



ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO
FACULTAD DE CIENCIAS
CARRERA ESTADÍSTICA

**MODELO DE SCORING PARA CRÉDITO DE CONSUMO EN LA
COOPERATIVA DE AHORRO Y CRÉDITO “MINGA” LTDA
UTILIZANDO TÉCNICAS DE MACHINE LEARNING**

Trabajo de Titulación

Tipo: Proyecto de Investigación

Presentado para optar al grado académico de:

INGENIERA ESTADÍSTICA

AUTORA:

ANA ELIZABETH CEPEDA GUAMINGA

Riobamba-Ecuador

2022



ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO
FACULTAD DE CIENCIAS
CARRERA ESTADÍSTICA

**MODELO DE SCORING PARA CRÉDITO DE CONSUMO EN LA
COOPERATIVA DE AHORRO Y CRÉDITO “MINGA” LTDA
UTILIZANDO TÉCNICAS DE MACHINE LEARNING**

Trabajo de Titulación

Tipo: Proyecto de Investigación

Presentado para optar al grado académico de:

INGENIERA ESTADÍSTICA

AUTORA: ANA ELIZABETH CEPEDA GUAMINGA

DIRECTORA: Ing. NATALIA ALEXANDRA PEREZ LONDO

Riobamba-Ecuador

2022

©2022, Ana Elizabeth Cepeda Guaminga

Se autoriza la reproducción total o parcial, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica del documento, siempre y cuando se reconozca el Derecho de Autor.

Yo, ANA ELIZABETH CEPEDA GUAMINGA, declaro que el presente trabajo de Titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este trabajo de Titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba 10, de noviembre de 2022

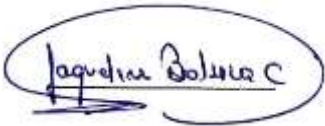

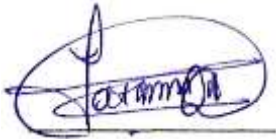


Ana Elizabeth Cepeda Guaminga

060456101-9

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO
FACULTAD DE CIENCIAS
CARRERA ESTADÍSTICA

El Tribunal del Trabajo de Titulación, certifica que: El Trabajo de Titulación; Tipo: Proyecto de Investigación: **MODELO DE SCORING PARA CRÉDITO DE CONSUMO EN LA COOPERATIVA DE AHORRO Y CRÉDITO “MINGA” LTDA UTILIZANDO TÉCNICAS DE MACHINE LEARNING**, realizado por la señorita: **ANA ELIZABETH CEPEDA GUAMINGA**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación. El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	FIRMA	FECHA
Dra. Jaqueline Elizabeth Balseca Castro, Mgs. PRESIDENTE DEL TRIBUNAL		2022-11-10
Ing. Natalia Alexandra Perez Londo, MSc. DIRECTOR DEL TRABAJO DE TITULACIÓN		2022-11-10
Ing. Johanna Enith Aguilar Reyes, Mgs. ASESORA DEL TRABAJO DE TITULACIÓN		2022-11-10

DEDICATORIA

Este trabajo es dedicado primeramente a Dios, porque ha estado siempre presente en cada paso que doy, cuidándome, guiándome y dándome fortaleza para no rendirme y así poder alcanzar una meta más en mi vida; a mis padres Luis Cepeda y Rosa Guaminga quienes a lo largo de mi vida me han apoyado siempre en todas las situaciones y han velado por mi bienestar, y de manera especial a toda mi familia, amigos y cada una de las personas que me han apoyado, sus consejos para lograr esta meta.

Elizabeth

AGRADECIMIENTO

Un agradecimiento especial a la Ing. Natalia Pérez, directora de mi tesis y a la Ing. Johanna Aguilar, quienes me han ayudado con sus conocimientos, tiempo y paciencia lo cual me permitió culminar con éxito mi tesis, así también por compartir sus conocimientos como docentes de algunas materias a lo largo de mi vida universitaria.

A todos los docentes de la carrera de Estadística que semestre a semestre fueron impartiendo valiosos conocimientos que me han servido para crecer como persona y profesional. También a la cooperativa de Ahorro y Crédito “MINGA” Ltda.

Elizabeth

TABLA DE CONTENIDO

ÍNDICE DE TABLAS.....	ix
ÍNDICE DE FIGURAS.....	x
ÍNDICE DE GRÁFICOS.....	xi
ÍNDICE DE ECUACIONES	xii
ÍNDICE DE ANEXOS	xiii
RESUMEN.....	xiv
SUMMARY	xv
INTRODUCCIÓN	1

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL	2
1.1. Antecedentes	2
1.2. Planteamiento del problema.....	3
1.3. Formulación (Incógnita).....	3
1.4. Hipótesis	3
1.4.1. <i>Hipótesis Nula</i>	3
1.4.2. <i>Hipótesis Alternativa</i>	3
1.5. Objetivos	4
1.5.1. <i>Objetivo General</i>	4
1.5.2. <i>Objetivos específicos</i>	4
1.6. Estado del arte.....	4
1.6.1. <i>Análisis exploratorio de datos</i>	4
1.6.1.1. <i>Identificaciones variables</i>	5
1.6.2. <i>Medidas de la estadística descriptiva</i>	5
1.6.2.1. <i>Medidas de tendencia central</i>	5
1.6.3. <i>Beneficios de los modelos de scoring</i>	8
1.6.4. <i>Tipos de Scoring</i>	9
1.6.4.1. <i>Score de organización</i>	9
1.6.4.2. <i>Score de comportamiento</i>	9
1.6.4.3. <i>Score de Bureau</i>	9
1.6.4.4. <i>Marketing Scores</i>	9
1.6.5. <i>Variables para desarrollar el scoring</i>	9

1.6.6.	<i>Machine Learning</i>	10
1.6.7	<i>¿Cómo funciona el Machine Learning?</i>	10
1.6.8.	<i>Algoritmos empleados por el Machine Learning</i>	11
1.6.9.	<i>Arboles de decisiones</i>	12
1.6.10.	<i>Random Forest</i>	12
1.6.10.1.	<i>Algoritmo de formación Random Forests</i>	13
1.6.10.2.	<i>OOB Out of Bag Error</i>	14
1.6.10.3.	<i>Overfitting</i>	14
1.6.11.	<i>Matriz de confusión</i>	14
1.6.11.1.	<i>Propiedades analíticas de la matriz de confusión</i>	14
1.6.12.	<i>Curva de Roc</i>	15
1.7.	Bases teóricas	15
1.7.1.	<i>¿Qué es cooperativas de ahorro y crédito?</i>	15
1.7.2.	<i>¿Qué es el crédito?</i>	16
1.7.3.	<i>Tipos de crédito</i>	16
1.7.3.1.	<i>El crédito de consumo</i>	16
1.7.4.	<i>Sistema financiero del Ecuador</i>	16
1.7.4.1.	<i>Activos financieros</i>	17
1.7.4.2.	<i>Pasivos financieros</i>	17
1.7.4.3.	<i>Patrimonio Financiero</i>	18
1.7.5.	Riesgo	18
1.7.5.1.	<i>Riesgo de Crédito</i>	18

CAPITULO II

2.	MARCO METODOLÓGICO	19
2.1.	Tipo de la Investigación	19
2.2.	Diseño de la investigación no experimental	19
2.2.1.	<i>Localización de estudio</i>	19
2.2.2.	<i>Población de estudio</i>	20
2.2.3.	<i>Método de muestreo</i>	20
2.2.4.	<i>Tamaño de la muestra</i>	20
2.2.5.	<i>Técnica de recolección de datos</i>	20
2.2.6.	<i>Identificación de variables</i>	20
2.2.7.	<i>Modelo estadístico</i>	21
2.3.	VARIABLES EN ESTUDIO	21

2.3.1.	<i>Variable Default</i>	21
2.3.2.	<i>Operacionalización de las variables</i>	22

CAPÍTULO III

3.	MARCO DE RESULTADOS Y DISCUSIÓN DE RESULTADOS	27
3.1.	Análisis Descriptivo	27
3.1.1.	<i>Variables cualitativas</i>	28
3.1.2.	<i>Variables Cuantitativas</i>	31
3.2.	Procesamiento de los datos	35
3.2.1.	<i>Filtrado de datos</i>	35
3.2.2.	<i>Análisis de correlación</i>	36
3.2.3.	<i>Identificación de valores atípicos</i>	37
3.2.4.	<i>Análisis de la variable Default: Impago</i>	38
3.2.4.1.	<i>Balanceo de clases</i>	39
3.2.5.	<i>Generación de muestras para los modelos</i>	40
3.3.	Modelado con árboles de decisión	40
3.3.1.	<i>Validación de los modelos</i>	42
3.3.2.	<i>Selección del mejor modelo</i>	43
3.3.3.	<i>Análisis de otras métricas de evaluación del modelo seleccionado</i>	44
3.3.4.	<i>Análisis de la variable respuesta: Variable impago</i>	45
3.3.5.	<i>Interpretación de los resultados de scoring de crédito</i>	46
	CONCLUSIONES	48
	RECOMENDACIONES	49
	BIBLIOGRAFÍA	
	ANEXOS	

ÍNDICE DE TABLAS

Tabla 1-2:	Operacionalización de variables.....	22
Tabla 2-2:	Codificación de la variable Educación.....	23
Tabla 3-2:	Codificación para la variable Calificación.....	23
Tabla 4-2:	Codificación de la variable Actividad económica.....	24
Tabla 5-2:	Codificación para la variable Oficina.....	26
Tabla 1-3:	Período de análisis.....	27
Tabla 2-3:	Resumen descriptivo de la variable Ingreso.....	31
Tabla 3-3:	Resumen descriptivo de la variable saldo.....	32
Tabla 4-3:	Resumen descriptivo de la variable Monto de operación.....	33
Tabla 5-3:	Resumen descriptivo de la variable Egresos.....	33
Tabla 6-3:	Resumen descriptivo de la variable Valor de la cuota.....	34
Tabla 7-3:	Resumen descriptivo de la variable Mora.....	34
Tabla 8-3:	Matriz de correlación.....	36
Tabla 9-3:	Análisis de la variable Default.....	39
Tabla 10-3:	Calificación de socios de la variable impago balanceado.....	39
Tabla 11-3:	Promedio de Gini de la importancia de las variables.....	41
Tabla 12-3:	Comparación de los resultados del <i>accuracy</i> de los modelos 1, 2 y 3.....	43
Tabla 13-3:	Mejor modelo.....	43
Tabla 14-3:	Resultados de test del modelo 2.....	45
Tabla 15-3:	Resultado de calificaciones de los socios.....	46
Tabla 16-3:	Resultado de calificaciones de los socios.....	46
Tabla 17-3:	Resultado de calificaciones de los socios.....	47
Tabla 18-3:	Resultado de calificaciones de los socios.....	47

ÍNDICE DE FIGURAS

Figura 1-1:	Algoritmo machine Learning	11
Figura 2-1:	Esquema de bosques aleatorios.....	12
Figura 3-1:	Estructura de sistema financiero del Ecuador	17
Figura 1-2:	Localización cooperativa Minga matriz.....	19

ÍNDICE DE GRÁFICOS

Gráfico 1-1:	Distribución simétrica, sesgada a la izquierda, sesgada a derecha	6
Gráfico 2-1:	Distribución simétrica, sesgada a la izquierda, sesgada a derecha	15
Gráfico 1-3:	Número de socios por Oficinas de la Cooperativa	28
Gráfico 2-3:	Número de socios por rango de edad de la Cooperativa	28
Gráfico 3-3:	Número de socios de la Cooperativa por sexo	29
Gráfico 4-3:	Número de socios de la Cooperativa por Nivel de estudio.....	29
Gráfico 5-3:	Número de socios de la Cooperativa por Cargas familiares.....	30
Gráfico 6-3:	Número de socios de la Cooperativa por Calificación de crédito	31
Gráfico 7-3:	Número de socios de la Cooperativa por rango del ingreso	32
Gráfico 8-3:	Variables con datos faltantes	36
Gráfico 9-3:	Gráfico de variables con datos atípicos	38
Gráfico 10-3:	Curva ROC	44

ÍNDICE DE ECUACIONES

Ecuación (1-1):	Media Aritmetica	6
Ecuación (2-1):	Mediana par.....	6
Ecuación (3-1):	Mediana impar	6
Ecuación (4-1):	Moda	7
Ecuación (5-1):	Medidas de posición.....	7
Ecuación (6-1):	Rango	8

ÍNDICE DE ANEXOS

ANEXO A: AVAL DE LA COOPERATIVA DE AHORRO Y CRÉDITO MINGA LTDA.

ANEXO B: CÓDIGO EN R.

RESUMEN

El objetivo de la investigación ha planteado crear un modelo *scoring* para la cooperativa de ahorro y crédito “MINGA” LTDA, que permita la elección eficiente del cliente de crédito de consumo. Para este estudio, se ha obtenido la información directamente de la cooperativa y se ha efectuado el análisis con una base de datos de 15108 registros. Posteriormente, fue establecido un modelo y se definió a los clientes buenos y malos en función a los días de mora. Para ello, se han preparado los datos, por lo que se ha reducido el conjunto de datos a 9065 registros y fueron trabajadas 14 variables. Una vez preparada la información se ha procedido a aplicarla en el modelado de técnicas de *Machine Learning* utilizando el método de árboles de decisión, con el algoritmo *Random Forest* por recomendaciones bibliográficas, entrenándolo con 100, 500 y 1000 árboles, con selección y sin selección de variables, con y sin categorización de las variables cualitativas. Los resultados apuntaron a demostrar la comparación de la precisión de los modelos que se aplicaron en la investigación, evidenciando como ganador al modelo 2 con 13 variables, con selección de características, sin categorización de las variables cualitativas con 500 árboles y una precisión del 97.85%, teniendo una tasa de error del 2.15%. Concluyendo que un buen pagador tiene score mayor o igual a 600, ingreso de 1000, y egreso de 250 dólares, entre 25 y 40 años, recomendando incluir nuevas categorías en la clase.

Palabras clave: < MODELO SCORING >, < CRÉDITO DE CONSUMO >, < MACHINE LEARNING >, < ÁRBOLES DE DECISIÓN >, < ALGORITMO RANDOM FOREST > SOFTWARE ESTADÍSTICO R >




2381-DBRA-UPT-2022


DBRA
Ing. Cristian Castillo

SUMMARY

The objective of the research has been to create a scoring model for the savings and credit cooperative "MINGA" Ltd, which allows the efficient choice of the consumer credit client. For this study, the information has been obtained directly from the cooperative and the analysis has been done with a database of 15108 records. Subsequently, a model was established and good and bad customers were defined based on days past due. For this, the data has been prepared, so the data set has been reduced to 9065 records and 14 variables were considered. Once the information was ready, it has been applied in the modeling of Machine Learning techniques by using the decision tree method, with the Random Forest algorithm based on bibliographic recommendations, training it with 100, 500 and 1000 trees, with and without selection of variables, with and without categorization of qualitative variables. The results aimed to demonstrate the comparison of the precision of the models that were applied in the study, evidencing as the winner model 2 with 13 variables, with selection of characteristics, without categorization of the qualitative variables with 500 trees and an accuracy of 97.85%, having an error rate of 2.15%. Concluding that a good payer has a score greater than or equal to 600, income of 1000, and expenditure of 250 dollars, between 25 and 40 years of age, it is recommended including new categories in the class.

Keywords: <SCORING MODEL>, <CONSUMER CREDIT>, <MACHINE LEARNING>, <DECISION TREES>, <RANDOM FOREST ALGORITHM> <R STATISTICAL SOFTWARE>



Edgar Mesias Jaramillo Moyano
0603497397

INTRODUCCIÓN

Hoy en día se ha determinado que es de gran importancia elaborar un análisis de información para el desarrollo y expansión de las instituciones financieras, por lo tanto, se busca establecer modelos o medidas que permitan alcanzar a un mercado objetivo en el campo económico (Verdezoto, 2016, p. 86). Los modelos de *score* crediticio son técnicas muy manejadas por las instituciones financieras para apoyarse en sus decisiones de concesión de créditos al consumo. Ahora bien, la idea detrás de estos modelos es identificar el efecto de ciertas características del solicitante del préstamo sobre su probabilidad de impago.

Las instituciones financieras enfocadas en el otorgamiento de créditos de consumo se enfrentan a dos tipos de decisiones: por un lado, deben decidir si conceder un crédito a un cliente nuevo y, por otro, deben resolver cómo gestionar los créditos que ya otorgaron. Las herramientas estadísticas usualmente utilizadas para tomar el primer tipo de decisiones son las asociadas a los modelos de *score*, que permita la elección eficiente del cliente de crédito de consumo en función a sus características (Dassatti , 2019, pp. 13-15).

En este trabajo se desarrolló la metodología para la elaboración de un modelo *score*, que permita definir el comportamiento de un buen y mal cliente. La información recopilada fue directamente de la base de datos de la cooperativa de ahorro y crédito “MINGA” LTDA con el objetivo de asignar un puntaje a cada cliente, permitiendo de esta forma elegir eficientemente si se concede o no el crédito.

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL

1.1. Antecedentes

En los últimos años, las entidades financieras presentan dificultades en cuanto a la concesión de créditos; el retraso de los clientes en las fechas de pago y en algunos casos el incumplimiento de la deuda es causantes de una morosidad variable, que no es más que el reflejo de una ineficiencia en la asignación de créditos y la necesidad de ajustar los criterios de evaluación para los mismos (Medina&Ulfe, 2015, p. 23).

Habiendo visitado la Universidad Andina Simón Bolívar he podido encontrar la tesis titulada: “Elaboración y evaluación de un score para crédito de consumo en la Cooperativa de Ahorro y Crédito COOPAD”, cuyo autora es Jessica Verdezoto que presentó y sustentó para obtener el título de Magister en Finanzas y Gestión de Riesgos, en el año 2016; cuyo trabajo de investigación manifiesta que la implementación de un modelo *score* nos permite otorgar o rechazar la solicitud de crédito de los clientes de cooperativa. Además, genera una mayor objetividad en la evaluación de riesgos sin embargo esta información permite minimizar las operaciones que no están dentro del valor agregado de la cooperativa mediante el proceso de la evaluación de riesgos. Por lo cual fue indispensable la construcción del modelo *score* es importante e indispensable para la verificación de la base de datos con la finalidad de mejor capacidad de la institución (Verdezoto, 2016, p. 87).

Asimismo, puedo indicar que se encuentra registrada la tesis “Modelo de evaluación de crédito – scoring- para la cartera de consumo de la Cooperativa de Ahorro y Crédito Riobamba” cuyo autor es Jairo Rivera, quien investigó para titularse como Magister en Finanzas y Gestión de Riesgos en el año 2011, cuyo trabajo también revela que un modelo de *scoring* depende exclusivamente de los datos con los que cuenta una cooperativa. En este sentido, las variables que se incluyen en el modelo son propias para la institución y muy probablemente no serán las mismas al ser aplicadas en otra institución financiera, por esta razón se debe tener en cuenta las variables que permitan categorizar a los clientes. Ante ello, es necesario utilizar la estadística para escoger técnicamente los niveles de división de clientes, y hay que tomarse el tiempo necesario para probar la efectividad del modelo a fin de aceptar o no aceptar a un cliente dependiendo de su probabilidad de incumplimiento (Rivera, 2011, pp. 78-79).

1.2. Planteamiento del problema

La cooperativa de Ahorro y Crédito “MINGA” LTDA es una entidad financiera con 22 años de trabajo en beneficio del país (Riobamba, Quito y Guayaquil). Su función es atraer clientes y socios estratégicos, también ofrecen servicios como: créditos de consumo y microcréditos. En el proceso de otorgamiento de créditos la cooperativa toma las decisiones a partir de su juicio experto para cada una de las solicitudes recibidas. Sin embargo, este método no puede satisfacer las demandas en los aspectos económicos y de eficiencia. Por tal razón se debe adoptar sistemas de calificación de créditos para facilitar y acelerar los procesos en la toma de decisiones.

En el presente trabajo de investigación se realizó un análisis de la base de datos de los clientes con crédito de consumo de la cooperativa de ahorro y crédito “MINGA” LTDA y se elaboró un modelo *scoring* que permitió mejorar el riesgo existente dentro de las operaciones de los créditos otorgados. Para ello se utilizaron técnicas de *machine learning*, que permitieron conocer el comportamiento de los clientes. Asimismo, la aplicabilidad de la investigación fue garantizada ya que la cooperativa no disponía de un *scoring* para su cartera de consumo.

1.3. Formulación (Incógnita)

¿El modelo *scoring* en la cartera de crédito de consumo de la cooperativa de ahorro y crédito “MINGA” LTDA permitirá la elección eficiente del cliente de crédito?

1.4. Hipótesis

1.4.1. Hipótesis Nula

El modelo *scoring* en la cartera de crédito de consumo de la cooperativa de ahorro y crédito “MINGA” LTDA permitirá la elección eficiente del cliente de crédito en función a sus características.

1.4.2. Hipótesis Alternativa

El modelo *scoring* en la cartera de crédito de consumo de la cooperativa de ahorro y crédito “MINGA” LTDA no permitirá la elección eficiente del cliente de crédito en función a sus características.

1.5. Objetivos

1.5.1. Objetivo General

- Elaborar un modelo *Scoring* para crédito de consumo en la cooperativa de ahorro y crédito “MINGA” LTDA utilizando técnicas de *Machine Learning*, permitiendo de forma adecuada y ágil el análisis de riesgo crediticio y el otorgamiento de créditos.

1.5.2. Objetivos específicos

- Analizar el estado actual del crédito de consumo en la cooperativa de ahorro y crédito “MINGA” Ltda.
- Realizar análisis descriptivo de cada una de las variables en estudio.
- Elaborar el modelo de *Scoring* para el crédito de consumo en la cooperativa de ahorro y crédito “MINGA” LTDA utilizando técnicas de *Machine Learning*.

1.6. Estado del arte

1.6.1. Análisis exploratorio de datos

La estadística descriptiva es la ciencia que permiten asociar en forma y característica un conjunto de datos con el fin de describir pertinentemente las diversas características de ese conjunto. Las medidas de tendencia central, describe el conjunto de datos. Entre los principales enfoques estadísticos son: media aritmética, mediana, y moda (Bautista, 2021, p. 34).

A nivel financiero es necesario la aplicación de cálculos estadísticos sobre la base de datos de crédito que tienen un enfoque, de acuerdo a los resultados arrojados. También dependerá las políticas de crédito que se aplique a las inspecciones crediticias, por lo tanto, es recomendable trabajar con la totalidad de la población de datos observados para llegar a obtener un resultado lo más cercano posible a la realidad (Bautista, 2021, p. 38).

1.6.1.1. Identificaciones variables

Variable independiente

La variable independiente es el objeto de estudio en la investigación, que permite al individuo describir, explicar y transformar, es decir son las que crean y revelan los cambios de la variable dependiente. Ejemplo: el método de enseñanza de lectura que un profesor utiliza para mejorar la comprensión lectora de sus alumnos (Freire, 2018).

Variable dependiente

La variable dependiente se transforma por la acción de la variable independiente. Asimismo, constituyen los efectos o consecuencias que dan origen a los resultados de la investigación. Ejemplo: los cambios o mejora que los alumnos experimentan en su comprensión lectora tras un periodo de entrenamiento (Freire, 2018).

1.6.2. Medidas de la estadística descriptiva

“La estadística está definida como la ciencia de recolectar, organizar, presentar, analizar e interpretar datos para ayudar a tomar decisiones más seguras, por lo tanto, su importancia dentro de la gestión actual de riesgos es incuestionable” (Verdezoto, 2016, p. 91).

Los datos recabados en una investigación dentro del ámbito financiero, de forma estado natural no manifiestan información útil para la toma de decisiones del crédito de consumo, por lo que se requiere someter a un proceso matemático que defina diversas medidas que puedan resumir la información para obtener resultados que expliquen su comportamiento, es por ello que dentro de las medidas a considerar existen cuatro clasificaciones: (Bautista, 2021, p. 45)

1.6.2.1. Medidas de tendencia central

Manifiestan cómo están distribuidos los datos alrededor de un valor central de la empresa, dentro de este grupo se encuentra:

Media Aritmética

La media aritmética es más conocida como el promedio. Se calcula sumando todas las observaciones de un conjunto de datos, dividiendo después ese total entre el número total de

elementos involucrados. Su símbolo es \bar{X} si la media aritmética es de una muestra y μ si la media aritmética es de una población.

Fórmula:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{n} = \frac{\sum_{i=1}^n X}{n} \quad (1-1)$$

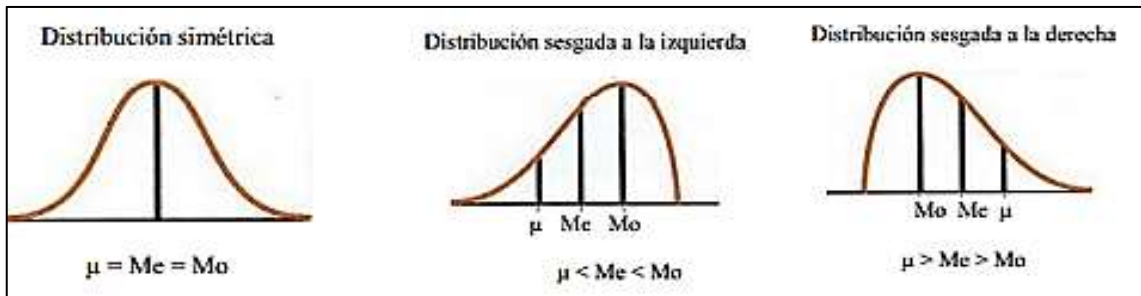


Gráfico 1-1: Distribución simétrica, sesgada a la izquierda, sesgada a derecha

Fuente: (Ribera, 2011).

Mediana

En un conjunto de datos ordenados de manera creciente, es el valor para el cual, la mitad de estos es menor que este valor y la otra mitad mayor.

Fórmula para un total de datos de número impar:

$$M_e = \frac{n + 1}{2} = \text{orden del valor de} \quad (2-1)$$

Fórmula para un total de datos de número par:

$$M_e = \frac{\frac{X_n}{2} + \frac{X_{n+1}}{2} + 1}{2} = \text{orden de datos a sumar para dividir a 2} \quad (3-1)$$

Moda

Es la observación que se presenta con mayor frecuencia en la muestra; además muestra hacia qué valor tienden los datos a agruparse

Fórmula para datos agrupados:

$$M_0 = L_1 + \left(\frac{D_1}{D_1 + D_2} \right) \times a \quad (4-1)$$

Dónde:

(i) = Intervalo de mayor frecuencia absoluta.

L_1 = Es el límite inferior real del intervalo que contiene a la moda.

D_1 = Diferencia entre frecuencia absoluta del intervalo de la moda y el intervalo anterior.

D_2 = Diferencia entre la frecuencia absoluta del intervalo moda y el intervalo posterior.

a = Es la amplitud del intervalo.

Medidas de posición.

La medida de posición permite identificar dentro del rango se encuentran ubicados los datos del de la investigación, de acuerdo a la división que se quiera dar al total de la población. Cuando se tiene datos agrupados en intervalos, estas medidas se consideran en cierta forma como una extensión de la mediana, entre los cuantiles que se usan con mayor frecuencia tenemos:

Formula:

$$Md_u = \frac{u(n + 1)}{Td} = N.r \quad (5-1)$$

Donde:

Td = Es el total de divisiones a ejecutar con la población de datos.

u = Ubicación dentro del número total de divisiones dadas.

n = Número total de la población de datos.

N = Número (z)entero a contar del recorrido en la población ordenada de los datos.

r = Decimal existente como distancia desde (N) y el valor que sigue.

Medidas de dispersión

A más de conocer la posición de los datos con relación a una división uniforme de la población, es necesario determinar qué tan cercanos se encuentran dichos datos entre sí. Asimismo, conocer, cual es la diferencia con relación a medida de posición.

Rango

Es la distancia existente entre el valor máximo, hacia el valor mínimo que tomen las variables estadísticas.

$$\text{Rango } (R) = V_{\max} - V_{\min} \quad (6-1)$$

P-Value

El valor p se define como la probabilidad de observar el valor dado del estadístico de prueba, o mayor, bajo la hipótesis nula. Tradicionalmente, el valor de corte para rechazar la hipótesis nula es 0,05, lo que significa que cuando no existe diferencia significativa, se espera un valor tan extremo para el estadístico de prueba en menos del 5 % de las veces (Ferreira, J; Patino, C, 2015, p. 41).

Modelos scoring

El modelo *credit scoring* es una herramienta para evaluar una operación en procesos de riesgo, siendo usada como una técnica para aceptar o rechazar las operaciones de créditos. La metodología cuantifica la calidad riesgo de una operación de un cliente mediante la asignación de un puntaje en función a sus características observables como lo son los datos socioeconómicos, operativos, de negocio, de comportamiento financiero o recursos externos (Medina&Ulfe, 2015, p. 18).

En la actualidad es importante la implementación modelo scoring que, permite asegurar una alta calidad en la eficiencia a nivel financiero, basado en el estudio de los principales atributos que intervienen en su proceso de construcción de los mismos, así como en una política de análisis y seguimiento periódico de las calificaciones resultantes (Exprián, 2021, pp. 1-5).

1.6.3. Beneficios de los modelos de scoring

De acuerdo al análisis del modelo *scoring* a nivel del sistema financiero nos otorga grandes beneficios y ventajas a las instituciones entre las principales tenemos:

- Es importante identificar a los buenos y malos pagadores.
- Se debe asignar la probabilidad de incumplimiento del pago a los clientes para otorgar o no un crédito al usuario.
- Permite que las instituciones establezcan un presupuesto provisional de acuerdo a su riesgo crediticio anhelado.
- Se debe fijar la tasa de interés de acuerdo al riesgo de crédito esperado y la meta establecida en función al ingreso financiero.
- Es confiable establecer la tasa de interés diferencial en función del riesgo de crédito de cada clientes.

1.6.4. Tipos de Scoring

1.6.4.1. Score de organización

El modelo de score de organización calcula una probabilidad estimada de incumplimiento de pago de un posible cliente otorgando a la empresa, permite decidir si acepta o no como posible consumidor de crédito a un cliente. Además, permite a la organización establecer el puntaje mínimo óptimo de aceptación (Medina&Ulfe, 2015, pp. 26-27).

1.6.4.2. Score de comportamiento

A diferencia del score de Organización éste predice la probabilidad de incumplimiento de aquellos que ya se les otorgó crédito en la institución. Por medio de las variables de comportamiento de las cuentas dentro de la propia institución es posible dar seguimiento al comportamiento de los clientes, permitiendo al departamento de cobranzas emplear técnicas para que un cliente siga siendo rentable para la empresa (Medina&Ulfe, 2015, pp. 26-27).

1.6.4.3. Score de Bureau

Está basado en el historial crediticio de las personas para predecir en el futuro el otorgamiento de un crédito. Asimismo, permite disponer de una primera calificación de riesgo del cliente sin solicitar información o referencias.

1.6.4.4. Marketing Scores

Este tipo de modelo *scoring* se centra principalmente en el análisis de la retención de clientes y en la adquisición de nuevos en la cartera de consumo de crédito, por ello, se están desarrollando *scoring* de retención responsable esto permite potenciar a los clientes (Juares, 2017, pp. 1-112).

1.6.5. Variables para desarrollar el scoring

Dentro del estudio existe una sin número de variables que entran en el análisis para el desarrollo de un modelo *scoring*. En la actualidad los datos del sociodemográfica podrían incluir variables cualitativas: estado civil, educación, vivienda, otros, y cuantitativas como el ingreso, edad, capacidad de pago, otros. Los resultados ser eficientes en la predicción de la probabilidad de incumplimiento. Un análisis descriptivo de estas variables siempre será necesario para poder

identificar sesgos e irregularidades, con el fin de encontrar posibles cortes poblacionales, siempre teniendo en cuenta una visión correcta del negocio (Medina&Ulfe, 2015, pp. 32-33).

Una condición importante que deben mostrar dichas variables es que no presenten una alta correlación significativa de esta forma se tendrán una mejor segmentación en la población. Primero, se deben establecer la población en grupos de segmentación y así encontrar una eficiente segmentación. También se puede utilizar técnicas de *machine learning*, árboles de decisión o algoritmos, para posteriormente utilizar algún tipo de modelo (Medina&Ulfe, 2015, pp. 32-33).

1.6.6. Machine Learning

El *machine learning* o aprendizaje automático es una técnica de análisis de datos que enseña a los ordenadores, aquello que es natural para todas las personas. Además, está involucrada en el campo de la inteligencia artificial que mediante algoritmos usa varios métodos computacionales capaces de manejar datos masivos y establecer predicciones (INTELLIGN, 2020, pp. 1-5).

1.6.7. ¿Cómo funciona el Machine Learning?

Machine learning recopila información de manera automática y mediante un análisis estadístico desarrolla programas que generalizan conductas. Es decir, las mejoras creadas día a día son realizadas automáticamente, sin la intervención del hombre. Asimismo, permite establecer predicciones futuras a través de algoritmos que identifican patrones de gran complejidad (SmartPanel, 2019).

El pronóstico permite que las instituciones mejoren sus fortalezas, oportunidades, debilidades y amenazas en el desarrollo de sus operaciones. De esta manera, se pueden optimizar automáticamente los procesos para producir ganancias y reducir costos. Este hecho es identificar patrones de comportamiento, los mismos que resultan ser muy complejos por esta razón se utilizan ordenadores de alta calidad con la finalidad de entregar resultados a un departamento específico (SmartPanel, 2019).

1.6.8. Algoritmos empleados por el Machine Learning

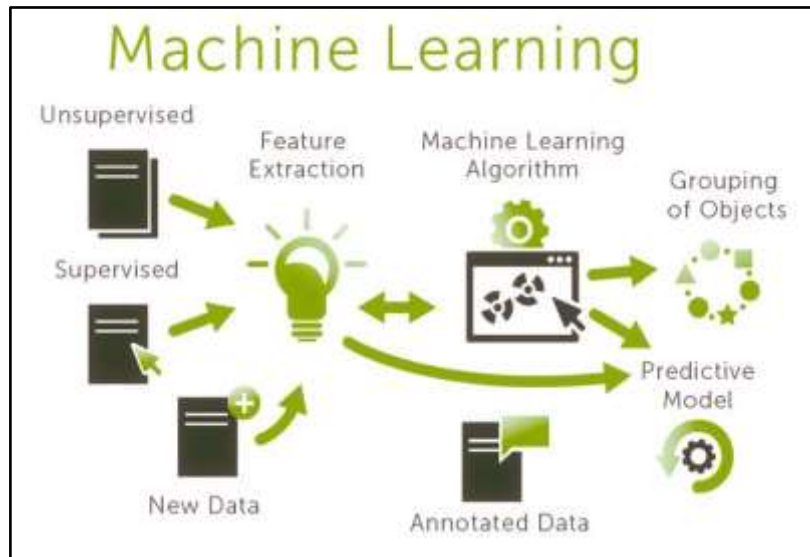


Figura 1-1: Algoritmo machine Learning

Fuente: (SmartPanel, 2019)

Los algoritmos empleados por el machine learning se clasifican en tres grandes conjuntos de aprendizaje:

- **Supervisado**

Se utiliza en aquellas circunstancias en las que una etiqueta se asocia a ciertos datos y requiere de la predicción para poder aplicarla en instancias diferentes.

- **No supervisado**

El aprendizaje no supervisado encuentra patrones ocultos o estructuras intrínsecas en los datos. Se utiliza para extraer inferencias de conjuntos de datos que consisten en datos de entrada sin respuestas etiquetadas.

Favorece el establecimiento de relaciones implícitas si se dispone de información no clasificada.

- **De refuerzo**

Es una solución intermedia entre las dos anteriores. Existe un modo de retroalimentar cada etapa o actividad de predicción, pero no se conoce la etiqueta particular o hay un mensaje que refleja un problema.

1.6.9. Árboles de decisiones

Los árboles de decisiones están constituidos por modelos predictivos formados por iteraciones que detallan el esquema de división y que obtiene decisiones jerárquicas de mayor relevancia en el análisis de *scoring*. Esta metodología o técnica está basada en dividir los datos utilizando una de las variables independientes para separar los datos en subgrupos homogéneos. La forma final del árbol puede traducirse a un conjunto de reglas If-Then-Else de la raíz a cada uno de los nodos de hoja. Por lo tanto, están estrechamente relacionados con los métodos de aprendizaje de reglas y sufren de la misma desventaja que ellos. Los árboles de decisión más conocidos son CART, C4.5 y PUBLIC (Ilbay, 2016, p. 36).

1.6.10. Random Forest

Random Forests es una popular herramienta de aprendizaje automático de conjunto basada en árboles que es altamente adaptable a los datos, se aplica a problemas, y es capaz de dar cuenta de la correlación, así como las interacciones entre las características (Chen, Ishwaran 2012).

Los bosques aleatorios son un conjunto de muestras de arranque del proceso de inducción del *tren*. La estrategia de los bosques aleatorios es seleccionar aleatoriamente subconjuntos para hacer crecer árboles, cada árbol crece en una muestra de arranque del conjunto de entrenamiento. Una vez construido el bosque, se utilizará para efectuar la predicción (García, 2018).

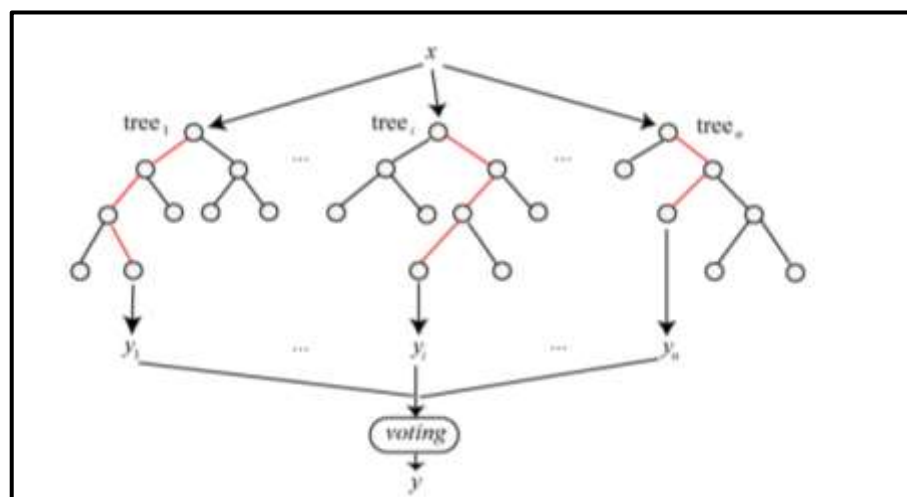


Figura 2-1: Esquema de bosques aleatorios

Fuente: (Chen, 2017).

1.6.10.1. Algoritmo de formación Random Forests

Cada árbol de decisión se constituirá de la siguiente forma:

- Asumiendo un conjunto N observaciones diferentes, se elegirá una muestra N aleatoria con reemplazamiento. Esta técnica, toma el nombre de **bootstrapping** y es utilizada en varios algoritmos de *machine learning*. También, introduce aleatoriedad al algoritmo, ya que cada árbol se forma de manera diferente (García, 2018).
- Dadas las M variables de entrada, en cada nodo se seleccionarán de forma aleatoria $p < M$ variables. Este número p , será constante en todo el proceso de formación del árbol e introducirá el segundo elemento de aleatoriedad en el algoritmo.
- Se dejará crecer el árbol, sin podar hasta la máxima extensión posible.

Por lo tanto, se puede establecer que los *random forest* dependen de dos parámetros fundamentales:

Ntree: Número de árboles que forman el bosque.

Mtree: Número de variables p que se seleccionan en cada nodo.

Entonces, en la práctica el valor de *Mtree* dependerá del problema. Al disminuir la correlación entre árboles, disminuirá la varianza, y por lo tanto, más preciso será el árbol.

Los valores recomendados son:

\sqrt{p} para un problema de clasificación.

$\frac{p}{3}$ para un problema de regresión.

El número de árboles, *Ntree*, también tiene efecto en la precisión de la predicción. Como es lógico, a mayor número de árboles mejor será la predicción, puesto que el número de datos para hacer el promedio es mayor. Sin embargo, existe un valor para el cual, el error ya no disminuye y se estanca, aumentando solo el tiempo del algoritmo.

1.6.10.2. *OOB Out of Bag Error*

El *Out of bag error* (OOB) es una medida de error aplicada a modelos que utilizan la técnica del *bootstrapping*. El OOB error, representa el error de predicción cometido por el bosque cuando se tienen en cuenta este conjunto de variables que han quedado “fuera de la bolsa”.

1.6.10.3. *Overfitting*

El *overfitting*, o sobreajuste como es conocido en castellano, es un término muy usado en estadística y en *machine learning*. Este problema ocurre cuando un algoritmo es capaz de realizar una buena predicción con los datos de partida, pero pierde mucha precisión con datos diferentes a la muestra inicial.

1.6.11. *Matriz de confusión*

En el campo de la inteligencia artificial y el aprendizaje automático una matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado, donde cada columna de la matriz representa el número de predicciones, mientras que cada fila representa a las instancias reales, es decir nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de pasar por el proceso de aprendizaje con los datos (Barrios, 2019, p. 19).

La matriz de confusión es una medida importante para evaluar la precisión de los modelos de calificación crediticia. Sin embargo, la literatura sobre matriz de confusión es limitada. Se ignoran las propiedades analíticas de Confusión Matrix. Además, el concepto de Matriz de confusión es confuso (Zeng, 2019, p. 27).

1.6.11.1. *Propiedades analíticas de la matriz de confusión*

- **La exactitud (o Exactitud general):** es la proporción de Verdadero Positivo y Verdadero Negativo con respecto al número total de cuentas.
- **Tasa de error (ERR):** es la proporción de falso positivo y falso negativo con respecto al número total de cuentas.
- **Sensibilidad:** es la precisión de los positivos o la tasa de verdaderos positivos.
- **Especificidad:** es la precisión de los negativos o la tasa de verdaderos negativos.

1.6.12. Curva de Roc

La curva ROC poblacional representa 1-especificidad frente a la sensibilidad para cada posible valor umbral o punto de corte en la escala de resultados de la prueba en estudio (Valle, 2017).

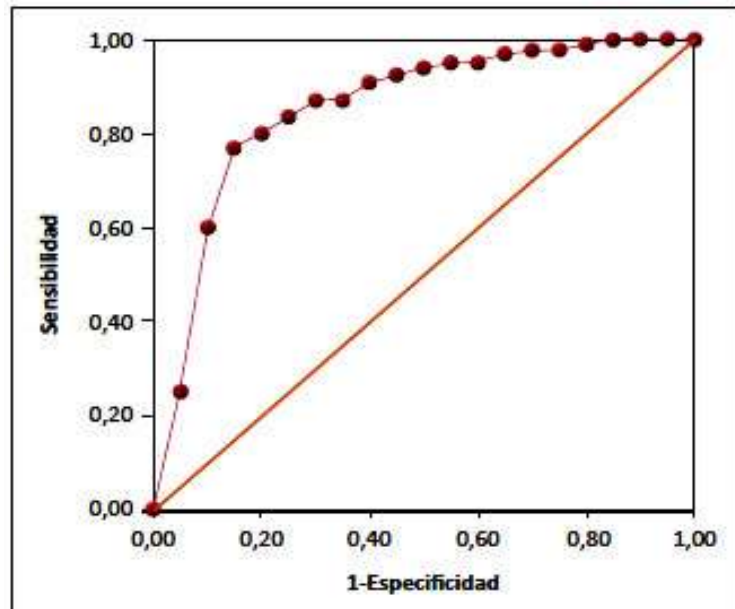


Gráfico 2-1: Distribución simétrica, sesgada a la izquierda, sesgada a derecha

Fuente: (Chen, 2017).

Mediante la curva ROC se representación los pares (1-especificidad, sensibilidad) obtenidos al considerar todos los posibles valores de corte de la prueba, la misma nos proporciona una representación global de la exactitud diagnóstica. La curva ROC es necesariamente creciente, propiedad que refleja el compromiso existente entre sensibilidad y especificidad: si se modifica el valor de corte para obtener mayor sensibilidad, sólo puede hacerse a expensas de disminuir al mismo tiempo la especificidad. La exactitud de la prueba aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo. Si la discriminación fuera perfecta (100% de sensibilidad y 100% de especificidad) pasaría por dicho punto (Valle, 2017, pp. 32-35).

1.7. Bases teóricas

1.7.1. ¿Qué es cooperativas de ahorro y crédito?

Las cooperativas de ahorro y crédito son organizaciones que prestan su servicio a socios sin fines de lucro, cabe recalcar que una cooperativa también ofrecen una variedad de servicios y beneficios a la sociedad dependiendo del servicio requerido como puede depósitos, prestamos, servicios

financieros donde cada cooperativa brinda una tasa de interés bajo que está regulado por el sistema financiero del Ecuador (National Credit Union Administration, 2021, pp. 1-2).

1.7.2. ¿Qué es el crédito?

El crédito significa confianza en una persona o institución que tenga la capacidad de cumplir la obligación contraída gracias a su propio compromiso. En general, la apertura de un crédito se basa en facilitar una cierta cantidad de dinero al solicitante donde cada uno de ellos está consientes de aceptar las condiciones que ofrece el prestamista. Por lo tanto, toda cooperativa o banco está en la obligación de facilitar y dar a conocer el valor de pago de intereses. Prestaciones, gastos y comisiones a quien postule por este servicio. Cabe recalcar a nivel del ámbito financiero el crédito es similar con diferencia que cada préstamo otorgado al solicitante por la banca son prestaciones activas (Acosta, 2010, p. 5).

1.7.3. Tipos de crédito

Para el análisis del estudio de investigación se va a analizar de forma general los tipos de crédito que ofrecen las cooperativas a nivel nacional, existe varios tipos de créditos, pero los más relevantes son los siguientes: (Acosta, 2010, pp. 5-8).

1. Crédito de consumo
2. Crédito empresarial o cooperativo
3. Crédito hipotecario para vivienda

1.7.3.1. El crédito de consumo

Es un crédito que esta abalizado por las entidades financieras y cooperativas, pero se otorga para tiempos cortos que específicamente se detalla para el consumo de bienes y servicios (Acosta, 2010, p. 5).

1.7.4. Sistema financiero del Ecuador

El sistema financiero ecuatoriano está conformado por un conjunto de instituciones como son entidades financieras y gubernamentales, medios donde interviene los activos financieros vigentes y mercados que hacen posible que el ahorro de los agentes económicos donde por cuestiones financieras vaya a recalar en demandas de crédito por instituciones de régimen general (Kisiryán.M., 2015, p. 21).

Dentro del sistema financiero de nuestro país está compuesto por bancos e instituciones de crédito donde están depositados los ahorros. Para que el sistema funcione es necesario que exista confianza en estas entidades, por lo que el Banco de España las regula y supervisa para garantizar que los individuos puedan recuperar su dinero cuando lo deseen. Se podría decir que el sistema financiero es el conjunto de mecanismos a través de los que se ponen en contacto ahorradores e inversores, y que permiten compatibilizar las preferencias y las necesidades de unos y otros en cuanto a importe, plazo, rentabilidad y riesgo (EDUFINEXT, 2019, p. 36).

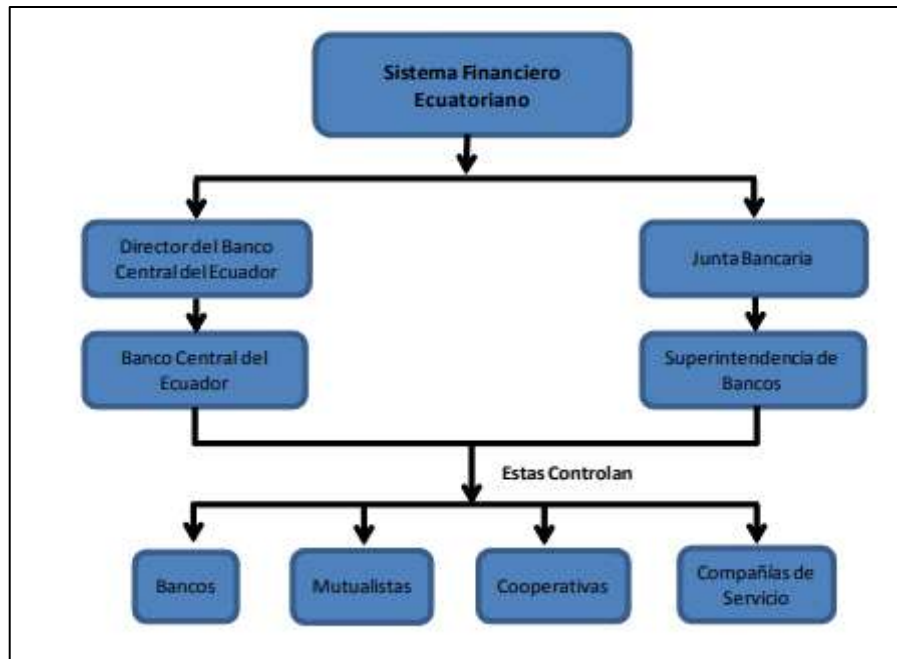


Figura 3-1: Estructura de sistema financiero del Ecuador

Fuente: (EDUFINEXT, 2019)

1.7.4.1. Activos financieros

Un activo es un recurso controlado por la entidad como resultado de sucesos pasados, del que la entidad espera obtener beneficios económicos en el futuro. Los beneficios económicos futuros de un activo consisten en el potencial para contribuir directa o indirectamente, a los flujos de efectivo, otros equivalentes al efectivo de la entidad (Rueda, 2016, pp. 36-37).

1.7.4.2. Pasivos financieros

Los pasivos son las obligaciones presentes contraídas por la entidad para el desarrollo del giro ordinario de su actividad (Resolución No. SBS- 2002-0297/ 29 de abril del 2002). Las obligaciones se originan de la captación de recursos del público en sus diferentes modalidades,

préstamos recibidos de instituciones financieras y otras entidades públicas o privadas y los recursos recibidos mediante la emisión de títulos valores (Rueda, 2016, p. 37).

1.7.4.3. Patrimonio Financiero

En concepto, el Patrimonio se refiere a la participación de los propietarios de los activos de la empresa, y es determinado por la diferencia entre activos y pasivos. Agrupa las cuentas que registran los aportes de los accionistas, socio o gobierno nacional, la prima o descuento en colocación de acciones, las reservas, otros aportes patrimoniales, superávit por valuaciones y resultados acumulados o del ejercicio. (Resolución No. SBS-2002- 0297/ 29 de abril de 2002) (Rueda, 2016, p. 38).

1.7.5. Riesgo

El riesgo a nivel del sistema financiero está definido como la contingencia, probabilidad o proximidad de daño producido a nivel económico a una institución, cabe recalcar que este riesgo a nivel financiero representa una pérdida en valor agregado a administración a la hora de realizar sus inversiones hacia el cliente, dependiendo las causas de perdida basado en los factores que interviene se puede clasificar los riesgos en diferentes tipos que continuación se detalla (Medina&Ulfe, 2015, pp. 45-46)..

- Riesgo de Mercado
- Riesgo de Crédito
- Riesgo de Liquidez
- Riesgo Operativo
- Riesgo Legal

1.7.5.1. Riesgo de Crédito

Se define como la variabilidad en los ingresos generados por el incumplimiento de un acreditado o contraparte que incluye las pérdidas por el importe adeudado y no pagado a las cooperativas por los acreditados (Medina&Ulfe, 2015, pp. 47-48).

CAPITULO II

2. MARCO METODOLÓGICO

2.1. Tipo de la Investigación

El método de investigación utilizado en este trabajo fue mixto, dado que cuenta con variables categóricas y numéricas. Variables categóricas como educación, calificación, actividad y oficina y variables numéricas como ingresos, egresos, mora entre otros. Este método se centró en un análisis profundo de la base de datos y en este caso se trabajó con la base de la cooperativa de ahorro y Crédito “MINGA” LTDA. Asimismo, el estudio fue descriptivo y exploratorio porque se caracterizó al cliente como bueno o malo para obtener predicciones y a su vez identificar posibles variables a estudiar en un futuro (Cazau, 2018, p. 15).

2.2. Diseño de la investigación no experimental

Se utilizó un método de investigación mixto y según la manipulación de variables fue un diseño no experimental debido a que la base de datos fue proporcionada por la Cooperativa de Ahorro y Crédito “MINGA” LTDA (Cazau, 2018, p. 25).

2.2.1. Localización de estudio

La cooperativa de Ahorro y Crédito “MINGA” LTDA de la provincia de Chimborazo empezó a brindar sus servicios en la ciudad de Riobamba, la matriz se encuentra en esta ciudad. Adicionalmente, presta sus servicios a través de seis agencias dentro de la ciudad y fuera de la provincia de Chimborazo (Quito y Guayaquil).



Figura 1-2: Localización cooperativa Minga matriz.

Fuente: Google maps (maps, 2021).

2.2.2. Población de estudio

Se realizó el análisis con información de 12 meses desde noviembre 2020 a octubre 2021, con una base de datos de 15108 registros del crédito de consumo de la matriz (Riobamba) y sucursales: Chimborazo, Guayaquil y Quito.

2.2.3. Método de muestreo

La información que se utilizó para desarrollar el trabajo es la base de datos proporcionada por la Cooperativa de Ahorro y Crédito “MINGA” LTDA, razón por la cual no se aplica un método de muestreo en el estudio.

2.2.4. Tamaño de la muestra

Se realizó el estudio sobre la base de datos de 15108 registros que corresponde a la base del segmento de crédito de consumo, mismo que cuenta con datos de 12 meses desde noviembre 2020 a octubre 2021.

2.2.5. Técnica de recolección de datos.

La información recolectada fue la base de datos que proporcionó el departamento de sistemas y negocios de la Cooperativa de Ahorro y Crédito “MINGA LTDA”, base que cuenta con los datos necesarios para el análisis del score de crédito de los socios.

2.2.6. Identificación de variables

- Fecha de corte
- Código del cliente
- Número de operación
- Cargas familiares
- Edad
- Educación
- Número de cuota
- Monto de operación
- Egresos
- Valor de la cuota

- Ingreso
- Ahorro
- Calificación
- Ciudad
- Sexo
- Actividad
- Saldo
- Mora
- Oficina

2.2.7. Modelo estadístico

Como parte de este modelo, se realizó un análisis descriptivo de las variables en estudio. También se utilizó modelos estadísticos de nueva generación como las técnicas de *Machine Learning* a través de los algoritmos de árboles de decisión para desarrollar el modelo de scoring de crédito.

2.3. Variables en estudio

2.3.1. Variable Default

Para la elaboración del modelo de *scoring*, se procedió a identificar a clientes malos y separarlos de los buenos, hubo la necesidad de establecer una definición de qué es bueno y qué es malo.

En esta definición, entran en juego dos factores: los días de atraso en un pago (días mora) para considerar a un cliente como moroso, y el desempeño, que es la ventana temporal en la que se observó la peor situación de dicho cliente. Los días de atraso determinaron si un cliente es moroso o no es moroso.

Finalmente, la variable default tendrá el valor de 1 si el cliente ha caído en mora (moroso) y 0 si no está en mora (no es moroso).

2.3.2. Operacionalización de las variables

Tabla 1-2: Operacionalización de variables

Nombre de la variable	Descripción	Tipo de variable	Escala de medición
Fecha de corte	Fecha en la que se genera la información de la cartera	Cualitativa	Nominal
Código del cliente	Número de identificación del cliente en la Cooperativa	Cualitativa	Nominal
Número de operación	Número asignado a cada operación que realiza el socio	Cualitativa	Nominal
Cargas familiares	Número de dependientes a cargo del socio	Cuantitativa	Discreta
Edad	Edad del socio	Cuantitativa	Intervalo
Educación	Nivel de educación del socio	Cualitativa	Nominal
Número de cuota	Número de cuotas según la tabla de amortización	Cuantitativa	Ordinal
Monto de operación	Valor de crédito otorgado	Cuantitativa	Intervalo
Egresos	Gastos mensuales estimados del socio	Cuantitativa	Intervalo
Valor de la cuota	Valor que debe pagar el cliente según el tiempo acordado con la Cooperativa	Cuantitativa	Intervalo
Ingreso	Valor mensual estimado de ingreso del socio	Cuantitativa	Intervalo
Ahorro	Variación entre el Ingreso y gasto mensual reportado por el socio	Cuantitativa	Intervalo
calificación	Riesgo de crédito según la Superintendencia de economía popular y solidaria. La calificación puede ser: A1, A2,B1,B2,C1,C2,D,E	Cualitativa	Ordinal
Ciudad	Lugar de residencia del socio	Cualitativa	Nominal
Sexo	Sexo	Cualitativa	Nominal
Actividad	Actividad económica del socio	Cualitativa	Nominal

Saldo	Valor adeudado a la fecha de corte	Cuantitativa	Intervalo
Mora	Días de retraso en el pago de una cuota de crédito	Cuantitativa	Intervalo
Oficina	Lugar otorgado el crédito al socio	Cualitativa	Nominal

Fuente: Base de Datos Coop "Minga" Ltda.

Realizado por: Cepeda, E., 2022

Teniendo en cuenta las variables categóricas: Educación, Calificación, Actividad y Oficina, estas también se codifican para realizar un mejor análisis.

Tabla 2-2: Codificación de la variable Educación

Educación	Codificación
Sin estudios	1
Primaria	2
Secundaria	3
Superior	4
Técnica	5
Masterado	6
Indefinida	7

Realizado por: Cepeda, E., 2022

Tabla 3-2: Codificación para la variable Calificación

Calificación	Codificación
A1	1
A2	2
A3	3
B1	4
B2	5
C1	6
C2	7
D	8
E	9

Realizado por: Cepeda, E., 2022

Tabla 4-2: Codificación de la variable Actividad económica

Actividad económica	Codificación
Educación primaria	1
Otro tipo de empresas comerciales	2
Educación superior	3
Educación secundaria	4
Ropa, prendas de vestir, boutique	5
Acabados para construcción	6
Salón de recepciones	7
Material eléctrico	8
Cooperativa	9
Servicios de seguridad	10
Avícola (crianza y proc.de aves)	11
Otros servicios personales	12
Oficinas	13
Educación preescolar	14
Otros servicios hoteleros	15
Secretario de finanzas	16
Comercio de legumbres y frutas.	17
Restaurante y comida rápida	18
Chofer en cooperativa de transporte	19
Otros servicios de educación	20
Sector publico	21
¡No definida...!	22
Servicios municipales	23
Artículos de hogar	24
Frigorífico y carnicería	25
Venta de accesorios de celular	26
Agricultor	27
Taller de reparac. mec. vehic. y automot	28
Electrodomésticos	29
Equipos de computación (hardw.sofw.)	30
Confecciones (ropa)	31
Otros productos alimenticios	32
Fuerzas armadas	33
Vehículos (compra venta de autom, c, b.)	34

Caña de azúcar	35
Papelería	36
Clínica y hospital	37
Banco	38
Servicio de mant. eléctrico y sanitario	39
Panadería y pastelería	40
Lubricadora y lavadora de vehículos	41
Vehículo (repuest.autopartes y acsr.)	42
Albañil	43
Lubricantes	44
Consultorio medico	45
Transporte pasajeros interprovincial	46
Artículos musicales	47
Productos agricolas	48
Material eq. y acabados construcc.	49
Venta de ropa	50
Electrónico	51
Urbanización	52
Víveres y abarrotés al por menor	53
Venta de legumbres	54
Transporte pasajeros taxi urbano	55
Salón de belleza y peluquería	56
Pollos y huevos	57
Farmacia y distrib. farmacéutico	58
Transporte terrestre de carga	59
Muebles	60
Asesoría jurídica	61
Insumos agrícolas y fertilizantes	62
Ganado porcino	63
Ferretería	64
Estación de gasolinera	65
Asesoría para el desarrollo	66
Espectáculo público en vivo	67
Supermercado	68
Depósito de bebidas (gas.crv.jugos)	69
Floristería	70

Vivienda (casas y departamentos)	71
Ganado leche	72
Imprentas	73
Hotel	74
Conserje privada.	75
Tienda naturista	76
Transporte urbano	77
Otras frutas	78
Calzado	79
Bazar y perfumería	80
Productos plasticos	81
Taller de reparación maq.eq. eléctricos	82
Aluminio y vidrio	83
Transporte terrestre para turismo	84
Decoración interior exterior	85
Agencias de viaje	86
Artículos de oficina y papelería	87
Juguetes	88

Realizado por: Cepeda, E., 2022

Tabla 5-2: Codificación para la variable Oficina

Oficina	Codificación
Matriz	1
Cajabamba	2
Dolorosa	3
Guayaquil Centro	4
Quito Centro	5
Guayaquil Norte	6
Uio Yaruqui	7

Elaborado por: Cepeda, E., 2022

CAPÍTULO III

3. MARCO DE RESULTADOS Y DISCUSIÓN DE RESULTADOS

3.1. Análisis Descriptivo

El presente trabajo se realizó con la base de 15108 registros, 1000 socios de la Cooperativa de Ahorro y Crédito “MINGA” LTDA del segmento de crédito de consumo y el período de análisis de noviembre 2020 a octubre 2021.

Como parte inicial de este análisis se consideró analizar la permanencia de los socios durante la ventana de tiempo mencionada, de los cuales 959 socios permanecen con créditos vigentes, y 41 socios finalizaron el pago de las cuotas durante este período. En la Tabla 1-3: Período de análisis, se puede visualizar la permanencia de 959 socios durante el período de análisis, este número fue la base para los análisis.

Tabla 1-3: Período de análisis

Fecha de corte	No. clientes
2020-11-30	959
2020-12-31	959
2021-01-31	959
2021-02-28	959
2021-03-31	959
2021-04-30	959
2021-05-31	959
2021-06-30	959
2021-07-31	959
2021-08-31	959
2021-09-30	959
2021-10-31	959

Realizado por: Cepeda, E., 2022

Posteriormente, se realizó un análisis descriptivo tanto para variables cuantitativas como cualitativas, para una mejor comprensión de las variables en estudio.

3.1.1. Variables cualitativas

En el **Gráfico 1-3: Número de socios por Oficinas de la Cooperativa**, se puede observar que aproximadamente el 35% de los socios de la Cooperativa de ahorro y crédito “MINGA” LTDA se encontraron registrados en la matriz; el 20% aproximadamente de los socios fue de la oficina de Cajabamba; el 13%, 12%, 8%, 8%, 6% aproximadamente se distribuyeron en las oficinas de Quito Centro, Quito Yaruqui, Dolorosa, Guayaquil Norte, y Guayaquil Centro respectivamente.

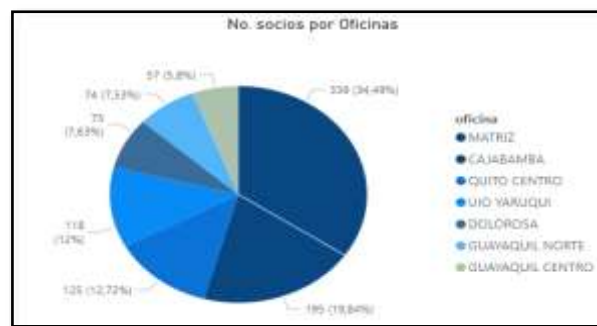


Gráfico 1-3: Número de socios por Oficinas de la Cooperativa

Realizado por: Cepeda, E., 2022

- **Variable edad**

Para el análisis del **Gráfico 2-3: Número de socios por rango de edad de la Cooperativa**, fue importante definir los rangos de edad: 1. Jóvenes: socios con edad mayor a 17 años y menores a 30 años; 2. Adultos: socios con edad mayor a 29 años y menores a 45 años; 3. Madurez: socios con edad mayor a 44 años y menores a 60 años; 4. Tercera edad: socios con edad mayor a 59 años.

Como resultado, la cooperativa tuvo 430 socios en los adultos; 334 socios en los jóvenes; 160 socios en la madurez y 35 socios de la tercera edad. Es decir, la mayor concentración de los socios se encuentra en el grupo de los adultos y en menor cantidad en el grupo de la tercera edad.

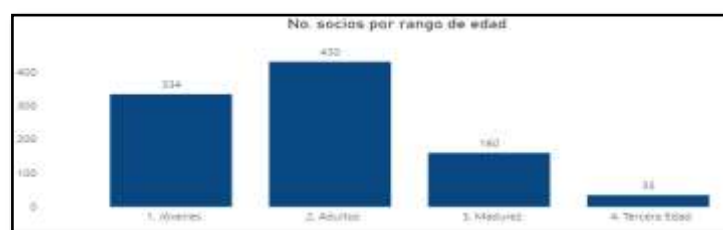


Gráfico 2-3: Número de socios por rango de edad de la Cooperativa

Realizado por: Cepeda, E., 2022

- **Variable sexo**

El **Gráfico 3-3:** Número de socios de la Cooperativa por sexo, el 66% aproximadamente de los socios fue del sexo masculino, mientras que el 34% de los socios fue del sexo femenino.



Gráfico 3-3: Número de socios de la Cooperativa por sexo

Realizado por: Cepeda, E., 2022

- **Variable nivel de educación**

En el **Gráfico 4-3:** Número de socios de la Cooperativa por Nivel de estudio, se observó que 435 clientes tuvieron el nivel de educación Secundaria; 333 socios tuvieron un nivel de educación Primaria; 138 socios tuvieron un nivel de educación de superior, 33 socios presentaron un nivel de educación indefinida, 19 socios sin nivel de estudio y 1 socios con el nivel de estudio intermedio.

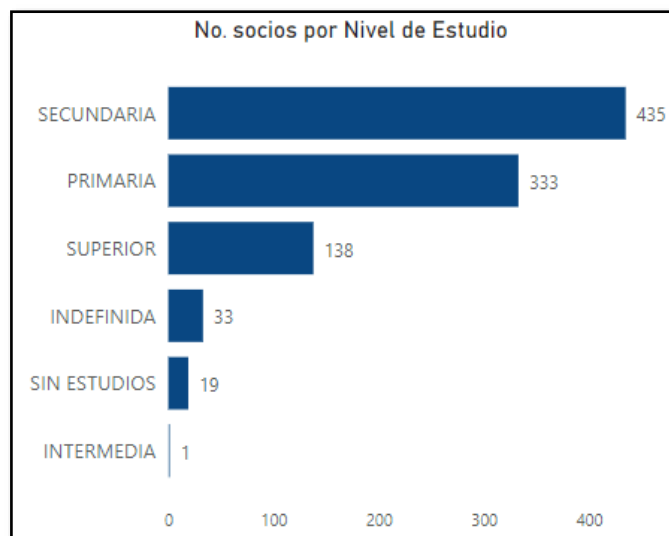


Gráfico 4-3: Número de socios de la Cooperativa por Nivel de estudio

- **Variable cargas familiares**

El **Gráfico 5-3:** Número de socios de la Cooperativa por Cargas familiares, representa el número de socios de la cooperativa según el número de cargas familiares. El 44% aproximadamente de los socios tuvieron 0 cargas familiares; el 28% aproximadamente de los socios tuvo 1 carga familiar; el 19% de los socios tuvieron 2 cargas familiares; y en menor proporción los socios se encontraban con 3, 4,5 y 6 cargas familiares.

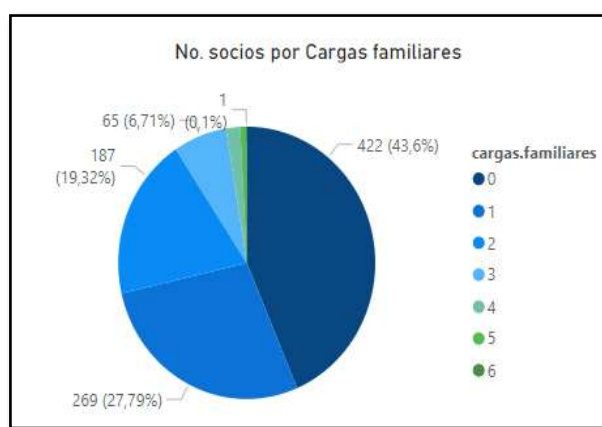


Gráfico 5-3: Número de socios de la Cooperativa por Cargas familiares

Realizado por: Cepeda, E., 2022

- **Variable calificación**

En el **Gráfico 6-3:** Número de socios de la Cooperativa por Calificación de crédito, se analizó la calificación del grupo de socios de la cooperativa, las calificaciones de crédito corresponden: A1, A2, A3, B1, B2, C1, C2, D y E, donde A, es la mejor calificación que un cliente adquiere por cancelar las cuotas según el tiempo establecido en la tabla de amortización de los créditos. Mientras que la letra E corresponde a la peor calificación, calificación dada por la cooperativa cuando el cliente no paga a tiempo las cuotas de los créditos, en esta calificación los socios tienen el número más alto en días mora.

En la mayoría de las instituciones, cooperativas, etc., se considera como cartera incobrable a la calificación E.

Según el gráfico se pudo apreciar que 959 socios de la cooperativa tienen un comportamiento muy bueno. 810, 97 y 30 socios tuvieron las calificaciones A1, A2 y A3, mientras que 29 socios se encontraban con la peor calificación. Es importante comentar que el número de socios no es

igual a la suma de los socios por calificación, puesto que, un socio puede tener más de un crédito y cada crédito tiene su propia calificación.

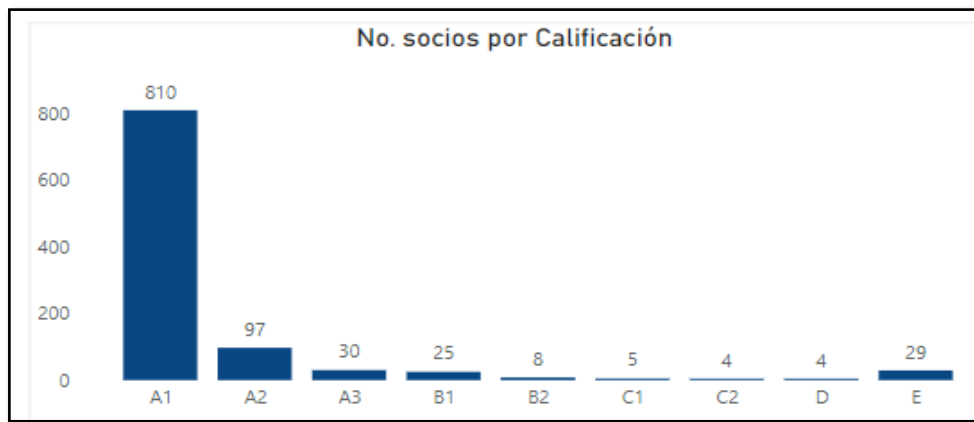


Gráfico 6-3: Número de socios de la Cooperativa por Calificación de crédito

Realizado por: Cepeda, E., 2022

3.1.2. Variables Cuantitativas

- **Variable ingreso**

Tabla 2-3: Resumen descriptivo de la variable Ingreso

INGRESO (\$)	
Media (\bar{x})	1084,13
Mediana(\tilde{x})	950
Desviación estándar(s)	502,95
Varianza de la muestra(s^2)	252956.2
Mínimo	500
Máximo	10000
Cuenta	959

Realizado por: Cepeda, E., 2022

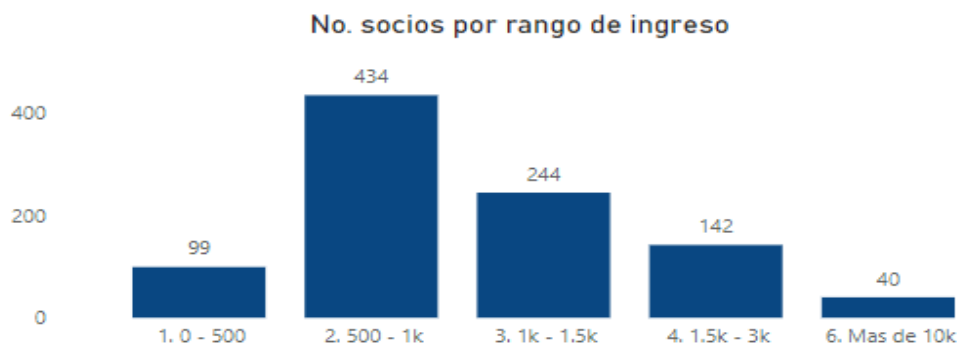


Gráfico 7-3: Número de socios de la Cooperativa por rango del ingreso

Elaborado por: Cepeda, E., 2022

En el **Gráfico 7-3:** Número de socios de la Cooperativa por rango del ingreso, se visualiza que la mayor concentración de los ingresos de los socios (434) de la cooperativa se encontró en el rango del ingreso de 500 dólares a 1000 dólares; 244 socios tuvieron un ingreso entre 1000 dólares y 1500 dólares; 142 socios percibieron un ingreso entre 1500 dólares y 3000 dólares y 40 socios contaron con un ingreso mayor a 10000 dólares

- **Variable saldo**

Tabla 3-3: Resumen descriptivo de la variable saldo

SALDO (\$)	
Media (\bar{x})	4525.66
Mediana(\tilde{x})	3397.755
Desviación estándar(s)	3777.161
Varianza de la muestra(s^2)	14266945
Mínimo	100
Máximo	15000
Cuenta	959

Realizado por: Cepeda, E., 2022

En el grupo de socios analizados se pudo apreciar que el saldo mínimo de los créditos de consumo fue de 100 dólares y el saldo máximo fue de 15000 dólares, la mediana del grupo cuenta con un saldo de 3397.75 dólares. Asimismo, la media del grupo de socios fue de 4525.66 dólares.

- **Variable monto de operación**

Tabla 4-3: Resumen descriptivo de la variable Monto de operación

MONTO DE OPERACIÓN (\$)	
Media (\bar{x})	6591.81
Mediana(\tilde{x})	5000
Desviación estándar(s)	4957.83
Varianza de la muestra(s^2)	24580143
Mínimo	350
Máximo	20000
Cuenta	959

Realizado por: Cepeda, E., 2022

En la **Tabla 4-3:** Resumen descriptivo de la variable Monto de operación, se puede visualizar que el monto mínimo otorgado en los préstamos de consumo fue de 350 dólares, el monto máximo fue de 20000 dólares, el promedio fue de 6591.81 dólares, la mediana del monto otorgado fue de 5000 dólares, con una desviación estándar de 4957.83 dólares y una varianza de 24580143 dólares.

- **Variable egreso**

Tabla 5-3: Resumen descriptivo de la variable Egresos

EGRESOS (\$)	
Media (\bar{x})	322.25
Mediana(\tilde{x})	250
Desviación estándar(s)	225.73
Varianza de la muestra(s^2)	50955.5
Mínimo	100
Máximo	1000
Cuenta	959

Realizado por: Cepeda, E., 2022

En la **Tabla 5-3:** Resumen descriptivo de la variable Egresos, se puede visualizar que el egreso mínimo de los clientes de la cooperativa fue de 100 dólares, el máximo valor en egresos fue de 1000 dólares, con una media de 322.25 dólares, la mediana de 250 dólares, con una desviación estándar de 225.73 dólares, la varianza de 50955.5 dólares.

- **Variable valor de la cuota**

Tabla 6-3: Resumen descriptivo de la variable Valor de la cuota

VALOR DE LA CUOTA (\$)	
Media (\bar{x})	279.31
Mediana(\tilde{x})	264.82
Desviación estándar(s)	137.80
Varianza de la muestra(s^2)	18987.78
Mínimo	54.40
Máximo	700
Cuenta	959

Elaborado por: Cepeda, E., 2022

En la **Tabla 6-3:** Resumen descriptivo de la variable Valor de la cuota, se puede visualizar que el valor de la cuota mínimo fue de 54.40 dólares, el máximo valor de las cuotas fue de 700 dólares, con una media de 279.31 dólares, la mediana de 264.82 dólares, con una desviación estándar de 137.80 dólares, y la varianza de 18987.78 dólares.

- **Variable mora**

Tabla 7-3: Resumen descriptivo de la variable Mora

MORA (\$)	
Media (\bar{x})	95
Mediana(\tilde{x})	29
Desviación estándar(s)	191
Varianza de la muestra(s^2)	36393
Mínimo	1
Máximo	861
Cuenta	959

Elaborado por: Cepeda, E., 2022

En la **Tabla 7-3:** Resumen descriptivo de la variable Mora, se puede visualizar que el mínimo de días mora fue de 1 día, el máximo de días vencidos fue de 861 días, con una media de 95 días, la media de 29 días, con una desviación estándar de 191 días, la varianza de 36393 días.

3.2. Procesamiento de los datos

El procesamiento de datos está formado por una serie de técnicas que tienen el objetivo de inicializar correctamente los datos que servirán de entrada para los algoritmos de machine learning (García , et al., 2015, p. 285).

El presente trabajo, basó el estudio en la base de 15108 registros, 20 variables: educación, calificación, actividad, oficina, Código del cliente, número de operación, cargas familiares, edad, numero de cuota, monto de la operación, egresos, valor de la cuota, ingreso, ahorro, ciudad, sexo, saldo, mora, producto, fecha de corte y 959 socios de crédito de consumo de la Cooperativa,

Posteriormente, se tomó la decisión de eliminar las variables número de operación y número de la cuota, dado que para el estudio se consideró como dato único el código del cliente. También, se eliminó la variable producto, puesto que el dato es constante (consumo).

3.2.1. Filtrado de datos

El filtrado, consistió en revisar la cantidad de datos faltantes tanto en variables como en instancias (registros) completas con el objetivo de filtrar aquellas que tuviesen en exceso de datos perdidos o faltantes, mismo que dificultaría los análisis posteriores.

Luego de realizar este proceso sobre las variables se obtuvieron los siguientes resultados:

- 15 variables: educación, calificación, actividad, oficina, Código del cliente, número de operación, cargas familiares, edad, número de cuota, monto de la operación, egresos, ingreso, ciudad, sexo, saldo, mora, producto, fecha de corte presentan el 0% de datos faltantes, es decir todas estas variables tiene datos.
- 1 variable: valor de la cuota presenta un 3% de datos faltantes
- 1 variable: variable ahorro un 9% de datos faltantes.

En el **Gráfico 8-3:** Variables con datos faltantes, se puede observar que las variables de color naranja presentan una pérdida de datos faltantes (marcado en el rectángulo rojo), mientras que las variables marcadas de color verde presentan 0% de datos faltantes.

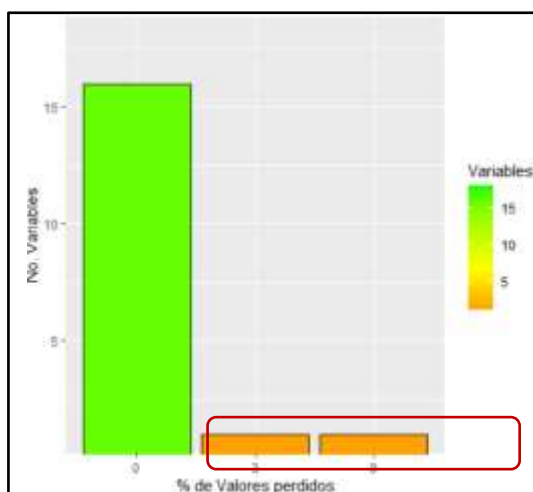


Gráfico 8-3: Variables con datos faltantes

Realizado por: Cepeda, E., 2022

Posteriormente, se tomó la decisión de eliminar las variables con datos faltantes mayor al 5%, la variable ahorro cumple con esta condición y ha sido eliminada para los siguientes procesos. Luego de este análisis la base de estudio está formada de 15108 registros y 16 variables.

3.2.2. *Análisis de correlación*

La matriz de correlación permite analizar qué tan correlacionadas están las variables, el valor oscila entre -1 y 1, el número negativo indica una correlación inversamente proporcional y el valor positivo indica una correlación directamente proporcional.

Tabla 8-3: Matriz de correlación

	Ingreso	Saldo	Monto operación	Egresos	Valor cuota	Mora
Ingreso	1.00	0.19	0.24	0.43	0.43	-0.04
Saldo	0.19	1.00	0.91	0.00	0.33	-0.14
Monto de operación	0.24	0.91	1.00	0.08	0.36	-0.14
Egresos	0.43	0.00	0.08	1.00	0.03	0.07
Valor cuota	0.43	0.33	0.36	0.03	1.00	-0.08
Mora	-0.04	-0.14	-0.14	0.07	-0.08	1.00

Realizado por: Cepeda, E., 2022

Como se observa en la **Tabla 8-3:** Matriz de correlación, la variable monto de operación y saldo son variables que están altamente correlacionadas con un valor de 0.91 y corresponde a una

correlación directamente proporcional, el resto de las variables tienen una correlación menor al 0.43 directamente proporcional y menor al -0.14 inversamente proporcional.

Para efectos del estudio se consideró eliminar a la variable monto de operación por estar altamente correlacionada con el saldo de los créditos de consumo de la Cooperativa.

Luego de este análisis la base de estudio está formada de 15108 registros y 15 variables.

3.2.3. Identificación de valores atípicos

El procedimiento de identificación de valores atípicos está diseñado para ayudar a determinar si una muestra de n observaciones numéricas contiene o no valores atípicos. Por “valor atípico” (outlier), es decir una observación que no proviene de la misma distribución que el resto de la muestra.

En el estudio se incluyó el método gráfico utilizando la función `uni.plot` del paquete `mvoutlier` del software estadístico R. La función `uni.plot` traza cada variable de x en paralelo en un gráfico de dispersión unidimensional y , además, marca valores atípicos multivariados.

El **gráfico 9-3**, muestra cada valor de los datos junto con la línea horizontal de 0 y más o menos (-5, +5) puntos, marcados de color verde. Y los puntos más allá de 5 puntos de color rojo, se considera que son valores atípicos potenciales y dignos de análisis o investigación.

En este análisis se consideró las variables cuantitativas como: cargas familiares, edad, monto de la operación, egresos, valor de la cuota, ingreso, saldo, y días mora.

Posteriormente, se tomó la decisión de excluir los datos atípicos, puntos marcados de color rojo que corresponden a los registros que distorsionarán los resultados de los análisis.

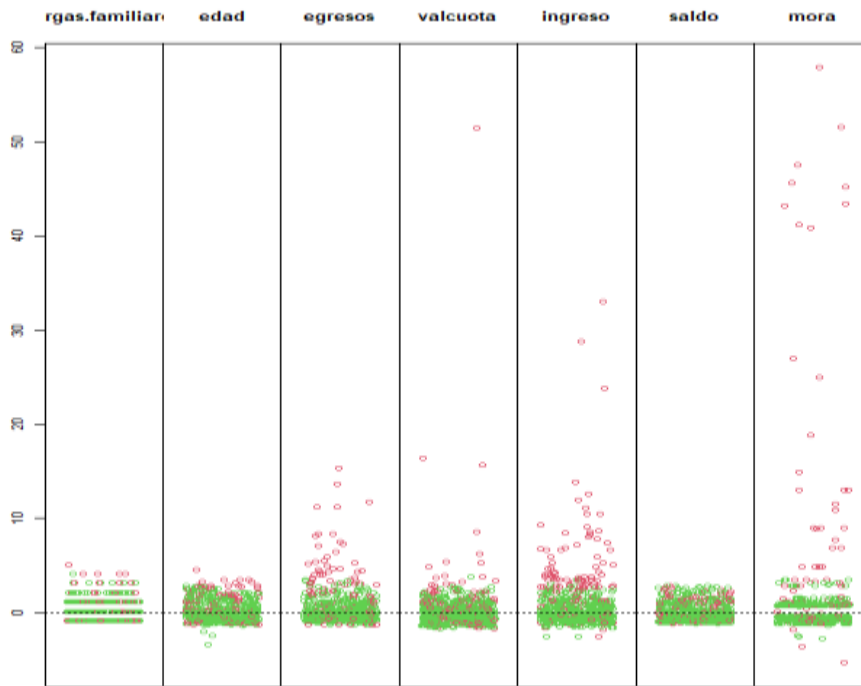


Gráfico 9-3: Gráfico de variables con datos atípicos

Realizado por: Cepeda, E., 2022

Finalmente, se excluyó el 40%, 6043 registros atípicos (puntos de color rojo) y nos quedamos con una base 9065 registros y 466 socios de crédito de consumo de la Cooperativa.

3.2.4. *Análisis de la variable Default: Impago*

En los capítulos anteriores se han mencionado que la variable default será la variable impago. Esta variable tiene dos opciones: *i)* Si el cliente paga las cuotas puntualmente según la fecha de corte de la tabla de amortización o paga las cuotas 20 días posteriores a la fecha de corte según la tabla de amortización para el estudio se considera socio bueno y el valor asignado es de cero (0); *ii)* si el cliente paga las cuotas 21 días posteriores a la fecha de corte de la tabla de amortización se considera socio malo y se asignó el valor de uno (1).

En la **Tabla 9-3:** Análisis de la variable Default, se puede visualizar que la cooperativa tuvo 466 (100%) total socios del segmento de consumo; 307 (66%) socios pagaron las cuotas de crédito dentro de los 20 días posteriores a la fecha de pago según la tabla de amortización y 159 (34%) socios pagaron las cuotas después de 21 días posteriores a la fecha de pago de la cuota según la tabla de amortización. Asimismo, en la tabla se puede ver que los socios buenos son más que los socios malos, en el estado del arte esta gran diferencia se conoce como desbalanceo de clases, con la finalidad de mitigar los riesgos de sobreajuste en los

resultados de los modelos recomiendan balanceo de clases que consiste en contar con el similar número de socios buenos y socios malos.

Tabla 9-3: Análisis de la variable Default

Socios	No	%
Bueno	307	66%
Malo	159	34%
Total	466	100%

Realizado por: Cepeda, E., 2022

En conclusión, la variable *Default* (probabilidad de pago) presentó un problema de desbalanceo de clases. El desbalanceo de clases surge cuando una o más clases se encuentran menos representadas en su número de muestras, en comparación con el número de muestras de otras clases (Pascual, et al., 2020, p. 176).

Posteriormente, se aplicaron técnicas de balanceo de clases y se describe de la siguiente manera:

3.2.4.1. Balanceo de clases

La calificación de los socios malos fue menor que el número de socios buenos. A partir de ahora a los socios malos se los llamará clase minoritaria y a los socios buenos clases mayoritaria. Este problema es considerado como sobre muestreo (*OverSampling*) que consiste el sobre muestreo aleatorio que duplica ejemplos de la clase minoritaria en el conjunto de datos (Ma & He, 2013).

Oversampling, se aplicó a través de la función *ovun.sample* del paquete *ROSE* del *software R*, con la finalidad de equilibrar el número de socios buenos y malos, en la Tabla 10-3: Calificación de socios de la variable impago balanceado se pudo visualizar que el 51% de los socios han sido calificados como socios que pagan las cuotas a tiempo (socios buenos), mientras que el 49% de los socios han sido considerados como socios que no pagan a tiempo las cuotas (socios malos). Por tanto, se cumple la condición que la clase minoritaria (socios malos) y la clase mayoritaria (socios buenos) la proporción es similar.

Tabla 10-3: Calificación de socios de la variable impago balanceado

Socios	No	%
Bueno	246	51%
Malo	234	49%

Total	480	100%
--------------	------------	-------------

Realizado por: Cepeda, E., 2022

Para posteriores análisis la base balanceada presentó 6225 registros, con 480 socios del segmento de consumo, 246 socios buenos y 234 socios malos.

3.2.5. Generación de muestras para los modelos

En los proyectos de machine learning según el estado del arte recomiendan separar el conjunto de datos inicial en 2: conjunto de entrenamiento (train) y conjunto de Pruebas (test). Por lo general se divide haciendo “80-20”. Y se toman muestras aleatorias (Hastie, et al., 2009).

La base datos inicial contempla 6225 registros, sobre esta base se calculó el 80% en conjunto de *train* 4980 registros y el 20% de datos para *test* 1245 registros.

La generación de estas muestras se realizó mediante el *software* R, utilizando librería *caret* y la función *createDataPartition*.

3.3. Modelado con árboles de decisión

En el presente trabajo se aplicó en el modelado técnicas de *Machine Learning* utilizando el método de árboles de decisión, con el algoritmo Random Forest, es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado (Braitman, 2016). El algoritmo se ejecutó en el software estadístico R.

La fórmula del algoritmo en R tiene los siguientes parámetros:

Algoritmo = *RandomForest*

Formula = La variable *defaul* que es la definición del cliente es buen o mal pagador

Data = conjunto de entrenamiento a considerar en el algoritmo

ntree= número de árboles a considerar en el modelo

Por ejemplo, la formula se desarrolló de la siguiente manera

Modelo = randomForest(formula = v_impago ~ ., data = train_balanceado, ntree = 100)
--

Posteriormente, se realizó la ejecución de este algoritmo utilizando distintos escenarios y se trabajó en los siguientes modelados:

- **Modelo 1**

El modelo 1 se consideró 14 variables: cargas familiares, edad, educación, egresos, val cuota, ingreso, calificación, ciudad, sexo, actividad, saldo, mora, oficina, v_impago. Se consideraron estas variables sin ninguna categorización, dado el modelado de árboles de decisión con *Random Forest* trabaja bien con todo tipo de variables: numéricas, ordinales, dicotómicas y cualitativas. No se consideró ninguna selección de características.

- **Modelo 2**

En el modelo 2 se consideró 14 variables: cargas familiares, edad, educación, egresos, val cuota, ingreso, calificación, ciudad, sexo, actividad, saldo, mora, oficina, v_impago. Se consideraron estas variables sin ninguna categorización, dado el modelado de árboles de decisión con *Random Forest* trabaja bien con todo tipo de variables: numéricas, ordinales, dicotómicas y cualitativas.

Este modelo se diferencia del anterior porque se aplicó selección de las mejores variables y se utilizó la técnica de selección de variables importantes a través del coeficiente de *Gini*, en la tabla 11-3, el coeficiente de Gini expresado en valores promedio.

Tabla 11-3: Promedio de Gini de la importancia de las variables

Variab les	Promedio Gini
Cargas familiares	0.87
Edad	35.49
Educación	10.69
Egresos	16.24
Valor de la cuota	39.25
Ingreso	122.34
Calificación	1.689.96
Ciudad	18.41
Sexo	1.15
Actividad	25.68
Saldo	103.25
Mora	1.805.54

Se procedió a eliminar las variables con bajos promedios de Gini, donde la variable con bajos promedios de Gini es la variable: Cargas familiares.

- **Modelo 3**

El modelo 3 se basó en las características del modelo 2 y se consideró la categorización de las variables cualitativas: educación, calificación, ciudad, sexo, actividad, oficina. Según lo presentado en el capítulo II.

3.3.1. Validación de los modelos

El presente trabajo utilizó la matriz de confusión para validar los modelos de score de crédito. La matriz de confusión es un término fundamental en Machine Learning que se utiliza para medir la precisión de un modelo comparando los valores predichos con los valores reales. La Matriz de confusión se ha utilizado en la calificación crediticia para medir la precisión de un modelo comparando el número de buenos y malos verdaderos con el número de buenos y malos predichos para una determinada puntuación de corte (Barrios, 2019).

La fórmula utilizada para la precisión (Accuracy) es:

$$\text{Precisión(Accuracy)} = \frac{VP}{VP + FP} \quad (7 - 1)$$

En la **tabla 12-3**, se presenta los resultados de los modelos: Modelo 1, con 14 variables sin aplicar selección de variables importantes para el modelo, variables cualitativas sin categorización, modelo entrenado con 100, 500 y 1000 árboles y los resultados de la precisión (accuracy) 83%, 96.1% y 96% respectivamente; Modelo 2, con 13 variables, se aplicó selección de variables importantes para el modelo, variables cualitativas sin categorización, modelo entrenado con 100, 500 y 1000 árboles y los resultados de la precisión (accuracy) 83%, 97.85% y 96.7% respectivamente; y Modelo 3, con 13 variables, se aplicó selección de variables importantes para el modelo, variables cualitativas categorizadas, modelo entrenado con 100, 500 y 1000 árboles y los resultados de la precisión (accuracy) 85%, 96% y 96.1% respectivamente.

Tabla 12-3: Comparación de los resultados del *accuracy* de los modelos 1, 2 y 3

Modelo	Variables	Selección de variables	Variables cualitativas categorizadas	Número de árboles	Precisión (Accuracy)
1	14	No	No	100	83.0%
				500	96.1%
				1000	96.0%
2	13	Si	No	100	83.0%
				500	97.85%
				1000	96.7%
3	13	Si	Si	100	85.0%
				500	96.0%
				1000	96.1%

Realizado por: Cepeda, E., 2022

3.3.2. Selección del mejor modelo

En la selección del mejor modelo se tomó la decisión de escoger los resultados de la precisión (*accuracy*) con el máximo valor.

El modelo que cumple esta condición fue el modelo 2 con 13 variables analizadas, donde se aplicó la selección de las variables importantes, sin categorizar las variables cualitativas, con 500 árboles y una precisión del 97.85% como se indica en la tabla 13-3.

Tabla 13-3: Mejor modelo

Modelo	Variables	Selección de variables	Variables cualitativas categorizadas	Número de árboles	Precisión (Accuracy)
2	13	Si	No	500	97.85%

Realizado por: Cepeda, E., 2022

Posteriormente, el resto del análisis se trabajaron de acuerdo al mejor modelo, como lo fue el modelo 2 con 13 variables, con selección de características, sin categorización de las variables cualitativas con 500 árboles y una precisión del 97.85%

3.3.3. Análisis de otras métricas de evaluación del modelo seleccionado

La curva ROC muestra la información en una sola curva graficando la distribución acumulada de los socios que pagan las cuotas a tiempo - socios buenos (eje Y) contra la distribución acumulada de los socios que no pagan a tiempo – socios malos (eje X) (Zeng, 2019, p. 23).

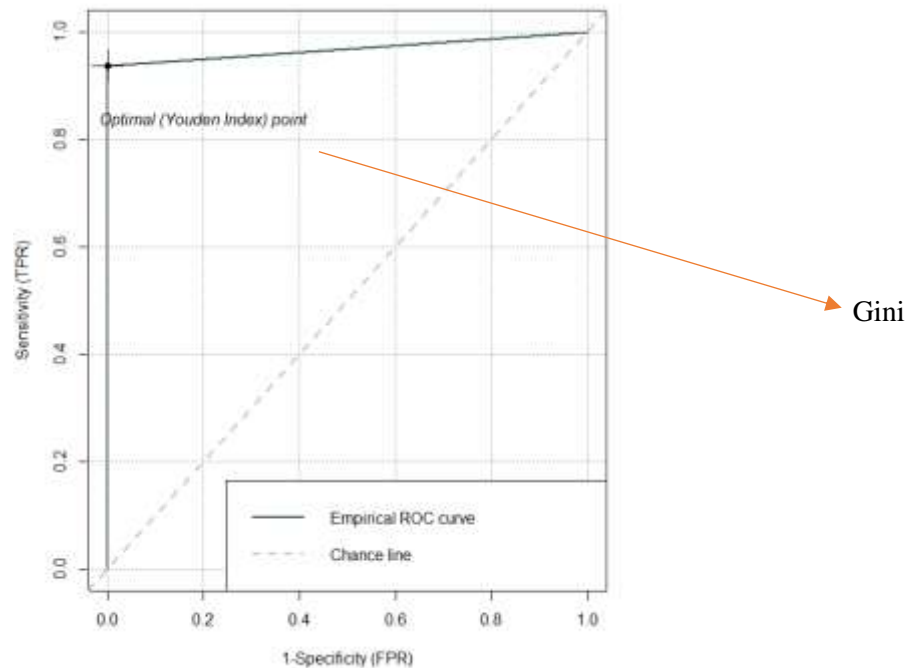


Gráfico 10-3: Curva ROC

Realizado por: Cepeda, E., 2022

En el **gráfico 10-3**, la curva describe la propiedad de clasificación del *score* en la medida en que cada punto de corte varía. El mejor score posible tendrá un ROC que va a lo largo del eje horizontal antes de subir por el eje vertical. Una curva ROC que pase por la línea entrecortada corresponderá a una en la que la probabilidad para cada *score* de ser bueno es exactamente igual a la probabilidad de ser malo. Note que la ocurrencia de esto no es lo más apropiado pues no representa ninguna ventaja en relación a clasificar aleatoriamente dado que se conoce el ratio buenos malos de la base de clientes. Por lo tanto, mientras más alejada de la diagonal esté la curva ROC es mejor el *score*. Si un *score* tiene una ROC que siempre es más alejada que otra, entonces el primero domina para todos los puntos de corte del *score*. Si mientras más lejos de la diagonal se encuentra una curva ROC es mejor el poder de clasificación del modelo, entonces mientras mayor sea el área formada por estas dos curvas mejor también será la clasificación del *score*. La medida que recoge el área bajo la curva se conoce como el coeficiente de Gini. Realmente este índice se define como dos veces el área formada por la diagonal y una curva. Este índice tendrá la propiedad de que en el caso de una perfecta clasificación su valor será de 1, mientras que una

clasificación aleatoria con una curva ROC sobrepuesta a la diagonal tendrá un valor de índice igual a 0. En si el Gini un número resumen del desempeño del *score* para todos los puntos de corte *score*.

3.3.4. Análisis de la variable respuesta: Variable impago

Para analizar la variable impago se utilizó el conjunto de datos de prueba (test) con 1245 registros. Sobre estos datos se aplicó el modelo de predicción utilizando la siguiente fórmula en el software R.

```
predict(modelo1, test, type = "prob")
```

Predicción = **predict (modelo2, data = test_balanceado, type = "prob")**

Con los resultados de esta predicción se construye la Tabla 14-3 y la descripción se detalla a continuación:

Tabla 14-3: Resultados de test del modelo 2

		PREDICCIÓN		
		Buenos pagadores	Malos pagadores	Total
ACTUAL	Buenos pagadores	832	33	865
	Malos pagadores	0	380	380
	Total	832	413	1245

Realizado por: Cepeda, E., 2022

- 832 verdaderos positivos, clientes que han sido “buenos pagadores” y el modelo los clasificó como “buenos pagadores”.
- 380 verdaderos negativos, clientes que han sido “malos pagadores” y el modelo los clasificó como “malos pagadores”.
- 33 falsos positivos, clientes que han sido “buenos pagadores”, pero fueron clasificados como “malos pagadores”.
- 0 falsos negativos, clientes que han sido “buenos pagadores”. Sin embargo, fueron clasificados como “malos pagadores”.

Al dividir los registros correctamente clasificados (la suma de 832 y 380) para el total de registros 1245 se obtiene una precisión para los “buenos pagadores” del 97.34%

3.3.5. Interpretación de los resultados de scoring de crédito

En la **tabla 15-3**, se visualiza los resultados del scoring de crédito del segmento de consumo de los socios de la cooperativa, de acuerdo con el mejor modelo seleccionado (modelo 2), los resultados del modelo arrojaron probabilidades de pago de los socios de cero (0) a uno (1), siendo 0 la peor calificación y 1 la mejor calificación, estos valores fueron categorizados de 0 a 1000 puntos y los puntos de corte se definieron según juicio de experto de Buró de Crédito Equifax.

También se puede ver que las calificaciones del score entre 0 y 300 están formadas por 275 socios y representan el 22.08%; las calificaciones del score entre 300 y 600 están formados por 138 socios y representan el 11.08%; las calificaciones del score entre 600 a 1000 corresponden a 832 socios y representan el 66.84%.

Finalmente, se concluyó que los socios con buena calificación de pago representan el 78% aproximadamente y los socios con puntuación menor a 600 son considerados como malos pagadores y representa el 22%.

Tabla 15-3: Resultado de calificaciones de los socios

Calificación	No	%
0 – 300	275	22.08%
300 – 600	138	11.08%
600 – 1000	832	66.84%
Total	1245	100%

Realizado por: Cepeda, E., 2022

En la **tabla 16-3**, se puede apreciar que los socios con calificación Buena tienen ingresos medios de 1000 dólares, 250 dólares en egresos y el monto de operación solicitado es de 5000 dólares. Mientras que los socios con calificación Mala tienen ingresos medios de 940 dólares, 225 dólares en egresos y 5000 dólares de monto de operación solicitada.

Tabla 16-3: Resultado de calificaciones de los socios

Calificación	Ingresos	Egresos	Monto de operación
Bueno	\$ 1000.00	\$ 250.00	\$ 5.000.00
Malo	\$ 940.00	\$ 225.00	\$5.000.00

Elaborado por: Cepeda, E., 2022

De acuerdo a los socios buenos pagadores se realizó un análisis de las variables cualitativas importantes como: sexo y edad.

En la **tabla 17-3**, se puede ver que los mejores pagadores son las personas del sexo masculino con un 54%, mientras que el sexo femenino es buen pagador con un 46%.

Tabla 17-3: Resultado de calificaciones de los socios

SEXO	%
Masculino	54%
Femenino	46%
Total	100%

Realizado por: Cepeda, E., 2022

En la **tabla 18-3**, se puede ver que los mejores pagadores son los socios comprendidos en la edad entre los 25 y 40 años y representa el 75%, las personas mayores a los 40 años y menores 25 años representan un 16% y 10% respectivamente.

Tabla 18-3: Resultado de calificaciones de los socios

EDAD	%
Menores 25 años	10%
Entre 25 y 40 años	75%
Mayores a 40 años	16%
Total	100%

Realizado por: Cepeda, E., 2022

Finalmente, un socio fue considerado como buen pagador si la calificación del score es mayor igual a 600, con ingreso mínimo de 1000 dólares, con egreso de 250 dólares, fue un socio del sexo masculino y con edad comprendida entre los 25 y 40 años.

CONCLUSIONES

- Una vez realizada la investigación de la literatura se encontraron artículos relacionados con la implementación de modelos de score crediticio utilizando técnicas de machine learning dentro del campo de clasificación de riesgo crediticio en entidades financieras. Esta información permitió encontrar metodologías y técnicas para la elaboración del modelo de clasificación con la información de la cooperativa.
- Para la elaboración del modelo score, se preparó el conjunto de datos, en donde se redujo el conjunto de datos de 15108 a 9065 registros, y se seleccionaron 14 variables para ingresarlas al modelo. Una vez preparada la información, se aplicó en el modelado técnicas de *Machine Learning* utilizando el método de árboles de decisión, con el algoritmo Random Forest por recomendaciones bibliográficas y se lo entrenó con 100, 500 y 1000 árboles, con selección y sin selección de variables, con y sin categorización de las variables cualitativas.
- Al realizar la comparación de la precisión de los modelos se evidenció como ganador el modelo 2 con 13 variables, con selección de características, sin categorización de las variables cualitativas con 500 árboles y una precisión del 97,85% y una tasa de error del 2.15%. Adicionalmente, el porcentaje de error para clasificar a los “malos” y “buenos” pagadores fue del 6.00% y 0.01% respectivamente. Estos resultados permiten clasificar a los “buenos” clientes. Sin embargo, se dificulta identificar a los “malos pagadores” y es necesario mejorar este porcentaje para tener mayor confiabilidad en el modelo.

RECOMENDACIONES

- La cooperativa debe ser más minuciosa al momento de registrar la información de sus socios, para elaborar un modelo scoring confiable; puesto que todo estudio posterior va a depender de la base de datos asentada en la institución financiera.
- El modelo scoring elaborado es aplicable exclusivamente para créditos de consumo, por esta razón la cooperativa, por esta razón no es aplicable para otras instituciones, puesto se analizaron solo las variables con las que cuenta la cooperativa, este tipo de estudios debe ser realizado en función a la información de cada institución financiera.
- Para futuras investigaciones, se debería evaluar la posibilidad de incluir una nueva categoría en la clase, el modelo actual únicamente cuenta con dos clases como lo son los “buenos” y “malos” pagadores, por lo que sería interesante segmentar el conjunto de datos.

BIBLIOGRAFÍA

ACOSTA, J. *Módulo II operaciones financieras fundamentales*, Quito: 2010.

BARRIOS ARCE, Juan Ignacio. *La matriz de confusión y sus métricas* [blog]. 2019. [Consulta: 03 enero 2022]. Disponible en: <https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/>

BAUTISTA ZUMBA, Wilmer Danilo. Diseño de estrategias de administración del riesgo crediticio para reducir el índice de morosidad en la Cooperativa de Ahorro y crédito Minga LTDA [En línea] (Trabajo de titulación). (Maestría) Escuela Superior Politécnica de Chimborazo, Facultad de Administración de Empresas, Maestría en Finanzas. Riobamba, Ecuador. 2021. pp. 1-131. [Consulta: 2022-01-03]. Disponible en: <http://dspace.espoch.edu.ec/handle/123456789/14629>

BONILLA MÉDICIS, Luis Oswaldo. Diseño y Validación del Scoring de aprobación de riesgo crediticio para la cooperativa de ahorro y crédito Progreso LTDA. Para las carteras de vivienda y consumo de los modelos estadísticos y econométricos [En línea] (Trabajo de titulación). (Ingeniería) Escuela Politécnica Nacional, Facultad de Ciencias, Ingeniería en Ciencias Económicas y Financieras. Quito, Ecuador. 2008. pp. 1-143. [Consulta: 2022-01-03]. Disponible en: <https://bibdigital.epn.edu.ec/handle/15000/608>

BRAIMAN, L. “Random Forests”. *Machine Learning* [En línea], 2001, 45, pp. 5-32. [Consulta: 2022-02-03]. Disponible en: <https://doi.org/10.1023/A:1010933404324>

CAZAU, Pablo. *Introducción a la Investigación*. 3 ed. Buenos Aires: 2018.

CHEN, Xi; & ISHWARAN, Hemant. “Random forests for genomic data analysis”. *Genomics* [En línea], 2012, (United States of America), 99 (6), pp. 323-329. [Consulta: 2022-02-03]. ISSN 08887543. Disponible en: <https://doi.org/10.1016/j.ygeno.2012.04.003>

DASSATTI, C. “Modelos de Score Crediticio: revisión metodológica y análisis a partir de datos de encuesta (Credit Score Models: Methodological Review and Analysis Based on Survey Data)”. *Social Science Research Network* [En línea], 2019, (Uruguay), pp. 1-33. [Consulta: 2022-02-03]. Disponible en: <https://doi.org/10.2139/ssrn.3443515>

EDUFINEXT. *Sistema financiero* [blog]. 2019. [Consulta: 2022-02-03]. Disponible en: <https://www.edufinet.com/>

ESPIN GARCÍA, O; & RODRÍGUEZ CABALLERO, Carlos V. “Metodología para un scoring de clientes sin referencias crediticias”. Cuadernos de Economía [En línea], 2013, (México) 32 (59), pp. 139-165. [Consulta: 04 abril 2022]. Disponible en: <https://revistas.unal.edu.co/index.php/ceconomia/article/view/38348>

EXPERIAN. *Evaluación del riesgo de crédito de particulares y empresas a través de modelos de scoring* [En línea]. Madrid: 2021. [Consulta: 04 abril 2022]. Disponible en: <https://www.experian.es/soluciones-empresas/necesidades-empresa/gestion-riesgo-credito/modelos-scoring>

FERREIRA, Juliana C; & PATINO, Cecilia M. “¿O que realmente significa o valor-p?”. J Bras Pneumo [En línea]. 2015, (Brasil), 41 (5), pp. 485-485. [Consulta: 20 marzo 2022]. Disponible en: <http://dx.doi.org/10.1590/S1806-37132015000000215>

ESPINOZA FREIRE, Eudaldo E. “Las variables y su operacionalización en la investigación educativa. Parte I”. Conrado [En línea], 2018, (Ecuador) 14 (1), pp. 39-49. [Consulta: 20 marzo 2022]. ISSN 1990-8644. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1990-86442018000500039

GARCÍA, S; LUENGO, J; & HERRERA, F. *Data Preprocessing in Data Mining* [En línea]. United States of America: Cham: Springer International Publishing, 2014. [Consulta: 23 marzo 2022]. Disponible en: [10.1007/978-3-319-10247-4](https://doi.org/10.1007/978-3-319-10247-4)

GARCÍA RUIZ DE LEÓN, Marta. Análisis de Sensibilidad mediante Randon Forests [En línea]. (Trabajo de titulación). (Ingeniería) Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Industriales, Grado de Ingeniería en Tecnologías Industriales, Madrid, España. 2018. pp. 1-104. [Consulta: 2022-04-21]. Disponible en: <https://oa.upm.es/53368/>

HASTIE, T; TIBSHIRANI, R; & FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* [En línea]. 2 ed. 2009. [Consulta: 2022-04-21]. Disponible en: <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiw5bHBneX6AhWNSzABHUp4Cm0QFnoECBIQAQ&url=https%3A%2F%2Flink.springer.com%2Fbook%2F10.1007%2F978-0-387-84858-7&usq=AOvVaw16HjyOnZepqyYmGOCbLefj>

ILBAY, E. *Análisis de asociaciones entre polimorfismos genéticos y fenotipos relacionados con actividad física mediante minería de datos*. Granada: 2016.

INTELLIGN. *Técnicas de machine learning*. Madrid: 2020

JUAREZ LERZUNDI, Luis Gustavo. Credit Scoring y su relación con el manejo del riesgo crediticio en la cartera pyme de la Cooperativa de Ahorro y Crédito San Pedro de Andahuaylas - Agencia Principal, Provincia Andahuaylas, Región Apurímac, 2017 [En línea] (Trabajo de titulación). [Licenciatura] Universidad Nacional José María Arguedas, Facultad de Ciencias de la Empresa, Escuela Profesional de Administración de Empresas, Andahuaylas, Perú. 2018. 1-113 [Consulta: 2022-04-21]. Disponible en: <https://repositorio.unajma.edu.pe/handle/20.500.14168/340>

KISIRYAN, M. *Sistema financiero*. 2015

MA, Yunqian; & HE, Haibo. *Imbalanced Learning: Foundations, Algorithms, and Applications*. 2013. ISBN: 978-1-118-07462-6

MEDINA RODRÍGUEZ, María del Pilar; & ULFE RENTERÍA, Henry Gustavo. Modelo de Credit Scoring para Predecir el Otorgamiento de Crédito Personal en Una Cooperativa de Ahorro y Crédito [En línea] (Trabajo de titulación). (Licenciatura) Universidad Nacional Pedro Ruiz Gallo, Facultad de Ciencias Físicas y Matemáticas, Escuela Profesional de Estadística, Lambayeque, Perú. 2017. pp. 1-121 [Consulta: 2021-10-20]. Disponible en: <https://repositorio.unprg.edu.pe/handle/20.500.12893/1339>

MOLINA ARIAS, M. “¿Qué significa realmente el valor de p?”. *Pediatría Atención Primaria* [En línea], 2017, España, 19 (76). pp. 377-381. [Consulta: 2021-10-20]. ISSN 1139-7632. Disponible en: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322017000500014

NARVÁEZ ZURITA, C, I; ORDÓNEZ GRANDA, E, M; & ERAZO ÁLVAREZ, J, C. “El sistema financiero en Ecuador. Herramientas innovadoras y nuevos modelos de negocio”. *Revista Arbitrada Interdisciplinaria Koinonía* [En línea]. 2020. (Ecuador), 5 (10). 195-225. [Consulta: 2021-10-20]. ISSN: 2542-3088. Disponible en: <https://doi.org/10.35381/r.k.v5i10.693>

NATIONAL CREDIT UNION ADMINISTRATION. *QUÉ ES UNA COOPERATIVA DE AHORRO Y CRÉDITO* [blog]. [Consulta: 10 enero 2022]. Disponible en: <https://espanol.ncua.gov/#:~:text=UU.,de%20ahorro%20y%20cr%C3%A9dito%20federales>.

PEÑA PALACIO, A; LOCHMULLER, C; MURILLO, J. G; PÉREZ, M. A; & VÉLEZ, C. A. “Modelo cualitativo para la asignación de créditos de consumo y ordinario - el caso de una cooperativa de crédito”. *Revista Ingenierías Universidad De Medellín* [En línea], 2011, (Colombia) 10 (19), pp. 101-111. [Consulta: 10 enero 2022]. Disponible en: <https://revistas.udem.edu.co/index.php/ingenierias/article/view/510>

PASCUAL MAURICIO, B; MARTÍNES de PIZÓN ASCACÍBAR, F. J; & VICENTE VIRCEDA, J. A. “Tratamiento de clases desbalanceadas con el método del cubo en problemas de credit scoring a través de la minería de datos”. *Cuadernos de economía* [En línea], 2020, (España) 43 (122), pp. 175-190. [Consulta: 10 enero 2022]. Disponible en: <https://dialnet.unirioja.es/servlet/articulo?codigo=7196160>

RIVERA VÁSQUEZ, Jairo Israel. Modelo de evaluación de crédito –scoring- para la cartera de consumo de la Cooperativa de Ahorro y Crédito “Riobamba” [En línea] (Trabajo de titulación). (Maestría) Universidad Andina Simón Bolívar, Área de Gestión, Maestría en Finanzas y Gestión de Riesgos, Quito, Ecuador. 2011. pp. 1-136. [Consulta: 2022-04-11], Disponible en: <https://repositorio.uasb.edu.ec/handle/10644/2865>

RUEDA ARIAS, Neyla Yirley. Bancarización, profundización y densidad financiera del sistema financiero ecuatoriano (2007-2013) [En línea] (Trabajo de titulación). (Economista) Pontificia Universidad Católica del Ecuador, Facultad de Economía, Escuela de Economía. Quito, Ecuador. 2016. pp. 1-133. [Consulta: 2022-04-11]. Disponible en: <http://repositorio.puce.edu.ec/handle/22000/12620>

SUPERINTENDENCIA DE BANCOS Y SEGUROS. *NORMAS GENERALES PARA LAS INSTITUCIONES DEL SISTEMA FINANCIERO.* Quito-Ecuador: 2019.

VALENCIA ECHEVERRÍA, Andrea. Modelo Scoring para el otorgamiento de crédito de las pymes [En línea] (Trabajo de titulación). (Maestría) Universidad EAFIT, Escuela de Economía y Finanzas. Medellín, Colombia. 2017. pp. 1-28. [Consulta: 2022-04-11]. Disponible en: <https://repository.eafit.edu.co/handle/10784/12295>

VALLE BENAVIDES, Ana Rocío. Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones [En línea] (Trabajo de titulación). (Trabajo Final de Grado) Universidad de Sevilla, Departamento de Estadística e Investigación Operativa, Sevilla, España. 2017. pp. 1-77. [Consulta: 2022-05-11]. Disponible en: <https://idus.us.es/handle/11441/63201>

VERDEZOTO CAMACHO, Jessica Jimena. Elaboración y evaluación de un score para crédito de consumo en la Cooperativa de Ahorro y Crédito COOPAD [En línea] (Trabajo de titulación). (Maestría) Universidad Andina Simón Bolívar, Área de Gestión, Programa de Maestría en Finanzas y Gestión de Riesgos. Quito, Ecuador. 2016. pp. 1-115. [Consulta: 20 marzo 2022]. Disponible en: <https://repositorio.uasb.edu.ec/handle/10644/4996>

ZENG, G. *On the confusion matrix in credit scoring and its analytical properties* [En línea]. Communications in Statistics - Theory and Methods, 2020. [Consulta: 2022-05-11]. Disponible en: <https://doi.org/10.1080/03610926.2019.1568485>



ANEXOS

ANEXO A. AVAL DE LA COOPERATIVA DE AHORRO Y CRÉDITO MINGA LTDA.



Oficio N° 099 -GG-CML-2021
Riobamba 08 de diciembre de 2021


Ing.
Pablo Flores Muñoz
DIRECTOR CARRERA ESTADISTICA ESPOCH
Presente. -


Jorge Vicente Chucho Lema en mi calidad de Gerente y Representante Legal de la Cooperativa de Ahorro y Crédito Minga Ltda., ante Usted muy respetuosamente comparezco y manifiesto:

La Cooperativa de Ahorro y Crédito Minga Ltda., autoriza a la Srta. **CEPEDA GUAMINGA ANA ELIZABETH** de c.c. 0604561019, estudiante de la Facultad de Ciencias carrera de Estadística de la Escuela Superior Politécnica de Chimborazo a desarrollar el trabajo de titulación "MODELO DE SCORING PARA CRÉDITO DE CONSUMO EN LA COOPERATIVA DE AHORRO Y CRÉDITO MINGA LTDA. UTILIZANDO TÉCNICAS DE MACHINE LEARNING", la Institución a la que regento brindará todas las facilidades en cuanto a la información para que pueda realizar el trabajo de titulación.

A su vez se recuerda que la información proporcionada a la estudiante es de uso CONFIDENCIAL.

Atentamente,


Cooperativa de ahorro y crédito
Minga Ltda.
GERENTE GENERAL



Ing. Jorge Vicente Chucho Lema
Gerente General
(+593) 03 3730810 Ext. 71205
jchucho@coopminga.com / www.coopminga.com
Riobamba - Ecuador

ANEXO B. CÓDIGO EN R.

1. definición de librerías

```
require(data.table)
require(stringr)
require(xlsx)
require(lubridate)
library(caret)
require(stringr)
require(mvoutlier)
library(ROSE)
library(randomForest)
library(e1071)
library(ROCit)
require(ggplot2)
```

2. Análisis de los datos

```
datos = read.csv( paste0(path_bases, "BASE_COOPERATIVA.csv"), header = T, sep = "," )
datos = unique(datos)
dim(datos) # análisis del conjunto de datos
colnames(datos) #nombre de las variables
str(datos) # definición de los tipos de datos
```

3. Preprocesamiento de los datos

```
res <- apply(datos , 1 , function (x) sum( is.na(x) ))/ncol(datos) * 100 # cálculo de los valores
perdidos en variables

mal = res > 5 # Aquellas filas con valor de porcentaje mayor al 5%, equivale a una variable.
filtrado = datos[ !mal, ] # se elimina las variables con porcentajes mayores al 5%

## eliminar variables que no agregan valor para el análisis
filtrado$num.opera=NULL
filtrado$numero.cuota=NULL

##variables con datos faltantes
col <- apply(filtrado , 2 , function (x) sum( is.na(x) ))/nrow(filtrado) * 100

val_nulos<-data.frame((table(round(col,digits = 0))))
min(val_nulos$Freq)
max(val_nulos$Freq)
mean(val_nulos$Freq)
```

```

colnames(val_nulos)<-c("Var1","Variables")

p <- ggplot(val_nulos,aes(x =Var1 , y=Variables,fill=Variables))+geom_bar(colour =
"black",stat="identity")
p +coord_cartesian(ylim = c(1, 18)) +
  labs(x = "% de Valores perdidos", y = "No. Variables",title="Valores faltantes por
variables")+ ### si ejecuto el codigo hasta aqui los gráficos tiene una escala del azul
  scale_fill_gradient2(limits=c(1, 18),midpoint =6.67 ,
    low="red",mid="yellow" , high="green")

## resultado después de eliminar las columnas con valores perdidos mayor al 8% de datos
faltantes
filtrado = filtrado[ , - which( col > 8)]
filtrado= filtrado[ filtrado$saldo > 0, ] # se eliminan los saldos mayores a 0

```

4. Análisis descriptivo : análisis por cada una de las variables cuantitativas

```

tmp = as.data.table(filtrado)
boxplot(filtrado$saldo, outline = T )

r = tmp[ tmp$saldo <= 15000, ]

mean(r$saldo)
median(r$saldo)
sd(r$saldo) # desviación estandar
var(r$saldo)
min(r$saldo)
max(r$saldo)

#

```

5. Análisis de correlación de las variables cuantitativas

```

filtrado = r

cols= c( "ingreso", "saldo" , "monto.de.operaci.n", "egresos", "valcuota", "mora")

tmp_corr= cor( filtrado[, c( "ingreso", "saldo" , "monto.de.operaci.n", "egresos", "valcuota",
"mora") ] )

# se eliminan las variables altamente correlacionadas con el saldo

filtrado = filtrado[, - "monto.de.operaci.n"]

```

6. Detección de datos anómalos

```
cols= c("cargas.familiares", "edad", "egresos", "valcuota", "ingreso", "saldo",
        "mora")

filtrado = as.data.frame(filtrado)

set.seed(123)
d_outliers = uni.plot( filtrado[, cols] )

print(d_outliers$outliers)

filtrado = filtrado[ !d_outliers$outliers, ] ## se elimina el 15% de los datos
filtrado$v_impago =ifelse(gsub(" ", 0, filtrado$mora) <= 25 , "Bueno", "Malo" )
```

7. Análisis de la variable Default: variable impago

```
t = as.data.table(filtrado)

r = unique(t[, .SD, .SDcols= c("v_impago", "cod.cliente")])

length(unique(t$cod.cliente))

table(r$v_impago)

table(filtrado$v_impago) ## tiene 3.5% de todas las transacciones

filtrado$v_impago = as.factor(filtrado$v_impago)

filtrado$fecha_corte=NULL
```

8. Particionar la base de datos en train y test

```
filtrado$producto= NULL
filtrado$cod.cliente= NULL
head(train_balanceado_cat,3)

set.seed(1337)
index <- createDataPartition(filtrado$v_impago,
                             p = 0.8, # % of data going to training
                             times = 1,
                             list = F)

train <- filtrado[ index,]
test   <- filtrado[-index,]
```

9. Balance de clases

```
table(train$v_impago)
set.seed(1337)
#train_balanceado= ovun.sample(v_impago~., data = train, method = "over", N =
621)$data
train_balanceado= ovun.sample(v_impago~., data = train, method = "over", N =
480)$data
```

10. Desarrollo del modelo 1

```
set.seed(123)
train_balanceado$v_impago = as.factor(train_balanceado$v_impago)

train_balanceado
modelo1 <- randomForest(v_impago~., data=train_balanceado, ntree= 500)

print(modelo1)

p1 <- predict(modelo1, test, type = "prob")
cal1 <-predict(modelo1, test, type = "response")
```

11. Desarrollo del modelo 2 con selección de características

```
importance(rf) # selección de las variables importantes del modelo 1 y aplicar en el
modelo 2

set.seed(123)
train_balanceado$v_impago = as.factor(train_balanceado$v_impago)

train_balanceado
modelo2 <- randomForest(v_impago~., data=train_balanceado, ntree= 500)

print(modelo2)

p1 <- predict(modelo2, test, type = "prob")
cal2 <-predict(modelo2, test, type = "response")
```

12. Desarrollo del modelo 3 con caracterización de las variables cualitativas

```
# caracterización de las variables
filtrado$V_educacion = ifelse(filtrado$educacion=="SINESTUDIOS",1,
                             ifelse(filtrado$educacion=="PRIMARIA",2,
                                     ifelse(filtrado$educacion=="SECUNDARIA",3,
                                             ifelse(filtrado$educacion=="SUPERIOR",4,
                                                    ifelse(filtrado$educacion=="TECNICA",5,
                                                           ifelse(filtrado$educacion=="MASTERADO",
                                                                6,
                                                                 ifelse(
filtrado$educacion=="INDEFINIDA",7,8))))))))))

filtrado$V_calif = ifelse(filtrado$calificacion=="A1",1,
                         ifelse(filtrado$calificacion=="A2",2,
                                 ifelse(filtrado$calificacion=="A3",3,
                                         ifelse(filtrado$calificacion=="B1",4,
                                                 ifelse(filtrado$calificacion=="B2",5,
                                                         ifelse(filtrado$calificacion=="C1",6,
                                                                 ifelse(filtrado$calificacion=="C2",7,
                                                                           ifelse(filtrado$calificacion=="D",8,
                                                                                 ifelse(filtrado$calificacion=="E",
                                                                                     9,10))))))))))
))))))

filtrado$V_actividad = as.factor(filtrado$actividad)
filtrado$V_oficina = as.factor(filtrado$oficina)

## desarrollo del modelo 3

set.seed(123)
train_balaceado$V_impago = as.factor(train_balaceado$V_impago)

train_balaceado
modelo3 <- randomForest(V_impago~., data=train_balaceado, ntree= 500)

print(modelo3)

p1 <- predict(modelo3, test, type = "prob")
cal3 <- predict(modelo3, test, type = "response")
```

13. Validación de los modelos

```
# Matriz de confusion
confusionMatrix(cal1, test$v_impago)
confusionMatrix(cal2, test$v_impago)
confusionMatrix(cal3, test$v_impago)

# Análisis de la curva ROC
#modelo 1
dres <- data.frame( pred=predict(modelo1, test, type="response"),
var=test$v_impago)
str(dres)
ROC <- rocit(as.numeric( dres$pred), as.numeric(dres$var))
plot(ROC)

#modelo 2
dres <- data.frame( pred=predict(modelo2, test, type="response"),
var=test$v_impago)
str(dres)
ROC <- rocit(as.numeric( dres$pred), as.numeric(dres$var))
plot(ROC)

#modelo 3
dres <- data.frame( pred=predict(modelo3, test, type="response"),
var=test$v_impago)
str(dres)
ROC <- rocit(as.numeric( dres$pred), as.numeric(dres$var))
plot(ROC)
```



epoch

**Dirección de Bibliotecas y
Recursos del Aprendizaje**

**UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y
DOCUMENTAL**

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 10 / 01 / 2023

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: Ana Elizabeth Cepeda Guaminga
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Estadística
Título a optar: Ingeniera Estadística
f. responsable: Ing. Cristhian Fernando Castillo Ruiz

2381-DBRA-UTP-2022