Machine Learning para la estimación del riesgo de crédito en una cartera de consumo

Wbeimar Ossa Giraldo

Verónica Jaramillo Marín

Curso de verano

Presentado como requisito parcial para obtener el título de

Magíster en Administración Financiera

Asesor: Brayan Rojas Ormaza. FRM. Msc

Universidad EAFIT

Escuela de Economía y Finanzas

Maestría en Administración Financiera

Medellín

2021

Agradecimientos

Deseo dar las gracias a todas aquellas personas que de forma directa o indirecta ayudaron durante este proceso formativo; este logro viene de la suma de todos y cada uno de sus aportes: a mi familia por darme soporte en estos últimos dos años, a mis profesores por sentar las bases de este trabajo, y a mi tutor en formación de ciencia de datos, Oscar Pérez, por el apoyo brindado a la hora de poder llevar un equilibrio entre trabajo, estudio de ciencia de datos y vida personal.

Wbeimar Ossa

No se puede esperar que se queme una casa para comprar un seguro contra incendio. No podemos esperar hasta que haya dislocaciones masivas en nuestra sociedad para prepararnos para la Cuarta Revolución Industrial.

Robert J. Shiller (2016)[1]

[1] Foro Económico Mundial de Davos, 2016.

Resumen

Las entidades financieras por su naturaleza de negocio se encuentran inherentemente expuestas al riesgo de crédito, por ello, están continuamente en búsqueda de nuevas formas para medir la probabilidad de incumplimiento de los clientes que solicitan un crédito. Esta investigación tiene como objetivo comparar la precisión de un modelo de regresión logística frente a algunos modelos de *Machine Learning,* para la estimación del riesgo de crédito en una cartera de consumo; dichas metodologías se perfilan como una herramienta clave para la estimación de riesgos, debido a su flexibilidad y capacidad de aprendizaje. Para ello, se utilizaron los modelos de Regresión logística, *Random Forest, Support Vector Machine* y *Multi-layer Perceptron*, haciendo una comparación en la eficiencia de la estimación de los clientes que van a entrar en mora, y obteniendo como resultado que el modelo más equilibrado al momento de la evaluación es el *Random Forest*, dado que fue el que presentó el mejor ajuste de acuerdo con las métricas de exactitud evaluadas.

Palabras claves: Inteligencia artificial, Riesgo de crédito, *Machine Learning*, Modelos predictivos, Regresión logística.

Abstract

Financial entities, due to their business nature, are inherently exposed to credit risk, for this reason, they are continually searching for new ways to measure the probability of default of clients requesting a loan. This research aims to comparing the precision of a logistic regression model against basic *Machine Learning* models for estimating credit risk in a consumer loan portfolio, these methodologies are emerging as a key tool for estimating risks due to their flexibility and learning capacity. For this, the Logistic Regression, Random Forest, Support Vector Machine and Multilayer Perceptron models were used, making a comparison in the efficiency of the estimation of the clients that are going to default, and obtaining as a result that the most balanced model at time of evaluation is the Random Forest.

Key words: Artificial Intelligence, Credit Risk, *Machine Learning*, Predictive Models, Logistic Regression.

Contenido

Lista de figuras

Lista de tablas

Lista de ecuaciones

1. Introducción

El riesgo de crédito, según el Comité de Supervisión Bancaria de Basilea, es definido como "la posibilidad de que un prestatario bancario o una contraparte no cumpla con sus obligaciones de acuerdo con los términos acordados" (Basel Committee on Banking Supervision, 1999, p. 1).

En Colombia, la Superintendencia Financiera define el riesgo de crédito como "la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que un deudor o contraparte incumpla sus obligaciones" (Superintendencia Financiera, 1995, p. 2).

Este riesgo es inherente a los bancos e instituciones financieras, los cuales deben hacer una gestión del mismo, con el objetivo de "maximizar la tasa de rendimiento ajustada al riesgo de un banco, manteniendo la exposición al riesgo de crédito dentro de parámetros aceptables" (Basel Committee on Banking Supervision, 1999, p. 1), y por lo tanto, además de sus políticas de otorgamiento de cartera, desarrollan modelos para medir la probabilidad de incumplimiento de sus clientes, y así poder definir con mayor precisión cuáles clientes son sujetos de crédito, según el apetito de riesgo[2] y el riesgo-retorno[3] esperado de cada entidad. Para tal fin y, específicamente en carteras masivas, como en la modalidad de consumo, comúnmente se han utilizado modelos estadísticos tradicionales, como por ejemplo, la regresión lineal y la regresión logística.

En la última década, se viene presentando un auge importante del uso de la Inteligencia Artificial (IA) y el Aprendizaje Automático (*Machine Learning*), para el desarrollo de modelos de toma de decisiones en diferentes sectores económicos. Las entidades financieras están adoptando a lo largo de toda su cadena de valor, la inteligencia artificial como parte de sus estrategias de transformación digital, para la mejora de procesos, reducción de costos, detección de fraudes, fidelización de clientes y toma de decisiones financieras y de riesgo, entre otros. Este auge se debe a dos factores importantes: el acceso a mayores fuentes y cantidades de datos y la evolución de la capacidad computacional para tratar los datos de forma más rápida y en mayor cantidad.

[2] Nivel máximo de tolerancia al riesgo dispuesto a asumir por una entidad para alcanzar sus objetivos.
[3] Relación entre nivel de riesgo asumido y su beneficio obtenido.

El *Machine Learning* como rama de la Inteligencia Artificial (AI), se ha convertido en una herramienta clave para el desarrollo de modelos de otorgamiento de crédito, puesto que su característica más importante es que son modelos que aprenden de los resultados para mejorar continuamente de forma autónoma, es decir, son flexibles y se adaptan a medida que los datos van entrando en el sistema para aprender de sus propias acciones.

A razón de lo anteriormente mencionado, el objetivo de este documento es comparar la precisión de un modelo de regresión logística frente a modelos de *Machine Learning,* para la estimación del riesgo de crédito en una cartera de consumo, determinando los modelos que resultan apropiados para evaluar el nivel de riesgo de esta cartera y hallar su nivel de precisión; también es importante destacar que durante este proceso, se buscó establecer las variables relevantes para ser utilizadas en modelos de *Machine Learning* en una cartera de consumo.

Con estos fines en mente, el documento realiza una revisión de los trabajos realizados en el área, que se acercan al caso evaluado; también efectúa un estudio de la fundamentación teórica de las metodologías utilizadas para la estimación del riesgo y brinda una descripción de la naturaleza del caso de estudio.

Posteriormente, se hace una detallada explicación de todos los procedimientos metodológicos necesarios para la correcta estimación de los modelos propuestos y cómo se llegó a estos; en este apartado, se hace también claridad sobre el alcance y las limitaciones presentadas por la información usada en el documento, así como los retos que su uso implica a la hora de aplicar los modelos propuestos.

Acto seguido, se presentan los resultados obtenidos después de la aplicación de todos los pasos metodológicos propuestos, donde se puede observar el conjunto de variables definidas como las más relevantes para la estimación de riesgo de crédito, así como las métricas de calidad de los modelos aplicados y cómo estos son efectivos para estimar el caso de estudio.

Como último apartado, el lector podrá encontrar las conclusiones derivadas de este trabajo, las cuales darán respuesta a las cuestiones propuestas por el caso de estudio, y se espera que se conviertan en un punto de apoyo para trabajos futuros relacionados con el tema.

2. Aproximación conceptual al problema

2.1 Estado del arte

En línea con los objetivos propuestos por el presente trabajo, se encuentran diversos artículos y documentos que hacen referencia a modelos de *Machine Learning,* para el análisis de riesgo de crédito; a continuación, se describe cómo ha sido abordado el tema.

El documento *Propuesta de Modelo para evaluación de Riesgo de Crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito LA,* del autor Juan Pablo Cuenca, (2019), manifiesta la necesidad de utilizar la información disponible y nuevas metodologías diferentes a modelos econométricos, utilizados convencionalmente para calcular riesgos de crédito. Se utiliza la regresión logística y, para la evaluación del modelo, se cuenta con curvas de características operativas del receptor (ROC), precisión, curva de elevación y valor de área bajo la curva (AUC). El autor concluye que la regresión logística obtiene el mejor rendimiento en la predicción del riesgo en lo que a métricas se refiere; el modelo planteado en el documento logra optimizar de manera significativa, la evaluación del riesgo que es tomada como referencia en la Cooperativa la Merced. Respecto al uso de técnica de *Machine Learning*, resalta su utilidad, pero reconoce que su efectividad puede estar sujeta a factores como: la calidad de los datos, la selección de variables y la aplicación de los algoritmos adecuados.

Por otro lado, Kruppa, Schwarz, Arminger y Ziegler (2013) plantean un marco para estimar riesgos crediticios de consumo, utilizando el *Machine Learning*; al igual que Cuenca, (2019), hacen referencia a la regresión logística como principal método para este ejercicio, sin embargo, mencionan que dicho método cuenta con una serie de limitantes asociadas a los supuestos subyacentes, entre las que están la correcta introducción de las variables al modelo y las interacciones contenidas, que adicionalmente no puede manejar problemas de multicolinealidad. Ante estas dificultades, el *Machine Learning* presenta otra serie de herramientas. Para este ejercicio presentado, utilizan estimación a través de regresiones no paramétricas y para estimar las probabilidades, se utilizan Árboles de Estimación de Probabilidades y *Random Forest*. Los resultados del ejercicio realizado con datos de prueba sobre créditos a plazos, muestran como resultado que la estimación utilizando *Random Forest* fue superior al modelo regresión logístico estándar, inclusive a una regresión logística ajustada.

A diferencia de los documentos citados anteriormente, Dastile, Celik y Potsane (2020) realizan una revisión sistemática de literatura, con el objetivo de analizar cómo se evalúa la solvencia crediticia de los prestatarios; reconocen que el método estadístico más utilizado es la regresión logística, sin embargo, la revisión realizada permitió determinar que existen métodos más sofisticados de *Machine Learning*, los cuales podrían reemplazar la regresión logística. Sin embargo, estos métodos podrían presentar problemas en la explicación de las predicciones y los conjuntos de datos desequilibrados. El estudio concluye que no existe un consenso en cuanto al rendimiento de los modelos estadísticos y los modelos de *Machine Learning*, para la evaluación de la calificación crediticia.

Por último, vale la pena mencionar que la temática abordada en el presente trabajo ha sido objeto de estudio en tesis; este es el caso de Trujillo (2017), un autor que presenta un recorrido por el concepto *Machine Learning* y los modelos que se adaptan para la gestión del riesgo, esto con la idea de conocer los alcances y las posibilidades, sin desconocer las dificultades que puede generar frente a la normativa que establece conocer plenamente el proceso, a través del cual son aprobados o rechazados los créditos; de esta manera, es posible determinar el grado de aplicabilidad y, adicionalmente, se hace un análisis de las metodologías tradicionales utilizadas para el cálculo del riesgo del crédito. Otro comparativo realizado está asociada al método de *credit score*, también utilizado para evaluar riesgo de crédito; en relación con este, se logra determinar que las técnicas *Machine Learning* aportan soluciones más eficientes a las estimaciones.

Para el caso de la aplicación de métodos de *Machine Learning* en la industria *Fintech,* la literatura académica es poca; si bien estos métodos forman parte del núcleo del desarrollo de este tipo de compañías, el detalle relacionado con la aplicación de estos métodos en el sector ha permanecido oculto a la luz de público, como lo indica Majid Bazarbash (Bazarbash, 2019); este autor plantea unas discusiones relativas acerca de las fortalezas y debilidades del uso de *Machine Learning* en la evaluación de los créditos; primero, hace un acercamiento a las técnicas más utilizadas; posteriormente discute los principales retos que tiene en el análisis de riesgo creditico en las *Fintech,* y como éstas pueden llegar a mejorar la inclusión financiera a través del uso de información de fuentes de datos poco comunes para la evaluación del riesgo creditico y, por último,

cómo podría facilitar esto la evaluación automatizada y de bajo costo de los clientes pequeños, que de otro modo, quedarían fuera del mercado de crédito tradicional.

Bazarbash destaca las garantías, las perspectivas de ingresos y los cambios generales de comportamiento crediticio, como los elementos de la evaluación crediticia que pueden verse en mayor medida beneficiados por la ampliación de fuentes de datos y el uso de *Machine Learning*; y, finalmente, expone la necesidad de ahondar en la calidad de la información, con el fin de que no se incurra en nuevos problemas, ya que en este tipo de análisis el uso de fuentes poco comunes hace que el tamaño de la muestra deba ser mayor que en las formas tradicionales de calificación crediticia, lo cual podría llegar a generar ruido y llevar a la exclusión financiera de solicitantes solventes.

Los estudios mencionados previamente permiten evidenciar que la temática propuesta en el presente trabajo es pertinente; se encuentra numerosa literatura asociada, exponiendo en algunos casos, métodos particulares de *Machine Learning* e, igualmente, se hallan distintas aplicaciones de dichos métodos y frecuentemente se hace contraste con las metodologías convencionales, tal como se propone en este ejercicio.

2.2 Marco teórico

Inteligencia artificial (IA) y Machine Learning (ML)

El reciente crecimiento y popularidad que ha tenido el uso de la inteligencia artificial y el *Machine Learning*, ha llevado a que estos dos conceptos se usen con frecuencia de forma indiscriminada; por lo tanto, es preciso indicar la diferencia entre ambos.

La Inteligencia Artificial es comúnmente definida como la forma de imitación de la inteligencia de los seres humanos mediante el uso de algoritmos, a través de un ordenador o cualquier otro sistema informático; Alan Turing, considerado como el padre de la inteligencia artificial, se aproximó a una primera definición en 1950, de la siguiente manera: “Si hay una máquina detrás de una cortina y un humano está interactuando con ella (por cualquier medio, por ejemplo, audio o vía escribiendo, etc.) y si el humano siente que está interactuando con otro humano, entonces la máquina es artificialmente inteligente” (Turing, 1950, citado por Joshi, 2020, p. 4).

En el año 1956 el informático estadounidense John McCarthy, hizo mención por primera vez del término *inteligencia artificial,* en la conferencia *Dartmouth Summer Research Project on Artificial Intelligence,* llevada a cabo en la universidad Dartmouth College en Estados Unidos.

Así pues, una inteligencia artificial estará dada por el hecho de que su comportamiento sea completamente indistinguible al de un humano; con base en dicha premisa, Russel y Norvig (2019) diferenciaron cuatro tipos de IA:

1. Sistemas que piensan como humanos: aquellos que realizan una imitación del funcionamiento del sistema nervioso mediante redes neuronales artificiales, con el fin de que las máquinas perciban la información, razonen y entreguen una respuesta que permita tomar decisiones y resolver problemas.
2. Sistemas que actúan como humanos: como lo vendrían siendo los robots y androides, que realizan procesos de forma similar a los humanos y de forma más eficiente.
3. Sistemas que utilizan la lógica racional: estos sistemas son capaces de percibir el entorno en el que se encuentran e intentan imitar de forma racional el comportamiento humano y actuar con base en esa información, como los sistemas expertos.
4. Sistemas que actúan racionalmente: buscan imitar de forma racional, es decir, en forma ideal el comportamiento humano, por ejemplo, los llamados agentes inteligentes. (Russel y Norvig, 2019, p. 2)

La diferencia fundamental entre IA y *Machine Learning* consiste en que la primera es la capacidad de los sistemas de mostrar un comportamiento racional, es decir, similar al de los seres humanos; mientras que la segunda es una técnica que se utiliza para producir y mejorar ese comportamiento. Por eso, el Aprendizaje Automático (*Machine Learning*) es considerado una rama de la inteligencia artificial, que se refiere a cuando un programa computacional puede aprender a producir un comportamiento que no está explícitamente programado por el autor del programa; por consiguiente, esta rama tiene como objetivo desarrollar técnicas que permitan que las computadoras aprendan de los datos en lugar de aprender de la programación explícita.

Se dice que un agente aprende cuando su desempeño mejora con la experiencia, cuando la habilidad que muestra no estaba presente en sus rasgos de nacimiento; dicho aprendizaje se da con base en tres factores:

- Datos que consume el programa.
- Una métrica que cuantifica el error o alguna forma de distancia entre el comportamiento actual y el comportamiento ideal.
- Un mecanismo de retroalimentación que utiliza el error cuantificado para guiar al programa a producir un mejor comportamiento en los eventos subsiguientes.

Como puede verse, el segundo y tercer factores rápidamente hacen que el concepto sea abstracto y enfatiza profundas raíces matemáticas del mismo, así los métodos de la teoría del aprendizaje automático son esenciales para construir sistemas artificialmente inteligentes (Joshi, 2020) y, a medida que los algoritmos ingieren datos de entrenamiento, se producen modelos más precisos; los cuales dependiendo de la naturaleza del problema a resolver, pueden tener diferentes enfoques de *Machine Learning* basados en el tipo y volumen de los datos.

Tabla 1. Enfoques del aprendizaje automático

Enfoque	Descripción
Aprendizaje supervisado	Consiste en enseñar al algoritmo lo que se quiere y lo que debe aprender a hacer por sí mismo. Ejemplos de uso: detección de anomalías, series de tiempo, estimación de riesgo, detección de fraudes, regresión y clasificación.
Aprendizaje no supervisado	Tiene la intención de clasificar datos con base en patrones o clústeres con base en patrones que se encuentren. Ejemplos de uso: identificar un correo electrónico no deseado, segmentación de mercado.
Aprendizaje por refuerzo	Está basado en la rama conductual de la psicología, en la que el aprendizaje se da cuando una acción particular es seguida por algo deseable, haciendo más probable que el ser humano repita dicha acción. En este tipo de aprendizaje, el algoritmo (modelo) es retroalimentado por el análisis de los datos, conduciendo hacia el mejor resultado; el sistema no se encuentra entrenado con el conjunto de datos de ejemplo, sino que aprende a través de prueba y error, como, por ejemplo: el elegir entre dos opciones en un videojuego y observar las consecuencias.
Aprendizaje Profundo	Hace uso de redes neuronales en varias capas sucesivas para aprender de los datos de forma iterativa y es útil cuando se busca aprender patrones de datos no estructurados. Estas redes neuronales están diseñadas para asemejarse al funcionamiento del cerebro humano, con el fin que las computadoras pueden ser entrenadas para atacar problemas mal definidos y realizar abstracciones; el Aprendizaje Profundo y las redes neuronales se utilizan mucho en

	el reconocimiento de voz y de imágenes como, por ejemplo, aplicaciones de visión de computadora.

Fuente: Elaboración propia, 2020.

Para la realización de esta investigación se hará uso del método de aprendizaje supervisado, mediante el cual se pueda detectar a los clientes, con un comportamiento riesgoso con base en la información histórica de estos; a continuación, se dará explicación de los modelos utilizados.

Regresión logística

La regresión logística es un método estadístico usado para resolver problemas de clasificación binaria, que usa una función logística para modelar una variable dependiente, que puede ser binomial (o binaria), ordinal o multinomial.

La clase binomial se ocupa de situaciones en las que el resultado observado para una variable dependiente solo puede tener dos tipos posibles; para efectos prácticos, es la que el resultado generalmente se expresa como "0" o "1", ya que esto conduce a una interpretación más sencilla (Hosmer & Lemeshow, 2000), estos números pueden representar, por ejemplo, "muerto" o "vivo" o "pasa el examen" o "no pasa el examen" y, en el caso específico de riesgo de crédito, "default" o "no default"; por su parte, la regresión logística multinomial se usa en situaciones en las que se pueden tener tres o más tipos de resultados posibles (por ejemplo, "enfermedad A", "enfermedad B", "enfermedad C", que no tiene un orden; y la regresión logística ordinal se usa cuando las variables dependientes están ordenadas.

Los modelos de regresión logística se utilizan principalmente como una herramienta de análisis e inferencia de datos, donde el objetivo es comprender el papel de las variables de entrada para explicar el resultado (Hastie, Tibshirani & Friedman, 2009).

La regresión logística es una técnica multivariante de dependencia, debido a que se basa en estimar la probabilidad de que ocurra un evento en función de la dependencia de otras variables; así pues, con la regresión logística, se puede explicar la probabilidad de que un cliente entre o no en default, en función de un conjunto de variables explicativas, es decir, relacionadas con su comportamiento crediticio.

Las variables independientes que utiliza la regresión logística pueden ser continuas o categóricas, pero a diferencia de la regresión lineal, la regresión logística tiene como variable dependiente variables que toman un número limitado de categorías, en lugar de un resultado continuo.

La función logística lleva en el núcleo de su método la función sigmoide; esta función es una curva en forma de S, que puede tomar cualquier número real y dar como resultado cualquier número entre cero y uno.

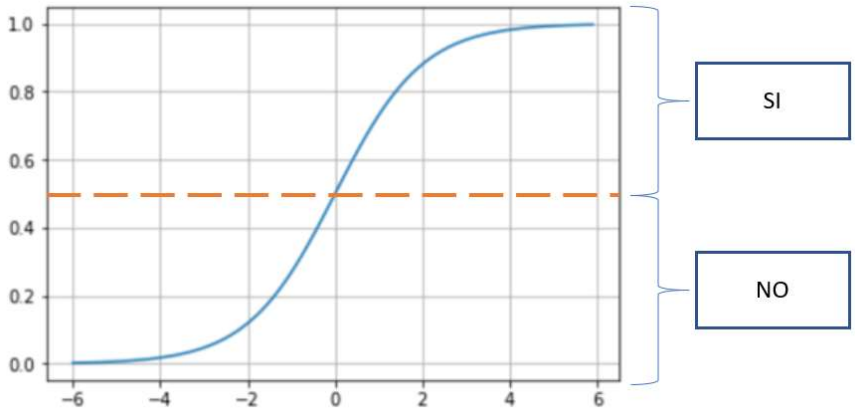
En el caso de *Machine Learning,* la función sigmoide relaciona la variable dependiente con las variables independientes; es una curva que puede tomar cualquier valor entre 0 y 1 y nunca valores por fuera de estos límites, así la ecuación que define la función sigmoidea es:

Ecuación 1. Función sigmoidea

$$f(x) = \frac{1}{1 + e^{-x}}$$

Donde x es un número real. En esta ecuación se puede observar que cuando x tiende a menos infinito, el cociente tiende a cero y cuando x tiende a infinito el cociente tiende a uno, tal como se muestra en la figura 1.

Figura 1. Gráfico de la función sigmoidea y su umbral



Fuente: Elaboración propia, basados en gráfico de Torres (2018)[4]

[4] https://torres.ai/deep-learning-inteligencia-artificial-keras/

Al tomar valores entre 0 y 1, el umbral de probabilidad de la función sigmoide es 0.5; por tanto, si la salida de la función es mayor a 0.5, se puede clasificar el resultado como 1 o "sí", y si es menor que 0.5, se puede clasificar como 0 o "no".

Por ejemplo, si el resultado de la función es 0.75, se puede decir que hay un 75% de probabilidad de que determinado cliente incurra en mora en sus pagos, lo cual lo colocaría en el grupo de categoría positiva.

Para entender la formulación del modelo *logit,* en primer lugar, es necesario partir de la regresión lineal. En estadística la regresión o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X, y un término aleatorio ε. Este modelo puede ser expresado como:

Ecuación 2. Modelo de regresión lineal

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i + \varepsilon$$

Donde:

β_0: es la intersección o término constante de la curva

y: variable dependiente o variable explicada

$x_1, x_2, \ldots, x_{i,}$: variables independientes o variables explicativas

$\beta_1, \beta_2, \ldots, \beta_{i,}$: parámetros que miden la influencia que las variables explicativas tienen sobre la variable dependiente

Si se aplica la función sigmoide en la regresión lineal, se obtiene la siguiente ecuación:

Ecuación 3. Cálculo del valor de p

$$p = \frac{1}{1 + e^{-(\boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i})}}$$

En este modelo, se violan los supuestos de regresión lineal, por ejemplo, los residuos no se pueden distribuir normalmente debido a que la función *logit* es lineal, pero las probabilidades en sí mismas no lo son (Gujarati & Porter, 2009); dado esto, la variable dependiente se debe convertir en una continua que pueda tomar cualquier valor real. Para esto, la regresión logística binomial primero calcula las probabilidades de que ocurra el evento para diferentes niveles de cada variable

independiente, y acto seguido, toma su logaritmo para crear un criterio que sea continuo, como una versión transformada de la variable dependiente; el logaritmo de dichas probabilidades es el *logit* de la probabilidad, el *logit* se define de la siguiente manera:

Ecuación 4. Modelo logit

$$logit(p) = log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$

Por su parte, la regresión lineal proporciona una salida continua, como por ejemplo: conocer la probabilidad de incumplimiento o el precio de una acción; mientras que la regresión logística proporciona una salida discreta, por ejemplo, conocer si un cliente incumplirá o no, o si el precio de una acción subirá o no. Por lo tanto, el modelo logístico expresa la variable dependiente como la ocurrencia o no de un acontecimiento en términos de probabilidad.

Entre los aspectos positivos de esta metodología, se encuentra que es un modelo simple y de fácil interpretación (Moral, 2016), es un proceso liviano desde el punto de vista de los recursos computacionales, además de permitir el uso de múltiples variables aun con pocos casos para cada una de ellas (Domínguez & Aldana, 2001), y permite obtener estimaciones consistentes de la probabilidad de incumplimiento, identificar los factores de riesgo que determinan dichas probabilidades, así como la influencia o peso relativo de éstos sobre las mismas.

Por otro lado, al ser una metodología lineal, no permite resolver directamente problemas no lineales, por ejemplo, en el caso de que la probabilidad tenga forma de U, es decir, se reduzca inicialmente al aumentar una característica y posteriormente suba la probabilidad al continuar aumentando la característica, un modelo logístico no logra reflejar este comportamiento de forma directa, esto obliga a transformar esta característica previamente para que el modelo pueda registrar este comportamiento no lineal, en este caso, es mejor utilizar otros modelos con capacidad polinomial.

Por último, el modelo necesita que la variable objetivo sea linealmente separable, es decir, en los datos deben existir dos "regiones" con una frontera lineal, lo cual se garantiza al dejar su aplicación solo para problemas de respuesta categórica, ya que, en caso de no serlo, el modelo de regresión logística no clasificará correctamente.

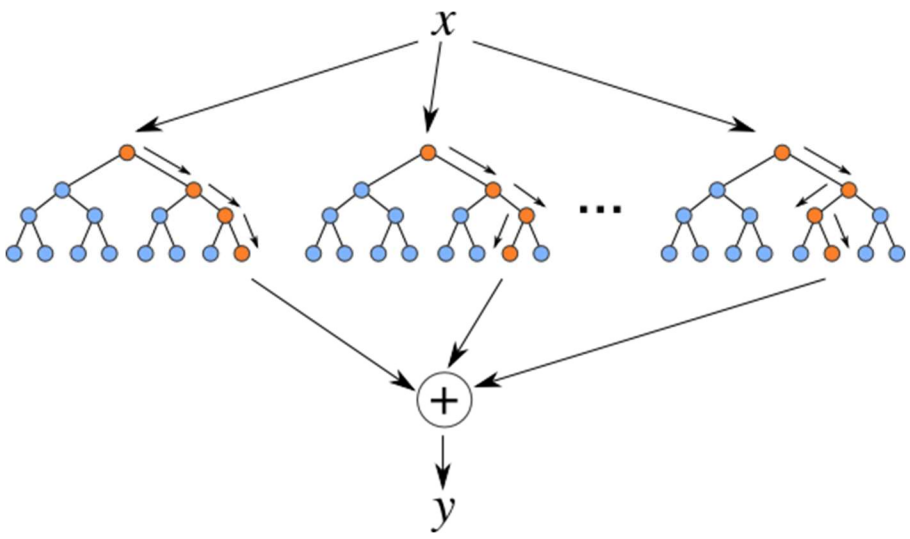
Random Forest

Los modelos de *Random Forest* fueron propuestos inicialmente por Ho (1995, 1998), y generalizados y llamados *Random Forest* por Breiman (1996), los cuales tienen una variedad de aplicaciones, tales como clasificación de imágenes, motores de recomendación y selección de características. También se pueden utilizar para identificar actividades fraudulentas, predecir enfermedades y clasificar a los solicitantes de préstamos.

Son modelos del tipo *ensemble*, esto quiere decir que es una estrategia de combinación de un conjunto de modelos de *Machine Learning,* en donde cada modelo produce una predicción diferente, y las predicciones de los distintos modelos se combinan para obtener una única predicción; en este tipo de modelos se usan varios tipos de criterios: *votación por mayoría, bagging, boosting y stacking.*

Puntualmente para el criterio Bagging, este consiste en combinar varios modelos de *Machine Learning*, y se entrena cada modelo con subconjuntos del conjunto de entrenamiento, formados por muestras aleatorias con repetición.

Figura 2. Diagrama esquemático de un *Random Forest*



Fuente: (Bakshi, 2020) en Gitconected Blog[5]

[5] https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Random Forest es un algoritmo de aprendizaje supervisado bajo la metodología *ensemble,* construido con el criterio *Bagging,* y que como lo indica su nombre, crea un "bosque" de una manera aleatoria. Para esto, crea múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable; en general, mientras más árboles en el bosque se tengan, más robusto es el bosque.

Un *Random Forest* selecciona al azar las observaciones y características, y con esta información construye varios árboles de decisión, luego obtiene el promedio de los resultados (regresión) o votación (clasificación) (Hastie et al., 2009).

En este algoritmo, la aleatoriedad permite que, en lugar de buscar la característica más importante al dividir un nodo, se busque la mejor característica entre un subconjunto aleatorio de características, dando como resultado una amplia diversidad de modelos que generalmente resulta en un mejor modelo; por lo tanto, en *Random Forest*, el algoritmo para segmentar un nodo solo tiene en cuenta un subconjunto aleatorio de las características iniciales; inclusive, se puede lograr que los árboles sean más aleatorios a través del uso adicional de umbrales aleatorios de probabilidad para cada función, en vez de como lo hace un árbol de decisión normal, que es buscando los mejores umbrales posibles.

El método funciona en tres etapas:

- Construir un árbol de decisión para cada base muestral y con ellos obtener un resultado de predicción.
- Realizar una votación para cada uno de los resultados previstos.
- Hacer la selección del resultado de la predicción con más votos como predicción final.

Como ya se mencionó, los *Random Forest* son una colección de árboles de decisión, pero existen varias diferencias entre ambos:

- Si se ingresa un conjunto de datos de entrenamiento con características y etiquetas en un árbol de decisión, formulará un conjunto de reglas que se utilizará para hacer las predicciones.

- A diferencia de los árboles de decisión que pueden sufrir de sobreajuste cuando son muy profundos, los *Random Forest* evitan el exceso de adaptación la mayor parte del tiempo, creando subconjuntos aleatorios de las características, construyendo árboles más pequeños en profundidad y utilizando estos subconjuntos.
- Cada modelo se entrena con subconjuntos del conjunto de entrenamiento y no con toda la información muestral, estos subconjuntos se forman eligiendo muestras aleatoriamente (con repetición) del conjunto de entrenamiento.

Los *Random Forest* tienen entre sus beneficios que son flexibles y fácil de usar para resolver problemas, tanto de clasificación como de regresión, produciendo un buen resultado de predicción debido al número de árboles de decisión que participan en el proceso; además, si hay suficientes árboles en el bosque, el algoritmo no se adaptará al modelo, evitando el sobreajuste, ya que tiene un límite teórico de sobreajuste (Hastie et al., 2009), soportan un gran número de variables y permiten datos omitidos sin que la predicción se vea afectada.

No obstante, entre los aspectos no tan convenientes de este tipo de modelos, se encuentra que este puede ser computacionalmente muy intensivo, ya que una gran cantidad de árboles puede hacer que el algoritmo sea lento, lo cual lo hace de difícil aplicación, en casos en los que el tiempo de respuesta del algoritmo sea prioritario. También es importante destacar que este tipo de modelos es difícil de interpretar en comparación con un árbol de decisión, dado que en un árbol de decisión se puede seguir fácilmente en la ruta del árbol, mientras que el *Random Forest* es un mecanismo predictivo mas no descriptivo.

Support Vector Machine

Los *Support Vector Machine* (SVM) son unos clasificadores discriminatorios que pertenecen al aprendizaje supervisado, los cuales consisten en un algoritmo que define un hiperplano o conjunto de hiperplanos de separación. Fue desarrollado por Benhard Boser (Boser, Guyon & Vapnik, 1992) y se aplica para problemas de clasificación y de regresión.

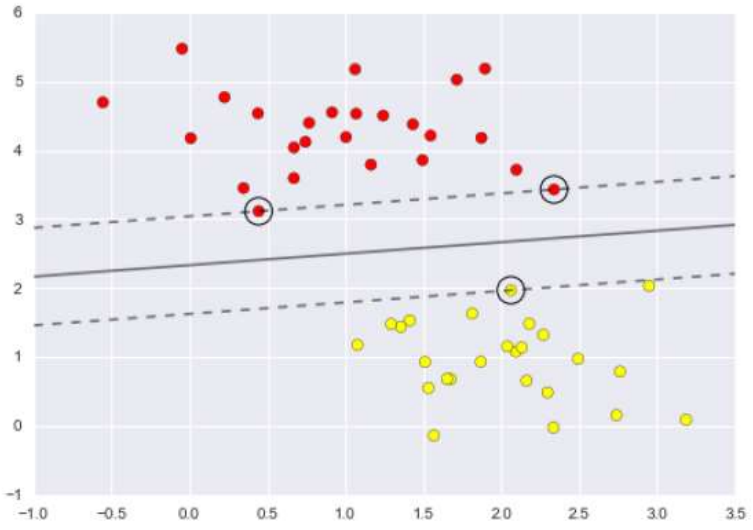
Dado un conjunto de datos de entrenamiento etiquetados, el algoritmo genera un hiperplano óptimo que clasifica los datos en dos partes o planos, lo más ampliamente posible. Posteriormente, se trazan dos líneas paralelas a cada uno de los lados de este, utilizando los vectores de soporte, los

cuales son los datos más cercanos al hiperplano y que son el soporte de estas dos líneas, creando una nueva banda. Como resultado, todos los datos que se encuentran en cada uno de los lados son clasificados igual a los datos correspondientes a ese hiperplano.

Cuando no es posible separar las clases mediante un hiperplano, se dice que no son linealmente separables; para esto, se usan las funciones matemáticas *Kernel*, las cuales agregan espacios para separar los datos en diferentes superficies de más de una dimensión. Los tipos más comunes de funciones *Kernel* son: Polinómica, Perceptrón, Base radial Gaussiana y Sigmoidal.

Los modelos SVM tienen como ventaja que no son propensos al sobreentrenamiento y son generalmente precisos. Entre sus desventajas se encuentra que requieren bastante capacidad computacional y en su entrenamiento deben realizar varias pruebas de funciones *Kernel* y ajuste de parámetros; así mismo, los resultados pueden resultar difíciles de interpretar.

Figura 3. Márgenes y vectores de soporte de un *linear SVM*



Fuente: (VanderPlas, 2016)

Multi-Layer Perceptron (MLP)

Los perceptrones fueron creados por el sicólogo estadounidense Frank Rosenblatt en 1957 (Rosenblatt, 1957), quien se considera uno de los más importante pioneros del campo de la inteligencia artificial; para su idea se basó en los conocimientos biológicos de la época, para definir

cómo este tipo de modelos pueden simular el comportamiento neuronal y aprender con base en la información.

El perceptrón multicapa es un tipo de red neuronal artificial que consta de al menos tres capas de nodos: una capa de entrada, una capa oculta y una capa de salida, en ocasiones es llamada red de retro propagación de capa única oculta, o perceptrón de una sola capa, y en términos coloquiales, como red neuronal vainilla (Hastie et al., 2009).

La capa de entrada es la que recibe la información para ser procesada y la capa de salida realiza la predicción o clasificación; las capas ocultas que están entre las capas de entrada y salida, realizan todo el procesamiento computacional de la red neuronal, su cantidad es un número arbitrario y están entrenadas con el algoritmo de retro propagación (Abirami & Chitra, 2020).

En el modelo *Multi-layer Perceptron* cada nodo es una neurona que utiliza una función de activación no lineal, (excepto en los nodos de entrada) y una llamada retro propagación para el entrenamiento (Frank Rosenblatt, 1963), esto significa que el modelo permite que la información se mueva desde la capa de salida a la de entrada, esto con el fin de que la capa anterior de la red pueda alimentarse también de la información de la capa de entrada y procesar no solo en función de las entradas pasadas sino también futuras (Sinha & Gupta, 2000).

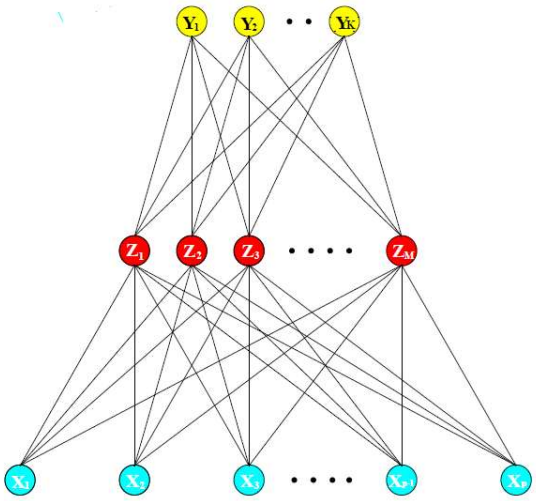
En el modelo MLP, hay K medidas de objetivo (resultados), por ejemplo, para el caso binario K=1, con base en esto las características derivadas (Z, o resultados de la capa oculta) son creadas a partir de combinaciones lineales de las entradas, y con base en combinaciones lineales de las características derivadas, se modela el objetivo (Y), y el resultado de este proceso arroja un vector de salida (T).

La función de salida es una transformación de los vectores de salida T para cada K-esima clasificación y está dada por la ecuación:

Ecuación 5. Función de salida MLP

$$G_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$

Figura 4. Esquema de red neuronal de una sola capa oculta



Fuente: (Hastie et al., 2009)

Entre sus ventajas se puede encontrar que puede separar clases no separables linealmente (Cybenko, 1989); puede funcionar para la estimación de valores único o múltiples, hace un mapeo lineal entre el vector de entrada y el vector de salida correspondiente, en un entorno estático, como por ejemplo, en el reconocimiento de caracteres (Sinha & Gupta, 2000) y es un modelo de red neuronal simple que permite comprender su funcionamiento fácilmente, de ahí su apodo de vainilla.

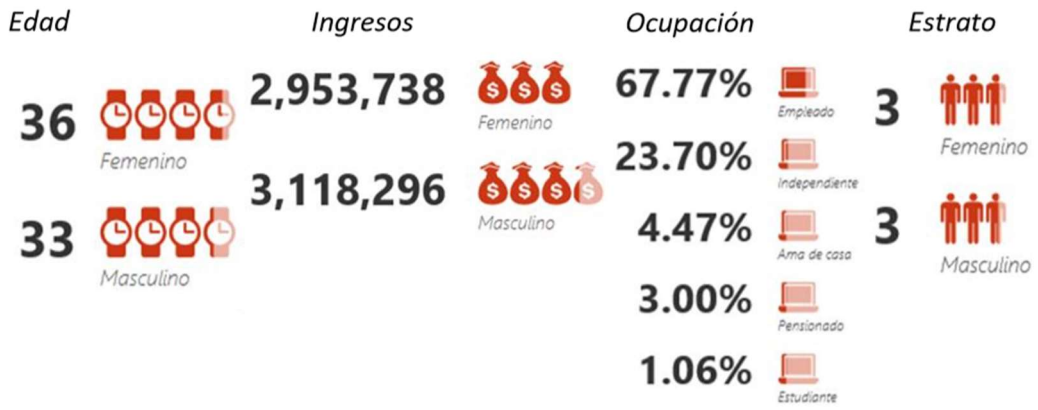
Por otro lado, una de sus principales desventajas es la ausencia de interpretación de resultados, dado que no permite conocer los pesos de las variables usadas. Entre otras desventajas se encuentra la imposibilidad de ajustar las funciones de activación de las neuronas para cada capa; en su lugar, debe ser para todo el modelo; también se encuentra que los resultados de salida de las neuronas pueden oscilar repentinamente de un estado a otro, a medida que hay cambios infinitesimales en los valores de entrada, así como presentar una convergencia de pesos muy lenta cuando hay muchos nodos ocultos en las capas (Davies, 2005).

Como dato curioso, Collobert & Bengio (2004) encontraron que bajo algunos supuestos presentados en su trabajo, los modelos *Multi-layer Perceptron* son equivalentes a los Support Vector Machine, ya que maximizan el margen espacial de los grupos en la capa oculta.

2.3 Caso de estudio

Para efectos de este estudio, se utilizó información proveniente de una compañía colombiana tipo *Fintech* fundada recientemente, la cual se encuentra enfocada en facilitar la cotidianidad de sus clientes colombianos, de estratos entre 2 y 4, con ingresos a partir de un salario mínimo, a través de alianzas con empresas líderes en el sector de la salud, mejoramiento del hogar, deporte, tecnología y movilidad; permitiendo acceder a la financiación de bienes y servicios que mejoran su calidad de vida de forma fácil y ágil, a través de su plataforma propia de originación de crédito, con la cual se evalúan las solicitudes y se otorgan[6] créditos de consumo en el mismo punto de venta en el que son adquiridos.

Figura 5. Perfil promedio de los clientes[7]



Fuente: Compañía de crédito, 2020.

El proceso de otorgamiento de crédito en la compañía es completamente digital, puesto que solo se requiere para su realización la presentación de la cédula del cliente, la lectura de su huella digital, y la firma electrónica de los documentos que instrumentan la obligación crediticia, con la ayuda de un asesor del punto de venta, usando un dispositivo como *tablet* o celular. Mediante el uso de múltiples fuentes de información, tales como centrales de riesgo, información sociodemográfica y de aportes al sistema de Seguridad Social, entre otros, se realiza la evaluación de la solicitud de

[6] Siempre y cuando se cumpla con las políticas de la entidad.
[7] Valores promedio para: edad, ingresos y estrato.

crédito para su rechazo o aprobación. También se tiene en cuenta en el proceso, el control de suplantación y la información de riesgo de lavado de activos.

Los montos financiados van desde $300.000 hasta $15.000.0000 de pesos colombianos y los plazos entre 6 y 60 meses. La edad mínima para acceder a un crédito es de 18 años y la máxima de 74 años.

Como se puede observar en la Figura 1, la compañía tiene como clientes a personas con edad promedio de 35 años, con un nivel de ingreso cercano a los 3 millones de pesos, los cuales tienen en su mayoría un empleo formal y son pertenecientes en promedio, al estrato socioeconómico 3 (medio-bajo).

Dadas las características actuales del proceso de otorgamiento de crédito, el cual es completamente digital y sin la intervención del análisis humano en el momento del estudio, se hace necesario contar con modelos de diversa índole, que sean precisos, que se adapten a la nueva información y aprendan de los mismos datos y de los resultados; perfilándose así como la principal herramienta con la cual llevar a cabo un proceso de otorgamiento robusto y con un control del riesgo adecuado, tomando en cuenta que este debe de estar alineado con el apetito de riesgo de la compañía y acorde a los objetivos de la misma.

Con base en lo mencionado anteriormente, definir cuál modelo de *Machine Learning* presenta las mejores prestaciones para determinar la probabilidad de incumplimiento de los clientes y la toma de decisión de aprobación o rechazo de la solicitud de crédito; esto con el fin de sentar las bases más idóneas para el proceso de crédito de la compañía.

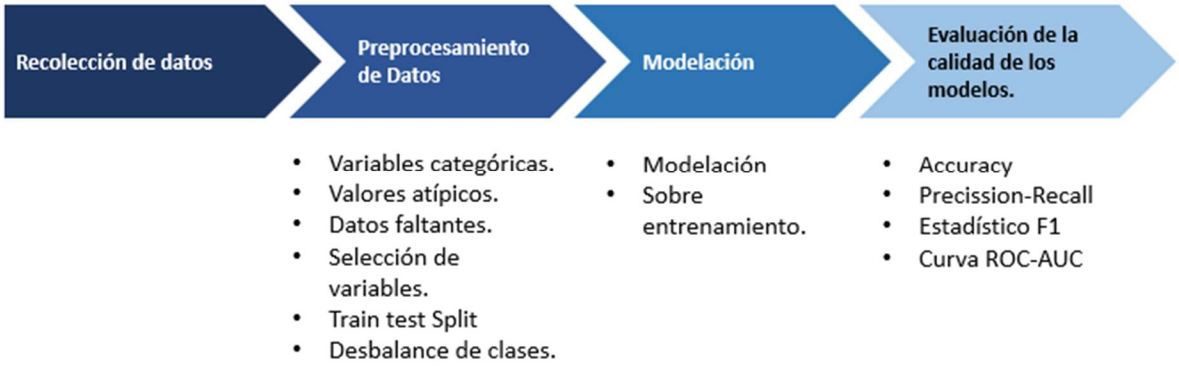
3. Metodología

La estimación del riesgo de crédito, así como la calificación crediticia del consumidor se puede abordar como una tarea de clasificación en la que los clientes reciben una marcación de estado crediticio bueno o malo, acorde con los criterios definidos por el prestamista y su comportamiento de pago; con base a esto, la evaluación puede realizarse bajo un análisis específico del desempeño y la condición del cliente o de una forma masiva por medio de técnicas cuantitativas.

Los métodos de aprendizaje automático hoy en día están fácilmente disponibles y su implementación puede ser fácil de usar, rápida y confiable, y pueden considerarse como competidores de la regresión logística, la cual ha sido la técnica más usada en la calificación crediticia masiva.

Con el fin de lograr una adecuada aplicación de los modelos de *Machine Learning* es necesario definir una cadena de procesos, que garantice que los modelos logren un adecuado nivel de calidad, dicha cadena comprende las etapas de comprensión de la base de datos, preprocesamiento de los datos y modelación y evaluación de la calidad de los modelos; con este fin se plantea el siguiente plan de trabajo metodológico:

Figura 6. Esquema de trabajo



Fuente: Elaboración propia, 2020.

Para el preprocesamiento, modelación y evaluación se utilizó el lenguaje de programación *Python*, a través de la consola de *Jupyter Notebook,* puntualmente, los módulos *Pandas, Numpy, y Sklearn.*

3.1 Recolección de datos

La propensión de los clientes al incumplimiento crediticio se estimó utilizando la información disponible al momento de la originación del crédito, y se compone de variables socioeconómicas, propias de las características del crédito y de información del buró de crédito, el cual indica la situación actual del endeudamiento del cliente. Para efectos de este trabajo, los datos fueron suministrados por la entidad, contenidos en dos bases de datos consolidadas a nivel de cliente, que contienen el historial de pago de las obligaciones desembolsadas durante el año 2019, correspondiente a 15.060 registros y 18 variables, y de una base de datos que contiene toda la información relacionada con la originación del crédito, con 15.096 registros y 180 variables**.**

Usando el historial de pago, se calculó la variable dependiente del modelo, la cual se definió como una variable binaria que presenta el valor de "1" cuando la obligación ha entrado en mora igual o superior a 30 días, al menos una vez después de su desembolso, y "0" cuando no. Esta variable es el discriminante entre los clientes identificados como incumplidos (o en *default*) y los que no. De la base de historial de pagos se tomó únicamente la información de las variables socioeconómicas, inherentes al crédito y centrales de riesgo; la información correspondiente a la identificación del cliente -la cual se considera sensible- y de listas de control de riesgo de lavado de activos y financiación del terrorismo no se usaron, puesto que contienen variables que no son de interés para la investigación.

Una vez fusionadas ambas bases de datos, se realizó una depuración de los registros que no fueron coincidentes a nivel de cliente, y también se eliminó la información de las obligaciones crediticias que no presentaban información financiera en el buró de crédito, ya que el caso de estudio de los clientes sin esta información, considerados en el argot bancario como "no bancarizados" está por fuera del análisis de interés. La base de datos resultante, con la cual se alimentó el plan de trabajo metodológico corresponde a 15.215 registros y 86 variables.

3.2 Preprocesamiento de datos

Las metodologías de aprendizaje automático o *Machine Learning* precisan para su ejecución unos buenos niveles de calidad de datos, así como volúmenes adecuados de observaciones; para garantizar esto, se debe realizar un adecuado preprocesamiento de datos que permita garantizar la

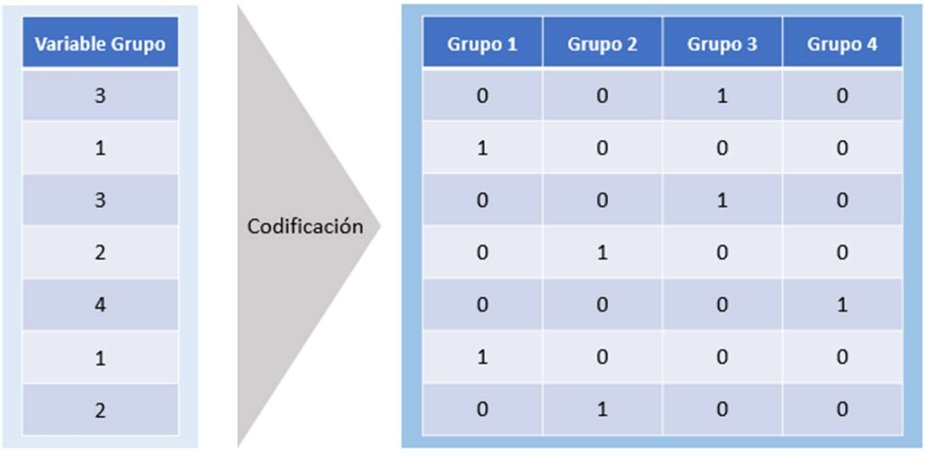
consistencia e integridad de estos; con base en lo anterior, se explican las metodologías de preprocesamiento utilizadas.

3.2.1 Tratamiento de variables categóricas

A partir de esta etapa, es necesario garantizar que la información sea adecuada para los diferentes modelos de estimación propuestos, debido a que tanto los modelos de regresión logística como de *Machine Learning,* precisan que las variables categóricas se encuentren codificadas de forma numérica ya sea ordenada (1, 2, 3, 4) o *dummy* (variable binaria que es 0 cuando no hay pertenencia al grupo, y 1 cuando existe pertenencia al grupo), así, al asignar a estas variables los niveles categóricos, permite tener en cuenta el hecho de que estos pueden tener efectos deterministas separados sobre la variable de respuesta (Draper & Smith, 1998).

En lo referente a las variables categóricas de la base de datos, se realizó un reemplazo, creando la categoría "sin información" para sustituir todos los valores vacíos de las variables categóricas; esto con el fin de no tener una pérdida de información, ya que "La ausencia de información también es información" (Pillai, 2020, notas de clase), y los clientes con campos vacíos podrían ser renuentes a compartir ciertos datos y recoger ese comportamiento puede resultar relevante.

Figura 7. Codificación Dummy



Fuente: Elaboración propia, 2020.

Debido a la alta dispersión de algunas de las variables categóricas, como por ejemplo el amplio número de puntos de venta, se realizó una agrupación en categorías para reducir la dimensionalidad

de dichas variables, al momento de realizar la codificación; para ello se usó un criterio de agrupación de riesgo, mediante el cual se definen grupos que presentan un comportamiento similar con respecto al cálculo de la variable de riesgo conocida como *First Payment Deafult (FPD)*, la cual ocurre cuando los solicitantes de préstamos se retrasan en los primeros[8] pagos de un préstamo (Koç & Sevgili, 2020); los grupos se ordenaron de forma descendente, colocando como primer grupo el de mayor riesgo para dicha categoría.

Figura 8. Ejemplo clasificación FPD

Porcentaje de FPD	Puntos de venta
0% a 0.5%	145
0.5% a 5%	57
5% a 10%	56
10% a 20%	36
de 20% a 25%	9
de 30% a 50%	7
Mas de 50%	3
Total puntos de Ventas	313

Porcentaje de FPD	Profesión
0% a 0.5%	199
0.5% a 5%	35
5% a 10%	43
10% a 20%	38
de 20% a 40%	20
Mas de 40%	7
Total profesión	342

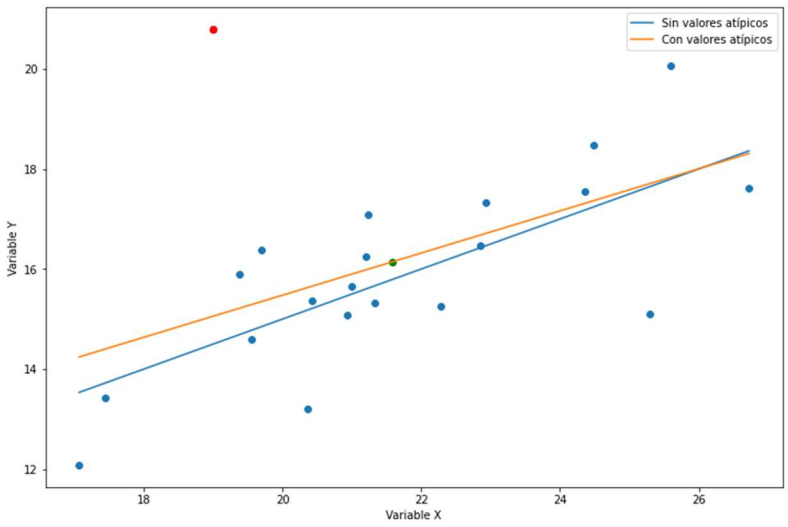
Fuente: Elaboración propia, 2020.

3.2.2 Tratamiento de valores atípicos

Una observación atípica es aquella que parece desviarse sustancialmente de otras observaciones de la muestra analizada, esta puede estar generada por una marcada varianza en la naturaleza de la variable, o por una desviación del procedimiento experimental, ya sea por errores de cálculo o en el registro de la información (Grubbs, 1969). Para el caso puntual de la información usada, solo se encontraron tres observaciones que presentaban valores atípicos, principalmente en las variables de los ingreso y gastos, con valores diez veces mayores al resto del grupo de observaciones y completamente fuera de la población objetivo de los productos de crédito; al considerarse errores en el registro de la información fueron eliminados, ya que estos pueden generar error en la convergencia de los modelos de índole lineal.

[8] Para efectos de este trabajo se tomaron los pagos de los tres primeros meses.

Figura 9. Ejemplo del efecto de un valor atípico en la línea de tendencia



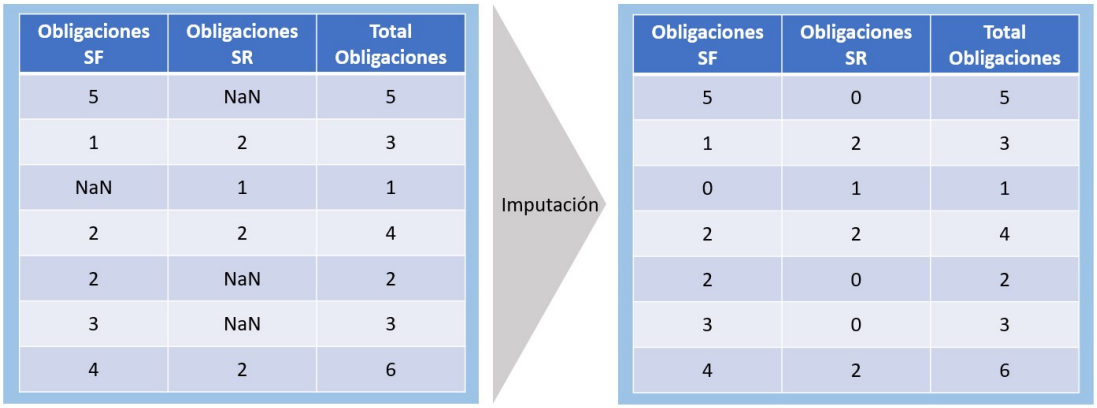
Fuente: Elaboración propia, 2020.

3.2.3 Tratamiento de datos faltantes

Los datos faltantes de índole numérica pueden corresponder a la ausencia de información para ese campo, o a dificultades en la medición de dicho campo; dicha situación puede convertirse en un inconveniente, ya que el número de registros con la información completa es una porción muy reducida de la base de datos. Ante esta situación y la imposibilidad de ignorar esos registros con datos faltantes, se debe definir una técnica que permita llenar los datos faltantes con valores consistentes con el comportamiento de la base de datos, dicha técnica se conoce como imputación (Useche & Mesa, 2006).

Para efectos de esta imputación se decidió usar la estrategia de valor constante, con el fin de que todos los valores vacíos de las variables numéricas que se encuentran en el set de datos sean convertidos a cero; se consideró esta estrategia como adecuada porque la mayor parte de esta información está contenida en variables correspondientes al buró de crédito, donde los clientes podrían tener endeudamiento en el sector financiero, pero no en el sector real o viceversa, por lo que un valor de cero es consistente.

Figura 10. Ejemplo de imputación de valores faltantes



Fuente: Elaboración propia, 2020.

3.2.4 Selección de variables

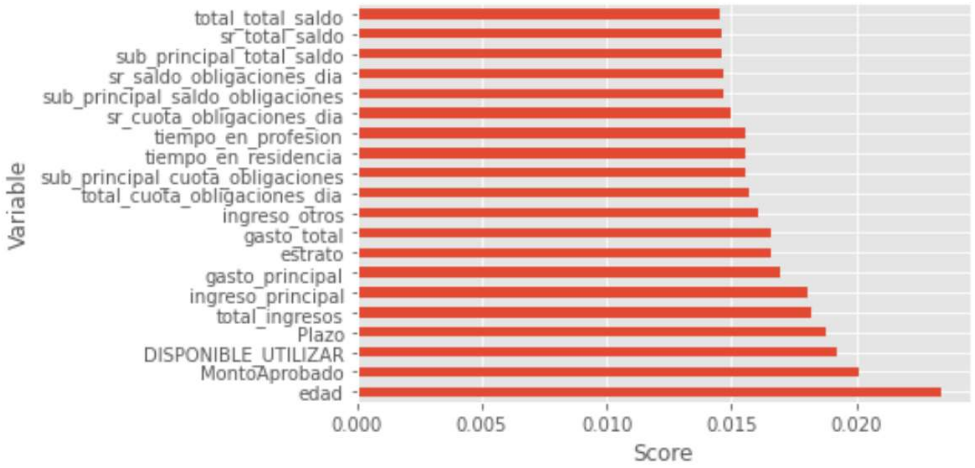
El conjunto de variables independientes que componen la base de datos posee relación con la variable dependiente, desde el punto de vista del negocio, pero no necesariamente presenta un impacto desde el punto de vista estadístico, o aun teniendo un nivel de relevancia estadística, su aporte a la estimación del modelo es marginal. De acuerdo con la teoría de componentes principales (Hotelling, 1933), reducir la dimensionalidad de datos multivariados permite identificar cuáles componentes presentan un mayor aporte a la variable de estudio.

Si bien en este estudio no se realizó una aplicación del análisis de componentes principales, es importante hacer una detección de las variables más relevantes para la estimación del modelo de *default*; para ello, se contempló la aplicación de algunos criterios de selección de variables que brindan información *a priori,* con respecto al conjunto de datos más adecuado a incluir en el modelo, así como permitir el uso de un modelo más parsimonioso, es decir, el que sea más explicativo con el menor número de variables.

Para este estudio, se aplicó la técnica de selección de variables de Árboles Extremadamente Aleatorios (Geurts, Ernst & Wehenkel, 2006), esta técnica de selección de variables es un método de *Machine Learning,* que al igual que los modelos de *Random Forest,* utiliza un subconjunto aleatorio de las características candidatas, seleccionando los umbrales para cada característica candidata al azar y elige como regla el mejor de estos umbrales generados aleatoriamente.

La técnica asigna una puntuación para cada variable, y cuanto mayor sea la puntuación, más importante o relevante es la variable. A continuación, se puede observar el gráfico donde se puntúan las 20 variables seleccionadas para la estimación de los modelos.

Figura 11. Gráfico de puntuación de las variables seleccionadas



Fuente: Elaboración propia, 2020.

Es importante destacar que esta técnica se aplicó en concordancia con la estimación de los modelos; así pues, las 20 variables seleccionadas son aquellas que permiten tener el mejor poder de predicción con el menor número de variables.

3.2.5 *Train test split*

Los algoritmos de *Machine Learning* se basan en la información contenida en las variables de la base de datos, para hacer la estimación de los mejores parámetros que permitan la estimación de la etiqueta de clase; este proceso se conoce como aprendizaje o entrenamiento (*Train*). A partir de dicho entrenamiento se realiza la generalización de un nuevo conjunto de datos, por medio de los parámetros estimados y se asigna una etiqueta de clase, dicha etapa es la que se conoce como predicción o prueba (*Test*) (VanderPlas, 2016).

Acorde con esto, la división de la base de datos en los conjuntos de entrenamiento y prueba es una división aleatoria del conjunto de datos originales, que permite tener un distribución aleatoria de los datos; no obstante, dada la naturaleza de fenómeno de estudio, al existir mayor número de clientes que hacen sus pagos versus el número de clientes que dejan de pagar y, por lo tanto,

materializan para la compañía el riesgo de crédito, es necesario realizar una división de bases por medio de un muestreo estratificado.

A diferencia del muestreo aleatorio, el cual no permite una correcta separación de clases desbalanceadas, el muestreo estratificado asegura que la distribución de clases se mantenga igual, tanto en el conjunto de prueba como en el conjunto de entrenamiento, esto permite entrenar los modelos sin que estos presenten el sesgo hacia una clase en particular (Joshi, 2020).

Para efectos de este trabajo, el conjunto de datos se dividió de forma estratificada en una proporción de 80% para el conjunto entrenamiento y 20% para el conjunto de prueba, como se muestra en la tabla 2. Dicha proporción fue elegida, debido a que la diferencia de distribución de clases y el poder clasificatorio del modelo se beneficiaría de tener un mayor número de observaciones en el conjunto de entrenamiento y, bajo la Ley Potencial de Pareto (Pareto, 1971), empíricamente se indica que en una relación funcional de cantidades, aproximadamente el 80% de las consecuencias proviene del 20% de las causas.

Tabla 2. División del conjunto de datos

Conjunto de datos	Observaciones sin Mora	Observaciones con Mora	Porcentaje sin Mora	Porcentaje con Mora
Base completa	13.111	2.104	86%	14%
Base entrenamiento	10.489	1.683	86%	14%
Base prueba	2.622	421	86%	14%
Porcentaje entrenamiento	80%	80%		
Porcentaje prueba	20%	20%		

Fuente: Elaboración propia, 2020.

3.2.6 Estandarización del conjunto de datos

Debido a que muchas de la técnicas de *Machine Learning* son sensibles a la escala, en la cual están los datos de entrada, cuando las dimensiones se encuentran en escalas diferentes, no son comparables entre sí (Grus, 2015); un ejemplo de ello sería que, al calcular las distancias entre dos variables que miden la misma información, pero en distinta escala, la distancia será diferente, lo cual añade ruido indeseado a los modelos.

La estandarización usada es a partir de la metodología Min-Max, la cual hace que el valor absoluto máximo de cada variable se lleve a la unidad (valor de 1), y a partir de ahí los demás se escalen hasta llegar al cero, este tipo de escalamiento permite que el estimador sea robusto frente a desviaciones estándar muy pequeñas de las variables.

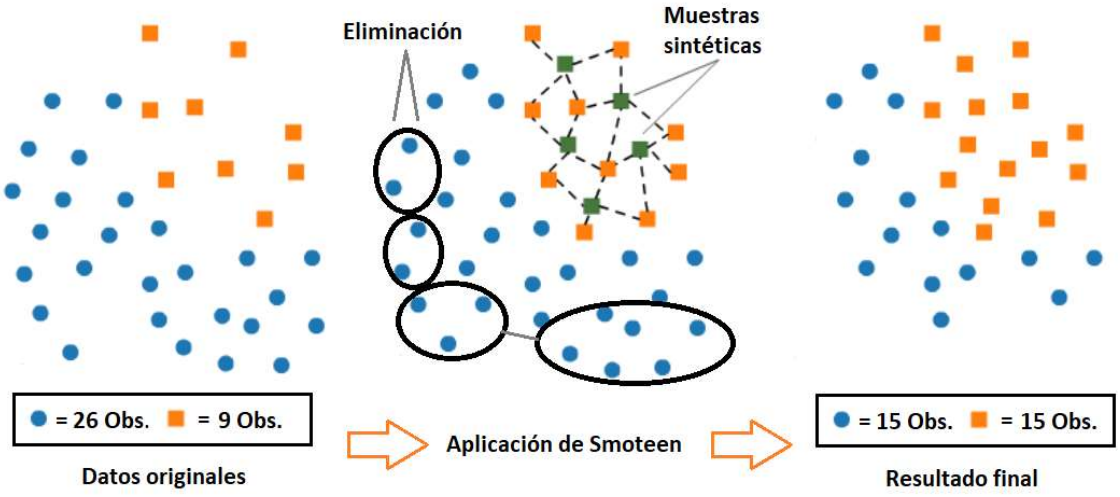
Ecuación 6. Fórmula de *Min-Max Scaler*

$$Xescaler = X\ desvest * (Xmax - Xmin) + Xmin$$

3.2.7 Desbalance de clases

La información de los créditos desembolsados presenta un desbalance natural entre los clientes que han entrado en mora y los que han realizado sus pagos correctamente, presentando una proporción de 6.2 a 1, entrando así en la categoría de datos de clases desequilibrados, los cuales son aquellos en los que una clase está subrepresentada en relación con otra (Mduma, Kalegele & Machuve, 2019). Debido a esta situación, un estimador de *Machine Learning* que prediga en la mayoría o en la totalidad de los casos la clase nula, tendrá un excelente nivel de exactitud, ya que catalogará bien la mayor cantidad de observaciones; no obstante, esto es contrario al objetivo de este trabajo en tanto la clase de interés es la positiva, es decir, cuando el cliente ha presentado mora.

Figura 12. Esquema de funcionamiento de SMOTEEN



Fuente: Elaboración propia, basados en gráfico de Alencar (2018)[9]

[9] https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

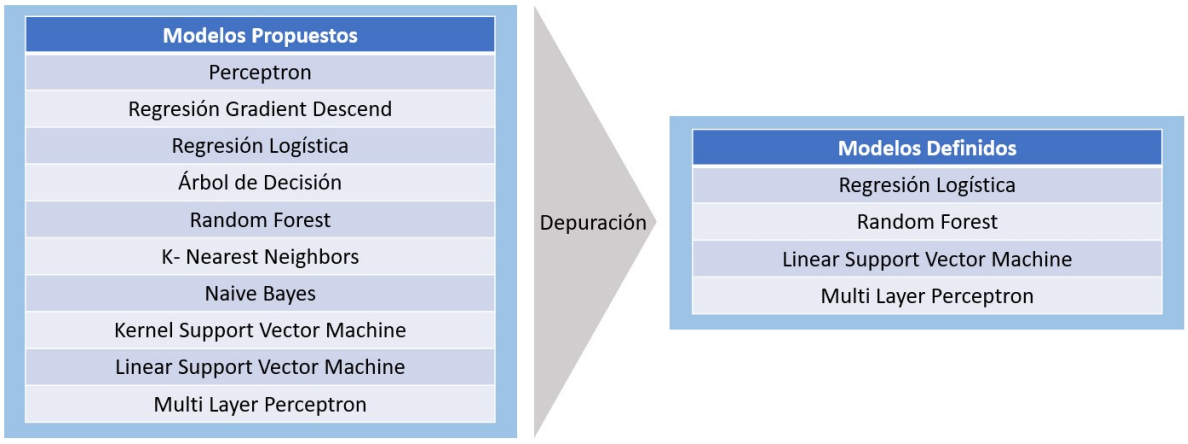
Con el fin de lograr que los algoritmos clasifiquen mejor la clase minoritaria, se hace uso del método de balanceo de muestras SMOTEEN en el conjunto de entrenamiento (Batista, Prati & Monard, 2004); esta técnica combina sobre muestreo con el algoritmo SMOTE y el submuestreo con el algoritmo ENN (*Edited Nearest Neighbor*).

La técnica de sobremuestreo de minorías sintéticas (SMOTE), forma nuevas observaciones sintéticas de la clase minoritarias, por medio de la interpolación de varias observaciones de la clase minoritaria cercanas; mientras que la técnica *Edited Nearest Neighbor* (ENN) elimina cualquier observación de la clase mayoritaria, cuya etiqueta de clase difiera de la clase de al menos dos de sus tres vecinos más cercanos, por lo tanto, la combinación de ambos métodos debería proporcionar una limpieza de datos más profunda y una mejor separación de las clases.

3.3 Modelación

Una vez realizados estos ajustes, se planteó la estimación de diez diferentes modelos de *Machine Learning,* para realizar una comparación en la eficiencia de la estimación de los clientes que van a entrar en mora; de este conjunto inicial de modelos, se tomaron los 4 modelos que mejor se adaptaron a la naturaleza del problema y que potencialmente lograron un mejor desempeño al momento de realizar la tarea de clasificación, estos son: Regresión Logística, *Random Forest, Linear Support Vector Machine* y *Multi-Layer Perceptron.*

Figura 13. Esquema de selección de modelos



Fuente: Elaboración propia, 2020.

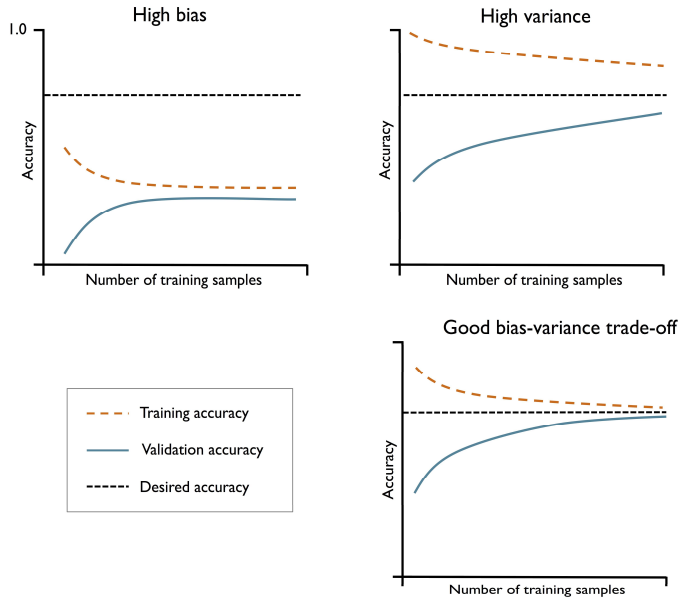
3.3.1 Sobreentrenamiento

Los modelos pueden tender a sobreajustarse en el conjunto de entrenamiento (*overfitting*), es decir, presentan un comportamiento en el cual se memorizan los datos de entrenamiento y no se logra una generalización del modelo, lo cual no permite un buen desempeño a la hora de clasificar las observaciones del conjunto de prueba.

Para poder determinar si el modelo presenta *overfitting* se analizan las curvas de aprendizaje, estas grafican el nivel de precisión en los conjuntos de entrenamiento versus el número de observaciones en el conjunto de datos, un valor de 1 o muy cercano a uno para el conjunto de entrenamiento a lo largo de todo el conjunto de tamaños muestrales, indica que el modelo presenta *overfitting*, también con las curvas de validación se puede detectar sesgo y varianza elevados en los estimadores.

El sesgo elevado indica que el modelo no tiene la capacidad de lograr una buena precisión, por lo que podría ser necesario recolectar más información. Por su parte, una varianza elevada, indica que no se da convergencia entre las estimaciones de los conjuntos de entrenamiento y prueba, lo cual puede indicar un modelo de alta complejidad, con un número muy alto de variables incorporadas o con datos de alta dispersión.

Figura 14. Ejemplo de curvas de aprendizaje



Fuente: Python *Machine Learning* (Raschka & Mirjalili, 2019)

3.4 Evaluación de la calidad de los modelos

En el caso de las clasificaciones, una máquina de aprendizaje debería mostrar una correcta discriminación entre los incumplimientos y los clientes normales. Para realizar la evaluación de la eficiencia de los modelos usaron los indicadores: *Accuracy*, Precisión, *Recall, F1-Score*, y AUC-Curva ROC.

3.4.1 *Accuracy*

La métrica *Accuracy* (exactitud) mide el porcentaje de observaciones en la que el modelo ha acertado; es la métrica más básica de desempeño de los modelos de *Machine Learning,* pero no es la más eficiente. Dado que la métrica es generalizada, se ve afectada por la paradoja de Simpson (Blyth, 1972), la cual es un resultado contraintuitivo o erróneo, que surge cuando los resultados agregados de un análisis muestran una tendencia diferente al comportamiento de los grupos analizados por separado (Wagner, 1982).

Ecuación 7. Fórmula *Accuracy*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde:

TP= Verdaderos positivos (*True Positives*)

TN= Verdaderos Negativos (*True Negatives*)

FP= Falsos Positivos (*False Positives*)

FN=Falsos Negativos (*False Negatives*)

En el caso de tener clases desbalanceadas, si el modelo predice siempre la clase mayoritaria, siempre tendría un excelente nivel de *Accuracy*, esta situación ocurre en el nivel de un análisis de datos descriptivo y puede confundir a un observador ingenuo; por eso es necesario analizar el nivel de *Accuracy* de la mano de otros indicadores, como los que se mencionan a continuación.

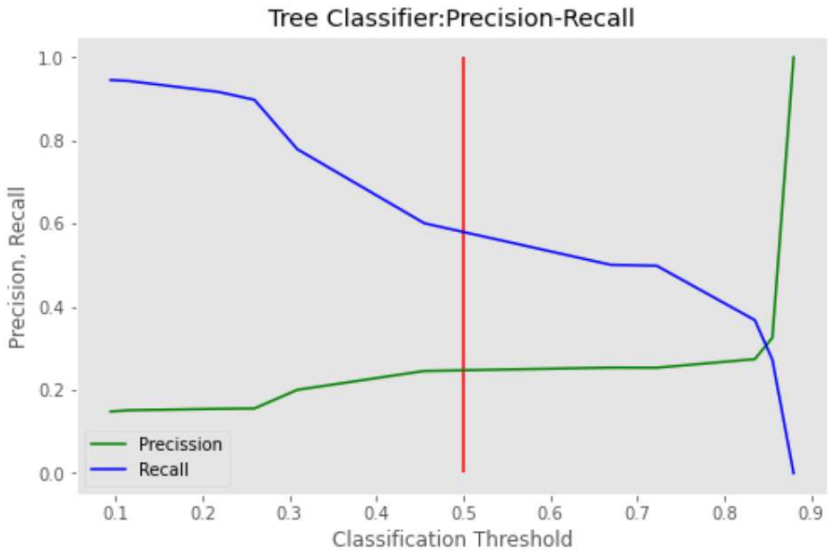
3.4.2 Precisión y *Recall*

Ecuación 8. Fórmulas Precisión y *Recall*

$$Precisión = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

La precisión es la tasa de observaciones positivas identificadas o correctamente clasificadas; el *Recall* o Exhaustividad es la tasa de positivos reales que se identificó o clasificó correctamente. Para todos los modelos de *Machine Learning* existe un *Trade-off*[10] entre estas dos mediciones; es decir, al aumentar la Precisión se reduce el *Recall* y al aumentar el *Recall* se disminuye la Precisión; esta situación es fácil de graficar al hacer los cálculos de estas dos métricas, dado el umbral de precisión (*Threshold*).

Figura 15. Ejemplo de Precisión-*Recall Trade-off*



Fuente: Elaboración propia, 2020.

En la figura 15 se observa que dada la tarea de clasificación binaria, el clasificador realiza el corte en un umbral de 50% de probabilidad de estar en la clase positiva, pero para cada nivel de probabilidad existe un nivel de precisión y *Recall*; así pues, para niveles bajos del umbral se tiene un clasificador ingenuo que clasifica a todas las observaciones en la clase positiva, por lo tanto el

[10] Se define como la situación conflictiva en la cual se debe perder o reducir cierta cualidad a cambio de otra cualidad.

Recall es alto y la precisión muy baja, y para niveles altos del umbral de probabilidad se tendría un clasificador demasiado exigente, el cual no clasificaría casi ninguna observación como de clase positiva, por lo tanto los niveles de precisión son muy altos y los de *Recall* caen fuertemente.

3.4.3 *F1-Score*

Ecuación 9. Fórmula *F1 -Score*

$$F1 = 2\frac{Precission * Recall}{Precision + Recall}$$

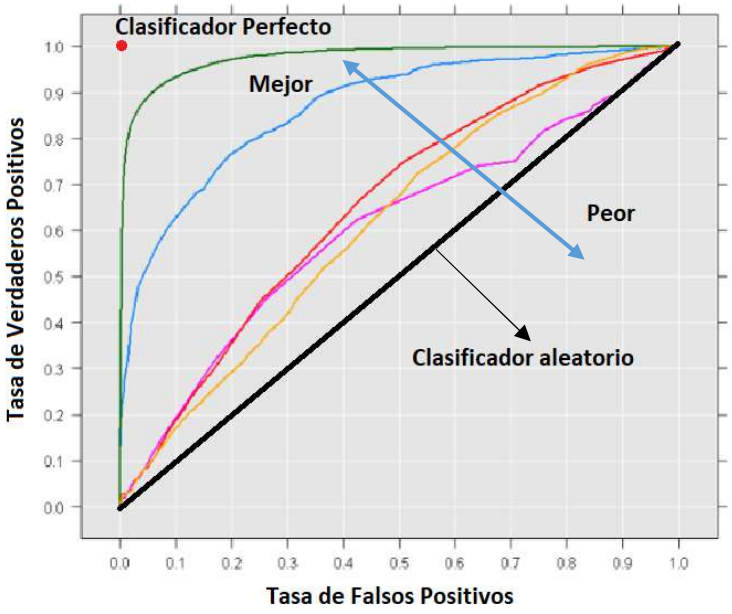
La medida *F1-Score* se define como una media armónica de Precisión y *Recall* (Sasaki, 2007); esta métrica tiene la capacidad de tener en cuenta tanto los falsos positivos como los falsos negativos, y así, un *F1-Score* de 1 indicará una precisión y un *Recall* perfectos, y, por lo tanto, un clasificador perfecto al momento de discriminar las clases. Si bien la intuición de esta métrica es un poco más compleja de entender que el *Accuracy*, Precisión y el *Recall*, esta resulta realmente útil en los casos donde existen clases distribuidas de forma desigual, para la aplicación del método de clasificación de *Machine Learning*, como ocurre en el caso de este trabajo.

3.4.4 Curva ROC-AUC

El último indicador del desempeño de los distintos métodos aplicados para la clasificación de clientes morosos y no morosos es la Curva Característica Operativa del Receptor, ROC (por sus siglas en inglés), el cual es un gráfico que muestra la capacidad de un sistema de clasificación binario para realizar un diagnóstico. Inicialmente el ROC se utilizó en los sistemas de detección de señales, por lo que sus primeras aplicaciones fueron en el campo militar.

Desde el punto de vista interpretativo, el ROC es una descripción del efecto del umbral de detección de la clasificación binaria, mostrando todas las posibles combinaciones de las frecuencia relativas de clasificaciones correctas e incorrectas (Metz, 1978); es decir, la tasa de acierto y la tasa de falsas alarmas (Fawcett, 2006).

Figura 16. Ejemplo de Curva ROC



Fuente: Elaboración propia, basados en gráfico de Chan (2018)[11]

En la línea media del gráfico se encuentra el clasificador aleatorio, el cual no tiene capacidad discriminatoria diagnóstica; los clasificadores debajo de la línea tienen un peor desempeño que una respuesta aleatoria, a medida que se desplaza la curva hacia el cuadrante superior izquierdo, se va obteniendo un mejor clasificador, hasta llegar al clasificador perfecto, el cual tiene una tasa de verdaderos positivos del 100% y una tasa de falsos positivos de 0%.

Esta interpretación gráfica va de la mano con la métrica AUC, la cual indica el área bajo la curva ROC del clasificador; esta métrica indica que, a un mayor valor, más área hay debajo de la curva, es decir, más se acerca la curva ROC al clasificador perfecto.

[11] https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

4. Resultados

Con base en la metodología expuesta, el objetivo del análisis de los modelos cuyos resultados se mostrarán en este apartado, es establecer cuál es el más adecuado para la estimación del riesgo de crédito en la compañía objeto de estudio, de acuerdo con su público objetivo y apetito de riesgo. El modelo a elegir debe hacer una discriminación categórica entre los clientes con mayor probabilidad de mora de los que no, según determinado umbral de probabilidad, entre "buenos" y "malos" y mantener un equilibrio adecuado en dicha estimación; puesto que un modelo que clasifique a todos los clientes como riesgosos no serviría para el desarrollo del negocio, ya que no se harían desembolsos de crédito, mientras que un modelo que no muestre ningún cliente como riesgoso sería un peligro para la estabilidad financiera de la entidad; el equilibrio en este aspecto garantiza un correcto riesgo-retorno y un mejor nivel de atención del público objetivo de la entidad. Las métricas de desempeño de un modelo adecuado deben ser suficientes para que sea elegible para la toma de decisiones, frente a la alternativa de no usar un modelo, es decir, aprobar los créditos aleatoriamente.

A continuación, se presentan los resultados obtenidos tras la selección de variables relevantes usadas para la estimación de los modelos, así como de la estimación de estos. Con este fin, se hace una revisión de las curvas de aprendizaje, para determinar el correcto equilibrio en el ajuste del modelo y evitar la sobre especificación (*overfitting*); e, igualmente, las cifras correspondientes a las diferentes métricas de calidad definidas en el apartado anterior y el análisis de las curvas Precisión-Recall y AUC-ROC derivadas de dichos estimadores.

Tras el uso de la técnica de selección de variables de Árboles Extremadamente Aleatorios, propuesta por Geurts et al. (2006), del subconjunto aleatorio de las características candidatas, las 20 mejores variables para la construcción del modelo son las indicadas en la tabla 3, donde se pueden observar que las variables están compuestas de cuatro variables socio-demográficas que son: la edad, el estrato socio-económico (donde está ubicada la vivienda del cliente, el tiempo en esa residencia y el tiempo en la profesión); y dos variables correspondientes al crédito desembolsado, la cuales son: el monto aprobado y el plazo al cual se va a diferir su pago.

Tabla 3. Variables usadas para la modelación

índice	Variable	Score
0	edad	0.023409
1	MontoAprobado	0.019836
2	Disponible_utilizar	0.019131
3	Plazo	0.018872
4	total_ingresos	0.018048
5	ingreso_principal	0.017869
6	gasto_principal	0.016917
7	gasto_total	0.016503
8	estrato	0.016367
9	ingreso_otros	0.016187
10	sub_principal_cuota_obligaciones	0.015874
11	tiempo_en_residencia	0.015613
12	total_cuota_obligaciones_dia	0.015530
13	tiempo_en_profesion	0.015255
14	sr_cuota_obligaciones_dia	0.015170
15	sr_total_saldo	0.014801
16	sr_saldo_obligaciones_dia	0.014695
17	sub_principal_saldo_obligaciones	0.014609
18	total_total_saldo	0.014559
19	sr_participacion_deuda	0.014470

Fuente: Elaboración propia, 2020.

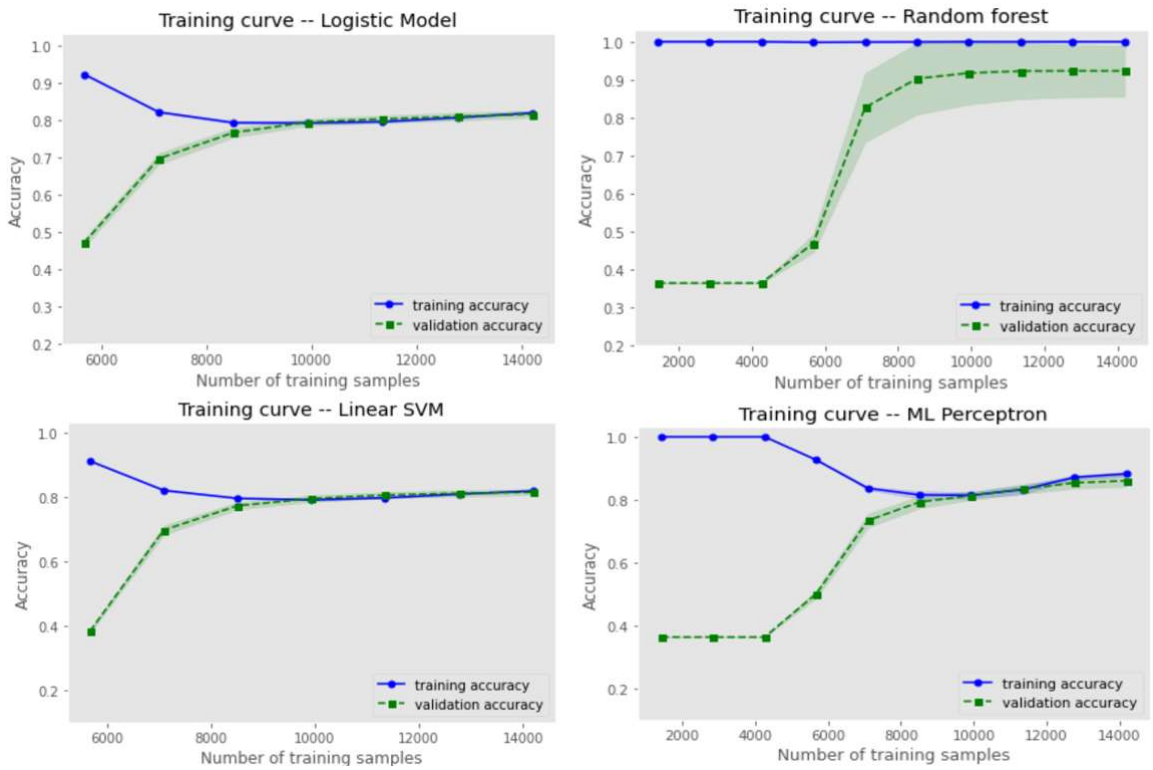
Los dos bloques restantes de variables están dados por seis variables que proveen información de los ingresos y gastos de cada cliente, y ocho variables obtenidas del buró de crédito, las cuales indican principalmente el número de obligaciones y los saldos en los cuales se es titular del crédito, y el número y total del saldo de obligaciones en las que se es codeudor o fiador.

Este número de variables, como se ha indicado en la metodología, son las que mostraron un nivel adecuado de consistencia en la estimación de los modelos de *Machine Learning* propuestos, ya que el conjunto inicial de variables era superior a 80, lo cual implica una tarea mucho más compleja de recolección de información y de procesamiento computacional.

Para proceder a analizar los resultados de los modelos propuestos para realizar la tarea de clasificación binaria, entre los clientes que presentan mora y los que no, se inició por el comportamiento de estos en sus curvas de entrenamiento. Como se puede observar en las gráficas, los modelos Regresión logística, *Linear Support Vector Machine* y *Multi-layer Perceptron*

presentaron una adecuada convergencia entre los conjuntos de entrenamiento y prueba para los diferentes tamaños de muestras; no obstante, el modelo *Random Forest* parece tener *overfitting,* lo cual es un comportamiento documentado y esperado (Hastie et al., 2009); con base en esto y en lo que se detallará más adelante en las métricas de evaluación del modelo, se consideró que este modelo es también consistente.

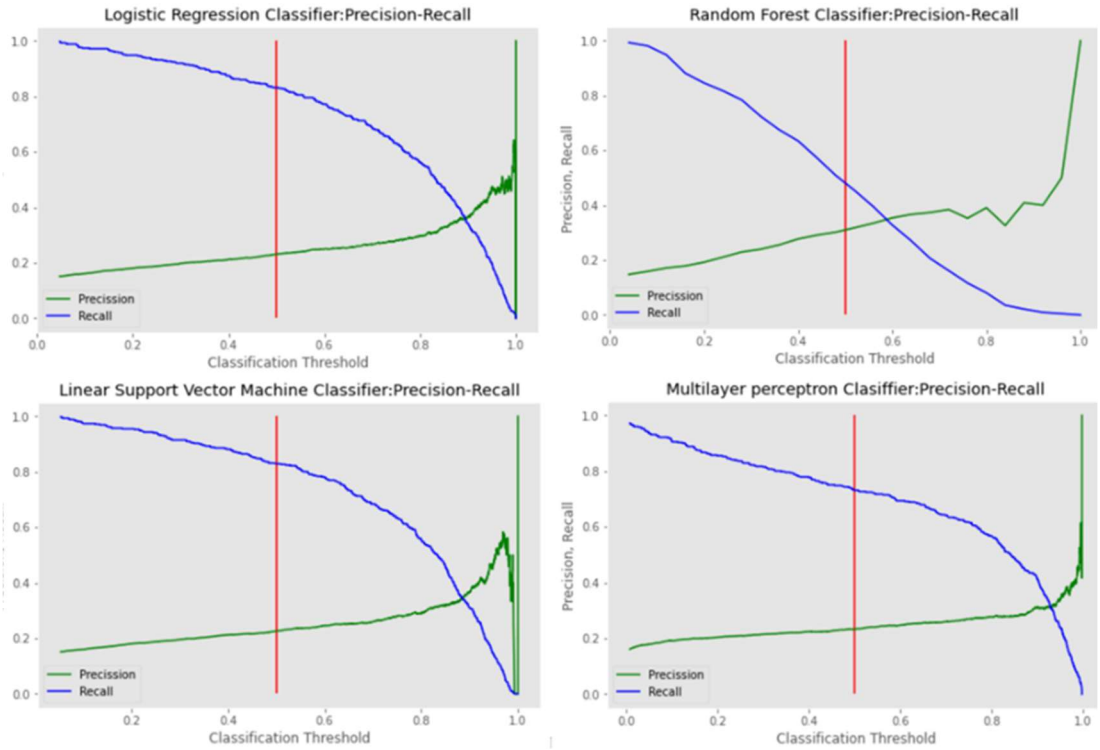
Figura 17. Curvas de entrenamiento de los modelos



Fuente: Elaboración propia, 2020.

Acto seguido, se puede observar que los modelos propuestos presentan un comportamiento muy similar en el *Trade-off* de Precisión y *Recall*, donde se evidencia una amplia separación de ellos para umbrales de probabilidad bajos; situación que se mantiene hasta umbrales de probabilidad de 80% o superior, lo cual indica que, dada la naturaleza de los datos, los clasificadores pueden presentar dificultades para lograr una discriminación de clases. Es importante destacar que la decisión de los modelos utilizados está tomada con un umbral de probabilidad de 50%.

Figura 18. Curvas Precisión-*Recall*-*Trade*-off de los modelos



Fuente: Elaboración propia, 2020.

El comportamiento del estimador se evalúa sobre el conjunto de datos de prueba, el cual representa el 20% de la base de datos; a continuación, se presentan las métricas resultantes de la estimación de los modelos de *Machine Learning* seleccionados:

Regresión logística

Se evidenció que su nivel de *Accuracy* es bajo, es decir, no logra predecir correctamente el 41% de los datos, pero al observar el valor del *Recall,* se puede observar que predice correctamente el 83% de los casos de mora, siendo uno de los modelos que presenta un mayor valor en esta métrica. No obstante, el marcado *Trade-off* entre precisión y *Recall* indica que el número de casos nulos (no morosos) correctamente predicho es bajo. Como se puede observar también, el nivel de precisión es del 23%; a pesar de esto, el modelo se encuentra en los límites de un clasificador adecuado, según lo indica el AUC-ROC, el cual es del 76%.

Tabla 4. Métricas de evaluación de los modelos

Modelo	Accuracy	Precisión	Recall	F1	AUC
Regresión logística	59%	23%	**83%**	36%	**76%**
Random Forest	**77%**	**30%**	51%	**38%**	74%
Linear SVM	58%	22%	**83%**	35%	**76%**
ML Perceptron	**63%**	23%	73%	35%	74%

Fuente: Elaboración propia, 2020.

Random Forest

Este modelo presenta el mejor nivel de *Accuracy*, puesto que logra predecir correctamente el 77% de los datos y el número de casos nulos (no morosos) correctamente predicho; es el mejor entre los modelos evaluados, como se puede observar en el nivel de precisión de 30%. Su *Recall* es el menor de todos los modelos, ya que solo predice correctamente el 51% de los casos de mora, mostrando la existencia del *Trade-off* entre precisión y *Recall,* a pesar de esto, el modelo presenta la mejor métrica F1, lo cual indica que es el modelo más balanceado entre las clases. El área bajo la curva AUC-ROC está ligeramente debajo del valor mínimo esperado de 75%, no obstante, se considera que este modelo tiene un comportamiento aceptable.

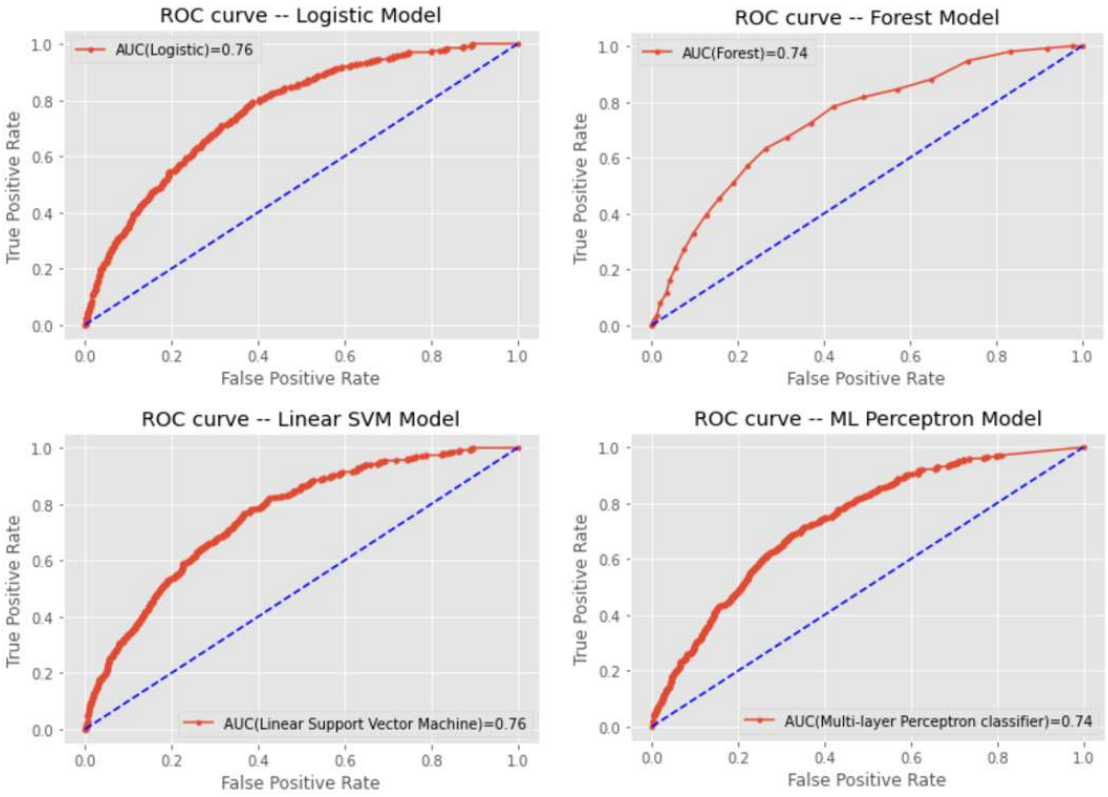
Linear Support Vector Machine

Dado que es un modelo lineal, este modelo presenta un comportamiento similar a la regresión logística; no obstante, tiene el menor *Accuracy* de todos, pues logra predecir correctamente el 58% de los datos, el número de casos nulos (no morosos) correctamente predicho es el más bajo de los modelos evaluados, con un 20% de precisión; no obstante, al igual que la regresión lineal, predice la mayor parte de los casos de mora con un *Recall* de 83% y el AUC-ROC está en un 75%, por lo que se considera que este modelo tiene un comportamiento aceptable.

Multi-layer Perceptron

Este modelo se muestra más equilibrado en sus métricas, logrando predecir correctamente el 63% de los datos; presenta un nivel de precisión 23% y un *Recall* de 73%; el AUC-ROC está en un 74%, ligeramente debajo del valor mínimo esperado de 75%; no obstante, se considera que este modelo tiene un comportamiento aceptable.

Figura 19. Curvas ROC de los modelos



Fuente: Elaboración propia, 2020.

Con base en los resultados obtenidos, se puede observar que los modelos de *Machine Learning* se perfilan como una alternativa valiosa para la estimación de riesgo de crédito, ya que los modelos presentados tienen un comportamiento similar a la regresión logística, el cual es el método más ampliamente utilizado por el sector financiero colombiano.

Sin embargo, de acuerdo con el apetito de riesgo de la entidad, se puede evaluar cuál modelo resulta mejor para sus objetivos, un modelo que identifique exhaustivamente posibles morosos, pero que disminuya el número de desembolsos, o un modelo que identifique mejor los no morosos, pero que pueda dejar pasar clientes con un comportamiento de riesgo y puedan engrosar la cartera vencida, o un modelo equilibrado que valore de manera similar ambos tipos de clientes.

5. Conclusiones

- La comparación de los modelos básicos de *Machine Learning* respecto al modelo de regresión logística, para la estimación del riesgo de crédito en una cartera de consumo, permite concluir que tienen un desempeño similar y que se perfilan como competidores de la metodología tradicional, con el valor agregado de que son teóricamente más recientes y tienen un potencial de refinamiento importante, ya que día a día se están creando nuevas metodologías de *Machine Learning*.
- Basándose en un clasificador equilibrado como el objetivo principal de una política de riesgo de crédito, los modelos de *Random Forest* tienen un mejor nivel de precisión que la regresión logística.
- Es posible concluir que la idoneidad de los modelos estará dada por el modelo, cuyo desempeño esté más acorde al apetito de riesgo de la entidad, y que sea mejor para sus objetivos de negocio y niveles de exposición de riesgo deseados.
- La incorporación de modelos predictivos en el proceso de análisis de crédito de una entidad tipo *Fintech* adquiere mayor relevancia, porque además de la necesidad primordial de gestionar el riesgo crediticio, también se requiere velocidad y oportunidad de respuesta por la forma como está definido su tipo de negocio, sin la intervención humana en la toma de la decisión; y también porque se necesitan modelos eficientes en términos de costo, puesto que normalmente los montos desembolsados son bajos. Estos modelos, al usar múltiples variables, dadas las características del negocio con la participación de los aliados comerciales, requieren que reflejen la interacción entre distintas variables y se encuentren constantemente actualizados; por lo tanto, es conveniente considerar *Machine Learning.* Así mismo, estas entidades se perfilan como futuras entidades vigiladas por la Superintendencia Financiera, dependiendo de su crecimiento futuro, de cambios en la regulación y de posibles alianzas con entidades vigiladas y, en consecuencia, es requisito que se encuentren alineados con las prácticas y exigencias regulatorias de estas, entre las cuales se encuentra la adopción de modelos de riesgo.
- Para la implementación de modelos de *Machine Learning* en la empresa objeto de estudio, es importante tener en cuenta que además de contar con personal y/o consultores externos

expertos en esta materia, el proceso debe iniciar con conocimiento del negocio, que permita antes del modelamiento, un entendimiento de los datos y, que posterior a su despliegue, haya constante retroalimentación para robustecer la capacidad predictiva y el tratamiento que se les da a los datos. No es recomendable que estos dos roles los ejerza quien está encargado del modelamiento, porque se pierde la especialidad de cada uno, pero sí es vital que haya coordinación constante entre ambos.

- Se puede establecer que una combinación adecuada de información para realizar la estimación de riesgo de crédito de un cliente de cartera consumo está compuesta por variables sociodemográficas, información inherente al crédito, información sobre su capacidad económica y el número de obligaciones y saldos que tengan en otras obligaciones externas a la entidad. Utilizar únicamente información sociodemográfica, sin tener en cuenta comportamiento crediticio, puede resultar en modelos de baja capacidad predictiva y no es recomendable para la entidad objeto de estudio, dadas las características de riesgo alto de sus clientes objetivo.
- La consecución de variables significativas para la estimación del riesgo y su correcto tratamiento previo a la incorporación en los modelos de *Machine Learning,* es un asunto crítico para garantizar la calidad y significancia de estos; se deben implementar políticas y controles que garanticen la calidad de la captura de las variables que sean relevantes en los modelos de riesgo. También es importante contar con máquinas de procesamiento computacional que tengan capacidades acordes al volumen de los datos y de los algoritmos del modelo a utilizar. En este caso de estudio, el *Random Forest* presentó un tiempo de procesamiento adecuado, comparado con los demás modelos usados.

Referencias

Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. In *Advances in Computers* (1st ed., Vol. 117, Issue 1). Elsevier Inc. https://doi.org/10.1016/bs.adcom.2019.09.007

Alencar, R. (2018). *Python notebook using data from Porto Seguro's Safe Driver Prediction*. Kaggle. https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets

Bakshi, C. (2020). *Random Forest Regression*. Levelup.Gitconnected. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Basel Committee on Banking Supervision (1999). Principles for the Management of Credit Risk - final document. *Basel Committee on Banking Supervision*, *2000*(July 1999), 4. https://www.bis.org/publ/bcbs54.pdf

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bazarbash, M. (2019). FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk. *IMF Working Papers*, *19*(109), 1. https://doi.org/10.5089/9781498314428.001

Blyth, C. R. (1972). On Simpson ' s Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association*, *67*(338), 364–366. http://www.jstor.org/stable/2284382

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). Training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, *August*, 144–152. https://doi.org/10.1145/130385.130401

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.1023/A:1018054314350

Chan, C. (2018). *What is a ROC Curve and How to Interpret It*. Display R Blog. https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Collobert, R., & Bengio, S. (2004). Links between Perceptrons, MLPs and SVMs. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 177–184. https://doi.org/10.1145/1015330.1015415

Cuenca, J. P. (2019). *Propuesta de Modelo para evaluación de Riesgo de Crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito la Merced* (Issue November) [Universidad Católica de Cuenca]. https://www.researchgate.net/profile/Juan_Cuenca5/publication/337480778_Propuesta_de_modelo_de_machine_learning_para_la_evaluacion_de_riesgo_de_credito_utilizando_algoritmos_de_prediccion_para_la_Cooperativa_de_Ahorro_y_Credito_La/links/5dda977b458515dc2f

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Davies, E. R. (2005). Machine Vision. In *Pattern Recognition Letters* (Tercera Ed). Elsevier. https://doi.org/10.1016/B978-0-12-206093-9.X5000-X

Domínguez, E., & Aldana, D. (2001). Logistic regression: An example of its use in Endocrinology. *Revista Cubana de Endocrinología*, *21*(1).

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis, 3rd Edition. In *John Wiley & Sons* (Tercera). Wiley & Sons, Inc. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471170828.html

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples.

Technometrics, *11*(1), 1–21. https://doi.org/10.1080/00401706.1969.10490657

Grus, J. (2015). *Data Science from Scratch First Principles with Python* (M. Beaugureau & M. Yarbrough (eds.); Primera ed.). O'Reilly.

Gujarati, D. N., & Porter, D. C. (2009). *Econometría* (J. M. Chacón (ed.); Quinta ed.). McGraw-Hill.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Segunda ed). Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Ho, T. K. (1995). Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *832 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *20*(8), 832–844.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression* (A. . C. Noel & N. I. Fisher (eds.); Segunda ed.). Wiley & Sons, Inc.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(7), 498–520. https://doi.org/10.1037/h0070888

Joshi, A. V. (2020). *Machine Learning and Artificial Intelligence* (Springer (ed.); 1st ed., Issue 1). Springer Nature Switzerland AG. https://doi.org/10.1021/ac00025a742

Koç, U., & Sevgili, T. (2020). Consumer loans' first payment default detection: A predictive model. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(1), 167–181. https://doi.org/10.3906/elk-1809-190

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, *40*(13), 5125–5131. https://doi.org/10.1016/j.eswa.2013.03.019

Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and

techniques for student dropout prediction. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-014

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Moral, I. (2016). Modelos de regresión: Lineal simple y regresión logística. *Revista Seden*, *14*, 195–214. https://www.revistaseden.org

Norvig, S. R. and P. (2019). Artificial Intelligence A Modern Approach 2nd Ed. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).

Pareto, V. (1971). *Manual of Political Economy: A Critical and Variorum Edition* (A. Montesano, A. Zanni, L. Bruni, J. S. Chipman, & M. McLure (eds.); Primera ed.). Oxford University Press.

Pillai, N. S. (2020). *Curso DS4A*. Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow* (Segunda ed.). Marcombo. https://books.google.com.co/books?id=c6a2wgEACAAJ

Rosenblatt, F. (1957). The Perceptron - A Perceiving and Recognizing Automaton. In *Report 85, Cornell Aeronautical Laboratory*. Cornell Aeronautical Laboratory.

Rosenblatt, Frank (1963). Principles of Neurodynamics. *The American Mathematical Monthly*, *70*(5), 586. https://doi.org/10.2307/2312103

Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, 1–5. http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Sinha, N. K., & Gupta, A. M. (2000). Soft Computing and Intelligent Systems. In N. K. Sinha & A. M. Gupta (Eds.), *Soft Computing and Intelligent Systems* (Primera Ed, Issue Ml). Elsevier. https://doi.org/10.1016/B978-0-12-646490-0.X5000-8

Superintendencia Financiera (1995). Capítulo II: Reglas relativas a la gestión del riesgo crediticio. In *Circular Externa 100* (pp. 1–31). https://www.superfinanciera.gov.co/descargas?com=institucional&name=pubFile1000224&downloadname=cap02riesgocrediticio.doc

Torres, J. (2018). *Deep learning introducción práctica con Keras (primera parte)*. Torres.Ai. https://torres.ai/deep-learning-inteligencia-artificial-keras

Trujillo, D. F. (2017). *Aplicación de Metodologías Machine Learning en la Gestión de Riesgo de Crédito*. Universidad Politécnica de Madrid.

Useche, L., & Mesa, D. (2006). Una introducción a la imputación de valores perdidos. *Terra Nueva Etapa*, *22*(31), 127–151. http://www.redalyc.org/articulo.oa?id=72103106

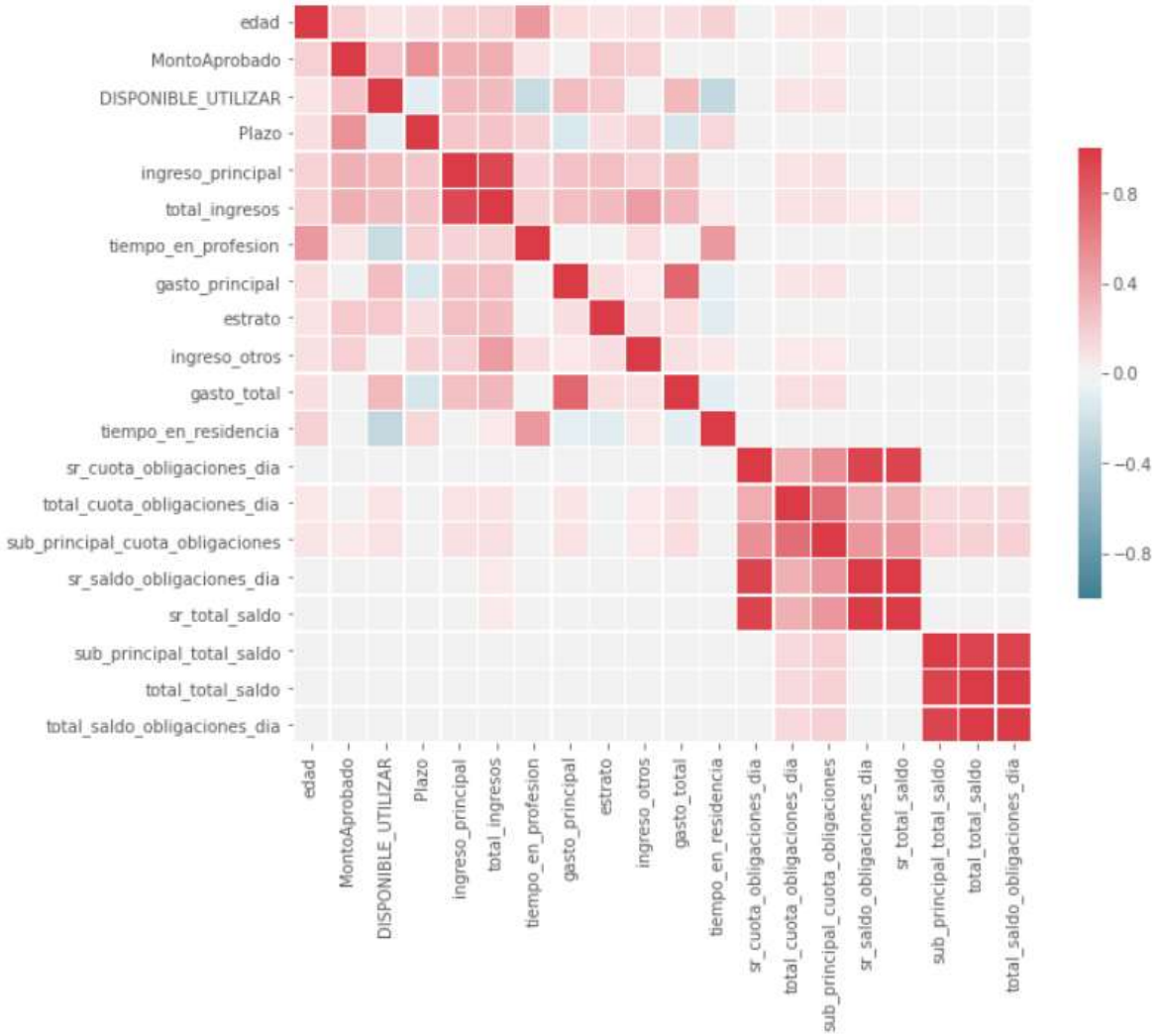
VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. In *O'Reilly*. http://shop.oreilly.com/product/0636920034919.do%0Ahttps://jakevdp.github.io/PythonDataScienceHandbook/05.01-what-is-machine-learning.html

Wagner, C. H. (1982). Simpson's Paradox in Real Life. *The American Statistician*, *36*(1), 46–48. http://www.jstor.org/stable/2684093

Anexos

Anexo A. Estadísticas descriptivas

Mapa colorético de la correlación entre variables dependientes



Fuente: Elaboración propia, 2020

Estadísticos descriptivos de las variables dependientes

	edad	MontoAprobado	DISPONIBLE_UTILIZAR	Plazo	ingreso_principal	total_ingresos	tiempo_en_profesion	gasto_principal	estrato	ingreso_otros
count	15.215	15.215	15.215	15.215	15.215	15.215	15.215	15.215	15.215	15.215
mean	36	1.938.397	508.160	24	2.539.251	3.213.913	5	293.343	3	420.855
std	13	1.724.222	783.557	14	2.339.267	2.766.309	7	366.040	1	844.641
min	18	783	0	3	0	0	0	0	1	0
25%	25	749.925	159.091	12	1.200.000	1.600.000	0	100.000	2	0
50%	32	1.400.000	285.714	24	2.000.000	2.500.000	2	200.000	3	0
75%	45	2.522.000	543.500	36	3.000.000	4.000.000	8	350.000	3	500.000
max	75	15.000.000	20.900.000	60	80.000.000	80.100.000	21	10.000.000	6	26.000.000

	gasto_total	tiempo_en_residencia	sr_cuota_obligaciones_dia	total_cuota_obligaciones_dia	sub_principal_cuota_obligaciones	sr_saldo_obligaciones_dia	sr_total_saldo	sub_principal_total_saldo	total_total_saldo	total_saldo_obligaciones_dia
count	15.215	15.215	15.215	15.215	15.215	15.215	15.215	15.215	15.215	15.215
mean	461.094	6	119	746	632	897	907	26.417	30.387	29.984
std	592.766	8	1.921	5.210	3.697	9.525	9.527	593.406	630.663	630.077
min	0	0	0	0	0	0	0	0	0	0
25%	150.000	1	0	82	65	0	0	304	425	405
50%	300.000	3	26	286	258	55	55	2.422	2.885	2.837
75%	514.500	10	127	748	680	466	488	12.617	15.199	15.036
max	17.000.000	21	234.711	424.370	361.615	1.081.340	1.081.340	44.936.154	44.936.154	44.936.154

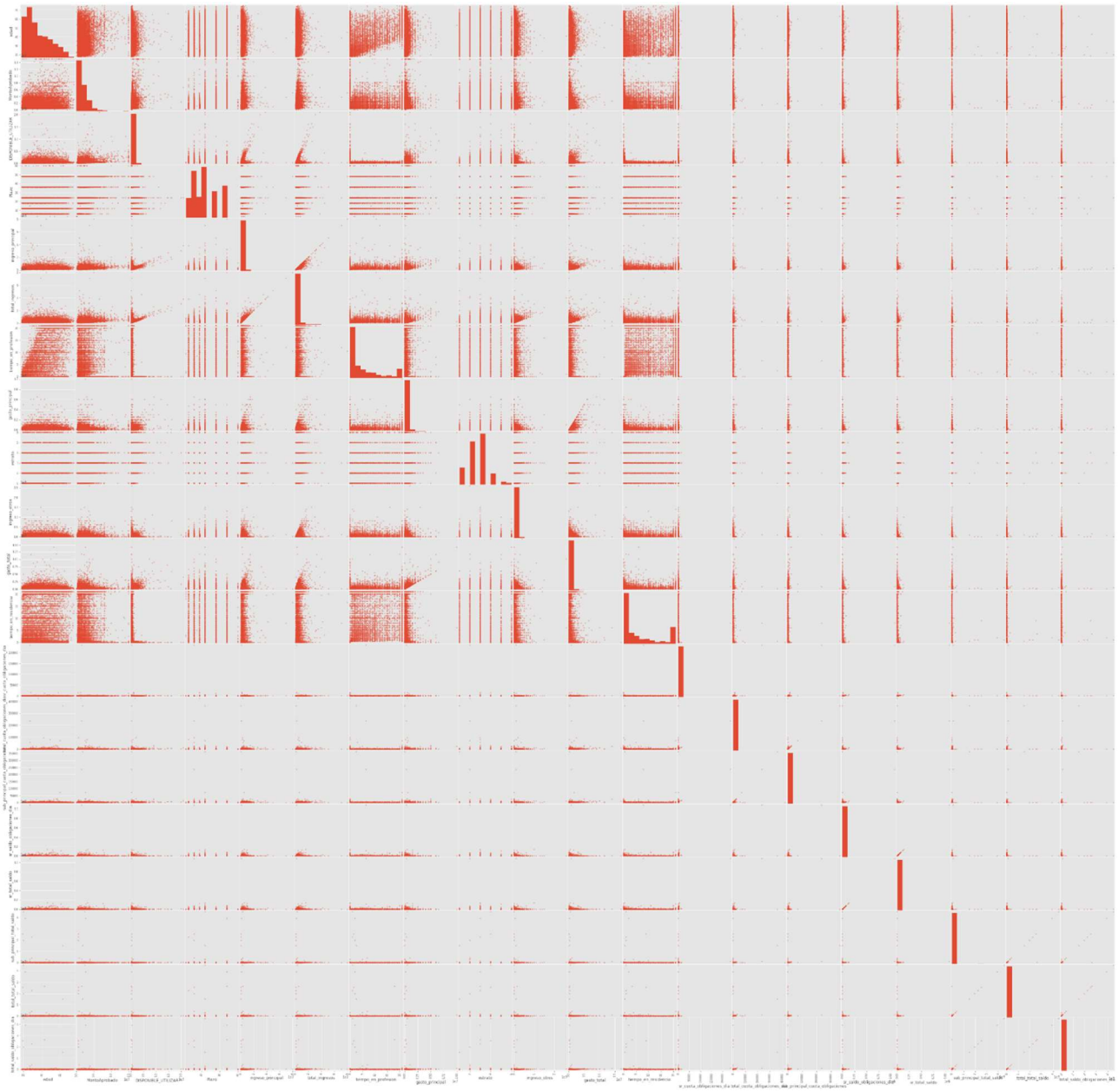
Fuente: Elaboración propia, 2020.

Matriz de correlaciones de las variables dependientes

	edad	MontoAprobado	DISPONIBLE_UTILIZAR	Plazo	ingreso_principal	total_ingresos	tiempo_en_profesion	gasto_principal	estrato	ingreso_otros	gasto_total	tiempo_en_residencia	sr_cuota_obligaciones_dia	total_cuota_obligaciones_dia	sub_principal_cuota_obligaciones	sr_saldo_obligaciones_dia	sr_total_saldo	sub_principal_total_saldo	total_total_saldo	total_saldo_obligaciones_dia
edad	100%	19%	8%	10%	17%	18%	48%	12%	8%	9%	11%	18%	0%	6%	7%	1%	1%	-1%	0%	0%
MontoAprobado	19%	100%	25%	52%	35%	36%	8%	1%	22%	18%	1%	1%	1%	4%	4%	2%	2%	1%	1%	1%
DISPONIBLE_UTILIZAR	8%	25%	100%	-9%	31%	30%	-24%	29%	22%	3%	31%	-27%	1%	7%	8%	2%	2%	1%	1%	1%
Plazo	10%	52%	-9%	100%	23%	25%	18%	-15%	11%	18%	-17%	14%	-1%	0%	0%	-1%	-1%	0%	1%	1%
ingreso_principal	17%	35%	31%	23%	100%	92%	16%	26%	28%	19%	27%	3%	1%	7%	9%	3%	3%	1%	1%	1%
total_ingresos	18%	36%	30%	25%	92%	100%	18%	28%	29%	47%	32%	5%	1%	8%	10%	4%	4%	1%	1%	1%
tiempo_en_profesion	48%	8%	-24%	18%	16%	18%	100%	2%	0%	12%	3%	48%	0%	2%	3%	1%	1%	0%	0%	0%
gasto_principal	12%	1%	29%	-15%	26%	28%	2%	100%	10%	5%	76%	-8%	1%	7%	8%	2%	2%	1%	1%	1%
estrato	8%	22%	22%	11%	28%	29%	0%	10%	100%	10%	11%	-12%	0%	3%	4%	1%	1%	2%	2%	2%
ingreso_otros	9%	18%	3%	18%	19%	47%	12%	5%	10%	100%	9%	6%	1%	4%	5%	3%	3%	0%	1%	1%
gasto_total	11%	1%	31%	-17%	27%	32%	3%	76%	11%	9%	100%	-9%	2%	9%	11%	3%	3%	2%	1%	1%
tiempo_en_residencia	18%	1%	-27%	14%	3%	5%	48%	-8%	-12%	6%	-9%	100%	0%	-2%	-1%	-1%	-1%	-1%	-1%	-1%
sr_cuota_obligaciones_dia	0%	1%	1%	-1%	1%	1%	0%	1%	0%	1%	2%	0%	100%	37%	52%	94%	94%	1%	1%	1%
total_cuota_obligaciones_dia	6%	4%	7%	0%	7%	8%	2%	7%	3%	4%	9%	-2%	37%	100%	71%	35%	35%	13%	13%	13%
sub_principal_cuota_obligaciones	7%	4%	8%	0%	9%	10%	3%	8%	4%	5%	11%	-1%	52%	71%	100%	50%	50%	18%	17%	17%
sr_saldo_obligaciones_dia	1%	2%	2%	-1%	3%	4%	1%	2%	1%	3%	3%	-1%	94%	35%	50%	100%	100%	2%	2%	2%
sr_total_saldo	1%	2%	2%	-1%	3%	4%	1%	2%	1%	3%	3%	-1%	94%	35%	50%	100%	100%	2%	2%	2%
sub_principal_total_saldo	-1%	1%	1%	0%	1%	1%	0%	1%	2%	0%	2%	-1%	1%	13%	18%	2%	2%	100%	94%	94%
total_total_saldo	0%	1%	1%	1%	1%	1%	0%	1%	2%	1%	1%	-1%	1%	13%	17%	2%	2%	94%	100%	100%
total_saldo_obligaciones_dia	0%	1%	1%	1%	1%	1%	0%	1%	2%	1%	1%	-1%	1%	13%	17%	2%	2%	94%	100%	100%

Fuente: Elaboración propia, 2020.

Diagramas cruzados de puntos de las variables dependientes



Fuente: Elaboración propia, 2020.