DETECCIÓN Y PREVENCIÓN DE DELINCUENCIA EN LA CIUDAD DE BOGOTÁ
Mediante Técnicas de Minería de Datos

Edwin Fernando González Martínez
Maestría en Ingeniería Analítica de Datos
Universidad Jorge Tadeo Lozano
Facultad de Ingeniería
Bogotá, Colombia
Email: edwinf.gonzalezm@utadeo.edu.co

Miguel Ángel Cortés Capera
Ma. Ingeniería Analítica de Datos
Universidad Jorge Tadeo Lozano
Facultad de Ingeniería
Bogotá, Colombia
Email: miguela.cortesc@utadeo.edu.co

Juan Camilo Ariza Torres
Ma. Ingeniería Analítica de Datos
Universidad Jorge Tadeo Lozano
Facultad de Ingeniería
Bogotá, Colombia
Email: juanc.arizat@utadeo.edu.co

***Abstract*— La Estadística y las Ciencias de la Computación emplean técnicas y metodologías robustas como son la Minería de Datos en inglés (Data Mining) y el Aprendizaje Automático en inglés (Machine Learning) para el proceso de identificación de patrones inusuales, además, explora y extrae información potencialmente nueva, útil y novedosa. El proceso de manipulación y calidad de sus resultados llamado el *Proceso estándar de la industria para la minería de datos* en inglés (Crisp-DM) que emplea seis fases que va desde la obtención de la información y reconocimiento del problema hasta generar nuevo conocimiento. También hay diversos métodos como el llamado *Ensemble* para obtener y generar nuevos resultados dado a su aprendizaje supervisado y no supervisado, además, en esta fase se encuentran métricas específicas para su validación y así obtener nuevo conocimiento.**

***Palabras claves*: Estadística, Minería de datos, Aprendizaje Automático, Algoritmos, Patrones, Delincuencia**

INTRODUCCIÓN

La delincuencia de la ciudad de Bogotá es alarmantes dado al número de denuncias por victimas que reportan a la Policía Nacional [1] y son publicadas dichas estadísticas, también, medios de comunicación y otras fuentes informan lo que sucede y describe las modalidades de hurto. Además, las pérdidas económicas y perdidas de vida son la infamia de los delincuentes que sin escrúpulos [2] y por conseguir su objetivo no importa con qué fin obtenerlo.

Por medio de la Estadística y el Machine Learning se emplearán técnicas robustas con el proceso Crispy - DM para identificar el problema a un conjunto de datos de delincuencia de la ciudad de Bogotá y a su vez realizar el proceso de modelamiento probabilístico con su análisis e informe; esto con el objetivo de emplear y tomar mejores decisiones en cuestión de seguridad y poder reducir las tasas de delincuencia en Bogotá.

20 de mayo de 2019

[1] https://www.policia.gov.co/revistacriminalidad

[2] El escrúpulo es la inquietud de ánimo provocada por la duda acerca de si algo es bueno o malo, correcto o incorrecto, verdadero o falso.
fuente de consulta: Wikipedia

I. MINERÍA DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

I-A. Crisp-Dm

Proceso estándar de la industria para la minería de datos en inglés (Crisp-DM) es un modelo estándar de seis fases que va desde el análisis y objetivo de la investigación, después la manipulación, transformación, análisis predictivos, validaciones hasta la última fase de generación de conocimiento e informe. Makhabel (2015)[8]

Comprenden las siguientes fases:

1. Comprensión del Negocio: Esta tarea incluye la determinación de los objetivos del negocio, la evaluación de la situación actual, el establecimiento de objetivos de la minería de datos, y el desarrollo de un plan.
2. Comprensión de Datos: esta tarea evalúa los requisitos de los datos e incluye la recopilación de datos inicial, descripción de los datos, exploración de datos y la verificación de la calidad de los datos.
3. Preparación de Datos: una vez disponibles, los recursos de datos se identifican en el último paso. Luego, los datos deben ser seleccionados, limpiados y luego incorporados en la forma y formato deseados.
4. Modelado: la visualización y el análisis de conglomerados son útiles para el análisis inicial. Las reglas de asociación iniciales pueden desarrollarse aplicando herramientas como la inducción de reglas generalizadas. Esta es una técnica de extracción de datos para descubrir el conocimiento representado como reglas para ilustrar los datos en la vista de la relación causal entre los factores condicionales y una decisión / resultado dado. Los modelos adecuados a la También se puede aplicar el tipo de datos.
5. Evaluación: los resultados deben evaluarse en el contexto especificado por los objetivos del negocio en el primer paso. Esto conduce a la identificación de nuevas

necesidades y, a su vez, vuelve a las fases anteriores en la mayoría de los casos.

6. Implementación: la minería de datos se puede utilizar para verificar hipótesis mantenidas previamente o para conocimiento.

I-B. Algoritmos de Aprendizaje Automático

Los algoritmos de aprendizaje automático consisten en evaluar la información y determinar la aleatoriedad de su variable dependiente. Están definidas por desarrollos matemáticos y estadísticos llamados algoritmos de aprendizaje que generan mediante un entrenamiento de maquina en inglés (Machine Learning) que generan nuevos parámetros probabilísticos estimados e hipotéticos para su aprendizaje supervisado y no supervisado. González (2018)[5]

Algoritmos de aprendizaje automático en 2 grupos :

1. Aprendizaje Supervisado: es básicamente un sinónimo de *clasificación*, la supervisión en el aprendizaje proviene de los ejemplos etiquetados en el conjunto de datos de entrenamiento.
2. Aprendizaje No Supervisado: es esencialmente un sinónimo de *agrupamiento*, el proceso de aprendizaje no está supervisado ya que los ejemplos de entrada no están etiquetados por clase, por lo general, podemos usar el agrupamiento para descubrir clases dentro de los datos.

I-C. Métodos de Ensemble

Los métodos de ensamble utilizan un conjunto de modelos entrenados $M_1, M_2, ..., M_k$ con el objetivo de crear un modelo mejorado de clasificación, M^p, dado a un k-ésimo conjunto de datos de entrenamiento donde se usa para generar un modelo de clasificación. Han et al. (2014) [6]. Estos modelos entrenados ayudan a mejorar la eficiencia para obtener una varianza mínima dada al conjunto de datos de entrenamiento y precisión en sus predicciones dado al conjunto de datos de prueba. Dada la flexibilidad del modelo aparece el problema de *Overfitting* que consiste en que para los datos de entrenamiento con los cuales se construye el modelo compuesto, se obtienen buenas predicciones, pero no se predice adecuadamente para los nuevos conjuntos de datos. Amat (2017) [1]

Los 3 algoritmos más implementados y potentes de métodos de ensamble:

1. Bagging es un diseño de muestreo *Bootstrapping* que genera un k-ésimo de muestras creando modelos entrenados $M_1, M_2, ..., M_k$ con el fin de obtener una varianza mínima dado al aprendizaje. El proceso de bagging se basa en el hecho de que se promedian un conjunto de modelos entrenados en que se busca reducir la varianza, la media $\bar{M}$ y la varianza de la media de los modelos $\frac{\sigma^2}{k}$. Amat (2017) [1]
2. Boosting Se ajusta de forma secuencial un conjunto de modelos entrenados $M_1, M_2, ..., M_k$ que aprenden en cadena a corregir los errores de un modelo débil dado a los anteriores. Campos (2017) [4] cita de su trabajo de grado que el modelo construido por Boosting es la suma ponderada de todos los modelos débiles dado a que el modelo final va a obtener una predicción eficiente y varianza mínima.
3. AdaBoost es un algoritmo de aprendizaje que pertenece a la clase de conjuntos de modelos, estos tipos de modelos son, en efecto, formados por un conjunto de modelos base en que contribuyen a la predicción del algoritmo utilizando los métodos de agregación y adaptativo. La construcción del modelo se obtiene de forma secuencial, cada nuevo miembro de la secuencia se obtiene mejorando los errores del modelo anterior de la secuencia, las mejoras se obtienen usando un esquema de ponderación que aumenta los pesos de los casos que están incorrectamente clasificados por el modelo anterior; esto significa que el aprendizaje base se usa en diferentes distribuciones de los datos de entrenamiento, las predicciones se obtienen mediante una media ponderada de las predicciones de los modelos base individuales, estos pesos se definen de modo que se otorguen valores mayores a los últimos modelos en la secuencia. Torgo (2011) [10]

II. Técnicas y Metodologías Minería de Datos

II-A. Random Forest

Random forest es un clasificador que consiste en una colección de árbol estructurado de clasificadores $\{h(x, \Theta_k), k = 1...\}$ donde $\{\Theta_k\}$ son independientes y distribuidos de forma idéntica. Además, cada árbol arroja una unidad de votación para la clase más popular en la entrada x. Breiman (2001) [2]

Random forest es también conocido como bosques aleatorios, son una combinación de predictores de árbol de modo que cada árbol depende de los valores de un vector aleatorio x e y, random forest consiste en un conjunto de árboles de decisión, árboles de regresión o de clasificación, se generan un número importante de árboles los cuales son entrenarlos y se calcula su promedio de salida. Torgo (2011) [10] cita en su libro que la predicción de estos se obtiene promediando las predicciones de cada árbol, para los problemas de clasificación, esto consiste en un mecanismo de votación, la clase que obtiene más votos en todos los árboles es la predicción del conjunto.

En los árboles de decisión y random forest se encuentran nodos, ramas y hojas. Los nodos son las variables de entrada, las ramas representan los posibles valores de las variables de entrada y las hojas son los posibles valores de la variable de salida. Como primer elemento de un árbol de decisión tenemos el nodo raíz que va a representar la variable de mayor relevancia en el proceso de clasificación. Todos los algoritmos de aprendizaje de los árboles de decisión obtienen modelos más o menos complejos y consistentes respecto a la evidencia, pero si los datos contienen incoherencias, el modelo se ajustará a estas incoherencias y perjudicará su comportamiento global en la predicción, es lo que se conoce como sobre ajuste. Parra (2017) [9]

Random forest y bagging utilizan el mismo algoritmo con la única diferencia de que el número de predictores son diferentes antes de cada división del nodo, bagging utiliza el número de predictores p y random forest utiliza un número indeterminado de predictores aleatoriamente m, se trata de promediar un conjunto de modelos probabilísticos para conseguir reducir la varianza y así poder obtener la eficiencia óptima del modelo. Breiman (2001) [2] cita en su artículo que el error converge a un límite a medida que aumenta la cantidad de número de árboles, el error de un bosque de clasificadores de árboles depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos; otro método de validación es el cuadrado medio de error mse en el cual se encuentra el valor óptimo del número de predictores y número de árboles dado a la validación iterativa del conjunto de modelos probabilísticos.

La calidad de los nodos está dada a las divisiones óptimas de los nodos. Existen varias alternativas para encontrar el nodo más puro y homogéneo posible, hay varias alternativas, pero las más utilizadas son el Índice de gini y entropía cruzada:

1. Índice de Gini: Se considera una medida de pureza del nodo, su valor de medida oscila entre (0) y (1) de tal manera que valores cercanos a cero indican pureza del nodo y cercano a uno impureza; es una medida de varianza total de las k-ésimas clases construidas del conjunto.
2. Entropía Cruzada: Es otra forma de cuantificar el desorden de un sistema. En el caso de los nodos, el desorden se corresponde con la impureza. Si un nodo es puro, contiene únicamente observaciones de una clase, su entropía es cero. Por el contrario, si la frecuencia de cada clase es la misma, el valor de la entropía alcanza el valor máximo de 1. Amat (2017) [1]

En forma resumida se sigue este proceso:

1. Se seleccionan individuos al azar (usando muestreo con reemplazo) para crear diferentes conjuntos de datos.
2. Se crea un árbol de decisión con cada conjunto de datos, obteniendo diferentes árboles, ya que cada conjunto contiene diferentes individuos y diferentes variables en cada nodo.
3. Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (es decir, sin podar).
4. Se predice los nuevos datos usando el ”voto mayoritario”, donde se clasificará como ”positivo”si la mayoría de los arboles predicen la observación como positiva. El proceso se resume en la figura 1.

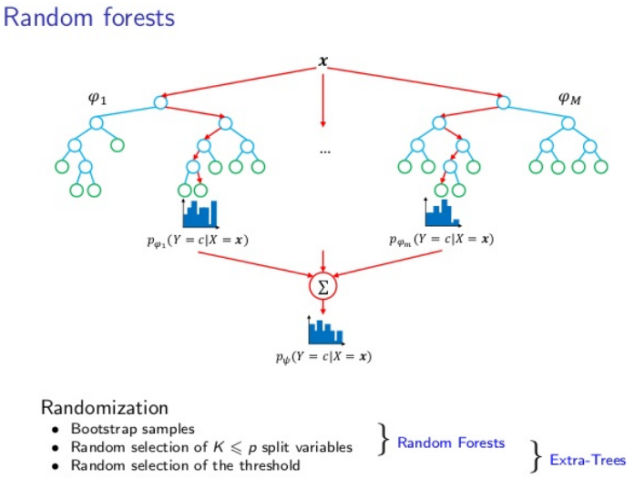


Fig. 1. *Random Forest*

II-B. Redes Neuronales Artificiales

La red neuronal artificial (RNA) es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro. Las redes neuronales intentan aprender mediante ensayos repetidos como organizarse mejor a sí mismas para conseguir maximizar la predicción. Un modelo probabilístico de una red neuronal se compone de nodos, que actúan como input, output o procesadores intermedios, y cada nodo se conecta con el siguiente conjunto de nodos mediante una serie de trayectorias ponderadas. Basado en un paradigma de aprendizaje, el modelo toma el primer caso, y toma inicial basada en las ponderaciones. Parra (2017) [9]

La red neuronal está estructurada por un número de capas de la siguiente forma dada a la figura 2

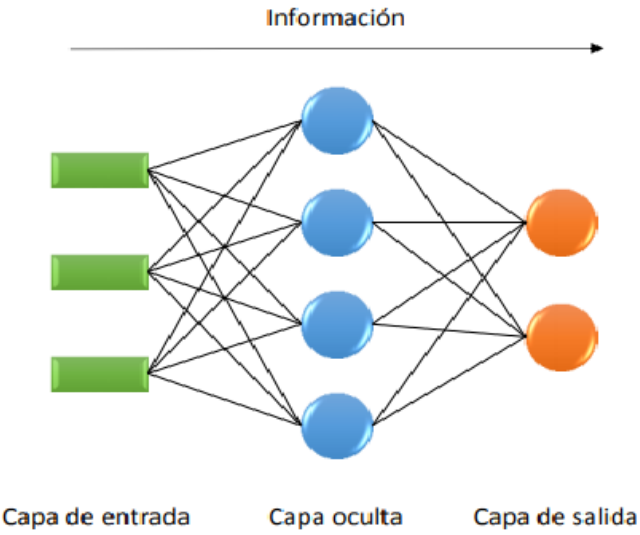


Fig. 2. *Capas de una Red Neuronal*

1. Capa de Entradas: Recepción de señales o información de su entorno
2. Capas Ocultas : Información recibida por los pesos sinápticos
3. Capas de Salida : Información procesada y transmitida

La primera red neuronal artificial fue elaborada en 1943 por el psiquiatra y neuroanatomista Warren McCulloch y el matemático Walter Pitts, con el fin de emular una función neuronal biológica por métodos psiquiátricos y matemáticos. Torgo (2011) [10]; a mediados de los años 80 hubo grandes desarrollos teóricos y a mediados 1990 fue desarrollado el algoritmo Backpropagation por *Werbos*.

II-B.1. Algoritmo Backpropagation: Backpropagation es un algoritmo de aprendizaje de redes neuronales, en el desarrollo de las redes neuronales fue originalmente activado por psicólogos y neurobiólogos que buscaban desarrollo de premisas computacionales en el desarrollo de las neuronas artificiales. Durante la fase de aprendizaje, la red aprende ajustando los pesos para poder predecir la etiqueta de clase correcta. Yanchang et al. (2013)[11]

Las ventajas de las redes neuronales incluyen su alta tolerancia a los datos ruidosos, así como su capacidad para clasificar los patrones en los que no han sido entrenados, se pueden usar cuando puede tener poco conocimiento de las relaciones entre los atributos y las clases, Los algoritmos de red neuronal son inherentemente paralelos; las técnicas de paralización se pueden usar para acelerar el proceso de cálculo. Además, varias técnicas se han desarrollado recientemente para la extracción de reglas de redes neuronales capacitadas. Estos factores contribuyen a la utilidad de las redes neuronales para la clasificación y la predicción numérica en la extracción de datos. Yanchang et al. (2013) [11]

La figura 3 muestra un ejemplo de un modelo neuronal con n entradas, que consta de:

- Un conjunto de entradas $x_1, ..., x_n$.
- Los pesos sinápticos $w_1, ..., w_n$, correspondientes a cada entrada.
- Una función de agregación, $\sum$.
- Una función de activación, f_x.
- Una salida y_i

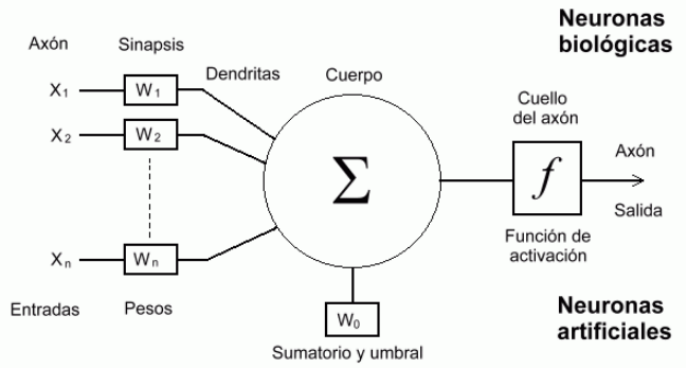


Fig. 3. *Modelo de Una Red Neuronal Artificial*

Las entradas son el estímulo que la neurona artificial recibe del entorno que la rodea, y la salida es la respuesta a tal estímulo. La neurona puede adaptarse al medio circundante y aprender de él modificando el valor de sus pesos sinápticos, y por ello son conocidos como los parámetros libres del modelo, ya que pueden ser modificados y adaptados para realizar una tarea determinada. Parra (2017) [9]

En este modelo, la salida neuronal y está dada por:

$$Y = f(\sum_{i=1}^{n} w_i x_i) \tag{1}$$

Un modelo de una red neuronal artificial realiza tareas de clasificación en el plano dado al número de entradas x_i y unos pesos w_i, y consideramos como función de activación a la función del signo definida, por lo tanto, la salida neuronal Y estará dada en este caso por:

$$Y = \begin{cases} 1 & \textit{sí } \sum_{i=1}^{n} x_i w_i \geq 0 \\ -1 & \textit{sí } \sum_{i=1}^{n} x_i w_i < 0 \end{cases} \tag{2}$$

La función de activación se elige de acuerdo con la tarea realizada por la red neuronal, se presentan las más comunes e implementadas y se destacan en la siguiente figura 4:

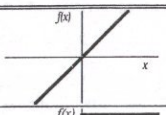
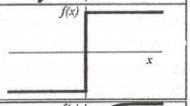
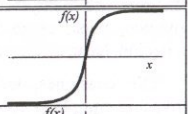
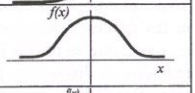
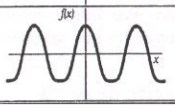
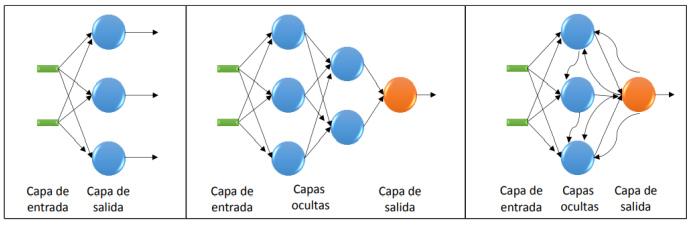
	Función	Rango	Gráfica
Identidad	$y = x$	[-∞, +∞]	
Escalón	$y = sign(x)$ $y = H(x)$	{-1, +1} {0, +1}	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$ $y = tgh(x)$	[0, +1] [-1, +1]	
Gaussiana	$y = Ae^{-Bx^2}$	[0,+1]	
Sinusoidal	$y = A\,\mathrm{sen}(\omega x + \varphi)$	[-1,+1]	

Fig. 4. *Funciones de Activación*

II-B.2. Arquitectura Neuronal: Una estructura neuronal está conformada por la forma en que están conectadas las diferentes formas de neuronas, dado a ello las conexiones o pesos sinápticos en que forman la topología de la red neuronal, en las que están definidas el tipo de estructura por número de capas, tipo de conexiones y grado de la conexión.

1. Número de Capas
   a) *Feedforward o Perceptrón Monocapa*: es un modelo Neuronal unidireccional, compuesto por dos capas de neuronas, una de entrada y otra de salida que realiza los diferentes tipos de cálculos; Manjarrez (2014) [7] este tipo de redes es útil en tareas relacionadas con la auto-asociación, es decir, regenera la información incompleta o distorsionada de patrones que se presentan a la red.
   b) *Feedforward o Perceptrón Multicapa*: Es un modelo Neuronal conformado por una capa de entrada, varias capas ocultas y una de salida, su transferencia o pesos sinápticos a cada nodo realiza el proceso iterativo.
2. Tipo de Conexión
   a) *Recurrentes*: Tipo de conexión en propagación y corrección de señales enlazadas entre las neuronas de una o varias capas.
   b) *No Recurrentes*: En este tipo de conexión la red de propagación se produce en un solo sentido, por lo que no realiza la corrección de la señal y estas no tienen memoria.
3. Grado de Conexión
   a) *Totalmente Conectadas*: Conexión entre las neuronas y el número de capas asignadas a la estructura.
   b) *Parcialmente Conectadas*: No se da la conexión total entre las neuronas y el número de capas asignadas a la estructura.



II-B.3. Mecanismo de Aprendizaje: El aprendizaje de una red neuronal artificial corresponde a la asignación de pesos sinápticos aleatorios o nulos y por el método de aprendizaje, al diseñar un modelo se especifica el tipo de estructura y un tipo de entrenamiento, el entrenamiento de la red neuronal se lleva a cabo en dos niveles:

1. Modelado por sinapsis: Consiste en modificar los pesos sinápticos siguiendo una cierta regla de aprendizaje, construida normalmente a partir de la optimización de una función de error, que mide la eficacia actual de la operación de la red. Si denominamos $w_{ij}(t)$ al peso que conecta la neurona presináptica j con la postsináptica i en la iteración t, el algoritmo de aprendizaje, en función de las señales que en el instante t llegan procedentes del entorno, proporcionará el valor $\Delta w_{ij}(t)$ que da la modificación que se debe incorporar en dicho peso, el cual quedará actualizado de la forma:

$$\Delta w_{ij}(t-1) = w_{ij}(t) + \Delta w_{ij}(t) \qquad (3)$$

   El proceso de aprendizaje es usualmente iterativo, actualizándose los pesos de la manera anterior, una y otra vez, hasta que la red neuronal alcanza el rendimiento deseado.
2. Modelado por aprendizaje: Dada a la arquitectura neuronal creada se realiza una modificación por el método de supervisión para la optimización deseada:
   a) Supervisado: Por el método supervisado presenta a la red las salidas que debe proporcionar ante los patrones de entrada. Se observa la salida de la red y se determina la diferencia entre ésta y la señal deseada. Para realizar esto es necesario presentar un conjunto de datos o patrones de entrenamiento para determinar los pesos o parámetros de diseño de las interconexiones de las neuronas. Posteriormente, los pesos de la red son modificados de acuerdo con el error cometido. Manjarrez (2014) [7]
   Este aprendizaje admite dos variantes:
      1) Aprendizaje por refuerzo: Sí la salida de la red corresponde o no con la señal deseada, es decir, la información es de tipo booleana verdadero o falso.
      2) Aprendizaje por corrección: Conocemos la magnitud del error y ésta determina la magnitud en el cambio de los pesos

b) No Supervisado: No se conoce la salida que debe presentar la red neuronal, la red en este caso se organiza ella misma agrupando, según sus características, los diferentes patrones de entrada. Estos sistemas proporcionan un método de clasificación de las diferentes entradas mediante técnicas de agrupamiento o clustering. Manjarrez (2014) [7]

II-B.4. Clasificación por Aprendizaje: El tipo y clasificación de modelos de redes neuronales dado a su estructura, algoritmo de aprendizaje y tipo de conexión se presenta en la siguiente figura 5 :

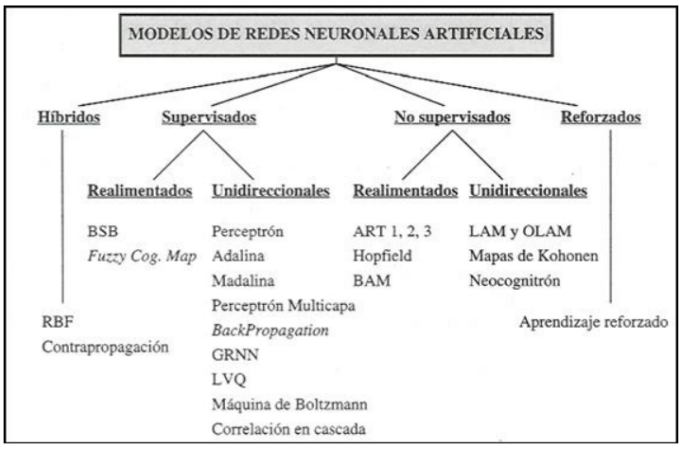


Fig. 5. *Clasificación Modelo de Algoritmo de Aprendizaje*

II-C. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial (MSV) es un método de aprendizaje supervisado por clasificación y regresión, es un modelo probabilístico avanzado. Para un método de clasificación, el modelo realiza un entrenamiento con un conjunto de datos en que realiza un mapeo de los datos, en que son clasificados a un alto espacio de características dimensionales separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano, y se muestra en la figura 6. Para un método de regresión, realiza un entrenamiento de con un conjunto de datos para el método de separación de linealidad, esto quiere decir que no realiza la clasificación dado a un hiperplano, entonces, el modelo probabilístico MSV realiza una curva de tendencia para la separación eficiente de clasificación dado a las diferentes funciones de kernel.

Las MSV fueron propuestas por *Vapnik* en la década de 1960 y su equipo en los laboratorios AT&T. Se han convertido en un área de intensa investigación debido a la evolución en el técnicas y teoría junto con extensiones a la regresión y la estimación de densidad. Burbidge & Buxton (2001) [3]

II-C.1. Aprendizaje Supervisado MSV: El problema general del aprendizaje automático es buscar un espacio generalmente muy grande de hipótesis potenciales para determinar cuál se ajustará mejor a los datos. Los datos pueden estar etiquetados o no etiquetados, si se dan etiquetas entonces el problema es uno de aprendizaje supervisado en el que la respuesta verdadera es conocida para un conjunto dado de datos, si las etiquetas son categóricas entonces el problema es de clasificación, si las etiquetas son de valor numérico el problema es uno de regresión. Si no se dan las etiquetas, entonces el problema es uno de aprendizaje no supervisado y el objetivo es caracterizar la estructura de los datos.

II-C.2. Método de Clasificación MSV: Los métodos de clasificación supervisada son datos de entrada vistos por vector p-dimensional; dado a un conjunto de datos de entrenamiento por un modelo probabilístico que busca en encontrar subconjuntos de datos y separarlos por categorías en un posible número de p - hiperplanos, además, por diferentes métodos y algoritmos se busca en predecir un punto y describir a que categoría pertenece. Parra (2017) [9]

El límite máximo hiperplanos busca en encontrar una separación óptima y la mayor distancia de separación del conjunto de datos de la superficie que son clasificados por una categoría dada a los vectores de soporte, estos soportes definen la calidad de clasificación y de la categoría dado a la distancia máxima en concepto de separación óptima como se muestra en la figura 6. Burbidge & Buxton (2001) [3] cita en su artículo que la formulación del aprendizaje y el entrenamiento de los datos cuando son linealmente separables entonces (w_i, b).

Donde w es el vector del peso y b es el sesgo que se denomina el límite tal que

$$\begin{cases} H_1 : w^T x_i + b \geq 1, & \textit{para todo } x_i \in P \\ H_2 : w^T x_i + b \leq -1, & \textit{para todo } x_i \in N \end{cases} \tag{4}$$

Dada la regla de decisión por

$$f_{w,b}(x) = signo(w^T x + b) \quad (5)$$

Las restricciones de desigualdad de la ecuación 5 se pueden combinar para dar

$$y_i \left(w^T x_i + b\right) \geq 1, \quad \textit{para todo } x_i \in P \cup N \quad (6)$$

Sin pérdida de generalidad, el par (w, b) pueden cambiar de escala de manera que

$$\min_{i=1...l} \mid w^T x_i + b \mid = 1$$

De esta forma, los puntos del vector que son etiquetados con una categoría en un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado.

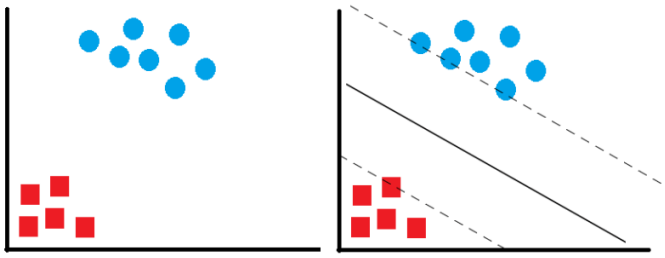


Fig. 6. *Método de Clasificación SVM*

II-C.3. Método de Regresión MSV: Se quiere determinar una probabilidad que se cometa un delito empleando una regresión; la regresión se basa en buscar la curva que modele la tendencia de los datos y, según ella, predecir cualquier otro dato en el futuro. Podremos definir siempre minimizando el error con las MSV en que garantizan una bondad de ajuste a la línea de tendencia.

En problemas no lineales siempre será posible utilizar la función del método de kernel, tras resolver un número de hipótesis del problema de dimensión en un hiperplano el conjunto de datos no es de separación lineales si no oblicuas, se busca que el modelo MSV por el método de kernel obtenga el ajuste de bondad de los datos. Burbidge & Buxton (2001) [3] cita que la flexibilidad de las propiedades de las funciones de kernel que permite a MSV a un buen ajuste de bondad.

1. Núcleo Polinómico de Grado

$$h : K(x_i, x_j) = ((x_i)(x_j + 1))^h \quad (7)$$

2. Núcleo de la Función Base Radial de Gauss:

$$K(x_i, x_j) = \exp\left\{-\frac{\mid\mid x_i, x_j \mid\mid^2}{2\sigma^2}\right\} \quad (8)$$

3. Núcleo Sigmoide :

$$K(x_i, x_j) = \tanh\left((kx_i)(x_j - \delta)\right) \quad (9)$$

III. Metodología

El conjunto de datos contiene los registros de denuncias que corresponden a los años 2017 y 2018 del departamento de Cundinamarca y del municipio de Bogotá, el número de registros son de 160446 individuos con 7 variables; se realiza la limpieza de datos faltantes y transformación de nuevas variables para procesamiento en Python. El objetivo es de identificar por medio de tres modelos supervisados que son un Random Forest, Red Neuronal Artificial y Máquinas de Soporte Vectorial para realizar la clasificación optima y confiable de los días de la semana, esta variable de interés representa la frecuencia que mayor tendencia en hurtos se presenta, además, las variables independientes son estadísticamente significativas para el análisis de clasificación:

1. Arma empleada
2. Método agresor
3. Método victima
4. Sexo
5. Estado Civil
6. Escolaridad

Se realizará un análisis descriptivo estadístico muy breve pero significativo para conocer su comportamiento y tendencia. También, para ello se realizará con un conjunto de datos de entrenamiento para el aprendizaje de los modelos y no tener sobre ajuste, además, se prueban con un conjunto de datos y a su vez se validarán sus pronósticos con las métricas de de clasificación llamadas Accuracy, Recall, Fbeta Score y Curva Roc.

III-A. Análisis Descriptivo

En la tabla I, representan la tasa de hurtos (%) por diversos métodos del empleo de armas por semana, pero como podemos analizar, que existe tasas altas en hurtos sin empleo de armas, esto quiere decir que

son timados por expertos estafadores y para ello no requieren de uso de la fuerza. En el segundo puesto y tercero están los que son empleados con arma cortopunzante y de fuego.

Método	Domingo	Jueves	Lunes	Martes	Miércoles	Sábado	Viernes
ARMA BLANCA	2.92	4.27	3.73	4.13	3.02	4.09	4.74
ARMA DE FUEGO	0.76	1.76	1.50	1.68	1.27	1.42	1.88
CONTUNDENTES	0.51	0.85	0.71	0.83	0.63	0.75	0.92
CORTANTES	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DIRECTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ESCOPOLAMINA	0.18	0.11	0.07	0.07	0.06	0.27	0.19
JERINGA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LLAVE MAESTRA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NO REPORTADO	0.16	0.26	0.21	0.24	0.22	0.24	0.28
PALANCAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PERRO	0.00	0.00	0.00	0.00	0.00	0.00	0.01
SIN EMPLEO DE ARMAS	4.76	9.28	7.48	8.60	6.73	8.00	10.16
SUSTANCIAS TOXICAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE I

Tasa (%) Arma Empleada por día Semanal

En la tabla II, medio de transporte predilecto por los delincuentes en la ciudad de Bogotá por semana, tasa de mayor hurto por medio de movilidad es a pie y de mayor preferencia para los jueves, viernes y sábados para los delincuentes. En segundo y tercer puesto están para los que prefieren la bicicleta y la moto.

Método	Domingo	Jueves	Lunes	Martes	Miércoles	Sábado	Viernes
A PIE	7.36	11.71	9.88	10.90	8.20	10.96	12.90
BICICLETA	0.49	0.82	0.67	0.85	0.63	0.71	0.91
CONDUCTOR BUS	0.01	0.03	0.03	0.02	0.02	0.02	0.03
CONDUCTOR MOTOCICLETA	0.25	0.79	0.66	0.73	0.65	0.57	0.83
CONDUCTOR TAXI	0.20	0.15	0.12	0.11	0.11	0.32	0.26
CONDUCTOR VEHÍCULO	0.16	0.24	0.19	0.24	0.19	0.28	0.30
PASAJERO BUS	0.37	1.73	1.31	1.72	1.36	1.08	1.80
PASAJERO MOTOCICLETA	0.26	0.81	0.61	0.73	0.57	0.57	0.84
PASAJERO TAXI	0.09	0.10	0.08	0.08	0.06	0.11	0.10
PASAJERO VEHÍCULO	0.09	0.17	0.17	0.19	0.15	0.16	0.20

TABLE II

Tasa (%) Método del Agresor por día Semanal

En la tabla III, los solteros presentan la mayor tasa de hurtos en Bogotá, por lo que esta población de personas en su mayoría son jóvenes, y son blancos muy fáciles para los delincuentes, esto es de mayor frecuencia cuando salen a sectores aptos para discotecas, tabernas entre otros. También, los de estado civil casados y de unión libre hay tasas muy parecidas entre ellas de denuncias presentadas por hurtos.

Estado Civil	Domingo	Jueves	Lunes	Martes	Miércoles	Sábado	Viernes
CASADO	1.70	3.41	2.84	3.28	2.51	2.83	3.72
DIVORCIADO	0.14	0.26	0.20	0.24	0.20	0.22	0.29
SEPARADO	0.19	0.39	0.34	0.35	0.28	0.35	0.41
SOLTERO	5.50	9.25	7.53	8.67	6.64	8.47	10.33
UNION LIBRE	1.67	3.08	2.67	2.89	2.22	2.78	3.27
VIUDO	0.08	0.15	0.13	0.15	0.10	0.11	0.15

TABLE III

Tasa (%) Estado civil por día de la Semana

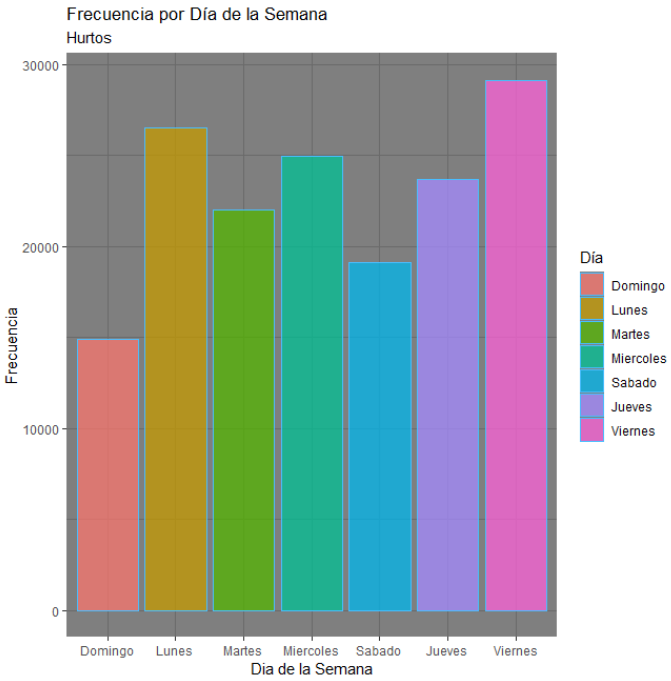


Fig. 7. *Frecuencia por Día de la Semana*

La figura 7, podemos observar que se presenta un mayor número de hurtos para los días de lunes a sábado, pero con mayor frecuencia significativa son los lunes y viernes, también los martes, miércoles y jueves están con la misma frecuencia de hurtos, esto conlleva a tomar políticas de seguridad para ciertos días y horas específicas.

La figura 8, presenta una mayor frecuencia de hurtos para los hombres los días correspondientes que son domingos, lunes y martes, además, para las mujeres tiene el mismo comportamiento de días que el de los hombres.

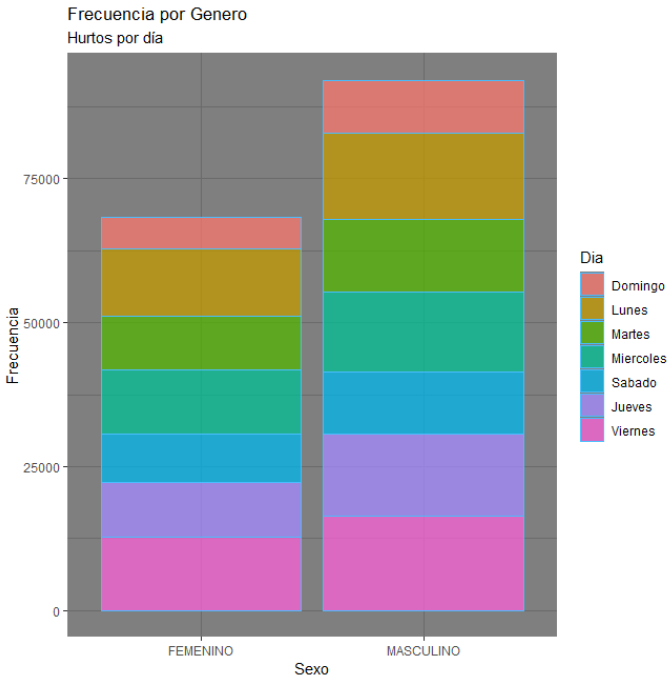


Fig. 8. *Frecuencia por Genero y Día de la Semana*

III-B. Modelos

Se realiza el proceso de aprendizaje de los modelos probabilísticos con el software *Python*, además, se toman dos conjuntos de datos 70 % entrenamiento y 30 % de prueba.

III-B.1. Random Forest: presenta una muy mala calidad de clasificación por día de la semana, su matriz de confusión de la tabla IV se puede observar como clasifica de mal, además, se realizó sus métricas de validación según tabla V accuracy 18 % tasa de error de clasificación del 82 %, f beta score del 14 % su promedio ponderado de recall con precisión y su curva Roc del 51 %.

	Domingo	Lunes	Martes	Miércoles	Sábado	Jueves	Viernes
Domingo	89	522	139	263	41	410	1462
Lunes	75	1059	271	527	98	597	2745
Martes	91	941	221	477	92	488	2127
Miércoles	66	1039	263	532	91	511	2503
Sábado	54	790	187	420	90	427	1883
Jueves	100	841	211	434	101	676	2321
Viernes	107	1134	295	545	117	633	2984

TABLE IV

Matriz de Confusión Random Forest

Modelo	Accuracy	FBeta Score	CURVA ROC
Random Forest	18	14	51

TABLE V

Tasa (%) Métricas de Validación Random Forest

III-B.2. Las Redes Neuronales Artificiales: presenta un mal desempeño y genera muy malas clasificaciones por día de la semana, su matriz de confusión según tabla VI se puede observar como clasifica de mal, además, sus métricas de validación según tabla VII accuracy 9.1 % tasa de error de clasificación del 90.9 %, f beta score del 1.5 % su promedio ponderado de recall con precisión y su curva Roc del 50 %.

	Domingo	Lunes	Martes	Miércoles	Sábado	Jueves	Viernes
Domingo	2939	0	0	0	0	0	0
Lunes	5277	0	0	0	0	0	0
Martes	4362	0	0	0	0	0	0
Miércoles	4990	0	0	0	0	0	0
Sábado	3749	0	0	0	0	0	0
Jueves	4840	0	0	0	0	0	0
Viernes	5933	0	0	0	0	0	0

TABLE VI

Matriz de Confusión RNA

Modelo	Accuracy	FBeta Score	Curva Roc
Red Nuronal	9.1	1.5	50.0

TABLE VII

Tasa (%) Métricas de Validación RNA

III-B.3. Máquinas de Soporte Vectorial: presenta un mal desempeño y genera muy malas clasificaciones por día de la semana, su matriz de confusión según tabla VIII se puede observar como clasifica de mal, además, sus métricas de validación según tabla IX accuracy 18.1 % tasa de error de clasificación del 81.9 %, f beta score del 11.9 % su promedio ponderado de recall con precisión y su curva Roc del 50 %.

	Domingo	Lunes	Martes	Miércoles	Sábado	Jueves	Viernes
Domingo	17	751	20	104	10	271	1837
Lunes	5	1274	33	239	13	242	3353
Martes	8	1092	39	214	14	228	2786
Miércoles	7	1267	34	271	21	243	3186
Sábado	8	958	26	212	14	196	2551
Jueves	22	1117	20	196	14	389	2984
Viernes	15	1322	43	272	15	330	3807

TABLE VIII

Matriz de Confusión Máquinas de Soporte Vectorial

Modelo	Accuracy	FBeta Score	Curva Roc
M.S.V	18.1	11.9	50

TABLE IX

Tasa (%) Métricas de Validación M.S.V

IV. Conclusiones

Los modelos probabilísticos de clasificacion no obtuvieron muy buenos resultados, dado a que los algoritmos utilizados para este conjunto de datos no fueron los adecuados. pero el modelo de pésimos resultados fueron las redes neuronales, según tabla VI la matriz de confusión clasifico todos los hurtos para el domingo, además sus métricas de validación son muy pobres.

Para el entrenamiento de los modelos son de un alto costo computacional en Python y tiempo de espera, sus predicciones de clasificación no fue la adecuada para el conjunto de datos de prueba.

List of Figures

List of Tables

REFERENCES

[1] Amat, Joaquin Rodrigo (2017) *Árboles de predicción: Bagging, Random Forest, Boosting y C5.0* https://rpubs.com/Joaquin_AR/255596

[2] Breiman Leo (2001) *Random Forests* Statistics Department, University of California, Berkeley, CA 94720

[3] Burbidge Robert & Buxton Bernard (2001) *An Introduction to Support Vector Machines for Data Mining* . Computer Science Dept., UCL, Gower Street, WC1E 6BT, UK

[4] Campos Yepes John Jairo (2017) *Modelos Apilados y factores que pueden afectar la eficiencia*. Universidad Santo Tomás sede Bogotá, Trabajo de Grado

[5] González Martínez Edwin Fernando (2018) *Detección de Fraude en Tarjetas de Crédito Mediante Técnicas de Minería de Datos*. Universidad Santo Tomás sede Bogotá, Trabajo de Grado

[6] Han Jiawei, Kamber Micheline & Pei Jian (2014) *Data Mining Concepts and Techniques* Third Edition, Elsevier Science, ISBN libro electrónico 9780123814807

[7] Manjarrez Lino (2014) *Relaciones Neuronales Para Determinar la Atenuación del Valor de la Aceleración Máxima en Superficie de Sitios en Roca Para Zonas de Subducción*. https://www.researchgate.net/publication/315762548

[8] Makhabel Bater (2015) *Learning Data Mining With R* ProQuest Ebook Central

[9] Parra Francisco (2017) *Estadística y Machine Learning con R* https://rpubs.com/PacoParra/293405

[10] Torgo Luis (2011) *Data Mining with R Learning With Case Studies* Chapman & Hall / CRC, ISBN 9781439810187

[11] Yanchang Zhao, Yonghua Cen, & Justin Cen (2013) *Data Mining Applications with R* Elsevier Science, ISBN libro electrónico 9780124115200