Escuela Politécnica Nacional

FACULTAD DE CIENCIAS

DISEÑO DE UNA ESTRATEGIA DE RECUPERACIÓN CREDITICIA TEMPRANA PARA CLIENTES PEQUEÑAS EMPRESAS DE UNA INSTITUCIÓN FINANCIERA DEL ECUADOR MEDIANTE LOS ALGORITMOS K-MEDIAS Y BOSQUES ALEATORIOS.

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

GUILLERMO MIGUEL CELI YANANGOMEZ
celi.guillermo@gmail.com

Director: SANDRA ELIZABETH GUTIÉRREZ POMBOSA
sandra.gutierrez@epn.edu.ec

Quito, marzo 2023

DECLARACIÓN

Yo GUILLERMO MIGUEL CELI YANANGOMEZ, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la ley de Propiedad Intelectual, por su reglamento y por la normativa institucional vigente.

Guillermo Miguel Celi Yanangomez

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por GUILLERMO MIGUEL CELI YANANGOMEZ, bajo mi supervisión.

SANDRA ELIZABETH GUTIÉRREZ POMBOSA, PhD.
Director del Proyecto.

AGRADECIMIENTOS

A mi tutor PhD. Sandra Gutiérrez por haberme ayudado a plasmar y concluir el presente trabajo de investigación, en especial por su paciencia, conocimiento y motivación. Mil gracias.

A mi familia: papi Abdón, mami Carmita, naña Vero y Grace, sobrinos Daryn y Alejo han sido mis mejores guías de vida y motivación para culminar el presente trabajo, gracias por ser quienes son, por creer en mí y extenderme su mano cuando más lo necesitaba.

A quantización: Diego, Oscar, Alex, Daniel, Esteban por sus consejos, recomendaciones y amistad.

DEDICATORIA

A mi esposa Moni por ser parte importante en mi vida, por el apoyo e inspiración para finalizar mis estudios.

Dahlia, Guillermo el mejor regalo de mi vida, y fuente de motivación.

ÍNDICE DE CONTENIDO

Índice de figuras

Índice de cuadros

Resumen

El objetivo de este estudio es proponer una estrategia de recuperación de crédito temprana antes de que los deudores alcancen su definición de "cliente malo" utilizando un enfoque moderno basado en algoritmos de aprendizaje automático, los cuales se implementaran a los clientes denominado como Pequeñas Empresas que realizan actividades crediticias en una institución financiera ecuatoriana.

Los clientes deben agruparse en función de sus características únicas con énfasis de aplicar una estrategia de recuperación antes que el cliente llegue a su condición de no pago de la obligación. Para lograr el objetivo se puede utilizar el algoritmo K-medias, el cual debe incluir indicadores que ayuden a comprender la influencia del deudor en la institución financiera. Las variables utilizadas en el trabajo en curso son: probabilidad de incumplimiento del préstamo, monto actual, plazo restante. La técnica de aprendizaje supervisado Bosque Aleatorio es utilizado para calcular la posibilidad de no pago del préstamo adquirido.

Los métodos de modelado se implementarán en el software estadístico R, que se utiliza para la manipulación, procesamiento y visualización gráfica de datos.

Palabras clave: Aprendizaje automático, K-medias, Bosques Aleatorios, programación R, probabilidad de incumplimiento, monto vigente, plazo remanente, pequeñas empresas.

Abstract

The purpose of this study is to show a strategy for early credit recovery, that is, before the clients reach a definition of “bad clients”. This is possible using modern methodologies based on machine learning algorithms. This model applies to small companies with current credit activity within a financial institution in Ecuador.

Clients must be grouped by their intrinsic unique characteristics, focused on general recovery strategies. The algorithm K-means is used for this purpose and should include variables that help to understand the impact of the client within the financial institution. In this case it will be used the probability of default, current amount and remaining term. The algorithm Random Forest calculates the probability of default.

The method used for the construction of the model will be implemented in the statistical software R, which is used for the manipulation, processing and graphic visualization of the data.

Keywords: statistical learning, machine learning, k-means, random forest, R programming, probability of default, current amount, remaining term, small companies.

CAPÍTULO 1

1 Introducción

En toda organización es de gran importancia generar acciones de seguimiento a cada uno de los resultados. Para una institución bancaria es mayor aun el nivel de control al respecto, especialmente en los riesgos en que esta incurre al momento de posicionar los créditos, que, por su propio giro del negocio, están expuestos constantemente al riesgo de que los clientes incumplan con sus obligaciones crediticias.

La cartera de préstamos de una institución financiera es un gran punto de referencia, es el balance de la cantidad ofrecida al prestatario, por lo que es conveniente evaluar al cliente contra el incumplimiento y tener claro un monto posible de no pago. Por ello, se crean protocolos de seguridad para brindar los productos crediticios basándose en lineamientos claros para facilitar la gestión.

Sin embargo, debido a la naturaleza del negocio, se genera riesgos altos de incumplimiento de los pagos por parte de los clientes, en conjunto con una falta de controles para la gestión de recuperación. Por lo que se recurre a diversas herramientas que permita la detección de esta situación de forma rápida y a tiempo. Cuando un cliente se declara insolvente o con posibilidad de no pago, la entidad financiera inicia otro proceso para rescatar el valor entregado, esto puede ser mediante reestructuraciones y refinanciaciones, ampliar el plazo de cobro, con el objetivo de recuperar la inversión y disminuir el nivel de morosidad.

1.1 Planteamiento del problema

En el Ecuador, las entidades financieras se encuentran obligados, por la Superintendencia de Bancos y Seguros, a desarrollar métodos que les permitan de forma eficiente avaluar los niveles de riesgo crediticio. Incluso mediante simulaciones que facilitan predecir el comportamiento del cumplimiento de las obligaciones, adaptando las condiciones de pago.

Actualmente, existe una entidad financiera que posee una cartera de clientes Pequeñas Empresas PES, que están incurriendo en retardo en el cumplimiento de las obligaciones, los cuales, por el perfil que presentan no poseen condiciones que faciliten la correcta evaluación con relación a los niveles de certificación del pago del crédito otorgado. Para este tipo de deudores, se requiere de un usuario o analista para llevar una apreciación precisa del riesgo del incumplimiento de la financiación realizada.

Una incorrecta evaluación del cliente puede causar que la institución financiera otorgue o plantee nuevas ofertas de préstamo a clientes con alta posibilidad de no pago. Por lo tanto, el índice de morosidad crediticia tiende a incrementarse si no se instaura una adecuada estrategia de recuperación del crédito de forma temprana.

Para la evaluación de la posibilidad de no pago del crédito, Bonini & Caivano (2018) menciona que existen estudios que han demostrado que los métodos de inteligencia artificial (IA) lograron un mejor rendimiento que las metodologías estadísticas tradicionales. El presente trabajo utilizará el algoritmo de Bosques Aleatorios[1] BA para la estimación de la probabilidad de incumplimiento (Kornfeld, S., 2020), dado que el modelo no presenta sobreajuste, no es afectado por los valores atípicos, y produce buenos resultados en la clasificación de los clientes (Breiman L., 2001). BA se encuentra dentro de los algoritmos supervisados del aprendizaje automático (AA) pertenecientes a la IA.

Adicional a obtener la probabilidad de no pago del crédito también es necesario agrupar a los PES basándose en sus características propias, en el trabajo de Carrasco O. (2020) menciona que el algoritmo K-medias[2] tiene la fama se ser uno de los “más simples y conocidos no solo dentro de los de tipo particional sino de los algoritmos en general, ya que sigue una forma fácil y sencilla para dividir un conjunto de datos en k grupos o clúster conocidos a priori” (pág. 14). Este análisis resulta ser muy útil, dado que la técnica K-medias permite realizar una agrupación de individuos y se aplicará en el presente estudio para la

[1] Algoritmo de Bosques Aleatorios se describe en el capítulo 2.2.1
[2] Algoritmo de K-medias se describe en el capítulo 2.2.2

segmentación de los clientes, el cuál se encuentra categorizado en los modelos no supervisados del aprendizaje automático de la IA.

En este trabajo se propone el diseño de una estrategia de recuperación crediticia temprana para pequeñas empresas, utilizando los algoritmos de K-medias y Bosques aleatorios, sustentados en métodos AA, que ajustan automáticamente la predicción de acuerdo con la información disponible.

Se desarrollará una acción dé recuperación temprana, enfocándose principalmente en clientes con alta posibilidad de no pago y mayor impacto en la institución.

1.2 Objetivo general

Clasificar a los clientes pequeñas empresas de una institución financiera del Ecuador en grupos mediante los algoritmos K-medias y Bosques aleatorios para determinar una estrategia de recuperación temprana del crédito otorgado, con la revisión de probabilidad de incumplimiento.

1.2.1 Objetivos específicos

- Describir a los clientes pequeñas empresas (PES) de una institución financiera en el Ecuador, para la determinación de la probabilidad de Incumplimiento a través del algoritmo de Bosques aleatorios.
- Catalogar a los clientes pequeñas empresas a través del algoritmo de K-medias, con las variables Probabilidad de Incumplimiento (PI), Plazo Remanente (PR), Monto Vigente (MV) para definir los grupos de riesgo, alto, medio y bajo de acuerdo con su clasificación.
- Proponer la estrategia de recuperación de la obligación más apropiado en cada grupo de riesgo, para minimizar atrasos en la recaudación de la cartera de crédito.

1.3 Justificación de la investigación

Con el fin de garantizar las utilidades de los accionistas dentro de la institución financiera y, el dinero de los depositantes que confían en la entidad es fundamental aplicar modelos matemáticos que mejoren la predictibilidad de incumplimientos de obligaciones de los clientes con el fin de gestionarlos de manera oportuna.

En las instituciones financieras del país, se suele segmentar el tamaño de las empresas basándonos en el nivel de ventas que las mismas reportan cada año con la finalidad de tener un panorama de su capacidad de pago. El financiamiento para Pymes son créditos productivos otorgados a personas naturales o jurídicas, que registren ingresos anuales entre USD 100 mil y USD 1 millón (Financiera, 2021, pág. 329).

Por otro lado, las Pymes en el Ecuador son sin duda un elemento dinamizador de la economía, pues su contribución alcanzó en el año 2010, el 67% del total de ingresos de la riqueza del Ecuador y generó el 75% de empleo a nivel nacional (Vélez, M. & Chamba, N., 2019, pág. 34). Con esto, puede afirmarse que es de vital importancia entender el comportamiento crediticio para los Bancos.

Según lo manifiesta Escalera Chávez (2011), las Pymes, por sus características, son empresas que continuamente están en peligro de cierre masivo por ser vulnerables a sus ambientes económicos y son entidades que necesitan de financiamiento externo para continuar su ejercicio (pág. 25). Para atender a este segmento de clientes se han incrementado diversos productos, servicios y políticas dentro de cada institución financiera (Delgado, J., 2015, pág. 49).

Las entidades financieras tienen la responsabilidad de administrar sus riesgos; por lo que es de suma importancia disponer de herramientas eficientes para el monitoreo, gestión y control del riesgo, mismas que se pueden desarrollar en sus propias metodologías, que deberán considerar criterios que estimen el impacto (Junta de Regulación Monetaria Financiera, 2017, artículo 3, pág. 263).

1.3.1 Justificación Teórica

El presente trabajo se enfocará en la aplicación práctica de dos técnicas de Aprendizaje Automático AA en las instituciones financieras. Estos modelos se sustentan en las siguientes bases teóricas: La probabilidad incumplimiento (PI); la cual se define como la posibilidad de que ocurra el no pago del crédito ya sea de forma parcial o total, en un tiempo determinado (Financiera, J. D. R. M., 2021, pág. 631). Para su cálculo se evidencia la utilización del algoritmo de bosques aleatorios (Kornfeld, 2020).

El algoritmo de Bosques Aleatorios (BA) está basado en las técnicas de Bootstrapping y Bagging. Bootstrapping es una técnica en el cual el conjunto de información para entrenamiento es re-muestreada y reemplazada por nuevas muestras sintéticas (Kornfeld, S., 2020, pág. 13). En cada grupo de datos donde se realiza el Bootstrapping se aplica un árbol de clasificación, es simplemente un árbol de decisión con variable de respuesta categórica; el cual tiene el objetivo de particionar el espacio en grupos homogéneos más pequeños. Posteriormente, se utiliza Bagging, el cual toma el promedio de todas las predicciones realizadas por el árbol de clasificación en cada grupo de información que se realiza el Bootstrapping (Kornfeld S., 2020, pág. 13).

La segunda parte de este proyecto pretende categorizar a los clientes, basado en el algoritmo de K-medias, el cual es una técnica de partición y tiene como objetivo colocar a cada cliente dentro de un grupo (Igual, L. & Seguí, S., 2017, pág. 121).

Para utilizar el algoritmo K-medias es necesario definir las variables que van a ser utilizadas. Para proponer los indicadores de agrupamiento, se utilizó la idea que proviene de una metodología que se utiliza en marketing, que segmenta los compradores mediante las variables: Reciente, Frecuencia y Monetario (RFM). Reciente (R) son las compras recientes, cantidad de días transcurridos desde la última transacción. Frecuencia (F) es el número de veces que el cliente realizó una compra. Monetario (M) significa la suma de todo el monto de las transacciones del periodo; esto quiere decir que se basa en el comportamiento de clientes observando el tiempo trascurrido de su última compra, la periodicidad que realiza las compras, y el total de dinero utilizado (Cuadros, Gonzalez, & Jiménez, 2017, pág. 43).

El presente trabajo realiza un análogo del análisis RFM, pero con indicadores que pueden alertar de un posible incumplimiento en las obligaciones y su impacto en la entidad financiera. Realizando un símil en el riesgo de crédito, se propone utilizar las variables: PI, Monto Vigente (MV) y Plazo Remanente (PR), las cuales permiten comprender la situación del cliente en la institución, observando su probabilidad de no pago, el capital de afectación en caso de que esto ocurra y el tiempo de permanencia de la obligación, el objetivo es agrupar clientes con características similares, para realizar una propuesta de gestión de recuperación temprana adecuada.

En función de las agrupaciones se define las maniobras de recuperación temprana. Las estrategias establecen las formas de cobrar el crédito, los criterios de negociación, condonaciones, entre otros; y todos ellos adecuados a cada grupo establecido y tendrá un efecto más preciso para lograr que los clientes cumplan con sus compromisos (Morales J. & Morales A., 2014, pág. 146).

Entonces es factible manifestar que existen suficientes bases teóricas que sustentan el desarrollo del presente trabajo, mismo que se ha de manejar de manera anónima para proteger los intereses de los clientes y de la propia entidad financiera, proporcionando al final un instrumento que se constituya en una herramienta de referencia para la clasificación de un cliente y la mitigación del riesgo crediticio que puede sufrir la empresa al depositar grandes sumas de dinero en los prestamistas.

1.3.2 Justificación Metodológica

Para lograr los objetivos del presente estudio, se utilizarán las mejores prácticas en técnicas matemáticas que actualmente están siendo utilizadas, como son los modelos de Aprendizaje Automático (Kornfeld S., 2020).

Las instituciones financieras deben estimar internamente la PI, para ello se emplea el Algoritmo de Bosques Aleatorios (Kornfeld S., 2020), el cual permite segmentar clientes, es utilizado en la detección de fraude, entre otros (Espinoza, J., 2020). Por otro lado, es una técnica supervisada que crea múltiples árboles de decisión sobre un conjunto de datos, en

los que los resultados conseguidos se ajustan con el propósito de conseguir un modelo único más firme en comparación a lo obtenido en cada árbol de decisión (Espinoza, J., 2020). Es una metodología adecuada para el cálculo de probabilidad de incumplimiento.

Para determinar las agrupaciones de los PES, resultan útiles las técnicas de segmentación, el análisis RFM es utilizado en la industria (McCarty, J. & Hastak, M., 2007) y, por ende, al realizar un símil y el uso el algoritmo K-medias se estará buscando una mejor agrupación a los clientes utilizando variables de riesgo de crédito, que permiten observar el impacto en la institución financiera.

Según lo manifestado en el trabajo de Carrasco O. (2020, pág. 14), el algoritmo K- medias tiene la fama de ser uno de los “más simples y populares no solo dentro de los de tipo particional sino de los algoritmos en general, ya que sigue una forma fácil y sencilla para dividir un conjunto de datos en k grupos o clúster conocidos a priori”. Este análisis resulta ser muy útil cuando hay grandes cantidades de información; para aplicar en investigaciones exploratorias en los que se desconoce el número ideal de conglomerados.

Una actividad primordial de la cobranza es reaccionar de manera adecuada y atinada mediante las estrategias establecidas, además, es fundamental segmentar a los clientes en función de características comunes para así determinar la estrategia de recuperación acertada (Morales J. & Morales A., 2014, pág. 146). El presente trabajo de titulación propone resolver el problema bajo dicho criterio.

Actualmente los métodos de AA se aplican en industrias que trabaja con grandes volúmenes de información, entre los principales son: marketing, que aplica este método para conocer más a profundidad a sus clientes, entender sus hábitos y ofrecer productos y servicios que satisfagan sus necesidades (Ortiz G., 2019). Por consiguiente, es imperioso en los tiempos actuales el empleo de la tecnología, con procesos computacionales que resultan económicos y eficientes, que hacen posible producir modelos de manera rápida y precisos y automática, y que puedan analizar datos complejos y con gran volumen.

1.3.3 Justificación Práctica

La necesidad de implementar una agrupación con enfoque de recuperación crediticia de forma temprana en la institución financiera es vital, se ha evidencia un incremento de los clientes en la entidad bancaria, por lo que ésta busca optimizar recursos.

El trabajo planteado en este proyecto de titulación permitirá analizar a los PES que actualmente poseen crédito y otorgar una clasificación de riesgo con orientación de la gestión de recuperación, enfocándose en los clientes que podrían causar mayor impacto, en caso de que se produzca el cumplimiento de sus obligaciones. Por otra parte, las empresas de enfoque financiero se ven forzadas a mejorar sus estrategias de cobranza con el propósito de reducir costos y aumentar el recobro de la obligación.

Adicionalmente, la metodología ayudará a que la entidad cumpla, ante la Superintendencia de Bancos, el requisito de tener un modelo propio de administración de crédito, dado que en esta investigación se propone realizar un monitoreo a Pequeñas Empresas, que son clientes vulnerables a factores externos (Chávez, M., 2011, pág. 25).

1.4 Hipótesis

Las hipótesis base del presente trabajo de investigación, quedan planteadas de la siguiente manera:

1) Se puede clasificar a los clientes Pequeñas Empresas en una empresa financiera mediante el algoritmo de K-medias.
2) Se puede diseñar una estrategia de recuperación crediticia temprana para clientes pequeñas empresas de una institución financiera del ecuador mediante los algoritmos K-medias y Bosques Aleatorios.
3) Se puede utilizar las variables PI, MV, PR para comprender el riesgo crediticio del cliente e impacto dentro de la institución financiera.

CAPÍTULO 2

2 Marco Teórico

En este capítulo se define las nociones y conceptos teóricos necesarios para comprender la metodología utilizada. Se empezará a describir la probabilidad de incumplimiento, el algoritmo BA es utilizado para su cálculo, también los criterios de evaluación del modelo, así como la técnica K-medias que permitirá agrupar a los clientes para establecer una estrategia de recuperación.

2.1 Probabilidad de Incumplimiento (PI)

El primer paso es realizar el cálculo de la probabilidad de incumplimiento, la cual se define como la posibilidad de que ocurra el no pago del crédito ya sea de forma parcial o total, en un tiempo determinado (Financiera, 2021). El proceso de obtención de la PI se puede obtenerlo mediante los métodos de Aprendizaje Automático haciendo uso del algoritmo BA (Kornfeld, S., 2020).

En el presente caso, se ha de considerar que un deudor se categoriza en un estado de no pago cuando al menos una de sus obligaciones presenta un vencimiento superior o igual a los 90 días mora (de Basilea, C. D. S. B., 2006). PI es entonces la posibilidad que en el efecto de un año el cliente no cumpla con más de 90 días sus compromisos crediticios, si cumple la condición el prestatario ingresa a un estado de incumplimiento. Ahora bien, en el marco de aprendizaje estadístico se tiene como finalidad el desarrollo de modelos basados en datos históricos, con el objetivo de realizar predicciones o clasificaciones.

De acuerdo con Kornfeld (2020) el modelo de PI es un problema de clasificación binaria $Y \in \{0,1\}$, donde 0 simboliza que el deudor no ingresó al estado de incumplimiento y 1 significa que cumple la condición de mal cliente; la variable aleatoria Y_i es la variable respuesta y tomará los valores originales y_i; dónde i representa el i-ésimo registro del conjunto de datos. (pág. 10)

Las variables utilizadas para entrenar al modelo estadístico se denominarán explicativas. Sea $X_i \in R^p$ un vector de entrada de valores reales aleatorios con información observada se denotados $x_i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}\}$ con p número total de predictores. (Kornfeld, S., 2020, pág. 10)

El conjunto de datos con N registros se lo expresa como D= $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$. Entonces, el modelo estadístico se representa por (Kornfeld, S., 2020, pág. 10):

$$Y_i = f(X_i) + \varepsilon_i, \text{ donde } \varepsilon_i \text{ es el error aleatorio.} \quad (2.1)$$

El modelo estadístico surge del campo del AA conocido como aprendizaje supervisado, el cual f aprende a través de un conjunto de observaciones. Los valores observados x_i alimentan el algoritmo y este produce salidas $f(x_i)$. El algoritmo de aprendizaje f modifica su información de entradas/salidas en respuesta a las diferencias entre la variable original y_i y la generada $f(x_i)$. La técnica se considera que es un clasificador, y se espera que los resultados esperados se acerquen a los reales (Kornfeld, S., 2020).

2.2 Modelos Aprendizaje Automático (AA)

Los modelos AA pueden ayudar enormemente a mejorar la toma de decisiones en riesgo crediticio, creando un puntaje de crédito [3]para cada uno de los clientes de la cartera, el cual, consiste en asignarle un puntaje al potencial deudor que, a su vez, representa una estimación del desempeño de la deuda para el banco.

Según lo señalan Igual & Seguí (2017), dentro del aprendizaje estadístico, se encuentran los modelos supervisados, los cuales son algoritmos que aprenden a partir de un conjunto de datos denominados base de entrenamiento; es decir, aprende a través de un grupo de observaciones iniciales. El algoritmo para utilizar en el presente trabajo es el llamado Bosques

[3] Puntaje de crédito, asigna un valor al cliente y este representa una estimación del desempeño del crédito para la institución financiera.

Aleatorios, el cual, en algunos estudios, es utilizado para el cálculo de probabilidad de incumplimiento.

2.2.1 Bosques Aleatorios

Es un algoritmo AA flexible y fácil de usar que da un buen resultado en la mayoría de los casos, incluso sin ajuste de parámetros. También es uno de los algoritmos más utilizados debido a su simplicidad y al hecho de que puede usarse tanto para tareas de clasificación como de regresión.

Breiman L. (2001) presenta una descripción formal del algoritmo de bosques aleatorios el cual se define como un clasificador que consta de una colección de árboles clasificadores estructurados $\{h(x, \Theta_k), k = 1, \dots\}$ donde $\{\Theta_k\}$ son vectores aleatorios independientemente distribuidos de manera idéntica y cada árbol emite un voto unitario para la clase más popular de la entrada x (pág. 6) donde:

- Θ_k: vector que genera al k-ésimo árbol del bosque, , $k = 1, \dots$
- x: vector de características.
- Voto: etiqueta de la clase, “1” mal cliente, “0” buen cliente.

El procedimiento común es que para cada k-ésimo árbol, un vector aleatorio Θ_k es generado, independiente de los vectores aleatorios pasados $\Theta_1, \dots, \Theta_{k-1}$ pero manteniendo la misma distribución, y el árbol se crece usando la base de entrenamiento y Θ_k, resultando un clasificador $h(x, \Theta_k)$, donde x es un vector de características de entrada. (Breiman L., 2001, pág. 5)

Kornfeld (2020) representa el algoritmo de Bosques Aleatorios con base en lo propuesto por Breiman (2001) y se describe a continuación:

El BA está basado en las técnicas de re-muestreo Bootstrapping y el algoritmo de Bagging. El Bootstraping es una técnica donde el conjunto de datos de entrenamiento $Z = \{(x_i, y_i)\}_{i=1}^{N}$ es re-muestreada y reemplazada por nuevas muestras sintéticas. Para cada

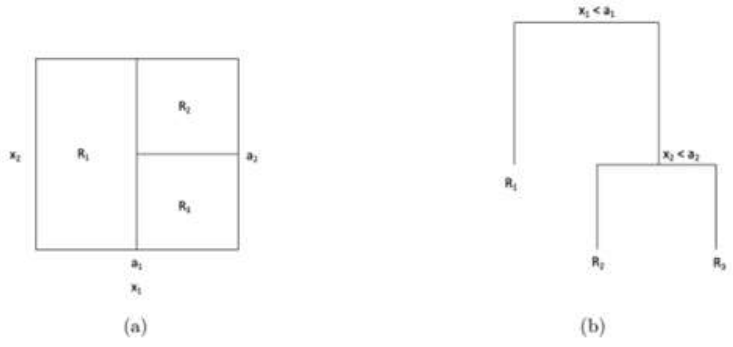
conjunto de datos donde se realiza el Bootstraping, denominada Z^*, se aplica un árbol de clasificación, produciendo una predicción $f(x)$ para cada entrada x.

El árbol de clasificación es simplemente un árbol de decisión con variable de respuesta categórica; el cual tiene el objetivo de particionar el espacio en grupos homogéneos más pequeños, mediante las sentencias SI-ENTONCES, las cuales funcionan como condiciones de división para realizar una clasificación. La figura 2.1 presenta gráficamente como un espacio o conjunto de dos variables es divido en subconjuntos más pequeños y son representados en el árbol mediante sus nodos. Para obtener el nodo, el algoritmo debe definir la variable que particiona el espacio y el valor o característica del corte. En cada nodo resultante o final se encuentran clientes que ingresaron a su estado de incumplimiento y cuáles no, obtenidas de $Z = \{(x_i, y_i)\}_{i=1}^{N}$, la proyección o estimación de cliente en cada nodo terminal, se define mediante:

$$\hat{y}_i = \begin{cases} 1, & si\ \hat{p} \geq c \\ 0, & si\ \hat{p} < c \end{cases} \tag{2.2}$$

Donde $\hat{p}$ es la proporción de clientes en estado de incumplimiento en el nodo terminal, del conjunto, c es un valor umbral el cual define como se clasifica el nodo terminal.

Figura 2.1: Diagrama de división de un árbol.



(a) Un espacio de dos variables (x_1, x_2) es divido en 3 subconjuntos a través de los puntos a_1 y a_2. (b) Resultado del árbol de clasificación con los nodos terminales R1, R2, R3.

Fuente: Adaptado de Kornfeld, 2020.

Obtenido el resultado del árbol de clasificación en cada Z^* se realiza el algoritmo Bagging, el cual toma el promedio de todas las predicciones realizadas por el árbol de clasificación en cada conjunto de datos que se realiza el Bootstraping. La estimación $\hat{f}_{bag}(x)$ del algoritmo se define:

$$\hat{f}_{bag}(x) = \frac{1}{B_{rf}} \sum_{n=1}^{B_{rf}} \hat{f}_n(x) \tag{2.3}$$

Donde B_{rf} es el número total de Bootstraping ejecutados, $\hat{f}_n(x)$ es la predicción de cada n-ésima muestra en la que se realiza Bootstraping.

BA hace predicciones en cada muestra Bootstraping y luego usa Bagging para la predicción final. A continuación, se describe el algoritmo basándose a lo descrito por Kornfeld (2020, pág. 14):

1. Se obtiene el conjunto de datos de entrenamiento $Z = \{(x_i, y_i)\}_{i=1}^{N}$ y una entrada x_k para proyectar.
2. De Z se realiza Bootstraping en B_{rf} muestras; es decir, se obtiene conjuntos $Z_1^*, Z_2^*, \dots, Z_{B_{rf}}^*$ los cuales producen submodelos $\hat{f}_1(Z_1^*), \hat{f}_2(Z_2^*), \dots, \hat{f}_{B_{rf}}(Z_{B_{rf}}^*)$.
3. Para cada división del submodelo $\hat{f}_n(Z_n^*)$, se selecciona q predictores de forma aleatoria de un total de p variables. De los q, para determinar el punto de corte se escoge el valor donde la división realiza una mayor proporción entre los clientes que ingresaron a su estado de incumplimiento y los que no.
4. Para cada submodelo $\hat{f}_n(Z_n^*)$ la variable x_k es usada para realizar la predicción $\hat{f}_1(x_k), \hat{f}_2(x_k), \dots, \hat{f}_{B_{rf}}(x_k)$
5. La predicción final está dada por:

$$\hat{f}_{rf}(x_k) = \frac{1}{B_{rf}} \sum_{n=1}^{B_{rf}} \hat{f}_n(x_k) \tag{2.4}$$

Basado en el algoritmo de bosques aleatorios, la clasificación del x_k esta dada por:

$$\hat{y}_k = \begin{cases} 1, & si \ \hat{f}_{rf}(x_k) \geq c \\ 0, & si \ \hat{f}_{rf}(x_k) < c \end{cases} \tag{2.5}$$

donde c es el punto de corte de la decisión.

2.2.1.1 Optimización de Hiperparámetros

En el caso de BA, la optimización de hiperparámetros es el proceso de elegir un conjunto óptimo de parámetros del modelo de AA para obtener el mejor rendimiento (Ortiz G., 2019, pág. 2). El beneficio del algoritmo depende de los parámetros escogidos y estos regirán el proceso de modelación.

En el presente estudios nos enfocaremos en los siguientes parámetros:

- Número de árboles (Ntree) que forman parte del BA, algunos experimentos indican que a mayor número de árboles menor problema de overfitting[4].
- Número de variables (Mtry) que se seleccionarán en cada árbol dentro del BA, y es uno de los parámetros más importantes.
- Cantidad de registros en cada nodo.

El criterio para evaluar la variable óptimo es el indicador “Error fuera de la muestra”.[5]

2.2.1.2 Evaluación del desempeño

BA posee evaluaciones estadísticas para medir su rendimiento y se lo aplica tanto en la muestra de construcción y prueba, en primera instancia debemos definir cuando el modelo

[4] Es el efecto de sobreentrenar el algoritmo de aprendizaje supervisado.
[5] La definición se describe en el capítulo 2.2.1.3

realiza una correcta estimación, Kubat M. (2017) define los criterios en el libro “An Introduction to Machine Learning” y la base para las definiciones se describe a continuación.

Matriz de confusión

La matriz de confusión es una herramienta comúnmente utilizada en la evaluación del desempeño de los algoritmos supervisados, y por ende en Bosques Aleatorios. La matriz es una tabla de 2x2, un eje representa las predicciones del modelo y el otro es la clasificación real, es útil observarlo de esta manera dado que permite identificar fortalezas y debilidades del modelo. En el cuadro 2.1 se ejemplifica la matriz con sus componentes.

Cuadro 2.1: Matriz de confusión.

		Estimación del Modelo	
		0	**1**
Clasificación Real	**0 = bueno**	TP	FN
	1 = Malo	FP	TN

Fuente: Elaboración propia.

TP= número de clientes buenos proyectados buenos.
TN= número de clientes malos proyectados malos.
FN= número de clientes buenas proyectados malos.
FP= número de clientes malos proyectados buenos.

A partir de la matriz se puede calcular indicadores de rendimiento del modelo.

Tasa de error (E)

Es el porcentaje de clientes clasificados erróneamente, en otras palabras, es la frecuencia de errores cometidos por el algoritmo sobre el conjunto de datos, este indicador es utilizado en los modelos de clasificación y se calcula a partir de la matriz de confusión, se lo obtiene dividendo el número de no aciertos, $FP + FN$, para el total de registros, $TP + FP + FN + TN$.

$$E = \frac{FP + FN}{TP + FP + FN + TN} \tag{2.6}$$

Exactitud (Acc)

Es el porcentaje de clientes clasificados correctamente por el algoritmo, $TP + TN$, sobre el total de registros, $TP + FP + FN + TN$, es útil para entender como el modelo está funcionando en términos generales, se calcula dividiendo el número de aciertos sobre el número total de la muestra.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.7}$$

Precisión (Pr)

Es la porción de clientes buenos correctamente estimados por la técnica, TP, entre todas las restimaciones que el algoritmo ha etiquetado, $TP + FP$; es decir, mide la capacidad del modelo para dar una respuesta correcta cuando la clasifica como positiva.

$$Pr = \frac{TP}{TP + FP} \tag{2.8}$$

Exhaustividad (Re)

Es el porcentaje de clientes buenos correctamente estimados por el clasificador, TP, sobre el total, $TP + FN$.

$$Re = \frac{TP}{TP + FN} \tag{2.9}$$

F1-score

Este indicador de rendimiento combina la Precisión y la exhaustividad especialmente útil en problemas donde las dos variables son importantes, además se utiliza cuando hay desproporción de muestras. Su cálculo es el siguiente:

$$F1 = \frac{2 * Pr * Re}{Pr + Re} \tag{2.10}$$

Sensibilidad (Se)

Conocida como tasa de verdaderos positivos (TPR) o eficacia, mide la capacidad del modelo para detectar correctamente los clientes buenos. Es la medida de **exhaustividad** para los deudores que no presentan problemas de pago. Se define como:

$$Se = \frac{TP}{TP + FN} \tag{2.11}$$

Especificidad (Sp)

Mide su capacidad para detectar correctamente los casos negativos. También se conoce como tasa de verdaderos negativos (TNR) o selectividad. Se calcula como el número de malos clientes correctamente identificados dividido por el número total de deudores clasificados como incumplidos. Es la medida de **exhaustividad** para los clientes malos.

$$Sp = \frac{TN}{TN + FP} \tag{2.12}$$

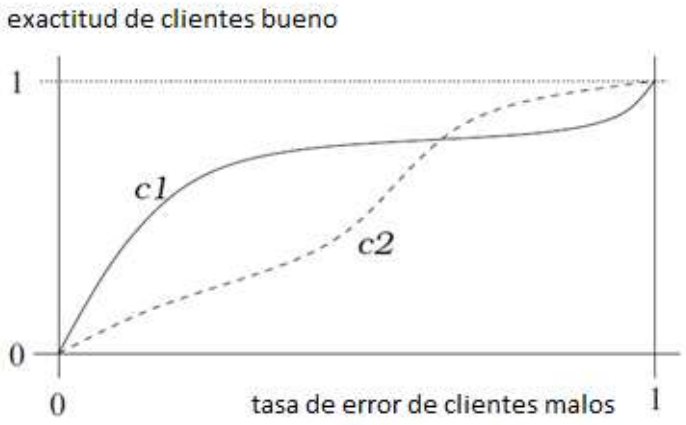
Curva Roc

Curva ROC (Receiver Operating Characteristic) es un instrumento utilizado para valorar el rendimiento de un algoritmo de clasificación binaria [6]. Se construye a partir de las tasas de verdaderos positivos y falsos positivos obtenidos al variar el umbral de decisión del modelo[7]. El eje x representa la tasa de error de los deudores en estado de incumplimiento, y el eje e incorpora la exactitud de los clientes buenos. El usuario puede escoger que prefiere mayor precisión o exhaustividad. La figura 2.2 se observa ejemplos de la curva Roc para dos clases, 0= buen cliente, 1= mal cliente, los parámetros de los calificadores se pueden utilizar para modificar los números de falso positivos y falsos negativos.

[6] Clasificación binaria es la tarea de clasificar los elementos de un conjunto en dos grupos.

[7] El umbral de decisión es el valor en el cual se decide si un caso es un buen o mal cliente en base al resultado del modelo de clasificación. Por defecto, la mayoría de los algoritmos utilizan un umbral de 0.5 de un puntaje entre 0 a 1, pero puede ser cambiado para adaptarse mejor al problema en cuestión.

Figura 2.2: Ejemplo de curva Roc



Fuente: Tomado de Kubat M. (2021)

AUC

Es el área bajo la curva ROC, es utilizada para evaluar el algoritmo, su valor se encuentra entre $0.5 \leq AUC \leq 1$, valores cercanos a 1 indica que el modelo de clasificación funciona bien.

2.2.1.3 Error en las observaciones no estimadas (error OOB)

Cada árbol de BA se construye utilizando bootstrap, lo que significa crear una muestra para entrenar el algoritmo, utiliza el 63% de la información, lo que resulta que existen registros no utilizados, las cuales se denominan observaciones OOB. Estas se utilizan para calcular el desfase del modelo entrenado calculando la falla de predicción denominado error OOB. Este procedimiento fue introducido originalmente por Breiman (1996) y se ha convertido en un método establecido para la estimación de errores en BA.

Velocidad de cálculo y ahorro de memoria son las ventajas de utilizar error OOB en el algoritmo BA (Janitza, S. & Hornung, R., 2018); por tales motivos utilizaremos este indicador para optimizar los hiperparámetros.

2.2.1.4 Importancia de las variables

BA al tratarse de una combinación múltiple de árboles no es factible dar una representación gráfica o intuitiva del modelo, para lo cual se eligen medidas para cuantificar la importancia de los predictores o variables dependientes, Dos indicadores son: “Disminución media de Gini” y “Disminución Media de exactitud”.

2.2.1.4.1 Disminución media de Gini

La importancia de una variable de un árbol de clasificación se puede medir utilizando el criterio de Gini, que cuando es utilizado como función de impureza se lo conoce como “Importancia de Gini” o “Disminución media de Gini”. Consiste en seleccionar la variable en cada partición en la construcción de los árboles y observar cual provoca una disminución de esta medida. “La importancia de una variable en un árbol se mide como la suma de los decrementos atribuidos a esa variable y la importancia resultante es el promedio en todos los árboles”. (Medina, R. & Ñique, C., 2017).

Índice de Gini

Mide el grado de “impureza” de un nodo en un árbol: índices Gini = 0 indican nodos puros, es decir, son datos que pertenecen a una sola categoría, mientras 0<Gini <=1 indican nodos con impurezas; es decir, con datos de más de una categoría. Se puede especificar de la siguiente manera:

$$Gini = 1 - (Probabilidad\ cliente\ bueno)^2 - (Probabilidad\ cliente\ malo)^2 \quad (2.13)$$

2.2.1.4.2 Disminución media de Exactitud

En el proceso de aprendizaje del BA se obtiene la muestra OOB, y en esta se puede evidenciar la jerarquía de las variables, de la cual se observa como disminuye la exactitud cuando se permutan las variables, la lógica se explica a continuación: primero se considera el error OOB inicial en la base de construcción, se selecciona una variable al azar y se permutan provocando un nuevo aprendizaje del modelo y nuevas proyecciones, posterior se

vuelve a calcular el error OOB del nuevo modelo y se lo compara con el inicial, si el error cambia se entiende que la variable escogida es significativa, este proceso se repite con todas las variables y se ordenan basándose a los cambios que produjeron cada uno del error OOB (Medina, R. & Ñique, C., 2017).

2.2.2 Algoritmo K-medias

Es un algoritmo supervisado que permite descubrir agrupamientos en conjuntos de información, y tiene como enfoque principal la partición de un conjunto de n observaciones en k grupos. Donde el grupo se encuentra representado por el promedio de los puntos que lo componen, el representante se denomina centroide. El número de conglomerados k es un parámetro que se fija desde el inicio. El procedimiento de agrupamiento comienza con k centroides colocados aleatoriamente y asigna cada observación al centroide más cercano. “Después de asignarlos, los centroides se mueven a la ubicación promedio de todos los datos asignados a él, y se vuelven a reasignar los puntos de acuerdo con las nuevas posiciones de los centroides” (MacQueen, J., 1967).

El algoritmo no siempre encuentra la configuración óptima, elije la configuración correspondiente al valor mínimo de la función objetivo, por lo que siempre termina. “Hallar un mínimo de la función, a pesar de que no se trate del mínimo absoluto, garantiza un agrupamiento en el que los grupos son poco dispersos y se encuentran separados entre sí” (MacQueen, J., 1967). K-media es significativamente sensible a los centroides que se seleccionan inicialmente de manera aleatoria.

Igual & Seguí (2017), describe la teoría base del algoritmo de K-medias, mismo que es una técnica de partición y tiene como objetivo colocar a cada dato dentro de un grupo, posterior se divide a un grupo de información X en k grupos distintos $c_{i,}$ $\mathrm{i} = 1,2,3,\ldots,\mathrm{k}$, los cuales están descrito por las medias μ_i explicado en las segmentaciones, la cual es llamada centroide, el agrupamiento se resuelve mediante la minimización del problema:

$$\arg min_c \sum_{j=1}^{k} \sum_{x \epsilon c_j} d(x, \mu_j) = \arg min_c \sum_{j=1}^{k} \sum_{x \epsilon c_j} ||x - \mu_j||_2^2 \quad (2.14)$$

Donde c_j es el conjunto de puntos dentro del subgrupo i y μ_i que es el centro de las clases c_i, el algoritmo para definir la similitud de los datos utiliza el trecho entre ellos, en este caso la distancia Euclidiana expuesta a continuación:

$$d(x, \mu_j) = ||x - \mu_j||_2^2 \quad (2.15)$$

El algoritmo de la K-medias también se llama algoritmo de Lloyd's y su objetivo es de forma interactiva encontrar los grupos de K-medias como lo describe (Igual, L. & Seguí, S., 2017):

1. Se define el número de grupo k.
2. Se establece k centroides de forma aleatoria.
3. Se asigna cada punto n del conjunto de datos a la agrupación cuyo centroide k está más cerca. Posterior, se calcula un nuevo centroide para cada grupo como la media de los puntos asignados a ese.
4. Se vuelve a estimar el centroide k, c_i, asumiendo que los grupos formados son correctos.
5. Si ninguno de los n datos cambian de clúster termina el algoritmo, caso contrario pase al paso 3.

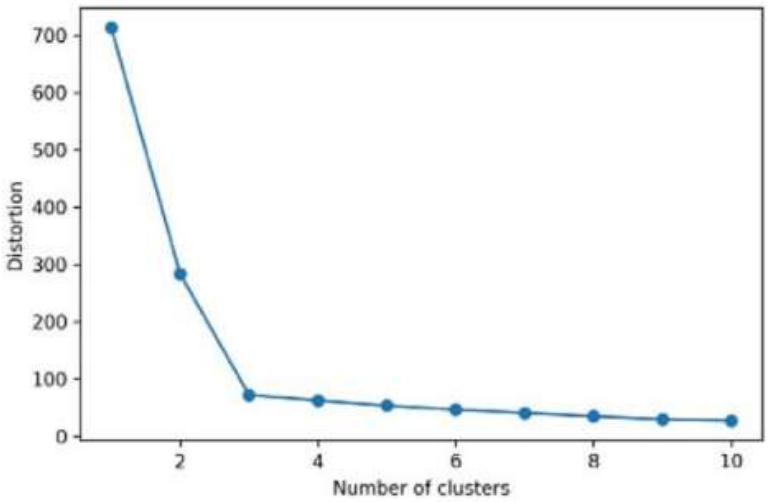
2.2.2.1 Método Elbow o codo

Es conveniente estimar el valor óptimo de agrupaciones en la muestra, para ello se utiliza la distorsión que produce cada grupo, por tal motivo se calcula la inercia obtenida tras aplicar el K-medias a diferentes clústeres, siendo la inercia la suma de las distancias al cuadrado en el objeto x_i del clúster a su centroide μ.

$$Inercia = \sum_{i=0}^{N} \|x_i - \mu\|^2 \tag{2.16}$$

Raschka & Mirjalili (2019) describe el método de Elbow como una herramienta gráfica que permite estimar el número óptimo de k grupos. La idea que hay detrás de la técnica es identificar el valor de k donde la distorsión empieza a aumentar rápidamente, que será evidente si representamos para diferentes valores de k, en el gráfico 2.3 se puede observar un ejemplo de lo descrito.

Figura 2.3: Gráfico Elbow



En el gráfico 2.3 se observa que los grupos óptimos son con $k = 3$, y es una buena opción para agrupar este conjunto de datos.
Fuente: Adaptado Raschka & Mirjalili (2019).

Para utilizar el algoritmo K-medias, es necesario definir las características que van a ser utilizadas, para proponer las variables de agrupamiento, se basó en la idea que proviene de una metodología; que se utiliza en marketing, que segmenta los clientes mediante los campos Reciente, Frecuencia y Monetario (RFM).

2.2.3 Reciente, Frecuencia y Monetario (RFM).

Reciente (R) es número de días transcurridos desde la última transacción. Frecuencia (F) es la cantidad de veces que el cliente realizó una compra. Monetario (M) significa la suma de

todo el dinero utilizado en las transacciones del periodo; esto quiere decir que se basa en el comportamiento de clientes observando el tiempo trascurrido de su última compra, la frecuencia que realiza las compras, y su monto de realización (Cuadros, Gonzalez, & Jiménez, 2017)

El presente trabajo realiza un símil del análisis RFM, pero con variables que pueden alertar de un posible riesgo crediticio y su impacto para la institución financiera. Se propone utilizar los indicadores: PI, MV y PR, las cuales permiten comprender la situación del cliente en la entidad bancaria, observando su probabilidad de incumplimiento, el monto de afectación en caso de que esto ocurra y el tiempo de permanencia del crédito, el objetivo es agrupar clientes con características similares, para realizar una propuesta de gestión de recuperación temprana adecuada.

En función de la agrupación se define la maniobra de recuperación del crédito- Las estrategias establecen los pasos a seguir para el cobro de la deuda, los criterios de negociación, condonaciones, entre otros; requiriendo que sean adecuados a cada segmento establecido y que tendrá seguramente un efecto más preciso de conseguir que el cliente cumpla con sus obligaciones (Morales J. & Morales A., 2014).

2.3 Estrategias de recuperación

Las estrategias que se usan para la recuperación del crédito se establecen de acuerdo con el grado de incumplimiento en los pagos del cliente. Los tipos de cobranza existentes según lo plantean Morales Castro & Morales Castro, (2014), son:

- ***Cobranza normal:*** al momento de recibir el pago, se emite su factura o estado de cuenta, con lo cual al cliente se informa de la evolución del crédito.
- ***Cobranza preventiva:*** en este caso se realiza un recordatorio de fechas de vencimiento próximas o recientes para los clientes, puede ser vía telefónica, correo, o bien por visitadores.

- ***Cobranza administrativa:*** el objetivo principal es contactarse con el cliente para obtener una promesa de pago.
- ***Cobranza domiciliaria:*** es atendido por un gestor, el cual realiza una visita domiciliaria.
- ***Cobranza extrajudicial:*** contacto personal con el deudor a fin de negociar la deuda. También pueden intervenir agencias de cobranza externas a la institución.
- ***Cobranza judicial:*** se inicia en cobro en los tribunales de justicia correspondiente, si el deudor no paga conforme al dictamen del juicio, el juez pude determinar otras acciones como embargo, liquidación de garantías entre otros.

CAPÍTULO 3

3 Metodología Analítica

En el presente capítulo presentan los pasos para aplicar la metodología presentada, en primera instancia se calcula la PI con BA posterior se utiliza K-medias para realizar la segmentación de clientes, se presenta resultados de cada algoritmo como los indicadores de calidad.

Se utilizará el software R para realizar el cálculo de los algoritmos e indicadores de calidad de os mismos, así como los gráficos presentados; también se utilizó la herramienta Excel.

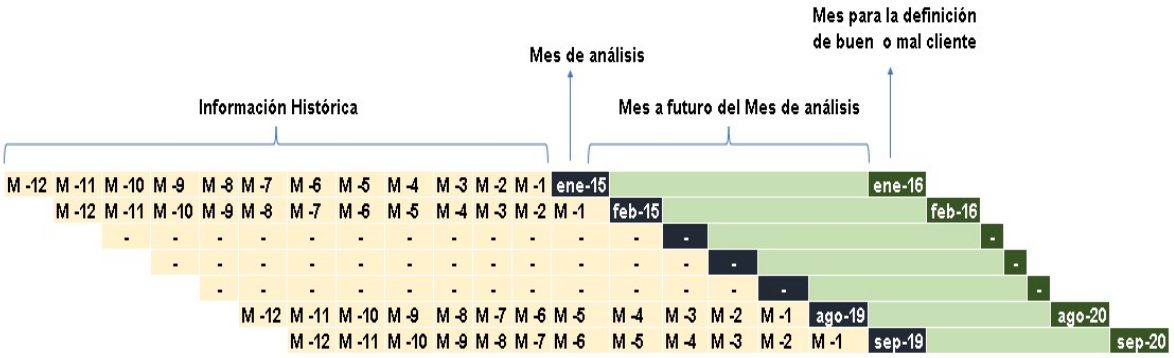
3.1 Descripción de la base de la información

La información utilizada, para desarrollar los algoritmos, pertenecen cartera de crédito de clientes PYME de una entidad financiera ecuatoriana, y contiene los siguientes bloques de información:

- **Información de Buró,** es un reporte que contiene las deudas registradas en el sistema financiero del Ecuador, en este se observa la morosidad histórica del cliente, vista desde los días de no pago de la obligación como en el segmento de colocación del crédito (comercial, consumo, vivienda, microempresa).
- **Información interna de la institución financiera,** contiene datos del deudor al corte de análisis, el monto adeudado, plazo del crédito, mora histórica un año atrás, e información para el cálculo de la definición del cliente malo.

Se posee 1.609.472 de registros de clientes, que comprenden una historia desde enero 2015 hasta septiembre 2019, operaciones menores a 90 días de mora, y en cada mes de análisis se observa doce meses a futuro para dar la definición de buen y mal cliente, la forma de construcción se grafica en la figura 3.1.

Figura 3.1: Esquema de información



Fuente: Elaboración propia.

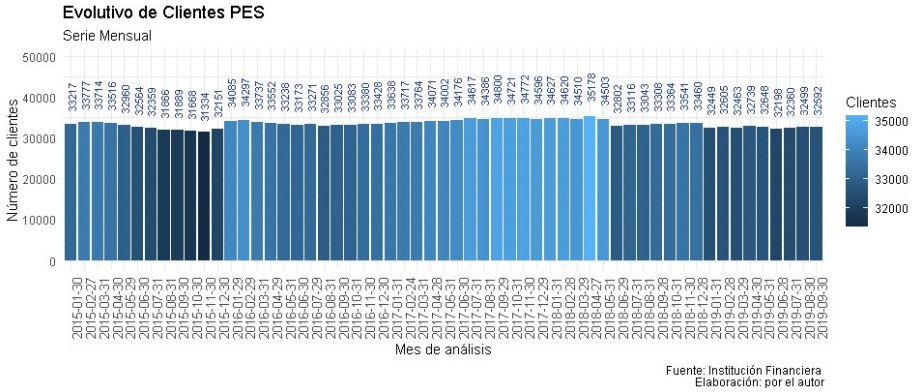
Mes de Análisis: mes donde se selecciona las operaciones para la construcción del modelo, cubre los meses de 2015 a septiembre 2019.

Información Histórica: a los clientes seleccionados se observa su comportamiento en meses anteriores al mes seleccionado. La ventana de estudio es hasta un año anterior.

Mes para la definición de buen y mal cliente: los clientes seleccionados se observan un año a futuro del mes de análisis, con el objetivo de observar si el deudor cumple con la condición de incumplimiento, y si lo hace se define como un mal cliente.

La evolución de los clientes se gráfica en la figura 3.2.

Figura 3.2: Serie Histórica del número de clientes



Fuente: Elaboración propia.

3.2 Cálculo de la Probabilidad de Incumplimiento

En primera instancia, se calcula la probabilidad de incumplimiento, el que se obtiene al aplicar el algoritmo de Bosques Aleatorios.

3.2.1 Comportamiento de la variable dependiente.

En el segundo capítulo se propone la definición de la variable dependiente que refleja el estado de los buenos y malos clientes, en la figura 3.3 se procede a graficar la tasa de deudores con la condición de mal pagador, y se analiza su comportamiento en el tiempo.

- **Cliente malo,** se define a todo deudor que tenga una operación de crédito con más de 90 días mora, o castigada, o demandada, después de 12 meses del mes de análisis.
- **Cliente Bueno,** es todo prestamista que no posee la definición de mal cliente en la ventana de estudio.

Se calcula la tasa de malos bajo la siguiente formula:

$$Tasa\ de\ malos = \frac{cantidad\ de\ malos\ en\ el\ mes\ de\ análisis}{Total\ clientes\ en\ el\ mes\ de\ análisis} \tag{3.15}$$

Figura 3.3: Evolutivo de la tasa de malos clientes

Fuente: Elaboración propia.

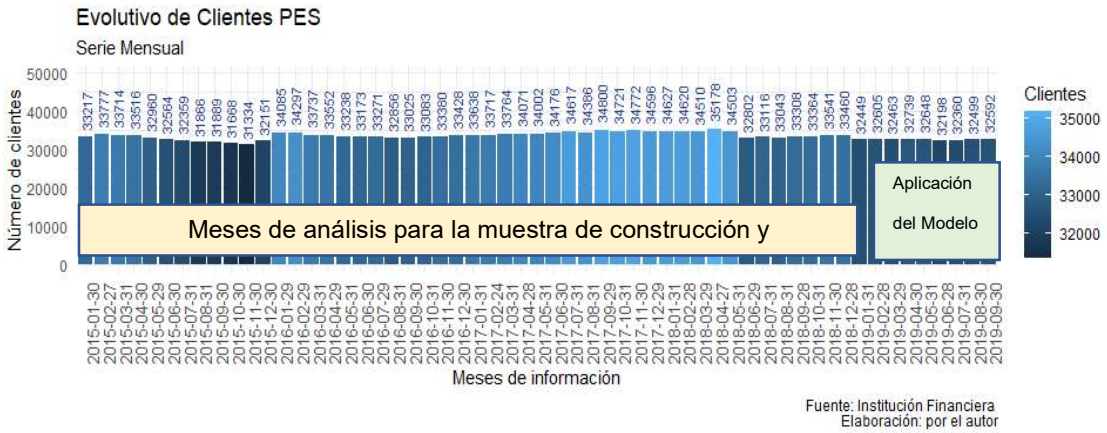
Se observa que la tasas de malos alcanzo su porcentaje mayoritario de 7.4% en julio 2015, e inicios del año 2019 empezó a tener unos crecimientos elevados, pero sin alcanzar su máximo; un mínimo de 2.4% en el mes de marzo del 2018. Adicional los deudores no superan el 6.6% en el año 2019, se evidencia que existe una desproporción en la muestra.

3.2.2 Muestra de entrenamiento, prueba y para la aplicación del modelo.

Se observa una cantidad de clientes buenos de 1,545,242 y de malos de 64,230, lo que resulta un porcentaje de tasa de malos de 3.99% en toda la fuente de información. Se procede a escoger meses para realizar la muestra de construcción, prueba, y meses para aplicar el modelo obtenido.

- Meses de análisis entre enero 2015 hasta diciembre 2018 se utilizarán en la muestra para la construcción del modelo y la prueba del modelo.
- Meses entre enero 2019 hasta septiembre 2019 son los meses para aplicar el modelo.

Figura 3.4: Selección base para construcción, prueba, aplicación del modelo.

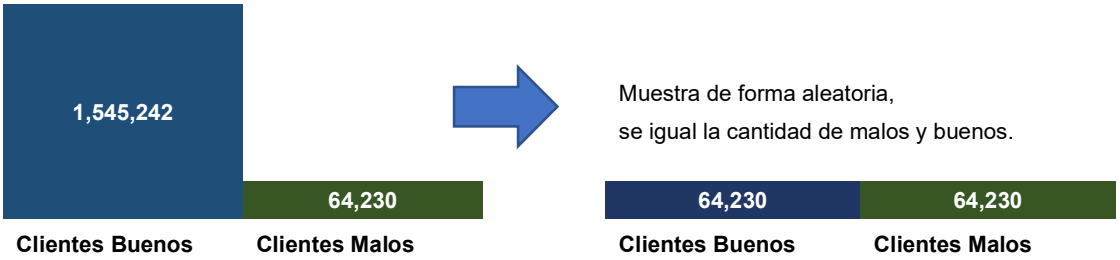


Dado la cantidad de clientes que se posee, y el modelado está asociado con altos costos computacionales se procederá a tomar una muestra, que es una sub-selección de todo el conjunto de datos disponibles.

En el cálculo de la probabilidad de incumplimiento es conveniente encontrar las características de los deudores malos, dado que es el objetivo identificar las particularidades que estos posean y prevenir antes que cumplan su condición de no pago de la obligación. Por tal motivo, se va a seleccionar todos los deudores malos y por la desproporción entre clientes se realiza submuestreo[8] de los buenos.

[8] Submuestreo es una técnica, en análisis de datos, utilizada para ajustar la distribución de clases de conjuntos de datos, en este caso la clase de buenos se igualarán a la clase de malos.

Cuadro 3.1: Submuestreo de Buenos a Malos Clientes



Fuente: Elaborado por el autor.

El resultado del submuestreo son 128,460 clientes, de estos se van a dividir en dos, una para muestra de construcción del modelo y otra para la prueba del algoritmo.

Puertas Medina & Martí Selva (2013) en su análisis de "score de crédito" utiliza una muestra de 50% para la base para construcción y el otro para la muestra de prueba del modelo, el autor considera que por la cantidad de clientes buenos es factible escoger el mismo porcentaje en el presente trabajo.

- La muestra para construcción es de 64,230 registros, contiene 32,115 clientes buenos y 32,115 deudores malos.
- La muestra para Prueba es de 64,230 registros, contiene 32,115 clientes buenos y 32,115 deudores malos.

3.2.3 Reducción de variables explicativas

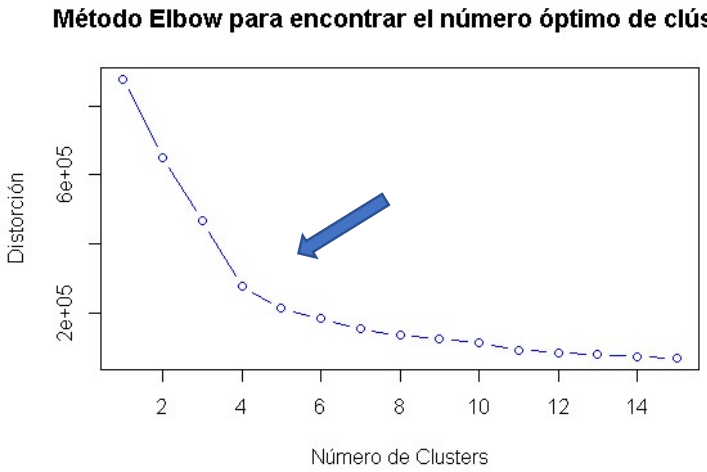
La base otorgada por la entidad posee 70 campos de información, las cuales corresponden a características de Buró e interna de la institución. Primero, necesitamos apreciar qué indicadores tienen el mayor impacto en la definición de buen y mal cliente. En el capítulo 2.2.1.4 se realizó la explicación de la importancia de las variables, se puede utilizar este método para reducirlas, tomando en consideración las medidas de Disminución media de Gini y de Exactitud, para ello se debe crear un primer modelo de bosques aleatorios donde se identifique en primera instancia la relevancia de las variables.

Los parámetros utilizados en un primer modelo BA son los siguientes:

- Número de árboles = 500
- Número de variables = 7^{9}
- Número de registros en el nodo = 10

Con el resultado del algoritmo BA inicial, se aplica el modelo de K-medias en los indicadores de disminución media de Gini y Exactitud, para lograr ese objetivo, en primera instancia se observa el número óptimos de clúster, y utilizando el método de Elbow se selecciona 4 grupos, dado que se evidencia en la figura 3.5 que la distorsión disminuye y forma un codo para los 4 conglomerados.

Figura 3.5: Método de Elbow aplicado en la reducción de variables



Fuente: elaborado por el autor.

En el cuadro 3.2 se puede observar la importancia de las variables, se evidencia que el que mayor aporta es la variable de Días mora iniciales y que menos influencia en la variable dependiente es X46.cerrado_mal_manejo_año, X47.monto_recuperado_año.

Cuadro 3.2: Importancia de las variables en el primer modelo BA.

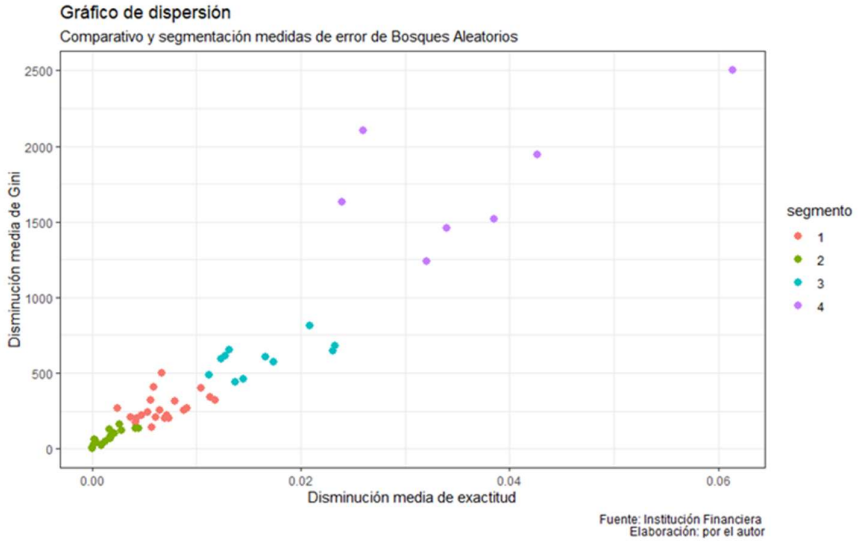
[9] El número de variables que se seleccionan en un modelo de Bosques Aleatorios es $\sqrt{p}$, con p números de variables, para nuestro caso p = 59, variables propuestas 7.

Variables	Disminución Media de Precisión	Disminución Media de Gini
X2.Region	0.001667195	527.4922000
X7.Calificacion_buro	0.007394884	1946.7186
X5.Dias_mora_iniciales	0.027293056	3811.0332
X10.obligaciones_buro	0.006792505	742.0569000
X12.obligaciones_consumo	0.005673701	577.5220000
X15.obligaciones_0_mora	0.108621800	1041.4710
X16.obligaciones_30_mora	0.003733529	256.3459000
X17.oblioac1 ones_60JnOra	0.005045759	382.0028000
X18.obligaciones_90_mora	0.005243699	369.4641000
X19.obligaciones_120_mora	0.022810215	2844.8891
X20.obligaciones_castigadas	0.004542100	360.7772000
X23.obligaciones_sector_financiero	0.004922507	503.6113000
X25.obligaciones_sector_real	0.002413327	448.8033000
X26.saldo_sector_financiero	0.010476329	1350.6891
X28.saldo_sector_real	0.004046448	872.7514000
X29.saldo_mora_financiero	0.023005815	4259.4162
X31.saldo_mora_sector_real	0.005221508	685.2681000
X32.cuota_sector_fianciero	0.009619032	1262.2358
X34.cuota_sector_real	0.001863921	548.5531000
X35.cuota_tarjeta	0.026231436	1407.4295
X36.utilizacion_tarjetas	0.025463950	1266.6308
X37.Tarjetas_buro	0.007308475	627.5590000
X38.Mora_tarjeta_buro	0.001384870	379.6449000
X39.Mora_bancaria_buro	0.019254463	1981.8165
X41.Mora_vivienda_buro	0.003123749	346.7405000
X42.Mora_30_año	0.002275935	398.7080000
X43.Mora_60_año	0.003161393	574.6561000
X44.Mora_90_año	0.004262424	608.6670000
X45.Mora_120_año	0.010861741	1805.7565
X60.Mora_11	0.002855838	434.0841000
X61.Mora_10	0.003311960	459.1964000
X62.Mora_9	0.003825163	487.9671000
X63.Mora_8	0.003367270	501.1310000
X64.Mora_7	0.004271859	617.7785000
X65.Mora_6	0.001736892	661.4036000
X66.Mora_5	0.004666401	774.9299
X67.Mora_4	0.005825681	995.8010
X68.Mora_3	0.007813242	1140.8725
X69.Mora_2	0.007498864	1484.8298
X70.Mora_1	0.014647862	2456.6120
X50.mora_consumo_IF	0.006454800	1288.0785

Fuente: Elaborado por el autor.

Mediante los 4 clúster sugerido por el método de Elbow, se segmenta las variables independientes y se selecciona los grupos que mayor contribuyen a la clasificación del cliente. En la figura 3.6 se observa que las variables que más aportan es el clúster categorizado 1, seguido por el 4, las cuales contienen 17 campos entro los dos, y son los indicadores escogidos para realizar el cálculo de la PI, y el que menos aporta como características son el grupo 3 y 2, el cual se procede a eliminar para posterior aplicar en el algoritmo de BA.

Figura 3.6: Gráfico de Dispersión de las variables en su respectivo grupo.



Fuente: Elaborado por el autor.

En el cuadro 3.3 se observa las variables finales, entre ellas se destaca que tienen importancia si el prestamista ha presentado mora en cualquier producto hasta 5 meses atrás, saldos de buró y morosidad del cliente, su forma de utilizar la tarjeta de crédito, si presenta créditos con destino consumo, de sus días de no pago de la deuda al momento de análisis, y la calificación de buró que este posee.

Cuadro 3.3: Variables con mayor importancia para BA

Variables	Disminución Media de Precisión	Disminución Media de Gini
X7.Calificacion_buro	0.007394884	1946.7186
X5.Dias_mora_iniciales	0.027293056	3811.0332
X15.obligaciones_0_mora	0.108621800	1041.4710
X19.obligaciones_120_mora	0.022810215	2844.8891
X26.saldo_sector_financiero	0.010476329	1350.6891
X29.saldo_mora_financiero	0.023005815	4259.4162
X32.cuota_sector_fianciero	0.009619032	1262.2358
X35.cuota_tarjeta	0.026231436	1407.4295
X36.utilizacion_tarjetas	0.025463950	1266.6308
X39.Mora_bancaria_buro	0.019254463	1981.8165
X45.Mora_120_año	0.010861741	1805.7565
X66.Mora_5	0.004666401	774.9299
X67.Mora_4	0.005825681	995.8010
X68.Mora_3	0.007813242	1140.8725
X69.Mora_2	0.007498864	1484.8298
X70.Mora_1	0.014647862	2456.6120
X50.mora_consumo_IF	0.006454800	1288.0785

Fuente: Elaboración propia.

3.2.4 Optimización de hiperparámetros BA

Con las 17 variables explicativas, con importancia demostrada, se procede a realizar la optimización de los hiperparámetros.

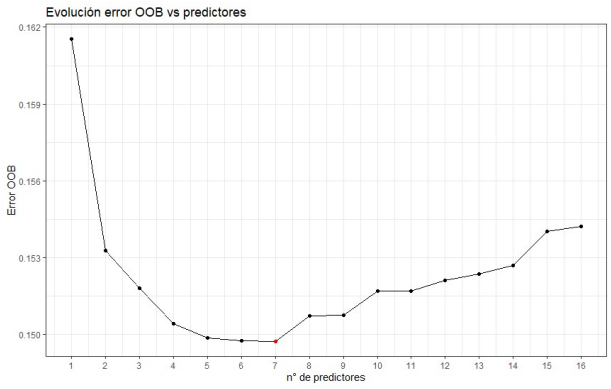
3.2.4.1 Optimización de número de predictores

Para la cantidad óptima de variables explicativas, se ira incrementado el número de variables y se aplica el algoritmo BA, se calcula el error OOB para cada cifra y se procede a graficarlo, se escoge el número de variables que posean el menor valor. Para nuestro caso es 7, se evidencia en el gráfico 3.7 que superior al mínimo se incrementa el error de predicción.

El modelo toma los siguientes parámetros para aplicar el algoritmo BA,

- Número de árboles = 500
- Número de nodos terminales = 10

Figura 3.7: Representación del OBB_error vs el número de predictores



Fuente: Elaboración propia.

.

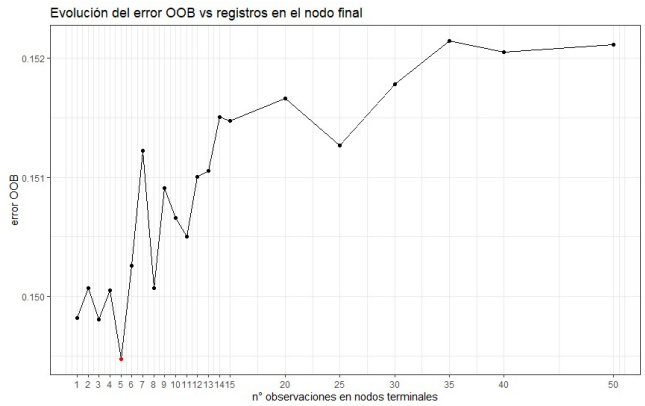
3.2.4.2 Optimización de número de registros en cada nodo final

Parte de la optimización es escoger la cantidad de clientes en cada nodo final, lo que promueve a disminuir el sobreajuste en los árboles, el proceso constituye en ir variando el número de registros en cada nodo final, empieza desde 1 registro hasta 15 en pasos de 1, posterior en incrementos de 5 hasta llegar a 50. En cada interacción se calcular el error OBB, y se elige el que menor. En este caso el valor óptimo es 5 en los nodos finales, como se evidencia en el gráfico 3.8.

Para la optimización se utilizó los siguientes parámetros:

- Número de árboles 500
- Número de variables 7

Figura 3. 8: Error OBB vs el número de registros en los nodos finales.



Fuente: Elaboración propia.

En el incremento de cada registro en el nodo final se observa que es estable entre 1 hasta 4, con 5 registros se obtiene el más bajo valor de error y si incrementa si este aumenta.

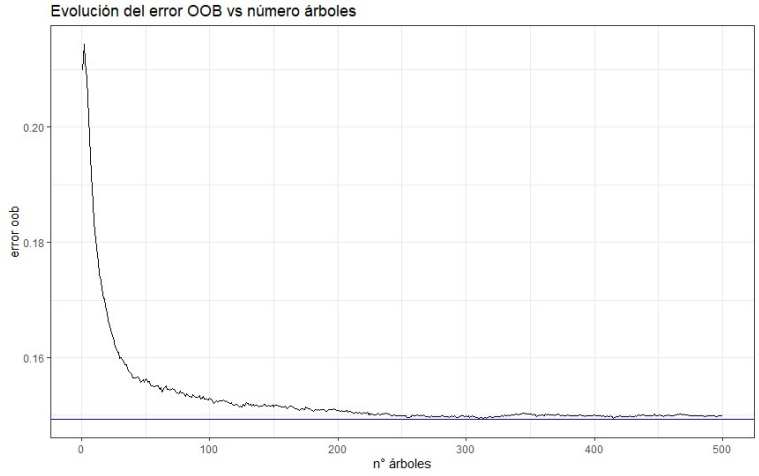
3.2.4.3 Optimización del número de árboles

La cantidad de árboles técnicamente no es un hiperparámetro, pero debe ser lo suficientemente grande para estabilizar la tasa de error. Más árboles representan una estimación de error más robusta y estable. Es posible que se necesiten menos árboles a medida que se ajustan los hiperparámetros, los cuales son: el número de variables y registros en los nodos terminales. Se observa en la Figura 3.9, el valor cae en 352 árboles, después de lo cual el error OBB comienza a estabilizarse.

Para la búsqueda de número de árboles con menor error se tomó los siguientes parámetros:

- 5 registros en los nodos terminales
- Número de variables 7

Figura 3.9: Representación del error OOB vs la cantidad de árboles



"mínimo error: 0.14910477969796 número de árboles con mínimo error 352"

Fuente: Elaboración propia.

3.2.4.4 Elección del modelo

Una vez obtenido los valores óptimos se procede a generar el algoritmo final de bosques aleatorios bajo los siguientes parámetros:

- Número de variables=7
- Registros en cada nodo final=5
- Número de árboles= 352

Figura 3.10: Imagen de la Ejecución del modelo BA en el programa R

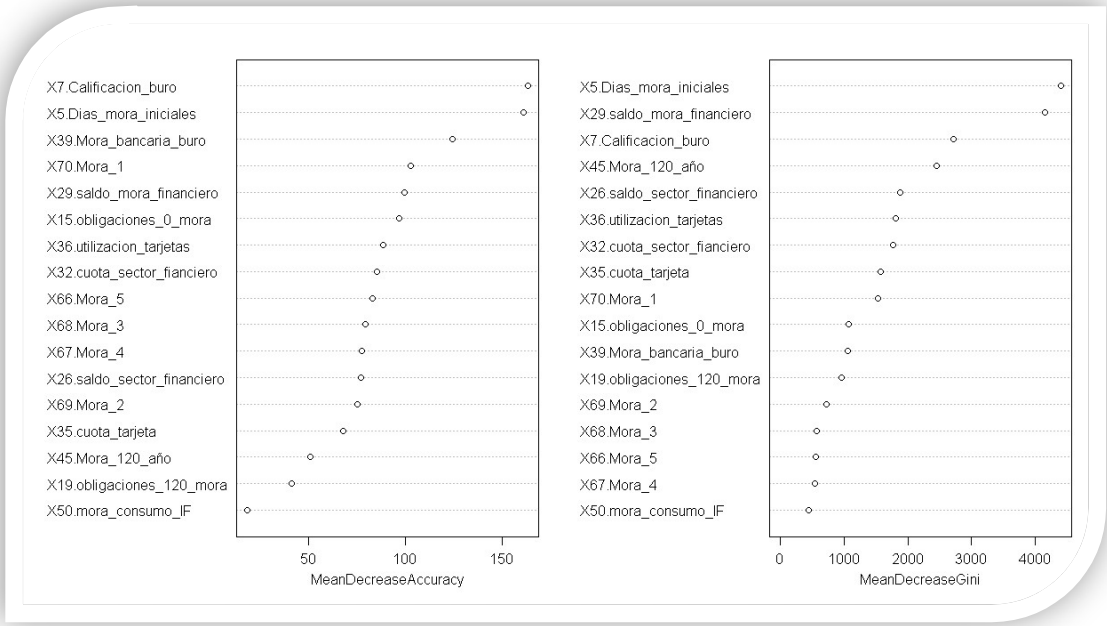
```
> modelo_rf

Call:
 randomForest(formula = as.factor(Construccion2[, 2]) ~ ., data = Construccion2[,      -2], mtry = 7, ntree = 352, nodesize = 5, importance =
 TRUE)
               Type of random forest: classification
                     Number of trees: 352
No. of variables tried at each split: 7

        OOB estimate of  error rate: 15.02%
```

Fuente: Elaboración propia.

En la figura 3.10 muestra la aplicación del modelo con los parámetros definidos, y se obtiene un Error OOB de 15.02% y se observa la importancia de las 17 variables al aplicar el algoritmo BA. Se identifica que los campos de Calificación de buró y días mora son las principales que aportan a la estimación de la variable objetivo, lo podemos apreciar en la figura 3.11.

Consolidando la información obtenida, se puede mencionar que los PES tienen variables que pueden predecir su conducta previo a su estado de incumplimiento, el principal es los días de no pago de la obligación que posee actualmente, seguido de su calificación de buró en junto con la cuota mensual que paga, dado que esto mide el comportamiento en todo el sistema financiero del Ecuador; el uso de tarjetas en los clientes es una variable significativa en conjunto con su cuota, y si el cliente presenta mora recurrente [10].

Figura 3.11: Importancia de las variables, modelo final BA



Fuente: Elaboración Propia.

3.2.5 Evaluación estadística del desempeño BA

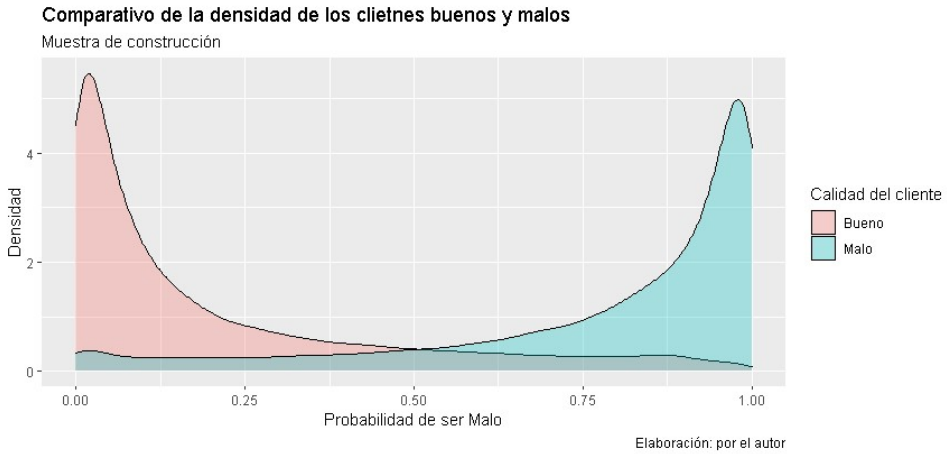
Dado la construcción del modelo se procede a evaluar el rendimiento de este, en primera instancia los indicadores de calidad se evalúan en la base de construcción, posterior se lo realiza en la muestra de prueba, pasando las validaciones se procede aplicar en el grupo que no pertenece al de construcción y muestra.

[10] Mora recurrente significa que el cliente presenta mora en sus obligaciones en todos los meses; clientes por sus preferencias de pago sus cuotas se encuentran atrasadas, pero no significa que no cumple con sus obligaciones.

3.2.5.1 Evaluación estadística en la muestra de construcción

El algoritmo presenta el resultado de una probabilidad de buen o mal cliente, previo a la validación se debe definir el valor que define a un deudore como bueno o malo; es decir, encontrar el corte óptimo donde se maximiza la distribución. Para nuestro caso es 0.515625 como se aprecia en el gráfico 3.12.

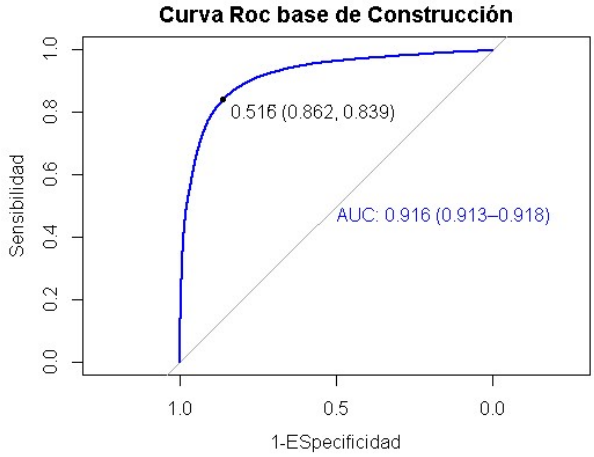
Figura 3.12: Distribución *de* clientes buenos y malos en *la base* de construcción.



Fuente: Elaboración propia.

En el gráfico de la curva ROC y el cálculo de AUC, se evidencia un AUC= 0.916 que es cercano a 1, lo que implica que el modelo esta discriminando correctamente. Adicional en la figura 3.13 presenta que el punto de corte que es 0.516.

Figura 3.13: Curva Roc en la muestra de validación



Fuente: Elaboración propia.

Procedemos a revisar los resultados de la matriz de confusión, con el punto de corte óptimo establecido. En el cuadro 3.4 se presenta los resultados para la muestra de construcción.

Cuadro 3.4: Matriz de confusión en la muestra de construcción

Matriz de Confusión		
	Real	
Predicción	Bueno	Malo
Bueno	27695	5194
Malo	4420	26921
Exactitud	0.8503	
Sensibilidad	0.8624	
Especificidad	0.8383	
Precisión	0.8421	
Exhaustividad	0.8590	
F1	0.8504	

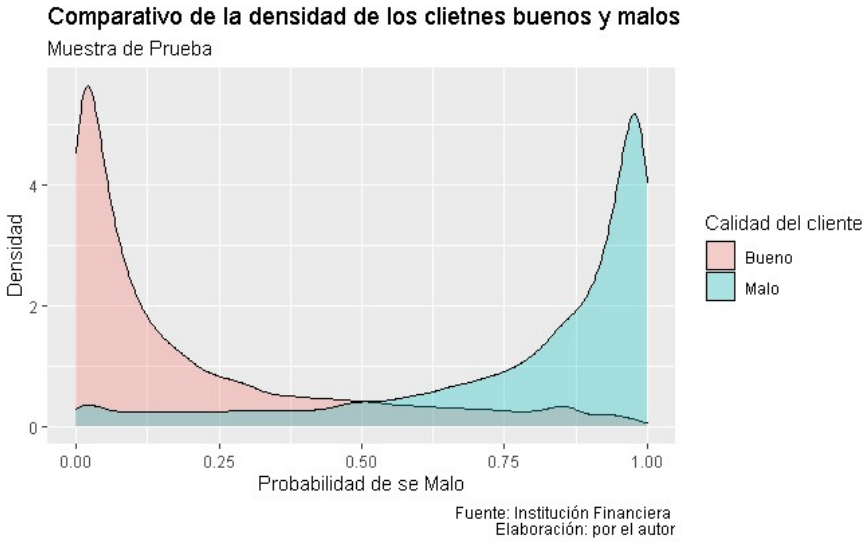
Fuente: elaboración propia.

La exactitud es de 0.85, es decir, el algoritmo predice en un 85.03%, la sensibilidad es del 86.24%, significa que los clientes buenos son evaluados correctamente en un 86.24%. Por otra parte, la cantidad de deudores malos estimados como malos se encuentran en un 83.83% de la muestra. El modelo es más sensible que especifico.

3.2.5.2 Evaluación estadística en la muestra de prueba

Se observa en la figura 3.14 la distribución de clientes buenos y malos en la muestra de prueba, si la comparamos con la figura 3.12 podemos decir que es similar gráficamente que al de la base de construcción. El punto de corte se identifica que el valor es del 48.2%.

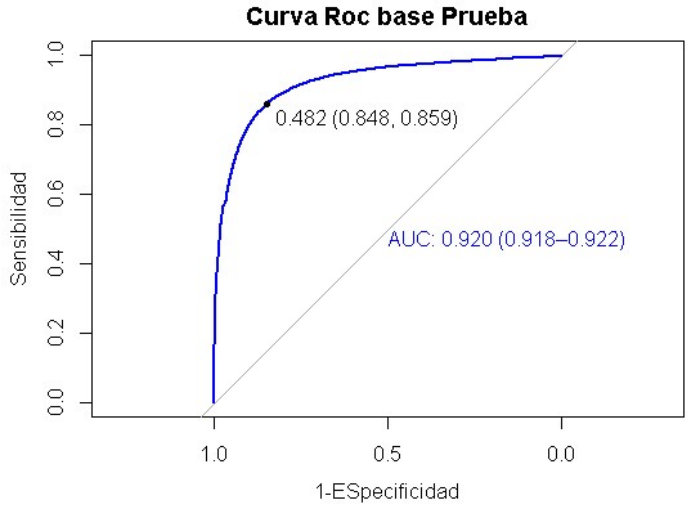
Figura 3.14: Distribución de los clientes, buenos y malos sobre su probabilidad en base de construcción



Fuente: elaboración propia.

La Curva Roc presenta una disminución en el corte óptimo, sin embargo, el AUC es 0.92, y no se aleja del valor obtenido en la muestra de construcción.

Figura 3.15: Curva Roc en la Base de Prueba



Fuente: Elaboración propia.

Sobre la matriz de confusión en la base de prueba se evidencia, que la precisión es de 0.8537 y es cercano al obtenido en la base de construcción, para este caso el algoritmo predice en un 85.37%, y su sensibilidad es del 84.47%. Por otra parte, la cantidad de clientes proyectados como malos se encuentran en un 83.83%. El algoritmo es más sensible que especifico, se mantiene lo mencionado en la muestra de construcción, se lo considera estable.

Cuadro 3.5: Matriz de confusión muestra de prueba

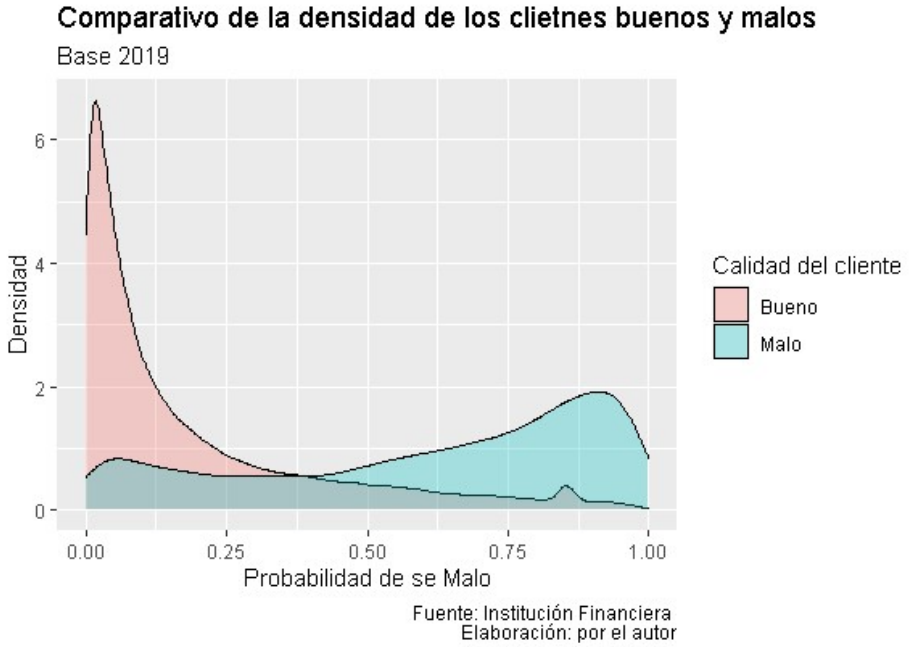
Matriz de Confusión		
	Real	
Predicción	Bueno	Malo
Bueno	27703	7988
Malo	4412	27127
Exactitud	0.8537	
Sensibilidad	0.8626	
Especificidad	0.8447	
Precisión	0.7762	
Exhaustividad	0.8601	
F1	0.8160	

Fuente: Elaboración propia.

3.2.6 Estudio del algoritmo en la base de aplicación del modelo.

En la Fase de modelamiento se separó las bases para aplicar el modelo y observar su efectivad en una muestra fuera del construcción y prueba, en esta base se encuentran 292,553 registros, corresponden a meses del 2019, y se aprecia en la figura 3.16 la distribución de clientes buenos vs malos, el cruce de las distribuciones se desplazó a la izquierda, y se debe a la desproporción de deudores cumplidos e incumplidos.

Figura 3.16: Distribución de los clientes sobre su probabilidad en la base de aplicación del modelo

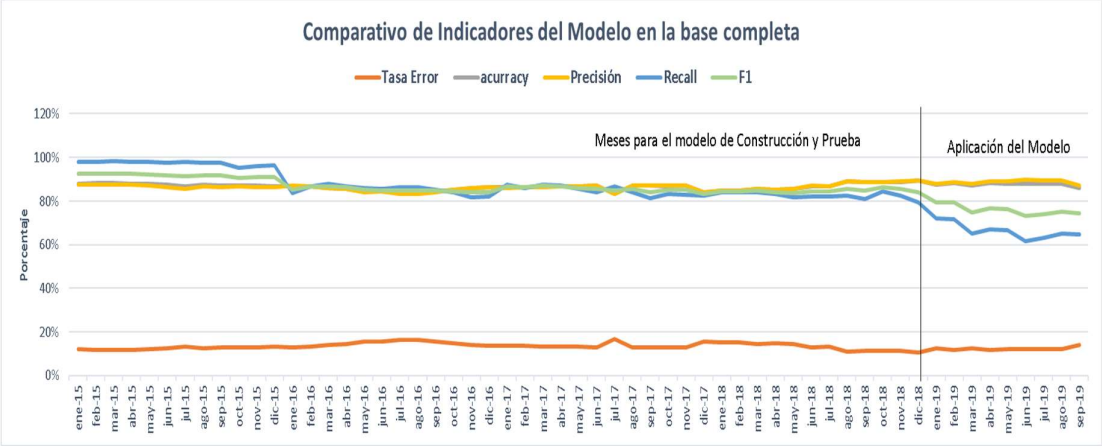


Para comparar el comportamiento de las medidas de calidad, se aplicó el modelo BA en toda la base de datos y se analiza por mes la evolución de los índices, se generó el gráfico 3.17 y se observa el histórico donde se incluye meses del 2019; se calculó los valores F1, y exhaustividad. Gráficamente comienza a disminuir las cifras, se intuye que esas operaciones comenzaron a ser influenciados por el COVID-2019[11], Ecuador estableció confinamiento lo que provocó el incremento de más clientes en condición de incumplidos, dado que sus vetas disminuyeron, y esto afecta las características de los deudores. Por otro lado, los indicadores de la Tasa de error, exactitud, precisión no se ven influenciados.

Se identifica que los indicadores son estables y se puede seleccionar el resultado del algoritmo de Bosques Aleatorios como probabilidad de incumplimiento.

[11] La pandemia del covid-19, obligo a poner restricciones de ventas provocando que algunos clientes PES disminuyan sus ventas, provocando el atraso de las obligaciones financieras.

Figura 3.17: Indicadores de calidad del algoritmo BA en la base completa



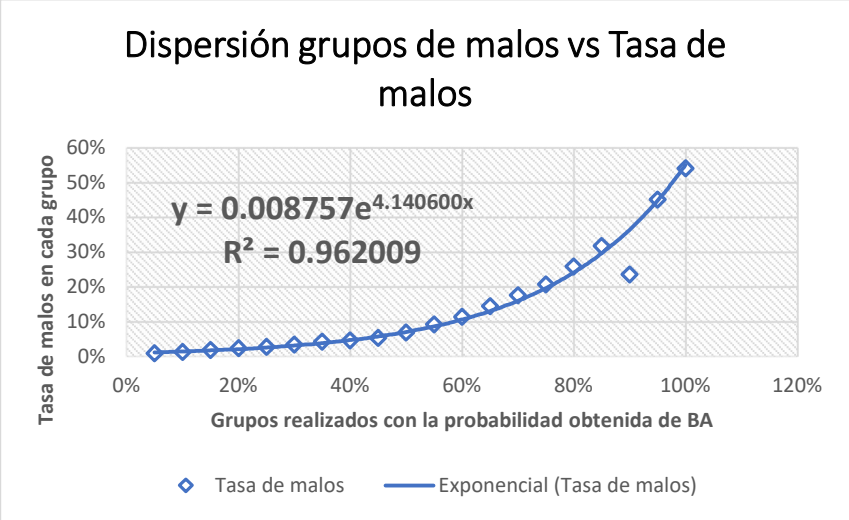
Fuente: Elaboración propia.

3.2.7 Corrección del cálculo de la PI

Se observa el gráfico 3.16 que el punto de corte se movió hacia la izquierda lo que provoca un error si se toma el valor como la probabilidad de incumplimiento, es necesario efectuar un ajuste y el presente trabajo propone realizarlo bajo el siguiente esquema:

- De los meses del 2019, donde se observa el cambio del punto de corte, se procede a realizar deciles de la muestra.
- Se procede a segmentar la probabilidad del BA en porcentajes de 5% hasta llegar a 100%, obteniendo 20 agrupaciones, y se calcula la tasa de malos para cada grupo.
- Se procede a realizar un gráfico de dispersión entre los grupos, para posterior encontrar la ecuación que realice la conversión.

Figura 3.18 Comparativo tasa de malos VS probabilidad obtenida en el algoritmo del BA en la base de aplicación del modelo.



Fuente: Elaboración propia

El mejor ajuste entre el resultado del algoritmo BA y la tasa de malos es una función exponencial, con ello se podría obtener la afinación del valor de la PI, dado que, en la muestra de aplicación del modelo, que asemeja a una ejecución mensual del BA en la institución financiera, el grupo que posee la probabilidad de incumplimiento desde el 90% al 100% realmente tiene una tasa de malos máxima del 50%.

3.3 Aplicación del algoritmo de K-medias

Con el valor del cálculo de probabilidad de incumplimiento tenemos la primera variable para realizar la segmentación, ahora incluimos las variables Monto y Plazo remanente. Para la agrupación de clientes se utilizará el resultado del número obtenido por el algoritmo BA.

3.3.1 Descripción de la información

La base para realizar la segmentación es el mes de septiembre del 2019, se selecciona solo una fecha por el motivo que se desea analizar cómo afectaría en un mes de ejecución o

de aplicación de la estrategia, se realiza en primera instancia una descripción de la información, se obtiene 32,591 registros para este análisis. Es primordial normalizar[12] las variables previo a la aplicación del algoritmo de K-medias.

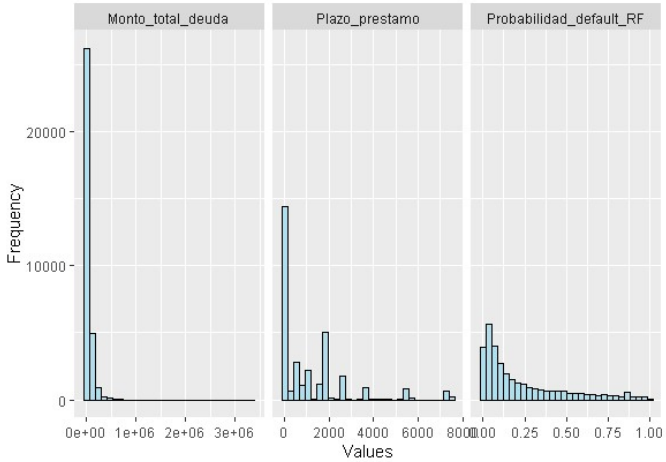
Cuadro 3.6: Estadística descriptiva de las variables para la segmentación.

Probabilidad_incumplimiento	Monto_total_deuda	Plazo_prestamo
Min. :0.00000	Min. : 0	Min. : 1
1st Qu.:0.04261	1st Qu.: 2429	1st Qu.: 30
Median :0.12216	Median : 10498	Median : 540
Mean :0.23175	Mean : 38264	Mean :1172
3rd Qu.:0.35511	3rd Qu.: 42381	3rd Qu.:1813
Max. :1.00000	Max. :3360858	Max. :7500

Fuente: Elaboración propia.

Se observa que los PES tienen deudas vigentes de USD 38,264 en promedio, plazos medios que bordean los 3 años con un máximo de 20 años, asimismo su probabilidad de incumplimiento promedio es del 23.17% y una media el 12.21%. Las distribuciones de las variables indican que hay mayor cantidad de clientes se ubican en valores bajos, así lo presenta la figura 3.19.

Figura 3.19: Distribuciones de las variables para segmentación:

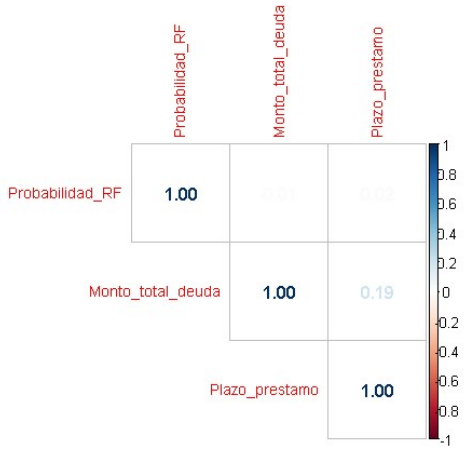


Fuente: Elaboración propia.

[12] Normalización es el cambio de escala de la variable, se resta su promedio y divide la desviación estándar, con ello se logra hacer comparables las variables.

Se observa en la figura 3.20 que no posee correlación las variables, dado que el indicador presenta valores menores al 0.5.

Figura 3.20: Correlación de las variables par a la segmentación de las variables.



Fuente: Elaboración propia.

3.3.2 Aplicación del algoritmo de K-medias

En primera instancia se procede a normalizar las variables, para posterior realizar el método de codo y encontrar los grupos óptimos de clientes. En el cuadro 3.7 se observa que el promedio de los campos es 0 dado que se encuentran normalizados, el monto total de la deuda evidencia un máximo de 160.97, y el 3 cuartil es un valor de 0.01992 lo que implica que pocos PES presentan grandes deudas.

Cuadro 3.7: Análisis descriptivo de las variables normalizadas.

```
 base3.Score      base3.X3.Monto_total_deuda base3.X9.Plazo_prestamo
Min.   :-0.8577   Min.   : -0.25802          Min.   :-0.7107
1st Qu.:-0.7191   1st Qu.: -0.24293          1st Qu.:-0.6931
Median :-0.4421   Median : -0.18941          Median :-0.3831
Mean   : 0.0000   Mean   :  0.00000          Mean   : 0.0000
3rd Qu.: 0.3775   3rd Qu.:  0.01992          3rd Qu.: 0.3902
Max.   : 3.2056   Max.   :160.97661          Max.   : 3.8482
> |
```

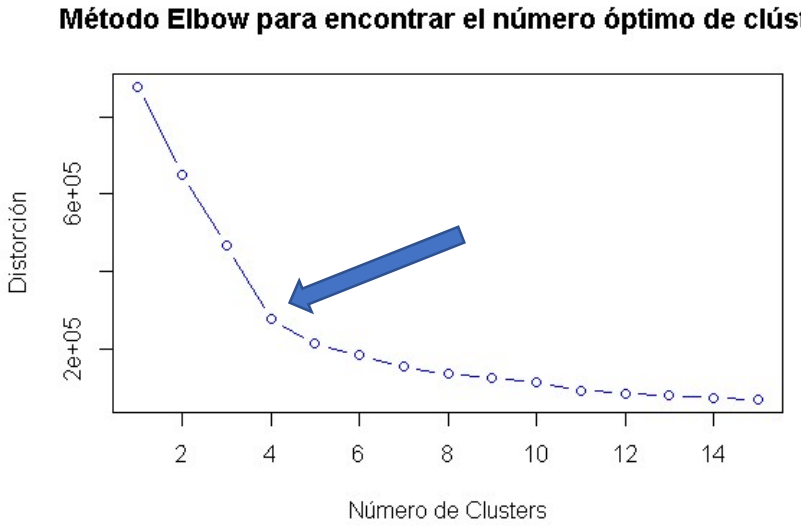
Fuente: Elaboración propia.

En el gráfico 3.21 se evidencia que con 4 grupos la distorsión se disminuye notablemente y para agrupaciones mayores las distorsiones son menores, se considera 4 como número

óptimo. Aplicando el modelo con 4 clúster se prestar atención a las siguientes estadísticas mostradas en el cuadro 3.8, donde se observa lo siguiente:

- Grupo1: se evidencia que tiene la más baja la probabilidad de incumplimiento, en media es de 11.4%, además en promedio el monto de deuda es USD 9,196, y los plazos de 221 días, que son bajos respecto a los otros grupos. Y es el clúster que contiene la mayor cantidad de registros.
- Grupo2: se observa en promedio que tiene los clientes con la mayor probabilidad de incumplimiento, 63.9%, sus montos son los segundos más bajos, USD 27,201, al igual que sus plazos, de préstamos es 849 días. Comparaciones respecto a los grupos realizados.

Figura 3.21: Número óptimo de clúster para la segmentación de los clientes.



Fuente: Elaboración propia.

- Grupo 3: Posee probabilidades de incumplimientos intermedios, 23.3%, y contiene clientes con mayor monto y plazos más extendidos.
- Grupo 4: es el segundo grupo con menor probabilidad de incumplimiento, 11.4%, y en la misma posición con la cantidad de monto, y sus plazos son intermedios, respecto a los demás.

Cuadro 3.8: Promedio de las variables en cada grupo obtenido.

	Promedios de los grupos obtenidos			
Grupo	Monto_deuda	Plazo Remante	Probabilidad default	Número de Registro
1	9,196	221	11.4%	16,066
2	27,201	849	63.9%	6,814
3	111,270	6,206	23.3%	2,094
4	92,411	2,082	11.5%	7,618

Fuente: Elaboración propia.

CAPÍTULO 4

4 Diseño de la estrategia para realizar una recuperación temprana.

Con el objetivo de efectuar la maniobra de recuperación temprana, es decir antes que el cliente alcance su condición de incumplimiento, se decide realizar 3 agrupaciones para proponer la estrategia, y se lo realizara en función a la segmentación descrita en el capítulo 3.3.2 y para poder unificar grupos se incluye el ratio de cartera vencida[13], se evidencia en el cuadro 3.9 la información y basándose a ello se plantea la unificación de grupos.

Cuadro 4. 1: Información descriptivo de los grupos de clientes.

Grupo	Ratio Vencido	Tasa de Malos	Número de Registro	Grupos
1	0.0%	2%	16,066	Riesgo Bajo
2	8.1%	17%	6,814	Riesgo Alto
3	0.9%	5%	2,094	Riesgo Medio
4	0.0%	5%	7,618	Riesgo Bajo

Fuente: Elaboración propia.

- Con la segmentación realizada mediante el algoritmo de K-medias se considera agrupar los grupos 1 y 4, y se establece como riesgo bajo, dado que tiene menor porcentaje de cartera vencida, y su proporción es del 73% de los PES. Adicional se evidencia la tasa de clientes malos que contiene, en el grupo 1 del 2% y del 4 su tasa es de 5%.
- El Grupo 3 se considera agrupación de riesgo medio, contiene un ratio de cartera vencida de 0.9%, tasa de malos del 5%, y 6.4 % de los clientes.
- El grupo 2 se considera segmento de riesgo alto, posee un ratio de cartera vencida del 8.1%, tasa de malos del 17%, y 20.9% de clientes.

[13] Saldo de la Cartera de Crédito vencida en el mes analizado / Saldo de la Cartera de Crédito total del mes analizado, indicador indica la cantidad de clientes que no cumplen sus obligaciones financieras en el mes de análisis.

El cuadro 4.2 proporciona una visión general de las agrupaciones de clientes, evidenciándose que está organizada por el porcentaje de la cartera vencida y su probabilidad de incumplimiento, y es muy adecuado para la gestión recuperación, a partir del cual se puede entender el impacto dentro de la institución financiera.

Cuadro 4.2: Distribución de los clientes sobre los grupos de riesgo.

Grupo	Ratio Vencido	Tasa de Malos	Número de Registro	Promedio Monto Deuda	Promedio Plazo Remante	Promedio Probabilidad Incumpimiento
Riesgo Bajo	0.0%	2%	23,684	$ 35,962	820	11.4%
Riesgo Medio	0.9%	5%	2,094	$ 111,270	6206	23.3%
Riesgo Alto	8.1%	17%	6,814	$ 27,201	849	63.9%

Fuente: Elaboración propia.

Para aplicar la estrategia de recuperación temprana sobre las categorías sugeridas, se toma las descritas por Morales Castro & Morales Castro, (2014), además se separa en dos grupos a los clientes, los que poseen desde un día no pago del crédito y los que no, la propuesta es la siguiente:

- Clientes en Riesgo bajo:

  Los clientes en esta agrupación poseen la más baja tasa de malos (2%) y PI (11.4%), la posibilidad de que la institución financiera se vea afectado por un incumplimiento del grupo es bajo, dado que su riesgo de no pago es menor respecto a los demás grupos observados en el cuadro 4.2. Por tal motivo, se propone una gestión de recuperación normal y preventiva y se lo plantea de acuerdo con los días de mora que el cliente posea.

  - Clientes con 0 días de mora, cobranza normal, recordatorio de las fechas de pago o correo de agradecimiento al momento del pago.
  - Clientes con más de 1 día de mora, Cobranza preventiva mediante vía telefónica o correo electrónico.

- Clientes en Riesgo medio:

Poseen una tasa de malos del 5% y su probabilidad de incumpliendo del 23.3%, los clientes tienen un monto promedio de deuda más alto con respecto a los demás grupos y una permanencia mayor en la institución, es necesario elevar la gestión de recuperación en el deudor, para ello se plantea una cobranza preventiva y domiciliaria basándose en los días mora que el cliente posea.

- Clientes con 0 días de mora, cobranza preventiva mediante vía telefónica o correo electrónico.
- Clientes con más de 1 día de mora, cobranza administrativa mediante vía telefónica con un acuerdo de pago en firme o domiciliaria.

- Clientes en Riesgo alto:

Clientes que poseen la mayor tasa de malos y PI, 17% y 63.9% respectivamente, ellos tienen una alta afectación a la institución por ende la maniobra de recuperación es más exigente, se propone una gestión administrativa y extrajudicial, siempre con el enfoque de evitar que el cliente llegue a su condición de no pago.

- Clientes con 0 días de mora, cobranza administrativa mediante vía telefónica con acuerdo de pago en firme[14].
- Clientes con más de 1 día de mora, Cobranza extrajudicial, se propone una negociación de la deuda, mediante refinanciamientos, reestructura u otros.

El resumen de la distribución de los clientes con la estrategia planteada se observa en el cuadro 4.3.

[14] Acuerdo de pago en firme significa que se llegó a un acuerdo y se pactó la fecha de pago de la cuota adeuda.

Cuadro 4. 3 Distribución de los clientes sobre los grupos de riesgo.

juchuza	Cliente en Mora	N. Clientes	% clientes	Estrategia Cobranza	Tasas de Malos
Riesgo Bajo	NO	22,859	70.1%	Normal	1.6%
	SI	825	2.5%	Preventiva (correo electrónico)	5.7%
Riesgo Medio	NO	1,680	5.2%	Preventiva (llamada telefónica)	2.7%
	SI	414	1.3%	Administrativa / Domiciliaria	16.7%
Riesgo Alto	NO	4,148	12.7%	Administrativa	10.8%
	SI	2,666	8.2%	Extrajudicial / Domiciliaria	25.5%

Fuente: Elaboración propia.

Se observa en el cuadro 4.3 que el 70.1% de PES se encuentran en categoría normal y se evidencia que su tasa de malos es el de menor valor con respecto al resto de grupos. Observando la maniobra de recuperación, la estrategia preventiva tiene participación del 7.7%, la administrativa de 14%, y la gestión fuerte de cobranza, extrajudicial y domiciliaria, se encuentra con una participación del 8.2%. Con ello se logra un enfoque de recuperación del crédito de forma temprana al portafolio PES dentro de la institución financiera.

CAPÍTULO 5

5 Conclusiones y Recomendaciones

El propósito de este trabajo es recomendar estrategias de cobranza de acuerdo con los grupos de riesgo definidos para PES, evitando así problemas de morosidad para las entidades financieras. Esto se hace utilizando las últimas herramientas estadísticas que nos permiten comprender el comportamiento del prestamista. Como resultado, las ganancias del banco aumentan al centrarse en los deudores que necesitan medidas de recuperación rápida, y no ofrecen productos adicionales a los clientes que se consideran con mayor posibilidad de no pago y generan un impacto considerable para la institución.

5.1 Conclusiones

Una vez desarrollado la metodología de alertas tempranas para los clientes PES, y haciendo uso de los modelos de aprendizaje automático, se enumeran los principales hallazgos encontrados a lo largo del presente trabajo:

1. El primer paso es encontrar la probabilidad de incumplimiento, para ello, de las 70 variables explicativas analizadas, se separó un grupo que explica mejor el comportamiento crediticio de los clientes PES en institución financiera, los resultados son 17 obtenidas por la aplicación del algoritmo K-medias a los indicadores de disminución media de Gini y Exactitud de una primera aplicación del algoritmo BA, permitiéndonos elegir el conjunto de variables con mayor significancia para nuestro objetivo. En la figura 3.6 muestra la agrupación de las variables según su importancia, en este caso se escogieron el grupo 4 y 3 que son los más representativos. En el cuadro 3.3 podemos identificar que:
    - El manejo de las tarjetas de crédito para los clientes PES es un factor significativo para su clasificación, en especial la cuota de la tarjeta y el monto que utiliza de su cupo, esto se representa por el ingreso de las variables: x35.cuota_tarjeta, x36.utilización_tarjeta.

- Los días de mora que posee el PES en los 5 meses anteriores a la otorgación de crédito influye en el acierto de buen y mal cliente dentro de la institución, y se observa por el ingreso de las variables: X66.Mora_5, X67.Mora_4, X68.Mora_3, X69.Mora_2, X70.Mora_1.
- La información de buró aporta significativamente con la calificación al cliente al momento de originar el crédito, cuantas obligaciones se encuentran sin inconvenientes, su mora actual y si posee mayores a 120 días de morosidad en el sistema bancario, los saldos en el sector financiero como el saldo en mora, y su pago mensual, dado que las variables que ingresaron son: X7.Calificacion_buro, X5.Dias_mora_iniciales, X15.obligaciones_0_mora, X19.obligaciones_120_mora, X26.saldo_sector_financiero, X29.saldo_mora_financiero, X32.cuota_sector_fianciero, X35.cuota_tarjeta, X36.utilizacion_tarjetas, X39.Mora_bancaria_buro, X45.Mora_120_año.
- Las variables que influyen más en la definición de no pago son: los días mora del cliente en la entidad al momento de otorgar su crédito y el saldo de deuda en el sector financiero.
- La aplicación del algoritmo de k-medias a las métricas de disminución media de Gini y exactitud nos permiten descubrir las variables más importantes de la variable objetivo.

2. Después de obtener las variables relevantes para calcular la PI, se aplica nuevamente el modelo de BA, pero con los hiperparámetros optimizados.
    a. La cantidad de predictores a seleccionar se muestra en la teoría que para un modelo de bosques aleatorios es $\sqrt{p}$, con p números de variables, aplicando el ejemplo para nuestro caso debe ser 4 considerando la aproximación de decimales. En el presente estudio se realiza una búsqueda propia del número de parámetros, en la figura 3.7 se observa que, al momento de la optimización del parámetro, el valor con menor Error OOB es 7 y se selecciona como

optimo. Además, se evidencia que la diferencia de error tomando 7 y 4 está aproximadamente entre 0.01.

b. La cantidad de registros óptimos son 5 en el nodo final, observado en la figura 3.8. El número de árboles óptimos es 352.

3. Los resultados de las métricas de evaluación en las bases de construcción y prueba no difieren significativamente, mencionando el AUC es de 0.920 y 0.916 respectivamente; por lo que se considera estable el modelo y es adecuado para otorgar una probabilidad de incumplimiento.
4. Al implementar el modelo final en la Muestra de Aplicación la cual se presenta en la figura 3.4, se observa una diferencia en los resultados desde el año 2019 percibidos en la figura 3.17. Una posible causa del desfase es la afectación de pandemia del covid-2019, el confinamiento produjo que los clientes PES bajen sus ventas y existe la posibilidad de provocar el no pago de las obligaciones crediticias, el presente trabajo no aborda el tema.
5. Los cálculos de probabilidad de incumplimiento se ajustaron principalmente debido a que en la Muestra de Aplicación del Modelo observado reveló el retraso del punto de corte el cual se observa en la Figura 3.16. Es causado por un desequilibrio entre buenos y malos clientes en la base de datos. Se propone ajustar aplicando la función exponencial como se expone en la Figura 3.18. La fórmula es $y = 0.008757e^{4.140600x}$.
6. Con la obtención de la PI se procede a realizar la agrupación de PES mediante la aplicación del algoritmo K-medias utilizando además las variables MV, PR; en el cuadro 3.8 se observa los 4 grupos que recomendó el método Elbow examinado en la figura 3.21, el grupo con mayor impacto negativo a la empresa es el 2, el cual manifiesta que 6,814 clientes deben tener una atención y gestión de recuperación prioritaria, dado que poseen la más alta probabilidad de incumplimiento con respecto a las demás asociaciones, en la institución financiera dichos deudores mantienen plazos de 2 años aproximadamente y montos promedios de USD 27,2010.

7. Se realiza una subagrupación, pasando de los 4 consolidaciones establecido por el algoritmo k-medias a 3, con el fin de aplicar la maniobra de recuperación, Los grupos realizados permiten dar una propuesta de recobro. Se observa el cuadro 4.3 que el 70.1% de los clientes necesitan ser gestionadas aplicando la estrategia normal; por ejemplo, se agrade por correo electrónico el pago de la obligación. Un 8.2% de los deudores tienen una necesidad de gestión alta, es conveniente realizar la visita de un gestor de cobranza al domicilio o plantear una reestructuración o refinanciamiento al deudor; dado que el cliente no llega a su condición de incumplimiento se identifica tempranamente al prestamista con posibles problemas, los asesores de crédito pueden considerar usar esta alerta cuando un PES está buscando nuevos préstamos y así evitar futuros inconvenientes en la entidad.
8. La utilización del algoritmo de BA y K-medias ayuda formar segmentos para establecer estrategia de recuperación. Esto lleva a mencionar que la metodología utilizada permite clasificar a los prestamistas antes de su estado de incumplimiento en categorías de alto, medio y bajo riesgo y entender la repercusión para la entidad financiera haciendo uso de las variables PI, MV, PR.

5.2 Recomendaciones

A partir de los objetivos alcanzados se pueden plantear algunas recomendaciones que servirán de ayuda para trabajos futuros y que permitirán mejorar el desempeño de la metodología propuesta.

1. En futuros estudios es fundamental profundizar en metodologías nuevas para la obtención de la probabilidad de incumplimiento, se utilizó BA, pero se puede explorar otros algoritmos supervisados y no supervisados para el cálculo de la PI y agrupación de clientes respectivamente, y comparar su eficiencia.
2. Enriquecer las bases de datos con características demográficas de clientes PES, con el objetivo de analizar si éstas permiten si mejorar el rendimiento del modelo, por

ejemplo, la antigüedad, tamaño de la empresa PES, actividad económica, información financiera.

3. Es útil para monitorear los resultados. En la Figura 3.3, observamos un incremento en el porcentaje de malos clientes en los meses de 2019, recordando que la definición de cliente malo se observa un año a futuro del mes de análisis, implica que la cantidad creció en el 2020, la posible causa del aumento es la influencia del covid-2019, para lo cual se recomienda que la entidad financiera analice que variables del modelo fue afectado por pandemia y realizar los debidos ajustes.
4. La utilización de la propuesta permite priorizar la gestión de recuperación, para lo cual se recomienda que se aplique en la institución financiera por parte de los asesores y esto llevaría a que tengan un espacio de tiempo adicional para otras actividades como colocación de nuevos productos en los clientes con riesgo bajo.

Apéndice A

Anexo 1: Código para la creación de los modelos

A.1. Librerías

```
#Librerías
library(dplyr)
library(ggplot2)
library(reshape2)
library(randomForest)
library(ROCR)
library(pROC)
library(corrplot)
library(tidyr)
library(factoextra)
library(NbClust)
```

A.2. Construcción de la Base de Construcción y de Prueba

```
Bueno<- subset(Base_modelo_menos_campos, Base_modelo_menos_campos[,3]
== "Bueno")
Malo<- subset(Base_modelo_menos_campos, Base_modelo_menos_campos[,3] ==
"Malo")

set.seed(54321)
indices <- sample( 1:nrow(Bueno), dim(Malo)[1], replace=FALSE)
Bueno_muestreado<-Bueno[indices,]
dim(Bueno_muestreado)

Base_para_modelo<-rbind(Bueno_muestreado,Malo)
dim(Base_para_modelo)
table(Base_para_modelo$'57.indicador')


index <- createDataPartition(Base_para_modelo$"57.indicador", p = 0.5, list =
FALSE)
Construccion<- Base_para_modelo[index, ]
Prueba<- Base_para_modelo[-index, ]
```

A.3. Reducción de Variables

A.3.1. Modelo Bosques Aleatorio

```
names (Construccion1) <- make. names(names(Construccion1))

modelo_rf <- randomForest(as.factor(Construccion1[,3]) ~ .,
                  data = Construccion1[,-3],
                                        mtry = 7, ntree = 500, nodesize = 10,
                                      importance=TRUE)
modelo_rf
```

A.3.2. Modelo k-medias

```
b <- modelo_rf$importance[,c(3,4)]
b <- data.frame(b)
SegmetnoNorm <- scale(b)

# gráfico Elbow

wssplot <- function(SegmetnoNorm, nc = 15, set.seed = 1234){
  wss <- (nrow(SegmetnoNorm) - 1)*sum(apply(SegmetnoNorm, 2, var))
  for(i in 2:nc) {
    set.seed(1234)
        wss[i] <- sum(kmeans(x = SegmetnoNorm, centers = i, nstart = 25)$withinss)
  }
plot(1:nc, wss, type = 'b', xlab = 'Número de clúster', ylab = 'Distorción', main = 'Método Elbow
para seleccionar el número de clúster óptimos', frame.plot = T,
     col = 'blue', lwd = 1.5)
}

wssplot(SegmetnoNorm)

#Modelo K-means

RF <- kmeans(scale(b), 4)
RF
b$segmento <- as.factor(RF$cluster)

#gráfico del modelo
ggplot(b, aes(x = b[,1], y = b[,2], colour=segmento)) +
  geom_point(size = 2.5) +
  theme_bw()+
  labs(    x = "Mean Decrease Accuracy",
           y = "Mean Decrease Gini",
           title = "Gráfico de dispersión",
           subtitle = "comparativo y segmentación medidas de error
                  Random Forest",
           caption = "Fuente: Institución Financiera
             Elaboración: por el autor",
           fill="Grupos de segmentación")
```

A.4. Optimización de hiperparámetros

A.4.1. Estimación de Número de Variables

```
max_predictores <- ncol(Construccion2)-2

n_predictores   <- rep(NA, max_predictores)
oob_mse         <- rep(NA, max_predictores)

names(Construccion2) <- make.names(names(Construccion2))
dim(Construccion2)
i=1

for (i in 1:max_predictores) {
  set.seed(789234)
  modelo_rf <- randomForest(as.factor(Construccion2[,2]) ~ ., data =
```

```
            Construccion2[,-2],  mtry = i, ntree = 500)
  n_predictores[i] <- i
  oob_mse[i] <- tail(modelo_rf$err.rate[,1], n = 1)
  print(i)
}
results <- data.frame(n_predictores, oob_mse)

ggplot(data = results, aes(x = n_predictores, y = oob_mse)) +
  scale_x_continuous(breaks = results$n_predictores) +
  geom_line() +
  geom_point() +
  geom_point(data = results %>% arrange(oob_mse) %>% head(1),
             color = "red") +
  labs(title = "Evolución del out-of-bag-error vs predictores",
       x = "nº predictores empleados") +
  theme_bw()
```

A.4.2. Estimación de Número de registros en los nodos finales

```
size <- c(1,2,3,4,5,6,7,8,9,10,11,12,14,13,15,20,25,30,35,40,50)

oob_mse <- rep(NA, length(size))
names(Construccion2)
names(Construccion2) <- make.names(names(Construccion2))


for (i in seq_along(size)) {
  set.seed(789654)
  modelo_rf <- randomForest(as.factor(Construccion2[,2]) ~ .,
                            data = Construccion2[,-2],
                            mtry = 7, ntree = 500, nodesize = i)
  oob_mse[i] <- tail(modelo_rf$err.rate[,1], n = 1)
  print(i)
}
results <- data.frame(size, oob_mse)


ggplot(data = results, aes(x = size, y = oob_mse)) +
  scale_x_continuous(breaks = results$size) +
  geom_line() +
  geom_point() +
  geom_point(data = results %>% arrange(oob_mse) %>% head(1),
             color = "red") +
  labs(title = "Evolución del out-of-bag-error vs nodesize",
       x = "nº observaciones en nodos terminales") +
  theme_bw()
```

A.4.3. Estimación de Número de registros en los nodos finales

```
modelo_rf <- randomForest(as.factor(Construccion2[,2]) ~ .,
                          data = Construccion2[,-2],
                          mtry = 7, ntree = 500, nodesize = 5,
                          importance=TRUE)

oob_mse <- data.frame(oob_mse = modelo_rf$err.rate[,1],
                      arboles = seq_along(modelo_rf$err.rate[,1]))
```

```
a <- oob_mse %>%
filter( oob_mse<=  0.1491048)
paste("mínimo error: ", a[,1], "número de árboles con mínimo error" ,
                    a[,2])

ggplot(data = ob_mse, aes(x = arboles, y = oob_mse )) +
  geom_line() +
  labs(title = "Evolución del out-of-bag-error vs número árboles",
       x = "nº árboles") +
  theme_bw()+ geom_hline(yintercept = round(min(oob_mse[,1]),7),
                      color="blue")+
  geom_vline(xintercept =  a[,2], color="blue")
```

A.4.4. Modelo Final

```
modelo_rf <- randomForest(as.factor(Construccion2[,2]) ~ .,
                      data = Construccion2[,-2],
                      mtry = 7, ntree = 352, nodesize = 5,
                      importance=TRUE)
modelo_rf
```

```
#gráfico de la importancia

imp <- modelo_rf$importance
varImpPlot(modelo_rf,n.var = 17)
```

A.5. Validación del Modelo

```
# gráfico de densidad clientes buenos malos

ggplot(Construccion2, aes(x = modelo_rf$votes[,2], fill =
as.factor(Construccion2[,2]))) +
  geom_density(alpha = 0.3) +
  labs(
    x = "Probabilidad de ser Malo",
    y = "Densidad",
    title = "Comparativo de la densidad de los clientes buenos y malos
          ",
    subtitle = "Muestra de construcción",
    caption = "Elaboración: por el autor",
    fill = "Calidad del cliente")


pred <- prediction( modelo_rf$votes[,2], as.factor(Construccion2[,2]) )
perf <- performance(pred, measure = "tpr", x.measure = "fpr" )

plot(perf, colorize = TRUE, type = "l",
     main = "Curva Roc base de Construcción" )
abline(a = 0, b = 1 )

# Calcular el área bajo la curva

AUC <- performance( pred, measure = "auc")
AUCaltura <- AUC@y.values

# Calcular el punto de corte óptimo
```

```
cost.perf <- performance( pred, measure = "cost" )
opt.cut <- pred@cutoffs[[1]][which.min(cost.perf@y.values[[1]])]
#coordenadas del punto de corte óptimo
x<-perf@x.values[[1]][which.min( cost.perf@y.values[[1]] ) ]
y<-perf@y.values[[1]][which.min( cost.perf@y.values[[1]] ) ]
points(x,y, pch=20, col="red")

cat( "AUC:", AUCaltura[[1]]) #out
cat("Punto de corte óptimo:", opt.cut) #out


# gráfico de la curva Roc

objroc <- roc(as.factor(Construccion2[,2]),
         modelo_rf$votes[,2],auc=T,ci=T)
objroc

plot.roc(objroc,print.auc=T,print.thres = "best",
       col="blue",xlab="1-ESpecificidad",ylab="Sensibilidad", main =
       "Curva Roc base de Construcción")

Construccion2$resultado <-
        as.factor(ifelse(modelo_rf$votes[,2]>=0.516,"Malo","Bueno"))

# Resultados de la matriz de confusión

confusionMatrix(Construccion2$resultado , as.factor(Construccion2[,2]
             ))
```

A.6. Aplicación en la muestra de Prueba

```
names(Prueba) <- make.names(names(Prueba))

# predice el modelo

Prueba$predi<-predict(modelo_rf, Prueba, type="prob")

# Gráfica la distribución buenos y malos

ggplot(Prueba, aes(x = predi[,2]  , fill = Prueba[,3])) +
  geom_density(alpha = 0.3) +
  labs(
    x = "Probabilidad de ser Malo",
    y = "Densidad",
    title = "Comparativo de la densidad de los clientes buenos y malos ",
    subtitle = "Muestra de Prueba",
    caption = "Fuente: Institución Financiera
            Elaboración: por el autor",
    fill = "Calidad del cliente")

#punto de corte aplicado en la muestra de prueba
Prueba$resultado <-
          as.factor(ifelse(Prueba$predi[,2]>=0.516,"Malo","Bueno"))

objroc <- roc(as.factor(Prueba[,3]), Prueba$predi[,2],auc=T,ci=T)
objroc

#gráfico de la curva roc
plot.roc(objroc,print.auc=T,print.thres = "best",
```

```
        col="blue",xlab="1-ESpecificidad",ylab="Sensibilidad", main =
                                "Curva Roc base Prueba")


#Matriz de confusión
        confusionMatrix(Prueba$resultado , as.factor(Prueba[,3] ))
```

A.7. Algoritmo K-medias para grupos de estrategia

```
# selección de las variables
        Segmento <- data.frame(base3$Score,
                        base3$"X3.Monto_total_deuda",
                        base3$"X9.Plazo_prestamo")

        names(Segmento)[1]<-"Probabilidad_default_RF"
        names(Segmento)[2]<-"Monto_total_deuda"
        names(Segmento)[3]<-"Plazo_prestamo"
        dim(Segmento)

# Un cliente registró un monto muy superior a la habitual se procede a eliminar
        Segmento <- filter(Segmento, Monto_total_deuda<=23002280 )

        summary(Segmento)
        str(Segmento)
        dim(Segmento)

        Segmento %>%
          gather(attributes, value, 1:3) %>%
          ggplot(aes(x = value)) +
          geom_histogram(fill = 'lightblue2', color = 'black') +
          facet_wrap(~attributes, scales = 'free_x') +
          labs(x="Values", y="Frequency")

# Correlación de las variables

        corrplot(cor(Segmento), type = 'upper', method = 'number', tl.cex =
                0.9)


# normallizar las variables previo al K-means

          SegmentoNorm <- as.data.frame(scale(Segmento))

# gráfico Elbow
           wssplot <- function(SegmetnoNorm, nc = 15, set.seed = 1234){
            wss <- (nrow(SegmetnoNorm) - 1)*sum(apply(SegmetnoNorm, 2, var))
            for(i in 2:nc) {
              set.seed(1234)
              wss[i] <- sum(kmeans(x = SegmetnoNorm, centers = i, nstart =
                   25)$withinss) }
            plot(1:nc, wss, type = 'b', xlab = 'Número de Clusters', ylab =
                 'Distorción',
                 main = 'Método Elbow para encontrar el número óptimo de
                        clúster', frame.plot = T,
                 col = 'blue', lwd = 1.5)
           }

           wssplot(SegmetnoNorm)
```

```
### Aplicación de los grupos
        set.seed(1234)


        final <- kmeans(SegmetnoNorm, centers = 4)
        print(final)

        a <- fviz_cluster(final, data = SegmetnoNorm, )
        a + ylim (-5, 5)+ xlim (-30, 30)

        Base_20191$segfinal <- final$cluster
        names(Base_2019)
        Base_20191$segfinal <- as.factor(Base_20191$segfinal)
```

Bibliografía

Bonini, S., & Caivano, G. (2018). Probability of default modeling: A Machine Learning approach. En *Mathematical and Statistical Methods for Actuarial Sciencie and Finance* (págs. 173-177). Springer, Cham.

Breiman L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L. (1996). Out-of-bag estimation. *Citesser*.

Chávez, M. (2011). *El impacto de las Características Organizacionales e Individuales de los Dueños o Administradores de las Pequeñas y Medianas Empresas en la Toma de Decisiones Financieras que influyen en la Maximización del Valor de la Empresa.* (Tesis Doctoral, Universidad Autónoma De San Luis Potosí).

Cuadros, Á, Gonzalez, C., & Jiménez, , P. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41-51.

de Basilea, C. D. S. B. (2006). Convergencia internacional de medidas y normas de capital. Marco Revisado-Versión integral. *actualizado a junio*.

Delgado, J. (2015). *Propuesta alternativa de medidas para el acceso de las PYMES a créditos bancarios.* (Tesis de Maestría, Flacso Ecuador).

Espinoza, J. (09 de 2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología, 21*(3).

Financiera, J. (2021). Codificación De Resoluciones Monetarias, Financieras, De Valores Y Seguros. Quito.

Financiera, J. D. R. M. (2021). Codificación de Resoluciones Monetarias, Financieras, de Valores y Seguros. En *Libro I: Sistema Monetario y Financiero.* Ecuador.

Igual, L., & Seguí, S. (2017). *Introduction to Data Science.* Springer, Cham.

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoSONE, 13*(8).

Kornfeld, S. (2020). *Predicting Default Probability in Credit Risk using Machine Learning Algorithms.* (Tesis Doctoral, KTH Royal Institute of Technology).

Kubat, M., & Kubat. (2017). *An Introduction to Machine Learning* (Vol. 2). Cham, Switzerland: Springer International Publishing.

L., B. (1996.). Out-of-bag estimation. *Citeseer*.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Sym. Math. Statist. Probability.*, 281-297.

McCarty, J., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of business research, 60*(6), 656-662. doi:https://doi.org/10.1016/j.jbusres.2006.06.015

Medina, R., & Ñique, C. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. *Interfases, (10)*, 165-189.

Morales J., & Morales A. (2014). Crédito y Cobranza. México: Grupo Editorial Patria.

Oberto, G. I. C. (2020). *Cluster no jerárquicos versus cart y biplot.* (Tesis Doctoral, Universidad De Salamanca).

Ortiz G. (2019). Optimización de hiperparámetros de algoritmos machine learning usados para el análisis de calidad del software desarrollado en IBM RPG.

Puertas, R., & Martí, M. (2013). Análisis del credit scoring. *RAE-Revista de Administració de Empresas, 53*(3), 303-315.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning, aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow.* Marcombo.

Vélez, M., & Chamba, N. (2019). Estructura de las pymes en la economía ecuatoriana. *Sur Academia: Revista Académica-Investigativa De La Facultad Jurídica, Social Y Administrativa, 4*(8).