Mundo Fesc, 11 (21), pp. 59-69 Enero-Junio, 2021, ISSN 2216-0353, 2216-0388

Comparación de técnicas de aprendizaje automático para la clasificación de pacientes con trastornos mentales y de comportamiento debido al consumo de psicotrópicos en la ciudad de Barranquilla

Comparison of machine learning techniques for the classification of patients with mental and behavioral disorders due to psychotropic consumption in the city of Barranquilla

[a]Harold Rafael Sarmiento-Gómez, [b]David Francisco Barrios-Marengo, [c]Roberto José Herrera-Acosta, [d] Kevin Rafael Palomino-Pacheco

aEstudiante de pregrado, hsarmiento@mail.uniatlantico.edu.co, Universidad del Atlántico, Barranquilla Colombia

bEstudiante de pregrado, dfbarrios@mail.uniatlantico.edu.co, Universidad del Atlántico, Barranquilla Colombia

c PhD. En estadística, robertoherrera@mail.uniatlantico.edu.co, Universidad del Atlántico, Barranquilla Colombia

d Estudiante Doctorado de ingeniería industrial, krpalomino@uninorte.edu.co, Universidad del Norte, Barranquilla Colombia

Recibido: Julio 10 de 2020 **Aceptado:** Diciembre 11 de 2020

Forma de citar: H.R. Sarmiento-Gómez, D.F. Barrios-Marengo, R.J. Herrera-Acosta, K.R. Palomino-Pacheco "Comparación de técnicas de aprendizaje automático para la clasificación de pacientes con trastornos mentales y de comportamiento debido al consumo de psicotrópicos en la ciudad de Barranquilla", *Mundo Fesc,* vol. 11, no. 21, pp. 59-69, 2021

Resumen

El trastorno por consumo de sustancia psicoactivas contribuye a una carga sustancial mundial de enfermedad, a pesar de los continuos esfuerzos de las entidades gubernamentales para mitigar esta problemática. Este problema es uno de los campos de investigación actuales más atractivo para desarrollar modelos de aprendizaje automático. Este estudio de investigación tuvo como objetivo comparar cuatro modelos de aprendizaje automático para la clasificación de pacientes con trastornos mentales y de comportamiento debido al consumo de psicotrópicos ubicados en la clase de intoxicación agudo o síndrome de dependencia en la ciudad de Barranquilla. El método utilizado consistió en entrenar, validar y comparar cuatro técnicas de aprendizaje automático con bases de datos de pacientes de Barranquilla. Los resultados revelaron que bosque aleatorio y regresión logística arrojaron la mejor precisión (72%). No obstante, red neuronal artificial es el mejor modelo para predecir la proporción de casos verdaderamente positivos entre los casos positivos detectados. Por otra parte, el mejor clasificador que predice la proporción de casos positivos que están bien detectadas es bosque aleatorio, asimismo el mejor clasificador que proporciona la más alta de casos negativo que están bien detectados es máquina de soporte vectorial. Finalmente, cabe mencionar que red neuronal artificial y bosque aleatorio son los clasificadores que mejor área bajo la curva registran con 80% cada uno. En términos generales, red neuronal artificial y bosque aleatorio mostraron indicios de ser un buen clasificador para discriminar entre pacientes que potencialmente estaría en un caso de intoxicación aguda o síndrome de dependencia, obteniendo valores promedios de desempeño entre 80 y 90%.

Palabras clave: Aprendizaje automático, máquinas de soporte vectorial, bosque aleatorio, red neuronal artificial, sustancia psicoactiva.

Autor para correspondencia:

*Correo electrónico:hsarmiento@mail.uniatlantico.edu.co

Abstract

Substance use disorder contributes to a substantial global burden of disease, despite ongoing efforts by government entities to mitigate this problem. This problem is one of the most attractive current research areas for developing machine learning models. This research study aimed to develop a Machine Learning model for the classification of patients with mental and behavioral disorders due to the consumption of psychotropic substances located in the acute intoxication class or dependency syndrome in the city of Barranquilla. The method used was to train, validate and compare four Machine Learning techniques with databases of patients in Barranquilla. The results revealed that Random Forest and Logistic Regression had the best accuracy (72%). However, Artificial Neural Network is the best model to predict the proportion of positive cases among the detected positive cases. On the other hand, the best predictor of the proportion of positive cases that are well detected is Random Forest, and the best predictor of the proportion of negative cases that are well detected is the Support Vector Machine. Finally, it is worth mentioning that Artificial Neural Network and Random Forest are the best classifiers that AUC records with 80% each. In general terms, Artificial Neural Network and Random Forest showed signs of being a good classifier to discriminate between patients who would potentially be in a case of acute intoxication or dependency syndrome, obtaining average performance values between 80% and 90%.

Keywords: Machine learning, Support Vector Machine, Random Forest, Artificial Neural Network, substance psychoactive.

Introducción

Los problemas de salud mental son una enfermedad con alta incidencia en la comunidad mundial y constituyen una de las cargas mundiales de morbilidades más altas del mundo [1]–[3]. El plan de acción de la organización mundial de la salud 2013-2020 [4], estima que el impacto mundial acumulado de los trastornos mentales en términos de pérdidas económicas será de US$ 16,3 billones entre 2011 y 2030. En suma, en la literatura se reconoce ampliamente que los problemas de salud mental entre los menores de edad es uno de los principales problemas sociales y de salud pública en el mundo [5], [6].

Dentro de los problemas que abarca la salud mental se mencionan principalmente trastorno de déficit de atención e hiperactividad (TDAH), depresión, sustancias psicoactivas y trastornos mentales. No obstante, esta investigación se limitará al campo de sustancias psicoactivas. En términos generales, la Organización Mundial de la Salud (OMS) define las sustancias psicoactivas como el consumo de sustancias que afectan los procesos mentales de tal forma que si no se controlan pueden desencadenar problemas personales, familiares, sociales, educativos, entre otros.

La evolución de la investigación sobre SPA ha sido creciente pero no constantes y se han orientado hacia distintos enfoques. Solo por mencionar algunos ejemplos, ciertos estudios se centran en el abuso de sustancias psicoactivas en adolescentes [7]–[13]. Otros estudios se relacionan la implicaciones sociales y geográficas del tráfico de drogas [14]–[19]. Consecuentemente, como en cualquier otro campo de investigación, surgen nuevas y diversas propuestas que constituyen pasos necesarios en el desarrollo del conocimiento científico. Entonces, surge la necesidad de seguir realizando investigaciones para entender mejor los comportamientos de los pacientes bajos un trastorno mental o del comportamiento debido al consumo de una sustancia psicoactiva o psicotrópica. Por lo tanto, esta investigación cuenta con un propósito principal: desarrollar cuatro modelos de Machine Learning para la Clasificación de pacientes con trastornos mentales y de comportamiento debido al consumo de psicotrópicos en la ciudad de Barranquilla

a partir de información sociodemográfica. Para cumplir con el anterior propósito, se construye los modelos de aprendizaje supervisado para posteriormente entrenarlos y validarlos, y luego encontrar cuál es el que mejor modelo que clasifica a los pacientes que pertenecen a la clase intoxicación aguda o síndrome de dependencia, tal y como se recomienda en la literatura [20]–[22].

Por último, el artículo está estructurado así: la sección 1 es introductoria, la sección 2 presenta tanto la metodología utilizada para la construcción de los modelos de aprendizaje automático. La sección 3 presenta resultados y discusiones; y la sección 4 las conclusiones.

Materiales y métodos

Para el desarrollo de la presente investigación se tomó en consideración las siguientes fases metodológicas: obtención de los datos, diseño e implementación de los algoritmos de aprendizaje automático, y últimamente el análisis y comparación de los resultados, tal como se muestra en la Figura 1. Dentro de la primera fase, se obtuvo la base de datos de paciente con trastornos mentales y del comportamiento debido a uso de psicotrópicos, luego se depuró y se particionó en dos bases de datos (entrenamiento y evaluación). A partir de lo anterior, se construyó los cuatro algoritmos de aprendizaje automático (máquina de soporte vectorial, red neuronal artificial, bosque aleatorio y regresión logística), para luego en la etapa de implementación, mediante las variables sociodemográficas entrenarlos y evaluarlos. Finalmente, Se analizó y comparó los cuatro modelos mediante las medidas de desempeño sensibilidad, especificidad, valor predictivo positivo (PPV), Valor predictivo negativo (NPV), tasa de falsos positivos (FDR), tasa de falsos negativos (FNR), precisión, Lift, y Área bajo la curva (AUC) con el fin de evaluar la eficiencia de cada clasificador para discriminar un paciente que se encuentra dentro de intoxicación aguda y síndrome de dependencia y escoger el mejor modelo ajustado a los datos recolectados. Este estudio es de tipo descriptivo de corte transversal, con representatividad en el distrito de Barranquilla, a partir de una base datos suministrada por once Empresas Prestadoras de Salud (EPS) de Colombia. Los datos obtenidos, se refirieron a la población civil desde los 2 hasta los 88 años, distribuida en cinco grupos etarios: Niños (2-14 años), Adolescentes (15-18 años), Jóvenes (18-26 años), Adultos (27-54 años), y Adulto Mayor (55- 88 años). Esta fuente recopiló información relacionada con aspectos sociodemográficos de la población afectada tales como estado civil, género, ubicación sociodemográfica, ámbito de atención, estrato socioeconómico, estado de intervención, y la tipología de trastornos mentales y de comportamiento debido al consumo de psicotrópicos de acuerdo con la codificación CIE-11: consumo de múltiples drogas y otros psicotrópicos (F19X), la cual corresponde a la variable dependiente de este estudio categorizada en dos clases (1: Intoxicación aguda, 0: síndrome de dependencia). La herramienta computacional que se usó fue Rstudio, debido a que permite mediante su interfaz y funciones matemáticas desarrollar los análisis pertinentes para este estudio,

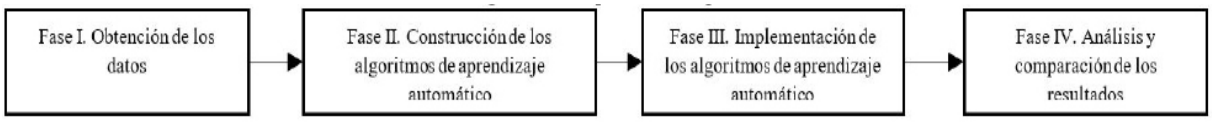

Figura 1. Enfoque metodológico

además porque es de libre acceso.

Regresión logística (LR)

La regresión logística es un modelo de predicción donde la variable dependiente es una variable aleatoria de la siguiente forma (1):

$$Y = \begin{cases} 1, & si\ la\ condicion\ esta\ presente. \\ 0, & en\ otro\ caso. \end{cases}$$

Y X=(x_1,…,x_n) los factores controlables. Entonces, se define la función (2):

$$\pi(X)=E(Y|x_1,\ldots,x_n) \quad (2)$$

Como la probabilidad de que una observación x pertenezca a uno de los dos grupos de la variable de respuesta. El modelo de regresión logística fue presentado por Homer and Lemeshow [23].

$$\pi(X) = \frac{Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}{1 + Exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}$$

Ahora definimos la función de enlace de la regresión logística (4):

$$logit(y) = \log\left[\frac{y}{1-y}\right] \quad (3)$$

Luego, al aplicar esta trasformación en $\pi(X)$, obtenemos mediante el método de estimación de la máxima verosimilitud los valores de $(\beta_0,\beta_1,\ldots,\beta_n)$ y así la probabilidad de que una nueva observación X* pertenezca a uno de los grupos definidos por la variable dependiente. (4)

Máquina de soporte vectorial (SVM)

Las maquinas vectoriales de soporte son consideradas como modelos de clasificación en el que la función discriminante es no lineal, además existe una función de núcleo a un espacio característico en el que los datos son linealmente separable [24]. En este nuevo espacio, cada punto de datos corresponde a un punto abstracto en un espacio p-dimensional, siendo p el número de variables en el conjunto de datos. Cuando Φ es aplicado a los datos originales, una nueva base de datos es originada, es decir, $\{\phi(X_i), y_i\}^n{}_{i=1}$; $y_i = \{-1,1\}$ indica los dos posibles casos (categorías), y el hiperplano equidistante al más cercano punto de cada clase en nuevo espacio es denotado por $W^T\phi(X) + b = 0$ Bajo el supuesto de separabilidad es posible encontrar un W y un b tal que $|W^T\phi(X) + b| = 1$ para todos los puntos cercanos al hiperplano [23]. Esto es (5):

$$W^T\phi(X) + b \begin{cases} \geq 1, & si\ y_i = 1 \\ \leq -1, & si y_i = 1 \end{cases}$$

Para todo i=1,…,n. Tal que la distancia marginal del punto más cercano a cada clase del hiperplano es 1/‖w‖ y la distancia entre dos grupos es 2/‖w‖. Ahora, para maximizar las distancias marginales implica resolver (6) y (7):

$$\min_{W,b} \|W\|^2 \quad (5)$$

Sujeto a

$$y_i(W^T\phi(X) + b) \geq 1 \quad (6)$$

Sea W^* y b^* la solución de la ecuación anterior que define el hiperplano. En un espacio característico, todos los valores de Φ (X_i) son llamados vectores de soporte. Del infinito número de hiperplanos que separan los datos, SVM suministra el hiperplano marginal óptimo, donde las clases son más distantes. (7)

Bosque aleatorio (RF)

Antes de definir el algoritmo de bosque aleatorio es pertinente definir primero los árboles de decisión dado que un bosque aleatorio es un número definido de árboles aleatorios. En términos generales, los árboles son métodos de clasificación que dividen el espacio covariable X en trozos separados y luego clasifican las observaciones según el elemento de partición en el que caen. Como

su nombre indica, el clasificador puede ser representado como un árbol [25]. Para La construcción matemática del árbol, primero supongamos que $\gamma \in Y=\{0,1\}$ y que sólo hay una única covariable X. Entonces, se elige un punto de división t que divide la línea real en dos conjuntos $A_1=(-\infty,T]$ y $A_2=[t,\infty)$. Ahora sea (p_s) (j) la proporción de observaciones en A_s tal que $Y_i = j$:

$$\widehat{p_s}(j) = \frac{\sum_{i=1}^{n} I(Y_i = j, X_i \in A_s)}{\sum_{i=1}^{n} I(X_i \in A_s)}$$

Para, s=1,2 y j=0,1. La impureza de la división t es definida como (9) o (10):

$$I(t) = \sum_{s=1}^{2} \gamma_s \quad (8)$$

Donde,

$$\gamma_s = 1 - \sum_{j=0}^{1} \widehat{p_s}(j)^2 \quad (9)$$

Esta medida particular de impureza se conoce como el índice de Gini. Si una partición de elemento A_s contiene todas las O's o todas las l's, entonces $\gamma_s=0$. De lo contrario, $\gamma_s>0$ se elige el punto de separación(10t) para minimizar la impureza. Por otra parte, cuando hay varias covariables, se elige cualquiera covariable y se divide teniendo en cuanta la que conduce a la más baja impureza. Este proceso se continúa hasta que se detienen bajo un criterio conocido. Solo por ejemplificar, se podría detener cuando cada elemento de la partición tiene menos que n_0 puntos de datos, donde n_0 es un número fijo. Los nodos inferiores del árbol se llaman hojas. A cada hoja se le asigna un 0 o un 1 dependiendo de si hay más puntos de datos con Y=0 o Y=1 en esa partición de elemento. Este procedimiento es fácilmente generalizado al caso donde $Y \in \{1,\ldots,K\}$. Porlo tanto, se define la impureza como (11):

$$\gamma_s = 1 - \sum_{j=1}^{k} \widehat{p_s}(j)^2$$

Donde (p_s) γ=(j) es la proporcion de la observación en el cual la partición de elemento es Y=j. Finalmente, luego de precisar matemáticamente un árbol de decisión, entonces se define un bosque (11) aleatorio como una combinación de predictores de árboles, de manera que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles del bosque [26].

Red neuronal artificial

Una red neuronal es un sistema compuesto de muchos elementos de procesamiento simples que operan en paralelo cuya función está determinada por la estructura de la red, la fuerza de la conexión, y el procesamiento realizado en elementos o nodos de computación [27], [28]. Entonces, el comportamiento de la red neuronal está determinado por su topología, los pesos de las conexiones y la función característica de las neuronas [29]. Así, una neurona o unidad procesadora sobre un conjunto de nodos se define matemáticamente como una tripleta (*X,f,Y*), donde X es un subconjunto de *N, Y* es un único nodo de *N y f* es una función neuronal (También llamada función activación) que calcula un valor de salida para Y basado en una combinación lineal de los valores de componente de *X*. Entonces la actividad lineal de x_i (12) dependiendo de los pesos w_i.

$$Y = f\left(\sum_{x_i \in X} w_i x_i\right)$$

Donde, los elementos X, Y f se denominan

conjunto de nodos de entrada, conjunto de nodos de salida y función neuronal de la unidad neuronal, respectivamente. Por otra parte, para incluir un calor umbral θ_i para la neurona x_i, se considera una n(1e2u) rona auxiliar de valor x_0=-1 y se conecta a x_i con un peso θ_i. Tal como se muestra a continuación (13):

$$u(w, x_i) = \sum_{j=1}^{n} w_{ij} x_j - w_{i0}\theta_i = W_i \cdot X \quad (13)$$

Resultados y discusión

Se analizaron 2,054 casos asociados a los trastornos mentales y de comportamiento debidos al consumo de sustancias psicoactivas (Códigos CIE F10X-F19X) en el periodo de enero de 2016 a mayo del 2019. El número de casos registrados por cada EPS en Barranquilla se encuentra en la Tabla 1, donde se observa que AMBUQ y Mutual Ser son las EPS que presentan mayor número de casos reportados, sumando un porcentaje de 71% aproximadamente del total de la muestra analizada.

Tabla 1. Casos confirmados por EPS

EPS	Número de Caso registrados	Porcentaje
AMBUQ	514	25.02%
CAJACOPI	44	2.14%
COMPARTA	114	5.55%
COOMEVA	20	0.97%
COOSALUD	6	0.29%
FAMISANAR	10	0.49%
FONCOLPUERTOS	8	0.39%
MEDIMAS	28	1.36%
MUTUALSER	957	46.59%
SALUDTOTAL	256	12.46%
SALUDVIDA	96	4.67%
SANITAS	1	0.05%

Fuente: Autores

Conforme a las características sociodemográficas de los individuos incluidos en este estudio, se encuentra que, de 2,059 casos registrados, son hombres 1,701 (82.6%) y mujeres 358 (17.4%). Un total de 1,116 (54.2%) son solteros, y la estratificación más predominante es estrato Medio-Bajo 969(47.1%). En cuanto al régimen de atención, 293 (14.2%) son contributivos y 231 (11.2%) subsidiados, del mismo modo, a lo que refiere al ámbito de atención, 1,294 (69.9%) fueron de carácter ambulatorio y 309 (22.1%) hospitalario (Ver Tabla 2). Por otra parte, es importante mencionar que la edad media de los individuos fue de 29.17 años con 70 (3.4%) niños (2-14 años), 417 (20.3%) adolescentes (15-18 años), 603 (29.3%) jóvenes (18-26 años), 814 (39.5%) adultos (27-54 años), y 155 (7.5%) adultos mayores (55-88 años) tal y como se muestra en la Figura 2.

Tabla 2. Resumen sociodemográfico de la población afectada

Género		Estrato		Ámbito		Régimen	
Femenino:	358	1:	150	Hospitalario:	309	Contributivo:	293
Masculino:	1.701	2:	131	Ambulatorio:	1.294	Subsidiado:	231
		3:	969				
		4:	32				
		5:	1				
		6:	2				

Fuente: Autores

Con respecto al consumo de Sustancias Psicoactivas (SPA) por ubicación geográfica en Barranquilla, se observó que tanto la localidad sur como la centro, mostraron altas concentraciones de habitantes reportados por trastornos mentales y comportamiento debido al consumo de Psicotrópicos (Ver Figura 3). Dicho mapa resulta importante

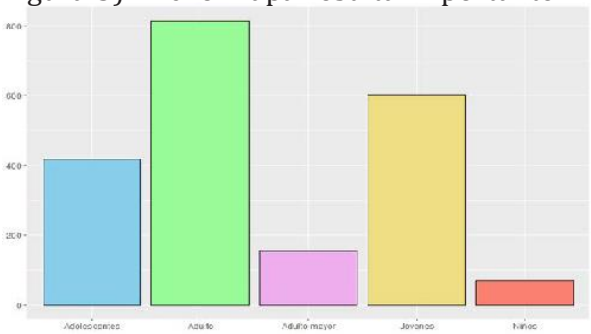

Figura 2. Distribución del grupo etario, **Fuente:** Autores

De un total de 1,581 casos identificado bajo la modalidad de trastorno mental y de comportamiento debido al consumo de múltiples drogas y otros psicotrópicos, la intoxicación aguda inicialmente registró

264 casos (17%) y síndrome de dependencia 332 (21%). Posteriormente con la depuración se logró configurar un base datos de 89 pacientes, donde 43 (48%) pertenecen a la clase síndrome de dependencia y 46 (52%) a la clase intoxicación aguda. Es importante mencionar que en algunos casos las variables independientes ("Estado Civil", "Ámbito De Atención", y "Estrato socioeconómico") no corresponden al total de los casos, por tal motivo se hizo necesario realizar una imputación de datos mediante el método de regresión lineal, en la cual la variable k (Factor de datos faltantes) se estima a partir de un grupo de covariables, es decir, se calculó los pesos de regresión de los datos observados y se devuelve los valores pronosticados como imputaciones. La literatura recomienda trabajar con este tipo de método cuando el porcentaje de datos faltante es bajo para que

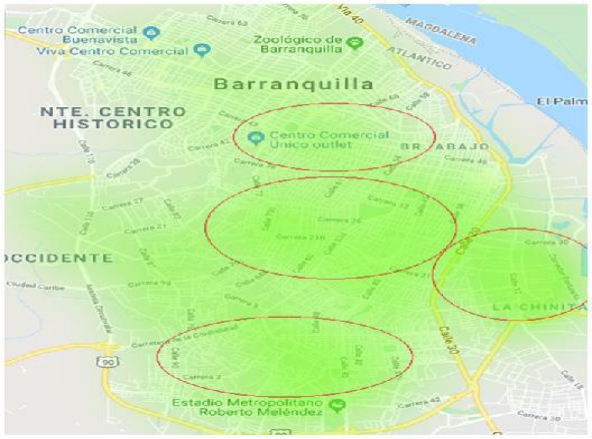

Figura 3. Distribución del consumo de sustancias psicoactivas en Barranquilla, **Fuente:** Autores

Como resultado de la imputación se obtiene una base de datos completa, es decir sin datos faltantes en cada variable independiente, luego se convierten en factores todos los datos para posteriormente seguir el proceso de entrenamiento de los datos. Ahora, siguiendo la metodología, se particionó la base datos para realizar el entrenamiento y la evaluación de los modelos, para lo cual se dividió los datos en 80% para el entrenamiento y 20% para la evaluación del modelo. Luego, se procedió a entrenar y evaluar los modelos con los cuatro algoritmos de aprendizaje automático (SVM, ANN, LR, RF). La rede neuronal artificial se sometió a entrenar mediante la función "***h2o.deeplearning***" del paquete h2o de R, donde se utilizó la función de ***activación Rectificador*** en cada neurona, mientras que el parámetro ***hidden*** sirvió para establecer la cantidad de capas ocultas, así como neuronas en cada una de ellas. En este caso, la red neuronal creada tuvo dos capas ocultas, cada una de ellas con 5 neuronas. Por su parte, el bosque aleatorio utilizó la función "randomForest" del paquete ***randomForest*** de R para el entrenamiento y evaluación del algoritmo. La máquina de soporte vectorial usó la función train de la librería caret, y se definió como parámetro un kernel ***linear***, un ***tuneLength*** de 10, un ***trControl*** a partir de un ***trainControl*** del ***crossvalidation*** para número de 10 con 3 repeticiones, y un ***tuneGrid*** a partir de una grilla de valores de 0 hasta 5. Finalmente, el último algoritmo supervisado utilizado es regresión logística el cual se entrenó y validó mediante la función "glm" de R. En el anexo 1 se puede visualizar el código utilizado para realizar esta investigación.

Como fase final de proceso, se compararon los clasificadores red neuronal artificial, máquina de soporte vectorial, bosque aleatorio, y regresión logística mediante las medidas de desempeño sensibilidad, especificidad, valor predictivo positivo (PPV), Valor predictivo negativo (NPV), tasa de falsos positivos (FDR), tasa de falsos negativos (FNR), precisión, Lift, y Área bajo la curva (AUC) con el fin de evaluar la eficiencia de cada clasificador para discriminar un paciente que se encuentra dentro de intoxicación aguda y síndrome de dependencia y escoger el mejor modelo ajustado a los datos recolectados. En la Tabla 3 se muestra el resumen de los cálculos de las medidas de desempeño, donde el bosque aleatorio y regresión logística tienen una

precisión del 72%, mientras que red neuronal artificial y máquinas de soporte vectorial registran valores de 67% de precisión. No obstante, red neuronal artificial es el mejor modelo para predecir la proporción de casos verdaderamente positivos entre los casos positivos detectados (Intoxicación aguda o Síndrome de dependencia) seguido de regresión logística con PPV de 89%. Por otra parte, el mejor clasificador que predice la proporción de casos positivos que están bien detectadas es el bosque aleatorio, asimismo el mejor clasificador que proporciona la más alta de casos negativo que están bien detectados es máquinas de soporte vectorial. Además, el mejor clasificador que indica la cantidad de veces que mejora la predicción del modelo con respecto al azar es máquina de soporte vectorial. finalmente, vale la pena analizar el área bajo la curva (AUC), donde la red neuronal artificial y el bosque aleatorio son los clasificadores que mejor AUC registran con 80% aproximadamente cada uno, lo que significa que son buenos algoritmos para discriminar entre personas que potencialmente estaría en un caso de intoxicación aguda o síndrome de dependencia. Por otra parte, es importante mencionar que la curva ROC brinda un mejor resumen de la capacidad predictiva del modelo, que una tabla de clasificación, porque presenta la potencia predictiva para todos los posibles valores de referencia β_0. Cuando β_0 es aproximadamente cero (0) casi todas las predicciones serán $\bar{y} = 1$, con lo cual la sensibilidad estar próxima a uno (1) y la especificidad estará cerca de 0, en ese orden de idea se destaca la red neuronal artificial como la que presenta mejor relación sensibilidad y especificidad para clasificar pacientes (Ver Figura 4).

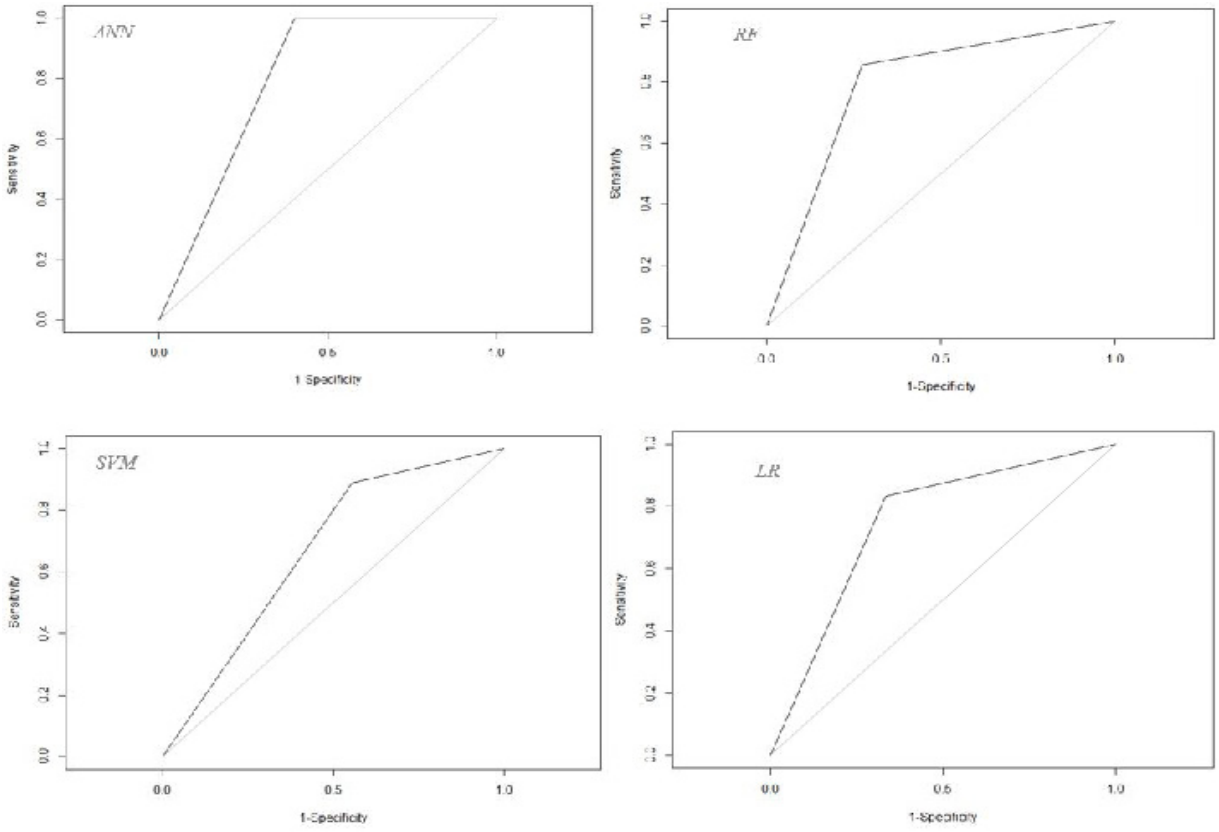

Figura 4. Curva ROC para cada clasificador, **Fuente:** Autores

Teniendo cuenta las entrenamiento y validacion de los modelos anteriores, en la Tabla 4 se muestran cinco corridas de predeccion que buscan ejemplificar la categoria a la que pertenecería un individuo dado a las variables sociodemograficas. A manera de ejemplo se observa el individuo 3 (ID: 3) el cual pertenece al género masculino, a la clase de adulto mayor, al regimen subsidiado, estado civil soltero y estratro socioeconomico 1 (bajo-bajo), fue calsificado dentro de la clase de variable de respuesta, sindrome de dependencia (*Y=0*) por los calsificadores ANN, SVM, RF mientras que LR clasifica este mismo individuo en la clase intoxicación aguda (*Y=1*).

Tabla 4. Predicción de la clase de la variable dependiente por cada clasificador

ID	Sexo	Edad	Régimen	Ámbito	Estado	Estrato	Estado Civil	ANN	SVM	RF	LR
1	2	3	2	1	2	1	2	0	0	0	0
2	2	3	2	1	2	1	2	0	0	0	0
3	2	5	2	1	2	1	2	0	0	0	1
4	2	2	1	1	1	3	2	0	0	0	0
5	2	4	1	1	7	4	1	0	0	0	0

Conclusiones

La necesidad de considerar una nueva forma de analizar los pacientes que sufren algún trastorno mental y del comportamiento por el uso de alguna sustancia psicoactiva o psicotrópico, se origina en utilizar las nuevas tendencias en inteligencia artificial para entender y estimar con mayor precisión las diferentes tipologías que se enfrentan los pacientes. Con estos cimientos, se esgrimió una propuesta para modelar algoritmos supervisados de Machine Learning para la clasificación de pacientes con trastornos mentales y de comportamiento debido al consumo de psicotrópicos en la subtipologia: intoxicación aguda y síndrome de dependencia en la ciudad de Barranquilla.

A partir de la preparación, entrenamiento y validación los datos con los clasificadores: red neuronal artificial, máquinas de soporte vectorial, bosque aleatorio, y regresión logística se encontró los mejores clasificadores dependiendo de la medida de desempeño. En términos generales, red neuronal artificial y bosque aleatorio mostraron indicios de ser buenos clasificador para discriminar entre pacientes que potencialmente estaría en un caso de intoxicación aguda o síndrome de dependencia obtenido valores promedios de desempeño entre 80 y 90%. Otro hallazgo encontrado, corresponde a que la mayoría de casos se reflejan en el género masculino, con un estado civil soltero y de estrato socioeconómico tres(medio-medio), además que la mayoría de los casos se encuentran concentrados en cuatro regiones de Barranquilla.

Referencias

[1] B.F. Grant *et al.*, "Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders - Results from the national epidemiologic survey on alcohol and related conditions," *Arch. Gen. Psychiatry*, vol. 61, no. 8, pp. 807–816, 2004

[2] R.C. Kessler, C.B. Nelson, K.A. McGonagle, M. J. Edlund, R. G. Frank, and P. J. Leaf, "The epidemiology of co-occurring addictive and mental disorders: Implications for prevention and service utilization," *Am. J. Orthopsychiatry*, vol. 66, no. 1, pp. 17–31, 1996

[3] D. A. Regier et al., "Comorbidity of Mental Disorders With Alcohol and Other Drug Abuse: Results From the Epidemiologic Catchment Area (ECA) Study," *JAMA*, vol. 264, no. 19, pp. 2511–2518, 1990

[4] WHO, "Mental Health Action Plan 2013-2020.," Ginebra, Suiza, 2013

[5] L.M. Squeglia et al., "Neural Predictors of Initiating Alcohol Use During Adolescence," *Am. J. Psychiatry*, vol. 174, no. 2, pp. 172–185, Feb. 2017

[6] T. Katsuki, T.K. Mackey, and R. Cuomo, "Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data," *J. Med. INTERNET Res.*, vol. 17, no. 12, 2015

[7] B. F. Grant, D. A. Dawson, F. S. Stinson, P. S. Chou, W. Kay, and R. Pickering, "The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample," *Drug Alcohol Depend.*, vol. 71, no. 1, pp. 7–16, Jul. 2003

[8] R. Ali et al., "The alcohol, smoking and substance involvement screening test (ASSIST): development, reliability and feasibility," *ADDICTION*, vol. 97, no. 9, pp. 1183–1194, Sep. 2002

[9] J.E. Schulenberg and J. L. Maggs, "A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood," *J. Stud. Alcohol*, no. 14, pp. 54–70, Mar. 2002

[10] H. W. Perkins, "Social norms and the prevention of alcohol misuse in collegiate contexts," *J. Stud. Alcohol*, no. 14, pp. 164–172, Mar. 2002

[11] M. Dennis et al., "The Cannabis Youth Treatment (CYT) Study: Main findings from two randomized trials," *J. Subst. Abuse Treat.*, vol. 27, no. 3, pp. 197–213, Oct. 2004

[12] R.W. Hingson, T. Heeren, R.C. Zakocs, A. Kopstein, and H. Wechsler, "Magnitude of alcohol-related mortality and morbidity among US college students ages 18-24," *J. Stud. Alcohol*, vol. 63, no. 2, pp. 136–144, Mar. 2002

[13] J.R. Greenmyer, M.G. Klug, C. Kambeitz, S. Popova, and L. Burd, "A Multicountry Updated Assessment of the Economic Impact of Fetal Alcohol Spectrum Disorder: Costs for Children and Adults," *J. Addict. Med.*, vol. 12, no. 6, pp. 466–473, 2018

[14] C. Potier, V. Laprevote, F. Dubois-Arber, O. Cottencin, and B. Rolland, "Supervised injection services: What has been demonstrated? A systematic literature review," *Drug Alcohol Depend.*, vol. 145, pp. 48–68, 2014

[15] L. Lu, Y. Fang, and X. Wang, "Drug abuse in China: Past, present and future," *Cell. Mol. Neurobiol.*, vol. 28, no. 4, pp. 479–490, Jun. 2008

[16] R.B. Felson and J. Staff, "Committing Economic Crime for Drug Money,"

CRIME Delinq., vol. 63, no. 4, pp. 375–390, 2017

[17]S. Metternich, S. Zoerntlein, T. Schoenberger, and C. Huhn, "Ion mobility spectrometry as a fast screening tool for synthetic cannabinoids to uncover drug trafficking in jail via herbal mixtures, paper, food, and cosmetics," *DRUG Test. Anal*., vol. 11, no. 6, pp. 833–846, Jun. 2019

[18]D.S. Dolliver, S. P. Ericson, and K. L. Love, "A Geographic Analysis of Drug Trafficking Patterns on the TOR Network," *Geogr. Rev*., vol. 108, no. 1, pp. 45–68, 2018

[19]E.-U. Nelson and I. Obot, "Beyond prohibition: responses to illicit drugs in West Africa in an evolving policy context," *DRUGS AND ALCOHOL TODAY*, 2020

[20]S. Gharaei-Manesh, A. Fathzadeh, and R. Taghizadeh-Mehrjardi, "Comparison of artificial neural network and decision tree models in estimating spatial distribution of snow depth in a semi-arid region of Iran," *Cold Reg. Sci. Technol*., vol. 122, pp. 26–35, Feb. 2016

[21]Y.S. Kim, "Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size," *Expert Syst. Appl*., vol. 34, no. 2, pp. 1227–1234, Feb. 2008

[22]E. Reza, R. Arash, and M. Behrouz, "Comparison of Classification Methods Based on the Type of Attributes and Sample Size," *J. Converg. Inf. Technol*., vol. 4, no. 3, pp. 94–102, Sep. 2009

[23]D.A. Salazar, J.I. Velez, and J.C. Salazar, "Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?," *Rev. Colomb. Estadística*, vol. 35, no. 2, pp. 223–237, 2012

[24]J. M. Moguerza and A. Muñoz, "Support Vector Machines with Applications," *Stat. Sci*., vol. 21, no. 3, pp. 322–336, 2006

[25]Wasserman, "All of Statistics : A Concise Course in Statistical Inference Brief Contents," *Simulation*, vol. C, p. 461, 2004

[26]L. Breiman, "Random forests," *Mach. Learn*., vol. 45, no. 1, pp. 5–32, Oct. 2001

[27]AFCEA International Press, *Darpa Neural Network Study*. 1988

[28]S. Haykin, *Neural networks: A comprehensive foundation. Pearson*, 1994

[29]J. Manuel Gutiérrez, "Introducción a las Redes Neuronales," España, 2016