# Modelado de riesgo crediticio: una aplicación con técnicas de balanceo de datos

Isabella Hernández Ochoa

Trabajo de grado


Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Matemáticas Aplicadas
Santiago de Cali
26 de febrero de 2024

Director: Andrés F. Amador Rodríguez
Co-director: Julián Gil González

# Agradecimientos

Quiero expresar mi profundo agradecimiento a Dios por brindarme la fortaleza y la guía necesaria para completar este proyecto académico. Agradezco a todos los docentes que han sido parte de mi trayectoria universitaria, y de manera especial, a mis profesores de tesis, Andrés Amador y Julián Gil. Su dedicación y conocimientos han sido pilar fundamental en el desarrollo de este proyecto, del cual he obtenido valiosos conocimientos para mi crecimiento profesional.

Doy gracias a mi familia que me ha brindado su apoyo desde el inicio de mi travesía por las Matemáticas Aplicadas. Quiero agradecer especialmente a mis padres, Liliana y Ruben, y a mi hermano Isaac, por su apoyo incondicional y aliento constante en cada etapa de mi vida. Sin duda alguna su presencia ha sido la fuente de mi fortaleza y éxito.

# Índice

# –1–
# Introducción

Las entidades financieras son aquellas que se dedican de manera habitual y profesional a la captación de recursos del público con el fin de realizar operaciones activas de crédito, inversión en valores o cualquier otra actividad financiera, con el objetivo de obtener beneficios [1]. Teniendo en cuenta la naturaleza de las actividades y las condiciones con las que operan las entidades financieras estas se encuentran expuestas a diversos tipos de riesgos bancarios dado que, cada una de sus operaciones contiene de manera implícita o explicita la incertidumbre.

Este estudio se centrará en el riesgo crediticio, el cual se refiere a la posibilidad de que un agente económico que ofrece servicios de crédito incurra en pérdidas debido al incumplimiento de pagos por parte de sus clientes en sus obligaciones contractuales o potenciales. Con base en esta premisa, se empleara el modelo de diversas técnicas, como el modelo de Regresión Logística, Random Forest y XGBoost, con el objetivo de comparar la técnica que mejor asegure la viabilidad de otorgar o denegar un préstamo. Se realizará este análisis utilizando una base de datos fija que presenta una particularidad significativa: el desbalance entre las clases, por lo que será importante utilizar técnicas como Random Undersampling, Random Oversampling, SMOTE y Tome-Links con el fin de minimizar el impacto de dicho desbalance.

## 1.1. Objetivos

### 1.1.1. Objetivo general:

Comparar el desempeño del modelo de regresión logística, Random Forest y XGBoots haciendo uso de una base de datos fija con el fin de determinar el riesgo crediticio asociado a un cliente considerando diferentes estrategias de balanceo de datos.

### 1.1.2. Objetivos específicos:

- Realizar un proceso de limpieza y procesamiento de la base de datos existente para mejorar la calidad y la integridad de los datos.
- Implementar diversas técnicas de balanceo de datos como random undersampling, random oversampling, SMOTE y Tome-Links sobre la base de datos original, caracterizada por la presencia de un desbalance entre las clases.
- Aplicar y evaluar el desempeño del modelo logístico y de diferentes modelos de Machine Learning haciendo uso de una misma base de datos.
- Evaluar y comparar los diferentes modelos de clasificación para seleccionar el más efectivo en la evaluación de riesgo crediticio.

## 1.2. Planteamiento del problema

El Comité de Supervisión Bancaria de Basilea fue creado en 1974 por el Banco de Pagos Internacional y es una organización mundial que reúne a las autoridades de supervisión bancaria. Su objetivo es reforzar la estabilidad de los sistemas financieros mediante la mejora de los mecanismos de supervisión internacional a través de enfoques y estándares comunes.

El Acuerdo de Basilea III, la más reciente iniciativa del Comité de Basilea, forma parte de una serie de acuerdos y estándares internacionales que han surgido desde sus inicios con el propósito de regular la actividad bancaria. Este acuerdo fue creado con el fin de darle respuesta a la crisis financiera mundial que se presentó entre el 2008 y el 2009, la cual fue ocasiona por el colapso de la burbuja inmobiliaria en Estados Unidos en el año 2006 [2].

La crisis mencionada anteriormente surgió debido a la disminución de los precios de los inmuebles y a la falta de cumplimiento de los deudores en el pago de sus préstamos hipotecarios. Posteriormente, al no conocerse la magnitud de los pasivos netos fuera del balance de las instituciones financieras, se paralizó el crédito y esto provocó una crisis sistémica [3]. En esta crisis, se vio afectado el sistema financiero estadounidense extendiéndose y perjudicando las economías de otros países [4].

Como consecuencia de la crisis y los fuertes impactos en los sistemas financieros y las economías mundiales, se realizó un periodo de evaluación de las regulaciones financieras a nivel global para determinar los cambios que debían realizarse frente a las regulaciones establecidas con anterioridad por el Comité de Basilea. Es por ello que, surge el acuerdo de Basilea III, en el cual se establecieron medidas más rigurosas en cuanto al capital mínimo que los bancos deben mantener para absorber pérdidas y protegerse contra riesgos. Sin embargo, también enfatiza en cuanto a la transparencia del sector bancario para prevenir

crisis financieras.

A pesar de que se tengan acuerdos y estándares internacionales de regulación bancaria, se presentan situaciones como la del Silicon Valley Bank (SVB), el cual quebró el pasado 10 de marzo de 2023. Debido a la falta de depósitos para hacerle frente a la subida de tasas de interés de Estados Unidos a causa de la inflación [5]. De hecho, las acciones de SVB se desplomaron un 80 % desde su máximo histórico a fines de 2021 y se estima que el 97 % de sus clientes tenían depósitos que superaban los 250.000 dólares asegurados [6].

El escenario anterior demuestra la importancia de la gestión de riesgo bancario dado que, no solo se está viendo afectada la entidad financiera sino también, los clientes asociados a estas los cuales podrían perder parte o la totalidad de los depósitos realizados. Como consecuencia de un mal manejo del riesgo y de las decisiones incorrectas por parte de las entidades financieras, sobre la inversión de sus activos.

Situaciones como la anterior, hace que muchos clientes se lleguen a preguntar sobre qué pasaría con sus ahorros si su banco quiebra, lo cual genera un impacto negativo en la reputación y confianza tanto de los clientes como del público en general frente a la estabilidad del sistema bancario de su país.

En particular, en Colombia y de acuerdo con el más reciente estudio 'Consumer Pulse' realizado por TransUnion [7] asociado al último trimestre del 2022, se encontró que el crecimiento en la adquisición de tarjetas de crédito y créditos de libre inversión sigue siendo fuerte a medida que los consumidores hacen frente al difícil contexto macroeconómico que se vive en el país. De hecho, la adquisición de estos productos aumentó un 12 % en comparación con el mismo trimestre del 2021.

Sin embargo, del anterior estudio también se obtuvo que la tasa de morosidad grave a nivel de consumidor (60 o más días de retraso) se situó en 9,7 % en el cuarto trimestre de 2022. Lo cual representa un aumento de 106 puntos básicos (pbs) en comparación con el mismo periodo de 2021, simbolizando un riesgo para las entidades financieras.

A partir de lo mencionado anteriormente, es importante resaltar el informe de política monetaria del Banco de la República. En el que se evidenció que los indicadores de endeudamiento en los hogares colombianos siguen creciendo mientras que su ahorro baja, reduciendo así la capacidad de respuesta de los hogares ante choques negativos de la actividad económica [8].

Es claro que de acuerdo con las necesidades de los consumidores en un momento en el que se tiene un aumento en el costo de vida, las entidades financieras dan prioridad a ofertar los productos que mejor se adapten a las necesidades de los consumidores. En algunas ocasiones, los usuarios buscan acceder a créditos con diferentes entidades con el fin de solventar sus deudas. Sin embargo, es importante que estos mismos conozcan las

repercusiones que podría llegar a tener el no pago oportuno de sus créditos.

En Colombia conforme a lo señalado en la Ley 1266 de 2008, las entidades financieras tienen la obligación de reportar a las centrales de riesgo (CR) la información sobre el comportamiento crediticio de sus deudores [9]. Las CR recopilan toda la información sobre el historial crediticio y financiero de una persona o empresa y, si estas han incurrido en incumplimiento de sus obligaciones, las entidades bancarias deben realizar la evaluación y recalificación de la cartera de créditos como también, del puntaje crediticio de los usuarios. Lo anterior, puede generar restricciones en el acceso a nuevos servicios financieros o dificultad de que a futuro pueda obtener nuevos créditos o préstamos pues ahora, el usuario representa un mayor riesgo de impago para la entidad.

De acuerdo con lo anterior, la gestión del riesgo bancario es de vital importancia dado que tanto usuarios como entidades financieras se ven afectadas en el momento de presentarse situaciones como las descritas anteriormente. En específico, el diseño de un modelo de riesgo de crédito en entidades que prestan el servicio de créditos es de vital importancia dado que, garantiza la viabilidad de conceder o no un préstamo a un cliente que cuenta con cierto historial financiero [10]. Además, le permite también a la entidad financiera evaluar, reducir tiempo, costo y el factor de subjetividad asociado a la valoración del riesgo de cada solicitud de crédito, previniendo de esta manera las pérdidas por impago y, además, generando mayor solidez financiera dado que la rentabilidad del banco aumentara.

De acuerdo con lo anterior, en este proyecto se le pretende dar respuesta a la siguiente pregunta de investigación:

¿Cuál de los enfoques, entre modelos probabilísticos tradicionales y técnicas de Machine Learning, son más eficientes para predecir el riesgo crediticio, especialmente cuando se aplican estrategias de balanceo de datos?

## 1.3. Estado del arte

A lo largo de los años, se ha venido trabajando en el desarrollo de modelos de gestión de riesgo de crédito con el fin de poder evaluar la capacidad de pago o el riesgo de no pago de los clientes y estos surgen de la década de los 30. Por ejemplo, Smith y Winakor en 1935 propusieron un modelo cuyo objetivo era evaluar el riesgo crediticio de una cartera de clientes basándose en información financiera contable. Este enfoque enfatizaba la importancia de que las empresas identificaran las causas fundamentales de sus dificultades financieras para poder abordarlas de manera eficaz [11].

Por otra parte, los modelos de score crediticio son técnicas muy utilizadas por las entidades financieras para apoyarse en sus decisiones de asignación de créditos. Los antecendentes de estos modelos se ubican en el trabajo pionero de Fisher en 1936 [12],

quien ideo la técnica estadística de análisis discriminante para diferenciar grupos dentro de una población a través de atributos medibles. Mas adelante, en 1941, Durand observo que el mismo enfoque que publico Fisher podría usarse con el fin de distinguir entre prestamos buenos y malos [13].

De acuerdo a la revisión realizada se identifica que los modelos de regresión logística son un punto de referencia en la industria del riesgo crediticio, principalmente por la facilidad de la interpretación de sus resultados. El modelo de regresión logística es una técnica de análisis estadístico que se utiliza para modelar la probabilidad de ocurrencia de un evento binario, como puede ser el impago de un préstamo.

Se resalta también el artículo de investigación realizado por Bermudez, Manotas y Olaya, el cual describe la construcción de un modelo de regresión logística para la estimación de la probabilidad de incumplimiento de sus asociados y valora el desempeño del modelo desde el poder de discriminación entre cumplidos e incumplidos mediante el método Hold Out de este trabajo se logra concluir que el uso de estos modelos es flexible y fácilmente se logra modelar la probabilidad de incumplimiento [14].

Posteriormente y teniendo en cuenta el aumento en la cantidad y la complejidad de los datos, estos modelos fueron enriquecidos a través de técnicas matemáticas, algoritmos de machine learning y de inteligencia artificial para encontrar los patrones ocultos de las credenciales de los clientes morosos y no morosos. De acuerdo a lo anterior, en la búsqueda de la literatura se encuentra el uso de modelos de aprendizaje automático tales como las redes neuronales.

Para este caso, Pérez y Fernández exponen la aplicación de un modelo de cuantificación del riesgo para una cartera comercial [15] donde a partir de varias pruebas y evaluaciones del modelo, llegan al mas adecuado para el problema de clasificación el cual es una red neuronal probabilística la cual arrojo el más alto porcentaje de aciertos.

El estudio realizado por Desai, Crook y Overstreet dan cuenta de la exploración que se estaba llevando a cabo sobre la capacidad de las redes neuronales, como los perceptrones multicapa para evaluar el riesgo crediticio. Este grupo de autores realizan comparación de los resultados obtenidos del uso de redes neuronales frente a técnicas tradicionales donde obtienen que el desempeño estos últimos no fue tan bueno como el de los modelos personalizados, particularmente cuando se trataba de clasificar correctamente los préstamos incobrables [16].

En [17] se presenta un modelo de calificación crediticia de mayor precisión basado en redes neuronales Multi-Layer Perceptron (MLP) que han sido entrenadas con el algoritmo de propagación hacia atrás. Realizaron el entrenamiento de 34 modelos con diferentes pesos iniciales, instancias de entrenamiento y cada uno tiene de 6 a 39 capas ocultas. Para este caso obtuvieron un modelo con una precisión de clasificación del 87 % lo cual

es un resultado relevante en comparación con resultados encontrados previamente y también, demostraron que la optimización de la estructura del conjunto de datos puede aumentar significativamente la precisión de un modelo en comparación con los métodos tradicionales.

En el articulo realizado por Khashman, el autor se enfoca en analizar el empleo de modelos de redes neuronales supervisadas como una metodología para evaluar el riesgo de crédito. Por medio de la base de datos seleccionada, realizó diferentes esquemas de aprendizaje con diferentes proporciones de datos de entrenamiento y validación con el fin de realizar una comparación entre los resultados de su implementación, identificando los índices de precisión que cada uno le ofrecía [18].

La maquina de vectores de soporte agrupados (CSVM) fue utilizada por Harris para el desarrollo de tarjetas de calificación crediticia donde se resalta, que este tipo de algoritmos logra niveles comparables de rendimiento de clasificación sin dejar de ser relativamente económico desde el punto de vista computacional en comparación con técnicas basadas en maquina de soporte vectorial (SVM) [19].

En el caso de caso de Xu, Zhou y Wang proponen un modelo de calificación crediticia que utiliza técnicas de clasificación de análisis de enlaces para preprocesar muestras en información ponderada y maquinas de soporte vectorial para crear clasificadores. De hecho a partir de este estudio se pudo identificar que la técnica de maquinas de soporte vectorial ponderado destaca por su capacidad de clasificación en comparación con los métodos convencionales [20].

En la tesis [21], se emplearon modelos como regresión logística, random forest y xgboost para la clasificación de clientes como riesgosos o no. La investigación abordó el análisis de una base de datos en Kaggle, que presentaba un desbalance entre las clases de la variable predictora. A pesar de esta desproporción, se llevó a cabo el entrenamiento de los modelos. En cuanto a la manipulación de los datos, se identificaron registros nulos y, al representar menos del 20 %, se optó por su eliminación.

En [22] utilizaron técnicas de inteligencia artificial y aplicaron la técnica de sobre-muestreo sintético (SMOTE) para mejorar tanto la estabilidad como el rendimiento de los modelos de aprendizaje automático. Estos modelos incluyeron árboles de decisión y redes neuronales artificiales (ANN) y se enfocaron en abordar el desequilibrio en los datos.

Caro y Rodas [23] decidieron trabajar con los modelos de Regresión Logística, Random Forest y Gradient Boosting. Debido al desbalance que manejaba la base de datos realizaron dos escenarios aplicando los modelos con la técnica de balanceo oversampling y luego con la técnica de balanceo undersampling. Dentro de los resultados obtenidos intuyen que los modelos se vieron afectados por la metodología de balanceo SMOTE debido a la generación de una gran cantidad de datos sintéticos. Posteriormente entrenaron el modelo

Gradient Boosting con la metodología Stratified K Fold e implementaron la validación cruzada.

Estudios como [24] han demostrado que un conjunto de datos equilibrado proporciona un rendimiento de clasificación general mejorado en comparación con un conjunto de datos desequilibrados. Por otro lado, resaltan que la técnica de sobremuestreo (SMOTE) es un método poderoso que ha mostrado una gran cantidad de éxito en varias aplicaciones.

En [25] realizaron un estudio sobre el algoritmo de machine learning maquinas de soporte vectorial. En este articulo, abordan el problema de aprender de un conjunto de datos desequilibrado sin embargo, realizan un proceso de sobremuestreo y submuestreo basado en bootstrapping con el fin de abordar el desbalance que presenta la base de datos.

# –2–
# Marco teórico

## 2.1. Riesgo

El riesgo es un concepto fundamental en múltiples campos de estudio, como la economía, las finanzas, la psicología, la sociología, entre otros. Sin embargo, para el diccionario de la Real Academia Española el riesgo se define como: contingencia o proximidad de un daño donde contingencia, se define como la posibilidad de que algo suceda o no suceda. Mientras que, Gallati define el riesgo como "una condición en la que existe la posibilidad de desviación de un resultado deseado que se espera o anhela" [26].

Dado que el concepto de riesgo ha sido ampliamente estudiando en diversos campos su definición puede variar teniendo en cuenta el contexto y la perspectiva teórica que se este utilizando para analizar el riesgo.

## 2.2. Riesgo bancario

Las entidades financieras son aquellas que se dedican de manera habitual y profesional a la captación de recursos del público para colocarlos en forma de créditos, inversión en valores o cualquier otra actividad financiera, con el objetivo de obtener beneficios. Por tanto, teniendo en cuenta la naturaleza de las actividades y las condiciones con las que operan las entidades financieras estas se encuentran expuestas a diversos tipos de riesgos bancarios puesto que, cada una de sus operaciones contiene de manera implícita o explicita la incertidumbre.

En este caso, el riesgo en entidades financieras se conoce como riesgo bancario el cual se refiere a la posibilidad de que un banco genere pérdidas financieras en el desarrollo de sus actividades comerciales debido a la variabilidad e incertidumbre de diferentes factores internos o externos.

Es importante resaltar que por las actividades comerciales que se tienen en las entidades financieras el riesgo bancario no puede ser completamente eliminado sin embargo, estas entidades pueden tomar las medidas pertinentes para lograr minimizar los riesgos y gestionarlos de una manera adecuada, para evitar que los bancos tenga un sistema financiero débil. De acuerdo a lo anterior, identificar los tipos de riesgo que se presentan es muy importante para poder gestionarlos de manera efectiva.

Con respecto a esto Gáytan distingue cinco clases de riesgo bancario las cuales abarcan cualquier tipo de eventos o variables que genera cierto grado de incertidumbre en los resultados financieros del banco. Dichas clases son: riesgo de mercado, riesgo de liquidez, riesgo operacional, riesgo legal y riesgo de crédito [27]. Sin embargo, para el desarrollo de este trabajo se hará énfasis en el riesgo de crédito.

1. **Riesgo de mercado**

   Riesgo de perdida en los valores de mercado de los activos, pasivos y operaciones fuera de balance debido a cambios desfavorables de factores de riesgo como las tasas de interés, los tipos de cambio, la inflación, los precios de los productos básicos y las tasas de crecimiento.

2. **Riesgo de liquidez**

   Surge de las dificultades temporales que se presentan en una entidad para hacer frente a sus compromisos de pago como consecuencia de desajustes temporales entre los flujos de efectivo de los activos y pasivos.

3. **Riesgo operacional**

   Se refiere a la perdida potencial resultante de diversos eventos inesperados, incompletos o fallidos, los cuales se pueden generar en los sistemas de información que maneja la entidad financiera, infraestructura, errores en el procesamiento, fallas, fraude o errores humanos. Las incidencia del riesgo operativo, y las pérdidas que conlleva, atraen fácilmente publicidad negativa y afectan la reputación de la entidad financiera por tanto, es importante que reconozcan y anticipen los probables eventos de riesgo operacional [28].

4. **Riesgo legal**

   Hace referencia a la posibilidad de que una entidad financiera sufra pérdidas financieras o daño a su reputación como resultado del incumplimiento involuntario o negligente de las leyes y regulaciones aplicables o por la imposición de multas y sanciones por parte de las autoridades reguladoras.

5. **Riesgo de crédito**

   El riesgo de crédito se origina por el incumplimiento de una contraparte, o de un deudor en su obligación. Puede ser un incumplimiento absoluto o un deterioro de la capacidad de pago de un deudor, lo que en última instancia hace que el incumplimiento sea más probable [28].

La valuación del riesgo de crédito se basa en la probabilidad de que el prestatario o emisor del bono incumpla con sus obligaciones (ocurra un default) ya que un banco asume un riesgo de crédito cuando otorga un préstamo con la expectativa de que el prestatario utilizara el préstamo de acuerdo con los términos y condiciones acordados y, en última instancia, pagara el préstamo en la fecha de vencimiento pactada sin embargo, esto puede o no realizarse.

Reconocer los factores que inciden en el riesgo de crédito es muy importante dado que el cumplimiento y la ineficiencia en la gestión de riesgo de crédito además de producir una alta mora afecta tanto la solvencia de las entidades financieras como la economía en general. El riesgo de crédito puede analizarse en tres dimensiones básicas [29]:

- *Riesgo de incumplimiento:* es la probabilidad de que se presente el no cumplimiento de una obligación de pago dentro de un período establecido o cuando este se realiza, con posterioridad a la fecha programada. Algunas entidades financieras establecen plazos de gracia antes de declarar el incumplimiento del pago.
- *Riesgo de exposición:* se entiende como la incertidumbre sobre los futuros pagos que se deben. Algunas entidades financieras brindan productos donde los desembolsos se otorgan sin fecha fija contractual y no se conoce con exactitud el plazo de liquidación; por ello se dificulta la estimación de los montos en riesgo
- *Riesgo de recuperación:* se origina por la existencia de un incumplimiento. No se puede predecir, puesto que depende del tipo de garantía que se haya recibido y de su situación al momento del incumplimiento.

Alguno de los términos mas utilizados para describir las pérdidas potenciales que una entidad financiera o inversor puede enfrentar son [30]:

- *Perdida esperada:* es la pérdida estimada que una entidad financiera espera enfrentar en su cartera de activos o inversiones, por medio del análisis y proyecciones de riesgo. La estimación de la pérdida esperada permite realizar provisiones para cubrir pérdidas futuras y resulta de la aplicación de la siguiente fórmula:

  $$PE = PD \times EA \times PEI$$

  Donde, PD es la probabilidad de default lo cual en otras palabras representa, la probabilidad de que en un lapso de tiempo los deudores incurran en incumplimiento. EA es la exposición del activo a un determinado riesgo y PEI representa la perdida esperada de valor activo dado el incumplimiento.
- *Perdida inesperada:* es la pérdida generada por sucesos no previsto inicialmente por la entidad. Es importante que las entidades financieras también ahorren dinero para cubrir este tipo de perdidas pues entre más amplia sea la dispersión de pérdidas inesperadas, mayor será el grado de riesgo de eventos crediticios inesperados.

Cada banco tiene su propia cultura financiera y los sistemas de calificación internos tienden a diferir significativamente de un banco a otro. En muchas ocasiones por razones de competitividad estos hacen parte del "patrimonio intelectualz hacen parte de la estrategia interna de los bancos para combatir el riesgo. Por tanto, contar con un esquema de calificación y asignación de créditos adecuado ayuda a evaluar el riesgo de prestar dinero a un cliente y así, determinar la viabilidad de concederlo o no un préstamo [10].

## 2.3. Reglamentaciones

El Comité de Supervisión Bancaria de Basilea es la organización mundial que reúne a las autoridades de supervisión bancaria, cuya función es fortalecer la solidez de los sistemas financieros [2] y actualmente está compuesto por 27 países de todo el mundo representados por miembros de la autoridad supervisora de cada sistema bancario.

Desde su surgimiento, el Comité ha sido un foro de discusión para impulsar la mejora de las prácticas y las normativas de supervisión bancaria, buscando siempre perfeccionar las herramientas de fiscalización internacional, a través de acercamientos y estándares comunes. Uno de los ejemplos más relevantes de la actividad del Comité de Basilea es el llamado Acuerdo de Basilea, que establece los estándares internacionales de regulación bancaria prudencial.

El primer Acuerdo de Basilea (Basilea I) surgió en 1988. Posteriormente en el año 2004, se publica Basilea II, en el cual se modificaba varios aspectos del anterior. Sin embargo, en diciembre de 2017, se finalizó el proceso de revisión en profundidad de Basilea II, acordándose un nuevo marco de regulación prudencial, conocido como Basilea III el cual además de ser un complemento a los estándares ya conocidos se centra en el fortalecimiento del capital, con el objetivo de que los bancos cuenten con la solvencia necesaria y acorde a los riesgos que toman.

Las autoridades regulatorias de los sistema financiero de cada país se basan en los estándares y regulaciones internacionales establecidas por el Acuerdo de Basilea. En Colombia específicamente, la Superintendencia Financiera de este país tiene por objetivo supervisar el sistema financiero colombiano con el fin de preservar su estabilidad, seguridad y confianza, así como promover, organizar y desarrollar el mercado de valores colombiano y la protección de los inversionistas, ahorradores y asegurados [9].

Dentro de la Circular Contable, la cual es una de las normas emitidas por la Superintendencia Financiera de Colombia, se establecen las directrices contables y financieras que deben seguir las entidades financieras con el fin de garantizar la transparencia y la veracidad de la información financiera [31]. Por ejemplo, el Capitulo II menciona las reglas relativas a la gestión del riesgo de crédito, las políticas deben precisar las características básicas de los sujetos de crédito de la entidad y los niveles de tolerancia frente al riesgo,

discriminar entre sus potenciales clientes para determinar si son sujetos de crédito y definir los niveles de adjudicación para cada uno de ellos.

## 2.4. Modelos

El análisis de la literatura realizado en la Sección 1.3 se exploraron diversas investigaciones relacionadas con el tema de interés. Sin embargo, se destacó que los modelos de regresión logística son fundamentales en el contexto del riesgo crediticio, siendo una herramienta de referencia ampliamente utilizada. A continuación, se presenta un análisis detallado:

### 2.4.1. Modelo de regresión logística

La regresión logística es probablemente el método más utilizado para desarrollar modelos de clasifación y calificación crediticia [32]. Esta técnica estima la probabilidad de que ocurra un evento, en función de un conjunto de datos determinado de variables independientes.

De acuerdo con [33] suponga un modelo de la forma

$$y_i = \mathbf{x}_i^{'}\beta + \epsilon_i,$$

en donde $\mathbf{x}_i^{'} = [1, x_{i1}, x_{i2}, ..., x_{ik}]$, $\beta^{'} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]$, y la variable de respuesta $y_i$ toma valor de cero o uno. Suponiendo que la variable de respuesta $y_i$ es una variable aleatorio de Bernoulli, su distribución de probabilidad esta dada por

| $y_i$ | **Probabilidad** |
|---|---|
| 1 | $P(y_i = 1) = \pi_i$ |
| 0 | $P(y_i = 0) = 1 - \pi_i$ |

Cuando la variable de respuesta es binaria, se observa empíricamente que la función de respuesta suele ser no lineal. Se emplea comúnmente una función logística, con forma en S, que puede ser creciente o decreciente, como se muestra en la siguiente figura:

$$E(y) = \frac{exp(\mathbf{x}^{'}\beta)}{1 + exp(\mathbf{x}^{'}\beta)},$$

o bien, lo que es igual

$$E(y) = \frac{1}{1 + exp(-\mathbf{x}^{'}\beta)}.$$

La función logística puede ser fácilmente linealizada al expresar la parte estructural del modelo en términos de la media de dicha función de respuesta. Sea

$$\eta = \mathbf{x}'\beta,$$

el predictor lineal, estando definida $\eta$ por la transformación

$$\eta = \ln\frac{\pi}{1-\pi}.$$

A la transformación anterior se le llama con frecuencia transformación logit de la probabilidad $\pi$ y la relación $\pi/(1-\pi)$ se denomina ventaja.

La forma general del modelo de regresión logística esta dada por

$$y_i = E(y_i) + \epsilon_i$$

donde las observaciones $y_i$ son variables aleatorias independientes de Bernoulli, cuyos valores esperados son

$$E(y_i) = \pi_i = \frac{\exp(\mathbf{x}'\beta)}{1+\exp(\mathbf{x}'\beta)}.$$

Cada observación de la muestra sigue la distribución de Bernoulli, por lo que la distribución de probabilidades de cada observación es

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \qquad i = 1, 2, ..., n$$

cada observación $y_i$ toma el valor de 0 o 1. Como las observaciones son independientes, la función verosimilitud no es mas que

$$\begin{aligned} L(y_1, y_2, ..., y_n, \beta) &= \prod_{i=1}^{n} f_i(y_i), \\ &= \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \end{aligned}$$

Aplicando logaritmo en la verosimilitud para mayor facilidad se obtiene:

$$\begin{aligned} \ln L(y_1, y_2, ..., y_n, \beta) &= \ln \prod_{i=1}^{n} f_i(y_i), \\ &= \sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n}\ln(1-\pi_i). \end{aligned}$$

Ahora bien, como $1-\pi_i = [1+\exp(\mathbf{x_i}'\beta)]^{-1}$, y $\eta_i = \ln[\pi_i/(1-\pi_i)] = \mathbf{x_i}'\beta$, el logaritmo de la verosimilitud se expresaria de la siguiente manera:

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i^{'} \beta - \sum_{i=1}^{n} \ln[1 + \exp(\mathbf{x}_i^{'} \beta)].$$

Teniendo en cuenta que en los modelos de regresión logística se tienen observaciones repetidas en cada nivel de la variables $x$. Sea $y_i$ la cantidad de 1 observado en $i$, y $n_i$ la cantidad de intentos en cada observación, el logaritmo de la verosimilitud se transforma en

$$\ln L(y, \beta) = \sum_{i=1}^{n} y_i \pi_i + \sum_{i=1}^{n} n_i \ln(1 - \pi_i) - \sum_{i=1}^{n} y_i \ln(1 - \pi_i). \tag{2.1}$$

La obtención de los estimados de máxima verosimilitud se pueden calcular a través de un proceso iterativo respaldado por un algoritmo de mínimos cuadrados iterativamente ponderados, también conocido por sus siglas en inglés como IRLS (Iteratively Reweighted Least Squares). Se proporcionan detalles exhaustivos de este procedimiento en el Apendice C.13 de la referencia [33] donde se deduce que el estimador de máxima verosimilitud resuelve una ecuación de la siguiente forma

$$X^{'}(y - \mu) = 0,$$

en donde $y^{'} = [y_i, y_2, ..., y_n]$ y $\mu^{'} = [n_1\pi_1, n_2\pi_2, ..., n_n\pi_n]$,. Este conjunto de ecuaciones se identifican comúnmente como las **ecuaciones de puntuación de máxima verosimilitud**.

No obstante, según lo indicado en [33], se menciona que el método para solucionar las ecuaciones de puntuación es el método de Newton-Raphson [34] debido a la naturaleza no lineal de las ecuaciones resultantes cuando se igualan a cero.

Conforme a lo anterior, sea $\beta$ el estimado final de los parámetros del modelo. Se puede demostrar que, en forma asintótica,

$$E(\hat{\beta}) = \beta \quad \text{y} \quad \text{Var}(\hat{\beta}) = (X^{'} V^{-1} X)^{-1}.$$

En última instancia, el valor estimado del predictor lineal se expresa como $\hat{n}_i = \mathbf{x}_i^{'} \hat{\beta}$, y el valor esperado del modelo de regresión logística se presenta de la siguiente manera:

$$\begin{aligned} \hat{y}_i = \hat{\pi}_i &= \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)}, \\ &= \frac{\exp(\mathbf{x}_i^{'} \hat{\beta})}{1 + \exp(\mathbf{x}_i^{'} \beta)}, \\ &= \frac{1}{1 + \exp(-\mathbf{x}_i^{'} \hat{\beta})}, \end{aligned}$$

$$\hat{y}_i = \frac{1}{1 + \exp(-\mathbf{x}_i'\hat{\beta})}. \tag{2.2}$$

Gráficamente la función logística se ve de la siguiente manera:

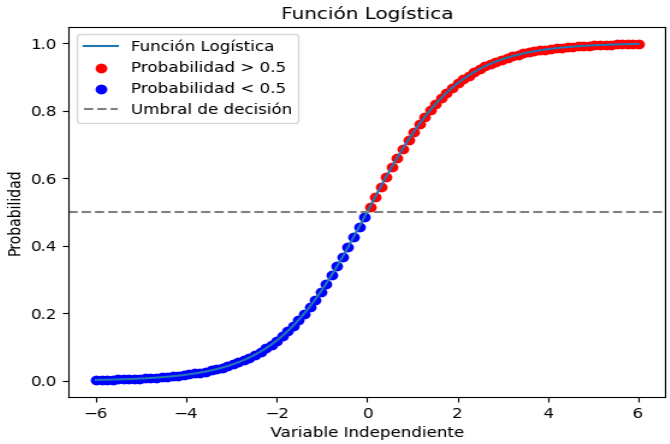



Figura 2.1: Curva logística. Elaboración propia.

Al examinar la ecuación 2.2, podemos observar que:

- Si $-\mathbf{x}_i'\hat{\beta}$ tiene un valor muy grande positivo $e^{-\mathbf{x}_i'\hat{\beta}}$ es aproximadamente 0 por tanto, $\hat{y}_i = 1$.
- Si $-\mathbf{x}_i'\hat{\beta} = 0$, $e^{-\mathbf{x}_i'\hat{\beta}} = 1$. En este contexto, la probabilidad resultante es exactamente 0.5, lo que marca el umbral crucial. Este punto de corte se utiliza para determinar si el registro i, caracterizado por esas específicas condiciones, pertenece a una clase particular o no.
- Si $-\mathbf{x}_i'\hat{\beta}$ tiene un valor muy grande negativo, $e^{-\mathbf{x}_i'\hat{\beta}} = \infty$ por tanto, $\hat{y}_i = 0$.

En el modelo de regresión logística cada variable aporta un peso a la calificación final del cliente por tanto, una de las principales fortalezas de este modelo es su facilidad en la interpretabilidad y facilidad de explicación [35]. Además, permite predecir el comportamiento de una variable categórica a partir de un conjunto de covariables discretas o continuas. Sin embargo, las variables independientes no deben estar relacionadas entre sí para obtener mejores resultados.

### 2.4.2. Modelos de Machine Learning

El Machine Learning (ML) o aprendizaje automático es una rama de la Inteligencia Artificial que se encarga de generar algoritmos que tienen la capacidad de aprender y no tener que programarlos de manera explícita. Esta herramienta ha sido de vital importancia

para encontrar patrones ocultos en las credenciales de los clientes morosos y no morosos con el fin de enriquecer la gestión en la detección de riesgo crediticio tal como se evidencio en la Sección 1.3. Siguiendo lo expuesto anteriormente, es fundamental destacar que los modelos de Machine Learning se dividen principalmente en dos categorías: supervisados y no supervisados.

- **Aprendizaje supervisado**
  En esta categoría para cada observación en el conjunto de datos, se tiene en cuenta las medidas o valores de las variables predictoras $x_i$ donde $i = 1, ..., n$ sin embargo, esta técnica tiene una medida asociada de la variable de respuesta conocida como $y_i$. Por tanto, el modelo se entrena con un conjunto de datos etiquetados, lo que le permite aprender a relacionar las salidas con los datos de entrenamiento y, por ende, realizar predicciones para nuevas entradas no vistas.

- **Aprendizaje no supervisado:** En este escenario para cada observación $i = 1, ...n$, se observa un vector de medidas $x_i$, pero no hay una variable de respuesta asociada $y_i$. En otras palabras, el modelo se entrena utilizando datos no etiquetados. A partir de esta información, el modelo se esfuerza por explorar de manera autónoma la estructura subyacente en los datos con el propósito de descubrir patrones de manera autónoma.

## Modelo de Random Forest

El algoritmo de bosque aleatorio fue desarrollado por Breiman y Cutler en 2001. De acuerdo con [36], el bosque aleatorio es un algoritmo versátil basado en árboles de decisión que puede elegir sus propias características y funcionar con diversos tipos de características sin necesidad de preparación previa, conservando la integridad de los datos numéricos.

Inicialmente es importante definir a que hace referencia un árbol de decisión. Un árbol de decisión comienza con un punto de partida llamado nodo raíz. Desde este nodo, se extienden ramas hacia nodos internos, que toman decisiones basadas en las características de los datos para dividirlos en grupos más homogéneos. Los nodos hoja representan los resultados finales o categorías en el conjunto de datos.

De acuerdo al procedimiento anterior, la representación gráfica de un árbol de decisión se presenta de la siguiente manera:

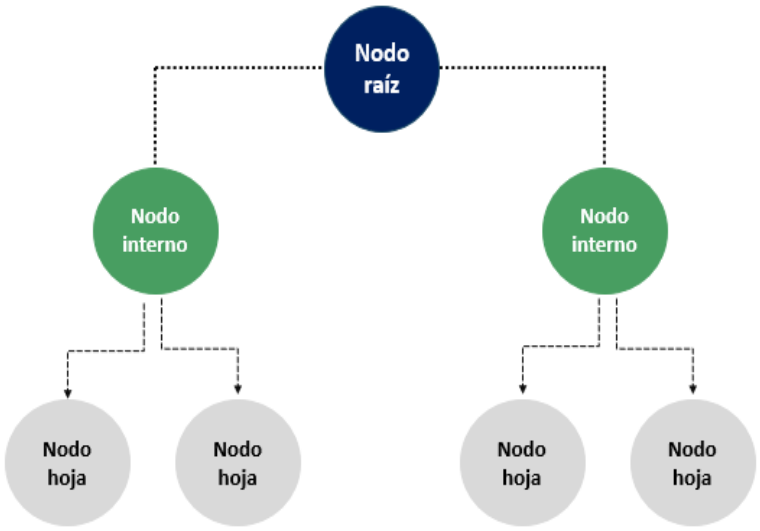



Figura 2.2: Árbol de decisión.

La ramificación en los arboles de decisión es fundamental ya que es el que permite que el modelo tome decisiones basadas en las características de los datos. Un algoritmo crucial para determinar como se lleva a cabo estas ramificaciones es mediante el uso del índice Gini [37] el cual mide la impureza de un conjunto de datos $D$. De acuerdo con esto, la formula del índice de Gini se tiene en la siguiente ecuación.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2, \tag{2.3}$$

donde $p_i$ representa la proporción de elementos en el conjunto de datos $D$ que pertenecen a la clase $C_i$ la cual se estima mediante $|C_{i,D}|/|D|$. La suma se calcula en $m$ clases. De acuerdo con [38] el proceso que se lleva a cabo para la construcción de un árbol de decisión teniendo en cuenta el índice de Gini es el siguiente:

- La impureza de un árbol se define como la suma de la impureza de cada nodo terminal del árbol, multiplicada por la proporción ($p_i p_j$) de casos que alcanzan ese nodo en el árbol.

- Seleccionamos la división que produce la mayor reducción en el índice de Gini. Este análisis implica considerar todas las opciones disponibles para dividir teniendo en cuenta atributos de valor continuo o discreto.

- Continuamos con las divisiones hasta que los nodos finales tengan un número muy reducido de casos o sean completamente homogéneos. El criterio de detención se alcanza cuando el índice de Gini en un nodo llega a cero, lo cual indica que todos los registros de datos en ese nodo han sido clasificados por completos.

La reducción de impurezas en la que se incurriría mediante una división binaria en un atributo $A$ de valor discreto o continuo es

$$\Delta Gini(A) = Gini(D) - Gini_A(D).$$

El atributo que maximiza la reducción en la impureza (o, en otras palabras, minimiza el índice Gini ponderado) se selecciona como el atributo de división óptimo. Esto significa que este atributo es el que mejor contribuye a la separación de las categorías en los subconjuntos resultantes.

Para un problema de dos clases, el gráfico de impurezas se mide en función de la probabilidad de la primera clase. La figura muestra que la impureza de Gini en el caso que nos interesa cuando la muestra es homogénea se obtiene en $I(A) = 0$ para ello $f$ debe ser cóncava con $f(0) = f(1) = 0$ en caso contrario, su valor máximo se obtiene cuando $f(0{,}5) = 0{,}5$.

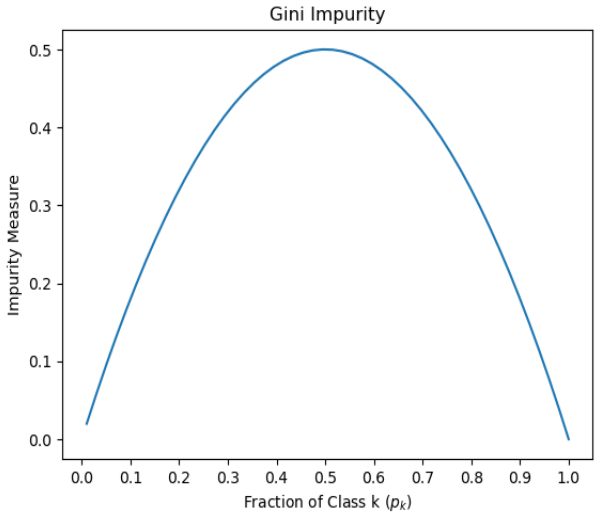



Figura 2.3: Impureza de Gini.

Teniendo una comprensión clara de la definición de un árbol de decisión, Random Forest es una técnica de aprendizaje supervisado desarrollado por Breiman [39] donde se generan diferentes árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto. Según la descripción de Breiman , podemos definir el bosque aleatorio de la siguiente manera:

**Definición:** Un bosque aleatorio es un clasificador que consta de una colección de clasificadores con estructura de árbol $h(\mathbf{x}, \Theta_k), k = 1, \ldots$ donde $\Theta_k$ son vectores aleatorios independientes distribuidos de forma idéntica y cada árbol emite un voto unitario para la clase más popular en la entrada $\mathbf{x}$.

El algoritmo de Random Forest sigue los siguientes pasos:

1. Se crea un grupo de observaciones aleatorias (mediante bootstrap, que es una técnica estadística para obtener muestras de una población donde una observación se puede considerar en más de una muestra). De acuerdo a lo anterior, las observaciones no estimadas en los árboles (conocidas como “out of the bag”) se utilizan para validar el modelo.

2. Cada nodo del árbol contiene un subconjunto aleatorio de variables predictoras.

3. Se busca la mejor división de los datos de entrenamiento teniendo en cuenta solo las variables seleccionadas en el paso 2 utilizando algun método como el índice de Gini.

4. Los pasos previos se replican de manera que se obtiene un conjunto de árboles de decisión entrenados con diversos conjuntos de datos y atributos.

5. Finalmente, una vez el algoritmo es entrenado, la evaluación de cada nueva entrada se realiza a partir del conjunto de árboles. Si se tiene un problema de clasificación, el bosque obtendrá como resultado el resultado de la votación mayoritaria de los arboles, y en caso de regresión por el valor promedio de los resultados.

En la Figura 2.4 se puede evidenciar gráficamente el proceso de construcción de este algoritmo.

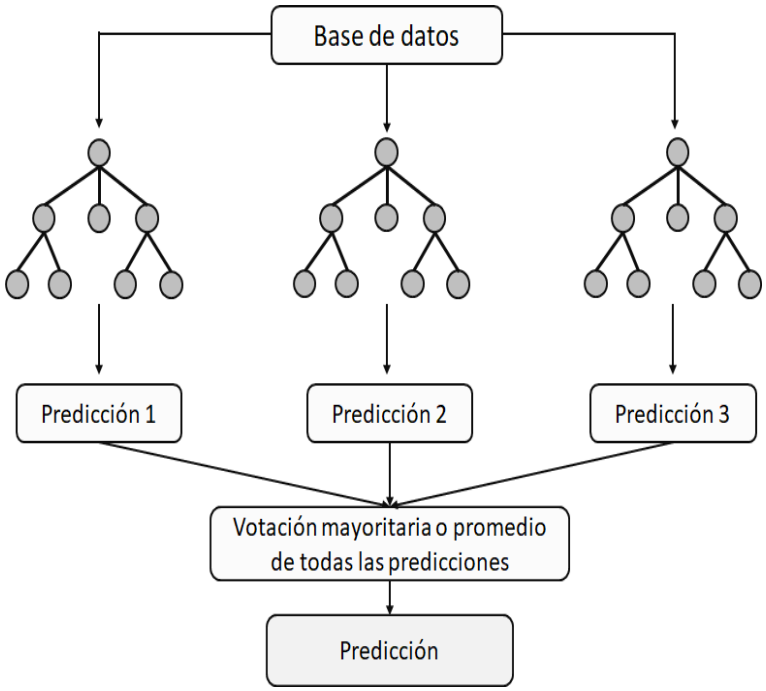



Figura 2.4: Elaboración propia.

## Ventajas del modelo Random Forest

Entre las principales ventajas para el modelo de Random Forest, se destacan las siguientes ([36], [40]):

- Este método puede usarse para casos de clasificación tomando el voto mayoritario del conjunto de árboles, y en caso de regresión realizando el valor promedio de los resultados.
- Es bastante resistente al sobre ajuste y no es necesario pensar demasiado en corregir sus hiper parámetros entre ellos el número de árboles y el número de atributos utilizados para hacer crecer cada árbol.
- Se resalta la capacidad de manejar datos de alta dimensión y utilizar una gran cantidad de árboles sobre el conjunto.
- En caso de valores faltantes este es un método eficaz para estimar estos datos y mantener la exactitud.
- El mecanismo de construcción de Random Forest también se utiliza para la reducción de dimensionalidad ya que permite permite establecer la importancia de cada variable en la predicción final.

## Desventajas del modelo Random Forest

Entre las principales desventajas para el modelo, se destacan las siguientes ([36], [40]):

- El modelo de Random Forest puede verse afectado por datos de entrenamiento desequilibrados. Al minimizar la tasa de error general, tiende a priorizar la precisión en la clase mayoritaria, lo que a veces resulta en una baja precisión para la clase minoritaria [41].
- La visualización gráfica de los resultados puede ser difícil de interpretar.
- Se tiene muy poco control sobre lo que hace el modelo por lo que puede parecer que tiene un enfoque de caja negra.
- Dado que utiliza la técnica de ensamblado de árboles de decisión aleatorios, puede requerir cierto procesamiento de los datos.

## XGBoost

El algoritmo XGBoost (Extreme Gradient Boosting) es una técnica de aprendizaje supervisado [42] también basada en árboles de decisión. Sin embargo, la principal diferencia entre los algoritmos XGBoost y Random Forest es que en el primero el usuario define la

extensión de los árboles mientras que en el segundo los árboles crecen hasta su máxima extensión.

Supongamos que tenemos un conjunto de datos con $n$ ejemplos y $m$ características, representado como $D = (\mathbf{x}_i, y_i)$, donde $|D| = n$, $\mathbf{x}_i \in \mathbb{R}^m$ y $y_i \in \mathbb{R}$. En el contexto de XGBoost, que utiliza clasificaciones y árboles de regresión por definición, consideremos un conjunto de árboles en los que se emplean $K$ funciones aditivas para realizar predicciones.

$$\hat{y} = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F}, \tag{2.4}$$

donde $\mathcal{F} = \{f(\mathbf{x}) = \omega_{q(\mathbf{x})}\}(q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$ es el espacio de los árboles de regresión. Aquí $q$ representa la estructura de cada árbol que asigna un ejemplo al índice de hoja correspondiente. $T$ es el número de hojas del árbol. Cada $f_k$ corresponde a una estructura de árbol independiente $q$ y pesos de hojas $w$.

Para decidir la forma de la función $f_k(\mathbf{x}_i)$ que se agrega en cada una de las $k$ iteraciones, el árbol base debe optimizar una función objetivo que consiste en una función de pérdida $l(\hat{y}_i, y_i)$ para medir la precisión entre el valor predictivo y el valor real. Al mismo tiempo, es importante incorporar un término de regularización $\Omega(f_t(x_i))$ para evitar el sobreajuste.

De acuerdo a lo anterior, la ecuación 2.5 muestra la función objetivo del algoritmo XGBoost:

$$\begin{aligned} \mathcal{L}(\phi) &= \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \\ \text{donde } \Omega(f) &= \omega T + \frac{1}{2}\lambda \left\|\omega\right\|^2 . \end{aligned} \tag{2.5}$$

La ecuación 2.5 incluye funciones como parámetros por tanto, no es posible utilizar métodos de optimización tradicionales. Lo anterior, lleva a entrenar el modelo de forma iterativa. Sea $\hat{y}_i^{(t)}$ la predicción de la i-ésima instancia en la k-ésima iteración, tendremos:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t+1)} + f_t(\mathbf{x}_i)) + \Omega(f_t). \tag{2.6}$$

Aplicando la expansión de Taylor de segundo orden.

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t+1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t), \tag{2.7}$$

donde $g_i = \partial_{\hat{y}(t-1)} l(y_i, \hat{y}^{(t-1)})$ y $h_i = \partial^2_{\hat{y}(t-1)} l(y_i, \hat{y}^{(t-1)})$ representan el gradiente de primer y segundo orden sobre la función de pérdida.

Definiendo $I_j = \{i|q(\mathbf{x}_i) = j\}$ como el conjunto de instancias de la hoja $j$ y expandiendo $\Omega$ se obtiene

$$\begin{aligned}\tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n}[g_i f_t(\mathbf{x}_i) + \frac{1}{2}h_i f_t^2(\mathbf{x}_i)] + \lambda T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \\ &= \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \lambda T.\end{aligned} \tag{2.8}$$

Para una estructura fija $q(\mathbf{x})$, podemos computar el peso óptimo $\omega_j^*$ de la hoja $j$ como

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} g_i h_i + \lambda}, \tag{2.9}$$

y calculando el correspondiente valor óptimo

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} g_i h_i + \lambda} + \lambda T. \tag{2.10}$$

En la mayoría de los casos, listar exhaustivamente todas las posibles estructuras de un árbol $q$ es un proceso demasiado complejo. En lugar de eso, se emplea un algoritmo codicioso que inicia con una única hoja y, de manera iterativa, añade ramas al árbol. Supongamos que $I_L$ e $I_R$ representan los conjuntos de instancias de los nodos izquierdo y derecho respectivamente, después de efectuar una división. Si $I = I_L \cup I_R$ entonces la reducción de la pérdida después de la división se calcula de la siguiente manera:

$$\mathcal{L}_{\text{split}} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} g_i h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} g_i h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} g_i h_i + \lambda}\right] - \gamma.$$

En este algoritmo de XGBoost cada árbol se entrena con una parte de los datos y utiliza características seleccionadas al azar. Durante el proceso, se calculan los residuos o errores de predicción de cada árbol en relación con los datos reales. Luego, se combinan las predicciones de todos los árboles, y el modelo final tiene en cuenta estos residuos para tomar decisiones más precisas y robustas. En la Figura 2.5 se tiene una representación gráfica que muestra cómo funciona este algoritmo.

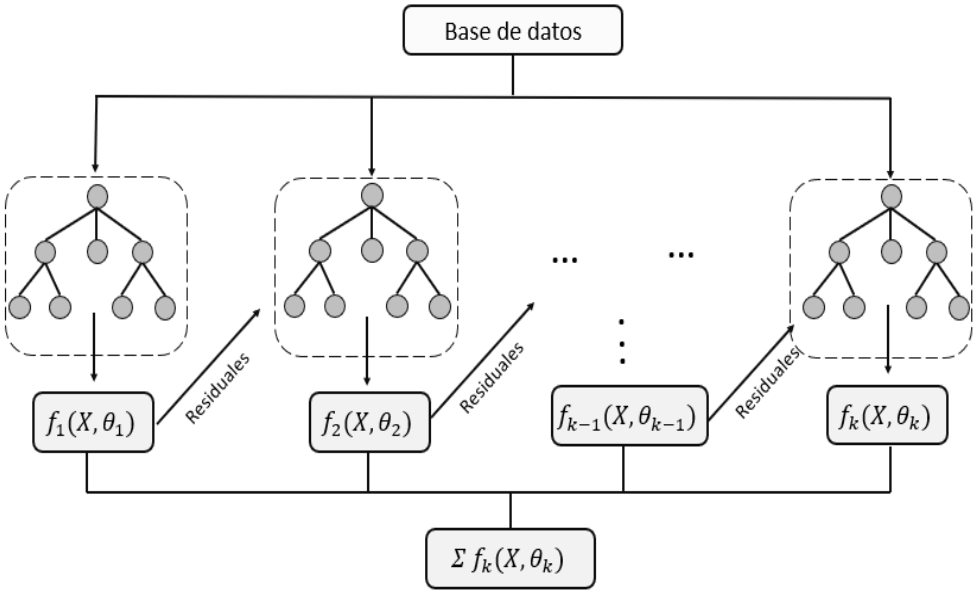



Figura 2.5: Modelo XGBoost.

## Ventajas del modelo XGBoost

Entre las principales ventajas para el modelo XGBoost, se destacan los siguientes aspectos [40]:

- XGBoost puede manejar valores faltantes en los datos sin requerir imputación y además, es eficaz para manejar relaciones no lineales y estructuras de datos complejas gracias a su capacidad de construir modelos secuenciales que se adaptan continuamente.
- Esta técnica permite entrenar modelos y realizar predicciones de manera eficiente.
- Mejora la velocidad del tiempo de ejecución al optimizar la clasificación mediante la ejecución en paralelo.
- Mejora el problema de sobreajuste mediante el uso de diferentes formas de regularización.
- Es un modelo que permite permite establecer la importancia de cada variable en la predicción final.

## Desventajas del modelo XGBoost

Entre las principales desventajas para el modelo XGBoost, se destacan los siguientes aspectos [40]:

- A pesar de que XGBoost maneja la ejecución en paralelo puede ser computacionalmente intensivo, especialmente para árboles profundos o grandes conjuntos de datos.
- La interpretación del modelo XGBoost tiende a volverse compleja debido a su estructura basada en la combinación de múltiples modelos.
- Para hacer uso de este modelo es importante identificar los hiperparámetros adecuados para lograr minimizar el error de precisión y evitar el sobreajuste del mismo.
- XGBoost opera exclusivamente con vectores numéricos, por lo que es necesario convertir previamente los datos no numéricos.

A partir de los modelos descritos anteriormente, se buscará contrastar modelos tradicionales y de machine learning haciendo uso de una misma base de datos (ver Sección 4.2) para determinar la mejor técnica que permite predecir si un solicitante de préstamo realizara o no el pago del mismo.

## 2.5. Técnicas de balanceo de datos

Cuando se trabaja con bases de datos y se pretende aplicar un algoritmo de aprendizaje, es esencial tener en consideración el potencial impacto del desequilibrio entre las clases de los datos. Este desequilibrio puede tener repercusiones en el rendimiento de los clasificadores, ya que los algoritmos que no abordan adecuadamente esta desigualdad tienden a dar prioridad a la clase mayoritaria mientras pasan por alto la clase minoritaria [43].

Según la revisión de la literatura realizada, se destaca que las técnicas de resampling son una de las soluciones más empleadas para abordar el desbalance de datos. Estas técnicas tienen un impacto directo en la distribución de las clases dentro del conjunto de datos y generalmente se dividen en dos categorías principales: las técnicas de resampling y undersampling.

Los métodos de resampling no heurísticos se enfocan en realizar submuestreo (undersampling) y sobre muestreo (oversampling) de manera aleatoria. Por otro lado, existen enfoques heurísticos que aplican técnicas con una estructura matemática y razonable durante el proceso de resampling [44]. A continuación, se presenta una descripción de estos métodos para una mejor comprensión de los conceptos mencionados.

### 2.5.1. Técnicas de undersampling

Hace referencia a aquellas técnicas que tiene como finalidad igualar las distribuciones desbalanceadas de datos eliminando instancias de la clase mayoritaria. Dentro de los

riesgos asumidos al utilizar este tipo de técnicas es la pérdida de información asociada a los datos de la clase mayoritaria.

Los algoritmos ha destacar de esta técnica son los siguientes:

- **Random Under Sampler** es una forma rápida y sencilla de equilibrar los datos seleccionando aleatoriamente un subconjunto de datos para las clases objetivo. Esta técnica elimina aleatoriamente las instancias de la clase mayoritaria tal como se muestra en la siguiente figura:

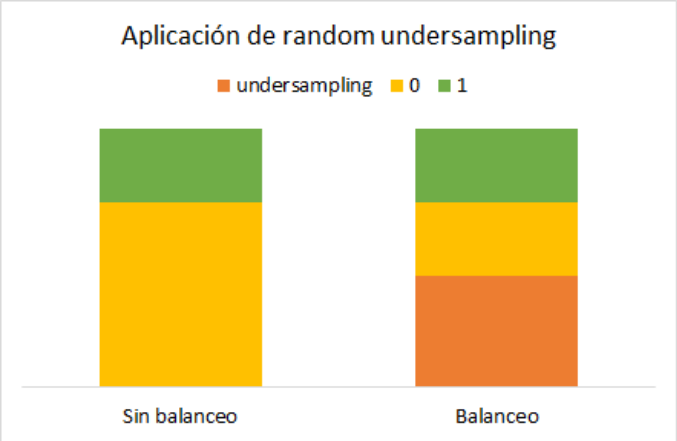



Figura 2.6: Técnica de Random Undersampling

- **Tomek Links** es una método desarrollado por Tomek en 1956 que surge a partir de una de una modificación de la técnica de submuestreo de vecinos más cercanos condensados (CNN) donde identifico ciertas desventajas sobre el mismo [45].

  En el estudio realizado por Tomek manifiesta que la aleatoriedad que maneja CNN podrían conservar instancias que no aportan información valiosa y que además, se podría conservar instancias internas de una clase en lugar de aquellas que se encuentran en los límites de decisión lo cual puede resultar beneficioso al momento de realizar el proceso de clasificación.

  De acuerdo a lo anterior, la técnica de Tomek Links surge de la siguiente manera:

  Sea $d(x_i, x_j)$ la distancia euclidiana entre $x_i$ y $x_j$, donde $x_i$ denota una muestra que pertenece a la clase minoritaria y $x_j$ denota una muestra que pertenece a la clase mayoritaria. Si no hay muestra, $x_k$ satisface la siguiente condición:

$$d(x_i, x_k) < d(x_i, x_j), \text{ o } d(x_j, x_k) < d(x_i, x_j),$$

  entonces el par de $(x_i, x_j)$ es un Enlace Tomek .

  Este método identifica los datos de la clase mayoritaria que está más cerca de los datos de la clase minoritaria y elimina los puntos cercanos al límite de decisión

entre ambas clases. Al eliminar estos puntos cercanos al limite, mejora la separación entre las clases y reduce la interferencia entre ellas. Una ilustración gráfica de esta técnica se encuentra a continuación:

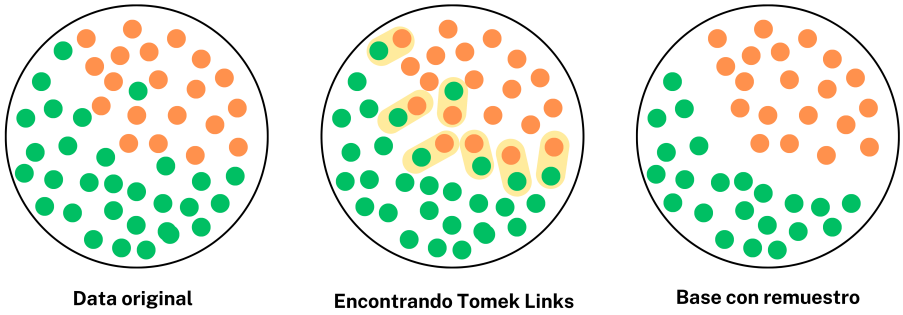



Figura 2.7: Técnica de Tomek Links. Elaboración propia.

La librería imbalanced-learn en Python proporciona la clase TomekLinks que se puede usar para llevar a cabo la eliminación de T̈omek linksën conjuntos de datos desequilibrados.

### 2.5.2. Técnicas de oversampling

Estas técnicas buscan equilibrar la distribución de las clases al aumentar artificialmente el número de instancias de la clase minoritaria o clases minoritarias, con el objetivo de mejorar el rendimiento del modelo en la predicción de dichas clases. Algunas de las técnicas de oversampling más comunes incluyen:

- **Random oversampling** es una técnica de oversampling que consiste en replicar aleatoriamente instancias de la clase minoritaria para equilibrar la distribución de clases en un conjunto de datos desbalanceado.

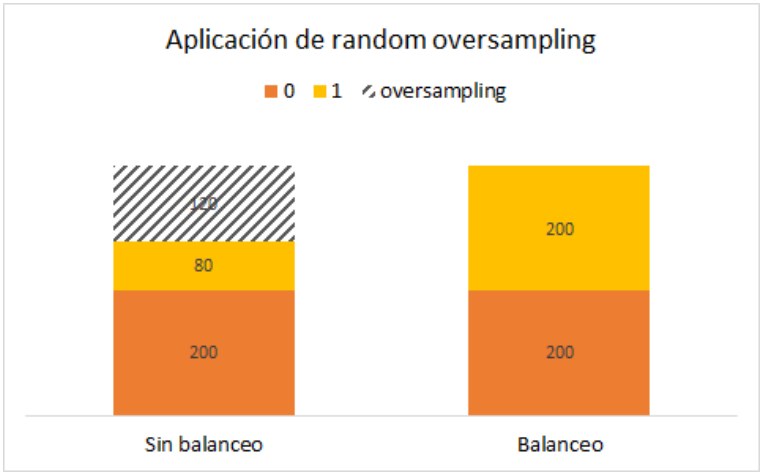



Figura 2.8: Técnica de Random Oversampling

- **Synthetic Minority Oversampling Technique (SMOTE)**

  Esta técnica de sobremuestro fue descrita por primera vez en el año 2002 [46] por Chawla y esta inspirada en investigaciones donde se realiza el sobremuestro con reemplazo. Sin embargo, al identificar que esta técnica no mejoraba significativamente el reconocimiento de la clase minoritaria entonces, surge SMOTE una técnica que selecciona aleatoriamente uno o más de los k-vecinos más cercanos para cada ejemplo en la clase minoritaria se eligen N de estas K instancias para interpolar nuevas instancias sintéticas [47].

  El algoritmo SMOTE funciona en cuatro sencillos pasos:

  - Elija una clase minoritaria como vector de entrada.
  - Encuentre sus k vecinos más cercanos (k-neighbors se especifica como argumento en la función SMOTE()).
  - Elija uno de estos vecinos y coloque un punto sintético en cualquier lugar de la línea que une el punto en consideración y su vecino elegido.
  - Repita los pasos hasta que los datos estén equilibrados.

Una ilustración gráfica del algoritmo descrito anteriormente se encuentra en la Figura 2.9 donde se puede evidenciar que el proceso de balanceo se ha realizado correctamente, puesto que el numero de muestras minoritarias ha igualado al de mayoritarias a partir de la creación de datos sintéticos al considerar los $k$ vecino más cercanos.

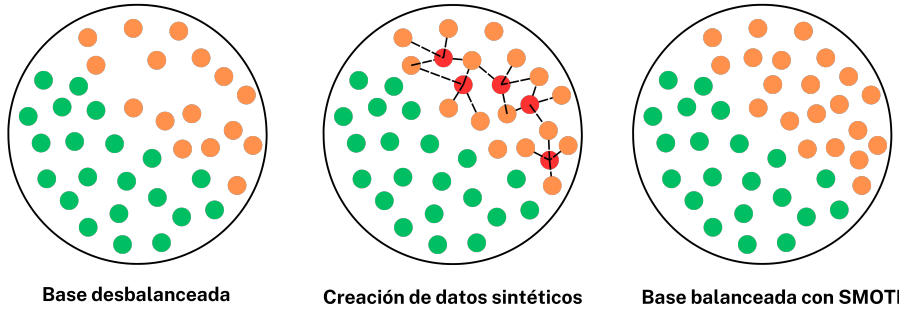



Figura 2.9: Técnica SMOTE. Elaboración propia.

A pesar de que los métodos aleatorios demuestran buenos rendimientos en los problemas de clasificación realmente los métodos heurísticos evitan y corrigen los sesgos de eliminación de información ya que estos, pretenden eliminar aquellos casos con escasos aporte a la variabilidad entre clases.

## 2.6. Hiperparámetros

De acuerdo con Bissuel [48] la optimización de hiper parámetros hace referencia al ajuste de los parámetros de un modelo de aprendizaje automático. Estos hiperparámetros

son esenciales en la configuración de un modelo, ya que representan aquellos parámetros que no se actualizan durante el proceso de aprendizaje. Estos parámetros son fijados antes de que el modelo comience a aprender a partir de los datos y desempeñan un papel crucial en en términos de rendimiento y precisión del modelo.

De acuerdo a lo anterior, existen diferentes técnicas para encontrar los mejores hiperparametros para el modelo que sea desea entrenar, entre ellos se encuentra:

- **Búsqueda aleatoria (Random Search)**: Define el espacio de búsqueda como un dominio acotado de valores de hiperparámetros y muestra aleatoriamente puntos en ese dominio.
- **Búsqueda de cuadrícula (Grid Search)**Define el espacio de búsqueda como una cuadrícula de valores de hiperparámetros y evalúa cada posición en la cuadrícula.

Sin embargo, el ajuste de hiperparámetros también puede llevarse a cabo de forma manual o mediante diversas técnicas diseñadas para prevenir sesgos o imprecisiones, teniendo en cuenta el factor de subjetividad que puede introducirse al realizar ajustes basados en la intuición o la experiencia del programador.

Existen diferentes puntos de vista sobre la selección del método de ajuste de hiperparametros Bergstra y Bengio en su articulo [49] muestran empírica y teóricamente que las pruebas elegidas al azar son más eficientes para la optimización de hiperparámetros que las pruebas en una cuadrícula. En el enfoque de búsqueda aleatoria, se destaca la ventaja de una mayor precisión en la identificación de óptimos para cada hiperparámetro cuando estos están poco correlacionados entre sí.

Para el desarrollo de este trabajo, se ha decidido implementar una búsqueda aleatoria con el objetivo de optimizar los parámetros de los modelos seleccionados. En este sentido, se empleará la función RandomizedSearch la cual es una clase de la biblioteca scikitlearn en Python. Pertenece al módulo sklearn.model_selection y se utiliza para realizar una búsqueda aleatoria de hiperparámetros mediante validación cruzada [50].

### 2.6.1. Hiperpárametros para Random Forest

Para el modelo de Random Forest se optimizaran los siguientes hiperpárametros:

- n_estimators: número de árboles que construye el algoritmo antes de promediar las predicciones.
- max_depth: Controla la profundidad máxima de los árboles individuales en el bosque.
- max_features: Número máximo de características que el bosque aleatorio considera al dividir un nodo.

- criterion: Técnica que usara para dividir cada nodo del árbol.
- bootstrap: Estrategia que implica generar múltiples muestras de datos a partir del conjunto de entrenamiento original mediante el muestreo con reemplazo.
- min_samples_leaf: Determina el número mínimo de hojas necesarias para dividir un nodo interno.
- min_samples_split: Controla el número mínimo de muestras requeridas para dividir un nodo interno (no hoja) en un árbol del bosque.

### 2.6.2. Hiperpárametros para XGBoost

Para el modelo de XGBoost se optimizaran los siguientes hiperpárametros:

- n_estimators: Define el número de árboles que se deben construir en el ensamble.
- max_depth: Profundidad máxima de un árbol. Se utiliza para controlar el sobreajuste.
- min_child_weight: Controla la cantidad mínima de instancias (muestras) necesarias en cada hoja del árbol.
- learning_rate: Controla la tasa de aprendizaje o la contribución de cada árbol al modelo general.
- subsample: Controla la fracción de las observaciones (muestras) que se utiliza para entrenar cada árbol.
- colsample_bytree: controla la fracción de características que se deben considerar al construir cada árbol en el conjunto.
- gamma: Especifica la reducción mínima de pérdida requerida para realizar una división.

## 2.7. Métricas de evaluación

Dado que no nos limitamos únicamente a probar un algoritmo para predecir la probabilidad de incumplimiento en el pago de las obligaciones crediticias de un prestatario, es fundamental recurrir a métricas de evaluación.

Estas métricas juegan un papel esencial al permitirnos medir la precisión de nuestro modelo y evaluar su desempeño a lo largo del proceso de entrenamiento para así mismo, tomar decisiones sobre qué modelo es más efectivo para el caso en estudio.

A continuación, se presentan diferentes métricas para evaluar modelos de aprendizaje automático:

- **Matriz de confusión:**
  Esta métrica brindan información sobre cuántos aciertos y errores cometió el modelo en sus predicciones y son esenciales para comprender cómo se desempeña en la tarea de clasificación.

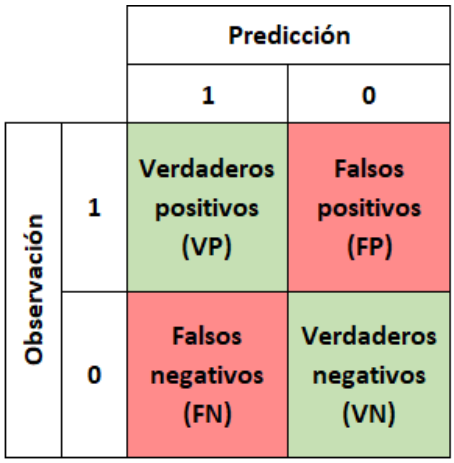



Figura 2.10: Matriz de confusión

  - Verdaderos positivos (VP): Son aquellos valores positivos predichos por el modelo que coinciden con los valores reales.
  - Verdaderos negativos (VN): Son aquellos valores negativos predichos por el modelo que coinciden con los valores reales.
  - Falsos positivos (FP): Casos en los que el modelo predijo incorrectamente que un elemento pertenece a la clase positiva cuando en realidad pertenece a la clase negativa.
  - Falsos negativos (FN): Casos en los que el modelo predijo incorrectamente que un elemento pertenece a la clase negativa cuando en realidad pertenece a la clase positiva.

  A partir de la matriz de confusión, se pueden calcular diversas métricas de evaluación que proporcionan información adicional sobre el rendimiento de un modelo de clasificación por ejemplo:

- **Precisión:** Mide la proporción de predicciones positivas correctas con respecto al número total de predicciones positivas.

$$\text{Precisión} = \frac{VP}{VP + FP}$$

- **Exhaustividad (Recall o Sensibilidad):** Mide la proporción de casos positivos reales que se predijeron correctamente con respecto al número total de casos positivos reales.

$$\text{Recall} = \frac{VP}{VP + FN}$$

- **Accuracy:** Proporciona una medida de cuántas de las predicciones totales realizadas por el modelo son correctas en relación con el número total de predicciones.

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Curva AUC-ROC:** Métrica numérica que permite cuantificar la calidad global de un modelo de clasificación y se crea trazando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR). Cuanto mayor sea el valor de AUC-ROC, mejor será el rendimiento del modelo en la tarea de clasificación.

- **F1-Score:** Es una métrica de evaluación en problemas de clasificación que combina la precisión (precision) y la exhaustividad (recall). Es útil cuando se busca un equilibrio entre la calidad de las predicciones positivas y la capacidad de identificar casos positivos reales.

$$\text{F1 Score} = \frac{2 \times \text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Se ha observado en varios estudios que utilizan el Área Bajo la Curva de Característica Operativa del Receptor (AUC-ROC) como métrica de evaluación para sus algoritmos de aprendizaje. Sin embargo, es importante señalar que esta métrica puede no ser la más apropiada cuando se enfrenta un desequilibrio significativo entre las clases. Lo anterior se debe a que la métrica se enfoca en la tasa de falsos positivos y verdaderos positivos, sin tener en cuenta la distribución real de las clases en el conjunto de datos, es decir, puede dar una impresión engañosa de un buen rendimiento cuando en realidad el modelo está fallando en la clasificación de la clase minoritaria (la tasa de verdaderos negativos) [51] en estos casos, se prefiere utilizar métricas como el F1-score ya que equilibra la precisión y la exhaustividad.

# –3–
# Metodología

La metodología de este Trabajo de Grado será desarrollada por medio CRISP-DM (Cross Industry Standard Process for Data Mining) o Modelo de Proceso Estándar para Minería de Datos, la estructura el ciclo de vida de un proyecto de explotación de datos en seis fases cuya sucesión no es rígida e interactúan entre ellas de forma iterativa durante el desarrollo del proyecto (ver [52] y [53] para más detalles). A continuación, se explican las fases en que se divide CRISP-DM:

- **Fase 1**: Comprensión del negocio o problema: Se centra en la comprensión de los objetivos y requisitos del proyecto. Importante desarrollar una estrategia inicial que no solo proporcione dirección, sino que también establezca las bases para la consecución de las metas predefinidas del proyecto.

- **Fase 2**: Comprensión de los datos: Comienza con una recopilación y exploración inicial de los datos con el objetivo de familiarizarse con ellos y establecer un primer contacto con el problema con el fin de garantizar la calidad y relevancia de los datos utilizados.

- **Fase 3**: Preparación de los datos: Esta fase de preparación trata de seleccionar, limpiar y generar conjuntos de datos correctos, organizados y preparados para la fase de modelado. Los errores en los datos que se pasan por alto y que no son resueltos en esta fase se trasladan hasta la fase de modelado, lo que genera una reducción en la exactitud de los modelos.

- **Fase 4**: Modelado de datos: Diversas técnicas de modelado son seleccionadas y aplicadas de acuerdo con el problema que se está trabajando para obtener calidad y eficacia antes de implementarlos para su uso. En esta etapa sus parámetros son calibrados a valores óptimos para obtener los mejores resultados posibles.

- **Fase 5**: Evaluación del modelo: Una evaluación detallada del modelo y la revisión de los pasos ejecutados para construir el modelo para asegurar que se han alcanzado los objetivos de negocio.

- **Fase 6**: Despliegue o implementación: El conocimiento o resultado adquirido debe ser documentada y presentada para que el cliente pueda usarla, se debe de elaborar el plan de implementación, monitoreo y mantenimiento.

De acuerdo con lo anterior, es importante resaltar que, teniendo en cuenta los alcances establecidos para este trabajo de grado, se adopto la metodología CRISP-DM hasta la fase de evaluación. Por consiguiente, la representación gráfica de la metodología se presenta de la siguiente manera:

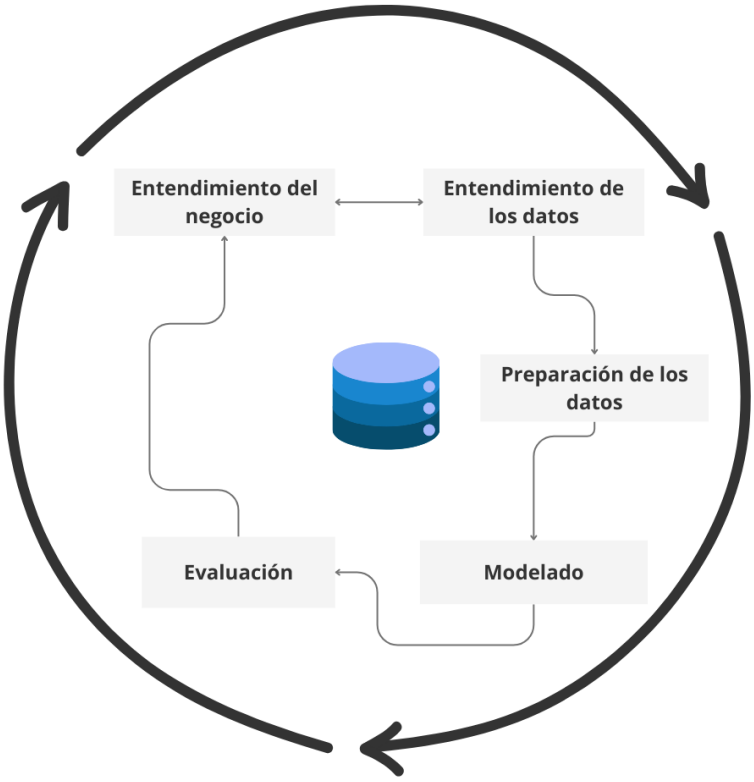



Figura 3.1: Metodología adoptada de CRISP-DM.

# –4–
# Desarrollo de la metodología

## 4.1. Entendimiento del negocio

Considerando la regulación bancaria de Basilea y tomando en consideración las normativas establecidas por la Superintendencia Financiera de Colombia (SFC), las cuales destacan la importancia de la gestión del riesgo crediticio, este estudio se centrará en modelos de clasificación. Estas técnicas son ampliamente empleadas por las entidades financieras para respaldar sus decisiones en la asignación de créditos. Se aplicarán diversas metodologías, como Regresión Logística, Random Forest y XGBoost, con el objetivo de evaluar la posibilidad de que un prestatario de una entidad financiera incumpla con sus obligaciones de pago.

## 4.2. Comprensión de los datos

En el ámbito de las entidades financieras, las bases de datos desempeñan un papel fundamental estas se convierten en pilares esenciales para la gestión de riesgos, la toma de decisiones y el diseño de estrategias efectivas. De acuerdo a las condiciones económicas cambiantes de los últimos tiempos la transformación digital dentro de las instituciones financieras es de vital importancia ya que a partir de un buen uso de estos se pueden tomar decisiones mas precisas y oportunas [54].

Para este estudio es importante contar con una base de datos que permita modelar el riesgo crediticio. Sin embargo, el acceso a estas bases de datos se encuentra restringido y regulado para garantizar la seguridad y privacidad de los datos ya que en estos albergan una gran cantidad de información sobre los clientes, transacciones financieras, historiales crediticio y más.

En cierto punto esta llega a ser una de las limitaciones del estudio sin embargo, a través de una revisión exhaustiva se logra obtener una base de datos a partir de la cual se pretende tener una representación precisa de la realidad donde se identifiquen las variables signifi-

cativas que influyen en la probabilidad de incumplimiento de pago por parte de los clientes.

La base de datos a utilizar fue tomada de la plataforma Kaggle y el notebook donde esta se puede encontrar recibe el nombre de analise de credito. El objetivo es identificar el perfil y las condiciones financieras del cliente, a partir de su información básica y datos financieros para determinar el riesgo crediticio del cliente.

En esta etapa se realizara una exploración descriptiva de los datos para entender la información de cada uno de sus atributos antes de realizar el entrenamiento de los modelos. Esta base de datos cuenta con 10.127 datos históricos de prestatarios y se divide en 8.500 registros etiquetados como no moros y 1.627 registros etiquetados como morosos. Este conjunto de datos cuenta con una variable predictora llamada default y 15 variables explicativas (características).

| Variable | Descripción | Tipo de dato |
|---|---|---|
| Id | Número de cuenta | Numérica |
| Edad | Edad de los clientes | Numérica |
| Sexo | Masculino (0) Femenino (1) | Categórica |
| Personas a cargo | Número de personas que el cliente tiene a cargo | Numérica |
| Escolaridad | Escolaridad del cliente | Categórica |
| Estado civil | Estado civil del cliente | Categórica |
| Salario anual | Monto total recibido en el año del cliente | Categórica |
| Tipo tarjeta | Tipo de tarjeta asociada al cliente | Categórica |
| Meses con tarjeta | Número de meses que el cliente ha tenido una cuenta | Numérica |
| Num productos | Número de productos que tiene el cliente con el banco | Numérica |
| Iteraciones 12m | Número de iteraciones del cliente con el banco | Numérica |
| Meses inactivos 12m | Número de meses que el cliente ha estado inactivo | Numérica |
| Valor transacción 12m | Promedio de las transferencias desde la cuenta del titular | Numérica |
| Valor disponible | Saldo que el cliente tiene disponible con el banco | Numérica |
| Num transacciones 12m | Número de transferencias en los últimos 12 meses | Numérica |
| Default | Clasificación de los clientes | Numérica |

Cuadro 4.1: Descripción de las variables de la base de datos

## 4.3. Preparación de los datos

Se realizó un proceso de renombramiento en la base de datos, que incluyó tanto las columnas como algunas variables categóricas, con el objetivo de cambiar los registros en portugués a español. Esto se llevo a cabo para facilitar la comprensión y manipulación de los datos de manera consistente.

Al analizar la base de datos a través de la variable de identificación (ID), se confirmó la ausencia de registros duplicados. Sin embargo, se presentan cierta cantidad de registros con valores nulos dentro de la base de datos. Como se puede evidenciar en la Tabla 4.1, la variable con mayor proporción de datos faltantes es escolaridad con un 15 %, salario anual con un 15 % y finalmente, el estado civil con un 7.4 % de datos faltantes.

| Variable | Valores faltantes | Porcentaje |
|---|---|---|
| escolaridad | 1519 | 15 % |
| salario_anual | 1112 | 11 % |
| estado_civil | 749 | 7.4 % |

Cuadro 4.2: Descripción de la tabla

Existen diferentes técnicas para el manejo de los valores nulos dentro de la base de datos y una de las más utilizadas es la imputación. Esta técnica implica la estimación o asignación de valores a las entradas faltantes en función de la información disponible en el conjunto de datos. Hay varias estrategias de imputación, como la imputación media, que utiliza la media de los valores conocidos para llenar los faltantes, o la imputación basada en técnicas mas avanzadas como KNN o la imputación de características multivariadas.

En el desarrollo de este proyecto, se utilizará el método de los k vecinos más cercanos (KNN), empleando como parámetro la consideración de los cinco vecinos más cercanos. Este enfoque se presenta como una estrategia efectiva para abordar la presencia de datos faltantes en nuestro conjunto de datos. La métrica de distancia seleccionada, denominada *'nan_euclidean'*, ha sido escogida con precisión para gestionar de manera eficiente los valores ausentes en el espacio euclidiano. Simultáneamente, al configurar los pesos en 'uniforme', se garantizará una imputación equitativa, asegurando que cada vecino contribuya de manera equitativa en el proceso de imputación.

A continuación, se presentan los métodos empleados para la imputación de valores faltantes en las tres variables mencionadas anteriormente. El objetivo fue evaluar la alteración en la distribución de los datos tras la incorporación de nuevos valores mediante estas técnicas de imputación, con la finalidad de determinar la metodología más adecuada.

En la figura 4.1 se puede observar que aplicando KNN se obtiene una distribución similar a la inicial pues, para otro métodos se observa un aumento en la escolaridad nivel 3 de los no morosos.

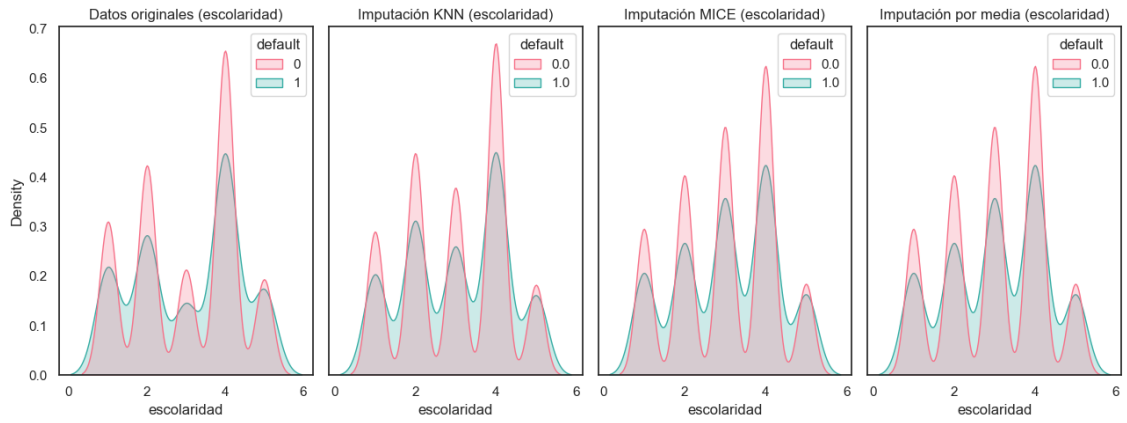



Figura 4.1: Imputación de valores faltantes en la variable escolaridad con diferentes métodos

De acuerdo con la figura 4.2 se puede identificar que los diferentes métodos utilizados logran mantener una distribución similar a la de los datos originales.

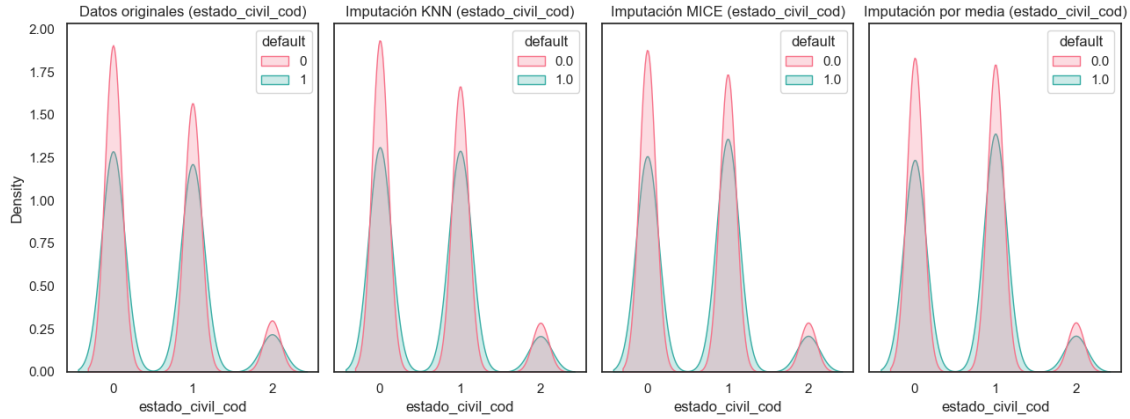



Figura 4.2: Imputación de valores faltantes en la variable estado civil con diferentes métodos

Según la figura 4.3, se observa que el método que preserva mejor la distribución es el basado en KNN. Se puede notar que en los otros métodos, hay un aumento o picos más pronunciados en los valores de las variables.

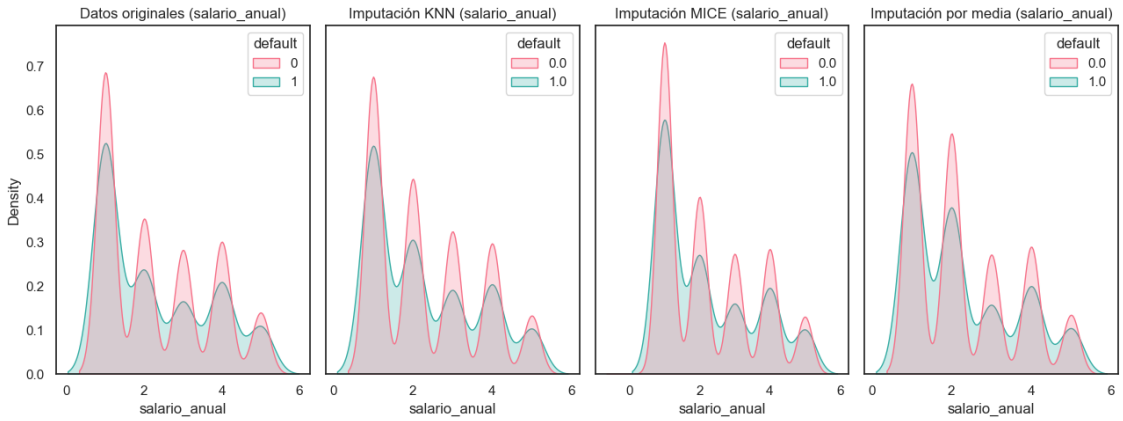



Figura 4.3: Imputación de valores faltantes en la variable salario anual con diferentes métodos.

### 4.3.1. Análisis descriptivo de los datos

- **default**: indica la clasificación del cliente entre moroso (1) y no moroso (0).

  Al examinar gráficamente la variable objetivo, podemos observar que el 83.93 % de los clientes son no morosos (categorizados como cero) mientras que, el 16.07 % de los clientes son no morosos (categorizados como uno).

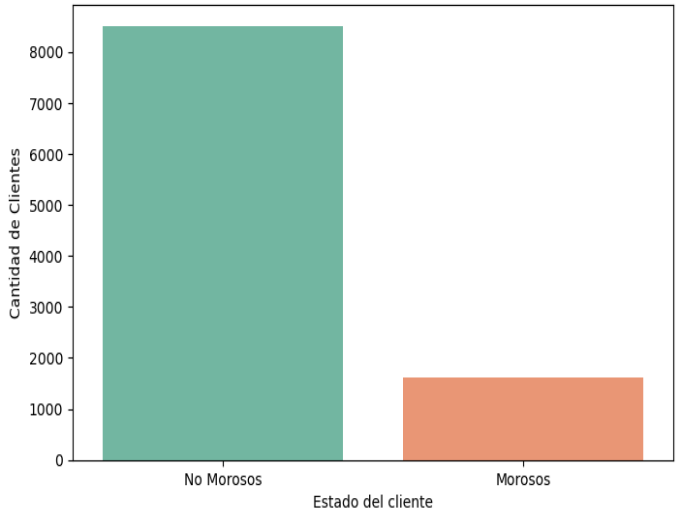



Figura 4.4: Clasificación de los clientes.

En la Figura 4.4 el comportamiento exhibido por la variable "default"será fundamental para llevar a cabo el desarrollo de este trabajo, ya que nos enfocaremos en la aplicación de técnicas de balanceo de datos mediante el modelo de Regresión Logística, Random Forest y XGBoost.

- **sexo**: indica la categoría de género de cada individuo.

| Categoría | Cantidad | Porcentaje |
|---|---|---|
| F | 5358 | 53 % |
| M | 4796 | 47 % |

Cuadro 4.3: Frecuencia observada por la variable sexo.

- **escolaridad**: se refiere al nivel educativo alcanzado por cada individuo en la base de datos. Se observa que la mayoría de los clientes han completado estudios de maestría, mientras que un 9 % de los clientes han alcanzado el nivel de doctorado.

| Variable | Cantidad | Porcentaje |
|---|---|---|
| maestría | 3422 | 34 % |
| escuela secundaria | 2301 | 23 % |
| graduado | 1938 | 19 % |
| sin educación formal | 1490 | 15 % |
| doctorado | 976 | 9 % |

Cuadro 4.4: Distribución de frecuencias según la variable escolaridad.

- **estado_civil**: Se observa que la mayoría de los clientes se encuentran casados o solteros mientras que, un 7 % de los clientes se encuentran divorciados.

| Variable | Cantidad | Porcentaje |
|---|---|---|
| casado | 4989 | 50 % |
| soltero | 4390 | 43 % |
| divorciado | 748 | 7 % |

Cuadro 4.5: Distribución de frecuencias según la variable estado_civil.

- **tipo_tarjeta**: esta variable representa diferentes niveles o categorías de tarjetas de crédito, cada una puede estar asociada con un grado particular de prestaciones o beneficios, lo cual es relevante para evaluar la capacidad de pago de los clientes.

| Variable | Cantidad | Porcentaje |
|---|---|---|
| blue | 9436 | 93.2 % |
| silver | 555 | 5.5 % |
| gold | 116 | 1.1 % |
| platinum | 20 | 0.2 % |

Cuadro 4.6: Distribución de frecuencias según la variable tipo_tarjeta

La información representada en la Figura 4.5 destaca que la mayoría de los clientes están asociados a una tarjeta de tipo blue. Sin embargo, es relevante señalar que, entre los clientes morosos, cinco de ellos están vinculados a una tarjeta platinum.

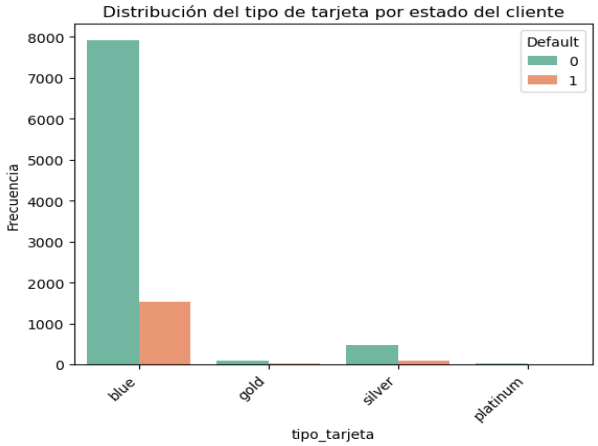



Figura 4.5: Distribución de frecuencias según la variable tipo_tarjeta.

- **salario anual**: esta variable permite identificar el ingreso total que un individuo recibe durante un año este campo es de vital importancia ya que permite analizar la situación financiera anual de los individuos dentro del conjunto de datos.

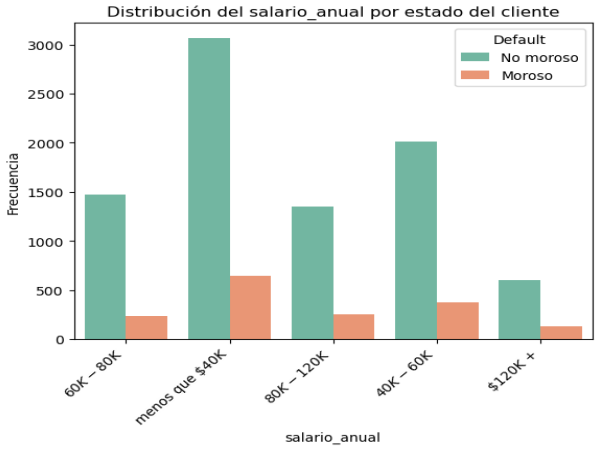



Figura 4.6: Distribución de frecuencias según la variable salario_anual.

En la Figura 4.6 se observa que la mayor parte de los clientes presenta ingresos anuales por debajo de las 40,000 unidades monetarias. Cabe destacar que, aunque los clientes morosos están dispersos en distintos niveles salariales, su presencia

es notablemente menor en la categoría de aquellos que ganan 120,000 unidades monetarias o más.

- **personas_a_cargo**: esta variable permite identificar el número de personas que dependen económicamente de los clientes de la base de datos. Además,

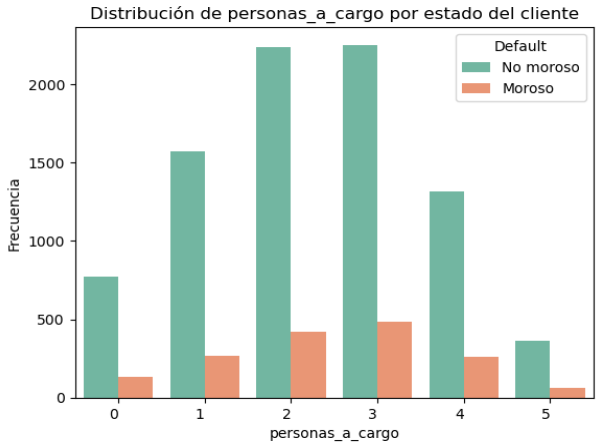



Figura 4.7: Distribución de las variables personas_a_cargo.

En la Figura 4.7, se destaca que la mayoría de los clientes tienen entre 2 y 3 personas dependientes económicamente de ellos. Curiosamente, los clientes morosos con 5 personas a su cargo exhiben la frecuencia más baja. Por lo tanto, no se puede concluir que aquellos clientes con menos personas a su cargo sean necesariamente mejores pagadores, ya que la presencia de morosidad varía incluso en segmentos de menor carga familiar.

- **Edad**:

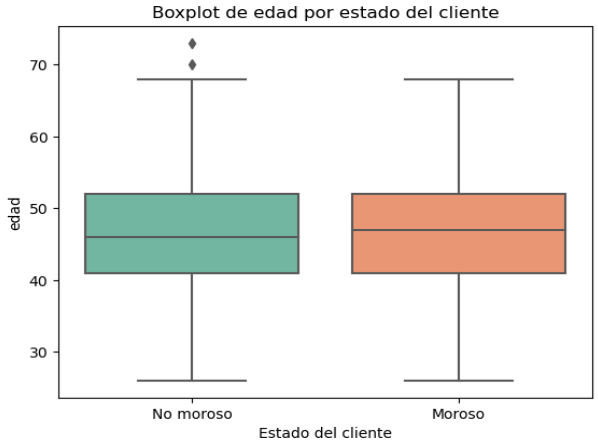



Figura 4.8: Boxplot de la edad con respecto al estado del cliente.

En la Figura 4.8 se presentan una serie de valores atípicos superiores aproximadamente a 65 años para clientes no morosos sin embargo, se mantienen estos registros ya que hacen parte de datos particulares de clientes de la base de datos.

- **meses_con_tarjeta**: se refiere al número de meses que el cliente ha tenido una cuenta con el banco.

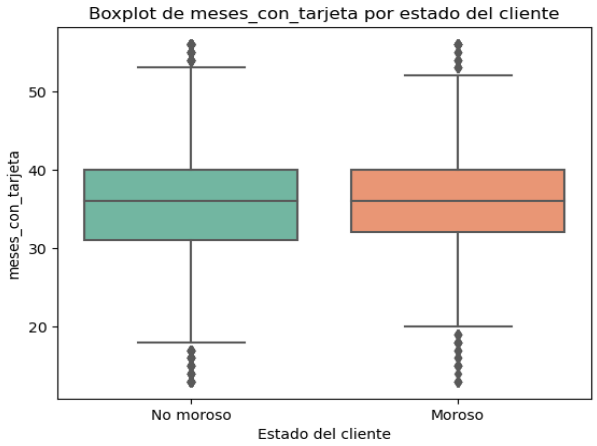



Figura 4.9: Boxplot de la edad con respecto al estado del cliente.

En la Figura 4.9 se puede evidenciar cierta variabilidad en la duración de la tarjeta la cual es coherente con nuestras expectativas, considerando las distintas características de los clientes. Algunos poseen tarjetas recientemente adquiridas, reflejando los 13 meses mínimos, mientras que otros demuestran una larga historia financiera con la entidad, evidenciada por el máximo de 56 meses.

- **valor_disponible**: hace referencia al saldo que el cliente tiene disponible con el banco

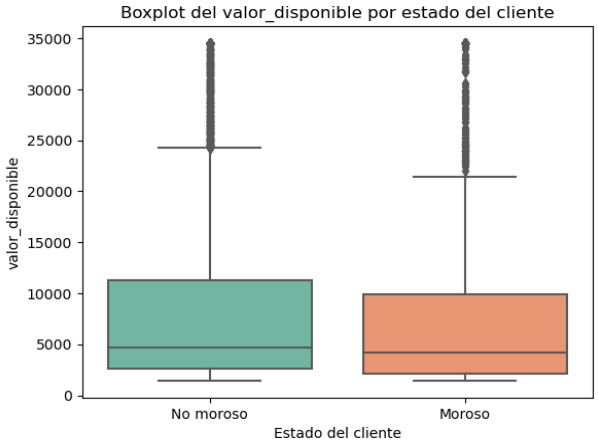



Figura 4.10: Boxplot de la edad con respecto al estado del cliente.

A partir de la información proporcionada en la Figura 4.10, se puede observar que, en promedio, los clientes no morosos disponen de un mayor saldo disponible con el banco. Aunque se presentan algunos valores atípicos, es importante tener en cuenta que estos casos particulares pueden deberse a situaciones en las que, a pesar de tener un límite de crédito total, el cliente no ha utilizado completamente dicho límite, posiblemente debido a su capacidad financiera disponible.

- **num_productos**: hace referencia al número de productos que tiene el cliente con el banco

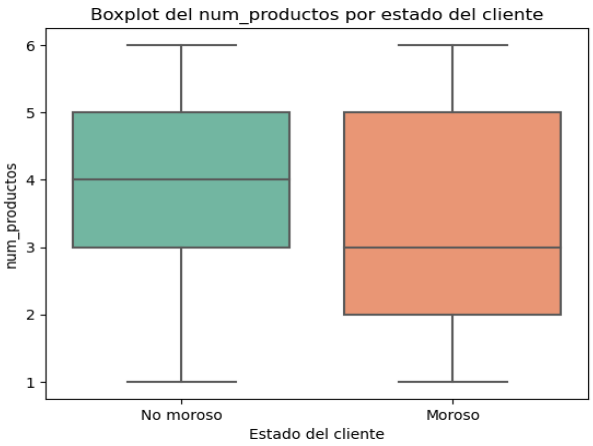



Figura 4.11: Boxplot del num_productos con respecto al estado del cliente.

La Figura 4.11 revela que, en promedio, los clientes sin morosidad cuentan con hasta cuatro productos vinculados al banco, en comparación con los clientes morosos que, en promedio, tienen hasta tres productos. Es importante destacar que el boxplot para la categoría de clientes morosos presenta una amplia distribución de los datos. Esto sugiere que, aunque la media es menor, la presencia de morosidad abarca un rango diverso de situaciones en cuanto al número de productos que los clientes mantienen con el banco.

- **iteraciones_12m**: hace referencia al número de iteraciones del cliente con el banco en los últimos 12 meses. La Figura 4.12 revela que los clientes morosos fueron aquellos que registraron una mayor cantidad de interacciones con el banco, a diferencia de los no morosos que presentaron un nivel inferior de interacción. Este patrón sugiere la posibilidad de que la intensidad de las interacciones con la entidad financiera esté vinculada a la presencia de morosidad, señalando que aquellos con una mayor participación podrían enfrentar mayores desafíos en términos de cumplimiento de pagos.

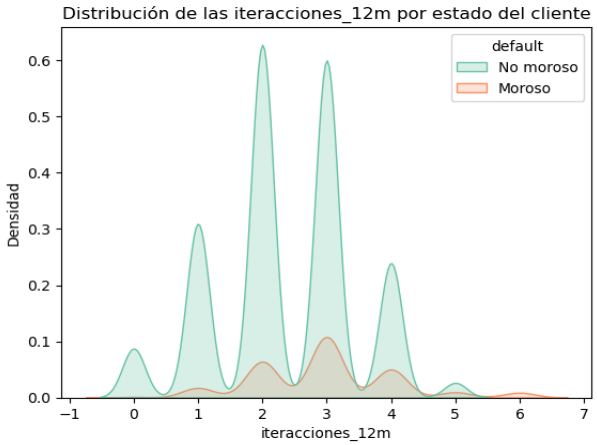



Figura 4.12: Distribución de las iteraciones de los últimos 12 meses con respecto al estado del cliente.

- **num_transacciones_12m**: hace referencia al número de transferencias en los últimos 12 meses realizados por el cliente.

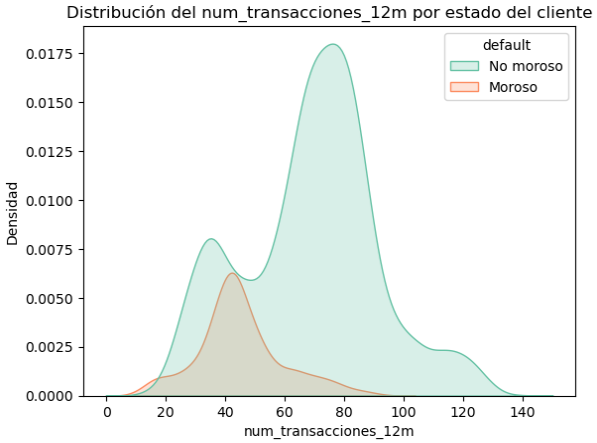



Figura 4.13: Boxplot del num_productos con respecto al estado del cliente.

Se observa en la Figura 4.13 que los clientes sin morosidad destacan por tener un mayor número de transacciones a lo largo del año. En contraste, los clientes morosos muestran un rango más limitado, concentrándose en realizar entre 35 y 50 transacciones.

### 4.3.2. Codificación

Se realiza un proceso de codificación a variables categóricas tales como el sexo, escolaridad, estado_civil, tipo_tarjeta y salario_anual con el fin de transformarlas en formatos

numéricos comprensibles para los algoritmos de modelado. Para la codificación de estas variables se realizo la implementación de diferentes técnicas:

- La **codificación one-hot** fue aplicada a la variable sexo dado que dicho campo en la base de datos seleccionada es de naturaleza categórica binaria, representada por dos categorías distintas,en este caso, F y M.

| Categoría | Codificación |
|---|---|
| F (Femenino) | 0 |
| M (Masculino) | 1 |

- La **codificación ordinal** fue aplicada a las variables de escolaridad, tipo_tarjeta y salario_anual ya que esta técnica es la mas apropiada cuando existe una jerarquía natural entre las categorías.

  **Variable escolaridad:**

| Categoría | Codificación |
|---|---|
| sin educación formal | 1 |
| escuela secundaria | 2 |
| graduado | 3 |
| maestría | 4 |
| doctorado | 5 |

  **Variable tipo_tarjeta:**

| Categoría | Codificación |
|---|---|
| blue | 0 |
| silver | 1 |
| gold | 2 |
| platinum | 3 |

  **Variable salario_anual:**

| Categoría | Codificación |
|---|---|
| menos que $40K | 1 |
| $40K - $60K | 2 |
| $60K - $80K | 3 |
| $80K - $120K | 4 |
| $120K | 5 |

- La **codificación de frecuencias** fue implementada en las variable de estado_civil dado que esta técnica asigna valores numéricos más altos a las categorías que aparecen con mayor frecuencia y valores numéricos más bajos a las categorías

menos frecuentes lo permite destacar la prevalencia de cada categoría en la base de datos.

**Variable estado_civil:**

| Categoría | Codificación |
|---|---|
| casado | 0 |
| soltero | 1 |
| divorciado | 2 |

## 4.4. Modelado

Las técnicas de modelado seleccionadas se encuentran detalladas en el Capítulo 2 de este trabajo. En la fase actual de modelado, la metodología a implementar implica una división de la base de datos, asignando un 70 % para el conjunto de entrenamiento (train) y reservando el 30 % restante para el conjunto de prueba (test).

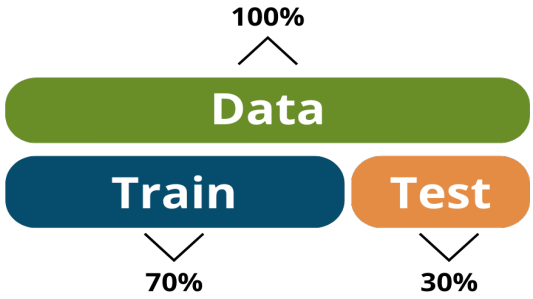



Figura 4.14: División de la base de datos.

Como resultado del proceso de división de la base de datos, las etiquetas de clientes morosos y no morosos en cada uno de los conjuntos de entrenamiento y prueba quedaron definidas de la siguiente manera:

| Base | Moroso (1) | No moroso (0) | Total |
|---|---|---|---|
| $Y_{train}$ | 1.156 | 5.932 | 7.088 |
| $Y_{test}$ | 471 | 2.568 | 3.039 |

### 4.4.1. Metodología para datos desbalanceados

Tal como se evidencio en la Sección 1.3, el desequilibrio entre las clases puede tener repercusiones en el rendimiento de los clasificadores, ya que los algoritmos que no abordan adecuadamente esta desigualdad tienden a dar prioridad a la clase mayoritaria mientras pasan por alto la clase minoritaria [43].

Por lo tanto, a continuación, se detallan las diversas técnicas de balanceo de datos que fueron seleccionadas para el desarrollo de este trabajo. Es importante aclarar que el balanceo de datos se llevó a cabo exclusivamente en la conjunto de entrenamiento.

## Oversampling

Para implementar esta metodología, se hizo uso de la librería imblearn.over_sampling y se aplicó la función RandomOverSampler. De acuerdo a su funcionamiento esta técnica iguala las clases teniendo en cuenta la clase mayoritaria, como resultado de este proceso, la división final de la base de datos se presenta de la siguiente manera:

| Moroso (1) | No moroso (0) | Total |
|---|---|---|
| 5932 | 5932 | 11.864 |

## Undersampling

Para implementar esta metodología, se hizo uso de la librería imblearn.over_sampling y se aplicó la función RandomOverSampler. De acuerdo a su funcionamiento esta técnica iguala las clases teniendo en cuenta la clase minoritaria, como resultado de este proceso, la división final de la base de datos se presenta de la siguiente manera:

| Moroso (1) | No moroso (1) | Total |
|---|---|---|
| 1156 | 1156 | 2.312 |

## SMOTE

Para implementar esta metodología, se hizo uso de la librería imblearn.under_sampling y se importo la función TomekLinks. De acuerdo a su funcionamiento esta técnica busca equilibrar las clases generando muestras sintéticas de la clase minoritaria por tanto, la división final de la base de datos se presenta de la siguiente manera:

| Moroso (1) | No moroso (0) | Total |
|---|---|---|
| 5932 | 5932 | 11.864 |

## Tome-Links

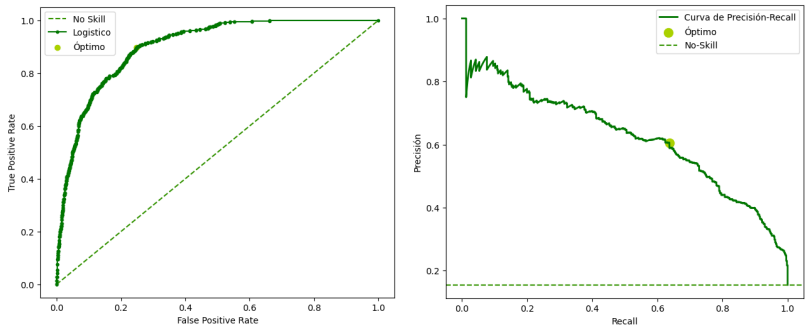
Para implementar esta metodología, se hizo uso de la librería imblearn.under_sampling y se importo la función TomekLinks. De acuerdo a su funcionamiento esta técnica identifica los datos de la clase mayoritaria que está más cerca de los datos de la clase minoritaria para posteriormente eliminarlos. Como resultado de este proceso la división final de la base de datos se presenta de la siguiente manera:

| Moroso (1) | No moroso (0) | Total |
|---|---|---|
| 1156 | 5696 | 6.852 |

Tras segmentar la base de datos con un 70 % para entrenamiento y un 30 % para prueba, manteniendo esta configuración tanto en condiciones estándar como en la aplicación de técnicas de balanceo, se inicia con proceso de modelado.

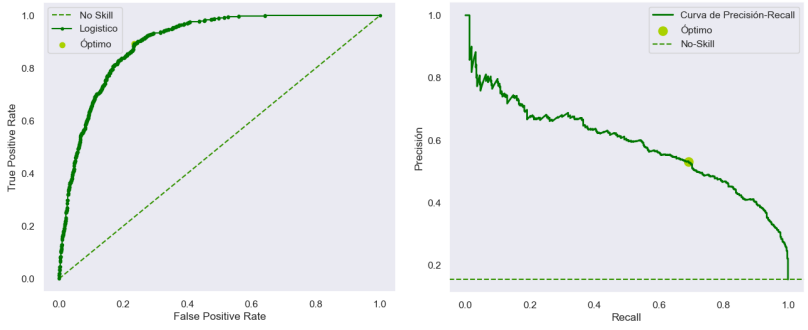
### 4.4.2. Entrenamiento del modelo de regresión logística

Se empleó la biblioteca Statsmodels en Python para implementar el modelo de regresión logística (ML). Al ajustar un modelo logit con Statsmodels, el procedimiento se enfoca en identificar los parámetros del modelo que maximizan la verosimilitud de los datos observados dentro del contexto del modelo logit. Para evaluar su rendimiento en cada configuración, se obtuvieron métricas como la Curva ROC y el F1-Score.



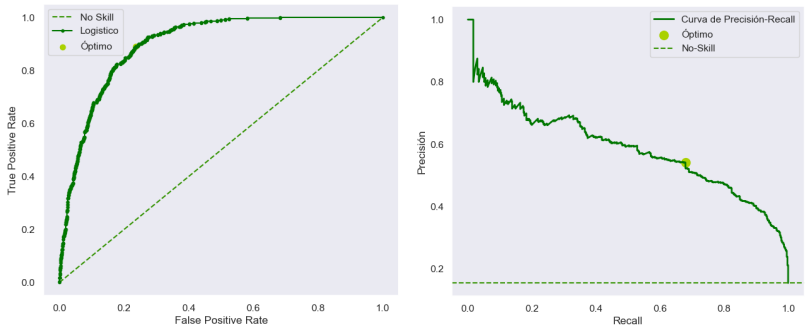

(a) Curva ROC. (b) Curva F1-Score.

Figura 4.15: Puntos de corte para el modelo de regresión logística.



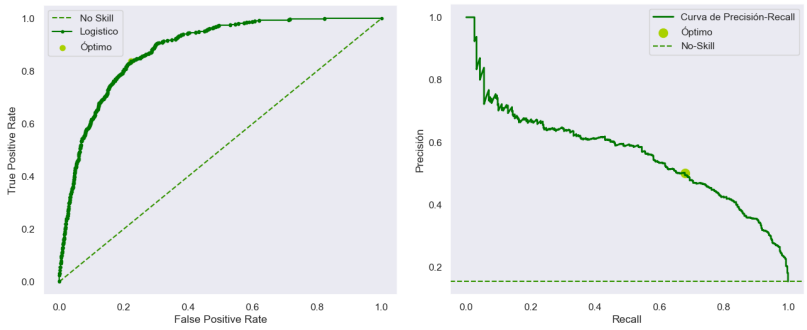

(a) Curva ROC. (b) Curva F1-Score.

Figura 4.16: Puntos de corte para el ML con la técnica de Random Oversampling.



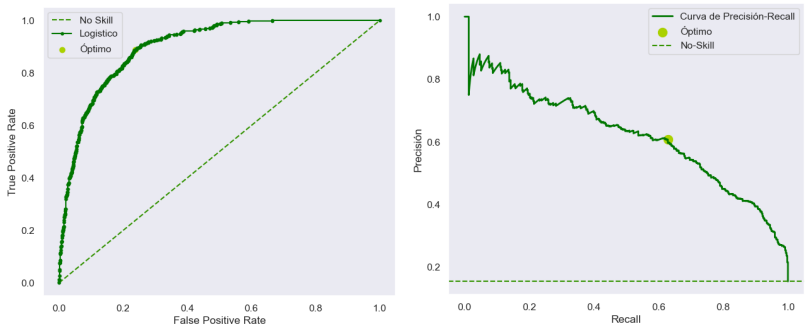

(a) Curva ROC. (b) Curva F1-Score.

Figura 4.17: Puntos de corte para el ML con la técnica de Random Undersampling.




(a) Curva ROC. (b) Curva F1-Score.

Figura 4.18: Puntos de corte para el ML con la técnica de SMOTE.




(a) Curva ROC. (b) Curva F1-Score.

Figura 4.19: Puntos de corte para el ML con la técnica de Tomek-Links.

Sin embargo, como se detalla en la Sección 2.7, el punto de corte seleccionado para las diversas configuraciones del modelo de regresión logística fue determinado por la F1-Score. Esta elección se justifica por nuestro interés en lograr un equilibrio entre precisión y

sensibilidad, como se evidencia en las consideraciones de evaluación del modelo. Por tanto, considerando ese punto de corte en especifico, los resultados obtenidos en la matriz de confusión son los siguientes:

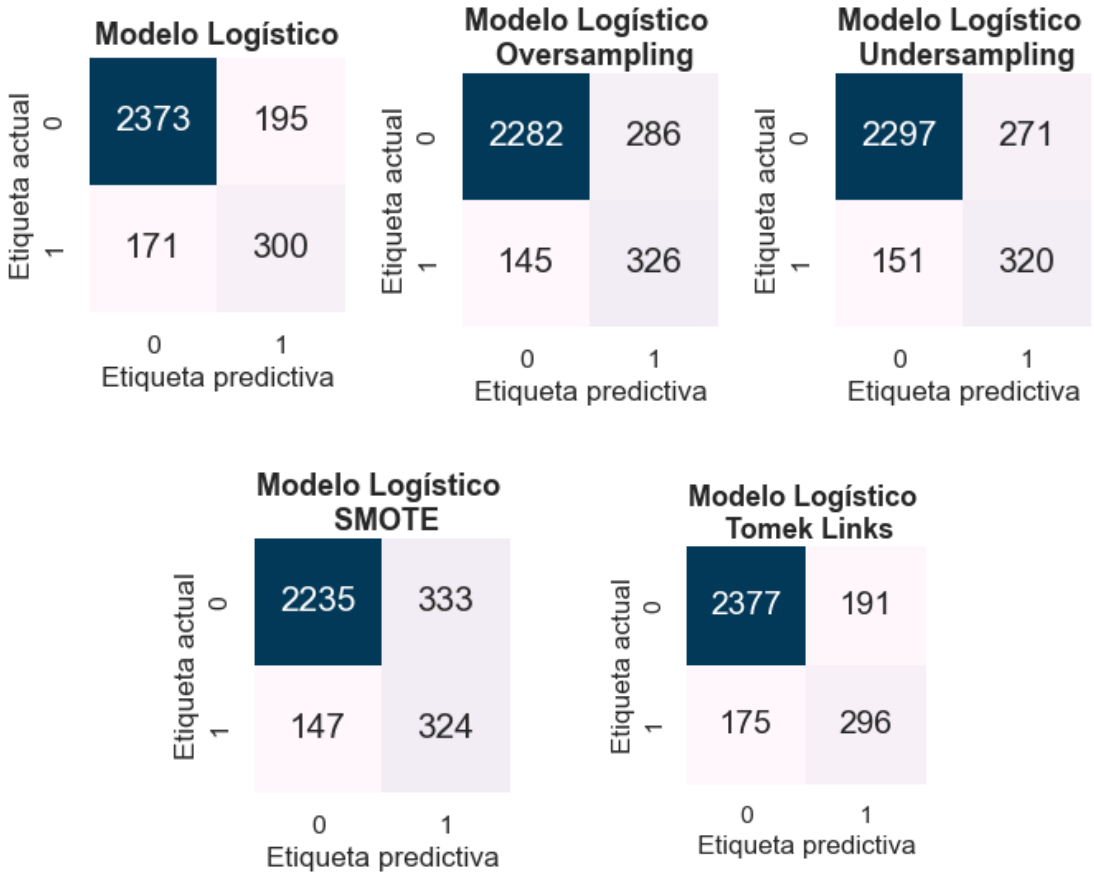



Figura 4.20: Matriz de confusión para el ML con diferentes condiciones.

De forma consolidada, al emplear métricas adicionales para evaluar el rendimiento de los modelos mencionados anteriormente, se obtuvieron los siguientes resultados:

| **Método** | **Precision** | | **Recall** | | **F1 Score** | | **Accuracy** |
|---|---|---|---|---|---|---|---|
| | **0** | **1** | **0** | **1** | **0** | **1** | |
| ML | 93 % | 61 % | 92 % | 64 % | 93 % | 62 % | 88 % |
| ML Oversampling | 94 % | 53 % | 89 % | 69 % | 91 % | 60 % | 86 % |
| ML Undersampling | 94 % | 54 % | 89 % | 68 % | 92 % | 60 % | 86 % |
| ML SMOTE | 94 % | 49 % | 87 % | 69 % | 90 % | 57 % | 84 % |
| ML Tome-Links | 93 % | 61 % | 93 % | 63 % | 93 % | 62 % | 88 % |

Cuadro 4.7: Consolidado de las métricas de clasificación para cada método.

Del Cuadro 4.7 que presenta las métricas para los distintos modelos implementados, resalta la métrica de F1 Score, que ofrece una evaluación equilibrada entre precisión y recall. De lo anterior, se evidencia:

- SMOTE demuestra ser eficaz en la identificación de clientes morosos al tener un recall del 69 %, la precisión es relativamente baja (49 %), lo que significa que hay una proporción significativa de falsos positivos.

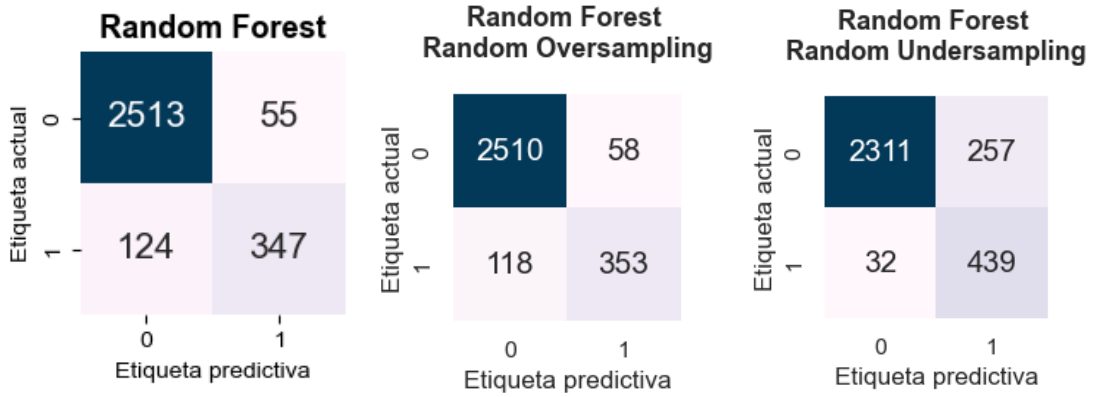
- El modelo de regresión logística (ML) y con la implementación de la técnica de balanceo de datos Tome-Links tienen los mejores resultados para ambas clases en términos de precisión, recall y F1 Score.

- Se puede observar que los modelos exhiben una buena precisión general, especialmente para la clase 0. Sin embargo, se identifica una oportunidad de mejora en términos de la precisión para la clase 1. Aunque los modelos logran un rendimiento sólido en general, enfocar esfuerzos específicos en la mejora de la precisión para la clase 1 podría ser beneficioso para adaptarse mejor a las necesidades específicas, particularmente en el contexto de la identificación de clientes morosos.

### 4.4.3. Entrenamiento del Modelo Random Forest

Se utilizó el modelo RandomForestClassifier de la biblioteca sklearn.ensemble. Para entrenar el modelo, se importó el módulo RandomizedSearchCV de sklearn.model_selection, ya que esta técnica facilita la exploración eficiente del espacio de búsqueda de hiperparámetros mediante el muestreo de combinaciones aleatorias.

A continuación, se muestran los resultados obtenidos en la matriz de confusión, proporcionando una visión detallada del desempeño del modelo de clasificación.



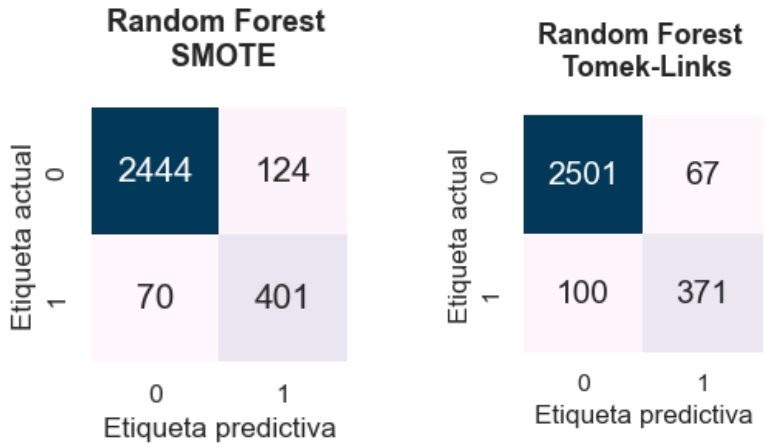



Figura 4.22: Matriz de confusión para Random Forest con diferentes condiciones.

De forma consolidada, al emplear métricas adicionales para evaluar el rendimiento de los modelos mencionados anteriormente, se obtuvieron los siguientes resultados:

| **Método** | **Precision** | | **Recall** | | **F1 Score** | | **Accuracy** |
|---|---|---|---|---|---|---|---|
| | **0** | **1** | **0** | **1** | **0** | **1** | |
| RF | 95 % | 98 % | 98 % | 72 % | 97 % | 80 % | 94 % |
| RF Oversampling | 95 % | 86 % | 98 % | 74 % | 97 % | 79 % | 94 % |
| RF Undersampling | 99 % | 64 % | 90 % | 93 % | 94 % | 76 % | 90 % |
| RF SMOTE | 97 % | 76 % | 95 % | 85 % | 96 % | 80 % | 93 % |
| RF Tome-Links | 96 % | 94 % | 97 % | 79 % | 97 % | 81 % | 94 % |

Cuadro 4.8: Consolidado de las métricas de clasificación para cada método.

Del Cuadro 4.8 que presenta las métricas para los distintos modelos implementados, resalta la métrica de F1 Score, que ofrece una evaluación equilibrada entre precision y recall. De lo anterior, se evidencia:

- RF Undersampling muestra la mayor precision para la clase 0 (99 %) y recall para la clase 1 (93 %). Sin embargo, el F1 Score para la clase 1 es más bajo en comparación con otros métodos.
- Los métodos como RF (Random Forest), RF SMOTE y RF Tome-Links muestran F1 Scores superiores para la clase 1, indicando un mejor equilibrio entre precision y recall en la identificación de clientes morosos, lo cual es crucial en la gestión de riesgo de crédito para minimizar tanto los falsos positivos como los falsos negativos.
- RF Tome-Links tiene un buen equilibrio entre precision, recall y F1 Score para ambas clases, y la accuracy es del 94 %.

### 4.4.4. Entrenamiento del Modelo XGBoost

Se empleó el modelo XGBoost (XG) a través del paquete de Python con el mismo nombre, xgboost. Para el entrenamiento del modelo, se importó el módulo RandomizedSearchCV de sklearn.model_selection, ya que esta técnica facilita la exploración eficiente del espacio de búsqueda de hiperparámetros mediante el muestreo de combinaciones aleatorias.

A continuación, se muestran los resultados obtenidos en la matriz de confusión, proporcionando una visión detallada del desempeño del modelo de clasificación.

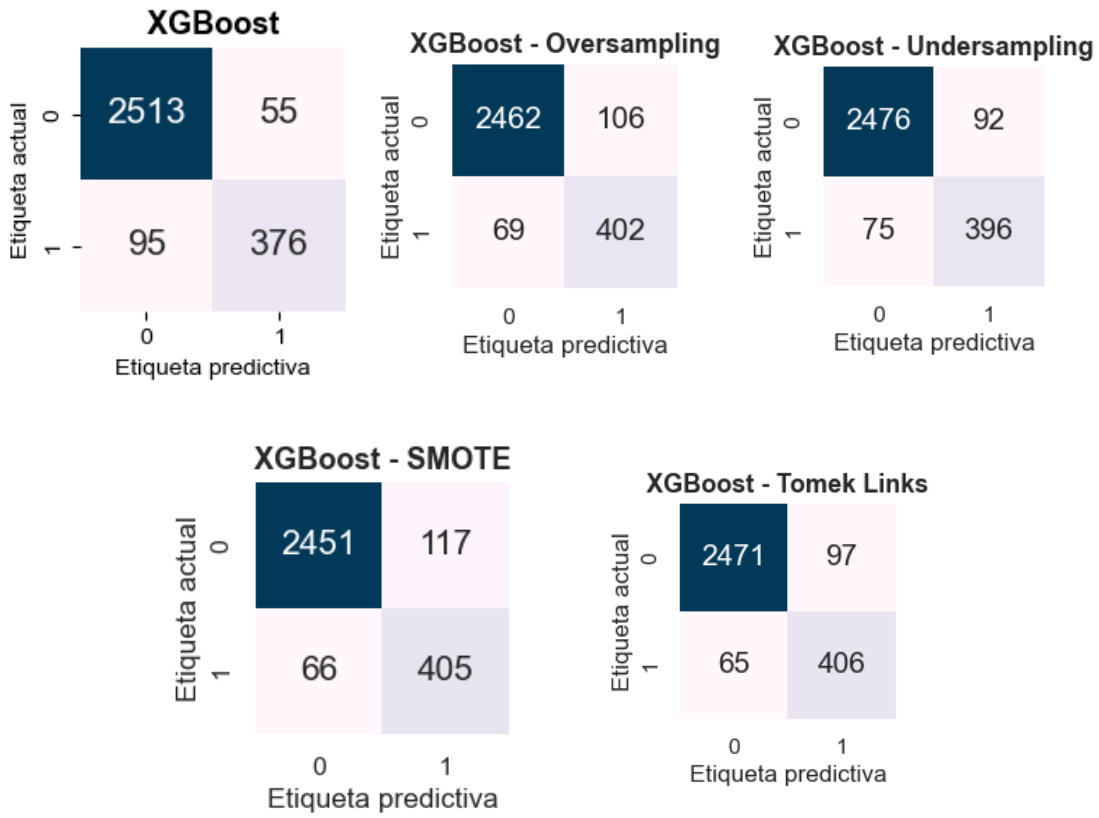



Figura 4.23: Matriz de confusión para XGBoost con diferentes condiciones.

De forma consolidada, al emplear métricas adicionales para evaluar el rendimiento de los modelos mencionados anteriormente, se obtuvieron los siguientes resultados:

| **Método** | **Precision** | | **Recall** | | **F1 Score** | | **Accuracy** |
|---|---|---|---|---|---|---|---|
| | **0** | **1** | **0** | **1** | **0** | **1** | |
| XG | 96 % | 87 % | 98 % | 80 % | 97 % | 83 % | 95 % |
| XG Oversampling | 97 % | 79 % | 96 % | 85 % | 97 % | 82 % | 94 % |
| XG Undersampling | 97 % | 81 % | 96 % | 84 % | 97 % | 83 % | 95 % |
| XG SMOTE | 97 % | 78 % | 95 % | 86 % | 96 % | 82 % | 94 % |
| XG Tome-Links | 97 % | 81 % | 96 % | 86 % | 97 % | 83 % | 95 % |

Cuadro 4.9: Consolidado de las métricas de clasificación para cada método.

Del Cuadro 4.9 que presenta las métricas para los distintos modelos implementados, resalta la métrica de F1 Score, que ofrece una evaluación equilibrada entre precision y recall. De lo anterior, se evidencia:

- Dado nuestro interés en minimizar los falsos positivos, es decir, la identificación incorrecta de clientes no morosos como morosos, el modelo XGBoost con la técnica de balanceo de datos utilizando Tomek Links se presenta como una opción destacada. Este enfoque no solo sobresale en la reducción de los errores de clasificación positiva falsa, sino que también demuestra eficacia en la identificación precisa de clientes auténticamente morosos.

- El modelo de XGBoost (XG) estándar es decir, sin aplicar ninguna de las técnicas de balanceo ofrece un buen equilibrio entre precision y recall para ambas clases, lo que resulta en un alto F1 Score general. Tiene buena capacidad para identificar clientes moroso y además, dado que precisión para la clase 0 es alta (96 %) tiene una buena capacidad para predecir correctamente clientes no morosos.

- XGBoost con la técnica de balanceo de datos Tome-Links logra un alto F1 Score, destacando el equilibrio entre precision y recall para ambas clases.

- El modelo de XGBoost con la técnica de Random Oversampling y SMOTE presentan un alto recall para la clase 1 (clientes morosos), lo que indica que son efectivos para identificar la mayoría de los clientes morosos en el conjunto de datos. Sin embargo, la precisión es relativamente baja a comparación de los otros modelos, lo que significa que también hay un riesgo más alto de clasificar incorrectamente a algunos clientes no morosos como morosos.

### 4.4.5. Importancia de las variables

A partir de las ventajas previamente resaltadas en los modelos Random Forest y XGBoost, se ha identificado que ambas técnicas ofrecen la capacidad de determinar la importancia relativa de cada variable en la predicción final. Este aspecto es crucial, ya que permite comprender qué factores ejercen una influencia significativa en el rendimiento

del modelo.

En el caso del modelo de regresión logística, se ha llevado a cabo la identificación de las variables independientes y los coeficientes correspondientes durante el proceso de entrenamiento, considerando diversas técnicas de balanceo de datos. La síntesis de los resultados de cada modelo se presenta en la Figura 4.24. Es importante destacar que la magnitud del coeficiente (positivo o negativo) desempeña un papel significativo en la interpretación de la influencia relativa de cada variable en la predicción de la probabilidad.

En base a lo anterior, se observa que las variables como num_productos, sexo,edad, num_transacciones_12m, meses_con_tarjeta poseen coeficientes negativos. Esto indica que un aumento en el valor de estas variables está asociado con una disminución en la probabilidad del evento de interés. Por otro lado, variables como tipo_tarjeta, iteraciones_12m y meses_inactivos_12m están asociadas con un aumento en la probabilidad del evento de interés.

Es relevante señalar que, para el modelo de regresión logística en cada una de las variaciones llevadas a cabo la variable valor_disponible parece no tener impacto en la clasificación del cliente, ya que su coeficiente se mantiene en cero para los diversos modelos evaluados.

| | Modelo | | | | |
|---|---|---|---|---|---|
| Coeficiente | Modelo Logistico | ML Oversampling | ML Undersampling | ML Tomek Links | ML SMOTE |
| sexo | -0,9775 | -0,8148 | -0,8263 | -0,9832 | -1,4032 |
| num_productos | -0,4361 | -0,3239 | -0,3299 | -0,4278 | -0,5889 |
| num_transacciones_12m | -0,1114 | -0,1273 | -0,1295 | -0,1130 | -0,1381 |
| meses_con_tarjeta | -0,0111 | -0,0153 | -0,0146 | -0,0109 | -0,0159 |
| edad | -0,0041 | -0,0023 | -0,0049 | -0,0078 | -0,0172 |
| valor_disponible | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| valor_transaccion_12m | 0,0004 | 0,0005 | 0,0005 | 0,0004 | 0,0005 |
| escolaridad | 0,0457 | 0,0360 | 0,0491 | 0,0429 | -0,1640 |
| personas_a_cargo | 0,1307 | 0,1310 | 0,0738 | 0,1191 | -0,1121 |
| salario_anual | 0,1714 | 0,1223 | 0,1613 | 0,1714 | -0,1135 |
| estado_civil | 0,4658 | 0,3708 | 0,4116 | 0,4319 | -0,2129 |
| meses_inactivos_12m | 0,4951 | 0,5357 | 0,5312 | 0,5120 | 0,3117 |
| iteraciones_12m | 0,5063 | 0,4753 | 0,5022 | 0,5110 | 0,3084 |
| tipo_tarjeta | 0,5078 | 0,4660 | 0,5203 | 0,4762 | -0,3937 |
| const | 2,0866 | 3,9330 | 4,1906 | 2,3847 | 8,8449 |

Figura 4.24: Importancia de las variables para el ML.

Con base en los resultados exhibidos en la Figura 4.25 para el modelo de Random Forest y las diferentes técnicas de balanceo de datos que fueron implementadas, se aprecia claramente que las variables asociadas a la información financiera del cliente poseen un peso significativo en la evaluación de este modelo. No obstante, es relevante señalar que las variables pertenecientes a la información sociodemográfica del cliente solo adquieren

notoriedad en quinto lugar, siendo la edad del cliente la variable más destacada en este contexto específico.

| | Modelo | | | | |
|---|---|---|---|---|---|
| Variable | RF | RF Oversampling | RF Undersampling | RF Tomek Links | RF SMOTE |
| valor_transaccion_12m | 33,78% | 31,74% | 33,68% | 32,64% | 31,23% |
| num_transacciones_12m | 31,44% | 27,35% | 35,24% | 28,56% | 31,40% |
| num_productos | 9,95% | 10,42% | 7,41% | 8,37% | 9,44% |
| valor_disponible | 5,72% | 6,66% | 6,20% | 7,02% | 7,41% |
| edad | 4,86% | 5,89% | 4,24% | 5,65% | 5,36% |
| iteraciones_12m | 3,53% | 4,00% | 2,84% | 3,75% | 2,15% |
| meses_inactivos_12m | 3,51% | 3,65% | 3,69% | 4,06% | 2,35% |
| meses_con_tarjeta | 2,32% | 3,11% | 2,49% | 3,37% | 2,74% |
| sexo | 1,40% | 1,91% | 0,93% | 1,45% | 2,90% |
| estado_civil | 0,97% | 1,31% | 1,05% | 1,18% | 0,89% |
| personas_a_cargo | 0,89% | 1,35% | 0,83% | 1,37% | 1,39% |
| salario_anual | 0,78% | 1,27% | 0,71% | 1,22% | 1,18% |
| escolaridad | 0,61% | 1,16% | 0,57% | 1,14% | 1,40% |
| tipo_tarjeta | 0,24% | 0,19% | 0,13% | 0,23% | 0,15% |

Figura 4.25: Importancia de las variables para Random Forest.

En relación al modelo XGBoost, la Figura 4.26 revela que las variables relacionadas con el estado financiero del cliente, como el número de transacciones en el último año, la cantidad de productos y el valor de transacciones realizadas durante el año, conservan su alta relevancia. Sin embargo, a diferencia de la importancia de las variables obtenidas para el Random Forest, en este escenario, las variables asociadas a la información sociodemográfica del cliente adquieren una importancia considerable. Se observa que el género se posiciona en el cuarto lugar en términos de importancia, mientras que la edad se ubica en el séptimo lugar.

| | Modelo | | | | |
|---|---|---|---|---|---|
| Variable | XGBoost | XGBoost Oversampling | XGBoost Undersampling | XGBoost Tomek Links | XGBoost SMOTE |
| num_transacciones_12m | 26,80% | 30,08% | 33,01% | 33,01% | 29,83% |
| num_productos | 20,96% | 11,65% | 13,27% | 13,27% | 11,34% |
| valor_transaccion_12m | 11,09% | 11,28% | 10,56% | 10,56% | 11,04% |
| sexo | 7,04% | 6,06% | 6,97% | 6,97% | 6,11% |
| meses_inactivos_12m | 6,16% | 6,24% | 5,45% | 5,45% | 6,95% |
| iteraciones_12m | 5,21% | 4,75% | 4,28% | 4,28% | 5,03% |
| edad | 5,20% | 5,42% | 5,26% | 5,26% | 5,44% |
| estado_civil | 4,83% | 5,50% | 5,34% | 5,34% | 5,71% |
| valor_disponible | 3,86% | 3,49% | 3,62% | 3,62% | 3,43% |
| meses_con_tarjeta | 3,16% | 2,80% | 2,61% | 2,61% | 2,57% |
| personas_a_cargo | 2,47% | 2,37% | 2,20% | 2,20% | 2,52% |
| escolaridad | 1,13% | 2,39% | 2,21% | 2,21% | 2,28% |
| tipo_tarjeta | 1,11% | 5,71% | 3,25% | 3,25% | 5,80% |
| salario_anual | 0,97% | 2,27% | 1,97% | 1,97% | 1,95% |

Figura 4.26: Importancia de las variables para XGBoost.

## 4.5. Evaluación y entregables

Se entrenaron, de forma inicial, un total de 15 modelos de clasificación entre ellos el Modelo Logistico, Random Forest y XGBoost de forma estándar y con técnicas de balanceo de datos, a los cuales se les evaluó el desempeño mediante métricas definidas para el problema de clasificación que se estaba llevando a cabo al considerar una base de datos que presentaba un desbalance entra las clases es decir, entre clientes morosos y no morosos.

En los archivos ejecutables [1], se encuentran las bases de datos en su versión inicial y procesada, junto con los scripts en Jupyter Notebook utilizados para llevar a cabo el procesamiento y análisis de variables. Asimismo, se proporcionan tres scripts que implementan el Modelo de Regresión Logística, Random Forest y XGBoost, tanto en su configuración estándar como aplicando diversas técnicas de balanceo de datos.

Finalmente al realizar un análisis exhaustivo del estado actual del campo y llevar a cabo este ejercicio, se concluye que las técnicas de machine learning pueden ofrecer soluciones más eficientes en la gestión del riesgo crediticio. Este enfoque se alinea con las regulaciones bancarias de Basilea y se ajusta a las normativas establecidas por la Superintendencia Financiera de Colombia (SFC), específicamente en la tarea de determinar la probabilidad o clasificación de que un prestatario incumpla con sus obligaciones de pago en una entidad financiera. Este enfoque no solo representa una evolución significativa en la evaluación del riesgo crediticio, sino que también se adapta de manera efectiva a los marcos normativos vigentes en el ámbito financiero.

[1] https://github.com/Isabellahernandezo/TrabajoDeGrado.git

# –5–
# Conclusiones

De acuerdo a los resultados obtenidos y despues de realizar el entrenamiento de diferente modelos, se detallan algunas conclusiones sobre este trabajo.

- La gestión del riesgo crediticio adquiere una importancia crucial al asegurar la viabilidad de la decisión de otorgar o rechazar un préstamo a un cliente, considerando su historial financiero y datos sociodemográficos.
- Al evaluar el riesgo de manera objetiva y basada en datos, se pueden tomar decisiones más justas, permitiendo que un grupo más amplio de personas acceda a oportunidades financieras evitando el factor de subjetividad asociado a la valoración del riesgo de cada solicitud de crédito.
- Al analizar y comparar el desempeño de los modelos, incluyendo el Modelo Logístico, Random Forest y XGBoost en condiciones estándar sin la implementación de técnicas de balanceo de datos, se destaca que el Modelo XGBoost logra un equilibrio notable entre precisión y recall para ambas clases, lo que resulta en un alto F1 Score con un 97 % al detectar instancias de la Clase 0 y con un 81 % al detectar instancias de la Clase 1. Al examinar los resultados de la matriz de confusión, se observa que el Modelo Logístico presenta un mayor número de casos de Falsos Positivos y Falsos Negativos en comparación con los otros dos modelos. Este hallazgo resalta la capacidad del Modelo XGBoost para ofrecer un rendimiento más equilibrado en la clasificación de ambas clases.
- La implementación de diversas técnicas de balanceo fue esencial dado el desbalance en la base de datos. Este proceso proporciono una visión más completa y detallada de cómo estas técnicas pueden influir en la calidad de los resultados de clasificación obtenidos y además, a conocer como estas técnicas respaldadas por implementaciones matemáticas, pueden ofrecer o no una oportunidad de mejora en los resultados de clasificación.
- En todas las evaluaciones del modelo, la aplicación de la técnica de Tomek Links se destacó al lograr un equilibrio notable entre precisión y recall para ambas clases.

Este enfoque condujo a un F1 Score más elevado, siendo la mejor estrategia para mejorar la capacidad de discriminación del modelo. No obstante, al implementar esta técnica en el modelo XGBoost, se observaron resultados notables, con un 97 % en la detección de instancias de la Clase 0 y un 83 % en la detección de instancias de la Clase 1. Estos resultados sugieren un desempeño significativo, respaldando la efectividad de la técnica de Tomek Links en el contexto del modelo XGBoost.

- En el contexto de este estudio en particular, la eliminación de puntos cercanos al límite contribuye significativamente a mejorar la separación entre las clases. Este proceso reduce la interferencia entre las clases, lo que se traduce en un mejor rendimiento general del modelo. Por tanto, se destaca el uso efectivo de la técnica de Tomek Links como una estrategia para abordar problemas de desbalance de datos.

- Al realizar el análisis sobre la importancia de las variables en cada uno de los modelos (Sección 4.4.5), se identifica que variables asociadas al estado financiero del cliente tienen mayor peso y un impacto significativo en el entrenamiento del modelo. Aunque la interpretabilidad puede ser un desafío en los modelos de Random Forest y XGBoost, entender cómo se realiza la toma de decisiones a nivel de variables ofrece una perspectiva valiosa sobre el funcionamiento interno del mismo.

# Bibliografía


[1] *Normatividad Decretos, Resoluciones y Leyes que rigen el sector*. URL: https://www.asobancaria.com/normatividad/.

[2] Sara Isabel Álvarez Franco, Christian Lochmüller y Alejandra Osorio Betancur. «La medición del riesgo crédito en Colombia y el Acuerdo de Basilea III». En: *Revista Soluciones de Postgrado* 4.7 (2011), págs. 49-66.

[3] Gloria Juárez, Alfredo Daza y Jesús González. «La crisis financiera internacional de 2008 y algunos de sus efectos económicos sobre México». Español. En: *Contaduría y Administración* 60 (2015), págs. 128-146. ISSN: 0186-1042. URL: https://www.redalyc.org/articulo.oa?id=39543183007.

[4] Alicia Bárcena. «Panorama de la inserción internacional de América Latina y el Caribe 2008-2009. Crisis y espacios de cooperación regional». En: *presentación realizada por Alicia Bárcena, Secretaria Ejecutiva de la CEPAL, Santiago* 25 (2009).

[5] «Silicon Valley Bank: por qué colapsó el banco estadounidense (y qué significa el rescate a sus clientes por parte de la Reserva Federal de EE.UU.)» En: *BBC News Mundo* (mar. de 2023).

[6] «Quién está detrás de Silicon Valley Bank y por qué ha llevado a la banca a su mayor crisis desde 2008». En: *EL DEBATE* (mar. de 2023).

[7] «Comportamientos y perspectivas de los consumidores respecto a los presupuestos, gastos y deudas actuales y futuras de los hogares». En: *Consumer Pulse* (dic. de 2022).

[8] «Endeudamiento de los hogares colombianos sigue creciendo, mientras el ahorro baja». En: *LA REPÚBLICA* (feb. de 2023). URL: https://www.larepublica.co/finanzas/endeudamiento-de-los-hogares-colombianos-sigue-creciendo-mientras-el-ahorro-baja-3546039.

[9] Superintendencia Financiera de Colombia. *Decreto No. 2555 de 2010*. 2010.

[10] A. Elizondo y E.I. Altman. *Medición integral del riesgo de crédito*. Area Contabilidad/Finanzas. Limusa, 2004. ISBN: 9789681863586. URL: https://books.google.es/books?id=UsK-1Ajo44UC.

[11] Alexi Ludovic Leal Fica, Marco Antonio Aranguiz Casanova y Juan Gallegos Mardones. «Análisis de riesgo crediticio, propuesta del modelo credit scoring». En: *Revista Facultad de Ciencias Económicas* 26.1 (2018), págs. 181-207.

[12] Ronald A Fisher. «The use of multiple measurements in taxonomic problems». En: *Annals of eugenics* 7.2 (1936), págs. 179-188.

[13] David Durand. *Risk elements in consumer instalment financing*. National bureau of economic research, 1941.

[14] Ivan Mauricio Bermudez Vera, Diego Fernando Manotas Duque y Javier Olaya Ochoa. «Modelo para la estimación del deterioro por riesgo de crédito». es. En: *Suma de Negocios* 11 (dic. de 2020), págs. 149-157. ISSN: 2215-910X. URL: http://www.scielo.org.co/scielo.php%20script=sci_arttext&pid=S2215-910X2020000200149&nrm=iso.

[15] Fredy Ocaris Pérez Ramírez y Horacio Fernández Castaño. «Las redes neuronales y la evaluación del riesgo de crédito». En: *Revista Ingenierías Universidad de Medellín* 6.10 (ene. de 1), págs. 77-91. URL: https://revistas.udem.edu.co/index.php/ingenierias/article/view/225.

[16] Vijay S. Desai, Jonathan N. Crook y George A. Overstreet. «A comparison of neural networks and linear scoring models in the credit union environment». En: *European Journal of Operational Research* 95.1 (1996), págs. 24-37. ISSN: 0377-2217. DOI: https://doi.org/10.1016/0377-2217(95)00246-4. URL: https://www.sciencedirect.com/science/article/pii/0377221795002464.

[17] Zongyuan Zhao y col. «Investigation and improvement of multi-layer perceptron neural networks for credit scoring». En: *Expert Systems with Applications* 42.7 (2015), págs. 3508-3516. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.12.006. URL: https://www.sciencedirect.com/science/article/pii/S0957417414007726.

[18] Adnan Khashman. «Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes». En: *Expert Systems with Applications* 37.9 (2010), págs. 6233-6239. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2010.02.101. URL: https://www.sciencedirect.com/science/article/pii/S0957417410001405.

[19] Terry Harris. «Credit scoring using the clustered support vector machine». En: *Expert Systems with Applications* 42.2 (2015), págs. 741-750. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.08.029. URL: https://www.sciencedirect.com/science/article/pii/S0957417414005119.

[20] Xiujuan Xu, Chunguang Zhou y Zhe Wang. «Credit scoring algorithm based on link analysis ranking with support vector machine». En: *Expert Systems with Applications* 36.2, Part 2 (2009), págs. 2625-2632. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.024. URL: https://www.sciencedirect.com/science/article/pii/S0957417408000456.

[21] Korin Alexa Idarraga. «EVALUACIÓN DE MODELOS DE CLASIFICACIÓN PARA LA PREDICCIÓN DE RIESGO CREDITICIO DE CLIENTES EN EL SECTOR FINANCIERO». 2023.

[22] Sihem Khemakhem y Younes Boujelbene. «Predicting credit risk on the basis of financial and non-financial variables and data mining». En: *Review of Accounting and Finance* 17 (ago. de 2018). DOI: 10.1108/RAF-07-2017-0143.

[23] Laura Cristina Caro y Lady Jhoana Rodas. «Modelos de aprendizaje supervisado para la clasificación de riesgo crediticio en la entidad financiera Home Credit». Tesis doct. 2022.

[24] Haibo He y Edwardo A. Garcia. «Learning from Imbalanced Data». En: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), págs. 1263-1284. DOI: 10.1109/TKDE.2008.239.

[25] Ali Zughrat, M. Mahfouf, Y.Y. Yang y S. Thornton. «Support Vector Machines for Class Imbalance Rail Data Classification with Bootstrapping-based Over-Sampling and Under-Sampling». En: *IFAC Proceedings Volumes* 47.3 (2014). 19th IFAC World Congress, págs. 8756-8761. ISSN: 1474-6670. DOI: https://doi.org/10.3182/20140824-6-ZA-1003.00794. URL: https://www.sciencedirect.com/science/article/pii/S1474667016429952.

[26] Reto R. Gallati. *Risk management and capital adequacy*. 1.ª ed. McGraw-Hill, 2003.

[27] Juan Gaytán Cortés. «Clasificación de los riesgos financieros». Español. En: *Mercados y Negocios* (2018). ISSN: 1665-7039. URL: https://www.redalyc.org/articulo.oa?id=571864088006.

[28] Leonard Onyiriuba. *Banking Processing and Operational Risk Management in Developing Economies*. Academic Press, 2016, págs. 569-587. ISBN: 978-0-12-805479-6. DOI: https://doi.org/10.1016/B978-0-12-805479-6.00029-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128054796000298.

[29] Máximo Jorge Saavedra Garcia Maria Luisa y Saavedra Garcia. «Modelos para medir el riesgo de crédito de la banca». es. En: *Cuadernos de Administración* 23 (jun. de 2010), págs. 295-319. ISSN: 0120-3592. URL: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-35922010000100013&nrm=iso.

[30] Ken Brown y Peter Moles. «Credit risk management». En: *K. Brown & P. Moles, Credit Risk Management* 16 (2014).

[31] Superintendencia Financiera de Colombia. *Circular Básica Contable y Financiera*. 1995.

[32] Michalis Doumpos, Christos Lemonakis, Dimitrios Niklis y Constantin Zopounidis. *Analytical Techniques in the Assessment of Credit Risk: An Overview of Methodologies and Applications*. Ene. de 2019. ISBN: 978-3-319-99410-9. DOI: 10.1007/978-3-319-99411-6.

[33] Douglas C. Montgomery, Elizabeth A. Peck y G. Geoffrey Vining. *Introducción al análisis de regresión lineal*. spa. México D.F: Compañía Editorial Continental, 2002. Cap. 13, págs. 399-404. ISBN: 970-24-0327-8.

[34] Timothy Sauer. *ANÁLISIS NUMÉRICO*. spa. 2ª edición. México: Pearson, 2013. Cap. 13, págs. 575-578. ISBN: 9786073220606.

[35] Mark G. Hand David J. y Kelly. «Superscorecards». En: *IMA Journal of Management Mathematics* 13.4 (oct. de 2002), págs. 273-281. ISSN: 1471-678X. DOI: 10.1093/imaman/13.4.273. eprint: https://academic.oup.com/imaman/article-pdf/13/4/273/2443979/130273.pdf. URL: https://doi.org/10.1093/imaman/13.4.273.

[36] Konrad Banachewicz, Luca Massaron y Anthony Goldbloom. *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing Ltd, 2022.

[37] Jiawei Han, Micheline Kamber y Jian Pei. «8 - Classification: Basic Concepts». En: *Data Mining (Third Edition)*. Ed. por Jiawei Han, Micheline Kamber y Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, págs. 327-391. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00008-3. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000083.

[38] S. Sivagama Sundhari. «A knowledge discovery using decision tree by Gini coefficient». En: *2011 International Conference on Business, Engineering and Industrial Applications*. 2011, págs. 232-235. DOI: 10.1109/ICBEIA.2011.5994250.

[39] Breiman Leo. «Random forests». En: *Machine learning* 45 (2001), págs. 5-32.

[40] Javier Jesús Espinosa Zúñiga. «Aplicación de algoritmos random forest y xgboost en una base de solicitudes de tarjetas de crédito». En: *Ingeniería Investigación y Tecnología* 21.3 (jul. de 2020), págs. 1-16. DOI: 10.22201/fi.25940732e.2020.21.3.022.

[41] Chao Chen, Andy Liaw y Leo Breiman. «Using Random Forest to Learn Imbalanced Data». En: *Department of Statistics, UC Berkeley* (2004).

[42] Tianqi Chen y Carlos Guestrin. «XGBoost: A Scalable Tree Boosting System». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, págs. 785-794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[43] Nitesh V. Chawla, Nathalie Japkowicz y Aleksander Kotcz. «Editorial: Special Issue on Learning from Imbalanced Data Sets». En: *SIGKDD Explor. Newsl.* 6.1 (jun. de 2004), págs. 1-6. ISSN: 1931-0145. DOI: 10.1145/1007730.1007733. URL: https://doi.org/10.1145/1007730.1007733.

[44] Joaquín García Abad. «Máster Universitario en Análisis de Datos para la Inteligencia de Negocios». Tesis doct. 2021.

[45] Ivan Tomek. «Two Modifications of CNN». En: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.11 (1976), págs. 769-772. DOI: 10.1109/TSMC.1976.4309452.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer. «SMOTE: Synthetic Minority Over-sampling Technique». En: *Journal of Artificial Intelligence Research* 16 (jun. de 2002), págs. 321-357. ISSN: 1076-9757. DOI: 10.1613/jair.953. URL: http://dx.doi.org/10.1613/jair.953.

[47] Swastik Satpathy. «SMOTE for Imbalanced Classification with Python». En: *Mercados y Negocios* (2023). URL: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/.

[48] Aloïs Bissuel. *Hyper-parameter optimization algorithms: a short review*. Abr. de 2019. URL: https://medium.com/criteo-engineering/hyper-parameter-optimization-algorithms-2fe447525903#e4fa.

[49] Y Bergstra J y Bengio. «Random Search for Hyper-Parameter Optimization». En: *Journal of Machine Learning Research* 13.10 (2012), págs. 281-305. URL: http://jmlr.org/papers/v13/bergstra12a.html.

[50] F. Pedregosa y col. «Scikit-learn: Machine Learning in Python». En: *Journal of Machine Learning Research* 12 (2011), págs. 2825-2830.

[51] Quan Zou y col. «Finding the Best Classification Threshold in Imbalanced Classification». En: *Big Data Research* 5 (ene. de 2016). DOI: 10.1016/j.bdr.2015.12.001.

[52] Toni Darmawan, Anggih Surya, Ery Eryanto3 y Tuga Mauritsius4. «Credit Classification Using CRISP-DM Method On Bank ABC Customers». En: *International Journal of Emerging Trends in Engineering Research* 8 (jun. de 2020), págs. 2375-2380. DOI: 10.30534/ijeter/2020/28862020. URL: https://www.warse.org/IJETER/static/pdf/file/ijeter28862020.pdf.

[53] Julio Carpio. «MODELO DE PREDICCIÓN DE LA MOROSIDAD EN EL OTORGAMIENTO DE CRÉDITO FINANCIERO APLICANDO METODOLOGÍA CRISP-DM». En: (2016). URL: https://alicia.concytec.gob.pe/vufind/Record/UANT_63ef929a6dd055996df6db3aeb8d7210.

[54] Eric Wheeler. «What's Next in Digital Transformation: Data-Driven Decision-Making». En: *THE FINANCIAL BRAND'S* (mayo de 2023). URL: https://thefinancialbrand.com/news/digital-transformation-banking/next-in-bank-digital-transformation-data-driven-decision-making-162571/.