Volumen 2
Número 1
Diciembre 2020

SERENGUETI

Revista de Estadística

"Una nueva propuesta para el aprendizaje"

Foto: Ricardo Alvarado B.

Universidad de Costa Rica, Escuela de Estadística
Ericka Valerio, Susana García, Andrea Vargas, Joshua Salazar, Maripaz Venegas, Nelson Torres

COMITÉ EDITORIAL

Estudiantes y egresados de bachillerato de la carrera de Estadística de la Universidad de Costa Rica, integran el comité editorial y participan en la elaboración de este volumen de la revista.

Este comité cumple la función de invocar a estudiantes activos que deseen participar en la divulgación de sus artículos, previamente recomendados por profesores de la carrera. Además, verifica que estos artículos cumplan con los lineamientos establecidos para su inclusión en la revista. Asimismo, el comité tiene entre sus funciones difundir y promocionar la revista en plataformas o medios de interés para la Escuela de Estadística y sus estudiantes.

Ericka Valerio Salas, Universidad de Costa Rica, c.e. *erickavs189@gmail.com*

Susana García Calvo, Universidad de Costa Rica, c.e. *asgc3093@gmail.com*

Andrea Vargas Montero, Universidad de Costa Rica, c.e. *avargas2398@gmail.com*

Joshua Salazar Obando, Universidad de Costa Rica, c.e. *joshua.salazar1692@gmail.com*

Maripaz Venegas González, Universidad de Costa Rica, c.e. *maripaz5199@hotmail.com*

Nelson Torres Chávez, Universidad de Costa Rica, c.e. *nelson0823@hotmail.com*

DEDICATORIA

Se dedica este número de la revista al profesor Ricardo Alvarado Barrantes, coordinador y evaluador principal de la selección de artículos. Muchas gracias por el apoyo y la motivación brindada durante todo el proceso y por gran su disposición para que este proyecto siga adelante.

CONTENIDO

INTRODUCCIÓN

En los últimos años, el estudiantado del Bachillerato en Estadística de nuestra escuela ha realizado trabajos con múltiples técnicas estadísticas, que se aplican a diversas áreas del conocimiento. Darnos cuenta de la riqueza que envuelve la Estadística como disciplina y como herramienta para la creación de conocimiento, es fundamental. A la vez se convierte en un reto, pues ahora más que nunca, nuestros estudiantes se ven enfrentados a técnicas más complejas que los coloca ante una situación de mayor dificultad para poder comprender las características de un estudio, y de esta forma poder aplicar un enfoque adecuado cuando se trata de cumplir con los objetivos del mismo.

Este nuevo número de la Revista Serengueti pone de manifiesto el papel tan importante de las aplicaciones de la Estadística en la formación de nuestro estudiantado. Se presenta un estudio mediante un diseño que se puede analizar mediante dos enfoques distintos. Se presenta un tema de actualidad como es el caso de la pandemia de Covid-19 en Costa Rica y el uso de un modelo lineal generalizado. En técnicas multivariadas se logra aplicar el agrupamiento a situaciones bastante innovadoras, tales como el análisis de sentimientos en tweets, la conformación de perfiles musicales a partir de Spotify y el uso de medidas antropométricas para caracterizar a los triatletas. También se usan técnicas de clasificación en casos más clásicos como créditos y casos de cáncer. Adicionalmente a los casos aplicados se presenta un estudio de simulación que permite partir de los conocimientos teóricos para determinar la importancia de usar una técnica como mínimos cuadrados ponderados.

I. DISEÑOS EXPERIMENTALES FACTORIALES

Un diseño experimental incluye todos los elementos que se van a considerar dentro de la ejecución y análisis de un experimento, de forma específica, un experimento factorial tiene un diseño que consta de dos o más factores, donde cada uno tiene distintos valores o niveles y se tienen tratamientos que se componen de la combinación entre los niveles de los distintos factores. Estos diseños fueron utilizados por primera vez en el siglo XIX por Henry J. Gilbert y John Bennet Lawes de la Estación Experimental de Rothamsted (Montes, 2018)[1].

[1] Montes, C. (2018). Análisis de diseños factoriales. Tecnológico Nacional de México-Instituto Tecnológico de Tapachula. Recuperado de academia.edu/37969293/Punto_No. _1. Qué_es_un_diseño?auto=download

Capacidad de explosión de granos de maíz: un análisis logístico y lineal

Emir Rojas Araya[2], Marielle Rodríguez Rodríguez[2], Fernando Céspedes Zamora[2]

emir.rojas@ucr.ac.cr, marielle.rodriguez@ucr.ac.cr, fernando.cespedeszamora@ucr.ac.cr

RESUMEN

Existen diferentes aspectos que influyen en que un grano de maíz explote. De la misma forma, hay distintas maneras de modelar el comportamiento de los granos ante diversos factores. El presente trabajo analiza la incidencia del tiempo de preparación, humedad y número de granos en los paquetes de palomitas de maíz. Para analizar este efecto se utiliza tanto un modelo logístico, considerando la respuesta como una distribución binomial; como un modelo lineal, interpretando la respuesta como una variable continua. Estos dos modelos se comparan mediante el uso de la estimación puntual y por intervalo. Se encuentra que todas las variables son relevantes para que una palomita explote, empero, el tiempo de preparación incide en mayor magnitud.

PALABRAS CLAVE: maíz, palomitas, modelo logístico, modelo lineal, Bartlett, Análisis de Varianza, ANOVA, razón de verosimilitud.

INTRODUCCIÓN

Entre algunos de los bocadillos más frecuentes alrededor del mundo se encuentran las palomitas de maíz, según Karababa (2004). La producción de palomitas no sólo es por sí misma una gran industria, sino que un importante porcentaje de los ingresos de los cines y lugares de entretenimiento proviene de su venta. Desde su descubrimiento se han implementado muchos métodos de cocción hasta llegar a los hornos microondas: probablemente, la técnica más popular en la actualidad.

De acuerdo con Gökmen (2004), aunque el uso de estos aparatos se ha vuelto uno de los más populares, los resultados que arroja no son los más óptimos en comparación con otros métodos convencionales. En su artículo comprobó que, aunque los hornos microondas producen el mayor tamaño de grano reventado, también producen el mayor porcentaje de granos sin explotar. Es por esto que aflora la interrogante de cómo ciertos factores pueden afectar la producción promedio, es decir, que el grano reviente y no se quede como grano, en hornos microondas.

Un alto porcentaje de granos sin reventar puede deberse al hecho de que una vez que varios se abren, estos disminuyen la intensidad de las ondas electromagnéticas para llegar a las semillas que aún no han explotado en el fondo y se alcanza a un punto en el que la temperatura sobrepasa un nivel crítico donde estas ya no explotan (Gökmen, 2004). Por esta razón es de importancia estudiar la cantidad de granos que se utilizan en la preparación de las palomitas de maíz.

Por otra parte, Villanueva Flores (2008) explica las características que permiten a un grano de maíz reventar. Menciona que estos se abren cuando la presión interna es mayor que la suma de la presión de explosión del grano y la atmosférica. En este caso debe distinguirse que la presión del grano al explotar está determinada por la fuerza necesaria para romper el pericarpio, o en términos comunes: la cáscara. Tomando esto en cuenta, concluye que el proceso promedio de reventar los granos de maíz depende del tiempo en el que

[2] Estudiantes de Estadística de la Universidad de Costa Rica

son expuestos a la inducción y un proceso aleatorio que depende de la presión crítica, temperatura y la actividad de agua interna.

Más específicamente, Hoseney y Delcour (2010) recalcan que el hecho de que las palomitas de maíz exploten, a diferencia de otros granos, se debe a su pericarpio, el cual actúa como un recipiente de alta presión y permite que el agua dentro del grano sea calentada a altas temperaturas. Debe notarse que otros granos, e incluso otros tipos de maíz no explotan, ya que su pericarpio es más poroso, entonces no deja que la presión se acumule.

De manera similar, Lusas y Rooney (2001) formulan que para que un grano de maíz explote tiene que haber un balance respecto a la velocidad que se aplica la temperatura, ya que, si el centro se calienta muy rápido, el pericarpio se va a romper antes de que el almidón del centro se suavice. En contraparte, si este se calienta muy lento entonces se va a perder presión lentamente por la punta donde el grano estaba pegado a la mazorca, por lo que nunca se va a alcanzar la presión crítica. No obstante, debe considerarse que esto no especifica el método de cocción, por lo que es probable que los resultados varíen entre técnicas. Aun así, esta es una consideración para el uso de bloques, debido a las diferencias de potencia entre los distintos microondas.

Por otra parte, Metzger, Ziegler y Bern (1989) analizaron el efecto de la humedad para rangos que variaron del 6% al 20%. Se encontró que el mayor nivel de explosión de granos sucede alrededor de 13% y 14.5%. Similarmente, Gökmen (2004) encontró que un 14% de humedad produjo los mayores tamaños de granos, volumen de explosión y menor porcentaje de granos sin reventar. Estos resultados reflejan que, a muy bajos o muy altos niveles de humedad, los granos no tienen un mismo comportamiento. Estos resultados muestran que la variable tiene un efecto considerable en el producto final de las palomitas.

Tomando en cuenta los anteriores artículos es posible percibir que la cantidad de granos, el tiempo de cocción y el nivel de humedad conforman factores determinantes en la cantidad de granos que explotan. Esto, aunado a la popularidad de la cocción en hornos microondas de dichos granos y los problemas que suele presentar, refleja la relevancia de estudiar su preparación a distintos niveles de los factores que afectan su expansión.

Se plantea como objetivo general hallar una combinación de factores que maximice la probabilidad de que un grano de maíz explote. Además, se establece como objetivo específico determinar si existe una relación entre la cantidad de granos de maíz utilizados, nivel de humedad, tiempo en horno microondas y la probabilidad de que un grano reviente.

METODOLOGÍA

Como unidad experimental se utilizan grupos de granos. La unidad de observación es el grano de maíz, de esta manera se observa si el grano explota o no. Se registra el número de granos no explotados (llámese éxito), granos explotados (fracaso) dado el tamaño de cada paquete, así como la proporción de éxitos sobre el tamaño del paquete.

Los factores a considerar son la humedad del grano, tiempo de cocción en microondas y la cantidad de granos que se colocan en cada corrida del experimento. La humedad es un factor de diseño con niveles: baja (colocando granos en un horno a 100ºC por 20 minutos y luego enfriando en un recipiente sellado), normal (sin tratamiento) y alta (dejando los granos en un recipiente con agua sellado por un día, sin que estos toquen el

agua). El segundo factor, tiempo de cocción en microondas, si bien es una variable continua se utilizan tres niveles: uno, dos y tres minutos. El tercer factor es la cantidad de granos colocados en cada grupo, utilizando grupos de 100 y 200 granos de maíz. Se tiene así un experimento con 18 tratamientos.

Se selecciona la marca Poppi de Jack's de la cual se obtienen los granos y mediante un procedimiento aleatorio se asigna a cada unidad experimental la cantidad de granos, luego el nivel de tiempo y por último el nivel de humedad. Los grupos de granos se colocan en bolsas de papel selladas, esto con la intención de que los granos sigan agrupados, ya que a la hora de reventar estos se mueven y pueden afectar el proceso de los otros que no hayan reventado. Posteriormente, la cocción de las palomitas se realiza en tres hornos distintos, cuyos modelos son GE JES1153SFE, Telster THE070310MD, Whirlpool WM1107D00.

Se trabaja con dos modelos para el análisis de los datos: uno logístico y uno lineal. Para el modelo logístico la variable respuesta es la cantidad de granos que no explotan dado el tamaño del paquete. Mientras que para el modelo lineal la variable respuesta es la proporción de los granos que no explotan entre el total de granos por paquete.

El modelo logístico es entonces:

$$log(\frac{\pi_{(ijk)}}{1-\pi_{(ijk)}}) = \beta_0 + \tau_i + \alpha_j + \gamma_k + \delta_l + (\tau\alpha)_{ij} + (\tau\gamma)_{ik} + (\alpha\gamma)_{jk} + (\alpha\tau\gamma)_{ijk}$$

Mientras que el modelo lineal:

$$E(y|\tau_i, \ldots, \delta_l) = \beta_0 + \tau_i + \alpha_j + \gamma_k + \delta_l + (\tau\alpha)_{ij} + (\tau\gamma)_{ik} + (\alpha\gamma)_{jk} + (\alpha\tau\gamma)_{ijk}$$

; donde:

π_{ijk}:probabilidad de que la palomita explote

β_0: media general

τ_i : efecto de humedad

α_j : efecto de tiempo

γ_k: efecto de cantidad de granos

δ_l: efecto del bloque

$(\tau\alpha)_{ij}$: efecto de interacción entre humedad y tiempo

$(\tau\gamma)_{ik}$: efecto de interacción entre humedad y cantidad de granos

$(\alpha\gamma)_{jk}$: efecto de interacción entre tiempo y cantidad de granos

$(\alpha\tau\gamma)_{ijk}$: efecto de interacción entre tiempo, humedad y cantidad de granos

Para el primer modelo, debido a que es logístico, no se tiene que evaluar supuestos, ya que se sabe con antelación que se van a incumplir teóricamente. Mientras que para el segundo modelo estos sí deben ser tomados en cuenta. Se evalúan los supuestos de homocedasticidad y normalidad, mediante el uso de las pruebas Bartlett y un gráfico cuantil-cuantil. Se realizan pruebas de Razón de Verosimilitud para el modelo logístico y

prueba de diferencia de medias a través de un ANOVA para el modelo lineal con el fin de analizar el aporte de las interacciones y variables.

El anterior procedimiento se realiza mediante el lenguaje de programación R (R Core Team, 2020) utilizando la versión 4.0.0 y los paquetes: *car: An {R} Companion to Applied Regression* (Fox y Weisberg, 2019), *ggplot2: Elegant Graphics for Data Analysis* (Wickham, 2016) y *lmtest: Diagnostic Checking in Regression Relationships* (Zeileis. y Hothorn, 2002).

Se define por convención que una diferencia en la razón de propensión de 0.5 es relevante, mientras que para las diferencias de promedios el delta se deja a conveniencia del lector. Esto no es antojadizo, ya que hay aspectos particulares que pueden hacer que este varíe, los cuales se salen del enfoque de este trabajo.

RESULTADOS

Primeramente, se analiza el modelo logístico. Al llevar a cabo la prueba de la Razón de Verosimilitud, revela que el modelo obtenido es el más parsimonioso, pues si se quitara alguna interacción o variable, su capacidad explicativa empeoraría de manera importante, por lo que se mantiene el modelo saturado.

Debido a la diferencia de potencias entre los hornos microondas se esperaba obtener diferencias entre los bloques y, efectivamente, se observa que los resultados varían entre personas (ver Figura 1, 2 y 3 en Anexos) por lo cual se justifica el uso de bloques en el modelo. Se procede entonces a analizar la razón de las propensiones (OR por sus siglas en inglés) y a realizar comparaciones entre medias.

Dado que el tiempo de cocción de los granos de maíz es el factor principal del experimento, se comparan los tres niveles de tiempo para cada nivel de humedad, manteniendo el número de granos en 100. Se fija esta variable para tener un número de contrastes más accesible, debido a la gran cantidad de tratamientos presentes. Debe notarse que según la Figura 3 (ver Anexos), parece haber un efecto donde a mayor cantidad de granos, mayor es la proporción.

Al realizar la prueba de hipótesis para estos contrastes se obtiene que no hay evidencia estadística para probar que la proporción de palomitas se mantenga igual con diferentes tiempos y humedad, por lo cual todos los contrastes resultan significativos. Se realizan entonces los respectivos intervalos de confianza y como se observa en la Tabla 1, todas las diferencias resultan relevantes pues son mayores a 1.5.

Tabla 1

Intervalos de confianza para contrastes significativos bajo el modelo logístico, donde h indica el nivel de humedad (alto (1), bajo(2), normal(3)) y t indica los minutos de cocción.

Modelo Logístico			
	L.Inf	OR	L.Sup
h1.t1t2	2.7404	4.6977	8.0529
h1.t1t3	34.7515	75.4691	163.894
h1.t2t3	7.7926	16.0651	33.1199
h2.t1t2	11.6505	21.8643	41.0324
h2.t1t3	24.2255	47.8424	94.4828
h2.t2t3	1.2393	2.1881	3.8635
h3.t1t2	5.0382	8.6804	14.9557
h3.t1t3	15.0887	27.9579	51.8031
h3.t2t3	1.8414	3.2208	5.6334

En la tabla 1 se aprecia que para los tres niveles de humedad la mayor razón de propensiones la genera la comparación entre un minuto y tres minutos. Por ejemplo, se observa que, para humedad alta, la propensión de no reventar el grano al cocinar las palomitas por un minuto es 75 veces mayor que al prepararlas por tres minutos (esta diferencia puede llegar a ser de hasta 163 veces), mientras que la propensión de no reventar por dos minutos es solamente 16 veces mayor que al elaborarlas por tres minutos, lo cual sigue siendo bastante alto. Esto refleja la relevancia del factor tiempo en la cocción de las palomitas y sugiere que se obtienen mejores resultados al calentarlas por tres minutos, aunque la diferencia entre dos y tres minutos no sea tan grande. También resulta interesante observar que para el nivel de humedad normal estas diferencias entre tiempos de cocción no son tan altas como sí lo son para humedad baja y humedad alta.

A continuación, se analiza el modelo lineal descrito anteriormente. Primeramente, se verifican los supuestos necesarios para el modelo lineal:

1) **Normalidad:** se realiza un gráfico de residuales (ver Figura 4 en anexo) que sugiere el cumplimiento del supuesto ya que la mayoría de puntos están cerca de la línea teórica; para formalizarlo se realiza la prueba Shapiro-Wilk la cual da un valor p de 0.10 con lo que se comprueba el supuesto bajo un nivel de significancia de 0.05.

2) **Homocedasticidad:** Dado que se cumple el supuesto de normalidad, se puede analizar la homocedasticidad sin mayores preocupaciones, por lo que se realiza la prueba Bartlett para igualdad de varianzas y se observa evidencia estadística de que hay heterocedasticidad, por lo que se recurre a Mínimos Cuadrados Ponderados.

3) **Linealidad:** Se comprueba con base en las Figuras 1,2 y 3.

Con el modelo resultante, después de realizar Mínimos Cuadrados Ponderados, se analizan las interacciones entre tiempo, humedad y número de granos. Tras realizar una prueba de igualdad de medias se obtiene que hay suficiente evidencia estadística para concluir que las medias son distintas, es decir, todas las interacciones son significativas, por lo que en este caso también se trabaja con el modelo saturado.

Se procede a realizar las comparaciones entre los niveles de tiempo para cada nivel de humedad, como se realizó con el modelo logístico. Bajo este modelo no se encuentra evidencia estadística para determinar que

haya diferencia entre las proporciones de dos minutos y tres minutos, tanto con humedad baja como con humedad normal. Sin embargo, las demás comparaciones resultan significativas. Es notable que esta es una diferencia respecto al modelo logístico, donde se realizan las mismas comparaciones y todas resultan significativas. Lo anterior ejemplifica ganancias marginales decrecientes conforme aumenta el tiempo, lo cual se debe a que suele haber palomitas que independientemente del tiempo no van a explotar por defectos en el grano. Debe señalarse, además, que un aspecto no contemplado en este experimento es que, en algunos hornos, gran número de granos se quemaron completamente. En la tabla 2 se puede observar los contrastes, así como sus respectivos intervalos de confianza.

Tabla 2

Intervalos de confianza para contrastes significativos bajo el modelo lineal, donde h indica el nivel de humedad (alto(1), bajo(2), normal(3)) y t indica los minutos de cocción.

	Modelo Lineal		
	L.Inf	Contraste	L.Sup
h1.t1t2	0.0181	0.3086	0.5991
h1.t1t3	0.4664	0.7598	1.0533
h1.t2t3	0.1558	0.4513	0.7467
h2.t1t2	0.3082	0.5997	0.8913
h2.t1t3	0.4376	0.73	1.0225
h2.t2t3			
h3.t1t2	0.1645	0.456	0.7475
h3.t1t3	0.3742	0.6671	0.96
h3.t2t3			

Como se mencionó, para el modelo logístico todas las diferencias son significativas y relevantes. No obstante, para el caso del modelo lineal se puede apreciar en la Tabla 2 que dos diferencias no son significativas. Asimismo, el primer contraste (entre uno y dos minutos para humedad baja) tiene un límite inferior de 0.0181, es decir, por poco resulta no significativo. Además, de estas tres comparaciones, los resultados del modelo lineal parecen confirmar los resultados del modelo logístico: las diferencias en la proporción son bastante grandes incluso para los límites inferiores, pues se está analizando la diferencia entre las proporciones.

Si se observa la tabla 2, puede notarse que en los intervalos de confianza para las diferencias hay dos valores mayores a 1, las cuales no son teóricamente posibles por tener una variable binomial. Sin embargo, no son motivo de alarma: se deben a la amplitud del intervalo de confianza e intentar aproximar una variable binomial asumiendo linealidad. Si bien dichos valores no son posibles, no son tan preocupantes como que el límite inferior incluya un negativo, pues indicarían que puede que no haya realmente una diferencia.

CONCLUSIONES

Con ambos modelos se encuentra que en los niveles seleccionados todas las variables logran explicar una parte importante, en otras palabras, ninguna resulta prescindible por lo que un modelo saturado es lo ideal. Se comprueba que existe una relación entre la cantidad de granos de maíz utilizados, el nivel de humedad, el tiempo en horno microondas y la probabilidad de que el grano reviente.

Queda en evidencia lo propuesto por Villanueva Flores (2008), ya que el factor que parece tener mayor influencia sobre la variable respuesta es el tiempo que pasan las palomitas en el horno microondas. Además, se observa que el tiempo tiene mayor incidencia sobre los granos que no explotan en humedades bajas y altas, pero no tanto cuando la humedad es normal. De esta manera, se podría sugerir que la combinación ideal entre los factores es con un tiempo de dos a tres minutos bajo una humedad normal.

Los dos modelos llegan a conclusiones similares, aunque hay diferencias puntuales que sugieren la pertinencia de que este estudio sea replicado en el futuro para esclarecer mejor el comportamiento de los datos ante las variables utilizadas. Debe notarse que el modelo lineal se usa para comparar los resultados del modelo logístico, ya que este último es el más adecuado por el tipo de respuesta con la que se trabaja, no obstante, se pueden plantear estudios futuros para considerar las ventajas de uno sobre el otro.

De la misma forma, debe señalarse que este estudio se limita solo a estudiar los granos de maíz que no explotan, sin tomar en cuenta factores como el costo de implementar tratamientos para que más palomitas revienten o la cantidad de granos que son perdidos al quemarse. Estas consideraciones afectarían la aplicabilidad de los resultados a una escala mayor y deberían ser estudiados con mayor profundidad.

BIBLIOGRAFÍA

Fox, J. y Weisberg, S (2019). car: An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. Recuperado de: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gökmen, S. (2004). Effects of moisture content and popping method on popping characteristics of popcorn. Journal of Food Engineering, 65, 357-362. Recuperado de: https://www.researchgate.net/publication/240430251_Effects_of_moisture_content_and_popping_method_on_popping_characteristics_of_popcorn

J. Delcour, C. Hoseney. (2010). Principles of Cereal Science and Technology. 3ra edición. St. Paul, Minnesota. AACC International Press.

Karababa, E. (2004). Physical properties of popcorn kernels. Journal of Food Engineering, 72, 100-107. 10.1016/j.jfoodeng.2004.11.028

Kutner, M., Nachtsheim,C., Neter, J., Li, W. (1974). Applied Linear Statistical Models. New York, USA. McGraw-Hill Irwin.

Lusas, E., Rooney, L. (2000) Snack Foods Processing. Boca Raton, Florida. CRC Press.

Metzger, D. D., Hsu, K. H., Ziegler, K. E., & Bern, C. J. (1989). Effect of moisture content on popcorn popping volume for oil and hot-air popping. Cereal Chemistry, 66, 247–248. Recuperado de: https://www.cerealsgrains.org/publications/cc/backissues/1989/Documents/66_247.pdf

Villanueva Flores, R. (2008). El maíz reventador como alternativa industrial. Ingeniería Industrial.pp. 113-124. Recuperado de: https://www.redalyc.org/articulo.oa?id=337428492007

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. New York, EU: Springer-Verlag.

Zeileis,A. y Hothorn,T. (2002). lmtest: Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. Recuperado de: https://CRAN.R-project.org/doc/Rnews/

ANEXOS

Figura 1

Proporción de granos que no revientan por el nivel de tiempo y bloque

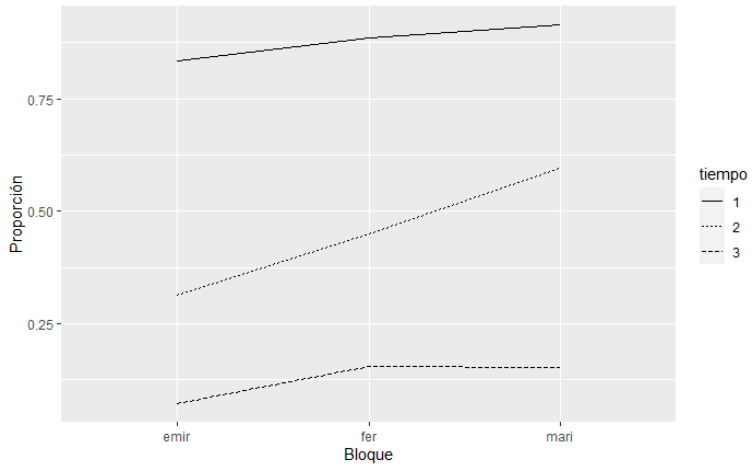


Figura 2

Proporción de granos que no revientan por el nivel de humedad y bloque.

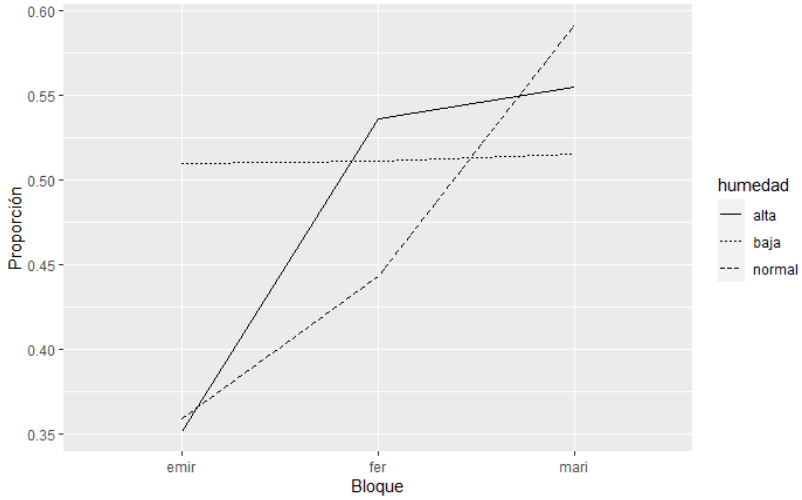


Figura 3

Proporción de granos que no revientan por número de granos y bloque

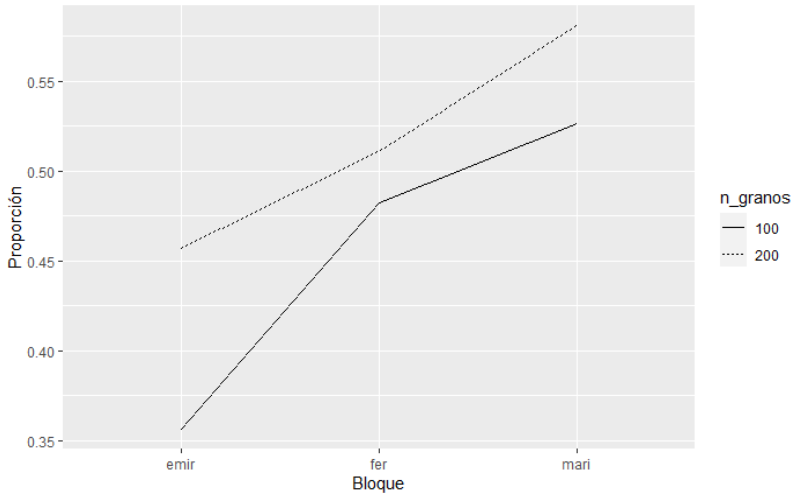
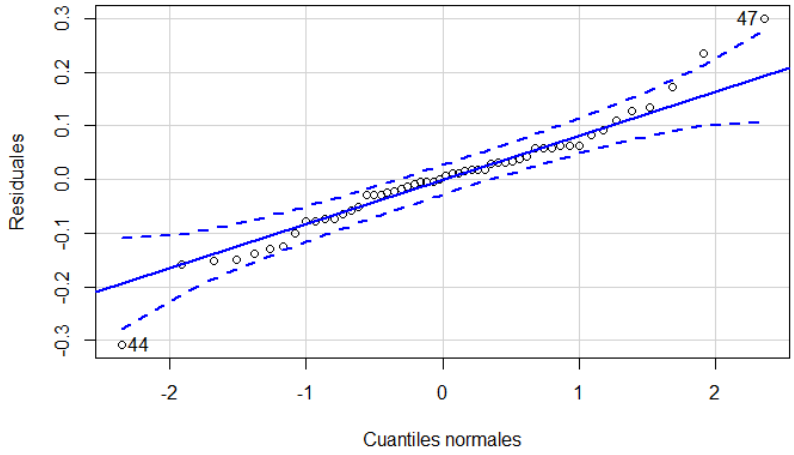


Figura 4

QQPlot de residuales del modelo lineal.



II. SIMULACIONES

En ocasiones interesa resolver una problemática o interrogante y no es posible lograrlo con pruebas estadísticas tradicionales, en casos como estos, se puede recurrir a los procesos conocidos como simulaciones para obtener respuesta de forma empírica. Robert Shannon define una simulación como “el proceso de diseñar un modelo de un sistema real y llevar a término experiencias con él, con la finalidad de comprender el comportamiento del sistema o evaluar nuevas estrategias" (Shannon y Johannes, 1976)[3].

[3] Shannon, Robert; Johannes, James D. (1976). Systems simulation: the art and science. IEEE Transactions on Systems, Man and Cybernetics. 6(10). pp. 723-724.

Efecto del uso de mínimos cuadrados ponderados en la potencia de la prueba de hipótesis para diferencias de medias, cuando se incumple el supuesto de homocedasticidad

Pedro Campos Jiménez[4], Steven Quirós Barrantes[4], Maripaz Venegas González[4]

pcamjmz14@hotmail.com, stevenqb84@gmail.com, maripaz5199@hotmail.com

RESUMEN

Cuando se estiman modelos es importante verificar que se cumplan los supuestos, pues si no se cumplen se deben tomar medidas remediales. En este trabajo se observa qué ocurre con la potencia para la prueba de igualdad de medias en un diseño factorial con tres tratamientos, cuando el supuesto de homocedasticidad no se cumple. Además, se quiere comprobar si al usar mínimos cuadrados ponderados (MCP) se obtiene una mejor potencia que cuando se ignora la heterocedasticidad presente. Se plantea una simulación que obtenga la potencia de la prueba (probabilidad de rechazar una hipótesis que es falsa) para distintos tamaños de muestra, ante cuatro grados de incumplimiento del supuesto de homocedasticidad. Como resultados, se obtiene que el incumplimiento del supuesto tiene poco efecto en la potencia cuando el nivel de heterocedasticidad es pequeño, en cambio, el efecto aumenta si las varianzas difieren más entre sí. Se encuentra, además, que si no se trata la heterocedasticidad con MCP la potencia se ve afectada levemente por diferencias pequeñas entre las varianzas, pero conforme el nivel de heterocedasticidad incrementa, el método genera una potencia de la prueba mayor.

PALABRAS CLAVE: supuesto de homocedasticidad, heterocedasticidad, comparación de medias potencia de la prueba, mínimos cuadrados ponderados, tamaño de muestra.

ABSTRACT

When a model is estimated, it is important to verify if assumptions are met. When they're not remedial actions should be taken. The current study analyses what occurs with the power of a test for a mean comparison in a factorial design with three treatments, when homoscedasticity is not satisfied. It also checks if using weighted least square (WLS) method improves test power in contrast to ignored heteroscedasticity. The study uses a simulation to obtain the power of a test (probability of failing to reject a hypothesis when it is false) with different sample sizes and four degrees of homoscedasticity violation. The results show that the assumption non-compliance has little effect in the power when the level of heteroscedasticity is small, instead, the effect grows if variance difference increases. Furthermore, when heteroscedasticity is not treated with WLS, power is slightly affected by small variance difference, however, as the level of heteroscedasticity increases, this remedial method results in a greater power of the test.

KEYWORDS: homoscedasticity assumption, heteroscedasticity, comparison of means. power of a test, weighted least squares, sample size

[4] Estudiantes de Estadística de la Universidad de Costa Rica

INTRODUCCIÓN

Muchos de los trabajos en estadística se basan en el cumplimiento de supuestos, por ejemplo, el supuesto de normalidad, de linealidad, de homocedasticidad, entre otros. La relevancia de que estos supuestos se ve reflejada en la calidad de las estimaciones. Un experimentador, como menciona Cochran (1974), raramente puede convencerse a sí mismo de que todos los supuestos de su modelo se han cumplido. Ante esto surge el interés de revelar los cambios en las estimaciones cuando se incumple alguno de los supuestos.

Cohen (1998) señala que, a pesar de ser un concepto importante, la potencia estadística suele entenderse bien ni es tomado en cuenta en muchas investigaciones. Al encontrarse con un estudio lo más común es hallar referencias al valor de la significancia (α), la probabilidad de cometer el error tipo I. Sin embargo, como menciona González-Lutz (2008) el concepto de error tipo II (β) no es tan utilizado. Ligado a este error, se encuentra el concepto de potencia, la probabilidad de rechazar cuando la hipótesis nula es falsa y tomar una decisión acertada.

González-Lutz (2008) declara que la potencia es vital cuando se investiga una relación de igualdad. Además, según Quezada (2007) la potencia estadística se puede utilizar para definir el tamaño muestral de un estudio, determinar la viabilidad o inviabilidad de una investigación y para controlar el error Tipo II. Menciona que cuando se tiene un valor de potencia suficientemente alto, "la investigación gana en rigor y, en posibilidades de publicación y aceptación". La potencia estadística, o poder estadístico, depende del nivel de significancia, el tamaño de la muestra y el tamaño del efecto (Cohen, 1992).

Por otra parte, Aguilar, Alvarado, Hernández (2019) abordan las consecuencias de la heterocedasticidad sobre la cobertura del intervalo de confianza para la diferencia entre dos medias. A partir de una simulación, analizan tres formas para trabajar heterocedasticidad de los datos: ignorándola, tomándola en cuenta con el método de mínimos cuadrados ponderados y considerándola a partir de la suma de las varianzas de las medias. Concluyen que los métodos en que se toma en cuenta la heterocedasticidad no difieren entre ellos, pero generan una mayor cobertura que cuando se ignora. La relación es más clara conforme aumenta el grado de incumplimiento del supuesto de igualdad de varianzas.

En este artículo, se abordará el efecto de la heterocedasticidad sobre la potencia de la prueba de igualdad de medias en un modelo factorial simple. En primer lugar, de forma general se comparan dos escenarios, uno en el que sí se cumple el supuesto y otro en el que no se cumple. Con base en la metodología utilizada por Aguilar et. al (2019), se tomará en cuenta la cantidad de réplicas por tratamiento y el 'grado' de incumplimiento del supuesto de igualdad de varianzas. Asimismo, se analiza el comportamiento de la potencia estadística cuando la heterocedasticidad es tomada en cuenta con método de mínimos cuadrados ponderados, respecto al caso en que es ignorada.

METODOLOGÍA

Primeramente, se define un modelo factorial simple con un factor de tres niveles. El modelo con el que se trabaja se puede escribir de la siguiente forma:

$$\mu_j = \mu + \tau_j$$

Donde, μ_j corresponde a la media del tratamiento *j*, μ a la media general y τ_j es el efecto del tratamiento *j*-ésimo que cumple con la restricción $\sum_{j=1}^{k-1} \tau_j = 0 \Rightarrow \tau_k = -\sum_{j=1}^{k-1} \tau_j$, ya que se trata de un modelo de suma nula. Como se tienen tres tratamientos, al expandir el modelo se obtiene $\mu_{trat} = \mu + \tau_1 C_1 + \tau_2 C_2$, donde C_1 y C_2 son variables auxiliares que se sustituyen por 1 y 0 según el tratamiento de interés.

En un primer caso, los datos para estimar los coeficientes del modelo provienen de distribuciones normales donde los tratamientos tienen varianzas iguales, es decir, cumplen el supuesto de homocedasticidad: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$. La estimación de los coeficientes se obtiene a partir de $\hat{\beta} = (X'X)^{-1}X^1Y$, donde X es la matriz de estructura y Y los valores de la variable respuesta. Para un segundo conjunto hay heterocedasticidad: al menos una de las varianzas de tratamiento difiere. En este, los datos son abordados de dos maneras. En primera instancia se ignora, se asume que hay homocedasticidad y se procede realizar las estimaciones de los coeficientes que minimizan la suma de cuadrados residual a partir de $\hat{\beta} = (X'X)^{-1}X^1Y$, donde X es la matriz de estructura y Y los valores de la variable respuesta.

En segundo lugar, se aplica el método de mínimos cuadrados ponderados (MCP) para tomar en cuenta la heterocedasticidad en las estimaciones. Este método se define como la ponderación de cada observación por el inverso de la variabilidad dado el tratamiento. Según Alvarado (2019), el procedimiento para realizar este método es inicialmente ajustar el modelo de regresión sin ponderar y analizar los residuales. Se estima la función de varianza y a partir de los valores ajustados se obtienen los ponderadores que en matricialmente corresponden a:

$$W = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

Donde $w_i = \frac{1}{(\hat{s}_i)^2}$ y $\hat{s}_i{}^2$ es la varianza de cada tratamiento. La estimación de los coeficientes de regresión utiliza los pesos en la matriz W, de forma que los estimadores se obtienen a partir de $\hat{\beta}_\omega = (X'WX)^{-1}X^1WY$

En seguida, se procede a realizar una Simulación de Montecarlo. En primer lugar, se generan datos muestrales para cada uno de los tres tratamientos provenientes de distribuciones normales con medias de 19, 25 y 31 unidades. Las distribuciones una vez cumplen el supuesto de homocedasticidad y en otro caso presentan heterocedasticidad.

A partir de la metodología de Aguilar et. al (2019), cuando se presenta heterocedasticidad se crean cuatro escenarios según el nivel de incumplimiento de la igualdad de varianzas: nivel “considerable” (una varianza representa 1.5 veces la otra), nivel “fuerte” (2 veces), nivel “alarmante” (4 veces) y nivel “muy alarmante” (8 veces). Se fija 25 como la varianza a partir de la cual se generan los distintos escenarios.

Se cuenta con un diseño balanceado, por lo que cada tratamiento tiene la misma cantidad de réplicas. Se generan muestras para números de réplicas de 2 a 50. El tamaño total de la muestra se obtiene al multiplicar la cantidad de réplicas por 3, el número de tratamientos. Con los datos muestrales generados bajo igualdad de varianzas, además del caso en que se ignora la heterocedasticidad, se crea el modelo factorial y se calcula la suma de cuadrados residuales (SCRes) y el cuadrado medio residual (CMRes) de la siguiente manera:

$$SCRes = \sum_{j=1}^{k} (r_j - 1)s_j^2 \quad y \quad CMRes = \frac{SCRes}{n-k} \quad ,$$

Donde *k* es el número de tratamientos, *n* es el tamaño de muestra total, *rj* es el número de réplicas por tratamiento y s^2j es la varianza de cada tratamiento.

Asimismo, se calcula la suma de cuadrados de tratamiento y el cuadrado medio de tratamiento:

$$SCTrat = \sum_{j=1}^{k} r_j(\bar{y}_j - \bar{\bar{y}}) \quad y \quad CMTrat = \frac{SCTrat}{k-1} \quad ,$$

Donde $\bar{y}_j$ es la media estimada para el tratamiento *j*, $\bar{\bar{y}}$ es la media general, r_j es el número de réplicas por tratamiento y *k* es el número de tratamientos.

Cuando hay heterocedasticidad y se toma en cuenta a partir del método de mínimos cuadrados ponderado, el cuadrado medio residual se estima de la siguiente forma:

$$CMRes = \frac{\sum_{j=1}^{k} w_j(y_j - \bar{y}_j)^2}{n-k}$$

Donde w_j es el ponderador de los valores ajustado del modelo, *k* es la cantidad de tratamientos y *n* el tamaño total de la muestra.

Una vez que se tienen esos valores, con el fin de realizar la prueba de hipótesis de igualdad de medias, se calcula el estadístico F que corresponde al cociente del CMTrat y el CMRes. Con la distribución F se obtiene la probabilidad de cometer el error tipo I asociada al valor de F calculado. Esta probabilidad se contrasta con el nivel de significancia que en este caso se define como 0.05.

Este proceso se realiza 10000 veces y se contabiliza las veces en que la hipótesis nula de igualdad de medias se rechaza. Efectivamente, la hipótesis nula es falsa, por lo tanto, la proporción de veces que es rechazada corresponde a la potencia de la prueba, cuyos valores oscilan entre 0 y 1. Entre mayor sea la potencia, quiere decir que se logró detectar en más ocasiones la diferencia entre medias.

El procedimiento se realiza para cada número de réplicas por tratamiento en cada uno de los cuatro escenarios de incumplimiento del supuesto de homocedasticidad.

Para efectos de la simulación estos valores se extraen de los modelos con homocedasticidad y heterocedasticidad, con y sin los pesos método de mínimos cuadrados ponderados, de la función de análisis de varianza (ANOVA por sus siglas en inglés) del software R versión 4.0.0 (2020) mediante RStudio (2016), que obtiene directamente la probabilidad asociada al error tipo I. Asimismo, para visualizar los resultados de manera gráfica, se utilizan los paquetes *ggplot2* (Wickham, 2009) y *car* (Fox y Weisberg, 2019).

RESULTADOS

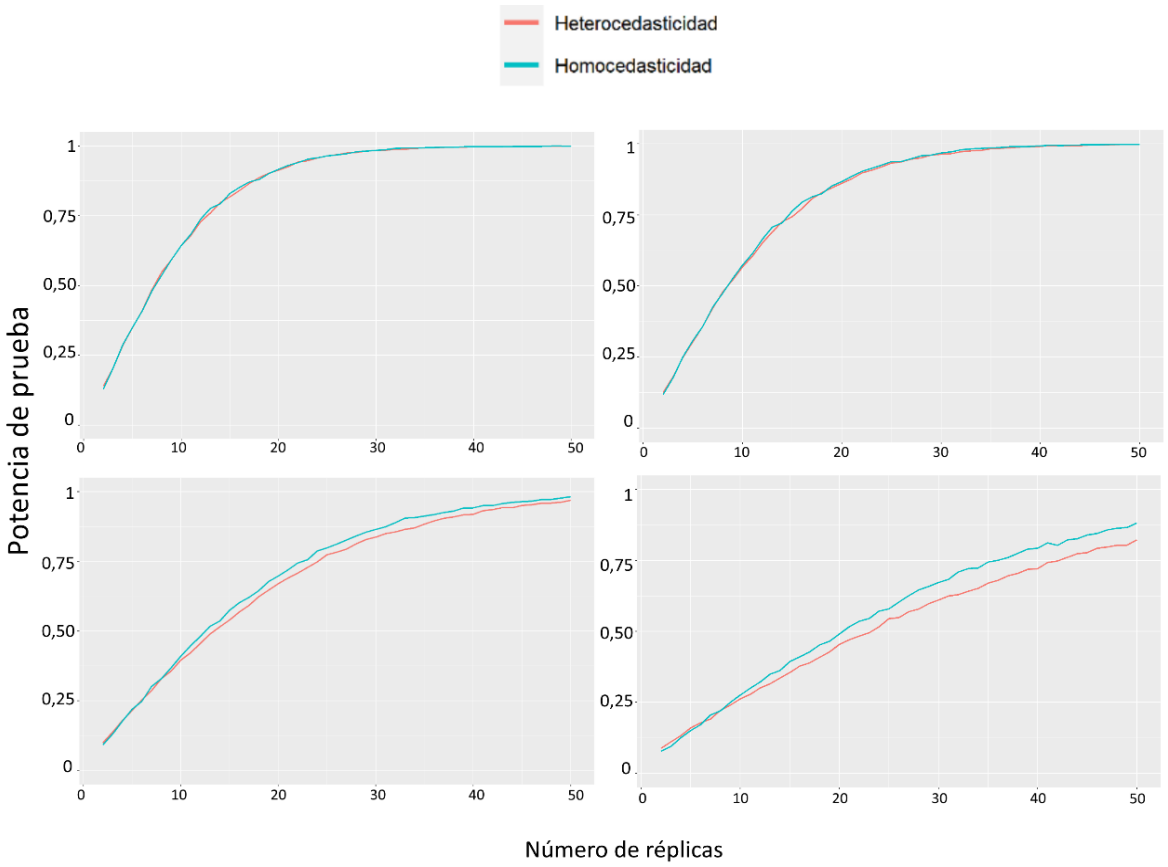
En seguida se presentan los resultados de la simulación. La diferencia entre la potencia cuando hay homocedasticidad y cuando hay heterocedasticidad se puede observar en la figura 1. Se nota cómo a mayor cantidad de réplicas (derecha a izquierda en el eje x), la potencia aumenta. En cuanto a los niveles de

heterocedasticidad considerable y fuerte, no se aprecia una diferencia relevante entre ambas curvas. Es decir, ignorar el incumplimiento del supuesto de homocedasticidad parece no afectar la potencia de la prueba.

Mientras tanto, cuando se de los niveles de heterocedasticidad alarmante y muy alarmante se tiene que, para ambos, a partir de 10 réplicas por tratamiento, la diferencia de potencias empieza a ser más notoria. Esto nos indica que cuando los datos presentan un nivel de heterocedasticidad de este calibre, las estimaciones no van a tener la misma potencia debido al incumplimiento del supuesto de homocedasticidad.

Figura 1

Potencia de la prueba de igualdad de medias, según cantidad de réplicas por tratamiento ante grados de incumplimiento del supuesto de homocedasticidad.



Visto de otra manera, por ejemplo, en el nivel de heterocedasticidad muy alarmante, para obtener una potencia de aproximadamente 0.75 si se tiene homocedasticidad se requieren unas 35 réplicas por tratamientos. Mientras tanto, para obtener la misma potencia de 0.75 cuando hay heterocedasticidad se requieren más de 40 réplicas por tratamiento.

Por otra parte, la figura 2 compara el efecto de ignorar la heterocedasticidad con el de incluirla a través del método de mínimos cuadrados ponderados. A grandes rasgos, el comportamiento de las líneas es muy similar a lo observado en la figura 1. El comportamiento de las líneas en la figura 2, que compara es muy similar a lo observado en la figura 1.

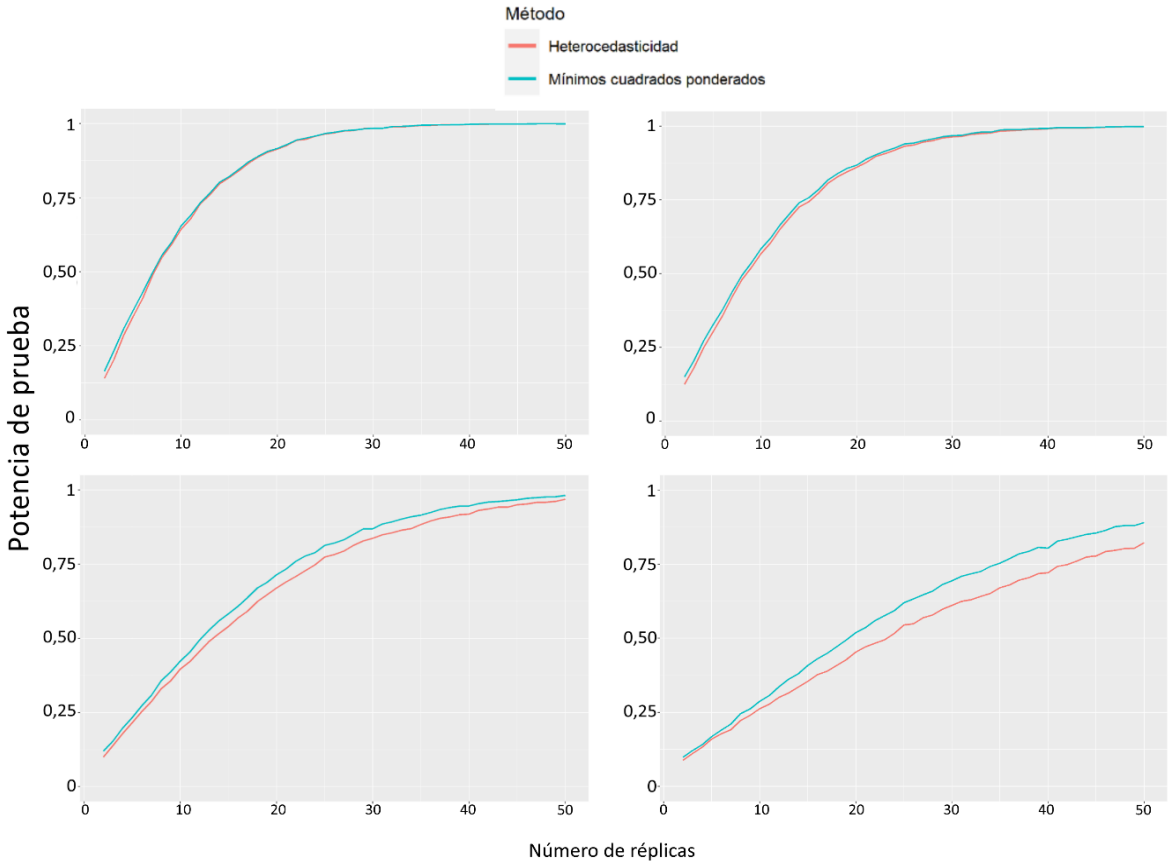
Se observa que para un escenario donde el supuesto de igualdad de varianzas es infringido en un nivel considerable la diferencia entre las líneas es mínima. En cambio, cuando el nivel de heterocedasticidad es fuerte se empiezan a notar cambios: el método de mínimos cuadrados ponderados genera una potencia mayor que cuando se ignora la heterocedasticidad.

Finalmente, en los gráficos inferiores que muestran la potencia cuando el nivel de heterocedasticidad es alarmante y muy alarmante, se observa cómo las líneas se van diferenciando más a medida que aumenta el tamaño de réplicas. Por ejemplo, en el caso de heterocedasticidad alarmante para obtener una potencia aproximada de 0.75, por el método de mínimos cuadrados ponderados necesitamos de 20 réplicas por tratamiento, y para el caso donde no fue tratada la heterocedasticidad, se requieren de aproximadamente unas 25 réplicas.

Estas diferencias son aún más notorias en el gráfico de heterocedasticidad muy alarmante, pues las líneas se encuentran muy separadas y las diferencias de réplicas por tratamiento para calcular una potencia dada, aumentan cada vez más.

Figura 2

Potencia de la prueba de igualdad de medias, ignorando la heterocedasticidad o considerándola por medio de MCP según cantidad de réplicas por tratamiento ante grados de incumplimiento del supuesto de homocedasticidad.



CONCLUSIONES

En primer lugar, el cumplimiento de la homocedasticidad otorga ventajas cuando se quiere incrementar el valor de la potencia. Empero, la potencia depende considerablemente de la cantidad de réplicas por tratamiento, pues aún con el cumplimiento del supuesto la potencia es relativamente baja cuando se tienen pocas réplicas. Además, esta ventaja no es relevante cuando el nivel de heterocedasticidad es considerable o fuerte (una de las varianzas es 1.5 o 2 veces la otra). Los incrementos en la potencia cuando se cumple el supuesto respecto a cuando no se cumple son más evidentes para diferencias de varianzas mayores y conforme aumenta el tamaño de muestra, tal como lo plantea Cohen (1992) quien afirma que la potencia depende del nivel de significancia, el tamaño del efecto y el tamaño de la muestra.

Por otro lado, si se incumple el supuesto de homocedasticidad, no significa que una potencia alta sea inalcanzable, simplemente se necesita un tamaño de muestra más grande. Quesada y Figueroa (2010) mencionan que, cuanto mayor tamaño muestral tengamos mayor representatividad tendrá la población que medimos, más fiable será nuestro análisis, y por lo tanto la potencia de la prueba estadística aumentará. Qué tanto debe aumentar la muestra depende del grado de incumplimiento del supuesto.

En el caso de un grado elevado de heterocedasticidad, es viable recurrir al método de mínimos cuadrados ponderados. Esto coincide con los resultados presentados por Alvarado et al. (2019), quienes hallan que al ignorar el incumplimiento del supuesto y conforme aumenta el grado de incumplimiento del supuesto de homocedasticidad, el intervalo de confianza tiene un desempeño por debajo del 95% deseado. Es decir, las estimaciones empeoran, con lo cual, esta tendencia se ve igualmente representada en la potencia de la prueba.

No es recomendable hacer la estimación de un modelo a base de datos con heterocedasticidad en la variable de respuesta de interés sin antes usar una medida remedial como puede ser hacer mínimos cuadrados ponderados, así como lo mencionan Osinki, Bruno y Costas (2000) cuando no sea posible trabajar con muestras numerosas, se deben plantear otras formas de análisis que permitan resultados más esclarecedores; debido a que con este método se va a obtener una potencia de prueba mayor. Este mejor desempeño se acentúa cuando las diferencias de varianzas son muy grandes y además si se tienen muchas réplicas por tratamiento.

Para estudiar más a fondo la problemática, se pueden incluir cambios en otros parámetros que suelen afectar el valor de la potencia. En este trabajo se fijó una media (25) y una diferencia entre medias (6), estos valores se pueden cambiar para observar el comportamiento de la potencia ante distintos efectos por tratamiento. Por otro lado, queda pendiente la aplicación de una simulación de esta índole en modelos no balanceados o más complejos, por ejemplo, más niveles por factor, más cantidad de factores, presencia de interacción, de bloques o la inclusión de variables continuas.

BIBLIOGRAFÍA

Aguilar, J. P., Alvarado, F. y Hernández, D. (2019). Afectación de la cobertura de los intervalos de confianza para diferencias de medias al incumplir el supuesto de homoscedasticidad, un enfoque de tres alternativas para analizar datos heteroscedásticos. *Serengueti 1*(2), 46-53. Recuperado de http://hdl.handle.net/10669/80141

Alvarado, R. (2019) Modelos de Regresión Aplicados: IV. Construcción del modelo. Universidad de Costa Rica. Presentaciones del curso.

Cochran, W. (1974). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3(1), 22-38. Recuperado de https://www.jstor.org/stable/3001535.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98-101. Recuperado de www.jstor.org/stable/20182143.

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences* (2da ed.). Nueva Jersey, NJ: Lawrence Erlbaum Associates. Recuperado de http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf.

Fox, J. y Weisberg, S. (2019). *car: An R Companion to Applied Regression*. Thousand Oaks, California.

González-Lutz, M. I. (2007). Potencia de prueba: la gran ausente en muchos trabajos científicos. *Agronomía Mesoamericana, 19*(2), 309-313. Recuperado de https://doi.org/10.15517/am.v19i2.5015.

Osinski, I. C., Bruno, A. S., y Costas, C. S. L. (2000). Estudio de la potencia de los contrastes de medias con dos y tres grupos con tamaño de efecto pequeño y en condiciones de no normalidad y homo-heterocedasticidad. *Psicothema*, 12(Suplemento), 114-116.

Quezada, C. (2007). Potencia estadística, sensibilidad y tamaño de efecto: ¿Un nuevo canon para la investigación? *Onomázein*, *16*, 159-170. Recuperado de http://onomazein.letras.uc.cl/Articulos/16/4_Quezada.pdf.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de https://www.R-project.org/.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA Recuperado de http://www.rstudio.com/.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. New York, EU: Springer-Verlag

III. MODELOS LINEALES GENERALIZADOS (MLG)

Los modelos lineales generalizados están formados por dos componentes: componente aleatorio (identifica la variable respuesta y la distribución de probabilidad) y el componente sistemático (especifica las variables explicativas utilizadas en la función predictora lineal). Fue introducida por primera vez por Nelder y Wedderburn y está constituida por modelos de regresión lineal de error normal y los modelos de regresión exponencial, logística y de Poisson no lineal, entre otros tipos de modelos, como los modelos log-lineales para datos categóricos (Kutner et al,2004)[5].

[5] Kutner, M., Nachtsheim, C., Neter, J. y Li, W. (2004). Applied Linear Statistical Models. 5th ed. New York: McGraw-Hill Companies, Inc.

Modelo de regresión Poisson para la predicción de muertes por la Covid-19 en Costa Rica en los meses de agosto y setiembre del año 2020.

Carlos Daniel Leandro Aguilar[6]

carlos.leandroaguilar@ucr.ac.cr

RESUMEN

La pandemia que se vive actualmente por la presencia de la COVID-19 ha cambiado el estilo de vida de todas las personas de alguna u otra forma desde el treinta de enero del 2020. Los modelos predictivos permiten predecir a cierta escala el comportamiento de la pandemia aplicado a la situación nacional y respaldar las medidas de regulación. Se utilizó el lenguaje de programación R con la versión RStudio-1.3.1073 a partir del cual se programó y generó un modelo de regresión de Poisson para visualizar mediante un gráfico el comportamiento de los casos activos en Costa Rica por la pandemia COVID-19. Al realizarse la prueba de hipótesis con un alfa de 0.05, los valores del intercepto, así como el de las pendientes son menores a este por lo que se puede afirmar que las variables tienen efecto sobre la variable respuesta. El comportamiento esperado según la proyección a finales del mes de agosto e inicios de septiembre es entre 15 y 19 muertes diarias aproximadamente. El modelo es un buen indicador de que pese a que se necesita la apertura comercial si no se toman medidas estrictas y consistentes, como el uso obligatorio de mascarillas y caretas el futuro es poco prometedor, el sistema de salud en algún punto colapsaría desgraciadamente como también se ha proyectado en otros estudios.

PALABRAS CLAVE: Coronavirus, COVID-19, modelo de predicción, Poisson.

ABSTRACT

The pandemic currently being experienced by the virus COVID-19 has changed the lifestyle of all people in one way or another since January 30, 2020. Predictive models allow having an idea of the behavior of the pandemic applied to a certain scale of the national situation and support government regulatory measures. The R programming language was used with RStudio-1.3.1073 version from which a linear Poisson regression model was programmed and generated to visualize through a graph the behavior of active cases in Costa Rica due to the COVID-19 pandemic. When the hypothesis test is carried out with an alpha of 0.05, the values of the intercept as well as the slopes are less than this, so it can be stated that the variables have an effect on the response variable. The expected behavior according to the projection at the end of August and the beginning of September is between 15 and 19 deaths per day approximately. The model is a good indicator that despite the need of economic openness if strict and consistent measures are nottaken, such as the mandatory use of masks and face masks, the future is not very promising and the health system unfortunately would collapse at some point as well as it has been screened in other studies.

KEY WORDS: Coronavirus, COVID-19, prediction model, Poisson.

[6] Estudiante de Estadística de la Universidad de Costa Rica

INTRODUCCIÓN

La pandemia que se vive actualmente por la presencia de la COVID-19 ha cambiado el estilo de vida de todas las personas de alguna u otra forma, desde que el treinta de enero del 2020 la Organización Mundial de la Salud declaró oficialmente al SARS-CoV2 como una situación de emergencia de salud pública. Para entender mejor las características de este virus es importante conocer que existen otros tipos de coronavirus humanos endémicos, tales como el HCoV-229E, HCoV-NL63, HCoV-HKU1 y HCoV-OC43, siendo estos coronavirus diferentes al nuevo coronavirus llamado SARS-CoV-2 que produce la enfermedad llamada COVID-19 (Ministerio de Salud, 2020), la cual es una enfermedad de dinámica respiratoria que puede comprometer la vida de las personas.

Si bien la salud ha sido el principal área de impacto de la pandemia, no ha sido el único ámbito que se ha visto involucrado ya que las medidas sanitarias como el distanciamiento social y el cierre de negocios ha impactado la economía del país (Hiscott, 2020), aumentando la tasa del desempleo y las reducciones de jornada, ha creado desesperación en la población e incluso ha llegado a afectar a nivel psicológico a aquellas personas que se han visto vulnerables ante las decisiones tomadas por el Gobierno.

Es así como la Estadística ha jugado un papel importante a lo largo de estos meses en donde abunda la incertidumbre ante los datos presentados diariamente en las noticias, siendo los modelos predictivos una base importante que permiten conocer a cierta escala el comportamiento de la pandemia aplicado a la situación nacional y respaldar las medidas de regulación que al aplicarse en los momentos oportunos, podrían eventualmente evitar un aumento exponencial de casos activos que colapsen los servicios de salud del país.

Actualmente estos modelos probabilísticos han sido posibles de realizar a través de softwares como R Studio el cual funciona con “una serie de reglas diseñadas para realizar procesos en una computadora” (Andina, 2018).

Los modelos regresión de Poisson son comúnmente usados para modelar eventos en los que su variable respuesta o dependiente son conteos, más específicamente contar datos discretos. Estos modelos ayudan a determinar el efecto de las variables predictoras o dependientes (X1, X2) en la variable respuesta o independiente (Y) (Dataquest, 2017).

METODOLOGÍA

Para llevar a cabo el presente trabajo se utilizó el lenguaje de programación R (RStudio Team, 2020) con la versión RStudio-1.3.1073 bajo la colaboración del Dr. Guaner Rojas Rojas, docente de la Escuela de Estadística de la Universidad de Costa Rica. A partir de esto se programó y generó un modelo de regresión de Poisson que permite a través de un gráfico visualizar el comportamiento de los casos activos en Costa Rica por la pandemia COVID-19 tomar los casos activos registrados desde el 7 de marzo del 2020, fecha en que se registró el primer caso activo en el país hasta el 28 de julio del 2020 se proyectó un modelo sobre el posible comportamiento hasta finales de agosto e inicios de septiembre del mismo año. Los datos de casos activos registrados se tomaron de la base de datos “Coronavirus source data” disponible en la página web Our World Source Data. Se utilizó un modelo de regresión de Poisson tomando como variable respuesta o predicha new deaths, como variables predictoras o independientes total cases (Y1) y new tests (Y2).

Se utilizó la función glm que reproduce un modelo lineal general en el cual la variable respuesta no sigue una distribución normal sino una distribución Poisson, contrario a los modelos lineales, esto debido a que no hay total certeza de que exista relación lineal entre las variables predictoras y la variable predicha. Para transformar esta relación no-lineal a una forma lineal se utilizó la función (link = "log") que indica la familia con la que se trabajó, en este caso la poisson. La ecuación empleada por el modelo es la siguiente:

$$log\ log\ (y) = \alpha + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp$$

Donde "y" es la variable respuesta, α y β son los coeficientes numéricos, siendo α el intercepto (β_0) y "x" las variables predictoras.

RESULTADOS

Primero se verifica que exista efecto de las variables predictoras en la variable respuesta, es decir, si el total de casos y las nuevas pruebas realizadas diariamente sirven para predecir el número de muertes en el país. Al realizarse la prueba de hipótesis con un alfa de 0.05, los valores del intercepto, así como el de las pendientes son menores a este. Por lo tanto, se puede afirmar que las variables tienen efecto sobre la variable respuesta.

En el análisis descriptivo previo se puede apreciar cómo el número de muertes en los últimos meses tuvo un incremento importante (Ver gráfico 2 y 3). Los casos bajos de muertes se pueden ver reflejados durante los primeros dos meses, esto se explica por las fuertes restricciones impuestas por el gobierno hasta cierto punto (Ver gráfico 2). El aumento lento pero progresivo de muertes se puede apreciar en el transcurso del mes de mayo, junio y julio donde finalmente se tiene un número mucho mayor de casos, sin embargo, no se alcanza un comportamiento exponencial en el cual exista un aumento de manera significativa, sino que se mantiene el comportamiento logístico para el caso de las muertes (Ver gráfico 2).

La proyección relativamente alta de muertes para finales de agosto concuerda con las medidas tomadas por las autoridades del país ya que se está en un momento de apertura comercial, disminución de restricciones vehiculares y por lo tanto hay un aumento de personas en la calle. Asimismo, se puede ver cómo en algunos puntos la predicción es más baja que la cantidad de muertes y este comportamiento se podría explicar al uso de mascarillas, caretas y demás medidas de protección impuestas por el gobierno (Ver gráfico 1).

CONCLUSIONES

Desde el punto de vista técnico el modelo es favorecedor, dado que permite de ver de manera simple una comparación ya que también se tiene a disposición una base de datos que se actualiza diariamente.

Según lo que indica el modelo a un ritmo de más de nueve muertes diarias en promedio y la tendencia a oscilar en un máximo de dos casos menos, no parece indicar que este crecimiento disminuye por lo menos a corto plazo ya que la línea de predicción finaliza en un punto alto, sin embargo, al ser un aumento con un ritmo muy lento proporciona el tiempo suficiente para que las autoridades tomen las acciones en el momento oportuno y eventualmente se evite un colapso del sistema de salud.

El modelo es un buen indicador de que pese a que se necesita la apertura comercial si no se toman medidas estrictas y consistentes, como el uso obligatorio de mascarillas y caretas el futuro es poco

prometedor, el sistema de salud en algún punto colapsaría como también se ha proyectado en otros estudios. Costa Rica cuenta aproximadamente con 227 camas en cuidados intensivos y aún con esta capacidad si aumentara la cantidad de muertes a las proyectadas se estarían estimando indicios de saturación hospitalaria.

La idea de este trabajo inició en el curso modelos probabilísticos discretos en el primer ciclo del año 2020.

BIBLIOGRAFÍA

Andina, M. 2018 *Introducción a la estadística con R.* [online] Disponible en: https://bookdown.org/matiasandina/R-intro/ [Accesado el 5 de septiembre del 2020]

Dataquest. 2017. Tutorial: Poisson Regression in R. [Online]. Disponible en: https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/#:~:text=The%20response%20variable%20yi,that%20follows%20the%20Poisson%20distribution. [Accesado el 5 de septiembre del 2020]

Hiscott, J., Alexandrini, M., Muscolini, M., Tassone, E., et.al. (2020) *The global impact of coronavirus pandemic.* Elseiver: Cytokine and Growth Factor Reviews (53):1–92. Doi: https://doi.org/10.1016/j.cytogfr.2020.05.010

Ministerio de Salud. (2020). *Lineamientos generales para la limpieza y desinfección de viviendas que alojan casos en investigación, probables o confirmados de COVID-19 en el marco de la alerta sanitaria por Coronavirus (COVID-19)* [PDF] (2da versión). San José, Costa Rica Disponible en: https://www.ministeriodesalud.go.cr

Our World in Data. 2020. *Coronavirus Source Data*. [online] Disponible en: https://ourworldindata.org/coronavirus-source-data [Accesado el 26 de agosto del 2020].

RStudio Team. 2020. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. Disponible en: https://rstudio.com/

ANEXOS

Figura 1

Comparación gráfica lineal entre las nuevas muertes (rojo) y las muertes predichas (azul)

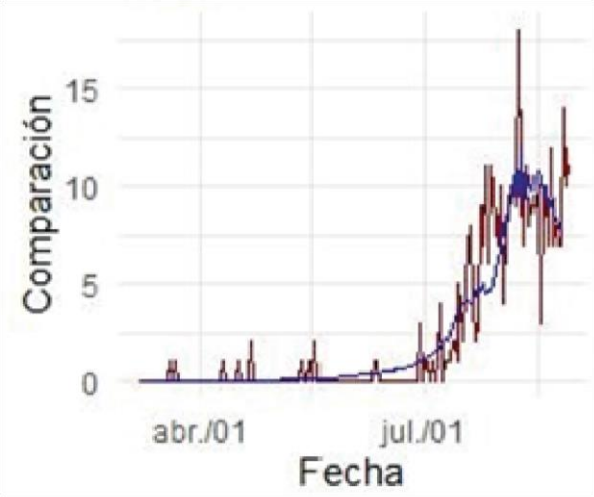


Figura 2

Comparación gráfica viendo la dispersión de las muertes predichas(azul) respecto a las nuevas muertes (rojo).

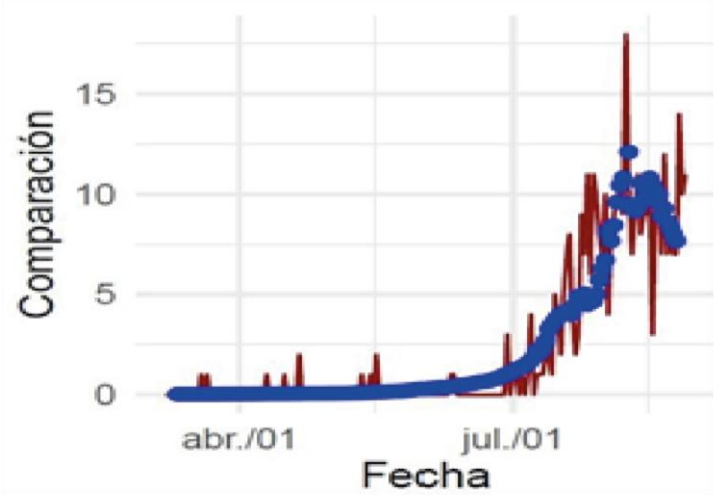
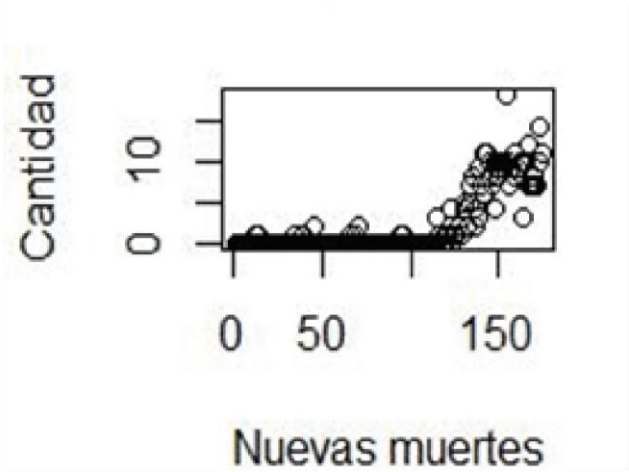


Figura 3

Análisis descriptivo del comportamiento del registro de muertes en modo de dispersión



IV. TÉCNICAS DE AGRUPAMIENTO

Las técnicas de agrupamiento o clustering son una colección de métodos estadísticos que buscan grupos de instancias con características similares mediante el análisis de los atributos que interesan estudiar. Son muy utilizadas en la minería de datos como una técnica de aprendizaje no supervisada, pero se utilizan en muchas áreas como la psicología, robótica y biología. Este tipo de técnicas fueron originadas en la antropología por Driver y Kroeber en 1932 (Driver y Kroeber, 1932)[7].

[7] Driver and Kroeber (1932). "Quantitative Expression of Cultural Relationships". University of California Publications in American Archaeology and Ethnology. Quantitative Expression of Cultural Relationships: 211–256 – via http://dpg.lib.berkeley.edu

Comparación de los conglomerados generados mediante un análisis de sentimientos sobre los tweets emitidos por los usuarios de Twitter Costa Rica en el periodo del 30 de abril al 6 de mayo del 2020

Josué Baltodano Leiva[8], Joshua Salazar Obando[8], Andrea Vargas Montero[8]

josuebaltodanoleiva@gmail.com, joshua.salazar1692@gmail.com, avargas2398@gmail.com

RESUMEN

El análisis de sentimientos es una manera de identificar emociones que son descritas en un texto por un individuo. En tiempos de pandemia, el tener que realizar distanciamiento social, se puede ver afectado el estado emocional de las personas y el cual es probable que se expresan a través de un medio escrito u oral. Bajo esta perspectiva, en esta investigación se extrajo sentimientos de los usuarios de Twitter Costa Rica a través de los tuits creados del 30 de abril al 6 de mayo del 2020. Una vez conformada la muestra, se procedió a realizar un análisis estadístico de conglomerados tomando como variables de agrupamiento 8 sentimientos, los cuales son furia, disgusto, miedo, felicidad, tristeza, sorpresa, anticipación y sorpresa. Además, se tienen variables que, una vez elaborados los grupos, se utilizaron para caracterizar a los individuos del estudio, estas corresponden a características demográficas como sexo, edad y la provincia de la persona, así como otras características pertenecientes a la red social Twitter como lo son la cantidad de seguidores y "likes" del usuario en el momento en que realiza una publicación. Dentro de los principales resultados, de los 3 grupos establecidos, en el caso del grupo 2 es el que presenta mayores sentimientos negativos, específicamente miedo tristeza, y el grupo 3 es el que obtiene mayores sentimientos positivos, relacionados a felicidad y confianza. Por último, no se tienen diferencias entre grupos para todas las características.

PALABRAS CLAVE: análisis de sentimientos, Twitter, caracterización, conglomerados

INTRODUCCIÓN

Cuando se expresan ideas o pensamientos de forma ya sea escrita u oral, a su vez se externan sentimientos, y esto puede presentarse de forma directa o indirecta. Al entablar una conversación se intenta descifrar los sentimientos que la otra persona está transmitiendo, para poder actuar conforme a estos. Sin embargo, descifrar sentimientos cuando la expresión es escrita muchas veces resulta más complejo, debido a que no se tiene interacción directa con el emisor del mensaje. Debido a que esta decodificación de sentimientos es de suma importancia para poder analizar, por ejemplo, la satisfacción de los clientes ante un servicio o producto, entre otros, se crea un tipo de análisis de texto conocido como análisis de sentimientos, extracción de opiniones, minería de opiniones, minería de sentimientos o análisis subjetivo. Este se define como el estudio computacional de opiniones, sentimientos y emociones expresadas en textos (Pang y Lee, 2008).

Este tipo de análisis está enmarcado dentro del campo del Procesamiento Natural y su objetivo principal es determinar la actitud de un escritor ante determinados productos, situaciones, personas u organizaciones, así como identificar los aspectos que generan opinión, quién posee estas opiniones y cuál es el tipo de emoción o su orientación semántica (Liu, 2010). Utilizando el análisis de sentimientos, en cualquiera

[8] Estudiantes de Estadística de la Universidad de Costa Rica

de los programas computacionales capaces, es posible identificar si la palabra utilizada expresa una emoción positiva o negativa. Además, de forma más específica y la más compleja dentro del marco del análisis, determinar el sentimiento "exacto" que transmite, como furia, enojo, frustración, sorpresa, por mencionar algunos ejemplos.

Las aplicaciones que se le han dado a este tipo de análisis son amplias, muchas veces alrededor de la publicidad y el estudio de la satisfacción ante un producto o servicio y debido al gran impacto que tienen las redes sociales en ambos aspectos, es común que la investigación se realice a textos extraídos de estas. Una de las formas más populares de hacer este tipo de análisis es mediante el análisis de tuits extraídos de la red social Twitter, ya que la expresión escrita es la principal forma en la que los usuarios de esta interactúan, contrario a otras redes sociales como Instagram o Facebook donde una parte visual juega un papel mucho más importante. El análisis de sentimiento utilizando la red social Twitter ha tenido un rápido crecimiento debido a sus utilidades, investigadores han explicado y predicho las consecuencias de distintos eventos, demostrando el valor que poseen las interacciones realizadas en esta red (Ghiassi, Skinner & Zimbra, 2013).

Por otro lado, dada la propagación de la enfermedad viral llamada COVID-19, su alto nivel de contagio y mortalidad, la realidad que viven muchas personas alrededor del mundo ha cambiado drásticamente. Costa Rica no es la excepción ya que, desde el 6 de marzo del presente año, el Ministerio de Salud, máxima autoridad en temas de salud del país, solicitó a los ciudadanos evitar salir a la calle; por lo que escuelas, colegios, universidad, comerciales, gimnasios, entre otros, tuvieron que cerrar o cambiar a una modalidad virtual. El cambio hacia un día a día con poco contacto con el exterior ha causado mayor vulnerabilidad al desarrollo de psicopatologías y otros aspectos negativos para la salud mental (Quezada, 2020).

En el presente trabajo se utiliza el resultado de un análisis de sentimientos aplicado a los tuits de personas usuarias de Twitter Costa Rica en la semana del 30 de abril al 6 de mayo, para elaborar grupos mediante un análisis de agrupamiento que permita caracterizar a las personas que se expresan en la red social según características demográficas como la edad, sexo y otras características relacionadas a la interacción de la persona usuaria en Twitter.

METODOLOGÍA

La recolección de los datos se realizó utilizando tanto el API[9] de Twitter como el software R versión 4.0.0 (R Core Team, 2020), específicamente la librería rtweet (Kearney, 2019). El uso del API de Twitter restringe el periodo del cual se pueden extraer datos a 1 semana previa al momento de la extracción. Esta limitante causó que todos los tuits extraídos para el análisis fueron emitidos en la semana del 30 de abril al 6 de mayo del presente año. Por otro lado, se agregó la limitante de que estos fueran emitidos por personas que viven en Costa Rica, esta limitante se agregó en las especificaciones de la extracción, pero también en la limpieza de datos se aseguró que se cumpliera. Asimismo, con el objetivo de utilizar ambas variables en el análisis posterior, se agregó

[9] API (siglas de 'Application Programming Interface') es un conjunto de reglas (código) y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas (Merino, 2014)

la edad y sexo del usuario, esto mediante la búsqueda manual, usuario por usuario, debido a que estas variables no son extraíbles automáticamente.

El análisis de sentimientos requiere que el texto que se analiza esté “limpio”. El proceso de limpieza de texto consiste en eliminar del texto todo aquello que no aporte información sobre su temática, estructura o contenido. No existe una única forma de hacerlo, depende en gran medida de la finalidad del análisis y de la fuente de la que proceda el texto. En este análisis, dado que el objetivo era analizar la forma en que se expresan las personas durante la cuarentena, se eliminaron siguientes elementos:

- Patrones no informativos (urls de páginas web)
- Etiquetas HTML
- Signos de puntuación
- Números
- Caracteres sueltos (por ejemplo, “a”)
- Hashtags y usuarios (por ejemplo, “#rstats” y “@carlosalvarado”)
- Stop words (palabras no relevantes para el análisis, por ejemplo, “una”)
- Texto en mayúscula.
- Emoticones

Una vez finalizada la limpieza se puede realizar el análisis de sentimientos como tal. Este tipo de análisis se puede implementar de distintas formas, sin embargo, la instrucción que se utilizó comprendía las siguientes emociones o sentimientos para hacer la clasificación:

- Furia
- Disgusto
- Miedo
- Felicidad
- Tristeza
- Sorpresa
- Anticipación
- Sorpresa.

Debido a que la extracción de tuits se hace de forma similar a un muestreo de tuits aleatorio y con reemplazo, un mismo usuario puede aparecer más de una vez dentro de la base de datos, por esta esta situación se cuenta con un total de 1884 tuits. Primeramente, se limitan los usuarios a aquellos de los que se obtuvieron de 4 a 10 tuits, lo cual causa que se manejaran 97 usuarios en total. Sin embargo, con el objetivo de tener una

cantidad suficiente de sentimientos que analizar para cada individuo se eliminaron aquellos usuarios de los que se obtuvieron menos de 7 sentimientos, asimismo, se eliminaron las cuentas pertenecientes a instituciones o empresas, por ejemplo, el Organismo de Investigación Judicial (OIJ). Por otra parte, se procedió a eliminar los valores extremos presentes para evitar que estos alteren el análisis. Al aplicar los filtros se redujo la cantidad de individuos a 89 y la de tuits a 896.

Con los individuos que resultan de ambos filtros se procedió a calcular la proporción de cada sentimiento en el total de sentimientos encontrados en los tuits del individuo. Esto se hace con objetivo de mitigar el efecto de tener individuos con mayor cantidad de tuits analizados y sentimientos encontrados.

Utilizando las proporciones mencionadas como variables de agrupamiento se realizó el análisis. Cabe recalcar que, con el fin de lograr realizar el mejor agrupamiento posible para el conjunto de datos disponible, se utilizaron varios métodos de agrupación y distancias distintas de forma que se pudieron comprar los distintos resultados para elegir el más adecuado. Tomando en cuenta que las variables por las que se quería agrupar eran variables continuas y que existía correlación entre algunas de estas, se utilizaron las siguientes distancias entre grupos y entre individuos como parte del método de agrupamiento jerárquico aglomerativo:

- Distancias entre individuos:
    - Distancia Euclídea: $d_{ij} = \sqrt{(X_i - X_j)^T (X_i - X_j)}$
    - Distancia de Mahalanobis: $d_{ij}{}^M = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$

  Donde X_i es la matriz con los valores del individuo "i" y Ses la matriz de varianzas y covarianzas.

- Distancias entre grupos:
    - Distancia Ward: $\delta(A,B) = \frac{n_A n_B}{n_A + n_B} \left|| g_A - g_B \right||^2$
    - Vecino más cercano: $\delta(A,B) = min\{d(x_i, x_j);\ x_i \epsilon A, x_j \epsilon B\}$
    - Vecino más lejano: $\delta(A,B) = max\{d(x_i, x_j);\ x_i \epsilon A, x_j \epsilon B\}$
    - Salto promedio: $\delta(A,B) = \frac{1}{n_A n_B} \sum_{x_i \epsilon A, x_j \epsilon B} d(x_i, x_j)$

Donde $g_A$representa el centroide del grupo A y $n_A$la cantidad de individuos en el grupo A. Asimismo, se aplicó el método de k-medias utilizando la Suma de Cuadrados de Grupo como medida de comparación.

Posterior al análisis de agrupamiento se utilizaron otras variables no incluidas en este para así categorizar los grupos formados. Las variables utilizadas con este propósito fueron: sexo, edad, cantidad de seguidores, cantidad de "likes", años utilizando Twitter y provincia de residencia.

Para completar el análisis se utilizaron las siguientes librerías de R: biotools (Da Silva et al, 2017), ggplot2 (Wickham, 2016), car (Fox y Weisberg, 2011), readxl (Bryan y Wickham, 2019), Rmisc (Ryan Hope,2016), tm (Feinerer, Hornik y Meyer, 2019), qdap (Rinker, 2020), readr (Wickham, Hester y Francois, 2018), tidyr (Wickham y Henry, 2020), lubridate (Grolemund y Wickham, 2011), tidytext (Robinson, 2016), purrr (Henry y Wickham,

2020), dplyr (Wickham, Francois, Henry y Müller,2019), ggcorrplot(Kassambara,2019), cluster(Maechler, Rousseeuw,2019) gridExtra(Baptiste,2017) y factorextra(Kassambara y Mundt,2017).

RESULTADOS

Se logró recolectar la información de 89 usuarios de Twitter en Costa Rica, provenientes de cuatro provincias: San José (65%), Alajuela (7%), Heredia (15%) y Cartago (13%). Dentro de esta muestra se encuentran 64 hombres (72%) y 25 mujeres (28%). Las personas usuarias muestreadas poseen edades entre los 17 y los 68 años, cuentan con un rango que va desde los 2 a los 32 341 seguidores totales en la red social.

Con el fin de llevar a cabo el análisis de conglomerados, se emplean las variables construidas por medio del análisis sentimental: Furia, Asco o Repulsión, Miedo, Tristeza, Confianza, Sorpresa, Anticipación y Felicidad.

Previo al análisis respectivo, es necesario revisar la composición existente de los datos, por lo que se mide inicialmente la correlación entre todas las variables antes construidas. A partir de la figura 1 se puede comentar que a nivel general no se presentan relaciones entre variables con una magnitud muy alta, pero de igual manera la correlación entre variables se da tanto en dirección positiva como negativa.

Seguidamente se procede a revisar valores extremos, esto debido a que si se mantienen este tipo de valores, no solo se distorsiona la medida de posición (media) o de dispersión (varianza), sino que también se ven afectadas las correlaciones entre las variables (Morillas, A., & Díaz, B, s.f).Los valores extremos pueden ser observaciones que no sean representativas de ningún grupo de la población, así como también puede darse la posibilidad de una sub-representación de algún grupo. Por tal motivo, se realiza un diagnóstico de regresión mediante la prueba *Leverage* para identificar valores extremos o atípicos. Como se observa en la figura 2, hay valores por encima del límite, son 10 de los cuales 8 sobrepasan la línea y otros 2 se ubican prácticamente sobre esta. Estos 2 últimos casos dado su proximidad al límite se decide contemplarlos. De esta manera, el registro de los datos propios de análisis cuenta con 89 observaciones.

Al haber separado los valores extremos, se procede a revisar nuevamente las correlaciones de las variables. Como se aprecia en la figura 3 hay una correlación fuerte, positiva y lineal entre las variables furia y asco, así como también entre sorpresa y anticipación.

Se puede comentar que, para el análisis posterior, es necesario contemplar la correlación para selección del método de distancia más adecuado, debido a que la distancia de Manhattan asume que las variables no están correlacionadas. Por tal motivo, no resulta conveniente proceder el análisis utilizando este método.

Así mismo, se hace una revisión previa de la variabilidad que posee cada proporción del sentimiento, estas resultan ser bajas y muy similares entre sí, por lo que no es necesario realizar algún tipo de estandarización de las variables para agrupar.

Se procede a efectuar el análisis de agrupamiento, para ello se realiza inicialmente el cálculo de las distancias entre los distintos individuos, se utilizaron los métodos de distancias: Euclídea y Mahalanobis; para cada una de ellas se evaluaron varios métodos de agrupación jerárquica, correspondientes a: vecino más cercano, vecino más lejano, salto promedio y distancia Ward. La tabla 1 muestra las distintas combinaciones obtenidas entre las distancias entre individuos y entre grupos.

Tabla 1

Combinaciones obtenidas entre las diferentes distancias.

Distancia entre Individuos	Distancia entre grupos	¿Se visualiza conformación de grupos?
Euclídea	Vecino mas cercano	✗
	Vecino mas lejano	✗
	Salto promedio	✗
	Distancia Ward	✓
Mahalanobis	Vecino mas cercano	✗
	Vecino mas lejano	✗
	Salto promedio	✗
	Distancia Ward	✓

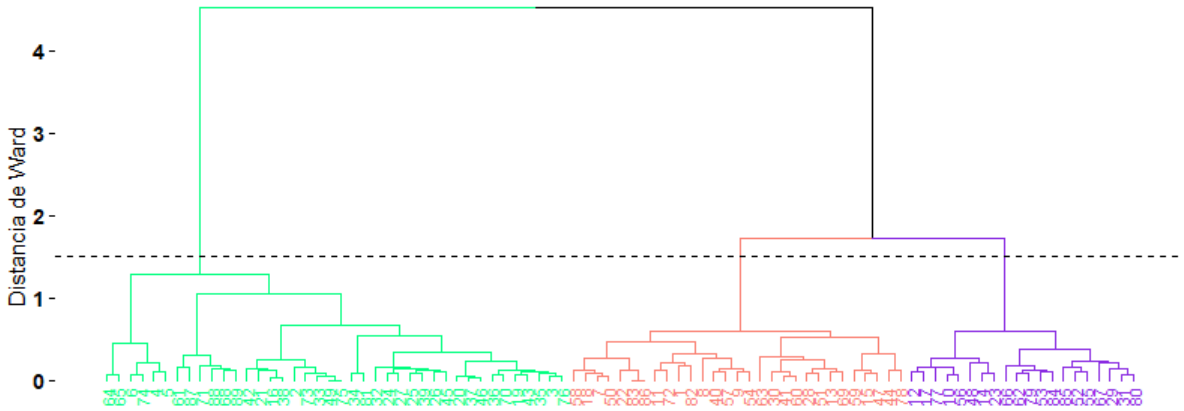
A partir de los resultados de la tabla 1, las combinaciones que arrojaron mejores resultados con respecto a la conformación de grupos fueron: método de Ward con distancia Euclídea y método de Ward con distancia Mahalanobis. Además, se empleó el método k-medias para evaluar la conformación de los grupos (Ver figura 4 en Anexos). En la combinación de método Ward con distancia Mahalanobis no fue posible observar una conformación apropiada de los grupos, debido a que no se logra identificar una distancia vertical considerable al momento de realizar cualquier corte, la división de los grupos resultantes no es clara, por lo que se descarta (Ver figura 5 en Anexos).

Finalmente se prefiere realizar el análisis con el método Ward y distancia Euclídea sobre el método k-medias, debido a que el primero proporciona una permanencia en las fusiones de los individuos en contraste con el segundo el cual tiende a realizar diferentes agrupaciones conforme se ejecuta. Las principales ventajas brindadas por usar el método Ward son: la formación de clústeres más compactos, con un tamaño similar y la minimización de la pérdida de información en el proceso de organización de los conglomerados (Cabello y Salama, 2012).

La figura 6 muestra los tres grupos conformados de acuerdo con el método escogido, la distribución sería la siguiente: 29 usuarios en el grupo 1, 40 usuarios en el grupo 2 y 20 en el grupo 3.

Figura 6

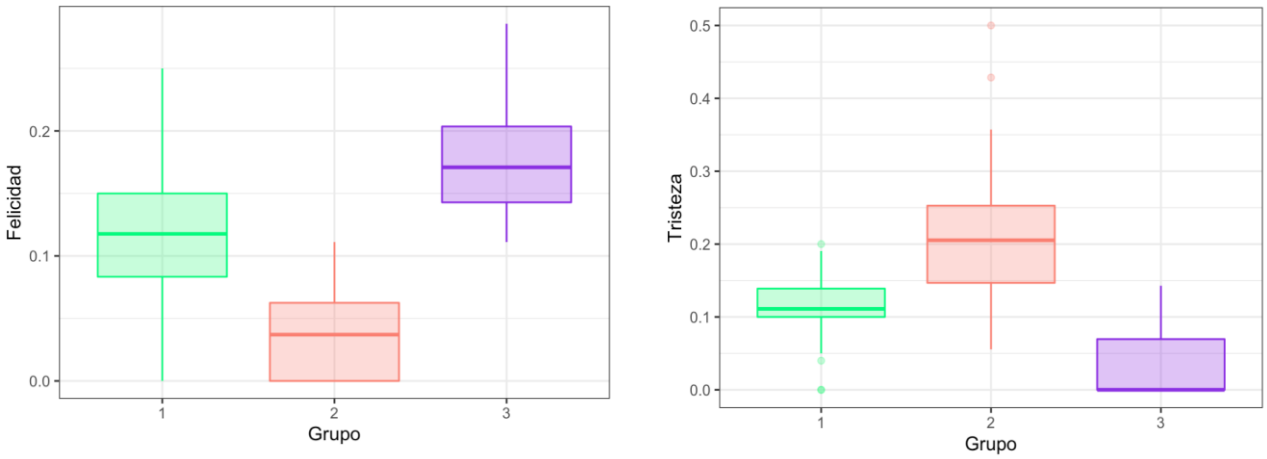
Dendograma de agrupamiento jerárquico según método de Ward y distancia Euclídea.



Una vez conformados los 3 grupos de la forma en la que se observó previamente en el dendograma, se hizo un análisis gráfico para determinar los sentimientos que tuvieron un mayor peso en la conformación de los grupos, para esto se utilizaron gráficos de cajas. En la figura 7 se logra observar que la separación de los grupos es clara para los sentimientos felicidad y tristeza. En el caso de felicidad se logra observar que el grupo 3 presenta una mayor proporción de este sentimiento, seguido por el grupo 1 y por último el grupo 2. En cuanto a la tristeza se observa que, como es de esperar, el comportamiento es opuesto, el grupo 2 es el que presenta mayor proporción de este sentimiento y el grupo 3 el que presenta menos.

Figura 7

Comparación de los grupos formados según los sentimientos de felicidad y tristeza.



No obstante, en el caso de confianza y miedo, cómo se logra observar en la figura 8, aunque la división no es tan clara, ya que algunas cajas se traslapan, igual se logra observar una división. Con respecto al sentimiento de confianza, aunque hay una variabilidad alta se puede identificar que el grupo 3 es el que presenta mayor

proporción y el grupo 2 el menor. En lo que concierne al sentimiento de miedo, el grupo 3 es el que presentó menor proporción y el 2 mayor.

Resulta pertinente mencionar el caso de los sentimientos asco y furia, donde el grupo 3 se separa por completo de los otros dos grupos, obteniendo, en su mayoría, una proporción sumamente baja. Por último, en el caso de sorpresa y anticipación se observa que no tuvieron mucho peso en la conformación de los grupos, puesto que no se logra observar mayor división de los grupos en cuanto a estos sentimientos, debido a esto se vuelve a realizar un análisis de agrupamiento sin tomar en cuenta estos sentimientos; los resultados de este nuevo análisis arrojan grupos levemente diferentes, sobre todo en el caso del grupo 1 y 2, sin embargo, se observa que la variabilidad de los distintos grupos aumenta en gran medida, por lo que se decide incluir estos sentimientos en el análisis aunque no sean decisivos en la conformación de los grupos.

Con el objetivo de corroborar la información descrita anteriormente, se realizó el procedimiento de componentes principales utilizando los sentimientos del análisis de agrupamiento. La figura 9 muestra gráficamente la comparación de los dos primeros componentes. Se logra observar que los individuos de los respectivos grupos se mantienen cerca y parecen formar los mismos grupos identificados anteriormente. Cabe recalcar que estos primeros dos componentes conjuntamente explican el 64% de la variabilidad original.

Seguidamente, se prosigue con la caracterización de los grupos según las variables mencionadas anteriormente en la metodología. Se hizo el análisis para lograr caracterizar los grupos según la edad promedio de los individuos que los conforman. Puntualmente se tiene que la edad promedio del grupo 1 es de 35.7 años, en el grupo 2 es de 29.8 y en el grupo 3 es de 36. Aunque pareciera que el grupo 3 es el de mayor edad, como se logra ver en la figura 10, la diferencia entre las edades promedio de los grupos no es relevante, ya que los intervalos de confianza se traslapan en gran medida, por lo tanto, utilizando la variable edad no se logró una caracterización clara de los grupos.

En el caso de la variable provincia mencionada entre las variables de caracterización previamente, la figura 11 muestra la forma en la que se conforman los tres grupos según las provincias donde residen los individuos, esto mediante un gráfico de barra 100%. Entre los usuarios extraídos de Twitter Costa Rica solo se obtuvieron personas que residen en las provincias Alajuela, Cartago, Heredia o San José, por lo que el gráfico solo muestra estas 4. Con esta comparación no se logran observar diferencias grandes entre grupos ya que se observa una distribución similar entre todos, con San José como la provincia donde residen la mayoría de las personas usuarias de todos los grupos. Sin embargo, se puede mencionar que el grupo 1 es el que está conformado por mayor proporción de personas que viven en San José mientras que el grupo 2 es en el que se clasificaron mayor proporción de individuos de Cartago y Heredia.

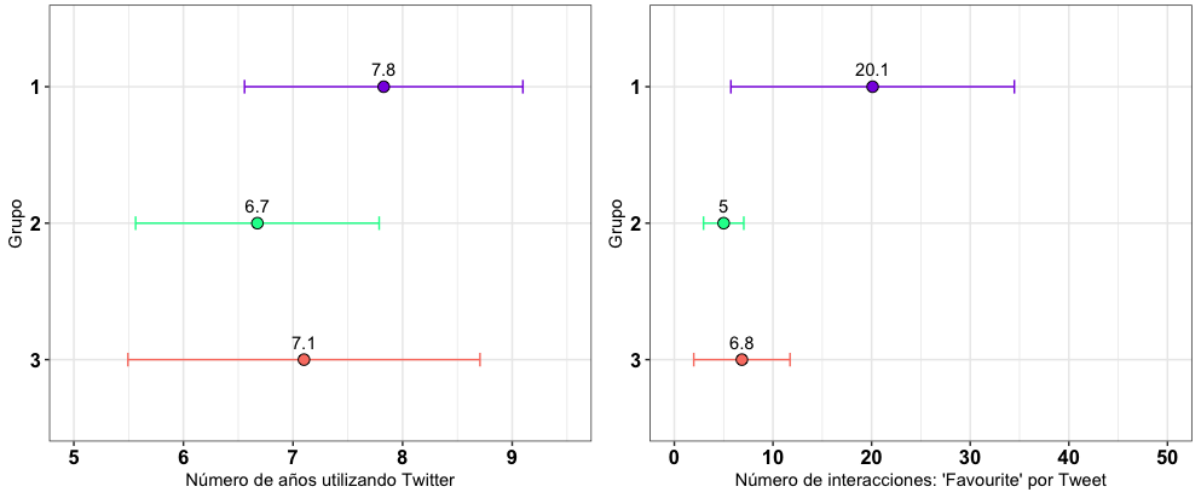
Para efectuar la comparación por sexo es necesario verificarlo por medio de la proporción de personas usuarias contempladas en los tres grupos elaborados previamente. En general, como se muestra en la figura 12, se puede comentar que se presentan más hombres, de los cuales en cada uno de los grupos acumulan más del 65 % de los usuarios. La mayor diferencia se da en el tercer grupo, ya que lo conforman 17 hombres y 3 mujeres. En este caso se logra mostrar una caracterización notoria, siendo los hombres lo que mayoritariamente están conformando cada uno de los grupos.

Por otro lado, se utilizaron las variables proporcionadas por Twitter para la caracterización. En la figura 13, es posible observar la comparación de los grupos según el número de años que los usuarios posean utilizando

la red social y de acuerdo con el número de interacciones “Favourite” recibidas por tweet. De acuerdo con el análisis gráfico es posible observar que no hay diferencias claras entre los grupos según estas variables; en la variable interacciones por tweet se observa que el grupo 1 presenta un número promedio de interacciones mayor en comparación con los demás grupos, a pesar de esto presenta mucha variabilidad, lo cual se puede deber a que, en la conformación de los grupos, este quedó como un grupo intermedio sin polarización de sentimientos. Para el caso del número de años utilizando Twitter, la variabilidad dentro de cada grupo es prácticamente la misma, además no se observa diferencia clara entre los grupos.

Figura 13

Intervalos de confianza del número de años utilizando Twitter e interacciones “Favourite” que posee cada grupo.



Otra de las variables que Twitter pone a disposición, es la cantidad de seguidores que tiene el usuario al momento en que se realiza la publicación respectiva en la red social. Precisamente con esto, se pretende observar el comportamiento de los grupos según el rango en el cual cada uno se encuentra y, con ello verificar si existen ciertas diferencias. Con base en la figura 14, el grupo 1 presenta el mayor rango de variación, aunado a que en promedio es el grupo con mayor cantidad de seguidores, contrario al caso del grupo 3 que es el que se percibe tener menor variabilidad dentro del grupo y el que en promedio presenta la menor cantidad de seguidores. Aun así, los tres intervalos de confianza se traslapan, por lo que da la impresión de que no deben hallarse diferencias por grupo dada la cantidad de seguidores. Según lo previsto, dicha variable no permite brindar una caracterización adecuada a los grupos.

CONCLUSIONES

Se realiza un análisis de agrupamiento jerárquico en el que, dado la naturaleza de los datos, se adecua de mejor manera el método de Ward con distancia de Euclídea. A partir de esta combinación resultan tres grupos, los cuales van a ser descritos de acuerdo con la proximidad con los diversos sentimientos, así como aquellas variables que fueron oportunas para caracterizar a los respectivos usuarios.

El primer grupo, con una edad promedio de 35.7 años y en su gran mayoría compuesto por personas usuarias procedentes de la provincia de San José (siendo el que posee mayor representación de esta provincia) tuvo como principal característica el ser un grupo intermedio. No es posible catalogarlo como un grupo con

sentimientos negativos ni como un grupo con sentimientos positivos, ya que posee un poco de ambos. Con respecto a las variables asociadas a la red social, este fue el que presentó mayor cantidad de seguidores y por ende mayor cantidad de interacciones, a pesar de esto no es posible asegurar que existan diferencias claras respecto a los demás grupos al compararlos según estas variables.

En cuanto al grupo 2 este fue conformado por 29.8 usuarios y en primer lugar se encontró que, en cuanto a los sentimientos utilizados para la agrupación, fue el grupo con menor proporción de felicidad y confianza expresada mediante sus tweets y consecuentemente, el grupo con mayor proporción de tristeza y miedo. Es decir, este grupo fue el que obtuvo mayor proporción de sentimientos negativos. Respecto a las variables de caracterización, se identifica que la edad promedio de los individuos de este grupo es de 36 años, y en comparación a los otros grupos, tiene una mayor proporción de individuos que residen en Heredia y Cartago. Lo que respecta a las variables cantidad de interacciones y seguidores fue el intermedio en ambas.

Respecto al grupo 3, en el análisis de sentimientos se destaca que es el grupo que posee una mayor proporción en el sentimiento de felicidad y confianza. Dado esta situación, es contrario a lo mostrado al grupo antes comentado, ya que el grupo al que pertenecen es el que presenta menor proporción en miedo y tristeza. Por tal motivo, de manera general se puede comentar que, de los 3 grupos elaborados, los usuarios del grupo 3 son aquellos que mediante los diferentes tweets describen más emociones positivas, específicamente de felicidad y confianza. En cuanto a su posible caracterización, dicho grupo posee una edad promedio de 33 años.

En el caso de las variables provincia de residencia, sexo y edad no se logran identificar mayores diferencias entre grupos. Se encuentra que, en todos los grupos, la mayoría de los usuarios residen en San José, son hombres y son jóvenes.

El período de análisis se establece precisamente, en el proceso de confinamiento que Costa Rica enfrenta debido a la enfermedad viral COVID-19. El poco contacto con el exterior y la poca interacción social parten como hechos concretos que implican inestabilidad emocional, situación que es valorada en los usuarios gracias al análisis de sentimientos de los tuits que han realizado.

Lo esperable y según lo comenta Quezada (2020) es que, bajo estas circunstancias, se desarrollen psicopatologías y otros aspectos negativos para la salud mental, lo cual, previendo los propósitos de la investigación, es probable encontrar una cantidad considerable de mensajes con un carácter negativo. Si se observa los grupos establecidos, el grupo 2 es aquel que tiende a presentar mayor relación con lo que se propone, puesto que son los usuarios que obtuvieron mayor proporción en sentimientos negativos, específicamente miedo y tristeza.

Por otra parte, es necesario comentar ciertas limitantes dentro del proceso y elaboración del análisis. Fundamentalmente los inconvenientes encontrados se dan a partir de la extracción de los datos proporcionados desde el API de Twitter, debido a que dentro de la recolección se deben especificar algunas características de acuerdo con el enfoque que se pretende y una de ellas es precisamente el periodo en el cual se dispone. Se tuvo que limitar el análisis a una sola semana, fue necesario modificar la propuesta inicial del estudio, ya que se contaba con tener un periodo más extenso. La situación comentada, se prevé que es debido a que solo se puede obtener información de una semana antes al día que se realiza la solicitud, probablemente producto de la gran cantidad de publicaciones que se efectúan diariamente.

Otra de las limitaciones presentadas, fue la obtención de características demográficas de los usuarios, como es el caso de la edad y sexo, la base de datos que se genera a partir del API de Twitter no brinda este tipo de características, por lo que requirió de una revisión manual por medio de las cuentas de estos usuarios. Esto requirió de mayor tiempo al que se había planeado en cuanto a la recolección de datos.

A raíz de estas limitantes, se pueden tener posibilidades para ciertas recomendaciones. Una de ellas se podría enfocar en generar un análisis de sentimientos con un periodo más extenso, con ello permite tener una mayor descripción de los usuarios. Observar si mediante las publicaciones realizadas, se perciben sentimientos que han sido más constantes y que, por lo tanto, pueden brindar más información del estado emocional de la persona en el lapso específico. Otra posibilidad recae en conformar un instrumento de medición cualitativa, esto puede ser mediante un cuestionario, en el que se pueda delimitar de mejor manera la información a recopilar, es decir, elaborar un instrumento que permita definir preguntas direccionadas a los sentimientos contemplados en el estudio.

BIBLIOGRAFÍA

Alboukadel Kassambara (2019). ggcorrplot:Visualization of a Correlation Matrix using 'ggplot2'.R package version 0.1.3.https://CRAN.R-project.org/package=ggcorrplot

Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Cabello, A. G., & Salama, A. (2012). Un estudio sobre la distribución regional de los préstamos en la Argentina por sector económico, 2000-2012. Una aplicación del análisis de cluster. *Analítika: revista de análisis estadístico*, (3), 43-57.

Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). *Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40(16), 6266–6282.* doi: 10.1016/j.eswa.2013.05.057

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Hadley Wickham and Lionel Henry (2020). tidyr: Tidy Messy Data. R package version 1.0.2. https://CRAN.R-project.org/package=tidyr

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

Ingo Feinerer and Kurt Hornik (2019). tm: Text Mining Package. Rpackage version 0.7-7. https://CRAN.R-project.org/package=tm

Kearney, M. W. (2019). rtweet: Collecting and analyzing Twitter data, Journal of Open Source Software, 4, 42. 1829. doi:10.21105/joss.01829 (R package version 0.7.0)

Lionel Henry and Hadley Wickham (2020). purrr: Functional Programming Tools. R package version 0.3.4. https://CRAN.R-project.org/package=purrr

Liu, B. (2010). Sentiment Analysis and Subjectivity. Department of Computer Science, University of Illinois at Chicago. Recuperado de https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0. https://cran.r-project.org/web/packages/ClusterR/index.html

Merino, M. (2014). ¿Qué es una API y para qué sirve? TICBeat. Recuperado de https://www.ticbeat.com/tecnologias/que-es-una-api-para-que-sirve/

Morillas, A., & Díaz, B. (s.f). EL PROBLEMA DE LOS OUTLIERS MULTIVARIANTES EN EL ANÁLISIS DE SECTORES CLAVE Y CLUSTER INDUSTRIAL. Recuperado de https://www.researchgate.net/publication/267415393_EL_PROBLEMA_DE_LOS_OUTLIERS_MULTIVARIANTES_EN_EL_ANALISIS_DE_SECTORES_CLAVE_Y_CLUSTER_INDUSTRIAL

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Department of Computer Science, Cornell University. Recuperado de http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

Quezada, V. (2020). Miedo y psicopatología la amenaza que oculta el Covid-19. Panamerican Journal of Neuropsychology. Recuperado de http://www.cnps.cl/index.php/cnps/article/view/394

R Core Team (2020). R: A language and environment for statistical computing. R

Rinker, T. W. (2020). qdap: Quantitative Discourse Analysis Package. 2.4.1. Buffalo, New York. http://github.com/trinker/qdap

Ryan M.Hope(2016). Rmisc: Ryan Miscellaneous. R package version 1.5 https://cran.r-project.org/package=Rmisc

Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

ANEXOS

Figura 1

Correlación de los sentimientos.

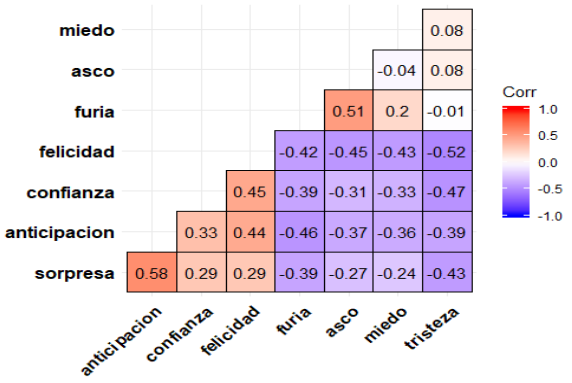


Figura 2

Identificar valores extremos.

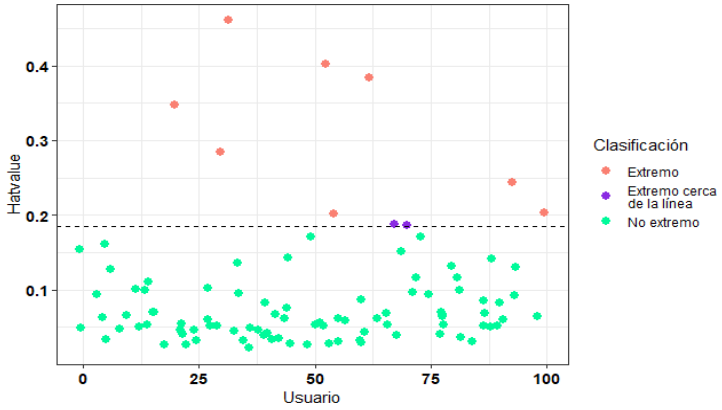


Figura 3

Correlación de los sentimientos quitando los valores extremos.

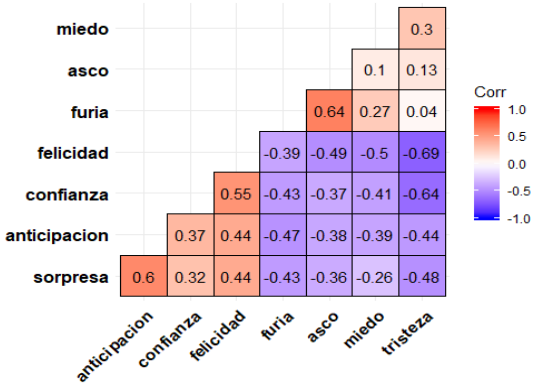


Figura 4

Comparación de distintos grupos por medio de la Suma de Cuadrados dentro de Grupo con el método de k-medias.

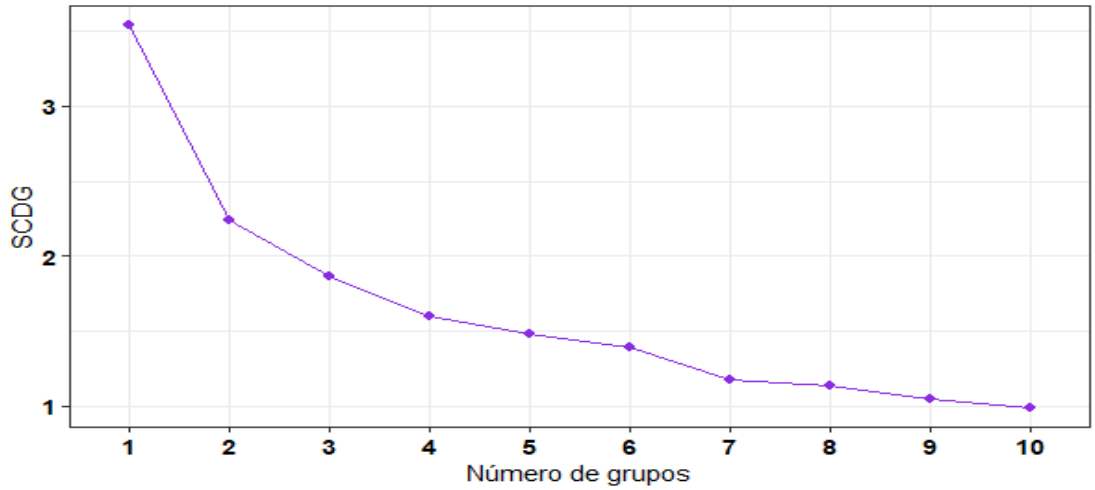


Figura 5

Dendograma de agrupamiento jerárquico según método de Ward y distancia Mahalanobis.

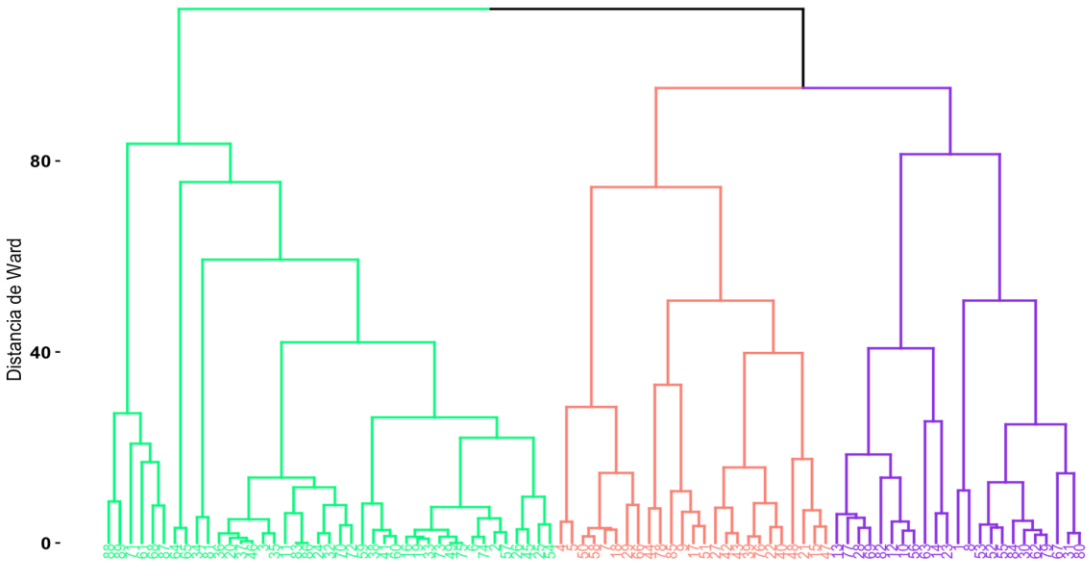


Figura 8

Comparación de los grupos formados según los sentimientos de confianza, miedo, asco, furia, anticipación y sorpresa.

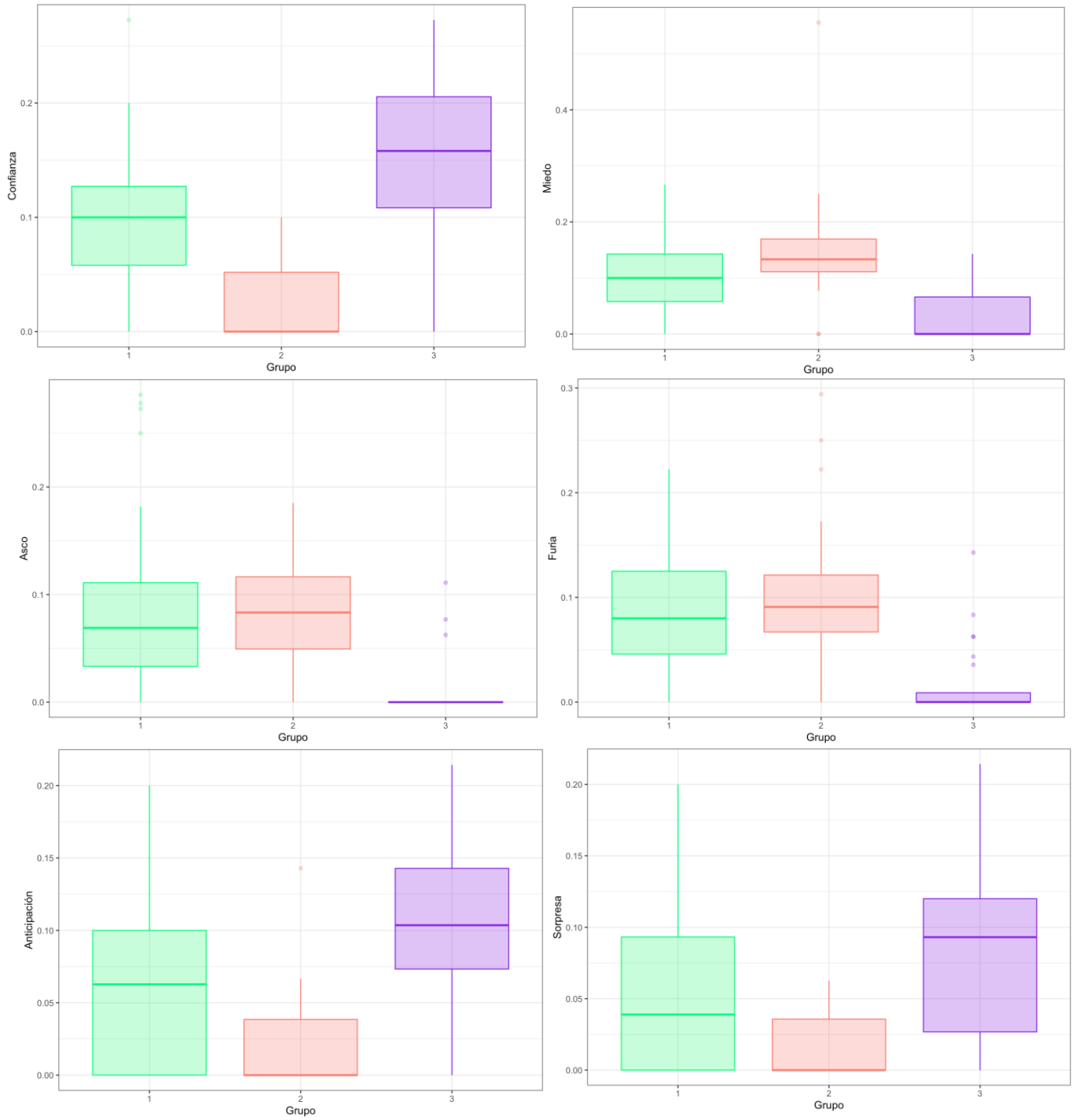


Figura 9

Puntajes del primer y segundo componente principal y agrupados según el método de Ward con distancia Euclídea.

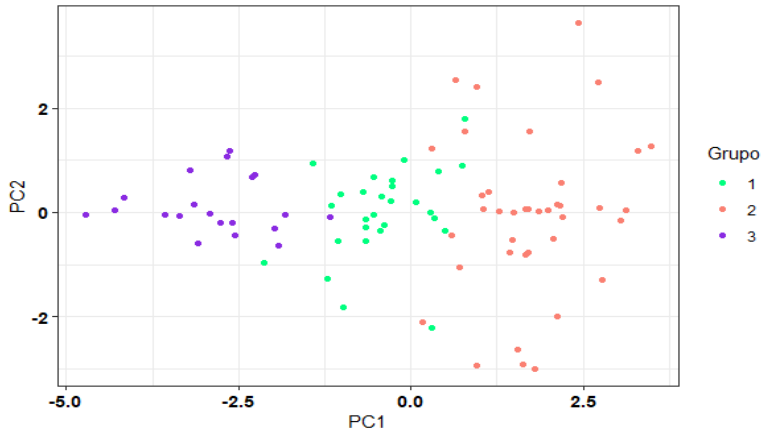


Figura 10

Edad promedio de los grupos.

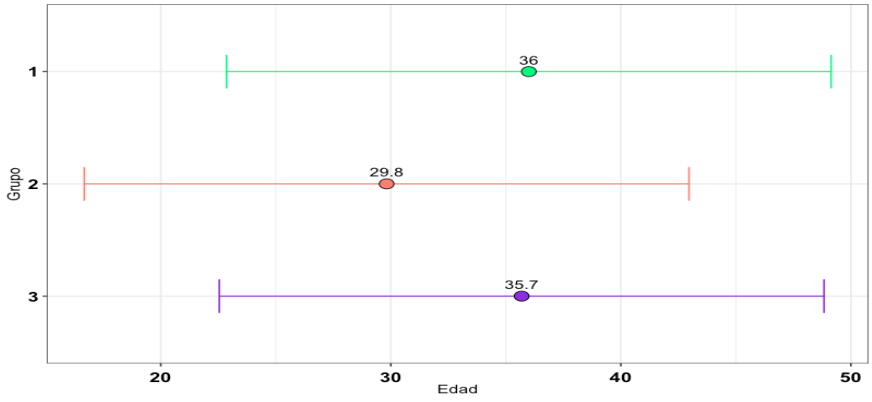


Figura 11

Caracterización de los grupos según provincia.

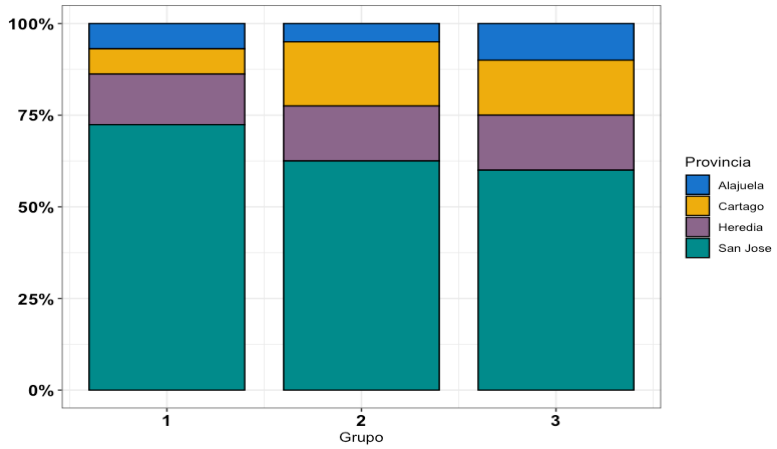


Figura 12

Caracterización de los grupos según sexo.

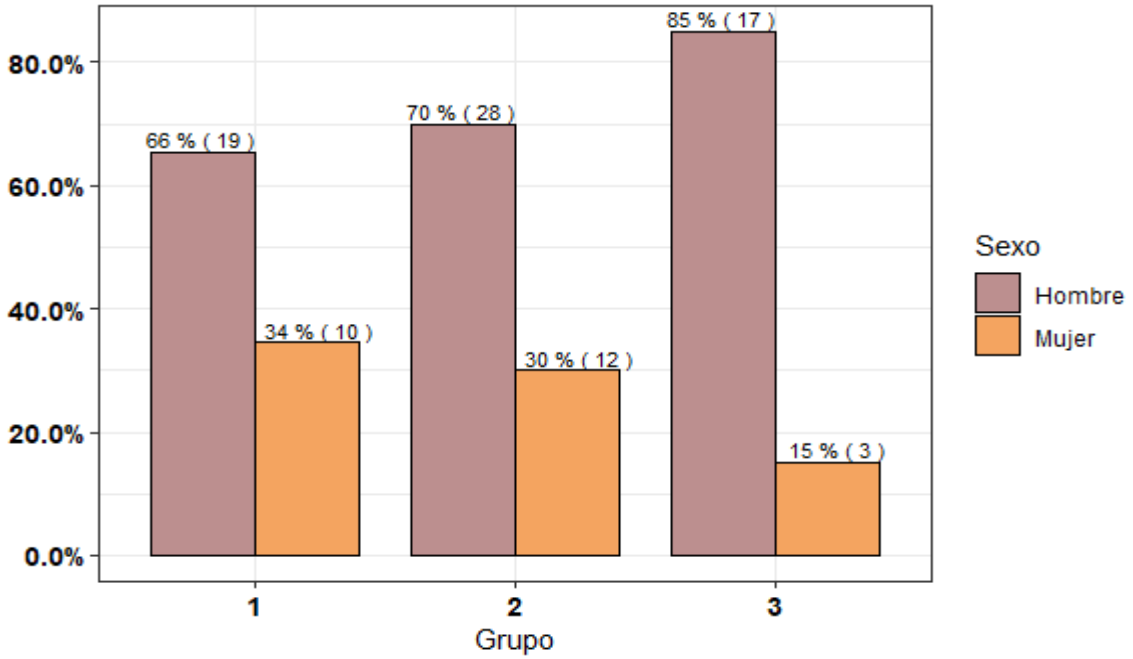
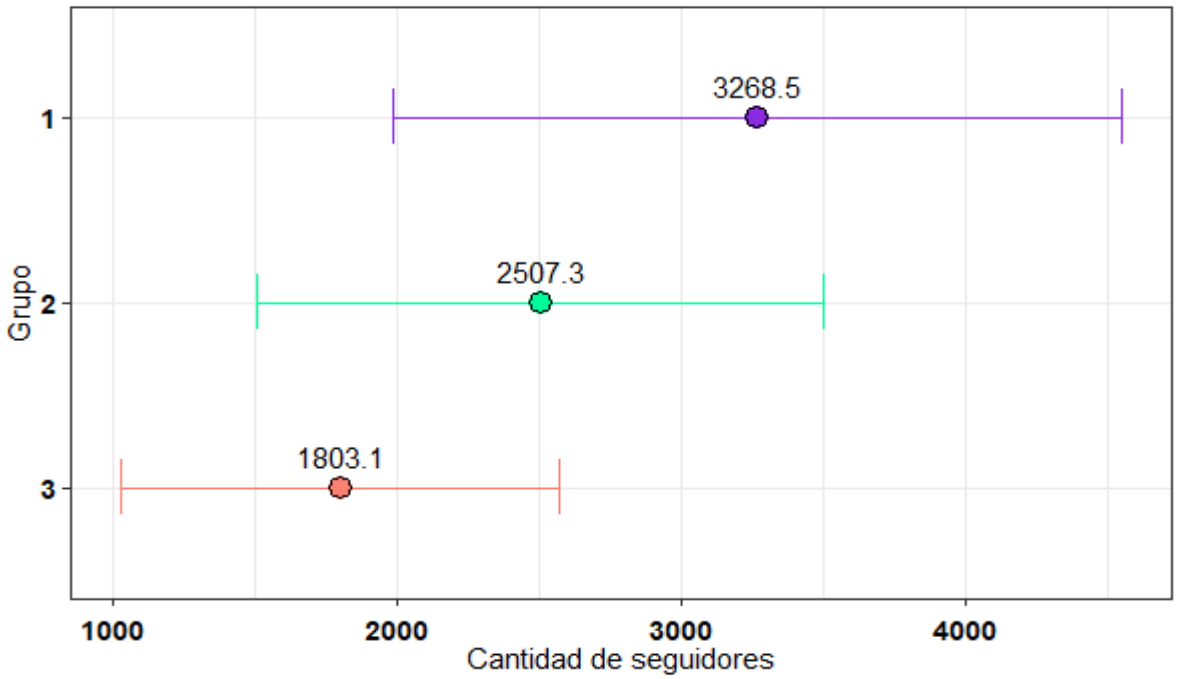


Figura 14

Intervalos de confianza de la cantidad de seguidores que posee cada grupo



Contraste del perfil musical de estudiantes y egresados de Estadística con el perfil musical de diferentes géneros musicales a partir de variables presentes en la plataforma Spotify

Juan José Jaikel Jiménez[10], José Pablo Aguilar Umaña[10], Fernando Alvarado Prado[10]

Juan.jaikel@ucr.ac.cr, jpaguma1995@gmail.com, feral1498@gmail.com

RESUMEN

La música como parte del día a día es un componente fundamental en la vida de las personas. Elementos importantes de las pistas musicales como la Bailabilidad, la cual representa qué tan bailable es una pista, la Energía, donde se toma como medida perceptiva de intensidad y actividad, el habla, que detecta la presencia de palabras habladas en una pista, entre otras, son útiles para definir un perfil musical, en el cual se ven determinados los gustos y preferencias de las personas. En el presente estudio, se propone describir a las personas según su gusto musical utilizando características de las canciones según la plataforma Spotify. Además, se busca definir si las preferencias son determinadas por uno o más géneros musicales. Para desarrollar los, se opta por un análisis multivariado de conglomerados a través de las distancias Euclídeas para los grupos. Se propone el concepto de "perfil musical" como medio para aproximar el gusto musical de las personas y como medida objetiva para describir las canciones. A partir del perfil musical se caracteriza a los individuos en cinco grupos. Además, se encuentra que cada género contemplado en la investigación presenta un perfil musical diferente entre sí. Por último, se encuentra relación entre los perfiles musicales de los grupos y los perfiles musicales de los géneros.

PALABRAS CLAVE: Bailabilidad, Valencia, Instrumentalidad, Acústica, Energía, Habla, Perfil Musical, Conglomerados.

ABSTRACT

Music as a daily activity is a fundamental component in people's lives. Important elements of musical tracks like Danceability, which represents how danceable a track is, Energy, which is used to measure track intensity, Speechiness, which detects the presence of spoken words in a track, among others, are useful in defining a musical profile that represents peoples' musical tastes and preferences. In the present study, we propose a way of describing people according to their musical taste, using song characteristics that are provided by Spotify. Additionally, we look to define if these preferences are determined by one or more musical genres. In order to achieve these objectives, we opt for Conglomerate Multivariate Analysis using Euclidean distances between groups. We also present the concept, "musical profile", in order to approximate peoples' musical tastes and as a structured way to describe songs. Using musical profiles, we characterize individuals by placing them into five clusters. Also, we find that every musical genre considered in this study presents a different musical profile. Finally, we find an association between the musical profile of the clusters and those of the genres.

KEY WORDS: *Danceability, Valence, Instrumentality, Acousticness, Energy, Speechiness, Musical Profile, Clusters*

[10] Estudiantes de Estadística de la Universidad de Costa Rica

INTRODUCCIÓN

La constante presencia de la música en la vida de las personas, se relaciona estrechamente con la mayoría de las situaciones que se realizan durante el día a día, lo que lleva a desarrollar un rasgo de identidad asociado con la sobreexposición a la música, principalmente los adultos jóvenes. (Sánchez, 2005).

Los jóvenes utilizan la música para relacionarse con los demás, socialmente hablando, entre otras actividades, de manera privada, la música es utilizada como acompañamiento, entre otros. La música funciona muy eficientemente para determinar posiciones sociales, tanto en jóvenes como en otros grupos de edad, aunque particularmente en los jóvenes por el tiempo flexible y amplio que poseen.

En cuanto a la definición de música, existen varias de estas, Castro (2003) recomienda seleccionar aquella que encierra una idea lo más completa posible acerca de esta rama del arte. Algunas definiciones de la música son: “Es el arte de los sonidos.”, “Es la expresión sonora de la belleza.” y “Es el arte de combinar sonidos de un modo agradable al oído.” (Castro, 2003), de manera que esta última se establece como la definición de la música para fines del estudio.

La música toma como herramientas el sonido, el silencio y el ruido para expresar sentimientos, pensamientos y funciona como un medio para transmitir algo. La música no tiene sentido para los seres humanos si estos no la experimentan como tal, señala Castro (2003). Cuanto mayor sea el conocimiento que se posee sobre la música, más completa será la concepción que se tenga de ella. (Castro, 2003)

Respecto al gusto musical de las personas, Gasser (2019) indica que se empieza a formar desde los primeros años de vida. La exposición a diferentes piezas musicales en una etapa temprana de la vida influencia lo que consideramos familiar. A su vez, de manera general, esta familiaridad con la música influencia nuestra disposición a exponernos a nuevas piezas de música a lo largo del tiempo y, según Chamorro-Premuzic (2011), esta se relaciona con nuestras personalidades por lo que el gusto musical no se puede considerar aleatorio.

Por otro lado, los gustos musicales se entrelazan con nuestras relaciones sociales y nuestra cultura social. Nuestra cercanía con otras personas y la formación de grupos sociales tienen una relación con nuestros gustos musicales. (Madsen, Hellmuth Margulis, Simchy-Gross, & Parra, 2019. Bakagiannis & Tarrant, 2006). Inclusive, Hou, Song, Hu, Pan, y Hu (2020) han encontrado que la actividad cerebral entre escuchantes de una pieza musical se sincroniza, en especial cuando estos disfrutan la música escuchada, de lo cual se confirma la socialización que ocurre a través de la música y cómo los gustos musicales unen a las personas.

A partir de lo anterior, interesa conocer si esta socialización se relaciona con los géneros musicales. El análisis de la música mediante los géneros ha sido el foco de mucho debate entre musicólogos. Existen estudios que han analizado el rol de los géneros en los gustos musicales (Eijck, 2001. Istók, Brattico, Jacobsen, Ritter, y Tervaniemi, 2013) como aquellos que han sostenido que los géneros no representan la única característica que define los gustos musicales de las personas (Chamorro-Premuzic, 2011. Wassenberg, 2019).

Hoy en día encontramos la música disponible en diferentes plataformas digitales de acceso público o privado (bajo subscripción), entre las cuales se encuentra Spotify, estas plataformas comparten lo que se conoce como “música en streaming”, que se refiere a la transmisión de música digital. Spotify es una aplicación web y de dispositivos móviles que cuenta con más de 207 millones de usuarios, la cual ofrece programas gratuitos y de subscripción para ser parte del envolvente mundo musical.

Esta plataforma cuenta con más de 40 millones de canciones, y clasifica sus canciones por artista, por género, popularidad, gustos del usuario, entre otros. Además, ofrece variables que describen características de audio de cada canción presente en su plataforma. Dichas variables se encuentran disponibles al público, se utilizan en la presente investigación como proxy de la definición de música. (Spotify AB, 2020).

Dado que las variables permiten caracterizar canciones tomando como referencia elementos de sonido, se considera que al analizar la combinación de ellas se puede aproximar la forma en que resulta más agradable para las personas la combinación de diferentes sonidos. De ahora en adelante se define como perfil musical al conjunto de los promedios de las variables a contemplar dentro de la investigación para una misma canción, persona, grupo de personas o género musical. Como primer objetivo, interesa caracterizar a los estudiantes y egresados de la carrera de Estadística según las preferencias en la música que escuchan.

Una vez definidos los grupos de personas, interesa también contrastar el perfil musical de los grupos con el perfil de géneros musicales, y determinar si cada grupo de personas presenta similitudes con uno o más géneros. Por ende, como segundo objetivo, se busca definir si dichas preferencias son determinadas por uno o más géneros musicales.

METODOLOGÍA

Descripción de las variables que determinan el perfil musical

Spotify pone a disposición de las personas un sitio denominado: "Spotify plataforma para desarrolladores". En este sitio se encuentran variables que son definidas y calculadas por dicha empresa con el fin de facilitar el aprender y realizar análisis acerca de las características de las canciones y otros aspectos (Spotify AB, 2020).

Para definir el perfil musical de las personas y los géneros se analizan las variables que, según Spotify, permiten caracterizar las canciones dadas las siguientes dimensiones: "Sentimiento", "Propiedades musicales" y "Contexto". Las variables a considerar se presentan en la tabla 1.

Tabla 1

Descripción de las variables que se pueden contemplar para caracterizar el perfil musical de las canciones, personas y géneros.

Variable	Escala	Descripción	Dimensión
Bailabilidad	0.00 a 1.00	Describe qué tan adecuada es una pista para bailar. Basada en una combinación de elementos musicales que incluyen tempo, estabilidad de ritmo, fuerza de ritmo y regularidad general. Un valor de 0.0 es menos bailable y 1.0 es más bailable.	Sentimiento

Energía	0.00 a 1.00		Representa una medida perceptiva de intensidad y actividad. Por lo general, las pistas energéticas se sienten rápidas, ruidosas y ruidosas. Por ejemplo, el death metal tiene alta energía, mientras que un preludio de Bach tiene puntajes bajos en la escala. Las características perceptivas que contribuyen a este atributo incluyen rango dinámico, volumen percibido, timbre, frecuencia de inicio y entropía general.	Sentimiento
Habla	0.00 a 1.00		Detecta la presencia de palabras habladas en una pista. Cuanto más exclusiva sea la grabación (como, por ejemplo, talk show, audiolibro, poesía), más cercano a 1.0 será el valor del atributo. Los valores superiores a 0,66 describen pistas que probablemente están hechas completamente de palabras habladas. Los valores entre 0.33 y 0.66 describen pistas que pueden contener música y discurso, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores por debajo de 0,33 probablemente representan música y otras pistas que no son del habla.	Sentimiento
Valencia	0.00 a 1.00		Describe la positividad musical transmitida por una pista. Las pistas con alta valencia suenan más positivas (por ejemplo, feliz, alegre, eufórica), mientras que las pistas con baja valencia suenan más negativas (por ejemplo, triste, deprimido, enojado).	Sentimiento
Instrumen-talidad	0.00 a 1.00		Predice si una pista no contiene voces. Los sonidos "Ooh" y "aah" se tratan como instrumentales en este contexto. Las pistas de rap o palabras habladas son claramente "vocales".Cuanto más cercano sea el valor de instrumentalidad a 1.0, mayor será la probabilidad de que la pista no contenga contenido vocal. Los valores superiores a 0.5 están destinados a representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1.0. Los valores superiores a 0.5 están destinados a representar pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1.0.	Propiedades musicales
Acústica	0.00 a 1.00		Medida de confianza acerca de si la pista es acústica. 1.0 representa una alta confianza de que la pista es acústica.	Contexto
Tempo	0.00 a ∞		El tempo general estimado de una pista en latidos por minuto (BPM). En terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y se	Propiedades musicales

		deriva directamente de la duración promedio del tiempo.	
Volumen	-∞ a ∞	El volumen general de una pista en decibelios (dB). Los valores de volumen se promedian en toda la pista y son útiles para comparar el volumen relativo de las pistas. Rango típico de valores entre -60 y 0 db.	Propiedades musicales
Duración	0.00 a ∞	La duración de la pista en milisegundos.	Contexto

Recolección de los datos

Con el fin de definir el perfil musical de estudiantes y egresados de la Escuela de estadística se procede a diseñar un cuestionario mediante la plataforma Google Forms, el cual fue difundido a través de redes sociales a personas estudiantes y egresadas de la escuela de Estadística de la Universidad de Costa Rica, el cual se observa en el anexo 1. El cuestionario permite recolectar las cinco canciones favoritas de dichas personas durante la última semana antes del momento de llenar el cuestionario. Además, se pregunta por aspectos sociodemográficos y de actividades para las cuales escuchó música en dicha semana, con el fin de caracterizar mejor los posibles agrupamientos resultantes bajo el perfil musical.

La participación total de manera voluntaria se atribuye a 60 observaciones y a través de un proceso de depuración de la información obtenida, se obtuvo un total de 54 observaciones, las cuales se utilizan en el estudio. La disminución en el tamaño de la muestra se debe a que se encuentran duplicados entre las observaciones, por lo que se procedió a eliminarlos.

Perfil musical de las personas

Una vez obtenidas las cinco canciones favoritas durante la última semana antes de llenar el cuestionario para los 54 participantes, se procede a extraer las variables descritas en la Tabla 1 mediante el paquete spotifyr (Thompson, Parry, Phipps, & Wolff, 2019) del software R (R Core Team, 2013). Con las variables se realiza un análisis descriptivo en donde se considera la variabilidad por persona y su correlación, con el fin de determinar cuáles son más útiles para conformar los grupos de personas de acuerdo a su perfil musical.

El analizar la variabilidad por persona permite evaluar qué tan útil es cada variable para definir los agrupamientos, pues es deseable considerar aquellas en donde la variabilidad por persona es poca. Esto indica que las variables son consistentes por persona, lo cual ayuda a obtener grupos de personas homogéneos a nivel interno y heterogéneos entre sí. Se consideró la correlación entre las variables para evitar tener aquellas que estén altamente correlacionadas, debido a que explicarían el mismo concepto y el peso de este concepto sería mayor para la conformación de los grupos.

Luego de analizar la variabilidad por persona y la correlación entre variables se procede a definir el perfil musical de cada persona. Para esto, se promedian las variables seleccionadas presentes en las cinco canciones que cada persona propone, y así obtener un registro único (el promedio) para cada variable por persona. De

esta manera, se define el perfil musical de las personas como el conjunto de promedios obtenidos correspondientes a cada variable.

Conformación de los grupos de personas

A partir de la definición de los perfiles musicales de las personas, se procede a definir las agrupaciones de las personas. Para lo anterior, se utiliza el método de distancias euclídeas, ya que todas las variables son continuas y permite definir una medida de distancia multidimensional entre dos unidades. Para las distancias entre grupos, se usa el método del dendrograma mediante "el vecino más cercano", "el vecino más lejano", "el salto promedio" y "la distancia de Ward". Las fórmulas matemáticas asociadas a los métodos mencionados se detallan en el anexo 2.

El vecino más cercano consiste en comparar las distancias entre el objeto de un grupo que está más cercano al de otro grupo, para todos los grupos. Paralelamente, el vecino más lejano representa la metodología opuesta: usando las distancias más lejanas entre los datos de los grupos. Por otro lado, el salto promedio indica la distancia promedio entre cada uno de los objetos de un grupo con el de otro grupo y la distancia de Ward compara los centroides de los grupos entre sí.

El dendrograma permite observar la agrupación, tanto de los datos como los grupos de ellos, este método permite formar criterios de agrupación a partir de subdivisiones visualizadas gráficamente. Se toma como punto para la partición, cambios en la altura que se consideren grandes, trazando una línea horizontal por dicho punto. Esto permite determinar la cantidad de grupos de personas, a partir de su perfil musical detallado anteriormente. Debido a que las agrupaciones se hicieron a partir del perfil musical de las personas, se procede a calcular los centroides de los grupos y estos se consideran como el perfil musical del grupo. Una vez definidos estos perfiles, es importante analizar la distribución de las variables por grupo, lo cual se realiza mediante gráficos de caja.

Perfil musical de los géneros

A partir de los resultados del cuestionario, se decide trabajar con siete géneros musicales, tomando en consideración cuántas veces aparecieron entre las canciones brindadas en el cuestionario, y verificando que estos géneros sean distintivos entre sí. Desde Spotify se extrajo, mediante el uso de spotifyr (Thompson, Parry, Phipps, & Wolff, 2019), una muestra de 100 canciones que son consideradas parte de cada género por la plataforma. Para cada una de las 100 canciones se obtienen las variables descritas en la tabla 1.

Para cada género se promedian las variables de las 100 canciones, de modo que por género se tiene como medida resumen el promedio para cada una (el perfil musical por género). De nuevo, resulta importante analizar la distribución de cada variable por género y para esto se plantea un análisis mediante gráficos de cajas.

Comparación de los grupos con características sociodemográficas y con los géneros

Por otro lado, se comparan los perfiles musicales de los grupos con las variables sociodemográficas consideradas en el cuestionario. Esto se hace con el fin de determinar si las personas tienen características sociodemográficas que son homogéneas por grupo y heterogéneas entre grupos. El análisis de lo anterior se realiza a través de tablas cruzadas entre las variables sociodemográficas y los grupos conformados.

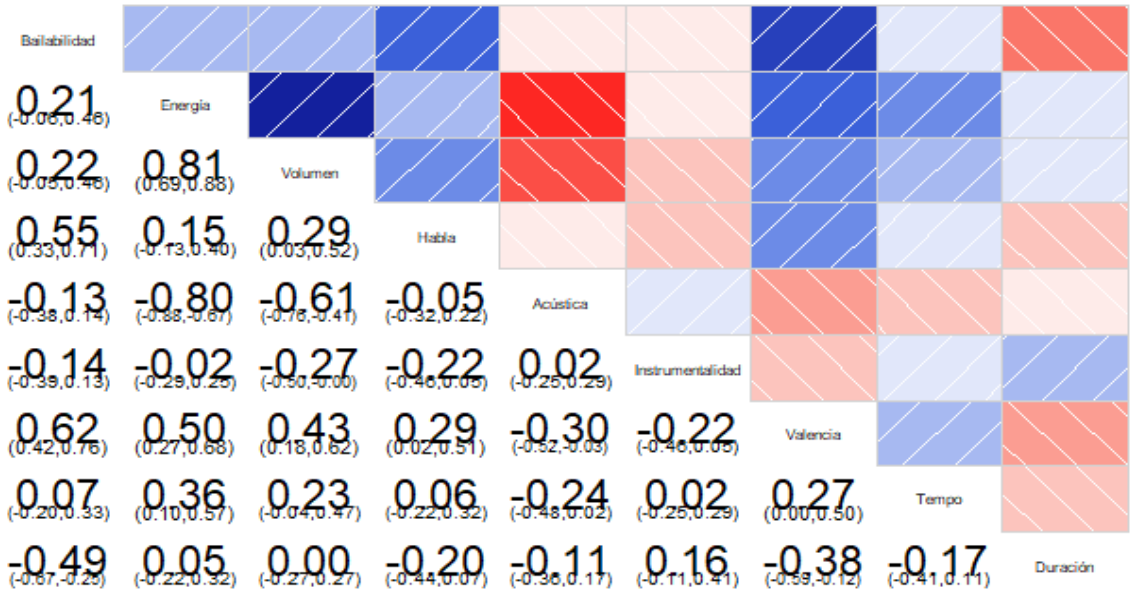
Finalmente, dado que las personas se agrupan según su perfil musical, y se ha determinado el perfil musical de los géneros seleccionados, se procede a comparar las distancias euclídeas entre los centroides de los grupos con el perfil musical de los géneros. Esta comparación, a través del Software R (R Core Team, 2013), utilizando las librerías *cluster* (Maechler, Rousseeuw, Struyf, Hubert & Hornik, 2019) y *biotools (*da Silva, Malafaia, & Menezes, 2017) del CRAN, permite determinar el género al que más se acerca cada grupo de acuerdo a las características musicales establecidas en su perfil.

RESULTADOS

Al analizar la correlación y la variabilidad por persona para determinar las variables que se consideran al conformar el perfil musical se seleccionan: “Bailabilidad”, “Energía”, “Habla”, “Valencia”, “Instrumentalidad” y “Acústica”. Las variables que no se consideraron son el “Tempo”, “Volumen” y “Duración” debido a que las tres tenían una alta variabilidad por persona (ver Anexo 3). En el caso de la variable “Tempo” esta presentó una alta correlación con la variable “Energía”, como se observa en la figura 1.

Figura 1

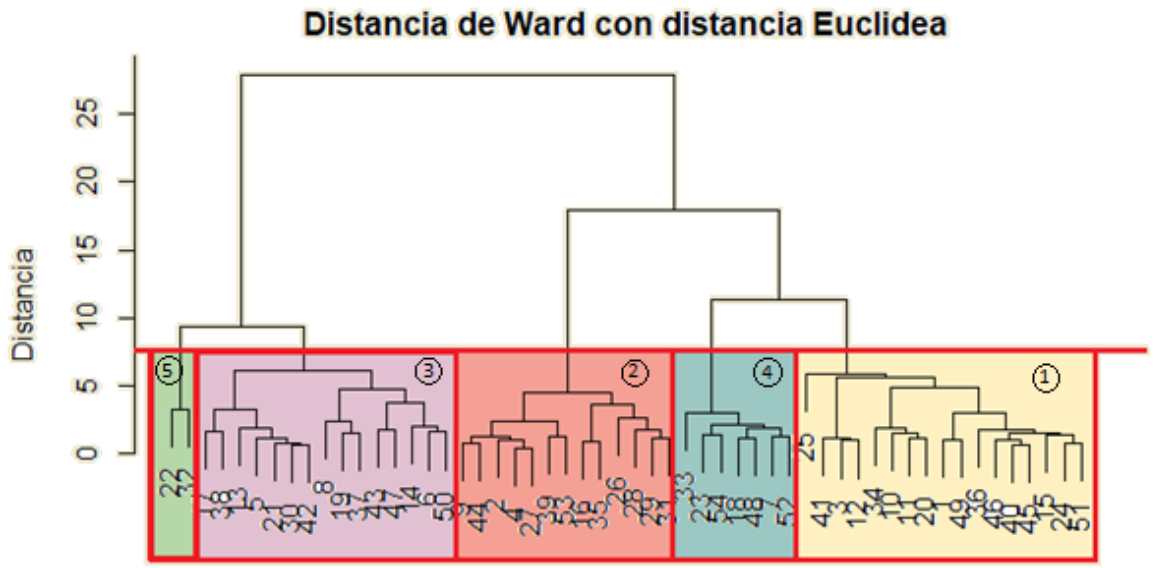
Correlación entre las variables del estudio.



Con respecto a la agrupación de las personas según su perfil musical, se determinan cinco grupos, los cuales están conformados por 17 personas en el Grupo 1, 13 personas en el Grupo 2, 15 en el Grupo 3, 7 personas en el Grupo 4 y por último el Grupo 5 conformado por tan solo 2 personas, lo cual se visualiza en la figura 2.

Figura 2

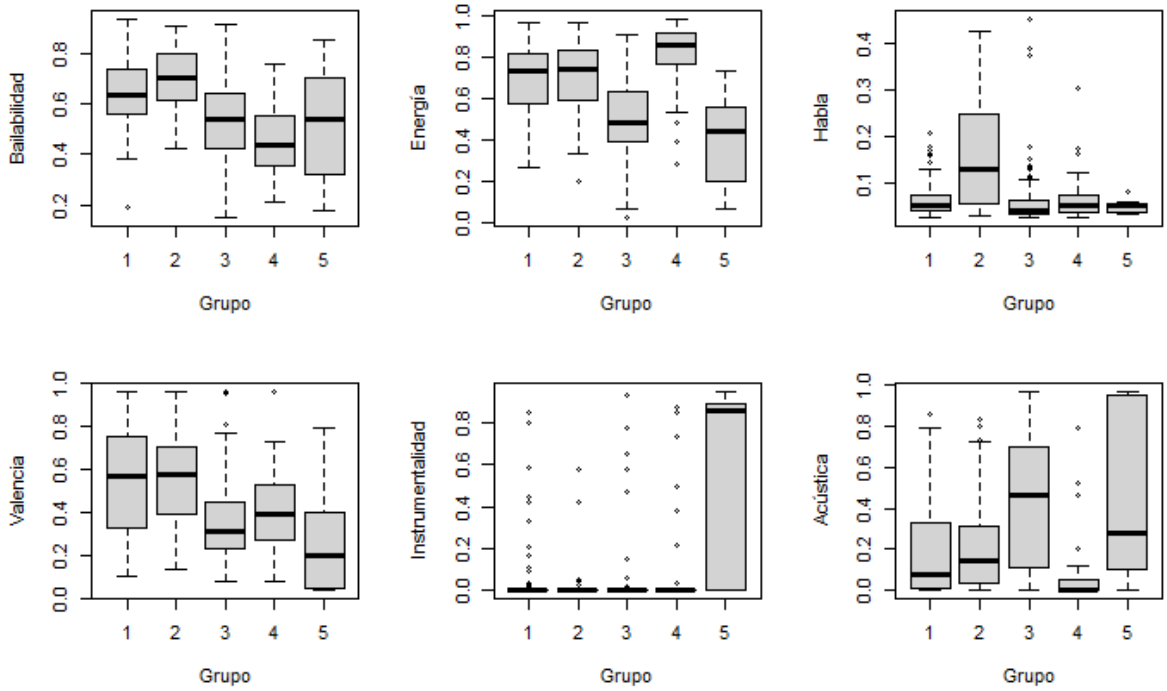
Dendrograma de Distancia de Ward con Distancia Euclídea.



El perfil musical para cada grupo se observa en el Anexo 4 pero para determinar las similitudes y diferencias entre los perfiles musicales de cada grupo es mejor utilizar la siguiente figura:

Figura 3

Distribución de cada variable presente en el perfil musical por grupo.

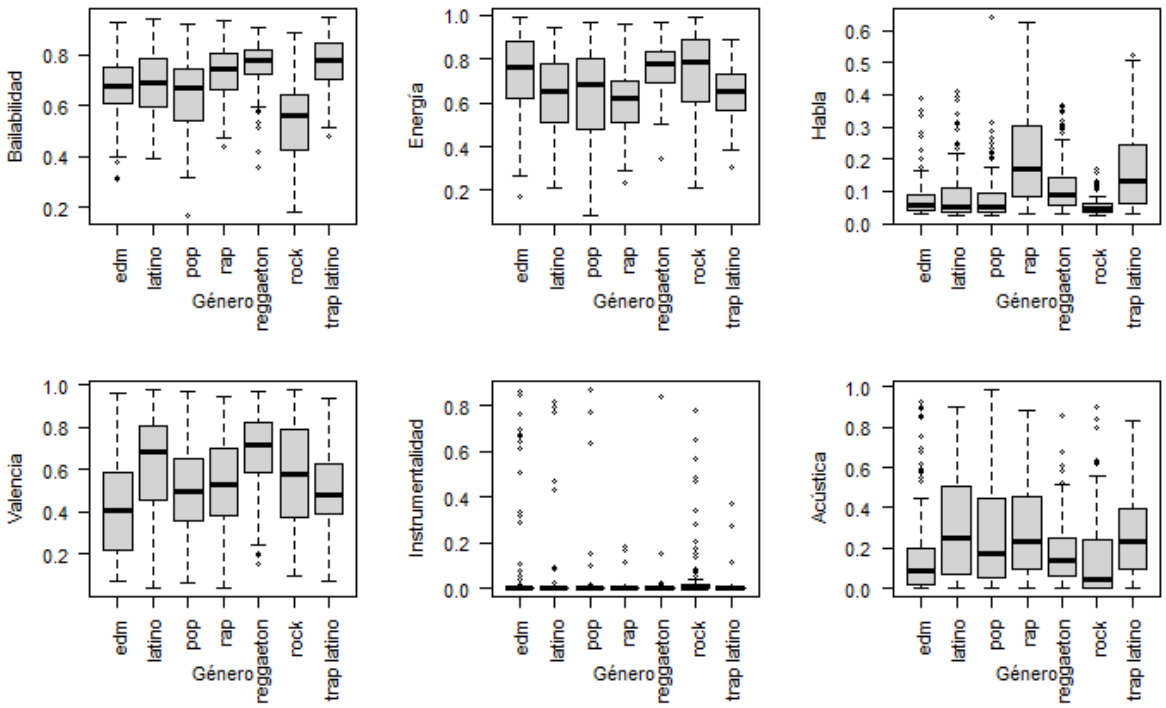


Se observa que el perfil musical del Grupo 1 y el Grupo 2 en general es muy similar en todas las variables, siendo ambos altamente bailables y energéticos, aunque el segundo presenta más habla y un poco más de Bailabilidad. También se tiene que el perfil musical del Grupo 3 es menos energético y bailable que los dos mencionados anteriormente, pero resulta más acústico. El perfil musical del Grupo 4 es el más energético de todos y también es el menos acústico. Por último, el perfil musical del Grupo 5 y el Grupo 3 son similares en Bailabilidad y la Acústica. Sin embargo, el Grupo 5 presenta menos valencia y energía que el 3, además es el más instrumental de todos.

De manera similar, para analizar las diferencias entre los perfiles musicales de los géneros seleccionados se utiliza la figura 4. Adicionalmente, en el anexo 5 se observa el perfil musical de los géneros seleccionados.

Figura 4

Distribución de cada variable presente en el perfil musical por género.



Basado en la figura 4, el Reggaeton y el Trap Latino son los que resultan más bailables mientras que el Rock es el menos bailable. En cuanto a Energía, el EDM, el Reggaeton y el Rock son los más energéticos, aunque en promedio, todos los géneros se consideran altamente energéticos. Al analizar el Habla, se encuentra que los perfiles musicales de Rap y Trap latino son los que en promedio utilizan más palabras. Para la Valencia, se tiene que los géneros musicales que en promedio transmiten mayor positividad son el Reggeaton y el Latino. Por último, en cuanto a la Acústica, si bien todos los géneros son muy similares en promedio, los perfiles musicales menos acústicos son Rock, EDM y Reggaeton. En la variable instrumentalidad se encontraron demasiados valores extremos, por lo que se dificulta su análisis.

Una vez establecidos los perfiles musicales para los grupos de personas y los géneros musicales se describe la relación entre ambos en la tabla 2.

Tabla 2.

Distancias entre el perfil musical de los grupos y el perfil musical de los géneros.

	EDM	LATINO	POP	RAP	REGGAETON	ROCK	TRAP LATINO
Grupo 1	0.20	0.16	0.14	**0.11**	0.19	**0.12**	0.21
Grupo 2	0.15	0.16	0.18	0.15	**0.12**	0.16	**0.14**
Grupo 3	0.54	0.36	0.37	**0.27**	0.35	0.54	0.35
Grupo 4	0.44	0.44	**0.24**	0.34	0.43	0.44	0.45
Grupo 5	0.86	0.72	0.67	**0.66**	0.72	0.74	0.72

Se obtiene que el perfil musical del Grupo 1 es muy cercano al perfil musical del Rap y el Rock. Al mismo tiempo, el perfil del Grupo 2 es muy cercano al perfil del Reggaeton y el Trap Latino. Para los perfiles de los Grupos 3 y 5 se obtienen distancias mayores, y resultan más cercanos al Rap. Para el Grupo 4, se encuentra más cercano al perfil musical del Pop, aunque no lo es tanto como el Grupo 1.

CONCLUSIÓN

Por medio del método de distancia de Ward, se logra agrupar a las 54 personas según su perfil musical en cinco grupos. Lo anterior implica que las variables que fueron utilizadas para la definición del perfil musical resultan útiles para identificar grupos de personas según sus gustos musicales.

La utilidad de las variables para caracterizar los grupos de personas de acuerdo a su perfil musical apoya lo encontrado por autores como Madsen, Hellmuth Margulis, Simchy-Gross & Parra (2019), Bakagiannis y Tarrant (2006) y Hou, Song, Hu, Pan, y Hu (2020) respecto al valor de la música en la socialización. Se observa a partir de los resultados que efectivamente las personas, quienes son seres sociales, se logran agrupar según sus gustos musicales, siguiendo lo planteado por Chamorro-Premuzic (2011). Este establece que el gusto musical no es aleatorio, sino que existen criterios, tales como la personalidad, para identificar la predisposición para disfrutar ciertos tipos de música en específico, lo cual apoya este estudio.

Así mismo, a partir del análisis realizado, se observan diferencias de manera descriptiva en los perfiles musicales de los géneros, estableciendo así que tiene utilidad en el análisis de preferencias musicales y reforzando la validez de estudios realizados por autores como Eijck (2001) y Istók, Brattico, Jacobsen, Ritter & Tervaniemi (2013) y de este mismo.

A partir de lo anterior, se rescata la relación que se encuentra entre los perfiles musicales de los grupos y los de los géneros. Se encuentra que los gustos musicales de cada grupo son similares a ciertos géneros. De hecho, entre los grupos no necesariamente se da una relación con los mismos géneros, sino que existen géneros que se acercan a un grupo y a otros no.

Entre las limitaciones encontradas en el desarrollo del estudio, se considera el tamaño de muestra, dado que es un formulario aplicado en línea para los y las estudiantes y egresados(as) de la carrera de Estadística que voluntariamente decidieron participar. Por lo tanto, se recomienda para investigaciones futuras, aumentar el tamaño de muestra y también la cantidad de canciones que brinde cada uno de los participantes, para que, de tal forma, el perfil musical de cada persona cuente con mayor precisión.

En cuanto a los géneros, en investigaciones futuras se recomienda seleccionar más de ellos, debido a que algunos grupos de personas parecen no ser similares a ninguno de los contemplados. Adicionalmente, sería valioso obtener una muestra de más canciones por género, para igualmente determinar el perfil musical de ellos con mayor precisión.

BIBLIOGRAFÍA

AB, S. (2020). *Spotify for Developers*. Obtenido de https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Bakagiannis, S., & Tarrant, M. (2006). Can music bring people together? Effects of shared musical preference on intergroups bias in adolescence. Obtenido de https://onlinelibrary-wiley-com.ezproxy.sibdi.ucr.ac.cr/doi/epdf/10.1111/j.1467-9450.2006.00500.x

Castro, L. (2003). *Música para todos: una introducción al estudio de la música.* San José, Costa Rica: Editorial de la Universidad de Costa Rica.

Chamorro, T. (2011). *The psychology of Musical Preferences*. Obtenido de https://www.psychologytoday.com/us/blog/mr-personality/201101/the-psychology-musical-preferences

da Silva, A.R., Malafaia, G., Menezes, I.P.P. (2017) biotools: an R function to predict spatial gene diversity via an individual-based approach. Genetics and Molecular Research, 16: gmr16029655.

Eijck, K. v. (2001). Social Differentiation in Musical Taste Patterns. *Oxford Journals*. Obtenido de https://www-jstor-org.ezproxy.sibdi.ucr.ac.cr/stable/pdf/2675621.pdf?refreqid=excelsior%3A1e8331c26879af66d24a09b8292d82ba

Gasser, N. (2019). *A musicologist explains the science behind your taste in music*. Obtenido de https://www.nbcnews.com/better/lifestyle/musicologist-explains-science-behind-your-taste-music-ncna1018336

Hou, Y., Song, B., Hu, Y., Pan, Y., & Hu, Y. (2020). *The averaged inter-brain coherence between theaudience and a violinist predicts the popularity of violin permance*. Obtenido de https://doi.org/10.1016/j.neuroimage.2020.116655

Istók, E., Brattico, E., Jacobsen, T., Ritter, A., & Tervaniemi, M. (2013). *'I love Rock 'n' Roll'—Music genre preference modulates brain responses to music*. Obtenido de https://doi.org/10.1016/j.biopsycho.2012.11.005

Madsen, J., Hellmuth Margulis, E., Simchy-Gross, R., & Parra, L. (2019). *Music synchronizes brainwaves across listeners with strong effects of repetition, familiarity and trainning*. Obtenido de https://www.nature.com/articles/s41598-019-40254-w

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

Sánchez., J. M. (2005). *Música, Jóvenes Generaciones y Medios de Comunicación*. Obtenido de https://books.google.es/books?hl=es&lr=&id=Wvepu33iBoMC&oi=fnd&pg=PA35&dq=gustos+musicales+en+los+jovenes&ots=ipjKjIXuDR&sig=RCUiClg6ZOOw_oy8LFios3Ce8eE#v=onepage&q&f=false

Tagg, P. (1982). *Analysing popular music: theory, method and practice*. Obtenido de https://doi.org/10.1017/S0261143000001227

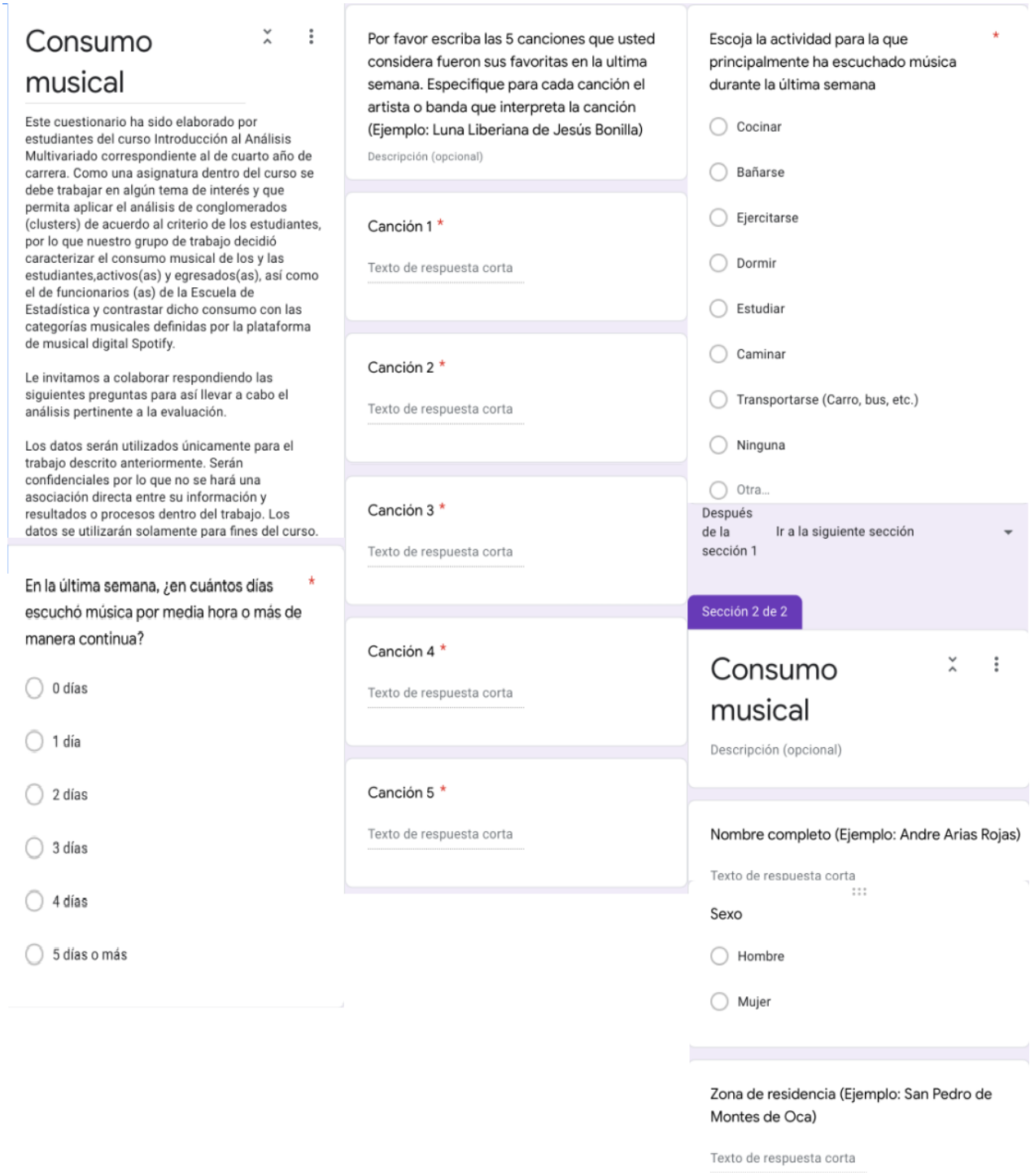
Thompson, C., Parry, J., Phipps, D., & Wolff, T. (2019). spotifyr: R Wrapper for the 'Spotify' Web API. R package version 2.1.1. https://CRAN.R-project.org/package=spotifyr

Wassenverg, A. (2019). *Why We Like Certain Music: The Brain and Musical Preference*. Obtenido de https://www.ludwig-van.com/toronto/2019/05/31/report-why-we-like-certain-music-the-brain-and-musical-preference/

ANEXOS

Anexo 1

Formulario para la recolección de los datos de estudiantes y egresados de la carrera Estadística de la Universidad de Costa Rica.



Anexo 2

Fórmulas de los métodos para calcular las distancias entre los individuos, entre los grupos y el método de agrupamiento.

Distancia entre individuos:

Distancia Euclídea

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(x_{ik} - x_{jk}\right)^2} = \sqrt{\left(X_i - X_j\right)^T \left(X_i - X_j\right)}$$

Donde: $d_{ij} = \ el\ valor\ que\ toma\ la\ distancia\ entre\ la\ i-ésima\ y\ la\ j-ésima\ unidad\ en\ la\ k-ésima$

x_{ik}= el valor que toma la i-ésima unidad en la k-ésima variable)

x_{jk}=el valor que toma la j-ésima unidad en la k-ésima variable)

X_i= el vector de q valores de la i-ésima unidad)

X_j= el vector de q valores de la j-ésima unidad)

Distancia entre grupos:

Distancia bajo el método del vecino cercano

$$\delta(A,B) = min\{d(x_i, x_j); x_i \in A, x_j \in B\}$$

Donde: $\delta(A,B) =$ el valor que toma la distancia entre el grupo A y el grupo B

x_i= el valor que toma la i-ésima unidad

x_j= el valor que toma la j-ésima unidad

Distancia bajo el método del vecino más lejano

$$\delta(A,B) = max\{d(x_i, x_j); x_i \in A, x_j \in B\}$$

Donde: $\delta(A,B) =$ el valor que toma la distancia entre el grupo A y el grupo B

x_i= el valor que toma la i-ésima unidad

x_j= el valor que toma la j-ésima unidad

Distancia bajo el método del Salto promedio

$$\delta(A,B) = \frac{1}{n_A n_B} \sum_{x_i \in A, x_j \in B} d(x_i, x_j)$$

Donde: $\delta(A,B) =$ el valor que toma la distancia entre el grupo A y el grupo B

x_i= el valor que toma la i-ésima unidad

x_j= el valor que toma la j-ésima unidad

n_A= el número de elementos en el grupo A

n_B= el número de elementos en el grupo B

Distancia bajo el método Ward

$$\delta(A,B) = \frac{n_A n_B}{n_A + n_B} \| g_A - g_B|^2$$

Donde: $\delta(A,B) =$ el valor que toma la distancia entre el grupo A y el grupo B

g_A= el valor que toma el centroide del grupo A

g_B= el valor que toma el centroide del grupo B

n_A= el número de elementos en el grupo A

n_B= el número de elementos en el grupo B

Método de agrupamiento:

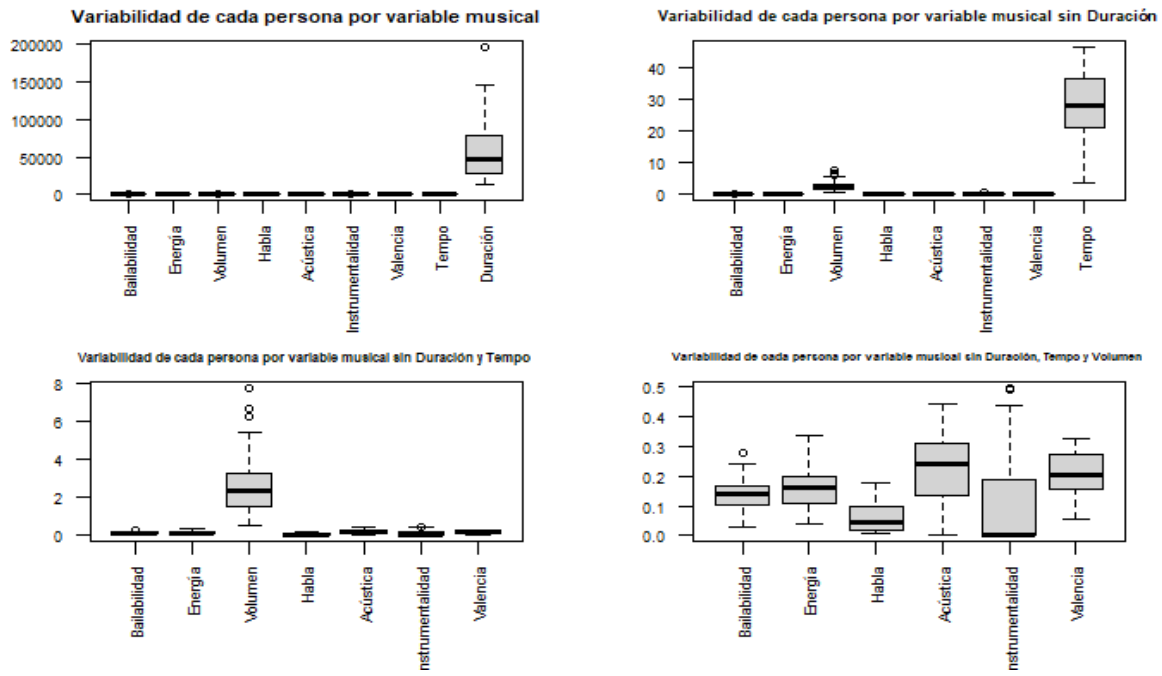
Método del dendograma

$$\delta(C_i, C_j) = min\{\delta(C_h, C_k); h \neq k\}$$

Donde: $\delta(C_i, C_j) =$el valor mínimo que toma la distancia entre el grupo C_i y el C_j

Anexo 3

Análisis de variabilidad de cada persona por variable musical.



Anexo 4

Perfil musical de cada grupo a partir de las variables de estudio.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Bailabilidad	**0.64**	**0.70**	**0.53**	0.46	0.51
Energía	**0.69**	**0.71**	**0.50**	**0.80**	0.41
Habla	0.06	**0.16**	0.07	0.07	0.05
Valencia	**0.55**	**0.56**	**0.36**	0.41	**0.25**
Instrumentallidad	0.06	0.02	0.05	0.10	**0.54**
Acústica	0.21	0.21	**0.43**	**0.07**	**0.48**

Anexo 5

Perfil musical de los géneros a partir de las variables de estudio.

	EDM	LATINO	POP	RAP	REGGAETON	ROCK	TRAP LATINO
Bailabilidad	0.67	0.68	0.64	0.73	0.76	0.54	0.77
Energía	0.73	0.64	0.63	0.61	0.76	0.73	0.64
Habla	0.08	0.09	0.08	0.21	0.12	0.05	0.16
Valencia	0.43	0.64	0.50	0.53	0.69	0.57	0.51
Instrumentalidad	0.08	0.03	0.03	0.00	0.01	0.05	0.01
Acústica	0.18	0.32	0.27	0.29	0.19	0.16	0.28

Caracterización de triatletas según medidas antropométricas, hidratación, recuperación alimenticia y consumo de suplementos

Julio Madrigal Sanabria[11], Nelson Torres Chávez[11]

juanmasa9704@gmail.com, nelson_0823@hotmail.com

RESUMEN

Existen factores que pueden definir el rendimiento que tienen los atletas en las competiciones de triatlón. Las variables antropométricas, la hidratación y el consumo de suplementos son unos de ellos. Debido a ello el objetivo principal de esta investigación es identificar distintos perfiles de triatletas de categoría olímpica según suplementación, hidratación, recuperación alimentaria post entrenamiento y composición corporal. Para ello se realizan dos análisis de componentes principales para variables de la misma naturaleza, uno para antropométricas y otro para consumo post-entrenamiento, con el fin de reducir la dimensión de variables. Luego se realiza un análisis jerárquico de conglomerados utilizando, para calcular la distancia entre individuos, el método de Gower y para calcular la distancia entre grupos: el vecino más cercano, vecino más lejano, salto promedio y Ward. Donde se encuentra que el mejor criterio que cumple el criterio de conglomerados homogéneos dentro de sí y heterogéneos entre sí es el método de Ward. Finalmente se obtienen 3 conglomerados con 14, 28 y 30 individuos respectivamente. Se procede a realizar un análisis descriptivo para caracterizar cada conglomerado y encontrar los perfiles de atletas.

PALABRAS CLAVE: Análisis de conglomerados, estadística, caracterización, triatlón, hidratación, suplementos.

INTRODUCCIÓN

En múltiples deportes se estudia la relación entre el rendimiento deportivo y la contextura física, manifestando cómo el porcentaje de grasa corporal afecta el tiempo y la sensación de esfuerzo en pruebas de resistencia (Ramos et al., 2015). Además, siendo el triatlón un deporte agotador, es importante el consumo de agua tanto durante como después de practicar una de las tres disciplinas, esto debido a que los triatletas presentan un porcentaje medio de deshidratación del 1% (Sellés et al., 2015). De hecho, la deshidratación progresiva en triatletas disminuye el rendimiento de resistencia y aumenta el riesgo de lesiones (Logan, 2019). Diversos investigadores, al estudiar triatletas, observaron que a mayor distancia más retención de líquidos y alteraciones en la composición corporal, ya que después de la carrera, los cambios significativos incluyeron reducciones en la masa corporal, masa grasa y porcentaje de grasa corporal (Baur, et al. 2016).

En la actualidad hay una creciente demanda de información sobre la suplementación, y el consumo de ayudas ergo génicas aumenta. (Martínez, 2017). Además, respecto al consumo de suplementos, se han realizado estudios en diferentes atletas para conocer qué tipo de sustancia consumen, porqué lo toman y con qué frecuencia, pero los estudios realizados en los deportes de resistencia, triatlón y carrera de montaña son insuficientes (Pitarch et al., 2013).

[11] Estudiantes de Estadística de la Universidad de Costa Rica

Además, en el caso de los triatletas, es importante analizar la eficiencia de su recuperación para la siguiente sesión, si esta no es adecuada en carbohidratos, proteínas, líquidos y electrolitos, las adaptaciones beneficiosas y el rendimiento pueden verse obstaculizados (Arley, 2020). Es decir, que de acuerdo a estos factores el provecho que tienen los atletas en una competencia, estos se pueden ver influenciados positiva o negativamente.

Por lo dicho anteriormente, el presente trabajo se enfoca en analizar las características que tienen una muestra de triatletas respecto a su hidratación, consumo post entrenamiento, composición física y utilización de suplementos, esto mediante un análisis estadístico de agrupamientos o clústeres, para así identificar los distintos perfiles que estos tienen. Esto como herramienta para todo aquel profesional que se involucre con estos deportistas y pueda ampliar su conocimiento para el manejo de estos, facilitando un adecuado seguimiento nutricional y aprovechando las ventajas del mismo para las competiciones.

Esta investigación tiene como objetivos contrastar conglomerados de una muestra de triatletas de categoría olímpica para encontrar qué características tienen los distintos grupos formados y enunciar las características que diferencian distintos conglomerados de una muestra de triatletas de categoría olímpica.

METODOLOGÍA

La investigación se desarrolla con una muestra de 73 triatletas de los equipos Tribu, Hypoxic, El Tim y ZoiTri. Para este estudio se toman en cuentan deportistas de ambos sexos con un rango de edad entre los 18 y 47 años. Los datos se recolectaron durante el tercer cuatrimestre 2019 y el primer cuatrimestre 2020, en Cartago, Costa Rica, y fue una selección no aleatoria. Las variables que se utilizarán para el análisis son: edad, sexo, consumo de suplementos (creatina, cafeína y proteína), cantidad de consumo de agua durante la práctica de un ejercicio (ciclismo, atletismo y natación), cantidad de consumo de agua al día (excluyendo durante práctica de ejercicios), cantidad de veces que entrena por semana, talla, peso, porcentaje de agua corporal, grasa corporal en kilogramos, masa corporal en kilogramos, nivel de grasa visceral y el consumo promedio post-entrenamiento (carbohidratos, proteínas, grasas y kilocalorías).

Para la edad, sexo, consumo de suplementos, consumo de agua y las veces de entrenamiento por semana se hizo una encuesta en los individuos, para calcular el consumo promedio post-entrenamiento, se utilizó la aplicación MyFitnessPal (Under Armour, 2015) durante tres días de entrenamiento, y se registraba el consumo de alimentos luego de entrenar, por último, las medidas antropométricas se midieron por medio de bioimpedancia.

Para empezar el análisis, se evalúa si existen valores extremos entre los datos, para ello se utiliza el criterio de los *leverages* o *hatvalues* (Cardinali, 2013). Estos valores corresponden a los elementos de la diagonal de la matriz hat que se calcula de la siguiente manera:

$$H = X(X^T X)^{-1} X^T$$

Donde

- La matriz X corresponde a la matriz de datos.

Una vez obtenidos los valores se establece el límite, que es igual a dos veces la media de los *leverages, a*quellos individuos que superen este límite son considerados valores extremos y deben omitirse de la investigación si no son muchos los casos. Luego de esto, se hace un análisis de correlaciones entre variables antropométricas y las de consumo post-entrenamiento por separado, con el fin de evaluar la posibilidad de realizar dos análisis de componentes principales para grupos de variables de su misma naturaleza, y con esto, verificar que se permita reducir la dimensión de las variables. Para ello se extrae la matriz de correlaciones de cada grupo de variables y se hace la descomposición espectral de la misma, para ello, se resuelve la siguiente ecuación matricial:

$$|A - \lambda I| = 0$$

Donde

- A corresponde a la matriz de correlaciones de A e I corresponde a la matriz identidad.

El objetivo es buscar la solución a esta ecuación, obteniendo así los valores λ, que corresponde a los valores propios de la matriz A. Una vez obtenidos los valores de λ se procede a buscar la solución de la siguiente ecuación para obtener los vectores propios de la matriz A:

$$|A - \lambda I| \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = 0$$

Finalmente, al tener tanto los vectores y valores propios de la matriz de correlación se realiza la siguiente operación:

$$Z_1 = a_{11}X_1^S + a_{12}X_2^S + \cdots + a_{1q}X_q^S$$

Donde

- Cada X_j^Ses un vector que contiene todos los valores de la j-ésima variable estandarizada. El vector de coeficientes a_1= (a_{11}, …, a_{1q})es el vector característico asociado con λ_1, que es el valor propio mayor de la matriz de correlaciones. Y Z_1 es la nueva variable o componente principal.

El procedimiento se repite j veces, el equivalente al número de variables utilizadas en el análisis y manteniendo siempre el orden de mayor a menor según el valor del λ.

Luego de obtener todos los componentes principales, se utiliza el criterio de varianza mínima para elegir la cantidad de componentes a utilizar. Este criterio dice que cuando se utilizan las variables estandarizadas, la cantidad de componentes a seleccionar son aquellos en los cuales la varianza de los mismos es mayor a uno (Peña, 2014).

Una vez conformados los componentes se procede a realizar el análisis de conglomerados, ésta es una técnica para agrupar observaciones similares en un número de conglomerados basado en los valores observados de un conjunto de variables para cada individuo (Sinharay, 2010). Para ello únicamente se toma en cuenta el

método jerárquico, debido a que, al tener variables categóricas, el método de K-Medias y K-Medoides se invalida (Clavijo M. & Granada D., 2016)**.** Lo primero a realizar es calcular la distancia entre individuos, para ello se utiliza la distancia de Gower (Esponda et al. 2010) pues esta toma en cuenta la existencia de variables categóricas; la fórmula que sigue este método es la siguiente:

$$d_{ij}^{G} = \frac{1}{q}\left[\sum_{h=1}^{p_1} \frac{|x_{ih} - x_{jh}|}{R_h} + \beta_{ij}\right]$$

Donde:

- R_h = rango de la *h-ésima* variable cuantitativa,
- b_{ij} = número de variables cualitativas que no son idénticas,
- q = número total de variables.

Luego de calcular las distancias entre individuos se procede a calcular las distancias entre grupos, para ello existen también diversos métodos, el método del vecino más cercano, el vecino más lejano, el salto promedio y finalmente el método de Ward, estos métodos siguen las siguientes fórmulas:

Vecino más cercano

$$\delta(A, B) = min\{d(x_i, x_j);\ x_i \epsilon\ A, x_j \epsilon B\}$$

Donde:

- d(x_i, x_j): corresponde a las distancias entre individuos

Vecino más lejano

$$\delta(A, B) = max\{d(x_i, x_j);\ x_i \epsilon\ A, x_j \epsilon B\}$$

Donde:

- d(x_i, x_j): corresponde a las distancias entre individuos

Salto promedio

$$\delta(A, B) = \frac{1}{n_A n_B} \sum_{x_i \epsilon A;\ x_j \epsilon B} d(x_i, x_j)$$

Donde:

- d(x_i, x_j): corresponde a las distancias entre individuos
- $n_a\ y\ n_b$: son el número de individuos en cada conglomerado.

Ward

$$\delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|g_A - g_B\|^2$$

Donde:

- $g_a \; y \; g_b$: centroides del conglomerado A y B
- $n_a \; y \; n_b$: son el número de individuos en cada conglomerado.

Posteriormente se crean los cuatro dendrogramas y se escoge cuál distancia entre grupos de las citadas anteriormente se utilizan, esto se hace bajo criterio de los autores, de manera que se observe la mejor categorización de individuos siguiendo el principio de que sean homogéneos dentro de cada conglomerado y heterogéneos entre conglomerados.

El software utilizado para el análisis es R (R Core Team, 2020) en su versión 4.0.0, utilizando los paquetes readr (Wickham et al. 2018) para la lectura de la base de datos, corrplot (Wei & Simko, 2017) para la visualización de correlaciones y conglomerados, cluster (Maechler et al. ,2019) para el cálculo de la distancia de gower, RColorBrewer Neuwirth (2014) para las escalas de colores de los gráficos, factoextra (Kassambara & Mundt, 2020) y dendextend (Galili, 2015) para la creación de dendrogramas**.**

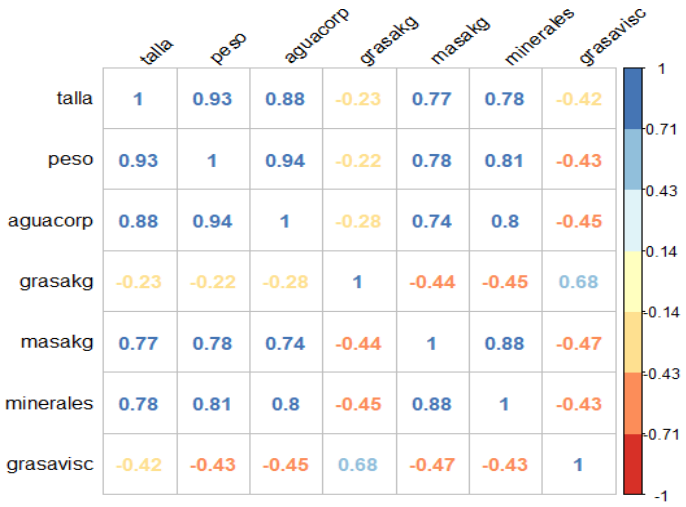
RESULTADOS

De manera inicial se buscaron valores extremos por el método de leverages ya mencionado, se encuentra un valor extremo en el individuo número 43 (Anexo 1), que decide no tomarse en cuenta para el análisis debido a que es solamente un dato.

Para examinar las correlaciones, se hizo según la naturaleza de las variables, es decir, se observaron las de variables antropométricas y las de consumo post-entrenamiento por aparte. En el caso de las correlaciones de las variables antropométricas (Figura 1), se observan ciertas correlaciones considerables, por ejemplo, los minerales están fuertemente correlacionados con la masa, agua corporal, el peso y talla de triatletas. Además, el agua corporal, peso y talla están correlacionados entre ellos, por esto se considera que es útil realizar un análisis de componentes principales y así lograr una reducción en la dimensión de variables.

Figura 1

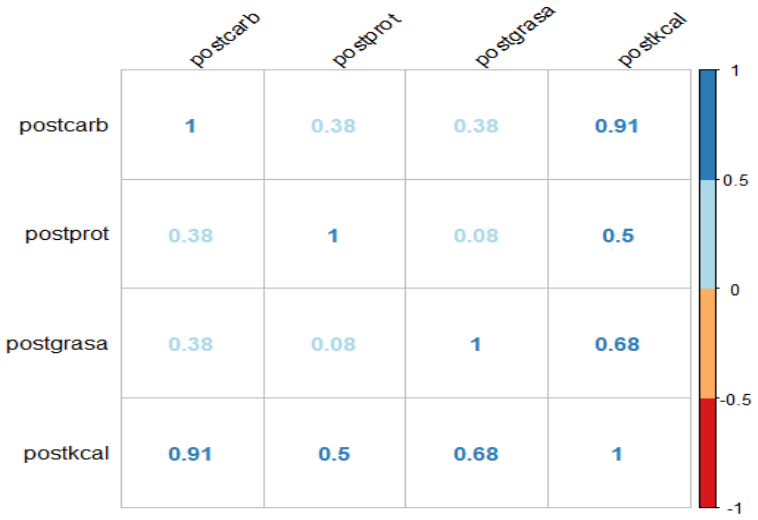
Correlaciones entre variables antropométricas.



En el caso de las variables de consumo post-entrenamiento se encuentran relaciones altas entre el consumo de carbohidratos y la cantidad de kilocalorías consumidas, mientras que se encuentran relaciones intermedias entre el consumo de grasas y proteínas con la cantidad de kilocalorías consumidas (Figura 2). Posteriormente y debido a la evidencia, se realiza un análisis de componentes principales separado del de variables antropométricas.

Figura 2

Correlación entre las variables de consumo post-entrenamiento.



Para realizar los dos análisis de componentes principales se estandariza la base de datos para atenuar las diferencias en las escalas de medición de las variables utilizadas. Luego, se procede a realizar el análisis de componentes principales, en este se logra observar cuáles son las variabilidades explicadas de cada componente. Debido a que los primeros dos componentes explican el 86.7% de la variabilidad total de las variables originales y que cada una de las varianzas de estos son mayor a 1, apegados al criterio de Kaiser (Anexo 2), se decide trabajar con dos componentes.

Una vez realizado el primer análisis de componentes principales, se procede a realizar el segundo que corresponde a las variables de post-entrenamiento. Donde se indica cuáles son las variabilidades explicadas de cada componente (Anexo 3), a pesar de que el primer componente sólo explica el 64% de la variabilidad total de las variables originales, este es el único componente cuya varianza es mayor a 1, debido a esto, se decide trabajar con un componente. Para ser claros se nombrarán como componentes 1 y 2 los formados por variables antropométricas y componente 3 al formado por las variables de consumo post-entrenamiento.

Para considerar la fuerza que tienen las variables originales sobre los 3 componentes seleccionados se observa la matriz de correlaciones entre los mismos (Figuras 3 y 4), donde se observa que con el primer componente están directamente correlacionadas todas las variables, exceptuando ambas grasas, en cambio, en el segundo componente la correlación es inversa en ambas grasas. Además, se observa una correlación fuerte con el consumo de carbohidratos y kilocalorías con el tercer componente.

Figura 3

Correlaciones entre las variables antropométricas y los componentes 1 y 2.

Variable	Componente 1	Componente 2
Talla	0.9070	-0.2747
Peso	0.9280	-0.2931
Porcentaje de agua corporal	0.9180	-0.2295
Grasa en kg	-0.4984	-0.8007
Masa en kg	0.8992	-0.0179
Minerales	0.9162	-0.0503
Grasa visceral	-0.6195	-0.6378

Figura 4

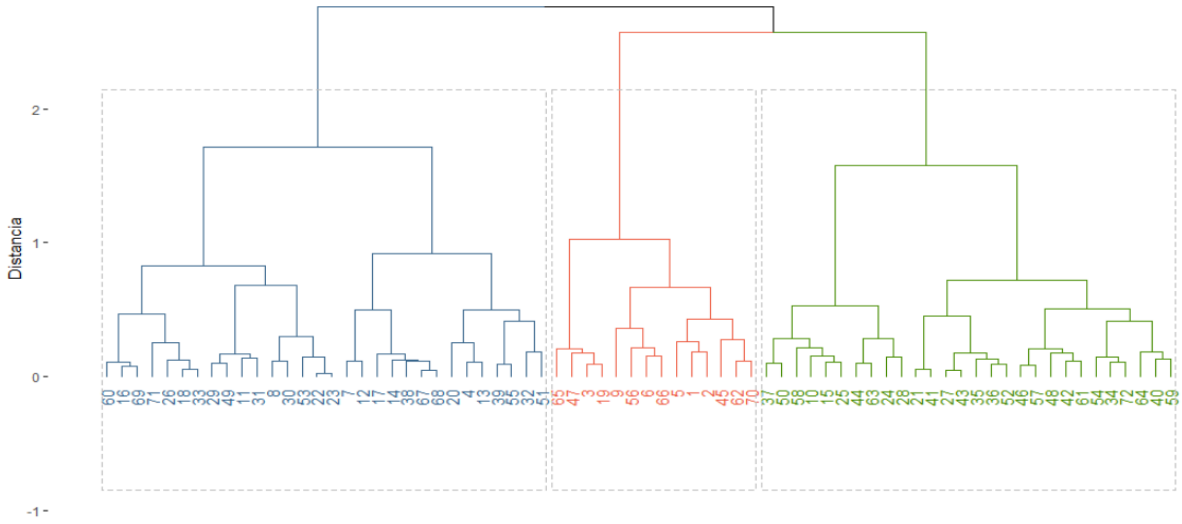
Correlaciones entre las variables de consumo post-entrenamiento y el componente 3.

Variable	Componente 3
Consumo de carbohidratos post-entrenamiento	0.8843
Consumo de proteínas post-entrenamiento	0.5730
Consumo de grasas post-entrenamiento	0.6807
Kilocalorías consumidas post-entrenamiento	0.9958

Se procede a hacer el análisis jerárquico de conglomerados con las siguientes variables: los 3 componentes ya descritos, el consumo de suplementos (cafeína, proteína y creatina), consumo de líquidos durante el entrenamiento en las distintas áreas (ciclismo, natación y atletismo) y cantidad de agua consumida diaria, para esto, se tienen los dendrogramas con la distancia entre individuos de Gower pues es el único método que permite calcular las mismas cuando existen variables categóricas. Para la distancia entre grupos se utilizaron los métodos del vecino más cercano, vecino más lejano, el salto promedio y Ward, el único que tenía una separación de conglomerados que permitía observar que fueran heterogéneos entre ellos y homogéneos dentro, fue la distancia de Ward, cuya representación está en la figura 5 (los demás gráficos se pueden observar en Anexos 6,7 y 8). En dicho dendrograma se escoge la formación de 3 conglomerados mediante el análisis y criterio propio. En el primer, segundo y tercer conglomerado hay 14, 30 y 28 observaciones, respectivamente (Anexo 9)

Figura 5

Dendrograma para el análisis jerárquico de conglomerados, distancia entre individuos de Gower y entre grupos de Ward.

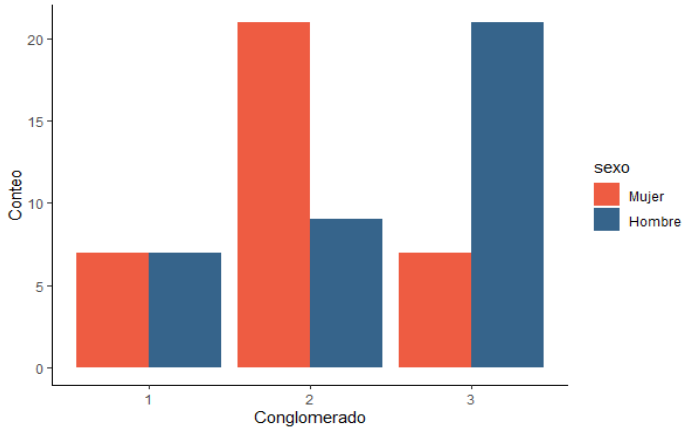


En la figura anterior se selecciona con color azul el conglomerado 1, con color rojo el conglomerado 2 y con color verde el conglomerado 3. Se nota visualmente pueden cumplir el criterio de ser homogéneos dentro de sí y heterogéneos entre sí.

Una vez realizado el dendrograma con sus respectivos conglomerados, corresponde en el análisis, realizar una descripción de los mismos según sus características, con el fin de identificar los perfiles de atletas de triatlón. A continuación, se muestra el análisis descriptivo por conglomerado.

Figura 6

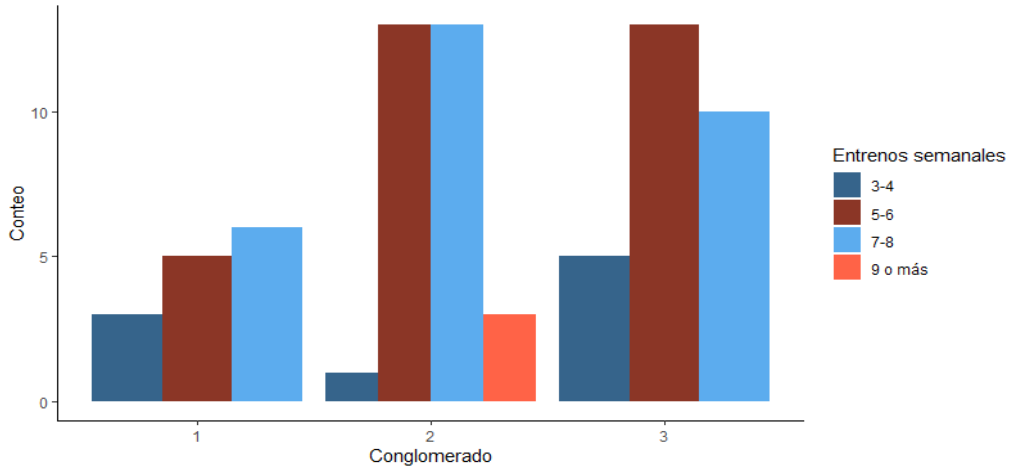
Distribución de los conglomerados según el sexo de los triatletas.



En la figura 6 se observa una misma cantidad de hombres y mujeres en el primer conglomerado, mientras que hay más mujeres en el segundo conglomerado (70% versus 30% de hombres), además, hay un total de 21 hombres (75%) en el tercer conglomerado contra 25% de hombres.

Figura 7

Distribución de los conglomerados según las veces que entrenan por semana los triatletas.



En la figura 7 se aprecia que en el conglomerado 2 están las personas que entrenan más veces por semana (están los únicos que entrenan 9 o más veces por semana), el tercer conglomerado está compuesto por un grupo de personas que en su mayoría entrena entre 5 a 8 veces por semana, mientras que el conglomerado 1 no tiene una gran diferencia en esta variable como los otros conglomerados.

Figura 8

Distribución de los conglomerados según la cantidad de líquido que consumen en cada categoría los triatletas.

Deporte	**Clúster 1**		**Clúster 2**		**Clúster 3**	
	n	**%**	**n**	**%**	**n**	**%**
Ciclismo						
301-500ml	8	57.14	10	33.33	3	10.71
501-700ml	0	00.00	7	23.33	**11**	39.29
Más de 700ml	6	42.86	13	43.33	**14**	50.00
Natación						
0-100ml	3	21.43	2	06.67	3	10.71
101-300ml	8	**57.14**	5	16.67	**10**	**35.71**
301-500ml	2	14.29	**13**	**43.33**	**11**	**39.29**
501-700ml	1	07.14	8	26.67	3	10.71
Más de 700ml	0	00.00	2	06.67	1	03.57

Atletismo						
101-300ml	6	42.86	17	56.67	5	17.86
301-500ml	5	35.71	13	43.33	17	**60.71**
501-700ml	3	21.43	0	**00.00**	6	21.43

En la figura 8 se puede observar cómo es el consumo de líquidos durante cada una de las disciplinas en cada conglomerado:

- conglomerado 1: en ciclismo toman mucho o poco líquido, no hay nadie que tome entre 501 ml - 700 ml, mientras que en natación la mayoría toma entre 101 ml - 300 ml.
- conglomerado 2: hay una cantidad considerable de personas que toman una cantidad de entre 301 ml - 700 ml en natación, mientras que, en atletismo, todos consumen entre 101 ml - 500 ml.
- conglomerado 3: Casi todos consumen 501 ml o más en ciclismo, mientras que en atletismo la mayoría toma entre 301 ml - 500 ml y, en natación, la mayoría toma entre 101 ml - 500 ml.

Figura 9

Distribución de los conglomerados según el consumo diario de agua por fuera de las sesiones de entrenamiento.

Consumo de agua	**Clúster 1**		**Clúster 2**		**Clúster 3**	
	n	**%**	**n**	**%**	**n**	**%**
0-1000ml	4	28.57	1	03.33	10	35.71
1001-2000ml	5	35.71	9	30.00	8	28.57
Más de 2000ml	5	35.71	20	**66.67**	10	35.71

En la figura 9, por otro lado, se analiza la cantidad del consumo diaria de agua por fuera del entrenamiento sólo se diferencia claramente en el conglomerado 2, donde la mayoría toma más de 2 litros de agua diaria.

Figura 10

Distribución de los conglomerados según el consumo de suplementos alimenticios de los triatletas.

Suplemento	Clúster 1		Clúster 2		Clúster 3	
	n	%	n	%	n	%
Cafeína						
Sí consume	10	71.43	8	26.67	28	100.0
No consume	4	28.57	22	73.33	0	00.00
Proteína						
Sí consume	11	78.57	18	60.00	18	64.29
No consume	3	21.43	12	40.00	10	35.71
Creatina						
Sí consume	14	100.0	2	06.67	0	00.00
No consume	0	00.00	28	93.33	28	100.00

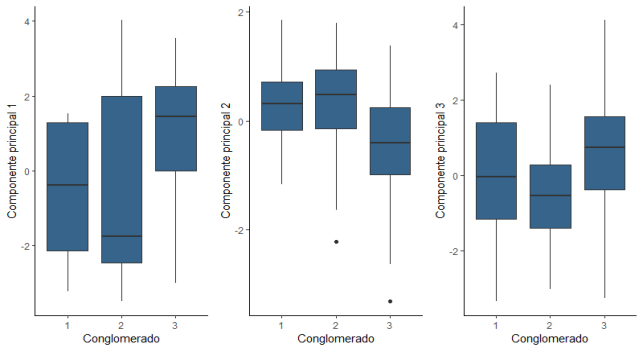
El consumo de suplementos para impulsar el rendimiento de los triatletas se analiza en la figura 10:

- conglomerado 1: la mayoría consume de los 3 suplementos
- conglomerado 2: la mayoría no consume ni cafeína ni creatina, mientras que una cantidad relevante consume proteína
- conglomerado 3: la mayoría consume cafeína y proteína, mientras que nadie consume creatina

Con el fin de observar el efecto que tienen de los 3 componentes principales (los dos formados por variables antropométricas y el formado por variables de consumo post-entrenamiento), sobre los conglomerados se realizaron gráficos de cajas (Figura 11).

Figura 11

Distribución de los conglomerados según componente.



En la figura anterior se aprecia como para el componente principal 1, que lo conforman principalmente la talla, el peso, el porcentaje de agua corporal, la masa en kg y los minerales (Figura 3), el conglomerado 3 parece tener la tendencia de tener un promedio mayor en el mismo, con respecto al conglomerado 1, mientras que el segundo, debido a la gran variabilidad, es complicado examinar si la media del componente 1 es mayor o menor que en los demás conglomerados. Para el componente 2, formado principalmente por la grasa en kg y el porcentaje de grasa visceral (Ver Figura 3), se contempla como el conglomerado 3 parece tener un promedio menor con respecto a los otros dos conglomerados. En contraste, los otros conglomerados parecen muy similares en términos de sus promedios. Por último, en el componente 3 está formado principalmente por el consumo de carbohidratos y el consumo de kilocalorías post-entrenamiento (Ver Figura 4). Parece que nuevamente el promedio del mismo en el conglomerado 3 es mayor, mientras que el conglomerado 1 parece tener un promedio cercano al general que es 0 y el número 2 el promedio menor. No obstante, puede ser que por términos de variabilidad sea necesario realizar un análisis de varianza para determinar si existe diferencia entre los promedios de cada conglomerado.

En resumen, las características que se lograron encontrar según cada conglomerado son los siguientes:

Conglomerado 1.

- Posee igual cantidad de hombres y mujeres
- Mayor cantidad entrena de 7 a 8 veces por semana
- Tienen un consumo bajo en líquido en natación.
- No presenta grandes diferencias respecto al consumo de agua diario
- La mayoría consume de los 3 suplementos.
- Componentes principales 1, 2 y 3 con valores promedio muy cercanos a la media general (0).

Conglomerado 2.

- Mayor cantidad de mujeres que de hombres
- Único conglomerado en donde se entrena 9 veces a la semana o más y donde la menor cantidad de personas entrenan solamente de 3 a 4 veces por semana.
- Hay una cantidad considerable de personas que toman gran cantidad de líquidos natación, mientras que, en atletismo, consumen poco.
- Son los que en su mayoría consumen más agua diariamente (más de 2 000 ml.)
- La mayoría no consume ni cafeína ni creatina.
- Componente principal 2 con un promedio superior a los demás conglomerados. Es decir, menor cantidad de grasas (debido a la alta correlación inversa con las variables de grasa en kg y grasa visceral)
- Componente principal 3 con un promedio menor a los demás conglomerados

Conglomerado 3.

- Mayor cantidad de hombres que de mujeres
- Está compuesto por un grupo de personas que en su mayoría entrena entre 5 a 6 veces por semana.
- Casi todos consumen mucho líquido en ciclismo (501 ml o más), mientras que en atletismo la mayoría toma una cantidad considerable pero no tanta (301 ml - 500 ml) y, en natación, la mayoría toma poca cantidad (101 ml - 500 ml).
- No presenta grandes diferencias respecto al consumo de agua diario

- La mayoría consume cafeína y proteína, mientras que nadie consume creatina.
- Componentes principales 1 y 3 con promedios superiores a la media y a los demás conglomerados, promedio de componente 2 más bajo de los tres conglomerados, es decir mayor cantidad de grasas (debido a la alta correlación inversa con las variables de grasa en kg y grasa visceral)

De acuerdo con las características encontradas los conglomerados se definen entonces los siguientes perfiles:

- Perfil 1 atleta de alto entrenamiento, misma cantidad de hombres y mujeres, baja hidratación, alto consumo los tres suplementos y con medidas antropométricas y consumo post-entrenamiento cercanos al promedio (conglomerado 1)
- Perfil 2 atleta de alto entrenamiento, mayoría de mujeres, alta hidratación, alto consumo de suplemento de proteína, con menor grasa visceral y corporal y bajo consumo post-entrenamiento.
- Perfil 3 atleta de entrenamiento regular, mayoría de hombres con hidratación regular, alto consumo de cafeína y proteína, grandes, pesados, con mucha masa y consumo post-entrenamiento mayores al promedio general y mayor cantidad de grasa visceral y total.

CONCLUSIONES

Un estudio elaborado en el 2016 sugiere que las personas adultas con índices de grasa corporal más altos pudieran también llevar a comportamientos que llevaran a una hidratación inadecuada (Chang et al., 2016). Esto explicaría el por qué a pesar de consumir suplementos, al tener una hidratación regular (no tan buena ni mala), los atletas con el perfil 1 tiene medidas antropométricas muy cercanas al promedio.

La composición corporal es un elemento que puede afectar el rendimiento de un deportista en diferentes pruebas, además, los descensos de los niveles de proteína están asociados a la pérdida muscular, que puede hacer reducir el rendimiento deportivo (Ramos et al., 2015). Debido a ello, se podría considerar la posibilidad de que los individuos del perfil 2 sean aquellos que logren desempeñarse mejor debido a que, al consumir en su mayoría suplementos de proteína disminuye la posibilidad de pérdida muscular, además de ser los que tienen en general un menor peso, talla, grasa y masa corporal (según las correlaciones de los componentes con las variables originales).

El proceso de bioimpedancia detecta menos grasa a mayor nivel de agua, además, metabólicamente, hay personas que tan sólo tener el hábito de tomar líquidos tiene menor grasa corporal y visceral, ya que la homeostasis interna del cuerpo es necesaria para controlar la grasa corporal (Arley, 2020), esta podría ser una razón del por qué el conglomerado 2 (que está compuesto en su mayoría por mujeres de alta hidratación), son las que tienen, en promedio, la menor cantidad de grasa corporal y visceral.

La cafeína es un potente estimulante del sistema nervioso central, y es normalmente usado para aumentar el estado de alerta-concentración, lo que promueve un trabajo con mayor fuerza, y, por ende, más estimulación muscular (Ortega, 2017). Además, la proteína es una suplementación reciente que aumenta dramáticamente la tasa de síntesis de proteína tras un esfuerzo y mejora la composición corporal al movilizar más masa grasa y aumenta la masa magra (Rodríguez, 2014). Todo lo anterior indica por qué en los atletas con el perfil tiene la agrupación indicada en el inciso 3. anterior.

En resumen, de acuerdo a los resultados encontrados se podría decir que los atletas con el perfil 2 son aquellos que, según la composición corporal, hidratación, entre otros, podrían llegar a tener un mejor rendimiento en las competencias de triatlón, mientras que los del perfil 3 al ser en general más pesados, grandes, tener mayor cantidad de grasa en general, son los que podrían tener un peor rendimiento. Por otro lado, los deportistas con el perfil 1 están muy cercanos al promedio general de todas las variables tomadas en cuenta en este estudio, por lo que no se puede mencionar que podría pasará con el rendimiento de estos. No obstante, la hidratación es un aspecto de suma importancia y puede, a largo plazo afectar el rendimiento de los mismos.

Se recomienda para futuros estudios relacionados con este tema evaluar el rendimiento de los atletas en las diferentes pruebas, ya sea mediante tiempos, esfuerzo, entre otras variables que pueden generar una mejor caracterización para los perfiles. Además, utilizar herramientas estadísticas como los análisis de varianza para determinar si existe diferencias entre los promedios de las variables antropométricas y de consumo de los distintos conglomerados o perfiles.

BIBLIOGRAFÍA

Arley, R. (2020). Relación entre la suplementación, hidratación, recuperación alimentaria post entrenamiento y la composición corporal en triatletas costarricenses categoría olímpica de 20 a 45 años, 2019. (Tesis de Licenciatura). Universidad Hispanoamericana, San José.

Baur, D. A., Bach, C. W., Hyder, W. J., & Ormsbee, M. J. (2016). Fluid retention, muscle damage, and altered body composition at the Ultraman triathlon. European Journal of Applied Physiology.

Cardinali, C. (2013). *Data Assimilation a data assimilation system*. Recuperado de: https://www.ecmwf.int/sites/default/files/elibrary/2013/16938-observation-influence-diagnostic-data-assimilation-system.pdf

Chang, N. Ravi, M. A. Plegue, K. R. Sonneville, M. M. Davis. (2016) Inadequate Hydration, BMI, and Obesity Among US Adults: NHANES 2009-2012. The Annals of Family Medicine, 14 (4)

Clavijo M., J. A., & Granada D., H. A. (2016). Una técnica de clasificación con variables categóricas. Ciencia En Desarrollo, 7(1), 15–20. https://doi.org/10.19053/01217488.4226

Esponda, D., Miranda, I., Nualles, M., & Fernández, L. (2010). Utilización del análisis de clusters con variables mixtas en la selección de genotipos de maíz. Revista Investigación Operacional, 30.

Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. DOI: 10.1093/bioinformatics/btv428

H. Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Kassambara & Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Logan-Sprenger*, H. M. (2019). Fluid balance and thermoregulatory responses of competitive triathletes. Journal of Thermal Biology, 79, 69–72.*

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.

Martínez, C. (2017). Nutrición y efectos de la suplementación ergonutricional en el fútbol. (Tesis doctoral). Instituto de biomedicina, Universidad de León, León.

Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

Ortega. (2017). ¿Cómo afecta la cafeína y el azúcar que contienen las bebidas energéticas, al rendimiento deportivo? (Tesis de grado). Universidad de las Islas Baleares, España.

Peña-Méndez, D. P. (2014). Análisis de componentes principales en la estimación de índices de empoderamiento en mujeres de Colombia. 1–49.

Pérez, M.J., Cabrera, W., Varela, G., Garaulet, M. (2010) Distribución regional de la grasa corporal. Uso de técnicas de imagen como herramienta de diagnóstico nutricional. Nutricion hospitalaria. Vol.25(2), pp.207-23

Pitarch, C., Lopez, C., & Castillo, E. (2013). Estudio de la alimentación y suplementación en jóvenes deportistas | Farmacéuticos Comunitarios. Recuperado el 23 de mayo de 2020, de: https://farmaceuticoscomunitarios.org/es/journal-article/estudio-alimentacion-suplementacion-jovenes-deportistas

Ramos, D.J., Rubio, J.A., Jiménez, J.F. (2015) Efectos sobre la composición corporal y la densidad mineral ósea de un programa de altitud simulada en triatletas. *Nutrición Hospitalaria, 32*(3), 1252-1260

Rodríguez, Y. (2014). Análisis de la suplementación con proteínas en el deporte: uso y efectos de la creatina y el suero de leche. (Tesis de grado). Universidad de León, León.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sellés, M.C., Martínez, J.M, Ayuso, J.M, Selles, S., Norte, A., Ortiz, R., Cejuela, R. (2015). Evaluación de la ingesta de líquido, pérdida de peso y tasa de sudoración en jóvenes triatletas. *Revista Española de Nutrición Humana y Dietética, 19*(3), 132-139.

Sinharay, S. (2010). An Overview of Statistics in Education. International Encyclopedia of Education. (3er edición, páginas 1-10). U.S: Editorial Board

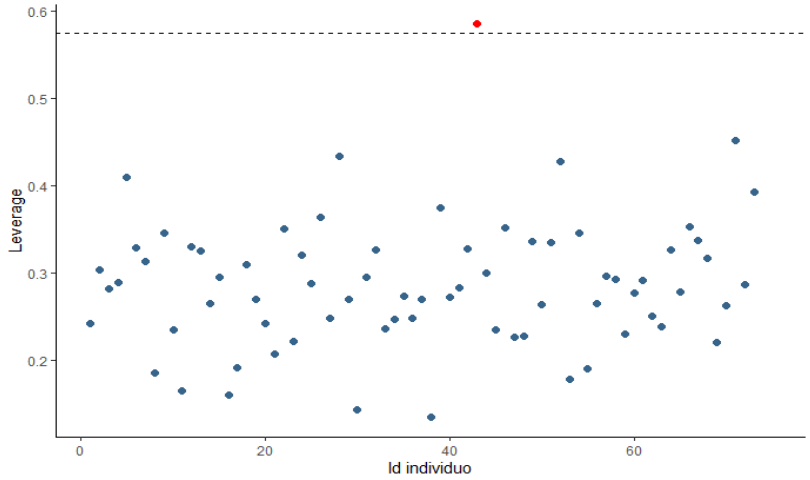
Under Armour. (2015). MyFitnessPal. (Versión 5.0) [Aplicación móvil]. Descargado de: https://play.google.com/store/apps/details?id=com.myfitnesspal.android&hl=es

Wright (2018). corrgram: Plot a Correlogram. R package version 1.13. https://CRAN.R-project.org/package=corrgram

ANEXOS

Anexo 1

Gráfico de leverages según individuo para detección de valores extremos



Anexo 2.

Resumen del análisis de componentes principales para las variables antropométricas.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Varianza explicada	4.78	1.28	0.45	0.22	0.13	0.09	0.04
Proporción de varianza explicada	0.69	0.18	0.06	0.03	0.02	0.01	<0.01
Proporción acumulada de varianza explicada	0.69	0.87	0.93	0.96	0.98	0.99	1

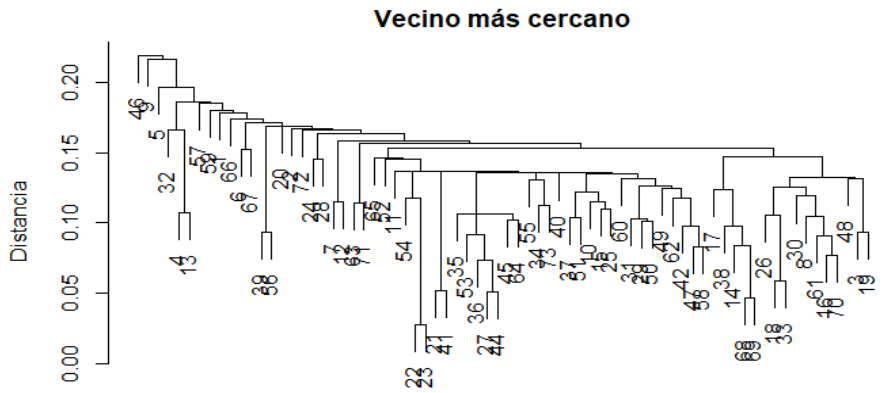
Anexo 3

Resumen del análisis de componentes principales para las variables de consumo post-entrenamiento.

	PC1	PC2	PC3	PC4
Varianza explicada	2.57	0.92	0.51	0.00
Proporción de varianza explicada	0.64	0.23	0.13	0.00
Proporción acumulada de varianza explicada	0.64	0.87	1.00	1.00

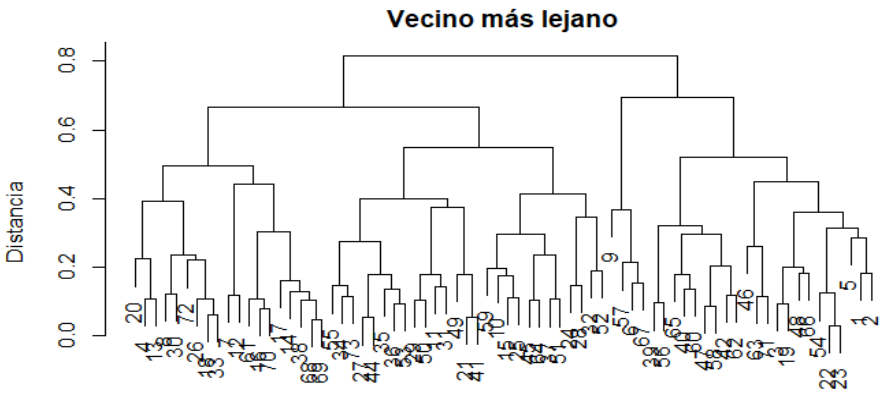
Anexo 4

Dendrograma con distancia entre individuos de gower y entre grupos del vecino más cercano.



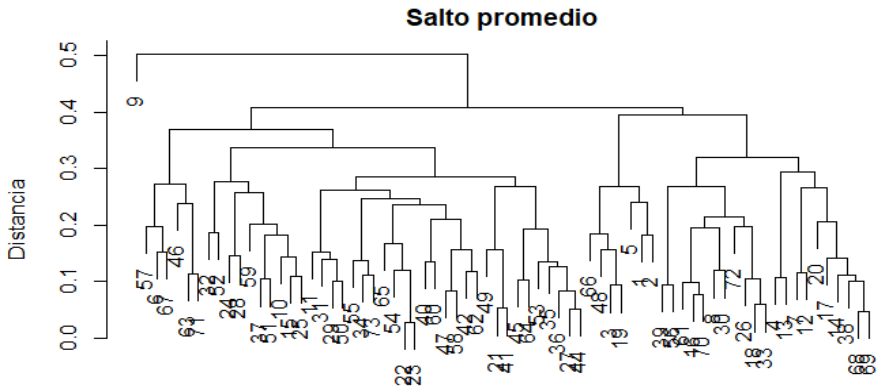
Anexo 5

Dendrograma con distancia entre individuos de gower y entre grupos del vecino más lejano.



Anexo 6

Dendrograma con distancia entre individuos de gower y entre grupos del salto promedio.



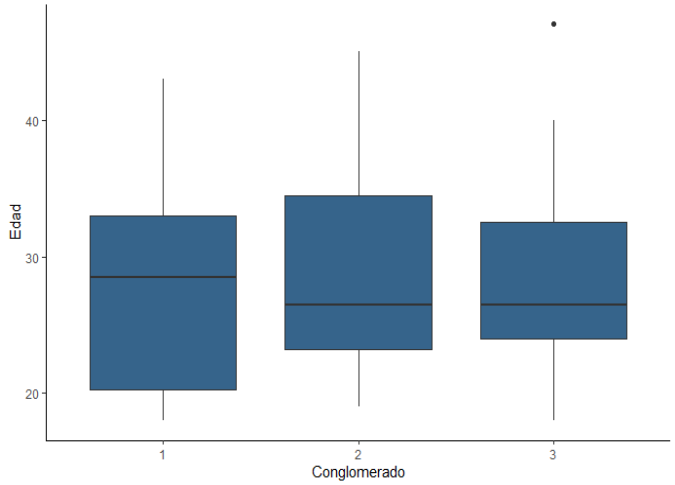
Anexo 7

Distribución absoluta y relativa de individuos entre los conglomerados.

Conglomerado	Tamaño	Porcentaje
Conglomerado 1	14	19.44
Conglomerado 2	30	41.67
Conglomerado 3	28	38.89
Total	**72**	**100**

Anexo 8

Comparación de edades según conglomerados.



V. MINERÍA DE DATOS PARA PREDICCIÓN

La minería de datos es el conjunto de técnicas y tecnologías que posibilitan manejar grandes bases de datos, de manera automática o semiautomática, con la intención de encontrar ciertos patrones repetitivos, tendencia o medidas que expliquen el comportamiento de un determinado contexto (Gibert, K., Ruiz, R., & Riquelme, J. ,2019)[12]. Los patrones desconocidos se clasifican en grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación). Cabe destacar que la minería de datos es muy utilizada en el análisis adicional y análisis predictivo. Dichas técnicas fueron apareciendo a principios de los años ochenta, por medio de Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, que empezaron a consolidar los términos de minería de datos (Virseda, F., & Román, J. 2010)[13].

[12] Gibert, K., Ruiz, R., & Riquelme, J. (2019). Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601, No. 29, 2006 (Ejemplar Dedicado a: Minería de Datos), Pags. 11-18. Recuperado de https://www.researchgate.net/publication/28140440_Presentacion_Mineria_de_Datos.

[13] Virseda, F., & Román, J. (2010). Minería de datos y aplicaciones. Revista Latinoamericana de Ingeniería de Software, 8. Recuperado de http://www.it.uc3m.es/jvillena/irc/practicas/06-07/22.pdf

Comparación de técnicas de clasificación para predecir el cumplimiento de pago de los créditos de 5 años en los primeros 2 años

Mónica Castrillo Gómez[14], José Flores Ramírez[14], Manrique Chacón Rojas[14]

moni_cg_97@hotmail.com, joseflore.jsf@gmail.com, man.c.r@hotmail.com

RESUMEN

La otorgación de créditos por parte de entidades financieras no es una tarea sencilla, por lo que clasificar a los solicitantes del crédito por medio de técnicas de clasificación es muy útil, con el fin de coadyuvar al área de control de riesgos y prever futuros problemas para la entidad. Para este trabajo se realizará un perfil del solicitante al crédito compuesto por seis variables asociadas al solicitante y una variable respuesta que indica si el solicitante cumplirá o no con el pago del crédito en dos años. Esto porque la entidad financiera determinó que los dos primeros años es el riesgo permitido para otorgar un crédito. Con el objetivo de clasificar a los solicitantes se utilizaron técnicas estadísticas de minería de datos (árboles de decisión, bagging, bosques aleatorios, k-vecinos más cercanos, máquinas vectoriales de soporte y regresión logística). Para cada una de las técnicas se realizó un proceso previo de calibración y seguidamente, un proceso de validación. Se encontró que la técnica más adecuada es la de árboles de decisión, debido al mejor rendimiento en los indicadores de error con un valor de 26.25%, AUC de 0.79 y de 46.95 para el KS.

PALABRAS CLAVE: crédito, minería de datos, árboles de decisión, k-vecinos más cercanos, máquinas vectoriales de soporte y regresión logística.

ABSTRACT

The provision of credit by financial institutions is not an easy task, so classifying credit applicants by rating techniques is very useful to assist the risk control area and to anticipate future problems for the institution. For this work, a profile of the applicant to the credit will be composed of six variables associated with the applicant and a variable response that indicates whether or not the applicant will fulfill the payment of the credit in two years; because the financial entity determined that the first two years is the risk allowed to grant a loan. Statistical data mining techniques (decision trees, bagging, random forests, k-nearest neighbors, vector support machines and logistic regression) were used to classify applicants. For each of the techniques, a pre-calibration process was carried out, followed by a validation process. Was found that the most suitable technique is decision trees, due to the best performance in error indicators with a value of 26.25%, AUC of 0.79 and 46.95 for the KS.

KEY WORDS: credit, data mining, decision trees, k-nearest neighbors, support vector machines and logistic regression.

INTRODUCCIÓN

Una entidad financiera se caracteriza por la toma de riesgos basada en la incertidumbre. Los riesgos, además de ser medidos, deben ser identificados y controlados, para evitar la quiebra de las instituciones

[14] Estudiante de Estadística de la Universidad de Costa Rica

financieras, con el objetivo de fijar las estrategias de mercado, que resulten en una ecuación favorable en el riesgo asumido y la recompensa obtenida midiendo así la rentabilidad del negocio. La valoración de estos riesgos consiste en medir el grado de variación de los resultados financieros de una empresa frente a los estimados, tomando en cuenta la volatilidad de los resultados, ya que entre más volátiles sean estos, mayor riesgo verá el ente financiero (Cardona, 2004).

Existen muchos tipos de riesgo como el de liquidez, legal, operativo, de mercado, de contraparte y de crédito. Para efectos de este trabajo se utilizará únicamente el riesgo crediticio, el cual define Cardona (2004) como: "la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos como consecuencia de que sus deudores fallen en el cumplimiento oportuno o cumplan imperfectamente los términos contractuales acordados" (Cardona, 2004, p.142).

En Costa Rica, el ente encargado de velar por el sistema financiero es la Superintendencia General de Entidades Financieras (SUGEF), cuya función principal es velar por la estabilidad y el funcionamiento eficiente del sistema financiero nacional (SUGEF, 2018). Cuando se tratase de riesgos financieros, este ente por medio del Consejo Nacional de Supervisión del Sistema Financiero (CONASSIF) rige las pautas que están publicadas en el Acuerdo SUGEF 12-10 Normativa para el Cumplimiento de la ley N° 8204ª (SUGEF, 2020). Una de las pautas para el otorgamiento de créditos compromete a las entidades a implementar una política denominada "conozca a su cliente".

Dicha política consiste en que, previo a iniciar la relación comercial, cada entidad debe obtener cierta información mínima del solicitante del crédito (como p. ej. nombre, fecha de nacimiento, cédula de identidad, entre otras), con el fin de salvaguardar el sistema financiero.

Adicionalmente, en el caso de que el solicitante desee obtener un crédito, una vez analizada y verificada su documentación, lo primero que deberá realizar es la solicitud del crédito para lo cual el ente financiero llevará a cabo un análisis con base en el perfil del solicitante, para determinar la existencia de riesgos.

Para efectos de este trabajo, el perfil de riesgo de cada solicitante de crédito estará compuesto por variables como la deuda de crédito, edad, salario, número de líneas crédito y préstamos, número de hipotecas, número de personas dependientes y si el solicitante cumplirá o no con el pago del crédito en dos años.

Por lo tanto, surge la importancia de clasificar a los solicitantes del crédito tomando en cuenta las variables mencionadas anteriormente, para determinar si un solicitante cumplirá con el pago del crédito. Esto es de gran importancia en una entidad financiera para identificar cuáles solicitantes cumplirán con el pago correspondiente, con el fin de coadyuvar al área de control de riesgos, prever futuros problemas, disminuir gastos administrativos, entre otros. Y a su vez, estar preparados para enfrentar las consecuencias y la incertidumbre sobre las variables que puedan afectar dichos resultados.

Para efectuar este análisis se utilizaron las siguientes técnicas de clasificación: árboles de decisión (bagging y bosques aleatorios), k-vecinos más cercanos, máquinas vectoriales de soporte y regresión logística.

METODOLOGÍA

En los últimos años, las financieras han dejado de lado los procesos empíricos y se han enfocado en metodologías apoyadas en procesos estadísticos para la toma de decisiones, por tanto, seguidamente se mostrarán algunas técnicas que se utilizaron en la medición y control del riesgo crediticio.

Para llevar a cabo el análisis se utilizó una base de datos de créditos con un total de 16 049 registros, con siete variables de las cuales seis son continuas y únicamente la variable respuesta binaria, entre ellas están: edad, salario, deuda de crédito, cantidad de líneas de crédito y préstamos, cantidad de hipotecas, número de personas dependientes y la variable objetivo del estudio, si el solicitante del crédito va a cumplir o no con el pago de un crédito de cinco años en los primeros dos años. En este caso, la entidad financiera determinó que los dos primeros años es el riesgo permitido para otorgar un crédito. Estos datos fueron suministrados por el economista Pedro Chacón, quien los recolectó en el año 2018 y son de carácter confidencial.

Cabe resaltar que el archivo no presenta valores faltantes y se observó que la mayoría de las variables están poco correlacionadas, siendo la más alta de 0.45 entre el número de hipotecas y el número de líneas de crédito y préstamo. También, se tiene que la distribución de la variable respuesta es de 43% para solicitantes que no podrán pagar el crédito, mientras que un 57% si lo podrá pagar.

Es importante mencionar que para cada una de las siguientes técnicas los mejores resultados para el AUC hacen referencia a valores altos (entre 0.6 y 0.9). Esto debido a que el AUC representa la probabilidad de que el modelo clasifique un positivo aleatorio más alto que uno negativo aleatorio. De igual manera, los mejores resultados de KS hacen referencia a valores que se encuentren alrededor 0.2 y 0.6 (especialmente entre más alto sea el valor dentro del rango es mejor). Es deseable obtener estos valores ya que es una medida que representa el ajuste del modelo con la distribución de los datos. Por último, los mejores resultados para el error hacen referencia a que es deseable obtener el menor valor de error posible.

La primera técnica de análisis que se utilizó fue árboles de decisión, este método es el de mayor uso en el sector financiero. Posee la gran ventaja de que es un método fácil de entender y de utilizar, especialmente para las personas que no cuentan con conocimientos avanzados en estadística, además que facilita la comprensión del conocimiento utilizado en la toma de decisiones y reduce el número de variables independientes (Pérez, 2011).

Primeramente, se realizó una partición de los datos, donde un conjunto corresponde a los datos de entrenamiento, el cual controla el proceso de predicción y otro conjunto que corresponde a los datos de evaluación que se encarga de predecir con respecto al perfil del solicitante del crédito si éste pagará o no en el plazo de los 2 años.

Cardona (2004) citando a Breiman, Friedman, Olshen y Stone (1984), define los árboles de decisión como un “método no paramétrico que no requiere de supuestos distribucionales, permite detectar interacciones, modela relaciones no lineales y no es sensible a la presencia de datos faltantes”, asimismo tiene como función principal generar particiones determinadas por ciertas reglas de clasificación de tipo “SI-ENTONCES”, hasta llegar a una clasificación final donde se podrá identificar cuál es la dimensión de la proporción de que un solicitante de crédito pague o no y de esta forma asignar la probabilidad que tienen los solicitante de crédito de pagar o no el crédito (Cardona, 2004).

Esta proporción se puede observar en los nodos terminales, que corresponden a las clasificaciones finales, por ejemplo, si se tiene 12 nodos finales esto quiere decir que existen 12 diferentes perfiles de riesgo y se podrá evaluar la distribución de los solicitantes del crédito entre si cumplirán o no con el pago del crédito y así poder realizar la toma de decisiones.

Para este método, en primer lugar, se realizó la calibración del modelo para buscar los parámetros más adecuados. De igual manera, el algoritmo de árboles de decisión permite utilizar la técnica de bosques aleatorios, que es un método que utiliza la agregación de bootstrap (bagging) para combinar diferentes árboles, donde cada árbol es construido con observaciones y variables aleatorias en cada nodo. Para esto, primeramente, se crea una muestra bootstrap del tamaño poblacional, se entrena una clasificación de árbol en la muestra en donde para cada nodo las variables predictoras son seleccionadas al azar entre todas las variables y la variable que proporciona la mejor división se utiliza para hacer una división binaria en el nodo. Realizando el ajuste de este modelo mediante la construcción de diferentes modelos. Posteriormente, se realizó la calibración de ambas técnicas.

La cuarta técnica empleada fue la de k-vecinos más cercanos (k-nn), un método de clasificación que se puede encontrar dentro del análisis de minería de datos. La idea básica de este algoritmo es que clasifica la clase más frecuente a la que pertenezcan los k-vecinos más cercanos (Izurieta & Moyano, 2019).

Al igual que árboles de decisión, este algoritmo utiliza un conjunto de entrenamiento para clasificar los nuevos ejemplos, para los cuales se calcula la distancia entre ellos. La distancia Euclidiana es la más utilizada para medir la proximidad entre el nuevo ejemplo y los que conforman los datos de entrenamiento (González *et al*., 2016).

Para llevar a cabo este algoritmo, se estandarizaron las variables. Luego, se realizó la calibración para determinar la cantidad de k-vecinos más adecuada, para lo que se consideró que, si k es muy grande el modelo se inclinara a asignar los elementos a la clase más grande, mientras que, si la cantidad k es pequeña el modelo será sensible a ruido o datos corruptos.

La quinta técnica que se utilizó fue máquinas vectoriales de soporte (SVM por sus siglas en inglés -Support Vector Machines-), tal como lo dice Sánchez (2005) citando a Vapnik y Cortés (1995), es una técnica de clasificación que utiliza algoritmos de aprendizaje supervisado y fue desarrollada por Vapnik, Cortés y su equipo AT&T.

En primer lugar, el algoritmo de SVM realiza un mapeo de los puntos de entrada a un espacio de características de una dimensión mayor (p. ej. si los puntos de entrada están en dos dimensiones entonces son mapeados por la SVM a tres dimensiones) y encuentra un hiperplano que los separe y maximice el margen m entre las clases en este espacio como se aprecia en la figura 1 (Betancourt, 2005). El mapeo es realizado por medio de un kernel (este puede ser sigmoidal, lineal, radial o polinomial), en otras palabras, a partir de unos inputs de entrada al modelo, se etiquetan las clases y se entrena una SVM construyendo un modelo que sea capaz de predecir la clase de los nuevos datos que se introduzcan al modelo. Es posible realizar la predicción, ya que los nuevos datos son introducidos al modelo, estos se colocan sobre el mismo eje y en función de la cercanía de los grupos antes separados, los cuáles serán clasificados en una u otra clase (Sánchez, 2015).

Entre las ventajas de este método, se tiene que no es necesario la totalidad de puntos disponibles para hallar una solución al problema de maximización de la separación entre clases; no hay óptimo local, como en las redes neuronales; el entrenamiento es relativamente fácil; se escalan bien para datos en espacios dimensionales altos; entre otros (Betancourt, 2005). Al igual que en los métodos anteriores, se realizó un procedimiento de calibración para elegir los parámetros antes de efectuar el análisis.

La sexta y última técnica utilizada fue regresión logística, una técnica igualmente útil debido a que la variable respuesta es dicotómica y hace referencia al número de éxitos (personas que si van a cumplir con el pago del crédito) y el número de fracasos (personas que no cumplirán). Está técnica se encuentra dentro de los modelos lineales generalizados, o GLM por sus siglas en inglés, debido a que generaliza la regresión ordinaria permitiendo que la variable respuesta tenga distribuciones diferentes a la normal y, por otro lado, incluyendo distintas funciones del enlace. El logaritmo consiste en predecir el resultado de la variable respuesta dependiendo de las variables independientes, donde se utiliza una función de enlace que permite conectar un predictor lineal a la media natural de la variable respuesta, con el fin de encontrar una función lineal de los parámetros para poder predecir la respuesta. Para llevar a cabo este algoritmo se utilizó la distribución binomial para la variable dependiente o respuesta, la función de enlace conocida como logit y el predictor lineal está compuesto por las 6 variables explicativas.

Por último, se procedió a evaluar cada uno de los modelos seleccionados mediante la validación cruzada para el error, la curva ROC (Característica Operativa del Receptor), por medio de la biblioteca ROCR (Sing, Sander, Beerenwinkel, N et al, 2005) que muestra gráficamente el rendimiento de un modelo de clasificación, al contrastar los falsos positivos contra la precisión positiva, y el método Kolmogorov-Smirnov (KS), que calcula la diferencia entre distribuciones acumuladas relativas de las clases, ordenando de manera ascendente según la probabilidad predicha por dicho modelo de clasificación.

Para analizar las seis técnicas de clasificación se utilizó el programa estadístico R, versión 4.0.0 (R Core Team, 2020), haciendo uso del entorno de programación R Studio, en la versión 1.1.447 y con ayuda de las bibliotecas: rattle (Williams, 2011), bitops (Mutuy & Maechler, 2013), tibble (Müller & Wickham, 2020), DT (Xie, Cheng & Tan, 2020), plotly (Sievert, 2020) y caret (Kuhn, 2020). Asimismo, la técnica de árboles de decisión se realizó por medio de la biblioteca rpart (Therneau & Atkinson, 2019), las técnicas bagging y bosques aleatorios fueron desarrollado mediante las bibliotecas adabag (Alfaro, Gamez & García, 2013) y randomForest (Liaw & Wiener, 2002), respectivamente. La técnica de K- vecinos más cercanos se desarrolló con ayuda de la biblioteca kknn (Schliep & Hechenbichler, 2016). Finalmente, para llevar a cabo el algoritmo de máquinas vectoriales de soporte se utilizó la biblioteca e1071 (Bennett & Campbell, 2000).

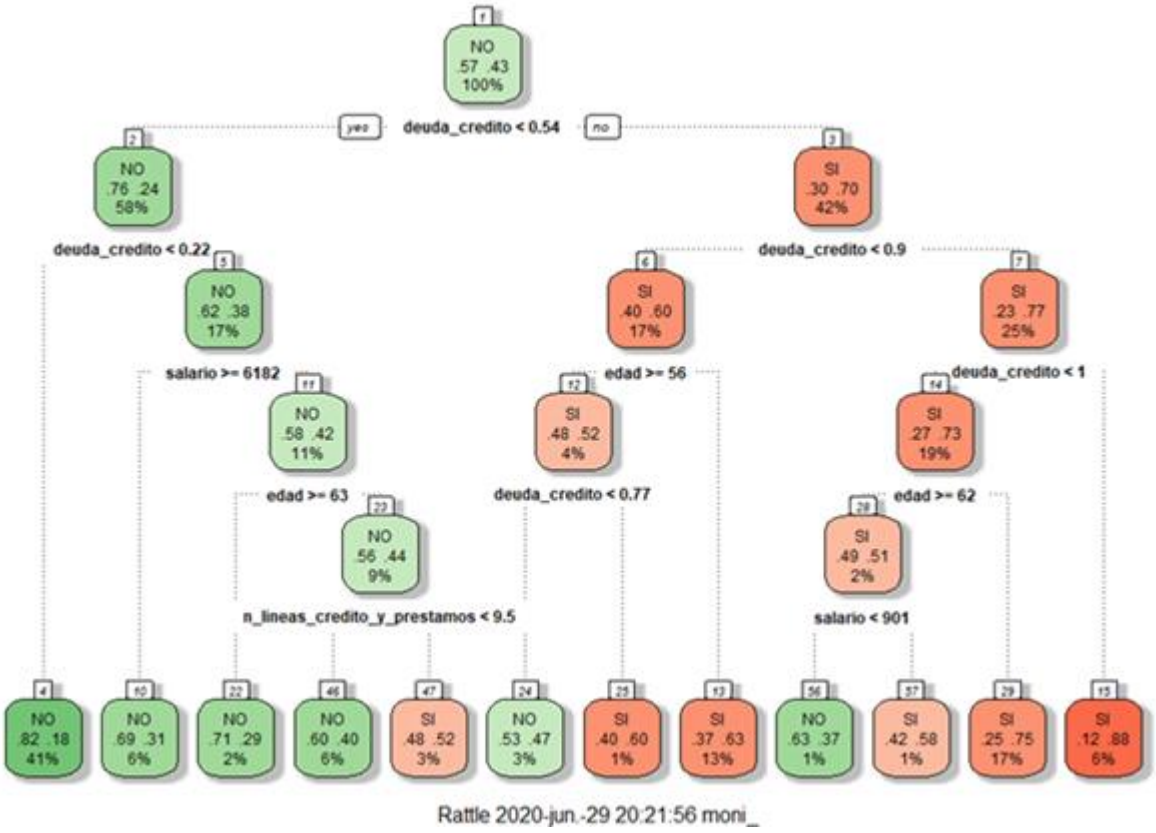
RESULTADOS

Para el análisis del método de árboles de decisión, primero se realizó la calibración del modelo fijando los parámetros del número mínimo de observaciones para que el nodo se pueda dividir (minsplit) en 1%* 12 839 = 128 y el número mínimo de observaciones que debe tener un nodo para ser considerado como terminal (minbucket) en 0.5% * 12 839 = 64, luego se procedió a la calibración del parámetro de complejidad (cp) (ver tabla 1), donde se encontró mejores resultados del error, el área bajo la curva (AUC) y Kolmogorov-Smirnov (KS) en el valor de cp = 0.0005 y por último se calibró el parámetro de profundidad máxima (maxdepth) (ver tabla 2) eligiendo un valor de 5, el cual arrojó los valores más altos para AUC y KS, en comparación con los otros.

Al calibrar, se obtuvo que las variables que aportan más al modelo fueron: deuda crédito, edad, salario y número de líneas de crédito y préstamos. En la siguiente figura se muestra la forma del árbol de decisión con las variables mencionadas anteriormente.

Figura 2

Forma del árbol de decisión.



Primeramente, en la figura anterior se observó por ejemplo que un solicitante del crédito cumple con el pago si su deuda crédito es menor a 0.22, donde un 41% de los solicitantes del crédito cumplen con el pago, sin embargo, de ese 41% existe una probabilidad de 0.82 de que el solicitante del crédito cumpla con el pago y una probabilidad de 0.18 de que incumpla con el pago. Por el contrario, si el solicitante del crédito tiene una deuda crédito mayor o igual a 0.22, pero menor a 0.54, un salario menor a $6182, una edad menor a 63 años y su número de líneas de crédito es mayor o igual a 9.5 entonces el solicitante del crédito va a incumplir con el pago, en este nodo se observa que un 3% de los solicitantes del crédito incumple con el pago donde existe una probabilidad de 0.48 de cumplir con el pago y un 0.52 de incumplir con este.

Posterior a la calibración, se realizó la matriz de confusión y se obtuvo una precisión de 73.98%, un error de 26.01%, un AUC de 0.7821 y un KS de 47.10 (ver figura 3 y 4), que en contraste con validación cruzada mediante la realización de 10 iteraciones para evaluar el rendimiento del modelo se obtuvo un AUC 0.7880. El cual indica que en el 78.80% de las veces un caso aleatorio seleccionado del grupo positivo (cumplimiento en el pago) tiene una probabilidad mayor de ser clasificado como positivo que como un caso negativo (incumplimiento en el pago), un KS de 46.95 y un error de 26.25% (ver figura 5) evidenciando así que el modelo

tiene un buen ajuste y presenta un rendimiento adecuado para la clasificación, es decir, puede predecir adecuadamente.

Seguidamente se realizó el método de agregación de bootstrap (bagging), iniciando con la calibración. Se utilizaron los mismos parámetros definidos en el método de árboles de decisión y se procedió a calibrar el parámetro de número de árboles (mfinal) (ver Tabla 3), donde se eligió el valor 19, ya que arrojó el mejor error, AUC y KS.

De igual manera, se obtuvo la matriz de confusión con una precisión de 73.89%, un error de 26.10%, un AUC de 0.7406 y un KS de 46.51 (ver figura 6 y 7); al realizar la validación cruzada se obtuvo un error de 26.64% lo que demuestra un aumento mínimo, un AUC de 0.7377 y un KS de 46.39 indicando una disminución mínima en comparación al modelo completo (ver figura 8). No obstante, queda en evidencia que este modelo está clasificando correctamente.

Para realizar la calibración en el método de bosques aleatorios, se utilizaron las reglas duras del tamaño del nodo terminal (nodesize) en 50 y el máximo de número de nodos (maxnodes) en NULL y se calibró el número de árboles (ntree) (ver tabla 4), el cual se fijó en 500, ya que se obtenía la mejor combinación de AUC, error y KS y para finalizar se agregó el número de variables aleatorias (mtry) donde fue la raíz cuadrada de siete.

Luego, se realizó la matriz de confusión y la precisión fue de 73.67%, el error de 26.33%, el AUC de 0.8809 y el KS de 56.58 (ver Figura 9 y 10), que en contraste con la validación cruzada presenta menores valores tanto de AUC (0.7970) como de KS (46.45) y un error de 26.70%; evidenciando que este modelo tiene un buen ajuste y está clasificando adecuadamente (ver figura 11).

Para el método de k-vecinos más cercanos, primeramente, se realizó la calibración utilizando los kernels rectangular, triangular, epanechnikov, gaussian, rank y optimal, siendo el kernel rectangular el que presentó los mejores resultados (ver figura 12). Seguidamente se procedió a seleccionar la cantidad de k-vecinos utilizando la curva ROC, para lo cual se concluyó emplear un k=13 (ver figura 13), debido a que presentaba los mayores valores de KS y AUC (ver figura 14 y 15).

Posteriormente, se procedió analizar la matriz de confusión del modelo y se encontró una precisión de 51.83%, un error de 48.16% y la medida del AUC correspondió a 0.7453 (ver Figura 14). Cuando se utilizó Kolmogorov-Smirnov para observar la máxima diferencia entre las distribuciones acumuladas relativas de las distintas clases, se obtuvo que esta fue de 35.98 (ver figura 15).

Al realizar la validación cruzada, se observó que el error resultante fue de 40.82%, comparándolo con el error obtenido con el modelo anterior, se observó como el de validación cruzada disminuyó 7.34%. El valor de AUC fue de 0.7463, teniendo un aumento de 0.001 con respecto al modelo anterior. Finalmente, el KS tuvo una disminución de 0.08 siendo el KS del modelo validado de 35.90 (ver figura 16).

Al analizar la técnica de máquinas vectoriales de soporte, lo primero que se realizó fue la calibración del modelo probando distintos tipos de kernels, esto con el fin de conocer cuál arrojaba mejores resultados. En otras palabras, se verificó cuál arrojaba valores menores de error y valores mayores tanto de KS como de AUC. Para esto se probaron cuatro tipos de kernel: sigmoidal, linear, polinomial y radial, como se puede observar en la tabla 5.

Tabla 5

Comparación de Kernels Mediante Indicadores de Desempeño.

KERNEL	AUC	ERROR	KS
Sigmoidal	0.47	46.84	3.64
Lineal	0.64	39.43	21.76
Radial	0.69	35.85	27.60
Polinomial	0.65	41.07	21.90

El kernel que proporciona mejores resultados al comparar el error, el KS y el AUC es el radial, por lo que se decidió continuar el análisis con este kernel. Seguidamente, al analizar el modelo se observó que la matriz de confusión arrojó una precisión de 63.46%, con un error de 36.54%. Además, la medida de clasificación de ROC fue de 0.6854, siendo una medida localizada en un punto medio de lo ideal (ver figura 17). También se observó que al utilizar Kolmogorov-Smirnov se obtuvo un valor de 27.59, el cual es relativamente bajo (ver figura 18).

Al realizar la validación cruzada y comparar el error, AUC y KS obtenidos en el modelo general, se aprecia que el valor del error permanece prácticamente igual pasando de 36.54% a 36.64%; el valor del AUC disminuye de 0.68 a 0.64 y en el caso del valor de KS, también disminuye, pasando de 27.59 a 24.34 (ver figura 19). Por lo tanto, la validación arrojó valores menores a los que se obtuvo en el modelo general.

Asimismo, para la última técnica de clasificación, la regresión logística, se procedió de una manera similar, lo único distinto es que para esta técnica no es necesario calibrar algún parámetro desde el inicio. En primer lugar, se analizó el modelo y al observar la matriz de confusión se encontró una precisión de apenas 61.74%, con un error del 38.25%.

Del mismo modo, la medida de clasificación de ROC fue de 0.6651, siendo una medida localizada en un punto medio de lo ideal (ver figura 20). También se observó que al utilizar Kolmogorov-Smirnov se obtuvo un valor de 25.19, el cual es relativamente bajo (ver figura 21).

A la hora de realizar validación cruzada y comparar el error, AUC y KS obtenidos en el modelo general, se aprecia que el valor del error aumenta pasando de 38.25% a 39.49%; para los valores de AUC y KS, ambos presentaron una disminución pasando de 0.66 a 0.64 y de 25.19 a 22.79, respectivamente (ver figura 22). Por lo tanto, la validación arrojó valores menores a los que se obtuvo en el modelo general.

Por último, se procedió a comparar las técnicas utilizadas, con el propósito de determinar cuál es la que clasifica de mejor manera los solicitantes del crédito de acuerdo al cumplimiento del pago del crédito. Para llevar a cabo la comparación y determinar cuál técnica es la mejor, se seleccionan tres indicadores de desempeño como se ha visto a lo largo del escrito (error, curva ROC: para la cual se utiliza el valor del AUC y Kolmogorov-Smirnov-KS-). Se presenta la tabla 6 con los valores del error, curva ROC y KS utilizando validación cruzada

Tabla 6

Comparación de Técnicas de Clasificación Mediante Indicadores de Desempeño Utilizando Validación cruzada.

TÉCNICA	ERROR	CURVA ROC	KS
Árboles de decisión	**26.25**	**0.79**	**46.95**
Bagging	26.64	0.74	46.39
Bosques aleatorios	26.70	0.79	46.84
K-Vecino más cercano	40.82	0.75	35.90
Máquinas Vectoriales de Soporte	36.64	0.64	24.34
Regresión Logística	39.49	0.64	22.79

CONCLUSIÓN

De las seis técnicas de clasificación utilizadas, la que mostró mejores resultados para predecir si el solicitante va a poder realizar el pago del crédito dentro de los primeros dos años, es la de árboles de decisión. Esto quiere decir que, este método es el que presenta resultados más bajos del error y que, además, posee valores más altos tanto para el área bajo la curva de ROC (AUC) como para el KS al compararlos con los cinco métodos restantes (ver Tabla 7.)

Tabla 7

Indicadores de Desempeño para el Método de Árboles de Decisión

ERROR	FP	FN	PRECISIÓN	PP	PN	AP	AN
26.01	21.96	31.36	73.99	68.64	78.04	70.31	76.65

Asimismo, tanto para el modelo completo como para validación cruzada se presentaron los mejores valores. Si se realizan conclusiones basadas en el modelo general sin ser validado se estaría cometiendo un error, ya que para calibrar es necesario validar. Esta validación tiene como objetivo ajustar el modelo mediante los valores de los indicadores de desempeño. De esta forma, se observó que tanto el error como el AUC aumentaron y el KS disminuyó, sin embargo, el modelo seleccionado siguió siendo el que proporcionaba los mejores resultados comparado con el resto.

Esto quiere decir que, al realizar la validación cruzada se corrigen los valores para los distintos indicadores. Por lo que, al tener como objetivo principal clasificar correctamente a los solicitantes del crédito, se estableció que la validación cruzada para el método de clasificación de árboles de decisión es la más adecuada para tratar la clasificación de los solicitantes en el cumplimiento del pago del crédito en dos años. Esto es de suma importancia ya que como menciona Cardona (2004) “la importancia de tener un modelo de cálculo de probabilidad de incumplimiento confiable y con una alta capacidad de discriminación radica en que esta impacta

considerablemente en el cálculo de provisiones, afectando directamente el balance y las utilidades que podría llegar a tener la entidad. Adicionalmente, como los modelos son empleados para el otorgamiento de créditos, hacen parte fundamental de la gestión de riesgo, por lo tanto, un cálculo u operación inapropiada podría llevar a una institución financiera a situaciones de insolvencia" (Cardona, 2004).

BIBLIOGRAFÍA

Alfaro, E., Gamez, M & García, N. (2013). Adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software, 54(2), 1-35. Recuperado de http://www.jstatsoft.org/v54/i02/

Betancourt, G. (2005). Las máquinas de soporte vectorial (SVMs). Scientia et Technica Año XI. 27. http://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895/4139

Cardona, P. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. Revista Colombiana de Estadística. 2(27). 139-151.

Dutky,S & Maechler,M (2013). bitops: Bitwise Operations. R package version 1.0-6. Recuperado de https://CRAN.R-project.org/package=bitops

González, H., Santos, G., Campos, F., & Morell, C. (2016). Evaluación del algoritmo KNN-SP para problemas de predicción con salidas compuestas. Revista Cubana de Ciencias Informáticas, 10(3), 119-129. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992016000300009&lng=es&tlng=es

Izurieta, G. & Moyano, R. (2019). Predicción de clientes potenciales utilizando el algoritmo k vecino más cercano en el área de negocios de la COAC "Riobamba" ltda. Universidad Nacional De Chimborazo, Riobamba, Ecuador. http://dspace.unach.edu.ec/bitstream/51000/6043/1/UNACH-EC-ING-SIT-COMP-2019-0010.pdf

Kuhn,M (2020). caret: Classification and Regression Training. R package version 6.0-86. Recuperado de https://CRAN.R-project.org/package=caret

Liaw,A & Wiener,M (2002). Classification and Regression by randomForest. R News 2(3), Recuperado de https://CRAN.R-project.org/doc/Rnews/

Meyer,D; Dimitriadou,E; Hornik,K; Weingessel,A & Leisch,F (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3. Recuperado de https://CRAN.R-project.org/package=e1071

Müller,K & Wickham,H (2020). tibble: Simple Data Frames.R package version 3.0.1. Recuperado de https://CRAN.R-project.org/package=tibble

Pérez, C. (2011). Técnicas de segmentación. Conceptos, herramientas y aplicaciones. Madrid: Gaceta Grupo Editorial.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de https://www.R-project.org/.

Sánchez, N. (2015). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. ODEON, 9, pp. 131-143. DOI: http://dx.doi.org/10.18601/17941113.n9.04

Schliep,K & Hechenbichler,K (2016). kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. Recuperado de https://CRAN.R-project.org/package=kknn

Sievert,C.(2020) Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020. Recuperado de https://plotly-r.com

Sing T; Sander O; Beerenwinkel N & Lengauer T (2005). "ROCR: visualizing classifier performance in R." _Bioinformatics_, *21*(20), 7881. Recuperado de http://rocr.bioinf.mpi-sb.mpg.de>.

SUGEF. (2018). Objetivos y funciones. https://www.sugef.fi.cr/sugef/objetivos_funciones.aspx#:~:text=de%20la%20colectividad.-,Funciones,las%20entidades%20bajo%20su%20control.&text=%2DPresentar%20informes%20de%20sus%20actividades,de%20Supervisi%C3%B3n%20del%20Sistema%20Financiero.

SUGEF. (2020). Acuerdo SUGEF 12-10 Normativa para el cumplimiento de la ley N°8204ª.https://www.sugef.fi.cr/normativa/normativa_vigente/SUGEF%2012-10%20(v17%2022%20de%20mayo%20de%202020)%202.pdf

Therneau,T & Atkinson,B (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. Recuperado de https://CRAN.R-project.org/package=rpart

Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!, Springer.Recuperado de http://www.amazon.com/gp/product/1441998896/ref=as_li_qf_sp_asin_tl?ie=UTF8&tag=togaware-20&linkCode=as2&camp=217145&creative=399373&creativeASIN=1441998896

Xie,Y; Cheng,J & Tan ,X (2020). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.13. Recuperado de https://CRAN.R-project.org/package=DT

ANEXOS

Tabla 1

Calibración del parámetro de complejidad (cp) para la técnica de árboles de decisión.

CP	ERROR	AUC	KS
0,0005	**0,2657**	**0,8048**	**51,2374**
0,0006	0,2699	0,8004	50,6048
0,0007	0,2667	0,7941	49,3647
0,0008	0,2676	0,7886	48,8161
0,0009	0,2655	0,7774	47,8483

Tabla 2

Calibración del parámetro de profundidad máxima (maxdepth) para la técnica de árboles de decisión.

MAXDEPTH	ERROR	AUC	KS
1	0,2660	0,7277	45,5394
5	**0,2669**	**0,7870**	**47,2785**
6	0,2653	0,7858	46,4379
7	0,2663	0,7857	46,2844
10	0,2667	0,7834	46,2877
15	0,2650	0,7837	46,4790

Tabla 3

Calibración del número de árboles (mfinal) para la técnica bagging.

MFINAL	ERROR	AUC	KS
5	0,2607	0,7371	46,9489
10	0,2617	0,7433	47,4270
15	0,2607	0,7448	47,5525
19	0,2617	0,7545	47,4657
23	0,2617	0,7454	47,4824
30	0,2611	0,7461	47,5525

Tabla 4

Calibración del número de árboles (ntree) para la técnica bosques aleatorios.

NTREE	ERROR	AUC	KS
500	**0,2647**	**0,8811**	**56,5565**
700	0,2646	0,8809	56.4531
600	0,2654	0,8809	56,5374

Figura 1

Frontera de decisión para Máquinas Vectoriales de Soporte.

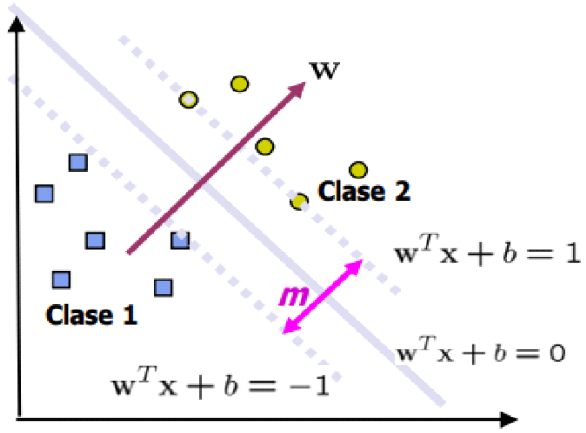


Figura 3

Indicador de desempeño AUC para el método de clasificación de árboles de decisión.

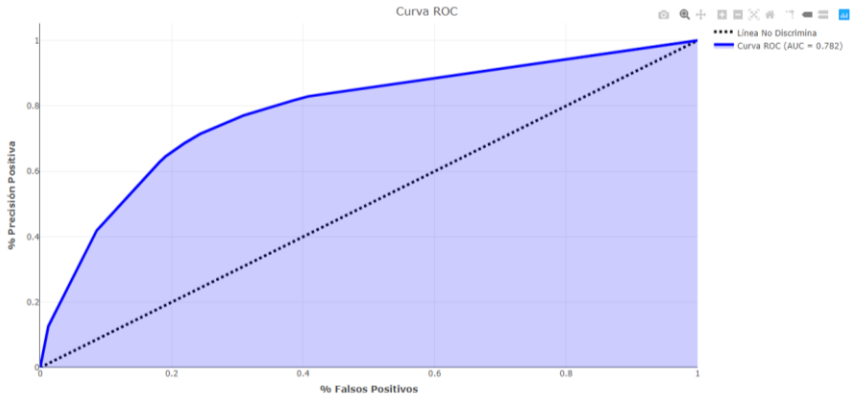


Figura 4

Indicador de desempeño KS para el método de clasificación de árboles de decisión.

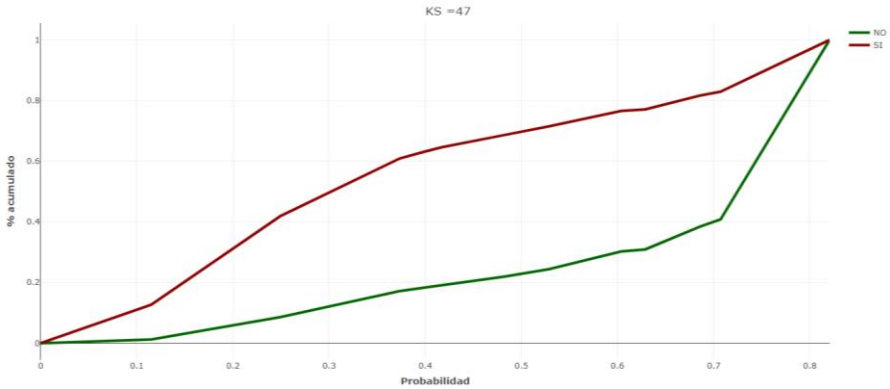


Figura 5

Indicadores de desempeño para el método de clasificación de Árboles de Decisión en validación.

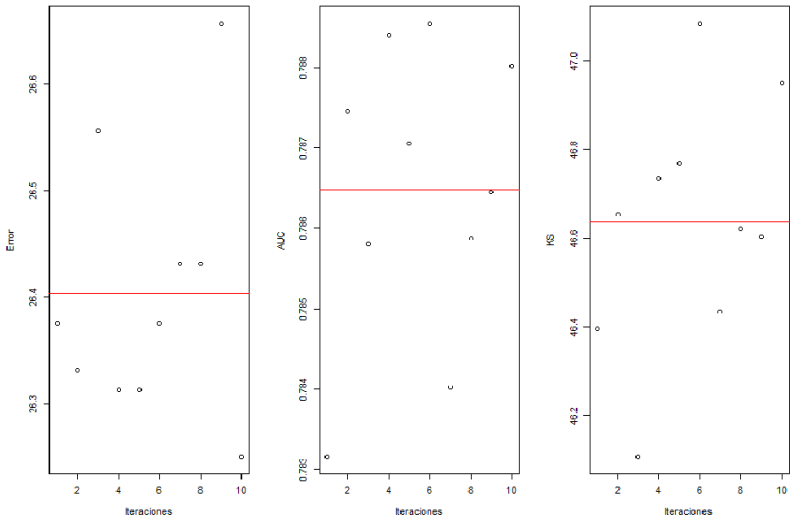


Figura 6

Indicador de desempeño AUC para el método de clasificación de Bagging.

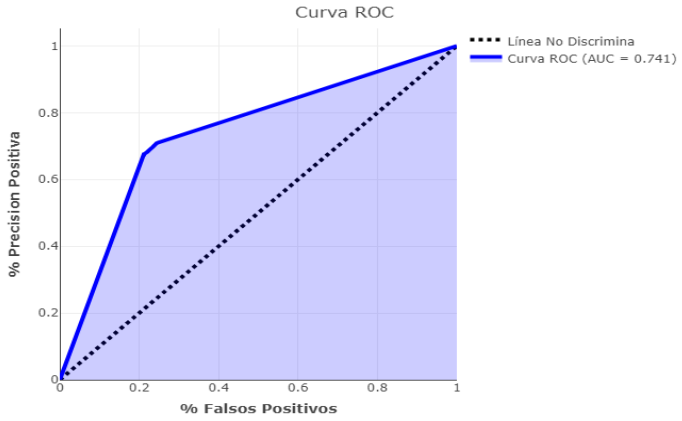


Figura 7

Indicador de desempeño KS para el método de clasificación de Bagging.

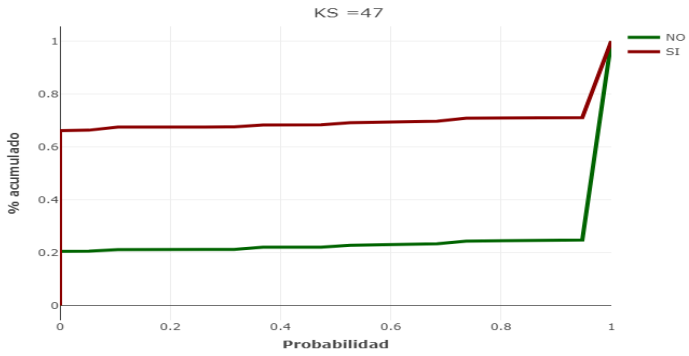


Figura 8

Indicadores de desempeño para el método de Bagging en validación cruzada con 10 iteraciones.

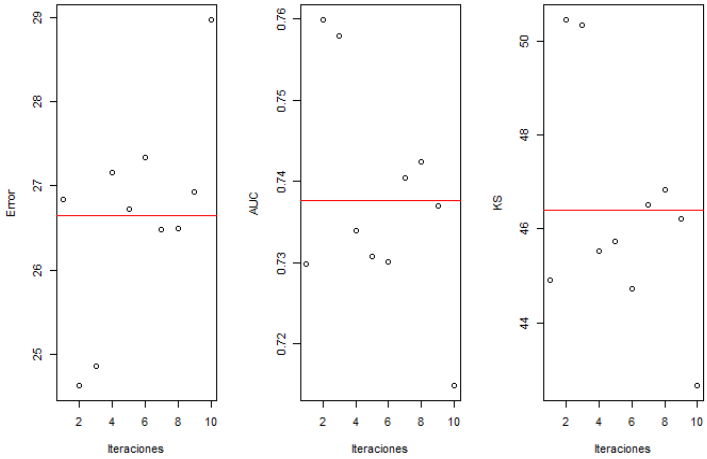


Figura 9

Indicador de desempeño AUC para el método de clasificación de Bosque Aleatorios.

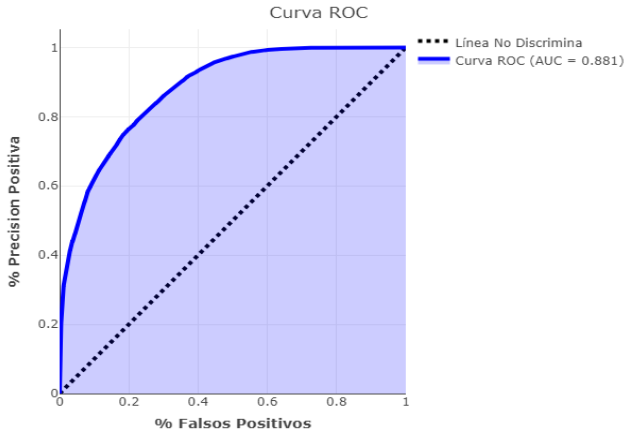


Figura 10.

Indicador de desempeño KS para el método de clasificación de Bosque Aleatorios.

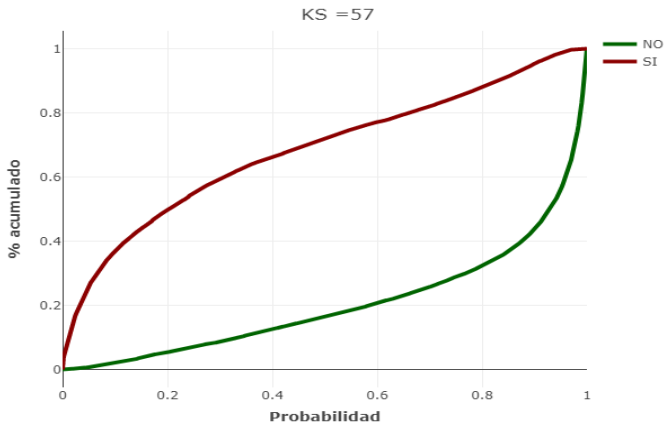


Figura 11

Indicadores de desempeño para el método de Bosques Aleatorios en validación cruzada con 10 iteraciones.

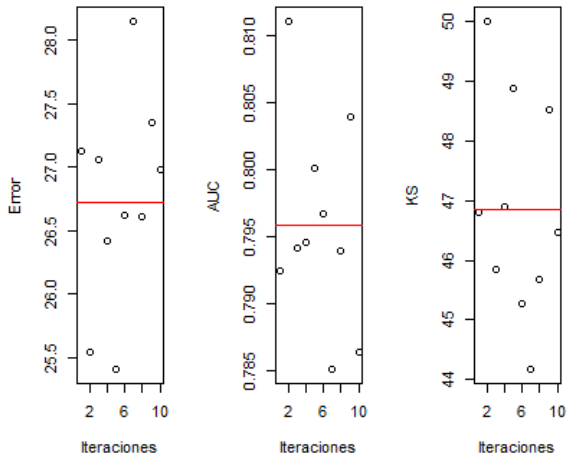


Figura 12

Comparación de Kernels.

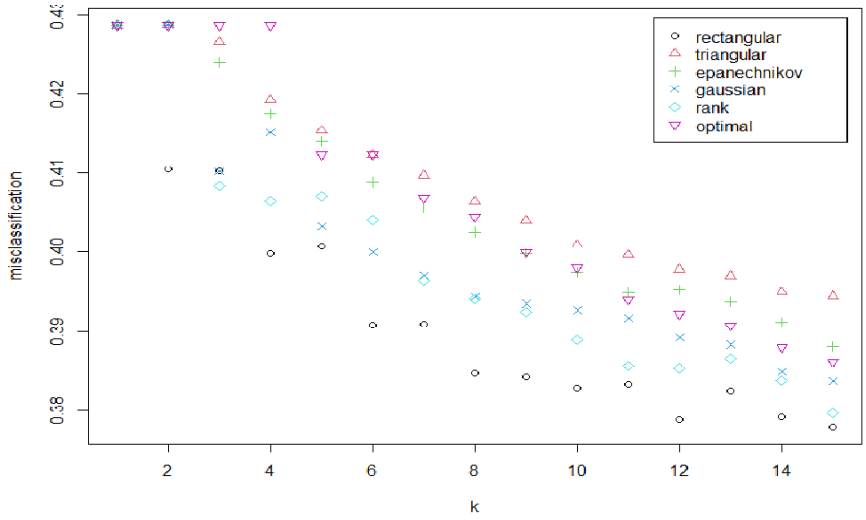


Figura 13

Comparación de los distintos valores de K.

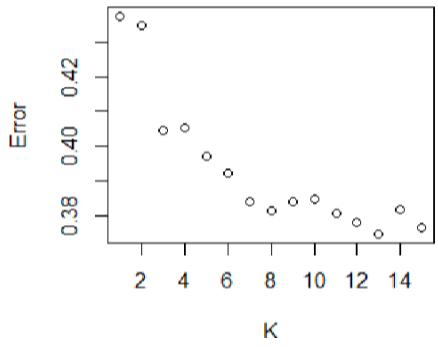


Figura 14

Indicador de desempeño de AUC para el método de clasificación K- vecinos más cercanos.

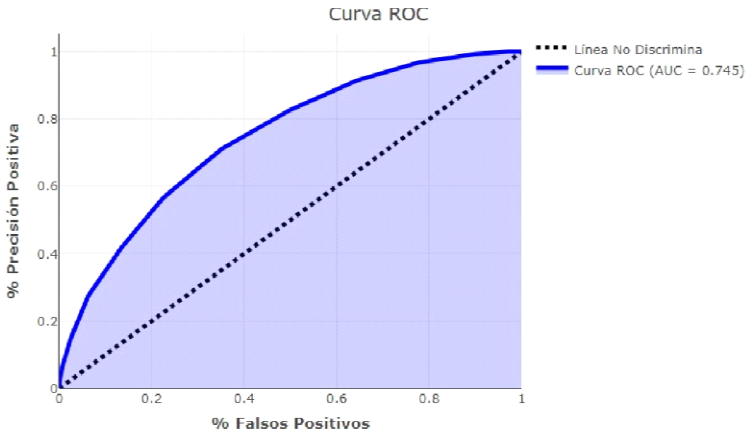


Figura 15

Indicador de desempeño de KS para el método de clasificación K vecinos más cercanos.

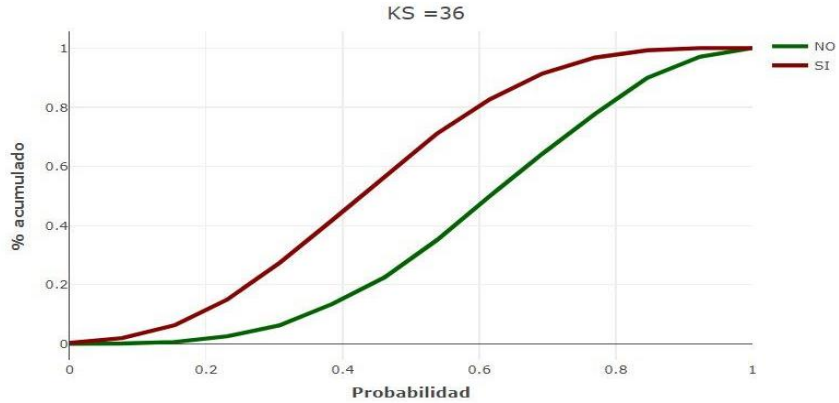


Figura 16

Indicadores de desempeño para el método K -vecinos más cercanos en validación cruzada con 10 iteraciones.

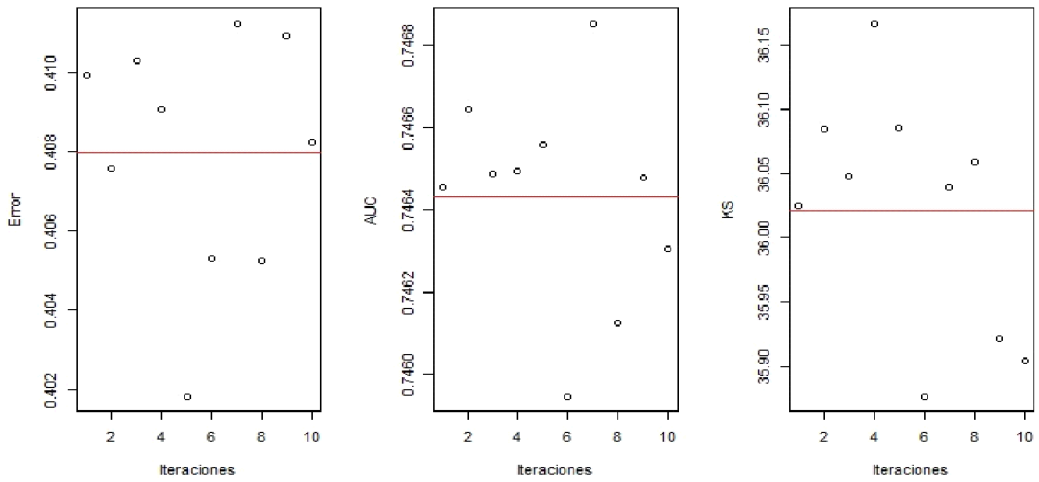


Figura 17

Indicador de desempeño AUC para el método de clasificación: Máquinas Vectoriales de Soporte.

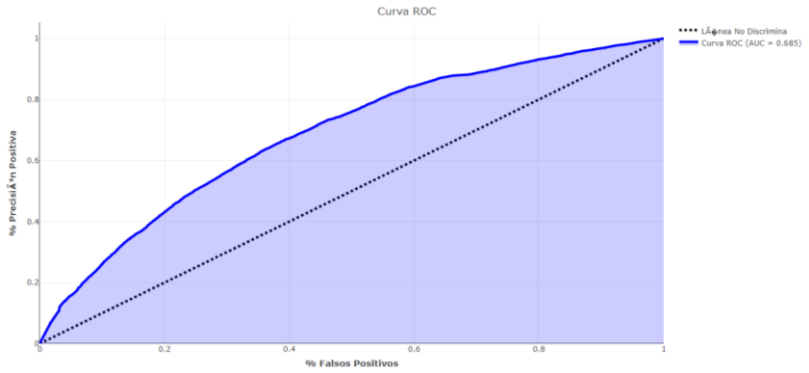


Figura 18

Indicador de desempeño KS para el método de clasificación: Máquinas Vectoriales de Soporte.

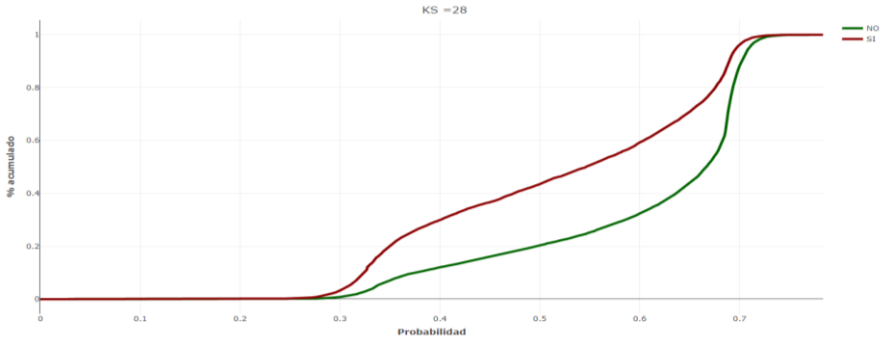


Figura 19

Indicadores de desempeño para el método Máquinas Vectoriales de Soporte con validación cruzada y 10 iteraciones.

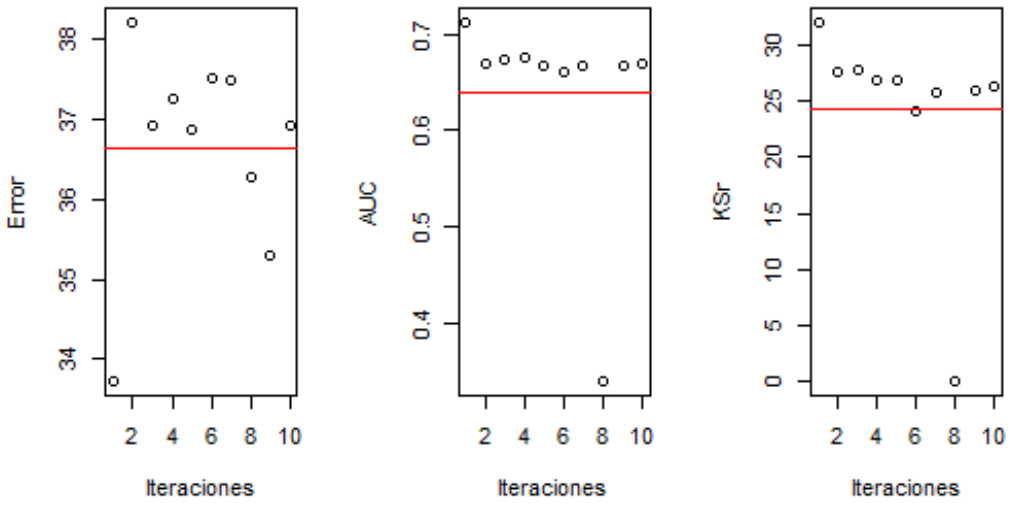


Figura 20

Indicador de desempeño AUC para el método de clasificación: Regresión Logística.

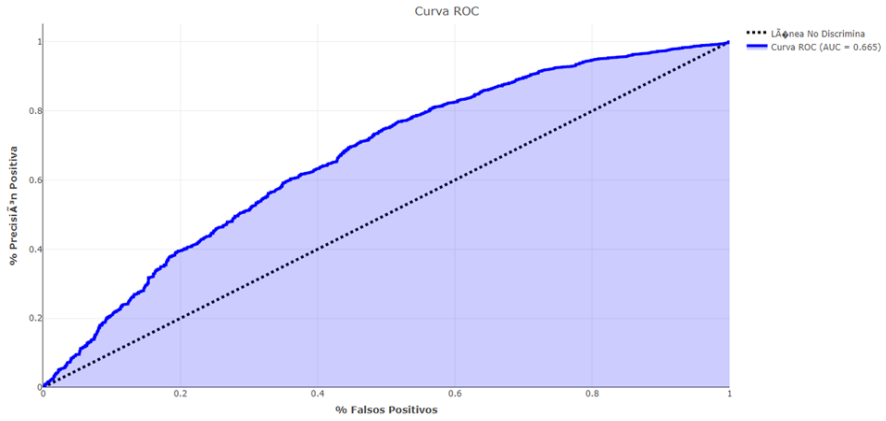


Figura 21

Indicador de desempeño KS para el método de clasificación: Regresión Logística.

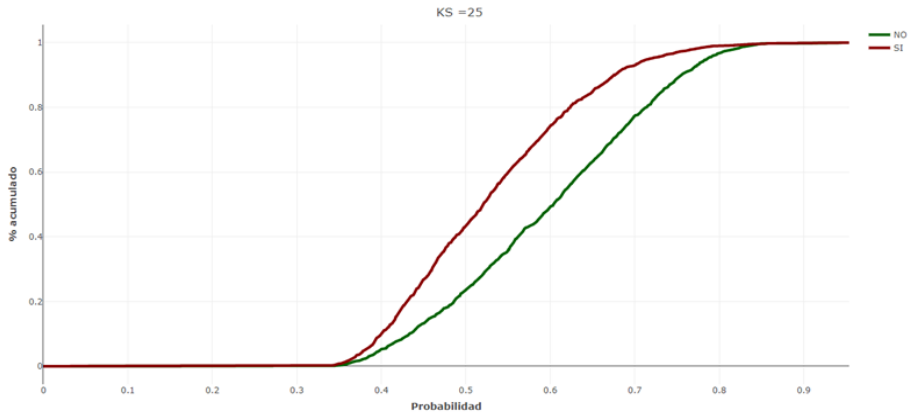
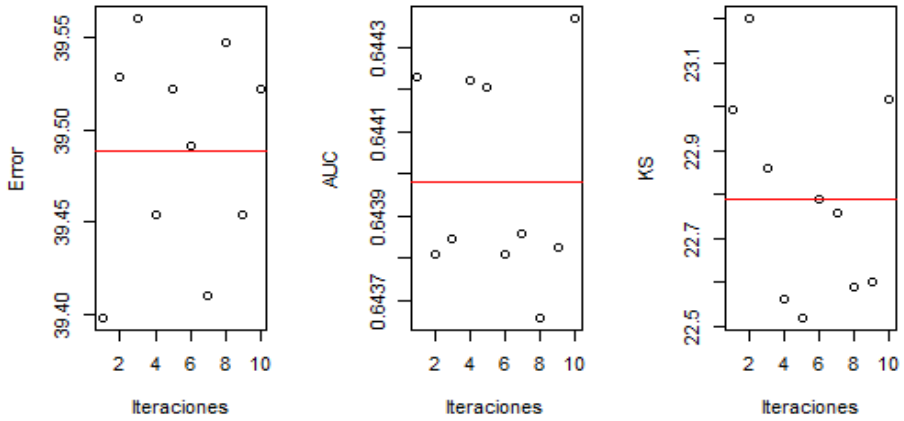


Figura 22

Indicadores de desempeño para el método Regresión Logística con validación cruzada y 10 iteraciones.



Técnicas de clasificación en datos de cáncer de mama para la confirmación del dictamen médico de especialistas en el área, mediante biopsia por aspiración con aguja fina, Wisconsin breast cancer data set

Andrea Alfaro Picado[15], Alcides Arroyo Arroyo[15], Daniely Hernández Orama[15]

andrea.alfaropicado@ucr.ac.cr, alcides.arroyo@ucr.ac.cr, daniely.hernandez@ucr.ac.cr

RESUMEN

Discriminar entre lesiones mamarias malignas y benignas de un paciente es una decisión crucial que afecta el éxito clínico de un paciente. Este tipo de cáncer ataca tanto a hombres como a mujeres, sin distinción de edad. Lograr superar un cáncer de mama va a depender considerablemente de un fácil acceso a servicios médicos con especialistas, principalmente en etapas de detección tempranas; motivo por el cual, mediante esta investigación se pretende hacer uso de técnicas de aprendizaje supervisado que acompañen el diagnóstico del médico para garantizar la atención temprana y más certera posible de cada uno de los pacientes que sufren esta enfermedad. Gracias a la base de datos Breast Cancer Wisconsin Diagnostic (WDBC) se cuenta con un total de 569 diagnósticos y se efectuaron seis métodos de clasificación: árboles de decisión, K vecinos más cercanos, Máquinas vectoriales de soporte, regresión logística, bosques aleatorios y esquema de aprendizaje automático; de este modo mediante la evaluación de los indicadores de desempeño, el cálculo del área bajo la curva de característica operativa del recepto y el estadístico Kolmogorv-Smirnov, se obtuvo que máquinas vectoriales de soporte (SVM) es el método con resultados mejores y más consistente para realizar el acompañamiento diagnóstico de un paciente con lesiones mamarias.

PALABRAS CLAVE: Clasificación, Diagnóstico, Calibración, Desempeño, Máquinas Vectoriales de Soporte, Árboles de decisión, Regresión Logística, Bosques aleatorios, esquema de aprendizaje automático, Indicadores de desempeño, Validación Cruzada

ABSTRACT

Discriminating between benign and malignant breast lesions of a patient is a crucial decision that affects the clinical success of a patient. This type of cancer strikes both men and women, regardless of age. Overcoming breast cancer will depend considerably on easy access to medical services with specialists, especially in the early detection stages; This is why, through this research, the aim is to make use of supervised learning techniques that accompany the doctor's diagnosis to guarantee the early and most accurate care possible for each of the patients suffering from this disease. Thanks to the Breast Cancer Wisconsin Diagnostic (WDBC) database, there are a total of 569 diagnoses and six classification methods were carried out: decision trees, K nearest neighbors, support vector machines, logistic regression, random forests and scheme machine learning; In this way, by evaluating the performance indicators, calculating the area under the receiver operating characteristic curve and the Kolmogorv-Smirnov statistic, it was obtained that support vector machines (SVM) is the method with the best and most consistent results. To perform the diagnostic follow-up of a patient with breast lesions.

[15] Estudiantes de Estadística de la Universidad de Costa Rica

KEY WORDS: Classification, Diagnosis, Calibration, Performance, Support Vector Machines, Decision Trees, Logistic Regression, Random Forests, Machine Learning Scheme, Performance Indicators, Cross Validation

INTRODUCCIÓN

El cáncer de mama es una enfermedad que provoca la forman células malignas (cancerosas) en los tejidos de la mama. Es el cáncer más común en las mujeres a nivel mundial y el segundo más común entre todos los cánceres, según la Fundación de Investigación sobre Cáncer de Mama, BCRF por sus siglas en inglés. En general, la razón de supervivencia para este cáncer varía entre países, pero ha mejorado a lo largo de los años.

En Costa Rica, según la Caja Costarricense del Seguro Social (Mairena M., J, 2019), se registra un 87% de sobrevida de cáncer de mama, siendo el país con mejor récord en el centro y sur de América Estados Unidos tiene la sobrevida más alta del mundo con un 90%, seguido de Israel y de Canadá con un 88%. Esto se presume que se debe al acceso comunitario de cuidado médico.

El Instituto Nacional de Salud de Estados Unidos (NIH) enumera varios tipos de exámenes para detección de este cáncer: mamografía (el más común), imágenes por resonancia magnética (para mujeres con alto riesgo), termografía y muestreo de tejido. Si estos exámenes determinan que existe una anomalía, lo habitual es que se realice una biopsia (extracción del tejido mamario anómala) para determinar, en conjunto con el diagnóstico de un patólogo, si realmente hay cáncer y es en esta última que se enfocará el estudio.

Se cuenta con la base de datos Breast Cancer Wisconsin (WDBC) creada por Dr. William H. Wolberg, W. Nick Street y Olvi L. Mangasarian (1995) y donada por Nick Street, que contiene los resultados de biopsias a tumores encontrados en pacientes, sin género ni edad especificado, a quienes se les diagnostica el tumor es maligno (reproduce células cancerosas de manera descontrolada) o benigno (las células no son cancerígenas).

Se plantea identificar el método de clasificación que más se asemeje a los diagnósticos de biopsia por aspiración con aguja fina (FNA) de una masa mamaria. Utilizando los métodos: árboles de decisiones, K-vecinos más cercanos, regresión logística, máquinas vectoriales de soporte, bosques aleatorios y esquema de aprendizaje automático; evaluando su rendimiento con indicadores de desempeño, curva de característica operativa del recepto (ROC por sus siglas en inglés) y Kolmogorov-Smirnov (KS), así como diferentes métodos para determinar la precisión de estos modelos de clasificación.

Por otra parte, se espera que estos métodos sean una herramienta para asegurar el diagnóstico que clasifica a los pacientes según el tipo de tumor mamario, por medio a las variables que arroja el examen clínico y el dictamen del especialista en la salud.

METODOLOGÍA

Para la realización de este análisis se cuenta con la información de 569 resultados clínicos de pacientes con cáncer de mama de la ciudad de Wisconsin, Estados Unidos. Dicha base cuenta con 11 variables que describen las características de una imagen digitalizada que se obtiene gracias a la aplicación de una biopsia por aspiración con aguja fina (FNA) de una masa mamaria.

Los datos proceden de la Universidad de Wisconsin, cabe resaltar que los datos son registros reales de personas con este tipo de cáncer.

Para la clasificación se tomaron en cuenta 11 variables:

- Diagnosis (diagnóstico): tipo de tumor presente es benigno o maligno (Variable respuesta). Donde 212 son clasificados como malignos y 357 benignos
- Radio: media de las distancias desde el centro a puntos del perímetro. Con valores desde 6,98 hasta 28,11
- Texture (textura): desviación estándar de valores de la escala gris. Con valores desde 9,71 hasta 39,28
- Perimeter (perímetro): Con valores desde 43,79 hasta 188,50
- Area (área): Con valores desde 143,5 hasta 2501
- Smoothness (suavidad): variación local del largo del radio. Con valores desde 0,05263 hasta 0,16340
- Compactness (compacidad): cuadrado del perímetro entre área menos 1.0. Con valores desde 0,01938 hasta 0,34540
- Concavity (concavidad): severidad de los puntos de concavidad del contorno. Con valores desde 0 hasta 0,42680
- Concave points (puntos de concavidad). Con valores desde 0 hasta 0,20120
- Symmetry (simetría). Con valores desde 0,1060 hasta 0,3040
- Fractal dimension (dimensión fractal): aproximación de los bordes. Con valores desde 0,04996 hasta 0,09744

Como parte de los métodos de clasificación aplicados se encuentran los **árboles de decisión**, que consisten en una herramienta utilizada para graficar y categorizar modelos de decisión, esta forma gráfica es similar a los diagramas de flujo, donde cada nodo o cuadro de decisión representan una prueba sobre un atributo específico de una variable y cada rama del árbol es un resultado de dicha prueba. Todos los nodos finales corresponden a la categoría que pertenece el sujeto u observación, para llegar a estos nodos finales es necesario ir desde la copa del árbol hasta la raíz, donde las reglas de clasificación son de forma sucesiva, fluida y ordenada.

Cabe resaltar la técnica de árboles de decisión, donde cada árbol contiene a una variable objetivo (dependiente) y la meta es obtener una función que permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos (Rajesh, 2018).

Como segunda técnica de clasificación se considera **K-vecinos más cercanos (KNN)**. Este es un algoritmo de aprendizaje supervisado. El algoritmo es un sistema de clasificación basado en la comparación de datos nuevos con datos ya presentes en la base de entrenamiento, según tenga k vecinos más cerca de un grupo o de otro. Es decir, calcula la distancia de la nueva observación a cada una de las ya existentes y ordena dichas distancias de menor a mayor para ir seleccionando al grupo al que pertenece. Se recomienda la estandarización

de la base de datos, ya que el vecino más cercano de una instancia es definido en términos de la distancia Euclidiana estándar (Ruiz, 2016).

Otra técnica es la **regresión logística**, esta permite calcular la probabilidad de que la variable dependiente u objetivo pertenezca a cada una de las categorías establecidas en función del valor que adquiera la variable independiente. Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas (Amat, 2016).

La **máquina vectorial de soporte** también ha sido utilizada como parte de las técnicas de clasificación. Este método busca encontrar la manera óptima de clasificar los elementos de un conjunto de datos según un hiperplano en forma de superficie de decisión que se realiza maximizando el margen de separación entre las clases, gracias a los vectores de soporte que definen el borde de esta separación (Martínez, 2020).

A partir de dicho hiperplano, que funciona como una línea que separa en dos grupos los datos, cada vez que ingresa una nueva observación va a ser clasificada mediante ese hiperplano se va a definir a cuál grupo perteneces. Igual que algunas de las otras técnicas anteriores, las máquinas de vectores también hacen uso de un subconjunto de observaciones de entrenamiento que se utiliza como soporte para la creación de la mejor superficie de decisión posible.

Bosques Aleatorios es una técnica de clasificación que consiste en usar datos de entrenamiento para seleccionar varias muestras aleatorias de forma simultánea, cada muestra aleatoria arroja un árbol de decisión, haciendo que los árboles se entrenen con distintas muestras para un mismo objetivo, y al combinar los resultados de todos los árboles unos errores se compensan con otros y así se obtiene una predicción que generaliza mejor los resultados (Martínez, 2020).

Por último, se hace uso de la técnica de **esquema de aprendizaje automático**. Este método busca mejorar un modelo predictivo gracias a un set de entrenamiento de donde se obtienen varias muestras aleatorias, de forma similar a árboles aleatorios, estas muestras deben de tener la misma cantidad de observaciones, requieren ser con reemplazo y el tamaño de la muestra es igual al tamaño de la base de entrenamiento, pero no contiene a todas las observaciones, ya que se repiten. Posterior a esto, se crea un modelo predictivo con cada una de las muestras aleatorias, para finalmente construir un modelo predictivo que es un promedio de todos los modelos antes construidos (Mehmet, 2009).

Para cada método explicado anteriormente, se utiliza, en primer lugar, un proceso de calibración con el objetivo de encontrar los parámetros adecuados para cada técnica de clasificación que permitan obtener los mejores resultados, minimizando el error de clasificación y maximizando los valores del área bajo la curva de característica operativa del receptor (ROC) y el estadístico Kolmogorv-Smirnov, para que el modelo obtenga un mejor desempeño y precisión en las predicciones.

Adicionalmente, cada una de las técnicas de clasificación fue evaluada mediante la validación cruzada, que tiene como objetivo poner a prueba la capacidad del modelo para predecir nuevos datos que no se utilizan en los entrenamientos de aprendizaje, con el fin de alertar sobre problemas como un exceso de ajuste o bien un sesgo de selección.

Esto se logra con una partición de la muestra de datos en subconjuntos, llamados datos de entrenamiento, que serán analizados en otro subconjunto de datos llamados validación, se hacen múltiples rondas de validación cruzada para reducir la variabilidad, usando diferentes particiones, y los resultados de la validación se combinan mediante rondas para dar una estimación del rendimiento predictivo del modelo (Delgado, 2018). En resumen, es como si se cosecharan medias de la aptitud en la predicción para derivar una estimación más precisa del rendimiento de predicción del modelo (Delgado, 2018).

Así mismo, se utilizan los indicadores de desempeño para evaluar y comparar la calidad de un modelo de clasificación, se toman en cuenta los indicadores de: Error, Falsos Positivos, Falsos Negativos, Falsos Positivos, Precisión, Precisión Positiva, Precisión Negativa, Asertividad Positiva y Asertividad Positiva. Cada uno de los indicadores anteriores se obtiene mediante una matriz de confusión, que contiene información sobre predicciones realizadas por un método de clasificación.

Otro método de evaluación es Kolmogorov-Smirnov, que tiene como objetivo identificar la máxima diferencia entre las distribuciones acumuladas de las clases, según la probabilidad predicha por los modelos de clasificación generados en cada una de las distintas técnicas de clasificación utilizadas en esta investigación (BCRF, s.f.). La curva de característica operativa del recepto (ROC) representa la relación entre la tasa de verdaderos positivos o sensibilidad y la tasa de falsos positivos (Mairena, 2019).

Por último, se destaca que el análisis se realizó utilizando el lenguaje de programación R 4.0.0 (R Core Team, 2020), mediante el software libre Rstudio en su versión 3.6.0, haciendo uso de las librerías: dplyr (Wickham et al, 2019), stringr (Wickham, 2019), rpart (Thernaeau & Atkinson, 2020), rattle (Williams, 2019), ROCR (Sing et al, 2015), plotly (Sievert, 2020), DT (Yihui Xie, Joe Cheng & Xianying Tan, 2019), caret (Kuhn, 2019), kknn (Schliep & Hechenbichler, 2016), adabag (Alfaro, Gamez & García, 2018), e1071 (Meyer et al., 2019), class (Ripley & Venables, 2020), randomForest (Liaw & Wiener, 2018), ggplot2 (Wikham, 2020).

RESULTADOS

Como se describe en la metodología, se ejecutan varias técnicas de clasificación para obtener el mejor método que clasifique el diagnóstico del tumor de mama (como benigno o maligno) con respecto a las variables tomadas por la aplicación de la biopsia por aspiración con aguja fina. En primera instancia se utilizaron ocho variables para realizar la clasificación del diagnóstico; esto debido a que las variables de radio, área y perímetro presentan una alta correlación, por lo que se decide excluir área y perímetro del análisis.

Se ejecutan de forma aleatoria bases de entrenamiento y de prueba con un total de 500 y 69 observaciones respectivamente; para realizar las evaluaciones del modelo mediante el cálculo de los indicadores de desempeño. Antes de dar inicio al desarrollo de cada método de clasificación se realiza la calibración de cada uno de estos con sus respectivas reglas duras de esta forma; para árboles de decisión se definen los parámetros en un minsplit de 100, minibucket de 50, maxdepth de 15 y un cp de 0,01; para el método de KNN se establece el número de vecinos (k) en 5; para máquinas vectoriales de soporte (SVM) el kernel empleado es el radial, para bosques aleatorios se instauraron los parámetros de ntree y mtry en 11 y 2 respectivamente; finalmente el esquema de aprendizaje automático se ejecuta con un total de 20 árboles de decisión.

Realizadas las calibraciones correspondientes se procede a la ejecución de cada técnica con la base de entrenamiento y el cálculo de los indicadores de desempeño con respecto a la base de prueba, de esta forma la

tabla 1 presenta los valores de los indicadores para cada método. Para la naturaleza de las clasificaciones se poseen cuatro posibles resultados: clasificar como benigno un tumor benigno, clasificar como maligno un tumor maligno, estas opciones son los ideales, pero se puede dar que se clasifique como maligno un tumor benigno y en un aspecto de mayor gravedad, clasificar como benigno un tumor maligno, este escenario se desea evitar a toda costa. Por lo que se busca que el indicador de falsos negativos sea bajo; pero no se debe restar importancia a la otra clasificación que puede "errar" por lo que el indicador de falsos positivos también se busca en una menor medida.

Observando conjuntamente los falsos negativos y falsos positivos, los métodos de SVM y esquema de aprendizaje automático presentan los valores más bajos. A pesar de que el indicador de falsos negativos del método de árboles de decisión es igual al de esquema de aprendizaje automático, el indicador de falsos positivos de árboles de decisión es el segundo más alto; y al tratarse de la salud de una persona definitivamente este aspecto no es adecuado.

También se debe prestar atención a los demás indicadores de desempeño, con respecto a la precisión los métodos de SVM y bosques aleatorios presentan el valor más alto, ambos con un 95,65; lo que establece que gran parte de las predicciones las realizan correctamente. Por otro lado, la precisión positiva y la negativa son complementarias a los falsos negativos y positivos, respectivamente, por lo que observando estos valores conjuntamente se vuelve a obtener que los métodos de SVM y esquema de aprendizaje automático presentan los mejores valores (para este escenario se desean los valores más altos posible).

Finalmente, con respecto a los indicadores de desempeño, la asertividad positiva y negativa son valores que se esperan estén altos, pues representan la proporción de valores a un diagnóstico predichos correctamente; para este caso los métodos del esquema de aprendizaje automático y árboles de decisión presentan los valores más altos en la asertividad negativa y bosques aleatorios y SVM en la asertividad negativa. En síntesis, los métodos que más resaltan por sus indicadores de desempeño resultan ser bosques aleatorios, SVM y esquema de aprendizaje automático.

Tabla 1

Indicadores de desempeño de cada técnica de clasificación utilizada.

Método	Falsos Positivos	Falsos Negativos	Precisión	Precisión Positiva	Precisión Negativa	Asertividad Positiva	Asertividad Negativa
Árboles decisión	4,26	**9,09**	94,20	90,09	95,74	90,91	95,74
KNN	2,27	12,00	94,20	88,00	97,77	95,65	93,48
Regresión Logística	6,38	13,64	91,30	86,37	93,62	86,37	93,62
SVM	**0,00**	9,67	**95,65**	**90,32**	**100**	**100**	92,68

Bosques aleatorios	**0,00**	18,19	94,20	81,81	**100**	**100**	92,16
Esquema de aprendizaje automático	2,12	**9,09**	**95,65**	90,90	97,87	95,24	**95,83**

Como parte del proceso de validación de las técnicas empleadas, se lleva a cabo el cálculo del error de clasificación, el área bajo la curva de característica operativa del recepto (ROC) y el estadístico Kolmogorov-Smirnov (KS), presentes en la tabla 2. De este modo, los métodos que presentan un error de clasificación más bajo son SMV y esquema de aprendizaje automático, la regresión logística presenta el error de clasificación más alto y los demás métodos son relativamente similares en este valor; se espera encontrar valores bajos en este indicador para garantizar una correcta clasificación de los diagnósticos.

Referente a los valores del área bajo la curva de ROC y KS es importante resaltar que sobrepasan el máximo valor esperado, esto sucede para todos los métodos, en otras palabras la mayoría de los valores del área bajo la curva están por encima de 0.9 y todos los valores de KS están por encima de 70; esto podría resultar preocupante ya que valores de esta índole podrían indicar un sobreajuste del modelo por su buena precisión, generando resultados que no son correctos, pero para este escenario valores de esta naturaleza son de esperar, pues el objetivo de la clasificación es encontrar una técnica equivalente al proceso de diagnóstico que se realiza, por lo que encontrar valores con las características mencionadas es lo más adecuado.

Aclarado el motivo del comportamiento de estos valores, se tiene que el área bajo la curva de ROC es más alta para el método SVM y KNN con un 0,95, para el KS solo el SVM supera el valor de 90. De los demás métodos, se puede destacar el esquema de aprendizaje automático con valores de área bajo la curva de ROC de 0,94 y un KS de 88,78 y el KNN con un KS de 89,16. En forma conjunta los tres métodos mencionados resultan ser los mejores. En los anexos se pueden observar los gráficos del área bajo la curva de ROC y el KS para cada uno de los métodos.

Tabla 2

Valores del error de clasificación, área bajo la curva de ROC y Kolmogorov-Smirnov para cada método de clasificación utilizado.

Método	Error de clasificación	Área bajo la curva de ROC	Kolmogorov-Smirnov
Árboles decisión	5,79	0,93	86,65
KNN	5,80	**0,95**	89,16
Regresión Logística	8,69	0,89	79,98
SVM	**4,35**	**0,95**	**90,32**
Bosques aleatorios	5,79	0,91	81,81
Esquema de aprendizaje automático	**4,35**	0,94	88,78

Al realizar el proceso de validación cruzada se vuelve a calcular el error de clasificación, el área bajo la curva de ROC y el KS. Al realizar este proceso se espera estabilidad por parte del modelo, es decir, que no existan cambios tan bruscos al emplear otra base de entrenamiento para realizar las predicciones con la base de prueba. Si se dan cambios muy variantes el modelo puede verse influenciado por el entrenamiento que recibe y memorizar el contenido de esta base en lugar de ejecutar las predicciones correctamente.

Se espera obtener de esta validación resultados similares a los vistos con la tabla 2, de tal modo, la tabla 3 presenta el cálculo de los valores antes dichos al realizar la validación cruzada. Para el error de clasificación se observa nuevamente que el método de SVM presenta el menor valor con un 0.05; el método de KNN presenta el siguiente error más bajo con un 0.057. Por el contrario, el método de esquema de aprendizaje automático presenta el error de clasificación más alto con 0.09 lo cual no es congruente con lo mencionado en la tabla 2.

Para el área bajo la curva de ROC y el KS el método de regresión logística presenta resultados mayores, pero dados los resultados anteriores a este no podría ser el método más adecuado, además el segundo método que presenta valores más altos en el área bajo la curva de ROC y el KS es el SVM y el KKN respectivamente, por lo que resultan ser buenos métodos de clasificación.

Tabla 3.

Valores del error de clasificación, área bajo la curva de ROC y Kolmogorov-Smirnov para cada método de clasificación utilizado al realizar el proceso de validación cruzada.

Método	Error de clasificación	Área bajo la curva de ROC	Kolmogorov-Smirnov
Árboles decisión	0,086	0,91	82,85
KNN	0,057	0,93	87,48
Regresión Logística	0,065	**0,98**	**90,60**
SVM	**0,050**	0,94	87,49
Bosques aleatorios	0,076	0,92	83,69
Esquema de aprendizaje automático	0,090	0,90	80,79

Como se observa en el desarrollo de los resultados, el método de máquinas vectoriales de soporte (SVM) ha presentado los mejores valores en términos generales para todos los cálculos efectuados. Dando los mejores indicadores de desempeño, valor de error de clasificación más bajo, área bajo la curva de ROC y KS altos, y estabilidad de estos resultados para con los efectuados al realizar validación cruzada. Por lo que este método sería el mejor para ser empleado en la clasificación del diagnóstico.

CONCLUSIONES

A partir de los análisis realizados con las 6 técnicas de clasificación desarrolladas a lo largo de esta investigación y los métodos para observar el buen desempeño de estas técnicas, se permite identificar que la

clasificación más precisa del tipo de tumor de cáncer de mama (Maligno o Benigno) es mediante el método de máquinas vectoriales de soporte (SVM).

Estos resultados son de suma importancia, ya que se busca que estos métodos sean de ayuda para acompañar el diagnóstico que da un experto en salud a los pacientes con herramientas tales como modelos de clasificación, gracias a la innovación y precisión que estos pueden brindar.

Al ser una decisión tan importante, se pretende que este tipo de diagnósticos se dé de la forma más certera posible para asegurar la mejor atención de los pacientes y velar por el resguardo de su salud en procesos tan difíciles como lo es el cáncer.

Máquinas vectoriales de soporte (SVM) puede ser usado para mejorar y asistir los diagnósticos de cáncer de mama, ya que, el número de casos de cáncer de mama en comparación a el número de especialistas es mucho menor, y para obtener una cita con dichos especialistas puede perderse tiempo valioso, dificultando una detección temprana de este padecimiento.

Por lo tanto, gracias a SVM el paciente puede realizarse el examen de biopsia por aspiración con aguja fina (FNA), y mediante las variables que arroja este resultado clínico ser calificado según su tumor e indicarle al especialista si hay señales de tumor maligno o benigno para priorizar su atención sin necesidad de que el experto deba hacer esta clasificación previa, sino que más bien pueda revisar el diagnóstico arrojado por el modelo de clasificación.

En comparación con otros estudios, que utilizan la misma base de datos presente, Mehmet (2009) al usar clasificación con algunos indicadores de desempeño y la curva de ROC llega a la conclusión de que el método de SVM es el que presenta más exactitud a la hora de clasificar los diagnósticos. Este autor llega a un valor más alto de exactitud que el presente trabajo, con un 99.51%, esto se debe a que él seleccionó las características que deseaba y no las que la base de datos proporciona.

Estos resultados con SVM se expresan también en la investigación de Patgiri, Nayak, Akutota, y Paul (2019) donde encuentran por medio de algoritmos de enseñanza de máquinas (*machine learning algorithms*) y duplicación de los datos (igual con WDBC) que tanto SVM como regresión logística tienen resultados mejores de manera más consistente.

De manera similar, en el estudio de Jiang et al (1999) donde su propósito es ver si el diagnóstico asistido por medio de computadora ayuda a mejorar el diagnóstico de cáncer de mama, quienes utilizan el criterio de la media de ROC y concluyen que, en efecto, el apoyo computacional mejora los diagnósticos de estos tumores.

A manera de recomendación, se plantea utilizar este modelo con bases de datos más grandes y con otras técnicas de detección de cáncer de mama, para obtener una asertividad mayor en la clasificación de diagnósticos, y que se pueda monitorear por especialistas para que aseguren el buen funcionamiento del modelo y eventualmente su aplicación.

BIBLIOGRAFÍA

Alfaro, E., Gamez, M. & García, N. (2018). adabag: An R Package for Classification with Boosting and Bagging. Journal of Statistical Software. https://cran.r-project.org/web/packages/adabag/adabag.pdf

Amat R., J. (2016) Regresión logística simple y múltiple. Ciencia de datos. https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple

BCRF (s.f.) Breast Cancer Statistics And Resources. Breast Cancer Research Foundation.https://www.bcrf.org/breast-cancer-statistics-and-resources

Delgado, R. (2018). Introducción a la validación cruzada.http://rstudio-pubs-static.s3.amazonaws.com/405322_6d94d05e54b24ba99438f49a6f8662a9.html

Jiang, Y., Nishikawa R. M., Schmidt, R. A., Metz, C. E., Giger, M. L. & Doi, K. (1999) Improving breast cancer diagnosis with computer-aided diagnosis. Academic Radiology. ScienceDirect. https://doi.org/10.1016/S1076-6332(99)80058-0

Kuhn, M. (2019). caret: Classification and Regression Training. R package version 6.0-86. https://cran.r-project.org/web/packages/caret/caret.pdf

Liaw, A. & Wiener, M. (2018). Classification and Regression by randomForest. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Mairena M., J. (2019) La sobrevida en cáncer de mama en Costa Rica se ubica entre las mejores del mundo. CCSS Noticias. https://www.ccss.sa.cr/noticias/salud_noticia?la-sobrevida-en-cancer-de-mama-en-costa-rica-se-ubica-entre-las-mejores-del-mundo

Martinez H., J. (2020) Máquinas de Vectores de Soporte (SVM). Inteligencia Artificial y Machine Learning en Castellano. https://iartificial.net/maquinas-de-vectores-de-soporte-svm/

Martinez H., J. (2020) Random Forest (Bosque Aleatorio): combinando árboles. Inteligencia Artificial y Machine Learning en Castellano. https://iartificial.net/random-forest-bosque-aleatorio/

Mehmet F., A. (2009) Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications. ScienceDirect. https://doi.org/10.1016/j.eswa.2008.01.009

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. (2019) e1071: Misc Functions of the Department of Statistics, Probability Theory Group. R package version 1.7-3 https://cran.r-project.org/web/packages/e1071/e1071.pdf

NIH (s.f.) Exámenes de detección del cáncer de seno (mama) (PDQ®) – Versión para pacientes. Instituto Nacional del Cáncer. https://www.cancer.gov/espanol/tipos/seno/paciente/deteccion-seno-pdq

Patgiri, R., Nayak, S., Akutota, T., & Paul, B. (2019). Machine Learning: A Dark Side of Cancer Computing. National Institute of Technology Silchar. https://arxiv.org/pdf/1903.07167.pdf

Rajesh S., B. (2018) Decision trees - A simple way to visualize a decision. Medium: GreyAtom. https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

Ripley, B. & Venables, W. (2020). Package 'class': Functions for Classification. https://cran.r-project.org/web/packages/class/class.pdf

Ruiz S., S. (2016) Algoritmos de clasificación: K-NN, Árboles de decisión simples y múltiples (random forest). https://rstudio-pubs-static.s3.amazonaws.com/237547_0171c04b6d2e4550aea58853c056d29d.html

Schliep, K. & Hechenbichler, K. (2016). kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. https://CRAN.R-project.org/package=kknn

Sievert C (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. https://plotly-r.com.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005). "ROCR: visualizing classifier performance in R." http://rocr.bioinf.mpi-sb.mpg.de

Therneau, T. & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

Wickham, H. (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham, H. (2020) ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. https://cran.r-project.org/web/packages/ggplot2/index.html

Wickham, H., François, R., Henry, L. & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.0. https://CRAN.R-project.org/package=dplyr

William H. Wolberg, W. Nick Street, Olvi L. Mangasarian (1995) Breast Cancer Wisconsin (Diagnostic) (WDBC) Data Set. Machine Learning Repository.https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Williams, G. (2019). rattle: Graphical User Interface for Data Science in R. https://cran.r-project.org/web/packages/rattle/index.html

Yihui Xie, Joe Cheng & Xianying Tan (2019). Package 'DT'. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.13. https://cran.r-project.org/web/packages/DT/DT.pdf

AGRADECIMIENTOS

El Comité Editorial de la Revista Serengueti brinda un profundo agradecimiento a todas las personas que formaron parte del diseño de la revista y el debido asesoramiento profesional en distintas áreas académicas.

Diseño: José Aguilar U. (estudiante de Estadística y Comunicación Colectiva).

Asesorías: Ricardo Alvarado B. (Profesor de Estadística), Raquel Arley M. (Nutrióloga), Pedro Chacón R. (Economista) y Karolina Anchía Ch. (Bachiller en Medicina).

Además, se resalta el esfuerzo realizado por los y las estudiantes para la elaboración de los artículos, dada la situación actual producto de la pandemia por el SARS-Cov-2, pues fue necesario acoplar las labores a una modalidad virtual, aun el desgate que esto implica, fueron partícipes del proceso de selección de la revista.