UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

ALGORITMOS DE APRENDIZAJE AUTOMÁTICO EN LA IDENTIFICACIÓN DE miARNS POTENCIALMENTE ASOCIADOS A DISCAPACIDADES INTELECTUALES NO SINDRÓMICAS

Tesis sometida a la consideración de la Comisión de Estudios de Biología para optar por el grado de Maestría Académica en Biología con énfasis en Genética y Biología Molecular

JULIÁN GONZÁLEZ BETANCUR

Ciudad Universitaria Rodrigo Facio San José, Costa Rica

2021

DEDICATORIA

A todas las personas que hicieron posible el esfuerzo que significó este trabajo,

a las personas con las que compartí mi amor por la ciencia,

a todos los que no volveré a ver.

A todas las personas que me dieron un “NO”,

cuando desesperadamente necesitaba un “SÍ”, sin ellos, no hubiera sido posible.

AGRADECIMIENTOS

Al Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA) por permitirme usar su cluster de servidores para la ejecución de los experimentos. Particularmente a Federico Muñóz, por su asistencia tan amable y paciente en el uso de los servidores. Agradezco a Henriette Raventos Vorst y a Adarli Romero Vásquez por sus consejos, su guía y aportes a la perspectiva biológica del trabajo.

"Esta Tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Biología de la Universidad de Costa Rica, como requisito parcial para optar al grado de Maestría Académica en Biología con énfasis en Genética y Biología Molecular".

Dra. Gabriela Chavarría Soley

Representante de la Decana Sistema de Estudios de Posgrado

Dra. Adarli Romero Vásquez

Profesora Guía

M.Sc. Henriette Raventós Vorst

Lectora

Dr. José Andrés Guevara Coto

Lector

Dr. Keilor Osvaldo Rojas Jiménez

Representante del Director del Programa de Posgrado en Biología

Julián González Betancur

Sustentante

Nota aclaratoria: la persona lector que consta en esta hoja de aprobación, cuya rúbrica física es omisa, se encuentra fuera de la República de Costa Rica, razón por la cual la grabación de la defensa de Tesis, debidamente resguardada por la Universidad de Costa Rica, espalda la aprobación y participación de su parte.

CONTENIDO

RESUMEN

Las Discapacidades intelectuales (IDs) son un grupo de trastornos del neuro-desarrollo con una alta heterogeneidad fenotípica y genotípica. La asociación de elementos genéticos, como genes, micro-ARNs o ARNs largos no codificantes, con IDs ha sido abordada de forma empírica. Sin embargo, recientemente, los modelos de aprendizaje automático (ML) han demostrado ser excelentes herramientas en la elucidación de estas asociaciones entre las causas genéticas y los fenotipos de ID. Los micro-ARNs son transcritos cortos no codificantes que participan en la regulación espacio-temporal de la expresión génica, lo cual los hace relevantes en el entendimiento de la causalidad genética de las IDs.

En la presente tesis, se utilizó la base de datos del cerebro en desarrollo BrainSpan, la cual contiene información de patrones de expresión espacio-temporal en el cerebro humano. Esta base de datos fue utilizada en el desarrollo de una serie de modelos de ML clasificadores de diferentes tipos: Máquinas de Soporte Vectorial (SVM), Bosques Aleatorios (RF) y Redes Neuronales Artificiales Pre-Alimentadas (FF-ANN). Estos modelos fueron entrenados para clasificar perfiles de expresión génica entre aquellos asociados y no asociados a IDs. El mejor modelo obtenido fue del tipo FF-ANN, con 0.78 de puntaje F1, 0.78 de rellamado y 0.78 de precisión. Este modelo fue utilizado para clasificar patrones de expresión de micro-ARNs, se filtraron todos aquellos micro-ARNs con una alta probabilidad de estar asociados a ID. Se realizó una búsqueda de literatura que vinculara estos micro-ARNs, o sus genes blanco, con neuro-desarrollo o con IDs.

ABSTRACT

Intellectual Disabilities (IDs) are a group of developmental disorders with high phenotypic and genotypic heterogeneity. The association of genetic elements, such as genes, microRNAs, and lncRNAs, with IDs has been empirically accomplished. However, recently, the Machine Learning (ML) models have proven to be an excellent instrument to elucidate these associations between genetic causes and ID phenotypes. The microRNAs are short non-coding RNAs involved in spatial-temporal regulation of genetic expression, the reason that makes them relevant genetic elements in the identification of causes of IDs.

In this Thesis, the Development Human Brain Database, BrainSpan, was used to develop a series of classifier ML Models. The models were able to classify over the spatio-temporal expression patterns of gene transcripts, registered in BrainSpan, to identify transcripts related with IDs. The algorithms used in the development of the models were: Support Vector Machine (SVM), Random Forest (RF), and Feed Forward Neural Networks (FF-ANN). The best model was obtained using the algorithm FF-ANN, this model presented a F1 of 0.78, a recall metric of 0.78, and a precision of 0.78. This best model, trained with gene expression patterns, was used to predict over microRNAs expression patterns, also available in BrainSpan. A set of microRNAs highly probable of being associated to IDs, were filtered from all the microRNAs classified as associated with ID by the best model obtained. Finally, a literature research was performed to find the target genes of the final set of microRNAs and their relationship with brain development and/or IDs.

LISTA DE CUADROS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

IDs: Discapacidades Intelectuales.

CNV: Copia en el Número de Variantes

ML: Aprendizaje Automático.

NS-IDs: Discapacidades Intelectuales no Sindrómicas.

miARNs: micro-ARNs.

ARNm: ARN mensajero.

RPKM: Lecturas por kilobase de transcrito por millón de lecturas.

Log2: Logaritmo en base 2.

AUC-ROC: Área bajo la curva ROC.

SVM: Máquinas de Soporte Vectorial.

RF: Bosques Aleatorios.

FF-ANN: Redes Neuronales Artificiales Pre-alimentadas.

SGD: Máquinas de Soporte Vectorial con Gradiente Estocástico de Descenso.

NE: No Encontrado.

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Julián González Betancur, con cédula de identidad 800930198, en mi condición de autor del TFG titulado Algoritmos de aprendizaje automático en la identificación miARNs potencialmente asociados a discapacidades intelectuales no sindrómicas

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.-

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

1.1 INTRODUCCIÓN

1.1.1 Discapacidades Intelectuales

Las discapacidades intelectuales (IDs, por sus siglas en inglés) son un grupo de trastornos del neurodesarrollo con una prevalencia estimada de un 1-3% de la población en países occidentales (Maulik, Mascarenhas, Mathers, Dua & Saxena, 2011), con un diagnóstico típicamente en edades tempranas. Se entiende por ID como la "reducción significativa del entendimiento de información nueva o compleja y el aprendizaje de nuevas habilidades" (Organización Mundial de la Salud, 2010), de severidad variable y que puede acarrear problemas serios a nivel educativo, social y de salud según los impedimentos cognitivos que provoque la ID. La severidad de la ID no solo es variable entre personas, sino que también puede presentarse en grados diferentes frente a entornos o circunstancias distintas en las que se encuentre la persona con la ID (Patja, Iivanainen, Vesala, Oksanen y Ruoppila, 2000). Esta definición incluye a las personas con autismo y su característica principal de diagnóstico es un coeficiente intelectual atípico, por debajo de 70 (Patja, Iivanainen, Vesala, Oksanen y Ruoppila, 2000).

Las IDs poseen una amplia heterogeneidad fenotípica y genotípica (Vissers, Gilissen y Veltman, 2016; Yang, Lui, He y Bai, 2020), y pueden separarse clínicamente según su carácter sindrómico (S-ID) o no sindrómico (NS-ID) (Siew, Tan, Babaei, Cheah, y Ling, 2013; Yang, Lui, He y Bai, 2020). Por sindrómicas se entienden aquellas IDs acompañadas de características fenotípicas visibles como malformación o atrofia de estructuras y tejidos, excluyendo aquellas difíciles de detectar como la atrofia de alguna estructura cerebral. Por otra parte, por no sindrómicas se entienden aquellas IDs que no se encuentran asociadas a características fenotípicas externas visibles y tipificables, es decir, características con morbilidad asociada.

Las IDs pueden estar relacionadas a factores ambientales o genéticos. Específicamente, se estima que alrededor del 50% de los casos pueden ligarse directamente a causas genéticas, y que este porcentaje se encuentra positivamente asociado con la severidad de la ID (McLaren y Bryson, 1987; Redon, Ishikawa, Fitch et al. 2006). A nivel genético, en el tipo sindrómico se encuentran aquellas IDs asociadas a grandes cambios cromosómicos o a unos pocos genes causantes definitivos de la condición, aunque dicho fenotipo es afectado positiva o negativamente por consecuencia de variantes en otros genes de menor efecto (Mefford, BatShaw & Hoffman, 2014; Anazi et al. 2016; Vissers, Gilissen & Veltman, 2016). Por otra parte, para el tipo no sindrómico el panorama genético es aún menos claro debido a la ausencia de signos físicos y su difícil caracterización fenotípica , sin embargo, se conoce el importante papel que juegan los genes del cromosoma X en la aparición de estas IDs y de las variantes en el número de copias (CNV) (Redon, Ishikawa, Fitch et al. 2006; Gécz, Shoubridge & Corbett, 2009; Hamdan et al. 2011; Rauch et al. 2012).

Las causas genéticas específicas de los casos de IDs son multifactoriales, heterogéneas, usualmente pleiotrópicas y permanecen desconocidas para la mayoría de los pacientes (Redon, Ishikawa, Fitch et al. 2006; de Ligt et al. 2012; Rauch et al. 2012; Mefford, BatShaw & Hoffman, 2014). Actualmente, las técnicas de secuenciación de nueva generación, con su producción masiva de datos genómicos y de expresión, son herramientas indispensables para el desarrollo de la investigación en IDs, especialmente si tenemos en cuenta que para las NS-IDs el descubrimiento de las causas genéticas suele ser con métodos de fuerza bruta de asociación de variantes (Kaufman, Ayub y Vincent, 2010; Redon, Ishikawa, Fitch et al. 2006; Martínez, et al. 2017). Tal heterogeneidad, sumada a la facilidad de obtención de grandes volúmenes de datos genómicos y exómicos, plantea problemas de análisis multifactoriales para los que el simple uso de los métodos clásicos de asociación resulta insuficiente en el estudio de estos desórdenes (Gilissen et al. 2014). Sacar el mayor provecho posible de

datos tan complejos como los obtenidos de secuenciación de genoma y exoma completos, requiere de la exploración de nuevas alternativas que complementen los métodos clásicos de análisis (Gilissen et al. 2014; Vissers, Gilissen & Veltman, 2016).

1.1.2 Aprendizaje Automático

El aprendizaje automático (ML, por sus siglas en inglés) es un área de estudio derivada de la inteligencia artificial, que integra varios campos del conocimiento humano para realizar reconocimiento de patrones en cualquier estructura de datos (Kotsiantis, 2007; Mitchell et al, 1990; Thabtah, 2017). Los modelos de ML se basan en algoritmos matemáticos y computacionales para aprender patrones y realizar predicciones a partir de grandes cantidades datos preprocesados y curados, y son utilizados en ciencia para producir conocimiento novedoso (Thabtah, 2017). Los modelos de ML pueden ser entrenados de tres formas generales. El Aprendizaje No Supervisado, se denomina como el aprendizaje en el cual no se conocen las clasificaciones de los datos a priori y se espera que los modelos generen agrupaciones basadas en las características de los datos. El Aprendizaje Semi-Supervisado, se define como el aprendizaje basado en datos de los que se conocen las clasificaciones que luego se utilizan como base para la clasificación de otros elementos similares con los cuales se puede reentrenar. El Aprendizaje Supervisado, se define como el aprendizaje basado completamente en elementos de clasificación conocida (Kotsiantis, 2007). En el presente trabajo se utilizó Aprendizaje Supervisado. Un modelo de ML es un sistema que aplica uno o varios algoritmos de ML con el objetivo de analizar nuevos datos.

Actualmente el uso de los modelos de ML ha tomado auge en el campo de la genómica médica por permitir el descubrimiento de patrones ocultos en las secuencias, en la relación entre variantes o en la expresión espacio-temporal de genes y en regulación de la expresión génica, asociados a patologías complejas (Brecher-Smith, Crawford y Escott-Price, 2020; Le, 2020). Se ha propuesto en la literatura que las IDs deben tener

causas genéticas comunes y, además, que la identificación de micro-ARNs (miARNs) asociados a estas discapacidades puede jugar un rol crítico en el entendimiento de su estructura genética y del desarrollo del cerebro humano por su papel en la regulación de la expresión génica (Owen, 2012; Zoghbi & Bear, 2012; Ziats & Rennert, 2013; Anazi et al. 2016).

Los modelos de ML han sido utilizados en mejorar el diagnóstico de IDs específicas y en la predicción de nuevos genes asociados a ellas (Cogill & Wang, 2016; Thabtah, 2017). Cogill y Wang (2016) utilizaron distintos modelos clasificadores basados en ML para predecir genes aún no reportados asociados con la aparición de autismo partiendo de una base de datos de expresión en tejido cerebral que es parte del atlas BrainSpan, que será utilizada en este proyecto. BrainSpan resulta idónea, al contar con datos de expresión de tejido cerebral, para el estudio de miARNs por su carácter tejido-específico.

1.1.3 miARNs

Los miARNs son transcritos cortos no codificantes de entre 21 y 23 ribonucleótidos que pueden regular la formación de heterocromatina, la escisión de ADN, la estructura y la transcripción del ARN mensajero (ARNm) al unirse por complementariedad a secuencias blanco y/o a proteínas específicas (Trivedi & Ramakrishna, 2009; Fiorenza & Barco, 2016). Además, los miARNs son específicos de tejidos o tipos celulares y tanto su presencia como su actividad puede variar en el tiempo (Krichevky et al. 2003; Willemsen et al. 2011; Qiao et al. 2013). Algunos miARN han sido clasificados como específicos de cerebro de mamíferos donde regulan genes involucrados en la división celular, en cascadas de señalización y en diferenciación celular (Krichevky et al. 2003). Esto sugiere que para estudiar su papel en la aparición de anomalías en el desarrollo cerebral debe utilizarse tejido nervioso, lo que dificulta su estudio en seres humanos. Más recientemente, se ha demostrado que los miARNs pueden tener un papel

importante en la aparición y severidad de IDs interviniendo en el retraso en el desarrollo del cerebro (Willemsen et al. 2011; Qiao et al. 2013; Wei-Hong, Kai-Leng, Abbaspour-Babaei, Pike-See & King-Hwa, 2013).

La disponibilidad de una base de datos de expresión en cerebro tan robusta como la disponible en BrainSpan nos permite partir de la hipótesis de que procesos relacionados con una misma condición comparten patrones de expresión espacio-temporales. Por esta razón es necesario partir de perfiles de expresión en el presente trabajo. Salta entonces a la vista la utilidad de emplear el ML para el reconocimiento de patrones que permitan identificar nuevos miARNs candidatos a asociados con IDs.

El objetivo principal de este trabajo fue el obtener una lista de miARNs con una alta probabilidad de estar asociados a NS-IDs, con el fin de generar información útil para el diseño y desarrollo de tratamientos personalizados para personas con NS-IDs (Fig. 1). Lo anterior se obtuvo a través del uso de algoritmos de ML, con el cual se generó una serie de modelos entrenados para identificar perfiles de expresión de genes asociados y no asociados con NS-IDs, que fueron utilizados para obtener la lista de candidatos (Fig. 1). Este estudio permitió (i) identificar miARNs asociados con IDs que no hayan sido previamente reportados, a través de sus patrones de expresión espacio-temporales en cerebro humano; (ii) determinar las características de su expresión espacio-temporal que los hacen candidatos a estar asociados a NS-IDs; y (iii) identificar los genes que podrían estar siendo regulados en el cerebro por los miARNs identificados. Esta información tiene un gran potencial en la dilucidación de la arquitectura genética de las NS-IDs, y, posiblemente, acelerar el desarrollo de tratamientos y atención adecuada para los pacientes, como lo sugieren Rauch y colaboradores (2012) o Zedníková y colaboradores (2020).

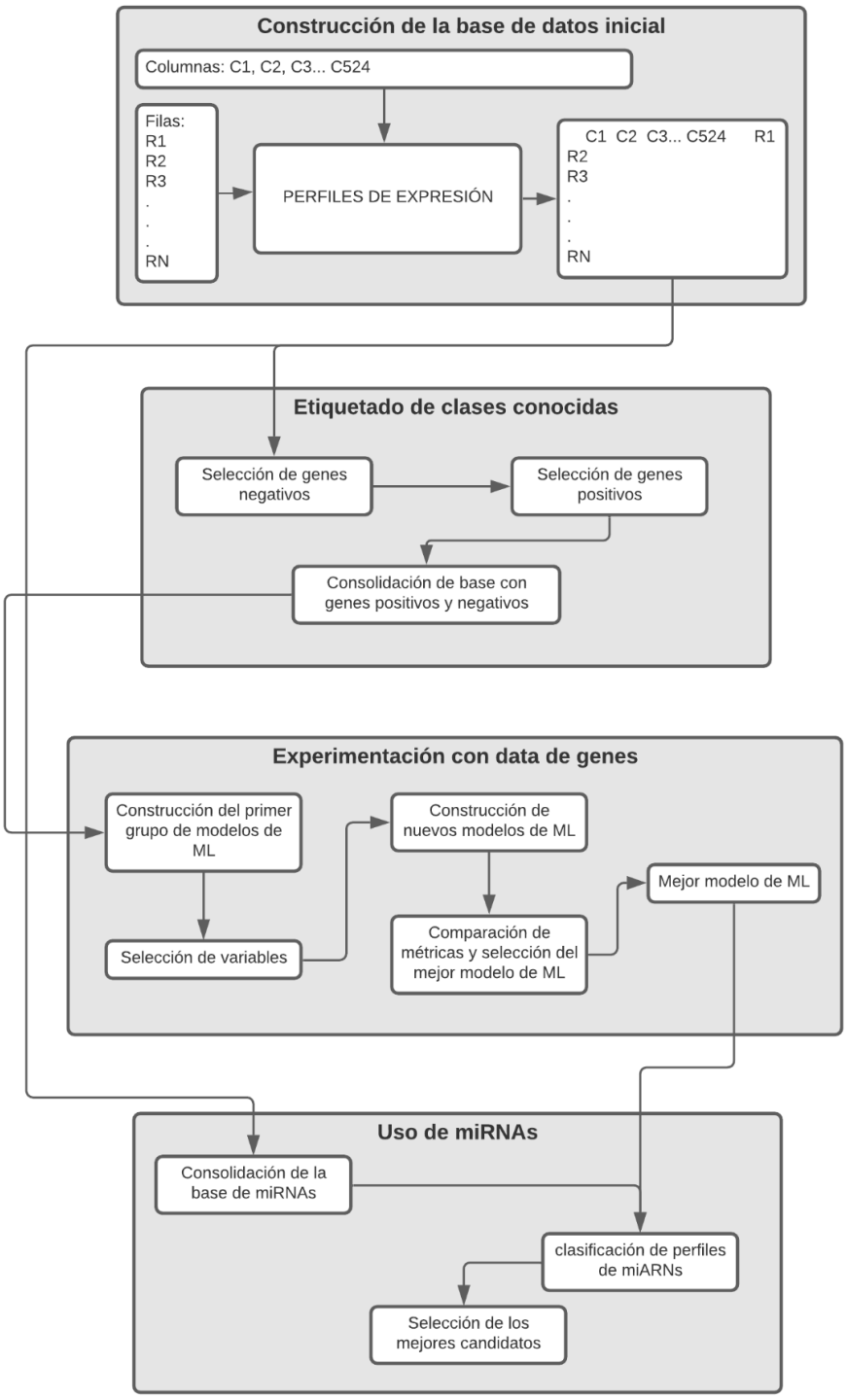

Fig. 1. Diagrama general de los pasos seguidos para obtener la lista de miARNs candidatos a estar asociados con NS-IDs.

1.2 METODOLOGÍA

1.2.1 Adquisición de datos

Tanto para el entrenamiento como para la clasificación de los mi-ARNs, se utilizó la base de datos de expresión génica del cerebro en desarrollo BrainSpan (BrainSpan, 2020). BrainSpan contiene perfiles de expresión de ARN-seq, en lecturas por kilobase de transcrito por millón de lecturas (RPKM), de secuencias génicas y no génicas en 16 áreas cerebrales, desde etapas prenatales hasta los 40 años de edad, con una profundidad promedio de 2 pacientes por cada combinación de área cerebral y estado del desarrollo. En esta base de datos, cada columna corresponde a la combinación de un paciente, un área cerebral y un estado del desarrollo. Esta información fue consolidada en una matriz al asociar los perfiles de expresión y la información de las variables espacio-temporales de los pacientes, que se encuentra en diferentes archivos disponibles en el sitio web de BrainSpan.

1.2.2 Preprocesamiento de datos

Se seleccionaron los patrones de expresión de 1823 genes sin variantes conocidas asociadas o causantes de fenotipos de ID. Este conjunto de 1823 genes fueron etiquetados con la clase “-1” para indicar la ausencia de asociación con ID (clase negativa). Por otra parte, se seleccionó un conjunto de 707 genes reportados en la literatura (Vissers, Gilissen y Veltman, 2016) como asociados a NS-ID. Este segundo conjunto fue etiquetado con la clase “1” para indicar su asociación positiva con fenotipos de NS-ID (clase positiva). El número de perfiles de expresión utilizados para cada clase de genes fue distinto al número de genes seleccionados respectivamente, debido a que BrainSpan cuenta con las secuencias identificadas a nivel de ARNm. Se tomó la decisión de no eliminar los perfiles de expresión duplicados según al gen al que corresponden los ARNm, ya que nos brindan información de distintas isoformas, con diferente actividad biológica. Para la correcta identificación de las secuencias, se hizo el emparejamiento del nombre de los genes con su identificador de Gencode con la

biblioteca BiomaRt (Bioconductor) (Durinck, Moreau, Kasprzyk, Davis, De Moor, Brazma y Huber, 2005), del lenguaje R versión 3.5.2. Se construyó una base de datos (nombrada como “base de datos inicial”) con únicamente los perfiles de expresión de genes etiquetados como pertenecientes a la clase positiva o a la clase negativa. Los valores de RPKM de la base de datos inicial fueron normalizados de tres formas: 1) aplicando el logaritmo con base 2 (Log2) de cada valor en los perfiles de expresión, 2) aplicando la normalización min-max, en la cual a todos los valores de cada variable se les resta el valor mínimo de la columna y se dividen entre el valor máximo, de forma que todos los valores quedan trasladados a una escala de 0 a 1, y 3) aplicando primero la normalización Log2 seguida de la aplicación de la normalización min-max. Para las normalizaciones que incluyen la aplicación de Log2, los valores de expresión de cero fueron reemplazados por el valor 0.00001, que corresponde a un valor de un orden de magnitud menor que el valor mínimo encontrado en todo el conjunto de datos. El uso de la normalización que combina las primeras dos normalizaciones, se basa en la noción de que la transformación logarítmica puede ser implementada para reducir las diferencias o ruido entre muestras y dentro de cada muestra, mientras que la normalización min-max puede ser utilizada para reubicar todas las muestras en la misma escala o plano, lo cual podría mejorar la convergencia de los algoritmos de aprendizaje. El uso de las tres metodologías de normalización tuvo como objetivo encontrar el modelo de ML con el mejor rendimiento posible en el problema de clasificación planteado. Todos los experimentos fueron ejecutados con las tres versiones normalizadas de la base de datos inicial (Fig.2).

1.2.3 Construcción y análisis de modelos de ML

Se utilizaron los siguientes algoritmos de ML para la construcción de los modelo: Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés), Bosques Aleatorios (RF, por sus siglas en inglés), Redes Neuronales Artificiales Pre-alimentadas (FF-ANN, por sus siglas en inglés). Los modelos de SVM fueron implementados en el lenguaje de

programación R, con la biblioteca “e1071” (Meyer, Dimitriadou, Hornik, Weingessel y Leisch, 2019), y en el lenguaje de programación Python, implementados como modelos clasificadores con descenso de gradiente estocástico (Clasificadores SGD), de la biblioteca “scikit learn” (sklearn) (Pedregosa, et al. 2011), que utiliza SVM lineales como algoritmo por defecto. Los modelos de RF fueron implementados en el lenguaje de programación R, con la biblioteca “randomForest” (Liaw y Wiener, 2002). Los modelos FF-ANN fueron implementados en el lenguaje R, con la biblioteca “neuralnet” (Fritsch, Guenther y Wright, 2019), y en el lenguaje Python, con la biblioteca “TensorFlow” (Abadi, 2015) y “Keras” (Chollet, 2015).

Se ejecutaron diez iteraciones con cada algoritmo de ML. Una iteración consistió en la selección de una semilla aleatoria y la creación de al menos diez modelos del algoritmo de ML correspondiente, con diferentes combinaciones de sus parámetros internos (Fig.2). En la creación de cada modelo de ML se utilizó el 70% (1771 perfiles de expresión génica) de los datos para el entrenamiento y 30% (759 perfiles de expresión génica) para el testeo y la validación. Esta división de los datos en un conjunto de entrenamiento y un conjunto de testeo y validación fue hecha con el objetivo de reducir la probabilidad de sobreajuste de los modelos. El sobre ajuste de un modelo corresponde al comportamiento en el cual el modelo es muy bueno prediciendo los resultados únicamente para los datos con los que fue entrenado, disminuyendo mucho su rendimiento con datos con los que no fue entrenado. Los datos de entrenamiento fueron balanceados para las clases positiva y negativa con la biblioteca “caret” del lenguaje R (Kuhn et al. 2018), y la biblioteca “sklearn” del lenguaje Python. Las matrices de confusión y sus respectivas semillas aleatorias y parámetros de los modelos fueron registrados para cada iteración.

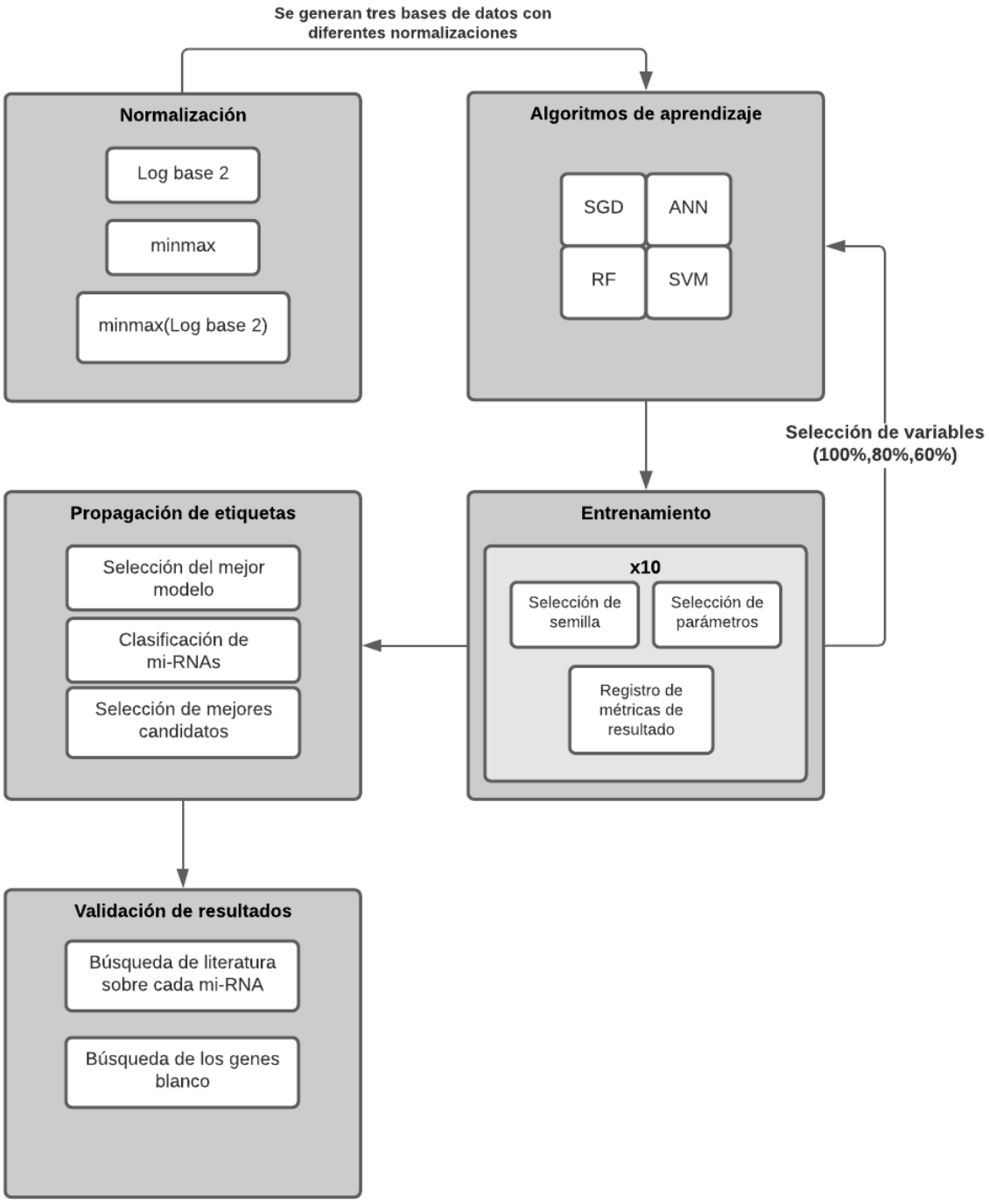

Fig. 2. Diagrama de ejecución de experimentos, reconocimiento de genes y proceso de clasificación de miARNs.

Los modelos de del tipo SVM fueron entrenados utilizando los kernels “radial”, “lineal” y “sigmoidal”, en el caso del kernel sigmoidal, los parámetros costo y gamma fueron ajustados con la función de ajuste de la biblioteca “e1071”. Los modelos del tipo RF

fueron entrenados con 1500 árboles de decisión por bosque, sin reemplazo, los parámetros restantes fueron utilizados con sus valores por defecto, y el conjunto de datos de entrenamiento fue balanceado para las clases positiva y negativa con la función SMOTE de la biblioteca “DMwR” del lenguaje R (Torgo, 2010), con k igual a 3 y sobremuestreo de la clase minoritaria del 190%. Los modelos del tipo FF-ANN fueron entrenados con todos los solucionadores disponibles en la biblioteca “keras” y en los clasificadores MLP de “sklearn”, con alfa entre 0.0001 y 1, variado en barridos de un orden de magnitud, con diferentes combinaciones de en el número de capas, de neuronas por capa y de funciones de activación, en las arquitecturas “uniforme” y “autoencoder”.

1.2.4 Selección de variables importantes

El modelo del tipo RF con el mayor rendimiento obtenido, medido como el puntaje F1 más alto, fue utilizado para dirigir la selección de variables de la base de datos inicial normalizada. Las variables fueron ordenadas según su disminución promedio de Geany, una medida de la importancia en la clasificación del modelo RF. Se seleccionaron dos grupos de variables, uno con el 80% de las variables con mayor importancia, y otro con el 60% de las variables con mayor importancia. A partir de esta selección, se obtuvieron dos nuevas bases: Una compuesta del primer grupo de variables, y otra compuesta con las variables del segundo grupo de variables. Tres iteraciones más fueron ejecutadas para todos los algoritmos de ML utilizando las nuevas bases de datos, normalizadas con el tercer algoritmo de normalización descrito anteriormente. Una vez ejecutada la primera nueva iteración de experimentos sobre la base de datos compuesta por el 60% de las variables más importantes, se descartó su uso para siguientes iteraciones dado a que el rendimiento obtenido por los modelos de ML entrenados con esta base de datos fue el menor obtenido entre todos los modelos de ML, incluyendo aquellos entrenados con la totalidad de las variables. El rendimiento más bajo obtenido de los modelos entrenados con la totalidad de los datos fue utilizado como umbral arbitrario para el

descarte de las bases de datos, esto con el objetivo de optimizar el uso de los recursos de tiempo y computacionales disponibles para la realización de este trabajo.

1.2.5 Clasificación de miARNs

Del total de modelos creados, se seleccionó aquel con el puntaje F1 más alto. El área bajo la curva ROC (AUC-ROC), fue utilizada como métrica de apoyo al puntaje F1. Los puntajes F1 fueron calculados utilizando la función “classification report” de la biblioteca “sklearn.metrics” del lenguaje Python. La probabilidad de cada miARN de pertenecer a cada clase, fue obtenida utilizando la función “predict_proba”, de la biblioteca sklearn. El AUC-ROC y la curva ROC fueron obtenidos usando la biblioteca ROSE del lenguaje R (Lunardon, 2014). El modelo con el puntaje F1 más alto fue considerado como el modelo que produciría los resultados más confiables de la lista de candidatos asociados a NS-IDs. El modelo seleccionado fue reentrenado con la totalidad de los perfiles de expresión génica de las clases positiva y negativa. Los perfiles de expresión de los miARNs fueron obtenidos de la misma base de datos fuente de los perfiles de expresión de los genes, disponible en el sitio web del proyecto BrainSpan. Una vez obtenidas las clasificaciones de los miARNs, se realizó una búsqueda de literatura que relacionara los miARNs etiquetados como clase positiva (miARNs candidatos) con NS-IDs o con neuro-desarrollo. Se filtraron los miARNs candidatos de acuerdo a su probabilidad de estar asociados con NS-IDs, según el modelo seleccionado, con el objetivo de obtener un grupo de candidatos con una alta probabilidad de estar efectivamente asociados. Se utilizó como valor de corte una probabilidad del 99%. Se obtuvieron los nombres de los mejores candidatos utilizando la biblioteca “biomaRt” del lenguaje R. El conjunto de genes blanco de cada miARN candidato fue obtenido utilizando la herramienta STARBASE, con búsquedas sobre el genoma humano y con alto rigor de CLIP (mayor o igual a 3, según las opciones de la herramienta). STARBASE es un sistema de identificación de interacciones ARN-ARN. STARBASE trabaja con una base de datos

relacional de ARNs y puede ser consultado vía web (Jun-Hao, Shun, Hui, Liang-Hu y Jian-Hua, 2014).

Información más detallada sobre el procesamiento, normalización y construcción de modelos de ML, pueden encontrarse en el siguiente enlace al repositorio de git https://github.com/JulianGonzalezB/ML_miRNA_IDs.git.

1.3 RESULTADOS

La base de datos de entrenamiento consistió en los perfiles de expresión de 27657 transcritos de genes de la clase negativa, y 11640 transcritos de genes de la clase positiva, descritos en 524 variables en combinaciones de paciente, área cerebral y estado del desarrollo.

El mejor algoritmo de normalización entre los utilizados, fue la combinación de los algoritmos de Log 2 y min-max. Los modelos entrenados con las bases de datos a las que se les aplicó selección de variables tuvieron un rendimiento superior a aquellos entrenados con la totalidad de los datos (CUADRO I). Sin embargo, los modelos entrenados con el 60% de las variables, tuvieron un rendimiento general menor a aquellos entrenados con el 80% de las variables (experimentos no continuados). Esto significa que al menos alrededor del 80% de las variables de expresión espacio-temporal son relevantes en la tarea de clasificación. El mejor modelo obtenido (CUADRO II) correspondió a un FF-ANN del tipo Perceptrón Clasificador Multicapas (Clasificador MLP) integrado por 5 capas, de 400 neuronas cada una, alfa de 0,01, solucionados SGD, y estado aleatorio de 121 y función de activación Logística. El valor de AUC del modelo con el mejor rendimiento fue de 0,799 (Fig. 3).

CUADRO I.

Métrica de rendimiento F1 según el tipo de modelo de ML y el porcentaje de variables utilizado en el entrenamiento y el testeo de cada modelo.

Modelo	100,00 % [1]		80,00 % [1]	
	promedio	desv.est	promedio	desv.est
SVM	0,222	0,148	0,142	0,148
RF	0,669	0,006	0,594	0,005
ANN	0,669	0,055	0,686	0,041
SGD	0,623	0,055	0,634	0,026

Máquinas de Soporte Vectorial (SVM), Bosques Aleatorios (RF), Redes Neuronales Artificiales Pre-alimentadas (FF-ANN), Máquinas de Soporte Vectorial con Gradiente Estocástico de Descenso (SGD).

[1]100%: Todas las variables originales. 80%: Únicamente las variables entre el 80% con importancia más alta según el mejor modelo del tipo RF.

CUADRO II.

Métricas de rendimiento del mejor modelo.

Ajuste	Métrica		
	Precisión	Rellamado	Puntaje F1
Exactitud	-	-	0,78
[1]**Promedio-Macro**	0,74	0,73	0,74
[2]**Promedio-ponderado**	0,78	0,78	0,78

[1]Promedio-Macro: Promedio no ponderado de las métricas calculadas para cada clase.

[2]Promedio-ponderado: Promedio ponderado de las métricas calculadas para cada clase.

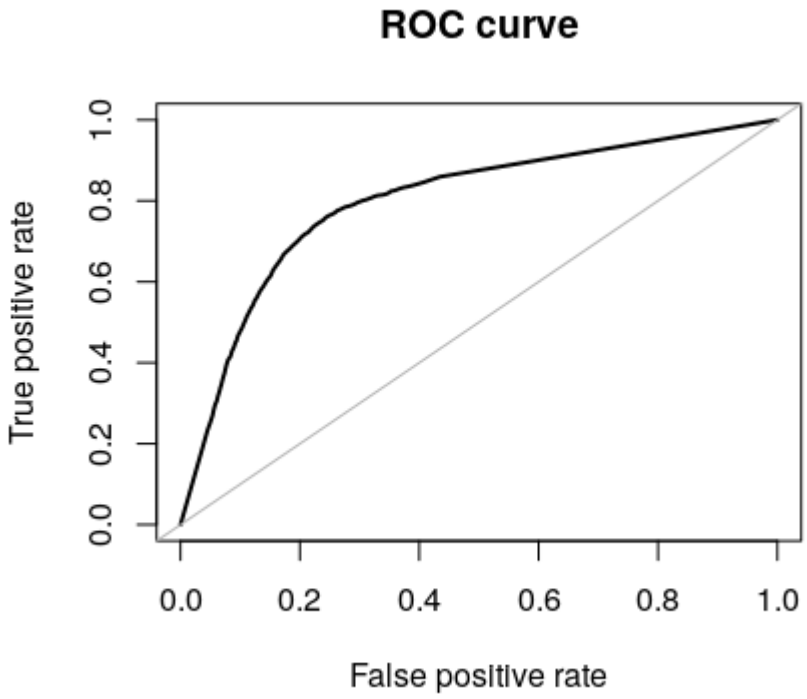

Fig. 3. Curva ROC del mejor modelo de ML obtenido. AUC = 0,799.

Una vez re-entrenado el mejor modelo, se llevó a cabo la clasificación de 544 miARNs, de los cuales 36 fueron clasificados como NS-IDs (CUADRO III), según su probabilidad de pertenecer a la clase positiva. En este grupo de 36 miARNs, la probabilidad promedio de estar asociados a NS-IDs fue de 0,933 (±0.119 desv. est.). Para los miARNs clasificados como clase negativa (no asociados a NS-IDs), la probabilidad de estar asociados a NS-IDs fue de 0,067 (±0.119 desv. est.). Un total de 23 miARNs clasificados como asociados a NS-IDs fueron seleccionados como los mejores candidatos dado que su probabilidad de estar asociados fue mayor o igual a 0,99 (CUADRO III). Para 9 de los 23 mejores candidatos, los nombres de mirbase no fueron encontrados con la herramienta “biomaRt”, probablemente debido a que BrainSpan utiliza Gencode 10 para el etiquetado de los transcritos de las bases de datos que ofrecen al público, versión que se encuentra desactualizada. Lo anterior también puede ser debido a que la técnica de ARN-seq tiene poca sensibilidad a los miARNs. De los mejores candidatos, 7 ya se encontraban reportados como asociados con desarrollo cerebral, desarrollo cognitivo o

NS-IDs (CUADRO III). Un total de 7 de los mejores candidatos tienen genes blanco reportados en literatura como asociados con ID y/o NS-ID (CUADRO III) según los resultados obtenidos en STARBASE. Un total de 5 de los mejores candidatos no fueron encontrados en el sistema de STARBASE, probablemente debido al uso del Gencode v10 en BrainSpan.

CUADRO III.

miARNs etiquetados como asociados a NS-IDs por el mejor modelo de ML obtenido.

Ensembl_gene_id	mirbase_id	probabilidad no asociado	probabilidad asociado
[1,3]ENSG00000211575	hsa-mir-760	0	1
[1]ENSG00000221643	NE	0	1
[1]ENSG00000207789	hsa-mir-26a-2	0	1
[1,3]ENSG00000207712	hsa-mir-627	0	1
[1,2]ENSG00000207863	hsa-mir-125b-2	0	1
[1,2,3]ENSG00000207550	hsa-mir-99b*	0	1
[1]ENSG00000207606	hsa-mir-554	0	1
[1]ENSG00000207945	NE	0	1
[1,3]ENSG00000207729	hsa-mir-556	0	1
[1,2]ENSG00000207991	Hsa-mir-601	0	1
[1]ENSG00000207610	NE	0	1
[1,2]ENSG00000199165	hsa-let-7a-1	0	1
[1]ENSG00000207709	NE	0,001	0,999
[1,3]ENSG00000207631	hsa-mir-641	0,001	0,999
[1]ENSG00000207690	NE	0,001	0,999
[1]ENSG00000221617	NE	0,001	0,999

[1,2]ENSG00000199017	Hsa-mir-1-1	0,002	0,998
[1,2,3]ENSG00000199179	hsa-let-7i	0,002	0,998
[1]ENSG00000253009	NE	0,003	0,997
[1,2,3]ENSG00000199043	Hsa-mir-335	0,005	0,995
[1]ENSG00000221604	hsa-mir-1293	0,005	0,995
[1]ENSG00000211995	NE	0,007	0,993
[1]ENSG00000222326	NE	0,01	0,99
[2]ENSG00000207608	Hsa-mir-127	0,016	0,984
ENSG00000215939	hsa-mir-873	0,023	0,977
[2]ENSG00000199135	Hsa-mir-101-1	0,03	0,97
ENSG00000207825	hsa-mir-519b	0,064	0,936
[2,3]ENSG00000208023	Hsa-mir-185	0,088	0,912
ENSG00000221657	NE	0,172	0,828
[2]ENSG00000207607	hsa-mir-200a	0,178	0,822
ENSG00000221541	NE	0,183	0,817
ENSG00000207819	NE	0,214	0,786
ENSG00000207626	hsa-mir-562	0,251	0,749
[2]ENSG00000211582	Hsa-mir-758	0,336	0,664
ENSG00000208022	hsa-mir-618	0,386	0,614
ENSG00000207979	hsa-mir-527	0,42	0,58

Los genes blanco fueron corroborados únicamente para los miARNs con una probabilidad de asociación mayor o igual a 0,99. No Encontrado (NE)

[1]Seleccionados como mejores candidatos (probabilidad mayor o igual a 0,99).

[2]Ya asociados con neuro-desarrollo, NS-IDs o desórdenes mentales.

[3]Candidatos con genes blanco asociados con neuro-desarrollo y/o NS-IDs.

1.4 DISCUSIÓN

El problema de asociar genes con desórdenes del neuro-desarrollo ha sido exitosamente abordado por varios autores, incluyendo a Cogil y Wang (2016), Gök (2019) y Wang y Wang (2020). Sin embargo, únicamente el mejor modelo obtenido por Wang y Wang fue del tipo ANN, con una arquitectura del tipo “autoencoder” (Wang y Wang, 2020). En este estudio también se obtuvo como mejor modelo de ML un modelo del tipo ANN, sin embargo, su arquitectura correspondió fue del tipo FF y no del tipo “autoencoder”. La diferencia entre los modelos FF-ANN y los demás tipos de modelos puede observarse claramente después de la aplicación de la selección de características, de forma que los FF-ANN tuvieron un rendimiento mayor al resto de tipos de modelo desarrollados. Esto sugiere que los patrones en los datos de expresión génica pueden ser mejor aprendidos siempre que el ruido sea filtrado. Debido a las limitaciones de tiempo y de recursos computacionales, se tomó la decisión de no realizar experimentos con el objetivo de encontrar el porcentaje óptimo de variables a considerar en la selección de variables. Buscar esta combinación óptima podría conllevar a un incremento importante del rendimiento del mejor modelo, sin embargo, no se consideró esto como un punto crítico dado el alto rendimiento obtenido en el mejor modelo construído. Se recomienda para futuros trabajos el realizar el estudio detallado del conjunto de variables óptimo para esta tarea de clasificación con la base de datos utilizada en este trabajo.

En este estudio se usó la totalidad de las variables iniciales para construir una serie de modelos de ML, de los cuales se escogió aquel con el mejor rendimiento (un modelo del tipo RF) para llevar a cabo la selección de variables. Esta técnica, ya reportada en la literatura y clasificada como un método de selección de variables embebido (Saeys, 2007) y utilizado por Wang y Wang (2020), nos permitió obtener una mejora significativa en el mejor modelo y en todos los desarrollados. Cabe resaltar que el tipo de modelo más beneficiado por la selección de variables fue el FF-ANN.

En problemas que requieren de la clasificación en dos categorías, la asignación aleatoria de clases obtendría un rendimiento de 0,5, es decir, un 50% de clasificaciones correctas, mientras la proporción de selección aleatoria se mantenga en 0.5. A partir de esto, cualquier aumento en el rendimiento por encima de 0,5 indica que se ha ganado información acerca de los patrones escondidos en los datos, lo cual es el caso del mejor modelo obtenido en este estudio, que obtuvo un rendimiento muy cercano a 0,8. Esta ganancia de información ha sido de gran valor en otros estudios en que se utilizan algoritmos de ML para asociar genes con condiciones, desórdenes y discapacidades. En estos estudios también se han obtenido rendimientos cercanos a 0,8, como fue el caso de nuestro mejor rendimiento obtenido (Cogill y Wang, 2016; Gök, 2019; Wang y Wang, 2020). Nuestro trabajo tuvo como fundamento la hipótesis de que los genes asociados a NS-IDs, y los miARNs que los regulan, tienen patrones de expresión espacio-temporales similares. En términos biológicos, la ganancia de información sobre los patrones significa que la hipótesis de estudio es correcta.

Debido a que el mejor modelo obtenido en este trabajo fue una red neuronal artificial, más precisamente del tipo FF-ANN, no fue posible obtener las variables espacio-temporales con mayor relevancia en la identificación de patrones de expresión de los miARNs. Las redes neuronales artificiales son consideradas, por lo general, como sistemas de caja negra, lo cual significa que la importancia de las variables de entrada sobre la predicción final es altamente compleja o imposible de determinar. Las ANN compuestas por múltiples capas y múltiples neuronas por capa, es particularmente adecuado considerarlas como sistemas de cajas negras, este es el caso del mejor modelo obtenido en este estudio (Géron, 2019).

1.5 CONCLUSIONES

Es importante notar que algunos de los miARNs reconocidos por el mejor modelo como positivos han sido reportados previamente como asociados a NS-IDs u otros desórdenes del neuro-desarrollo. Según el mejor modelo de ML obtenido, todos los miARNs escogidos como los mejores candidatos tienen una alta probabilidad de estar efectivamente asociados a NS-IDs, basados en sus patrones de expresión espacio-temporal en cerebro. El hecho de que todos los mejores miARNs candidatos cuentan con genes blanco ya reportados como asociados a NS-IDs o a desórdenes del neuro-desarrollo, apoya el futuro estudio de los miARNs candidatos propuestos por nuestro mejor modelo de ML, además de su alta probabilidad de estar involucrados en procesos importantes en la emergencia de NS-IDs.

1.6 REFERENCIAS

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & X. Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org. http://tensorflow.org/

BrainSpan (2020, Febrero 18). Atlas of the Developing Human Brain. https://www.brainspan.org/

Brecher-Smith, M., Crawford, K. & V. Escott-Price. 2020. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Molecular Psychiatry (Epub ahead of print June 26, 2020).

Chollet, F., et al. 2015. Keras. GitHub. https://github.com/fchollet/keras

Cogill S. & L. Wang. 2016. Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. Bioinformatics, 32(23), 3611–3618.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & W. Huber. 2005. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16), 3439–3440.

Gök, M. 2019. A novel machine learning model to predict autism spectrum disorders risk gene. Neural Computing and Applications, 31, 6711-6717.

Fiorenza, A. & A. Barco. 2016. Role of Dicer and the miRNA system in neuronal plasticity and brain function. Neurobiology of Learning and Memory, 135, 3-12.

Fritsch, F., Guenther, F., & M. N. Wright. 2019. Neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet

Géron, A. 2019. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly.

Jun-Hao, L., Shun, L., Hui, Z., Liang-Hu, Q. & Y. Jian-Hua. 2014. Starbase v2.0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. Nucleic Acids Research, 42, 92–97.doi:10.1093/nar/gkt1248

Kaufman, L., Ayub, M. & J. B. Vincent. 2010. The genetic basis of non-syndromic intellectual disability: a review. Journal of Neurodevelopmental Disorders, 2, 182-209.

Kotsiantis, S. B. 2007. Supervised machine learning: A review of classification techniques. Informatica, 31, 249-268.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. & T. Hunt. 2018. Caret: Classification and Regression Training. R package version 6.0-81. https://CRAN.R-project.org/package=caret

Le, D. H. 2020. Machine learning-based approaches for disease gene prediction. Briefings in Functional Genomics, 00(00), 1-14.

Liaw, A. & M. Wiener. 2002. Classification and regression by randomforest. R News, 2(3), 18–22.

Lunardon, N., Menardi, G., & N. Torelli. 2014. Rose: a package for binary imbalanced learning. R Journa,l 6(1), 82–92.

Martínez, F., Caro-Llopis, A., Roselló, M., Oltra, S., Mayo, S., Monfort, S. & C. Orellana. 2017. High diagnostic yield of syndromic intellectual disability by targeted next-generation sequencing. Journal of Medical Genetics, 54, 87-92.

Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T. & S. Saxena. 2011. Prevalence of intellectual disability: a meta-analysis of population-based studies. Developmental Disabilities, 32, 419–436.

McLaren, J. & S. E. Bryson. 1987. Review of recent epidemiological studies of mental retardation: prevalence, associated disorders, and etiology. American Journal of Mental Retardation, 92, 243–54.

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

Patja, K., Iivanainen, M., Vesala, H., Oksanen, H. & I. Ruoppila. 2000. Life expectancy of people with intellectual disability: a 35-year follow-up study. Journal of Intellectual Disability Research, 44(5), 591-599.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. 2011. Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct), 2825–2830.

Redon, R., Ishikawa, S., Fitch, K. et al. 2006. Global variation in copy number in the human genome. Nature, 444, 444–454. https://doi.org/10.1038/nature05329

Saeys, Y., Inza, I. & P. Larrañaga. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23, 2507-2517.

Siew, W. H., Tan, K. L., Babaei, M. A., Cheah, P. S. & K. H. Ling. 2013. MicroRNAs and intellectual disability (ID) in Down syndrome, X-linked ID, and Fragile X syndrome. Frontiers in Cellular Neuroscience, 7, 1-11.

Torgo, L. 2010. Data Mining with R, Learning with Case Studies. Chapman and Hall/CRC, ???. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

Vissers, L. E. L. M., Gilissen, C. & J. A. Veltman. 2016. Genetic studies in intellectual disability and related disorders. Nature Reviews Genetics, 6, 9-16.

Wang, J., & L. Wang. 2020. Prediction and prioritization of autism-associated long non-coding rnas using gene expression and sequence features. BMC Bioinformatics, 21(505). doi:10.1186/s12859-020-03843-5

Yang, J., Lui, A., He, I. & Y. Bai. 2020. Bioinformatics Analysis Revealed Novel 3'UTR Variants Associated with Intellectual Disability. Genes, 11(9), 998-1012.

Zedníková, I., Chylíková, B., Šeda, O., Korabečná, M., Pazourková, E., Břešťák, M., Krkavcová, M., Calda, P. & A. Hořínek. 2020. Genome-wide miRNA profiling in plasma of pregnant women with down syndrome fetuses. Molecular Biology Reports, 47, 4531–4540.

Ziats, M. N. & O. M. Rennert. 2014. Identification of differentially expressed microRNAs across the developing human brain. Molecular Psychiatry, 19, 848–852.