes entre sí y que los datos están censurados no informativamente, es decir, que la probabilidad de censura no está relacionada con el tiempo real hasta el evento. En su versión paramétrica extendida, el modelo Weibull puede ser formulado tanto bajo el marco de riesgos proporcionales (PH) como bajo el marco de tiempo acelerado de falla (AFT), por lo que se espera que se cumplan los supuestos asociados a ambos enfoques, como la proporcionalidad de los riesgos o la multiplicatividad del tiempo en función de las covariables, dependiendo del tipo de formulación utilizada. Por tanto, la elección del modelo Weibull requiere validar tanto su adecuación funcional a los datos

observados como el cumplimiento de los supuestos estructurales asociados al enfoque paramétrico.

Los modelos Weibull presentan la siguiente función de log-verosimilitud que se muestra en la siguiente Ecuación E.35 (Kleinbaum & Klein, 2005; Liu, 2012). Mientras que sus funciones de supervivencia $S(t)$ y de riesgo $h(t)$ se presentaron anteriormente en las Ecuaciones E.21 y E.22.

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} \left[\delta_i(\log(\lambda) + \log(p) + (p-1)\log(t_i)) - \lambda t_i^p\right] \qquad \text{(Ecuación E.35)}$$

donde:

- $\lambda = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$ si se modela con covariables,
    - β_k es el coeficiente de cada covariable X_k,
- p es el parámetro de forma,
- δ_i indica si la observación está censurada (0) o no (1),
- t_i es el tiempo observado del sujeto i.

Modelo Exponencial

Por su parte, como es de esperar, el modelo de supervivencia Exponencial parte del supuesto fundamental de que los tiempos hasta el evento siguen una distribución exponencial, lo que implica que la función de riesgo $h(t)$ es constante en el tiempo. Es decir, la probabilidad instantánea de que ocurra el evento, dado que el individuo ha sobrevivido hasta el tiempo t, no varía con el paso del tiempo. Esta característica lo convierte en uno de los modelos más simples dentro del análisis de supervivencia, y en un caso particular del modelo Weibull y Gamma cuando el parámetro de forma p es igual a 1.

A pesar de su simplicidad, el modelo exponencial puede extenderse para incluir covariables, lo cual se logra incorporando una estructura log-lineal sobre el parámetro de tasa λ, permitiendo así modelar la relación entre variables explicativas y el riesgo constante.

El modelo exponencial puede formularse tanto bajo el enfoque de riesgos proporcionales (PH), dado que el cociente entre las funciones de riesgo de dos individuos es constante en el tiempo, como bajo el enfoque de aceleración de falla (AFT), ya que puede reinterpretarse como un caso particular de este marco. Por tanto, su elección como modelo implica verificar tanto su validez estructural (PH o AFT) como su adecuación a los datos observados.

Los modelos exponenciales presentan la siguiente función de log-verosimilitud, mostrada en la siguiente Ecuación E.36 (Kleinbaum & Klein, 2005; Liu, 2012)

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} [\delta_i \log(\lambda) - \lambda t_i] \qquad \text{(Ecuación E.36)}$$

donde:

- $\lambda = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$ si se incluyen covariables,
    - β_k es el coeficiente de cada covariable X_{k},
- t_i es el tiempo observado del sujeto i.
- δ_i indica si la observación está censurada (0) o no (1),
- n: número total de observaciones.

Las funciones de supervivencia $S(t)$ y de riesgo $h(t)$ para este modelo exponencial se presentaron en las Ecuaciones E.19 y E.20 de este documento.

Modelo Log-logístico

De forma similar, el modelo de supervivencia Log-logístico parte de la suposición de que el logaritmo del tiempo hasta el evento sigue una distribución logística, lo que implica que la

función de riesgo ($h(t)$) presenta un comportamiento no monótono caracterizado por una fase inicial creciente y, posteriormente, decreciente. Esta forma en "campana" de la tasa de riesgo lo convierte en una alternativa flexible para modelar situaciones en las que el riesgo de ocurrencia del evento aumenta al inicio, pero luego disminuye, comportamiento común en contextos como la reincidencia o el uso de plataformas digitales. A diferencia de otros modelos como el de Cox o el de Weibull, el modelo Log-logístico no cumple el supuesto de riesgos proporcionales (PH), ya que la razón de riesgos entre dos individuos no permanece constante en el tiempo. En cambio, se enmarca dentro del enfoque de modelos de tiempo acelerado de falla (AFT), lo que implica que las covariables afectan multiplicativamente al tiempo hasta el evento. En este modelo, como en otros modelos de supervivencia, se asume que las observaciones son independientes entre sí y que la censura es no informativa, es decir, que la probabilidad de que un individuo sea censurado no debe depender del tiempo de ocurrencia del evento.

El modelo Log-logístico presenta la siguiente función de log-verosimilitud (Kleinbaum & Klein, 2005; Liu, 2012), la cual es mostrada en la siguiente Ecuación E.37.

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} \left[\delta_i \log\big(f(t_i)\big) + (1-\delta_i) \log\big(S(t_i)\big)\right] \qquad \text{(Ecuación E.37)}$$

donde:

- $f(t)$ es la función de densidad de probabilidad del modelo log-logístico,
- $S(t)$ es la función de supervivencia del modelo log-logístico, presentada anteriormente en la Ecuación E.23.
- δ_i es el indicador de censura (1 si el evento ocurrió, 0 si está censurado),
- t_i es el tiempo observado para el individuo i.

Dicha función de densidad se presenta en la siguiente ecuación E.38.

$$f(t) = \frac{\lambda p(\lambda t)^{p-1}}{[1 + (\lambda t)^{p}]^{2}} \qquad \text{(Ecuación E.38)}$$

Como se había mencionado, λ es el parámetro de escala (está relacionado con la mediana del tiempo hasta el evento), mientras que p es el parámetro de forma. Determina el comportamiento de la función de riesgo:

- Si $p > 1$, la función de riesgo tiene forma de campana (incrementa y luego decrece).
- Si $p = 1$, la función de riesgo es constante.
- Si $p < 1$, la función de riesgo es decreciente con respecto al tiempo.

Modelo Log-normal

De manera coherente con su formulación, el modelo de supervivencia Log-normal asume como premisa central que el logaritmo del tiempo hasta el evento sigue una distribución normal. Esto conlleva que la distribución de los tiempos de supervivencia sea asimétrica y pueda modelar adecuadamente situaciones donde la tasa de riesgo primero aumenta y luego disminuye a lo largo del tiempo. En otras palabras, si T representa el tiempo hasta el evento, entonces $\log(T) \sim \mathcal{N}(\mu, \sigma^2)$. Esta característica lo diferencia de otros modelos paramétricos como el Weibull, que no permiten una función de riesgo ($h(t)$) no monótona.

En el caso Log-normal, la función de riesgo no es monótona, sino unimodal: parte de valores muy bajos al inicio, aumenta hasta alcanzar un máximo en un tiempo intermedio y posteriormente decrece de nuevo. En términos sustantivos, esto supone la existencia de un "periodo crítico" en el que la probabilidad instantánea de ocurrencia del evento es mayor; antes de ese punto el riesgo todavía está en fase de aumento y, después, tiende a reducirse conforme sobreviven los individuos más resistentes o más comprometidos. Por ello, el modelo Log-normal resulta especialmente útil para describir procesos donde el riesgo de abandono o falla se concentra en una etapa intermedia del seguimiento, en lugar de aumentar o disminuir de forma sostenida a lo largo del tiempo.

El modelo Log-normal se enmarca exclusivamente dentro del enfoque de tiempo acelerado de falla (AFT), por lo que asume que las covariables afectan de manera multiplicativa al tiempo hasta el evento, acelerando o desacelerando su ocurrencia en función de los valores de las variables explicativas. Como en otros modelos de supervivencia, se requiere que las observaciones sean independientes entre sí y que la censura sea no informativa, lo que significa que los mecanismos de censura no deben estar relacionados con la probabilidad del evento de interés. Asimismo, se asume una correcta especificación del modelo en cuanto a la relación funcional entre las covariables y el log-tiempo, y la ausencia de colinealidad severa entre predictores.

La función de log-verosimilitud para el modelo Log-normal (Liu, 2012, p. 129) se presenta en la siguiente Ecuación E.39.

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} \left[\delta_i \log\big(f(t_i)\big) + (1-\delta_i) \log\big(S(t_i)\big)\right] \qquad \text{(Ecuación E.39)}$$

donde, $f(t)$ es la función de densidad Log-normal que se presenta en la siguiente Ecuación E.40.

$$f(t) = \frac{1}{t\,\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(t)-\mu)^2}{2\sigma^2}\right) \qquad \text{(Ecuación E.40)}$$

Además, $S(t)$ es la función de supervivencia Log-normal, que se mostró anteriormente en la Ecuación E.25.

- μ: es la media del logaritmo del tiempo hasta el evento.
- σ: es la desviación estándar del logaritmo del tiempo hasta el evento, con $\sigma > 0$.
- t_i: tiempo observado de la observación i.
- δ_i: indicador de censura (1 si se observó el evento, 0 si está censurado),
- $\Phi(\cdot)$: la función de distribución acumulada de la normal estándar.

Modelo Gamma

Por su parte, el modelo de supervivencia Gamma asume, como su nombre lo indica, que el tiempo hasta el evento sigue una distribución Gamma, caracterizada por su flexibilidad para modelar diferentes formas de la función de densidad y de riesgo. Dependiendo de sus parámetros de forma y escala, este modelo permite representar funciones de riesgo crecientes, decrecientes o unimodales, por lo que puede adaptarse a una gran variedad de situaciones empíricas.

A diferencia de modelos como Weibull o Gompertz, la función de supervivencia y la función de riesgo del modelo Gamma no tienen expresiones cerradas tan sencillas, lo que dificulta una interpretación directa de algunas cantidades (por ejemplo, la forma exacta del riesgo instantáneo a lo largo del tiempo) y suele requerir el apoyo de software para su evaluación numérica.

En términos de marcos de regresión, en este trabajo el modelo Gamma se utiliza principalmente dentro del enfoque de tiempo acelerado de falla (AFT): las covariables actúan multiplicando el tiempo hasta el evento (acelerándolo o desacelerándolo), de modo que un cambio en las covariables se interpreta como un factor que acorta o prolonga la supervivencia esperada.

Como ocurre con los demás modelos de supervivencia considerados, se asume que las observaciones son independientes entre sí y que los mecanismos de censura son no informativos; es decir, la probabilidad de que un individuo sea censurado no depende de su tiempo real de ocurrencia del evento. Debido a su flexibilidad y a la ausencia de una estructura interpretativa tan simple como la de otros modelos, el modelo Gamma se utiliza principalmente como alternativa cuando modelos paramétricos más restrictivos (por ejemplo, el Weibull o el Exponencial) no ofrecen un ajuste adecuado a los datos. Cabe recordar que el modelo Exponencial puede entenderse como un caso particular del modelo Gamma cuando el parámetro de forma es igual a 1, de modo que la elección entre ambos responde al grado de flexibilidad que se requiera para representar la función de riesgo.

Para este modelo, se cuenta con la siguiente función de log-verosimilitud (Liu, 2012) que se presenta en la siguiente Ecuación E.41.

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} \delta_i[\alpha \log(\lambda) + (\alpha - 1)\log(t_i) - \lambda t_i - \log\Gamma(\alpha)] + (1 - \delta_i)\log\left[1 - \frac{\gamma(\alpha, \lambda t_i)}{\Gamma(\alpha)}\right] \qquad \text{(Ecuación E.41)}$$

donde:

- t_i: Tiempo observado hasta el evento (o censura) para el individuo i.
- δ_i: Indicador del evento:
    - $\delta_i = 1$: el evento fue observado (no censurado);
    - $\delta_i = 0$: el dato está censurado.
- $\alpha > 0$: Parámetro de forma de la distribución Gamma.
- $\lambda > 0$: Parámetro de tasa (es decir, el inverso del parámetro de escala).
- $\Gamma(\alpha)$: Función Gamma completa evaluada en α.
- $\gamma(\alpha, \lambda t)$: Función Gamma incompleta inferior.

Modelo Gompertz:

De acuerdo con Liu (2012), el modelo Gompertz parte de la premisa fundamental de que la tasa de riesgo (*hazard rate*) aumenta exponencialmente con el tiempo, lo que implica una función de riesgo siempre creciente y nunca cóncava. Esto significa que no solo la tasa de riesgo aumenta con el tiempo, sino que también la tasa de cambio del riesgo aumenta continuamente a lo largo del periodo estudiado. Por lo tanto, el supuesto principal es una tasa de riesgo que crece exponencialmente, lo cual se puede verificar mediante transformaciones gráficas lineales de la función de riesgo en escala logarítmica.

Por lo tanto, su supuesto principal, como se puede deducir de la Ecuación E.30, consisten en presentar una tasa de riesgo creciente exponencialmente: El riesgo relativo aumenta constantemente en el tiempo. O, dicho de otra forma, existe una relación log-lineal entre el riesgo y el tiempo.

El modelo Gompertz se formula bajo el marco de riesgos proporcionales (PH), en el que los efectos de las covariables pueden expresarse como razones de riesgo constantes a lo largo del tiempo. Esto se debe a que la función de riesgo presenta un crecimiento exponencial con el tiempo, manteniendo constante la razón de riesgos entre dos individuos que difieren solo en sus covariables. Como en otros modelos, se asume que las observaciones son independientes entre sí y que la censura es no informativa.

Este modelo Gompertz, presenta la función de Log-verosimilitud, para datos sin censura, presente en la siguiente Ecuación E.42 (Liu, 2012).

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} \left[\log(h_0) + rt_i - \frac{h_0}{r}(e^{rt_i} - 1) \right] \qquad \text{(Ecuación E.42)}$$

Mientras que, en el caso que se cuenta con observaciones censuradas, el modelo Gompertz presenta la siguiente función de Log-verosimilitud de la Ecuación E.43.

$$\log(\boldsymbol{L}) = \sum_{i=1}^{n} \delta_i [\log(h_0) + rt_i] - \frac{h_0}{r} \sum_{i=1}^{n} (e^{rt_i} - 1) \qquad \text{(Ecuación E.43)}$$

donde:

- n: Número de observaciones en la muestra.
- h_0: Parámetro de escala de la función de riesgo.
- r: Parámetro de crecimiento exponencial del riesgo. Si $r > 0$, el riesgo aumenta con el tiempo.
- t_i: Tiempo observado para el individuo i.
- δ_i: Indicador de censura para el individuo i; toma el valor 1 si el evento fue observado (no censura) y 0 si está censurado.
- rt_i: Término que introduce el crecimiento del riesgo en el tiempo.

- $e^{rt_i} - 1$: Parte de la integral del riesgo acumulado, utilizada en la construcción de la verosimilitud.

Herramientas para el cálculo de modelos paramétricos en análisis de supervivencia

La generación, ajuste y análisis de modelos de supervivencia de tipo paramétrico puede realizarse de forma eficiente mediante el uso del software estadístico R (R Core Team, 2023), ampliamente reconocido por su versatilidad en investigación y análisis estadístico avanzado.

En particular, R dispone de paquetes especializados como *survival* (Therneau, 2024) y *survminer* (Kassambara et al., 2024), los cuales proporcionan herramientas robustas y flexibles para la estimación de modelos, diagnóstico, evaluación y visualización gráfica de curvas de supervivencia.

No obstante, para el caso específico de modelos paramétricos, también destaca el paquete *flexsurv* (Jackson, 2016), que ofrece una mayor variedad de distribuciones y opciones de modelado. Este paquete se ha consolidado como una de las mejores alternativas disponibles para el ajuste, representación y análisis detallado de modelos paramétricos en el contexto del análisis de supervivencia.

2.5. EVALUACIÓN DE LOS MODELOS

Antes de proceder a la comparación o interpretación de los modelos de análisis de supervivencia, es indispensable verificar el cumplimiento de los supuestos estadísticos propios de cada tipo de modelo. Esta verificación garantiza la validez de las estimaciones obtenidas, así como la correcta interpretación de los efectos de las covariables sobre el tiempo hasta la ocurrencia del evento.

Complementario a la verificación de los supuestos, la evaluación del rendimiento de los modelos se puede abordar desde dos perspectivas complementarias: el ajuste del modelo a los datos observados y la capacidad predictiva del modelo.

Para evaluar qué tan bien se ajusta un modelo a los datos observados, es común utilizar criterios de información como el Criterio de Información de Akaike (AIC), el cual penaliza la complejidad del modelo e incentiva la parsimonia, siendo preferible el modelo con menor valor de AIC; y el Criterio de Información Bayesiano (BIC), el cual es similar al AIC, pero con una penalización mayor para modelos más complejos, especialmente cuando el tamaño muestral es grande. Ambos criterios permiten comparar distintos modelos estimados sobre los mismos datos, incluso cuando estos modelos no sean anidados entre sí (Burnham & Anderson, 2004; Bütepage et al., 2022).

El AIC fue introducido por Hirotugu Akaike en 1974 (Akaike, 1974) como una medida basada en la teoría de la información para comparar modelos estadísticos. Está definido por medio de la Ecuación E.44.

$$AIC = 2k - 2\,ln(\hat{L}) \qquad \text{(Ecuación E.44)}$$

donde, k es el número de parámetros estimados en el modelo, y $\hat{L}$ es el valor máximo de la función de verosimilitud del modelo.

El AIC estima la cantidad de información perdida al utilizar un modelo para representar el proceso generador de datos. Un menor valor de AIC indica un modelo que pierde menos información, es decir, un modelo que se ajusta mejor a los datos (Burnham & Anderson, 2004; Bütepage et al., 2022). Sin embargo, el AIC puede favorecer modelos más complejos, especialmente en muestras pequeñas, lo que puede llevar al sobreajuste.

El BIC, también conocido como Criterio de Schwarz, fue propuesto por Gideon Schwarz en 1978 (Schwarz, 1978) como una aproximación al factor de Bayes para la selección de modelos. Está definido por medio de la Ecuación E.45.

$$BIC = k * ln(n) - 2 * ln(\hat{L}) \qquad \text{(Ecuación E.45)}$$

donde, n es el tamaño de la muestra, y k y $\hat{L}$ se definieron anteriormente.

El BIC penaliza más fuertemente la complejidad del modelo que el AIC, especialmente en muestras grandes, debido al término $k * \ln(n)$. Esto lo hace más conservador, favoreciendo modelos más simples cuando hay poca evidencia para justificar la inclusión de parámetros adicionales (Burnham & Anderson, 2004; Bütepage et al., 2022).

En análisis de supervivencia, el BIC se utiliza de manera similar al AIC para comparar modelos. Su penalización más fuerte puede ser beneficiosa para evitar el sobreajuste, pero también puede llevar a seleccionar modelos demasiado simples si la penalización es excesiva.

En el caso del AIC, de acuerdo con Burnham y Anderson (2002), cuando la diferencia en el valor de dos modelos (ΔAIC) es menor a 2, se interpreta que los modelos son prácticamente indistinguibles según esta métrica, es decir, ninguno de los dos puede considerarse claramente superior. Cuando la diferencia se ubica entre 2 y 4 ($2 \leq \Delta AIC < 4$), existe una evidencia leve a favor del modelo con menor AIC. Si la diferencia se encuentra en un rango de 4 a 7 ($4 \leq \Delta AIC < 7$), se puede afirmar que hay evidencia considerable a favor del modelo con el AIC más bajo. Finalmente, cuando la diferencia es mayor o igual a 10 ($\Delta AIC \geq 10$), la evidencia es fuerte, de modo que, en este escenario, la recomendación es seleccionar únicamente el modelo con menor AIC.

De manera análoga, los criterios para interpretar el BIC siguen una lógica similar. Cuando la diferencia en el valor del BIC entre dos modelos es menor a 2 ($\Delta BIC < 2$), se considera que la evidencia es débil y que los modelos son esencialmente comparables. Si la diferencia se sitúa entre 2 y 6 ($2 \leq \Delta BIC < 6$), se interpreta como evidencia positiva a favor del modelo con menor BIC. En el rango de 6 a 10 ($6 \leq \Delta BIC < 10$), la evidencia se considera fuerte, mientras que diferencias iguales o superiores a 10 ($\Delta BIC \geq 10$) constituyen evidencia muy fuerte en favor del modelo con menor BIC (Raftery, 1996; Burnham & Anderson, 2002).

Por su parte, otra forma de evaluar la calidad de un modelo estadístico de supervivencia consiste en estimar qué tan bien predice los tiempos de supervivencia. Para esto, se pueden emplear varios indicadores. Entre los más utilizados se encuentra el Estadístico C de Harrell (Harrell et al., 1982), también conocido como índice de concordancia (*C-index*), el cual es una medida no paramétrica que evalúa la capacidad discriminativa de un modelo de supervivencia. Específicamente, estima la probabilidad de que, para un par aleatorio de

individuos, aquel con mayor riesgo predicho experimente el evento de interés antes que el otro individuo. Por lo tanto, esta métrica mide la concordancia entre los tiempos observados y los predichos por el modelo. Un valor cercano a 1 indica una excelente capacidad predictiva, mientras que un valor cercano a 0.5 sugiere una predicción aleatoria (Harrell et al., 1982; Harrell, 2015; Hosmer & Lemeshow, 2000; Newson, 2010; Uno et al., 2011).

El cálculo del Estadístico C de Harrell (*C-index*) implica los siguientes pasos:

1. Identificación de pares comparables: Se consideran todos los pares posibles de individuos en el conjunto de datos donde al menos uno de ellos ha experimentado el evento de interés.
2. Evaluación de la concordancia:
    - Par concordante: Si el individuo con mayor riesgo predicho por el modelo experimenta el evento antes que el otro.
    - Par discordante: Si el individuo con menor riesgo predicho experimenta el evento antes.
    - Empate: Si ambos individuos tienen el mismo riesgo predicho o el evento ocurre al mismo tiempo.
3. Cálculo del estadístico: El C-index se calcula como se muestra en la siguiente Ecuación E.46 (Harrell, 2015).

$$C = \frac{\textit{Número de pares concordantes} + 0.5 \times \textit{Número de empates}}{\textit{Número total de pares comparables}} \qquad \text{(Ecuación E.46)}$$

Esta fórmula representa la proporción de pares correctamente ordenados por el modelo en relación con todos los pares posibles comparables (al menos uno debe haber experimentado el evento).

Por otro lado, otro indicador ampliamente utilizado para evaluar la precisión predictiva de los modelos estadísticos de supervivencia es el *Brier Score* (Brier, 1950). Este indicador representa el error cuadrático medio entre las probabilidades predichas por el modelo y los

eventos observados. En otras palabras, mide cuán cercanas están las probabilidades pronosticadas del evento respecto a lo que realmente ocurrió. Por lo que un menor valor del *Brier Score* indica una mayor precisión del modelo, e implica que las predicciones se ajustan mejor a los datos observados (Gerds & Schumacher, 2006).

El *Brier Score* es una medida de precisión para predicciones probabilísticas. Fue introducido por Glenn W. Brier en 1950 (Brier, 1950) como una forma de cuantificar la exactitud de las predicciones de eventos binarios. Posteriormente, ha sido adaptado al contexto del análisis de supervivencia, para evaluar la precisión de las probabilidades de supervivencia predichas por un modelo en momentos específicos en el tiempo (Gerds & Schumacher, 2006; Kvamme & Borgan, 2023).

La fórmula general del *Brier Score* (BS) para datos sin censura es la presente en la Ecuación E.47 (Kvamme & Borgan, 2023).

$$BS(t) = (1/n)\sum_{i=1}^{n}\left(1_{T_i^*>t} - \pi_i(t)\right)^2 = (1/n)\sum_{i=1}^{n}\left[\pi_i(t)^2 \cdot 1_{T_i^*\leq t} + \left(1-\pi_i(t)\right)^2 \cdot 1_{T_i^*>t}\right] \quad \text{(Ecuación E.47)}$$

donde, n es el número total de individuos en el conjunto de datos, T_i^* es el tiempo real del evento para el individuo i (sin censura), $1_{T_i^*>t}$ representa una función indicadora: vale 1 si $T_i^* > t$, y 0 si no, $\pi_i(t)$ es la predicción de la probabilidad de supervivencia para el individuo i en el tiempo t. $BS(t)$ es el valor del *Brier Score* en el tiempo t. Este score toma valores entre 0 y 1, donde 0 indica una predicción perfecta.

Sin embargo, en análisis de supervivencia, los datos suelen estar sujetos a censura a la derecha, lo que significa que para algunas observaciones no se conoce el tiempo exacto del evento. Para manejar esta censura, se utiliza una versión ajustada del *Brier Score* que incorpora pesos de probabilidad inversa de censura (IPCW) (Kvamme & Borgan, 2023). Este BS con corrección por censura se muestra en la Ecuación E.48. Esta fórmula evalúa el error

cuadrático entre las probabilidades de supervivencia predichas y los eventos observados, considerando adecuadamente la presencia de datos censurados.

$$BS_{IPCW}(t) = \frac{1}{\tilde{n}(t)} \sum_{i=1}^{n} \left[\frac{\pi_i(t)^2 \cdot 1_{\{T_i \leq t,\, D_i=1\}}}{G_i(T_i^-)} + \frac{\left(1-\pi_i(t)\right)^2 \cdot 1_{\{T_i > t\}}}{G_i(t)} \right] \quad \text{(Ecuación E.48)}$$

$$\tilde{n}(t) = \sum_{i=1}^{n} \left[\frac{1_{\{T_i \leq t,\, D_i=1\}}}{G_i(T_i^-)} + \frac{1_{\{T_i > t\}}}{G_i(t)} \right] \quad \text{(Ecuación E.49)}$$

donde, $\pi_i(t)$ es la probabilidad predicha de supervivencia para el individuo *i* en el tiempo t, es decir, $\hat{\mathrm{S}}(t|X_i)$; T_i es el tiempo observado (mínimo entre el tiempo de evento T_i^* y de censura C_i^*). Además, D_i es el indicador de evento observado (1 si ocurrió el evento antes de censura; 0 si está censurado); $G_i(t)$ es la función de supervivencia del tiempo de censura para el individuo *i*. Mientras que, $G_i(T_i^-)$ es el límite por la izquierda de la función de supervivencia en T_i (para eventos exactos); y $\tilde{n}(t)$ es el factor de normalización, que corresponde a la suma de los pesos IPCW aplicados a cada observación en el tiempo t, y asegura que $BS(t) \in [0,1]$. Donde, la fórmula para obtener $\tilde{n}(t)$ se presenta en la Ecuación E.49.

El peso IPCW se calcula utilizando el estimador de Kaplan-Meier para la función de supervivencia de la censura. Esto significa que, para calcular los pesos necesarios para ajustar el *Brier Score* ante la presencia de datos censurados, es necesario conocer la probabilidad de que un individuo no haya sido censurado en un tiempo determinado t. Esa probabilidad se obtiene mediante el estimador de Kaplan-Meier, que se aplica no sobre el evento de interés, sino sobre la distribución de censura (es decir, se trata como si la censura fuera el evento principal a modelar).

Para interpretar este indicador, Gerds y Schumacher (2006) señalan que valores bajos del *Brier Score* reflejan una mayor precisión predictiva del modelo. Por el contrario, valores cercanos a 0.25 suelen interpretarse como una predicción poco informativa, especialmente en situaciones donde la proporción de eventos es aproximadamente del 50%.

El *Brier Score* es una métrica que evalúa simultáneamente dos aspectos importantes del desempeño predictivo de un modelo: la calibración (qué tan bien las probabilidades predichas coinciden con las frecuencias observadas) y la discriminación (capacidad del modelo para distinguir correctamente entre individuos que experimentan el evento y los que no).

Adicionalmente, es posible calcular el *Brier Score Integrado* (IBS), el cual corresponde al promedio del *Brier Score* a lo largo de un intervalo de tiempo. Esta medida ofrece una visión global de la precisión predictiva del modelo durante todo el período de seguimiento.

Formalmente, el IBS se calcula como el promedio temporal del error de predicción (PEC) estimado en una malla de tiempos, ponderando el área bajo la curva de errores. En este estudio, el IBS fue estimado mediante la función *pec()* del paquete *pec* de R (Gerds, 2025), aplicada tanto a los modelos semi-paramétricos como a los paramétricos (Exponencial, Weibull, Log-normal, Log-logístico, Gamma y Gompertz). La fórmula del IBS se presenta a continuación mediante la Ecuación E.50.

$$\mathrm{IBS} = \frac{1}{\mathrm{T_{max}} - \mathrm{T_{min}}} \sum_{\mathrm{i=1}}^{\mathrm{m-1}} \frac{\mathrm{B(t_i)} + \mathrm{B(t_{i+1})}}{2} \, (\mathrm{t_{i+1}} - \mathrm{t_i}). \qquad \text{(Ecuación E.50)}$$

donde, $B(t_i)$ corresponde al valor del *Brier Score* ajustado en el tiempo t_i.

En conjunto, el *Brier Score* y su versión integrada (IBS) constituyen herramientas valiosas para la evaluación de modelos de supervivencia, en especial cuando se busca una métrica que capture tanto la calibración como la discriminación. Su capacidad para ajustarse a la presencia de datos censurados lo convierte en un instrumento particularmente útil en contextos donde la censura es frecuente, como ocurre comúnmente en estudios de supervivencia.

Sumado a los anteriores indicadores, el Área Bajo la Curva ROC (AUC) es una métrica ampliamente utilizada para evaluar la capacidad discriminativa de modelos predictivos. En el caso de los análisis de supervivencia, donde los eventos pueden ocurrir en diferentes momentos y los datos pueden estar censurados, se requiere una adaptación de esta métrica,

la cual se denomina "AUC dependiente del tiempo" (AUC(t)) (Blanche et al., 2013; Heagerty et al., 2000).

El AUC tradicional mide la capacidad de un modelo para distinguir entre dos clases (por ejemplo, presencia o ausencia de una enfermedad) en un momento fijo. Sin embargo, en estudios de supervivencia, el interés radica en predecir la ocurrencia de un evento a lo largo del tiempo, y los datos pueden estar censurados (es decir, no se observa el evento durante el período de estudio para algunos individuos). El AUC dependiente del tiempo (AUC(t)) extiende el concepto tradicional para manejar estas características, evaluando la capacidad del modelo para discriminar entre individuos que experimentan el evento antes de un tiempo específico t y aquellos que no (Blanche et al., 2013; Heagerty et al., 2000; Uno et al., 2007).

Sin embargo, el índice de concordancia (C-index) y el área bajo la curva ROC (AUC, por sus siglas en inglés) son métricas estrechamente relacionadas, ambas utilizadas para evaluar la capacidad discriminativa de los modelos predictivos. En el caso de problemas de clasificación binaria, ambas medidas son equivalentes (Uno et al., 2011). El AUC, como se mencionó anteriormente, se interpreta como la probabilidad de que, al seleccionar aleatoriamente un par de individuos, el modelo asigne un puntaje de riesgo mayor al que pertenece a la clase positiva frente al de la clase negativa (Heagerty et al., 2000). Mientras que, el estadístico C, por su parte, constituye una generalización del AUC y se utiliza con mayor frecuencia en análisis de supervivencia, dado que puede manejar datos censurados y resultados de tiempo hasta el evento (Uno et al., 2011). En este contexto, como ya se había mencionado, el estadístico C mide la probabilidad de que, para un par aleatorio de individuos, aquel que experimenta el evento antes haya recibido un puntaje de riesgo mayor por parte del modelo (Newson, 2010; Uno et al., 2011). Al igual que el AUC, sus valores oscilan entre 0.5 (equivalente al azar) y 1.0 (discriminación perfecta).

Por lo tanto, para variables binarias el C-index y el AUC(t) son esencialmente equivalentes (Pölsterl, s. f.). No obstante, en el análisis de supervivencia, donde los datos suelen estar sujetos a censura, el C-index ofrece una ventaja decisiva al extender la lógica del AUC a este tipo de escenarios (Pölsterl, s. f.). En consecuencia, cuando se utiliza el C-index en estudios de supervivencia, el cálculo adicional del AUC(t) resulta redundante, ya que el primero

contiene la información discriminativa que el segundo provee, con la ventaja de estar adaptado a datos censurados.

Sumado a las anteriores métricas, en el análisis de supervivencia, es importante evaluar la robustez, estabilidad y capacidad predictiva de los modelos ajustados. Para esto, se pueden utilizar técnicas de validación interna como la validación cruzada (*cross-validation*), que permite estimar el rendimiento del modelo en muestras distintas a aquellas usadas para su ajuste, minimizando el sobreajuste y proporcionando estimaciones más realistas del error de predicción (Steyerberg et al., 2001; Efron & Tibshirani, 1997).

La validación cruzada consiste en dividir el conjunto de datos en k subconjuntos (o *folds*). En cada iteración, uno de los subconjuntos se utiliza como conjunto de validación, mientras que los $k - 1$ restantes se usan para entrenar el modelo. Este proceso se repite k veces y los resultados se promedian para estimar el rendimiento del modelo (Hastie et al., 2009).

En el contexto de supervivencia, este procedimiento puede adaptarse a la presencia de censura. Por ejemplo, se recomienda conservar la proporción de eventos censurados en cada partición mediante la técnica de *stratified k-fold cross-validation* (Almonteros & Matias, 2024; Bennis et al., 2021). Esta técnica busca asegurar que cada *fold* contenga una representación equilibrada tanto de eventos ocurridos como de censurados, lo que es especialmente importante para evitar estimaciones sesgadas del rendimiento del modelo. En lugar de realizar particiones aleatorias simples, el procedimiento estratificado divide los datos de forma que se mantenga la proporción observada de eventos en todas las particiones. Esto contribuye a una evaluación más estable y representativa de la capacidad predictiva del modelo. Además, se sugiere utilizar métricas específicas para datos de supervivencia, como el índice de concordancia (C-index) o el IBS en cada iteración (Blanche et al., 2013).

Además, una variante más robusta es la validación cruzada repetida. En esta, se realiza el mismo proceso anterior varias veces con diferentes divisiones aleatorias de los datos. Esto reduce la variabilidad debida a una sola partición y mejora la estimación del error predictivo (Steyerberg, 2019).

El uso del estadístico C de Harrell (C-index) o del Índice de Brier Integrado (IBS), en combinación con un esquema de validación cruzada estratificada repetida, permite evaluar si las diferencias observadas en las métricas de desempeño entre modelos son estadísticamente significativas. Para ello, es posible aplicar pruebas de hipótesis basadas en comparaciones pareadas en cada repetición y pliegue (*fold*) de la validación cruzada.

Las pruebas de hipótesis mencionadas pueden realizarse mediante una prueba *t* de Student para medias de diferencias pareadas. El procedimiento de estas pruebas comienza con el cálculo, en cada partición, de la diferencia en el rendimiento entre dos modelos. Posteriormente, a partir del conjunto de diferencias obtenidas en cada *fold* y repetición, se estiman la media $\bar{\mathrm{d}}$ y la desviación estándar s_d de dichas diferencias. Como estadístico de prueba es apropiado utilizar la t de Student pareada:

$$\mathrm{t}_0 = \bar{\mathrm{d}} \Big/ \mathrm{s}_\mathrm{d}/\sqrt{\mathrm{n}} \qquad \text{(Ecuación E.51)}$$

donde n corresponde al número total de pares (*folds* × repeticiones).

El contraste se realiza mediante una prueba bilateral, bajo la hipótesis nula de que no existe diferencia en la métrica promedio entre los modelos. De esta forma, un valor-p inferior al umbral de significancia ($\alpha = 0.05$) indica diferencias significativas en el desempeño medio entre modelos. Así, un modelo se considera superior si presenta sistemáticamente mayor C-index o menor IBS, y además resulta estadísticamente mejor, para un nivel de significancia establecido, en las comparaciones pareadas. Mientras que, en ausencia de diferencias significativas, los modelos se reportan como equivalentes en desempeño predictivo.

CAPÍTULO III. METODOLOGÍA

El presente capítulo describe de forma detallada el diseño metodológico adoptado para analizar la permanencia o abandono de los usuarios de la plataforma digital de contratación de personal de enfermería mediante técnicas de análisis de supervivencia. Esta metodología permite modelar el tiempo hasta la ocurrencia del evento de abandono (*churn*) e identificar los factores asociados a dicho comportamiento, tomando en cuenta tanto usuarios que abandonaron la plataforma como aquellos que permanecieron activos al final del periodo de observación.

A lo largo del capítulo se expone la fuente de datos utilizada, los criterios de inclusión de la muestra, la construcción de las variables relevantes para el análisis, y las técnicas estadísticas aplicadas. Asimismo, se detalla el enfoque secuencial seguido para la generación, ajuste y evaluación de los modelos de supervivencia no paramétricos, semi-paramétricos y paramétricos, así como los procedimientos para la selección del mejor modelo y la validación de sus resultados. Este enfoque garantiza un análisis robusto y fundamentado del fenómeno de estudio.

3.1. FUENTE DE DATOS

Los datos recolectados consisten en la información de los usuarios de un servicio digital de contratación "*per diem*" de personal de enfermería de los Estados Unidos. Se recolectó información en el periodo entre el 1 de diciembre de 2021 y el 1 de octubre de 2024. La información se encuentra almacenada en el servidor web de bases de datos de la plataforma digital en estudio. Dicha información se descargó y almacenó en un archivo digital tipo "RData", formato de datos utilizado por el software R (R Core Team, 2023), para ser luego manipulada con la ayuda del mismo software R. En particular, se emplearon paquetes como *dplyr* (Wickham et al., 2023) y *lubridate* (Grolemund & Wickham, 2011) para realizar tareas de limpieza, transformación, y generación de nuevas variables, para posteriormente, ajustar los modelos de supervivencia a dichos datos.

Los datos consisten en la actividad que los usuarios realizan en la plataforma, donde lo más importante para este estudio es si hizo o no una "solicitud" para trabajar en alguno de los trabajos publicados en la plataforma, ya que esto demuestra el interés del usuario por conseguir un trabajo y, por ende, por utilizar el servicio. Por lo tanto, en este estudio se está asumiendo que un usuario se encuentra activo (*no churn*) cuando realiza "solicitudes" de trabajo. Por otro lado, también se recopilaron datos personales y demográficos de cada uno de los usuarios de la plataforma para ser usadas como variables explicativas del fenómeno. Estas variables se explicaron en la sección "2.3. Explicación de las variables".

De esta forma, se obtuvo la cantidad de "solicitudes" por día que cada usuario realiza entre la fecha en que se registró a la plataforma y la última fecha del estudio (1 de octubre de 2024). Por lo tanto, la información primaria consiste en el identificador del usuario (ID), la fecha en que se registró a la plataforma, y la cantidad de solicitudes de trabajo que realizó en cada uno de los días en estudio. Luego, con dicha información se puede obtener la fecha de la última solicitud de trabajo de cada usuario.

Para identificar los datos censurados, se tomó como fecha de censura el 1 de agosto de 2024 (2024-08-01) (2 meses antes de la fecha final de recolección de los datos), por lo que aquellos usuarios que hayan realizado al menos una solicitud de trabajo después de dicha fecha se consideran como censurados (ya que no han incurrido en "*churn*" aún), y por tanto, su tiempo de vida es desconocido. De esta forma, este tiempo de vida se calcula como los días entre el 1 de octubre de 2024 (última fecha del estudio) y la fecha en que el usuario se registró a la plataforma (pero es un tiempo censurado). En este caso, la proporción de datos censurados es del 29.0%, mientras que la proporción de datos no-censurados es del 71.0%.

Una vez organizados los datos, la información principal para este análisis de supervivencia se presenta en una tabla cuyas columnas incluyen: el identificador del usuario (ID), la fecha en que el usuario se registró en la plataforma, la fecha de la última solicitud de trabajo, un indicador de si el usuario está censurado o no censurado, y el tiempo de vida del usuario. Además, se incluyen las variables independientes, cuya construcción se explica más adelante.

3.2. TAMAÑO Y SELECCIÓN DE LA MUESTRA

El presente estudio utilizó como fuente de información a los usuarios de la plataforma digital que realizaron al menos una solicitud laboral entre el 1 de diciembre de 2021 y el 1 de octubre de 2024, periodo definido como ventana de observación. Además, para ser incluidos en el estudio, los usuarios debían haberse registrado en la plataforma durante ese mismo intervalo de tiempo. Este periodo fue definido porque es en el que se cuenta con la información más confiable y, además, presenta una cantidad de datos sustancial que permite realizar un análisis robusto.

Con base en los criterios de inclusión definidos previamente, la muestra inicial estuvo compuesta por 51,496 usuarios. Sin embargo, antes de proceder con los análisis, fue necesario realizar un proceso de limpieza y depuración de los datos con el objetivo de garantizar la calidad y consistencia de la información.

En primer lugar, se excluyeron los registros que presentaban valores faltantes en variables clave, como la región geográfica. Adicionalmente, se eliminaron los usuarios cuya región estaba categorizada como "West Coast and Desert", ya que este grupo contenía muy pocos casos, lo que limita su representatividad y relevancia para los objetivos del estudio. Posteriormente, se depuraron los registros en los que la variable de tiempo de supervivencia presentaba valores iguales a cero o negativos, los cuales imposibilitan un análisis válido bajo los supuestos del modelo de supervivencia. En total, se identificaron y eliminaron 3,021 registros con tiempo igual a cero.

Como resultado de este proceso de depuración, la muestra final quedó conformada por 48 437 usuarios únicos, los cuales cumplen con todos los criterios requeridos para el análisis estadístico.

3.3. CONSTRUCCIÓN DE LAS VARIABLES

Con el objetivo de aplicar técnicas de análisis de supervivencia, fue necesario construir cuidadosamente dos variables fundamentales: la variable de tiempo y la variable de censura,

además de definir y construir las covariables explicativas que serían incorporadas posteriormente en los modelos estadísticos.

Variable de tiempo de supervivencia

La variable de *tiempo de supervivencia* se definió como el número de días transcurridos entre la fecha de registro de cada usuario en la plataforma y su última actividad registrada, entendida como la fecha de la última solicitud de empleo enviada. Para su cálculo, se identificaron los usuarios activos (aquellos con al menos una solicitud registrada) y se determinó, para cada uno, la fecha más reciente en la que realizaron dicha acción. En los casos en que el usuario permaneció activo más allá del 1 de agosto de 2024 (2024-08-01), considerada como la fecha de censura, se asignó como fecha de última actividad el 1 de octubre de 2024 (2024-10-01). Por tanto, el tiempo de supervivencia se calculó como la diferencia entre la fecha de última actividad (2024-10-01 en el caso de los censurados) y la fecha de creación de su cuenta.

Posteriormente, este tiempo se expresó en semanas dividiendo el número de días entre 7.

Variable de Censura

Dado que no todos los usuarios presentan un evento observable de "abandono" dentro del periodo de estudio, se incorporó una variable indicadora de censura, esencial para los modelos de supervivencia. El evento de interés fue definido como la inactividad sostenida del usuario, y se operacionalizó como la ausencia total de solicitudes durante los 60 días anteriores al cierre del periodo de observación, establecido en el 1 de octubre de 2024.

Bajo esta definición, se consideró que un usuario había experimentado el evento (abandono) si su última solicitud de empleo ocurrió antes del 1 de agosto de 2024. Es decir, si no había tenido actividad en los últimos 60 días del periodo de estudio. Por el contrario, si el usuario presentó actividad durante ese intervalo, se le clasificó como censurado, ya que se asume que su tiempo exacto de abandono no pudo ser observado dentro del marco temporal del estudio.

Es importante señalar que el umbral de 60 días fue elegido con base en criterios prácticos sobre el comportamiento de uso en plataformas digitales, y posteriormente fue sometido a un análisis de sensibilidad que se detalla más adelante en los resultados del estudio, con el fin de evaluar la robustez de esta decisión.

Variables independientes

En los modelos estadísticos de supervivencia, las variables independientes (también denominadas covariables explicativas) representan características que se presume influyen en el tiempo hasta la ocurrencia del evento de interés. Estas covariables permiten modelar cómo distintos factores individuales, demográficos, profesionales o conductuales afectan el riesgo de que ocurra dicho evento, proporcionando así una comprensión más precisa y detallada del fenómeno en estudio.

Para el presente análisis, se seleccionaron diversas covariables relevantes, tanto por su respaldo teórico como por su disponibilidad en los registros administrativos de la plataforma. Entre ellas se incluyen: una variable demográfica correspondiente a la edad del usuario al momento del registro; una variable profesional referida al tipo de licencia laboral del usuario; una variable geográfica que describe la zona de residencia o de operación principal del usuario; y una variable que refleja el nivel de éxito económico alcanzado por el usuario durante sus primeros meses de uso del servicio.

Las covariables fueron tratadas como variables continuas o categóricas según su naturaleza, y se sometieron a un análisis exploratorio preliminar para evaluar sus distribuciones, detectar posibles valores atípicos, examinar colinealidad entre variables, y asegurar su pertinencia en los modelos de supervivencia aplicados más adelante.

Construcción de la variable *Edad*

La variable *Edad* fue derivada a partir de la fecha de nacimiento registrada para cada usuario en la base de datos. Como primer paso, se convirtió el campo de fecha de nacimiento al formato de fecha estándar para su correcto procesamiento. Posteriormente, se aplicaron

criterios de limpieza para asegurar la validez de los valores observados: se consideraron como valores faltantes o inválidos todas aquellas fechas anteriores al 1 de enero de 1920, por considerarse inverosímiles, así como aquellas posteriores al 1 de junio de 2006, ya que implicarían edades inferiores a 18 años al momento del análisis, lo cual no es compatible con el perfil laboral del estudio. Para las fechas comprendidas entre el 1 de enero de 1920 y el 1 de enero de 1938, que podrían indicar registros válidos, aunque poco frecuentes, se unificaron todas esas observaciones asignando la fecha de 1 de enero de 1938 como fecha de nacimiento estándar.

Una vez depurada la variable, se calculó la edad de cada usuario como la diferencia, en años, entre el año de corte del estudio (1 de enero de 2024) y la fecha de nacimiento. Esta diferencia se redondeó al número entero más cercano. Con el fin de mantener la integridad de la base de datos y evitar eliminar registros, las observaciones cuya edad resultó como valor faltante fueron reemplazadas por 0, permitiendo su inclusión posterior como una categoría diferenciada. Para distinguir estos casos, se creó una variable indicadora (*dummy*) que toma el valor de 1 cuando la edad original era desconocida o inválida, y 0 en caso contrario (a esta variable se le denominó "sin_edad"). Este enfoque permite controlar por posibles sesgos asociados a la falta de información demográfica en los modelos de análisis subsiguientes.

Construcción de la variable *Licencia*

Una de las covariables explicativas clave incorporadas en los modelos de supervivencia fue el tipo de licencia profesional del usuario. Esta variable busca reflejar el perfil ocupacional del profesional de enfermería registrado en la plataforma, lo cual puede tener implicaciones directas sobre su comportamiento de uso y su permanencia activa en el sistema.

La construcción de esta variable implicó una depuración y transformación de los datos provenientes de los registros presentes en la plataforma digital. Primeramente, se generó una variable que clasifica cada usuario en una de cinco categorías mutuamente excluyentes, siguiendo un criterio jerárquico y práctico de codificación. Si el campo de licencia resultaba ausente (NA) o contenía cadenas de texto vacías como "NULL", el usuario se clasificó en la categoría "Sin licencia", correspondiente a personas sin licencia registrada. En segundo

lugar, se identificaron los casos que mencionaban una licencia de tipo RN (Registered Nurse), quienes fueron codificados como "RN". De manera similar, los usuarios con licencia LPN (Licensed Practical Nurse) fueron clasificados como "LPN"; aquellos con licencia de tipo CNA (Certified Nursing Assistant) o GNA (Geriatric Nursing Assistant) fueron agrupados bajo una misma categoría denominada "CNA / GNA", debido a su alta similitud funcional. Finalmente, cualquier otra forma de licencia que no encajara en las categorías anteriores fue incluida bajo la clasificación "Otras".

Este proceso de recodificación permite transformar dicha variable de tipo texto poco estructurada en una covariable categórica adecuada para el análisis de este trabajo, asegurando una representación clara, consistente y analíticamente útil del perfil profesional de los usuarios dentro de la muestra.

Construcción de la variable *Región*

La variable Región fue incorporada como una covariable categórica que indica la zona geográfica principal en la que cada usuario opera o reside. Esta información ya se encontraba previamente estructurada y codificada dentro de las bases de datos administrativas de la plataforma digital, por lo que no fue necesario modificar la información obtenida. La variable había sido generada internamente utilizando criterios consistentes de segmentación regional, lo cual facilitó su uso directo en este análisis.

No obstante, se realizó una depuración mínima para asegurar la calidad y relevancia de los datos. En primer lugar, se eliminaron aquellos registros en los que la región no estaba especificada (valores NA) los cuales eran únicamente 10 usuarios. En segundo lugar, se excluyeron los usuarios cuya región correspondía a "West Coast and Desert", debido a que únicamente 12 individuos pertenecían a esta categoría, lo que representaba una frecuencia insuficiente para un análisis estadístico robusto y podría afectar la estabilidad de las estimaciones.

En el Anexo A se presenta el Cuadro A.1., que contiene los estados incluidos en cada región.

Construcción de la variable *Ingresos*:

Con el objetivo de incorporar información sobre la participación económica de los usuarios en la plataforma, se construyó una covariable derivada a partir de los ingresos registrados por cada profesional. Esta variable permite evaluar si el usuario generó ingresos durante los primeros días de actividad laboral y, por tanto, si logró concretar oportunidades laborales mediante la plataforma digital. Específicamente, se generó la variable *Ingresos*, que representa de forma dicotómica si el usuario obtuvo algún ingreso económico durante sus primeros 60 días activos utilizando la plataforma (0 = no generó ingresos, 1 = sí generó ingresos).

El proceso de construcción de esta variable comenzó con la integración de los registros financieros al conjunto de datos principal, a través de una unión por identificador único de usuario (ID). Luego, se definió el intervalo de tiempo relevante: desde la fecha mínima de actividad laboral registrada para el usuario hasta 60 días después de esta. Dentro de este periodo, se sumaron todos los ingresos asociados a trabajos que comenzaron dentro del rango de fechas especificado. El resultado de esta suma fue almacenado en una variable cuantitativa que representa el total de ingresos obtenidos por el usuario durante sus dos primeros meses de actividad.

Posteriormente, con el fin de facilitar su uso en modelos estadísticos e interpretación, se transformó esta variable continua en una variable dicotómica, denominada *Ingresos*. Esta fue definida de la siguiente manera: se asignó un valor de 1 si el usuario generó ingresos mayores a cero en ese periodo inicial, y un valor de 0 si no se registraron ingresos. Esta codificación permite analizar de manera binaria la relación entre los ingresos tempranos del usuario y su comportamiento de permanencia en la plataforma.

Como paso complementario, todos los valores faltantes (*NA*) en la variable de *Ingresos* fueron reemplazados por ceros, bajo la lógica de que la ausencia de información representa la no generación de ingresos durante el periodo analizado.

3.4. ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio de los datos (AED) constituye una fase crucial en todo estudio empírico, ya que permite comprender la estructura general del conjunto de datos, detectar posibles inconsistencias, valores atípicos, y explorar relaciones preliminares entre las variables (James et al., 2021; Tukey, 1977). En este estudio, el AED se llevó a cabo en cuatro ejes principales: análisis descriptivo, evaluación de las relaciones entre variables, análisis exploratorio de las variables mediante curvas de supervivencia, y caracterización de valores extremos.

Este análisis exploratorio constituye una etapa esencial para validar la calidad del conjunto de datos, identificar patrones relevantes y guiar la construcción e interpretación de los modelos estadísticos que se presentan en las siguientes secciones del estudio.

3.4.1. Análisis descriptivo de las variables

En primer lugar, se realizaron análisis estadísticos univariados para todas las variables incluidas en el estudio. Las variables continuas, como la edad y el tiempo de supervivencia (medido en semanas), fueron descritas mediante medidas de tendencia central (media, mediana), de dispersión (desviación estándar), y de posición (mínimos y máximos), además de histogramas para observar sus distribuciones. Para las variables categóricas, como tipo de L*icencia*, *Región*, *Ingresos* y estado de *Censura*, se calcularon frecuencias absolutas y relativas, por medio de tablas y gráficos con las distribuciones porcentuales de cada categoría.

3.4.2. Análisis de relaciones entre variables independientes

Con el objetivo de explorar asociaciones entre variables categóricas, se elaboraron tablas de contingencia que describen la distribución proporcional conjunta entre pares de variables. Estas tablas permitieron observar visualmente posibles patrones de asociación entre las categorías de las distintas variables.

Posteriormente, se aplicaron pruebas de independencia basadas en el estadístico chi-cuadrado (χ^2) (Agresti, 2007), complementadas con la estimación del coeficiente *V* de Cramer (Cramer, 1946), utilizado como medida de la fuerza de asociación entre variables nominales.

Asimismo, se utilizó análisis de varianza (ANOVA) (Montgomery, 2013) para examinar diferencias en la media de edad entre los distintos niveles de variables categóricas. Esta técnica permitió evaluar si la edad promedio varía significativamente según las distintas categorías.

3.4.3. Análisis exploratorio de la variable dependiente: tiempo de supervivencia

Con el objetivo de explorar el comportamiento de la variable dependiente, tiempo de supervivencia en la plataforma, se examinó su distribución mediante histogramas, lo cual permite identificar asimetrías, acumulaciones de eventos en determinados intervalos y posibles valores atípicos.

Posteriormente, se aplicaron estimadores no paramétricos de Kaplan-Meier para la muestra total, así como para distintos subgrupos definidos según características relevantes de los usuarios. Las curvas de supervivencia generadas permiten visualizar de manera clara el patrón de abandono a lo largo del tiempo, evidenciando momentos con mayor o menor riesgo de deserción durante el periodo de análisis.

Asimismo, se elaboraron curvas de Kaplan-Meier estratificadas según las variables de *Ingresos*, *Región*, y *Licencia*. Esta estratificación facilita la comparación entre categorías y permite identificar diferencias relevantes en la dinámica de permanencia de los usuarios, aportando insumos valiosos para la posterior modelación de los datos.

3.4.4. Identificación y caracterización de valores extremos

Como complemento al análisis descriptivo, se procedió a identificar valores extremos (VE) mediante el análisis de los residuos de deviancia (Therneau & Grambsch, 2000; Hosmer et al., 2008), obtenidos a partir del modelo extendido de Cox (modelo de Cox ajustado con

interacciones temporales logarítmicas, $log\, t + 1$). Ya que, como señala Therneau y Grambsch (2000), los residuos de deviancia fueron diseñados para mejorar a los residuos de Martingala en la detección de valores atípicos individuales, particularmente en aplicaciones gráficas.

Para esto, se definieron dos grupos de valores extremos:

- VE altos: individuos con residuos de deviancia mayores a 2.
- VE bajos: individuos con residuos de deviancia menores a -2.

Se realizó una caracterización detallada de ambos grupos en comparación con la muestra total, considerando variables como la fecha de registro, *Edad*, *Región*, *Ingresos*, estado de censura y tiempos de falla. Esta comparación se llevó a cabo mediante gráficos y tablas descriptivas y comparativas.

Esta caracterización permite detectar perfiles atípicos que podrían influir en la estabilidad de los modelos estadísticos, y constituye un insumo valioso para decisiones posteriores relacionadas con la inclusión o análisis separado de estos casos en los modelos de supervivencia.

3.5. MODELOS GENERADOS

En esta sección se describe detalladamente la metodología utilizada para la generación, evaluación, validación y obtención de los resultados de los distintos modelos de análisis de supervivencia aplicados en el estudio. Se adoptó un enfoque secuencial, comenzando por los modelos no paramétricos, continuando con los modelos semi-paramétricos y concluyendo con los modelos paramétricos. Cada modelo fue sometido a un proceso riguroso de verificación de supuestos estadísticos, validación interna y evaluación del desempeño predictivo, siguiendo las prácticas metodológicas recomendadas por autores como Harrell (2015) y Steyerberg (2019), quienes destacan la importancia de combinar criterios estadísticos y teóricos en la selección y ajuste de modelos.

3.5.1. Modelos no paramétricos

El análisis de los datos por medio de modelos de supervivencia se realizó, en primer lugar, utilizando el estimador no paramétrico de Kaplan-Meier, el cual permite estimar la función de supervivencia sin necesidad de asumir una distribución específica para el tiempo hasta el evento. Entre las principales ventajas de este modelo se encuentran su flexibilidad, su simplicidad de interpretación y el hecho de que requiere pocos supuestos estadísticos, siendo los más relevantes el de independencia entre los tiempos de supervivencia y el mecanismo de censura (Hosmer et al., 2008).

Para el ajuste de los modelos de Kaplan-Meier se utilizó, como variable dependiente, el tiempo del usuario entre su registro en la plataforma hasta que abandonó el servicio (tiempo de supervivencia), medido en semanas. Se utilizó una variable de censura para distinguir entre usuarios que abandonaron la plataforma en el periodo de observación (evento observado) y aquellos cuya actividad final no fue observada dentro del periodo de estudio (censurados). Como variables explicativas o de estratificación se consideraron las siguientes: *Edad*, el tipo de *Licencia* profesional, la *Región* geográfica y la generación de *Ingresos*.

Los modelos fueron ajustados en el entorno de R utilizando el paquete *survival* (Therneau, 2024), mediante la función *survfit()*, que permite calcular la función de supervivencia estratificada por una o más covariables.

Este enfoque permite identificar diferencias relevantes en los patrones de abandono de los usuarios a lo largo del tiempo, facilitando la comparación de curvas de supervivencia entre distintos grupos definidos por covariables claves. En particular, resultó útil para visualizar de manera clara si ciertas características de los usuarios están asociadas con mayores o menores tasas de deserción en la plataforma.

Las curvas de supervivencia de Kaplan-Meier fueron acompañadas por intervalos de confianza al 95%, calculados mediante el método de Greenwood, que es el estándar implementado por la función *survfit()* para estimar la varianza de la función de supervivencia. También, se realizaron pruebas de Log-Rank para evaluar diferencias significativas entre grupos. Asimismo, se generaron gráficos estratificados que permiten visualizar la

variabilidad en los patrones de retención según las características de los usuarios. Entre las ventajas de estos gráficos se encuentra que permiten detectar si existe heterogeneidad en el comportamiento de supervivencia entre los grupos.

Cuando se realizaron múltiples comparaciones entre curvas (por ejemplo, al contrastar varias categorías dentro de una misma covariable), se aplicó la corrección de Bonferroni (Bonferroni, 1936) con el fin de controlar el error por comparaciones múltiples. En particular, esta corrección controla la tasa de error familiar (Family-Wise Error Rate, FWER), es decir, la probabilidad de cometer al menos un falso positivo al considerar simultáneamente un conjunto de pruebas. El método consiste en ajustar el nivel de significancia dividiéndolo entre el número de comparaciones realizadas, de forma que cada prueba individual se evalúa con un umbral más estricto.

Una ventaja de esta corrección de Bonferroni es su simplicidad y su capacidad para mantener controlada la FWER de manera general. Sin embargo, su principal desventaja es que tiende a ser conservadora, especialmente cuando el número de comparaciones es grande, lo cual puede reducir la potencia estadística y aumentar la probabilidad de no detectar diferencias reales. Esta prueba se puede presentar por medio de la siguiente Ecuación E.52.

$$\alpha^* = \frac{\alpha}{m} \qquad \text{Ecuación E.52}$$

donde, α es el nivel de significancia global (por ejemplo, 0.05); m es el número total de comparaciones realizadas; y, α^*es el nivel de significancia ajustado que se usa en cada prueba individual.

Por otro lado, para la representación gráfica de la función de riesgo instantáneo (*hazard*) a lo largo del tiempo, se empleó una estimación no paramétrica mediante el paquete *muhaz* de R (Hess & Gentleman, 2021). Este procedimiento parte de la información de tiempos observados de falla o censura y del indicador de evento, a partir de los cuales se construye un estimador suavizado del *hazard*, sin asumir una forma funcional o distribución paramétrica específica para el riesgo.

La función de riesgo instantáneo $\lambda(t)$ describe la probabilidad instantánea de experimentar el evento en un intervalo infinitesimal alrededor de t, dado que el individuo ha sobrevivido hasta ese momento. Dado que $\lambda(t)$ no se puede observar directamente, se utiliza un estimador tipo *kernel*, que aplica técnicas de suavizado sobre los tiempos de evento. En particular, el estimador de *hazard* suavizada implementado por *muhaz* corresponde al método propuesto por Müller y Wang (1994), que utiliza ventanas de suavizado (*bandwidths*) para obtener una curva continua y estable del riesgo en función del tiempo, reduciendo la variabilidad inherente a estimadores empíricos no suavizados.

Este procedimiento es no paramétrico, en el sentido de que no impone una forma funcional predeterminada para el *hazard* (como ocurre en modelos paramétricos), sino que se adapta a la estructura de los datos observados. El suavizado controla el grado de detalle de la curva: valores pequeños del parámetro de suavizado producen curvas más sensibles a fluctuaciones puntuales, mientras que valores grandes generan curvas más estables y suaves, aunque con menor capacidad de reflejar cambios locales en el *hazard*.

En síntesis, las curvas obtenidas permiten visualizar la evolución del riesgo instantáneo a lo largo del tiempo para cada conjunto de datos, proporcionando un diagnóstico exploratorio útil para identificar periodos de mayor riesgo de ocurrencia del evento y complementar el análisis realizado posterior con modelos paramétricos y semi-paramétricos.

3.5.2. Modelos semi-paramétricos

Posteriormente, se ajustó un modelo semi-paramétrico de riesgos proporcionales de Cox, el cual permite evaluar el efecto de múltiples covariables sobre el riesgo de abandono sin necesidad de asumir una distribución específica para la función de riesgo basal. Este modelo se implementó en R utilizando también el paquete *survival* (Therneau, 2024), y específicamente la función *coxph()*.

Como en el modelo de Kaplan-Meier, la variable dependiente fue el tiempo registrado hasta el abandono (medido en semanas), y, también, se utilizó la variable de censura como

indicador del estado del evento. Sumado a esto, se utilizaron como covariables (variables independientes): la *Edad*, el tipo de *Licencia*, la *Región* y la variable dicotómica de *Ingresos*.

Antes de interpretar los resultados del modelo, se evaluaron rigurosamente sus supuestos estadísticos (mencionados en la sección *"2.4.2. Modelos semi-paramétricos"*). En particular, se examinó el supuesto de proporcionalidad de riesgos mediante el análisis de residuos de Schoenfeld (Grambsch & Therneau, 1994), tanto en su representación gráfica como a través de pruebas estadísticas formales. Adicionalmente, se utilizaron residuos de Martingala para verificar la linealidad de las covariables continuas como la *Edad*.

Dado que se encontraron violaciones al supuesto de riesgos proporcionales, se optó por ajustar una versión extendida del modelo de Cox (modelo de Cox extendido), el cual permite que los coeficientes varíen con el tiempo. Este enfoque incorpora efectos no constantes en el tiempo mediante interacciones entre covariables y funciones del tiempo, lo que facilita capturar dinámicas cambiantes en el riesgo de abandono durante el período de observación.

Este modelo "Cox extendido" se construyó utilizando la función *coxph()* en combinación con el término *tt()* (time transform) del paquete *survival* (Therneau, 2024). El cual permite especificar una transformación del tiempo para cada covariable cuyo efecto se presume variable. En este caso, se utilizó la transformación logarítmica $log(t + 1)$ dentro de la fórmula del modelo para capturar los cambios dinámicos del efecto de la covariable sobre el riesgo. Esta metodología permitió modelar de forma más realista situaciones donde el efecto de una covariable no es constante en el tiempo.

Para seleccionar las variables del modelo extendido de Cox, se aplicó una estrategia de selección jerárquica basada en comparaciones mediante pruebas de verosimilitud (*Likelihood Ratio Test*). Esta técnica, ampliamente utilizada en modelos de regresión, resulta especialmente útil cuando se trabaja con modelos anidados, es decir, cuando un modelo es una versión simplificada del otro. Inicialmente, se incluyeron todas las covariables consideradas y analizadas en las secciones previas, así como las interacciones sugeridas en el análisis de riesgos proporcionales. Posteriormente, se fueron excluyendo aquellas variables o términos que no aportaban significativamente al ajuste del modelo, evaluando en cada paso si la simplificación afectaba el desempeño estadístico del modelo.

Este procedimiento se fundamenta en el principio de parsimonia, el cual sugiere que, entre dos modelos con un ajuste similar, se debe preferir el más sencillo. Aunque incluir variables no significativas en el modelo no representa un error en sí mismo, su permanencia puede aumentar la complejidad del modelo innecesariamente, reducir la interpretabilidad de los coeficientes, y provocar inestabilidad en las estimaciones, especialmente en muestras pequeñas o cuando existe colinealidad. Por esto, se recomienda mantener únicamente aquellas variables o interacciones cuyo efecto se respalda con evidencia estadística.

Los coeficientes del modelo se estimaron mediante máxima verosimilitud utilizando la función antes mencionada (*coxph*) del paquete *survival*. La interpretación de estos coeficientes depende de si corresponden a efectos proporcionales (constantes en el tiempo) o a efectos extendidos (interacciones con el tiempo). Posteriormente, se realizó una revisión detallada de estos coeficientes, su significancia estadística, y su interpretación.

Luego, para la construcción de las curvas de supervivencia $S(t)$ y de riesgo $h(t)$, se emplearon herramientas propias basadas en la estructura del modelo de Cox extendido. En particular, se implementaron funciones personalizadas que: (i) calculan los incrementos acumulados de la función de riesgo de base para cada instante de evento observado; (ii) evalúan la contribución del perfil del usuario sobre el riesgo a lo largo del tiempo; y (iii) obtienen la función de supervivencia a partir del riesgo acumulado utilizando la relación $S(t) = \exp[-H(t)]$. A partir de estas curvas se construyeron intervalos de confianza del 95%, obtenidos mediante remuestreo paramétrico de los coeficientes del modelo (asumiendo su distribución normal asintótica y recalculando las curvas en cada réplica).

Por último, se llevó a cabo una evaluación de la bondad de ajuste y de la capacidad predictiva del modelo mediante diversos indicadores. En primer lugar, se calcularon el AIC y el BIC, que permiten comparar modelos alternativos penalizando la complejidad paramétrica. En segundo lugar, se estimó el índice de concordancia (C-index) con validación cruzada (seis particiones (*folds*) y cuatro repeticiones), el cual mide la capacidad del modelo para discriminar correctamente el orden relativo de los tiempos de evento entre individuos. Sumado a lo anterior, se calculó el *Brier Score Integrado* (IBS), también mediante validación cruzada con seis particiones (*folds*) y cuatro repeticiones, que cuantifica la precisión global

de las predicciones de supervivencia a lo largo del tiempo, penalizando las discrepancias entre las probabilidades estimadas y los resultados observados.

3.5.3. Modelos paramétricos

Se exploraron, ajustaron y evaluaron modelos paramétricos de supervivencia, los cuales asumen una distribución específica para el tiempo hasta la ocurrencia del evento. En particular, en este estudio se consideraron las siguientes distribuciones: Exponencial, Weibull, Log-normal, Log-logística, Gamma, y Gompertz.

En todos los casos, la variable dependiente de los modelos correspondió al tiempo registrado hasta el abandono (medido en semanas), mientras que la variable de censura indicó el estado del evento (censurado o no censurado). Además, como covariables en cada uno de los modelos se incluyeron: la *Edad*, el tipo de *Licencia*, la *Región* y la variable dicotómica de *Ingresos*.

Los modelos fueron ajustados, con el software R, empleando las librerías *survival* (Therneau, 2024), para los modelos: Exponencial, Weibull, Log-normal, y Log-logístico, y *flexsurv* (Jackson, 2016), para los modelos: Gamma y Gompertz. Posteriormente, se revisaron los supuestos estadísticos asociados a cada modelo y se evaluó su desempeño en términos de ajuste a los datos y capacidad predictiva.

Dado el no cumplimiento del supuesto de riesgos proporcionales de las variables *región* e *ingresos*, lo cual también afecta a algunos de los modelos paramétricos, se decidió estratificar la base de datos en múltiples subconjuntos, definidos a partir de las combinaciones de estas variables *región* e *ingresos*. Garantizando de esta manera el cumplimiento de los supuestos de estos modelos paramétricos. El procedimiento dio lugar a 14 subconjuntos de datos que representan configuraciones específicas de la población de estudio (Por ejemplo: usuarios "con ingresos" y de la región "Utah & Arizona", o usuarios "sin ingresos" y de la región "California & Nevada", creando de esta forma los 14 subconjuntos).

Una vez construidos los subconjuntos, se procedió a ajustar, de manera independiente en cada uno de ellos, los distintos modelos paramétricos de supervivencia mencionados. El

ajuste de los modelos se realizó utilizando los métodos de máxima verosimilitud disponibles en las librerías mencionadas del software estadístico R, asegurando la estimación eficiente de los parámetros y la obtención de medidas de bondad de ajuste. Posteriormente, para cada uno de los 14 subconjuntos se seleccionó el modelo paramétrico más adecuado de acuerdo con los criterios de comparación descritos más adelante en la sección "3.5.3.1. Selección del mejor modelo paramétrico", los cuales incluyen medidas de información (AIC y BIC), capacidad predictiva y validación interna.

De esta manera, la metodología adoptada permite mitigar la violación de los supuestos de los modelos, y obtener estimaciones más realistas de la calidad de cada uno de los modelos propuestos.

Una vez identificados, para cada subconjunto, los modelos con mejor desempeño según las métricas de ajuste y predicción, se realizó una síntesis de dichos resultados que incluyó también el ajuste con la base de datos completa. A partir de esta comparación global se eligió el modelo Log-normal como modelo paramétrico de referencia, dado su buen ajuste a los datos y su rendimiento competitivo en las métricas predictivas. Complementariamente, se seleccionó el modelo Gamma como un candidato sólido para las predicciones, debido a su buen desempeño en este ámbito.

El modelo Log-normal se volvió a ajustar utilizando la base de datos completa, con el fin de obtener estimaciones de los coeficientes de las covariables más estables y útiles para la toma de decisiones en la plataforma digital (por ejemplo, en ámbitos operativos y de mercadeo). Sobre este ajuste definitivo se evaluaron los supuestos del modelo y se estudiaron a profundidad los coeficientes asociados a las covariables.

En particular, para el modelo Log-normal se verificó la plausibilidad de la distribución Log-normal de los tiempos hasta el evento, la adecuación de la formulación bajo el enfoque de tiempo acelerado de falla (AFT), la linealidad de las covariables continuas en la escala del log-tiempo y la ausencia de patrones sistemáticos en los residuos. El resto de las distribuciones consideradas se mantuvieron como alternativas de comparación, pero no se sometieron al mismo nivel de escrutinio, ya que no se utilizaron para la interpretación principal de los efectos de las covariables.

Entre estas alternativas, el modelo Gamma destacó por su buen desempeño en términos predictivos, especialmente por obtener de forma sistemática valores más bajos de IBS en las pruebas t pareadas de validación cruzada. Por este motivo, aunque no se analizaron con el mismo detalle sus supuestos ni se usaron sus coeficientes para extraer conclusiones sustantivas, el modelo Gamma se conservó como modelo paramétrico con fines esencialmente predictivos.

En cuanto a las predicciones de tiempos de supervivencia, se emplean los dos modelos paramétricos seleccionados, Log-normal y Gamma, para obtener, a partir de las funciones de supervivencia estimadas, los tiempos (t_S) y sus intervalos de confianza asociados a distintos niveles de probabilidad de supervivencia (S) y a diversos perfiles de usuarios. Estos resultados se contrastan posteriormente con los obtenidos mediante el modelo de Cox extendido y el estimador de Kaplan-Meier, lo que permite evaluar de forma conjunta la coherencia y el desempeño predictivo de los distintos enfoques. Esta estrategia se detalla con mayor profundidad en la sección "3.7. Cálculos de tiempos de supervivencia".

3.5.3.1. Selección del mejor modelo paramétrico

La selección del mejor modelo de supervivencia constituye una fase crítica en el proceso metodológico, ya que permite identificar el modelo que ofrece el equilibrio óptimo entre ajuste estadístico, cumplimiento de supuestos, capacidad predictiva y parsimonia. En el presente estudio, se evaluaron diversos modelos de supervivencia paramétricos, con el fin de seleccionar el más adecuado para explicar el fenómeno de abandono de la plataforma digital.

Para realizar esta selección, se siguió un procedimiento estructurado compuesto por las etapas explicadas en las siguientes secciones.

Comparación del ajuste a los datos: AIC y BIC

En esta etapa del análisis se comparó el ajuste estadístico de los modelos utilizando los criterios de información de Akaike (AIC) y Bayesiano (BIC). Ambos criterios penalizan la complejidad del modelo, aunque en distinta magnitud, y permiten la comparación entre

modelos, incluso cuando no son anidados, siempre que hayan sido ajustados sobre la misma muestra de datos y con la misma variable dependiente.

En una primera instancia, estos criterios se calcularon para los modelos ajustados sobre la base de datos completa. Adicionalmente, con el fin de explorar posibles heterogeneidades y reducir el impacto de eventuales violaciones del supuesto de riesgos proporcionales (PH), la comparación mediante AIC y BIC se repitió en cada uno de los 14 subconjuntos de datos, definidos por las combinaciones de *Región* e indicador de *Ingresos*. De esta forma se dispuso tanto de una evaluación global como de una evaluación estratificada por subpoblaciones. En todos los casos, se consideró preferible el modelo con el menor valor de AIC y, de manera complementaria, el que presentara el menor valor de BIC. La interpretación final de los resultados se basó en la evidencia conjunta aportada por ambos criterios, tanto en el análisis con la muestra completa como en los 14 subconjuntos.

Evaluación de la capacidad predictiva con validación cruzada

La capacidad predictiva de los modelos fue evaluada mediante el cálculo del estadístico C de Harrell, y el IBS ajustado por censura. Estas métricas permiten valorar tanto el poder discriminativo como la precisión de las predicciones realizadas por cada modelo.

- Estadístico C de Harrell (C-index): Evalúa la capacidad del modelo para discriminar correctamente entre pares de individuos. Específicamente, mide la concordancia entre las predicciones de riesgo y el orden real de ocurrencia del evento. Un valor de C cercano a 1 indica una excelente capacidad de discriminación, mientras que un valor de 0.5 sugiere un desempeño equivalente al azar. Este indicador es especialmente útil en presencia de censura. En este estudio, la métrica C-index fue estimada mediante la función *Score()* del paquete *riskRegression* de R (Gerds et al., 2025).
- El Índice de Brier Integrado (IBS) constituye una extensión del Brier Score que resume, en una sola medida, la precisión predictiva de un modelo de supervivencia a lo largo de todo el horizonte temporal de análisis. Mientras que el Brier Score en un

instante específico t mide el error cuadrático medio entre las probabilidades de supervivencia predichas y los eventos observados, el IBS integra esta información a través del tiempo, proporcionando una medida global del desempeño predictivo.

En el contexto de datos censurados, como es habitual en análisis de supervivencia, el IBS se ajusta mediante técnicas como el peso de probabilidad inversa (IPCW, *Inverse Probability of Censoring Weighting*), lo que permite incorporar adecuadamente la información parcial de los individuos censurados. De esta forma, el IBS ajustado refleja no solo la capacidad del modelo para discriminar entre individuos con diferentes riesgos, sino también su calibración global frente a la censura presente en los datos. En este estudio, la métrica IBS fue estimada mediante la función *pec()* del paquete *pec* de R (Gerds, 2025).

Para robustecer la evaluación, tanto el C-index como el IBS se calcularon utilizando validación cruzada, la cual es una técnica de validación interna que pretende corregir el posible sesgo optimista que se produce al evaluar un modelo en los mismos datos con los que fue entrenado. Este procedimiento resulta esencial, ya que permite garantizar que las métricas de desempeño obtenidas reflejen de manera más realista la capacidad predictiva del modelo.

En particular, se implementó la técnica de validación cruzada estratificada (*stratified k-fold cross-validation*), la cual constituye una estrategia robusta para verificar la estabilidad de los modelos. Este método consiste en dividir la muestra original en varios subconjuntos o "*folds*", procurando que cada uno mantenga una proporción balanceada de eventos y censuras. Posteriormente, en cada iteración, el modelo se entrena con un conjunto de datos y se valida en el subconjunto restante, garantizando que la evaluación no dependa de una única partición de los datos.

Este procedimiento aporta diversos beneficios, entre los que se encuentra: (i) permite obtener estimaciones más consistentes y menos sesgadas de las métricas de desempeño, y (ii) facilita la comparación entre diferentes modelos al reducir la variabilidad asociada a la selección de una sola partición. De esta manera, la validación cruzada estratificada se convierte en una

herramienta fundamental para evaluar la capacidad predictiva real de los modelos de supervivencia de este estudio.

Luego, para comparar un modelo contra otro según las métricas C-index y IBS con validación cruzada, se emplearon pruebas de hipótesis. Donde, se aplicó una prueba t de Student para medias de diferencias pareadas, la cual se menciona en el marco teórico. Bajo la hipótesis nula de que no existe diferencia en la métrica promedio entre los modelos, y con un umbral de significancia de 0.05 ($\alpha = 0.05$) se averiguó si existen o no diferencias significativas en el desempeño medio entre los modelos. Así, un modelo se considera superior si presenta sistemáticamente mayor C-index o menor IBS, y además resulta estadísticamente mejor, para dicho nivel de significancia.

Evaluación del cumplimiento de supuestos

Seguidamente, se evaluó el cumplimiento de los supuestos fundamentales de las distribuciones paramétricas seleccionadas, en particular de la distribución Log-normal, de acuerdo con lo descrito en la sección de supuestos de los modelos paramétricos. El objetivo de esta etapa fue verificar y comprender los modelos que, aun mostrando buenas métricas de ajuste, resultaran discordantes con la estructura teórica requerida por cada familia de distribución, así como comprender mejor sus limitaciones.

En primer lugar, se llevó a cabo una evaluación gráfica de los supuestos estructurales asociados a las distribuciones con enfoque de riesgos proporcionales (PH) y de tiempo acelerado de falla (AFT). Para ello se utilizaron gráficos de $\log[-\log S(t)]$contra $\log(t)$, comúnmente conocidos como gráficos log–log (Kleinbaum & Klein, 2005). La forma y disposición de las curvas para distintos grupos permiten interpretar, de manera aproximada, lo siguiente:

- Líneas aproximadamente rectas y paralelas sugieren que una distribución Weibull describe razonablemente los datos, lo que implica que el modelo puede formularse coherentemente bajo enfoques PH y AFT dentro de esta familia.

- Líneas aproximadamente rectas y paralelas con pendiente cercana a 1 indican que el modelo Exponencial (caso particular del Weibull) es adecuado.
- Líneas aproximadamente paralelas, pero con curvatura evidente, pueden sugerir que el supuesto de riesgos proporcionales (PH) es razonable, aunque la forma Weibull/AFT no se ajuste bien.
- Líneas claramente no paralelas o que se cruzan indican que el supuesto de riesgos proporcionales no se cumple, lo que cuestiona la validez de modelos formulados estrictamente en el marco PH.

Además de estos gráficos log–log, en el caso particular del modelo Log-normal, se verificó de forma específica la plausibilidad de que el logaritmo de los tiempos hasta el evento siguiera una distribución aproximadamente normal, así como la adecuada linealidad de las covariables continuas en la escala del log-tiempo y el supuesto de homocedasticidad que estipula la ausencia de patrones sistemáticos en los residuos.

En primer lugar, el supuesto de forma Log-normal de los tiempos de falla se evaluó mediante gráficos de diagnóstico sobre $log(T)$, incluyendo histogramas y gráficos de probabilidad normal (Q–Q plots), con el fin de verificar la simetría, la alineación de los cuantiles observados con la recta teórica y la ausencia de colas excesivamente pesadas.

Posteriormente, el supuesto de homocedasticidad se revisó a través de la homogeneidad y ausencia de patrones sistemáticos en los residuos, mediante un diagnóstico gráfico de residuos. En datos con censura, los residuos habituales no son apropiados; por ello se recomienda utilizar los residuos de devianza del modelo ajustado. Por lo que, para analizar la homocedasticidad y detectar posibles malas especificaciones, se graficaron los residuos de devianza frente al predictor lineal ($X\hat{\beta}$).

Finalmente, el supuesto de linealidad de las covariables continuas se revisó por medio de gráficos que muestran la relación entre los residuos estandarizados del log-tiempo y las covariables continuas, con el objetivo de identificar posibles tendencias o curvaturas que indicaran desviaciones de la linealidad asumida en el modelo.

3.6. ANÁLISIS DE SENSIBILIDAD DE LA FECHA DE CENSURA

En los análisis de supervivencia, la censura corresponde a la situación en la cual el tiempo hasta la ocurrencia del evento de interés no se observa completamente para algunos individuos. En este estudio, la censura es exclusivamente por la derecha, y ocurre cuando los usuarios aún permanecen activos al finalizar el periodo de observación, de modo que no se conoce con exactitud su tiempo de abandono de la plataforma. Esta característica es especialmente relevante debido a que, en el servicio analizado, los usuarios no mantienen una suscripción formal que deba cancelarse, sino que pueden interrumpir y retomar su actividad en cualquier momento, dificultando así la determinación precisa de la fecha de abandono. Por lo tanto, tanto la definición de la censura como la determinación del tiempo de abandono deben entenderse como aproximaciones sujetas a incertidumbre, ya que pueden contener un margen de error no observable en los datos disponibles.

Para operacionalizar la censura, se estableció que un usuario se considera censurado si realizó al menos una solicitud de empleo en la plataforma en el intervalo comprendido entre el primero de agosto de 2024 y el primero de octubre de 2024 (fecha de finalización de la recopilación de datos), lo cual corresponde a un intervalo de 60 días. Bajo esta definición, el tiempo de supervivencia de los usuarios censurados se calculó como el número de semanas transcurridas entre la fecha de creación de la cuenta y el 1 de octubre de 2024. En el caso de los usuarios no censurados, su tiempo de supervivencia se definió como el intervalo comprendido entre la fecha de creación de la cuenta y la última solicitud de empleo registrada. No obstante, debe considerarse que algunos de estos usuarios clasificados como censurados podrían retomar su actividad en la plataforma después de la fecha de cierre de la recopilación de datos, lo cual introduce un error en los datos y una limitación metodológica que es la que se intenta medir y mitigar el error mediante este análisis de sensibilidad.

Esta elección del intervalo de 60 días previo al cierre de la información para seleccionar los usuarios censurados constituye una decisión metodológica que podría influir en la estimación de los modelos de supervivencia. Por esta razón, se planteó un análisis de sensibilidad, cuyo propósito es evaluar la robustez de los resultados obtenidos frente a variaciones en la definición de este intervalo o fecha de censura.

Por lo tanto, este análisis consistió en modificar de manera sistemática la fecha de corte utilizada para determinar el estatus de censura de los usuarios. Para ello, se generaron diversas bases de datos en donde lo que se cambió fue la fecha de censura, esta se desplazó progresivamente hacia atrás en el tiempo. En particular, se utilizaron como puntos de referencia las siguientes fechas: 1 y 15 de junio, 1, 10 y 20 de julio, 1, 10 y 20 de agosto, y 1 de septiembre, todas del año 2024. Cada una de estas definiciones alternativas permite identificar distintos conjuntos de usuarios como censurados, y en consecuencia, recalcular los tiempos de supervivencia de acuerdo con la nueva fecha de corte.

Posteriormente, con cada una de estas bases de datos se volvieron a ajustar los modelos de supervivencia: tanto el modelo semi-paramétrico (modelo de Cox extendido con covariables dependientes del tiempo) como el modelo paramétrico con distribución Log-normal.

Finalmente, se compararon los coeficientes estimados y sus intervalos de confianza al 95% en todos los escenarios considerados. Esta estrategia metodológica permite identificar hasta qué punto los resultados y las conclusiones del estudio dependen de la decisión inicial de establecer un intervalo de 60 días, y si este criterio puede considerarse adecuado para representar de forma válida la condición de censura en los usuarios de la plataforma y sus respectivos tiempos de supervivencia.

Adicionalmente, se desarrolló un análisis complementario cuyo propósito fue evaluar en qué medida el criterio de censura adoptado inicialmente (60 días antes de la fecha de cierre de la información) puede llevar a clasificar de forma incorrecta a algunos usuarios. Para ello, se utilizó información adicional que incluye todas las solicitudes de empleo realizadas por los usuarios hasta el 15 de agosto de 2025, es decir, casi un año después de la fecha de corte original del estudio (1 de octubre de 2024), esta información corresponde a una nueva base de datos que se pudo obtener posteriormente al estudio original. Esta extensión temporal permite comprobar cuáles de los usuarios que fueron considerados como no censurados en el análisis original, es decir, aquellos que se supuso habían abandonado la plataforma, en realidad volvieron a interactuar con el sistema en fechas posteriores al cierre oficial de los datos. Dichos usuarios constituyen casos de "falsos abandonos", ya que fueron tratados como si hubieran concluido su tiempo de vida en la plataforma cuando, en la práctica, continuaron

utilizándola. Además, varios usuarios que fueron considerados originalmente como "censurados", podrían no haber seguido realizando solicitudes de trabajo, es decir, se deberían haber considerado como *churn* y no como censurados (falsos censurados).

Con esta información extendida se construyó un procedimiento de sensibilidad que consiste en redefinir progresivamente la fecha de censura, desplazándola hacia atrás desde el 1 de octubre de 2024 en intervalos de cinco días, hasta un máximo de 200 días. Cada valor de este desplazamiento define un "intervalo de censura" distinto (por ejemplo, 5, 10, 60, 150 o 200 días antes de la fecha final del estudio). Para cada escenario, primero se identifican los usuarios que, utilizando únicamente la información disponible hasta el 1 de octubre de 2024, serían clasificados como "no censurados" (*churn*) o como "censurados". Luego, con la información adicional (base de datos extendida: entre el 1 de octubre de 2024 y el 15 de agosto de 2025), se verifica si dichas clasificaciones fueron correctas o no.

De este modo, se cuantifican dos tipos de errores de clasificación. El primero corresponde a los usuarios que, bajo una determinada fecha de censura, son tratados como "no censurados" (abandonos) pero que, en el periodo posterior, vuelven a realizar al menos una solicitud en la plataforma; estos casos representan falsos abandonos o errores de clasificación como no censurados (error no-censurados). El segundo tipo corresponde a los usuarios que el modelo trata como "censurados" (se asume que siguen activos), pero que, considerando todo el horizonte de observación, no retoman su actividad y, por tanto, podrían haber sido clasificados como *churn*; estos casos representan falsos censurados o errores de sub–detección de *churn* (error censurados).

Con los resultados de este procedimiento se genera un gráfico de sensibilidad en el que, en el eje horizontal, se representa la distancia en días entre la fecha de censura considerada y la fecha final del estudio, y en el eje vertical se muestra el porcentaje de usuarios mal clasificados respecto del total de usuarios de la base original. El gráfico incluye dos curvas: una para el porcentaje de error asociado a los usuarios inicialmente clasificados como "no censurados" (falsos abandonos) y otra para el porcentaje de error correspondiente a los usuarios tratados como "censurados" pero que, a la luz de la información adicional, debieron haber sido considerados *churn* (falsos censurados). Este análisis permite evaluar la

sensibilidad del criterio de 60 días frente a cambios en la definición de censura y explorar empíricamente la existencia de un intervalo de censura que logre un compromiso razonable entre ambos tipos de error y el nivel de censura en los datos.

Con este análisis se pretende evaluar qué tan sensible es el criterio de 60 días frente a cambios en la definición de censura y determinar si este intervalo representa un balance adecuado entre minimizar errores de clasificación y evitar un exceso de censura; en consecuencia, el procedimiento permite explorar de forma empírica la existencia de un punto de equilibrio, es decir, un intervalo de censura lo suficientemente amplio como para reducir al mínimo los errores de clasificación, pero lo bastante corto como para no generar un exceso de censura.

3.7. CÁLCULOS DE TIEMPOS DE SUPERVIVENCIA

Una vez finalizado el ajuste y la evaluación de los distintos modelos de supervivencia (no paramétricos, semi-paramétricos y paramétricos), se procedió a generar estimaciones de tiempos de supervivencia asociados a distintas probabilidades de retención dentro de la plataforma digital.

Este procedimiento tuvo como objetivo complementar el análisis interpretativo de los modelos con una visión cuantitativa del tiempo esperado de uso, expresado en términos de percentiles de la función de supervivencia. En particular, se calcularon los tiempos correspondientes a las probabilidades de supervivencia del 75%, 50% (mediana), y 25%. Esto permite identificar, por ejemplo, cuántas semanas se espera que transcurran hasta que el 25%, 50% o 75% de los usuarios hayan abandonado la plataforma.

Para asegurar una comparación integral, estos tiempos fueron estimados a partir de cuatro enfoques distintos:

1. El modelo de Kaplan-Meier, como referencia no paramétrica.
2. El mejor modelo semi-paramétrico, el cual fue el modelo de Cox extendido, ya que presentó un buen ajuste y cumplimiento de supuestos.

3. Los dos mejores modelos paramétricos, seleccionados con base en criterios de información (AIC y BIC) y desempeño predictivo. Los cuales fueron el modelo Log-normal, y el modelo Gamma.

Las estimaciones se presentaron principalmente mediante representaciones gráficas comparativas, que permiten visualizar de forma más clara las diferencias entre modelos y la magnitud de la incertidumbre asociada a cada tiempo de supervivencia.

En particular, se generó un gráfico que muestra, para el "usuario promedio", los tiempos t_S y sus intervalos de confianza al 95% para un conjunto de probabilidades de supervivencia seleccionadas ($S = 0.25, 0.50$ y 0.75) y para cada uno de los cuatro modelos considerados. Se trata de un gráfico de puntos con barras de error que permite comparar de manera directa los valores puntuales de los tiempos predichos y la amplitud de sus intervalos de confianza, resaltando las discrepancias entre modelos en niveles específicos de S.

Sumado al anterior gráfico, se generó un segundo que presenta las curvas completas de supervivencia para los mismos cuatro modelos y para el mismo perfil de usuario. En este caso se representa $S(t)$ en todo el horizonte de seguimiento, junto con bandas de confianza, lo que facilita evaluar diferencias globales en la forma de las curvas, así como la coherencia relativa entre los enfoques paramétricos, semi-paramétrico y no paramétrico. La combinación de ambos gráficos ofrece una visión complementaria: el primero se centra en tiempos concretos de interés operativo, mientras que el segundo resume el comportamiento de la supervivencia a lo largo de todo el periodo de observación.

En todos los casos, el cálculo de los tiempos t_S siguió la misma lógica general. Primero, para cada modelo y perfil de usuario se obtuvo la curva de supervivencia estimada $\hat{S}(t \mid \mathrm{x})$. En el caso de Kaplan-Meier, esta curva se deriva directamente del estimador escalonado basado en los tiempos observados y los estados de censura. Para el modelo de Cox extendido, la curva se reconstruyó a partir de los incrementos estimados de la función de riesgo base y del efecto del perfil sobre el riesgo, de acuerdo con los coeficientes del modelo. En los modelos paramétricos Log-normal y Gamma, la supervivencia se calculó a partir de las expresiones cerradas de $S(t \mid \mathrm{x})$ propias de cada distribución.

Este mismo procedimiento se puede aplicar tanto al "usuario promedio" como a distintos perfiles definidos por combinaciones de *Región*, tipo de *Licencia*, *Edad* y condición de *Ingresos*, lo que permite explorar diferencias en los tiempos esperados de permanencia entre subpoblaciones de interés.

Además de las estimaciones puntuales, se calcularon intervalos de confianza del 95% para los tiempos t_S. Para el estimador de Kaplan-Meier, los intervalos se obtuvieron directamente con la función *survfit()* del paquete *survival*, que calcula el error estándar de $S(t)$ mediante la fórmula de Greenwood y construye los intervalos usando la transformación log–log. En el modelo de Cox extendido, los intervalos se obtuvieron mediante remuestreo paramétrico de los coeficientes del modelo a partir de su matriz de varianzas y covarianzas, reconstruyendo la supervivencia y los tiempos t_S en cada réplica y usando percentiles de la distribución empírica resultante. En el caso del modelo Gamma se aplicó un esquema análogo de bootstrap paramétrico, mientras que para el modelo Log-normal se utilizó la aproximación normal asintótica de los cuantiles, que puede interpretarse como un caso particular del mismo enfoque de remuestreo.

Finalmente, esta metodología se implementó en una aplicación *Shiny* (Chang, 2025; Sievert, 2020) (descrita en la sección 4.6 y en el Anexo C), que permite reproducir de forma interactiva estos cálculos para distintos perfiles de usuario. En dicha aplicación, la persona usuaria puede seleccionar uno o varios modelos (Gamma, Log-normal y Cox extendido), indicar los valores de S de interés y definir hasta tres perfiles de usuarios según *Región*, tipo de *Licencia*, *Edad* e *Ingresos* iniciales. La herramienta genera automáticamente un gráfico interactivo de puntos con barras de error para los tiempos t_S y ofrece la opción de descargar la tabla de resultados en formato "CSV", lo que facilita su uso operativo por parte del personal de la empresa y la integración de estas predicciones en procesos de análisis y toma de decisiones.

CAPÍTULO IV. RESULTADOS

En este capítulo se presentan los principales hallazgos derivados del análisis de los datos, organizados de acuerdo con las etapas metodológicas previamente descritas. En primer lugar, se realiza un análisis exploratorio de las variables incluidas en el estudio, con el propósito de caracterizar la muestra y describir las relaciones iniciales entre los factores de interés. Posteriormente, se exponen los resultados de los modelos de supervivencia ajustados, incluyendo tanto las estimaciones no paramétricas como los modelos semi-paramétricos y paramétricos, lo que permite contrastar y complementar los diferentes enfoques de modelación.

A lo largo del capítulo se ofrece una visión integrada del comportamiento de los tiempos de supervivencia en la población estudiada, destacando patrones relevantes y analizando la influencia de las variables independientes sobre el riesgo de abandono. Los resultados que se presentan en las siguientes secciones sirven de base para las conclusiones finales del presente trabajo.

4.1. ANÁLISIS EXPLORATORIO DE LOS DATOS

En esta sección se presentan los análisis descriptivos de las variables, su caracterización, y el análisis exploratorio de las relaciones entre ellas. Además, se incluye una exploración de los tiempos de supervivencia utilizando modelos de Kaplan-Meier.

Como se ha mencionado anteriormente, las variables independientes consideradas en este estudio son las siguientes: *Censura*, *Edad*, *Licencia*, *Región* e *Ingresos*. Por otro lado, la variable dependiente es el *Tiempo de Supervivencia*.

4.1.1. Análisis descriptivo de las variables

En cuanto a la variable *Censura,* la cual es dicotómica (unos para los usuarios no censurados y 0 para los usuarios censurados), se encontró un porcentaje de 71.01% de datos no

censurados y un 28.99% de censurados, donde el total de valores es de 48 437 usuarios de la plataforma.

Otra de las variables consideradas es la edad de los usuarios de la plataforma digital, el siguiente Cuadro 3 presenta un resumen estadístico descriptivo de esta variable *Edad*, diferenciando entre los grupos definidos por la variable de Censura (Censurado y No censurado). Se incluyen medidas de tendencia central (media y mediana), de dispersión (desviación estándar), así como los valores extremos (mínimo y máximo). En general, la variable *Edad* abarca un rango de entre 18 y 86 años, con una media global de 35.52 años y una mediana de 34 años, así como una desviación estándar de 10.55 años. Al desagregar según el estado de censura, se observa que la media de edad para los usuarios censurados es de 35.62 años y la mediana de 34 años; mientras que para los no censurados la mediana se mantiene en 34 años y la media es de 35.48 años, ligeramente menor a la categoría de censurados. Las diferencias en todas estas medidas estadísticas del Cuadro 3 son bastante pequeñas, lo que sugiere una distribución similar entre ambos grupos en términos de edad. Además, señalar que esta variable presenta un total de 327 valores faltantes, lo cual representa un 0.67% del total de los datos disponibles.

Cuadro 3.

Estadísticos descriptivos de la variable *Edad* según las categorías de Censura

Censurado	Edad				
	Media	Mediana	Desviación Est.	Mínimo	Máximo
Censurado	35.62	34	10.58	18	80
No censurado	35.48	34	10.53	18	86
Para el Total	35.52	34	10.55	18	86

Fuente: Plataforma digital anónima (2024).

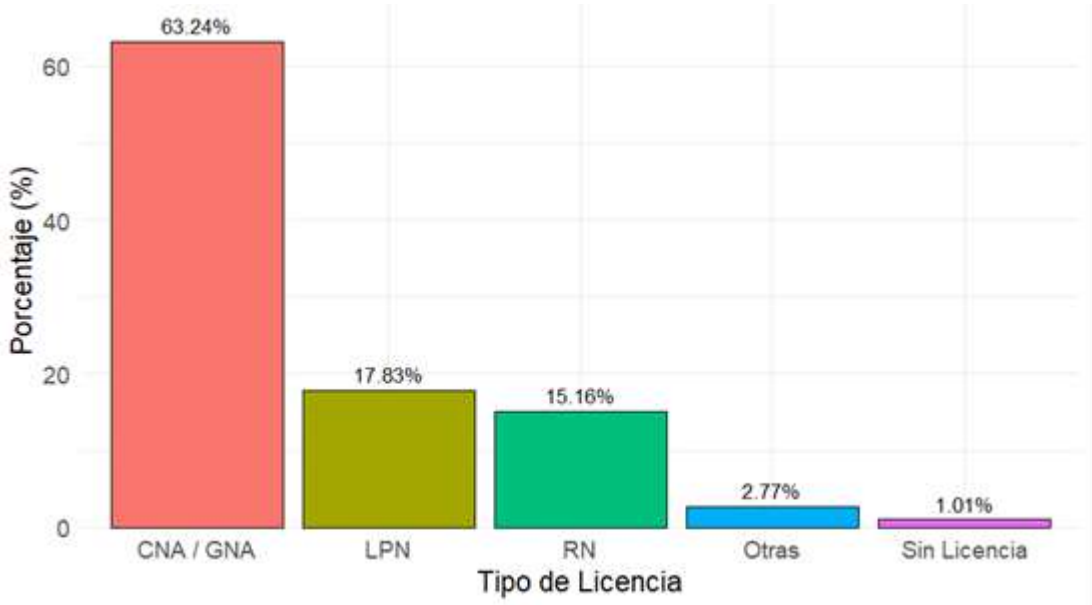
Por su parte, la variable *Licencias* se encuentra dividida en cinco categorías principales, cuya distribución porcentual se detalla en el Gráfico 1. Este gráfico presenta de manera clara la proporción de usuarios que poseen cada tipo de licencia de enfermería dentro de la muestra analizada.

De forma similar, el Cuadro 4 muestra la distribución proporcional de la variable *Censura* en función de las categorías de la variable *Licencia*. Los valores presentados representan proporciones y reflejan la relación entre el estado de censura y los diferentes tipos de licencias de los usuarios de la plataforma.

En dicho Cuadro 4 se observa que, la categoría "LPN" presenta la mayor proporción de censurados, con un 33%, seguida de "CNA / GNA" y "RN", con valores de 29% y 27%, respectivamente. Por otro lado, la categoría "Sin licencia" tiene la menor proporción de censurados, con apenas un 11%, indicando una diferencia notable respecto a las demás categorías. Se revela que las diferencias en las proporciones de usuarios censurados entre las categorías de *licencias* son relativamente pequeñas, con excepción de la categoría "Sin licencia". Este resultado sugiere que dicha categoría se comporta de manera distintiva en relación con la censura, lo cual podría estar relacionado con características específicas de este grupo.

Gráfico 1.

Distribución porcentual de las categorías de la variable Licencias



Total de datos: 48.437

Fuente: Plataforma digital anónima (2024).

Cuadro 4.

Porcentaje de Censura según categorías de *Licencia*

Licencias	Porcentaje de censura (%)
CNA / GNA	29
RN	27
LPN	33
Sin licencia	11
Otras	23

Fuente: Plataforma digital anónima. (2024).

Con respecto a la relación entre las edades y las categorías de *licencias*, en el Cuadro 5 se muestran los estadísticos descriptivos de la variable *Edad* según las distintas categorías de la variable *Licencias*. Adicionalmente, el Gráfico 2 resume visualmente las medias de edad por cada categoría de licencia. Se aprecia que los usuarios con licencias que requieren mayor formación académica y experiencia profesional, como "RN" y "LPN", presentan las edades promedio más elevadas (39.12 y 39.06 años, respectivamente). En contraste, los usuarios con licencias de menor nivel, como "CNA" o aquellos clasificados en la categoría "Otras", así como quienes no cuentan con licencia, presentan edades promedio más bajas (33.71, 35.94 y 35.84 años, respectivamente).

Aunque se observan diferencias entre las categorías de *licencia* en todos los estadísticos descriptivos de la variable Edad, dichas diferencias son relativamente pequeñas en términos absolutos para cada una de las medidas analizadas (media, mediana, desviación estándar, mínimo y máximo). No obstante, estas pequeñas variaciones son coherentes con el nivel de preparación y responsabilidad asociado a cada tipo de licencia, ya que a mayor formación o experiencia requerida, mayor tiende a ser la edad de los profesionales registrados en la plataforma.

Cuadro 5.

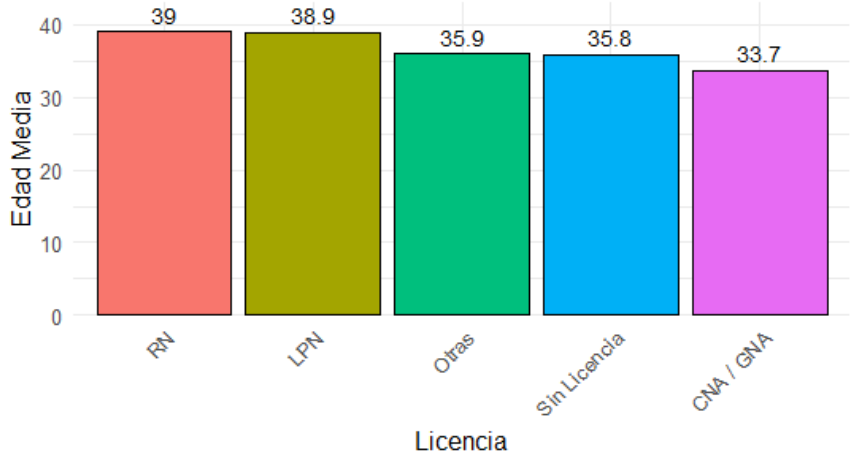
Estadísticos descriptivos de la variable *Edad* según las categorías de *Licencias*

Licencia	Edad				
	Media	Mediana	Desviación Est.	Mínimo	Máximo
CNA / GNA	33.71	32	10.44	18	80
LPN	38.92	38	9.96	19	86
Otras	35.94	34	10.80	18	72
RN	39.00	38	9.72	20	85
Sin licencia	35.84	34	10.53	20	79

Fuente: Plataforma digital anónima (2024).

Gráfico 2.

Media de la *Edad* según *Licencia* de Enfermería



Fuente: Plataforma digital anónima (2024).

En relación con la variable Región, el Gráfico 3 presenta la distribución porcentual de la muestra de usuarios según las distintas categorías geográficas consideradas en el estudio. Se observa que la región de "Rockies & Red River" concentra la mayor proporción de casos (28.97%), mientras que la región "Southeast" presenta la participación más baja en la muestra (6.99%). Asimismo, el Cuadro 6 muestra el porcentaje de censura para cada una de las diferentes categorías de la variable *Región*, y revela diferencias notables en las proporciones

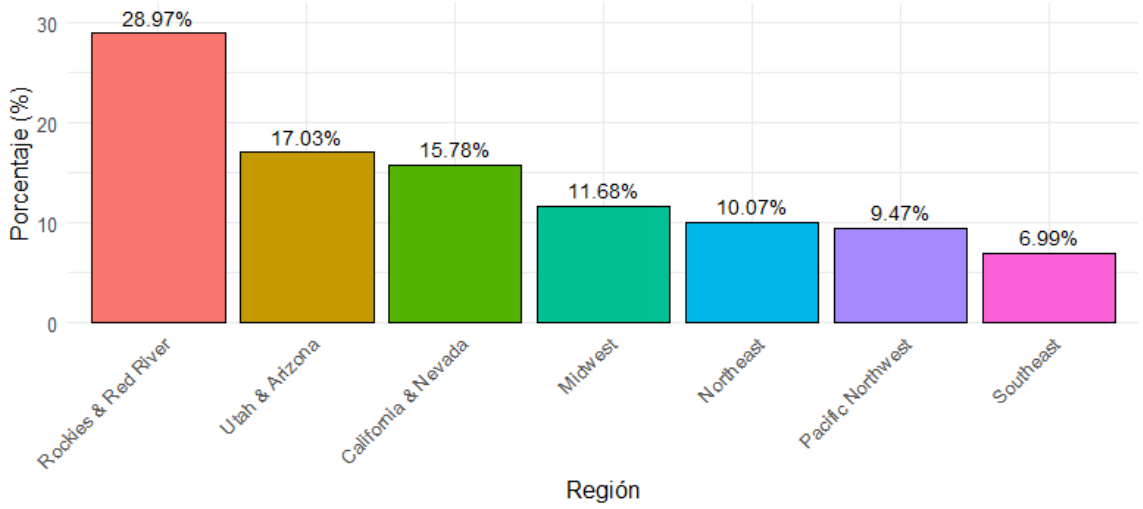
de usuarios censurados entre las distintas regiones. Donde, la categoría "Pacific Northwest" presenta la mayor proporción de usuarios censurados, con un 41.0%, seguida por "California & Nevada", con un 38.9%. Estas regiones destacan por tener un porcentaje de censura significativamente mayor en comparación con otras.

Por otro lado, la región "Rockies & Red River" muestra la menor proporción de usuarios censurados, con un 19.7%, lo que representa una diferencia considerable frente a las categorías con mayores proporciones de censura. Otras regiones con proporciones relativamente bajas son "Utah & Arizona" (20.9%) y "Midwest" (29.2%).

Para esta variable, las diferencias en la proporción de usuarios censurados entre regiones son apreciables, lo que sugiere que el patrón de censura no es homogéneo en el territorio y podría estar asociado a factores contextuales propios de cada área geográfica que conviene considerar.

Gráfico 3.

Distribución porcentual de las categorías de la variable *Región*



Fuente: Plataforma digital anónima. (2024).

Cuadro 6.

Porcentaje de Censura según categorías de *Región*

Región	Porcentaje de censura (%)
Utah & Arizona	20.90
California & Nevada	38.90
Midwest	29.20
Northeast	36.80
Pacific Northwest	41.00
Rockies & Red River	19.70
Southeast	36.80

Fuente: Plataforma digital anónima. (2024).

El Cuadro 7 presenta la distribución porcentual de las categorías de *Licencias* según las distintas regiones consideradas en el estudio. En general, las regiones muestran una composición relativamente similar, con predominio de la licencia "CNA/GNA". No obstante, se observan algunas diferencias puntuales: "Utah & Arizona" destaca por una mayor proporción de "RN" y "Otras", además de una menor participación de "LPN", mientras que "California & Nevada" concentra una mayor proporción de "CNA/GNA" y una menor de "RN". En conjunto, estas variaciones son moderadas y coherentes con la distribución global observada en el Gráfico 1.

Cuadro 7.

Distribución porcentual de *Licencias* según las categorías de *Región*

Región	Licencias				
	CNA / GNA	RN	LPN	Sin licencia	Otras
Utah & Arizona	55.97	23.72	9.78	1.28	9.25
California & Nevada	69.63	10.26	19.33	0.73	0.05
Midwest	64.58	15.06	19.28	0.80	0.28
Northeast	62.55	14.90	20.93	0.78	0.84
Pacific Northwest	66.88	13.80	17.92	1.20	0.20
Rockies & Red River	64.66	11.59	19.35	0.90	3.51
Southeast	58.02	20.53	19.68	1.62	0.15

Los valores presentados en este cuadro son porcentajes (%)

Fuente: Plataforma digital anónima. (2024).

Asimismo, el Cuadro 8 presenta los estadísticos descriptivos de la variable *Edad* según las diferentes categorías de la variable *Región*. Permitiendo identificar variaciones en la distribución de la edad entre las regiones. Se aprecia como las diferencias de la variable edad según las regiones son pequeñas, por ejemplo, las medias y medianas son muy similares para cada una de las regiones. La región "Rockies & Red River" presenta los valores de media y mediana más altos (36.58 años y 35 años, respectivamente), indicando que los usuarios de esta región tienden a ser mayores en comparación con otras. Por otro lado, la región "Midwest" tiene los valores más bajos en la media (34.66 años) y mediana (33 años), lo que sugiere que sus usuarios tienden a ser ligeramente más jóvenes.

Cuadro 8.

Estadísticos descriptivos de la variable *Edad* según las categorías de *Región*

Región	Edad				
	Media	Mediana	Desviación Est.	Mínimo	Máximo
Utah & Arizona	34.78	33	11.57	17	78
California & Nevada	35.55	34	10.80	18	76
Midwest	34.66	33	9.73	17	69
Northeast	36.04	35	10.02	18	76
Pacific Northwest	35.23	34	10.40	18	85
Rockies & Red River	36.58	35	10.34	15	86
Southeast	36.11	35	9.72	18	73

Fuente: Plataforma digital anónima. (2024).

En cuanto a la variable de *Ingresos*, se encontró una proporción de 68.2% de los usuarios no generaron ingresos en sus primeros 2 meses realizando solicitudes de trabajo, mientras que 31.7% del total de usuarios analizados sí generaron ingresos en sus primeros meses utilizando la plataforma.

El Cuadro 9 presenta los estadísticos descriptivos de la variable *Edad* según las categorías de la variable *Ingresos*. Se observa que las diferencias en la variable *Edad* entre los usuarios

"con ingresos" y aquellos "sin ingresos" son muy pequeñas. No obstante, los usuarios "con ingresos" presentan una edad promedio ligeramente mayor en comparación con los de la categoría "sin ingresos".

Cuadro 9.

Estadísticos descriptivos de la variable *Edad* según las categorías de *Ingresos*

Ingresos	Edad				
	Media	Mediana	Desviación Est.	Mínimo	Máximo
Con ingresos	36.18	35	10.90	18	78
Sin ingresos	35.21	34	10.37	18	86

Fuente: Plataforma digital anónima. (2024).

Por su parte, el Cuadro 10 presenta el porcentaje de censura para las categorías de la variable *Ingresos*. Del cuadro se desprende que los usuarios "sin ingresos" tienen una proporción de censura del 24.70%, considerablemente menor que el 38.10% observado en los usuarios "con ingresos". Esto indica que los usuarios que generan ingresos tienen una mayor probabilidad de estar censurados, lo que refleja diferencias en los patrones de uso o en la duración de actividad en la plataforma entre ambos grupos. Siendo que, se espera que los usuarios "con ingresos" continúen utilizando la plataforma por periodos de tiempo mayores a su contraparte.

Cuadro 10.

Porcentaje de Censura según categorías de *Ingresos*

Ingresos	Porcentaje de censura (%)
Sin ingresos	24.70
Con ingresos	38.10

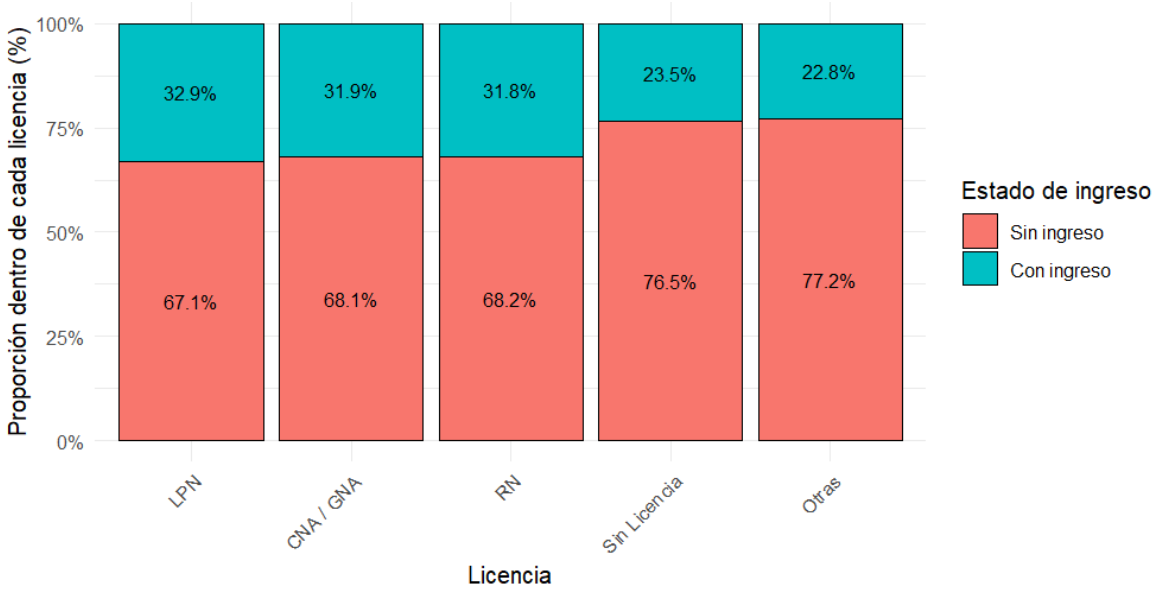
Fuente: Plataforma digital anónima. (2024).

Complementariamente, el Gráfico 4 muestra la distribución proporcional de ingresos generados por los usuarios según las categorías de la variable *Licencias*. En el Gráfico se

observa que los usuarios con licencias "LPN" presentan la proporción más alta de usuarios "con ingresos" generados (32,90%) en comparación con otras categorías, mientras que las licencias "CNA / GNA" y "RN" tienen una proporción ligeramente menor (31,90%). Por otro lado, los usuarios "Sin licencia" y aquellos en la categoría de "Otras" licencias muestran la proporción más baja de usuarios que generaron ingresos en sus primeros dos meses utilizando la plataforma (23,50% y 22,80% respectivamente). Estas diferencias sugieren que el tipo de licencia puede estar relacionado con la probabilidad de generar ingresos, aunque las diferencias entre las categorías son moderadas.

Gráfico 4.

Distribución porcentual de *Ingresos* según las categorías de *Licencias*



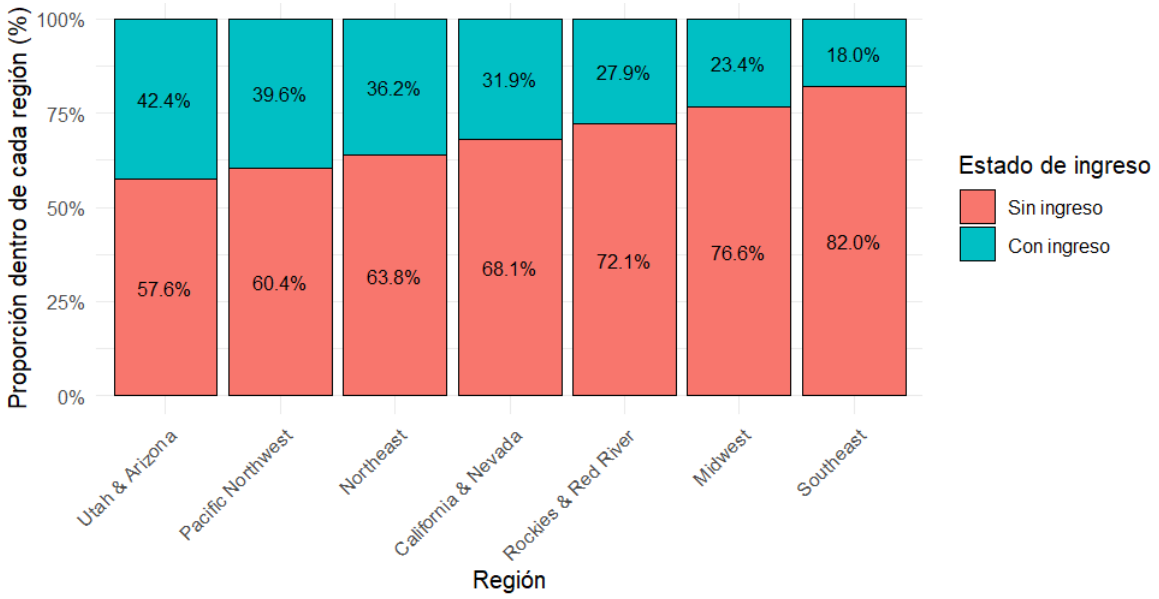
Fuente: Plataforma digital anónima. (2024).

De forma similar, el siguiente Gráfico 5 presenta la distribución proporcional de ingresos generados por los usuarios de la plataforma, desglosada según las diferentes regiones de Estados Unidos. De acuerdo con el gráfico, las regiones de "Utah & Arizona" y "Pacific Northwest" presentan las mayores proporciones de usuarios "con ingresos" generados (42,40% y 39,60%, respectivamente). Estas diferencias que se aprecian en el gráfico podrían

reflejar factores regionales que afectan la probabilidad de generar ingresos a través de la plataforma, donde los contrastes entre algunas regiones son considerablemente marcados.

Gráfico 5.

Distribución porcentual de *Ingresos* según las categorías de *Región*



Fuente: Plataforma digital anónima. (2024).

4.1.2. Asociaciones entre variables independientes

Con el objetivo de explorar posibles asociaciones entre las variables categóricas incluidas en el estudio, se realizaron pruebas estadísticas de independencia. El Cuadro 11 resume los resultados obtenidos, los cuales se fundamentan en el estadístico chi-cuadrado (X^2) y el Coeficiente *V* de Cramer. Estas pruebas permiten identificar relaciones significativas entre las variables, así como estimar la magnitud de dichas asociaciones. En particular, el valor-p asociado al chi-cuadrado permite evaluar la significancia estadística de cada relación, mientras que el coeficiente *V* de Cramer proporciona una medida estandarizada de la fuerza de asociación, facilitando su interpretación comparativa. Los resultados presentados en el Cuadro 11 contribuyen a identificar posibles patrones de dependencia entre las covariables,

lo cual resulta relevante para la correcta especificación de los modelos de supervivencia en etapas posteriores del análisis.

Cuadro 11.

Pruebas chi-cuadrado y coeficientes *V* de Cramer para evaluar la asociación entre variables categóricas

Variable 1	Variable 2	Estadístico (X^2)	Grados de libertad	Valor-p	*V* de Cramer
Ingresos	Censura	911.2	1	< 0,0001	0.137
Región	Censura	1783.1	6	< 0,0001	0.192
Región	Ingresos	1178.7	6	< 0,0001	0.156
Licencias	Región	3147.5	24	< 0,0001	0.127
Licencias	Censura	156.1	4	< 0,0001	0.057
Licencias	Ingresos	70.4	4	< 0,0001	0.038

V de Cramer: Coeficiente *V* de Cramer.

Estadístico de prueba (X^2): estadístico chi-cuadrado.

Fuente: Plataforma digital anónima. (2024).

Los resultados presentes en el Cuadro 11 indican que todas las relaciones analizadas son estadísticamente significativas ($p<0.0001$), lo que sugiere que existe dependencia entre las variables. Sin embargo, la magnitud de estas asociaciones varía considerablemente, como se refleja en los valores del *V* de Cramer.

La relación más fuerte se observa entre las variables Región y la de Censura, con un *V* de Cramer de 0.192, lo que indica una asociación moderada. Esto implica que la condición de censura está asociada con la región del profesional de enfermería. Por otro lado, la relación entre *Región* e *Ingresos* presenta un *V* de Cramer de 0.156, que también sugiere una asociación significativa aunque un poco menos marcada. Asimismo, las relaciones entre *Ingresos* y *Censura* (V=0.137) y entre *Licencias* y *Región* (V=0.127) muestran asociaciones débiles, pero no despreciables, lo que refleja que la distribución regional tiene pequeños impactos en otras variables.

Finalmente, las relaciones más débiles se observan entre *Licencias* y *Censura* (V=0.057) y entre *Licencias* e *Ingresos* (V=0.038). Aunque estas asociaciones son estadísticamente significativas, la magnitud de la dependencia entre estas variables es muy poca, indicando que las categorías de licencias tienen poca influencia en los niveles de censura y en si el usuario genera o no genera ingresos con la ayuda del servicio.

De forma similar, en el Cuadro 12 se presentan los resultados de las pruebas ANOVA realizadas para analizar las diferencias en las medias de la variable Edad según las categorías de las variables independientes *Región*, *Licencia*, *Ingresos* y *Censura*. Este análisis busca determinar si existen diferencias significativas en la media de edad en función de estas otras variables. Los resultados incluyen el estadístico F, los grados de libertad (G.L.) y el valor-p para evaluar la significancia estadística de las diferencias.

Cuadro 12.

Comparación de medias de la variable *Edad* según las covariables categóricas mediante ANOVA

Variable 1	Variable 2	Estadístico (F)	Grados de libertad	Valor-p	eta
Edad	Censura	45.88	1	< 0.0001	0.0317
Edad	Región	30.61	6	< 0.0001	0.0634
Edad	Licencia	642.8	4	< 0.0001	0.2312
Edad	Ingresos	135.9	1	< 0.0001	0.0545

Fuente: Plataforma digital anónima. (2024).

El análisis ANOVA mostrado en el Cuadro 12 muestra que todas las variables consideradas (*Región*, *Licencia*, *Ingresos* y *Censura*) están asociadas con las medias de edad (valores $p < 0.001$ en todos los casos). No obstante, como se vio en los Cuadros 3, 5, 8, y 9, y en el Gráfico 2, no parecen hacer diferencias importantes en el comportamiento de las edades de los usuarios según cada una de las categorías presentes en cada una de las variables categóricas.

Más adelante, al generar los modelos semi-paramétricos de análisis de supervivencia (Sección 4.3.1.), se presenta la prueba del Factor de Inflación de la Varianza Generalizado

(GVIF, por sus siglas en inglés), con el fin de evaluar la posible existencia de colinealidad entre las covariables incluidas. Esta verificación permitirá asegurar la estabilidad de las estimaciones y la validez de las inferencias realizadas a partir de los modelos ajustados.

4.1.3. Análisis descriptivo de la variable dependiente: *tiempo de supervivencia*

En esta sección se analiza el comportamiento de uso y abandono del servicio por parte de los usuarios en la Plataforma digital, tomando como variable dependiente el Tiempo de Supervivencia. A través de cuadros y gráficos descriptivos, se explora la distribución general del tiempo hasta el abandono del servicio. Inicialmente, se presentan histogramas para visualizar los patrones generales, seguidos por las curvas de supervivencia Kaplan-Meier, donde también se incluyeron las variables independientes como las *licencias*, *regiones* geográficas, *edades*, y la variable de *ingresos*. Este enfoque ofrece una perspectiva sobre cómo estas características influyen en los patrones de uso y abandono, proporcionando una visión clara y descriptiva de la relación entre el tiempo de supervivencia y las variables estudiadas.

El Gráfico 6 presenta la distribución del tiempo de supervivencia, medido en semanas, de la muestra de usuarios de la Plataforma digital, mediante histogramas. Dicho gráfico se divide en tres subgráficos: El Gráfico 6-A, que incluye a todos los usuarios, el Gráfico 6-B para únicamente los usuarios con datos no censurados, y un tercer gráfico, Gráfico 6-C, para únicamente los usuarios con datos censurados.

En el gráfico general, Gráfico 6-A, que incluye a todos los usuarios, se observa que la mayoría de los individuos tienen tiempos de supervivencia cortos, lo cual se refleja en una alta densidad en las primeras semanas, la cual decrece rápidamente conforme aumenta el tiempo. Este comportamiento sugiere que gran parte de los usuarios abandonan la plataforma en las primeras semanas o meses, indicando un patrón de alta deserción temprana.

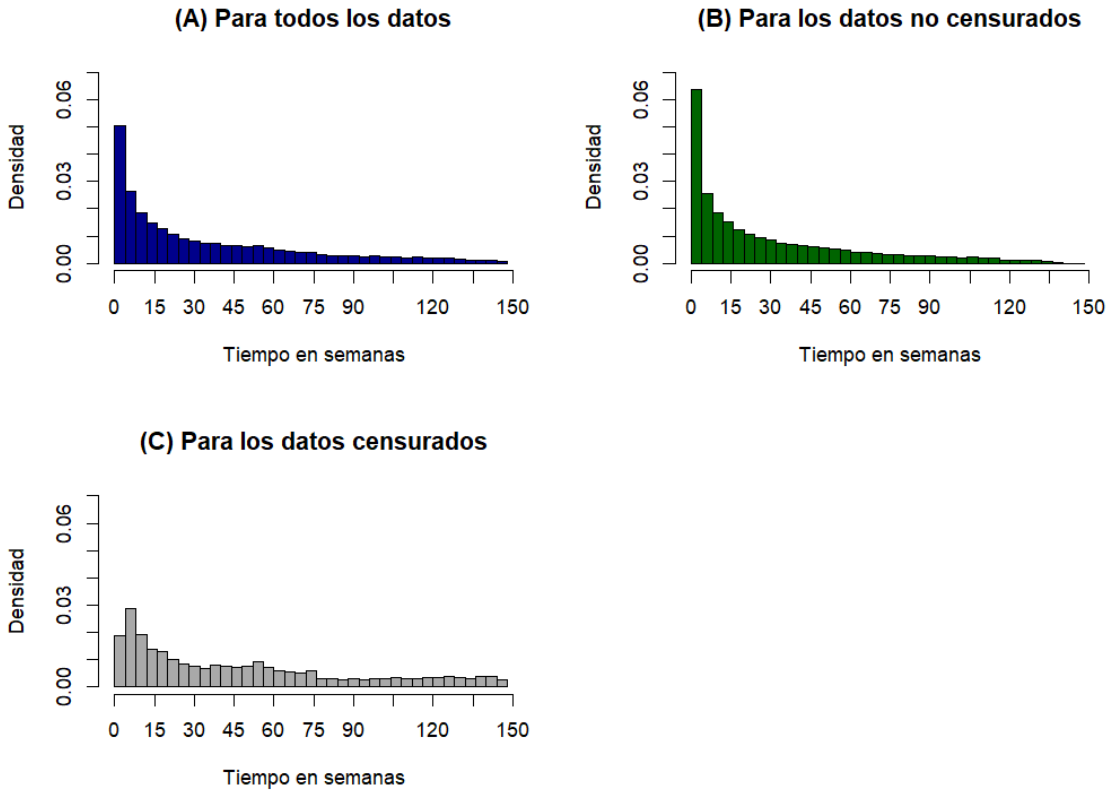
Cuando se analizan únicamente los datos no censurados, Gráfico 6-B, el patrón se mantiene, aunque la densidad en los tiempos más largos es aún menor. Esto indica que, dentro de este grupo, los usuarios que efectivamente abandonaron tienden a hacerlo en un plazo más corto,

reforzando la conclusión de una menor retención a largo plazo entre los usuarios no censurados. Por otro lado, el análisis de los datos censurados, Gráfico 6-C, revela una distribución más uniforme en los tiempos intermedios y largos. Este comportamiento sugiere que los usuarios censurados, aquellos cuyo tiempo de supervivencia no concluyó dentro del periodo de observación, presentan una mayor probabilidad de permanecer activos en la plataforma por un periodo más prolongado.

En conjunto, estos resultados reflejan diferencias importantes en el comportamiento entre usuarios censurados y no censurados, así como un patrón general de abandono temprano en la plataforma.

Gráfico 6.

Distribución del tiempo de supervivencia de la muestra de usuarios de la Plataforma digital mediante histogramas



Fuente: Plataforma digital anónima. (2024).

Esta sección se ve complementada con los análisis posteriores con los distintos modelos de supervivencia, por ejemplo, la sección 4.2. Modelo no-paramétrico (Modelo Kaplan-Meier) brinda una gran aproximación exploratoria a la variable tiempo y su relación con las distintas variables independientes de este estudio.

4.1.4. Valores extremos

En esta sección se presentan los resultados del análisis de valores extremos (VE) identificados a partir del modelo extendido de Cox, utilizando los residuos de deviancia como criterio diagnóstico. Tal como se indicó en la sección de Metodología, estos residuos permiten detectar observaciones atípicas individuales de forma más eficaz que otros enfoques tradicionales, especialmente en contextos gráficos (Therneau & Grambsch, 2000). Para ello, se clasificaron como VE aquellos casos con residuos mayores a 2 o menores a -2. A continuación, se describe la proporción de casos detectados, así como una caracterización comparativa de los VE en relación con la muestra general, con el fin de explorar patrones particulares que puedan incidir en la estabilidad o interpretación de los modelos de supervivencia.

Es importante destacar que estos residuos de deviancia fueron calculados a partir de una base de datos transformada al formato denominado "semana-persona". Este tipo de estructura se genera mediante una segmentación del tiempo de seguimiento de cada individuo en intervalos discretos (en este caso, semanas), de forma que cada fila de la base representa una observación para una persona en una semana específica de su seguimiento. Este enfoque se suele utilizar cuando se aplican modelos de supervivencia con covariables dependientes del tiempo, ya que permite capturar adecuadamente los cambios en la exposición o en las covariables a lo largo del tiempo. La transformación fue realizada utilizando la función *survSplit()* del paquete *survival* en R (Therneau, 2024), dividiendo los tiempos de supervivencia en múltiples registros por individuo. Así, un usuario con un tiempo de supervivencia de, por ejemplo, 10 semanas, genera 10 filas en esta base, lo que permite modelar de forma más flexible las asociaciones entre las covariables y el tiempo por medio

del modelo Cox extendido. En este contexto, los residuos calculados reflejan el ajuste del modelo para cada segmento temporal de cada individuo.

El Cuadro 13 presenta un resumen estadístico de los residuos de deviancia generados a partir del modelo Cox extendido ajustado sobre la base de datos en formato semana-persona. En total se obtuvieron 1 703 578 valores residuales, correspondientes a las observaciones semanales de todos los usuarios en el estudio. A partir de estos residuos, se identificaron como valores extremos (VE) aquellos cuya magnitud superó el umbral absoluto de 2, dando como resultado un total de 32 709 observaciones clasificadas como VE, lo cual representa un 1.92% del total.

Como se puede observar en el Cuadro 13, la distribución de los residuos presenta una media de -0.137 y una mediana ligeramente menor (-0.180), lo que sugiere una ligera asimetría hacia valores negativos. Por otro lado, el rango abarca desde -0.723 hasta 3.114, con una desviación estándar de 0.399. Destaca el hecho de que el valor máximo supera el umbral superior de 2, lo que valida la presencia de residuos atípicos positivos.

Cuadro 13.

Residuos de deviancia del modelo extendido de Cox

Media	Mediana	Desviación Est.	Mínimo	Máximo
-0.137	-0.180	0.399	-0.723	3.114

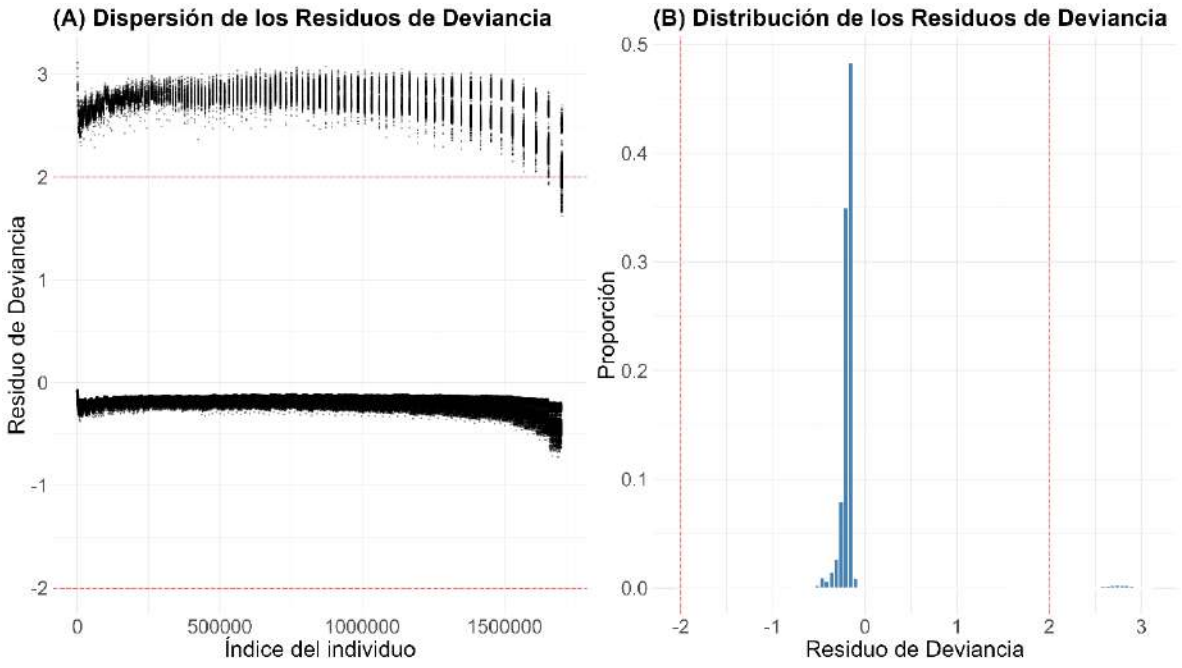
Por su parte, el Gráfico 7 presenta una visualización diagnóstica de los residuos de deviancia obtenidos a partir del modelo Cox extendido ajustado con datos en formato semana-persona. Este tipo de análisis gráfico permite detectar observaciones atípicas que podrían influir en la estabilidad del modelo, complementando el análisis estadístico mostrado en el Cuadro 13. En la parte izquierda (Gráfico 7.A) se muestra la dispersión de los residuos de deviancia. La mayoría de los residuos se concentran entre los valores de -1 y 1, mientras que un subconjunto reducido supera el umbral de 2, indicado mediante líneas punteadas rojas, lo cual es indicativo de un grupo de valores extremos.

De forma complementaria, el histograma de la derecha (Gráfico 7.B) muestra la distribución agregada de los residuos. Se observa una forma simétrica y centrada cerca de cero, lo cual es coherente con los supuestos del modelo, aunque, de igual forma que con el Gráfico 7.A, se identifican algunos residuos de gran magnitud en la cola derecha. Esta visualización también respalda la identificación de un grupo de valores extremos con valores mayores a 2.

En el caso de los modelos de supervivencia, los residuos de deviancia permiten identificar discrepancias entre el tiempo observado de ocurrencia del evento y el tiempo estimado por el modelo. En particular, residuos de deviancia positivos indican que el evento ocurrió antes de lo que el modelo predecía como más probable, lo que significa que el individuo presentó una menor supervivencia de la esperada. La presencia de valores extremos en estos residuos sugiere la existencia de casos atípicos, cuya supervivencia difiere de manera considerable de lo estimado.

Gráfico 7.

Residuos de Deviancia del modelo extendido de Cox



Las líneas rojas punteadas son indicadores de los valores ±2

Por su parte, el Cuadro 14 presenta la distribución de usuarios censurados según el grupo de análisis considerado: la muestra completa original, la base expandida tipo semana-persona, y el subconjunto de usuarios clasificados como valores extremos (VE), identificados con base en los residuos de deviancia del modelo extendido de Cox.

Cuadro 14.

Porcentaje de usuarios en los grupos de *censura* según el grupo de datos, toda la muestra personas, toda la muestra semana-persona, y valores extremos

	Cantidad total de datos	Cantidad de valores No Censurados	Porcentaje de valores No Censurados
Toda la Muestra (personas)	48 437	34 397	71.01%
Toda la Muestra (semana-personas)	1 705 528	34 397	2.02%
Valores Extremos	32 709	32 709	100%

Como se puede observar, el 100% de los casos clasificados como valores extremos corresponden a usuarios no censurados, es decir, aquellos para los cuales se observó el evento de interés (en este caso, la salida de la plataforma). En total, se identificaron 32 709 valores extremos, lo cual representa un 1.92% del total de observaciones utilizadas para el cálculo de los residuos (1 703 578 observaciones en la base tipo semana-persona).

Por su parte, en la muestra completa tipo semana-persona, únicamente un 2.02% de los registros corresponden a observaciones no censuradas. Esta baja proporción se debe a la propia estructura de este tipo de base, en la que cada fila representa una unidad de tiempo (semana) en la que el usuario aún se encuentra activo, lo que implica que la mayoría de registros reflejan periodos previos a la ocurrencia del evento (casos censurados).

En contraste, la muestra original incluye 48 437 individuos, de los cuales 34 397 (71.01%) no están censurados. Esto significa que, si bien los valores extremos representan una pequeña fracción del total de observaciones en la base semana-persona, concentran un número importante de usuarios no censurados.

Cabe destacar que cada uno de los 32 709 valores extremos corresponde al último registro (fila) de un individuo no censurado, es decir, al momento en que se observa su salida de la plataforma. Estos registros representan el intervalo final de seguimiento para cada usuario, definido por dos semanas consecutivas: la penúltima semana de exposición y la semana en la que ocurre el evento. Por ejemplo, para un usuario cuya duración total fue de 9 semanas, el valor extremo se asocia al intervalo que abarca desde la octava hasta la novena semana.

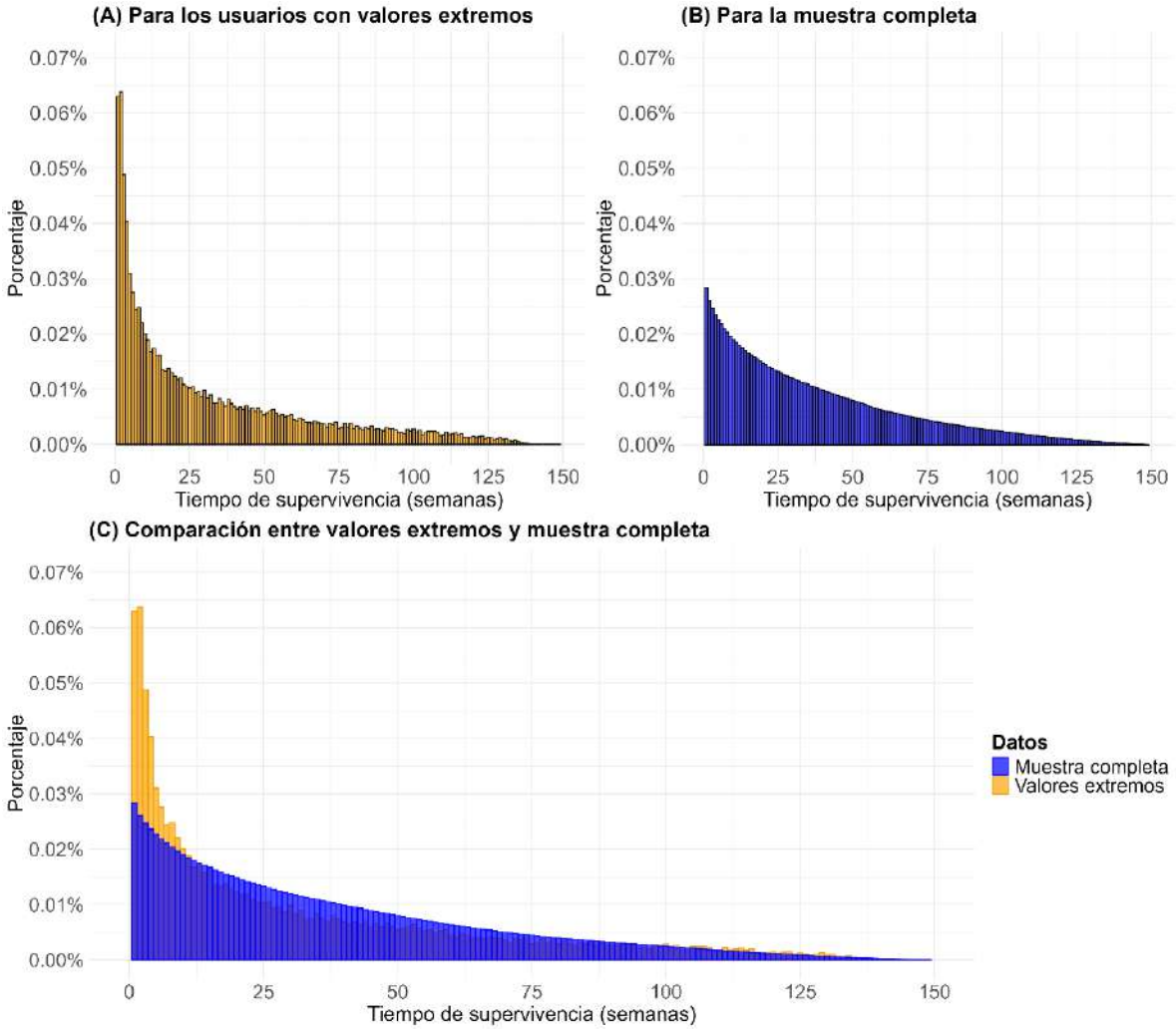
Una excepción notable está constituida por 1 688 usuarios no censurados que no forman parte de los valores extremos, cuyos tiempos de abandono fueron particularmente breves: 1 680 de ellos dejaron la plataforma tras una sola semana, y 8 usuarios después de dos semanas. A pesar de su corta duración, estos individuos presentan residuos de deviancia positivos pero menores a 2.

A continuación, se presenta el Gráfico 8. el cual muestra las distribuciones del tiempo de supervivencia en semanas (la duración entre la creación de los usuarios y el momento de falla de los individuos), tanto para la muestra completa de usuarios como para el subconjunto de observaciones clasificadas como valores extremos (VE) según los residuos de deviancia del modelo extendido de Cox. Se compara la distribución del tiempo de supervivencia entre ambas poblaciones, donde los valores provenientes de la muestra completa semana-persona se representan en azul, y los VE en anaranjado.

En el panel (A) del Gráfico 8, correspondiente a los VE, se observa una fuerte concentración de observaciones en las primeras semanas de seguimiento, lo que indica que una proporción considerable de usuarios abandonó la plataforma muy pronto tras su ingreso. Esta acumulación en valores bajos del tiempo de supervivencia es un hallazgo relevante, ya que sugiere que los VE están asociados a salidas tempranas que resultan inesperadas para el modelo.

Gráfico 8.

Distribución comparativa del tiempo de supervivencia entre la muestra completa semana-persona y los valores extremos



En contraste, la distribución del panel (B) del Gráfico 8, que representa a la muestra completa, muestra una mayor dispersión de los tiempos de supervivencia a lo largo del eje temporal, con una densidad más suavemente decreciente. Esto se debe a la estructura de la base tipo semana-persona, en la cual cada fila representa una semana en la que el usuario aún se encuentra activo. Por tanto, la mayoría de los registros corresponden a periodos previos a la ocurrencia del evento, lo que explica la forma más aplanada y extendida de la distribución.

Una diferencia fundamental entre ambos grupos es que, para los VE, cada uno de los 32 709 registros corresponde al último intervalo observado para un usuario no censurado, es decir, al momento preciso de su salida de la plataforma. Esta característica estructural explica por qué su distribución se limita a un único registro por individuo y se concentra en los momentos cercanos a la falla. Además, tiene sentido que existan varios valores de permanencia breve de los usuarios, ya que estos usuarios presentan VE positivos, lo que indica que, desde la perspectiva del modelo, su salida ocurrió antes de lo esperado. En términos estadísticos, un residuo de deviancia positivo refleja una discrepancia en la cual el tiempo observado hasta el evento fue menor al tiempo que el modelo estimaba como más probable. Es decir, estos casos de abandono ocurrieron antes de lo que el modelo esperaba.

Para profundizar en la caracterización del grupo de usuarios clasificados como valores extremos (VE), se presenta el Gráfico 9, el cual muestra la distribución de las fechas de creación de los usuarios, comparando a aquellos con VE frente al total de la muestra. La muestra completa incluye un total de 48 437 usuarios, de los cuales 32 709 (67.5%) han sido clasificados como casos que presentan al menos un residuo de deviancia considerado extremo, según el modelo de Cox extendido.

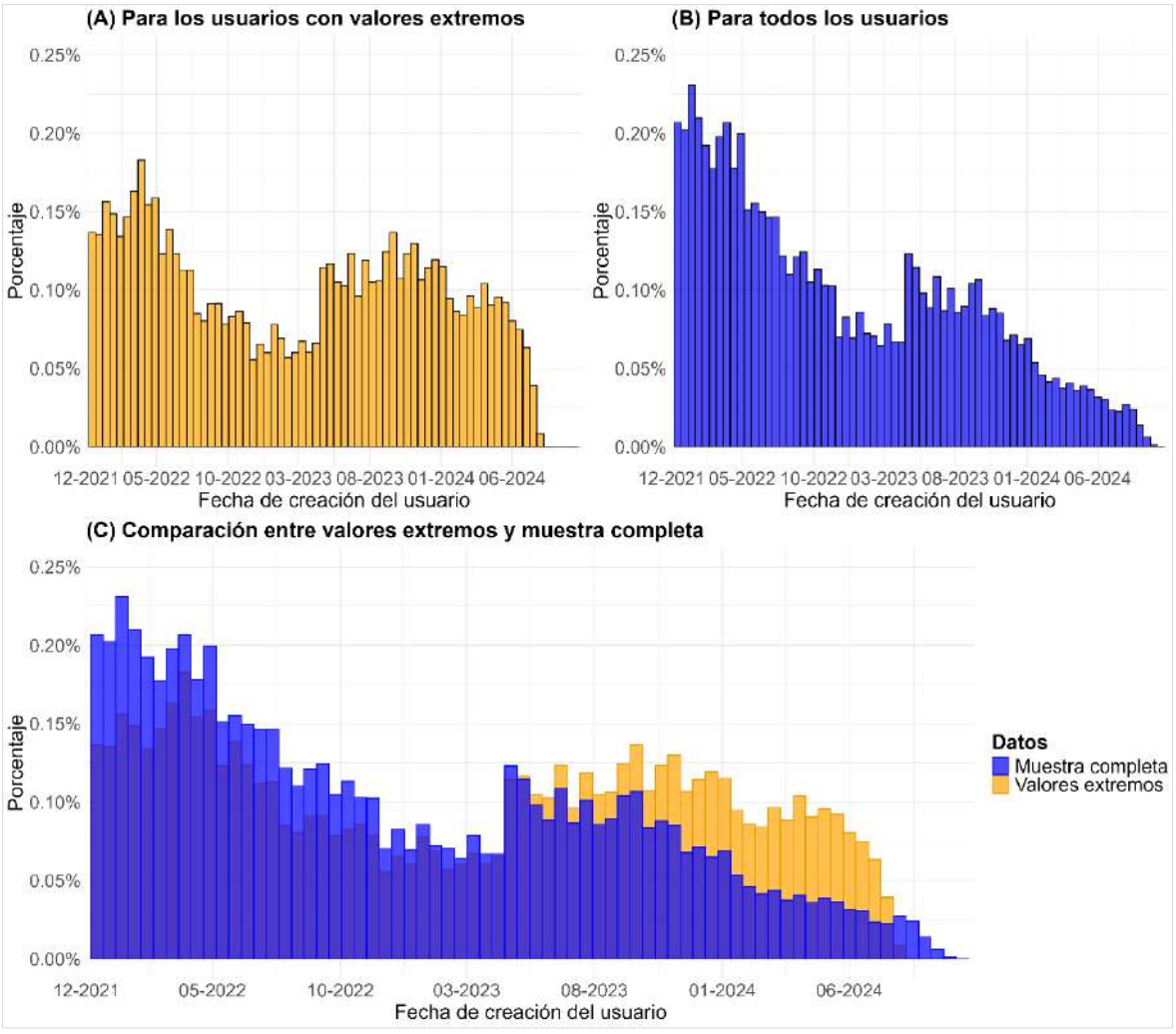
En el panel (A) del Gráfico 9, se observa la distribución de fechas de creación exclusivamente para los usuarios con valores extremos, mientras que el panel (B) del Gráfico 9 se representa la misma variable pero para la muestra completa. El panel (C) de dicho gráfico permite una visualización superpuesta que facilita el contraste directo entre ambas distribuciones. Una diferencia notable entre los paneles (A) y (B) radica en que, la distribución de la muestra completa presenta una clara tendencia decreciente a lo largo del tiempo, lo cual sugiere que la mayoría de los usuarios ingresaron a la plataforma durante los primeros meses del período de observación, la distribución correspondiente a los usuarios con VE es considerablemente más uniforme y presenta incluso picos más notorios en los tramos finales del periodo de observación.

Este patrón indica que los valores extremos tienden a estar sobrerrepresentados entre los usuarios más recientes. Dicho hallazgo puede estar vinculado a varios factores, entre ellos, la dificultad del modelo para anticipar de forma precisa los tiempos de salida de usuarios

cuya permanencia en la plataforma ha sido inusualmente breve en comparación con lo que se esperaría dada su fecha de ingreso reciente. En otras palabras, aunque estos usuarios han tenido relativamente poco tiempo desde su incorporación, abandonan la plataforma antes de lo estimado por el modelo, generando residuos de deviancia positivos. Este tipo de desajuste sugiere que el modelo sobreestima el tiempo de supervivencia para ciertos perfiles de usuarios recientes, los cuales terminan exhibiendo un comportamiento atípico desde el punto de vista del modelo de riesgos proporcionales.

Gráfico 9.

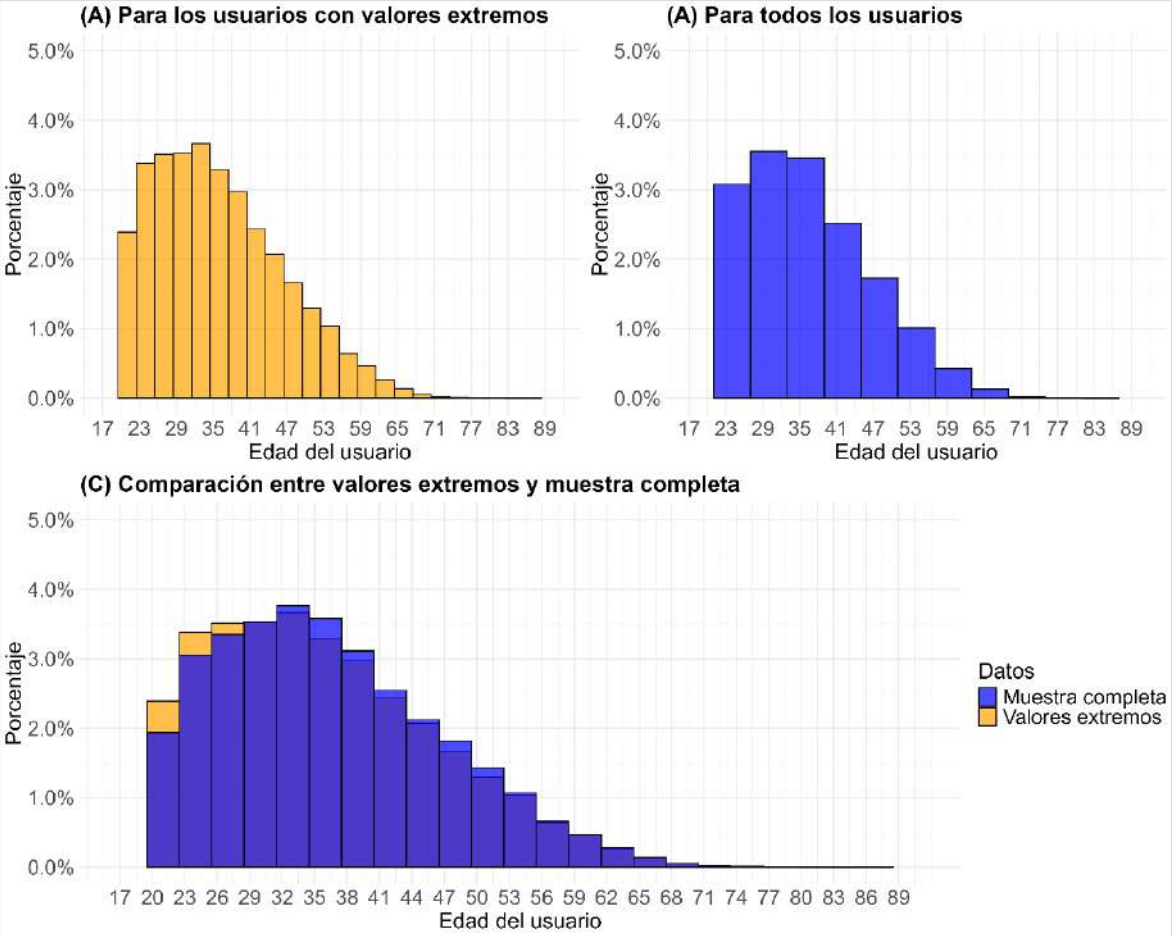
Comparación de las distribuciones de las fechas de creación de los usuarios entre toda la muestra y la muestra de valores extremos



De forma similar, con el objetivo de explorar posibles diferencias en las características demográficas de los usuarios identificados como valores extremos (VE), el Gráfico 10 presenta la distribución de las edades de los usuarios, comparando a aquellos pertenecientes al grupo de VE contra el total de la muestra.

Gráfico 10.

Comparación de las distribuciones de las Edades de los usuarios entre toda la muestra y la muestra de valores extremos



A partir del Gráfico 10, se aprecia que la distribución de edades entre los usuarios con VE es similar a la del total de la muestra, lo cual sugiere que el comportamiento atípico en el tiempo de supervivencia observado para este grupo no se debe exclusivamente a factores etarios. No obstante, se aprecian algunas diferencias sutiles: por ejemplo, en los usuarios más jóvenes

(entre 17 y 27 años), el grupo de VE tiende a estar ligeramente sobrerrepresentado, como se observa en el panel (C), donde la barra azul (VE) sobresale levemente sobre la barra anaranjada (muestra completa) en esos rangos de edad. Esta diferencia sugiere que algunos individuos jóvenes presentan un patrón de retención inesperado para el modelo, permaneciendo menos tiempo en la plataforma de lo que se anticiparía.

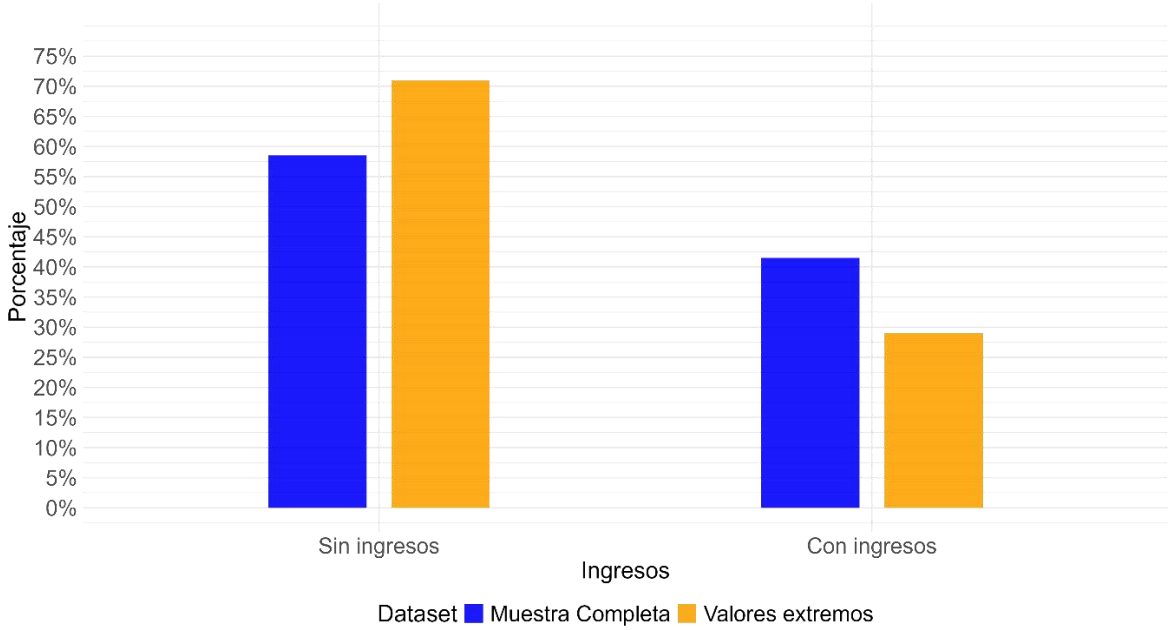
A partir de los 30 años en adelante, las distribuciones comienzan a converger y se mantienen en proporciones muy similares, lo cual refuerza la idea de que la edad, por sí sola, no es un factor diferenciador determinante entre los usuarios con valores extremos y los del resto de la muestra.

Continuando con la caracterización, el Gráfico 11 presenta una comparación de la distribución de la variable *Ingresos* entre los usuarios identificados como VE y el total de la muestra de usuarios. Esta variable clasifica a los usuarios en dos categorías: aquellos que reportaron ingresos (con ingresos) y aquellos que no (sin ingresos), lo cual permite explorar si la presencia de *ingresos* económicos se asocia de alguna manera con el comportamiento atípico capturado por el modelo de supervivencia.

El análisis de dicho Gráfico 11 revela una diferencia clara entre ambos grupos. En la muestra completa de usuarios, aproximadamente el 58% de los individuos no reporta ingresos, mientras que el 42% sí lo hace. Sin embargo, en el subconjunto de valores extremos, esta proporción se amplía notablemente: cerca del 71% de los usuarios con valores extremos pertenecen a la categoría "sin ingresos", mientras que solo alrededor del 29% reporta ingresos.

Gráfico 11.

Comparación de las distribuciones de la variable *Ingresos* de los usuarios de toda la muestra y la muestra de valores extremos



Esta disparidad sugiere que los usuarios "sin ingresos" tienen una mayor probabilidad de presentar tiempos de supervivencia inesperados desde la perspectiva del modelo, lo cual los lleva a ser clasificados como valores extremos. Una posible interpretación de este hallazgo es que, al no generar ingresos en la plataforma, estos usuarios deciden retirarse antes de lo anticipado, lo cual resulta comprensible, ya que es esperable que, al no experimentar un éxito económico temprano, opten por abandonar la plataforma de manera prematura. En contraste, los usuarios "con ingresos" parecen ajustarse en mayor medida a las predicciones del modelo, lo que se refleja en su menor representación dentro del grupo de valores extremos.

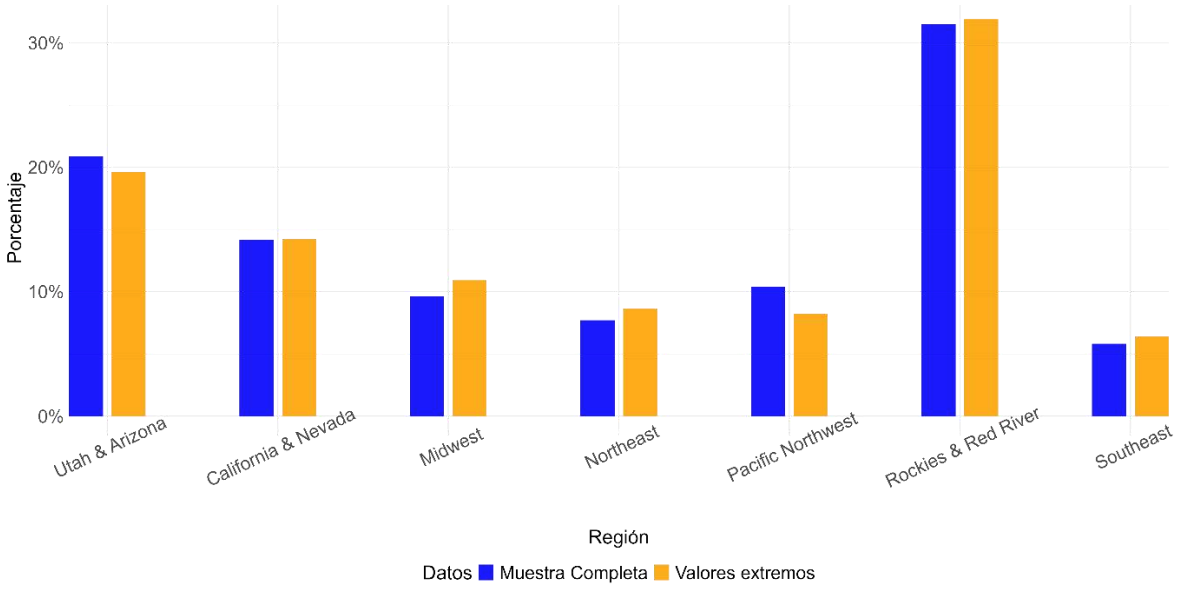
En conjunto, estos resultados sugieren que la variable *Ingresos* desempeña un papel importante en la diferenciación entre usuarios con patrones de supervivencia esperados y aquellos con comportamientos atípicos desde el punto de vista del modelo estadístico.

Por otra parte, a continuación, se presentan los gráficos 12 y 13, los cuales muestran la comparación de la distribución de las variables categóricas *Región* y *Licencias* respectivamente, tanto para los individuos clasificados como valores extremos (VE) como para el total de la muestra tipo semana-persona. Estas comparaciones permiten evaluar si

existen diferencias sustantivas en la composición geográfica o profesional entre los usuarios identificados como VE y el conjunto total de usuarios.

Gráfico 12.

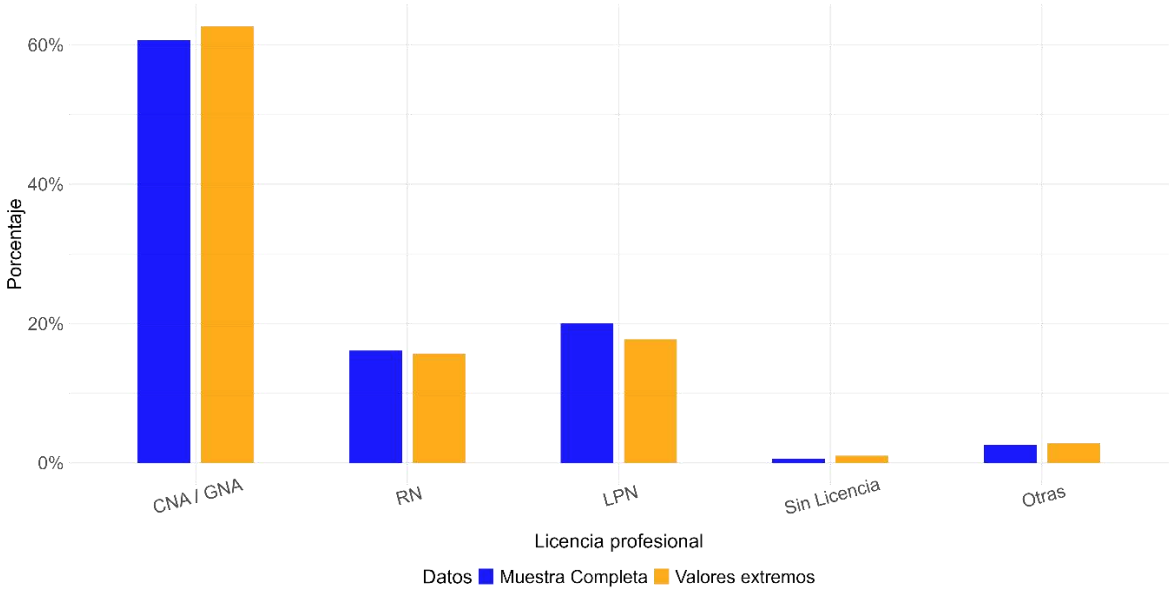
Comparación de las distribuciones de la variable Región de los usuarios de toda la muestra y la muestra de valores extremos



En el caso del Gráfico 12, correspondiente a la variable *Región*, se observa que la distribución geográfica de los usuarios con VE es bastante similar a la de la muestra completa. Las proporciones por región mantienen patrones consistentes en ambos grupos, con una alta concentración de usuarios en la región de "Rockies & Red River" (superior al 30% en ambos casos), seguida por las regiones de "Utah & Arizona" y "California & Nevada". En este caso, no se evidencian diferencias marcadas que sugieran una sobre o subrepresentación de los VE en alguna región específica.

Gráfico 13.

Comparación de las distribuciones de la variable Licencias de los usuarios de toda la muestra y la muestra de valores extremos



De forma similar, el Gráfico 13, que muestra la distribución según la variable *Licencias*, también revela una alta similitud entre ambos conjuntos de datos. En particular, la categoría "CNA/GNA" concentra la mayor proporción de usuarios en ambos casos (más del 60%), seguida de las licencias "LPN" (Licensed Practical Nurse) y "RN" (Registered Nurse). Las categorías de "Sin licencia" y "Otras" representan proporciones pequeñas tanto en la muestra total como entre los VE.

En conjunto, el análisis de estos dos gráficos sugiere que no existen diferencias relevantes en la distribución de los usuarios por región ni por tipo de licencia profesional entre los usuarios con valores extremos y la muestra completa. Por tanto, estas variables no parecen estar relacionadas de manera directa con la probabilidad de que un usuario presente un valor extremo según los residuos de deviancia del modelo extendido de Cox.

En conclusión, la variable explicativa que incide en mayor medida en la generación de valores extremos en este modelo extendido de Cox es la variable *Ingresos*. Los resultados muestran que los usuarios "sin ingresos" presentan con mayor frecuencia discrepancias entre

los tiempos observados y los estimados por el modelo, reflejadas en residuos de deviancia positivos que indican un abandono más temprano de lo previsto. Este hallazgo resulta lógico, ya que es razonable esperar que quienes no logran generar ingresos en la plataforma tiendan a retirarse de manera más rápida al no percibir beneficios económicos. En consecuencia, la ausencia de ingresos se consolida como un factor clave que explica de manera natural la sobre-representación de este grupo dentro del conjunto de valores extremos.

4.2. MODELO NO-PARAMÉTRICO

En esta sección se presentan los resultados obtenidos mediante la aplicación del modelo no paramétrico de Kaplan-Meier, el cual permitió estimar las funciones de supervivencia para la muestra de usuarios de la plataforma digital de contratación "*per diem*". Donde, el objetivo es describir cómo varía la probabilidad de permanencia de los usuarios en la plataforma a lo largo del tiempo, así como identificar posibles diferencias entre grupos definidos por variables sociodemográficas y laborales.

De forma complementaria, además de las curvas de supervivencia, en esta sección también se incluyen los gráficos del riesgo instantáneo $h(t)$ estimados no paramétricamente. En conjunto, estos gráficos permiten realizar un análisis descriptivo global y estratificado por las variables consideradas en la sección metodológica (*edad*, tipo de *licencia* profesional, *región* geográfica, e *ingresos*). Se evaluaron diferencias entre curvas de supervivencia con la prueba de Log-Rank y se contrastaron los perfiles de riesgo para localizar ventanas temporales de mayor probabilidad condicional de abandono. Esta aproximación permite caracterizar con mayor precisión los patrones de permanencia y cuantificar sus diferencias entre subgrupos, aportando evidencia útil para definir estrategias de intervención en los usuarios de la plataforma en estudio.

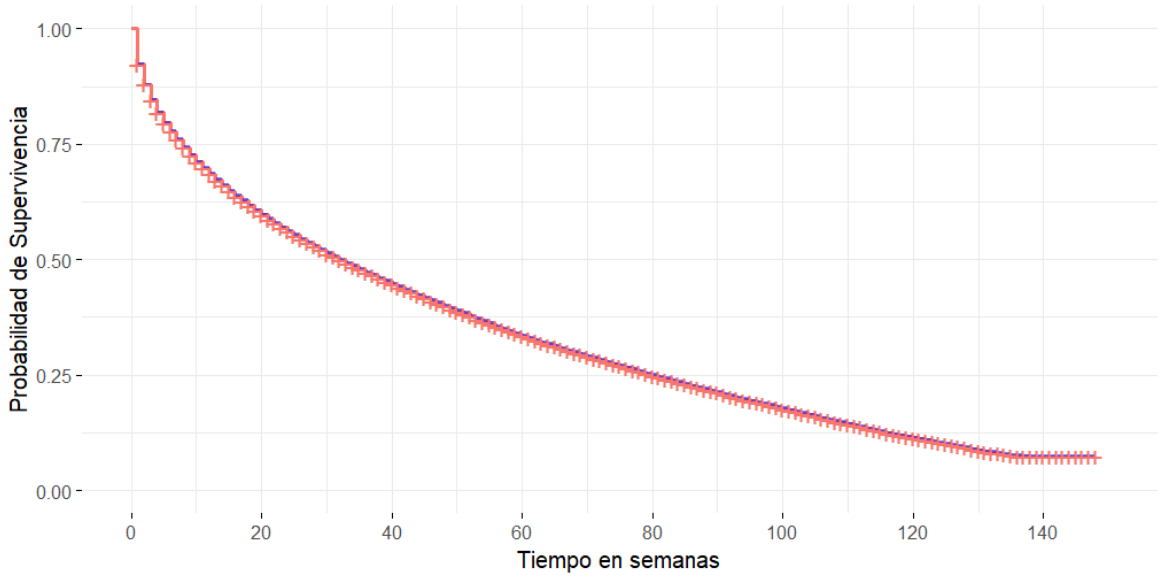
4.2.1. Gráficos y tiempos de supervivencia no paramétricos

El siguiente Gráfico 14 presenta la curva de supervivencia Kaplan-Meier para la muestra completa de usuarios de la plataforma. Esta curva permite observar la probabilidad

acumulada de supervivencia de los usuarios a lo largo del tiempo, medida en semanas, y está acompañada de intervalos de confianza al 95%, representados en color azul.

Gráfico 14.

Curva de supervivencia Kaplan-Meier para toda la muestra de usuarios



El intervalo de confianza del 95% se presenta en color azul.

Fuente: Plataforma digital anónima. (2024).

La curva de supervivencia presentada en el Gráfico 14 evidencia una disminución progresiva en la probabilidad de que los usuarios permanezcan activos en la plataforma a lo largo del tiempo. Este patrón decreciente es consistente con lo observado en los histogramas (Gráfico 6), donde se registró una alta concentración de eventos de abandono durante las primeras 25 semanas. Inicialmente, la curva desciende con mayor pendiente, lo que indica una tasa de abandono acelerada en las primeras etapas del uso del servicio. Sin embargo, conforme transcurren las semanas la pendiente se suaviza, sugiriendo que aquellos usuarios que superan los primeros meses tienden a mantener una mayor permanencia. A partir de aproximadamente la semana 20, la curva adopta una forma más estable y cercana a la

linealidad, lo cual refleja una menor tasa de abandono entre los usuarios que han permanecido más tiempo en la plataforma.

Por su parte, el Cuadro 15 presenta los tiempos estimados de supervivencia, expresados en semanas, obtenidos mediante el uso de las curvas Kaplan-Meier, para distintos niveles de probabilidad de supervivencia (20%, 50%, 75% y 90%). Estos resultados se muestran tanto para la muestra general de usuarios como para los subgrupos clasificados según el tipo de *licencia*. Y, se puede apreciar que, para la curva general, con todos los usuarios de la muestra, la media (probabilidad del 50%) es de 32 semanas, equivalente a aproximadamente 7.4 meses.

De forma complementaria, el Gráfico 15 muestra las curvas de supervivencia estimadas mediante el método de Kaplan-Meier para las distintas categorías de *licencia* profesional registradas en la plataforma. Además, cada curva incluye intervalos de confianza al 95%, lo que permite valorar la precisión de las estimaciones y comparar la robustez de las diferencias observadas entre grupos.

Cuadro 15.

Tiempos de Supervivencia según Kaplan-Meier para las distintas categorías de Licencia

Licencias	Probabilidades de Supervivencia			
	0.2	**0.5**	**0.75**	**0.9**
General	93	32	8	2
Licencia LPN	108	40	10	2
Licencia RN	96	34	9	2
Licencia CNA / GNA	89	30	7	1
Otras Licencias	79	25	5	1
Licencia Sin licencia	50	12	2	1

Se presentan los tiempos en número de semanas.

Fuente: Plataforma digital anónima. (2024).

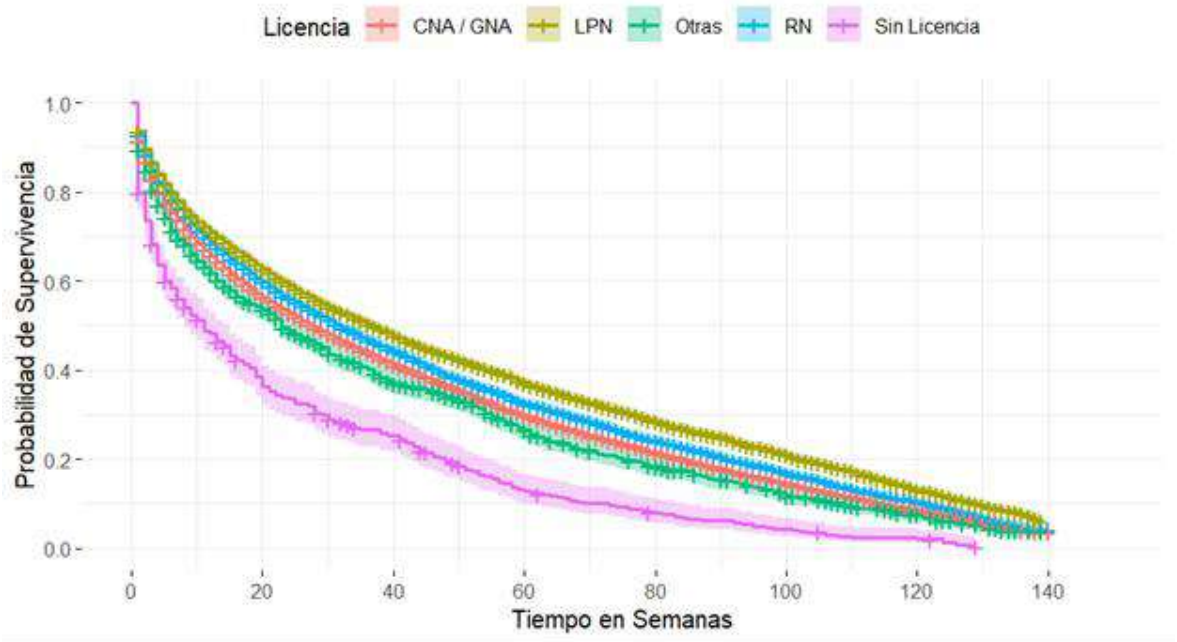
Se puede apreciar en el Cuadro 15 que, en general, los usuarios con licencia "LPN" presentan los tiempos de supervivencia más prolongados para todas las probabilidades de falla, alcanzando hasta 40 semanas para una probabilidad del 50%. En contraste, los usuarios "Sin licencia" tienen los tiempos más cortos, como es esperado, con tan solo 12 semanas para la misma probabilidad de supervivencia. Este patrón sugiere que el tipo de *licencia* presenta una influencia importante en la duración del uso de la plataforma, con los usuarios más profesionalizados, mostrando mayor compromiso y uso sostenido de las herramientas de contratación per-diem.

De manera consistente, el siguiente Gráfico 15 confirma que los usuarios con licencia "LPN" presentan la mayor probabilidad de supervivencia a lo largo del tiempo, lo que se traduce en una retención más alta en el uso de la plataforma. Por su parte, los usuarios con licencias "RN", "CNA/GNA" y "Otras" siguen patrones de supervivencia similares, aunque ligeramente inferiores. En contraste, los usuarios "Sin licencia" tienen las probabilidades más bajas de supervivencia desde el inicio, y su curva desciende de forma más abrupta en las primeras semanas en comparación con los usuarios con otras licencias, lo que refleja un abandono más temprano del servicio.

En conjunto, los resultados del Gráfico 15 y el Cuadro 15 destacan que el nivel de profesionalización, representado por el tipo de licencia, está estrechamente relacionado con el compromiso y la duración del uso de la plataforma. Este hallazgo resalta la importancia de las licencias como un factor determinante en el comportamiento de los usuarios.

Gráfico 15.

Curvas de Supervivencia Kaplan-Meier para las distintas categorías de *Licencia*



Las curvas presentan intervalos de confianza al 95%

Fuente: Plataforma digital anónima. (2024).

Por su parte, el siguiente Cuadro 16 presenta los resultados de la prueba de Log-Rank, utilizada para evaluar si existen diferencias estadísticamente significativas entre las curvas de supervivencia según las distintas categorías de la variable *Licencia*. Se reporta tanto la comparación global entre todas las categorías como las comparaciones par a par. Dado que se realizan múltiples contrastes (con cinco categorías se generan 10 comparaciones por pares), se aplicó la corrección de Bonferroni para controlar el error por comparaciones múltiples. En consecuencia, para un nivel de significancia $\alpha = 0.05$, el umbral ajustado es $\alpha^* = \alpha/m = 0.05/10 = 0.005$.

Cuadro 16.

Prueba de Log-Rank para la Comparación de Curvas de Supervivencia según categorías de *Licencia*

Comparación	Chi²	G. L.	Valor-p.	Interpretación
RN vs CNA / GNA	21.92	1	< 0.0001	Significativa
RN vs LPN	43.12	1	< 0.0001	Significativa
RN vs Sin licencia	163.20	1	< 0.0001	Significativa
RN vs Otras	27.68	1	< 0.0001	Significativa
CNA / GNA vs LPN	175.39	1	< 0.0001	Significativa
CNA / GNA vs Sin licencia	133.09	1	< 0.0001	Significativa
CNA / GNA vs Otras	11.21	1	0.0008	Significativa
LPN vs Sin licencia	234.11	1	< 0.0001	Significativa
LPN vs Otras	79.84	1	< 0.0001	Significativa
Sin licencia vs Otras	59.48	1	< 0.0001	Significativa
Entre todas las categorías de licencia	359.18	4	< 0.0001	Significativa

Nota: G.L. corresponde a los grados de libertad empleados en la prueba de Log-Rank.

Como se observaba en el Cuadro 16 los resultados de las pruebas Log-Rank para *Licencias* evidencian diferencias estadísticamente significativas en la supervivencia entre la mayoría de las comparaciones de grupos, con valores p inferiores a 0.005 (con corrección por Bonferroni). Además, la prueba global confirma diferencias significativas entre todas las categorías de licencia ($Chi^2 = 359.18$; G.L. = 4; $p < 0.001$), lo que indica que el tipo de licencia profesional se asocia de manera relevante con la permanencia de los usuarios en la plataforma.

Por su parte, relacionado con las regiones geográficas donde residen y trabajan los usuarios de la plataforma, en el Cuadro 17 y el Gráfico 16 se describen los tiempos de supervivencia según la metodología Kaplan-Meier para estos usuarios según las diferentes regiones geográficas de Estados Unidos. Estos resúmenes en tabla y representaciones gráficas permiten explorar cómo las características geográficas de los usuarios influyen en los patrones de uso y abandono de la plataforma.

Cuadro 17.

Tiempos de Supervivencia según Kaplan-Meier para las distintas Regiones geográficas

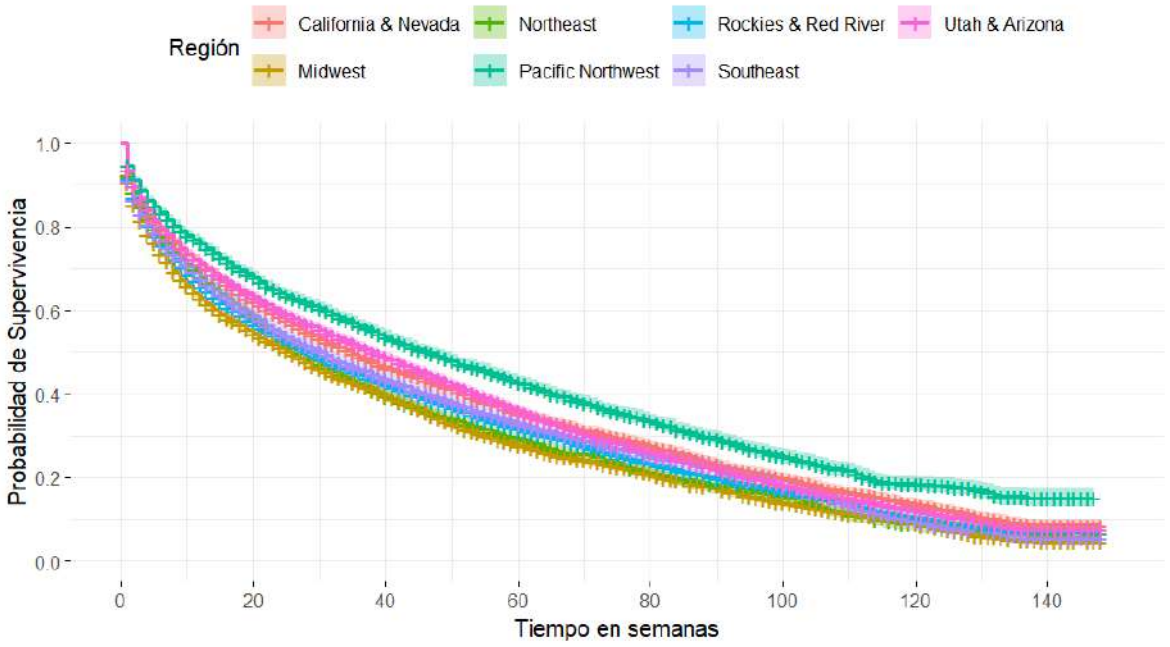
Regiones	Probabilidades de Supervivencia			
	0.2	0.5	0.75	0.9
General	93	32	8	2
Región Pacific Northwest	112	46	13	3
Región Utah & Arizona	95	38	9	2
Región California & Nevada	99	35	9	2
Región Rockies & Red River	89	28	6	1
Región Southeast	94	30	7	1
Región Northeast	83	27	8	2
Región Midwest	82	25	5	1

Se presentan los tiempos en número de semanas

Fuente: Plataforma digital anónima. (2024).

Gráfico 16.

Curvas de Supervivencia Kaplan-Meier para las distintas Regiones geográficas



Las curvas presentan intervalos de confianza al 95%

Fuente: Plataforma digital anónima. (2024).

Del análisis conjunto del Cuadro 17 y el Gráfico 16 se observa que, aunque la tendencia general de las curvas de supervivencia es similar para todas las regiones, existen diferencias notables en los tiempos de permanencia. En términos generales, todas las regiones muestran una disminución progresiva en la probabilidad de supervivencia a lo largo del tiempo, con una pendiente inicial más pronunciada que se estabiliza gradualmente, lo cual es consistente con el comportamiento observado en la curva general de Kaplan-Meier. Este patrón podría indicar que las dinámicas de uso de la plataforma son relativamente homogéneas entre regiones; sin embargo, factores geográficos o contextuales parecen tener un efecto sobre la duración del compromiso de los usuarios.

Destaca especialmente la región "Pacific Northwest", que presenta los mayores tiempos de supervivencia, con 112 semanas para una probabilidad de falla del 20% y 46 semanas para una probabilidad del 50%. Esto sugiere que los usuarios de esta región tienden a permanecer más tiempo en la plataforma antes de abandonar, en comparación con los usuarios de otras zonas geográficas. Por el contrario, las regiones "Midwest" y "Northeast" muestran los menores tiempos de permanencia, con medianas de supervivencia de 25 y 27 semanas respectivamente, lo que refleja una propensión al abandono más temprano en estos territorios.

Complementariamente, el Cuadro 18 presenta los resultados de la prueba Log-Rank, empleada para comparar las curvas de supervivencia entre las distintas categorías de la variable Región. Además de la comparación global, que evidencia diferencias estadísticamente significativas entre las regiones ($Chi^2 = 387.67$, G.L. = 6, $p < 0.001$), también se incluyen comparaciones par a par con el fin de identificar qué pares de regiones difieren de manera relevante. Dado que se realizan múltiples contrastes (con 7 categorías se generan 21 comparaciones por pares), se aplicó la corrección de Bonferroni para controlar el error por comparaciones múltiples. En consecuencia, para $\alpha = 0.05$, el nivel de significancia ajustado es $\alpha^* = \alpha/m = 0.05/21 = 0.0024$.

Los resultados del Cuadro 18 muestran diferencias estadísticamente significativas en múltiples comparaciones, especialmente entre la región "Pacific Northwest" y las demás

regiones, lo que evidencia una mayor retención de usuarios en dicha zona. Sin embargo, algunas comparaciones no mostraron diferencias significativas ($p > 0.0024$), concretamente: "California & Nevada vs Utah & Arizona", "Northeast vs Rockies & Red River", "Northeast vs Midwest", "Northeast vs Southeast", y "Rockies & Red River vs Southeast". Estos resultados indican que, aunque se observa una tendencia general de disminución en la probabilidad de supervivencia, existen pares de regiones con patrones de permanencia similares, mientras que otras zonas, como "Pacific Northwest", sobresalen por una mayor permanencia en la plataforma.

Sumado a lo anterior, el Gráfico 17 presenta las curvas de supervivencia de Kaplan-Meier para la variable Edad, categorizada en cinco grupos (18–25, 26–35, 36–45, 46–55 y 55 años o más) con el fin de facilitar la comparación de los patrones de permanencia entre rangos etarios. El gráfico incluye intervalos de confianza al 95%, lo que aporta información adicional sobre la incertidumbre de las estimaciones a lo largo del tiempo.

Cuadro 18.

Prueba de Log-Rank para la Comparación de Curvas de Supervivencia según categorías de Región

Comparación	Chi²	G. L.	Valor-p.	Interpretación
California & Nevada vs Northeast	46.95	1	< 0.0001	Significativa
California & Nevada vs Rockies & Red River	54.09	1	< 0.0001	Significativa
California & Nevada vs Midwest	104.92	1	< 0.0001	Significativa
California & Nevada vs Pacific Northwest	65.80	1	< 0.0001	Significativa
California & Nevada vs Southeast	15.18	1	0.0001	Significativa
California & Nevada vs Utah & Arizona	0.02	1	0.8936	No significativa
Northeast vs Rockies & Red River	1.85	1	0.1917	No significativa
Northeast vs Midwest	7.12	1	0.0094	Significativa
Northeast vs Pacific Northwest	184.42	1	< 0.0001	Significativa
Northeast vs Southeast	4.10	1	0.0502	No significativa
Northeast vs Utah & Arizona	63.50	1	< 0.0001	Significativa
Rockies & Red River vs Midwest	23.53	1	< 0.0001	Significativa
Rockies & Red River vs Pacific Northwest	222.35	1	< 0.0001	Significativa
Rockies & Red River vs Southeast	1.28	1	0.2715	No significativa
Rockies & Red River vs Utah & Arizona	54.92	1	< 0.0001	Significativa
Midwest vs Pacific Northwest	277.44	1	< 0.0001	Significativa
Midwest vs Southeast	19.20	1	< 0.0001	Significativa
Midwest vs Utah & Arizona	117.80	1	< 0.0001	Significativa
Pacific Northwest vs Southeast	103.84	1	< 0.0001	Significativa
Pacific Northwest vs Utah & Arizona	72.73	1	< 0.0001	Significativa
Southeast vs Utah & Arizona	16.52	1	< 0.0001	Significativa
Entre todas las categorías de regiones	387.67	6	< 0.0001	Significativa

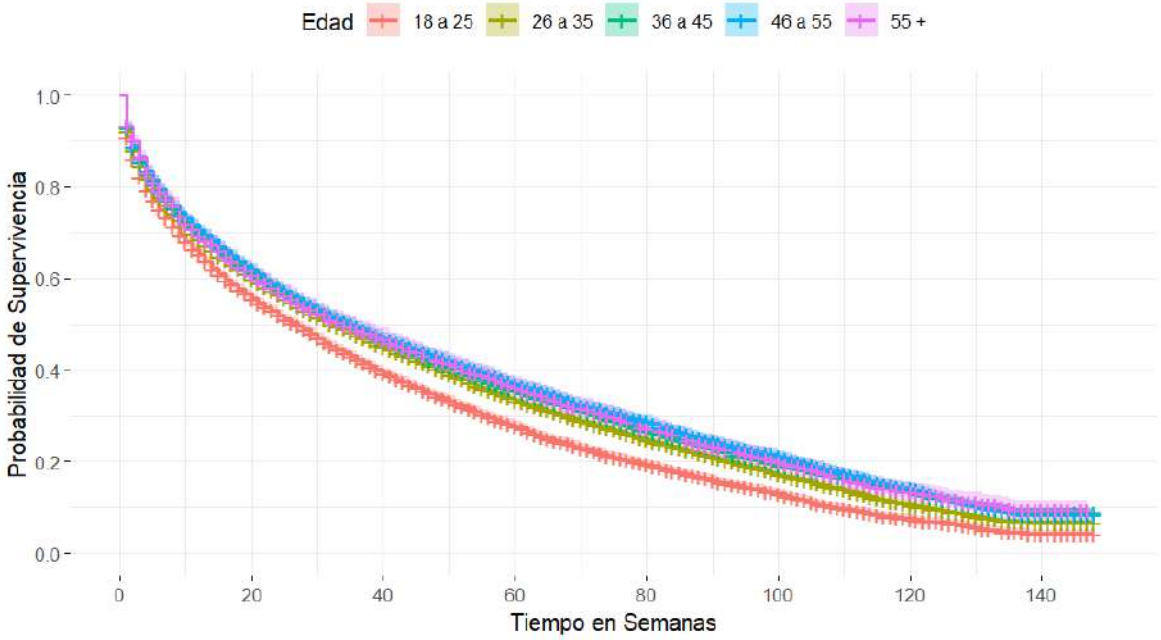
Nota: G.L. corresponde a los grados de libertad empleados en la prueba de Log-Rank.

El Cuadro 19 complementa este análisis mediante la prueba Log-Rank. En primer lugar, la comparación global evidencia diferencias estadísticamente significativas entre los grupos de edad ($Chi^2 = 228.89$, G.L. = 4, $p < 0.001$). Además, se reportan comparaciones par a par para identificar entre qué rangos etarios se presentan diferencias relevantes. Dado que se definieron cinco categorías de edad, se realizan 10 comparaciones por pares; por ello, se aplicó la corrección de Bonferroni para controlar el error por comparaciones múltiples, utilizando un umbral ajustado de $\alpha^* = 0.05/10 = 0.005$. Bajo este criterio, las diferencias más consistentes se observan principalmente en las comparaciones que involucran a los grupos más jóvenes, mientras que varios contrastes entre grupos de mayor edad no muestran evidencia suficiente de diferencias. En particular, las comparaciones "26–35" vs "55+" ($p = 0.0084$), "36–45" vs "46–55" ($p = 0.157$), "36–45" vs "55+" ($p = 0.7019$) y "46–55" vs "55+" ($p = 0.6197$) no resultaron significativas tras el ajuste de Bonferroni, lo que sugiere que, en los tramos de edad más altos, las probabilidades de permanencia tienden a ser más similares.

Estos hallazgos sugieren que las diferencias en los tiempos de permanencia son más pronunciadas entre los usuarios más jóvenes, especialmente aquellos de 18 a 25 años, mientras que entre los grupos de edad más avanzada la probabilidad de abandono tiende a ser más homogénea.

Gráfico 17.

Curvas de Supervivencia Kaplan-Meier para los distintos grupos de *Edades*



Las curvas presentan intervalos de confianza al 95%

Fuente: Plataforma digital anónima. (2024).

Cuadro 19.

Prueba de Log-Rank para la Comparación de Curvas de Supervivencia según categorías de Edad

Comparación	Chi²	G. L.	Valor-p.	Interpretación
18-25 vs 26-35	98.34	1	< 0.0001	Significativa
18-25 vs 36-45	169.62	1	< 0.0001	Significativa
18-25 vs 46-55	157.69	1	< 0.0001	Significativa
18-25 vs 55+	63.41	1	< 0.0001	Significativa
26-35 vs 36-45	20.33	1	< 0.0001	Significativa
26-35 vs 46-55	28.04	1	< 0.0001	Significativa
26-35 vs 55+	7.58	1	0.0084	No significativa
36-45 vs 46-55	2.35	1	0.157	No significativa
36-45 vs 55+	0.15	1	0.7019	No significativa
46-55 vs 55+	0.34	1	0.6197	No significativa
Entre categorías de Edades	228.89	4	< 0.0001	Significativa

Nota: G.L. corresponde a los grados de libertad empleados en la prueba de Log-Rank.

De manera adicional, el Cuadro 20, Cuadro 21, y el Gráfico 18 presentan de forma complementaria el análisis de los tiempos de supervivencia según las categorías de *Ingresos* de los usuarios de la plataforma digital. Para este análisis, se diferenciaron dos grupos: "sin ingresos", correspondiente a los usuarios que no generaron ingresos mediante la plataforma durante sus primeros dos meses de uso, y "con ingresos", que incluye a quienes sí lograron generar ingresos en ese periodo inicial.

Los resultados muestran diferencias sustanciales entre ambos grupos, tanto a nivel gráfico como estadístico. El Gráfico 18 permite visualizar de manera clara estas diferencias, mostrando que la curva de supervivencia para los usuarios "con ingresos" se mantiene consistentemente más alta a lo largo del tiempo, con un descenso más gradual, especialmente durante las primeras semanas. En contraste, la curva de los usuarios "sin ingresos" presenta una disminución más pronunciada en las etapas iniciales, reflejando una mayor probabilidad de abandono temprano. Ambas curvas incluyen intervalos de confianza al 95%, lo que refuerza la solidez de las estimaciones.

Estas diferencias gráficas se ven reflejadas también en los tiempos estimados de supervivencia presentados en el Cuadro 20. Por ejemplo, el tiempo necesario para alcanzar una probabilidad de supervivencia del 50% es de 54 semanas para los usuarios "con ingresos", mientras que para los usuarios "sin ingresos" se reduce a tan solo 23 semanas, evidenciando una clara brecha en los patrones de permanencia según la generación de ingresos.

Desde el punto de vista estadístico, el Cuadro 21 complementa este análisis mediante la prueba global de Log-Rank, la cual confirma la existencia de diferencias significativas entre ambos grupos ($Chi^2 = 2{,}251.96$; G.L. = 1; $p < 0.001$). Este resultado valida numéricamente lo observado en las curvas y los tiempos de supervivencia, confirmando que la condición de *ingresos* es un factor diferenciador relevante en la permanencia de los usuarios dentro de la plataforma.

En conjunto, el análisis de los Cuadros 20 y 21 junto al Gráfico 18 permite concluir que los usuarios que logran generar ingresos durante sus primeras semanas de uso muestran una mayor retención y una menor probabilidad de abandono temprano. Estos hallazgos sugieren

que la capacidad de generar ingresos a corto plazo podría estar estrechamente vinculada a una mayor satisfacción o motivación para permanecer activos en la plataforma, destacando la relevancia de este factor como determinante clave en los patrones de uso y deserción.

Cuadro 20.

Tiempos de Supervivencia según Kaplan-Meier para las distintas categorías de *Ingresos*

Variable Ingresos	Probabilidades de Supervivencia			
	0.2	**0.5**	**0.75**	**0.9**
General	93	32	8	2
Sin Ingresos	78	23	5	1
Con Ingresos	117	54	19	6

Se presentan los tiempos en número de semanas

Fuente: Plataforma digital anónima. (2024).

Cuadro 21.

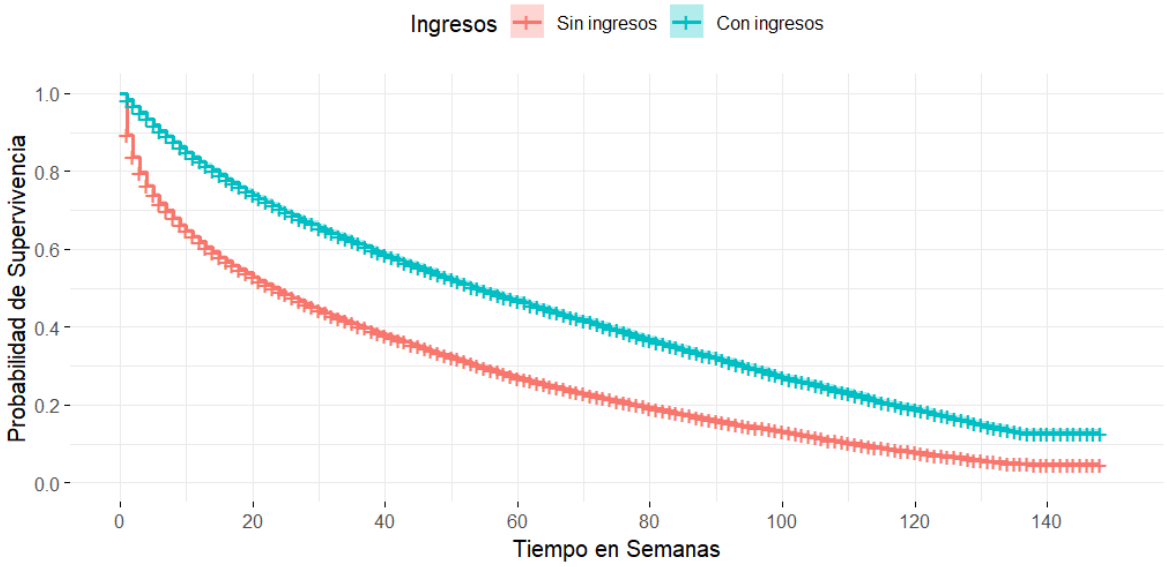
Prueba de Log-Rank para la Comparación de Curvas de Supervivencia según categorías de *Ingresos*

Comparación	Chi²	G. L.	Valor-p.	Interpretación
Entre las categorías de Ingresos	2,251.96	1	< 0.0001	Significativa

Nota: G.L. corresponde a los grados de libertad empleados en la prueba de Log-Rank.

Gráfico 18.

Curvas de Supervivencia Kaplan-Meier para los distintos grupos de Ingresos



Las curvas presentan intervalos de confianza al 95%

Fuente: Plataforma digital anónima. (2024).

4.2.2. Gráficos del riesgo de supervivencia

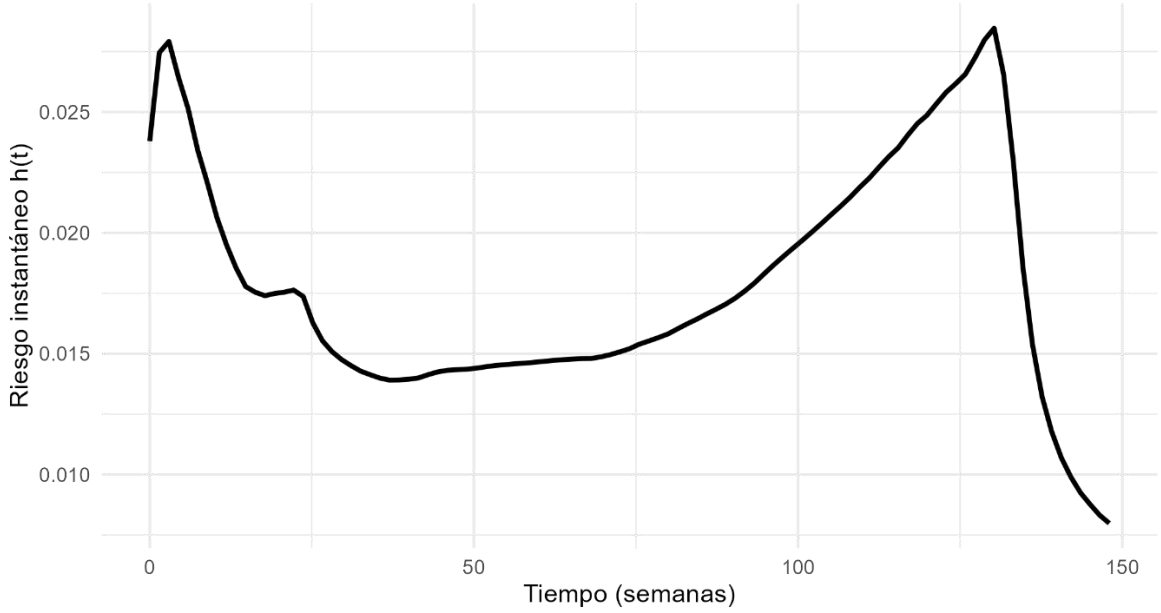
En esta sección se presentan distintos gráficos del riesgo instantáneo $h(t)$ de los datos, tanto a nivel global como para diversos subconjuntos de usuarios.

Para su construcción se empleó la función *muhaz* (Hess & Gentleman, 2021), tal como se describió en la metodología. Este procedimiento es no paramétrico y estima $h(t)$ mediante suavizado por núcleo, siguiendo el enfoque de Müller y Wang (1994), lo que permite obtener curvas de riesgo más estables y legibles en el tiempo.

A continuación, se presenta el Gráfico 19, el cual muestra la tasa de riesgo instantáneo $h(t)$ para el conjunto de datos completo, estimada de forma no paramétrica. Esta representación permite identificar periodos donde la probabilidad condicional de abandono es mayor o menor entre quienes continúan activos.

Gráfico 19.

Riesgo instantáneo estimado no paramétricamente con suavizado



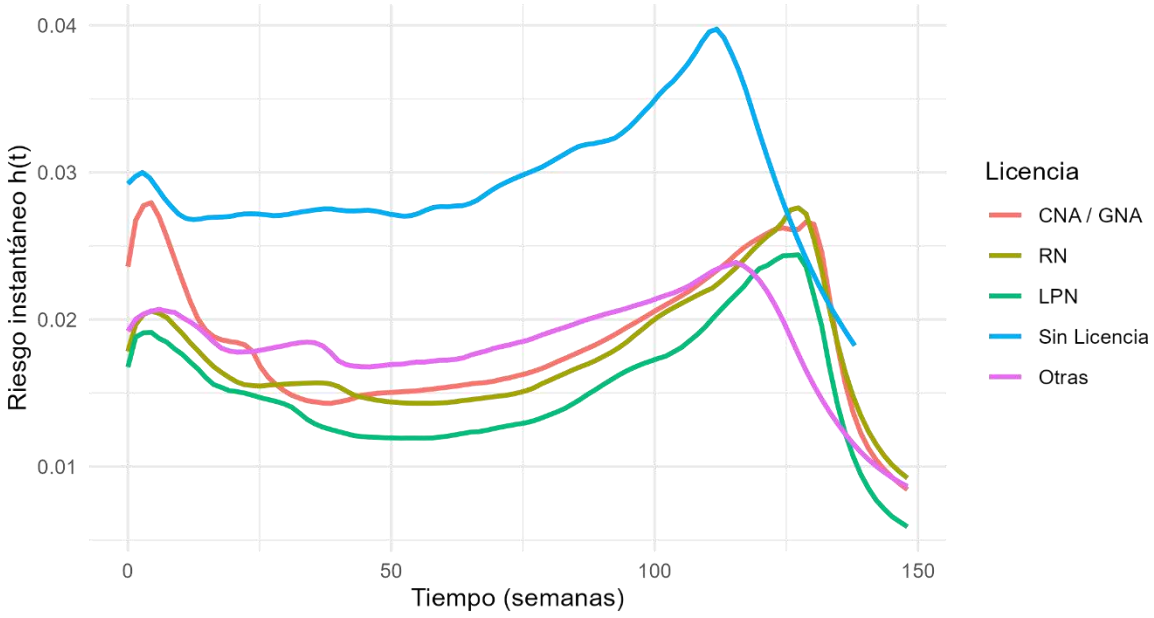
En el Gráfico 19, se observa un nivel de riesgo relativamente alto al inicio del periodo de estudio, que desciende de forma sostenida durante las primeras de semanas hasta alcanzar un valle cercano a 0.015 (alrededor de la semana 30). A partir de aproximadamente 90–100 semanas, el riesgo vuelve a incrementarse gradualmente, sugiriendo una nueva aceleración del abandono en horizontes más largos, con un pico pronunciado alrededor de 120–130 semanas. Hacia el extremo final la curva muestra un descenso brusco, que debe interpretarse con cautela por la escasez de individuos en riesgo en esas semanas (efecto típico de cola), producido por ser el periodo de finalización del estudio. En conjunto, la trayectoria indica dos ventanas de atención: una temprana, donde conviene reforzar la retención inicial, y otra tardía, donde podrían ser útiles estrategias de reactivación o fidelización para quienes han permanecido por largos periodos.

Seguidamente, el Gráfico 20 divide explícitamente los datos por tipo de licencia profesional de los usuarios de la plataforma ("CNA/GNA", "RN", "LPN", "Sin licencia" y "Otras") y, para cada estrato, estima la tasa de riesgo instantáneo $h(t)$ mediante el procedimiento no paramétrico anteriormente mencionado. Esta estratificación permite comparar cómo

evoluciona la probabilidad condicional de abandono a lo largo del tiempo dentro de cada categoría de *licencia*, manteniendo el mismo horizonte temporal y la misma metodología de estimación para asegurar comparabilidad directa entre curvas.

Gráfico 20.

Riesgo instantáneo estimado no paramétricamente con suavizado para los distintos grupos de *Licencias*



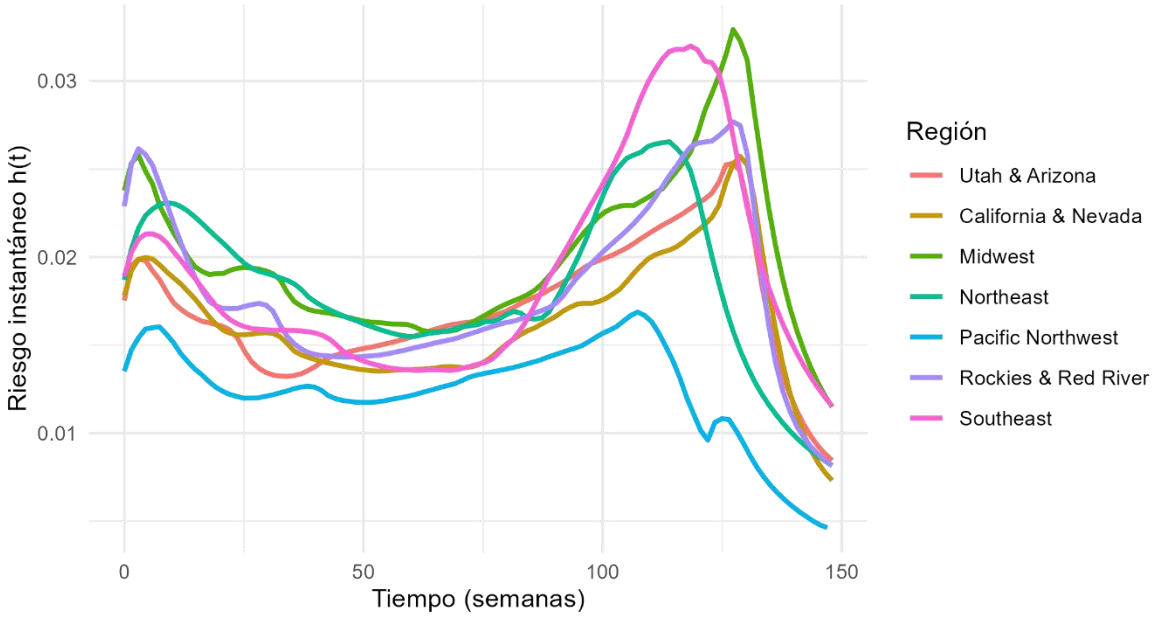
En el Gráfico 20 se observa un patrón común de pico temprano seguido de un valle intermedio (alrededor de 30–50 semanas) y un incremento paulatino hacia el tramo de 100–125 semanas, tras el cual aparece un descenso final que debe leerse con cautela por la reducción de individuos en riesgo dado el periodo de finalización del estudio. La categoría "Sin licencia" presenta el riesgo más elevado de forma sostenida y alcanza el pico más pronunciado en el tramo tardío (alrededor de las semanas 110–125), lo que sugiere mayor propensión al abandono tanto al inicio como en horizontes largos. En contraste, "LPN" mantiene el perfil de riesgo más bajo durante casi todo el periodo, con una curva claramente por debajo del resto y un pico tardío más moderado, indicando que los usuarios con esta licencia presentan menores riesgos de abandonar la plataforma. "RN" y "CNA/GNA" se ubican en una banda intermedia, con "CNA/GNA" mostrando un pico inicial relativamente

alto antes de descender, y "RN" con niveles menores durante buena parte del horizonte. La categoría "Otras" se mantiene entre el grupo intermedio y, en el tramo final, converge hacia niveles similares a "RN" y "CNA-GNA". En conjunto, la evidencia sugiere que los perfiles con licencia formal (en especial "LPN" y, en segundo término, "RN") exhiben menores riesgos de abandono, mientras que los usuarios sin licencia concentran las ventanas críticas de intervención tanto tempranas como tardías.

El Gráfico 21, de forma similar al anterior, estratifica los usuarios según la región geográfica en la que trabajan ("Utah & Arizona", "California & Nevada", "Midwest", "Northeast", "Pacific Northwest", "Rockies & Red River" y "Southeast") y, para cada estrato regional, estima la tasa de riesgo instantáneo $h(t)$ mediante el procedimiento no paramétrico con suavizado por núcleo. Esta desagregación permite comparar la evolución temporal de la probabilidad condicional de abandono entre regiones bajo un mismo horizonte y una metodología de estimación homogénea, garantizando la comparabilidad directa de las curvas.

Gráfico 21.

Riesgo instantáneo estimado no paramétricamente con suavizado para los distintos grupos de Regiones



En dicho Gráfico 21, se aprecia un patrón compartido de pico temprano, valle intermedio (cerca de las 30–60 semanas) y alza tardía hacia las 100–125 semanas, aproximadamente, seguido de un descenso final que debe interpretarse con cautela por la reducción del número de sujetos en riesgo. Destacan "Midwest y Southeast" con los picos tardíos más pronunciados, señalando mayor probabilidad condicional de abandono en horizontes largos; "Rockies & Red River" también muestra un incremento elevado en ese tramo. En contraste, "Pacific Northwest" mantiene la trayectoria más baja durante casi todo el periodo, con un valle profundo y un nivel de riesgo final claramente inferior. "Utah & Arizona" y "Northeast" se ubican en una banda intermedia, similar a "California & Nevada" la cual presenta un perfil relativamente estable y moderado. En conjunto, las diferencias regionales sugieren ventanas de intervención tardías especialmente relevantes en "Midwest", "Southeast", y "Rockies & Red River", y un riesgo persistentemente menor en "Pacific Northwest", coherente con mejores perspectivas de retención en esa región.

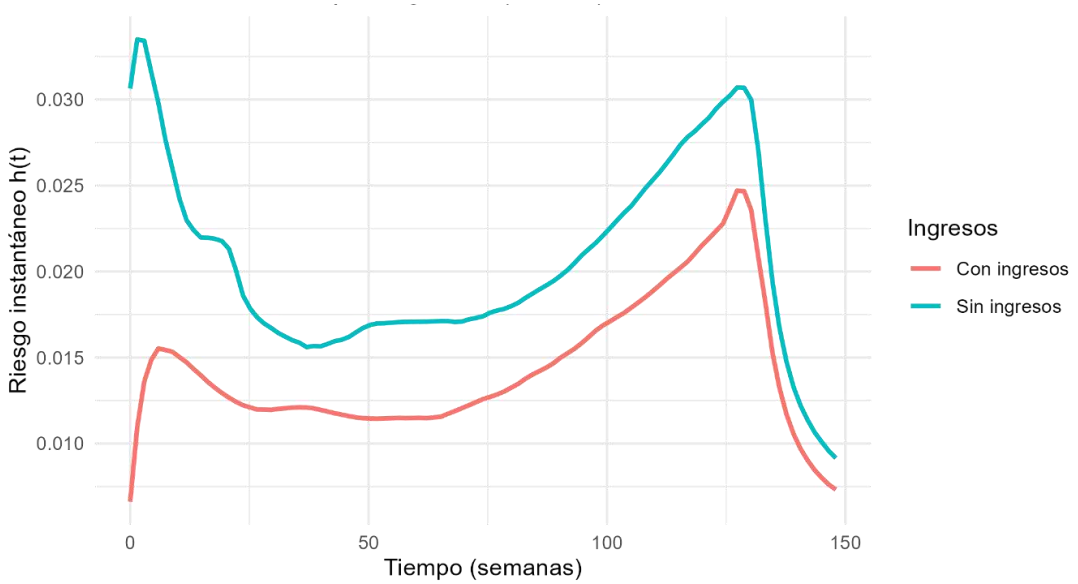
El Gráfico 22 estratifica los datos según la condición de ingresos en los primeros 2 meses de haberse registrado en la plataforma ("con ingresos" vs. "sin ingresos") y estima, para cada estrato, la tasa de riesgo instantáneo $h(t)$ mediante el mismo procedimiento no paramétrico. Esta desagregación permite comparar la evolución de la probabilidad condicional de abandono entre ambos grupos, garantizando comparabilidad directa de las curvas.

En dicho Gráfico 22, se observan trayectorias con pico temprano, valle intermedio (cerca de las 30–50 semanas) y alza tardía hacia las 100–125 semanas, similares a los de los gráficos anteriores. Sin embargo, en este caso las diferencias entre los dos grupos de *Ingresos* son más marcadas. A lo largo de casi todo el periodo, el grupo "sin ingresos" mantiene un riesgo claramente superior: exhibe un pico inicial bastante alto (primeras semanas), desciende hasta un mínimo alrededor de 40 semanas y luego vuelve a acelerar de forma marcada hacia el tramo tardío, donde alcanza su máximo antes de caer al final. En contraste, el grupo "con ingresos" presenta un pico inicial más moderado, un valle más estable y un incremento tardío menos pronunciado, manteniéndose claramente por debajo de la curva de "sin ingresos" en todo el intervalo. En conjunto, los resultados sugieren ventanas de intervención temprana y tardía más críticas en el grupo "sin ingresos", mientras que la presencia de ingresos se asocia

con un perfil de riesgo sistemáticamente menor, consistente con mejores perspectivas de retención.

Gráfico 22.

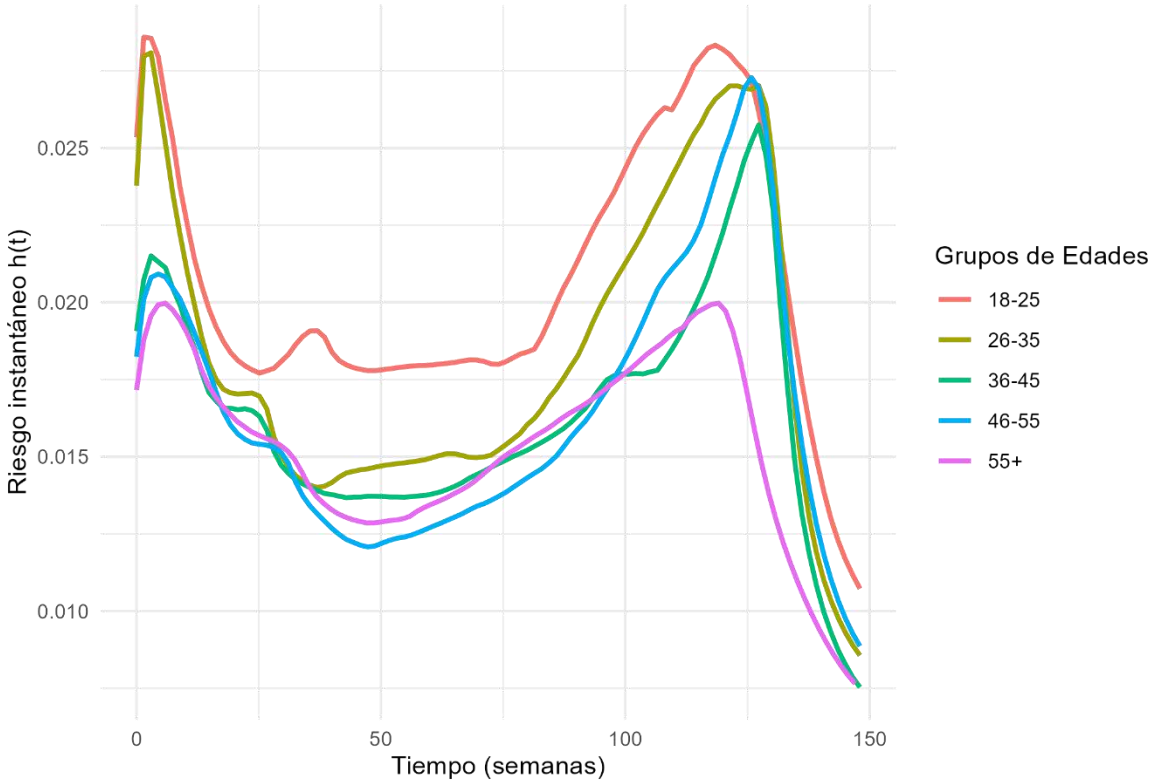
Riesgo instantáneo estimado no paramétricamente con suavizado para los distintos grupos de *Ingresos*



El Gráfico 23 muestra la tasa de riesgo instantáneo $h(t)$ estratificada por grupos de edad. La variable Edad (originalmente cuantitativa) se categorizó para facilitar la interpretación y mantener coherencia con el análisis de supervivencia del Gráfico 17 (Edades: 18–25, 26–35, 36–45, 46–55 y 55+). Para cada estrato se estimó $h(t)$ mediante el procedimiento no paramétrico mencionado anteriormente.

Gráfico 23.

Riesgo instantáneo estimado no paramétricamente con suavizado para los distintos grupos de Edades



Las curvas exhiben el patrón ya observado: pico temprano, valle intermedio, y alza tardía en las últimas semanas, seguido de un descenso final. En cuanto a orden relativo, el grupo 18–25 mantiene el riesgo más alto durante prácticamente todo el periodo; le sigue 26–35, también con niveles elevados de riesgo. Los grupos 36–45 y 46–55 conforman una franja intermedia con trayectorias muy cercanas entre sí. Finalmente, 55+ presenta de forma consistente el riesgo más bajo, con un ascenso tardío más moderado. En suma, el riesgo decrece con la edad: jóvenes (18–35) > intermedios (36–55) > mayores (55+), lo que es congruente con perfiles de permanencia relativamente mejores a edades más avanzadas.

4.3. MODELO SEMI-PARAMÉTRICO

En este capítulo se presentan los principales resultados obtenidos a partir del análisis con modelos semi-paramétricos de supervivencia, específicamente el modelo de riesgos proporcionales de Cox y su extensión, el modelo de Cox extendido. El objetivo de esta sección es evaluar la aplicabilidad de estos modelos para estudiar los factores asociados al riesgo de abandono de los usuarios en la plataforma digital.

En primer lugar, se realizó una evaluación rigurosa de los supuestos fundamentales del modelo de Cox, con especial atención al supuesto de riesgos proporcionales, así como a la linealidad de los efectos de las covariables continuas y a la ausencia de colinealidad entre covariables. Estas verificaciones resultaron clave para determinar la validez del modelo y la fiabilidad de sus estimaciones.

Como resultado de este análisis preliminar, se identificaron violaciones importantes del supuesto de riesgos proporcionales, lo cual limitó la interpretación de los efectos bajo el modelo clásico de Cox. Ante esta situación, se decidió no reportar resultados derivados directamente de dicho modelo, como coeficientes, razones de riesgos (*hazard ratios*) o medidas globales de ajuste, debido al riesgo de obtener estimaciones sesgadas.

Para atender esta limitación metodológica, se optó por ajustar un modelo de Cox extendido, el cual permite incorporar efectos no proporcionales mediante interacciones entre covariables y funciones del tiempo. Esta aproximación proporciona una mayor flexibilidad para capturar dinámicas temporales en el riesgo de abandono, facilitando una representación más precisa de los patrones observados durante el periodo de seguimiento.

En síntesis, este capítulo expone de manera estructurada el proceso de validación de los supuestos del modelo de Cox y presenta los resultados obtenidos mediante el modelo extendido, ofreciendo una perspectiva más robusta y mejor ajustada a la naturaleza dinámica del fenómeno bajo estudio.

4.3.1. Análisis de los supuestos del modelo

Antes de interpretar los resultados del modelo de riesgos proporcionales de Cox, es indispensable verificar que se cumplan los supuestos estadísticos que sustentan su validez. Estos supuestos permiten garantizar la confiabilidad de las estimaciones y la correcta interpretación de los efectos de las covariables sobre el riesgo de abandono.

El primer modelo ajustado, y cuyos supuestos serán evaluados a continuación, se representa mediante la siguiente expresión matemática, identificada como Ecuación E.53.

$$
\begin{aligned}
h\big(t,\, X(t)\big) \;=\; & h_0(t) * exp(\, \beta_1 \cdot \{Región\ California\ \&\ Nevada\} \\
& + \beta_2 \cdot \{Región\ Midwest\} \\
& + \beta_3 \cdot \{Región\ Northeast\} \\
& + \beta_4 \cdot \{Región\ Pacific\ Northwest\} \\
& + \beta_5 \cdot \{Región\ Rockies\ \&\ Red\ River\} \\
& + \beta_6 \cdot \{Region\ Southeast\} + \beta_7 \cdot Licencia_{LPN} \\
& + \beta_8 \cdot Licencia_{RN} + \beta_9 \cdot Licencia_{Sin\_Licencia} \\
& + \beta_{10} \cdot Licencia_{Otras} + \beta_{11} \cdot Ingreso \\
& + \beta_{12} \cdot Edad + \beta_{13} \cdot Edad\_Dicotomica
\end{aligned}
\qquad \text{Ecuación E.53}
$$

donde, $h_0(t)$ denota la función de riesgo basal, es decir, el riesgo instantáneo de abandono para un usuario de referencia, cuando todas las covariables toman el valor de la categoría de referencia o son iguales a cero. Los coeficientes $\beta_i (i = 1, \dots, 13)$ representan los efectos log-multiplicativos de cada covariable sobre el riesgo de abandono: al exponenciarlos, exp (β_i), se obtienen las razones de riesgos asociadas a cada categoría o unidad de incremento de la covariable, manteniendo constantes las demás.

El primer supuesto que se analiza en esta sección es la ausencia de colinealidad entre las covariables incluidas en el modelo. Este supuesto establece que las variables independientes no deben estar altamente correlacionadas entre sí, ya que una alta colinealidad puede generar inestabilidad en la estimación de los coeficientes, aumentar sus errores estándar, e incluso dificultar la identificación del efecto específico de cada covariable. En términos prácticos,

cuando dos o más variables explicativas contienen información redundante, el modelo tiene dificultades para asignar adecuadamente el peso de cada una en la predicción del riesgo.

Este supuesto ya fue abordado parcialmente en la sección "*4.1.2. Interacciones entre variables independientes*", donde se exploraron asociaciones entre covariables categóricas mediante pruebas chi-cuadrado y coeficientes *V* de Cramer. En ese análisis se identificaron algunas asociaciones estadísticamente significativas entre pares de variables (por ejemplo, entre *Región* e *Ingresos*), sin embargo, las magnitudes de las relaciones fueron generalmente bajas o moderadas. Estos hallazgos preliminares sugieren que, si bien hay cierto nivel de dependencia entre algunas variables, no se evidencian relaciones suficientemente fuertes como para anticipar colinealidad severa.

No obstante, en esta sección se profundiza en la verificación de la colinealidad mediante el uso del Factor de Inflación de la Varianza generalizado (GVIF), especialmente adecuado cuando el modelo incluye covariables categóricas. Esta revisión es importante para garantizar que el modelo de Cox, y posteriormente su versión extendida, no presenten colinealidades severas entre covariables, lo que podría afectar la precisión y estabilidad del análisis de supervivencia.

Para evaluar este supuesto en el modelo de riesgos proporcionales de Cox, se calculó el GVIF para cada término del modelo (ver Cuadro 22). Este indicador extiende el VIF tradicional al caso de términos con más de un grado de libertad, como las variables categóricas con varios niveles, y permite cuantificar la inflación de la varianza debida a la colinealidad.

Además, dado que el GVIF tiende a aumentar con el número de grados de libertad del término, se utilizó la versión ajustada, definida como $GVIF^{1/(2\cdot gl)}$, donde gl denota los grados de libertad de la covariable. Este reescalamiento sitúa todos los términos en una escala comparable, análoga a la del VIF clásico, y facilita la identificación de posibles problemas de colinealidad.

Cuadro 22.

Evaluación de la Colinealidad entre Covariables mediante el GVIF en el modelo de riesgos proporcionales de Cox

Variable	GVIF	G.L.	$\text{GVIF}^{1/(2\cdot \text{gl})}$
Región	1.11	6	1.02
Licencia	1.13	4	1.01
Ingresos	1.03	1	1.02
Edad	1.16	1	1.08
Sin_Edad	1.10	1	1.05

Como se observa en el Cuadro 22, los valores ajustados del GVIF para todas las covariables son considerablemente bajos, con un rango que va de 1.01 a 1.08. En general, valores menores a 2 se consideran aceptables, mientras que valores superiores a 5 podrían indicar problemas importantes de colinealidad.

En conclusión, los resultados obtenidos indican que no existen problemas relevantes de colinealidad entre las covariables incluidas en el modelo de Cox. Esto sugiere que las variables pueden ser incorporadas conjuntamente en el análisis sin necesidad de ser eliminadas o transformadas debido a redundancia o dependencia excesiva.

Luego, continuando con la evaluación de los supuestos fundamentales del modelo, se analiza el supuesto más relevante: el de riesgos proporcionales, el cual establece que la razón de riesgos (*hazard ratio*) entre los grupos permanece constante a lo largo del tiempo. Su incumplimiento puede sesgar los resultados e invalidar la interpretación de los efectos de las covariables.

Para evaluar este supuesto se aplicaron pruebas estadísticas, como la prueba global de proporcionalidad basada en residuos de Schoenfeld (Grambsch & Therneau, 1994), y métodos gráficos, incluyendo: Gráficos de residuos de Schoenfeld frente al tiempo para detectar patrones temporales; y Gráficos del logaritmo del riesgo acumulado (*hazard acumulado H(t))* $\left[\log\left(-\log\left(S(t)\right)\right) = \log\left(H(t)\right)\right]$ contra el logaritmo del tiempo (también llamados gráficos log-log), donde la proporcionalidad se verifica si las curvas son aproximadamente paralelas (Kleinbaum & Klein, 2005).

A continuación, en el Cuadro 23 se presentan los resultados de la prueba global de proporcionalidad de Grambsch y Therneau, aplicada al modelo de riesgos proporcionales de Cox, mediante los residuos de Schoenfeld para cada una de las covariables incluidas en dicho modelo.

Cuadro 23.

Prueba de Grambsch-Therneau para evaluar los riesgos proporcionales

Variable	Chi²	G. L.	Valor-p.
Región California & Nevada	0.142	1	0.7059
Región Midwest	14.086	1	0.0002
Región Northeast	6.275	1	0.0122
Región Pacific Northwest	5.527	1	0.0187
Región Rockies & Red River	22.290	1	< 0.0001
Región Southeast	1.595	1	0.2066
Licencia LPN	0.027	1	0.8696
Licencia RN	5.408	1	0.0200
Licencia Sin licencia	4.132	1	0.0421
Otras Licencias	0.490	1	0.4838
Ingresos	623.994	1	< 0.0001
Edad	1.662	1	0.1973
Sin_Edad	0.084	1	0.7713
Global	663.168	13	< 0.0001

De acuerdo con los resultados mostrados en el Cuadro 23, algunas categorías de la variable *Licencias* presentan valores p ligeramente significativos en la prueba de Grambsch-Therneau, siendo la categoría "CNA / GNA" la categoría de referencia para esta variable. Específicamente, las categorías "RN" ($p = 0.0200$) y "Sin licencia" ($p = 0.0421$) muestran niveles de significancia apenas por debajo del umbral de 0.05, lo cual podría indicar alguna variación en el efecto sobre el riesgo de abandono a lo largo del tiempo. Por el contrario, las categorías "LPN" ($p = 0.8696$) y "Otras" ($p = 0.4838$) no presentan significancia estadística, lo que sugiere un cumplimiento adecuado del supuesto de proporcionalidad.

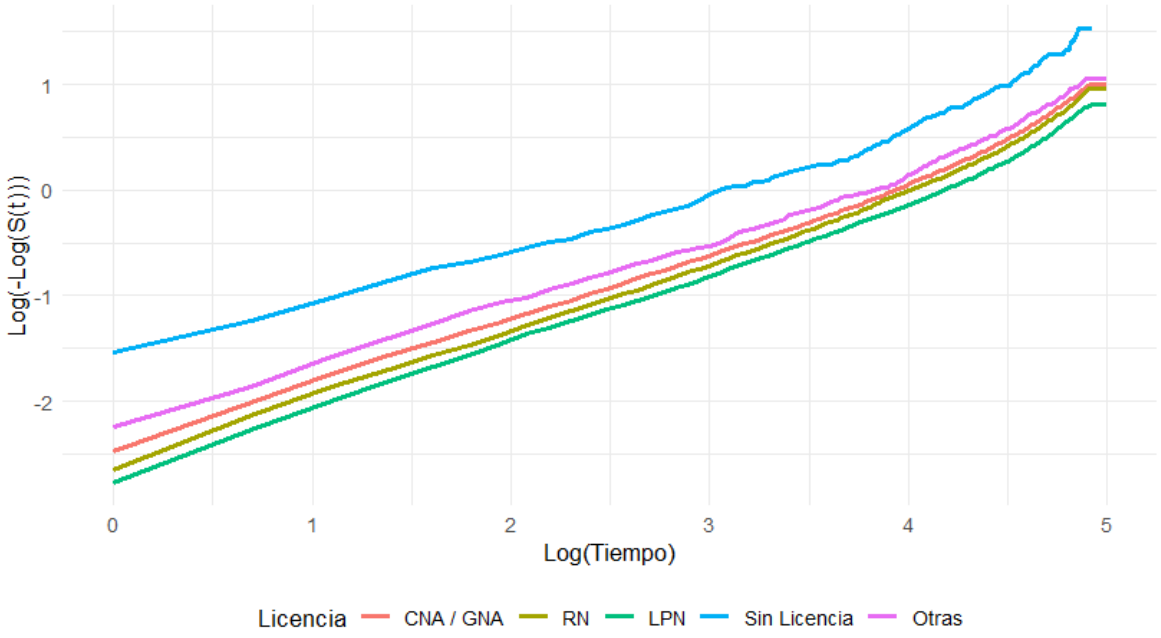
Complementando el análisis estadístico anterior de la variable *Licencias*, el Gráfico 24 muestra las curvas del logaritmo del riesgo acumulado contra el logaritmo del tiempo para

las distintas categorías de *Licencias* (generadas con la metodología de Kaplan-Meier). A nivel visual, las curvas presentan trayectorias en general paralelas, con pendientes similares, lo cual es consistente con el cumplimiento del supuesto de riesgos proporcionales. Sin embargo, se observa una leve divergencia en las últimas semanas para las categorías "RN" y "Sin licencia", aunque esta desviación es moderada y no evidencia un patrón sistemático de violación del supuesto.

En conjunto, tanto la prueba estadística como la evaluación gráfica sugieren que el supuesto de proporcionalidad de riesgos se cumple en términos generales para la variable *Licencias*, ya que las diferencias encontradas son pequeñas y las curvas mantienen una pendiente similar durante la mayor parte del periodo observado.

Gráfico 24.

Curvas del logaritmo del riesgo acumulado contra el logaritmo del tiempo por *Licencia* para evaluar la proporcionalidad de riesgos



De forma similar, en los resultados presentados en el Cuadro 23, algunas categorías de la variable Región muestran valores p que indican violaciones claras al supuesto de riesgos

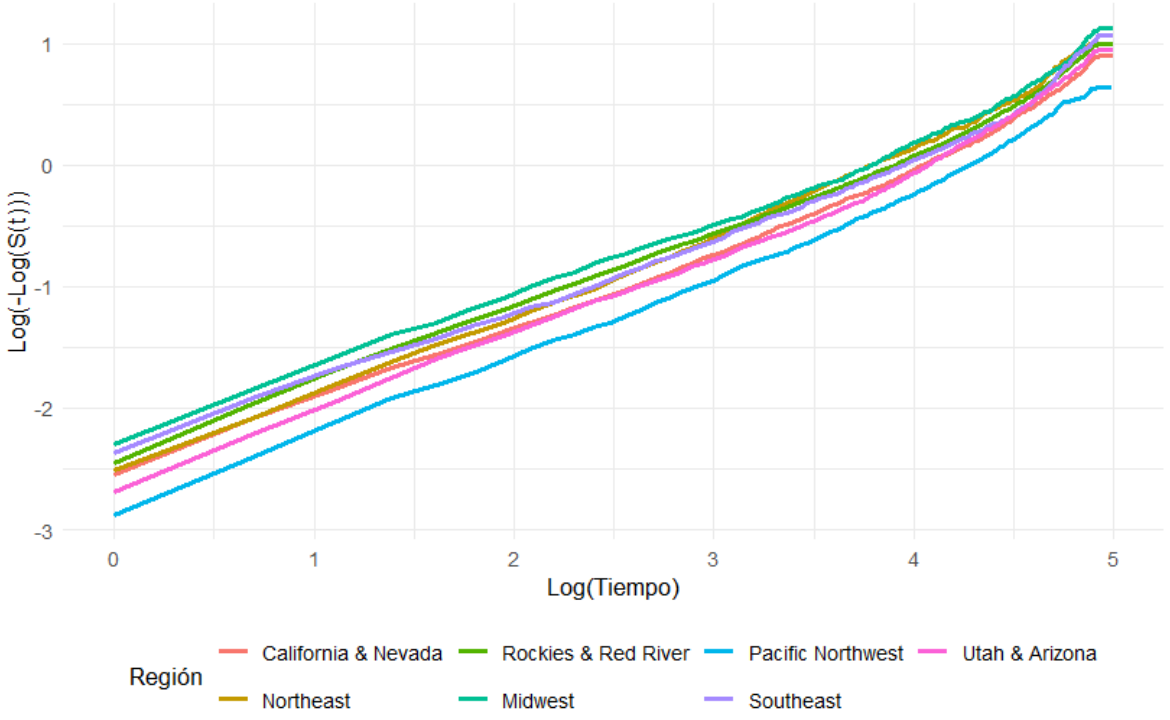
proporcionales. Específicamente, las categorías "Midwest" ($p = 0.0002$), "Northeast" ($p = 0.0122$), "Pacific Northwest" ($p = 0.0187$) y "Rockies & Red River" ($p < 0.001$) presentan valores estadísticamente significativos ($p < 0.05$), lo que sugiere que el efecto de estas regiones sobre el riesgo de abandono varía a lo largo del tiempo y, por tanto, no cumplen con el supuesto de proporcionalidad de riesgos. Por el contrario, las categorías "California & Nevada" ($p = 0.7059$) y "Southeast" ($p = 0.2066$) no presentan significancia estadística, lo cual indica un mejor cumplimiento del supuesto en estos grupos. Cabe señalar que la categoría de referencia para la variable Región es "Utah & Arizona", razón por la cual no aparece en el Cuadro 23.

Al complementar este análisis con el Gráfico 25, que muestra las curvas del logaritmo del riesgo acumulado frente al logaritmo del tiempo para las distintas regiones, se observa que la mayoría de las curvas mantienen trayectorias relativamente paralelas durante buena parte del período observado. Sin embargo, es posible identificar desviaciones más notables en algunas regiones, especialmente en las etapas iniciales y finales del seguimiento, lo cual es coherente con los resultados estadísticos obtenidos para "Midwest", "Northeast", "Pacific Northwest" y "Rockies & Red River". Estas categorías muestran variaciones en la pendiente de las curvas que respaldan la evidencia de incumplimiento del supuesto.

En conjunto, tanto los resultados estadísticos como los análisis gráficos sugieren que el supuesto de proporcionalidad de riesgos se incumple para las categorías "Midwest", "Northeast", "Pacific Northwest" y "Rockies & Red River", mientras que se mantiene razonablemente para "California & Nevada" y "Southeast". Esto señala la necesidad de considerar posibles ajustes al modelo, como la incorporación de términos de interacción con el tiempo o el uso de modelos extendidos, cuando se evalúan los efectos de las distintas regiones.

Gráfico 25.

Curvas del logaritmo del riesgo acumulado contra el logaritmo del tiempo por Región para evaluar la proporcionalidad de riesgos



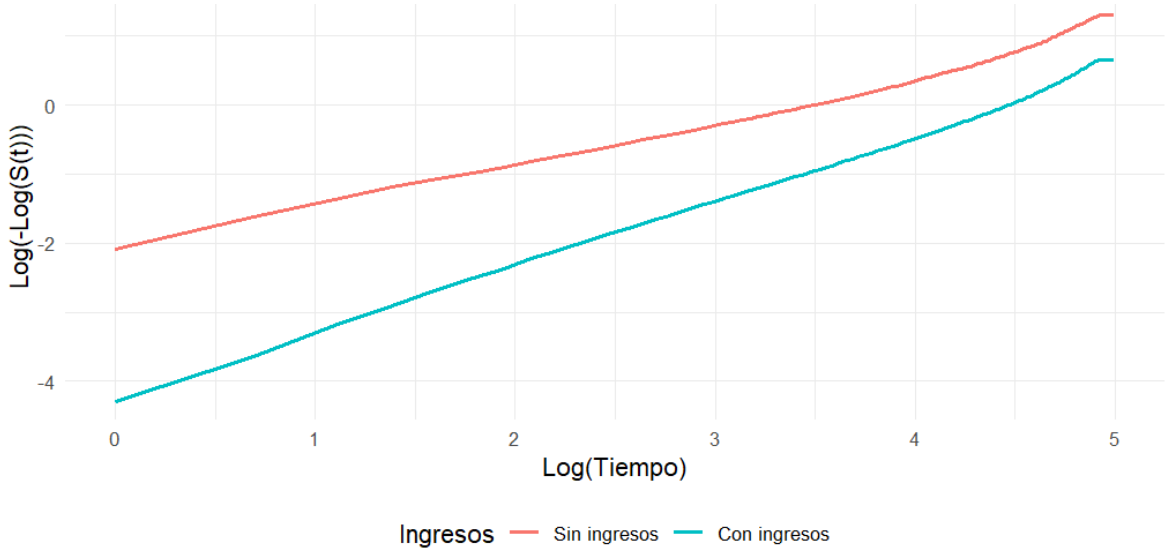
Por su parte, la variable Ingreso presenta evidencia clara de violación del supuesto de riesgos proporcionales. En el Cuadro 23, se observa un valor-p extremadamente bajo ($p < 0.001$) en la prueba de Grambsch-Therneau, lo cual indica que el efecto de esta variable sobre el riesgo de abandono varía significativamente a lo largo del tiempo.

Este resultado se complementa con lo mostrado en el Gráfico 26, donde se representan las curvas del logaritmo del riesgo acumulado frente al logaritmo del tiempo para las categorías "con ingresos" y "sin ingresos". Visualmente, las curvas no son paralelas, evidenciando pendientes claramente diferenciadas durante todo el periodo observado. Esta divergencia gráfica es coherente con la prueba estadística, confirmando que la razón de riesgos entre los grupos "con ingresos" y "sin ingresos" no se mantiene constante en el tiempo.

En conjunto, tanto el análisis estadístico como el gráfico respaldan que la variable Ingreso no cumple con el supuesto de proporcionalidad de riesgos, por lo que es recomendable considerar un modelo extendido o incorporar términos de interacción con el tiempo para capturar adecuadamente su efecto.

Gráfico 26.

Curvas del logaritmo del riesgo acumulado contra el logaritmo del tiempo para la variable *Ingreso* para evaluar la proporcionalidad de riesgos



En el caso de la variable Edad, el Cuadro 23 muestra un valor-p de 0.1973, el cual no es estadísticamente significativo al nivel habitual de 0.05. Esto sugiere que no hay evidencia suficiente para rechazar el supuesto de proporcionalidad de riesgos para esta covariable. Por tanto, con base en los resultados estadísticos, se concluye que la variable Edad cumple con el supuesto de proporcionalidad de riesgos, ya que no presenta variación significativa en su efecto sobre el riesgo de abandono a lo largo del tiempo.

En los resultados anteriores, se evidencia y concluye que el modelo de riesgos proporcionales de Cox presenta violaciones al supuesto de proporcionalidad de riesgos en varias covariables incluidas en el modelo inicial. Estas evidencias sugieren que los efectos de ciertas covariables

sobre el riesgo de abandono no se mantienen constantes a lo largo del tiempo, lo cual incumple uno de los principios básicos del modelo de Cox clásico.

Dado que el supuesto de riesgos proporcionales es esencial para garantizar la validez de las estimaciones derivadas de dicho modelo, y considerando que su incumplimiento podría conducir a sesgos en los coeficientes, interpretaciones erróneas de los efectos covariables y un mal ajuste global del modelo, se decidió no continuar con la interpretación de los resultados obtenidos bajo la formulación tradicional del modelo de Cox.

En consecuencia, se optó por la utilización de un modelo de Cox extendido, también conocido como modelo de Cox con efectos que varían en el tiempo. Esta alternativa metodológica permite capturar la dinámica temporal de los efectos de las covariables sobre el riesgo de abandono, mediante la incorporación explícita de términos de interacción entre covariables y funciones del tiempo.

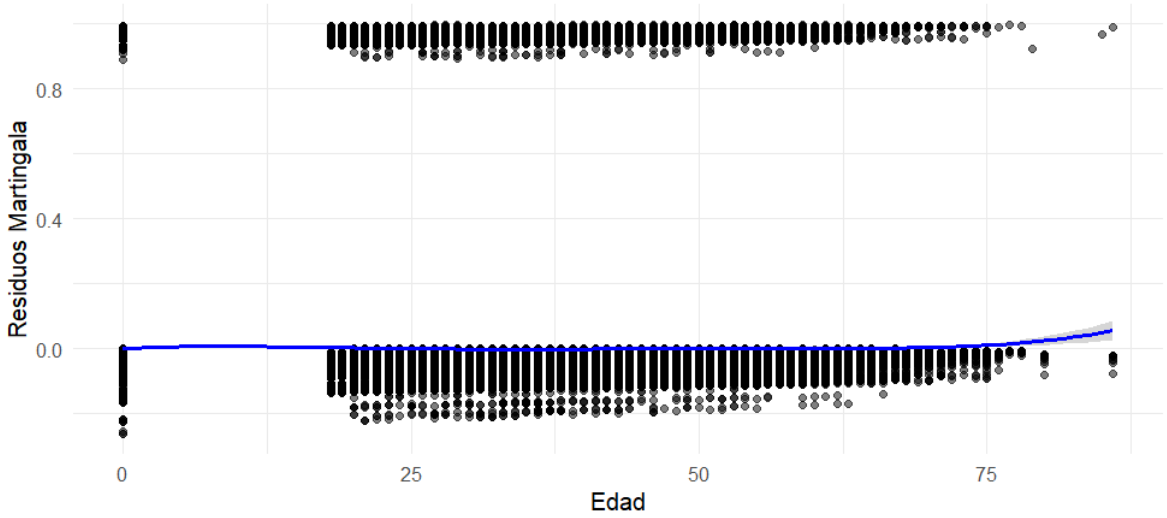
Como consecuencia directa de este cambio de enfoque, a partir de este punto todas las verificaciones estadísticas, evaluaciones de supuestos, y demás procedimientos de validación del modelo semi-paramétrico se realizarán sobre la base del modelo de Cox extendido. Esto incluye la revisión de la linealidad de las covariables continuas en relación con el logaritmo del riesgo, así como las verificaciones estadísticas del ajuste global del modelo.

Continuando con la verificación de los supuestos del modelo extendido de Cox, se analiza el supuesto de linealidad entre la covariable continua *Edad* y el logaritmo del *hazard acumulado*. Este supuesto establece que, para que los coeficientes estimados sean válidos, debe existir una relación lineal entre la covariable numérica y la función de riesgo, específicamente en la escala del predictor lineal del modelo.

En el presente análisis, la variable *Edad* es la única covariable independiente de tipo numérico incluida en el modelo, por lo que el cumplimiento del supuesto de linealidad se verifica mediante un gráfico que relaciona los residuos de Martingala con los valores observados de *Edad* (Gráfico 27). La curva suavizada azul, incluida en el Gráfico 27, fue ajustada utilizando un polinomio de quinto grado con bandas de error estándar, con el fin de facilitar la detección visual de patrones no lineales.

Gráfico 27.

Relación entre los residuos de Martingala y la variable *Edad* para evaluar el supuesto de linealidad del modelo extendido de Cox



Nota: La curva suavizada azul, fue ajustada utilizando un polinomio de quinto grado con bandas de error estándar.

El Gráfico 27 muestra que la línea azul suavizada permanece en su mayoría próxima a cero y no evidencia patrones sistemáticos de curvatura a lo largo del eje de *Edad*. Aunque se presenta una leve desviación hacia el extremo superior derecho, esta no es lo suficientemente marcada como para indicar una violación significativa del supuesto.

En conjunto, la ausencia de una tendencia claramente no lineal sugiere que la relación entre la variable *Edad* y el logaritmo del riesgo acumulado puede considerarse aproximadamente lineal, cumpliéndose así con este supuesto fundamental del modelo de Cox.

4.3.2. Análisis de los Coeficientes del Modelo Extendido de Cox

En esta sección se presentan los resultados del análisis de los coeficientes estimados a partir del modelo de Cox extendido. Se examina su significancia estadística y se interpretan sus implicaciones en términos del riesgo de abandono (razón de riesgos). Asimismo, se incluyen las interacciones entre covariables y funciones del tiempo, con el objetivo de capturar

posibles variaciones temporales en los efectos de estas variables, lo cual enriquece la comprensión del modelo.

Se consideran tanto los coeficientes base (β) como aquellos asociados al término dependiente del tiempo (δ), evaluando su contribución al ajuste del modelo y la magnitud de su efecto. Además, se emplea la transformación exponencial de los coeficientes ($\exp(\beta)$) como medida de razón de riesgos (*hazard ratio*), lo que facilita una interpretación más directa y práctica de los resultados.

Para seleccionar las variables del modelo de Cox extendido, se aplicó una estrategia de selección jerárquica basada en comparaciones mediante pruebas de verosimilitud (*Likelihood Ratio Test*). Esta técnica, ampliamente utilizada en modelos de regresión, resulta especialmente útil cuando se trabaja con modelos anidados, es decir, cuando un modelo es una versión simplificada del otro. Inicialmente, se incluyeron todas las covariables consideradas y analizadas en las secciones previas, así como las interacciones sugeridas en el análisis de riesgos proporcionales. Posteriormente, se fueron excluyendo aquellas variables o términos que no aportaban significativamente al ajuste del modelo, evaluando en cada paso si la simplificación afectaba el desempeño estadístico del mismo.

Por consiguiente, la interacción entre la covariable región "Pacific Northwest" y el término dependiente del tiempo log(t+1) presentó un valor-p de 0.87, por lo que fue eliminada en la primera iteración. Seguidamente, se encontró que la covariable región "Southeast", aunque su valor-p fue de 0.61, se decidió mantenerla en el modelo debido a que aporta información sustantiva y relevante sobre la ubicación laboral de los usuarios en la plataforma, lo cual resulta clave para el análisis. Posteriormente, la interacción entre la covariable región "Northeast" y el término dependiente del tiempo log(t+1) obtuvo un valor-p de 0.16, por lo que fue eliminada del modelo.

A continuación, en la Ecuación E.54 se presenta la ecuación del modelo extendido de Cox finalmente ajustado.

$$
\begin{aligned}
h\big(t,\, X(t)\big) \;=\; & h_0(t) * exp(\, \beta_1 \\
& \cdot \{Región\ California\ \&\ Nevada\} \;+\; \beta_2 \\
& \cdot \{Región\ Midwest\} \;+\; \delta_2 \;\cdot \{Región\ Midwest\} \\
& \cdot \; log\ (t+1) \;+\; \beta_3 \;\cdot \{Región\ Northeast\} \;+\; \beta_4 \\
& \cdot \{Región\ Pacific\ Northwest\} + \; \beta_5 \\
& \cdot \{Región\ Rockies\ \&\ Red\ River\} \;+\; \delta_5 \\
& \cdot \{Región\ Rockies\ \&\ Red\ River\} \;\cdot\; log\ (t+1) \\
& +\; \beta_6 \cdot \{Region\ Southeast\} + \beta_7 \cdot Licencia_{LPN} \\
& +\; \beta_8 \cdot Licencia_{RN} +\; \beta_9 \cdot Licencia_{Sin\_Licencia} \\
& + \beta_{10} \cdot Licencia_{Otras} + \beta_{11} \cdot Ingreso + \delta_{11} \\
& \cdot Ingreso \cdot \log(t+1) + \beta_{12} \cdot Edad + \beta_{13} \\
& \cdot Sin\_Edad
\end{aligned}
$$

Ecuación E.54

Donde, de forma análoga al modelo de la Ecuación E.53, $h_0(t)$ denota la función de riesgo basal, los coeficientes β_i representan los efectos log-multiplicativos "promedio" de cada covariable sobre el riesgo de abandono, y los coeficientes δ_j cuantifican los efectos que varían en el tiempo mediante las interacciones entre ciertas covariables y $log\ (t+1)$.

Seguidamente, en el Cuadro 24, se presentan los resultados del modelo de Cox extendido, en este se resume los coeficientes estimados, sus errores estándar (ee), los estadísticos z, y los valores p.

Cuadro 24.

Resultados para los coeficientes del modelo de Cox extendido

Variable o interacción	coef	exp(coef)	ee(coef)	z	Valor-p
R. California & Nevada	-0.035	0.966	0.019	-1.88	0.060
R. Midwest	0.238	1.268	0.041	5.74	< 0.001
R. Midwest*log$(t+1)$	-0.041	0.960	0.014	-2.97	0.003
R. Northeast	0.167	1.181	0.021	7.82	< 0.001
R. Pacific Northwest	-0.187	0.829	0.022	-8.43	< 0.001
R. Rockies & Red River	0.1823	1.200	0.030	5.97	< 0.001
R. Rockies & Red River*log$(t+1)$	-0.039	0.961	0.009	-4.21	< 0.001
R. Southeast	-0.012	0.988	0.025	-0.46	0.646
Licencia LPN	-0.174	0.840	0.015	-11.59	< 0.001
Licencia RN	-0.047	0.954	0.016	-3.00	0.003
Licencia "Sin licencia"	0.520	1.682	0.049	10.62	< 0.001
Otras Licencias	0.070	1.073	0.033	2.16	0.0306
Ingresos	-1.373	0.253	0.034	-39.77	< 0.001
Ingresos*log$(t+1)$	0.276	1.317	0.011	26.30	< 0.001
Edad	-0.005	0.995	0.001	-9.15	< 0.001
Sin_Edad	0.120	1.128	0.063	1.91	0.056

coef: coeficiente; ee(coef): error estándar del coeficiente.

Como se evidencia en el Cuadro 24, la gran mayoría de las variables e interacciones presentan valores p bajos (menores a 0.05), lo cual constituye evidencia estadística de que sus efectos son significativamente distintos de cero. Esto justifica su inclusión en el modelo y su consideración en el análisis. Es particularmente relevante la significancia de las interacciones con la función logarítmica del tiempo $[log(t+1)]$, ya que indican que el efecto de ciertas covariables, como las regiones "Midwest" y "Rockies & Red River", así como la variable *ingresos*, varía a lo largo del tiempo. Esta característica respalda el uso de un modelo de Cox extendido, el cual permite modelar efectos no proporcionales.

Por otro lado, dos variables presentan valores p un poco mayores a 0.05: "Sin_Edad" y la "Región California & Nevada". Al no alcanzar el umbral de 0.05 comúnmente utilizado de significancia estadística, sus efectos no se consideran significativos en este modelo. Sin

embargo, sus valores p (0.0564 y 0.060, respectivamente) son cercanos al umbral de 0.05. En este contexto, su inclusión en el modelo puede justificarse por su relevancia teórica y práctica, ya que ambas variables se asocian con características importantes de los usuarios de la plataforma digital, y ayudan a la interpretación más completa de los resultados.

Adicionalmente, se decidió mantener en el modelo la variable dicotómica "Región Southeast", a pesar de presentar un valor-p de 0.646, dado que constituye un componente esencial dentro de la variable más general de *Región*. Su permanencia resulta pertinente no solo por la importancia sustantiva de esta categoría geográfica, sino también porque su inclusión no altera la significancia y los valores de los coeficientes de las demás covariables. En consecuencia, se consideró más beneficioso para este estudio conservarla en el modelo en lugar de excluirla.

En el modelo de Cox extendido, los coeficientes estimados representan el efecto logarítmico de cada covariable sobre el riesgo de abandono (*hazard*). Por su parte, estos coeficientes exponenciados corresponden a las razones de riesgo (*hazard ratios*, HR), que permiten una interpretación más intuitiva en términos relativos.

La región tomada como referencia en este modelo es "Utah & Arizona", por lo cual todas las demás regiones se interpretan en comparación con dicha categoría base.

En el caso específico de la región "California & Nevada", se observa una razón de riesgo (HR) de 0.966, lo que apunta a un efecto protector pequeño frente a la región de referencia. No obstante, la estimación presenta incertidumbre estadística: con $p = 0.06$, no se rechaza la hipótesis nula de coeficiente igual a cero (o, de forma equivalente, $HR = 1$) al nivel de significancia establecido ($\alpha = 0.05$), por lo que la evidencia disponible no permite afirmar que exista una diferencia real en el riesgo de abandono para esta región.

De forma similar, la región "Southeast" presenta una razón de riesgo ($HR = 0.988$) muy cercana a 1, con un valor-p de 0.646, lo que indica que la diferencia respecto a la categoría de referencia "Utah & Arizona" no es estadísticamente significativa. En otras palabras, los usuarios que trabajan en la región "Southeast" muestran un comportamiento de abandono

prácticamente equivalente al de la región base, sin evidencias claras de un mayor o menor riesgo.

En cuanto a la región "Northeast", se obtiene una razón de riesgo (HR) de 1.181, revelando que los usuarios en esta zona tienen un riesgo 18.1% mayor de abandonar la plataforma en comparación con la categoría de referencia, resultado estadísticamente significativo ($p < 0.001$). Por otro lado, la región "Pacific Northwest" muestra una HR de 0.829, lo que implica una reducción del riesgo del 17.1% en relación con la región base, resultado que también es estadísticamente significativo ($p < 0.001$).

Por otra parte, las regiones "Midwest" y "Rockies & Red River" presentan un efecto dependiente del tiempo, al igual que la variable *Ingresos*. En el caso de "Midwest", la razón de riesgos puede expresarse como $HR(t) = \exp\{0.238 - 0.041 * \log(t+1)\}$, donde tanto el coeficiente principal como el término de interacción con $\log(t+1)$ resultan estadísticamente significativos. Al inicio del periodo, "Midwest" exhibe un riesgo de abandono mayor que la región de referencia ("Utah & Arizona"); sin embargo, el coeficiente temporal negativo indica que este exceso de riesgo disminuye gradualmente conforme transcurren las semanas.

Para facilitar la interpretación de los efectos que varían en el tiempo, a continuación se reportan los HR de estas tres variables en cuatro momentos representativos: t = 0 (inicio), 8 semanas (aproximadamente 2 meses), 52 semanas (1 año) y 104 semanas (2 años).

Variable	HR Base	HR (8 semanas)	HR (52 semanas)	HR (104 semanas)
Midwest	1.268	1.159	1.078	1.048
Rockies & Red River	1.200	1.101	1.028	1.000
Ingresos	0.253	0.465	0.758	0.915

Por su parte, a partir de los coeficientes mostrados en el Cuadro 24, la región "Rockies & Red River" presenta un efecto que varía con el tiempo, cuya razón de riesgo puede expresarse como $HR(t) = \exp\{0.182 - 0.039 * \log(t+1)\}$, la cual también se representa en la tabla anterior. En términos interpretativos, al inicio del periodo los usuarios de esta región muestran un riesgo de abandono aproximadamente 20.0% mayor que el de la región de

referencia ("Utah & Arizona"), diferencia que resulta estadísticamente significativa ($p < 0.001$). Sin embargo, el término de interacción con el tiempo indica que este riesgo disminuye de forma gradual conforme transcurren las semanas. En otras palabras, aunque el riesgo inicial sea relativamente alto, la asociación se debilita paulatinamente a medida que los usuarios acumulan más tiempo en la plataforma, lo cual respalda la inclusión de esta interacción en el modelo extendido.

En resumen, las regiones con mayor riesgo de abandono, al inicio de uso de la plataforma, son "Midwest" (HR base = 1.268) y "Rockies & Red River" (HR base = 1.200); no obstante, en ambos casos este efecto se atenúa con el tiempo, por lo que la brecha frente a las otras regiones tiende a reducirse conforme avanza el seguimiento. En contraste, "Pacific Northwest" muestra un riesgo claramente menor (HR = 0.829), lo que sugiere una mayor permanencia relativa de sus usuarios. Estos resultados indican que las estrategias de retención podrían priorizar, especialmente en las primeras etapas de uso de la plataforma, a los usuarios de "Midwest" y "Rockies & Red River", mientras que en "Pacific Northwest" las condiciones actuales parecen favorecer naturalmente la permanencia de los usuarios en la plataforma.

La variable *Ingresos* es una covariable dicotómica: toma el valor "0" (cero) si el usuario no generó ingresos durante sus primeros dos meses de actividad y "1" (uno) si sí los generó en ese periodo. En el modelo, este efecto varía con el tiempo y se expresa como: $HR(t) = \exp\{-1.373 + 0.276 * \log(t+1)\}$, lo cual también se resume en la tabla anterior para los tiempos seleccionados. En términos interpretativos, al inicio del periodo los usuarios que generaron ingresos presentan un riesgo de abandono aproximadamente 74.7% menor que quienes no generaron ingresos, constituyéndose en uno de los efectos más relevantes del modelo ($p < 0.001$). No obstante, el término de interacción con el tiempo indica que este efecto protector se debilita conforme transcurren las semanas. Es decir, aunque inicialmente los usuarios "con ingresos" muestran una mayor permanencia en la plataforma, esa ventaja relativa tiende a atenuarse gradualmente con el paso del tiempo.

En relación con las variables asociadas al tipo de licencia laboral, la categoría de referencia considerada en este modelo es Licencia "CNA/GNA", por lo que todas las demás categorías de licencias se interpretan en comparación con este grupo.

El grupo de usuarios con licencia "LPN" presentan una razón de riesgo (HR) de 0.8402, lo que implica un riesgo de abandono 15.98% menor respecto al grupo base, efecto significativo ($p < 0.001$). De manera similar, los usuarios con licencia "RN" muestran una HR de 0.9539, equivalente a una reducción leve del 4.6% en el riesgo de abandono, y estadísticamente significativa ($p = 0.003$).

En contraste, los usuarios clasificados como "Sin licencia" exhiben un riesgo de abandono sustancialmente mayor (HR = 1.6825), es decir, 68.3% superior al de la categoría de referencia, lo que refleja una clara vulnerabilidad frente al abandono temprano y constituye un efecto significativo ($p < 0.001$). Por su parte, la categoría "Otras" licencias muestra una HR de 1.0729, que representa un incremento moderado del 7.3% en el riesgo de abandono, resultado también significativo ($p = 0.0306$).

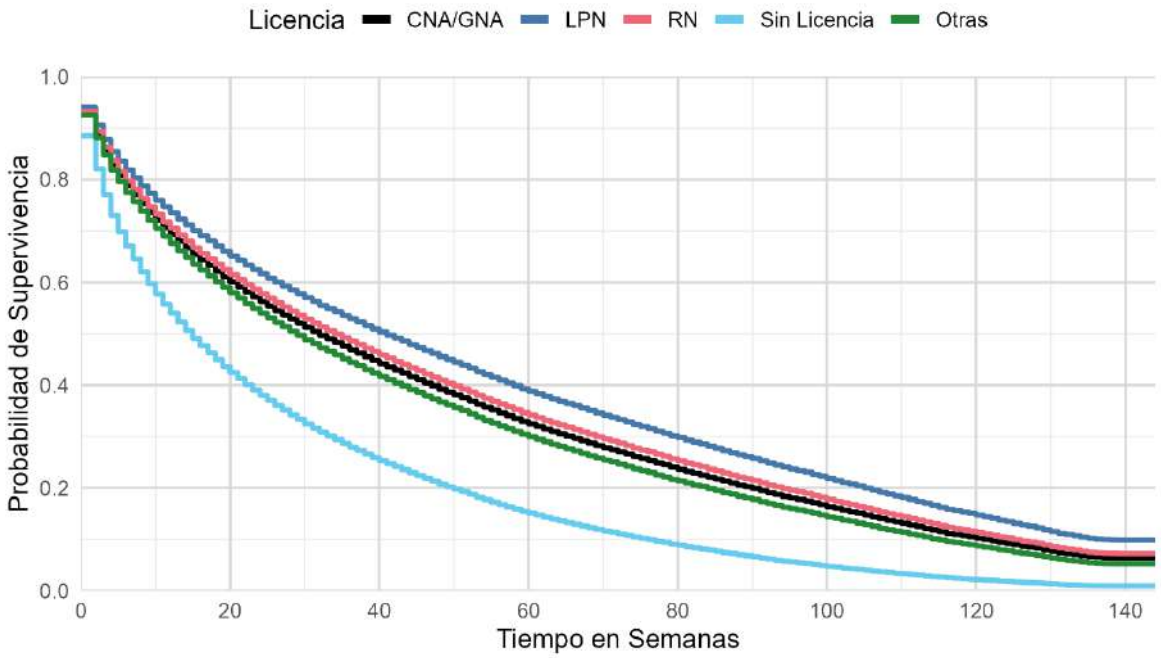
En síntesis, los usuarios "Sin licencia" son quienes enfrentan el mayor riesgo de abandono, mientras que el grupo con licencia "LPN" presenta los riesgos más bajos, lo cual sugiere una menor propensión al *churn*. Finalmente, los usuarios con licencias "RN", "CNA/GNA", y "Otras" ocupan una posición intermedia, con riesgos de abandono similares. En términos prácticos, estos resultados resaltan la importancia de considerar el tipo de licencia como un factor determinante en la retención de usuarios dentro de la plataforma, sugiriendo la necesidad de diseñar estrategias diferenciadas de acuerdo con el perfil profesional de cada grupo.

4.3.3. Gráficos de las funciones de supervivencia y riesgo (Modelo extendido de Cox)

A continuación, se presenta el Gráfico 28, el cual muestra las curvas de supervivencia estimadas para cada tipo de licencia profesional, obtenidas con el modelo extendido de Cox. Estas curvas permiten comparar, para cualquier semana, la probabilidad de que los usuarios permanezcan activos más allá de dicho tiempo. Curvas más altas indican mayor permanencia esperada, y diferencias sostenidas entre curvas sugieren efectos diferenciados por licencia a lo largo del tiempo.

Gráfico 28.

Curvas de supervivencia según tipo de *licencia* obtenidas con el Modelo de Cox extendido

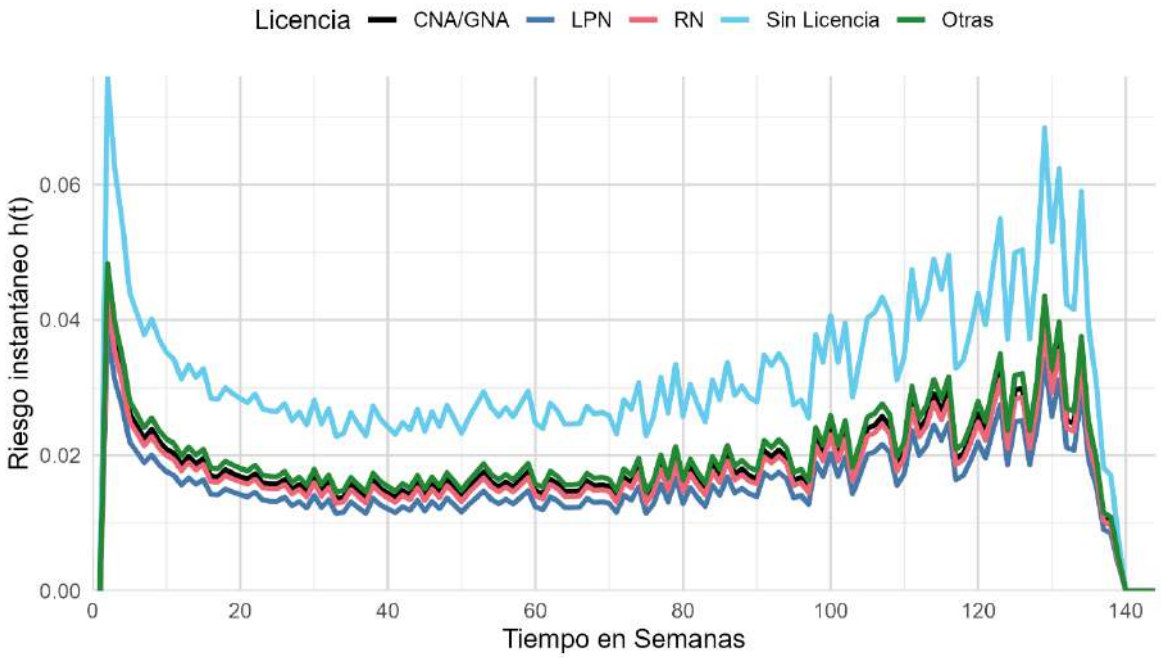


En el mencionado Gráfico 28 se aprecia que los usuarios con licencia "LPN" presentan, de forma consistente, probabilidades de supervivencia más altas a lo largo del horizonte analizado; en contraste, el grupo "Sin licencia" exhibe probabilidades considerablemente más bajas en comparación con las demás categorías, reflejando mayor propensión al abandono. Las licencias "RN" y "CNA/GNA" muestran trayectorias intermedias y próximas entre sí, mientras que la categoría de "Otras" tiende a situarse levemente por debajo de la base. Este patrón es coherente con la evidencia no paramétrica previa (Kaplan-Meier) que ubica a "LPN" como el grupo de mejor permanencia y a "Sin licencia" como el de mayor abandono.

Por su parte, el Gráfico 29 presenta el riesgo instantáneo $h(t)$ para cada tipo de licencia. Esta función representa la tasa de abandono en el instante t, condicionada a seguir activo justo antes de t. Picos en $h(t)$ identifican ventanas críticas de mayor abandono, mientras que mesetas o descensos reflejan atenuación del riesgo con el tiempo.

Gráfico 29.

Riesgo instantáneo según tipo de *licencia* obtenido con el Modelo extendido de Cox



En dicho Gráfico 29 se observan valores de $h(t)$ más elevados para la categoría "Sin licencia", con picos tempranos que sugieren necesidad de intervenciones en las primeras semanas. El grupo "LPN" mantiene los niveles más bajos de $h(t)$ a lo largo del periodo, mientras que "RN" y "CNA/GNA" presentan riesgos moderados y más estables. Por su parte, la categoría "Otras" licencias se sitúa ligeramente por encima de la base, pero muy por debajo de "Sin licencia".

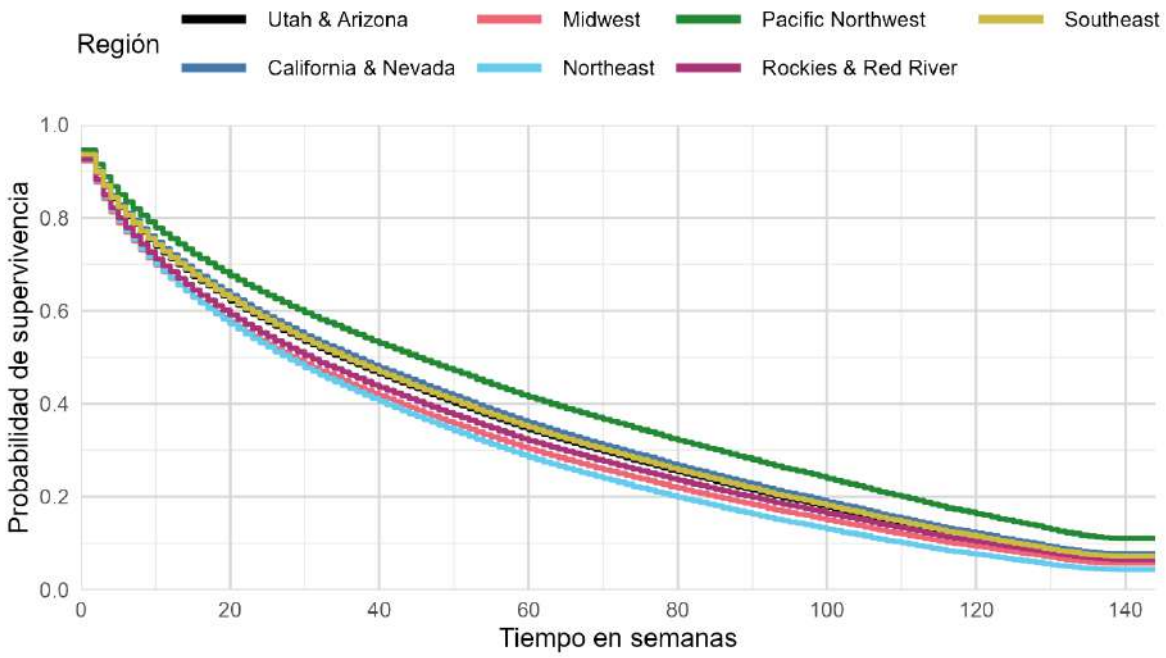
Se evidencia que los periodos de mayor riesgo de abandono se concentran en dos ventanas temporales. La primera ocurre muy temprano (alrededor de las semanas 0–6), con un pico pronunciado en el grupo "Sin licencia"; esta fase sugiere la necesidad de acciones de incorporación y acompañamiento inicial de los usuarios. La segunda corresponde a un repunte tardío (alrededor de las semanas 100–135), donde el riesgo aumenta para todas las licencias.

Los periodos de menor riesgo instantáneo se observan en el tramo intermedio (alrededor de las semanas 20–80), cuando las tasas se estabilizan en niveles bajos (del orden de 0.02–0.03 por semana), con "LPN" manteniéndose sistemáticamente como el grupo de menor riesgo, seguido por "RN" y "CNA/GNA". El descenso final cercano a la semana 140 debe interpretarse con cautela, pues probablemente responde a un efecto de borde del periodo de estudio y no a un cambio real en el comportamiento de los usuarios de la plataforma.

Por otra parte, el siguiente Gráfico 30, compara las curvas de supervivencia $S(t)$ entre regiones. Su lectura es análoga al Gráfico 28: curvas más altas indican mayor retención de los usuarios y diferencias persistentes a lo largo del tiempo sugieren efectos regionales en la permanencia de los mismos.

Gráfico 30.

Curvas de supervivencia según Región obtenidas con el Modelo extendido de Cox



En el Gráfico 30 se observan diferencias claras entre regiones: "Northeast" muestra la menor supervivencia relativa (curva más baja) durante la mayor parte del periodo, mientras que

"Pacific Northwest" se ubica en niveles más altos de supervivencia a lo largo de prácticamente todo el horizonte analizado. Por su parte, "Southeast" aparece muy similar a "Utah & Arizona", coherente con la ausencia de significancia estadística de su efecto temporal en las pruebas de proporcionalidad. Además, aunque no es fácil apreciarlo solo con el gráfico, "Rockies & Red River" inicia con niveles de supervivencia comparativamente bajos, pero conforme transcurren las semanas sus probabilidades de supervivencia aumentan y convergen hacia valores intermedios del conjunto. Esto se debe a que su efecto varía en el tiempo. En "Rockies & Red River" la razón de riesgos relativa está dada por $\mathrm{HR(t)} = \exp\{\, 0.182 - 0.039\ \log(t+1)\}$.

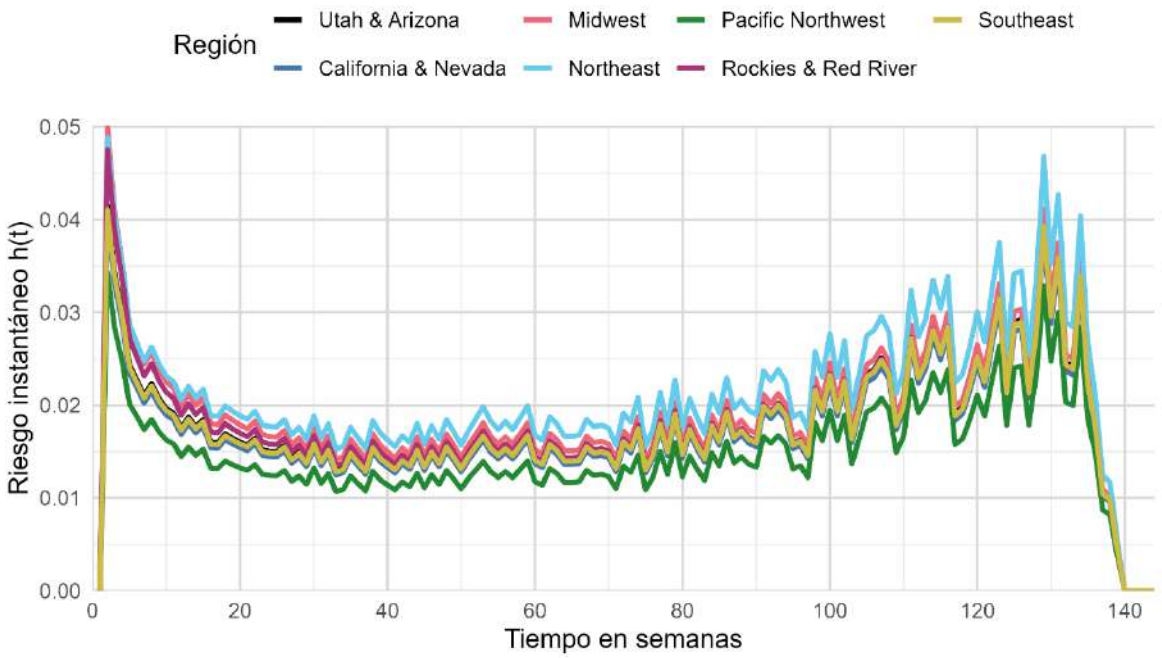
Al inicio $(t \approx 0)$ $HR \approx \mathrm{e}^{0.182} \approx 1.20$ (mayor riesgo y, por tanto, menor supervivencia respecto a la referencia). Sin embargo, el término temporal negativo $-0.039 * log(t+1)$ reduce el riesgo relativo conforme pasa el tiempo. La HR cruza 1 alrededor de la semana 106 y hacia el final del periodo se sitúa ligeramente por debajo de 1, lo que explica que su curva se acerque, e incluso quede levemente por encima, de la de la referencia (Utah & Arizona).

De forma similar, la región "Midwest" también presenta un efecto que varía con el tiempo. En este caso, la razón de riesgos relativa está dada por $HR(t) = \exp\{0.238 - 0.041\log(t+1)\}$, donde tanto el coeficiente principal como el término de interacción con $\log(t+1)$ resultan estadísticamente significativos. Al inicio del periodo, "Midwest" exhibe un riesgo mayor que la región de referencia (Utah & Arizona), pero el coeficiente temporal negativo indica que ese exceso de riesgo se atenúa gradualmente conforme avanzan las semanas.

De forma similar, se presenta el Gráfico 31, que muestra el riesgo instantáneo $h(t)$ por región. Este gráfico permite detectar picos temporales de mayor abandono y valorar si el riesgo relativo entre regiones cambia con el tiempo (consistente con el uso del modelo extendido).

Gráfico 31.

Riesgo instantáneo según Región obtenido con el Modelo extendido de Cox

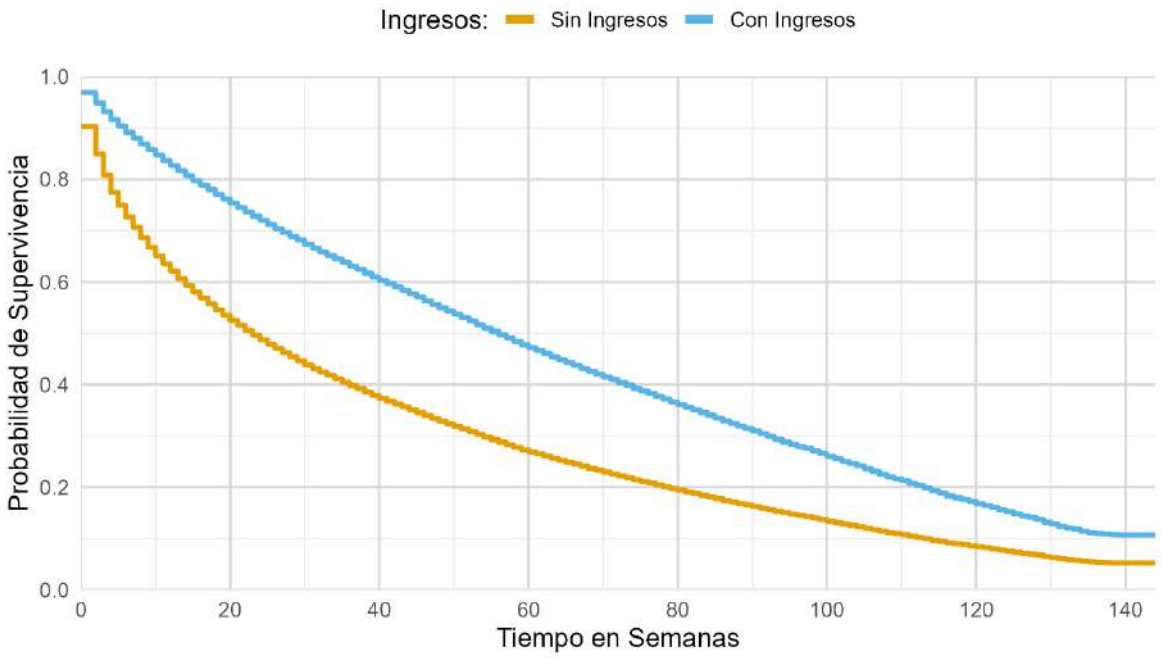


En el Gráfico 31, la región "Northeast" y "Midwest" concentran los picos más altos de $h(t)$ en la fase inicial (alrededor de las semanas 0–10), "Northeast" se mantiene como la región de mayor riesgo relativo durante la mayor parte del periodo. Tras el pico temprano, las tasas descienden y se estabilizan en un tramo intermedio (alrededor de las semanas 20–80), con diferencias moderadas entre regiones. Donde, "Pacific Northwest" exhibe de forma persistente los niveles más bajos. Hacia el final (alrededor de las semanas 100–135) aparece un repunte general en el que se preserva el orden relativo, "Northeast" arriba y "Pacific Northwest" abajo, mientras "Southeast" y "Utah & Arizona" permanecen cercanas entre sí.

Seguidamente, se presenta el Gráfico 32, que compara las curvas de supervivencia para la variable *Ingresos* (con las categorías "con ingresos" y "sin ingresos"). La distancia vertical entre ambas curvas informa sobre la ventaja relativa en permanencia asociada a generar ingresos y cómo dicha ventaja evoluciona con el tiempo.

Gráfico 32.

Curvas de supervivencia según *Ingresos* obtenidas con el Modelo extendido de Cox



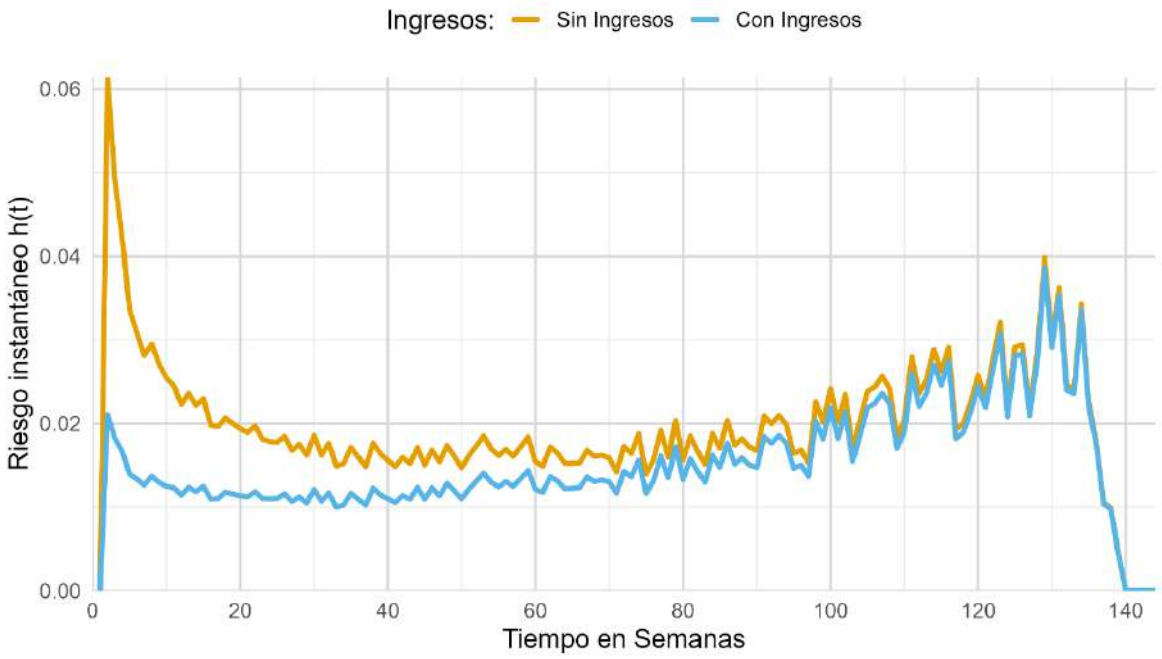
En dicho Gráfico 32, el grupo "con ingresos" inicia con mayor supervivencia (curva superior) y la brecha se atenúa conforme avanza el tiempo, lo cual concuerda con la interacción de la variable con el tiempo: el efecto protector disminuye a lo largo del seguimiento (HR de la interacción > 1).

$$\frac{h(\,t \mid \text{Ingresos} = 1\,)}{h(\,t \mid \text{Ingresos} = 0\,)} = \exp\{-1.373 + 0.276\,\log(t+1)\} = 0.253\,(t+1)^{0.276}$$

Continuando con la variable *Ingresos*, el Gráfico 33, presenta el riesgo instantáneo $h(t)$ para esta variable ("con ingresos" y "sin ingresos"). Este gráfico permite ubicar cuándo el riesgo es mayor y si la diferencia entre grupos se mantiene o cambia con el tiempo.

Gráfico 33.

Riesgo instantáneo según *Ingresos* obtenido con el Modelo extendido de Cox



En el Gráfico 33, se observa que, en el tramo inicial (alrededor de las semanas 0–6) el grupo "sin ingresos" exhibe un pico marcado de $h(t)$ (0.06 aproximadamente), muy por encima de "con ingresos". Esto indica mayor probabilidad condicional de abandono en las primeras semanas y concuerda con una HR inicial protectora para "con ingresos" (0.25 aproximadamente). En el tramo intermedio, el cual es bastante estable (alrededor de las semanas 20–80) las tasas se nivelan en valores bajos (cerca de 0.010–0.020 por semana) y la brecha entre grupos se atenúa; por lo que el efecto protector de "con ingresos" disminuye progresivamente, coherente con la interacción positiva con $log(t + 1)$.

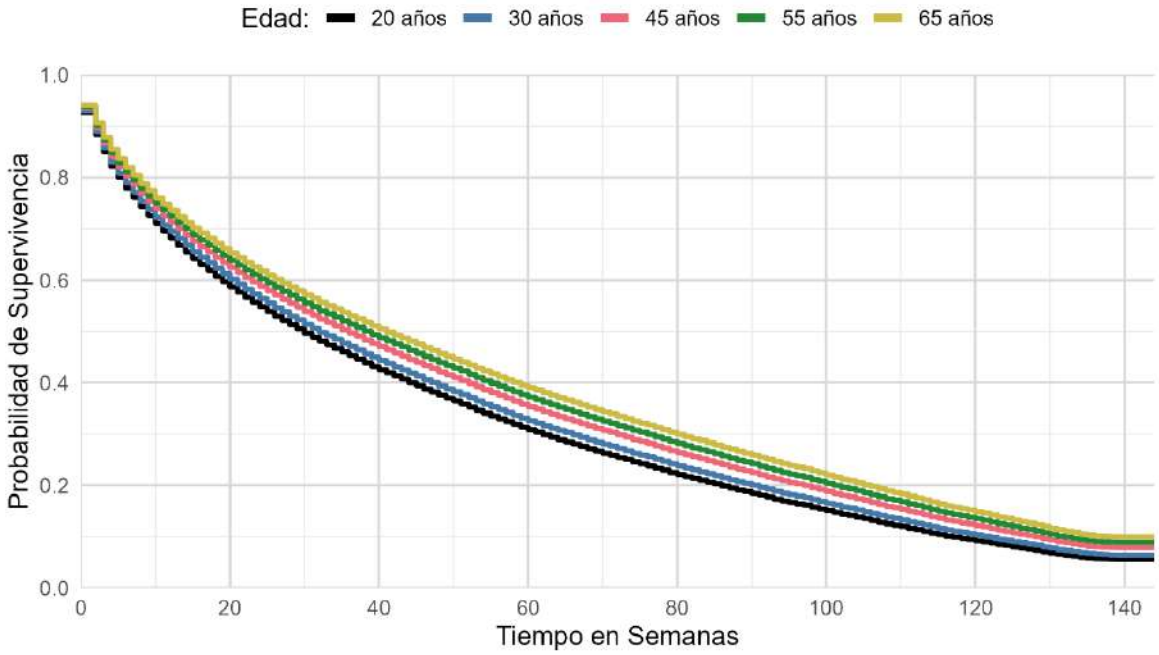
Uno de los últimos tramos, es un tramo creciente (alrededor de las semanas 100–135) en el que se observa un aumento del riesgo para ambos grupos. Este ascenso significa que, entre quienes han permanecido activos hasta esas semanas, la probabilidad condicional de abandonar se incrementa. En esta parte, las dos curvas parecen converger. Finalmente, el descenso abrupto del cierre (alrededor de las semanas 139–144) debe interpretarse con cautela: responde al borde del periodo de estudio y no a un cambio real en el riesgo.

Continuando con las curvas de supervivencia, como se ha venido presentando, en el siguiente Gráfico 34 se comparan las curvas de supervivencia $S(t)$ entre grupos de edad. Específicamente se graficaron las $S(t)$ promedio para las edades de 20, 30, 45, 55 y 65 años. Este gráfico permite identificar si existe un gradiente de permanencia asociado a la edad y en qué medida las diferencias se mantienen o cambian a lo largo del horizonte temporal.

En el mencionado Gráfico 34 se observan diferencias moderadas entre tramos etarios, sin un gradiente pronunciado y con curvas relativamente próximas entre sí, pero claramente diferenciadas. Este comportamiento es coherente con el coeficiente de *Edad* presentado en el Cuadro 24 ($\beta = -0.005;\ \exp(\beta) = 0.995;$ valor p < 0.001): por cada año adicional, el riesgo instantáneo disminuye aproximadamente 0.5%. En términos acumulados, 10 años más implican 5% menos riesgo. Y si se compara 20 años contra 65 años se encuentra que el primero presenta 20% mayor riesgo de abandono, esto de acuerdo con el modelo aquí analizado. Así, aunque las diferencias visuales son discretas, son estadística y sustantivamente consistentes con el modelo.

Gráfico 34.

Curvas de supervivencia según Edades obtenidas con el Modelo extendido de Cox

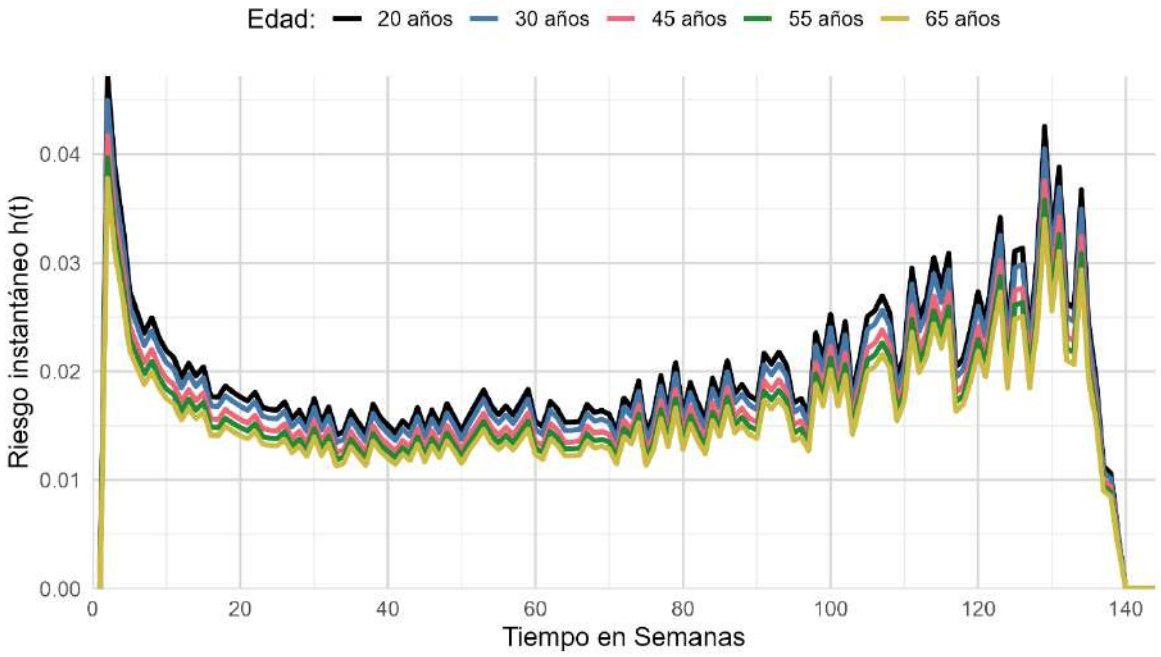


Por último, en esta sección, se presenta el Gráfico 35, que traza el riesgo instantáneo $h(t)$ por grupos de edad, con el mismo objetivo mencionado para los pasados gráficos de $h(t)$. En este caso, las curvas de $h(t)$ por edad no evidencian picos persistentes claramente diferenciados entre tramos; las variaciones observadas son moderadas y acordes con los gráficos anteriores.

Al igual que con los gráficos anteriores para $h(t)$, en el Gráfico 35 se distinguen cuatro fases: (i) inicio con pico y rápida caída (alrededor de las semanas 0–6), (ii) tramo intermedio estable en niveles bajos (alrededor de las semanas 20–80), (iii) incremento gradual hacia el final (alrededor de las semanas 100–135), conservando el orden por edad, y (iv) descenso final (alrededor de las semanas 139–144) atribuible al cierre del horizonte de observación. Además, el orden observado entre las categorías es coherente con el coeficiente de *Edad* del Cuadro 24 y con lo mostrado en el Gráfico 34 anterior.

Gráfico 35.

Riesgo instantáneo según Edades obtenido con el Modelo extendido de Cox



4.3.4. Evaluación de la bondad de ajuste y de la capacidad predictiva del modelo

En el siguiente Cuadro 25 se presenta un resumen de las métricas utilizadas para evaluar el desempeño de los modelos semi-paramétricos de supervivencia ajustados en este estudio, en este caso particular se está evaluando el modelo extendido de Cox. En este cuadro se incluyen los criterios de información AIC y BIC, empleados para comparar la bondad de ajuste penalizada por la complejidad del modelo; el índice de concordancia de Harrell (C-index), que mide la capacidad discriminativa del modelo; y el Brier Score Integrado (IBS), que evalúa la precisión de las predicciones del modelo. Además, el C-index y el IBS fueron calculados no solo sobre la muestra completa, sino también bajo un esquema de validación cruzada (VC) con 6 *folds* y 4 repeticiones, con el fin de obtener estimaciones más robustas y evitar sesgos asociados al sobreajuste. En conjunto, estos indicadores ofrecen una visión integral de la capacidad predictiva, el ajuste y la estabilidad del modelo de supervivencia.

Cuadro 25.

Métricas de la bondad de ajuste y de la capacidad predictiva del Modelo Extendido de Cox

Sin Validación Cruzada				Con Validación Cruzada. 6 *folds* y 4 repeticiones.	
AIC	**BIC**	**C-index**	**IBS**	**C-index**	**IBS**
680 588.5	680 715.1	0.6011	0.1810	0.5805	0.1811

A partir del Cuadro 25 se aprecia que este modelo extendido de Cox presentó un AIC igual a 680,588.5 y un BIC de 680,715.1. Estos valores se pueden utilizar para comparar especificaciones alternativas dentro de la misma familia del modelo de Cox (por ejemplo, con/sin interacciones o con diferentes formas funcionales). Al estar este modelo extendido de Cox basado en verosimilitud parcial, sus AIC/BIC no son directamente comparables con los de modelos paramétricos (basados en verosimilitud completa). Por ello, cuando se contraste el desempeño de este modelo frente a modelos paramétricos, el énfasis debe ponerse en otras métricas predictivas, como el C-index, o IBS.

En cuanto a la discriminación global, el C-index de Harrell del modelo resultó en 0.6011 sin validación cruzada, y en 0.5805 con validación cruzada, lo que indica una capacidad discriminativa modesta: en pares comparables, el sujeto con mayor riesgo predicho fallece antes en aproximadamente 58% de los casos. Este nivel de desempeño sugiere que el ranking de riesgo es útil pero mejorable.

Respecto a la precisión global a lo largo del tiempo, el Brier Score Integrado (IBS) sin validación cruzada dio 0.1810 y, bajo validación cruzada con 6 *folds* y 4 repeticiones, resultó en 0.1811. La diferencia es poca entre estos dos valores, lo que evidencia muy poco sobreajuste y respalda la estabilidad del modelo en términos de error de predicción agregado. Donde, valores de IBS cercanos o por debajo de 0.20 suelen considerarse aceptables en escenarios con incidencia moderada.

En síntesis, el modelo extendido de Cox presenta una capacidad de discriminación global modesta (C-index de aproximadamente 0.58), que deja un margen importante para mejorar la separación entre individuos de mayor y menor riesgo. En contraste, el nivel de error promedio en la predicción a lo largo del tiempo puede considerarse aceptable (IBS de aproximadamente 0.18). La similitud entre las métricas obtenidas con y sin validación cruzada sugiere que el ajuste es estable y que el sobreajuste, de existir, es limitado. En conjunto, estos resultados indican que el modelo es útil como herramienta de estratificación de riesgo, aunque su capacidad discriminativa sigue siendo moderada.

4.4. MODELOS PARAMÉTRICOS

En esta sección se presentan los resultados del ajuste de los modelos paramétricos de supervivencia. El análisis se orienta a identificar las distribuciones que mejor combinan ajuste y capacidad predictiva y, a partir de ellas, examinar los efectos de las covariables sobre el riesgo de abandono, así como los tiempos de supervivencia estimados y sus variaciones entre distintos perfiles de usuarios.

Como se expuso en la metodología, se consideraron seis distribuciones paramétricas: Exponencial, Weibull, Log-normal, Log-logística, Gamma y Gompertz. Además, debido a

la violación del supuesto de riesgos proporcionales en las covariables *Región* e *Ingresos* (documentada en secciones previas), se dividió la base de datos en 14 subconjuntos independientes, producto de la división de cada una de las siete regiones geográficas con la combinación de la condición de *Ingresos* del usuario ("con ingresos" y "sin ingresos"). Esta estratificación favorece el cumplimiento de supuestos y permite obtener estimaciones más específicas y representativas para cada grupo de análisis.

4.4.1. Comparación de las métricas de los distintos modelos paramétricos

En cada uno de los 14 subconjuntos (combinaciones de *Región* e *Ingresos*) se ajustaron los seis modelos paramétricos, para un total de 84 modelos. Para cada modelo se calcularon medidas de ajuste y parsimonia (AIC y BIC) y métricas predictivas de uso frecuente en supervivencia, incluyendo el índice de concordancia (C-index) y el Brier Score Integrado (IBS) con validación cruzada. La validación cruzada se implementó con seis particiones (*folds*) y cuatro repeticiones, generando 24 estimaciones por métrica; el promedio de dichas estimaciones se utilizó como valor representativo.

Con el fin de evaluar si las diferencias entre modelos eran estadísticamente significativas, se aplicó la prueba *t* de Student para medias de diferencias pareadas sobre los valores obtenidos en cada fold y repetición (cuando se utilizó validación cruzada). Este procedimiento permite contrastar modelos bajo condiciones controladas de variabilidad y constituye una práctica estándar en la comparación de métricas de predicción.

De esta forma, en el Cuadro 26 se presenta, para cada uno de los 14 subconjuntos, el tamaño muestral, el mejor modelo (distribución) según el AIC y su valor, y el mejor modelo según el BIC y su valor. Por su parte, el Cuadro 27 muestra, también para cada subconjunto, el mejor modelo según el C-index y su valor, así como el mejor modelo según el IBS y su valor correspondiente. En este último caso (Cuadro 27), las métricas provienen de la validación cruzada.

Cabe señalar que los resultados del Cuadro 27 se fundamentan en pruebas t pareadas (no mostradas en ese cuadro, pero sintetizadas en el texto). En varios subconjuntos no se

observaron diferencias estadísticamente significativas que permitieran definir un único ganador. De forma similar, los modelos señalados como "mejores" en el Cuadro 26 se determinan conforme a los criterios de interpretación propuestos por Burnham y Anderson (2002) para el AIC, y por Raftery (1996) y Burnham y Anderson (2002) para el BIC, tal como se expone en la Sección 2.4 del Marco Teórico.

Adicionalmente, el Cuadro 28 resume los resultados de ajuste y predicción obtenidos al considerar la base de datos completa. Los resultados de las pruebas t pareadas (sobre C-index por *fold* y repetición en validación cruzada) para estos casos se presentan en el Anexo B, Cuadros B.1 y B.2.

Con base en estos criterios, se seleccionó un modelo de referencia para interpretar los coeficientes asociados a las covariables y, asimismo, se identificaron los dos modelos mejor posicionados en términos predictivos para la estimación de los tiempos de supervivencia.

De acuerdo con lo presentado en el Cuadro 26, se evidencia que, de manera consistente, el modelo Log-normal se posiciona como el mejor ajuste en la mayoría de los conjuntos de datos, tanto según el criterio de AIC como de BIC. En particular, se observa que esta distribución predomina en las bases más grandes, como California & Nevada-Sin Ingresos, Midwest-Sin Ingresos, Northeast-Sin Ingresos, Pacific Northwest-Sin Ingresos y Rockies & Red River-Sin Ingresos, lo que refuerza su robustez y estabilidad como modelo de referencia. Esta regularidad en los conjuntos de mayor tamaño otorga un peso adicional a su superioridad, dado que en muestras más extensas los resultados tienden a ser más confiables y menos sensibles a la variabilidad aleatoria.

En segundo lugar, el modelo Weibull aparece de forma reiterada como la mejor opción en algunos subconjuntos, especialmente en los grupos Midwest-Con Ingresos, Southeast-Sin Ingresos y Utah & Arizona-Sin Ingresos. Aunque su presencia es menor en comparación con el Log-normal, su relevancia se ve respaldada porque emerge como la mejor alternativa en bases de datos de tamaño considerable, lo que lo convierte en el segundo modelo más competitivo.

En síntesis, el cuadro evidencia que el Log-normal es el modelo con mejor desempeño global y más consistente entre los diferentes subconjuntos de datos, mientras que el Weibull se posiciona como la segunda mejor opción, especialmente al considerar el tamaño de las bases en las que resulta favorecido.

Cuadro 26.

Comparación de ajuste y parsimonia (AIC y BIC) entre modelos paramétricos por subconjunto: mejor distribución identificada

Conjunto de datos	Tamaño	Mejor modelo AIC	Valor AIC	Mejor modelo BIC	Valor BIC
California & Nevada-Con Ingresos	2438	Exponencial	12311.7	Exponencial	12352.3
Midwest-Con Ingresos	1323	Weibull	8490.8	Exponencial	8529.0
Northeast-Con Ingresos	1764	Log-normal	9361.2	Log-normal	9405.0
Pacific Northwest-Con Ingresos	1816	Exponencial	9523.0	Exponencial	9561.5
Rockies & Red River-Con Ingresos	3916	Gompertz	29659.0	Exponencial	29704.4
Southeast-Con Ingresos	610	Log-normal	4040.6	Log-normal	4075.9
Utah & Arizona-Con Ingresos	3498	Gompertz	27064.2	Gompertz	27113.5
California & Nevada-Sin Ingresos	5207	Log-normal	32676.6	Log-normal	32729.1
Midwest-Sin Ingresos	4335	Log-normal	28326.0	Log-normal	28377.0
Northeast-Sin Ingresos	3115	Log-normal	19191.3	Log-normal	19239.6
Pacific Northwest-Sin Ingresos	2771	Log-normal	17805.9	Log-normal	17853.3
Rockies & Red River-Sin Ingresos	10116	Log-normal	77344.6	Log-normal	77402.3
Southeast-Sin Ingresos	2775	Weibull	16277.5	Weibull	16324.9
Utah & Arizona-Sin Ingresos	4753	Weibull	36899.3	Weibull	36951.0

El siguiente Cuadro 27, muestra los resultados comparativos de los modelos de supervivencia en términos del índice de concordancia (C-index) y del Brier Score Integrado (IBS). En lo que respecta al C-index, los modelos que aparecen con mayor frecuencia como mejores son el Weibull y el Log-normal, seguidos en algunos subconjuntos por el Exponencial y el Log-logístico. Sin embargo, no se observa una dominancia clara y consistente de un único modelo

en todos los conjuntos de datos, lo que refleja que, según esta métrica, el desempeño es muy similar para cada una de las distribuciones entre los distintos subconjuntos de datos.

No obstante, al analizar los resultados con base en el IBS, se obtiene una conclusión mucho más clara. El modelo Gamma se posiciona sistemáticamente como el mejor en la gran mayoría de los subconjuntos, presentando los valores de error de predicción promedio más bajos. En algunos pocos casos, el modelo Gompertz también aparece como la mejor alternativa, pero siempre en menor frecuencia. Esta consistencia a favor del modelo Gamma es particularmente relevante porque el IBS, al integrar tanto la discriminación como la calibración de los modelos a lo largo de todo el horizonte temporal, ofrece una medida más robusta y global de la capacidad predictiva que el C-index, el cual se centra principalmente en la discriminación de riesgos entre individuos.

En consecuencia, si bien el C-index muestra cierta variabilidad en la elección del mejor modelo, la evidencia basada en el IBS es contundente: el modelo Gamma es el que ofrece de manera más consistente el mejor desempeño predictivo en los distintos subconjuntos de datos. Por lo tanto, considerando la mayor robustez del IBS como métrica de evaluación, la conclusión final es que la distribución Gamma debe reconocerse como el mejor modelo global para realizar predicciones, respaldada tanto por su desempeño promedio como por su coherencia a través de los diferentes escenarios evaluados.

Cuadro 27.

Comparación de capacidad predictiva (C-index y IBS-VC) entre modelos paramétricos por subconjunto: mejor distribución identificada

Conjunto de datos	Mejor modelo C-index	Valor C-index	Mejor modelo IBS	Valor IBS
California & Nevada-Con Ingresos	Log-normal	0.5645	Gamma	0.2026
Midwest-Con Ingresos	Weibull	0.5452	Gamma	0.1928
Northeast-Con Ingresos	Log-normal	0.5178	Gompertz	0.2060
Pacific Northwest-Con Ingresos	Exponencial	0.5438	Log-normal	0.3052
Rockies & Red River-Con Ingresos	Log-logística	0.5408	Gamma	0.2009
Southeast-Con Ingresos	Weibull	0.5459	Gompertz	0.2025
Utah & Arizona-Con Ingresos	Exponencial	0.5459	Gompertz	0.1955
California & Nevada-Sin Ingresos	Weibull	0.5197	Gamma	0.1791
Midwest-Sin Ingresos	Log-logística	0.5394	Gamma	0.1710
Northeast-Sin Ingresos	Log-normal	0.5179	Gamma	0.1652
Pacific Northwest-Sin Ingresos	Weibull	0.5262	Gamma	0.1891
Rockies & Red River-Sin Ingresos	Log-logística	0.5293	Gamma	0.1728
Southeast-Sin Ingresos	Exponencial	0.5194	Gamma	0.1895
Utah & Arizona-Sin Ingresos	Weibull	0.5482	Gamma	0.1718

Seguidamente, se presenta el Cuadro 28, que como se había mencionado, resume para la base de datos completa, el desempeño de cada distribución paramétrica (Exponencial, Weibull, Log-normal, Log-logística, Gamma y Gompertz). El cuadro reporta los criterios de información AIC y BIC, así como las métricas predictivas C-index e IBS promediadas mediante validación cruzada (junto con sus errores estándar). Los contrastes estadísticos entre modelos (pruebas t pareadas sobre C-index en VC) se detallan en los Cuadros B.1 y B.2 del Anexo B.

Cuadro 28.

Comparación general de los modelos según AIC, BIC, C-index, y IBS

Distribución	AIC	BIC	Promedio C-index	C-index EE	Promedio IBS	IBS EE
Log-normal	330247.7	330379.54	0.6006	0.0026	0.2009	0.0009
Weibull	330630.9	330762.72	0.6007	0.0048	0.1979	0.0010
Gamma	330959.6	331091.42	0.6006	0.0027	0.1827	0.0012
Log-logístico	331725.9	331857.73	0.6006	0.0026	0.2006	0.0009
Gompertz	332575.5	332707.29	0.6006	0.0026	0.1836	0.0013
Exponencial	333671.7	333794.70	0.6003	0.0027	0.1999	0.0011

De acuerdo con los resultados presentados en el Cuadro 28, y siguiendo los criterios de interpretación propuestos por Burnham y Anderson (2002) para el AIC, así como por Raftery (1996) y Burnham y Anderson (2002) para el BIC, se observa de manera consistente que el modelo con mejor ajuste a los datos corresponde a la distribución Log-normal.

En primer lugar, al analizar el AIC, se identifica que el valor más bajo corresponde al modelo Log-normal (330,247.7), lo que lo posiciona como la mejor opción según este criterio. El segundo mejor modelo es el Weibull (330,630.9); sin embargo, la diferencia entre ambos es de $\Delta AIC = 383.2$, valor que supera ampliamente el umbral de 10 y que, por ende, constituye evidencia muy fuerte a favor del modelo Log-normal frente al Weibull. En el caso del modelo Gamma (330,959.6), la diferencia con respecto al Log-normal asciende a $\Delta AIC = 711.9$, lo cual refuerza la superioridad del Log-normal y coloca al Gamma, junto con los modelos Log-logístico, Gompertz y Exponencial, en una posición claramente inferior. En síntesis, la conclusión derivada del AIC es inequívoca: el modelo Log-normal ofrece el mejor ajuste, sin que exista ningún empate o comparabilidad estadística con el resto de modelos, dado que todas las diferencias superan con creces el umbral de 10.

De forma análoga, la evaluación con el BIC conduce a la misma conclusión. El valor más bajo de BIC se alcanza con el modelo Log-normal (330,379.54), seguido por el modelo Weibull (330,762.72). La diferencia entre ambos es de $\Delta BIC = 383.18$, lo cual también se ubica por encima del umbral de 10 y constituye evidencia muy fuerte en favor del modelo Log-normal. Las diferencias observadas con los modelos restantes (Gamma, Log-logístico, Gompertz y Exponencial) son incluso mayores, consolidando aún más esta conclusión.

En consecuencia, tanto el AIC como el BIC señalan de manera consistente e indiscutible que la distribución Log-normal es la que mejor explica los datos. No se identifican empates ni modelos cercanos en términos de ajuste o parsimonia, dado que todas las diferencias superan de manera holgada los umbrales establecidos en la literatura especializada.

Por otra parte, los resultados de la comparación basada en el C-index mostraron que, aunque el modelo Weibull aparece como el mejor posicionado en el Cuadro 28, las pruebas t pareadas (ver Cuadro B.1 en Anexos B) revelaron que las diferencias con los modelos Log-normal y Log-logístico no alcanzan significación estadística estricta ($p = 0.0515$ y $p = 0.0647$, respectivamente). Esto sugiere que, en términos de discriminación (C-index), no existe un modelo claramente superior.

Por el contrario, al analizar los resultados de la prueba *t* pareada para el IBS, se encontró que los modelos Gamma y Gompertz presentan sistemáticamente valores más bajos, lo que indica un mejor desempeño predictivo en términos de error de predicción promedio. Esta tendencia se refleja también en los resultados del Cuadro 28 y se confirma en el Cuadro B.2 del Anexo B, donde se presentan los contrastes detallados. En particular, de acuerdo con esta métrica IBS, el modelo Gamma emergió como el más consistente en cuanto a capacidad predictiva.

La interpretación conjunta de los resultados conduce a una conclusión matizada. El modelo Weibull se mantiene competitivo para la predicción de tiempos de supervivencia, especialmente cuando la función de riesgo (*hazard*) es monótona (creciente o decreciente). Sin embargo, el modelo Gamma mostró una superioridad sistemática en el IBS, métrica que evalúa globalmente la exactitud de las predicciones. A su vez, los modelos Log-normal y Log-logístico exhibieron un desempeño adecuado según el C-index; en particular, el Log-normal destacó además por su mejor ajuste de acuerdo con AIC y BIC y por su consistencia en los subconjuntos de mayor tamaño.

En este marco, se adoptó la distribución Log-normal como modelo de referencia para el análisis de los coeficientes y la interpretación de las covariables debido a su buen desempeño en comparación con las otras distribuciones. Además, este modelo permite la derivación e interpretación de tiempos de supervivencia esperados y probabilidades asociadas a distintos

perfiles de usuarios; por ello, también se utilizará para la predicción de supervivencia de los usuarios de la Plataforma.

No obstante, con el propósito de ampliar el alcance predictivo, se seleccionó también la distribución Gamma como modelo complementario para la estimación de tiempos de supervivencia, ya que mostró un gran rendimiento principalmente con la métrica IBS (esto se puede observar tanto en el Cuadro 27 como en el Cuadro 28). Además, es pertinente señalar que tanto la distribución Weibull como Gamma pueden verse como extensiones de la Exponencial, la cual se obtiene como caso particular cuando el parámetro de forma es igual a 1. Cuando dicho parámetro difiere de 1, ambas familias permiten riesgos monotónicos dependientes del tiempo (crecientes o decrecientes), relajando el supuesto de riesgo constante propio del modelo Exponencial. En este sentido, estas distribuciones ofrecen mayor flexibilidad para representar la evolución temporal del riesgo, aunque restringida a patrones monotónicos.

En síntesis, a continuación, se trabaja con dos modelos complementarios: el Log-normal, priorizado para la inferencia, análisis de coeficientes, y la presentación de fórmulas de predicción por su superioridad en ajuste e interpretabilidad; y el modelo Gamma, enfatizado por su desempeño predictivo. Esta combinación permite una visión más integral y robusta del fenómeno de abandono en la plataforma.

4.4.2. Análisis de los supuestos del modelo Log-normal

El modelo paramétrico ajustado en esta sección, cuyos supuestos se evalúan a continuación, se representa mediante la Ecuación E.55, que corresponde a un modelo de tiempo de falla acelerado (AFT) con distribución Log-normal. En este modelo, se asume que el logaritmo del tiempo hasta el evento ($\log(T_i)$) se distribuye normalmente, con media determinada por una combinación lineal de las covariables incluidas en el modelo y varianza constante.

$$\begin{aligned}\log(T_i) = \beta_0 &+ \beta_1 \cdot \{\text{Región California \& Nevada}\}_i \\ &+ \beta_2 \cdot \{\text{Región Midwest}\}_i \\ &+ \beta_3 \cdot \{\text{Región Northeast}\}_i \\ &+ \beta_4 \cdot \{\text{Región Pacific Northwest}\}_i \\ &+ \beta_5 \cdot \{\text{Región Rockies \& Red River}\}_i \\ &+ \beta_6 \cdot \{\text{Región Southeast}\}_i + \beta_7 \cdot \{\text{Licencia}_{\text{LPN}}\}_i \\ &+ \beta_8 \cdot \{\text{Licencia}_{\text{RN}}\}_i + \beta_9 \cdot \{\text{Licencia}_{\text{Sin licencia}}\}_i \\ &+ \beta_{10} \cdot \{\text{Licencia}_{\text{Otras}}\}_i + \beta_{11} \cdot \text{Ingreso}_i \\ &+ \beta_{12} \cdot \text{Edad}_i + \beta_{13} \cdot \text{Edad\_Dicotómica}_i + \sigma\,\varepsilon_i\end{aligned}$$

Ecuación E.55

donde, $\varepsilon_i \sim N(0,1)$, y $\sigma > 0$ representa el parámetro de escala del modelo. Mientras que, los β_j corresponden a los coeficientes asociados a cada covariable incluida en el modelo.

En este contexto, el término $\exp(X_i^{\top}\beta)$ actúa como un factor de aceleración, indicando el efecto proporcional de cada covariable sobre el tiempo de supervivencia. Valores de $\exp(\beta_j) > 1$ sugieren que la covariable asociada aumenta el tiempo esperado hasta el evento (mayor supervivencia), mientras que valores inferiores a uno implican una reducción del tiempo esperado (menor supervivencia).

Como ya se ha mencionado, la validez de un modelo de supervivencia depende en gran medida del cumplimiento de los supuestos teóricos sobre los cuales se construye. En el caso del modelo Log-normal, estos supuestos garantizan la consistencia y la interpretabilidad de los parámetros estimados, así como la fiabilidad de las inferencias realizadas. En esta sección se examinan los principales supuestos asociados a este modelo, incluyendo la ausencia de multicolinealidad entre predictores, la correcta especificación funcional del modelo, la normalidad del logaritmo de los tiempos de supervivencia, la homocedasticidad y la linealidad de la covariable continua respecto al predictor lineal.

Ausencia de multicolinealidad severa entre predictores:

Este supuesto ya fue abordado en capítulos anteriores, tanto en la sección exploratoria de los datos como en el análisis del modelo semi-paramétrico. Dichas evaluaciones preliminares no evidenciaron relaciones fuertes entre las covariables incluidas. No obstante, en esta sección se realiza nuevamente el análisis mediante el cálculo del Factor de Inflación de la Varianza (VIF), con el fin de corroborar que el modelo Log-normal ajustado no presenta problemas de colinealidad que pudieran afectar la estabilidad de los coeficientes o la interpretación de los resultados.

Para evaluar la posible presencia de multicolinealidad entre las covariables incluidas en el modelo Log-normal, se calculó el Factor de Inflación de la Varianza (VIF). Los valores obtenidos se presentan en el Cuadro 29. En general, todos los VIF se mantuvieron muy por debajo del umbral crítico de 5, con un rango entre 1.03 y 2.01, lo que indica una baja correlación entre predictores.

Cuadro 29.

Valores del Factor de Inflación de la Varianza (VIF) para las covariables calculados con un modelo lineal auxiliar equivalente al modelo Log-normal ajustado

Covariable	VIF
R. California & Nevada	1.706
R. Midwest	1.554
R. Northeast	1.472
R. Pacific Northwest	1.448
R. Rockies & Red River	2.013
R. Southeast	1.358
Licencia LPN	1.094
Licencia RN	1.108
Sin licencia	1.009
Otras Licencias	1.063
Ingresos	1.031
Edad	1.146
Sin_Edad	1.081

Como se observa en el Cuadro 29, todos los valores de VIF son considerablemente bajos (menores a 2.1). De acuerdo con las recomendaciones de Kutner et al. (2005) y Kleinbaum & Klein (2005), valores inferiores a 5 indican que no existe colinealidad relevante entre las covariables. En consecuencia, se concluye que el modelo Log-normal cumple con el supuesto de ausencia de multicolinealidad severa entre predictores. Estos resultados confirman que no existe colinealidad severa entre las variables explicativas, por lo que las estimaciones de los coeficientes del modelo son estables y confiables.

Enfoque AFT (*Accelerated Failure Time*)

En el modelo Log-normal, el efecto de las covariables se interpreta en la escala del tiempo: se asume que multiplican (aceleran o desaceleran) el tiempo hasta el evento por un factor constante a lo largo de toda la distribución de supervivencia. Formalmente, si T es el tiempo al evento y X el vector de covariables, el modelo postula que $\log \mathrm{T} = X\beta + \sigma\varepsilon$, con ε normal estándar. En consecuencia, los coeficientes se interpretan como razones de tiempo (time ratios, TR) iguales a $\exp(\beta)$: un $\mathrm{TR} > 1$ indica tiempos más largos (ralentización del fallo) y un $\mathrm{TR} < 1$ indica tiempos más cortos (aceleración del fallo). La premisa AFT clave es que ese factor de aceleración es constante para cualquier cuantil de la distribución, de modo que los perfiles de supervivencia de dos grupos difieren principalmente por un desplazamiento horizontal en el eje del tiempo.

La verificación de este supuesto ya se abordó parcialmente en secciones anteriores, a partir del análisis de los gráficos log(-log(S(t))) contra log(t), conocidos comúnmente como gráficos log-log. En ellos, líneas aproximadamente rectas y paralelas indican que el supuesto AFT se cumple.

De acuerdo con los Gráficos de secciones anteriores, específicamente los Gráficos 24, 25 y 26 (tipo log-log), se observa que la variable *Ingresos* y las variables de las regiones "Midwest", "Northeast", "Pacific Northwest" y "Rockies & Red River" presentan desviaciones respecto a la paralelidad esperada, lo que sugiere violaciones parciales del supuesto AFT en comparación con sus categorías de referencia.

De igual manera, al analizar las curvas de supervivencia de los Gráficos 15, 16, 17 y 18, bajo la perspectiva del modelo AFT, se espera que las curvas correspondientes a distintos grupos mantengan la misma forma (pendiente general y curvatura), diferenciándose únicamente por un desplazamiento horizontal en el eje del tiempo, es decir, que una curva parezca "estirada" o "trasladada" hacia la derecha en relación con la otra. Por lo tanto, si las curvas son aproximadamente paralelas en la escala del log-tiempo o presentan formas similares pero desplazadas, se puede considerar que el supuesto AFT se cumple razonablemente bien. En este caso, las conclusiones obtenidas a partir de los gráficos de supervivencia coinciden con las observadas en los gráficos log-log, confirmando que el supuesto AFT no se cumple de forma completa para las variables ya mencionadas (variable *Ingresos* y las variables de las regiones "Midwest", "Northeast", "Pacific Northwest" y "Rockies & Red River"). No obstante, las discrepancias observadas son de baja magnitud, especialmente para las variables de tipo región, aparecen como pequeños desplazamientos y leves variaciones de forma respecto del patrón esperado, por lo que él su incumplimiento no parece ser severo para estas variables de tipo región.

Supuesto de distribución Log-normal del tiempo al evento

El modelo de supervivencia Log-normal asume que, condicionado a las covariables incluidas en el modelo, el logaritmo del tiempo hasta el evento sigue una distribución Normal. Bajo este supuesto, los residuos estandarizados del modelo deberían seguir una distribución Normal estándar $N(0,1)$.

Sin embargo, en muestras grandes, como es el caso del presente estudio, las pruebas formales de normalidad tienden a rechazar incluso desviaciones mínimas respecto a la distribución Normal, que no necesariamente comprometen la validez del modelo. Por ello, el diagnóstico visual mediante el gráfico Q-Q (quantile–quantile) es una mejor herramienta para evaluar la adecuación de la suposición de Log-normalidad en este caso.

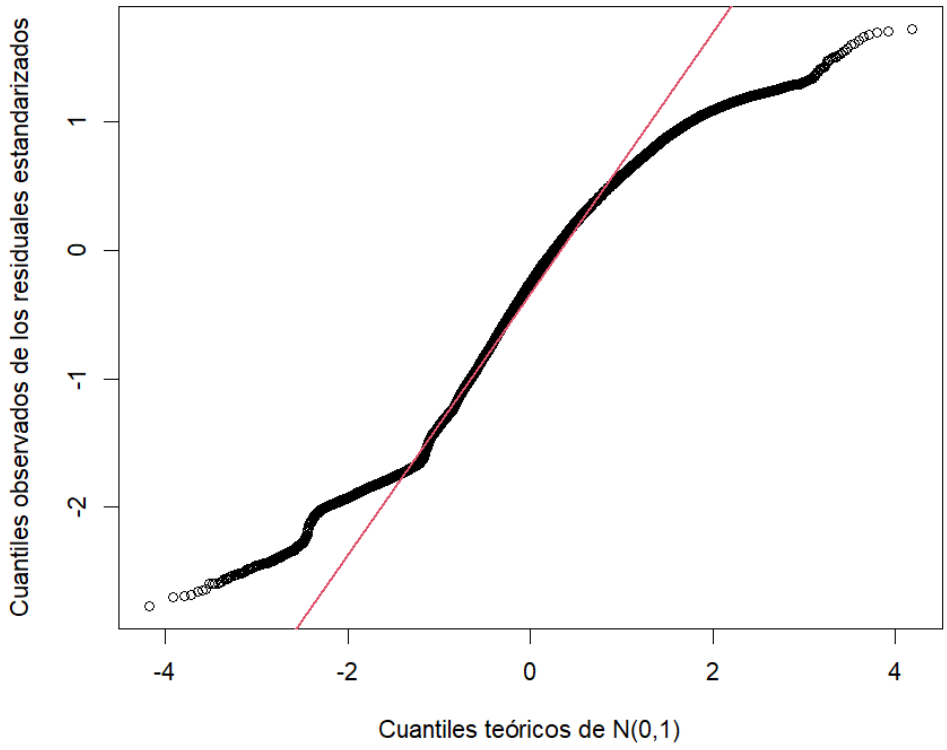
En el Gráfico 36 se presenta el *Q-Q plot* de los residuos estandarizados del log-tiempo bajo el modelo Log-normal. Si el supuesto de Log-normalidad se cumple, los puntos deberían

alinearse aproximadamente sobre la recta diagonal, indicando que los cuantiles empíricos de los residuos coinciden con los cuantiles teóricos de una Normal estándar $N(0,1)$.

La alta correlación observada entre los cuantiles teóricos y los observados ($R = 0.986$) sugiere que, aunque las pruebas de normalidad detectan desviaciones estadísticamente significativas, la aproximación global a la Normal estándar es adecuada. Esto indica que la distribución Log-normal describe razonablemente bien la variabilidad de los tiempos de supervivencia.

Gráfico 36.

Q-Q plot de residuos estandarizados del log-tiempo bajo el modelo Log-normal



Analizando el Gráfico 36, se observa que en el centro los puntos se ajustan estrechamente a la línea de referencia, lo que evidencia un buen cumplimiento del supuesto de normalidad en la zona central de la distribución del log-tiempo. Sin embargo, en los extremos o colas, los

puntos muestran una ligera desviación respecto a la línea diagonal: en la cola inferior se observa una subestimación, mientras que en la cola superior se aprecia una sobreestimación de los cuantiles teóricos.

Estas desviaciones en las colas pueden atribuirse a que los valores extremos del tiempo de supervivencia (muy pequeños o muy grandes) presentan colas más pesadas (mayor presencia de extremos) de lo que asume la normalidad del log-tiempo bajo el modelo Log-normal. Este comportamiento es habitual en datos de supervivencia, donde la distribución real puede exhibir una mayor probabilidad de observaciones extremas que la prevista por una distribución normal. En términos prácticos, esto indica que el modelo podría subestimar la probabilidad de eventos muy tempranos y también subestimar (o representar con menor precisión) la ocurrencia de tiempos muy tardíos, afectando principalmente la exactitud de las predicciones en los extremos de la distribución temporal.

No obstante, dado que el ajuste es adecuado en la región central, donde se concentra la mayor parte de las observaciones, el supuesto de Log-normalidad puede considerarse razonablemente cumplido. Esto permite realizar predicciones y análisis de inferencia con el modelo, siempre que se reconozca que las estimaciones asociadas a los tiempos más extremos deben interpretarse con cautela.

Supuesto de homocedasticidad para el modelo Log-normal: verificación gráfica con residuos

En los modelos AFT con distribución Log-normal se asume que la varianza del término de error es constante entre observaciones. Este decir, la varianza del error en la escala logarítmica del tiempo es la misma para todos los niveles de las covariables.

Este supuesto es crucial porque respalda la validez de la inferencia: cuando se cumple, las estimaciones de los coeficientes son insesgadas y eficientes, y los errores estándar reflejan adecuadamente la incertidumbre. Si no se cumple, las pruebas de significancia y los intervalos de confianza pueden resultar poco confiables. En términos prácticos, la homocedasticidad implica que las covariables actúan como factores de aceleración o desaceleración sobre los tiempos de supervivencia, sin modificar la dispersión de dichos

tiempos. Si existiera heterocedasticidad, se podría subestimar la incertidumbre en subgrupos con mayor variabilidad o sobreestimar efectos al confundirlos con cambios en la varianza.

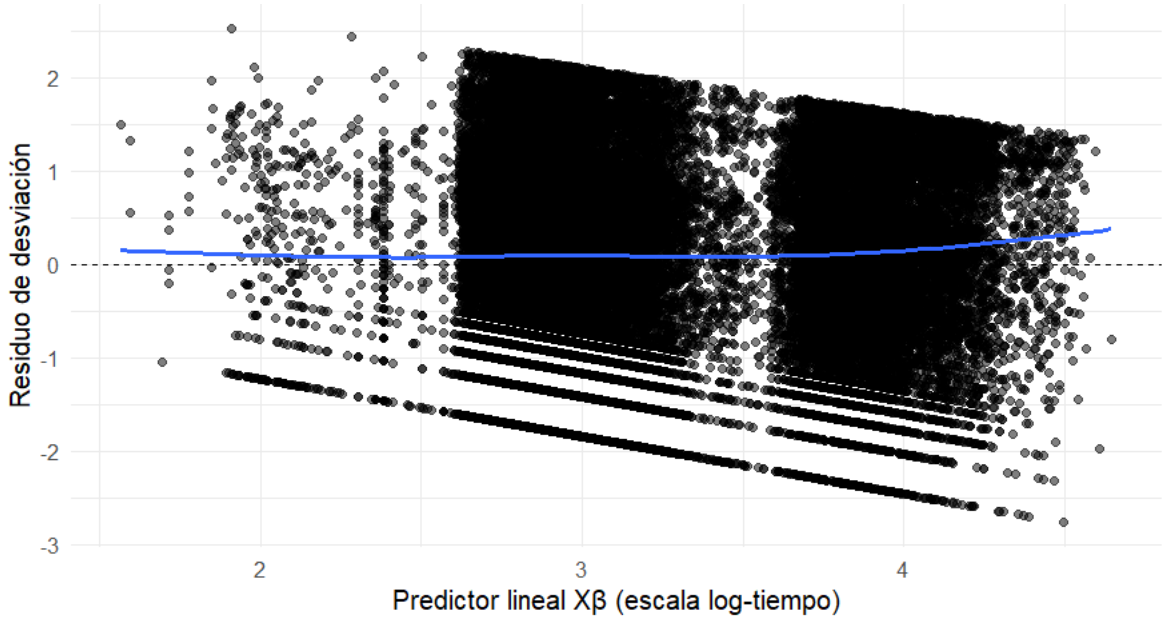
Una manera sencilla de evaluar este supuesto es mediante un diagnóstico gráfico de residuos. En datos con censura, los residuos habituales no son apropiados; por ello se recomienda utilizar los residuos de devianza del modelo ajustado. Para analizar homocedasticidad y detectar posibles malas especificaciones, se graficó los residuos de devianza frente al predictor lineal ($X\widehat{\beta}$).

En estos gráficos, si la nube de puntos muestra dispersión aproximadamente constante a lo largo del eje x, y la curva de suavizamiento se mantiene próxima a la horizontal en torno a 0, hay apoyo para la homocedasticidad. En cambio, patrones de "abanico" (dispersión creciente o decreciente con $X\hat{\beta}$) o tendencias sistemáticas de la curva sugieren heterocedasticidad o mala especificación.

En el Gráfico 37 se representan los residuos de deviancia versus el predictor lineal (escala log-tiempo) del modelo Log-normal. La línea punteada indica el nivel cero, y la curva de suavizamiento (*LOESS*) resume la tendencia media de los residuos. Este gráfico es útil para evaluar el supuesto de homocedasticidad: bajo varianza constante, los residuos deben dispersarse de forma aproximadamente uniforme alrededor de cero en todo el rango del eje x, sin "abanicos" ni tendencias sistemáticas.

Gráfico 37.

Residuos de Deviancia versus predictor lineal para el modelo Log-normal



En términos generales, la dispersión vertical de los residuos es bastante similar a lo largo de la mayor parte del rango del eje x, lo que respalda homocedasticidad razonable en la zona central. La curva *LOESS* se mantiene cerca de 0, aunque muestra una ligera curvatura ascendente en el extremo derecho (valores altos de $X\hat{\beta}$), lo que sugiere un leve sesgo o una posible pequeña expansión de la varianza en esa región.

Se observan además bandas o franjas de puntos en la parte negativa de los residuos, esperables con censura y con la discretización del tiempo; no necesariamente implican mal ajuste. Existen algunos valores atípicos (residuos > 2 o < −2), pero su número es pequeño y no parecen alterar el patrón global. Estos valores atípicos fueron analizados en la sección del análisis exploratorio de los datos.

Linealidad de Covariables Continuas

En el modelo de supervivencia Log-normal, se asume que la relación entre cada covariable continua y el logaritmo del tiempo al evento es lineal. Este supuesto implica que el efecto de la covariable sobre el tiempo de supervivencia actúa de manera proporcional y constante en la escala logarítmica, de modo que una unidad de cambio en la covariable produce un cambio fijo en el log-tiempo esperado, independientemente de su valor inicial.

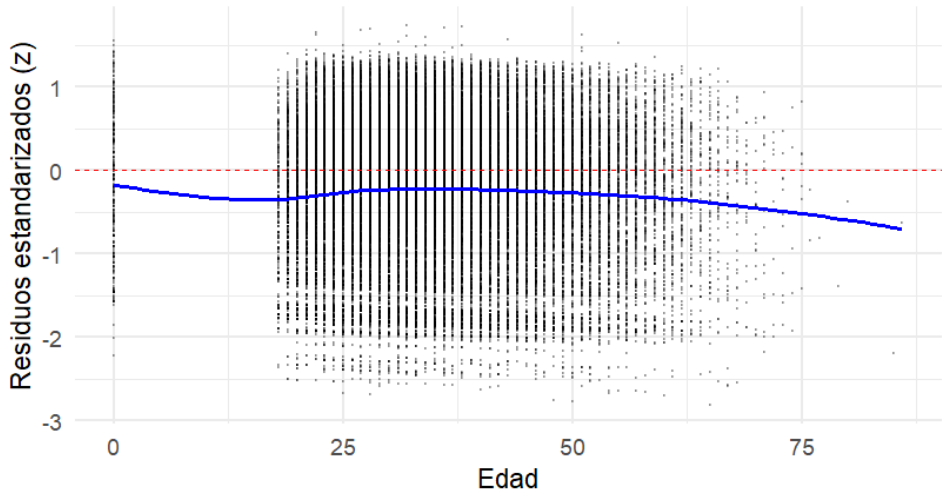
El incumplimiento de este supuesto puede generar sesgos en los coeficientes estimados y una mala especificación funcional del modelo, lo que afectaría tanto la interpretación de los parámetros como la validez de las inferencias derivadas. Por ello, resulta esencial verificar si las covariables continuas presentan una relación aproximadamente lineal con el log-tiempo o si, por el contrario, es necesario considerar transformaciones no lineales o funciones más flexibles, como términos polinómicos o *splines*.

En este estudio, la única covariable continua incluida en el modelo es la *Edad*, por lo que la verificación de este supuesto se centra exclusivamente en dicha variable. Para ello, se analizó la relación entre los residuos estandarizados del log-tiempo y la covariable *Edad*. La presencia de una tendencia sistemática o curvatura en esta relación indicaría una posible violación del supuesto de linealidad, mientras que una dispersión aleatoria de los residuos alrededor de cero respaldaría el cumplimiento del mismo.

De esta forma, el siguiente Gráfico 38 muestra la relación entre los residuos estandarizados del log-tiempo y la covariable *Edad*, con el propósito de evaluar el cumplimiento del supuesto de linealidad de las covariables continuas en el modelo Log-normal. En este gráfico, la línea azul suavizada (*LOESS*) representa la tendencia promedio de los residuos a lo largo del rango de edades, mientras que la línea roja discontinua indica el valor de referencia cero. Si la relación entre *Edad* y el log-tiempo es aproximadamente lineal, la línea azul debería mantenerse cercana a cero y sin mostrar curvaturas sistemáticas.

Gráfico 38.

Residuos estandarizados del modelo Log-normal versus Edad (Verificación de la linealidad)



Como se observa en el Gráfico 38, los residuos se distribuyen de manera simétrica alrededor de cero y la línea suavizada azul se mantiene relativamente plana en la mayor parte del rango de *Edad*, con una ligera desviación descendente hacia los valores más altos. Este comportamiento indica que, en general, la relación entre la *Edad* y el logaritmo del tiempo al evento es aproximadamente lineal, aunque podría existir una leve tendencia no lineal en los extremos superiores del rango etario.

Dado que esta desviación es pequeña y no evidencia un patrón sistemático pronunciado, se considera que el supuesto de linealidad se cumple de forma razonable, permitiendo mantener la variable *Edad* en su forma lineal dentro del modelo Log-normal.

La evidencia presentada en esta sección sugiere que los supuestos analizados del modelo AFT con distribución Log-normal no se cumplen de manera estricta. Sin embargo, esta situación no es atípica a aplicaciones con datos reales, incluso podría decirse que es particularmente frecuente en contextos sociales, donde la complejidad del fenómeno excede las simplificaciones propias de estos modelos. No obstante, los resultados siguen siendo útiles siempre que se interpreten con prudencia, entendiendo el modelo como una

aproximación al proceso subyacente y no como su representación exacta. En consecuencia, la incertidumbre de las estimaciones no queda íntegramente reflejada por los errores estándar e intervalos de confianza: existe además un componente de error de especificación (o de ajuste) que conviene ponderar de forma empírica y práctica. Por ello, las conclusiones deben acompañarse de diagnósticos, análisis de sensibilidad y comparaciones con especificaciones alternativas, a fin de evaluar la robustez de los hallazgos y acotar el alcance inferencial de las estimaciones.

4.4.3. Análisis de los Coeficientes del Modelo Log-normal

En esta sección se presentan y analizan los resultados del modelo paramétrico de supervivencia con distribución Log-normal. El objetivo es evaluar el efecto de cada covariable sobre el tiempo hasta el evento. Este modelo pertenece a la familia AFT (tiempo de falla acelerado), por lo que los coeficientes se interpretan como cambios multiplicativos en la escala del tiempo: $\exp(\beta_j)$ es el factor de aceleración. Valores de $\exp(\beta_j)$ mayores que 1 indican un alargamiento del tiempo esperado hasta el evento asociado a la covariable correspondiente; valores menores que 1 reflejan una reducción de dicho tiempo, manteniendo constantes las demás covariables.

A continuación, en el Cuadro 30 se presentan los coeficientes estimados (coef), los coeficientes exponenciados (exp(coef)), sus errores estándar (ee), el estadístico z, los niveles de significancia estadística (valor-p), y los intervalos de confianza al 95% (IC_bajo_95 y IC_alto_95) para cada coeficiente exponenciado. Asimismo, se incluye el parámetro de escala (σ), que representa la desviación estándar del término de error en la escala logarítmica del tiempo y proporciona una medida de la dispersión de los tiempos de supervivencia.

Cuadro 30.

Resultados para los coeficientes del modelo Log-normal

Variable o parámetro	coef	exp(coef)	ee(coef)	z	Valor-p	IC_bajo_95	IC_alto_95
Intercepto (β_0)	2.72	15.14	0.03	82.5	< 0.01	14.20	16.15
Log(Escala (σ))	0.49	1.62	0.00	124.2	< 0.01	1.61	1.64
R. California & Nevada (β_1)	0.06	1.07	0.03	2.3	0.02	1.01	1.13
R. Midwest (β_2)	-0.20	0.82	0.03	-6.6	< 0.01	0.77	0.87
R. Northeast (β_3)	-0.15	0.86	0.03	-4.6	< 0.01	0.81	0.92
R. Pacific Northwest (β_4)	0.29	1.34	0.03	8.9	< 0.01	1.25	1.42
R. Rockies & Red River (β_5)	-0.12	0.89	0.02	-5.0	< 0.01	0.85	0.93
Southeast (β_6)	-0.02	0.98	0.04	-0.6	0.56	0.91	1.05
Licencia LPN (β_7)	0.24	1.27	0.02	11.1	< 0.01	1.22	1.33
Licencia RN (β_8)	0.08	1.08	0.02	3.3	< 0.01	1.03	1.13
Licencia "Sin licencia" (β_9)	-0.79	0.45	0.08	-10.3	< 0.01	0.39	0.53
Otras Licencias (β_{10})	-0.10	0.90	0.05	-2.1	0.03	0.82	0.99
Ingresos (β_{11})	0.97	2.63	0.02	55.6	< 0.01	2.54	2.72
Edad (β_{12})	0.01	1.01	0.00	7.3	< 0.01	1.00	1.01
Sin_Edad (β_{13})	-0.21	0.81	0.10	-2.2	0.03	0.67	0.97

coef: coeficiente; ee(coef): error estándar del coeficiente.

IC_bajo_95: intervalo de confianza bajo del 95%.

IC_alto_95: intervalo de confianza alto del 95%.

La siguiente Ecuación E.56 se obtiene al sustituir los coeficientes estimados presentados en el Cuadro 30 dentro de la Ecuación E.55, correspondiente al modelo paramétrico con distribución Log-normal. Esta expresión representa el modelo ajustado para el logaritmo del tiempo hasta el evento ($log\ (T)$), e incluye las covariables analizadas. Cada término de la ecuación refleja el efecto estimado de la respectiva covariable sobre el log-tiempo de supervivencia, manteniendo constantes las demás variables.

$$
\begin{aligned}
log(T) = 2.718 &+ 0.065 \cdot \{Región\ California\ \&\ Nevada\} \\
&- 0.200 \cdot \{Región\ Midwest\} \\
&- 0.147 \cdot \{Región\ Northeast\} \\
&+ 0.290 \cdot \{Región\ Pacific\ Northwest\} \\
&- 0.120 \cdot \{Región\ Rockies\ \&\ Red\ River\} \\
&- 0.021 \cdot \{Región\ Southeast\} \\
&+ 0.240 \cdot \{Licencia_{LPN}\} + 0.077 \cdot \{Licencia_{RN}\} \\
&- 0.789 \cdot \{Licencia_{Sin\ licencia}\} - 0.103 \cdot \{Licencia_{Otras}\} \\
&+ 0.966 \cdot Ingreso \\
&+ 0.006 \cdot Edad - 0.2150 \cdot Sin\_Edad
\end{aligned}
$$

Ecuación E.56

Continuando con el análisis de los parámetros y coeficientes del modelo, como se observa en el Cuadro 30, tanto el intercepto del modelo (β_0) como el parámetro de escala (σ) presentan valores altamente significativos ($p < 0.001$), lo que evidencia que ambos parámetros contribuyen de manera sustancial al ajuste general del modelo. El intercepto ($\beta_0 = 2.718$) representa el logaritmo del tiempo medio de supervivencia esperado para la categoría de referencia, es decir, cuando todas las covariables toman el valor cero. En la escala original del tiempo (T_i), este valor corresponde a un tiempo medio de aproximadamente 15.14 semanas, lo que constituye el punto de partida sobre el cual actúan los factores explicativos del modelo.

En cuanto al parámetro de escala, el valor de $log(\sigma) = 0.486$ ($p < 0.001$) implica que la desviación estándar del término de error en la escala logarítmica es $\sigma = \exp(0.486) = 1.63$. Este resultado sugiere una dispersión moderada en los tiempos de supervivencia alrededor del promedio estimado, reflejando cierta heterogeneidad entre los individuos.

Respecto a las variables de tipo región, indicar primero que la categoría de referencia es Utah & Arizona, por lo que las interpretaciones se realizan en comparación con dicha zona. Lo primero que se puede apreciar es que los resultados señalan una heterogeneidad geográfica en la duración esperada del uso de la plataforma, ya que los valores de los coeficientes son bastante variados.

En primer lugar, la región California & Nevada muestra un coeficiente positivo ($\beta_1 = 0.065$, $p = 0.0215$), $\exp(\beta_1) = 1.067$. Esto sugiere que, manteniendo constantes las demás variables, los usuarios de esta región presentan un tiempo de supervivencia esperado aproximadamente 6.7% mayor que el de la región de referencia. Aunque el efecto es estadísticamente significativo, su magnitud es moderada, por lo que se esperan tiempos similares de supervivencia entre esta región y la región de referencia.

Por el contrario, las regiones Midwest ($\beta_2 = -0.200$, $p < 0.001$) y Northeast ($\beta_3 = -0.147$, $p < 0.001$) exhiben coeficientes negativos y altamente significativos, con factores de aceleración ($\exp(\beta)$) de 0.819 y 0.864, respectivamente. Esto implica que los tiempos de supervivencia esperados en estas regiones son 18.1 % y 13.6 % más cortos que en la región base, evidenciando una mayor propensión al abandono temprano de la plataforma.

En contraste, la región Pacific Northwest presenta un coeficiente positivo ($\beta_4 = 0.290$, $p < 0.001$) con un factor de aceleración de 1.336, lo cual se traduce en un incremento del 33.6 % en el tiempo esperado de uso respecto a la región de referencia. Este hallazgo sugiere una permanencia considerablemente mayor de los usuarios en esta zona en comparación con la región de referencia.

Por su parte, la región Rockies & Red River ($\beta_5 = -0.120$, $p < 0.001$) muestra un factor de aceleración de 0.887, indicando un tiempo esperado 11.3 % menor que el de la región base, lo que refleja una menor retención relativa. Finalmente, la región Southeast ($\beta_6 = -0.021$, $p = 0.5635$) presenta un coeficiente no significativo, con un factor de aceleración de 0.979, prácticamente indistinguible de 1, lo que sugiere que el comportamiento de los usuarios en esta región no difiere significativamente del de la región de referencia.

En conjunto, estos resultados evidencian diferencias regionales sustantivas en la duración del uso de la plataforma. Las zonas con mayor permanencia esperada son Pacific Northwest y, en menor medida, California & Nevada, mientras que las regiones Midwest, Northeast y Rockies & Red River se asocian con tiempos de supervivencia más cortos, lo que apunta a un comportamiento de abandono más rápido en dichas áreas.

En relación con las variables vinculadas al tipo de licencia profesional, la categoría de referencia en este modelo corresponde a las licencias "CNA/GNA", de modo que las demás categorías se interpretan comparativamente con respecto a este grupo.

Los resultados del Cuadro 30 muestran diferencias sustanciales entre los distintos tipos de licencia. En primer lugar, la variable de licencia "LPN" presenta un coeficiente positivo y altamente significativo ($\beta_7 = 0.240$, p < 0.001), con un factor de aceleración $\exp(\beta_7) = 1.271$. Esto implica que, manteniendo constantes las demás covariables, los usuarios con licencia "LPN" registran un tiempo de supervivencia esperado aproximadamente 27.1% mayor que el del grupo base ("CNA/GNA"). Dicho resultado sugiere que este tipo de profesionales tiende a permanecer más tiempo activos en la plataforma antes de abandonar su uso.

De manera similar, la licencia "RN" presenta también un efecto positivo y estadísticamente significativo ($\beta_8 = 0.077$, p = 0.001), con un factor de aceleración de $\exp(\beta_8) = 1.080$. Esto indica que los usuarios con licencia "RN" presentan un tiempo esperado de uso 8% superior al del grupo de referencia. Si bien la magnitud del efecto es menor que la observada para los profesionales con licencia "LPN", ambas categorías reflejan una mayor estabilidad y permanencia en la plataforma en comparación con los usuarios "CNA/GNA".

En contraste, la categoría "Sin licencia" muestra un comportamiento opuesto: su coeficiente negativo ($\beta_9 = -0.789$, p < 0.001) y su factor de aceleración de $\exp(\beta_9) = 0.454$ evidencian un tiempo de supervivencia esperado 54.6 % menor que el del grupo base. Este resultado indica una marcada propensión al abandono temprano, lo cual podría explicarse por una menor vinculación profesional con la práctica de enfermería o por mayores barreras de acceso a las oportunidades ofrecidas por la plataforma.

Por su parte, la categoría "Otras" licencias también presenta un efecto negativo y significativo ($\beta_{10} = -0.103$, p = 0.0342), con un factor de aceleración de $\exp(\beta_{10}) = 0.902$. Esto implica que los usuarios con licencias clasificadas en esta categoría muestran un tiempo esperado de permanencia 9.8% menor al del grupo "CNA/GNA", aunque la magnitud del efecto es moderada.

En conjunto, los resultados indican un gradiente claro de permanencia asociado al nivel de licencia profesional: los usuarios con licencias "LPN" y "RN" exhiben los mayores tiempos esperados de supervivencia en la plataforma, seguidos por el grupo "CNA/GNA", mientras que los individuos con otras licencias y sobre todo sin licencia presentan los menores niveles de retención. Estos hallazgos sugieren que el grado de acreditación profesional está positivamente asociado con la continuidad de uso del sistema, posiblemente debido a una mayor integración laboral, estabilidad contractual, mejores condiciones de remuneración económica, y una mayor motivación profesional entre los grupos con licencia formal.

En cuanto a la variable de *Ingresos*, esta presenta un coeficiente positivo ($\beta_{11} = 0.966$) y un valor-p altamente significativo ($p < 0.001$), con un factor de aceleración $\exp(\beta_{11}) = 2.627$. Este resultado indica que, manteniendo constantes las demás covariables, los usuarios "con ingresos" al inicio del uso de la plataforma exhiben un tiempo de supervivencia esperado aproximadamente 162.7 % mayor que aquellos "sin ingresos" (categoría de referencia). En otras palabras, un mayor nivel de ingresos se asocia con una mayor permanencia en la plataforma y una menor probabilidad de abandono en el corto plazo. Este hallazgo puede interpretarse en función de la estabilidad laboral y económica que caracteriza a los profesionales con mayores ingresos, quienes tienden a mantener mayor uso del sistema como parte de su rutina de trabajo o fuente de oportunidades laborales estables.

Por su parte, la variable *Edad* también muestra un coeficiente positivo ($\beta_{12} = 0.006$) y altamente significativo ($p < 0.001$), con un factor de aceleración $\exp(\beta_{12}) = 1.006$. Esto implica que, por cada unidad adicional en la edad del usuario (medida en años), el tiempo de supervivencia esperado en la plataforma aumenta aproximadamente un 0.6 %. Por lo tanto, cada 10 años de edad aumenta el tiempo de supervivencia esperado en aproximadamente 6.2%, esta consistencia estadística sugiere que los usuarios de mayor edad tienden a ser más estables en su participación, posiblemente debido a un mayor sentido de compromiso profesional. Cabe señalar que la variable dicotómica de *Edad* se incluyó exclusivamente como mecanismo de control para el manejo de valores faltantes en la variable continua *Edad*; por ello, no es objeto de interpretación sustantiva, ni extraer conclusiones a partir de su estimación.

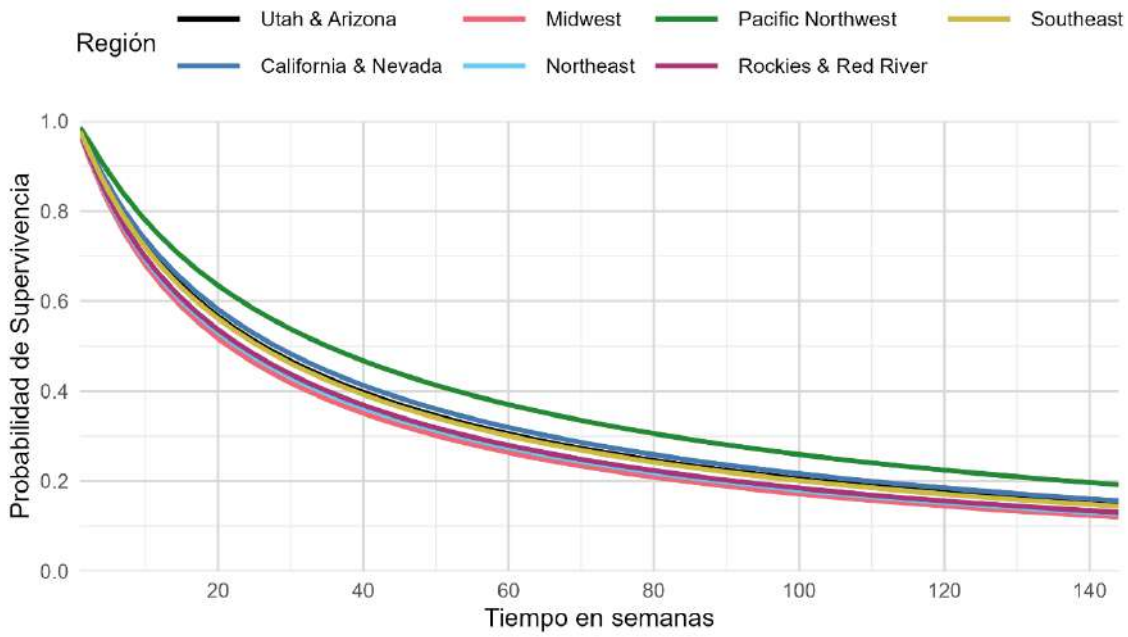
4.4.4. Gráficos de las funciones de supervivencia y riesgo (Modelo Log-normal)

En esta sección se presentan las curvas de supervivencia $S(t)$ y las tasas de riesgo instantáneo $h(t)$ estimadas a partir del modelo Log-normal (enfoque AFT). A diferencia del modelo de Cox, aquí los efectos se interpretan principalmente como aceleraciones o desaceleraciones del tiempo de supervivencia. No obstante, la lectura cualitativa de las curvas es análoga: curvas de $S(t)$ más altas indican mayor permanencia esperada a lo largo del horizonte temporal, y picos en $h(t)$ señalan ventanas de mayor probabilidad condicional de abandono.

A continuación, se presenta el Gráfico 39, que muestra las curvas de supervivencia por región estimadas con el modelo Log-normal. En él se comparan, en un mismo horizonte temporal, las probabilidades de permanecer activas en la plataforma para cada una de las siete regiones en estudio, lo que permite visualizar diferencias sistemáticas en la retención esperada entre perfiles geográficos.

Gráfico 39.

Curvas de supervivencia según la Región, obtenidas con el Modelo Log-normal

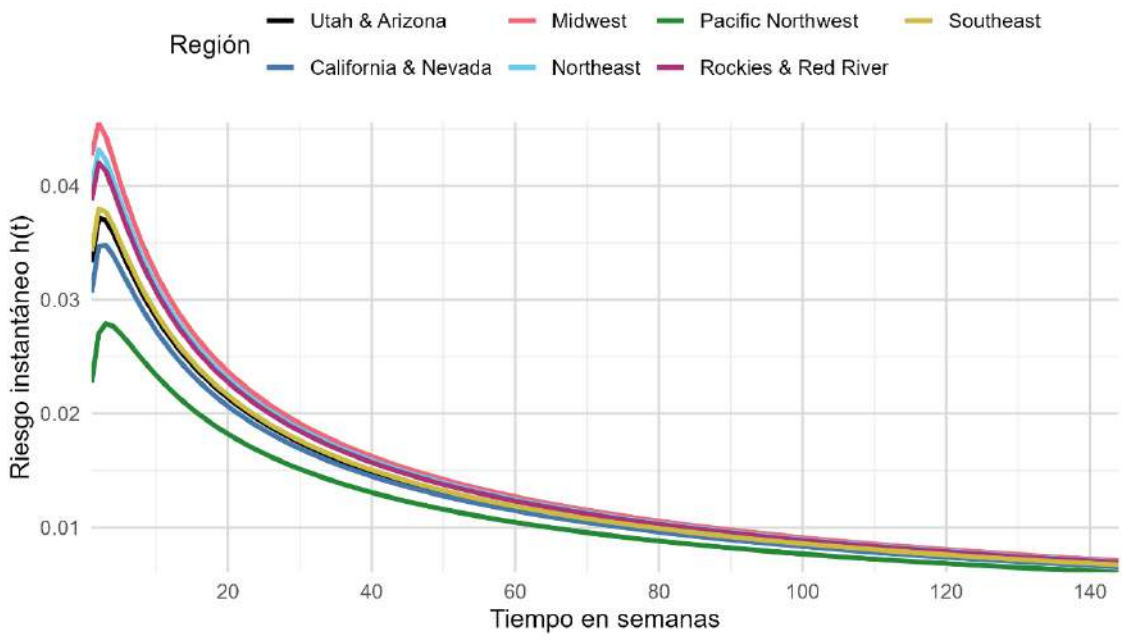


Como se observa en el Gráfico 39, las siete curvas se mantienen relativamente próximas, pero exhiben un orden estable y claro. Pacific Northwest destaca con la mayor supervivencia (curva sistemáticamente por encima del resto), lo que sugiere, de acuerdo con este modelo Log-normal, tiempos de uso más prolongados. En el extremo opuesto, Midwest se ubica como la región con la curva más baja durante el periodo de estudio, evidenciando menor permanencia de los usuarios en la plataforma. Southeast se sitúa muy cercana a la región de referencia Utah & Arizona, con diferencias no significativas, y California & Nevada presenta tiempos de supervivencia ligeramente superiores a los de Utah & Arizona.

De forma análoga, el Gráfico 40 presenta las curvas de riesgo instantáneo $h(t)$ para cada región. Las tasas exhiben la forma característica del modelo Log-normal: un pico temprano (alrededor de la semana 5) seguido de una caída monótona y prolongada.

Gráfico 40.

Riesgo instantáneo según la Región, obtenido con el Modelo Log-normal

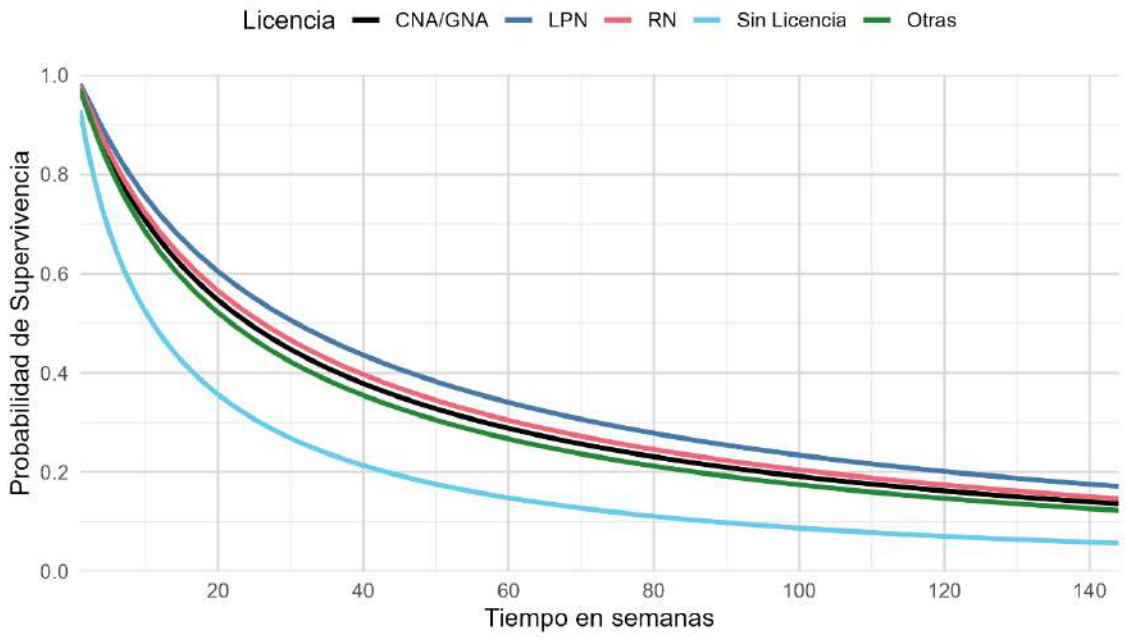


En el Gráfico 40 se observa que Midwest, seguida de Northeast, presenta los picos iniciales más altos, lo que señala un mayor riesgo de abandono en las primeras semanas; en contraste, Pacific Northwest que mantiene el riesgo más bajo, especialmente al inicio. En las primeras 5 semanas aproximadamente, se aprecia el tramo más pronunciado de las curvas; posteriormente, todas descienden y convergen hacia niveles bajos y decrecientes, y alrededor de las 100 semanas prácticamente tienden a coincidir. En este contexto, los picos identifican ventanas críticas de intervención temprana, mientras que los descensos reflejan la atenuación progresiva del riesgo entre quienes permanecen activos.

Seguidamente, se presenta el Gráfico 41, el cual muestra las curvas de supervivencia por tipo de licencia estimadas con el modelo Log-normal. De igual forma, esta comparación permite observar diferencias claras en la permanencia esperada de los usuarios según su licencia profesional de enfermería, especialmente durante las primeras semanas y a lo largo de todo el periodo de estudio.

Gráfico 41.

Curvas de supervivencia según la Licencia, obtenidas con el Modelo Log-normal

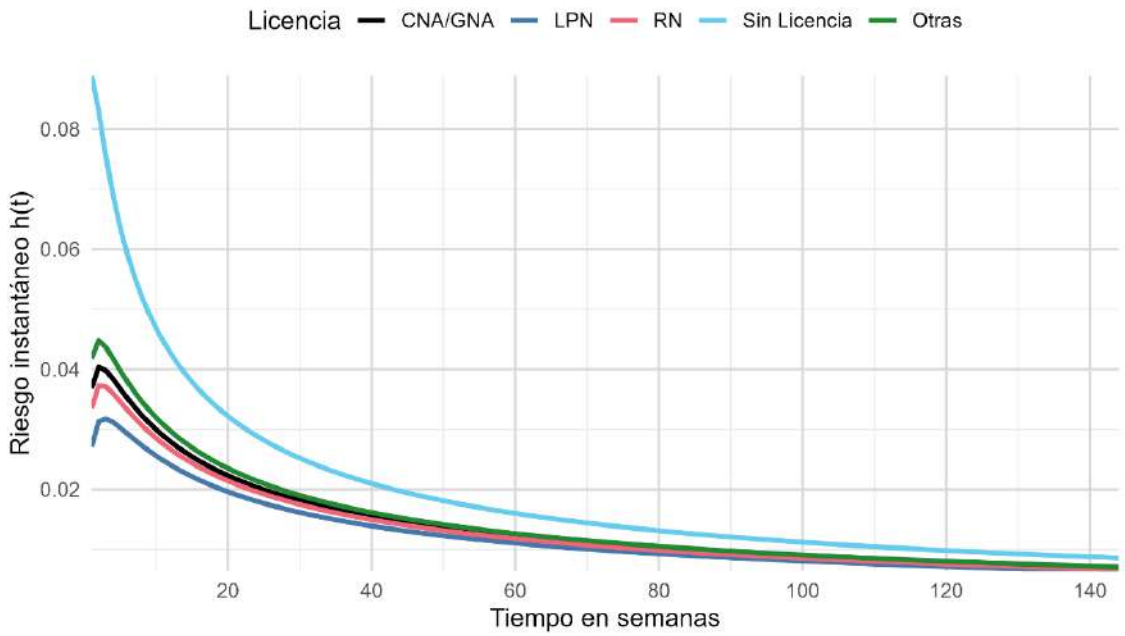


Analizando el Gráfico 41, se aprecia que la licencia "LPN" exhibe la mayor supervivencia de forma sostenida. Las categorías "RN", "CNA/GNA" (referencia) y "Otras" se ubican en un nivel intermedio, próximas entre sí y en dicho orden. En contraste, la categoría "Sin licencia" presenta claramente la supervivencia más baja, con un descenso muy pronunciado al inicio que se mantiene como brecha durante el horizonte, lo que sugiere tiempos de uso notablemente más cortos para esta categoría.

También, se presenta el Gráfico 42, que muestra el riesgo instantáneo $h(t)$ por tipo de *licencia* bajo el modelo Log-normal. Las curvas comparten un pico temprano (aproximadamente en la semana 5 o antes) seguido de un descenso prolongado, lo que facilita identificar ventanas de riesgo elevado al inicio y su posterior atenuación.

Gráfico 42.

Riesgo instantáneo según la Licencia, obtenido con el Modelo Log-normal



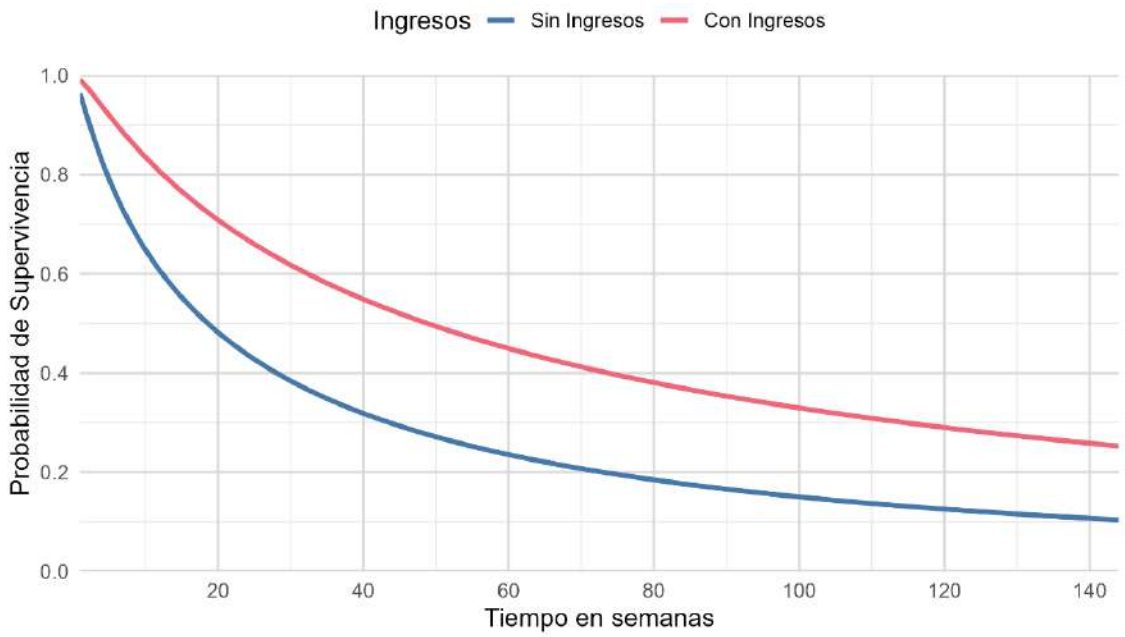
En el Gráfico 42, se observa que "Sin licencia" concentra el riesgo más alto, con un pico inicial particularmente marcado. En el extremo opuesto, "LPN" mantiene el riesgo más bajo

a lo largo del periodo, seguida de "RN"; "CNA/GNA" y "Otras" se sitúan en una franja intermedia y muy próxima entre sí. Tras las primeras 5 semanas aproximadamente, todas las curvas descienden de manera sostenida y, hacia las 80–100 semanas, tienden a converger en niveles bastante bajos, evidenciando una reducción progresiva de la probabilidad condicional de abandono entre quienes continúan activos.

Posteriormente, se presenta el Gráfico 43, que compara las curvas de supervivencia según la condición de *Ingresos* bajo el modelo Log-normal. Este gráfico permite visualizar la probabilidad de permanencia de quienes reportan ingresos frente a quienes no en los primeros 2 meses de haberse registrado ("con ingresos" y "sin ingresos").

Gráfico 43.

Curvas de supervivencia según *Ingresos*, obtenidas con el Modelo Log-normal



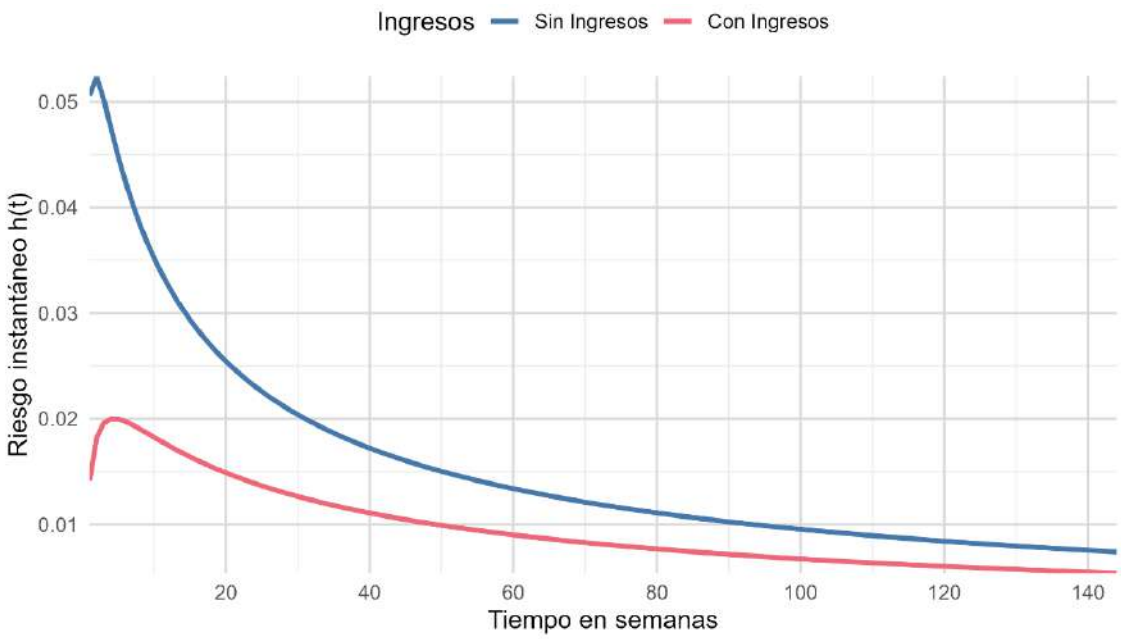
De esta forma, en el Gráfico 43 se aprecia una separación clara y sostenida: el grupo "con ingresos" exhibe mayor supervivencia durante todo el periodo, mientras que "sin ingresos" muestra una caída más pronunciada desde las primeras semanas. La brecha se mantiene

estable a lo largo del tiempo, lo que sugiere tiempos de uso más prolongados para quienes cuentan con ingresos en los primeros 2 meses del estudio.

Por su parte, el Gráfico 44 muestra las curvas del riesgo instantáneo $h(t)$ para ambas categorías de la variable *Ingresos*, generadas a partir del modelo Log-normal. A partir de este Gráfico 44 se puede concluir que la categoría "sin ingresos" concentra el riesgo más alto, con un pico inicial marcado y valores superiores en todo el intervalo. En contraste, "con ingresos" mantiene un riesgo más bajo y menos pronunciado; hacia las últimas semanas ambas curvas se acercan a niveles bajos, aunque persiste una ventaja clara para el grupo "con ingresos", coherente con la mayor supervivencia observada en el Gráfico 43.

Gráfico 44.

Riesgo instantáneo según Ingresos, obtenido con el Modelo Log-normal

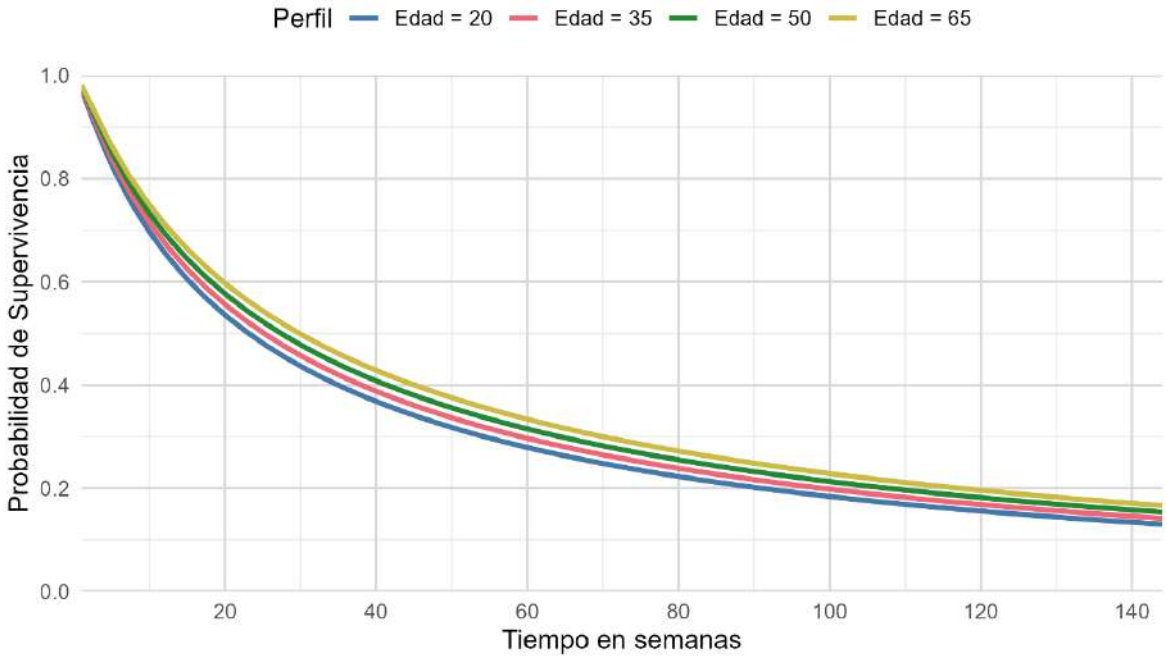


Seguidamente, se presenta el Gráfico 45, que compara las curvas de supervivencia para cuatro perfiles etarios (20, 35, 50 y 65 años), escogidos para mostrar el comportamiento de esta variable continua. Estas curvas son estimadas a partir del modelo Log-normal. La

visualización permite contrastar cómo varía la probabilidad de permanencia esperada a medida que cambia la edad de los usuarios de la plataforma.

Gráfico 45.

Curvas de supervivencia según las Edades, obtenidas con el Modelo Log-normal



En el Gráfico 45 se aprecia un orden consistente entre las curvas: 65 años exhibe la mayor supervivencia en todo el periodo, seguido de 50 y 35 años, mientras que 20 años presenta la probabilidad más baja. Estas curvas, sugieren tiempos de uso más prolongados a edades mayores, acorde con lo mencionado en el análisis de los coeficientes de este modelo.

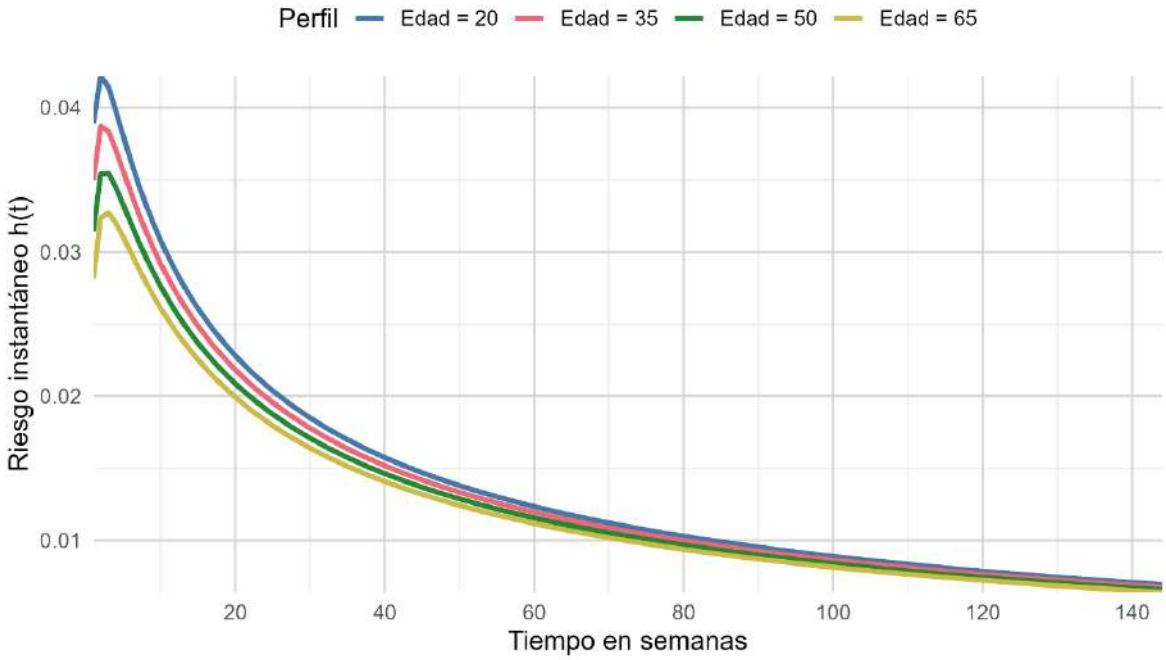
Por su parte, el Gráfico 46 muestra el riesgo instantáneo $h(t)$ para los mismos perfiles de edades. Como en los gráficos anteriores de $h(t)$, se genera un pico temprano, seguido de un descenso continuo hasta el final del periodo, lo que facilita identificar el tramo de mayor probabilidad condicional de abandono y su posterior atenuación.

A partir de este Gráfico 46 se desprende que 20 años concentra el pico inicial más alto, seguido por 35 y 50 años, mientras que 65 años mantiene el riesgo más bajo desde el inicio.

Además, tras las primeras 5 semanas aproximadamente, las curvas decrecen de manera sostenida y tienden a acercarse hacia las últimas semanas. Indicando que los riesgos instantáneos más elevados de abandono del uso de la plataforma se encuentran en las primeras 5 semanas después de haberse inscrito en la misma.

Gráfico 46.

Riesgo instantáneo según las Edades, obtenido con el Modelo Log-normal



4.5. ANÁLISIS DE SENSIBILIDAD DE LA FECHA DE CENSURA

Con el fin de evaluar la solidez de los resultados, se realizó un análisis de sensibilidad respecto a la definición de la fecha de censura. En este estudio, se estableció un intervalo de 60 días antes del cierre de la información (del 1 de agosto al 1 de octubre de 2024) para identificar a los usuarios censurados y, a partir de ello, calcular los tiempos de abandono. Sin embargo, esta elección metodológica puede influir en la estimación de los modelos, por lo que resulta necesario examinar cómo varían los resultados al modificar dicho criterio.

Para esto, se construyeron escenarios alternativos que consideraron distintos umbrales para las fechas de censura, desplazando el punto de corte hacia atrás en el tiempo, con intervalos de referencia comprendidos entre el 1 de junio de 2024 y el 1 de octubre de 2024. Cada uno de estos escenarios permite generar nuevas bases de datos recalculando los tiempos de supervivencia bajo definiciones alternativas de censura. Posteriormente, se ajustaron nuevamente los modelos de supervivencia, específicamente el semi-paramétrico de Cox extendido con covariables dependientes del tiempo, así como el modelo paramétrico con distribución Log-normal, y se compararon los coeficientes estimados junto con sus intervalos de confianza al 95%.

Los gráficos que se presentan a continuación sintetizan este procedimiento metodológico y ofrecen una visión general de la estabilidad de los resultados del estudio frente a cambios en la definición de censura. De este modo, es posible determinar si las conclusiones dependen fuertemente de la elección del intervalo de 60 días o si, por el contrario, se mantienen consistentes bajo diferentes supuestos.

El Gráfico 47 presenta el análisis de sensibilidad de los coeficientes estimados del Modelo Cox extendido al modificar el umbral definido para la fecha de censura. En este gráfico se muestran los resultados correspondientes a la primera mitad de las covariables incluidas en el modelo, mientras que en el Gráfico 48 se reportan los resultados para la segunda mitad de las covariables.

Cada color del coeficiente y su intervalo de confianza representa un conjunto de datos generado bajo una fecha de partida distinta para establecer la censura (identificados en la leyenda como "Datos"), la cual corresponde al día inicial del intervalo de censura (la etiqueta corresponde al formato MM-DD, donde el primer número es el mes y el segundo el día). Por ejemplo, el código "06-01" indica que la censura se definió a partir del 1 de junio de 2024, mientras que "08-20" corresponde al 20 de agosto de 2024, y así sucesivamente.

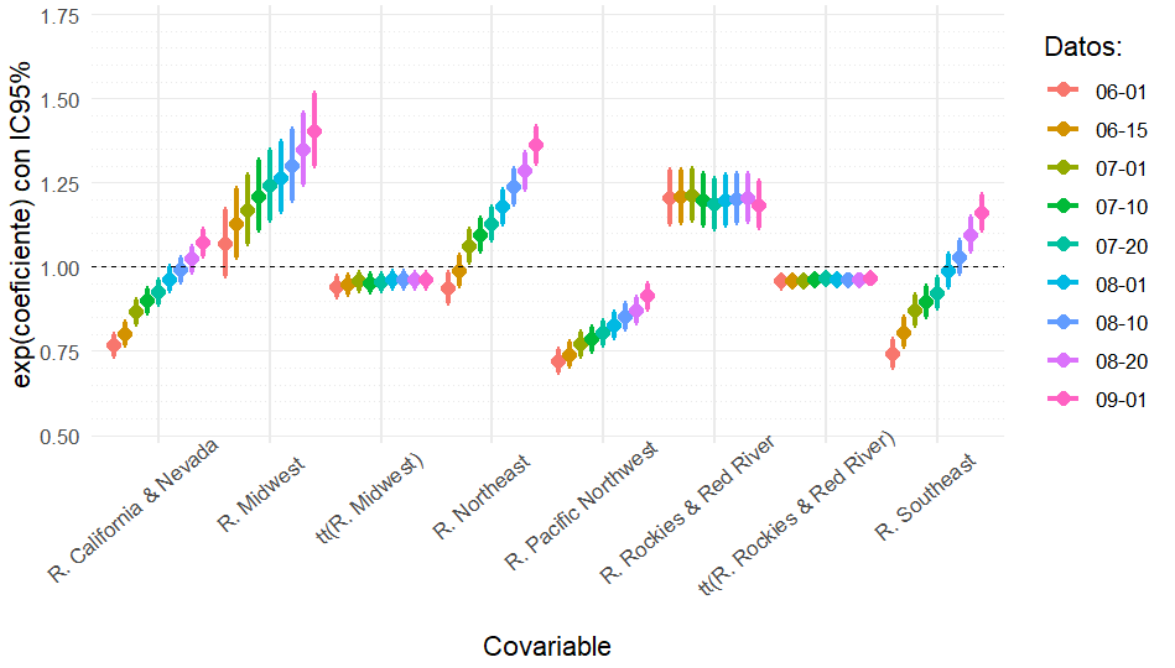
En el eje "x" se muestran las covariables y en el eje "y" el valor del coeficiente exponenciado del modelo de Cox, es decir, la Razón de Riesgos (*Hazard Ratio*, HR), acompañado de sus intervalos de confianza al 95%. De este modo, la comparación entre los distintos escenarios

permite evaluar cómo varían los estimadores y sus intervalos de confianza al modificar la definición temporal de la censura.

Gráfico 47.

Análisis de la sensibilidad de los coeficientes del Modelo extendido de Cox según la definición de la variable de Censura – Primera mitad de las covariables

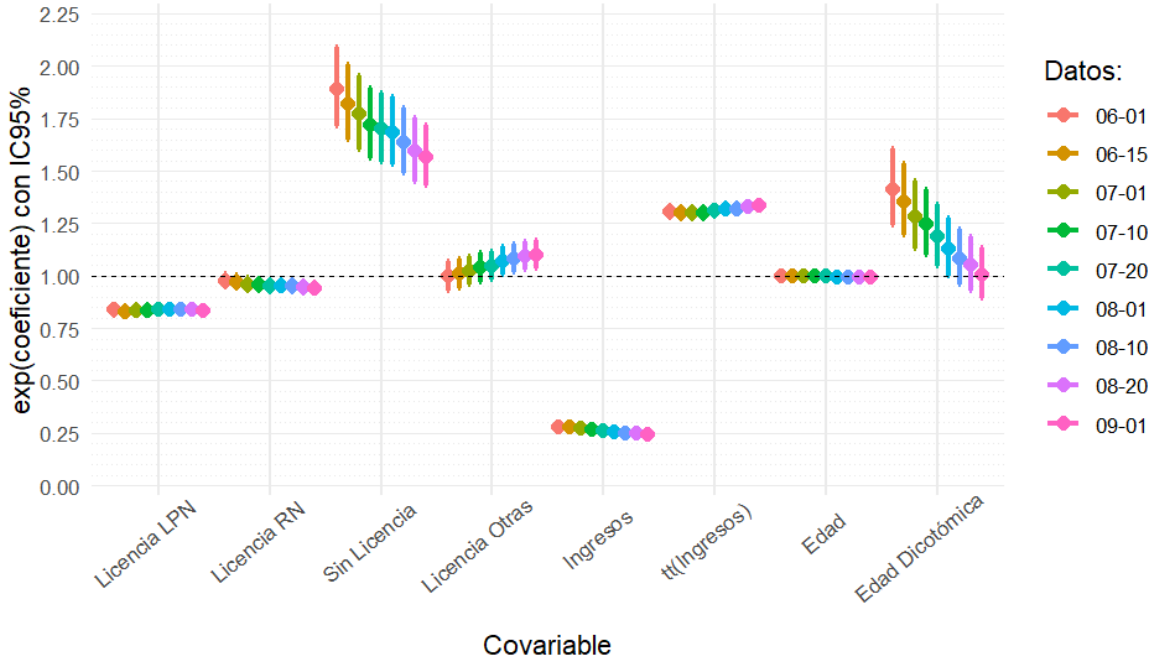
Su parte



R: Región

Gráfico 48.

Análisis de la sensibilidad de los coeficientes del Modelo extendido de Cox según la definición de la variable de Censura – Segunda mitad de las covariables



En el Gráfico 47, primera mitad de las covariables, correspondientes principalmente a las variables regionales, se observan patrones diferenciados según cada escenario de censura. Para el caso de la variable región "California & Nevada", la mayoría de los coeficientes exponenciados de los modelos se sitúan por debajo de uno, lo que equivale a HR menores que uno. Sin embargo, cabe destacar que existen variaciones entre escenarios de censura y que los intervalos de confianza se traslapan en algunos casos, aunque en la mayoría no ocurre así. Entre los escenarios, el que presenta el HR más bajo corresponde a la fecha de censura más temprana (06-01), lo que implica un riesgo aún más reducido en ese escenario. Es claro que, para esta variable, conforme la fecha de censura se acerca a fechas más recientes, el HR aumenta de forma progresiva.

Por su parte, en la región "Midwest", los HR también tienden a aumentar conforme la fecha de corte se mueve hacia escenarios más recientes. En la mayoría de los casos, estos valores son ligeramente mayores a 1, lo que refleja que los usuarios de esta región presentan un

riesgo de abandono un poco superior al de la categoría de referencia ("Utah & Arizona"). En cuanto al término de interacción tt(Midwest), los HR se mantienen muy similares entre los diferentes escenarios de censura y ligeramente menores a 1, lo que sugiere un efecto estable en esta covariable. La región "Northeast" también muestra un comportamiento similar al de "California & Nevada", a medida que la fecha de censura se desplaza hacia escenarios más recientes, los HR aumentan. En los escenarios más tempranos (06-01 y 06-15), los HR son menores o muy cercanos a 1, mientras que en los escenarios posteriores los valores superan este umbral. Al igual que en "California & Nevada", los intervalos de confianza se traslapan en algunos casos, pero no en la mayoría.

En cuanto a la región "Pacific Northwest", los HR se ubican por debajo de 1 en todos los escenarios, lo que implica un efecto protector consistente: los usuarios de esta región presentan un menor riesgo de abandono frente a la categoría de referencia. Aunque el comportamiento sigue el patrón general de otras variables, un aumento del HR en escenarios de censura más recientes. En el caso de la región "Rockies & Red River" y su término de interacción temporal, los coeficientes se mantienen muy estables entre los diferentes escenarios de censura. Esto significa que los usuarios de esta región no presentan diferencias sustanciales en su riesgo de abandono conforme varía la definición de censura. La región "Southeast" presenta un comportamiento mixto. Conforme el intervalo de censura disminuye hacia fechas más recientes, los HR tienden a aumentar, pero en algunos escenarios los valores se ubican por encima de 1 y en otros por debajo, siempre cercanos al valor de referencia. Esto sugiere una relación más inestable y dependiente del escenario de censura considerado.

En cuanto a la segunda mitad de las covariables (Gráfico 42), correspondientes a *Licencias*, *Ingresos* y *Edad*, los patrones son, en general, más estables. Las categorías de *Licencia* "LPN", "RN" y "Otras" presentan HR muy similares en todos los escenarios de censura, además de intervalos de confianza al 95% relativamente estrechos, lo que indica una estimación más precisa y consistente. Por su parte, la categoría "Sin licencia" muestra intervalos más amplios, los cuales se traslapan entre escenarios de censura, y estos HR son en todos los casos mayores a 1. Para esta covariable, conforme la fecha de censura avanza hacia escenarios más recientes, los HR tienden a disminuir.

La covariable *Ingresos*, junto con su término de interacción $tt(Ingresos)$, y la variable *Edad*, muestran patrones comparables a los de las *licencias*. En estos casos, los HR varían muy poco entre escenarios, lo que refleja estabilidad. Finalmente, la variable "Sin_Edad" (dicotómica) presenta ligeras variaciones en los HR según los escenarios de censura, donde se observa que, conforme la fecha de corte de censura se mueve hacia fechas más recientes, los valores del HR tienden a disminuir ligeramente.

En conjunto, el análisis muestra que, aunque algunas covariables, como "California & Nevada", "Pacific Northwest", "Southeast" y "Sin licencia", evidencian cierta sensibilidad a la fecha de censura, en la mayoría de los casos los HR y sus intervalos de confianza se mantienen relativamente estables. Esto sugiere que las conclusiones principales del modelo de Cox extendido no se ven sustancialmente afectadas por variaciones razonables en la definición de censura. No obstante, la elección de dicha fecha debe declararse y considerarse explícitamente al interpretar resultados o formular conclusiones sobre el comportamiento de los usuarios, pues se trata de una decisión metodológica que introduce variaciones en las estimaciones y cuyo impacto conviene documentar mediante un análisis de sensibilidad como el presente.

De forma similar al análisis de sensibilidad de la definición de la censura para el modelo de Cox extendido, se realizó un procedimiento equivalente para el modelo Log-normal. En este caso, las primeras 6 covariables y sus coeficientes con intervalos de confianza al 95% se muestran en el Gráfico 49, mientras que las 7 covariables restantes se presentan en el Gráfico 50. Para cada fecha de corte del intervalo de censura se generó un modelo distinto; estos se indican en la leyenda como "Datos" y se diferencian por color. La etiqueta corresponde al formato MM-DD, donde el primer número es el mes y el segundo el día.

Gráfico 49.

Análisis de la sensibilidad de los coeficientes del Modelo Log-normal según la definición de la variable de Censura – Primera mitad de las covariables

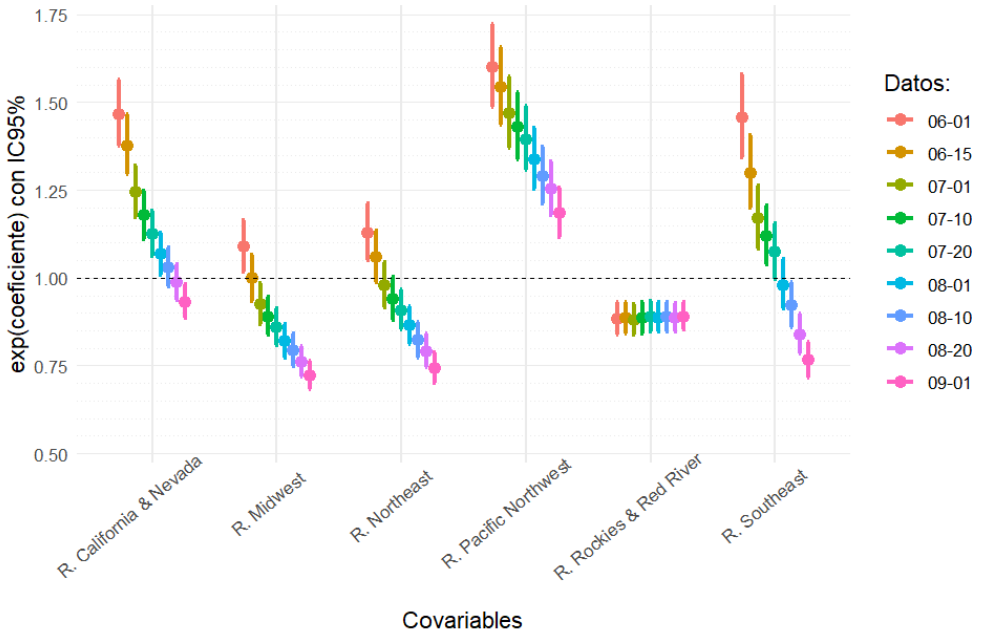
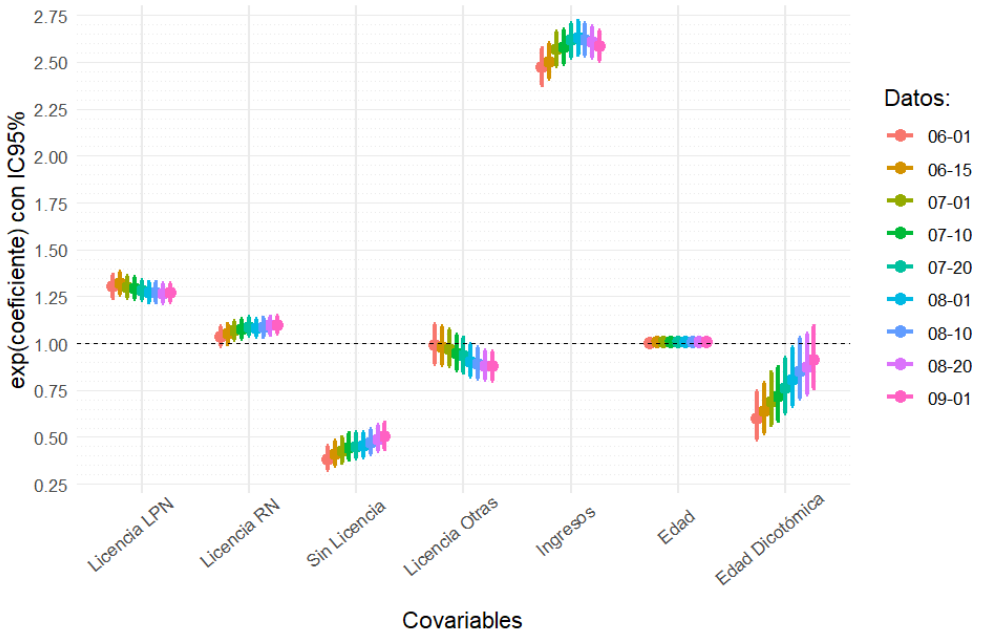


Gráfico 50.

Análisis de la sensibilidad de los coeficientes del Modelo Log-normal según la definición de la variable de Censura – Segunda mitad de las covariables

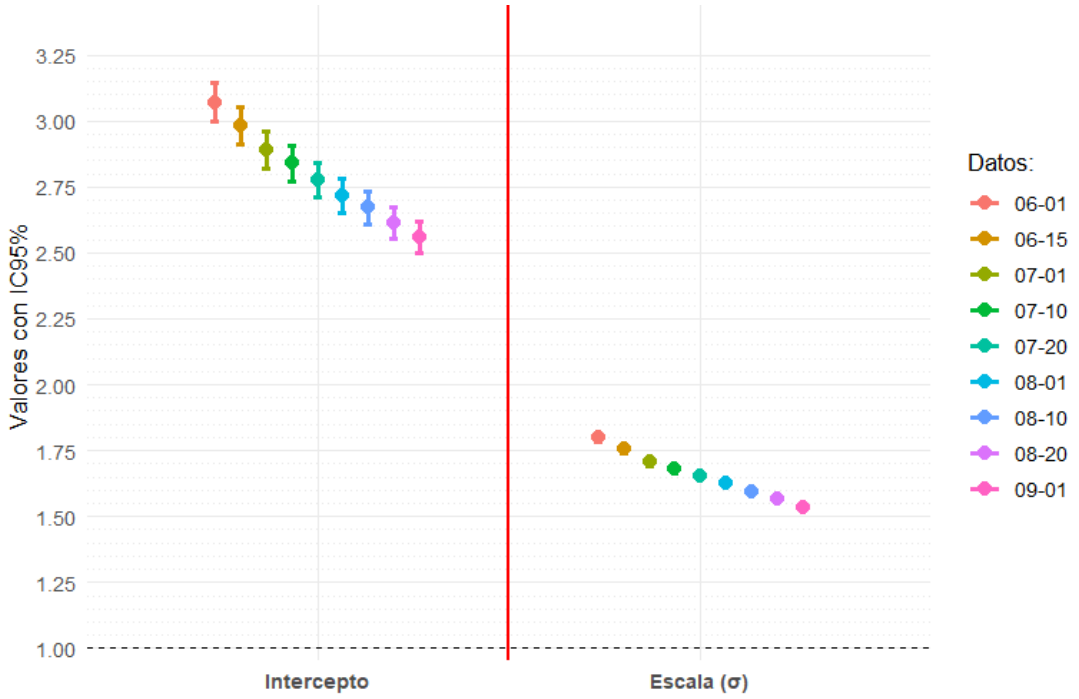


En complemento a los gráficos anteriores, el Gráfico 51 presenta, para cada uno de los nueve escenarios de censura, los valores estimados del intercepto del modelo Log-normal y del parámetro de escala (σ), junto con sus intervalos de confianza al 95%.

En el modelo Log-normal (AFT) el coeficiente exponenciado se interpreta como razón de tiempos (Time Ratio, TR). Valores > 1 indican tiempos de permanencia más largos (evento más tardío) respecto de la categoría de referencia; valores < 1, tiempos más cortos.

Gráfico 51.

Análisis de la sensibilidad del Intercepto y de la escala (σ) del Modelo Log-normal según la definición de la variable de Censura



De esta forma, observando el Gráfico 49, se aprecia que para la variable de la región "California & Nevada", los TR disminuyen al pasar del modelo con censura definida en 06-01 a 09-01. Entre escenarios consecutivos existe un traslape de los IC, sin embargo, no así entre escenarios no consecutives, lo que sugiere diferencias apreciables entre los valores de TR para diferentes escenarios de censura. Por su parte, para las variables de las regiones "Midwest", "Northeast", "Pacific Northwest", y "Southeast", los TR también disminuyen conforme el intervalo de censura disminuye hacia fechas más recientes. Y, únicamente la variable de región "Rockies & Red River", mantiene valores de TR bastante estables entre distintos escenarios de censura.

En el Gráfico 50 se observa que, para las licencias "LPN" y "RN", los TR se mantienen prácticamente constantes a lo largo de los distintos escenarios de censura, con intervalos de confianza estrechos y en gran medida traslapados; en consecuencia, el efecto de estas

covariables es estable y poco sensible a la fecha de corte de censura. Por su parte, la categoría "Sin licencia" muestra un patrón ligeramente ascendente: los TR aumentan conforme la censura se desplaza hacia fechas más recientes, con un traslape de intervalos importante. De forma similar, la categoría de "Otras" licencias, también presenta un traslape importante de los IC, pero muestra una tendencia opuesta, con TR que disminuyen al acortarse la ventana de censura. Lo que sugiere sensibilidad moderada a la definición de la censura.

Respecto de la covariable de *Ingresos*, esta presenta un pequeño incremento de los TR a medida que la fecha de corte de la censura también aumenta hacia fechas más recientes, donde los intervalos se traslapan bastante entre los distintos escenarios.

En cuanto a la covariable Edad, esta mantiene valores de TR muy estables para los diferentes escenarios de la censura. Finalmente, la variable "Sin_Edad" muestra un aumento de los TR conforme la censura se hace más reciente; el traslape de los intervalos es alto entre escenarios indicando cambios graduales pero coherentes.

En el Gráfico 51, que resume el intercepto del modelo Log-normal y el parámetro de escala (σ) para cada escenario, se aprecia que el intercepto disminuye de manera aparentemente monótona al mover la fecha de corte de la censura hacia fechas más recientes. Los intervalos de confianza se traslapan parcialmente entre escenarios adyacentes y menos entre los más distantes, lo que indica que el nivel base del tiempo de supervivencia (esto es, la mediana del tiempo en el grupo de referencia) se ajusta sistemáticamente conforme cambia la definición de la censura, de modo que dicho tiempo aumenta cuando se amplía el intervalo de censura (fechas de corte más antiguas).

Por su parte, la escala (σ) también desciende suavemente a medida que la fecha de corte de censura aumenta a fechas más recientes. Además, los intervalos de confianza son muy estrechos para este parámetro, lo que refleja baja variabilidad en su estimación y, por ende, alta precisión del parámetro bajo los distintos escenarios, no obstante, el valor de σ varia un poco según cada uno de los escenarios de censura (los intervalos no se traslapan). Valores mayores de σ implican tiempos de supervivencia más dispersos (mayor heterogeneidad entre individuos). Estas variaciones en los valores para los distintos modelos (distintos escenarios

de censura) señala una importante sensibilidad de estos dos valores (Intercepto y σ) a la elección de la fecha de corte de la censura.

En síntesis, las variables que presentan mayor sensibilidad a la definición de las fechas de la censura son las variables de las regiones, excluyendo a "Rockies & Red River". Además, el intercepto muestra ser sensible a esta definición, así como el parámetro de Escala (σ). Lo que indica que conforme el intervalo de censura se amplía con fechas de corte más antiguas, los tiempos de supervivencia también aumentan y se tornan más dispersos.

Además del análisis de sensibilidad realizado a partir de los coeficientes de los modelos de supervivencia, se desarrolló un análisis complementario orientado a evaluar en qué medida la definición del intervalo de censura adoptado inicialmente (60 días antes de la fecha de cierre de la información) pudo conducir a errores de clasificación de los usuarios. En particular, se buscó identificar los casos en los que usuarios considerados como "no censurados" en el análisis original, es decir, clasificados como si hubieran abandonado la plataforma, en realidad continuaron activos y realizaron nuevas solicitudes después de la fecha final de recopilación de datos inicial (1 de octubre de 2024). Estos casos representan "falsos abandonos" que afectan la correcta estimación del tiempo de supervivencia.

De forma similar, también se calcularon los casos que fueron clasificados como "Censurados" en la base de datos, pero podrían haber sido clasificados como "no censurados" o *churn*. Este error de sub-detección de *churn* provoca entre otras cosas, como se mencionó en el análisis anterior, que los tiempos de supervivencia sean más dispersos y con intervalos de confianza más amplios o menos precisos.

Para esto, se amplió la información del estudio con registros de solicitudes hasta el 15 de agosto de 2025, lo que permitió verificar la actividad posterior de los usuarios y, con base en ello, recalcular de manera progresiva las fechas de censura. Dicho procedimiento consistió en desplazar la fecha de corte hacia atrás desde el 1 de octubre de 2024 en intervalos de cinco días, hasta un máximo de 200 días, generando así múltiples escenarios de censura. Luego, una vez calculados los "no censurados" y "censurados" con la información anterior al 1 de

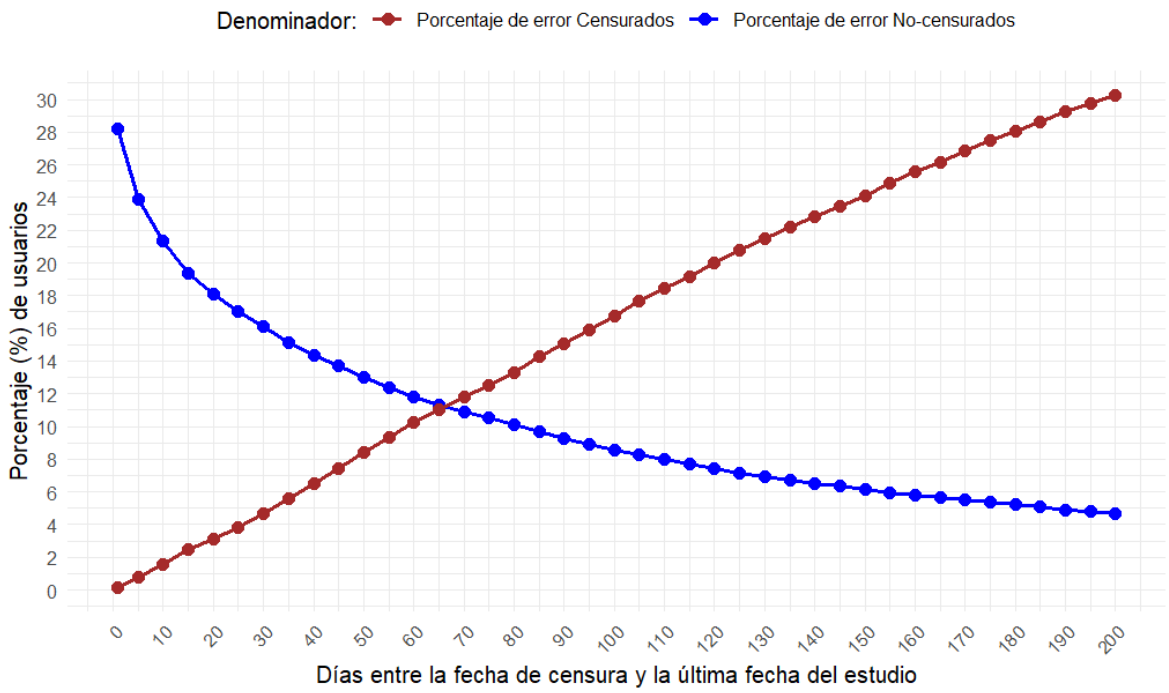
octubre de 2024, se comprueba con la información que va desde esa fecha hasta el 15 de agosto de 2025 si efectivamente esos usuarios debieron haberse considerado como "no censurados" o "censurados".

El Gráfico 52 resume visualmente este ejercicio. En el eje horizontal (x) se representa la distancia en días entre la fecha de censura considerada y la fecha final del estudio (intervalo de censura), mientras que en el eje vertical (y) se muestra el porcentaje de usuarios mal clasificados (usuarios clasificados erróneamente entre el total de usuarios de la base original). El gráfico presenta dos curvas: el primer caso corresponde a la curva en color azul, que representa los casos considerados como No-censurados (*churn*) pero que luego se descubrió que realizaron una solicitud (es decir, no estaban realmente *churn*), y como es un porcentaje de error, este valor se dividió entre el total de usuarios. Es decir, este caso en azul, es el porcentaje de usuarios que, bajo la fecha de censura considerada, deben ser clasificados como "no censurados" (con *churn*), pero que posteriormente vuelven a generar al menos una solicitud en la plataforma, por lo que representan un error de clasificación como "no censurados". Por otro lado, el caso de la línea café, representa el porcentaje de usuarios que fueron tratados como censurados, pero que podrían haber sido correctamente clasificados como *churn* a la luz de la información disponible posterior al periodo de estudio original; es decir, son la proporción de usuarios que el modelo trata como censurados (se asume que siguen activos), pero que, viendo todo el horizonte de observación, probablemente ya habían abandonado y quedaron mal etiquetados como censurados, por lo que esto constituye un error de sub-detección de *churn*.

Esta representación permite evaluar la magnitud de los errores de clasificación bajo diferentes definiciones de censura y valorar si el criterio de 60 días constituye un balance adecuado entre minimizar dichos errores y evitar un exceso de censura.

Gráfico 52.

Porcentajes de error en la clasificación de usuarios como abandonos (*churn*) y censurados según el intervalo de censura



La interpretación de este resultado (Gráfico 52) revela una tensión metodológica importante. Cuando el intervalo de censura es muy corto, por ejemplo, de solo 10 o 15 días antes del cierre, el porcentaje de usuarios mal clasificados como "no censurados" (línea azul) es elevado, ya que en un periodo tan breve es frecuente que los usuarios considerados como abandonos aún realicen nuevas solicitudes en fechas posteriores al cierre. En contraste, conforme el intervalo de censura se amplía, por ejemplo, a 100 o 150 días antes del cierre, el porcentaje de usuarios mal clasificados como "no censurados" disminuye de manera significativa, pues resulta poco probable que quienes permanecieron inactivos durante tanto tiempo retomen su actividad después del cierre de la recopilación de información.

Sin embargo, extender demasiado el intervalo de censura no es una solución libre de costos, ya que, al aumentar el intervalo, también se incrementa el número de usuarios tratados como censurados, lo que reduce la cantidad de eventos observados (abandonos efectivos) y limita la precisión y el poder estadístico de los modelos de supervivencia. En otras palabras, un

intervalo muy corto incrementa los errores de clasificación (falsos abandonos), mientras que un intervalo de censura demasiado largo aumenta la censura y restringe la información útil disponible para el análisis.

Por este motivo, además del porcentaje de error asociado a los falsos abandonos, se evaluó también el porcentaje de error correspondiente a los usuarios clasificados como "censurados", que en realidad podrían haber sido considerados como *churn*, dado que no volvieron nunca a retomar su actividad en la plataforma. Como se aprecia en el Gráfico 52, el equilibrio entre estos dos tipos de error (clasificar como "no censurado" a quien no lo es, o como "censurado" a quien probablemente ya había abandonado) parece alcanzarse cuando ambos se sitúan alrededor de un 11%, lo cual ocurre aproximadamente con un intervalo de 65 días (es decir, considerar como censurados a los usuarios cuya última solicitud se registró entre el 27 de julio de 2024 y el 1 de octubre de 2024).

En este sentido, este análisis de sensibilidad permite explorar de forma empírica la existencia de un punto de equilibrio: un intervalo de censura lo suficientemente amplio como para reducir los errores de clasificación en *churn*, pero lo bastante corto como para no generar un exceso de censura innecesaria. La evidencia mostrada en el Gráfico 52 indica que este equilibrio se alcanza en torno a los 65 días, donde las curvas presentan una pendiente más suave (disminución más lenta del porcentaje de errores) y los porcentajes de usuarios mal clasificados se sitúan cerca del 11%, una magnitud considerada razonable. Además, en dicho punto el porcentaje de error asociado a los casos tratados como "censurados" se mantiene en niveles moderados, evitando una sub–detección excesiva de *churn*.

En consecuencia, la elección de un intervalo de 60 días en el análisis principal se justifica como un balance metodológico adecuado: reduce de forma sustancial la proporción de falsos abandonos sin sacrificar en exceso la cantidad de eventos observados, lo que se traduce en modelos más robustos y conclusiones más confiables. Se optó por 60 días, y no por 65, porque ofrece un nivel de error muy similar al punto de equilibrio identificado, pero resulta más sencillo de interpretar y comunicar, ya que corresponde a dos meses.

4.6. PREDICCIONES DE LOS TIEMPOS DE SUPERVIVENCIA

En esta sección se traducen los modelos ajustados en secciones previas en predicciones operativas de interés para la plataforma: tiempos de supervivencia esperados asociados a probabilidades de supervivencia específicas (S), por ejemplo, la mediana ($S = 0.50$), para perfiles definidos de usuarios. Además, se realizan comparaciones entre enfoques paramétricos, semi-paramétricos y no paramétricos a fin de evaluar la consistencia y el desempeño predictivo de estos resultados. En particular, se emplean como modelos paramétricos de referencia el Log-normal y Gamma, seleccionados por su buen desempeño global en el estudio: el Log-normal por su ajuste superior según AIC/BIC, y el Gamma por su desempeño sistemáticamente mejor en el error de Brier integrado (IBS) en las pruebas pareadas de validación.

Para una mejor estimación y comparación de los tiempos de supervivencia calculados, también se incluyen: (a) la curva Kaplan-Meier, estimador no paramétrico ampliamente utilizado que no requiere suponer una forma funcional para los tiempos y maneja adecuadamente la censura; y (b) el modelo de Cox extendido, que permite obtener curvas de supervivencia ajustadas a covariables que varían en el tiempo. Esta combinación posibilita contrastar los resultados de una forma más rigurosa.

En el caso del modelo Log-normal, las predicciones del tiempo de supervivencia para un perfil de usuarios x se pueden obtener de manera directa a partir de los parámetros del modelo. De forma que:

$$\mu = \beta_0 + \sum_j \beta_j x_j\,, \qquad \sigma > 0. \tag{Ecuación E.57}$$

$$\log T \sim \mathcal{N}(\mu,\ \sigma^2) \tag{Ecuación E.58}$$

$$t_S(x) = \exp\{\mu(x) + \sigma\, z_{1-S}\}, \qquad z_{1-S} = \Phi^{-1}(1-S) \tag{Ecuación E.59}$$

donde, T es la variable aleatoria "tiempo hasta el evento"; σ es el parámetro de escala (desviación estándar) de $\log T$ (controla la dispersión en la escala logarítmica); β_0 es el

intercepto del modelo en escala log–tiempo; β_j son los coeficientes asociado a las covariables; $\mathcal{N}(\mu, \sigma^2)$ es la distribución Normal con media μ y varianza σ^2; $t_S(x)$ es el tiempo asociado a la probabilidad de supervivencia S para el perfil de usuarios x; z_{1-S} es el cuantil 1 menos S de la Normal estándar; $\Phi^{-1}(\cdot)$ es la inversa de la función de distribución acumulada de la Normal estándar.

De esta forma, para realizar las predicciones con el modelo Log-normal, se debe seleccionar el perfil del o los usuarios que se desean investigar. A manera de ejemplo, se va a escoger un usuario que trabaje en la región "Utah & Arizona" (región de referencia) con la licencia "RN", con una Edad de 40 años, y "con ingresos" en sus primeros 2 meses de trabajar con la plataforma. De esta forma, utilizando los coeficientes presentes en el Cuadro 30 y en la Ecuación E.56, como se muestra a continuación, se puede obtener el tiempo esperado de supervivencia con una probabilidad (S) del 0.50 y del 0.30 con la ayuda de las Ecuaciones E.57, E.58, y E.59.

Resolviendo la Ecuación E.57, se obtiene lo siguiente:

$$\mu = 2.718 + 0.065 \cdot 0 - 0.200 \cdot 0 - 0.147 \cdot 0 + 0.290 \cdot 0 - 0.120 \cdot 0 - 0.021 \cdot 0 + 0.240 \cdot 0 + 0.077 \cdot 1 - 0.789 \cdot 0 - 0.103 \cdot 0 + 0.966 \cdot 1 + 0.006 \cdot 40 - 0.2150 \cdot 0 = 4.001$$

Para probabilidad del 0.50, donde σ =1.625; $z_{1-0.50}$ =0.00; el tiempo es:

$$t_{0.50} = exp(4.001 + 1.625 * 0) = \mathbf{54.65}\ \textbf{semanas}.$$

Para probabilidad del 0.30, donde σ =1.625; $z_{1-0.30} = 0.524$; el tiempo es:

$$t_{0.30} = exp(4.001 + 1.625 * 0.524) = \mathbf{128.06}\ \textbf{semanas}.$$

Mientras que, para el modelo Gamma, los tiempos $t_S(x)$ se derivan de la función cuantil (inversa de la función de distribución acumulada) de la distribución Gamma, evaluada con los parámetros de forma κ, y tasa $\lambda(x)$ del perfil x considerado. En la práctica, dicho cuantil se obtiene numéricamente. De forma compacta se tienen las siguientes ecuaciones:

$$t_S(x) = Q_{Gamma}\big(1 - S;\ \text{forma} = \kappa,\ \text{tasa} = \lambda(x)\big) \qquad \text{Ecuación E.60}$$

$$\lambda(x) = \exp\{\eta(x)\} \qquad \text{Ecuación E.61}$$

$$\eta(x) = x^{\top} * \beta \qquad \text{Ecuación E.62}$$

donde, κ es el parámetro de forma de la Gamma (controla la curvatura del *hazard*, si $\kappa = 1$, el modelo se reduce al Exponencial); $\lambda(x)$ es la tasa (rate); $\eta(x)$ es el predictor lineal para el perfil de usuarios x; β son los coeficientes asociados a las covariables x. Importante mencionar que en este caso la fórmula de $\eta(x)$ también incluye un intercepto, al cual se le suele llamar "tasa", el cual corresponde al parámetro que se muestra en el Cuadro 31 con el nombre de "Tasa". Por su parte, $Q_{\text{Gamma}}(p; \kappa, \lambda)$ es la función cuantil de la distribución Gamma, definida como el valor t que satisface $F_{\text{Gamma}}(t; \kappa, \lambda) = p$, donde F_{Gamma} es la función de distribución acumulada de Gamma.

Con el objetivo de realizar las predicciones, en el siguiente Cuadro 31 se presentan los coeficientes estimados (coef) para el modelo Gamma ajustado con los datos que se utilizan en este estudio. En dicho cuadro también se presentan los coeficientes exponenciados (exp(coef)), sus errores estándar (ee), el estadístico z, los niveles de significancia estadística (valor-p), y los intervalos de confianza al 95% (IC_bajo_95 y IC_alto_95) para cada coeficiente exponenciado.

Cuadro 31.

Resultados para los coeficientes del modelo Gamma

Variables	coef	exp(coef)	ee(coef)	z	Valor-p	IC_bajo_95	IC_alto_95
shape	-0.31	0.74	0.005	-67.9	< 0.001	0.730	0.743
rate	-3.84	0.02	0.001	-6317.9	< 0.001	0.022	0.022
R. California & Nevada	-0.06	0.95	0.024	-2.4	0.0155	0.902	0.989
R. Midwest	0.14	1.15	0.024	5.8	< 0.001	1.098	1.209
R. Northeast	0.17	1.19	0.027	6.5	< 0.001	1.127	1.251
R. Pacific Northwest	-0.24	0.79	0.028	-8.7	< 0.001	0.744	0.830
R. Rockies & Red River	0.07	1.07	0.019	3.8	< 0.001	1.035	1.115
Southeast	-0.02	0.99	0.030	-0.4	0.6683	0.930	1.047
Licencia LPN	-0.21	0.82	0.018	-11.3	< 0.001	0.788	0.846
Licencia RN	-0.05	0.95	0.019	-2.7	0.006	0.916	0.986
Licencia Sin licencia	0.61	1.84	0.058	10.6	< 0.001	1.642	2.058
Otras Licencias	0.09	1.10	0.039	2.5	0.0138	1.020	1.188
Ingresos	-0.64	0.53	0.015	-43.4	< 0.001	0.513	0.543
Edad	-0.01	0.99	0.001	-9.2	< 0.001	0.993	0.995
Sin_Edad	0.19	1.2079	0.074	2.5	0.0109	1.0444	1.3971

coef: coeficiente; ee(coef): error estándar del coeficiente.

IC_bajo_95: intervalo de confianza bajo del 95%.

IC_alto_95: intervalo de confianza alto del 95%.

Con esta información, y de forma similar a como se realizaron las predicciones con el modelo Log-normal, a continuación, se procede a realizar los cálculos de los tiempos de supervivencia para el mismo perfil de usuario y con las mismas probabilidades de supervivencia ($S = 0.50\ y\ S = 0.30$), pero en este caso utilizando el modelo Gamma.

En primer lugar, con base en los valores de los coeficientes presentes en el Cuadro 31, el perfil anteriormente definido, y la Ecuación E.62, se obtiene:

$$\eta(x) = x^{\top} * \beta = -3.8348 - 0.0569 \cdot 0 + 0.1415 \cdot 0 + 0.1719 \cdot 0 - 0.2413 \cdot 0 + 0.0719 \cdot 0 - 0.0130 \cdot 0 - 0.2025 \cdot 0 - 0.0510 \cdot 1 + 0.60887 \cdot 0 + 0.0959 \cdot 0 - 0.6389 \cdot 1 - 0.0060 \cdot 40 + 0.1889 \cdot 0 = \mathbf{-4.7653}$$

Luego, a partir de la Ecuación E.61, se obtiene: $\lambda = \exp(-4.7653) = 0.0085$.

Con estos datos, y la Ecuación E.60, para la probabilidad del 0.50, el tiempo es

$t_{0.50}(x) = Q_{Gamma}(1 - 0.50;\ \text{forma} = 0.7363,\ \text{tasa} = 0.0085) =$ **51.80 semanas**.

Mientras que, para una probabilidad de supervivencia de 0.30, el tiempo es:

$t_{0.30}(x) = Q_{Gamma}(1 - 0.30;\ \text{forma} = 0.7363,\ \text{tasa} = 0.0085) =$ **100.78 semanas**.

Por otra parte, las incertidumbres de las predicciones (intervalos de confianza para $t_S(x)$ o $S(t \mid x)$) se pueden obtener mediante la propagación de la variabilidad de los estimadores del modelo (por ejemplo, a partir de la matriz de varianzas y covarianzas de $\hat{\beta}$ y de los parámetros de escala). Sin embargo, dada la complejidad algebraica de este procedimiento, es recomendable utilizar algún software para realizar estos cálculos. Por esta razón, no se detallan aquí los aspectos computacionales de este procedimiento. Cabe señalar que, una vez disponibles los detalles del modelo (coeficientes, parámetros y matrices de varianzas y covarianzas) y definido el perfil de usuario de interés, la obtención de estos intervalos es sencilla y directa con la ayuda de software. Es de esta forma que se obtuvieron las estimaciones para el "usuario promedio" mostradas en el Gráfico 53.

A continuación, se presenta el Gráfico 53, que muestra los tiempos de supervivencia del "usuario promedio" para distintas probabilidades de supervivencia (S). En el eje "x" se representa S (probabilidad de continuar más allá de t) y en el eje "y" el tiempo en semanas asociado a cada $S(t_S)$. Este Gráfico 53 se trata de un gráfico de puntos con barras de error: cada punto corresponde a la predicción de uno de los enfoques considerados y las barras de error muestran intervalos de confianza del 95 % obtenidos con el procedimiento específico de cada modelo. Se incluyen los modelos paramétricos Gamma y Log-normal, el semi-paramétrico Cox-extendido y el estimador no paramétrico Kaplan-Meier.

Este gráfico cumple un propósito descriptivo y comparativo: en primer lugar, permite comprender los tiempos medianos de supervivencia de los "usuarios promedios" de la plataforma digital, además, facilita visualizar el patrón de los tiempos predichos t_S en distintos niveles de S, contrastar la posición relativa de cada modelo (desplazamientos verticales como indicio de sesgo sistemático), comparar la incertidumbre entre enfoques mediante la longitud de los intervalos de confianza (95%), identificar discrepancias localizadas en valores extremos de S, y evaluar el solapamiento de los intervalos como señal de compatibilidad entre estimaciones.

Además, de forma muy similar al Gráfico 53, el Gráfico 54 también presenta los tiempos de supervivencia del "usuario promedio" para distintas probabilidades de supervivencia (S). Pero en este caso, en formato de curvas de supervivencia para cada uno de los cuatro modelos considerados.

En el Gráfico 54, cada curva corresponde a uno de los cuatro modelos y el sombreado alrededor de ella indica los intervalos de confianza del 95 %. Como en este caso se representa un espectro mucho más amplio de probabilidades de supervivencia y tiempos t_S, el detalle de los intervalos resulta algo más difícil de apreciar que en el Gráfico 53. No obstante, precisamente por abarcar todo el rango de S, el Gráfico 54 permite visualizar con mayor claridad el contraste global entre los cuatro modelos y ofrece una visión más rica del patrón de t_S a lo largo de todo el espectro de probabilidades de supervivencia, mostrando cómo se separan o solapan las curvas según el modelo. Esto facilita la comparación conjunta de la forma de las curvas y del grado de solapamiento entre enfoques en todo el rango de probabilidades, proporcionando un resumen sintético del comportamiento de los tiempos de supervivencia en función de S.

Gráfico 53.

Tiempos de supervivencia del usuario promedio para distintas probabilidades (S): contraste entre modelos paramétricos, el modelo semi-paramétrico Cox-extendido y Kaplan-Meier, con intervalos de confianza al 95%

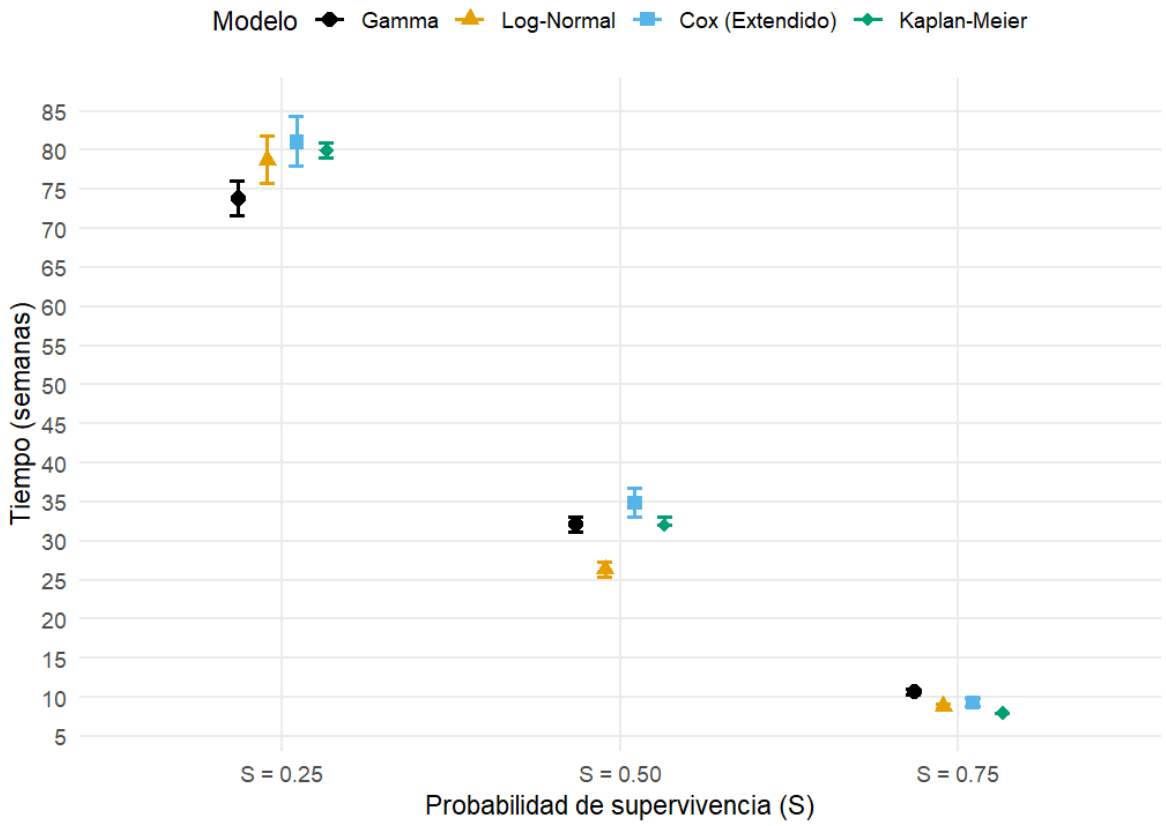
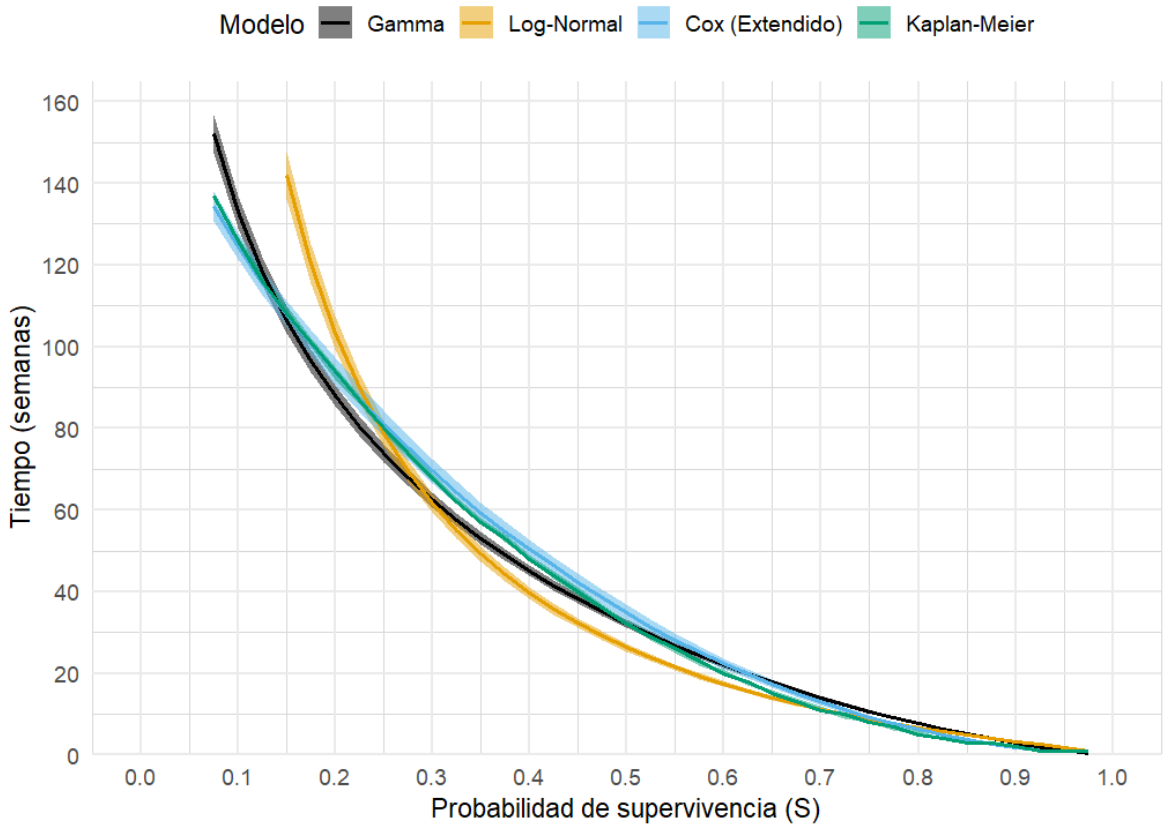


Gráfico 54.

Curvas de supervivencia para el usuario promedio (Tiempos de supervivencia para distintas probabilidades de supervivencia): contraste entre modelos paramétricos, el modelo semi-paramétrico Cox-extendido y Kaplan-Meier, con intervalos de confianza al 95%



Observando los Gráficos 53 y 54, no se identifica un modelo que sea sistemáticamente el más alto o el más bajo en todos los niveles de S. Esta ausencia de desplazamientos verticales persistentes es una buena señal: sugiere que las conclusiones no dependen en gran medida del supuesto de modelo (paramétrico vs. no/semi-paramétrico), reduce el riesgo de sesgo sistemático y refuerza la robustez de las predicciones para los usuarios de la plataforma digital.

En cuanto a los tiempos medianos de supervivencia $t_{0.50}$ (esto es, para $S = 0{,}50$) del usuario promedio que se aprecian en el Gráfico 53, Cox-extendido se sitúa alrededor de 33-37, Kaplan-Meier cerca de 32–33 semanas, Gamma en torno a 31–33 semanas, y Log-normal

entre 25–27 semanas. En conjunto, el orden relativo es $\text{Cox-ext.} \gtrsim \text{KM} \gtrsim \text{Gamma} > \text{Log} - \text{normal}$, con diferencias máximas del orden de 6 a 12 semanas (Cox-ext vs. Log-normal). Estos tiempos, que se acompañan de los IC del 95% mostrados en las barras de error, permiten identificar diferencias significativas entre modelos, sin embargo, las mismas son pequeñas. A partir de estos resultados, y siendo conservadores, se puede inferir que los tiempos medianos de los usuarios promedio de la plataforma digital se encuentran entre 25 y 37 semanas desde que empezaron a utilizarla.

En general, el modelo de Cox-extendido presenta los intervalos de confianza (IC) más amplios, mientras que Log-normal y Gamma presentan IC similares, por su parte, el modelo KM presenta los intervalos más estrechos. Observando el Gráfico 53, se observan solapamientos importantes entre los IC de los modelos, lo que sugiere diferencias pequeñas a moderadas en las predicciones puntuales. Cabe recordar que las predicciones de estos enfoques tienden a ser menos confiables en las colas de la distribución (extremos altos y bajos del tiempo).

Además, el gráfico muestra que la incertidumbre (longitud de las barras) es mayor en S bajos (tiempos largos) y menor en S altos (tiempos cortos), como es esperable por la variabilidad creciente en la cola superior de las distribuciones.

Para $S = 0.25$ se obtienen, como es esperable, tiempos más altos en todos los modelos, siendo Cox-extendido y Kaplan-Meier los que muestran los valores mayores, mientras que Gamma reporta los menores valores. Mientras que, para $S = 0.75$, Gamma presenta los valores más altos (alrededor de las semanas 10-11), mientras que Log-normal, Cox-extendido y Kaplan-Meier se agrupan algo más abajo (alrededor de las semanas 8–10). El orden relativo entre modelos cambia con S, reforzando que no hay un modelo con los mayores tiempos en todo el rango.

Dado que las diferencias entre los distintos modelos para el "usuario promedio", en especial en la mediana $t_{0.50}$, parecen estables, esto favorece decisiones consistentes (p. ej., metas de retención o planificación operativa). Sin embargo, las discrepancias se acentúan en los extremos de S (especialmente en los valores bajos, donde los t son mayores), donde pequeñas variaciones pueden tener impacto operativo (p. ej., estimar la persistencia de usuarios "muy

fieles" o de "alto riesgo"). Por ello, en esas zonas conviene reportar conjuntamente varios enfoques; esta estrategia puede ayudar a reducir el riesgo de decisiones basadas en estimaciones inestables.

Sumado a los resultados anteriores, como anexo a este documento, se desarrolló una aplicación *Shiny* (Chang, 2025) que permite: (i) seleccionar el o los modelos con los que se desea predecir (modelo Gamma, Log-normal, o Cox-extendido); (ii) definir el perfil del usuario; y (iii) elegir los niveles de supervivencia S de interés. La aplicación calcula, mediante procedimientos similares a los descritos en esta sección, los tiempos de supervivencia t_S y sus intervalos de confianza del 95% para cada combinación solicitada, y genera un gráfico de puntos con barras de error que facilita la visualización de los resultados, y el contraste entre modelos y perfiles seleccionados.

Más detalles sobre esta aplicación *Shiny* se indican en el Anexo C.

CAPÍTULO V. CONCLUSIONES

En términos generales, los resultados muestran que el riesgo instantáneo de abandono de los usuarios en la plataforma digital de contratación *per diem* de personal de enfermería en Estados Unidos no es constante en el tiempo. Tanto los estimadores no paramétricos como los modelos semi-paramétricos y paramétricos, coinciden en un patrón con tres fases principales: un pico de riesgo de abandono elevado en las primeras semanas tras el ingreso, un tramo intermedio relativamente estable con riesgos de abandono bajos y, finalmente, un incremento tardío del riesgo de abandono entre los usuarios que permanecen activos durante horizontes prolongados. Este comportamiento respalda la idea de que las decisiones de abandono se concentran especialmente en las etapas iniciales de la relación con la plataforma, pero que existe también un segundo momento de vulnerabilidad en horizontes largos, cercanos aproximadamente al año y medio de permanencia. Esta forma no monótona del riesgo sugiere que la retención de usuarios requiere tanto intervenciones tempranas (para evitar la deserción inicial) como estrategias de seguimiento para quienes han permanecido por periodos extendidos.

Este patrón, con un riesgo de abandono inicial elevado seguido de una disminución, concuerda con lo observado en estudios previos: Lai y Zeng (2014) señalaron que en una biblioteca digital la mayoría de los usuarios abandonaba dentro de los primeros tres meses de uso, y de forma similar Pérez Pizarro (2022) reportó una pérdida considerable de suscriptores en los primeros meses de la suscripción digital (especialmente tras expirar las promociones iniciales).

En cuanto al análisis de las covariables de los modelos, se encontró evidencia de que todas las variables consideradas resultaron estadísticamente significativas: tipo de licencia de enfermería, región de trabajo, edad e ingresos en los primeros dos meses de uso de la plataforma (variable dicotómica).

El análisis por covariables permite identificar perfiles claramente diferenciados. En cuanto al tipo de licencia, los usuarios con credenciales más especializados, especialmente quienes poseen licencia "LPN" y, en menor medida, "RN", presentan los menores riesgos de

abandono, mientras que el grupo "Sin licencia" concentra el riesgo más elevado y un patrón de deserción temprana más marcado. A nivel regional, las zonas "Midwest" y "Rockies & Red River" exhiben los mayores riesgos de abandono, mientras que "Pacific Northwest" se distingue por una mayor permanencia relativa, coherente con mejores perspectivas de retención.

Por lo tanto, el orden de usuarios con mayor probabilidad de permanecer más tiempo en la plataforma, según el tipo de licencia, sería: "LPN", "RN", "CNA/GNA", "Otras" y, finalmente, "Sin licencia". Por su parte, según la región de trabajo, el orden esperado aproximado de mayor a menor permanencia sería: "Pacific Northwest", "California & Nevada", "Utah & Arizona", "Southeast", "Northeast", "Rockies & Red River", y finalmente "Midwest", concentrando los riesgos más altos. No obstante, este orden puede variar un poco a medida que aumenta el tiempo de permanencia en la plataforma (son variables dependientes del tiempo).

De acuerdo con el análisis del supuesto de riesgos proporcionales y los resultados del modelo de Cox extendido, se identificó que los efectos de algunas covariables no se mantienen constantes en el tiempo. En particular, las regiones "Midwest" y "Rockies & Red River", así como la variable dicotómica de *Ingresos* en los primeros dos meses, presentan interacciones significativas con la función logarítmica del tiempo, lo que evidencia riesgos no proporcionales. En las regiones "Midwest" y "Rockies & Red River" el riesgo de abandono es inicialmente más alto que en la región de referencia ("Utah & Arizona"), pero este exceso de riesgo disminuye gradualmente conforme transcurren las semanas, llegando incluso en el caso de "Rockies & Red River" a situarse en niveles similares o ligeramente inferiores a los de la región base hacia el final del periodo de observación. Por su parte, los usuarios que generan ingresos en los primeros meses presentan al inicio un riesgo de abandono sustancialmente menor que aquellos "sin ingresos" en dicho periodo; sin embargo, la interacción positiva con el tiempo indica que este efecto protector se atenúa progresivamente, de modo que la diferencia en el riesgo de abandono entre ambos grupos se reduce a medida que aumenta el tiempo de permanencia en la plataforma.

Estos resultados indican que las estrategias de retención podrían priorizar, especialmente en las primeras etapas de uso de la plataforma, a los usuarios de "Midwest" y "Rockies & Red River", mientras que en "Pacific Northwest" las condiciones actuales parecen favorecer naturalmente la permanencia de los usuarios en la plataforma. De forma complementaria, desde la perspectiva del tipo de licencia, conviene reforzar las acciones de retención en los perfiles con mayor riesgo de abandono, particularmente en el grupo "Sin licencia". En contraste, los usuarios con licencia "LPN" y "RN" presentan, en términos generales, una mayor permanencia esperada, por lo que no sería prioritario enfocar en ellos este tipo de estrategias.

El hecho de que las personas con licencias "LPN" y "RN" presenten menores riesgos de abandonar la plataforma, en comparación con quienes poseen otras licencias o ninguna, es esperable. Ya que, en general, se trata de personal más calificado, con mayor nivel de formación y, en consecuencia, con tarifas o salarios por hora más altos dentro de la plataforma.

Es razonable suponer, además, que este grupo de usuarios con licencias "LPN" y "RN", tiene una mayor responsabilidad profesional y un mejor posicionamiento para obtener turnos a través de la aplicación que quienes cuentan con credenciales de menor nivel o carecen de licencia. En otras palabras, estos usuarios tienen más probabilidades de alcanzar rápidamente el éxito en el uso de la plataforma (conseguir y completar turnos con mayor facilidad y con una mejor remuneración), y ese éxito temprano favorece que se mantengan activos durante más tiempo. En este sentido, resulta esperable que los usuarios cuyas credenciales se asocian con mejores remuneraciones presenten tiempos de permanencia más largos. Todo lo anterior se traduce en una mayor percepción de utilidad de la plataforma y en un mayor interés sostenido por continuar utilizándola como una fuente relevante de oportunidades laborales.

En contraste, para la variable Región de trabajo no es posible formular conclusiones con la misma claridad, ya que no resulta evidente, *a priori*, cuáles zonas son más favorables para este tipo de trabajos de enfermería. Comprender en detalle las causas de las diferencias observadas constituye un tema amplio, que amerita un estudio específico, pues los mecanismos potenciales son numerosos.

Por ejemplo, en algunas regiones las tarifas por hora pueden ser más altas que en otras, pero al mismo tiempo podrían enfrentarse costos de vida más elevados. También influyen la relación entre oferta y demanda de turnos de enfermería, la disponibilidad y calidad del transporte, los tiempos de traslado, así como el estado de las instalaciones y el ambiente laboral de los centros de salud asociados a la plataforma en esas regiones. Estos factores, entre otros, pueden variar de manera considerable tanto entre regiones como dentro de una misma región, lo que puede hacer que ciertas zonas resulten más atractivas y, por ende, retengan mejor a los usuarios que otras.

En cuanto a la variable de *Ingresos* en los primeros meses de uso de la plataforma emerge como uno de los determinantes más importantes del comportamiento de supervivencia. Los usuarios que logran generar ingresos tempranamente presentan, en las primeras semanas, un riesgo de abandono sustancialmente menor que aquellos que no registran ingresos, si bien este efecto protector se atenúa gradualmente con el tiempo. Este resultado es coherente con la interpretación de los ingresos como un indicador dinámico de vinculación, aprendizaje y éxito operativo en la plataforma, más que como una característica fija del usuario. Esta variable constituye un claro determinante del "éxito" en el uso de la plataforma: a mayor éxito inicial, medido en términos de poder obtener ingresos, mayor es la probabilidad de que el usuario continúe utilizando la plataforma en el mediano y largo plazo.

Este hallazgo es consistente con la literatura previa; por ejemplo, Lai y Zeng (2014) segmentaron a los usuarios de una biblioteca digital según su comportamiento de uso y observaron que aquellos con menor actividad inicial presentaban mayor propensión a la fuga, una tendencia análoga a la relación aquí identificada entre bajos ingresos iniciales (indicativo de menor involucramiento) y mayor riesgo de *churn*. De manera similar, Pérez Pizarro (2022) destacó que una frecuencia de uso más baja elevaba considerablemente la probabilidad de cancelación en la suscripción digital, reforzando la noción de que un mayor *engagement* temprano se traduce en una mayor permanencia del cliente en diversos entornos digitales.

Asimismo, el análisis por edad confirma un gradiente claro: los usuarios más jóvenes presentan mayores riesgos de abandono, mientras que los grupos de mayor edad muestran trayectorias de permanencia más estables y favorables. Esta variable puede relacionarse con

factores como la responsabilidad, la experiencia laboral, la estabilidad en el proyecto de vida profesional y el compromiso con la profesión. En general, a mayor edad es razonable suponer mayores niveles de experiencia, estabilidad y compromiso, características que, de manera similar a lo observado con las licencias, tienden a asociarse con mejores tarifas, y mayor facilidad para conseguir trabajos. Todo ello redunda en un mayor interés por mantener el vínculo con la plataforma.

La combinación de estos factores permite delinear perfiles de riesgo. De manera sintética, los perfiles con menor probabilidad de *churn* corresponden a usuarios con licencia formal (en particular LPN), de mayor edad, que generan ingresos en los primeros meses y que se ubican en regiones con menor riesgo estructural, como "Pacific Northwest". En el extremo opuesto, los mayores riesgos de abandono se concentran en usuarios jóvenes, sin licencia, sin ingresos en los primeros meses y residentes en regiones con *hazard* más elevado (por ejemplo, "Midwest" o "Rockies & Red River"). Estos resultados ofrecen una base empírica clara para la estratificación de usuarios según su probabilidad de retención y para el diseño de estrategias focalizadas de intervención.

Desde el punto de vista metodológico, este estudio pone de manifiesto la utilidad de combinar enfoques no paramétricos (Kaplan-Meier), semi-paramétricos (modelo de Cox extendido con covariables dependientes del tiempo) y paramétricos. En particular, el modelo de Cox extendido presenta una capacidad de discriminación global moderada, con un C-index con validación cruzada de 0.581 y un IBS con validación cruzada de 0.181, sin indicios importantes de sobreajuste. Sus criterios de información son AIC = 680,588.5 y BIC = 680,715.1; dado que se basan en la verosimilitud parcial, estos valores solo son comparables con otras especificaciones dentro de la misma familia de modelos de Cox y no con los modelos paramétricos.

La comparación sistemática de las seis distribuciones paramétricas consideradas muestra que la Log-normal ofrece el mejor ajuste según los criterios de información (AIC = 330,247.7; BIC = 330,379.5), además presentó un C-index con validación cruzada de 0.601 y un IBS con validación cruzada de 0.201. Por su parte, la distribución Gamma presenta un ajuste ligeramente inferior (AIC = 330,959.6; BIC = 331,091.4), pero alcanza el mejor desempeño

en términos de error de predicción integrado, con un C-index con validación cruzada también de 0.601 y un mejor IBS con validación cruzada de 0.183. Considerando conjuntamente estas métricas, no se identifica un único modelo claramente superior en todas las dimensiones: la Log-normal destaca por su mejor ajuste global a los datos, la Gamma por su menor error predictivo entre los modelos paramétricos, y el modelo de Cox extendido por su flexibilidad para incorporar covariables que varían en el tiempo, y un menor error de predicción (IBS con VC) incluso que el modelo Gamma. La coherencia de los resultados obtenidos con estos tres enfoques complementarios refuerza la solidez de las conclusiones sustantivas del estudio.

Las curvas de supervivencia estimadas para el "usuario promedio" por estos tres modelos, muestran trayectorias muy similares, con un solapamiento importante de los intervalos de confianza para amplios rangos de probabilidad. En particular, las medianas de supervivencia para dicho usuario promedio "con ingresos" en los primeros dos meses se sitúan, según el modelo empleado, en un rango aproximado de 47.5 a 58.4 semanas, Cox extendido ofrece los valores más altos (55.7 ± 2.7 semanas), seguido de Log-normal (49.1 ± 1.5 semanas), y Gamma (48.7 ± 1.2 semanas) ligeramente por debajo. Esta proximidad en las predicciones centrales sugiere que, aun bajo supuestos de modelo distintos, el mensaje práctico es consistente: un usuario típico permanece activo en la plataforma del orden de varios meses (alrededor de 11 o 13 meses), lo que proporciona un horizonte temporal útil para el diseño de estrategias de retención y seguimiento. Por otro lado, de acuerdo con las funciones de riesgo (*hazard*) estimadas, la intervención temprana sobre los usuarios debería realizarse en las primeras semanas (primeras 6 semanas) desde su incorporación a la plataforma, ya que es en ese periodo donde sus riesgos instantáneos de *churn* son más altos.

Un aporte adicional relevante es el análisis de sensibilidad de la definición de censura, que permitió evaluar distintos intervalos para distinguir entre usuarios realmente inactivos y aquellos potencialmente observados de forma incompleta (censurados). Los resultados sugieren que un intervalo cercano a 60 días ofrece un compromiso razonable entre reducir la clasificación errónea de abandonos y evitar un exceso de censura, mejorando así la calidad de la variable de tiempo al evento y la estabilidad de los modelos. No obstante, esta elección implica aceptar cierto grado de sesgo residual asociado a la definición operativa de *churn*,

por lo que las estimaciones deben interpretarse con cautela y entender que este supuesto genera incertidumbre en los modelos.

En conjunto, estos hallazgos metodológicos respaldan el uso del análisis de supervivencia como marco adecuado para estudiar el *churn* en plataformas digitales de contratación *per diem* de personal de enfermería y ofrecen herramientas concretas para que los responsables de la toma de decisiones estratégicas puedan utilizar los tiempos y probabilidades de supervivencia en la planificación de metas de retención, la priorización de recursos y la evaluación de intervenciones dirigidas a los segmentos de mayor riesgo.

En conjunto, los modelos de supervivencia empleados presentaron un desempeño predictivo aceptable y permitieron identificar segmentos de usuarios de alto riesgo de abandono, brindando insumos concretos para diseñar estrategias de retención y optimizar la gestión de la plataforma. Estos resultados reafirman lo planteado por Lai y Zeng (2014) acerca de la utilidad del análisis de supervivencia para comprender la duración de la relación cliente-servicio en entornos digitales, y coinciden con la aplicación exitosa de modelos de supervivencia en el estudio de Pérez Pizarro (2022) para guiar acciones de fidelización en un servicio de suscripción digital.

Finalmente, este estudio presenta algunas limitaciones que deben ser consideradas al interpretar los resultados. En primer lugar, el horizonte temporal analizado es relativamente acotado, por lo que sería deseable ampliarlo en futuras investigaciones para capturar con mayor precisión la dinámica de largo plazo del abandono de usuarios. En segundo lugar, la información disponible se restringe principalmente a variables administrativas de la plataforma y no incorpora de forma explícita factores contextuales externos, como las condiciones del mercado laboral local, las características de los empleadores o las diferencias en la demanda de turnos entre regiones, ni variables comportamentales más detalladas de los usuarios, tales como los patrones de búsqueda de turnos, la frecuencia e intensidad de uso de la plataforma, el nivel de interacción con la aplicación o indicadores más directos de compromiso y satisfacción. La incorporación de este tipo de variables podría aportar información adicional a los modelos, lo que permitiría segmentar de manera más precisa a los usuarios y obtener estimaciones de los tiempos de supervivencia más robustas.

Asimismo, la incorporación de nuevas covariables y la construcción de variables derivadas, por ejemplo, indicadores más finos de actividad en la plataforma, patrones de búsqueda de turnos o métricas de compromiso digital, representan una línea de trabajo prometedora y con amplio potencial de desarrollo en este campo de estudio.

De igual forma, es posible que persista heterogeneidad no observada entre los usuarios, es decir, variabilidad en el riesgo de *churn* asociada a factores no medidos, como motivaciones individuales, preferencias laborales, percepciones sobre la plataforma o condiciones particulares del entorno institucional. Frente a ello, futuras investigaciones también podrían complementar el enfoque aquí utilizado mediante modelos más flexibles, como los modelos de fragilidad, los modelos multiestado, los modelos de eventos recurrentes o métodos basados en *deep learning* para datos de supervivencia. En particular, los modelos de fragilidad permitirían capturar parte de esa heterogeneidad no observada mediante efectos aleatorios sobre la función de riesgo, mientras que los modelos multiestado resultarían útiles para representar trayectorias más complejas de los usuarios dentro de la plataforma, por ejemplo, la transición entre estados de actividad, inactividad temporal y abandono definitivo, evitando tratar la experiencia del usuario como un proceso binario. Así mismo, los modelos de eventos recurrentes permitirían estudiar situaciones en las que un mismo usuario alterna períodos de actividad e inactividad a lo largo del tiempo. De esta forma, modelos alternativos a los utilizados en este estudio podrían contribuir a refinar las estimaciones y predicciones, así como a mejorar la comprensión de la variabilidad residual no explicada por las covariables observadas.

También se recomienda que estudios posteriores evalúen el impacto de intervenciones de retención diseñadas a partir de la evidencia aquí presentada (por ejemplo, estrategias focalizadas en usuarios sin licencias, o acciones específicas en las regiones con mayores riesgos de *churn*).

A pesar de estas restricciones, los resultados aquí obtenidos ofrecen una base sólida y metodológicamente rigurosa para comprender y gestionar el abandono de usuarios en plataformas digitales de contratación *per diem* de personal de enfermería. En particular, este trabajo constituye uno de los primeros acercamientos cuantitativos a la problemática de la

retención y el *churn* en este tipo de plataformas digitales, proporcionando elementos de análisis y evidencia empírica sobre el comportamiento de abandono de los usuarios de enfermería, así como insumos útiles para el diseño de políticas y estrategias de retención más informadas.

BIBLIOGRAFÍA

Adigital & Boston Consulting Group. (2020). *La economía digital en España ya representa un 19% del PIB*. https://elpais.com/economia/2020-06-23/la-economia-digital-en-espana-ya-representa-un-19-del-pib.html.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Wiley-Interscience.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Almonteros, J. R., & Matias, J. B. (2024). Integration of Stratified K-Fold Cross Validation to Enhance Prediction Accuracy: A Comparison Study. *2024 5th International Conference on Data Analytics for Business and Industry (ICDABI),* 81–85. https://doi.org/10.1109/ICDABI63787.2024.10800425.

Araujo Delgado, M. (2024). *El impacto de las plataformas digitales en entornos de trabajo*. Universidad Rey Juan Carlos. https://hdl.handle.net/10115/37393.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley.

Bennis, A., Mouysset, S., & Serrurier, M. (2021). PseudoNAM: A pseudo value based interpretable neural additive model for survival analysis. *CEUR Workshop Proceedings*, *3068*, 1–6. Recuperado de https://ceur-ws.org/Vol-3068/short3.pdf.

Berson, A., Smith, S., & Thearling, K. (2000). *Building data mining applications for CRM*. McGraw-Hill.

Blanche, P., Dartigues, J. F., & Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under ROC curves for censored event times with competing risks. *Statistics in Medicine, 32*(30), 5381–5397. https://doi.org/10.1002/sim.5958.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 1–62.

Boote, J. (1998). Towards a comprehensive taxonomy and model of consumer complaining behaviour. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, 11*, 140–151.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer. https://doi.org/10.1007/b97636.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Bütepage, G., Vitor, C., & Carlqvist, P. (2022). Performance of AIC and BIC for the extrapolation of survival data. *Value in Health, 25*(6), 1003–1010.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2025). *shiny: Web application framework for R* (Versión 1.11.1) [Paquete de R]. https://doi.org/10.32614/CRAN.package.shiny.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–220.

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Datta, P., Masand, B., Mani, D. R., & Li, B. (2000). Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review, 14*(6), 485–502.

Efron, B., & Tibshirani, R. J. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association, 92*(438), 548–560. https://doi.org/10.2307/2965703.

Equihua, J. P., Nordmark, H., Ali, M., & Lausen, B. (2023). *Modelling customer churn for the retail industry in a deep learning based sequential framework* (arXiv Preprint No. 2304.00575). arXiv. https://doi.org/10.48550/arXiv.2304.00575.

Gallardo Allen, E., Molina-Delgado, M., & Cordero Cantillo, R. (2016). Aplicación del análisis de sobrevivencia al estudio del tiempo requerido para graduarse en educación superior: el caso de la Universidad de Costa Rica. *Revista Páginas de Educación, 9*(1).

Gerds, T. A., & Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6), 1029–1040. https://doi.org/10.1002/bimj.200610301.

Gerds, T. A. (2025). *pec: Prediction Error Curves for Risk Prediction Models in Survival Analysis* (R package). Recuperado de https://cran.r-project.org/web/packages/pec/pec.pdf.

Gerds, T. A., Ohlendorff, J., & Ozenne, B. (2025). *riskRegression: Risk regression models and prediction scores for survival analysis with competing risks* (R package). Recuperado de https://cran.r-project.org/web/packages/riskRegression/index.html.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81*(3), 515–526. https://doi.org/10.1093/biomet/81.3.515.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software, 40*(3), 1–25. https://www.jstatsoft.org/v40/i03/.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA*, 247(18), 2543–2546.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics, 56*(2), 337–344. https://doi.org/10.1111/j.0006-341x.2000.00337.x.

Hess, K., & Gentleman, R. (2021). *muhaz: Hazard Function Estimation in Survival Analysis* (R package version 1.2.6.4) [Computer software]. CRAN. https://doi.org/10.32614/CRAN.package.muhaz.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: Regression modeling of time-to-event data* (2nd ed.). John Wiley & Sons.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(4), 515–524. https://doi.org/10.1016/j.eswa.2005.09.080.

Jackson, C. (2016). *flexsurv: A platform for parametric survival modeling in R*. *Journal of Statistical Software, 70*(8), 1–33. https://doi.org/10.18637/jss.v070.i08.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R* (2nd ed.). Springer.

Kassambara, A., Kosinski, M., & Biecek, P. (2024). *survminer: Drawing Survival Curves using 'ggplot2'* (R package version 0.5.0). https://CRAN.R-project.org/package=survminer.

Keaveney, S. M., & Parthasarathy, M. (2001). Customer switching behavior in online service: An exploratory study. *Journal of the Academy of Marketing Science, 29*(4), 374–390.

Khattak, A., Mehak, Z., Ahmad, H., Asghar, M. U., Asghar, M. Z., & Khan, A. (2023). *Customer churn prediction using composite deep learning technique*. *Scientific Reports, 13*, Article 17294. https://doi.org/10.1038/s41598-023-44396-w.

Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis: A self-learning text* (2nd ed.). Springer.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.

Kvamme, H., & Borgan, Ø. (2023). The Brier Score under Administrative Censoring: Problems and a Solution. *Journal of Machine Learning Research*, 24(2), 1–26. https://www.jmlr.org/papers/v24/19-1030.html.

Lai, Y., & Zeng, J. (2014). Analysis of customer churn behavior in digital libraries. *Program: Electronic Library and Information Systems, 48*(4), 370–382. https://doi.org/10.1108/PROG-08-2011-0035.

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 27(2), 277–285. https://doi.org/10.1016/j.eswa.2004.02.002.

Liu, X. (2012). *Survival analysis: Models and applications*. Wiley.

Masarifoglu, M., & Hakan Buyuklu, A. (2019). Applying survival analysis to telecom churn data. *American Journal of Theoretical and Applied Statistics, 8*(6), 261–275. https://doi.org/10.11648/j.ajtas.20190806.18.

Mavri, M., & Ioannou, G. (2008). Customer switching behavior in Greek banking services. *Managerial Finance, 34*(3), 186–197.

Monika, A. V., Indahwati, & Aidi, M. N. (2021). Churn Analysis in Telecommunication Industry Customers Using Semiparametric and Non Parametric Survival Method. *Journal of Physics: Conference Series*, 1863(1), 012034. https://doi.org/10.1088/1742-6596/1863/1/012034.

Montgomery, D. C. (2013). *Design and analysis of experiments* (8th ed.). John Wiley & Sons.

Müller, H.-G., & Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1), 61–76. https://doi.org/10.2307/2533197.

Newson, R. (2010). Comparing the predictive power of survival models. *The Stata Journal, 10*(3), 339–358.

OECD. (2020a). *The digitalisation of science, technology and innovation: Key developments and policies*. OECD Publishing. https://doi.org/10.1787/b9e4a2c0-en.

OECD. (2020b). *OECD Digital Economy Outlook 2020*. OECD Publishing. https://doi.org/10.1787/bb167041-en.

O'Quigley, J. (2021). *Survival analysis: Proportional and non-proportional hazards regression*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-33439-0.

Organización Internacional del Trabajo. (2021). *El papel de las plataformas digitales en la transformación del mundo del trabajo*. https://www.ilo.org/es/publications/el-papel-de-las-plataformas-digitales-en-la-transformaci%C3%B3n-del-mundo-del.

Ortiz Robles, I. de J. (2022). *Análisis de Sobrevivencia para Predecir el Retiro de Funcionarios de Ministerios de Gobierno & Análisis de Cohortes para determinar el Impacto Salarial de un Modelo de Salario Único para los Empleados de Ministerios de Gobierno*. Trabajo final de Maestría. Universidad de Costa Rica, Sistema de Estudios de Posgrado.

Pérez Pizarro, C. E. (2022). *Prospección de fuga de clientes de un servicio de suscripción digital* [Tesis de maestría, Universidad del Desarrollo]. Repositorio Institucional Universidad del Desarrollo.

Pölsterl, S. (s. f.). *Evaluating survival models*. En *scikit-survival documentation*. Recuperado el 11 de septiembre de 2025, de https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation. https://www.R-project.org/.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika, 83*(2), 251–266. https://doi.org/10.1093/biomet/83.2.251.

Reichheld, F. F., & Sasser, W. E. J. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 68*(5), 105–111.

Sánchez Brenes, K. (2021). *Duración de la reincidencia y su factores asociados en la población penitenciaria de Costa Rica, perfil cantonal a partir de agrupaciones de cantones mediante la construcción de un índice de seguridad ciudadana en Costa Rica, 2016-2018.* Trabajo final de Maestría. Universidad de Costa Rica, Sistema de Estudios de Posgrado.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Boca Raton, FL: Chapman and Hall/CRC.

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0.

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., & Habbema, J. D. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology, 54*(8), 774–781. https://doi.org/10.1016/s0895-4356(01)00341-9.

Therneau, T. (2024). *A package for survival analysis in R* (Version 3.7-0). https://CRAN.R-project.org/package=survival.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*. Springer.

Therneau, T., Crowson, C., & Atkinson, E. (2024). *Using time dependent covariates and time dependent coefficients in the Cox model*. Mayo Clinic. Recuperado de https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

U.S. Bureau of Labor Statistics. (2024). *Occupational outlook handbook*. https://www.bls.gov/ooh/.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating adequacy of prediction procedures. *Statistics in Medicine, 30*(10), 1105–1117.

Uno, H., Cai, T., Tian, L., & Wei, L. J. (2007). Evaluating prediction rules for t-year survivors. *Journal of the American Statistical Association, 102*(478), 527–537. https://doi.org/10.1198/016214507000000149.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (R package version 1.1.4). CRAN. https://CRAN.R-project.org/package=dplyr.

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. G. M. (2018). Time-varying covariates and coefficients in Cox models. *Annals of Translational Medicine, 6*(7), 121. https://doi.org/10.21037/atm.2018.02.12.

ANEXOS

Anexo A.

Cuadro A.1.

Estados de Estados Unidos presentes en cada una de las Regiones

Región	Estados
Utah & Arizona	Arizona (AZ), Utah (UT)
California & Nevada	California (CA), Nevada (NV)
Midwest	Illinois (IL), Indiana (IN), Iowa (IA), Kansas (KS), Kentucky (KY), Michigan (MI), Minnesota (MN), Missouri (MO), Nebraska (NE), North Dakota (ND), Ohio (OH), South Dakota (SD), Wisconsin (WI)
Northeast	Connecticut (CT), Delaware (DE), District of Columbia (DC), Maine (ME), Maryland (MD), Massachusetts (MA), New Hampshire (NH), New Jersey (NJ), New York (NY), Pennsylvania (PA), Rhode Island (RI), Vermont (VT), Virginia (VA), West Virginia (WV)
Pacific Northwest	Oregon (OR), Washington (WA)
Rockies & Red River	Colorado (CO), Idaho (ID), Montana (MT), New Mexico (NM), Oklahoma (OK), Texas (TX), Wyoming (WY)
Southeast	Alabama (AL), Arkansas (AR), Florida (FL), Georgia (GA), Louisiana (LA), Mississippi (MS), North Carolina (NC), South Carolina (SC), Tennessee (TN)

Nota: La abreviatura de cada estado se presenta entre paréntesis.

Anexo B.

El siguiente Cuadro B.1 presenta las comparaciones pareadas del C-index entre todas las distribuciones paramétricas evaluadas. Para cada par de modelos se calcula la diferencia de medias del C-index (definida como *primera distribución – segunda distribución*), utilizando las 24 estimaciones obtenidas por validación cruzada (6 folds × 4 repeticiones). Se reportan además el estadístico t, el valor-p y el número de pares (n = 24) obtenidos en cada contraste.

La interpretación es directa: diferencias positivas indican que la primera distribución presenta, en promedio, un C-index mayor que la segunda (mejor concordancia), mientras que diferencias negativas indican lo contrario. Los valores p permiten juzgar si la diferencia observada es estadísticamente significativa bajo el esquema de validación utilizado.

Cuadro B.1.

Comparaciones pareadas del C-index entre distribuciones paramétricas mediante prueba *t*

Comparación	Diferencia en medias	Valor t	Valor-p	n pares
Exponencial - Weibull	-0.00036	-3.75	0.0010	24
Exponencial - Log-normal	-0.00024	-2.36	0.0271	24
Exponencial - Log-logística	-0.00024	-2.70	0.0127	24
Exponencial - Gamma	-0.00030	-3.36	0.0027	24
Exponencial - Gompertz	-0.00028	-4.21	0.0003	24
Weibull - Log-normal	0.00012	2.05	0.0515	24
Weibull - Log-logística	0.00012	1.94	0.0647	24
Weibull - Gamma	0.00006	3.93	0.0007	24
Weibull - Gompertz	0.00008	2.45	0.0225	24
Log-normal - Log-logística	0.00000	0.09	0.9300	24
Log-normal - Gamma	-0.00006	-0.83	0.4126	24
Log-normal - Gompertz	-0.00004	-0.65	0.5215	24
Log-logística - Gamma	-0.00006	-0.84	0.4085	24
Log-logística - Gompertz	-0.00004	-0.79	0.4402	24
Gamma - Gompertz	0.00002	0.66	0.5134	24

Por su parte, el siguiente Cuadro B.2 resume las comparaciones pareadas del IBS (Brier Score integrado) entre las mismas distribuciones, también a partir de validación cruzada con 6 folds y 4 repeticiones (24 estimaciones por modelo). La diferencia de medias se define igualmente como primera distribución – segunda distribución. Dado que menor IBS implica mejor desempeño predictivo, una diferencia negativa indica que la primera distribución, en promedio, obtiene menor error (mejor) que la segunda; una diferencia positiva sugiere lo contrario. Se incluyen el estadístico t, el valor-p y n = 24 pares por contraste, lo que permite valorar la significancia estadística de las diferencias en error de predicción entre distribuciones.

Cuadro B.2.

Comparaciones pareadas del IBS entre distribuciones paramétricas mediante prueba *t*

Comparación	Diferencia en medias	Valor t	Valor-p	n pares
Exponencial - Weibull	0.0020	74.1	0.000000	24
Exponencial - Log-normal	-0.0010	-15.2	0.000000	24
Exponencial - Log-logística	-0.0006	-10.8	0.000000	24
Exponencial - Gamma	0.0172	88.7	0.000000	24
Exponencial - Gompertz	0.0164	81.7	0.000000	24
Weibull - Log-normal	-0.0030	-71.7	0.000000	24
Weibull - Log-logística	-0.0026	-70.1	0.000000	24
Weibull - Gamma	0.0153	78.7	0.000000	24
Weibull - Gompertz	0.0144	72.0	0.000000	24
Log-normal - Log-logística	0.0003	39.6	0.000000	24
Log-normal - Gamma	0.0182	87.3	0.000000	24
Log-normal - Gompertz	0.0173	81.2	0.000000	24
Log-logística - Gamma	0.0179	86.3	0.000000	24
Log-logística - Gompertz	0.0170	80.0	0.000000	24
Gamma - Gompertz	-0.0009	-51.1	0.000000	24

Anexo C.

Con el fin de complementar los resultados presentados en el cuerpo principal de este trabajo y facilitar su uso por parte de analistas y personal de la empresa, se desarrolló una aplicación interactiva en *Shiny* (Chang, 2025) que permite calcular y visualizar tiempos de supervivencia t_S y sus intervalos de confianza para distintos perfiles de usuarios y diferentes modelos de supervivencia. Esta herramienta traduce varios resultados de este estudio en un entorno gráfico sencillo, de modo que el usuario pueda explorar distintos escenarios (perfiles de usuario) sin necesidad de trabajar directamente con código en R ni realizar cálculos manuales complejos.

Shiny (Chang, 2025) es un paquete de R (R Core Team, 2023) que permite construir aplicaciones web interactivas a partir de código en este lenguaje. En una aplicación *Shiny* el usuario interactúa con controles (por ejemplo, menús desplegables, casillas de selección, campos numéricos) y, cada vez que modifica una entrada, el servidor de R recalcula de forma reactiva los resultados y actualiza las salidas correspondientes (gráficos, tablas, resúmenes numéricos, y más) (Sievert, 2020). Estas aplicaciones pueden ejecutarse localmente o en un servidor y son especialmente útiles para comunicar modelos estadísticos de manera intuitiva a usuarios no especializados en programación.

La aplicación desarrollada, ilustrada en este anexo, implementa la metodología descrita en los capítulos anteriores. En primer lugar, se cargan los paquetes necesarios y se leen desde disco los modelos previamente ajustados: el modelo paramétrico Gamma, el modelo Log-normal y el modelo de Cox-extendido, también se carga la base de datos utilizada para reconstruir la función de riesgo base de este último. Sumado a esto, se importa un archivo auxiliar con funciones específicas (por ejemplo, para limpiar nombres de variables y calcular t_S e intervalos de confianza mediante remuestreo paramétrico para diferentes probabilidades).

En términos generales, el remuestreo que se utiliza para calcular los intervalos de confianza de los tres modelos se basa en aproximar la distribución muestral de los parámetros y, a partir de ella, la distribución de los tiempos t_S. El procedimiento puede describirse, de forma simplificada, en cuatro pasos: (i) se parte de los estimadores puntuales de los coeficientes del

modelo y de su matriz de varianzas y covarianzas; (ii) se generan muchas réplicas plausibles de esos parámetros, simulándolos como si fueran nuevas estimaciones obtenidas de muestras similares a la observada (remuestreo o *bootstrap* paramétrico); (iii) para cada conjunto simulado de parámetros se calculan los tiempos t_S correspondientes al perfil de usuario definido; y (iv) se forma un intervalo de confianza tomando percentiles adecuados de la distribución empírica de todos los tiempos simulados (por ejemplo, el percentil 2.5 y el 97.5 cuando se usa un nivel de confianza del 95%). En los modelos Gamma y de Cox-extendido este esquema de remuestreo se implementa explícitamente; en el modelo Log-normal se utiliza la aproximación normal asintótica de los cuantiles, que puede interpretarse como un caso particular en el que la variabilidad de t_S se obtiene a partir de la distribución aproximada de sus errores de estimación.

Luego, se definen constantes (nivel de confianza fijo en 95 % y número de réplicas de *bootstrap* paramétrico fijado en 6000) y vectores de ponderaciones para representar perfiles promedio por región y por tipo de licencia cuando el usuario elige la opción "*All regions*" o "*All licenses*". Luego se incluyen funciones de utilidad que transforman las entradas del formulario (región, licencia, edad e indicador de ingresos en los primeros dos meses) en un perfil de covariables compatible con la estructura de los modelos y que construyen los *data frames* a utilizar en cada uno de ellos. Sobre esta base se implementan las funciones que calculan los tiempos t_S y sus intervalos de confianza a partir de los parámetros, coeficientes y matrices de varianzas y covarianzas de los modelos.

En las figuras que se muestran más adelante en esta sección (Figuras C1, C2, C3, C4, C5, C6, y C7), se incluyen capturas de pantalla representativas de la aplicación *Shiny* descrita, las cuales ilustran ejemplos de uso con uno (Figuras C2 y C3), dos (Figuras C4 y C5) y tres perfiles de usuario (Figuras C6 y C7), y permiten apreciar el tipo de salidas gráficas que genera la herramienta para cada una de las selecciones de perfiles realizadas en el panel lateral de la aplicación. Mientras que en la Figura C1, se muestra de manera global toda la interfaz de usuario de la aplicación.

Como se muestra en la Figura C1, la interfaz de usuario (UI) se organiza mediante un panel lateral (también mostrado en las Figuras C2, C4, y C6) y un panel principal que presenta los

gráficos (también mostrado en las Figuras C3, C5, y C7). En el panel lateral el usuario puede: (i) seleccionar los modelos con los que desea trabajar (Gamma, Log-normal o Cox-extendido); (ii) introducir entre 1 y 5 valores de la probabilidad de supervivencia S a calcular; y (iii) especificar entre uno y tres perfiles de usuario, eligiendo para cada perfil la región, el tipo de licencia, la edad y si generó ingresos en los primeros dos meses. El botón "*Compute*" dispara el cálculo y la opción "*Download CSV*" permite descargar en formato tabular los tiempos t_S y sus intervalos de confianza para cada probabilidad de supervivencia seleccionada.

En la parte del servidor, la aplicación valida los valores de S, construye la lista de perfiles solicitados y, para cada combinación perfil–modelo, calcula los tiempos de supervivencia y sus intervalos de confianza para cada probabilidad seleccionada mediante las funciones antes mencionadas. Todos los resultados se combinan en un único *data frame* que alimenta un gráfico generado con *ggplot2* (Wickham, 2016; Kassambara et al., 2024) y convertido en objeto interactivo mediante *plotly* (Sievert, 2020). Este gráfico presenta, para cada perfil y cada valor de S, un punto con barras de error que representan el intervalo de confianza del 95%, diferenciando los modelos por color y forma del marcador. Asimismo, se habilitan mensajes emergentes (*tooltips*) con los valores numéricos detallados y se ofrece la posibilidad de hacer *zoom* o desplazar el gráfico. Esta interfaz y gráficos generados por la aplicación *Shiny* se ilustran con las Figuras C1, C3, C5, y C7.

Figura C1.

Interfaz completa (UI) de la aplicación *Shiny* para seleccionar las probabilidades de supervivencia, y el perfil de usuario a graficar

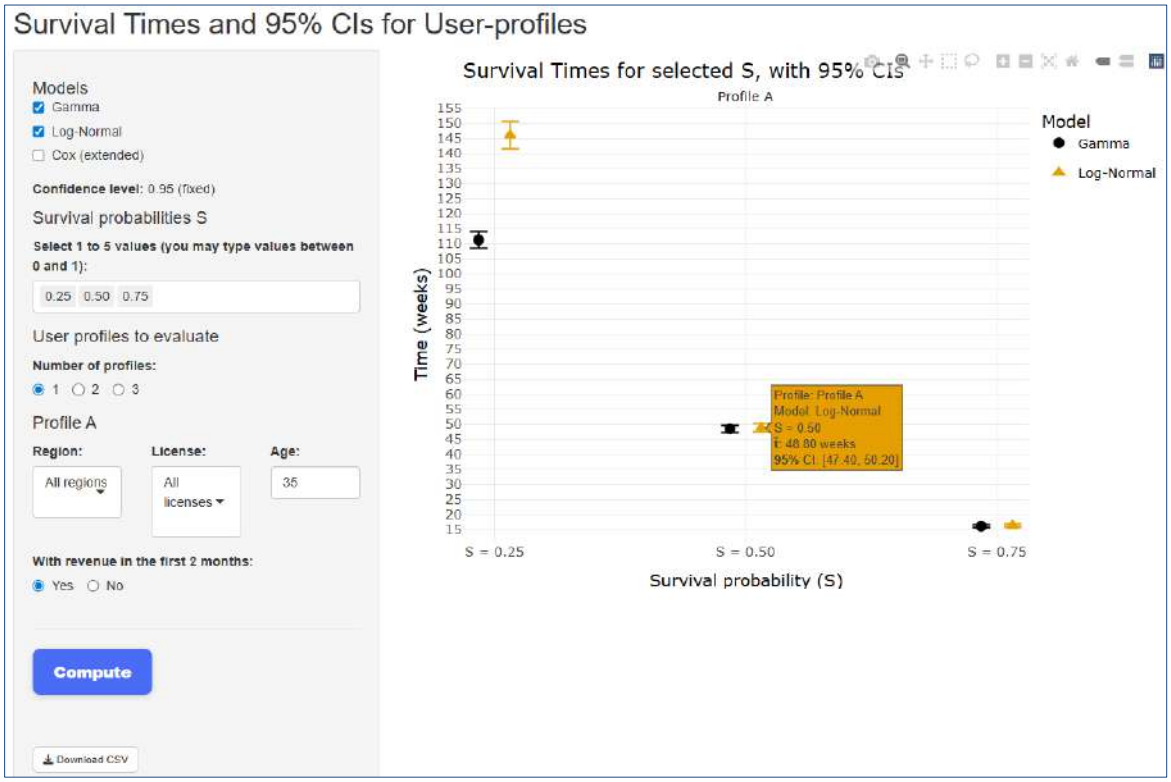


Figura C2.

Panel lateral de la Interfaz de la aplicación *Shiny* para seleccionar el perfil de usuario a graficar: seleccionando un único perfil y dos modelos

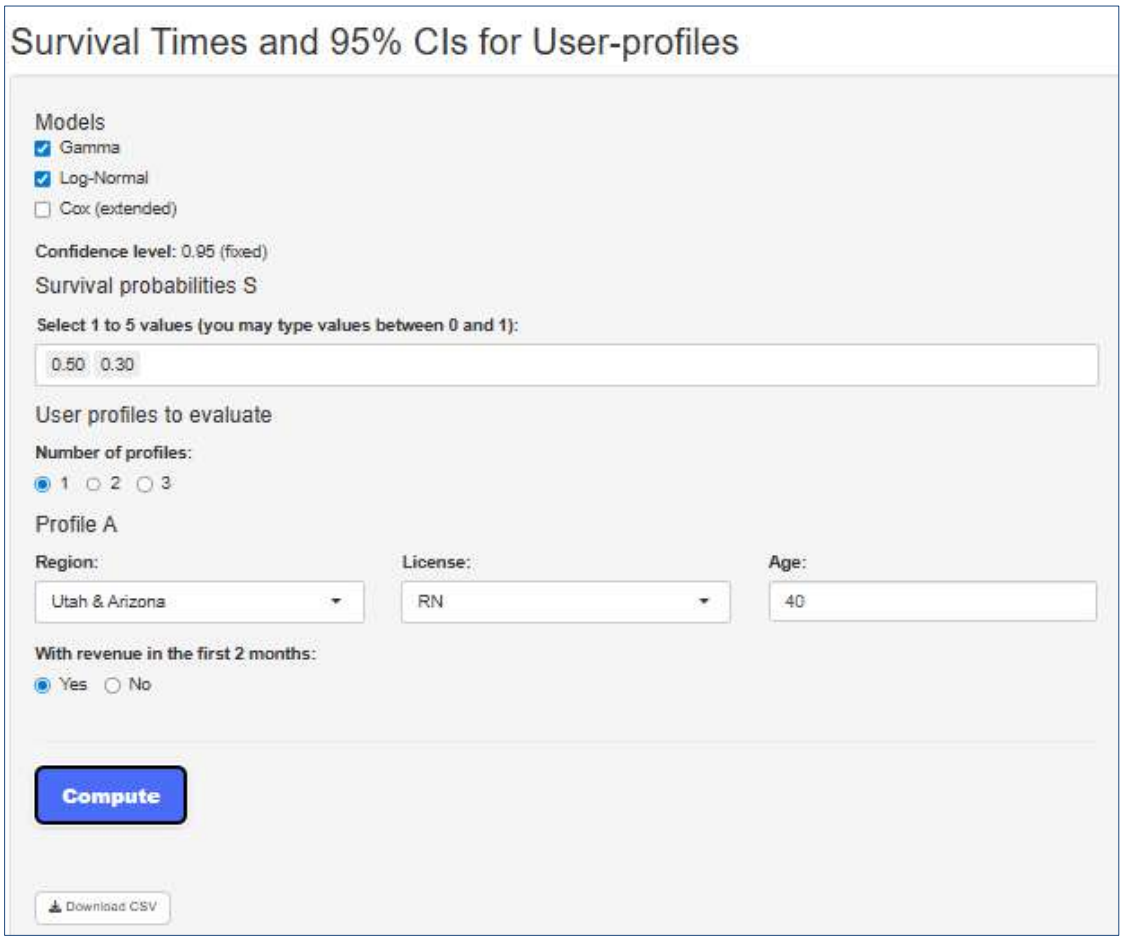
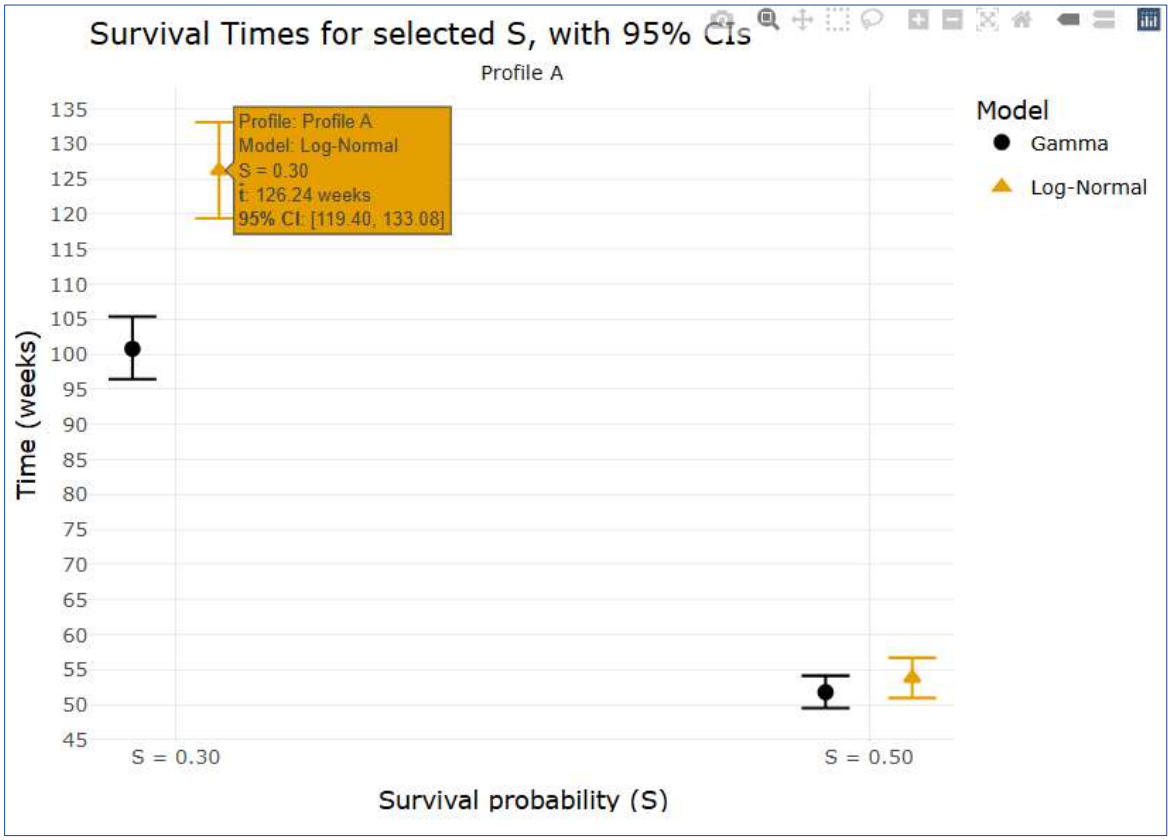


Figura C3.

Panel de los Gráficos de la Interfaz de la aplicación *Shiny*: seleccionando un único perfil de usuario y dos modelos a graficar



Nota: El perfil y modelos seleccionados son los que se muestran en la Figura C2.

Figura C4.

Panel lateral de la Interfaz de la aplicación *Shiny* para seleccionar el perfil de usuario a graficar: seleccionando dos perfiles y tres modelos

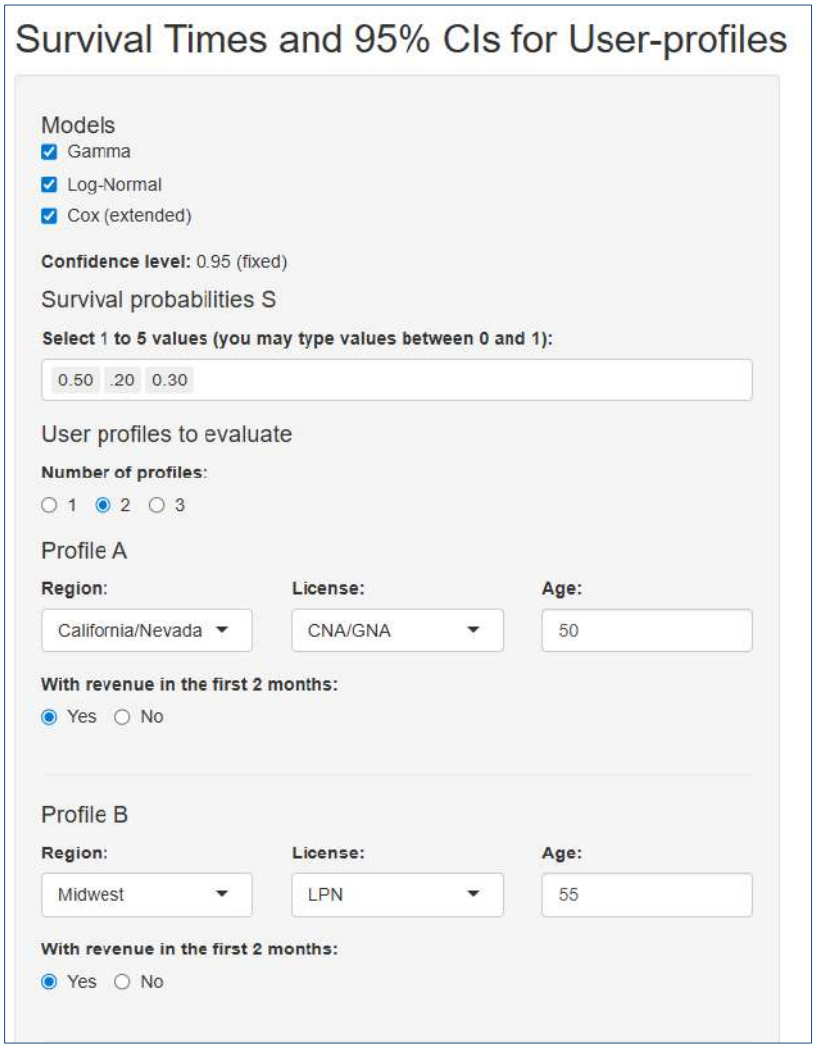
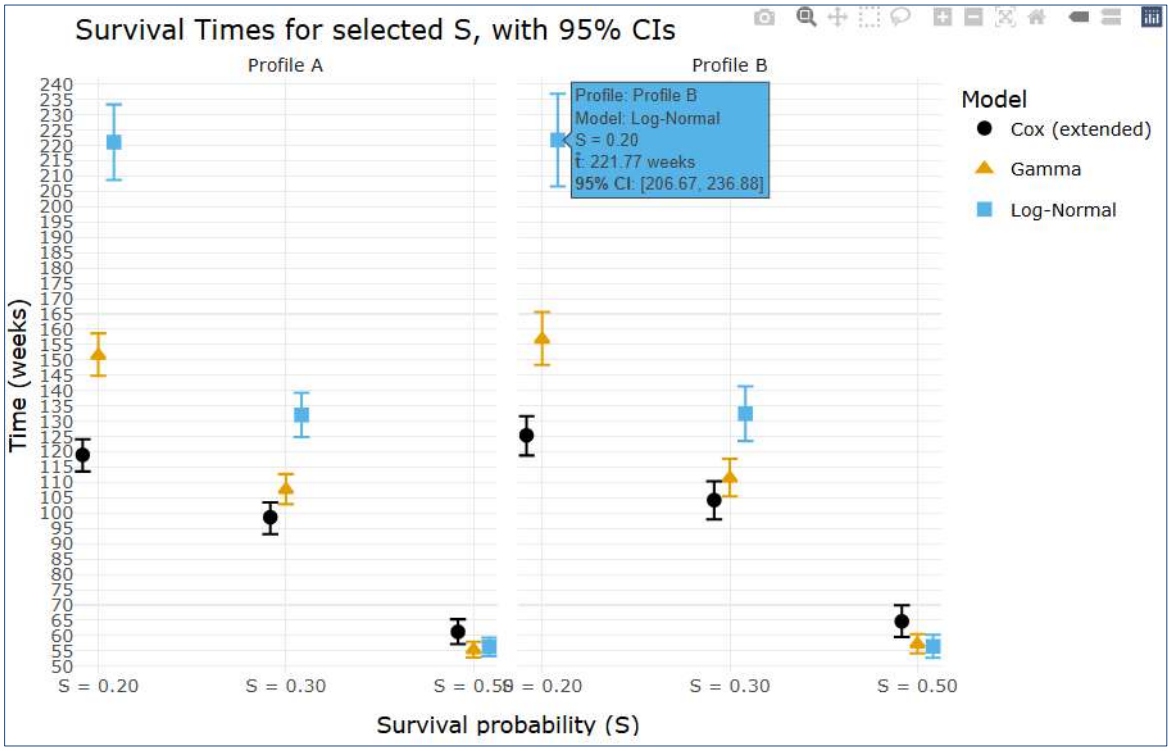


Figura C5.

Panel de los Gráficos de la Interfaz de la aplicación *Shiny*: seleccionando dos perfiles de usuarios y tres modelos a graficar



Nota: El perfil y modelos seleccionados son los que se muestran en la Figura C4.

Figura C6.

Panel lateral de la Interfaz de la aplicación *Shiny* para seleccionar el perfil de usuario a graficar: seleccionando tres perfiles de usuarios y dos modelos

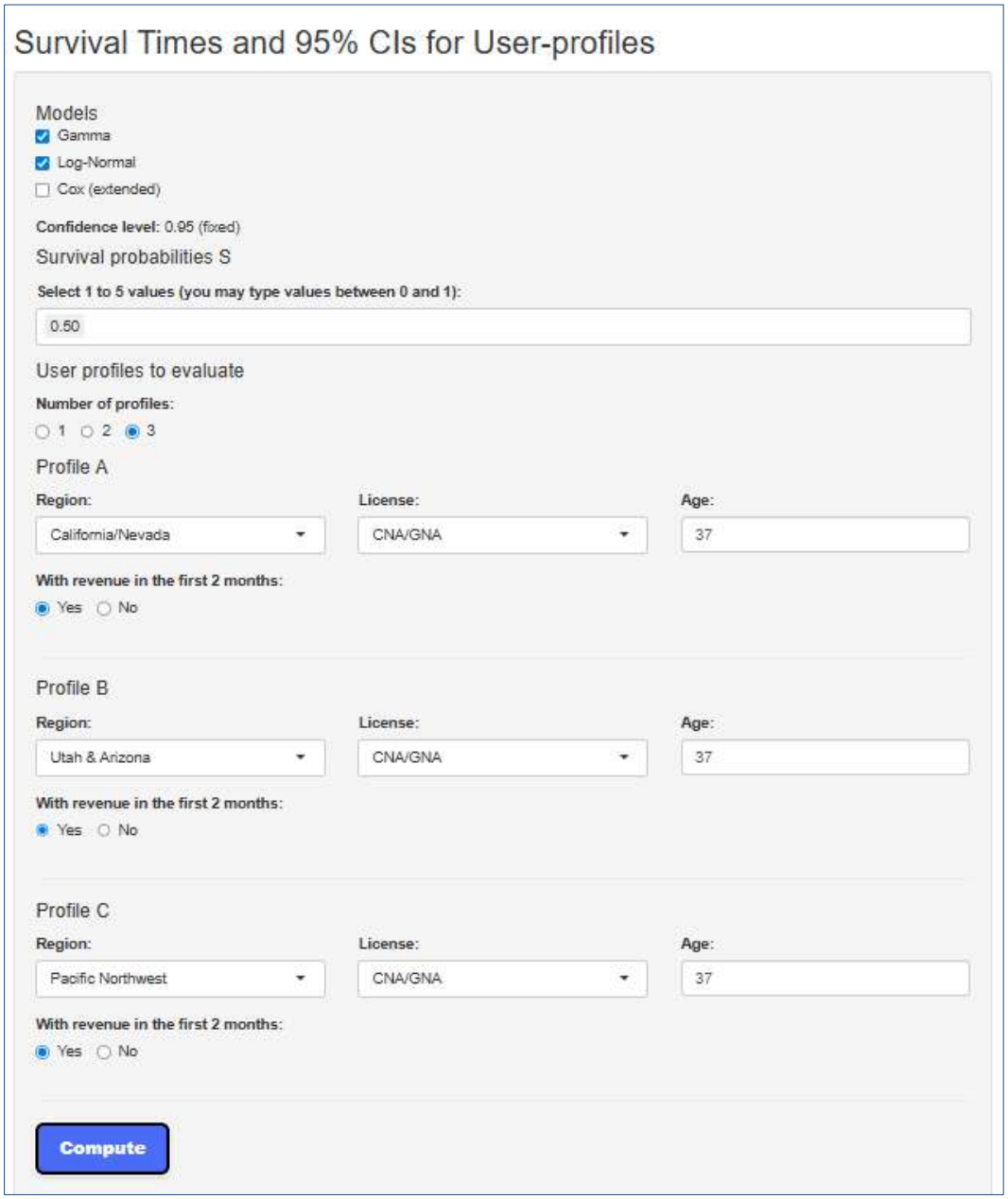
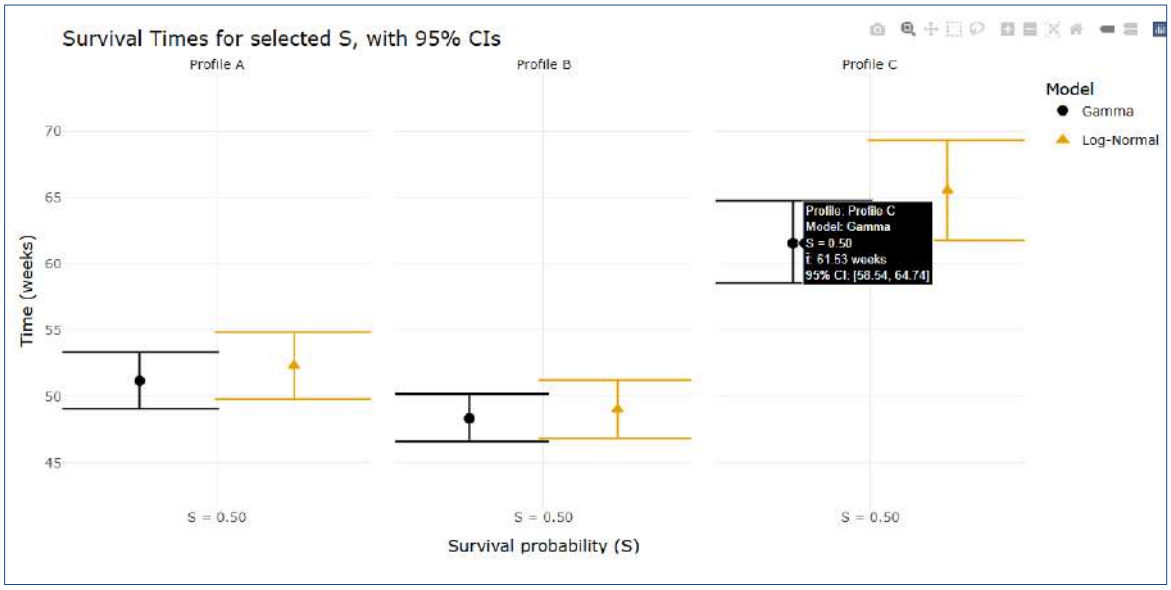


Figura C7.

Panel de los Gráficos de la Interfaz de la aplicación *Shiny*: seleccionando tres perfiles de usuarios y dos modelos a graficar



Nota: El perfil y modelos seleccionados son los que se muestran en la Figura C6.