UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

**Facultad de Estadística**
Maestría en Estadística Aplicada
Trabajo de grado
Marzo 2022

# MODELOS DE MACHINE LEARNING PARA CLASIFICAR LA CARTERA EN UN FONDO DE PENSIONES

## MACHINE LEARNING MODELS TO CLASSIFY THE PORTFOLIO IN A PENSION FUND

Ricardo Gil Rubio[a]
ricardo.gil@usantotomas.edu.co

Edwin Andrés Cruz Perez[b]
dir.estadisticaaplicada@usantotomas.edu.co

Oscar Perdomo Charry[c]
oscarjulian.perdomo@gmail.com

**Resumen**

El presente trabajo tiene como objetivo, a través de la aplicación de diferentes técnicas de Machine Learning y diagnósticos estadísticos e inferenciales, proponer modelos de análisis predictivos que permitan identificar, clasificar y procesar oportunamente cuáles son las empresas que no pagan los aportes de pensión a sus trabajadores afiliados al fondo de pensiones, y así implementar diferentes estrategias de cobro encaminadas a recuperar los dineros adeudados.

En el proceso de evaluación de rendimiento de los modelos se logró evidenciar que la técnica Árboles de Decisión presenta excelentes resultados: no requirió estandarización de los datos al lograr un porcentaje de certeza excelente y clasificó de forma rápida y eficiente la variable predictora en una base de datos con un número adecuado de registros. Las demás técnicas mostraron buenos resultados en la clase tipo 0, 3 y 4 con porcentajes superiores al 96,8 % tanto en exhaustividad como en medida-F, mientras se redujo el desempeño para las técnicas Regresión Logistica 71,8 % y Máquinas de Vectores de Soporte 69,2 % en exhaustividad y Redes Bayesianas 18,5 % en medida-F, lo anterior para la clase tipo 1. En la técnica Redes Bayesianas para la clase tipo 2 se redujo en 24,7 % y 29,3 % tanto en exhaustividad como en medida-F y Máquinas de Vectores de Soporte en 59,4 % para medida-F. Lo anterior se abordo con el tratamiento de clases desbalanceadas y con los algoritmos de refuerzo o conjunto.

El desequilibrio de clases es una problema bastante frecuente cuando se trabaja con datos reales; cuando muestras de una o de múltiples clases están sobrerepresentadas en un conjunto de datos. Existen varios ámbitos en los que puede ocurrir, como el filtrado de spam, detección de cancer, la identificación de fraude o la detección de enfermedades. Las estrategias para tratar el desequilibrio de clases incluyen el muestreo ascendente de la clase minoritaria, el muestreo descendente de la clase mayoritaria y la generación de muestras de entrenamiento sintéticas mediante el algoritmo más utilizado (SMOTE, por sus siglas en Inglés).

Una vez evaluados los modelos con la segmentación propuesta se generaron las estrategias que permitieron identificar los mecanismos de gestión de cobro dependiente del tipo de deudor, esto va, desde una visita comercial, gestión de contact center para cobro preventivo o un extracto con información de pagos, para deudores de baja criticidad, pasando por una carta de cobro persuasivo, asesoramiento en los puntos de atención o mensajes de texto para deudores de criticidad media, hasta el proceso de cobro coactivo, embargos y demás medidas para los deudores que son renuentes al pago.

[a]Estudiante
[b]Director
[c]Director

***Palabras clave**: machine learning, regresión logística, máquinas de vectores de soporte, árboles de decisión, redes neuronales, redes bayesianas, cartera, fondos de pensiones, mora.*

**Abstract**

The present paper has as objective, the application of different Machine Learning techniques as well as statistical and inferential diagnostics, to propose predictive analysis models that allow to in due time identify, classify and process the companies that are not paying pension contributions to their employees affiliated to the pension fund, and thus to implement different collection strategies to recover contributions owed.

In the process of evaluating the performance of the models, it was possible to show that the Decision Trees technique presents excellent results: it did not require standardization of the data by achieving an excellent percentage of certainty and it quickly and efficiently classified the predictor variable in a database with an adequate number of records. The other techniques showed good results in class type 0, 3 and 4 with percentages above 96.8 % both in completeness and in measure-F, while the performance decreased for Logistic Regression 71.8 % and Support Vector Machines 69.2 % in completeness and Bayesian Networks 18.5 % in measure-F, the above for class type 1. In the Bayesian Networks technique for class type 2 it was reduced by 24.7 % and 29.3 % both in completeness and F-measure and Support Vector Machines at 59.4 % for F-measure. This was addressed with the treatment of unbalanced classes and with the reinforcement or ensemble algorithms.

Class imbalance is a fairly common problem when working with real data; when samples from one or multiple classes are over represented in a data set. There are several areas in which it can occur, such as spam filtering, cancer detection, fraud identification or disease detection. Strategies to deal with class imbalance include minority class up sampling, majority class down sampling, and generation of synthetic training samples using the most commonly used algorithm (SMOTE).

Once the models with the proposed segmentation were evaluated, the strategies were generated that allowed identifying the collection management mechanisms depending on the type of debtor, this ranges from a commercial visit, contact center management for preventive collection or an extract with payment information, for debtors of low criticality, going through a persuasive collection letter, advice at service points or text messages for debtors of medium criticality, to the coercive collection process, embargoes and other measures for debtors who are reluctant to pay.

***Keywords**: machine learning, logistic regression, decision trees, neural networks, bayesian networks, support vector machines, portfolio, pension funds, arrears.*

# 1. Introducción

La seguridad social, en su sentido más amplio, significa la obligación de un estado de garantizar al individuo y a la comunidad el goce de una calidad de vida alta, entendiendo por ello los mecanismos para que las personas obtengan condiciones básicas de subsistencia e igualdad de recursos económicos, sociales y culturales (COLFONDOS,2003).

En Colombia, el Sistema de Seguridad Social fue instituido por la ley 100 de 1993 y reúne de manera coordinada un conjunto de entidades, normas y procedimientos a los cuales pueden tener acceso las personas y la comunidad con el fin principal de garantizar una calidad de vida que esté acorde con la dignidad humana. De acuerdo con dicha ley, el Sistema de Seguridad Social se compone de los sistemas de pensiones, de salud y de riesgos laborales y del Servicio Social Complementario de Beneficios Económicos Periódicos (COLFONDOS,2003).

La ley 100 creó en materia de pensiones dos únicos regímenes: uno público, el Régimen de Prima Media con prestación definida, y otro privado, el Régimen de Ahorro Individual con solidaridad. Los trabajadores colombianos están en la libertad de escoger cualquiera de estos y realizar las respectivas contribuciones en su etapa laboral.

Las cotizaciones son los aportes en dinero que deben hacer mensualmente los trabajadores independientes o dependientes por medio de sus empleadores. Ellos pagan un porcentaje con el objetivo de que en un futuro puedan tener derecho al pago de las pensiones de invalidez, vejez y sobrevivencia. Tanto el empleador como el trabajador dependiente están obligados a efectuar aportes a uno de los dos regímenes de acuerdo con el ingreso base de cotización sobre el cual se hacen las cotizaciones al Sistema de Seguridad Social en Colombia (COLFONDOS,2003).

De esa manera, el derecho a la seguridad social está consagrado como un servicio público a cargo del Estado. El instrumento jurídico establecido por la ley 100 de 1993 determinó precisamente la estructura contributiva del sistema para garantizar el reconocimiento y pago de las prestaciones económicas a los sujetos protegidos. Cuando se detecta una evasión de aportes por parte del empleador, al no realizar los pagos a la seguridad social de sus trabajadores, podrá ser objeto de multas y sanciones impuestas por el Ministerio de Trabajo, la Superintendencia Financiera o la UGPP. En caso de no realizar el cobro de aportes oportunamente a la empresa, el fondo de pensiones puede acarrear sanciones jurídicas y estará obligado a responder por los aportes de sus afiliados.

En cifras registradas por la Unidad de Gestión Pensional y Parafiscal (UGPP), para el 2012 más de 900.000 empleadores evadieron el pago de al menos uno de sus empleados, lo que equivale al 26 %, es decir, 14,6 billones de pesos anuales. En la actualidad, la morosidad que se presenta es de un 17 %, esto es, alrededor de 10 billones de pesos; además, el número de obligados a aportar subió a 800.000 personas y la población que realmente pagó creció en dos millones, lo que revela una caída de 1.2 millones en morosos durante los últimos cinco años (Olarte,2016).

En el marco del crecimiento de las deudas en los fondos de pensiones públicos en Colombia, generado por el incumplimiento por parte de las empresas al no pagar obligaciones pensionales a sus empleados, las técnicas de *Machine Learning* se presentan como una herramienta de gran utilidad en este campo. Pues no solamente permiten detectar a las empresas morosas, sino también mitigar ese tipo de riesgos clasificando la cartera: ya sea mediante la automatización de procesos o por intermedio del conocimiento que puedan adquirir las máquinas de comportamientos pasados en el pago de los aportes.

De hecho, en las técnicas de *Machine Learning*, los métodos de aprendizaje supervisado presentan una relevancia significativa capaz de entrenar el algoritmo a partir de datos que vienen etiquetados con la respuesta correcta, la cual se construye en base al conocimiento de los expertos del sector pensional. Se trata, por lo tanto, de un sistema que por sí mismo sin intervención humana y en forma automatizada-aprende a descubrir patrones, tendencias y relaciones en los datos, y gracias a dicho conocimiento, en cada interacción con información nueva, ofrecer mejores perspectivas, soluciones.

En las páginas que siguen, con la aplicación de este tipo de técnicas, se busca, por un lado, obtener conocimiento sobre el perfilamiento del deudor y, por el otro, proporcionar información sobre si es considerado moroso o no, y si lo es tener una mayor certeza de cómo enfocar los recursos en el cobro de los aportes. La incorporación metodológica del análisis estadístico, por otra parte, pretende disminuir el riesgo de tomar decisiones en la gestión de cobro, bien sea mediante mecanismos de percepción o a discreción. Esto permite establecer reglas para llevar a cabo un proceso de detección claro y confiable que, a su vez, posibilite focalizar estrategias de recuperación de cartera eficientes, lo que se traduce en una disminución de los costos en la operación.

Tradicionalmente, la clasificación de la cartera histórica se realiza teniendo en cuenta la demanda que pueda existir para cobrar los dineros sobre ciertas empresas, que posiblemente no están priorizadas y que la administradora enfoca esfuerzos y recursos en depuración de cartera, más no en recuperación. Además, se hace necesario ejecutar el modelo con herramientas computacionales robustas que permitan detectar esos patrones de comportamiento en grandes volúmenes de información como son las bases de empresas y pagos de la entidad.

# 2. Marco teórico y revisión de literatura

En este capítulo se estudia la teoría sobre las técnicas de *Machine Learning* aplicadas en problemas con grandes volúmenes de información al clasificar la cartera en un fondo de pensiones. La finalidad de esta primera parte del trabajo es describir los métodos de clasificación de empleadores morosos con algoritmos de *Machine Learning* como herramienta metodológica para descubrir patrones en bases de datos de pagos y deudas que a simple vista no son posibles de encontrar.

## 2.1. Machine Learning

El *Machine Learning* (aprendizaje automático) es una disciplina de las ciencias informáticas relacionada con el desarrollo de inteligencia artificial, así como también con el análisis predictivo o aprendizaje estadístico. Básicamente, hace referencia a la capacidad de un software, una máquina o un aparato de aprender y predecir comportamientos futuros mediante la adopción de ciertos algoritmos de su programación respecto a cierta entrada de datos en su sistema. En otras palabras, a la aplicación de un conjunto extenso de reglas o heurísticas que ofrecen un amplio abanico de soluciones para casos en los que se dispone de un conjunto de datos muy grandes y difíciles de interpretar y clasificar.

La variedad de aplicaciones y usos que se le han dado van desde el mercadeo, análisis de datos y filtros *antispam* para correo electrónico, hasta la conducción automática de vehículos y el reconocimiento de voz e imágenes. Es más, para muchas industrias del mundo:

*Machine Learning* se está convirtiendo en un factor que detona una profunda innovación, dado que su aplicación contribuye a mejorar procesos, generar nuevas oportunidades de negocio (antes imperceptibles o cuya detección no era sencilla) y definir nuevos parámetros empresariales (por ejemplo, la forma en que se consigue la satisfacción del cliente). Todo esto desde la propia solución tecnológica, sin que una instancia humana tenga que estudiar los datos, valorarlos y generar una respuesta.

Entre los sectores que ya aprovechan esta innovación analítica, se destacan las empresas de servicios, que detectan insights en la información de sus clientes y sus operaciones. Esto les permite en forma automatizada, sin la intervención de un equipo de ejecutivos recomendar productos financieros al usuario indicado y en el momento preciso (por ejemplo, una oportunidad de inversión), detectar a clientes con un alto perfil de riesgo, descubrir operaciones anómalas que podrían esconder fraudes, entre otras posibilidades (Statistical Analysis System Institute, 2019).

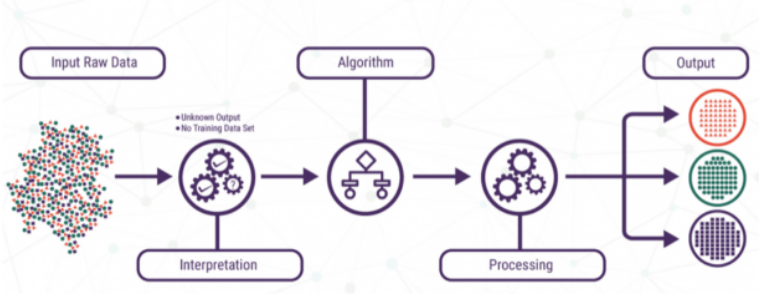
Como veremos particularmente en este trabajo, otro de los usos más relevantes de los modelos de *Machine Learning* es el de implantar controles que mitiguen riesgos relacionados con la seguridad de la información. El sector de seguridad social, específicamente el de los fondos de pensión, no es la excepción, sobre todo si tenemos en cuenta el alto volumen de datos de información existente allí que es difícil de identificar y procesar (aproximadamente 60 millones de registros con deudas que alcanzan los 14 billones de pesos). Con ellos, incluso, es posible generar patrones para la toma de decisiones, así como también estrategias de retención y fidelización de las empresas. Así, las empresas morosas se pueden clasificar de forma oportuna mes a mes y ser categorizadas de acuerdo al nivel de criticidad e importancia que tienen para el fondo de pensiones.

Los métodos de *Machine Learning* presentan diferentes categorías las cuales se presentan a continuación.

## 2.2. Aprendizaje no supervisado

Los métodos no supervisados son algoritmos que basan su proceso de entrenamiento en un conjunto de datos sin etiquetas o clases previamente definidas. Es decir, a priori no se conoce ninguna variable objetivo o predictora, ya sea categórica o numérica. En líneas generales como se ve representado en la figura 1, el aprendizaje no supervisado está basado en tareas de agrupamiento o clustering, donde el objetivo central es encontrar grupos homogéneos entre sí y heterogéneos entre los otros (López,2015).

Figura 1: Aprendizaje no supervisado

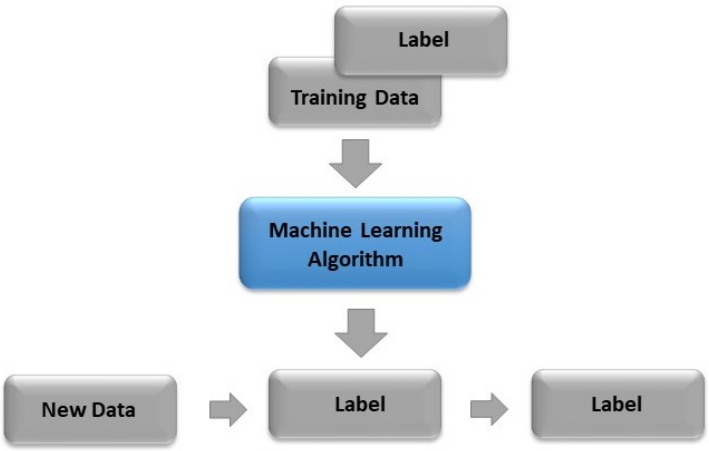



*Fuente. Adaptado de ¿En qué consiste el Machine Learning?, por Ironhack, 2015*

## 2.3. Aprendizaje supervisado

En los problemas de aprendizaje supervisado se enseña o entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta como se ilustra en la figura 2. Cuanto mayor es el conjunto de datos, mayor es la posibilidad de que el algoritmo aprenda sobre el tema. Una vez concluido el entrenamiento, se le brindan nuevos datos, ya sin las etiquetas de las respuestas correctas, y el algoritmo de aprendizaje utiliza la experiencia pasada que adquirió durante la etapa de entrenamiento para predecir un resultado (López,2015).

Figura 2: Aprendizaje Supervisado




*Fuente. Adaptado de Machine Learning con Python, por R. López, 2015*

### 2.3.1. Regresión Logística binomial y multinomial

De acuerdo con Agresti (2002), la Regresión Logística es el modelo más importante para los datos de respuesta categórica. Se utiliza cada vez más en una amplia variedad de aplicaciones. Los primeros usos fueron en estudios biomédicos, pero en los últimos veinte años también ha tenido un uso intensivo

en la investigación de las ciencias sociales y el marketing. Recientemente, la Regresión Logística se ha convertido en una herramienta popular en las aplicaciones empresariales. Algunas aplicaciones de credit scoring la utilizan para modelar la probabilidad de que un sujeto pueda acceder a un crédito. Por ejemplo, la probabilidad de que un sujeto pague una factura a tiempo puede utilizar predictores tales como el valor de la factura, el ingreso anual, la ocupación, las obligaciones hipotecarias y de deuda, porcentaje de facturas pagadas a tiempo en el pasado, y otros aspectos del historial de crédito de un solicitante.

El modelo de Regresión Logística no requiere de los supuestos de la Regresión Lineal, como son el supuesto de normalidad de los errores y homocedasticidad, además de que las variables involucradas sean continúas (Nieto, 2010). En este sentido, la Regresión Logística se aplica tanto a datos que se distribuyen con normalidad en sus residuos como también a datos que no lo son. Por lo tanto, el modelo de regresión logística es útil cuando la variable respuesta Y no está distribuida normalmente y tanto las variables predictoras como de respuesta tienen valores discretos, categóricos, ordinales o no ordinales.

La capacidad predictiva del modelo logístico se valora mediante la comparación entre el grupo de pertenencia observado y el pronosticado. El modelo debe ser capaz de clasificar a los individuos en cada uno de los grupos, basado en las variables que definen las características de estos. Si una variable cualitativa con dos niveles se codifica como 0 y 1, matemáticamente es posible ajustar un modelo de Regresión Lineal por mínimos cuadrados transformando el valor devuelto por la regresión lineal:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

Incluye: $y$, variable dependiente,

$\beta_0$, la ordenada al origen,

$\beta_1$, la pendiente de la recta de regresión,

$x$, variable independiente,

$\varepsilon$, el término del error, la diferencia entre cada uno de los valores observados de la variable $y$, y los valores estimados para cada valor de $x$, de acuerdo con la ecuación de regresión de mínimos cuadrados.

En la Regresión se emplea una función cuyo resultado está siempre comprendido entre 0 y 1, una de las más utilizadas es la función logística (también conocida como función sigmoide):

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

Amat (2016) señala que la Regresión Logística múltiple es una extensión de la Regresión Logística simple. Se basa en los mismos principios, pero ampliando el número de predictores. Los predictores pueden ser tanto continuos como categóricos. La función logística que se utiliza en este modelo es la siguiente:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i \tag{3}$$

$$logit(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i \tag{4}$$

El valor de la probabilidad de Y se puede obtener con la inversa del logaritmo natural:

$$p(Y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i}} \tag{5}$$

Donde $\beta_0, \beta_1, \beta_2, ..., \beta_i$ son los parámetros del modelo, $e$ se refiere a la función exponencial.La expresión del lado derecho de la ecuación se conoce con el nombre de *función logística multivariada.*

### 2.3.2. Árboles de Decisión

Aruna y Nirmala (2013) señalan que un Árbol de Decisión es una estructura de árbol similar a un diagrama de flujo donde un nodo interno representa una característica o atributo; la rama representa una regla de decisión y cada nodo hoja representa el resultado. El nodo superior en un árbol de decisión en Machine Learning se conoce como el nodo raíz, ya que aprende a particionar en función del valor del atributo, divide el árbol de una manera recursiva llamada partición recursiva, esta estructura tipo diagrama de flujo permite la toma de decisiones y es una visualización de cómo un diagrama de flujo imita fácilmente el pensamiento a nivel humano.

En la figura 3 se observa como cada nodo en el árbol actúa como un caso de prueba para algún atributo, y cada borde que desciende de ese nodo corresponde a una de las posibles respuestas al caso de prueba. Este proceso es recursivo y se repite para cada subárbol enraizado en los nuevos nodos.

1. Nodo raíz (nodo de decisión superior): representa a toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.

2. División: es un proceso de división de un nodo en dos o más subnodos.

3. Nodo de decisión: cuando un subnodo se divide en subnodos adicionales, se llama nodo de decisión.

4. Nodo de hoja/terminal: los nodos sin hijos (sin división adicional) se llaman hoja o nodo terminal.

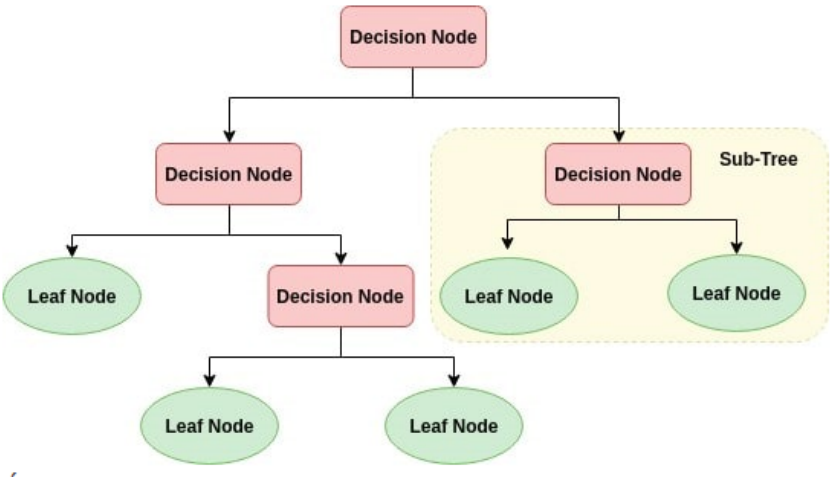
5. Poda: cuando reducimos el tamaño de los árboles de decisión eliminando nodos (opuesto a la división), el proceso se llama poda.

6. Rama/Subárbol: una subsección del árbol de decisión se denomina rama o subárbol.

7. Nodo padre e hijo: un nodo, que se divide en subnodos se denomina nodo principal de subnodos, mientras que los subnodos son hijos de un nodo principal.

Para construir un Árbol de Decisión o clasificación es necesario definir una función que relaciona una variable categórica dependiente (factor) con n variables independientes que pueden ser categóricas o numéricas (Ruiz,2016).

Figura 3: Árbol de Decisión




*Fuente. Adaptado de Árboles de decisión simples y múltiples (random forest), por S. Ruiz, 2016, r-studio-pubs-static*

El algoritmo de clasificación busca cuál es la variable que permite obtener una submuestra más diferen-

ciada para la variable dependiente e identifica también qué intervalo o agrupación de categorías de las variables independientes permitirían maximizar dicha división.

Una vez identificada la variable independiente que permite obtener la clasificación con una mayor capacidad de diferenciación, el proceso se repite reiterativamente en cada uno de los nodos obtenidos hasta que el algoritmo no encuentra diferencias significativas que le permitan seguir profundizando en los nodos.

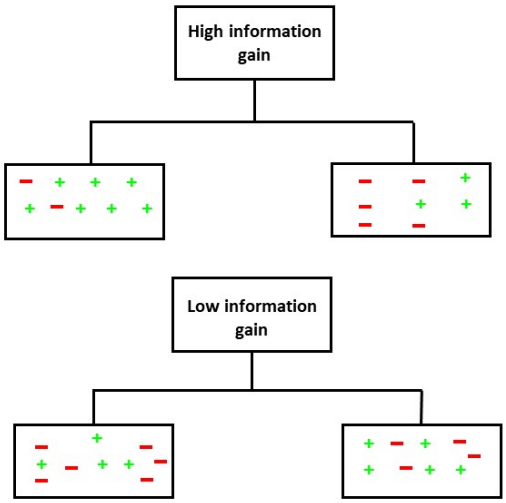
Los Arboles de Decisión están compuestos por nodos interiores, nodos terminales y ramas que emanan de los nodos interiores. Cada nodo interior en el árbol contiene una prueba de un atributo, y cada rama representa un valor distinto del atributo. Siguiendo las ramas desde el nodo raíz hacia abajo, cada ruta finalmente termina en un nodo terminal creando una segmentación de los datos (López,2015).

En palabras de Aruna y Nirmala(2013), la medida de selección de atributos es una heurística para seleccionar el criterio que divide los datos de la mejor manera posible. También se conoce como reglas de división porque ayuda a determinar puntos de interrupción para tuplas en un nodo dado, en el caso de un atributo de valor continuo, los puntos de división para las ramas también deben definirse. Las medidas de selección más populares son ganancia de información, proporción de ganancia e índice de Gini.

### Ganancia de información

Aruna y Nirmala (2013) plantean que la ganancia de información es una propiedad estadística que mide cuán bien un atributo dado separa los datos de entrenamiento de acuerdo con su clasificación objetiva. Tal como se puede apreciar en la figura 4, un ejemplo paralelo considerando la ganancia de información se puede ver que un atributo con baja ganancia de información (abajo) divide los datos de manera relativamente uniforme y como resultado no se acerque más a una decisión. Mientras que un atributo con alta ganancia de información (arriba) divide los datos en grupos con un número desigual de positivos y negativos y, como resultado, ayuda a separarlos entre sí.

Figura 4: Ganancia de Información




*Fuente. Adoptado de Construction of Decision, por R. Aruna K. Nirmala, 2013, International Journal of Advancements in Research Technology*

Distribución de puntos en caso de alta y baja ganancia de información. Para poder calcular la ganancia de información, primero se debe introducir el término entropía de un conjunto de datos.

### Entropía

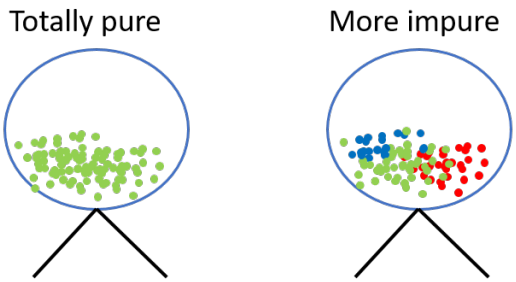
El concepto de entropía mide la impureza del conjunto de entrada (Aruna y Nirmala, 2013). En física y matemáticas, la entropía se conoce como aleatoriedad o impureza en el sistema. En teoría de la

información, se refiere a la impureza en un grupo de ejemplos. La ganancia de información es una disminución de la entropía.

La idea detrás de la entropía es, en términos simplificados, la siguiente: se tiene una rueda de lotería como se ilustra en la figura 5, que incluye 100 bolas verdes. Se puede decir que el conjunto de bolas dentro de la rueda de lotería es totalmente puro porque sólo se incluyen bolas verdes.

Para expresar esto en la terminología de entropía, este conjunto de bolas tiene una entropía de 0 (impureza cero). Considerando ahora, 30 de estas bolas son reemplazadas por bolas rojas y 20 por bolas azules.

Figura 5: Entropía



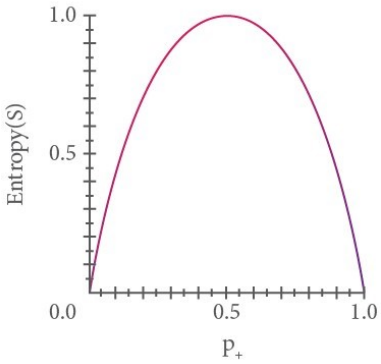

*Fuente. Adoptado de Construction of Decision, por R. Aruna K. Nirmala, 2013, International Journal of Advancements in Research Technology*

Si ahora se saca otra bola de la rueda de loteria,la probabilidad de recibir una bola verde se ha reducido de 1.0 a 0.5. Como la impureza aumentó, la pureza disminuyó, por lo tanto, también aumentó la entropía.

En ese sentido, se puede decir que cuanto más "impuro" sea un conjunto de datos, mayor será la entropía; y cuanto menos "impuro" sea un conjunto de datos, menor será la entropía. Hay que tener en cuenta que la entropía es 0 si todos los miembros de S pertenecen a la misma clase. Por ejemplo, si todos los miembros son positivos, entropía (S) = 0. La entropía es 1 cuando la muestra contiene el mismo número de ejemplos positivos y negativos. Si la muestra contiene un número desigual de ejemplos positivos y negativos, la entropía estará entre 0 y 1.

La figura 6 muestra la forma de la función de entropía en relación con una clasificación booleana, ya que la entropía varía entre 0 y 1.

Figura 6: Función de entropía en relación con una clasificación booleana




*Fuente. Adoptado de Construction of Decision, por R. Aruna K. Nirmala, 2013, International Journal of Advancements in Research Technology*

Se puede calcular la entropía usando la fórmula:

$$Entropy = -p.log_2(p) - q.log_2(q) \tag{6}$$

Aquí, $p$ y $q$ es la probabilidad de éxito y fracaso respectivamente en ese nodo. La entropía también se usa con la variable objetivo categórica, elige la división que tiene la entropía más baja en comparación con el nodo principal y otras divisiones, cuanto menor sea la entropía, mejor será. La ganancia de información calcula la diferencia entre la entropía antes de la división y la entropía promedio después de la división del conjunto de datos en función de los valores de atributo dados.

**Gini**

Según Garcia (2020), el índice de Gini mide el grado de pureza de un nodo y la probabilidad de no sacar dos registros de la misma clase del nodo. A mayor índice de Gini menor pureza, por lo que se selecciona la variable con menor Gini ponderado. Suele seleccionar divisiones desbalanceadas, donde normalmente aísla en un nodo una clase mayoritaria y el resto de clases los clasifica en otros nodos.

Se define el índice de Gini como:

$$1 - \sum_{i=1}^{n} (P_i)^2 \tag{7}$$

Donde $P_i$ es la probabilidad de que un ejemplo sea de la clase $i$.

Aruna y Nirmala (2013) afirman que, si se seleccionan dos elementos de una población al azar, entonces deben ser de la misma clase y la probabilidad de esto es 1 si la población es pura:

1. Funciona con la variable objetivo categórica “éxito” o “fracaso”.

2. Realiza solo divisiones binarias.

3. Cuanto mayor sea el valor de Gini, mayor será la homogeneidad.

4. CART (árbol de clasificación y regresión) utiliza el método Gini para crear divisiones binarias.

### 2.3.3. Redes neuronales

Una red neuronal artificial, es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo estas se organizan formando la estructura del cerebro (Sancho, 2017). El cerebro puede considerarse un sistema altamente complejo, donde se calcula que hay aproximadamente 100.000 millones de neuronas en la corteza cerebral, que forman un entramado de más de 500 billones de conexiones neuronales (una neurona puede llegar a tener 100.000 conexiones, aunque la media se sitúa entre 5000 y 10000 conexiones).

Respecto a su funcionamiento, el cerebro puede ser visto como un sistema inteligente que lleva a cabo tareas de manera distinta a como lo hacen las computadoras actuales. Si bien estas últimas son muy rápidas en el procesamiento de la información, existen tareas muy complejas, como el reconocimiento y clasificación de patrones, que demandan demasiado tiempo y esfuerzo aún en las computadoras más potentes de la actualidad (Sancho, 2017).

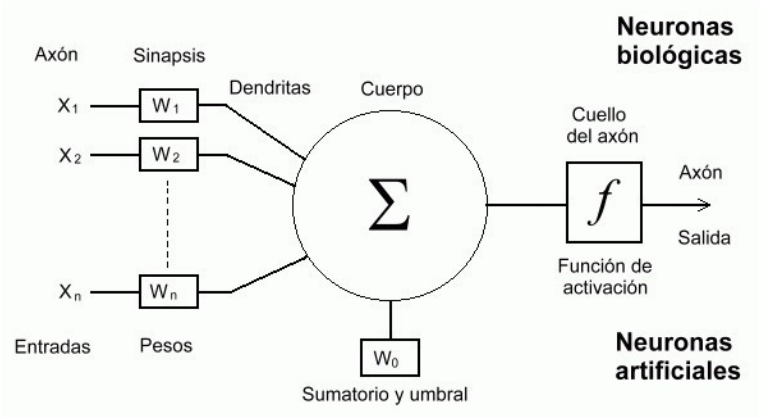
La función principal de las neuronas es la transmisión de impulsos nerviosos. Estos viajan por toda la neurona comenzando por las dendritas hasta llegar a las terminaciones del axón, donde pasan a otra neurona por medio de la conexión sináptica. La manera en que respondemos ante los estímulos del mundo exterior, y el aprendizaje que podemos realizar del mismo, está directamente relacionado con las conexiones neuronales del cerebro, y las RNAs son un intento de emular este hecho (Sancho, 2017).

### Modelo neuronal de McCulloch-Pitts

Sancho (2017) afirma que el primer modelo matemático de una neurona artificial, creado con el fin de llevar a cabo tareas simples, fue presentado en el año 1943 en un trabajo conjunto entre el psiquiatra y neuro anatomista Warren McCulloch y el matemático Walter Pitts. Como se detalla en la figura 7, un modelo neuronal con nn entradas consta de:

- Un conjunto de entradas $X_1, X_2...X_n$
- Los pesos sinápticos $W_1, W_2...W_n$ ,correspondientes a cada entrada.
- Una función de agregación, $\sum$
- Una función de activación, $f$
- Una salida, Y.

Figura 7: Modelo neuronal de McCulloch-Pitts




*Fuente. Adoptado de Modelo neuronal de McCulloch-Pitts, por F. Sancho, 2017*

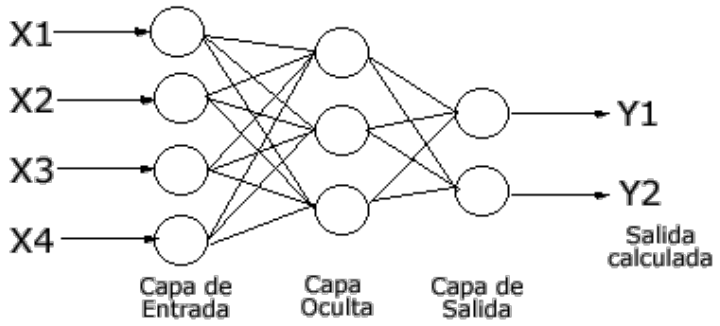
Las entradas son el estímulo que la neurona artificial recibe del entorno que la rodea, y la salida es la respuesta a tal estímulo. La neurona puede adaptarse al medio circundante y aprender de él modificando el valor de sus pesos sinápticos. Por ello son conocidos como los parámetros libres del modelo,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

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

$$Y = f(\sum_{i=1}^{n} W_i X_i) \tag{8}$$

En términos de Sancho (2017), un caso particular de perceptrón múltiple se puede visualizar en la figura 8, este se puede formar organizando sus neuronas en capas. Así, tenemos la capa de entrada formada por las entradas a la red, la capa de salida formada por las neuronas que constituyen la salida final de la red y las capas ocultas formadas por las neuronas que se encuentran entre los nodos de entrada y de salida. Una RNA puede tener varias capas ocultas o no tener ninguna de ellas. Las conexiones sinápticas (las flechas que llegan y salen de las neuronas) indican el flujo de la señal a través de la red, y tienen asociadas un peso sináptico correspondiente. Si la salida de una neurona va dirigida hacia dos o más neuronas de la siguiente capa, cada una de estas últimas recibe la salida neta de la neurona anterior. La cantidad de capas de una RNA es la suma de las capas ocultas más la capa de salida. En el caso de existir capas ocultas nos referimos a la RNA como un perceptrón multicapa.

Figura 8: Perceptrón multicapa




*Fuente. Adoptado de Modelo neuronal de McCulloch-Pitts, por F. Sancho, 2017*

El problema habitual con este tipo de redes multicapa es el de dado un conjunto de datos ya clasificados, de los que se conoce la salida deseada, proporcionar los pesos adecuados de la red para que se obtenga una aproximación correcta de las salidas si la red recibe únicamente los datos de entrada (Sancho, 2017).

### 2.3.4. Redes Bayesianas

El teorema de Bayes nos permite actualizar las probabilidades de variables cuyo estado no hemos observado dada una serie de nuevas observaciones. Las redes bayesianas automatizan este proceso, permitiendo que el razonamiento avance en cualquier dirección a través de la red de variables. Las redes bayesianas están constituidas por una estructura en forma de grafo, en la que cada nodo representa variables aleatorias (discretas o continuas) y cada arista representa las conexiones directas entre ellas. Estas conexiones suelen representar relaciones de causalidad. Adicionalmente, las redes bayesianas también modelan el peso cuantitativo de las conexiones entre las variables, permitiendo que las creencias probabilísticas sobre ellas se actualicen automáticamente a medida que se disponga de nueva información (López, 2015).

**Nodos y variables**

Lo primero que debemos hacer es identificar las variables de interés. Sus valores deben ser mutuamente excluyentes y exhaustivos. Los tipos de nodos discretos más comunes son:

- Nodos booleanos, que representan proposiciones tomando los valores binarios verdadero (V) y falso (F). En el dominio del diagnóstico médico, por ejemplo, un nodo llamado “cáncer” podría representar la proposición de que el paciente tenga cáncer.

- Valores ordenados. Por ejemplo, un nodo “contaminación” podría representar la exposición de un paciente a la contaminación del ambiente y tomar los valores (alta, baja).

- Valores enteros. Por ejemplo, un nodo llamado “edad” puede representar la edad de un paciente y tener valores posibles de 1 a 120.

Lo importante es elegir valores que representen el dominio de manera eficiente, pero con suficiente detalle para realizar el razonamiento requerido (López, 2015).

### Estructura de la red

La estructura o topología de la red debe captar las relaciones cualitativas entre las variables. En particular, dos nodos deben conectarse directamente si uno afecta o causa al otro, con la arista indicando la dirección del efecto (López, 2015). Por ende, en nuestro ejemplo de diagnóstico médico, podríamos preguntarnos qué factores afectan la probabilidad de tener cáncer. Si la respuesta es "contaminación y fumar", entonces deberíamos agregar aristas desde "contaminación","fumador"hacia el nodo "cáncer". Del mismo modo, tener cáncer afectará la respiración del paciente y las posibilidades de tener un resultado positivo de rayos X. Por lo tanto, también podemos agregar aristas de "cáncer", "disnea".

Es deseable construir redes bayesianas lo más compactas posibles por tres razones. Primero, mientras más compacto es el modelo, más fácil será de manejar. Segundo, cuando las redes se vuelven demasiado densas, fallan en representar la independencia en forma explícita. Y tercero, las redes excesivamente densas no suelen representar las dependencias causales del dominio.

### Probabilidades condicionales

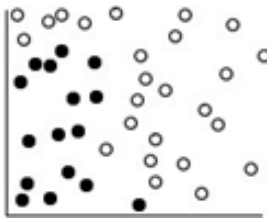
Una vez que tenemos definida la estructura de la red bayesiana, el siguiente paso es cuantificar las relaciones entre los nodos interconectados. Esto se hace especificando una probabilidad condicional para cada nodo (López, 2015). Primero, para cada nodo necesitamos mirar todas las posibles combinaciones de valores de los nodos padres. Por ejemplo, continuando con el ejemplo del diagnóstico del cáncer, si tomamos el nodo "cáncerçon sus dos nodos padres "contaminación", "fumador", podemos calcular los posibles valores conjuntos (A, V), (A, F), (B, V), (B, F). La tabla de probabilidad condicional específica para cada uno de estos casos podría ser la siguiente: 0,05, 0,02, 0,03, 0,001.

## 2.3.5. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés, soport vector machines) son una técnica de clasificación y regresión que aprovecha al máximo la precisión de las predicciones de un modelo sin ajustar excesivamente los datos de entrenamiento. Es ideal para analizar datos con un gran número de campos de predictores y tiene aplicaciones en multitud de disciplinas, incluyendo la gestión de relaciones con los clientes, el reconocimiento facial y otras imágenes, bioinformática, extracción de conceptos de minería de texto, detección de intrusiones, predicción de estructura de proteínas y reconocimiento de la voz (International Business Machines [IBM], 2019).

Básicamente, funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro (IBM, 2019). Por ejemplo, imagine la figura 9, en la que los puntos de datos corresponden a dos categorías diferentes.

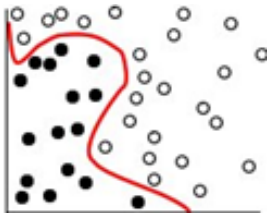
Figura 9: Puntos de datos con dos categorías diferentes



*Fuente. Adoptado de Conjunto de datos original, por IBM, 2019*

Las dos categorías se pueden separar con una curva, como se muestra en la figura 10.

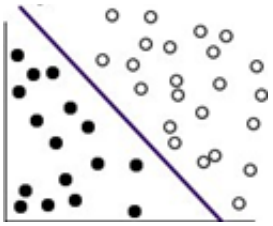
Figura 10: Puntos de datos con dos categorías separados por una curva



*Fuente. Adoptado de Datos con un separador añadido, por IBM, 2019*

Tras la transformación, el límite entre las dos categorías se puede definir por un hiperplano, como se muestra en la siguiente figura.

Figura 11: Puntos de datos con dos categorías definidas por un hiperplano



*Fuente. Adoptado de Datos con un separador añadido, por IBM, 2019*

La función matemática utilizada para la transformación se conoce como función kernel con los siguientes tipos: Lineal, Polinómico, Función de base radial (RBF) y Sigmoide. Una función kernel lineal es recomendable si la separación lineal de los datos es sencilla. En otros casos, se debe utilizar una del resto de las funciones (IBM,2019).

## 2.4. Caso de éxito en la implementación de Machine Learning en el sector financiero

Según Brito & Artes (2018), los bancos en Brasil cada vez más requieren del aprendizaje automático a través de Machine Learning para desarrollar modelos de calificación crediticia. El uso de árboles de regresión aditiva bayesiana, por ejemplo, se ha convertido en una herramienta importante en el análisis crediticio debido a la necesidad de estandarización y agilidad en la toma de decisiones. Es más, ahora hay situaciones en las que la aprobación o rechazo crediticio está totalmente automatizado. Identificar con mayor precisión a los clientes con alta probabilidad de incumplimiento permite reducir costos y aumentar los ingresos.

Según Husejinovic et al. (2018), los estudios de predicción de pagos predeterminados de tarjetas de crédito son muy importantes para cualquier institución financiera que se ocupe de las tarjetas de crédito. El propósito del trabajo que señala el autor con nombre Aplicación de algoritmos de aprendizaje automático en la predicción de pagos predeterminados de tarjetas de crédito, fue evaluar el rendimiento de los métodos de aprendizaje automático en la predicción de pagos predeterminados de tarjetas de crédito mediante regresión logística, árbol de decisión C4.5, máquinas de vectores de soporte (SVM), Naive Bayes, algoritmo k-vecinos más cercanos (k-NN) y métodos de aprendizaje como voting, bagging y boosting. El rendimiento de los algoritmos se evalúo a través de las siguientes métricas de rendimiento:

precisión, sensibilidad y especificidad. El mejor resultado entre todos los algoritmos para la tasa de precisión general se logró mediante el modelo de regresión logística con una tasa de 0,820. El modelo con mejor desempeño para la detección de clientes de tarjetas de crédito en mora, con un éxito del 71,3 %, fue el modelo Naive Bayes.

Según Srinath Gururaja (2022), el aprendizaje automático se está convirtiendo rápidamente en una de las soluciones centrales para varios problemas del mundo real. Gracias al potente hardware y los grandes conjuntos de datos, entrenar un modelo de aprendizaje automático se ha vuelto más fácil y gratificante. Sin embargo, un problema inherente en varios modelos de aprendizaje automático es la falta de comprensión de lo que sucede "bajo el capó". La falta de explicabilidad e interpretabilidad conduce a niveles más bajos de confianza en las predicciones del modelo, lo que significa que no se puede utilizar en aplicaciones sensibles como el diagnóstico de dolencias médicas y la detección de terrorismo. Esto ha llevado a varios avances para hacer que el aprendizaje automático sea explicable. En el estudio planteado por el autor se utilizan varios modelos de caja negra para clasificar a los morosos de tarjetas de crédito. Estos modelos se comparan utilizando diferentes métricas de rendimiento, y las explicaciones de estos modelos se proporcionan utilizando un explicador agnóstico del modelo.

Las conclusiones dadas por el autor Srinath Gururaja (2022) hacen referencia a la investigación de varios modelos de aprendizaje automático para clasificar a los morosos de tarjetas de crédito. Los resultados experimentales parecen indicar que el modelo XGBoost con DALEX basado en árboles proporcionan los mejores resultados, con el modelo Random Forest y SVM no muy lejos. Sin embargo, donde XGBoost se distingue es en su toma de decisiones sorprendentemente similar a la humana. Este modelo utiliza una cadena lógica comprensible para tomar sus decisiones, lo que puede ayudar a mejorar los procesos de pensamiento humano. Por lo tanto, el autor propone como algoritmo de potenciación del árbol XGBoost junto con DALEX como la solución para identificar a los morosos de tarjetas de crédito.

Según Oñate (2016), este trabajo presenta el estudio de minería de datos en la educación para modelar la perdida de la condición académica para estudiantes matriculados en los programas de Ingeniería Electrónica e Ingeniería de Sistemas de la Universidad Popular del Cesar. Se utilizaron dos tareas de minería de datos. En primer lugar, una tarea descriptiva basada en el algoritmo K-medias, que fue utilizado para seleccionar varios grupos de estudiantes. En segundo lugar, una tarea de clasificación soportada en dos técnicas conocidas como árbol de decisión y Naïve Bayes para predecir la pérdida de la condición académica debido a los malos resultados durante los cuatro primeros semestres de un estudiante. Para el entrenamiento y prueba de los modelos, se utilizaron los expedientes académicos y los datos recogidos durante el proceso de admisión de los estudiantes y se evaluaron utilizando la técnica de validación cruzada. Los resultados experimentales han demostrado que la predicción de la pérdida de la condición académica se mejora cuando se añaden los datos de la matrícula académica anterior.

## 2.5. Desbalanceo de clases

Según Raschka & Mirjalili(2019) el desbalanceo de clases es una problema bastante frecuente cuando se trabaja con datos reales; cuando muestras de una o de múltiples clases están sobrerepresentadas en un conjunto de datos. Por intuición, podemos pensar en varios ámbitos en los que puede ocurrir, como el filtrado de spam, la detección de fraude o la detección de enfermedades.

Es importante evaluar correctamente el conjunto de datos con estas caracteristicas a partir de la medición con distintas métricas, tiene sentido que centrarse en otras métricas distintas a la certeza cuando se comparan diferentes modelos, como la exactitud, la exhaustividad o la curva ROC, la que mejor vaya para nuestra aplicación. Existen ejemplos como poder identificar a la mayoria de los pacientes con cáncer maligno para recomendarles una detección adicional, por lo que la métrica más diciente seria la exhaustividad. En el filtrado de spam, donde no nos interesa etiquetar ciertos correos electrónicos como spam si el sistema no está muy seguro, la métrica más adecuada sería la exactitud.

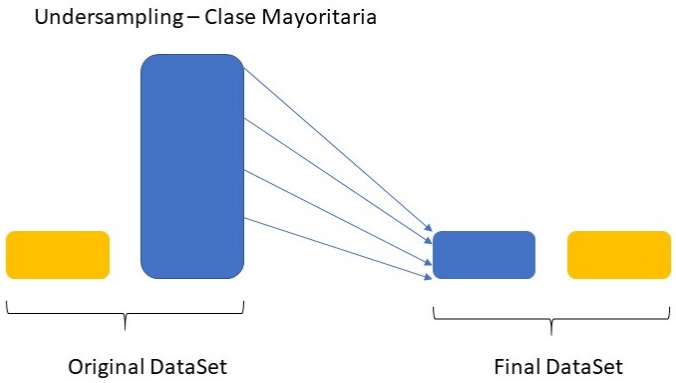
El autor Raschka & Mirjalili(2019) menciona que una manera de tratar las proporciones de clases desequilibradas durante el ajuste del modelo es asignar una penalización mayor a a las predicciones erróneas en

las clases minoritarias. Con scikit-learn, ajustar una penalización así es tan conveniente como configurar el parámetro class_weight a class_weight = 'balanced', implementado para la mayoria de los clasificadores.

Otras estrategias que menciona el autor para tratar el desequilibrio de clases incluyen el muestreo ascendente de la clase minoritaria, el muestreo descendente de la clase mayoritaria y la generación de muestras de entrenamiento sintéticas. Es importante revisar la literatura sobre las funciones resample, también el algoritmo más utilizado para la generación de muestras de entrenamiento sintéticas (SMOTE,por sus siglas en Inglés).

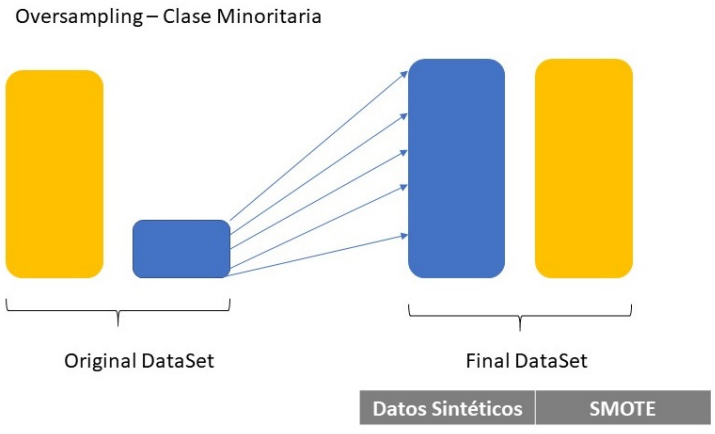
A continuación se ilustran los métodos Undersampling de clase mayoritaria figura 12 y Oversampling de clase minoritaria figura 13 que se utilizan para el balanceo de clases.

Figura 12: Undersampling – Clase Mayoritaria




*Fuente. Elaboración propia*

Figura 13: Oversampling – Clase Minoritaria




*Fuente. Elaboración propia*

## 2.6. Modelos de aprendizaje automático de conjunto

El autor Naviani (2018) menciona que existen los algoritmos de refuerzo que combinan múltiples modelos de baja precisión (o débiles) para crear modelos de alta precisión (o fuertes). Se puede utilizar en varios ámbientes, como crédito, seguros, marketing y ventas. Los algoritmos de impulso como AdaBoost, Gradient Boosting y XGBoost son algoritmos de aprendizaje automático ampliamente utilizados que trabajan bajo un conjunto compuesto, estos combinan una serie de clasificadores de bajo rendimiento con el objetivo de crear un clasificador mejorado. Aquí, se devolvió el voto del clasificador individual y la etiqueta de predicción final que realiza la votación mayoritaria. Los conjuntos ofrecen más precisión que el clasificador individual o base. Los métodos de conjunto pueden paralelizarse asignando a cada alumno base a diferentes máquinas diferentes. Finalmente, puede decir que los métodos de aprendizaje ensemble son meta algoritmos que combinan varios métodos de aprendizaje automático en un solo modelo predictivo para aumentar el rendimiento. Los métodos de conjunto pueden disminuir la varianza mediante el enfoque de embolsado, el sesgo mediante un enfoque de impulso o mejorar las predicciones mediante el enfoque de apilamiento.

Según Mendoza(2020) para conseguir un modelo más fuerte a partir de estos modelos débiles, se emplea un algoritmo de optimización, en este caso (XGBoost, por sus siglas en inglés). Durante el entrenamiento, los parámetros de cada modelo débil son ajustados iterativamente tratando de encontrar el mínimo de una función objetivo, que puede ser la proporción de error en la clasificación, el área bajo la curva (AUC), la raíz del error cuadrático medio (RMSE) o alguna otra. Cada modelo es comparado con el anterior. Si un nuevo modelo tiene mejores resultados, entonces se toma este como base para realizar modificaciones. Si, por el contrario, tiene peores resultados, se regresa al mejor modelo anterior y se modifica ese de una manera diferente. Qué tan grandes son los ajustes de un modelo a otro es uno de los hiperparámetros que debe definir el usuario. Este proceso se repite hasta llegar a un punto en el que la diferencia entre modelos consecutivos es insignificante, lo cual indica que se ha identificado el mejor modelo posible, o cuando se llega al número de iteraciones máximas definido por el usuario.

## 2.7. Medidas de desempeño del modelo

El desempeño del modelo se midió con la matriz de confusión que se ilustra en Tabla 1. Con esta matriz se puede calcular el desempeño del algoritmo en términos de exactitud, precisión, recall, error y curva ROC. Cada columna representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real (Alpaydin, 2004).

Tabla 1: Matriz de confusión

| Predicción | Positivo | Negativo |
|---|---|---|
| Positivo | Verdadero Positivo (VP) | Falso Negativo (FN) |
| Negativo | Falso Positivo (FP) | Verdadero Negativo (VN) |

Exactitud (Accuracy): es la proporción del número total de predicciones que son correctas; donde N es el total de ejemplos del conjunto de validaciones.

$$\text{Exactitud (Accuracy)} = \frac{(VP + VN)}{N} \tag{9}$$

Error de Clasificación: es la proporción del número total de predicciones que son incorrectas.

$$\text{Error de Clasificación} = \frac{(FP + FN)}{N} \tag{10}$$

Exhaustividad (Recall): es la proporción de casos positivos que fueron clasificados correctamente.

$$\text{Exhaustividad (Recall)} = \frac{VP}{(VP + FN)} \tag{11}$$

Precisión: es la predicción de casos positivos que fueron clasificados correctamente.

$$\text{Precisión} = \frac{VP}{(VP + FP)} \tag{12}$$

Medida – F (fmeasure): es una combinación de la precisión y el recall; Donde f, r y p son fmeasure, recall y precisión, respectivamente.

$$\text{Medida – F (fmeasure)} = \frac{2pr}{(p + r)} \tag{13}$$

Especificidad (Specificity): es la proporción de casos negativos que fueron clasificados correctamente.

$$\text{Especificidad (Specificity)} = \frac{VN}{(FP + VN)} \tag{14}$$

Sensibilidad (Sensitivity): es la proporción de casos positivos que fueron clasificados correctamente; este parámetro es el mismo que el Recall.

$$\text{Sensibilidad (Sensitivity)} = \frac{VP}{(VP + FN)} \tag{15}$$

Tasa de Falsos Negativos (FN): es la proporción de casos positivos que son clasificados incorrectamente como negativos.

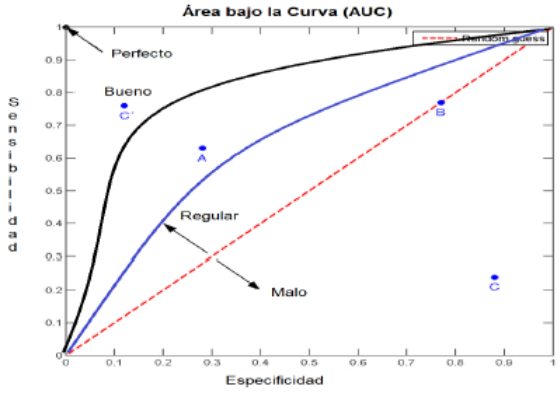
$$\text{Tasa de Falsos Negativos (FN)} = \frac{FN}{(VP + FN)} \tag{16}$$

Tasa de Falsos Positivos (FP): es la proporción de casos negativos que son clasificados incorrectamente como positivos.

$$\text{Tasa de Falsos Positivos (FP)} = \frac{FP}{(VN + FP)} \tag{17}$$

En la figura 14 se observa el área bajo la curva (AUC): es una representación gráfica de la sensibilidad de los verdaderos positivos (VP/VP + FN) vs. los falsos positivos (FP/FP + VN). El punto (0,1) se llama una clasificación perfecta. La línea diagonal que divide el espacio de la ROC en áreas de la clasificación buena o mala. Los puntos por encima de la línea diagonal indican buenos resultados de clasificación, mientras que los puntos por debajo de la línea indican resultados equivocados. En la figura 14 se presenta una representación gráfica de una curva ROC.

Figura 14: Curva ROC y representación gráfica área bajo la curva (AUC)




*Fuente. Adoptado de ROC curve, Alpaydin, 2004, Copyright 2004 por The MIT press Cambridge*

Tabla 2: Discriminación de la curva ROC

| | |
|---|---|
| Si ROC <0.5 | Sugiere No Discriminación |
| Si 0.5 <= ROC <0.7 | Discriminación Regular |
| Si 0.7 <= ROC <0.8 | Discriminación Aceptable |
| Si 0.8 <= ROC <0.9 | Discriminación Buena |
| Si ROC >= 0.9 | Discriminación Excelente |

Error Cuadrático Medio: mide la cantidad de error que hay entre dos conjuntos de datos. En otras palabras, compara un valor predicho y un valor observado o conocido.

$R^2$(coeficiente de determinación): puede tomar valores entre 0 y 1. El valor de aumenta cuando se incluyen nuevas variables explicativas o independientes en el modelo. Se incrementa hasta cuando son poco significativas o tienen poca correlación con la variable dependiente.

Kappa de Cohen: una estadística que mide el acuerdo entre anotadores. Esta función calcula una puntuación que expresa el nivel de acuerdo entre dos anotadores en un problema de clasificación. Se define como:

$$k = \frac{(P_0 - P_e)}{(1 - P_e)} \tag{18}$$

Dónde $P_0$ es la probabilidad empírica de acuerdo en la etiqueta asignada a cualquier muestra (la relación de acuerdo observada), y $P_e$ es el acuerdo esperado cuando ambos anotadores asignan etiquetas al azar. $P_e$ se estima utilizando un previo empírico por anotador sobre las etiquetas de clase (Cohen, 1960).

# 3. Metodología

A continuación se describen brevemente los procedimientos que se utilizaron para resolver el problema planteado, explicando cómo se realizó la investigación y las técnicas a utilizar, entre otros.

Según Parra (2017), la técnica utilizada se basa en la clasificación supervisada, que es una de las tareas que más frecuentemente son llevadas a cabo por los denominados sistemas de inteligencia artificial. Por lo tanto, un gran número de paradigmas bien desarrollados por la Estadística (Regresión Logística, Análisis Discriminante) o bien por la Inteligencia Artificial (Redes Neuronales, Inducción de reglas, Árboles de Decisión, Redes Bayesianas) son capaces de realizar las tareas propias de la clasificación.

Paso previo a aplicar un método de clasificación, se requiere primero analizar y generar las variables independientes, las cuales se distinguen en dos tipos: cuantitativas y cualitativas, y una variable dependiente (objetivo) que segmentará las empresas no deudoras y deudoras, y estas últimas en cuatro clases. Tanto el análisis descriptivo como el análisis de significancia permitirá determinar las variables que le suministrarán información al modelo.

Para Aruna y Nirmala (2013), el modelo debe aprender de los datos antes de aplicar un método de clasificación. En virtud de ello, se particionará el conjunto de datos: 80 % entrenamiento y 20 % test. El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se emplea para comprobar el comportamiento del modelo estimado. Cada registro de la base de datos debe de aparecer en uno de los dos subconjuntos. Lo ideal es entrenar el modelo con un conjunto de datos independientemente de los datos con los que realizamos el entrenamiento, para así evaluar a su vez la generalización del modelo.

## 3.1. Datos de estudio

Las bases de datos que se utilizaron corresponden a un fondo público del Régimen de Prima Media con prestación definida. La primera base está conformada por las empresas que cotizaron los últimos 24 meses de febrero de 2020 hacia atrás por sus trabajadores afiliados al fondo de pensiones. El mes se seleccionó teniendo en cuenta que no fuera junio o diciembre, donde se consignan bonificaciones, comisiones y salarios extralegales. Una vez las empresas realizan el pago de los aportes pensionales a través del operador de la planilla integrada de liquidación de aportes (PILA), estos se almacenan en el sistema de información llamado bodega de datos. Posteriormente pasan a un proceso de imputación en la historia laboral de los afiliados; allí, la primera base de datos contiene información referente al recaudo, las variables consultadas fueron: ingreso base de cotización, recaudo, número de trabajadores con recaudo, ingreso base de cotización promedio, número de periodos a pagar, número de periodos pagados, índice de pago. La segunda base contiene información referente a las deudas, las variables consultadas fueron: valor de deuda, número de periodos con deuda y la tercera base contiene los atributos de la empresa como segmento, tamaño, sector empleador, ciudad, departamento, regional. La información se encuentra almacenada en una tabla de datos sociodemográfica.

## 3.2. Tipo de investigación

La presente investigación es de tipo descriptiva porque está orientada a la realidad tal como se presenta. Esto es, a describir algunas características fundamentales de un fenómeno, hecho o situación determinada.

Para este estudio se propone un modelo de aprendizaje con Machine Learning utilizando técnicas como Regresión Logística, Árboles de Decisión, Redes Neuronales, Máquinas de Vectores de Soporte y Redes Bayesianas, que serán aplicadas a la clasificación de las empresas deudoras y no deudoras. Sin embargo, es importante realizar un análisis descriptivo con el objeto de evaluar el comportamiento de los datos y una prueba de significancia estadística para medir la bondad de ajuste del modelo. Esta prueba determinará si existe una relación entre la variable respuesta y las variables regresoras, como un acercamiento al modelo adecuado.

## 3.3. Método de investigación

En este trabajo se emplea la metodología Knowledge Discovery in Databases (KDD), que se basa en la búsqueda de información relevante dentro de una base de datos. Sobre este gran volumen de información se aplicarán técnicas de Machine Learning con el fin de detectar las empresas morosas y clasificarlas para enfocar estrategias que permitan la recuperación de la cartera de manera eficiente. Siguiendo la metodología planteada se realizaron las siguientes etapas anteriores al procesamiento:

1- Recolectar los datos: la fuente de almacenamiento de la información se encuentra en la bodega de datos, la cual se extrajo sobre tres consultas: recaudo, deudas e información sociodemográfica.

2- Preprocesar los datos: se transformaron los nombres de las variables categóricas, a través de una codificación, como requisito para la entrada de datos al modelo.

3- Explorar los datos: se realizó un análisis de correlación de variables cuantitativas, un análisis descriptivo de variables cualitativas a través de gráficos de barras, una prueba de significancia que permitió medir la bondad de ajuste del modelo y de esta forma construirlo para una mejor predicción.

4- Entrenar el algoritmo: en esta etapa alimentamos las técnicas de Machine Learning seleccionadas con los datos que se vienen procesando en las etapas anteriores. Se particionó el conjunto de datos 80 % entrenamiento y 20 % test. El subconjunto de datos de entrenamiento se utilizó para estimar los parámetros del modelo y el subconjunto de datos de test se empleó para comprobar el comportamiento del modelo estimado. Lo ideal es que el algoritmo logre buscar patrones importantes en los datos con la información que le suministramos en las predicciones.

5- Evaluar el algoritmo: en esta etapa se puso a prueba la información o conocimiento que se obtuvo del algoritmo del paso anterior. Se evaluó cuán preciso fue el algoritmo en sus predicciones, y como no se estuvo muy conforme, se ajustó el modelo mediante la estandarización del conjunto de datos para conseguir un mejor rendimiento. Las medidas utilizadas para evaluar la calidad del modelo fueron: certeza, error, exhaustividad/sensibilidad, precisión, medida – F, tasa de falsos negativos/positivos, error cuadrático medio, R, Kappa y el gráfico "curva ROC", parámetro random State (seed) y validación cruzada.También se incluyó el balanceo de clases para generar un mejor resultado.

6- Aplicar el modelo: en esta etapa utilizamos los resultados del modelo "clasificación de la cartera" para generar la segmentación y crear estrategias que permitan realizar las acciones de cobro de una manera más eficiente.

## 3.4. Descripción de las variables

Una vez consolidada la base de datos con las consultas a las tablas de recaudo, deudas y datos sociodemográficos que se encuentran en el sistema bodega de datos, se generaron 235.475 registros, que corresponden a empresas que cotizaron el mes de febrero de 2020. Esta base tiene un tamaño de 19.8 megabytes, óptimo para el procesamiento de la información con la aplicación Python. Respecto a la variable índice de pago se generó con los períodos pagados y con las deudas de las empresas en una ventana de 24 meses, tomando como referencia los períodos de cotización de febrero de 2020 hacia atrás. Es importante tener en cuenta que los períodos pagados más las deudas no dan 24 meses en todos los casos, debido a la existencia de novedades de retiro aplicadas que pueden hacer que una empresa presente menos periodos cotizados, aunque estos casos pueden ser mínimos.

En la Tabla 3 se observan las variables que fueron incluidas en el modelo como variables independientes o predictoras: Ingreso base de cotización (IBC), valor pagado, valor presunta, IBC Promedio, número de trabajadores, índice de pago, segmento, tamaño y sector empleador, las cuales se describen de la siguiente forma:

Tabla 3: Variables Independientes del Modelo

| Variables Independientes o Predictoras | | |
|---|---|---|
| Variable | Tipo de variable | Descripción |
| Número aportante | Cuantitativa Continua | Identificación del aportante |
| Ingreso Base de Cotización | Cuantitativa Continua | Suma del Ingreso Base de Cotización |
| Valor Pagado | Cuantitativa Continua | Suma de cotizaciones a pensión pagadas |
| Deuda Presunta | Cuantitativa Continua | Suma de cotizaciones a pensión con deuda |
| IBC Promedio | Cuantitativa Continua | Media entre el IBC y el No. de trabajadores |
| No Trabajadores | Cuantitativa Discreta | Suma de los trabajadores que cotizan a pensión |
| Índice de pago | Cuantitativa Continua | Periodos a pagar / Periodos pagados |
| No. de Periodos a Pagar | Cuantitativa Discreta | Suma de periodos a pagar por la empresa |
| No. de Periodos Pagados | Cuantitativa Discreta | Suma de periodos pagados por la empresa |
| No. de Periodos con Deuda | Cuantitativa Discreta | Suma de periodos de cotización con deuda |
| Segmento | Categórica Ordinal | 1. VIP: >= 43.012.348 en IBC |
| | | 2. Especial: entre 8.778.031 y 43.012.347 en IBC |
| | | 3. Clásico: entre 3.511.213 y 8.778.030 en IBC |
| | | 4. Pionero: entre 877.804 y 3.511.212 en IBC |
| | | 5. No objetivo: entre 0 y 877.803 en IBC |
| Tamaño | Categórica Ordinal | 1. Corporativo: >= 50 afiliados |
| | | 2. Mediana: entre 11 y 49 afiliados |
| | | 3. Pyme: entre 5 y 10 afiliados |
| | | 4. Micro: entre 2 y 4 afiliados |
| | | 5. No objetivo: 1 afiliado |
| Sector empleador | Categórica Nominal | 1. Sector privado |
| | | 2. Sector público |
| Código departamento aportante | Categórica Nominal | Codificación DANE |
| Código ciudad aportante | Categórica Nominal | Codificación DANE |
| Regional | Categórica Nominal | 1. Capital |
| | | 2. Centro |
| | | 3. Antioquia |
| | | 4. Santander |
| | | 5. Occidente |
| | | 6. Eje Cafetero |
| | | 7. Caribe |
| | | 8. Sur |

La variable objetivo o dependiente permitirá clasificar la cartera en el fondo de pensiones y dependerá de los patrones que pueda encontrar el sistema en los datos en conjunto con las variables predictoras. Para que responda adecuadamente, esta variable se encuentra categorizada de la siguiente forma:

1. Clase 0: grupo de empresas que no presentan deudas con una ventana de tiempo de 24 meses.

2. Clase 1: grupo de empresas que presentan deuda y tienen una excelente categorización en sus variables.

3. Clase 2: grupo de empresas que presentan deuda y tienen una buena categorización en sus variables.

4. Clase 3: grupo de empresas que presentan deuda y tienen una aceptable categorización en sus variables.

5. Clase 4: grupo de empresas que presentan deuda y tienen una baja categorización en sus variables.

La categorización se generó con base en el análisis de las diferentes variables que ingresaron al modelo y su relevancia para la administradora, teniendo en cuenta la importancia que tiene la recuperación de los dineros adeudados para financiar el Sistema de la Seguridad Social. En ese sentido se agruparon por categorías formando así la variable independiente Mora.

# 4. Desarrollo y aplicación del modelo

Normalmente, la gestión de cobro de las empresas que presentan mora se basa en las solicitudes o requerimientos que radican los afiliados al fondo de pensiones, al no contar con las cotizaciones que le permitan acceder a una pensión para su vejez o por solicitud de las empresas que requieren ponerse al día en sus aportes, ya que pueden estar reportadas en el boletín de deudores morosos y sin este paz y salvo no podrían contratar con el estado, por solicitud de entes regulatorios o de control que exigen gestión en la recuperación de los dineros adeudados por las empresas.

Para determinar las acciones de cobro a seguir se debe acudir al análisis de las bases de datos de los casos reportados y al conocimiento que tienen los gestores sobre la normatividad vigente, dejando de lado empresas que por su criticidad requieren de acciones inmediatas de cobro como, por ejemplo, afiliados próximos a pensionarse, empresas en proceso concursal o empresas reincidentes con el no pago de las obligaciones pensionales. El proceso requiere poder descubrir patrones de comportamiento, tendencias y relaciones en los datos de forma automática sin intervención humana y gracias a ese aprendizaje poder orientar y decidir las políticas de recuperación y las estrategias de cobro de la administradora de pensiones.

## 4.1. Modelo predictivo

### 4.1.1. Técnicas de clasificación

En esta sección se presentan cinco técnicas de clasificación para predecir las empresas morosas con base en las variables independientes o predictoras: Regresión Logística, Árboles de Decisión, Redes Neuronales, Máquinas de Vectores de Soporte y Bayes gaussianos.

Después de la selección de las técnicas y la construcción del modelo se procedió a describir cómo se divide el conjunto de datos disponible. El primer conjunto corresponde a los datos de entrenamiento que se parametriza para que el sistema escoja por muestreo aleatorio simple el 80 % de los datos, el modelo aprende de estos datos ajustando sus parámetros y la validación se realiza sobre el 20 % de los datos de test, los cuales evalúan el rendimiento del modelo. La razón de dividir el modelo en dos partes es para evitar pérdida de información. Si el conjunto de datos de entrenamiento presenta poco volumen de datos, el sistema no podrá generalizar y el modelo no será efectivo en su tarea de predecir, por ende, el conjunto de datos de validación no generará los resultados esperados. Es importante que no se incluyan observaciones del conjunto de entrenamiento en el conjunto de validación, si esto ocurre será difícil evaluar si el algoritmo ha aprendido a generalizar a partir del conjunto de entrenamiento o

simplemente lo ha memorizado.

### 4.1.2. Diseño del modelo

En el diseño del modelo se utilizó el software Python, que es un sistema para el aprendizaje automático, cuenta con una gran variedad de librerías de primer nivel que extienden la funcionalidad del lenguaje. Para el estudio la principal librería que utilizamos fue Scikit-Learn, la cual incorpora una gran variedad de algoritmos de aprendizaje. Esta librería permite y facilita la tarea de clasificación, evaluación, diagnóstico y validación cruzada. El operador para realizar la partición de la base de datos en entrenamiento y pruebas fue Split, los operadores y parámetros utilizados en los modelos predictivos se ilustran en la siguiente tabla:

Tabla 4: Parámetros del Modelo

| Técnica | Algoritmo | Parametros Utilizados | Descripción |
|---|---|---|---|
| Regresión Logística | LogisticRegression | C=10.0 | Inverso de la fuerza de regularización; debe ser un flotador positivo. Al igual que en las máquinas de vectores de soporte, los valores más pequeños especifican una regularización más fuerte. |
| Árbol de decisión | tree. DecisionTreeClassifier | Entropy | Se selecciono la función Entropy para la ganancia de información. |
| Redes Neuronales | MLPClassifier | Random_State = 42 max_iter = 100 | Siempre que la aleatorización sea parte de un algoritmo de aprendizaje de Scikit Learn, random_state, puede proporcionar un parámetro para controlar el generador de números aleatorios utilizado. Para los estimadores que involucran optimización iterativa, esto determina el número máximo de iteraciones que se realizarán. |
| Máquinas de vectores de soporte | LinearSVC | C=30.0 | Parámetro de regularización. La fuerza de la regularización es inversamente proporcional a C. Debe ser estrictamente positiva. |
| Redes Bayesianas | GaussianNB | Predeterminado default=1e-9 | Porción de la varianza más grande de todas las características que se agrega a las varianzas para la estabilidad del cálculo. |

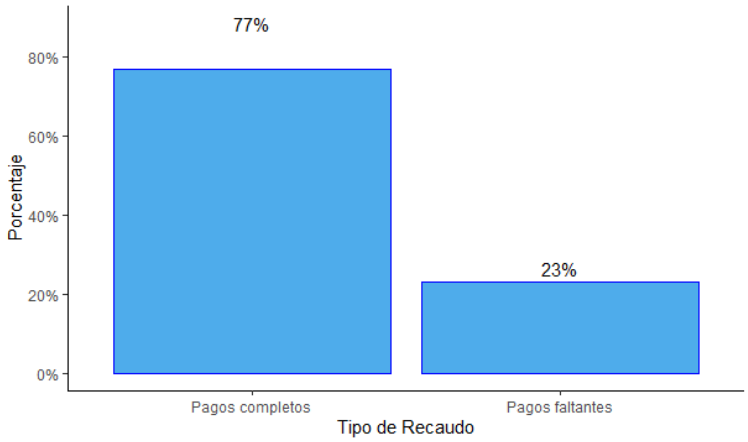
### 4.1.3. Análisis de las empresas con y sin pagos

El índice de pago permite identificar las empresas que presentan mora o las empresas que han realizado los pagos correctamente. El cálculo se realiza con una ventana de tiempo de 24 meses, teniendo como fecha de referencia el periodo de cotización a pagar: febrero de 2020 hacia atrás. Mes seleccionado porque no presenta pagos o variaciones adicionales sobre el recaudo que realizan las empresas, como bonificaciones, primas y demás prestaciones que hacen parte del factor salarial en ciertas épocas del año, y que pueden distorsionar de alguna forma el resultado.

$$\text{Índice de Pago} = \frac{\sum Número\,de\,periodos\,pagados}{\sum Número\,de\,periodos\,a\,pagar} \tag{19}$$

Este índice permite identificar si las empresas realizaron el pago de todas las cotizaciones con una ventana de tiempo de 24 meses, el cual presenta como resultado 1, si pago todos los periodos de cotización o en caso de no contar con todos los pagos, menor a 1. La figura 15 permite visualizar que, de las 235.475 empresas, 181.112 presentan pagos completos lo que corresponde al 77 %, mientras que 54.363 presentan mora, lo cual representa el 23 % de empresas morosas.

Figura 15: Gráfico de pagos de aportes pensionales

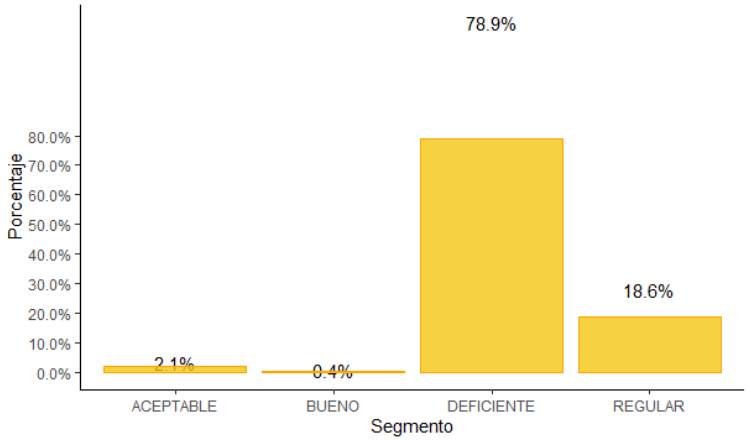



### Clasificación empresas con mora

Las empresas que presentan mora se clasifican en cuatro grupos como se observa en la figura 16. En el primer grupo están las empresas que presentan una buena categorización en las variables índice de pago, ingreso base de cotización, ingreso base de cotización promedio, valor pagado, número de periodos a pagar, deuda presunta, sector empleador, segmento, tamaño y número de trabajadores, correspondiente al 0,4 % con 205 empresas. En el segundo grupo están las empresas que presentan una categorización aceptable de las diez variables anteriormente mencionadas, correspondiente al 2,1 % con 1.157 empresas. En el tercer grupo se encuentran las empresas que presentan una categorización regular en las variables previamente señaladas, correspondiente al 18,6 % con 10.129. En el cuarto grupo están las empresas que presentan una categorización deficiente en sus variables, correspondiente al 78,9 % con 42.872 empresas.

Esta clasificación permite establecer las estrategias de cobro por cada grupo de acuerdo con los canales dispuestos por el fondo de pensiones. Sin embargo, debido al volumen de empresas morosas es difícil ejercer una acción de cobro de forma personalizada, por lo que se plantea la clasificación para optimizar recursos y ser más eficientes en las acciones de cobro.

Figura 16: Gráfico de empresas que presentan mora



### Análisis descriptivo de las variables de estudio

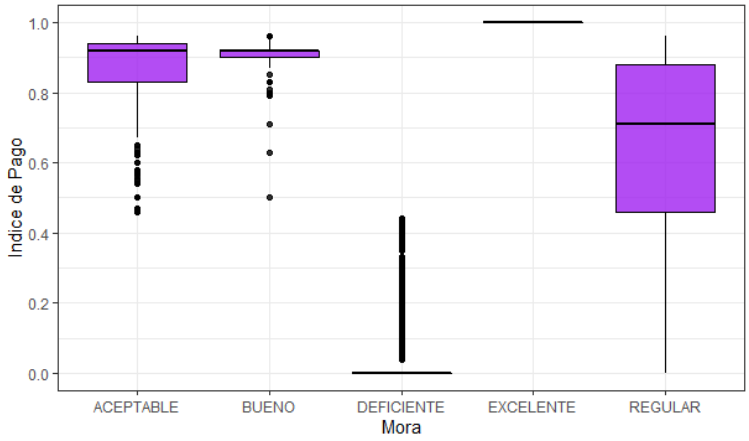
El análisis y entendimiento de las variables que participarán en el modelo permitirán clasificar los empleadores morosos con un nivel de certeza mayor, porque son significativas y le aportan información al modelo o bien porque al modificar sus características se pueden obtener mejores resultados. Como aspectos relevantes en las variables cuantitativas se puede observar en la Tabla 5 que la base de datos contiene 235.475 empresas, la media del ingreso base de cotización corresponde a $20.295.514 y la desviación estándar a $303.746.188, el valor pagado promedio por empresa corresponde a $3.308.326. En promedio, las empresas tienen 16 trabajadores, una empresa puede tener un máximo de trabajadores de 28.906 y la media del índice de pago fue del 80 %.

Tabla 5: Estadísticos descriptivos de las variables cuantitativas

| Variable | Media | Desviación Estándar | Mínimo | Máximo | No. |
|---|---|---|---|---|---|
| Ingreso Base de Cotización | 20.295.514 | 303.746.188 | 29.260 | 48.538.641.776 | 235.475 |
| Valor Pagado | 3.308.326 | 60.570.561 | 0 | 14.930.177.933 | 235.475 |
| Deuda Presunta | 67.201 | 505.652 | 0 | 169.028.778 | 235.475 |
| IBC Promedio | 1.021.544 | 3.364.631 | 443 | 713.803.556 | 235.469 |
| Número de Trabajadores | 16 | 86 | 1 | 28.906 | 235.475 |
| Índice de Pago | 0,8 | 0,39 | 0 | 1 | 235.475 |
| No. Periodos Por Pagar | 20 | 7 | 1 | 24 | 235.475 |
| No. Periodos Pagados | 16 | 10 | 0 | 24 | 235.475 |
| No. Periodos con Deuda | 4 | 9 | 0 | 24 | 235.475 |

La figura 17 ilustra un diagrama de caja y bigotes que representa gráficamente una serie de datos numéricos a través de sus cuartiles. En el gráfico se observa que la clasificación EXCELENTE presenta el mismo valor igual a 1 en la mediana y el cuartil, quiere decir esto que las empresas que se encuentran en este grupo presentan pagos completos por sus trabajadores. En la clasificación catalogada como BUENO presenta algunos valores atípicos entre los que se encuentra el valor mínimo que corresponde a un índice de pago 0.5, que puede corresponder a novedades que no ha sido reportadas por la empresa, mientras que el índice de Pago en la mayoría de estas empresas se encuentra con una mediana de 0.91. En la clasificación ACEPTABLE presenta una mayor dispersión con el valor mínimo en el índice de pago de 0.46, el valor máximo 0.96, el primer cuartil se sitúa en 0.82 y la mediana se encuentra en 0.90, quiere decir que estas empresas, aunque tienen valores atípicos y una mayor dispersión, en su mayoría presentan entre aceptable y buen Índice de Pagos. En la clasificación REGULAR presenta una mayor variabilidad y dispersión de los datos comparado con las demás clasificaciones, con un valor máximo en el índice de pagos de 0.96 y un valor mínimo de 0, la mediana se sitúa en el 0.71, quiere decir esto que las empresas no han realizado el pago de aportes pensionales en su totalidad por sus trabajadores, es importante precisar que la gran parte de los datos se sitúa entre el primer cuartil 0.48 y el tercer cuartil 0.89. En la clasificación DEFICIENTE la mayor parte de los datos se sitúan con un índice de pago de 0, mientras que los demás datos se presentan como valores atípicos entre 0.04 y 0.44, en esta clasificación las empresas presentan un índice de pagos muy bajo, las empresas en su mayoría tienen un nivel alto de morosidad, no realizan el pago cumplido de los aportes pensionales a seguridad social por sus trabajadores.

Figura 17: Gráfico de empresas que presentan o no mora por Índice de Pago

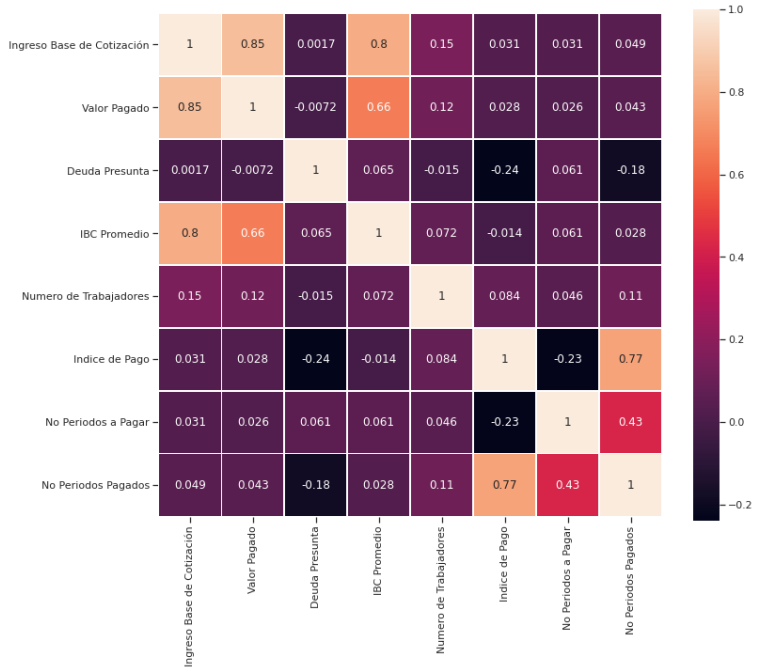



### Análisis de correlación variables cuantitativas

El algoritmo de aprendizaje automático básico busca patrones a partir de la correlación entre las variables de los datos. Dos variables tienen correlación cuando el cambio en una de ellas causa un cambio en la otra, del mismo signo o del signo opuesto. Por ejemplo, la altura y el peso de una persona están correlacionados: a más altura se tiende a pesar más. La correlación se expresa en un coeficiente de correlación que varía entre -1 y +1. La correlación negativa es cuando un incremento en una variable causa un decremento en la otra. La correlación cero es que no hay ninguna relación entre las dos variables y la correlación cercana a 1 existe una relación muy cercana.

Es importante tener cuidado porque algunos modelos (LM, GLM) se ven perjudicados si incorporan predictores altamente correlacionados, por lo que no podrían predecir la variable objetivo del modelo tipo deudor. En la figura 18 se puede observar que las variables que presentan una correlación buena son el ingreso base de cotización con el valor pagado (0,85). Esto quiere decir que si aumenta el ingreso base de cotización producto del aumento del IBC por cambio de año o por el incremento del salario de los empleados podría incrementar el recaudo. El ingreso base de cotización presenta una correlación "buena" con el ingreso base de cotización promedio (0,80); si aumenta el ingreso de la empresa también aumentaría el ingreso base de cotización promedio por trabajador. Si el valor pagado y el ingreso base de cotización promedio (0,66) presentan una aceptable correlación posiblemente pueda deberse a que el ingreso base de cotización de un trabajador no necesariamente redundará en un recaudo en la misma proporción, ya que puede cotizar en determinado mes por un número menor de días.

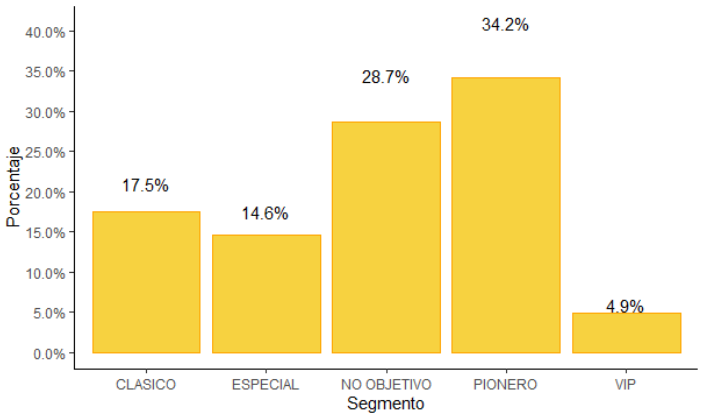
Figura 18: Correlación de variables cuantitativas




### Análisis descriptivo variables cualitativas

En la figura 19 se observa que el 34,2 % de las empresas están representadas por el segmento pionero, el 28,7 % de las empresas están representadas por el segmento no objetivo, el 17,5 % de las empresas están en el segmento clásico, el 14,6 % de las empresas están en el segmento especial y el 4,9 % de las empresas están en el segmento VIP. Las empresas que tienen menos ingresos base de cotización relacionados en el pago de sus afiliados son las más representativas en la segmentación, mientras que las empresas con más ingresos base de cotización únicamente son 11.583. Respecto a estas empresas es importante hacer campañas de fidelización y de mantenimiento como medidas de retención para que sigan afiliando y aportando al fondo de pensiones público y de educación respecto a las obligaciones pensionales.

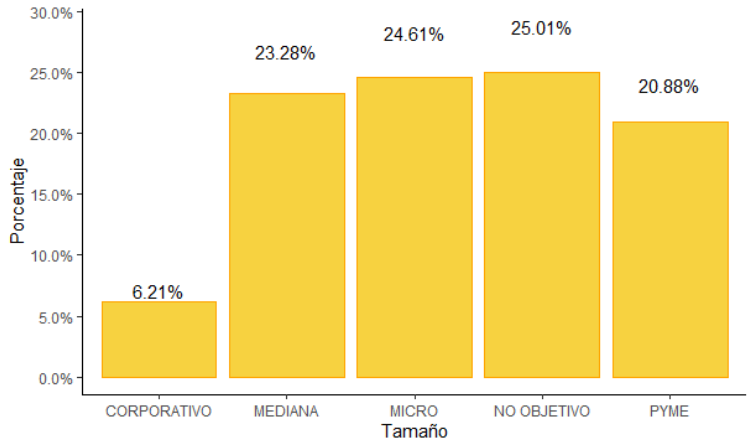
Figura 19: Gráfico de empresas por segmento




El número de empresas en los tamaños mediana, micro y no objetivo son muy similares, las empresas

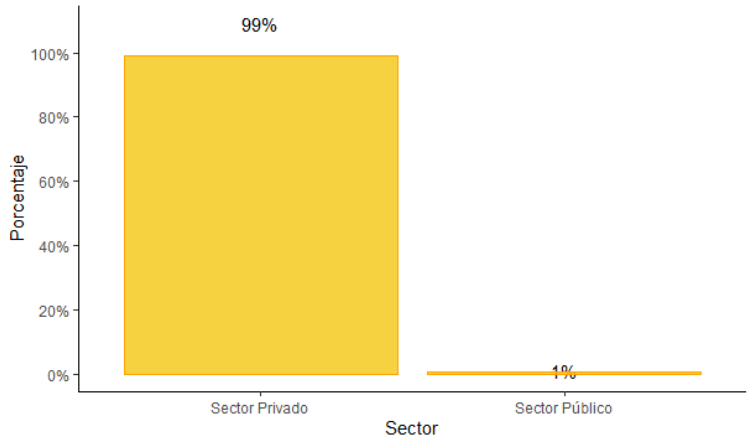
no objetivo son las más representativas con el 25,01 %, son empresas pequeñas que cotizan por un solo trabajador. Mientras que las empresas corporativas representan el 6,21 % y son empresas que cotizan por más de 50 afiliados. Tal como se ilustra en la figura 20.

Figura 20: Gráfico de empresas por tamaño



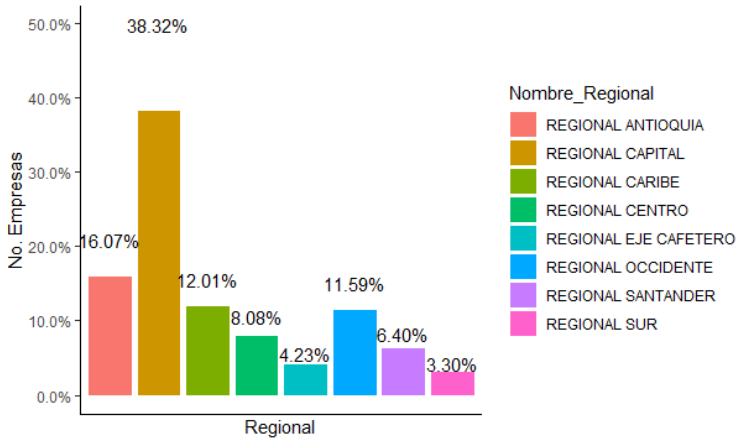

En la figura 21 se observa que las empresas del sector público representan el 1 %, cuyo cobro de aportes requiere de un tratamiento especial, mesas de trabajo que permitan revisar las planillas de autoliquidación y si se evidencian novedades pendientes por tramitar, realizar la labor operativa que garantice la normalización de la cartera o por el respectivo pago de la seguridad social. Las empresas privadas, por su parte, la mayoría (99 %), requieren ser clasificadas para lograr focalizar los recursos de cobro de cartera.

Figura 21: Gráfico sector de la empresa




Las empresas aportantes o deudoras realizan sus aportes a pensiones en diferentes ciudades que se encuentran agrupadas por regionales. Los datos de la figura 22 muestran que el 38,32 % de las empresas hace sus pagos en la regional Capital representada por Bogotá; el 16,07 % de las empresas realiza pagos a la seguridad social en la regional Antioquia, seguidos por la regional Caribe y la regional Occidente con el 12,01 %. La proporción de empresas más bajo está representada por la regional Sur con el 3,30 %. Conocer está información permite realizar una planeación de recursos más eficiente para el cobro de la cartera, pues el costo de la correspondencia para una ciudad como Bogotá es diferente al de otras ciudades, específicamente por los procesos de centralización de esta.

Figura 22: Gráfico de empresas por regional




### Análisis de significancia variables de empresas con mora

El método stepwise se utilizó para encontrar un conjunto de variables independientes que influyan significativamente en la variable dependiente. El sistema automáticamente fue incluyendo las variables significativas a medida que iba iterando a través de una serie de pruebas t y eliminando aquellas que no son estadísticamente significativas. El paquete estadístico utilizado para este ejercicio fue SAS MINER, el cual se parametrizo con 50 iteraciones.

En la Tabla 6 se observa la columna Pr $\geq$ ChiSq que mide el nivel de significancia, el parámetro utilizado corresponde a los valores menores a 0.05 para la inclusión de las variables independientes que le aportan información al modelo.

Tabla 6: Significancia de variables empresas con mora

| Variable | DF | Wald Chi-Square | Pr >ChiSq |
|---|---|---|---|
| Ingreso Base de Cotización | 3 | 266,1517 | <.0001** |
| IBC Promedio | 3 | 45,4493 | <.0001** |
| Índice de Pago | 3 | 15,0584 | 0.0018** |
| Valor Pagado | 3 | 51,2683 | <.0001** |
| Deuda Presunta | 2 | 64913,9204 | <.0001** |
| Sector empleador | 2 | 65,838 | <.0001** |
| Segmento | 2 | 52171060,2 | <.0001** |
| Tamaño | 2 | 14,6561 | 0.0007** |
| No de Trabajadores | 3 | 140,1555 | <.0001** |

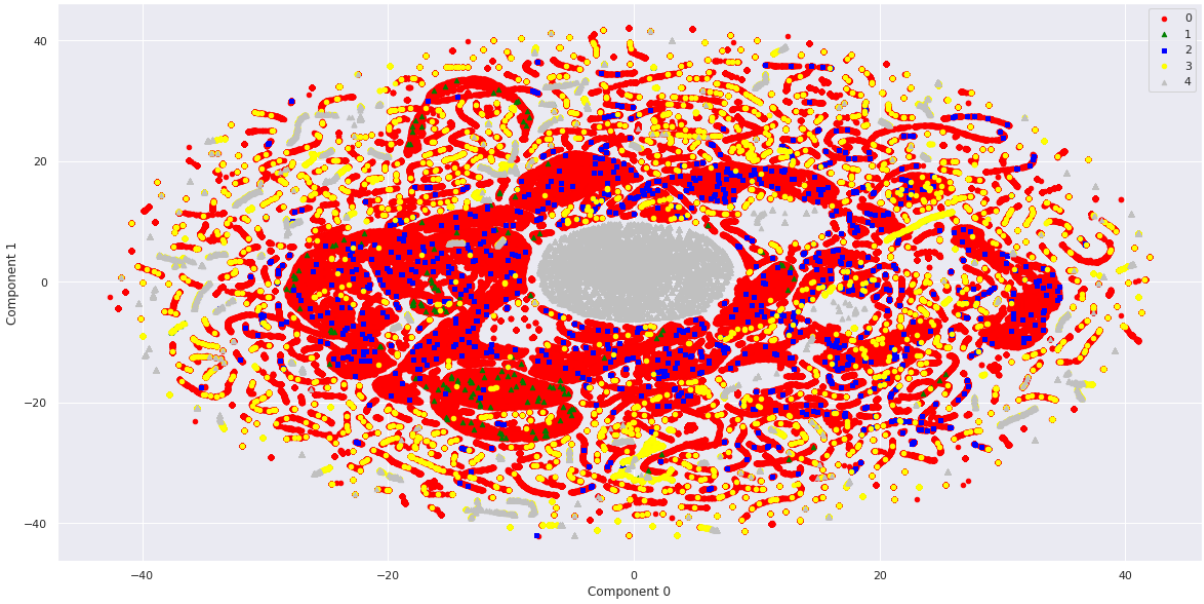
La deviance es una medida de ajuste del modelo a los datos, como la deviance del modelo 16.540 es menor que la deviance null 16.978, el modelo con las variables seleccionadas se ajusta mejor que el modelo con el intercepto.

### 4.1.4. Visualizaciones

#### Embebido de vecinos estocasticos distribuidos en t (t-SNE, por sus siglas en Inglés)

Como se observa en la figura 23, los datos para la variable objetivo mora 0 están muy cercanos. Esto quiere decir que son empresas que presentan un alto índice de pagos y, en efecto, no presentan deudas. Para la variable objetivo mora 1 y 2 están muy cercanos a los puntos de la variable objetivo 0, ya que son empresas con índice de pagos bueno o aceptable, con un tamaño corporativo, mediana o pyme y un segmento clásico, especial o VIP. Los datos clasificados como 3 reflejan el comportamiento regular de pagos de las empresas en su gran mayoría presentan un tamaño no objetivo, micro y pyme y un segmento pionero y no objetivo; por último, se encuentra la clasificación 4, la deficiente, representada de color gris, en donde las empresas presentan un índice de pagos bajo, se trata de empresas con un número de trabajadores reducido y un segmento no objetivo por el bajo recaudo que consignan mensualmente.

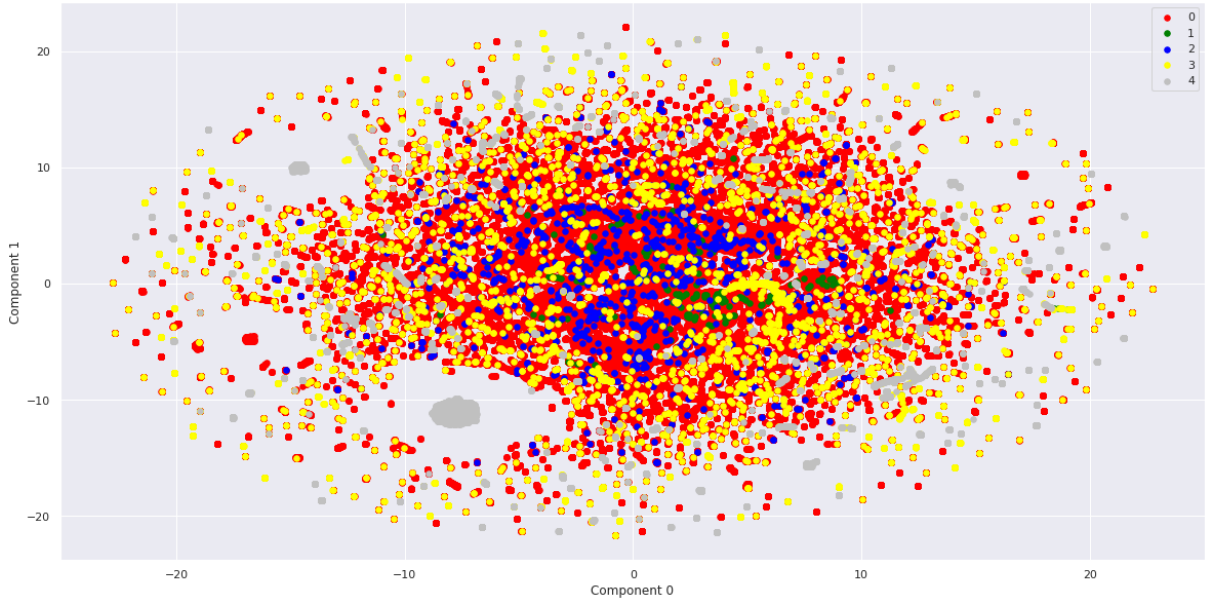
Figura 23: Embebido de vecinos estocasticos distribuidos en t (t-SNE)




#### Aproximación y proyección de una distribución uniforme (UMAP, por sus siglas en inglés)

En la figura 24, se pueden observar los grupos que discriminan las diferentes clasificaciones de la variable objetivo. La agrupación de puntos en el espectro define las empresas morosas y no morosas, el grupo con mayor variabilidad en los datos está representado por la categoría 3, ya que presentan un índice de pagos regular reflejados en algunos segmentos y tamaños. La clasificación de las empresas representada en la categoría 1 de color verde se puede observar en un pequeño grupo, correspondiente a empresas medianas o grandes, que han realizado pagos y posiblemente las deudas que presentan no reportaron las novedades de retiro. El mayor volumen de puntos se ve reflejado en las empresas que han realizado los pagos cumplidamente representados en color rojo con el número 0. El grupo de color azul, aquellas agrupadas con el número 2, son empresas con un aceptable comportamiento de pago, se encuentran representadas en su mayoría por empresas medianas y se encuentran en un segmento especial o clásico. En el grupo de color gris, representadas con el número 4, están las empresas con un deficiente comportamiento de pago en sus variables, además de ser empresas no objetivo para el fondo de pensiones tanto en segmento como en tamaño.

Figura 24: Aproximación y proyección de una distribución uniforme (UMAP)

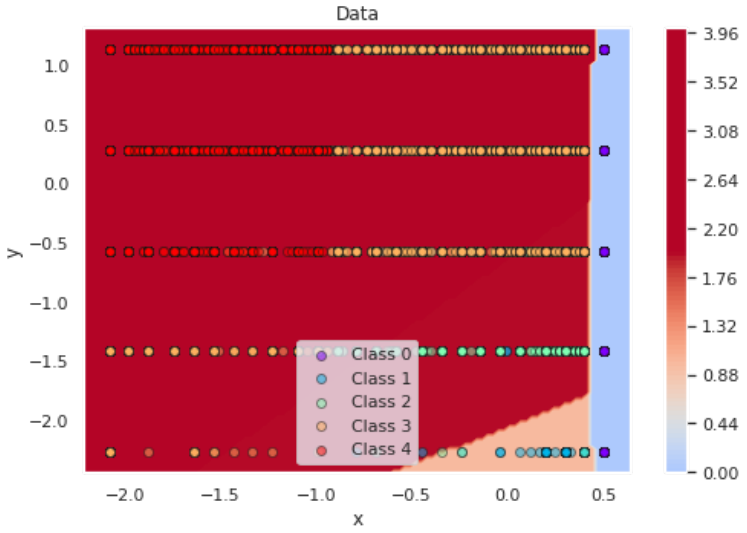



### 4.1.5. Representación gráfica de las técnicas de Machine Learning

Se representó gráficamente las técnicas de Machine Learning en un plano X, Y como variables independientes y la clase como variable dependiente.

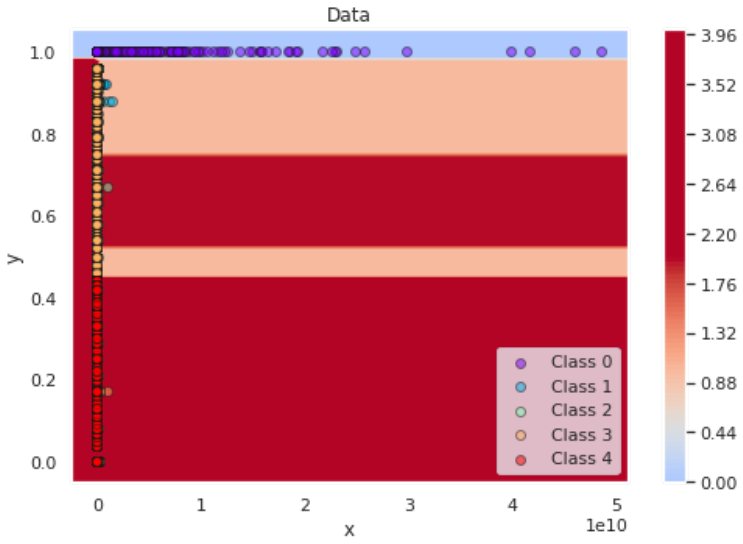
**Regresión Logística**: se representó gráficamente en la figura 25 un modelo de Regresión Logística con las variables índice de pago y segmento, el grado de discriminación con la base total de los datos fue 98,86 %.

Figura 25: Gráfica Regresión Logística (x, y)



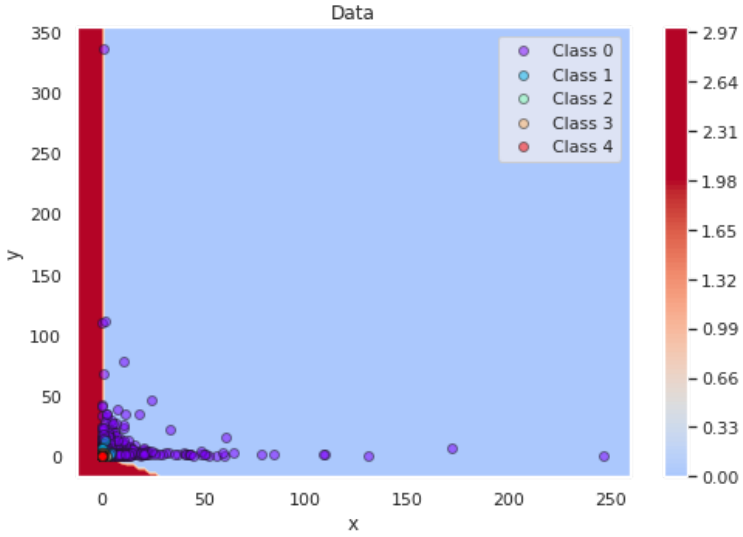
**Árboles de Decisión**: se representó gráficamente en la figura 26 un modelo de Árboles de Decisión con las variables ingreso base de cotización e índice de pago, el grado de discriminación con la base total de los datos fue 99,64 %.

Figura 26: Gráfica Árboles de Decisión (x, y)



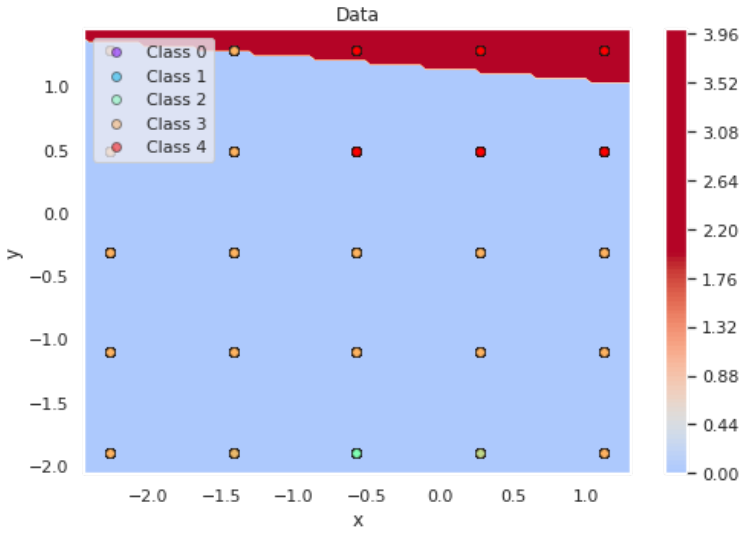

**Redes Neuronales**: se representó gráficamente en la figura 27 un modelo de Redes Neuronales con las variables valor pagado y número de trabajadores, el grado de discriminación con la base total de los datos fue 95 %.

Figura 27: Gráfica Redes Neuronales (x, y)



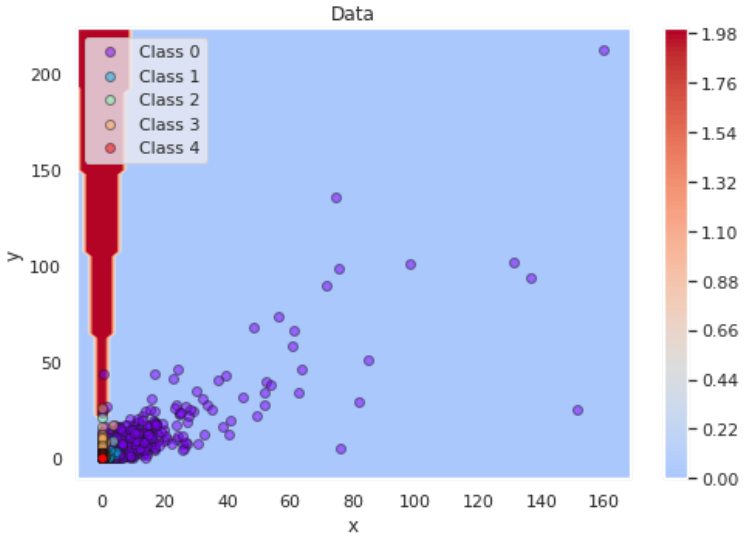

**Máquinas de Vectores de Soporte**: se representó gráficamente en la figura 28 un modelo de Máquinas de Vectores de Soporte con las variables segmento y tamaño, el grado de discriminación con la base total de los datos fue 85 %.

Figura 28: Gráfica Máquinas de Vectores de Soporte (x, y)




**Redes Bayesianas**: se representó gráficamente en la figura 29 un modelo de Redes Bayesianas con las variables ingreso base de cotización e ingreso base de cotización promedio, el grado de discriminación con la base total de los datos fue 28 %.

Figura 29: Gráfica Redes Bayesianas (x, y)




### 4.1.6. Resultados de desempeño de las empresas con mora

La estandarización de los conjuntos de datos es un requisito común para muchos estimadores de aprendizaje automático; podrían comportarse mal si las características individuales no se parecen a los datos estándar distribuidos normalmente: gaussiano con media cero y varianza unitaria. En la práctica, a menudo se ignora la forma de la distribución y simplemente se transforman los datos para centrarlos

eliminando el valor medio de cada entidad, luego se escala dividiendo entidades no constantes por su desviación estándar. Para realizar el escalado de los datos se utilizó el paquete preprocessing de sklearn con la función scale. La instancia del escalador se puede usar en nuevos datos para transformarla, acción seguida con lo hecho en el conjunto de entrenamiento.

Se realizó un análisis comparativo de los datos no escalados y escalados, tomando como medida de calidad el indicador de certeza. La mejora más drástica se ve reflejada en la técnica Redes Neuronales con un 6 %, seguida de Redes Bayesianas con un 4 % y Regresión Logistica con un 3 % de mejora.

A continuación se representa a través de la Tabla 7, un comparativo de certeza con escalado y sin escalado de los datos.

Tabla 7: Comparativo de escalado – certeza del modelo

| | Base sin Escalado | | Base con Escalado | |
|---|---|---|---|---|
| Método/Tipo | Entrenamiento (80 %) | Test (20 %) | Entrenamiento (80 %) | Test (20 %) |
| Regresión Logística | 96 % | 96 % | 99 % | 99 % |
| Arboles de Decisión | 99 % | 99 % | 99 % | 99 % |
| Redes Neuronales | 94 % | 93 % | 99 % | 99 % |
| SVM | 98 % | 98 % | 98 % | 98 % |
| Redes Bayesianas | 94 % | 94 % | 98 % | 98 % |

El modelo se trabajó con las variables escaladas, a excepción de Árboles de Decisión que no hubo necesidad de realizar el escalado por su buen desempeño.

En el preprocesamiento de las bases de datos aplicando los algoritmos de Machine Learning se fraccionó la base total en una base de entrenamiento con el 80 % de los datos y otra base para pruebas con el 20 % de los datos.

En la Tabla 8, se observa la certeza de las técnicas predictivas por cada variable objetivo y el número de casos que fueron predichos tanto correcta como incorrectamente, tomando como referencia la medida de predicción de exhaustividad.

En la clase tipo 0 con una base de entrenamiento de 144.748 empresas y de pruebas con 36.364 empresas, tuvo una certeza superior al 99 % en todas las técnicas.

En la clase tipo 1 con una base de entrenamiento de 166 empresas y de pruebas con 39 empresas, tuvo una certeza del 99,9 % en Árboles de Decisión, 94,9 % en Redes Neuronales, como los mejores resultados, mientras que la certeza de Regresión Logistica 71,8 % y Máquinas de Vectores de Soporte 69,2 % fue aceptable.

En la clase tipo 2 con una base de entrenamiento de 922 empresas y de test de 235 empresas, tuvo una certeza del 98,7 % en Árboles de Decisión, 94 % en Redes Neuronales como los mejores resultados, mientras que la certeza de Regresión Logistica 80,4 % y Máquinas de Vectores de Soporte 80 % fue buena en el conjunto de pruebas para validar la calidad del modelo. La técnica Redes Bayesianas no alcanzó a lograr una aceptable discriminación con un 24,7 %.

En la clase tipo 3 con una base de entrenamiento de 8.126 empresas y de pruebas con 2.003 empresas, tuvo una excelente certeza en todas la técnicas al lograr un porcentaje superior al 90 %, a excepción de Máquinas de Vectores de Soporte que logró una certeza de 77,4 % calificada como aceptable discriminación.

En la clase tipo 4 con una base de entrenamiento de 34.418 empresas y de pruebas con 8.454 empresas, tuvo una certeza superior al 99 % en todas las técnicas discriminando correctamente.

Tabla 8: Matriz de confusión certeza en la clasificación de la deuda

| | | | Entrenamiento (80 %) | | | | | Test (20 %) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Etiqueta predicha | | | | | | | | | |
| | | | 0 | 1 | 2 | 3 | 4 | 0 | 1 | 2 | 3 | 4 |
| Regresión Logística | Etiqueta Real | 0 | 144743 | 0 | 0 | 5 | 0 | 36360 | 0 | 4 | 0 | 0 |
| | | 1 | 0 | 108 | 58 | 0 | 0 | 0 | 28 | 11 | 0 | 0 |
| | | 2 | 24 | 13 | 756 | 129 | 0 | 7 | 5 | 189 | 34 | 0 |
| | | 3 | 2 | 0 | 64 | 7775 | 285 | 0 | 0 | 11 | 1906 | 86 |
| | | 4 | 0 | 0 | 0 | 285 | 34133 | 0 | 0 | 0 | 71 | 8383 |
| Árboles de Decisión | | 0 | 144744 | 0 | 4 | 0 | 0 | 36361 | 0 | 3 | 0 | 0 |
| | | 1 | 0 | 166 | 0 | 0 | 0 | 0 | 39 | 0 | 0 | 0 |
| | | 2 | 0 | 1 | 920 | 1 | 0 | 0 | 1 | 232 | 2 | 0 |
| | | 3 | 0 | 0 | 0 | 8126 | 0 | 0 | 0 | 0 | 2003 | 0 |
| | | 4 | 0 | 0 | 0 | 0 | 34418 | 0 | 0 | 0 | 1 | 8453 |
| Redes Neuronales | | 0 | 144742 | 0 | 0 | 6 | 0 | 36359 | 1 | 4 | 0 | 0 |
| | | 1 | 7 | 157 | 2 | 0 | 0 | 2 | 37 | 0 | 0 | 0 |
| | | 2 | 11 | 5 | 872 | 34 | 0 | 4 | 3 | 221 | 7 | 0 |
| | | 3 | 4 | 0 | 10 | 8110 | 2 | 5 | 0 | 1 | 1996 | 1 |
| | | 4 | 0 | 0 | 0 | 23 | 34395 | 0 | 0 | 0 | 9 | 8445 |
| SVM | | 0 | 144650 | 0 | 0 | 98 | 0 | 36345 | 0 | 0 | 19 | 0 |
| | | 1 | 0 | 106 | 13 | 47 | 0 | 0 | 27 | 7 | 5 | 0 |
| | | 2 | 18 | 56 | 719 | 129 | 0 | 1 | 12 | 188 | 34 | 0 |
| | | 3 | 57 | 10 | 823 | 6230 | 1006 | 8 | 2 | 199 | 1551 | 243 |
| | | 4 | 0 | 0 | 6 | 153 | 34259 | 0 | 0 | 3 | 39 | 8412 |
| Redes Bayesianas | | 0 | 143231 | 1113 | 149 | 255 | 0 | 36002 | 277 | 33 | 52 | 0 |
| | | 1 | 0 | 125 | 13 | 28 | 0 | 0 | 34 | 2 | 3 | 0 |
| | | 2 | 0 | 35 | 272 | 615 | 0 | 0 | 17 | 58 | 160 | 0 |
| | | 3 | 0 | 7 | 140 | 7815 | 164 | 0 | 1 | 44 | 1913 | 45 |
| | | 4 | 0 | 0 | 105 | 953 | 33360 | 0 | 0 | 24 | 246 | 8184 |

En la Tabla 9, al revisar las medidas de predicción del conjunto de datos de entrenamiento, se observa que la medida de predicción certeza que mide el número de predicciones correctas lograron buenos resultados en las técnicas Regresión Logística (99,5 %), Árboles de Decisión (99,8 %), Redes Neuronales (99,6 %), Máquinas de Vectores de Soporte (98,7 %) y Redes Bayesianas (98,1 %); el error más alto se produjo con la técnica Redes Bayesianas, aunque por las buenas predicciones el error es mínimo (1,9 %).

En la medida de exhaustividad/sensibilidad, que es la proporción de casos positivos que fueron clasificados correctamente, las técnicas en la clase tipo 0 en general tuvieron buenas predicciones. Las técnicas que tuvieron los mejores porcentajes en la clase tipo 1 fueron: Árboles de Decisión (99,8 %), Redes Neuronales (94,6 %) y los porcentajes más bajos se presentaron en las técnicas Regresión Logistica (65,1 %) y Máquinas de Vectores de Soporte (63,9 %) ; en la clase tipo 2 los máximos porcentajes fueron: Árboles de Decisión (99,8 %), Redes Neuronales (94,6 %), Regresión Logistica (82 %) y el menor porcentaje fue Redes Bayesianas (29,5 %); en la clase tipo 3 los mejores resultados se organizan de la siguiente forma: Árboles de Decisión (99,8 %), Redes Neuronales (99,8 %), Redes Bayesianas (98,2 %) y Regresión Logística (95,7 %) y el menor porcentaje fue Máquinas de Vectores de Soporte (76,7 %) ; y en la clase tipo 4: Árboles de Decisión (99,1 %), Redes Neuronales (99,9 %), Máquinas de Vectores de Soporte (99,5 %), Regresión Logística (99,2 %) y Redes Bayesianas (96,9 %) dieron excelentes resultados.

En la medida de desempeño precisión, que es la predicción correcta de casos, las técnicas en la clase tipo 0 tuvieron buenas predicciones. Las técnicas que tuvieron los mejores porcentajes en la clase tipo

1 fueron: Árboles de Decisión (99,7 %), Redes Neuronales (96,9 %), Regresión Logística (89,3 %) y el menor porcentaje fue Redes Bayesianas (9,8 %); en la clase tipo 2: Árboles de Decisión (99,5 %), Redes Neuronales (98,6 %) y Regresión Logística (86,1 %) y los menores porcentaje fueron Máquinas de Vectores de Soporte (46,1 %) y Redes Bayesianas (40,1 %); en la clase tipo 3: Árboles de Decisión (99,6 %), Redes Neuronales (99,3 %), Máquinas de vectores de soporte (93,6 %) y Regresión Logística (94,9 %); por último, en la clase tipo 4 todas las técnicas tuvieron buenas predicciones.

En la medida –F, que es la combinación de la Precisión y el Recall, o sea, la proporción de casos positivos que fueron clasificados correctamente, las técnicas en la clase tipo 0 tuvieron buenas predicciones. Las técnicas que obtuvieron los mejores porcentajes en la clase tipo 1 se organizan así: Árboles de Decisión (99,7 %) y Redes Neuronales (95,7 %) y el menor porcentaje fue Redes Bayesianas (17,3 %); en la clase tipo 2 las mejores técnicas fueron: Árboles de Decisión (99,8 %) y Redes Neuronales (96,6 %) y los menores porcentaje fueron Máquinas de Vectores de Soporte (57,9 %) y Redes Bayesianas (34 %); en la clase tipo 3 se destacan las técnicas: Árboles de Decisión (99,6 %), Redes Neuronales (99,6 %) y Regresión Logística (95,3 %); y en la variable clase tipo 4 todas las técnicas tuvieron buenas predicciones.

En la tasa de clasificación de falsos negativos/positivos, que es la proporción de casos clasificados incorrectamente más baja, la técnica Árboles de Decisión arrojó 0,4 %,0,2 %,0,2 %,0,2 % y 0,9 % en las clases tipo 0, 1, 2, 3 y 4; mientras que el porcentaje más alto estuvo en la clase tipo 2 Redes Bayesianas (70,5 %), en la clase tipo 1 Máquinas de Vectores de Soporte (36,1 %) y Regresión Logística (34,9 %).

El error cuadrático medio más alto fue Redes Bayesianas (0,035) y Máquinas de Vectores de Soporte (0,010) y el mejor R como estadístico que tiene la mejor correlación entre la variable independiente con las variables dependientes es el modelo Regresión Logistica 0,998, Árboles de Decisión 0,999 y Redes Neuronales con 0,999. El mejor indicador de Kappa de cohen, que mide la probabilidad empírica de acuerdo entre la etiqueta real y la etiqueta predicha, está dado por Árboles de Decisión (0,998), Redes Neuronales (0,998), Regresión Logística (0,987), Máquinas de Vectores de Soporte (0,965) y el indicador más bajo de Kappa de cohen fue Redes Bayesianas (0,950).

Tabla 9: Modelo de predicción datos de entrenamiento – medidas de calidad del modelo

| Medidas de Predicción | Etiqueta | Regresión Logística | Árboles de Decisión | Redes Neuronales | SVM | Redes Bayesianas |
|---|---|---|---|---|---|---|
| Certeza | 0,1,2,3,4 | 99,5 % | 99,8 % | 99,6 % | 98,7 % | 98,1 % |
| Error | 0,1,2,3,4 | 0,5 % | 0,2 % | 0,4 % | 1,3 % | 1,9 % |
| Exhaustividad/Sensibilidad | 0 | 99,8 % | 99,9 % | 99,9 % | 99,9 % | 99 % |
| | 1 | 65,1 % | 99,8 % | 94,6 % | 63,9 % | 75,3 % |
| | 2 | 82 % | 99,8 % | 94,6 % | 78 % | 29,5 % |
| | 3 | 95,7 % | 99,8 % | 99,8 % | 76,7 % | 98,2 % |
| | 4 | 99,2 % | 99,1 % | 99,9 % | 99,5 % | 96,9 % |
| Precisión | 0 | 99,5 % | 99,4 % | 99,8 % | 99,9 % | 99,8 % |
| | 1 | 89,3 % | 99,7 % | 96,9 % | 61,6 % | 9,8 % |
| | 2 | 86,1 % | 99,5 % | 98,6 % | 46,1 % | 40,1 % |
| | 3 | 94,9 % | 99,6 % | 99,3 % | 93,6 % | 80,9 % |
| | 4 | 99,2 % | 99,7 % | 99,8 % | 97,1 % | 99,5 % |
| Medida – F | 0 | 99,8 % | 99,7 % | 99,8 % | 99,9 % | 99,5 % |
| | 1 | 75,3 % | 99,7 % | 95,7 % | 62,7 % | 17,3 % |
| | 2 | 84 % | 99,8 % | 96,6 % | 57,9 % | 34 % |
| | 3 | 95,3 % | 99,6 % | 99,6 % | 84,3 % | 88,7 % |
| | 4 | 99,2 % | 99,5 % | 99,6 % | 98,3 % | 98,2 % |
| Falsos Negativos/Positivos | 0 | 0,2 % | 0,1 % | 0,1 % | 0,1 % | 1 % |
| | 1 | 34,9 % | 0,2 % | 5,4 % | 36,1 % | 24,7 % |
| | 2 | 18 % | 0,2 % | 5,4 % | 22 % | 70,5 % |
| | 3 | 4,3 % | 0,2 % | 0,2 % | 23,3 % | 1,8 % |
| | 4 | 0,8 % | 0,9 % | 0,1 % | 0,5 % | 3,1 % |
| Error cuadrático medio | 0,1,2,3,4 | 0,005 | 0,005 | 0,0008 | 0,010 | 0,035 |
| R | 0,1,2,3,4 | 0,998 | 0,999 | 0,999 | 0,995 | 0,986 |
| Kappa | 0,1,2,3,4 | 0,987 | 0,998 | 0,998 | 0,965 | 0,950 |

En la Tabla 10, al revisar las medidas de predicción del conjunto de datos de entrenamiento, se observa que la medida de predicción certeza que mide el número de predicciones correctas presenta una excelente discriminación en las técnicas Regresión Logística (99,5 %), Árboles de Decisión (99,8 %), Redes Neuronales (99,6 %), Máquinas de Vectores de Soporte (98,8 %) y Redes Bayesianas (98,0 %); el error más alto se produjo con la técnica Redes Bayesianas, aunque por las buenas predicciones el error es mínimo (2 %).

En la medida de exhaustividad/sensibilidad, que es la proporción de casos positivos que fueron clasificados correctamente, las técnicas en la clase tipo 0 en general tuvieron buenas predicciones. Las técnicas que tuvieron los mejores porcentajes en la clase tipo 1 fueron: Árboles de Decisión (99,2 %), Redes Neuronales (94,9 %) y los porcentajes más bajos se presentaron en las técnicas Regresión Logistica (71,8 %) y Máquinas de Vectores de Soporte (69,2 %); en la clase tipo 2 los máximos porcentajes fueron: Árboles de Decisión (98,7 %), Redes Neuronales (94,0 %), Regresión Logistica (80,4 %) y el menor porcentaje fue Redes Bayesianas (24,7 %); en la clase tipo 3 los mejores resultados se organizan de la siguiente forma: Árboles de Decisión (99,3 %), Redes Neuronales (99,7 %), Redes Bayesianas (97,7 %), Regresión Logística (95,2 %) y en menor porcentaje Máquinas de Vectores de Soporte (77,4 %); y en la clase tipo 4: Árboles de Decisión (99,4 %), Redes Neuronales (99,9 %), Máquinas de Vectores de Soporte (99,5 %), Regresión Logística (99,2 %) y Redes Bayesianas (96,8 %).

En la medida de desempeño precisión, que es la predicción correcta de casos, las técnicas en la clase tipo 0 tuvieron buenas predicciones. Las técnicas que tuvieron los mejores porcentajes en la clase tipo 1 fueron: Árboles de Decisión (97,5 %), Redes Neuronales (90,2 %), Regresión Logística (84,8 %) y en

menor porcentaje Redes Bayesianas (10,3 %); en la clase tipo 2 las mejores técnicas fueron: Árboles de Decisión (99,4 %), Redes Neuronales (99,5 %), Regresión Logística (89,6 %) y en menor medida Máquinas de Vectores de Soporte (47,4 %) y Redes Bayesianas (36 %); en la clase tipo 3 los porcentajes más altos se organizan de la siguiente manera: Árboles de Decisión (99,9 %), Redes Neuronales (99,2 %), Máquinas de vectores de soporte (94,1 %) y Regresión Logística (94,8 %); por último, en la clase tipo 4 todas las técnicas tuvieron buenas predicciones.

En la medida –F, que es la combinación de la Precisión y el Recall, o sea, la proporción de casos positivos que fueron clasificados correctamente, las técnicas en la clase tipo 0 tuvieron buenas predicciones. Las técnicas que obtuvieron los mejores porcentajes en la clase tipo 1 se organizan así: Árboles de Decisión (98,7 %), Redes Neuronales (92,5 %) y en menor porcentaje Redes Bayesianas (18,5 %); en la clase tipo 2 las mejores técnicas fueron: Árboles de Decisión (99,4 %) y Redes Neuronales (96,7 %) y en menor medida Máquinas de Vectores de Soporte (59,5 %) y Redes Bayesianas (29,3 %); en la clase tipo 3 se destacan las técnicas: Árboles de Decisión (99,9 %), Redes Neuronales (99,4 %) y Regresión Logística (95 %); y en la clase tipo 4 todas las técnicas tuvieron buenas predicciones.

En la tasa de clasificación de falsos negativos/positivos, que es la proporción de casos clasificados incorrectamente, la más baja fue la técnica Árboles de Decisión arrojó 0,1 %, 0,8 %, 1,3 %, 0,7 % y 0,6 % en las variables 0, 1, 2, 3 y 4; mientras que el porcentaje más alto estuvo en la clase tipo 2 Redes Bayesianas (75,3 %) y en la clase tipo 1 Máquinas de Vectores de Soporte (30,8 %) y en Regresión Logística (28,2 %).

El error cuadrático medio más alto fue Redes Bayesianas (0,031) y Máquinas de Vectores de Soporte (0,010) y el mejor R como estadístico que tiene la mejor correlación entre la variable independiente con las variables dependientes es el modelo Árboles de Decisión con 0,999. El mejor indicador de Kappa de cohen, que mide la probabilidad empírica de acuerdo entre la etiqueta real y la etiqueta predicha, está dado por Árboles de Decisión (0,999), Redes Neuronales (0,998), Regresión Logística (0,987), Máquinas de Vectores de Soporte (0,967) y el indicador más bajo de Kappa de cohen fue Redes Bayesianas (0,949).

Es importante incluir en la etapa de validación diferentes métricas a parte de la certeza cuando se comparan diferentes modelos como por ejemplo: la precisión, la exhaustividad, medida-F o la curva ROC, la que mejor vaya a nuestra aplicación. Esto debido a que las clases minoritarias presentan muy pocas muestras en comparación con las clases mayoritarias, presentando problemas en el proceso de generalización y no ser detectables en la utilización de una sola medida de calidad.

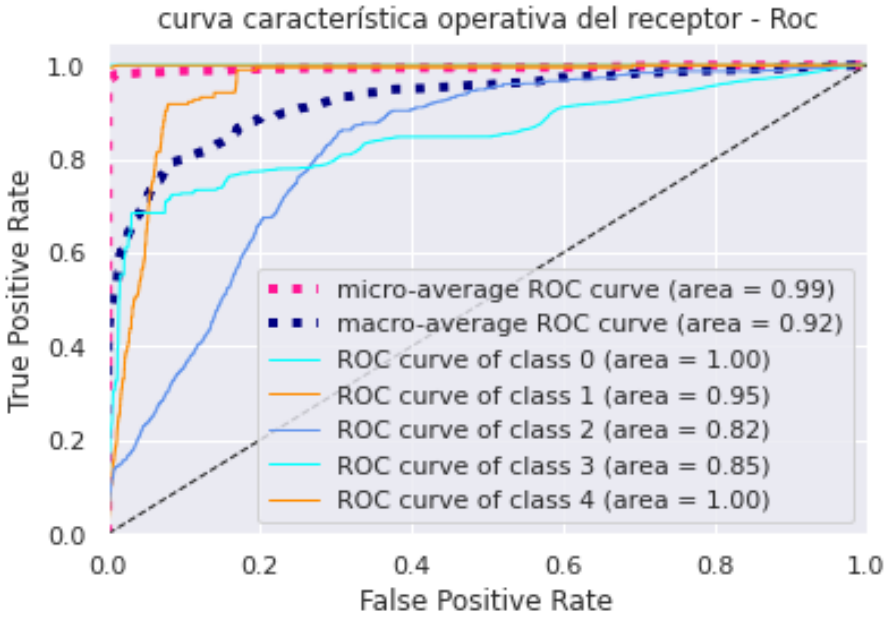
Tabla 10: Modelo de predicción datos de pruebas – medidas de calidad del modelo

| Medidas de Predicción | Etiqueta | Regresión Logística | Árboles de Decisión | Redes Neuronales | SVM | Redes Bayesianas |
|---|---|---|---|---|---|---|
| Certeza | 0,1,2,3,4 | 99,5 % | 99,8 % | 99,6 % | 98,8 % | 98,0 % |
| Error | 0,1,2,3,4 | 0,5 % | 0,2 % | 0,4 % | 1,2 % | 2 % |
| Exhaustividad/Sensibilidad | 0 | 99,8 % | 99,9 % | 99,8 % | 99,9 % | 99,2 % |
| | 1 | 71,8 % | 99,2 % | 94,9 % | 69,2 % | 87,2 % |
| | 2 | 80,4 % | 98,7 % | 94 % | 80 % | 24,7 % |
| | 3 | 95,2 % | 99,3 % | 99,7 % | 77,4 % | 97,7 % |
| | 4 | 99,2 % | 99,4 % | 99,9 % | 99,5 % | 96,8 % |
| Precisión | 0 | 99,4 % | 99,2 % | 99,3 % | 99,3 % | 99,2 % |
| | 1 | 84,8 % | 97,5 % | 90,2 % | 65,9 % | 10,3 % |
| | 2 | 89,6 % | 99,4 % | 99,5 % | 47,4 % | 36 % |
| | 3 | 94,8 % | 99,9 % | 99,2 % | 94,1 % | 80,6 % |
| | 4 | 94,8 % | 99,4 % | 99,2 % | 97,2 % | 99,5 % |
| Medida – F | 0 | 99,2 % | 99,2 % | 99,4 % | 99,4 % | 99,5 % |
| | 1 | 77,8 % | 98,7 % | 92,5 % | 67,5 % | 18,5 % |
| | 2 | 84,8 % | 99,4 % | 96,7 % | 59,5 % | 29,3 % |
| | 3 | 95 % | 99,9 % | 99,4 % | 85 % | 88,3 % |
| | 4 | 96,9 % | 99,4 % | 99,9 % | 98,3 % | 98,2 % |
| Falsos Negativos/Positivos | 0 | 0,2 % | 0,1 % | 0,2 % | 0,1 % | 0,8 % |
| | 1 | 28,2 % | 0,8 % | 5,1 % | 30,8 % | 12,8 % |
| | 2 | 19,6 % | 1,3 % | 6 % | 20 % | 75,3 % |
| | 3 | 4,8 % | 0,7 % | 0,3 % | 22,6 % | 2,3 % |
| | 4 | 0,8 % | 0,6 % | 0,1 % | 0,5 % | 3,2 % |
| Error cuadrático medio | 0,1,2,3,4 | 0,005 | 0,0008 | 0,0018 | 0,010 | 0,031 |
| R | 0,1,2,3,4 | 0,997 | 0,999 | 0,998 | 0,996 | 0,987 |
| Kappa | 0,1,2,3,4 | 0,987 | 0,999 | 0,998 | 0,967 | 0,949 |

La curva ROC se usa típicamente en la clasificación binaria o de varias clases para estudiar la salida de un clasificador y el poder discriminatorio de este. Para extender la curva y el área ROC a la clasificación de etiquetas múltiples, como es nuestro caso, es necesario binarizar la salida. Dibujamos la curva ROC por tipo de variable objetivo. En la figura 30 se ve expresada por 0,1,2,3,4. La clase tipo 0 dio un valor de 1.0, es decir, presenta una excelente discriminación; la clase tipo 1 que dio un valor de 0.95 también mostró una excelente discriminación; la clase tipo 2 dio una buena predicción con un valor de 0.82; la clase tipo 3 presentó una buena discriminación con un valor de 0.85; y la clase tipo 4 presentó, de igual manera, una excelente discriminación de 1.0. También se realizó considerando cada elemento de la matriz del indicador para cada grupo como predicción binaria (micro promedio), la cual dio como resultado 0.99, y como una clasificación de grupos múltiples (macro promedio), que otorga el mismo peso a la clasificación de cada etiqueta, la cual dio como resultado 0.92.

Las curvas ROC suelen presentar una tasa de verdaderos positivos en el eje Y y una tasa de falsos positivos en el eje X. Esto significa que la esquina superior izquierda de la gráfica es el punto ïdeal": una tasa de falsos positivos de cero y una tasa de verdaderos positivos de uno. Significa que un área más grande bajo la curva (AUC) suele ser mejor. Esto quiere decir que para nuestro estudio el mejor clasificador fue la clase tipo 0 y 4.

Figura 30: Curva ROC por cada grupo de deudores




### 4.1.7. Validación del modelo con diferentes semillas

La validación del modelo con diferentes semillas permitió garantizar que las divisiones generadas sean reproducibles: las filas se asignan a la base de entrenamiento de forma aleatoria, lo cual asegura que los conjuntos de datos sean una muestra representativa del conjunto de datos original. Al asignar una semilla o random state (número aleatorio) cada algoritmo fue entrenado y evaluado de la misma manera en las mismas submuestras de los datos.

La presente validación tiene como objetivo modificar la semilla con diez datos diferentes para evaluar la incertidumbre inherente dentro del modelo. Los resultados de la Tabla 11 validación del modelo de entrenamiento y Tabla 12 validación del modelo de pruebas, muestran que en el método de Regresión Logística, tanto para entrenamiento como test, variaron los porcentajes de certeza con diferentes semillas al oscilar entre 99,5 % y 99,6 % con escalado y entre 95,3 % y 95,8 % sin escalado. En el método Árboles de Decisión los porcentajes de certeza de entrenamiento con diferentes semillas fueron entre 99,1 % y 99,7 % con y sin escalado y de test entre 99,1 % y 99,7 % con y sin escalado. En el método Redes Neuronales en entrenamiento hubo una variación con la técnica de escalado al generar un porcentaje que se encuentra entre 96,9 % y 98,9 % y en test oscila entre 99,1 % y 99,8 % y hubo una variación en los porcentajes de certeza con diferentes semillas al oscilar entre 92,60 % y 95,9 % sin escalado, sin embargo hubo una disminución significativa en la semilla 15 y 100 con un porcentaje de 76,9 % y 80 % tanto en entrenamiento como en test sin escalado. En el método Máquinas de Vectores de Soporte tanto en entrenamiento como en test hubo una variación leve que osciló entre 98,5 % y 98,8 % con la técnica de escalado, sin embargo hubo una variación en los porcentajes de certeza con diferentes semillas al oscilar entre 91,40 % y 98,8 % sin escalado, es importante tener en cuenta que hubo una disminución significativa en la semilla 35, 50 y 100 con un porcentaje de 76,9 %, 79,7 % y 76,9 % tanto en entrenamiento como en test sin escalado. En el método Bayes Gaussianos tanto en entrenamiento como en test hubo una variación leve que osciló entre 97,8 % y 98,1 % con la técnica de escalado, sin embargo hubo una variaron en los porcentajes de certeza con diferentes semillas al oscilar entre 94,10 % y 94,2 % sin escalado, es importante tener encuenta que hubo una disminución significativa en la semilla 20, 35, 50 y 100 con un porcentaje de 44,1 %, 44,3 %, 84 % y 44,1 % tanto en entrenamiento como en test sin escalado.

A modo de resumen, se puede apreciar que al aplicar las iteraciones con diez semillas, los diferentes métodos no tuvieron variaciones significativas con la técnica escalado, mientras sin la técnica de escalado algunos métodos como Máquinas de Vectores de Soporte y Redes Bayesianas disminuyeron el rendimiento para algunas semillas.

Tabla 11: Validación del modelo de entrenamiento con diferentes semillas

| Modelo/Semilla | Regresión Logistica | | Arboles de Decisión | | Redes Neuronales | | SVM | | Redes Bayesianas | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sin escalado | Con escalado | Sin escalado | Con escalado | Sin escalado | Con escalado | Sin escalado | Con escalado | Sin escalado | Con escalado |
| 10 | 95,3 % | 99,5 % | 99,1 % | 99,6 % | 95,70 % | 98,9 % | 94,90 % | 98,8 % | 94,1 % | 97,9 % |
| 15 | 95,8 % | 99,5 % | 99,4 % | 99,5 % | 76,90 % | 99,0 % | 91,40 % | 98,7 % | 94,2 % | 98,0 % |
| 20 | 95,4 % | 99,5 % | 99,4 % | 99,5 % | 92,60 % | 97,9 % | 94,40 % | 98,6 % | 44,1 % | 97,9 % |
| 30 | 95,8 % | 99,6 % | 99,2 % | 99,3 % | 93,80 % | 96,9 % | 98,80 % | 98,7 % | 94,1 % | 98,1 % |
| 35 | 95,8 % | 99,6 % | 99,4 % | 99,5 % | 95,00 % | 96,9 % | 76,90 % | 98,5 % | 44,3 % | 98,0 % |
| 40 | 95,8 % | 99,5 % | 99,5 % | 99,6 % | 95,00 % | 97,9 % | 92,40 % | 98,6 % | 94,2 % | 97,9 % |
| 42 | 95,5 % | 99,6 % | 99,6 % | 99,7 % | 95,40 % | 97,9 % | 92,40 % | 98,7 % | 94,1 % | 97,8 % |
| 50 | 95,7 % | 99,6 % | 99,4 % | 99,5 % | 95,40 % | 96,9 % | 79,70 % | 98,7 % | 84,0 % | 97,9 % |
| 60 | 95,7 % | 99,5 % | 99,3 % | 99,4 % | 95,60 % | 97,9 % | 94,90 % | 98,8 % | 94,1 % | 98,0 % |
| 100 | 95,7 % | 99,5 % | 99,5 % | 99,6 % | 80,00 % | 96,9 % | 76,90 % | 98,8 % | 44,1 % | 97,9 % |

Tabla 12: Validación del modelo de pruebas con diferentes semillas

| Modelo/Semilla | Regresión Logistica | | Árboles de Decisión | | Redes Neuronales | | SVM | | Redes Bayesianas | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sin escalado | Con escalado | Sin escalado | Con escalado | Sin escalado | Con escalado | Sin escalado | Con escalado | Sin escalado | Con escalado |
| 10 | 95,5 % | 99,6 % | 99,1 % | 99,6 % | 95,90 % | 99,8 % | 95,1 % | 98,7 % | 94,2 % | 98,0 % |
| 15 | 95,7 % | 99,6 % | 99,3 % | 99,4 % | 76,70 % | 99,8 % | 91,3 % | 98,6 % | 94,3 % | 97,9 % |
| 20 | 95,2 % | 99,5 % | 99,4 % | 99,5 % | 92,30 % | 99,7 % | 94,3 % | 98,7 % | 44,0 % | 98,0 % |
| 30 | 95,8 % | 99,6 % | 99,1 % | 99,2 % | 93,80 % | 99,6 % | 98,8 % | 98,8 % | 94,2 % | 98,0 % |
| 35 | 95,6 % | 99,6 % | 99,4 % | 99,5 % | 94,80 % | 99,5 % | 76,7 % | 98,6 % | 46,4 % | 97,9 % |
| 40 | 95,6 % | 99,6 % | 99,5 % | 99,6 % | 94,90 % | 99,4 % | 92,4 % | 98,7 % | 94,2 % | 98,0 % |
| 42 | 95,3 % | 99,6 % | 99,6 % | 99,7 % | 95,30 % | 99,9 % | 92,4 % | 98,6 % | 94,1 % | 97,8 % |
| 50 | 95,8 % | 99,5 % | 99,5 % | 99,6 % | 95,40 % | 99,4 % | 80,0 % | 98,7 % | 84,0 % | 97,9 % |
| 60 | 95,8 % | 99,5 % | 99,2 % | 99,3 % | 95,70 % | 99,1 % | 95,0 % | 98,8 % | 94,3 % | 98,0 % |
| 100 | 95,7 % | 99,5 % | 99,5 % | 99,6 % | 80,00 % | 99,2 % | 77,1 % | 98,8 % | 43,7 % | 97,9 % |

### 4.1.8. Validación cruzada

En el proceso de validación cruzada permitió medir la calidad de la predicción en las técnicas. Este procedimiento se realizó al interior del conjunto de datos de entrenamiento, al dividir los datos de forma aleatoria en cinco grupos que son aproximadamente del mismo tamaño. Este proceso se repite cinco veces utilizando un grupo distinto como validación en cada iteración, mientras que el resto de los datos se tomó como conjunto de entrenamiento. Una vez finalizadas las iteraciones, se calculó la certeza con el promedio de los cinco modelos entrenados, en primera instancia se generaron los modelos sin la técnica de escalado de los datos y en la segunda parte de esta sección se generaron los resultados con la técnica de escalado.

Al revisar los resultados obtenidos de la validación cruzada sin datos escalados de la Tabla 13 permitió evidenciar que la técnica de Regresión Logistica presentó una media de 95,5 %, similar a la certeza del conjunto de datos de entrenamiento 95,7 % y de test 95,8 %, de igual forma la técnica Árboles de Decisión presentó una media de 98,5 %, similar a la certeza del conjunto de datos de entrenamiento 98,9 % y de test 97,9 % y la técnica Máquinas de vectores de soporte tuvo una media de 94,8 %, parecida a la certeza del conjunto de datos de entrenamiento y de test 93,8 %.

La técnica Redes Neuronales presentó una variación menor en la media 80,7 % respecto a las métricas de entrenamiento 93,7 % y test 93,8 % y en la técnica Redes Bayesianas también se observa una reducción pero los resultados siguen siendo buenos con una media de 87,5 % y en las métricas de entrenamiento 94,1 % y test 94,2 %.

Es importante tener en cuenta que, aunque la media del porcentaje de certeza de los 5 folds en la

validación cruzada para algunos métodos como Redes Neuronales y Redes Bayesianas, se redujo un poco el resultado sigue siendo satisfactorio.

Tabla 13: Validación cruzada con datos sin escalar

| **Métricas cross_validation** | **Regresión Logística** | **Árbol de Decisión** | **Redes Neuronales** | **SVM** | **Redes Bayesianas** |
|---|---|---|---|---|---|
| **Fold1** | 95,2 % | 98,9 % | 86,3 % | 94,9 % | 94,0 % |
| **Fold2** | 95,4 % | 97,9 % | 80,9 % | 94,9 % | 94,1 % |
| **Fold3** | 95,8 % | 98,9 % | 45,2 % | 93,7 % | 60,9 % |
| **Fold4** | 95,3 % | 98,9 % | 95,3 % | 94,8 % | 93,9 % |
| **Fold5** | 95,8 % | 98,9 % | 95,8 % | 95,6 % | 94,3 % |
| **Media cross_validation** | 95,5 % | 98,5 % | 80,7 % | 94,8 % | 87,5 % |
| **Métrica en Test** | 95,8 % | 97,9 % | 93,8 % | 93,8 % | 94,2 % |
| **Métrica en Train** | 95,7 % | 98,9 % | 93,7 % | 93,8 % | 94,1 % |

A continuación se ilustran los datos de la validación cruzada con la técnica escalado de los datos en la Tabla 14.

En general permitio observar que para las diferentes técnicas la media en la validación cruzada es muy parecida a los datos de entrenamiento y de test. Respecto a los resultados comparando las diferentes técnicas se observa que las técnicas de Máquinas de Vectores de Soporte y Redes Bayesianas redujeron la media en 0,8 puntos porcentuales y 1,6 puntos porcentuales respecto a las otras técnicas pero la calidad de los modelos sigue siendo buena.

Tabla 14: Validación cruzada con datos escalados

| **Métricas cross_validation** | **Regresión Logística** | **Árbol de Decisión** | **Redes Neuronales** | **SVM** | **Redes Bayesianas** |
|---|---|---|---|---|---|
| **Fold1** | 99,5 % | 99,4 % | 99,8 % | 98,7 % | 98,0 % |
| **Fold2** | 99,6 % | 99,5 % | 99,4 % | 98,8 % | 98,2 % |
| **Fold3** | 99,5 % | 99,9 % | 99,9 % | 98,8 % | 98,0 % |
| **Fold4** | 99,6 % | 99,6 % | 99,3 % | 98,7 % | 97,9 % |
| **Fold5** | 99,6 % | 99,8 % | 99,6 % | 98,8 % | 97,9 % |
| **Media cross_validation** | 99,6 % | 99,6 % | 99,6 % | 98,8 % | 98,0 % |
| **Métrica en Test** | 99,5 % | 99,8 % | 99,6 % | 98,8 % | 98,0 % |
| **Métrica en Train** | 99,6 % | 99,8 % | 99,8 % | 98,7 % | 98,1 % |

### 4.1.9. Tratamiento de clases desbalanceadas

Los datos desbalanceados corresponden a las clases minoritarias que presentan muy pocas muestras en la variable objetivo, la clase tipo 1 tiene 205 empresas, la clase tipo 2 tiene 1.157 empresas, la clase tipo 3 tiene 10.129 empresas, la clase tipo 4 tiene 42.872 empresas, mientras que la clase tipo 0 tiene 181.112 empresas. Este tipo de eventos puede afectar a los algoritmos en su generalización perjudicando a las clases minoritarias, por tal razón es importante evaluar correctamente el conjunto de datos con estas caracteristicas a partir de la medición de métricas diferentes a la certeza cuando se comparan diferentes modelos, como la precisión, la exhaustividad, medida-F o la curva ROC, la que mejor vaya para nuestra aplicación.

Un método para abordar conjuntos de datos desequilibrados es sobremuestrear la clase minoritaria, que consiste en generar multiples registros en la clase minoritaria hasta equilibrar las diferentes clases, aunque estos no agregan ninguna información nueva al modelo. Este es un tipo de aumento de datos para la clase minoritaria se desarrollo bajo la técnica de sobremuestreo de minorías sintéticas o SMOTE.

En la Tabla 15 se observa el comparativo de los modelos sin balancear y después de realizar el balanceo de las clases minoritarias, los siguientes son los resultados más relevantes:

En la técnica Regresión Logística se incrementó el porcentaje de exhaustividad en 10 % para la clase tipo 1, la cual fue de (82 %), después de realizar el balanceo. En está misma técnica disminuyó la precisión en 25 % para la clase tipo 1 quedó en (60 %) y 19 % para la clase tipo 2 dio (71 %).

En la técnica Árboles de Decisión y Redes Neuronales no hubo variaciones significativas después de realizar el balanceo de clases.

En la técnica Máquinas de Vectores de Soporte se incremento el porcentaje de exhaustividad en 11 % para la clase 1 fue (80 %), después de realizar el balanceo. En está misma técnica disminuyó la precisión en 59 % para la clase 1 dio (7 %) y 10 % para la clase 2 quedó en (37 %), también disminuyó medida - F un 55 % en la clase tipo 1 fue (37 %).

En la técnica Bayes Gaussianos disminuyó el porcentaje de exhaustividad en 7 % para la clase tipo 1 fue (80 %), se incrementó el porcentaje de exhaustividad en 52 % para la clase tipo 2 dio (77 %), después de realizar el balanceo. En está misma técnica aumentó la precisión en 20 % para la clase tipo 2 quedó en (56 %) y en medida - F se incrementó 36 % para la clase 2 dio (65 %).

Tabla 15: Comparativo aplicando la técnica de sobremuestreo de clases desbalanceadas - datos de pruebas

| | Etiqueta | Regresión Logística | | | Árboles de Decisión | | | Redes Neuronales | | | SVM | | | Redes Bayesianas | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif |
| **Certeza** | **0,1,2,3,4** | 98 % | 99 % | 1 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 98 % | 98 % | 0 % | 98 % | 98 % | 0 % |
| **Exhaustividad** | **0** | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % |
| | **1** | 72 % | 82 % | 10 % | 98 % | 97 % | -1 % | 95 % | 95 % | 0 % | 69 % | 80 % | 11 % | 87 % | 80 % | -7 % |
| | **2** | 80 % | 87 % | 7 % | 98 % | 99 % | 1 % | 94 % | 97 % | 3 % | 80 % | 78 % | -2 % | 25 % | 77 % | 52 % |
| | **3** | 95 % | 95 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 77 % | 86 % | 9 % | 98 % | 97 % | -1 % |
| | **4** | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 97 % | 95 % | -2 % |
| **Precisión** | **0** | 99 % | 99 % | 0 % | 99 % | 98 % | -1 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % |
| | **1** | 85 % | 60 % | -25 % | 98 % | 97 % | -1 % | 90 % | 97 % | 7 % | 66 % | 7 % | -59 % | 10 % | 11 % | 1 % |
| | **2** | 90 % | 71 % | -19 % | 98 % | 98 % | 0 % | 99 % | 97 % | -2 % | 47 % | 37 % | -10 % | 36 % | 56 % | 20 % |
| | **3** | 95 % | 94 % | -1 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 94 % | 95 % | 1 % | 81 % | 80 % | -1 % |
| | **4** | 95 % | 99 % | 4 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 97 % | 99 % | 2 % | 99 % | 100 % | 1 % |
| **Medida – F** | **0** | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % |
| | **1** | 78 % | 69 % | -9 % | 98 % | 97 % | -1 % | 93 % | 96 % | 4 % | 68 % | 13 % | -55 % | 18 % | 19 % | 1 % |
| | **2** | 85 % | 78 % | -7 % | 98 % | 99 % | 1 % | 97 % | 97 % | 0 % | 59 % | 50 % | -9 % | 29 % | 65 % | 36 % |
| | **3** | 95 % | 94 % | -1 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 85 % | 90 % | 5 % | 88 % | 88 % | 0 % |
| | **4** | 97 % | 99 % | 2 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 98 % | 99 % | 1 % | 98 % | 97 % | -1 % |

Otro método para abordar conjuntos de datos desequilibrados es submuestrear la clase mayoritaria, que consiste en eliminar muestras de clases sobre representadas hasta nivelar el número de muestras de la clase minoritaria. El algoritmo utilizado fue RandomUnderSampler.

En la Tabla 16 se observa el comparativo de los modelos sin balancear y después de realizar el balanceo de las clases mayoritarias, los siguientes son los resultados más relevantes:

En la técnica Regresión Logística se incrementó el porcentaje de exhaustividad en 8 % para la clase tipo 1 fue de (80 %), después de realizar el balanceo. En está misma técnica disminuyó la precisión en 63 % para la clase tipo 1 quedó en (22 %) y 52 % para la clase tipo 2 dio (38 %).

En la Árboles de Decisión se incrementó el porcentaje de exhaustividad en 4 % para la clase tipo 3 fue (95 %), después de realizar el balanceo. En está misma técnica disminuyó la precisión en 11 % para la clase tipo 1 dio (87 %) y 21 % para la clase tipo 2 quedó en (77 %), también disminuyó medida - F un 5 % en la clase tipo 1 fue (93 %) y un 12 % en la clase tipo 2 dio (86 %) .

En la técnica Redes Neuronales se incremento el porcentaje de exhaustividad en 4 % para la clase tipo 1

fue (99 %), después de realizar el balanceo. En está misma técnica disminuyó la precisión en 82 % para la clase tipo 1 dio (8 %) y 62 % para la clase tipo 2 quedó en (37 %), también disminuyó medida - F un 78 % en la clase tipo 1 fue (15 %) y (44 %) en la clase tipo 2 fue (53 %).

En la técnica Máquinas de Vectores de Soporte se incrementó el porcentaje de exhaustividad en 11 % para la clase tipo 1 fue (70 %),después de realizar el balanceo. En está misma técnica disminuyó la precisión en 59 % para la clase tipo 1 dio (6 %) y 10 % para la clase tipo 2 quedó en (42 %), también disminuyó medida - F un 55 % en la clase tipo 1 fue (11 %).

En la técnica Bayes Gaussianos disminuyó el porcentaje de exhaustividad en 15 % para la clase tipo 1 fue (80 %), se incrementó el porcentaje de exhaustividad en 61 % para la clase tipo 2 dio (86 %), después de realizar el balanceo. En está misma técnica aumentó la precisión en 18 % para la clase tipo 2 quedó en (54 %) y en medida - F se incrementó 38 % para la clase tipo 2 dio (67 %).

Tabla 16: Comparativo aplicando la técnica de submuestreo de clases desbalanceadas

| | Etiqueta | Regresión Logística | | | Árboles de Decisión | | | Redes Neuronales | | | SVM | | | Redes Bayesianas | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif | Sin Balanceo | Con Balanceo | Dif |
| **Certeza** | **0,1,2,3,4** | 99 % | 98 % | -1 % | 99 % | 99 % | 0 % | 99 % | 98 % | -1 % | 98 % | 97 % | 0 % | 98 % | 98 % | 0 % |
| **Exhaustividad** | **0** | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 98 % | -1 % | 99 % | 98 % | 0 % | 99 % | 99 % | 0 % |
| | **1** | 72 % | 80 % | 8 % | 98 % | 99 % | 1 % | 95 % | 99 % | 4 % | 69 % | 70 % | 11 % | 87 % | 72 % | -15 % |
| | **2** | 80 % | 85 % | 5 % | 98 % | 97 % | -1 % | 94 % | 95 % | 1 % | 80 % | 87 % | -2 % | 25 % | 86 % | 61 % |
| | **3** | 95 % | 91 % | -4 % | 99 % | 95 % | -4 % | 99 % | 97 % | -2 % | 77 % | 90 % | 9 % | 98 % | 95 % | -3 % |
| | **4** | 99 % | 98 % | -1 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 96 % | 0 % | 97 % | 94 % | -3 % |
| **Precisión** | **0** | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % |
| | **1** | 85 % | 22 % | -63 % | 98 % | 87 % | -11 % | 90 % | 8 % | -82 % | 66 % | 6 % | -59 % | 10 % | 7 % | -3 % |
| | **2** | 90 % | 38 % | -52 % | 98 % | 77 % | -21 % | 99 % | 37 % | -62 % | 47 % | 42 % | -10 % | 36 % | 54 % | 18 % |
| | **3** | 95 % | 92 % | -3 % | 99 % | 94 % | -5 % | 99 % | 96 % | -3 % | 94 % | 85 % | 1 % | 81 % | 78 % | -3 % |
| | **4** | 95 % | 99 % | 4 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 97 % | 99 % | 2 % | 99 % | 99 % | 0 % |
| **Medida – F** | **0** | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % |
| | **1** | 78 % | 12 % | -66 % | 98 % | 93 % | -5 % | 93 % | 15 % | -78 % | 68 % | 11 % | -55 % | 18 % | 13 % | -5 % |
| | **2** | 85 % | 66 % | -19 % | 98 % | 86 % | -12 % | 97 % | 53 % | -44 % | 59 % | 57 % | -9 % | 29 % | 67 % | 38 % |
| | **3** | 95 % | 89 % | -6 % | 99 % | 95 % | -4 % | 99 % | 97 % | -2 % | 85 % | 87 % | 5 % | 88 % | 86 % | -2 % |
| | **4** | 97 % | 98 % | 1 % | 99 % | 99 % | 0 % | 99 % | 99 % | 0 % | 98 % | 98 % | 1 % | 98 % | 97 % | -1 % |

### 4.1.10. Algoritmos de conjunto o refuerzo

El algoritmo AdaBoost combina múltiples clasificadores para aumentar la certeza del conjunto de datos al combinar varios clasificadores de bajo rendimiento y de está forma obtener un clasificador fuerte de alta precisión. Es importante establecer los pesos de los clasificadores y entrenar el conjunto de datos en cada iteración para garantizar predicciones precisas de observaciones inesperadas. Cualquier algoritmo de aprendizaje automático puede usarse como clasificador base si acepta pesos en el conjunto de entrenamiento.Los algoritmos deben cumplir dos condiciones: el clasificador debe ser entrenado interactivamente en varios modelos de entrenamiento pesados y en cada iteración, intenta proporcionar un buen ajuste minimizando el error de entrenamiento (Naviani,2018).

En la Tabla 17 se puede observar que al aplicar el algoritmo AdaBoost en Árboles de Decisión presenta excelentes resultados en todas las métricas con un 99,9 % de eficiencia. Se realizó el ejercicio con el algoritmo Ada Boost parametrizando los estimadores base como predeterminados por el sistema y un número de estimaciones de 50, en la Tabla 17 se identifica un porcentaje de certeza del 97 %, para las clases 0, 2 y 4 logró buenas estimaciones en exhaustividad y para las clases 0 y 4 en precisión que contrasta con el bajo desempeño logrado con la clase 1 en todas las métricas, según la literatura este tipo de métodos son óptimos en conjuntos de datos mayores a 1000.

El algoritmo AdaBoost también fue utilizado para mejorar el desempeño y el balanceo de los datos en las técnicas empleadas en el alcance de este trabajo, en la técnica de Regresión Logistica no hubo mejoría al pasar de un porcentaje de certeza del 99 % al 86 %, en las demás métricas el comportamiento fue similar, en Máquinas de Vectores de Soporte no hubo optimización sobre las métricas y en Redes Bayesianas se observo el mayor rendimiento en la clase tipo 2 al pasar del 25 % al 42 % en exhaustividad, en precisión

al pasar del 10 % al 69 % para la clase tipo 1 y del 36 % al 66 % para la clase tipo 2 y en medida-F al pasar del 18 % al 77 % en la clase tipo 1 y del 29 % al 52 % en la clase tipo 2.

Extreme Gradient Boosting (XGBoost, por sus siglas en Inglés) es un algoritmo predictivo supervisado que utiliza el principio de boosting. La idea detrás del boosting es generar múltiples modelos de predicción "débiles" secuencialmente, y que cada uno de estos tome los resultados del modelo anterior, para generar un modelo más "fuerte", con mejor poder predictivo y mayor estabilidad en sus resultados (Mendoza,2020).

XGBoost usa como sus modelos débiles árboles de decisión de diferentes tipos, que pueden ser usados para tareas de clasificación, al observar los resultados de la Tabla 17, los porcentajes de certeza en su mayoria están en el 99 %, aunque para la clase 1 se reduce a 85 % pero sigue conservando buenos resultados.

Tabla 17: Comparativo algoritmos AdaBoost y XGBoost

| Métrica/Técnica | Etiqueta | AdaBoost (Arboles) | XGBoost | AdaBoost (Total) | Regresión Logistica | | SVM | | Redes Bayesianas | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Origen | AdaBoost | Origen | AdaBoost | Origen | AdaBoost |
| **Certeza** | **0,1,2,3,4** | 99 % | 99 % | 97 % | 99 % | 86 % | 98 % | 98 % | 98 % | 98 % |
| **Exhaustividad** | **0** | 99 % | 99 % | 99 % | 100 % | 83 % | 99 % | 99 % | 99 % | 99 % |
| | **1** | 99 % | 85 % | 0 % | 72 % | 82 % | 69 % | 62 % | 87 % | 87 % |
| | **2** | 99 % | 99 % | 87 % | 80 % | 44 % | 80 % | 63 % | 25 % | 42 % |
| | **3** | 99 % | 99 % | 72 % | 95 % | 77 % | 77 % | 76 % | 98 % | 84 % |
| | **4** | 99 % | 99 % | 90 % | 99 % | 99 % | 99 % | 99 % | 97 % | 99 % |
| **Precisión** | **0** | 99 % | 99 % | 99 % | 99 % | 99 % | 99 % | 99 % | 99 % | 99 % |
| | **1** | 99 % | 99 % | 0 % | 85 % | 2 % | 66 % | 18 % | 10 % | 69 % |
| | **2** | 99 % | 97 % | 53 % | 90 % | 52 % | 47 % | 77 % | 36 % | 66 % |
| | **3** | 99 % | 99 % | 63 % | 95 % | 24 % | 94 % | 97 % | 81 % | 90 % |
| | **4** | 99 % | 99 % | 95 % | 95 % | 96 % | 97 % | 95 % | 99 % | 97 % |
| **Medida – F** | **0** | 99 % | 99 % | 99 % | 99 % | 91 % | 99 % | 99 % | 99 % | 99 % |
| | **1** | 99 % | 92 % | 0 % | 78 % | 5 % | 68 % | 28 % | 18 % | 77 % |
| | **2** | 99 % | 98 % | 66 % | 85 % | 48 % | 59 % | 69 % | 29 % | 52 % |
| | **3** | 99 % | 99 % | 67 % | 95 % | 37 % | 85 % | 85 % | 88 % | 87 % |
| | **4** | 99 % | 99 % | 93 % | 97 % | 98 % | 98 % | 98 % | 98 % | 98 % |

## 4.2. Segmentación y estrategias

La segmentación generada permitió focalizar las estrategias de gestión de cartera sobre los deudores morosos. En el primer segmento se encuentran las empresas que presentan una clase o grupo alto asociado a variables como índice de pago, ingreso base de cotización, ingreso base de cotización promedio, valor pagado, sector empleador, segmento, tamaño, número de trabajadores y un bajo nivel de deuda presunta; en este grupo las estrategias deben ir enfocadas a tener un contacto amable y de servicio con la empresa, para que normalice sus aportes o reporte las novedades correspondientes. En el segundo segmento se encuentran las empresas con clase o grupo aceptable asociado a las variables ya mencionadas; en este grupo las estrategias deben ir enfocadas a realizar un proceso de cobro disuasivo para que de forma voluntaria el deudor realice el pago de sus aportes. En el tercer segmento están las empresas que presentan una clase o grupo regular asociado a las variables ya acotadas, las cuales son reiterativas en el no pago de los aportes a pensión, a estas se les debe implementar estrategias de cobro persuasivo. En el cuarto segmento se encuentran las empresas con una clase o grupo deficiente asociado a las variables ya descritas; en este grupo las estrategias deben ir enfocadas a realizar procesos de cobro coactivo y decretar medidas cautelares para que -mediante embargos- se pueda recuperar las obligaciones adeudadas.

Teniendo en cuenta la segmentación presentada, la evaluación y optimización de los modelos con las

técnicas de Regresión Logística, Árboles de Decisión, Redes Neuronales, Máquinas de Vectores de Soporte y Redes Bayesianas se plantearon varias estrategias de cobro de aportes pensionales, las cuales permitirán, en el mediano o largo plazo, reducir el nivel de mora en las empresas. En la siguiente tabla detallamos las estrategias generadas de acuerdo con la segmentación.

Tabla 18: Segmentación y estrategias de cobro

| Segmento | Estrategia | Segmento | Estrategia |
|---|---|---|---|
| 0 - 1 | Visita de cobro comercial | 2 | Gestión de contact center aviso de incumplimiento |
| | Brochere electrónico o extracto de deudas | | Mensajes de texto |
| | Gestión de contact center cobro preventivo | | Carta disuasiva |
| | Recepción de trámites y asesoría prioritaria en puntos de atención | | Recepción de trámites y asesoría en puntos de atención |
| 3 | Gestión de contact center cobro persuasivo | 4 | Liquidación certificada de deuda |
| | Mensaje de texto | | Ingreso a cobro coactivo (Mandamiento de Pago) |
| | Carta de cobro persuasivo | | Medidas cautelares |
| | Recepción de trámites y asesoría en puntos de atención | | Recepción de trámites y asesoría en puntos de atención |

# 5. Conclusiones y recomendaciones

## 5.1. Conclusiones

A lo largo del presente trabajo de investigación se planteó la aplicación de métodos de *Machine Learning* como herramienta base para clasificar las empresas deudoras y no deudoras, y así maximizar los recursos y hacer más oportuna la gestión de cobro de los aportes a pensión. Para ello, se determinaron las principales variables predictoras, se implementaron las diferentes técnicas de aprendizaje supervisado y se evaluó el modelo con indicadores de calidad para evidenciar la eficiente generalización de los algoritmos de supervisión sobre la base de datos.

Los principales resultados hallados fueron los siguientes:

- Se evidencia que en la base de datos del fondo de pensiones público, el 77 % de las empresas presentan pagos completos, mientras que el 23 % se encuentran en mora; el 0,4 % de esos deudores corresponde a empresas que presentan un índice de pagos alto y sus variables son objetivo del fondo de pensiones: se encuentran bastantes trabajadores cotizando al fondo de pensiones y el nivel de recaudo es bueno, posiblemente estas empresas se encuentran en mora por inconsistencias en el reporte de las novedades por sus trabajadores afiliados al fondo de pensiones. En contraste, el 78,9 % de las empresas con mora presentan un índice de pago muy bajo: son empresas con segmentos y tamaños no objetivo para el fondo, cuyo recaudo y número de trabajadores es bajo, de allí la necesidad de generar acciones de cobro persuasivo y coactivo que permitan recuperar los dineros adeudados. El porcentaje restante de los deudores se encuentra clasificado en aceptables 2,1 % y regulares 18,6 %; para estos se deben generar estrategias diferenciadas de cobro, que permitan obtener oportunamente el pago de estas obligaciones pensionales.

- En el proceso de evaluación de rendimiento de los modelos se logró evidenciar que la técnica Árboles de Decisión presenta excelentes resultados: no requirió estandarización de los datos al lograr un porcentaje de certeza excelente y clasificó de forma rápida y eficiente la variable objetivo en una base de datos con un número adecuado de registros. Las demás técnicas mostraron buenos resultados en la clase tipo 0, 3 y 4 tanto en exhaustividad como en medida-F, mientras se redujo el desempeño para las técnicas Regresión Logistica 71,8 % y Máquinas de Vectores de Soporte 69,2 % en exhaustividad y Redes Bayesianas 18,5 % en medida-F, lo anterior para la clase tipo 1. En la técnica Redes Bayesianas para la clase tipo 2 se redujo en 24,7 % y 29,3 % tanto en exhaustividad como en medida-F y Máquinas de Vectores de Soporte en 59,4 % para medida-F.

- Durante el desarrollo del modelo los parámetros utilizados dieron un buen desempeño al obtener una certeza que estuvo por encima del 93 % para todas las técnicas, esto aunado a la mejora incremental con el escalado de los datos que finalmente generó un resultado mayor al 98 %. En las técnicas de Regresión Logística y Máquinas de Vectores de Soporte los parámetros de regularización entre más pequeños, fueron más fuertes y presentaron mejores resultados. En la técnica Árboles de Decisión el método entropía permitió medir cuanta información brindó la característica sobre la clase a través de la ganancia de la información, en ese sentido dio a conocer el orden de los nodos en los atributos. En la técnica de Redes Neuronales un número de iteraciones de 100 y un generador de números aleatorios de 42 para la inicialización de ponderaciones y sesgos generó buenos resultados y en Redes Bayesianas un parámetro predeterminado por el algoritmo agregó la varianza para la estabilidad del cálculo.

- Al comparar el desempeño de varios algoritmos de aprendizaje supervisado en aplicaciones de clasificación, se realizó el escalamiento de las variables independientes tanto del conjunto de datos de entrenamiento como de test para unificar la dimensionalidad de las variables, y de esta forma mejorar ostensiblemente el poder de discriminación del modelo. Los resultados obtenidos dieron como resultado un micro-average ROC curve (área= 0,99) y un macro-average ROC curve (área= 0,92), con excelentes resultados para las clases 0 y 4 con una ROC curve de 1.0 y para la clase 1 una ROC curve de 0.95.

- La validación del modelo con diferentes semillas permitió garantizar que las divisiones generadas fueran reproducibles, las filas fueron asignadas a la base de entrenamiento de forma aleatoria. Esto aseguró que el conjunto de datos fuera una muestra representativa del conjunto de datos original. En las técnicas Regresión Logística, Árboles de Decisión y Máquina de Vectores de Soporte fueron muy estables los resultados al realizar la modificación en el parámetro random state, mientras que para los modelos Redes Neuronales y Redes Bayesianas la certeza se redujo para algunas semillas. Sin embargo es importante evaluar el modelo con otro tipo de métricas como exhaustividad, precisión y medida-F que permita observar la calidad de la predicción en las clases tanto minoritarias como mayoritarias de forma independiente, para saber que también están generalizando los algoritmos en las técnicas utilizadas.

- La evaluación del modelo con validación cruzada permitió medir la calidad de la predicción en las técnicas, al dividir los datos de forma aleatoria en cinco grupos que son aproximadamente del mismo tamaño. Esto dio como resultado que las técnicas de Regresión Logística y Árboles de Decisión en sus cinco folds obtuvieran una media excelente en el porcentaje de certeza, muy parecida a los resultados de la métrica de entrenamiento y test de la base de datos. Para las técnicas Redes Neuronales, Máquina de Vectores de Soporte y Redes Bayesianas hubo diferencias, pero no considerables en algunos folds. Lo anterior evidencia que, aunque se redujo la efectividad, el modelo sigue conservando una buena certeza: no presenta overfitting, está generalizando y aprendiendo correctamente para los nuevos registros que puedan ingresar al modelo mes a mes.

- El conjunto de datos presenta una distribución desigual entre sus clases porque el número de empresas que pagan correctamente sus aportes es mayor al número de deudores y estos últimos a su vez presentan una distribución de acuerdo al tipo de deudor. Se utilizaron dos métodos para abordar el conjunto de datos desbalanceados, uno referente al sobremuestreo de las clases minoritarias, el cual dio buenos resultados en el indicador de exhaustividad aumentando los porcentajes de certeza en las clases tipo 1 y 2, en las demás clases no hubo variaciones significativas en las técnicas. En el indicador de precisión hubo una reducción de desempeño para la clase tipo 1 y 2 en las técnicas de Regresión Logistica y Máquinas de Vectores de Soporte en comparación con las técnicas sin balancear. El otro método hace referencia al submuestreo de la clase mayoritaria, presentando buenos resultados en todas las técnicas para el indicador de exhaustividad, mientras que pierde desempeño en las técnicas de las clases tipo 1 y 2 en el indicador de precisión y medida F, para las demás clases no existen variaciones significativas.

- El método AdaBoost ofrecio buenos resultados en la técnica de Árboles de Decisión y Redes Bayesianas, este algoritmo puede ser una solución para aumentar el rendimiento de modelo de baja precisión a modelos de alta precisión. El método XGBoost ofrecio buenos resultados al mejorar el desempeño de los Árboles de Decisión débiles, este algoritmo puede ser una buena alternativa cuando se tiene modelos muy complejos o con variables heterogéneas.

- Se propone una segmentación que permite clasificar las empresas que pagan sus aportes oportunamente y las que no, estas empresas a su vez se agrupan por tipo de deudor. Esto permite generar estrategias diferenciadas de cobro para maximizar los recursos por el alto volumen de entidades que entran en mora mensualmente, estas actividades van desde una visita comercial, gestión de contact center para cobro preventivo o un extracto con información de pagos, para deudores de baja criticidad, pasando por una carta de cobro persuasivo, asesoramiento en los puntos de atención o mensajes de texto para deudores de criticidad media, hasta el proceso de cobro coactivo, embargos y demás medidas para los deudores que son renuentes al pago.

### 5.2. Recomendaciones o trabajos futuros

Para esta investigación se utilizó la base de empresas, pero es de vital importancia para estudios posteriores incluir la base de afiliados independientes, de conformidad con lo establecido en el Sistema General de Pensiones. Allí se hace referencia a que es obligatoria tanto para trabajadores dependientes como independientes: la afiliación y su cotización no podrá ser sobre una base inferior al monto equivalente al salario mínimo mensual, por lo que es de estricto cumplimiento las cotizaciones de los afiliados independientes en el Sistema General de Pensiones.

Otra propuesta para futuras investigaciones podría ser la utilización de un algoritmo no supervisado de clustering como k-mean u otros métodos, que posiblemente permitan generar la variable objetivo o dependiente de forma automática, con esto se ahorraría tiempo en la generación de esta etiqueta.

## 6. Referencias bibliográficas

Agresti, A. (2002). Análisis de datos categóricos. Segunda edición, John Wiley  Sons, Inc., Nueva York. En línea. Recuperado de: http://dx.doi.org/10.1002/0471249688.

Alpaydin, E. (2004). Introduction to Machine Learning. The MIT press Cambridge, MA.

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

Aruna, R. & Nirmala, K. (2013). Construction of Decision Tree: Attribute Selection Measures. International Journal of Advancements in Research  Technology, Volume 2, Issue 4. Recuperado de: http://www.ijoart.org/docs/Construction-of-Decision-Tree–Attribute-Selection-Measures.pdf.

Brito, F. & Artes, R. (2018). Aplicación de árboles de regresión aditiva bayesiana en el desarrollo de modelos de calificación crediticia en Brasil. Producción, 28., https://doi.org/10.1590/0103-6513.20170110.

Cohen, J. (1960). Un coeficiente de acuerdo con las escalas nominales. Medida educativa y psicológica, 20 (1), pp. 37-46. Doi: 10.1177 / 001316446002000104.

Colfondos. (2013). Manual del participante, Ley 100 de 1993.En línea. Recuperado de: https://www.colfondos.com.co/dxp/documents/20143/37693/LEY+100+DE+1993.pdf/c2be65aa-08dd-decc-447c-647409ce4f12.

International Business Machines Corporation (2019). Funcionamiento de SVM. Recuperado de: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=models-how-svm-works.

Garcia, N. (2020).Qué son los árboles de decisión y para que sirven.Recuperado de: https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

Husejinovic et al. (2018). Aplicación de algoritmos de aprendizaje automático en la predicción de pagos predeterminados de tarjetas de crédito.Recuperado de: https://www.researchgate.net/publication/328026972-Application-of-Machine- Learning-Algorithms-in-Credit-Card-Default-Payment-Prediction.

Ironhack (2015). ¿En qué consiste el Machine Learning?. En línea. Recuperado de: https://www.ironhack.com/es/data-analytics/que-es-machine-learning.

López, R. (2015). Machine Learning con Python. En línea. Recuperado de: https://relopezbriega.github.io/blog/2015/10/10/machine-learning-con-python/.

Mendoza, J. (2020). XGBoost en Python. En linea. Recuperado de: https://medium.com/@jboscomendoza/tutorial-xgboost-en-python-53e48fc58f73.

Müller, A. & Guido, S. (2017). Introduction to Machine Learning with Python. A Guide for Data Scientists. O'reilly, United States of America.

Naviani, (2018).Clasificador AdaBoost en Python. Recuperado de: https://www.datacamp.com/tutorial/adaboost-classifier-pythonrdl.

Nieto, S. (2010). Crédito al Consumo: La estadística aplicada a un problema de riesgo crediticio [Tesis de Maestría]. Universidad Autónoma Metropolitana. Recuperado de: http://mat.izt.uam.mx/mcmai/documentos/tesis/Gen.07-O/Nieto-S-Tesis.pdf.

Olarte, N. (8 de abril de 2016). El pequeño dato que puede arruinar su futuro. Revista Semana. En línea. Recuperado de: http://www.finanzaspersonales.co/pensiones-y-cesantias/articulo/que-hacer-cuando-la-empresa-no-hace-aportes-a-pension/59958.Fecha de consulta: noviembre de 2018.

Oñate (2016). Análisis de la Deserción y Permanencia Académica en la Educación superior Aplicando Minería De Datos. Universidad Nacional de Colombia.

Parra, F. (2017). Estadística y Machine Learning con R. Rpubs. Recuperado de: https://rpubs.com/PacoParra/293405, Fecha de consulta: noviembre de 2018.

Raschka & Mirjalili (2019). Python Machine Learning. Aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow. Marcombo.

Resolución 2082 de 2016 (2017). Principales cambios o ajustes. Proceso de extracción de conocimiento. Unidad de Gestión Pensional y Parafiscales.

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

Sancho, F. (2017). Redes Neuronales: una visión superficial. En línea. Recuperado de: http://www.cs.us.es/ fsancho/?e=72.

Srinath & Gururaja (2022). Aprendizaje automático explicable en la identificación de morosos de tarjetas de crédito.Recuperado de: https://www.sciencedirect.com/science/article/pii/S2666285X22000619.

Statistical Analysis System (SAS) Institute. (2019). Machine Learning, una expresión de la Inteligencia Artificial. En línea. Recuperado de: https://www.sas.com/content/dam/SAS/es$_m x/doc/whitepaper$1/109075$_0$917$.pdf$.