Elaboración de un modelo de Scoring para el otorgamiento de créditos de bajos montos en una cooperativa de ahorro y crédito en Colombia.

Yony Javier Gaviria Orozco y Dairo Josue Díaz Meléndez

Trabajo de Grado para Optar el Título de Especialista en Estadística

Director

Henry Lamos Diaz

Ph. D en Física- Matemáticas

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Bucaramanga

2020

DEDICATORIA

A Dios por su infinito amor y misericordia, por darnos salud y sabiduría para lograr esta meta, por colocar en nosotros fortaleza y perseverancia para continuar sin importar los obstáculos que se nos presentó en el camino.

A nuestros padres por creer en nosotros, por todo el apoyo, sus consejos y el amor que nos brindaron en el transcurso de esta especialización, por formarnos con principios y valores y hacer de nosotros mejores personas para la sociedad.

A financiera Comultrasan por contar con nosotros, brindándonos la posibilidad de poder realizar una meta más en nuestro ciclo profesional.

AGRADECIMIENTOS

A Dios por su infinito amor y misericordia, por darnos salud y sabiduría para lograr esta meta, por colocar en nosotros fortaleza y perseverancia para continuar sin importar los obstáculos que se nos presentó en el camino.

A nuestras familias por creer en nosotros, por todo el apoyo, sus consejos y el amor que nos brindaron en el transcurso del posgrado, por formarnos con principios y valores y hacer de nosotros mejores personas para la sociedad.

A Financiera Comultrasan por permitirnos cumplir con un sueño más en el ámbito profesional, por confiar en nosotros, esperamos retribuir de forma satisfactoria lo aprendido y poder aplicarlo en el diario laboral.

A las Universidad Industrial de Santander, UIS, a todos sus directivos y especialmente a los docentes que nos acompañaron en este proceso de especialización, por compartir con nosotros todos sus conocimientos y darnos las herramientas necesarias para nuestra vida laboral.

A nuestros compañeros de estudio, amigos que estuvieron en cada momento de alegría y frustración, que fueron nuestro apoyo para continuar, por cada chiste, por cada palabra de ánimo, por cada consejo, por todas las noches donde no se podía dormir por estudiar, a ustedes, Gracias.

Yony Javier Gaviria Orozco,

Dairo Josue Díaz Meléndez

Contenido

Lista de tablas

Lista de figuras

Lista de Apéndices

Resumen

Título: Elaboración de un Modelo de Scoring para el Otorgamiento de Créditos de Bajos Montos en una Cooperativa de Ahorro y Crédito en Colombia.

Autores: Yony Javier Gaviria Orozco, Dairo Josue Díaz Meléndez

Palabras claves: Default - Riesgo de crédito - Scoring de crédito - Indicador de mora - Capacidad de pago

Descripción:

La herramienta del uso del scoring se ha potenciado desde los años noventa con el fin de mitigar el riesgo de crédito. El presente proyecto se trata de proponer un modelo de scoring para una entidad financiera en el departamento de Santander – Colombia, vigilada por la Súper Intendencia Solidaria, a un segmento de clientes que no tengan experiencia crediticia en el sector financiero y cuyos montos de créditos no superen los 2 salarios mínimos legales vigentes. (Smmlv). Se implementara un modelo de regresión logística con variables cuantitativas, cualitativas, sociodemográficas, este tipo de modelo busca estimar una probabilidad de incumplimiento para lograr distinguir o discriminar entre clientes buenos o clientes malos. Para el cálculo del modelo se utilizaron datos propios de la entidad financiera a la cual se está implementando el estudio con datos desde el año 2016 hasta el años 2019, la creación de este modelo permite tener un parámetro objetivo y cuantitativo para cada cliente. Las herramientas utilizadas para el desarrollo del proyecto fueron SPSS que nos permitió obtener las salidas de resultados y la herramienta Excel, que nos permitía graficar los resultados obtenidos. La implementación del presente proyecto quedara a disposición de la entidad financiera a la cual se está realizando el estudio.

[*] Trabajo de grado

[**] Facultad de Ciencias. Escuela de Matemáticas Especialización en Estadística. Director: Henry Lamos Diaz

Abstract

Title: Elaboration of a scoring model for granting of low- cost credits in a credit unión in Colombia

Authrs: Yony Javier Gaviria Orozco, Dairo Josue Díaz Meléndez

Key Words: Default - Credit Risk - Credit Scoring - Default Indicator - Payment capacity

Description:

The scoring tool has been strengthened since the 1990s in order to mitigate credit risk. This project is about proposing a scoring model for a financial institution in the Santander - Colombia department, supervised by the Superintendency of Solidarity, to a segment of clients who do not have credit experience in the financial sector and whose loan amounts do not exceed the 2 legal minimum wages in force. (Smmlv). A logistic regression model will be implemented with quantitative, qualitative, sociodemographic variables, this type of model seeks to estimate a probability of default in order to distinguish or discriminate between good clients and bad clients. For the calculation of the model, data from the financial institution to which the study is being implemented with data from 2016 to 2019 were used, the creation of this model allows having an objective and quantitative parameter for each client. The tools used for the development of the project were SPSS that allowed us to obtain the results outputs and the Excel tool, which allowed us to graph the results obtained. The implementation of this project will be available to the financial entity to which the study is being carried out.

[*] Trabajo de grado

[**] Facultad de Ciencias. Escuela de Matemáticas Especialización en Estadística. Director: Henry Lamos Diaz

Introducción

Las entidades financieras en Colombia que tienen entre sus funciones principales el otorgamiento de créditos a personas se denominan Establecimientos de Crédito en la modalidad de Bancos o Cooperativas especializadas en Ahorro y Crédito (Presidencia de la República, 1993).

Esta actividad de prestar dinero, a la espera de que sea devuelto en un tiempo posterior pagando un mayor valor como consecuencia del cobro de intereses lleva a un riesgo de pérdidas para las entidades crediticias, en caso de que los clientes a los que se les otorgaron los recursos no los devuelvan en las condiciones pactadas, es decir, cuando los clientes incumplen sus compromisos (Superintendencia Financiera de Colombia, 1995). Por lo anterior resulta necesario para dichas instituciones establecer mecanismos que permitan cuantificar la probabilidad de incumplimiento de los clientes desde el momento mismo en que solicitan los préstamos de dinero, de forma que la decisión de aprobación se pueda tomar utilizando una medición de la exposición a riesgo inherente a cada uno de los créditos por otorgar. Según los modelos para medir los riesgos de crédito en la banca escrito por Saavedra García María y Saavedra García Máximo, en su texto nos indica que los modelos para estimar la probabilidad de incumplimiento surgieron de manera formal durante la década de los setenta e ilustran varios modelos de riesgo como: Modelo KMV, modelo de valuación de Merton, modelo Credimetrics de J.P. Morgan, modelo de retorno sobre capital ajustado al riesgo Falkenstein y modelo CyRCE (Saavedra García & Saavedra García, 2010).

En particular, la situación a abordar en este proyecto es la operación de otorgamiento de crédito de bajo monto, donde las cooperativas apuntan a atender este segmento del mercado que no ha sido explorado por parte del sector financiero dado que actualmente esta necesidad viene siendo atendida por lo que comúnmente se denominan prestamos “gota a gota”, en los cuales los

microempresarios y trabajadores independientes acceden a créditos con poca exigencia debido a sus pocos requisitos solicitados, con desembolso inmediato y envío al lugar donde desarrollan sus actividades, cobro personal diario, entre otros. Estas y otras particularidades se tendrán en cuenta en este trabajo para la elaboración de los perfiles basados en una metodología de aprobación de las operaciones.

En tal sentido, resulta relevante abordar la problemática relacionada tal como se especifica en el texto Building credit scorecards using SAS and Python donde se establece la necesidad de esquemas de medición de riesgo para el otorgamiento de créditos de bajos montos, los cuales se pueden evaluar con tarjetas de puntaje y el valor de la calificación crediticia. Los esquemas se materializan en cuadros de mando, los cuales a su vez se clasifican en cuadros de comportamiento y cuadros de mando de la aplicación (Building credit scorecards using SAS and Python, 2016).

Los cuadros de mando de comportamiento se ocupan más de predecir o puntuar a los clientes actuales y su probabilidad de incumplimiento. Los cuadros de mando de la aplicación se utilizan cuando los nuevos clientes solicitan préstamos para predecir su probabilidad de ser clientes rentables y para asociarles un puntaje. Para las entidades financieras, la calificación crediticia ayuda a gestionar el riesgo, depende de la empresa evaluar la solvencia crediticia y los puntajes crediticios de los consumidores para identificar soluciones de productos optimas basadas en el riesgo, tiempos de respuesta, negaciones de créditos incorrectos y más. El uso de este modelo de scoring puede optimizar el riesgo y maximizar la rentabilidad para las entidades financieras. La cooperativa Financiera Comultrasan, entidad en la que se pretende desarrollar el presente proyecto, posee modelos de scoring diseñados para créditos de otras líneas, sin embargo hay poca participación en operaciones de bajo monto debido a la operatividad que genera el otorgamiento

de este tipo de créditos y que escasamente representan el 3% del total de la cartera según la cifras otorgados por la misma entidad (Financiera Comultrasan, 2020).

Ante este panorama, se planteó llevar a cabo el presente proyecto cuyo objetivo es diseñar un modelo de scoring que aporte un criterio que apoye la toma de decisiones en relación con el otorgamiento de créditos de bajo monto, menos de dos salarios mínimos, basados en el uso de técnicas estadísticas, de este modo, se exploraran varios modelos de scoring para establecer el que mejor se ajusta a las características del perfil de riesgo de los asociados y clientes de la entidad financiera sobre la que se apoyará este trabajo e identificar las variables cualitativas y cuantitativas disponibles en la población en estudio que pueden incidir en el incumplimiento de las personas que han contratado créditos de consumo con dicha entidad.

1. Antecedentes

Los scoring de crédito son herramientas estadísticas usadas desde los años setenta para asignar una probabilidad de incumplimiento a los solicitantes de crédito, realizando una ponderación de sus variables cualitativas y cuantitativas y a su vez, con inclusión de las condiciones del crédito solicitado. El uso de scoring se ha potenciado desde los años noventa con el fin de mitigar el riesgo de crédito de las solicitudes y buscar los clientes que se ajustan al perfil de la entidad (Gutiérrez Girault, 2007). Así también estos modelos han mostrado contribuir con los tiempos de respuesta para aprobar o denegar el crédito, cuando no se cuenta con dichos scoring se pueden negar equivocadamente un buen prospecto de crédito, en conclusión, se pueden perder clientes ante competidores (Bulding and Implementing Better, 2016).

En Financiera Comultrasan desde el año 2010 se vienen implementando modelos de scoring para los segmentos de crédito de microfinanzas que corresponde a una cartera clasificada como microcrédito, donde el volumen de operaciones es considerable y años después se aplicó a líneas de créditos dirigidas a Empleados, pensionados, Banca Pyme cartera, estos son clasificados como Consumo y Comercial, dichos scoring no es común encontrar metodologías para aplicarlos a clientes sin referencias crediticias, es decir clientes que carecen de información en las entidades nacionales de crédito. Este trabajo presenta una metodología general usando información sociodemográfica para construir un modelo sencillo de crédito scoring enfocado justamente a esa población la cual ha venido tomando una mayor importancia en el sector crediticio con el fin de mitigar y controlar las probabilidades de incumplimiento y así mismo con el fin de dar cumplimiento a lo establecido por los entes reguladores (Cuaderno Economico, 2013) en cuanto a la implementación de un sistema de riesgo que disminuya posibles pérdidas económicas por medio del cuidado de cada una de las etapas que comprenden el ciclo de créditos en este caso, de manera específica el otorgamiento.

2. Justificación

Las entidades financieras de hoy día buscan desarrollar nuevos productos innovadores en mercados no explorados que no han sido aún atendidos por múltiples razones. Se considera que existe una gran cantidad de trabajadores independientes e informales con bajos ingresos que tienen que recurrir a diferentes modalidades de préstamos por fuera de la banca tradicional al no tener un historial crediticio y respaldo económico; de otro lado, la cooperativa de ahorro y crédito Financiera Comultrasan entre el portafolio de servicio no tiene diseñado una metodología para la

evaluación de créditos de bajo monto para atender está población, por ello ha decidido explorar un nuevo servicio que se ajuste a las necesidades de los trabajadores independientes.

En este sentido, teniendo en cuenta que se requiere de un proceso fácil y rápido que no genere traumatismo en los actuales procesos y que a su vez los costos operativos de personal no se aumenten, ha decidido desarrollar una metodología de créditos basada en métodos estadísticos para la línea de créditos que no superen los dos salarios mínimos mensuales legales vigentes para personas independientes sin experiencia crediticia. Por tal razón el presente trabajo está dirigido a la construcción de un modelo de scoring que ayudaría a la organización a mejorar el proceso de decisión de aprobar o negar una operación futura de crédito a un segmento con actividad económica independiente sin experiencia crediticia en las centrales de riesgo; se espera que el modelo contribuya a que se tenga un control sobre el riesgo de cartera y una mejor calidad del mayor activo de la cooperativa, el cual es la cartera colocada en las operaciones crediticias a los asociados.

3. Objetivos

3.1. Objetivo General

Proponer un modelo de scoring para la aprobación de créditos de bajos montos a partir de información propia de la cooperativa.

3.2. Objetivos Específicos

- Determinar las variables que afectan el scoring de un cliente para la aprobación del crédito.

- Editar una base de datos con información propia de la cooperativa Financiera Comultrasan para la construcción del modelo.
- Aplicar el proceso de transformación y limpieza de la base de datos construida.
- Ajustar un modelo de regresión logística con los datos recopilados.
- Validación de supuestos del modelo de scoring propuesto.

4. Alcances y Delimitaciones Del Proyecto

Para la elaboración del modelo de scoring de aprobación de bajos montos se cuenta con información sociodemográfica con un gran número de clientes, dicha información se encuentra en una base de datos de la cooperativa Financiera Comultrasan, la cual nos aportan información de vital importancia desde el año 2016 hasta el 2019. La calidad de la información es buena y fue relevante a la hora de trabajar las variables seleccionadas dado que no fue necesario la eliminación de alguna de ellas por inconsistencias o ausencia de campos.

Este modelo no considera información del mercado, como información reportada por entidades estatales, externos o indicadores macroeconómicos que pudiera explicar aún más el modelo propuesto, ahora bien, si se decidiera incluir, se debe tener en cuenta que el mercado a atender es muy informal y se considera que con la información suministrada se logra una construcción adecuada de un modelo de predicción.

Los cálculos y manejos de la información, así como los registros fueron resumidos y tratados en el programa estadístico IBM SPSS junto con EXCEL para una optimización de los procesos, ya que nos permite graficar y observar con una mayor claridad dicha información.

5. Marco Teórico

5.1. Riesgo De Crédito

El riesgo crediticio es la probabilidad de que una organización solidaria incurra en pérdidas y se disminuya el valor de sus activos como consecuencia de que sus deudores incumplan con el pago de sus obligaciones en los términos acordados (Samaniego Medina, 2008).

La superintendencia y el direccionamiento de riesgos de crédito, nos indica ciertas funciones o responsabilidades a cargo las cuales son: ejecutar la supervisión de los riesgos de crédito de acuerdo a las políticas, metodologías, y procedimientos aprobados por el superintendente, realizar seguimiento, monitoreo y evaluación de los riesgos, participar en la práctica de visitas con el fin de obtener conocimientos, verificar que las entidades vigiladas cumplan con las normas y cuenten con sistemas y procesos adecuados para clasificar, valorar y contabilizar la cartera de créditos (Superintendencia Financiera de Colombia, 2017).

Por otra parte, según la reforma de Basilea III y en términos más generales, nos brinda una respuesta ante la crisis financiera mundial, permitiendo al sistema apoyar a la economía real a lo largo del ciclo de la economía, obteniendo una mejora en la calidad del capital regulador bancario, aumentar el nivel de los requerimientos de capital con el fin de que los entidades sean resilientes, mejorando la cuantificación del riesgo, mejora la solidez y sensibilidad al riesgo de crédito, pueden optar por implementar requerimientos más conservadores y/o disposiciones transitorias aceleradas, puesto que el marco de Basilea constituye exclusivamente un conjunto de normas mínimas (Comité de Supervisión Bancaria de Basilea, 2017).

Una de las estrategias o mecanismos para la mitigación del riesgo crediticio es el Sistema de Administración de Riesgos Crediticios SARC, según el cual las instituciones financieras deben vigilar de manera permanente el historial crediticio de un cliente atreves de ciertos aplicativos, con el fin de monitorear todas las actividades crediticias de una entidad para así gestionar y vigilar respectiva a sus clientes para ver si son solventes (Rankia & Trecet, 2020). Otra estrategia para la mitigación del presente riesgo es el Sistema de Administración de Riesgo Operativo SARO, en el cual permite identificar los diferentes riesgos que sean actuales y potenciales que puedan ocasionar o generar algún conflicto en los diferentes procesos de la organización, obteniendo un control y adaptación de diferentes acciones para la mitigación de los riesgos, teniendo en cuenta la importancia de realizar monitoreos, inspecciones y seguimientos para lograr mejoras al sistema o acciones correctivas (Ministerio de Hacienda, 2013)

Como bien sabemos el riesgo es una variable que se debe tener en cuenta por la sensibilidad que presenta ante un resultado financiero de una entidad con actividad de colocación de créditos, por tanto, es de vital importancia medir la probabilidad que tiene un deudor frente a un acreedor de cumplir con sus obligaciones de pago, ya sea durante la vigencia de los recursos otorgados (Peiro Ucha, 2018).

Dentro de las actividades de la Gestión de Riesgos, se puede definir el riesgo de crédito como la probabilidad que puede presentar un cliente de al momento del vencimiento de sus obligaciones, una entidad no recupere o logre retornar la devolución en su totalidad de los recursos otorgados mas los rendimientos acordado sobre un instrumento financiero, debido a quiebra, iliquidez o alguna otra razón (Ealde, 2018).

En resumen, el riesgo se concibe como la posibilidad de perder algo o de tener un resultado que implica perdida de recursos o riesgo de materialización, Dicho esto podemos decir que el riesgo de una actividad puede tener dos componentes: la posibilidad o probabilidad de que un resultado negativo ocurra y el tamaño de ese resultado. Por lo tanto, mientras mayor sea la probabilidad y la pérdida potencial, mayor será el riesgo (Perez López C. , 2004).

5.2. Factores Determinantes Del Riesgo De Crédito

Los autores Restrepo y Arango nos indican que los factores determinantes en la medición de riesgos y tomando como base el modelo de Scoring diseñado por Samaniego establecen que (Arango & Restrepo, 2017):

- La probabilidad de incumplimiento es aquella en la cual la persona que solicita el crédito no cumple con sus responsabilidades.
- La exposición es el valor de perdida en el que incurre en el momento de incumplimiento de la contraparte.
- Porcentaje de perdida resultante luego del incumplimiento y la recuperación, la recuperación depende de la garantía del crédito puesto que una garantía disminuye el riesgo de crédito si la cobertura de la deuda que proporciona es adecuada y es de fácil realización, es decir, puede convertirse en dinero con facilidad (Saavedra García & Saavedra García, 2010).

5.3. Tipos De Riesgo De Crédito

Para los riesgos de crédito contemplados se es necesario diferenciar 4 tipos de estos los cuales se han podido establecer según (Ealde, 2018):

- Riesgo de impago: Es la posibilidad de incurrir en una perdida si la contraparte no cumple plenamente las obligaciones financieras, acordadas en el contrato, también llamado riesgo fallido.
- Riesgo de migración: Grado en que puede mejorar o deteriorarse la calidad crediticia o calificación del crédito.
- Riesgo de exposición: se entiende como la incertidumbre sobre los futuros pagos que se deben, este riesgo puede estar asociado a la actitud del cliente o bien a la evolución de variables del mercado.
- Riesgo colateral: Conocido como el riesgo de la tasa de recuperación, que varía según haya o no garantías colateral de la operación.

5.4. Glosario

A continuación, se presentarán algunas de las definiciones más relevantes para nuestro proyecto en la elaboración de un modelo de scoring para el otorgamiento de créditos de bajos montos son:

- **Riesgo de crédito:** Definido por la Superintendencia financiera de Colombia como la posibilidad de qué una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que un deudor o contraparte incumpla sus obligaciones.
- **SARC:** Se define como el sistema de administración del riesgo del crédito y es el conjunto de metodologías, procedimientos y políticas a través de las cuales se evalúa, asume, califica, controla y administra el riesgo crediticio.
- **Cosechas:** Se define como el conjunto de desembolso de obligaciones realizadas en un periodo de tiempo definido. El análisis de cosechas se fundamenta en identificar los periodos de colocación de cartera y como a través del tiempo este ha presentado resultados

óptimos y deficientes en cuanto a calidad de la cartera, castigo, seguimiento y recuperación, las causas que dieron lugar a ciertos comportamientos y el contexto sobre el cual se desenvuelven las fases de colocación, seguimiento y recuperación.

- **Indicador de mora:** Indicador financiero que busca medir la calidad de la cartera en términos de la proporción de saldos con días de mora superiores a 30 días sobre el saldo vigente de la cartera.
- **Capacidad de pago:** Indicador que evalúa la suficiencia que tiene un posible deudor para atender una obligación de manera adecuada.
- **Incumplimiento:** Estado en el que entra un deudor cuando deja de realizar los pagos de la deuda y la entidad financiera considera que el recaudo de los recursos no se va a realizar por parte del deudor.
- **Scoring de crédito:** Es un sistema de evaluación que permite, a partir de un conjunto de variables cualitativas y cuantitativas, predecir la probabilidad de pago de un deudor.

5.5. Métodos Multivariados

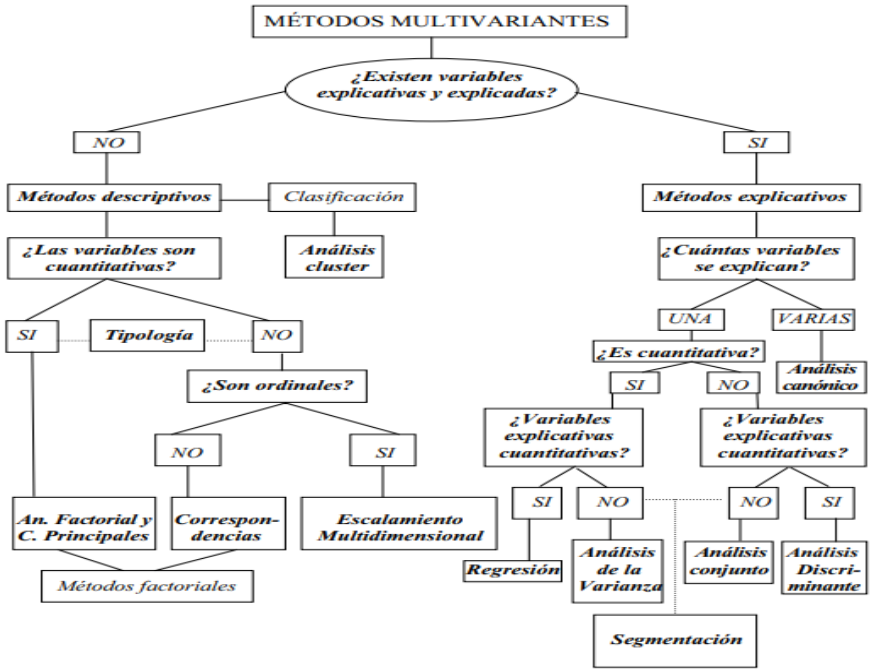
El análisis multivariado permite la resolución de problemas y la toma de decisiones mediante métodos estadísticos y matemáticos sobre todas las variables que influyen en el problema que se trata. En el análisis multivariado existe una gran variedad de métodos que ayudan al estudio e interpretación de la información. Es frecuente clasificar los métodos multivariantes en funcionales o dependientes y en funcionales o interdependientes de acuerdo con la finalidad que persiga.

Las técnicas o métodos funcionales o dependiente construyen un modelo, ecuación o función formada por el conjunto de variables involucradas. En estas técnicas se utiliza una variable respuesta (dependiente) y unas variables independientes. Los métodos multivariantes

(multivariados) estructurales o interdependiente tiene con objetivo el resumir la información. Es importante decir que en estos métodos todas las variables se manejan como independientes. En la Figura 1 se puede observar la clasificación dependiendo de las características que tiene el conjunto de datos analizar.

Figura1.

Métodos Multivariantes.



(Perez López C. , 2004).

Algunos de los análisis que se llevaran a cabo en el presente trabajo con el propósito de identificar o segmentar mejor a los clientes y construir el modelo scoring se detallan a continuación.

5.6. Regresión Logística

Para el presente proyecto de elaboración del modelo de scoring para la aprobación de créditos de bajos montos se utilizara esta técnica estadística denominada regresión logística la cual mide factores multivariantes y es destinada al análisis de una relación de dependencia entre una variable dependiente y un conjunto de variables independientes, de forma similar a como actúa el análisis de regresión lineal clásico. Dicha técnica nos determinara si existe un comportamiento que determine factores asociados al ingreso de los clientes a un default.

La finalidad de poder contar con predicciones de los rasgos con los que se identifique el patron de los clientes, dicho de otra forma se pueden evaluar las probabilidades de un suceso que ocurre o se define por la variable dependiente en función de otras variables que permiten identificar los patrones del sujeto a identificar. Para identificar las variables en un modelo clásico de regresión lineal la variable dependiente es cuantitativa, también es importante conocer que en estas variables independientes se puede incorporar variables cualitativas construidas a través de una condición tipo dummy como, por ejemplo: SI/NO, Bueno/Malo, Presente/Ausente etc. En la regresión logística se predice una variable cualitativa o categórica, que dificultan su utilización y sus posibilidades, en particular sobre esta definición nos apoya el texto de (López, 2015).

$$P(Y = 1) = \frac{e^z}{1 + e^z}$$

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n)}}$$

5.6.1. Razón, odd y odd ratio

También es importante integrar al utilizar la técnica de regresión logística una adecuada definición e interpretación de los conceptos razón o ratio y/o los odds, podríamos iniciar definiendo de acuerdo al texto de Camarero Rioja que una razón o ratio es el cociente entre dos cantidades y señala cuantas veces una cantidad es mayor o menor respecto a la otra (Camarero Rioja, Almazan Llorente, & Mañas Ramirez, 2018).

El término Odd se refiere a la razón que se establece entre la ocurrencia o su probabilidad de un suceso respecto a su no ocurrencia. El término de Odd se debe interpretar en términos de probabilidad. Para ilustrar un caso de nuestro proyecto en curso podríamos señalar que la probabilidad de encontrar aleatoriamente un asociado que ingresa en default en una agencia de Bajo riesgo es la quinta parte respecto a la de encontrar un asociado que pertenece a las agencias de alto riesgo. Ahora bien cuando se interprete el concepto de Odd Ratio el cual encontraremos abreviadamente identificado con las siglas OR- y el resultado de este se debe leer cuando alcanza el valor 1 quiere decir que no existen diferencias.

5.6.2. R-cuadrado de cox y snell, y r-cuadrado de nagelkerkeb

Al momento de revisión por la adecuación del modelo a los datos podemos utilizar de acuerdo a los fundamentos de regresión logística dos coeficientes que miden de forma similar la asociación entre las variables independientes y dependientes.

La aplicación estadística que utilizaremos para aplicar a nuestra data SPSS-IBM ofrecen distintos coeficientes con este propósito, en este caso, se ha obtenido el de Cox y Snell, este coeficiente toma valores entre 0 y 1 de forma que 0 indicaría un efecto muy bajo de las variables independientes, mientras que en la proximidades de 1 mostraría un efecto

considerable. Sin embargo, este coeficiente no puede llegar a valer 1. Por eso se utiliza el R2 de Nagelkerke, que es el valor del R2 de Cox y Snell estandarizado sobre el valor máximo que éste podría tomar. De esta forma se garantiza que se pueda interpretar su valor entre 0 y 1. El valor obtenido en este caso, los resultados o valores cercanos a 0 señala el "pésimo" ajuste que han tenido nuestros datos (Camarero Rioja, Almazan Llorente, & Mañas Ramirez, 2018).

5.7. Arboles De Decisiones

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no encontraríamos con estadísticos más tradicionales.

Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discriminante de variables continuas. Crean un modelo de clasificación basado en diagramas de flujo.

Clasifican casos en grupos o pronostican valores de una variable dependiente (criterio) basada en valores de variables independientes (predictoras). En el texto consultado Técnicas de Análisis Multivariante de Datos Aplicaciones con SPSS se puede establecer las siguientes ventajas de un árbol de decisión son (Perez López, Técnicas de Análisis Multivariante de Datos Aplicaciones con SPSS., 2004):

- Facilita la interpretación de la decisión adoptada.

- Facilita la comprensión del conocimiento utilizado en la toma de decisiones.
- Explica el comportamiento respecto a una determinada decisión.
- Reduce el número de variables independientes.

5.8. Discriminancia Estadística

Existen numerosos test estadísticos para medir el poder predictivo de la variable dependiente, en este caso para calcular el poder discriminante de la variable en estudio, existen varios métodos dentro de los cuales los más conocidos o frecuentemente utilizados en riesgo de créditos son el test de KS y la curva de ROC .

5.8.1. Test kolmogorov – smirnov.

El test Kolmogorov – Smirnov (KS) es un test de hipótesis que se utiliza para determinar si dos muestras independientes tienen la misma distribución. La finalidad de este test se basa en las diferencias entre las frecuencias relativas acumuladas para los mismos puntos de corte en cada muestra.

En este caso nos interesa comparar la distribución de una variable entre dos muestras: aquella con la que se calibro el modelo y una nueva muestra de datos, con el objetivo de determinar si la variable en estudio ha sufrido cambios significativos en su distribución que pudiesen afectar de manera negativa un modelo construido en base a la primera muestra de datos (Amat Rodrigo, 2020)

5.8.2. Curva de roc (receiver operating characteristic)

Estudia la probabilidad o predicción de otorgar una clasificación negativa de los clientes en un modelo de caracterización de comportamiento, utilizada para localizar el punto de corte donde se maximiza. En los errores que se pueden cometer en la clasificación de los clientes se pueden tipificar Clientes buenos son clasificados como posibles Negativos o Clientes que ingresan en Default son clasificados como posibles positivos.

Figura 2.

Matriz de Confusión para errores en la clasificación de comportamiento.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Nota: Laboratorio e infectolodía (2012).

6. Metodología

A continuación, se detallan las actividades a realizar para la captura, reprocesamiento y procesamiento de la información objeto de estudio. Las actividades se presentan por etapas y fases las cuales son:

- Fase 1. Desarrollo y entendimiento, se busca el conocimiento relevante y los objetivos del usuario final, llamado selección de los datos en la figura 3.

- Fase 2. Creación del conjunto de datos objetivos, seleccionando el subconjunto de variables, llamado en la figura 3 como Datos seleccionados.

- Fase3. Pre-procesado de los datos, elimina el ruido, estrategias para manejar valores ausentes.

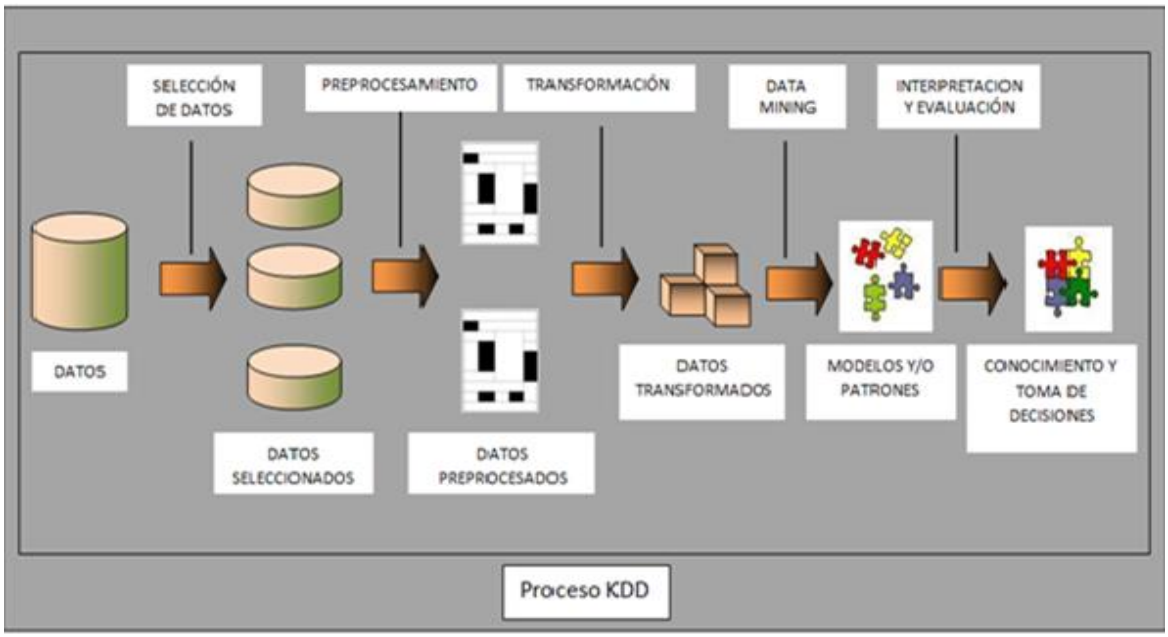
- Fase 4. Transformación y reducción de los datos, incluye la búsqueda de características útiles de los datos según sea el objetivo final.

- Fase 4. Elección de tipo de sistema para los modelos y patrones, esto depende de si el objetivo del proceso de KDD es la clasificación, regresión, agrupamiento.

Se propone una metodología para el descubrimiento de conocimiento en bases de datos. Indica que constituye el primer modelo que define el conocimiento en bases de datos como un proceso compuesto por distintas etapas y fases que van desde la preparación de los datos hasta la interpretación y difusión de los resultados (Moine, 2013). Los pasos principales del proceso iterativo del campo del descubrimiento de conocimiento en Bases de Datos, denominado Knowledge Discovery in Data Bases en inglés y usualmente abreviado KDD las cuales equivalen a (Ver Figura 3)

Figura 3.

Representación de las fases o proceso KDD.



Nota Beltrán Martínez (2014).

6.1. Selección De Variables

El set de datos se construye a partir de las variables socio-demográficas que el área de crédito en conjunto con riesgos de Financiera Comultrasan definen son datos relevantes a capturar de los clientes que se requieren para la elaboración del modelo de scoring. A continuación, se describen las principales variables según el conocimiento del negocio.

- Genero.
- Estado Civil.
- Tipo de Vivienda.
- Edad (métrica).

- Nivel de Escolaridad.
- Estrato.
- Tipo de vivienda.
- Profesión.
- Ciudad de residencia.
- Activos (métrica).
- Pasivos (métrica).
- Ingresos (métrica).
- Desembolso (métrica).
- Valor cuota (métrica).

Todos los datos de las variables necesarias serán suministrados por la Cooperativa Financiera Comultrasan (Financiera Comultrasan, 2020).

6.2. Pre Procesamiento De Datos

El pre procesado de los datos es una etapa importante ya que se basa en la preparación y limpieza de los datos extraídos desde las fuentes de datos en un formato manejable necesario para las fases posteriores. Se considera viable realizar las siguientes etapas para su desarrollo.

- Imputar valores faltantes.
- Identificar y eliminar datos que se pueden considerar ruido.
- Corregir inconsistencias.

6.3. Transformación

En esta etapa se realizará la transformación de las variables para el desarrollo del modelo, se debió recategorizar algunas variables categóricas a numéricas y el cambio de variables numéricas a categóricas según se tenga la idea para el desarrollo del modelo. Incluye el uso de técnicas estadísticas, correlación entre las variables, análisis de componentes principales, esto con el fin de tomar las variables que sean necesarias para la construcción del modelo.

6.4. Aplicación De Técnicas Estadísticas

En esta etapa se realiza el modelamiento para minería de datos, en nuestro caso utilizamos métodos estadísticos como Regresión logística, Árbol de decisión, con el fin de conocer patrones desconocidos potencialmente útiles.

6.5. Interpretación Del Conocimiento

Se identifican los patrones obtenidos, con posibilidad de iterar de nuevo desde el primer paso; la obtención de resultados aceptables dependerá de la definición de las medidas del conocimiento que permitan filtrar de forma automática; existen técnicas de visualización para facilitar la valoración de los resultados o búsqueda manual de conocimiento útil entre los resultados obtenidos.

Muchas veces sobre los pasos que constituyen el proceso de KDD que no están claramente diferenciados, pequeños cambios en una parte pueden afectar fuertemente el resto del proceso; sin quitarle importancia a las fases del proceso KDD, se puede decir que la minería de datos es parte fundamental del proceso y en la que más esfuerzo se realiza (Beltrán Martínez, 2014).

7. Resultados

7.1. Población Y Muestra

Para proceder a realizar el análisis de las variables, se debe considerar que se debe trabajar con información de calidad y bien definida para que al final los resultados sean los más adecuados posibles. Para el análisis de las variables es importante conocer y detallar la base con la cual se trabajará, conocer si presentan algunas incoherencias o anomalías. Por ello es la importancia de realizar un análisis univariado de las variables que se tendrán en cuenta para la construcción del modelo, con el fin de identificarlas a fondo y observar que tan relevantes pueden ser. En esta etapa, como se menciona anteriormente, es necesario aclarar que la base suministrada por la cooperativa Financiera Comultrasan se encuentran en buenas condiciones sin datos faltantes en las variables o valores atípicos, por tanto no fue necesario aplicar técnicas de imputación o eliminación/reemplazo. Se cuenta con un total de 19.689 registros que corresponden a créditos desembolsados desde el año 2016 hasta el año 2019 en las líneas de crédito independiente y montos de créditos aprobados y desembolsados máximos hasta 2 SMMLV por año.

Tabla 1.

Default.

Default	N	Porcentaje
0	17.028	86,48%
1	2.661	13,52%
Total	**19.689**	**100%**

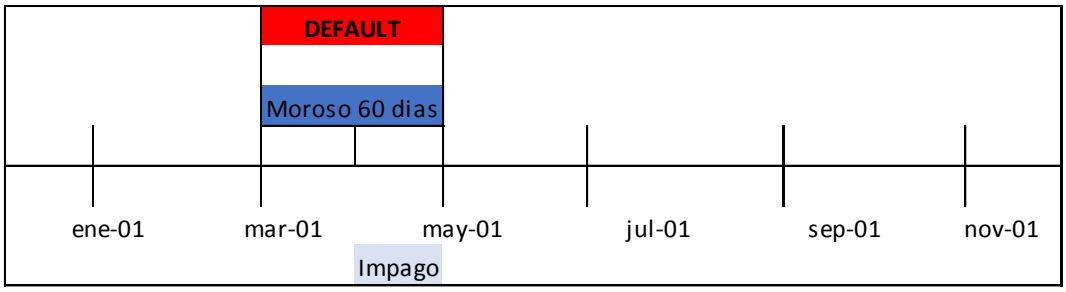
Tabla 1

En la Tabla 1 se observa el total de registros 19.689, de los cuales 17.028 han pagados sus créditos sin entrar en Default lo cual corresponde a un 86,48% y 2.661 han entrado en Default, lo cual corresponde a un 13.52 %. La variable default se determinó con un 0 para las personas que no entraron en default y determinada en 1 para las personas que entraron en default.

DEFAULT. Esta variable indica si un crédito entró en algún momento en mora igual o mayor a 60 días en un horizonte de 12 meses (Figura 4), esto indica si una persona presenta buen hábito de pago. El default está determinado a 60 días para la línea independiente para la cooperativa financiera Comultrasan. El default es nuestra variable objetivo y será explicado a través de las siguientes variables.

Figura 4.

Gráfico Default.



7.2. Descripción De Las Variables Frente Al Default

GÉNERO. La variable describe el género que tienen las personas que tomaron un crédito para nuestra población. En la Tabla 2 podemos observar el comportamiento del género por default.

Tabla 2.

Comportamiento del género por Default.

Genero	% Participación	Default 60	
		0	1
Femenino	55,50%	88,41%	11,59%
Masculino	44,50%	84,08%	15,92%
Total	100%		

Tabla 2

En la Tabla 2, se observa la participación de población por género y por default. Evidenciando que la mayoría de los solicitantes de crédito son las mujeres con participación del 55,50 y una participación de los hombres del 44,50%. Se observa también que las mujeres con respecto a los hombres tienen una menor entrada en default, siendo en 11,59 % para las mujeres y un 15,92 % para los hombres, la diferencia entre el género masculino y femenino no es tan amplio, pero se considera que las mujeres cancelan mejor que los hombres, por tal razón al momento de realizar el modelo de regresión logística en caso de pertenecer un cliente al género "Masculino" se identifica con el Numero 0 y si pertenece al género "Femenino con el Numero 1.

TIPO DE VIVIENDA. La variable describe el tipo de vivienda en el cual viven las personas que tomaron un crédito, para nuestros datos el tipo de vivienda se divide en las siguientes. Alquiler, familiar y propia. En la Tabla 3 podemos observar el comportamiento de default por tipo de vivienda.

Tabla 3.

Comportamiento de default por tipo de vivienda.

Tipo Vivienda	% Participación	Default 60	
		0	1
Alquiler	10,26%	82,43%	17,57%
Familiar	58,80%	84,71%	15,29%
Propia	30,94%	91,20%	8,80%
Total	**100%**		

Tabla 3

En la Tabla 3, se observa que la participación por tipo de vivienda de la población es mayor en tipo de vivienda familiar con un 58,80%, seguido de tipo de vivienda propia con un 30,94 % de participación y por último el tipo de vivienda en alquiler con un 10.26%. Se observa que el mayor porcentaje de entrada en default para este tipo de variable es la vivienda en alquiler con un 17,57% es una variable a tener en cuenta ya que esta misma representa la menor participación en comparación a las demás, las personas con tipo de vivienda en alquiler y vivienda familiar entran más en default que las personas que viven en tipo de vivienda propia.

Por tal razón al momento de analizar unidimensionalmente esta variable frente al default para la elaboración del modelo de regresión logístico se identificará a los clientes que posean esta variable “Vivienda propia” con el número 1 y para los clientes que poseen las otras variables como lo son “Alquiler” y “Familiar” se identificara con el número - 0.

ESTADO CIVIL. La variable describe el estado civil de la población, esta variable se encuentra dividida en los siguientes estados, Casado, Divorciado, Separado, Soltero, Unión Libre, Viudo. En la Tabla 4 podemos observar el comportamiento de default por estado civil.

Tabla 4.

Comportamiento de default por estado civil.

Estado Civil	% Participación	Default 60	
		0	1
Casado	11,35%	95,26%	4,74%
Divorciado	14,40%	86,24%	13,76%
Soltero	48,75%	83,86%	16,14%
Union Libre	23,41%	87,42%	12,58%
Viudo	2,09%	91,26%	8,74%
Total	**100%**		

Tabla 4

En la Tabla 4 se observa que la mayor participación por estado civil es para las personas solteras con un 48,75%, seguido de las personas que se encuentran en estado unión libre con un 23,41%, las personas en estado civil divorciadas con un 14,40%, las personas casadas con un 11,35% y el menor porcentaje de participación son las personas que se encuentran en estado civil viudo con un 2,09%. Se observa que las personas con estado civil soltero con un 16,14 % son las que más caen en default seguido de las personas en estado civil Divorciada con un 13,76%. Las personas que menos caen en default son las que se encuentran en estado civil casado con un 4,74%. Para la elaboración del modelo de regresión logística basados en esta información realizaremos dos segmentaciones de los clientes de acuerdo a esta variable así: los clientes que posean el atributo

"Casado" se identificaran con el número 1, la otra variable a segmentar serán los clientes que posean el atributo "Soltero" los cuales se identificaran en otro campo con el número 1.

ZONA. Esta variable muestra la zona del país en la cual se encuentra la agencia en la cual las personas tomaron el crédito, la cooperativa financiera Comultrasan tiene dividida sus zonas en 7 las cuales son. Boyacá, Bucaramanga y área metropolitana, Centro sur y oriente santandereano, Cesar, Cundinamarca, Magdalena medio, Norte de Santander. En la Tabla 5 observamos el comportamiento de default por zona.

Tabla 5.

Comportamiento de default por zona.

Zona	% Participación	Default 60	
		0	1
Boyaca	8,41%	90,76%	9,24%
Bucaramanga, Area Metropolitana y Zonas Aledañas	25,04%	87,20%	12,80%
Centro, Sur y Oriente Santandereano	10,05%	90,80%	9,20%
Cesar	26,99%	82,43%	17,57%
Cundinamarca	0,30%	94,92%	5,08%
Magdalena Medio	15,28%	88,23%	11,77%
Norte de Santander	13,93%	85,27%	14,73%
Total	**100%**		

Tabla 5

En la Tabla 5 se observa la participación total por zona, César con un 26,99% es la de mayor participación, seguidas de Bucaramanga área metropolitana y zonas aledañas con 25,04%, Magdalena Medio con 15,28%, Norte de Santander con 13,93%, Centro sur y oriente

santandereano con 10,05%, Boyacá con 8,41%, y la de menor participación es la zona de Cundinamarca con un porcentaje de 0,30%. Se observa que las zonas que más entraron en default fueron Cesar con un 17,57% Norte de Santander con un 14.73%, y Bucaramanga área metropolitana y zonas aledañas con un 12,80%. Las zonas con menor tasa de incumplimiento y de menor entrada en default son Boyacá con un 9,24%, la zona centro sur y oriente santandereano con un 9,20%. Cundinamarca también presenta un porcentaje bajo de entrada en default, pero también tiene una muy baja participación.

Para la elaboración del modelo de regresión se tendrá en cuenta para los clientes que pertenezcan a las zonas de "CESAR" y "NORTE DE SANTANDER" Se identificaran con el número - 1 y las otras zonas se identificaran con el numero 0 donde se identifica ausencia de riesgo representativo definido por esta variable.

ESTRATO. Esta variable define el estrato socioeconómico del solicitante del crédito. Los estratos van de 1 a 6, siendo 1 el de menor estrato y 6 el de mayor estrato. En la Tabla 6 observamos el comportamiento de default por tipo de estrato.

Tabla 6.

Comportamiento de default por tipo de estrato.

Estrato	% Participación	Default 60	
		0	1
1	21,79%	86,27%	13,73%
2	48,82%	85,06%	14,94%
3	23,66%	88,41%	11,59%
4	5,35%	91,46%	8,54%
5	0,32%	90,48%	9,52%
6	0,05%	100,00%	0,00%
Total	**100%**		

Tabla 6

En la Tabla 6 observamos la participación total por estrato de la población, las personas de estrato 2 con un 48,82% es la de mayor participación, seguidas de estrato 3 con 23,66%, estrato 1 con 21,79%, estrato 4 con un porcentaje de 5,35%, estrato 5 con 0,32% y la de menor porcentaje es para las personas de estrato 6 con 0,05%. La relación de la Tabla 6 tiene coherencia con la población que estamos trabajando, teniendo en cuenta que son créditos otorgados para personas independientes con montos máximos hasta 2 SMMLV. Y en su gran mayoría estas personas son microempresarios y en algunos de los casos son créditos por primera vez.

Se observa también que las personas de estrato 2 son las que mayor entran en default 14,94% seguido de los estratos 1 y 3 con 13,73% y 11,59% respectivamente. Las personas de estrato 5 y 4 entran en un porcentaje de incumplimiento menor y las personas de estrato 6 no entraron en incumplimiento. Teniendo en cuenta que por esta variable no se logra determinar un comportamiento marcado, solo para el Estrato 6 pero no es significativo, por lo cual la misma no se tendrá en cuenta para la elaboración del modelo de regresión.

DESEMBOLSO.

Esta variable al ser cuantitativa se procede a realizar un análisis descriptivo, con el fin de conocer su comportamiento.

Figura 5.

Resumen de procesamiento de casos variable Valor de desembolso.

Resumen de procesamiento de casos

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
VALOR_DESEMBOLSO	19689	100,0%	0	0,0%	19689	100,0%

Figura 6.

Descriptivos Variable valor de desembolso.

Descriptivos

			Estadístico	Desv. Error
VALOR_DESEMBOLSO	Media		1001720	1841
	95% de intervalo de confianza para la Media	Límite inferior	998110	
		Límite superior	1005329	
	Media recortada al 5%		999699	
	Mediana		1000000	
	Varianza		66765525541	
	Desv. Desviación		258390	

Figura 7.

Histograma variable valor de desembolso.

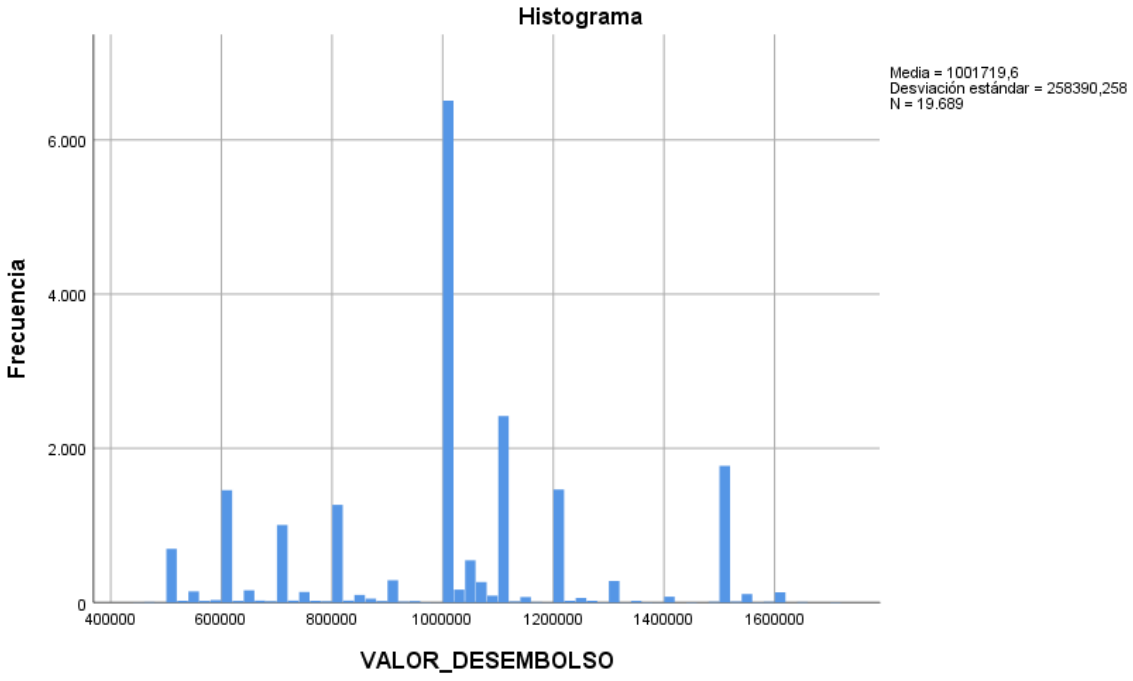


Figura 8.

Pruebas de normalidad variable valor de desembolso.

Pruebas de normalidad			
	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
VALOR_DESEMBOLSO	0,214	19689	0,000
a. Corrección de significación de Lilliefors			

Nota Elaboración propia en herramienta SPSS.

La base de datos consta de 19.689 registros; de la prueba de Kolgomorov- Smirnov se puede confirmar que los datos no provienen de una distribución normal. La media del valor de desembolso es de $1.001.719, con una desviación estándar de 258.390, esta variable presenta una alta variación en sus datos y su mediana en es de $1.000.000

La variable valor desembolso indica el monto por el cual fue aprobado el crédito. En la Tabla 7 se muestra el cruce entre la variable monto desembolso y el default.

Tabla 7.

Cruce entre el variable monto desembolso y el default.

Desembolso	% Participación	Default 60	
		0	1
< 1 Smmlv	26,64%	82,18%	17,82%
(De 1 Smmlv a 1.5 Smmlv)	62,37%	87,43%	12,57%
> 1.5 Smmlv	10,99%	91,54%	8,46%
Total	**100%**		

Tabla 7

En la anterior Tabla, se observa que la mayor participación la tienen los créditos desembolsados entre 1 SMMLV y 1.5 SMMLV con un 62,37 %, seguido de los créditos inferiores a 1 SMMLV con un 26,64%; además, la mayor entrada en incumplimiento son los créditos inferiores a 1 SMMLV con 17,82% y los de menor entrada en incumplimiento son los superiores a 1.5 SMMLV

con un 8,46%. Para la elaboración del modelo se creara una variable ordinal donde 1 identificara < 1 SMMLV, 2 (De 1 SMMLV a 1.5 SMMLV) y 3 > 1.5 SMMLV.

VALOR DE CUOTA. Para la variable rango cuota se realizó el siguiente análisis descriptivo.

Figura 9.

Resumen de procesamiento de casos variable valor de cuota.

Resumen de procesamiento de casos

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
VALOR_CUOTA	19689	100,0%	0	0,0%	19689	100,0%

Nota Elaboración propia en herramienta SPSS.

Figura 10.

Descriptivos variable valor de cuota.

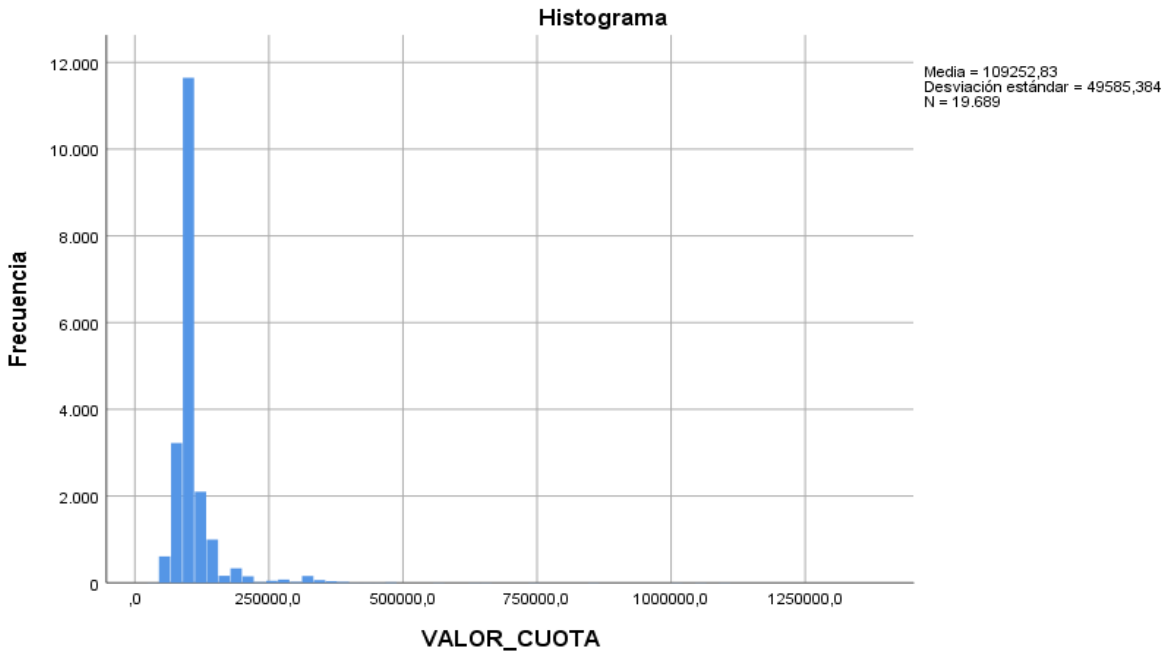
Descriptivos

			Estadístico	Desv. Error
VALOR_CUOTA	Media		109253	353
	95% de intervalo de confianza para la	Límite inferior	108560	
		Límite superior	109945	
	Media recortada al 5%		103095	
	Mediana		100756	
	Varianza		2458710341	
	Desv. Desviación		49585	

Fuente: Elaboración propia en herramienta SPSS.

Figura 11.

Histograma variable valor de cuota.



Nota: Elaboración propia en herramienta SPSS.

Figura 12.

Pruebas de normalidad Kolmogorov-Smirnov variable valor de cuota.

Pruebas de normalidad			
	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
VALOR_CUOTA	0,272	19689	0,000
a. Corrección de significación de Lilliefors			

Nota: Elaboración propia en herramienta SPSS.

Hay un total de registros de 19.689, el valor del estadístico de prueba de 0.272, por consiguiente, a un nivel de significancia de 0.05 se puede afirmar que los datos no provienen de una distribución normal. La media del valor de cuota es de $109.252, con una desviación estándar de 49.585. Se

observa en la (Figura 11) que para la variable valor de cuota hay asimetría a la derecha, esto afirma que los datos no tienen una distribución normal.

Esta variable indica el rango del valor de cuota que cancelan las personas que desembolsan los créditos, en la Tabla 8 se observa el comportamiento de default por el rango de cuota.

Tabla 8.

Comportamiento de default por el rango de cuota.

Cuota	% Participación	Default 60	
		0	1
< 0,11 Smmlv	26,49%	82,05%	17,95%
(De 0,11 Smmlv a 0,16 Smmlv)	63,71%	87,42%	12,58%
> 0,16 Smmlv	9,80%	92,38%	7,62%
Total	**100%**		

Tabla 8

En la Tabla 8 observamos el comportamiento de los rangos por valor de cuota, el porcentaje de participación más alto son los siguientes de 0,11 SMMLV a 0,16 SMMLV con un 63,71% seguido de los inferiores a 0,11 SMMLV con un 26,41%, el de menor participación son los mayores a 0,16 SMMLV con un 9,80%. Se observa que los que mayor incurre en incumplimiento son los inferiores a 0,11 SMMLV y los que menor incurre son los superiores a 0,16 SMMLV. A un mayor valor de cuota mejor cancelan. Para la elaboración del modelo se crea una variable ordinal donde 1 serán < 0,11 SMMLV, 2 (De 0,11 SMMLV a 0,16 SMMLV), y 3 > 0,16 SMMLV.

PLAZO. Esta variable indica el plazo por el cual fue desembolsado el crédito, en la Tabla 9 se observa el comportamiento de default por plazo.

Tabla 9.

Comportamiento de default por plazo.

Plazo	% Participación	Default 60	
		0	1
< 6 meses	7,43%	92,82%	7,18%
(6 a 12 meses)	77,84%	85,37%	14,63%
(13 a 18 meses)	14,07%	89,21%	10,79%
(19 a 24 meses)	0,64%	88,10%	11,90%
> a 24 meses	0,02%	100,00%	0,00%
Total	**100%**		

Tabla 9

En la Tabla 9 observamos la participación por plazo, los créditos desembolsados entre 6 a 12 meses tiene un porcentaje de 77,84%, seguido del rango entre 13 a 18 meses con un porcentaje de 14,07%, los créditos desembolsados para pagar antes de 6 meses tienen una participación de 7,43%, de 19 a 25 meses una participación de 0,64% y por último los créditos mayores a 24 meses con una participación de 0,02%. Para la elaboración del modelo se creará una variable ordinal codificada numéricamente así: 1 < 6 meses, 2 (De 6 a 12 meses), 3 (13 a 18 meses), 4 (19 a 24 meses) 5 > a 24 meses.

NIVEL EDUCATIVO. Esta variable indica el nivel educativo de la población al momento de tomar el crédito, los niveles son los siguientes, personas que no tienen ningún tipo de estudio "ninguno", primaria, secundaria, técnico, universidad y posgrado, en la Tabla 10 se observa el comportamiento de default por nivel educativo.

Tabla 10.

Comportamiento de default por nivel educativo.

Nivel Educativo	% Participación	Default 60	
		0	1
NINGUNA	1,16%	89,08%	10,92%
POSGRADO	0,03%	100,00%	0,00%
PRIMARIA	32,67%	87,64%	12,36%
SECUNDARIA	56,28%	85,41%	14,59%
TECNICO	7,13%	87,46%	12,54%
UNIVERSIDAD	2,74%	91,09%	8,91%
Total	**100%**		

Tabla 10

En la Tabla 10 se observa la participación del nivel educativo en la población, el 56,28% tienen un nivel educativo en secundaria, seguido de primaria con un 32,67%, técnico con 7,13%, universidad 2,74%, ningún tipo de escolaridad con 1,16% y por último nivel educativo posgrado con un 0,03%.

Nivel educativo con mayor entrada en default, secundaria con 14,59% técnico 12,54% primaria 12,36%, causa curiosidad que el nivel educativo ninguno con una participación tan baja tiene una entrada en default de 10,92%. Los niveles educativos con menor entrada en default universidad y posgrado. Para la variable nivel educativo se define no tener en cuenta para la elaboración del modelo de regresión dado que sus datos son muy similares y no definen un comportamiento marcado en los clientes que ingresaron al default

AGENCIA. Esta variable indica la agencia en la cual fue desembolsado el crédito, la cooperativa tiene participación en 6 departamentos como lo son Santander, Boyacá, Cesar,

Atlántico, Norte de Santander y Cundinamarca. En cada uno de los departamentos nombrados hay presencia de agencias.

En la Tabla 11 se observa el comportamiento de default por agencia.

Tabla 11.

Comportamiento de default por agencia.

Agencia	% Participación	Default 60	
		0	1
Bosconia	3,82%	70,35%	29,65%
San Alberto	2,53%	74,70%	25,30%
Codazzi	2,55%	77,09%	22,91%
Calle 35	0,82%	79,50%	20,50%
Centenario	1,04%	80,98%	19,02%
Cañaveral	0,58%	81,74%	18,26%
Curumaní	2,29%	82,26%	17,74%
San Francisco	1,51%	82,55%	17,45%
San Rafael	1,21%	83,19%	16,81%
Cúcuta	7,12%	83,87%	16,13%
Puerto Wilches	1,61%	83,91%	16,09%
Duitama	1,52%	84,00%	16,00%
Barrancabermeja (Comercio)	1,59%	84,98%	15,02%
Barrancabermeja (Nororiente)	2,85%	85,03%	14,97%

Floridablanca	2,22%	85,13%	14,87%
Plaza Satelite	1,74%	85,13%	14,87%
Aguachica	5,50%	85,30%	14,70%
Valledupar	3,73%	85,83%	14,17%
Lebrija	1,99%	85,93%	14,07%
San Gil	2,35%	85,96%	14,04%
La Cumbre	2,50%	86,38%	13,62%
Ocaña	3,53%	86,64%	13,36%
Carrera 11	2,06%	86,67%	13,33%
Pamplona	3,28%	86,82%	13,18%
Cimitarra	1,05%	86,89%	13,11%
Tunja Norte	1,05%	88,41%	11,59%
El Playón	1,51%	89,23%	10,77%
San Vicente de Chucurí	1,98%	89,23%	10,77%
Kennedy	2,17%	89,25%	10,75%
Málaga	2,32%	89,28%	10,72%
Poblado	1,85%	89,32%	10,68%
Barrancabermeja (Torcoroma)	0,86%	89,41%	10,59%
San Martín	2,76%	89,50%	10,50%
Socorro	1,48%	89,73%	10,27%
Corresponsal Gamarra	0,78%	90,20%	9,80%
Pelaya	3,05%	90,83%	9,17%
Sogamoso	2,30%	91,83%	8,17%
Rionegro	1,75%	91,86%	8,14%
Piedecuesta	3,28%	92,40%	7,60%
Sabana De Torres	2,56%	93,27%	6,73%
Tunja	3,18%	93,29%	6,71%
Corresponsal Madrid	0,24%	93,75%	6,25%
Vélez	1,39%	93,77%	6,23%
Barbosa	2,41%	95,57%	4,43%
Zapatoca	0,64%	96,03%	3,97%
El Carmen de Chucurí	0,92%	96,70%	3,30%
Corresponsal Paipa	0,35%	97,10%	2,90%
Cabecera	0,03%	100,00%	0,00%
Chia	0,03%	100,00%	0,00%
Chiquinquira	0,05%	100,00%	0,00%
Corresponsal Funza	0,03%	100,00%	0,00%
Puente Nacional	0,05%	100,00%	0,00%
Total	**100%**		

Tabla 11

En la Tabla 11 observamos que las agencias que tienen mayor participación Cúcuta 7,12%, Aguachica 5,50%, Bosconia 3,82%, Valledupar 3,73%, Ocaña 3,53%, Pamplona 3,28%,

Piedecuesta 3,28%, Tunja y Pelaya 3,18% y 3,05% respectivamente. Las agencias que menor participación tienen Corresponsal Funza, Cabecera, Chía cada una con 0,03%, Puente Nacional y Chiquinquirá 0,05% cada una, Corresponsal Madrid 0,24%, Corresponsal Paipa, 0,35%, Cañaveral 0,58%, Zapatoca 0,64%, Corresponsal Gamarra 0,78%, Calle 35 0,82%, Torcoroma 0,86%, El Carmen del chucuri 0,92%.

Las oficinas con mayor entrada en default son las siguientes, Bosconia, San Alberto, Codazzi, Calle 35, Centenario, Cañaveral, Curumani, San Francisco, San Rafael, Cúcuta, Puerto Wilches, Duitama, Barranca Centro. Todas con un porcentaje por encima del 15 %.

Las oficinas con menor porcentaje de clientes en default son las siguientes. Corresponsal Funza, Cabecera, Chía, Puente Nacional, Chiquinquirá, Corresponsal Paipa, El Carmen de chucuri, Zapatoca, Barbosa, estas oficinas tienen un porcentaje de incumplimiento por debajo del 5%.

Para la elaboración del modelo de regresión, para la variable agencia se utilizaran nuevas variables dicotómicas de tipo dummie; para la primera segmentación se seleccionaran las agencias con bajo nivel de riesgo de entrar en default de acuerdo a su participación dentro de la totalidad de las agencias, se seleccionaron las que se encuentran con un indicador por debajo del 8,15% de clientes que ingresaron al default las cuales serían: Puente Nacional, Cabecera, El Carmen de Chucuri, Zapatoca, Barbosa, Velez, Tunja, Sabana de Torres, Piedecuesta, Rionegro; así mismo la otra variable a segmentar dentro de las agencias serán las agencias que presentan un riesgo alto por el comportamiento de clientes que ingresaron al default la cual se definirá por la que posean indicadores superiores a 16,10% dentro de las cuales se encontraran las agencias de: Bosconia, San Alberto, Codazzi, Calle 35, Centenario, Cañaveral, Curumani, San Francisco, San Rafael, Cúcuta.

EDAD. La variable categórica tiene 4 modalidades, a saber está dividida por rangos de edad, la población se encuentra entre los 18 a mayores a 50 años de edad. Esta variable corresponde a la edad de la población al momento del desembolso del crédito.

En la Tabla 12, podemos observar el comportamiento de default por rango de edad.

Tabla 12.

Comportamiento de default por rango de edad.

EDAD	% Participación	Default 60	
		0	1
(De 18 a 30 años)	37,41%	82,20%	17,80%
(De 31 a 40 años)	22,04%	86,45%	13,55%
(De 41 a 50 años)	19,25%	89,74%	10,26%
> 50 Años	21,30%	91,10%	8,90%
Total	**100%**		

Tabla 12

La distribución de los clientes por edad es la siguiente, rangos de edad entre 18 a 30 años, con una participación de 37,41%, seguido del rango entre 31 a 40 años, con una participación de 22,04%, los mayores de 50 años con un 21,30% y por último el rango entre 41 a 50 años con una participación del 19,25%, así que el mayor riesgo son clientes entre 18 y 30 años (37.41%) siendo una población muy joven. También, se observa que a medida que se incrementa la edad menor es el incumplimiento en los préstamos.

ACTIVOS. Para la variable rango activo se realizó el siguiente análisis descriptivo.

Figura 13.

Resumen de procesamiento de casos Variable Rango Activos.

Resumen de procesamiento de casos

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
TOT_ACTIVO	19689	100,0%	0	0,0%	19689	100,0%

Nota Elaboración propia en herramienta SPSS.

Figura 14.

Descriptivos Variable Total Activos.

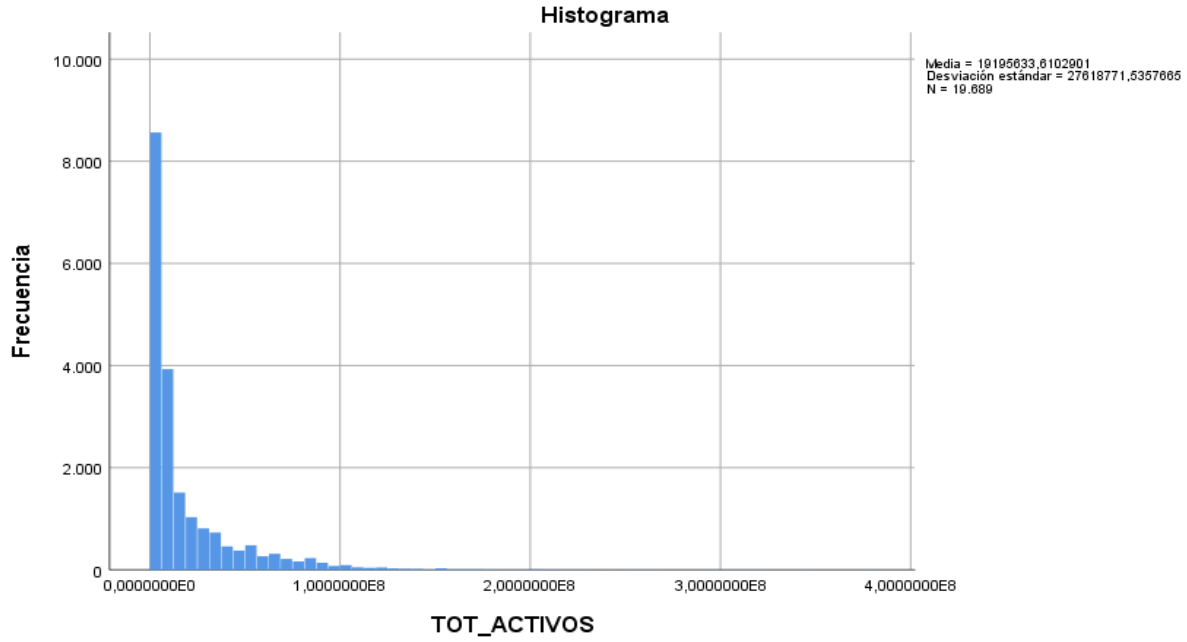
Descriptivos

			Estadístico	Desv. Error
TOT_ACTIVOS	Media		19195634	196831
	95% de intervalo de confianza para la Media	Límite inferior	18809829	
		Límite superior	19581438	
	Media recortada al 5%		15272742	
	Mediana		7550000	
	Varianza		762796541144867	
	Desv. Desviación		27618772	

Nota Elaboración propia en herramienta SPSS.

Figura 15.

Histograma variable Rango de activos.



Nota Elaboración propia en herramienta SPSS.

Figura 16.

Pruebas de normalidad Rango de activos.

Pruebas de normalidad

	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
TOT_ACTIVOS	0,244	19689	0,000

a. Corrección de significación de Lilliefors

Nota Elaboración propia en herramienta SPSS.

En la (Figura 14) observamos un análisis descriptivo para la variable total activos, donde la media de los datos es de $19.195.633 y una desviación estándar de $27.618.771. En la (Figura 15) se evidencia una asimetría a la derecha de los datos.

Esta variable muestra el valor de activos de la población al momento del desembolso del crédito. En la Tabla 13 se observa el comportamiento de default por rango de activos.

Tabla 13.

Comportamiento de default por rango de activos.

ACTIVOS	% Participación	Default 60	
		0	1
< 5 Smmlv	31,24%	81,92%	18,08%
(De 5 Smmlv a 10 Smmlv)	23,27%	84,61%	15,39%
(De 10 Smmlv a 15 Smmlv)	10,01%	86,20%	13,80%
(De 15 Smmlv a 20 Smmlv)	5,59%	88,18%	11,82%
> 20 Smmlv	29,89%	92,49%	7,51%
Total	**100%**		

Tabla 13

En la Tabla 13 observamos una participación alta en los activos inferiores a 5 SMMLV con un 31,24% seguido de los superiores a 20 SMMLV con un 29,89% de participación, los de menor participación son los de 15 SMMLV hasta 20 SMMLV con una participación del 5,59%, se observa que a un mayor valor de activos menor entrada en incumplimiento y a un menor valor en los activos una mayor entrada en incumplimiento. Para la elaboración del modelo se creara una variable ordinal. 1 activos < a 5 SMMLV, 2 (De 5 SMMLV a 10 SMMLV), 3 (De 10 SMMLV a 15 SMMLV), 4 (15 SMMLV a 20 SMMLV), 5 (> 20 SMMLV)

DEPARTAMENTO DE RESIDENCIA. Esta variable indica el lugar de residencia de la población en la cual vivían cuando tomaron el crédito, en algunos de los lugares de residencia de la población no hay presencia de las agencias que dispone la cooperativa como lo es Antioquia y Bolívar. En la tabla 14 se muestra el comportamiento de default por departamento de residencia.

Tabla 14.

Comportamiento de default por departamento de residencia.

Departamento Residencia	% Participación	Default 60	
		0	1
Santander	47,81%	88,21%	11,79%
Cesar	25,70%	82,33%	17,67%
Norte de Santander	15,27%	85,16%	14,84%
Boyaca	9,53%	90,99%	9,01%
Bolivar	1,29%	85,43%	14,57%
Cundinamarca	0,31%	95,08%	4,92%
Antioquia	0,09%	88,89%	11,11%
Total	**100%**		

Tabla 14

En la Tabla 14 se observa que la gran parte de la población vivía al momento del desembolso del crédito en el departamento de Santander con una participación de 47,81%, seguido del departamento del Cesar con un 25,70%, Norte de Santander tiene una participación del 15,27%, Boyacá 9,53%, Bolívar 1,29% Cundinamarca y Antioquia con 0,31% y 0,09% respectivamente.

Los departamentos que más entran en incumplimiento son Cesar con un 17,67%, seguido de Norte de Santander con un 14,84%, Bolívar que es un departamento donde no hay presencia de la cooperativa con una entrada en default de 14,57%. Los departamentos que como Boyacá y Cundinamarca no entran en gran cantidad a default. Esta variable no será tenida en cuenta en la elaboración del modelo.

INGRESO. Para la variable rango ingreso se realizó el siguiente análisis descriptivo.

Figura 17.

Resumen de procesamiento de casos Variable Ingresos.

Resumen de procesamiento de casos						
	Válido		Casos Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
TOTAL INGRESO	19689	100,0%	0	0,0%	19689	100,0%

Nota Elaboración propia en herramienta SPSS.

Figura 18.

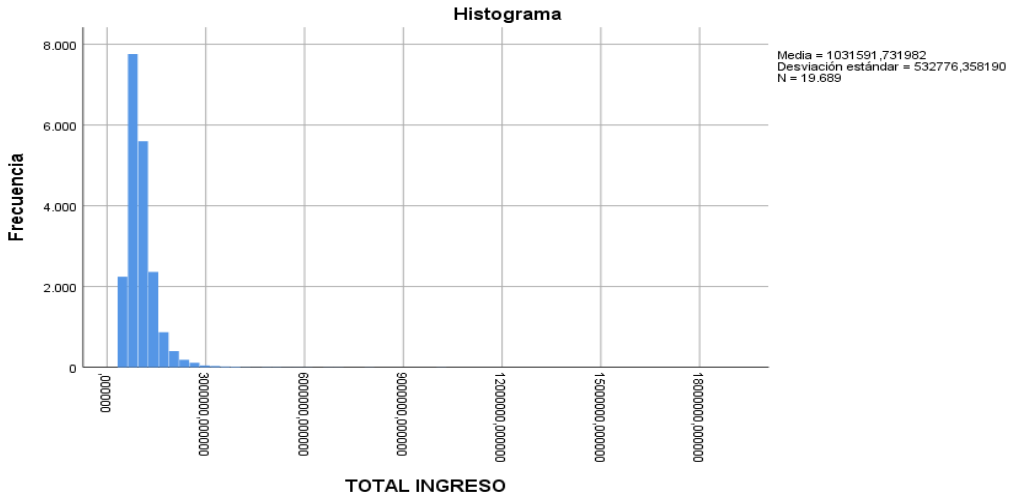
Descriptivos Variable Ingresos.

Descriptivos				
			Estadístico	Desv. Error
TOTAL INGRESO	Media		1031592	3797
	95% de intervalo de confianza para la	Límite inferior	1024149	
		Límite superior	1039034	
	Media recortada al 5%		978726	
	Mediana		930000	
	Varianza		283850647846	
	Desv. Desviación		532776	

Nota Elaboración propia en herramienta SPSS.

Figura 19.

Histograma Variable Ingresos.



Nota Elaboración propia en herramienta SPSS.

Figura 20.

Pruebas de normalidad Variable Ingresos.

Pruebas de normalidad

	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
TOTAL INGRESO	0,153	19689	0,000

a. Corrección de significación de Lilliefors

Nota Elaboración propia en herramienta SPSS.

Se puede concluir que no hay valores perdidos, hay un total de registros de 19.689, se puede confirmar que los datos no provienen de una distribución normal. La media del valor de cuota es de $19.195.633, con una desviación estándar de $27.618.771, esta variable muestra alta variación en sus datos y en la (Figura 19) observamos también asimetría a la derecha.

Esta variable muestra los ingresos de la población al momento del desembolso del crédito. En la Tabla 15 se observa el comportamiento de default por ingresos.

Tabla 15.

Comportamiento de default por ingresos.

Ingresos	% Participación	Default 60	
		0	1
< 1.2 SMMLV	74,18%	85,89%	14,11%
> 1.2 SMMLV	25,82%	88,20%	11,80%
Total	**100%**		

Tabla 15

En la Tabla 15 observamos una alta participación en ingresos inferiores a 1.2 SMMLV con un 74.18% y los ingresos superiores a 1.2 SMMLV tienen una participación de 25,82%. Se observa que los ingresos inferiores a 1.2 SMMLV entran más en incumplimiento. Para la elaboración del modelo se crea una variable ordinal donde, 1 (< 1.2 SMMLV), 2 (>1.2 SMMLV).

PASIVOS. Esta variable indica los pasivos que presentaba la población al momento del desembolso del crédito, en la Tabla 16 se observa el comportamiento de default por pasivos.

Tabla 16.

Comportamiento de default por pasivos.

PASIVOS	% Participación	Default 60	
		0	1
0	70,66%	86,79%	13,21%
> $5.000.000	23,41%	84,66%	15,34%
De $5.000.001 a $10.000.000	3,81%	88,68%	11,32%
De $10.000.001 a $15.000.000	1,17%	91,30%	8,70%
De $15.000.001 a $20.000.000	0,49%	93,81%	6,19%
De $20.000.001 a $25.000.000	0,18%	91,43%	8,57%
De $25.000.001 a $30.000.000	0,15%	93,33%	6,67%
< $30.000.000	0,12%	100,00%	0,00%
Total	**100%**		

Tabla 16

En la Tabla 16 observamos que el 70,66% de la población no presentaba pasivos al momento del desembolso y que esta misma población presenta un nivel de incumplimiento de 13,21%, la población con pasivos inferiores hasta $5.000.000 presentan una participación de 23,41% y presenta el porcentaje de incumplimiento más alto de la población con un 15,34%. La población con pasivos entre $5.000.001 hasta $10.000.000 presenta una participación de 3,81% con un porcentaje de incumplimiento de 11,32%.

No se considera una variable viable a tener en cuenta en el modelo ya que hay más del 70 % de la población con pasivos cero

8. Análisis Multivariado

Para proceder con el desarrollo del análisis multivariado y elaboración del modelo, se realiza una matriz de correlación para determinar qué relación hay entre las variables numéricas seleccionadas para el modelo.

8.1. Matriz De Correlaciones

Tabla 17.

Matriz de correlaciones variables.

Correlaciones					
	EDAD	**TOT_ACTIVOS**	**TOTAL INGRESO**	**VALOR_DESEMBOLSO**	**VALOR_CUOTA**
EDAD	1	,372**	,113**	,040**	,032**
TOT_ACTIVOS	,372**	1	,265**	,122**	,128**
TOTAL INGRESO	,113**	,265**	1	,188**	,183**
VALOR_DESEMBOLSO	,040**	,122**	,188**	1	,319**
VALOR_CUOTA	,032**	,128**	,183**	,319**	1
TOTAL	19689	19689	19689	19689	19689

**. La correlación es significativa en el nivel 0,01 (bilateral).

Fuente: Elaboración propia en herramienta SPSS.

Para las variables seleccionadas se observa que no hay correlaciones importantes entre ellas ni tampoco son significativas a un nivel de significancia del 5 %

8.2. Árbol De Clasificación

Se procede hacer un análisis mediante Árbol de CLASIFICACIÓN para conocer cuales variables son las que tienen una mayor influencia, como variable dependiente se encuentra el default y como variables explicativas se encuentran. ZONA_C_NS, AGENCIA_AR, AGENCIA_BR, GENERO FEMENINO, ESTADO CIVIL SOLTERO, CASADO, VIVIENDA PROPIA, EDAD_R, ACTIVOS_R, INGRESOS_R, DESEMBOLSO_R, CUOTA, PLAZO. (ANEXO 1)

La Tabla de resumen del modelo ilustrado en la Figura 21, proporciona información general sobre las especificaciones utilizadas para crear ambos diseños, tanto como el modelo original sobre el modelo resultante. La sección Especificaciones, ofrece información sobre los valores de configuración utilizados con el fin de generar el modelo de árbol, teniendo en cuenta las variables utilizadas implementadas en el análisis, de igual forma la sección Resultados muestra información sobre el número de nodos totales y terminales, la profundidad del árbol (número de niveles por debajo del nodo raíz) y las variables independientes incluidas en el modelo final

Figura 21.

Resumen del modelo árbol de clasificación.

Resumen del modelo		
Especificaciones	Método de crecimiento	EXHAUSTIVE CHAID
	Variable dependiente	DEFAULT_60
	Variables independientes	ZONA_C_NS, PLAZO, AGENCIA_AR, AGENCIA_BR, FEMENINO, SOLTERO, CASADO, VIVIENDA_PROPIA, EDAD_R, ACTIVOS_R, INGRESOS_R,
	Validación	Ninguna
	Máxima profundidad del árbol	3
	Casos mínimos en nodo padre	100
	Casos mínimos en nodo hijo	50
Resultados	Variables independientes incluidas	ACTIVOS_R, AGENCIA_AR, DESEMBOLO_R, AGENCIA_BR, FEMENINO, EDAD_R,
	Número de nodos	25
	Número de nodos terminales	15
	Profundidad	3

Nota Elaboración propia en herramienta SPSS.

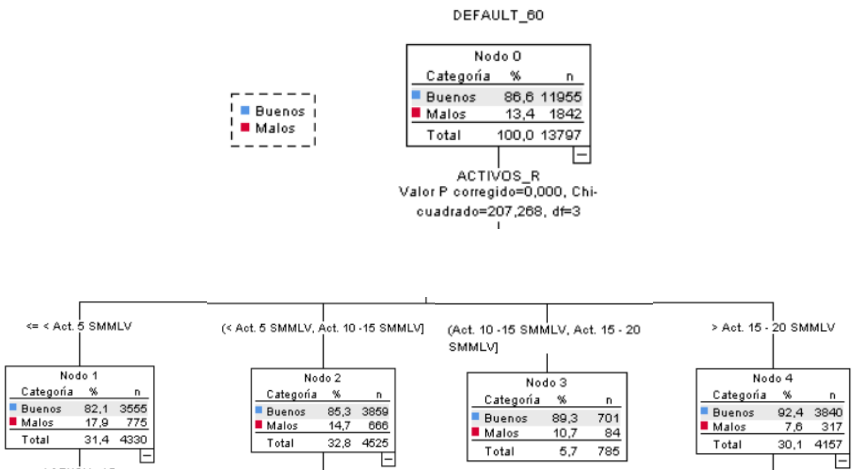
El Diagrama de árbol obtenido es la representación gráfica del resultado. En el anexo 1 "árbol de decisión" todas las variables son tratadas como nominales y cada nodo contiene una tabla de frecuencias que muestra el número de casos (frecuencia y porcentaje) para cada categoría de la variable dependiente DEFAULT_60. También incluye el gráfico de frecuencias.

La categoría “pronosticada”, que es la categoría con el mayor valor de frecuencia en cada nodo, aparece resaltada con una franja gris.

Con el fin de perfilar los distintos niveles de riesgo en la nueva segmentación de las variables seleccionadas que más se ajustan al modelo en construcción, se utiliza la técnica de árboles de decisión, donde el nodo matriz seria nuestra variable dependiente DEFAULT_60, mientras que los nodos terminales serían la probabilidad del score final encontrado.

Figura 22.

Árbol de clasificación Rango de activos.



Nota Elaboración propia en herramienta SPSS.

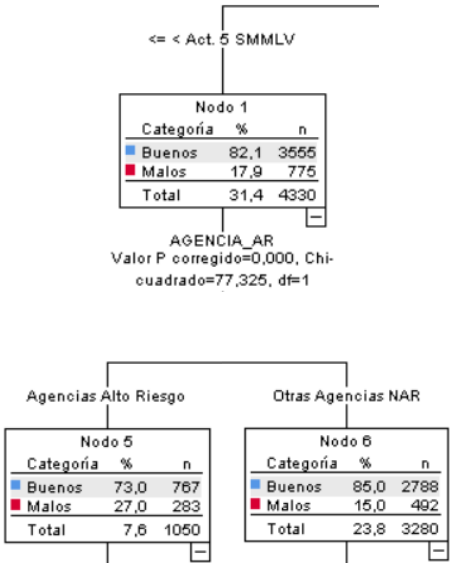
En primer lugar, nos fijamos en el nodo 0 que describe la variable dependiente: porcentaje de los clientes que ingresan en DEFAULT_60 “Malos” que equivalen al 13,4% representados en 1942 clientes y de los que no ingresan “Buenos” que equivalen al 86,6% representados en 11955 clientes.

La técnica de árbol de decisión indica que la principal variable ó “Variable predictora” porque predice el comportamiento frente a nuestra variable dependiente son los Rangos de los activos,

observamos que la variable dependiente se ramifica en cuatro nodos: Nodo 1 – 2 – 3 – 4 pertenecientes a la variable Rango de activos, indicando que ésta es la variable principal predictora, donde los clientes que poseen el atributo Activos < 5SMMLV ubicados en el Nodo 1 son los más relevantes dado que son los que entran en default y corresponden al 17,9%.

Figura 23.

Árbol de clasificación Agencias de alto riesgo.

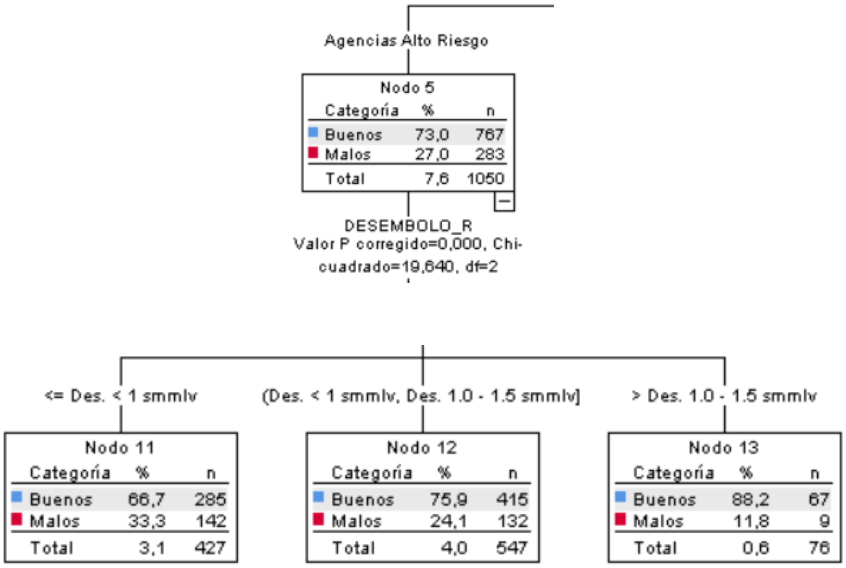


Nota Elaboración propia en herramienta SPSS.

Siguiendo la ruta del árbol dentro de la participación de los Activos menores a 5 SMMLV encontramos otra variable que define el comportamiento de los clientes que ingresan a default dentro del cual se relaciona la variable Agencias de Alto Riesgo representada con el nodo 5, el cual asciende al 27% de los clientes que ingresaron al default.

Figura 24.

Árbol de clasificación Rango de desembolso.



Nota Elaboración propia en herramienta SPSS.

Siguiendo el árbol dentro del 27% de cliente seleccionados con estos atributos, otra variable significativa que permite definir el comportamiento de los clientes que ingresan a default son los rangos de los desembolsos de crédito, donde se establece que existe una mayor probabilidad de entrar en default para las operaciones de crédito con desembolsos menores a 1 SMMLV que se ubican en el Nodo 11 con un 33,3% de los clientes que ingresaron.

Por tanto, a modo resumen, los nodos que definen el perfil de los clientes que ingresan en Default (variables que influyen en el ingreso en Default) son: Nodo 0 -Nodo 1 - Nodo 5 - Nodo 11 – Nodo 16 – Nodo 18 y Nodo 22 Es decir, influyen las siguientes variables: Rangos de Activos <5SMMLV, Agencias de Alto Riesgo, Rango de desembolso <1SMMLV, Variable Casados, Rango de edad 18 – 30 años.

Analizado el árbol ilustrado en el Anexo 1 se puede concluir las siguientes: interpretaciones de los datos:

- La variable Rango de Activos es el mejor predictor para el conocer el ingreso de clientes en default, con cuatro categorías o rangos: Activos <5SMMLV, Activos entre >5SMMLV y 10-15SMMLV, Activos entre 10-15 SMMLV y 15-20 SMMLV, Activos >15-20 SMMLV.

- La probabilidad más alta de ingresar en default (17.9%) se da entre los clientes que poseen activos inferiores a 5SMMLV, que pertenecen a agencias catalogadas de alto riesgo y su valor de desembolso es menor a 1SMMLV.

- La probabilidad más baja de ingresar en default (7.6%) se da entre los clientes que poseen activos mayores a 15-20 SMMLV, que no se encuentran incluidos en agencias de alto riesgo, Si sus rangos de edades se encuentran entre >31-40 años la probabilidad disminuye 5,8%.

- Entre los clientes que poseen activos entre >5SMMLV y 10-15SMMLV que registran una probabilidad de (14,7%) de ingresar en default, la probabilidad aumenta cuando los clientes pertenecen a agencias catalogadas de alto riesgo en un (21,5%), así mismo presenta un incremento en la posibilidad de ingresar en default cuando pertenece a este tipo de agencias y registra estado civil diferente a casado en un (22,6%).

9. Modelo De Regresión Logística Binaria

La regresión logística tienen como objetivo comprobar hipótesis o relaciones causales cuando la variable dependiente es nominal, en el proyecto se busca determinar las variables predictoras que tienen mayor pesos en el pronóstico de clientes que pertenecen al grupo default 1, es decir aportar al modelo o scoring de aprobación de créditos de bajo monto.

Se realiza un modelo de regresión logística, técnica que muestra el comportamiento de una población, para la cual utilizaremos una muestra de entrenamiento que será el 70% de los datos y una muestra de comprobación o prueba con el restante 30% de los datos, con el fin de validar que utilizando un menor número de datos nuestro modelo tiene similar comportamiento; recordamos que nuestra variable dependiente DEFAULT es binaria y tiene dos opciones, 1 caer en default, 0 no caer en default razón por la cual estamos aplicando esta regresión logística de tipo binario.

Se procede a realizar el modelo con el total de variables (13 variables) seleccionadas las cuales son: Zona Norte Santander y Cesar, Agencias Alto Riesgo, Agencias Bajo Riesgo, Femenino, Soltero, Casado, Vivienda Propia, Edad, activos, Ingresos, Valor de desembolso, cuota, plazo.

9.1. Modelos

Para la creación del modelo se tendrán en cuenta las siguientes variables. Zona, Agencia Alto riesgo, Agencia Bajo riesgo, Género Femenino, Estado civil soltero- casado, Tipo de vivienda, Edad, Activos, Ingresos, Desembolso, Cuota, Plazo.

9.1.1. Modelo 1

Para la realización del modelo 1 se incluyen la totalidad de las variables seleccionas descritas en el párrafo anterior (13 variables) y se selecciona aleatoriamente el 70,1% del total de la población de los datos de la base (13.798 registros).

Figura 25.

Resumen de procesamiento de casos regresión logística.

Regresion Logística

Resumen de procesamiento de casos

Casos sin ponderar[a]		N	Porcentaje
Casos seleccionados	Incluido en el análisis	13798	70,1
	Casos perdidos	0	0,0
	Total	13798	70,1
Casos no seleccionados		5891	29,9
Total		19689	100,0

a. Si la ponderación está en vigor, consulte la tabla de clasificación para el número total de casos.

Nota Elaboración propia en herramienta SPSS.

Validando el modelo se procede a seleccionar una muestra aleatoria del total de la población 70.1 % que corresponde a 13.798 registros con el fin de evaluar el comportamiento del modelo, quedando 5.891 registros sin seleccionar que corresponde al 29.9% con el fin de realizar la muestra de comprobación.

Figura 26.

Codificación variable dependiente.

Codificación de variable	
Valor original	Valor interno
Buenos	0
Malos	1

Nota Elaboración propia en herramienta SPSS.

De acuerdo con la figura 26 la Codificación de la variable dependiente, 1 para los malos y 0 para los buenos.

Sobre la bondad de ajuste del modelo:

Significación de Chi-cuadrado del modelo 1 en la prueba ómnibus.

Figura 27.

Pruebas ómnibus de coeficiente de modelo.

Pruebas ómnibus de coeficientes de modelo				
		Chi-cuadrado	gl	Sig.
Paso 1	Paso	749,176	13	0,000
	Bloque	749,176	13	0,000
	Modelo	749,176	13	0,000

Nota Elaboración propia en herramienta SPSS.

La prueba ómnibus nos indica que el Modelo 1 explica el evento, es decir las variables independientes explican la variable dependiente dado que el resultado del P valor es menor que 0,05, procedemos a validar mediante el siguiente paso:

R-cuadrado de Cox y Snell, y R-cuadrado de Nagelkerkeb y tablas de clasificación

Esta prueba indica la parte de la varianza de la variable dependiente explicada por el modelo.

Figura 28.

Resumen del modelo R cuadrado de Cox y Snell – R cuadrado de Nagelkerke.

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	10251,565[a]	0,053	0,096

a. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.

Nota Elaboración propia en herramienta SPSS.

Mediante la prueba de R cuadrado de Cox y Snell y R cuadrado de Nagelkerke se ha obtenido un valor de R cuadrado de Cox = 0.053. Teniendo en cuenta que este coeficiente toma valores entre 0 y 1 siendo 0 un efecto bajo de las variables independientes. El resultado obtenido de R cuadrado Nagelkerke = 0.096 señala un bajo ajuste de los datos esto puede darse por la cantidad de variables tomadas.

Prueba de Hosmer y Lemeshow.

Figura 29.

Prueba de hosmer y lemeshow.

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	16,481	8	0,036

Nota Elaboración propia en herramienta SPSS.

En el contraste se dividen los datos en deciles en base a las probabilidades predichas. Se construye el siguiente estadístico de prueba

$$HL = \sum_{i=1}^{10} \frac{[O_i - N_i \bar{p}_i]^2}{N_i \bar{p}_i (1 - \bar{p}_i)}$$

Los términos de la fórmula se describen a continuación:

O_i Es el número de unos en la decila *i*-ésima

$\bar{p}_i$ Es la media de las probabilidades predichas en la decila *i*-ésima

N_i Es el número de observaciones en la decila *i*-ésima

La prueba de Hosmer es un método para estudiar la bondad de ajuste del modelo. La hipótesis nula del test es que no hay diferencias entre los valores observados y los valores pronosticados, se tiene que HL se distribuye como una chi cuadrado con 8 de grados de libertad. La conclusión es que a un nivel de significación de 0.05 el modelo 1 no ajusta bien los datos, dado que el P valor se encuentra por debajo de 0.05.

Figura 30.

Análisis de las variables de la ecuación.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	ZONA_C_NS	0,017	0,059	0,085	1	0,771	1,017
	AGENCIA_AR	0,585	0,060	95,811	1	0,000	1,796
	AGENCIA_BR	-0,735	0,096	58,686	1	0,000	0,479
	FEMENINO	-0,349	0,051	46,073	1	0,000	0,705
	SOLTERO	0,093	0,056	2,770	1	0,096	1,098
	CASADO	-1,013	0,137	54,864	1	0,000	0,363
	VIVIENDA_PROPIA	-0,220	0,079	7,796	1	0,005	0,802
	EDAD_R	-0,139	0,026	29,721	1	0,000	0,870
	ACTIVOS_R	-0,121	0,022	31,553	1	0,000	0,886
	INGRESOS_R	0,072	0,064	1,243	1	0,265	1,074
	DESEMBOLO_R	-0,315	0,078	16,316	1	0,000	0,730
	CUOTA	-0,108	0,074	2,141	1	0,143	0,898
	PLAZO	0,232	0,077	9,078	1	0,003	1,261
	Constante	-0,889	0,196	20,551	1	0,000	0,411

a. Variables especificadas en el paso 1: ZONA_C_NS, AGENCIA_AR, AGENCIA_BR, FEMENINO, SOLTERO, CASADO, VIVIENDA_PROPIA, EDAD_R, ACTIVOS_R, INGRESOS_R, DESEMBOLO_R, CUOTA, PLAZO.

En el modelo 1 podemos observar los contrastes del modelo mediante la prueba Chi cuadrado de Wald, permitiendo comparar la significancia individual de cada una de las variables presentes en el modelo. Se evidencia que las variables Zona_Nor (0,085), SOLTERO (2,770), INGRESOS_R (1,074) y CUOTA (0,898), no resultan ser significantes a un nivel de 0,05, por tanto, se concluye que se debe correr un nuevo modelo excluyendo estas variables.

9.1.2. Modelo 2

Se excluyen las variables que no fueron significativas en el modelo 1, Zona_Nor, SOLTERO, INGRESOS_R y CUOTA.

Se procede a realizar el modelo con el total de variables (9 variables) seleccionadas las cuales son: Agencias Alto Riesgo, Agencias Bajo Riesgo, Femenino, Casado, Vivienda Propia, Edad, activos, Valor de desembolso, plazo las cuales según el ejercicio validado del escenario Numero 1 indican que poseen una mejor predicción de nuestra variable dependiente.

Para la realización del modelo se incluyen las variables seleccionas descritas en el párrafo anterior (9 variables) y se selecciona aleatoriamente el 70,1% del total de la población de los datos de la base (13798 registros).

Significación de Chi-cuadrado del modelo en la prueba ómnibus.

Figura 31.

Pruebas ómnibus de coeficientes de modelo regresión logística.

Pruebas ómnibus de coeficientes de modelo				
		Chi-cuadrado	gl	Sig.
Paso 1	Paso	742,824	9	0,000
	Bloque	742,824	9	0,000
	Modelo	742,824	9	0,000

La prueba ómnibus nos indica que el Modelo 2 explica el evento, es decir las variables independientes explican la variable dependiente dado que el resultado del P valor es menor que 0,05, procedemos a validar mediante el siguiente paso:

R-cuadrado de Cox y Snell, y R-cuadrado de Nagelkerke y tablas de clasificación

Esta prueba indica la parte de la varianza de la variable dependiente explicada por el modelo.

Figura 32.

Resumen del modelo R cuadrado de Cox y Snell – R cuadrado de Nagelkerke.

Resumen del modelo			
Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	10257,917[a]	0,052	0,095
a. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.			

Nota Elaboración propia en herramienta SPSS.

Mediante la prueba de R cuadrado de Cox y Snell y R cuadrado de Nagelkerke se ha obtenido un valor de R cuadrado de Cox = 0.052. Teniendo en cuenta que este coeficiente toma valores entre 0 y 1 siendo 0 un efecto bajo de las variables independientes. El resultado obtenido de R cuadrado Nagelkerke = 0.095 señala un bajo ajuste de los datos esto puede darse por la cantidad de variables tomadas incluso menos variables que en el modelo 1.

Prueba de Hosmer y Lemeshow.

Según el R cuadrado de Cox y Snell y R cuadrado de Nagelkerke se concluye que las variables independientes están explicando de manera correcta la variable dependiente.

Figura 33.

Prueba de Hosmer y Lemeshow.

Prueba de Hosmer y Lemeshow			
Paso	Chi-cuadrado	gl	Sig.
1	10,096	8	0,258

Nota Elaboración propia en herramienta SPSS.

La hipótesis nula del test es que no hay diferencias entre los valores observados y los valores pronosticados. La conclusión es que a un nivel de significancia del 5 % se concluye que el modelo 2 ajusta bien los datos, dado que el P valor se encuentra por encima de 0,05.

Figura 34.

Análisis de las variables de la ecuación.

Variables en la ecuación							
		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	AGENCIA_AR	0,579	0,056	108,274	1	0,000	1,784
	AGENCIA_BR	-0,731	0,093	61,546	1	0,000	0,481
	FEMENINO	-0,348	0,051	46,039	1	0,000	0,706
	CASADO	-1,059	0,134	62,701	1	0,000	0,347
	VIVIENDA_PROPIA	-0,243	0,078	9,785	1	0,002	0,784
	EDAD_R	-0,145	0,025	34,102	1	0,000	0,865
	ACTIVOS_R	-0,118	0,021	31,494	1	0,000	0,889
	DESEMBOLO_R	-0,397	0,048	67,066	1	0,000	0,672
	PLAZO	0,293	0,060	23,975	1	0,000	1,340
	Constante	-0,900	0,133	46,169	1	0,000	0,406

a. Variables especificadas en el paso 1: AGENCIA_AR, AGENCIA_BR, FEMENINO, CASADO, VIVIENDA_PROPIA, EDAD_R, ACTIVOS_R, DESEMBOLO_R, PLAZO.

Nota Elaboración propia en herramienta SPSS.

En el modelo 2 podemos observar los contrastes del modelo mediante la prueba Chi cuadrado de Wald para cada variable permitiendo contrastar la significancia individual de cada una de las

variables presentes en el modelo, además se evidencia que realizadas las diferentes pruebas arrojan que se ajustan dado que explican las variables dependientes utilizadas en el modelo 2 y resultan ser significantes. En general el modelo es aceptable.

9.2. Ecuación Del Modelo

Con la información obtenida se procede a construir la ecuación del modelo.

$$log\ (odds) = intercep + b_1 x_1 + b_2 x_2 + \ldots + b_{11} x_{11}$$

$$\mathbf{Z} = -0{,}900 + 0{,}579 * AGENCIA\_AR - 0{,}731 * AGENCIA\_BR - 0{,}348 * FEMENINO - 1{,}059 * CASADO - 0{,}243 * VIVIENDA\_PROPIA - 0{,}145 * EDAD\_R - 0{,}118 * ACTIVOS\_R - 0{,}397 * DESEMBOLSO\_R + 0{,}293 * PLAZO$$

DONDE

Probabilidad de Incumplimiento $$P(Y = 1) = \frac{e^z}{1 + e^z}$$

EJEMPLO SOBRE EL USO DEL MODELO. Para explicar el funcionamiento del modelo se realizarán los siguientes ejemplos. 2 personas de diferente característica.

PERSONA 1

Crédito solicitado por un hombre de 22 años, perteneciente a la agencia de Cúcuta, cuyo estado civil es soltero, vivienda familiar, posee en activos $4.000.000, monto del crédito $800.000 a un plazo de 12 meses.

Dada la categorización de las variables y los coeficientes para cada una de las categorías de la tabla resultantes del modelo 2, los valores en estas variables se multiplican por los parámetros estimados.

La suma de cada uno de los valores se constituye como el valor de Z para el presente ejemplo el valor de Z es -0,396 que se remplaza en la formula probabilidad de incumplimiento. Como resultado para la persona 1 se obtiene una probabilidad de incumplimiento de 0,402 que indica que es una persona que se encuentra en un alto riesgo de caer en default y según nuestro score distribution (Figura 35) es una persona no apta para tomar el crédito.

PERSONA 2

Crédito solicitado por un hombre de 51 años, perteneciente a la agencia de Zapatoca, cuyo estado civil es casado, vivienda propia, posee en activos $50.000.000, monto del crédito $1.500.000 a un plazo de 18 meses.

Dada la categorización de las variables y los coeficientes para cada una de las categorías de la tabla resultantes del modelo 2. Las variables se multiplican por las respectivas estimaciones de los parámetros.

La suma de cada uno de los valores se constituye como el valor de Z para el presente ejemplo el valor de Z es -4,764 que se remplaza en la formula probabilidad de incumplimiento. Como resultado para la persona 2 se obtiene una probabilidad de incumplimiento de 0,0085 que indica que es una persona en bajo riesgo de caer en default y según nuestro score distribution (Figura 35) es una persona apta o buena para tomar el crédito.

10. Tabla Distribución De Frecuencias

En la Figura 35 se observa una tabla de frecuencias de las probabilidades calculadas bajo el modelo de regresión logística discriminado por buenos y malos pagadores.

Esta distribución de frecuencia se realiza para determinar un punto de corte en la probabilidad, para determinar quién es bueno y quien es malo, se utiliza como herramienta visual con distribución de percentiles del 2,5 % para obtener un desglose de como varía la probabilidad por percentiles. Por ejemplo en la Figura 35 observamos que el 10% de la población tiene una probabilidad de 0,0428 % basado en esa probabilidad se realiza una tabla de frecuencia de personas que pagan bien o que pagan mal para poder determinar un punto de corte aceptable. La columna default acumulado determina el nivel de tolerancia que se está dispuesta asumir la cooperativa.

Figura 35.

Score distribution del modelo de otorgamiento.

SCORE DISTRIBUTION																				
Distribución probabilidades defualt regresión logística			NÚMERO ASOCOCIADOS			PROPORCIÓN ASOCIADO x 100		NÚMERO ACUMULADO ASOCOCIADOS			DEFAULT ACUMULADO		PROPORCIÓN ACUMULADO ASOCOCIADOS			INDICADORES DE RIESGO		CALIFICACIÓN	
Banda	Percentil	Prob.	Totales	Buenos	Malos	Buenos	Malos	Totales	Buenos	Malos	Buenos	Malos	Totales	Buenos	Malos	KS	Odds Ratio		
1	2,5	0,0222	487	482	5	98,97	1,03	487	482	5	98,97	1,03	2,5	3	0	2,6	15,1	AA	BUENO
2	5	0,0287	498	489	9	98,19	1,81	985	971	14	98,58	1,42	5,0	6	1	5,2	10,8		
3	7,5	0,0359	467	450	17	96,36	3,64	1.452	1.421	31	97,87	2,13	7,4	8	1	7,2	7,2		
4	10	0,0428	490	472	18	96,33	3,67	1.942	1.893	49	97,48	2,52	9,9	11	2	9,3	6,0		
5	12,5	0,0491	500	470	30	94,00	6,00	2.442	2.363	79	96,76	3,24	12,4	14	3	10,9	4,7		
6	15	0,0537	327	313	14	95,72	4,28	2.769	2.676	93	96,64	3,36	14,1	16	3	12,2	4,5		
7	17,5	0,0591	671	641	30	95,53	4,47	3.440	3.317	123	96,42	3,58	17,5	19	5	14,9	4,2		
8	20	0,0630	498	473	25	94,98	5,02	3.938	3.790	148	96,24	3,76	20,0	22	6	16,7	4,0		
9	22,5	0,0701	484	456	28	94,21	5,79	4.422	4.246	176	96,02	3,98	22,5	25	7	18,3	3,8	A	
10	25	0,0743	488	447	41	91,60	8,40	4.910	4.693	217	95,58	4,42	24,9	28	8	19,4	3,4		
11	27,5	0,0778	420	388	32	92,38	7,62	5.330	5.081	249	95,33	4,67	27,1	30	9	20,5	3,2		
12	30	0,0837	576	528	48	91,67	8,33	5.906	5.609	297	94,97	5,03	30,0	33	11	21,8	3,0		
13	32,5	0,0882	490	449	41	91,63	8,37	6.396	6.058	338	94,72	5,28	32,5	36	13	22,9	2,8		
14	35	0,0934	493	446	47	90,47	9,53	6.889	6.504	385	94,41	5,59	35,0	38	14	23,7	2,6		
15	37,5	0,0979	465	424	41	91,18	8,82	7.354	6.928	426	94,21	5,79	37,4	41	16	24,7	2,5		
16	40	0,1038	485	428	57	88,25	11,75	7.839	7.356	483	93,84	6,16	39,8	43	18	25,0	2,4		
17	42,5	0,1078	529	475	54	89,79	10,21	8.368	7.831	537	93,58	6,42	42,5	46	20	25,8	2,3	B	
18	45	0,1135	489	439	50	89,78	10,22	8.857	8.270	587	93,37	6,63	45,0	49	22	26,5	2,2		
19	47,5	0,1191	450	405	45	90,00	10,00	9.307	8.675	632	93,21	6,79	47,3	51	24	27,2	2,1		
20	50	0,1232	524	466	58	88,93	11,07	9.831	9.141	690	92,98	7,02	49,9	54	26	27,8	2,1		
21	52,5	0,1300	501	437	64	87,23	12,77	10.332	9.578	754	92,70	7,30	52,5	56	28	27,9	2,0		
22	55	0,1352	482	429	53	89,00	11,00	10.814	10.007	807	92,54	7,46	54,9	59	30	28,4	1,9		
23	57,5	0,1377	505	413	92	81,78	18,22	11.319	10.420	899	92,06	7,94	57,5	61	34	27,4	1,8		
24	60	0,1457	486	414	72	85,19	14,81	11.805	10.834	971	91,77	8,23	60,0	64	36	27,1	1,7		
25	62,5	0,1520	468	407	61	86,97	13,03	12.273	11.241	1.032	91,59	8,41	62,3	66	39	27,2	1,7		
26	65	0,1545	525	454	71	86,48	13,52	12.798	11.695	1.103	91,38	8,62	65,0	69	41	27,2	1,7		
27	67,5	0,1632	376	299	77	79,52	20,48	13.174	11.994	1.180	91,04	8,96	66,9	70	44	26,1	1,6		
28	70	0,1689	592	488	104	82,43	17,57	13.766	12.482	1.284	90,67	9,33	69,9	73	48	25,1	1,5		
29	72,5	0,1799	491	398	93	81,06	18,94	14.257	12.880	1.377	90,34	9,66	72,4	76	52	23,9	1,5	C	MALOS
30	75	0,1840	254	208	46	81,89	18,11	14.511	13.088	1.423	90,19	9,81	73,7	77	53	23,4	1,4		
31	77,5	0,1919	745	585	160	78,52	21,48	15.256	13.673	1.583	89,62	10,38	77,5	80	59	20,8	1,3		
32	80	0,2024	468	361	107	77,14	22,86	15.724	14.034	1.690	89,25	10,75	79,9	82	64	18,9	1,3		
33	82,5	0,2074	507	407	100	80,28	19,72	16.231	14.441	1.790	88,97	11,03	82,4	85	67	17,5	1,3	D	
34	85	0,2165	466	353	113	75,75	24,25	16.697	14.794	1.903	88,60	11,40	84,8	87	72	15,4	1,2		
35	87,5	0,2284	525	418	107	79,62	20,38	17.222	15.212	2.010	88,33	11,67	87,5	89	76	13,8	1,2		
36	90	0,2460	448	344	104	76,79	23,21	17.670	15.556	2.114	88,04	11,96	89,7	91	79	11,9	1,1		
37	92,5	0,2722	542	426	116	78,60	21,40	18.212	15.982	2.230	87,76	12,24	92,5	94	84	10,1	1,1	E	
38	95	0,2868	423	306	117	72,34	27,66	18.635	16.288	2.347	87,41	12,59	94,6	96	88	7,5	1,1		
39	97,5	0,3208	557	416	141	74,69	25,31	19.192	16.704	2.488	87,04	12,96	97,5	98	93	4,6	1,0		
40	100	0,4485	497	324	173	65,19	34,81	19.689	17.028	2.661	86,48	13,52	100,0	100	100	0,0	1,0		
TOTAL			19689	17028	2661														

Esta tabla de frecuencias resulta de aplicar el modelo 2 a todos los elementos de la base utilizada para ajustarlo.

Para este segmento se establece el límite máximo de probabilidad de incumplimiento del 0,1689 que corresponde al 69.9% del total de créditos. Se determina hasta este punto de corte teniendo en cuenta que el indicador de default seria del 9.33 % algo que sería favorable teniendo en cuenta la línea a la cual estamos realizando el estudio.

Definición de las calificaciones.

- AA = Este tipo de cliente tiene una muy alta probabilidad de cancelar muy bien sus créditos dado que el indicador de mora para esta calificación se estima máximo del 3.76%.

- A = Este tipo de cliente tiene alta probabilidad de cancelar bien sus créditos, dado que el indicador de mora para esta calificación se estima máximo del 6.16%.

- B = Este tipo de cliente tiene una media probabilidad de cancelar bien sus créditos, dado que el indicador de mora para esta calificación se estima máximo del 9.33%, que es hasta donde se considera de acuerdo a los parámetros de la empresa tolerable.

- C = Este tipo de cliente tiene una baja probabilidad de cancelar bien sus créditos.

- D = Este tipo de cliente tiene muy baja probabilidad de cancelar bien sus créditos.

- E = Este tipo de cliente pagaría muy mal sus créditos, dado que la probabilidad de incurrir en mora es muy alta.

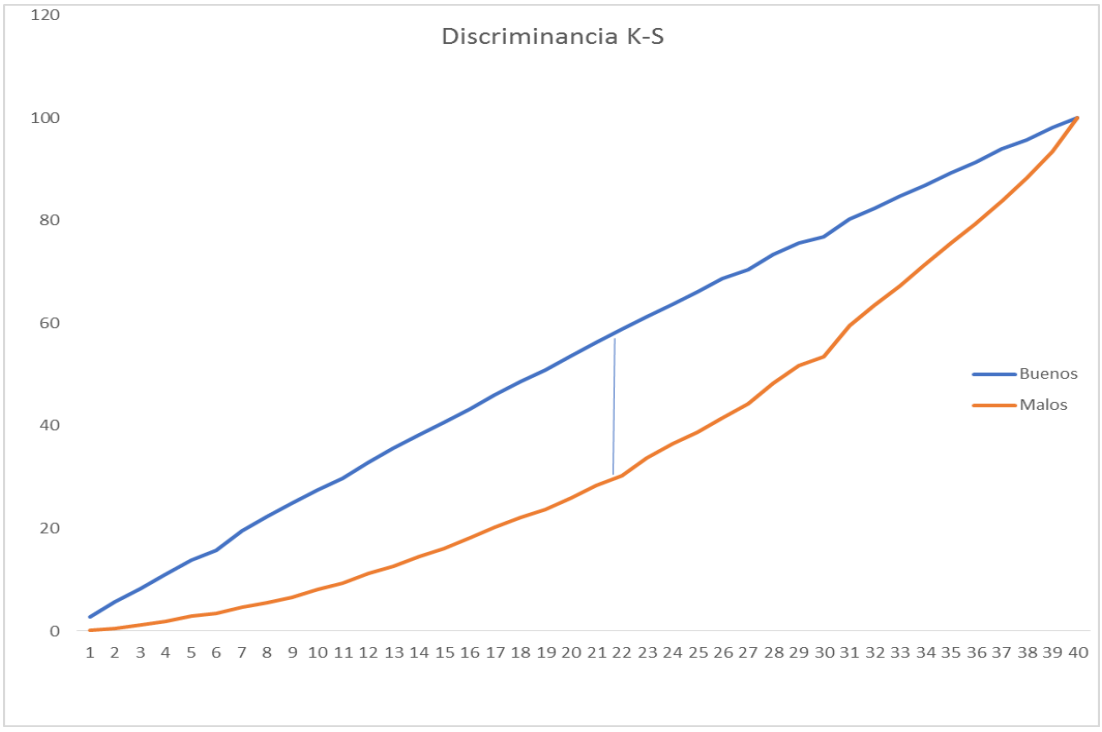
11. Poder Predictivo Del Modelo

11.1. Test Kolgomorov – Smirnov (ks)

El test Kolgomorov – Smirnov (KS) ilustrado en la figura 36 muestra el poder discriminante que tiene el modelo entre buenos y malos pagadores, se grafican las frecuencias acumuladas de (buenos y malos) y después se calcula la máxima diferencia que existe entre ambas en términos absolutos. En este caso los que presentan mora mayor a 60 días y los que no. Para este caso el KS en la muestra del modelo arroja un resultado de 28.4% que corresponde al 55% de la población. El modelo discrimina aceptablemente entre los buenos y los malos pagadores.

Figura 36.

Test Kolgomorov – Smirnov (KS) – Discriminancia.



Nota Elaboración propia en herramienta SPSS.

11.2. Tabla De clasificación

Podemos evidenciar en la Figura 37 que el modelo se comporta similar al 70% de la muestra seleccionada o muestra de entrenamiento (70,2) y al 30 % de la muestra no seleccionada o muestra de prueba. (71.5) El modelo explica bien para los que pagan bien, pero no tan bien para los que pagan mal, esto podría tener consecuencia de no tomar otras variables que podrían tener mayor potencia en el modelo como son centrales de riesgo, pero el perfil analizado es para personas sin experiencia.

Figura 37.

Tabla de clasificación.

Tabla de clasificación[a]

			Pronosticado					
			Casos seleccionados[b]			Casos no seleccionados[c]		
			DEFAULT_60		Porcentaje correcto	DEFAULT_60		Porcentaje correcto
	Observado		Buenos	Malos		Buenos	Malos	
Paso 1	DEFAULT_60	Buenos	8705	3209	73,1	3811	1303	74,5
		Malos	909	975	51,8	377	400	51,5
	Porcentaje global				70,2			71,5

a. El valor de corte es ,169

b. Casos seleccionados Aprox. 70% Casos EQ 1

c. Casos no seleccionados Aprox. 70% Casos NE 1

Nota Elaboración propia en herramienta SPSS.

11.3. Tabla Cruzada Por Calificación

Con la tabla cruzada se concluye que el modelo calcula muy similar a la muestra del 70% como al 30 % de muestra de prueba. Basados en las calificaciones.

Figura 38.

Tabla cruzada calificación según probabilidad 70% casos comprobación.

Tabla cruzada Calificación según probabilidad*DEFAULT_60*Aprox. 70% Casos

Aprox. 70% Casos				DEFAULT_60 Buenos	Malos	Total
Comprobación	Calificación según probabilidad	AA	Recuento	1193	46	1239
			% dentro de Calificación según probabilidad	96,3%	3,7%	100,0%
		A	Recuento	1030	100	1130
			% dentro de Calificación según probabilidad	91,2%	8,8%	100,0%
		B	Recuento	1562	227	1789
			% dentro de Calificación según probabilidad	87,3%	12,7%	100,0%
		C	Recuento	442	127	569
			% dentro de Calificación según probabilidad	77,7%	22,3%	100,0%
		D	Recuento	363	112	475
			% dentro de Calificación según probabilidad	76,4%	23,6%	100,0%
		E	Recuento	417	138	555
			% dentro de Calificación según probabilidad	75,1%	24,9%	100,0%
	Total		Recuento	5007	750	5757
			% dentro de Calificación según probabilidad	87,0%	13,0%	100,0%

Nota Elaboración propia en herramienta SPSS.

Figura 39.

Tabla cruzada calificación según probabilidad Entrenamiento.

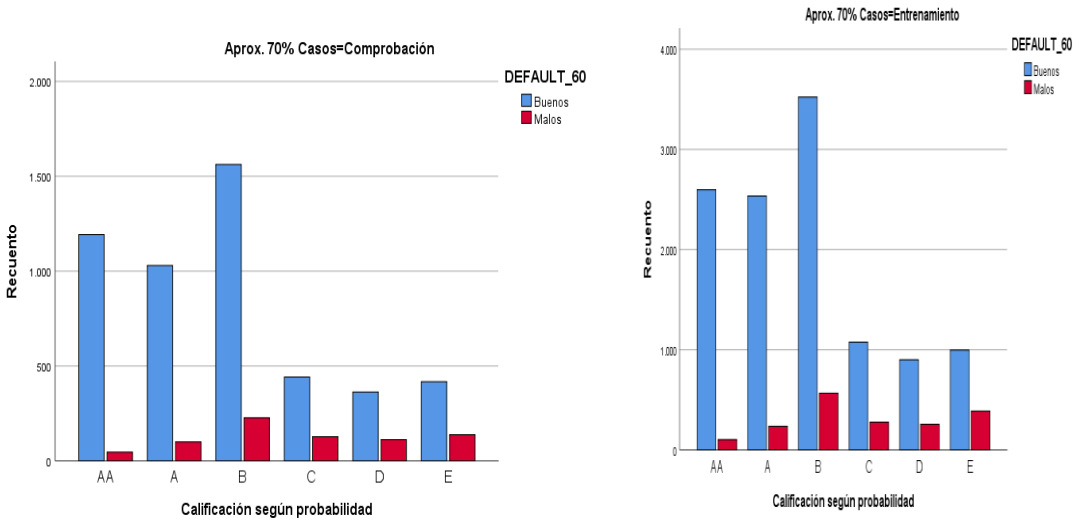
Entrenamiento	Calificación según probabilidad	AA	Recuento	2597	102	2699
			% dentro de Calificación según probabilidad	96,2%	3,8%	100,0%
		A	Recuento	2535	235	2770
			% dentro de Calificación según probabilidad	91,5%	8,5%	100,0%
		B	Recuento	3522	566	4088
			% dentro de Calificación según probabilidad	86,2%	13,8%	100,0%
		C	Recuento	1076	277	1353
			% dentro de Calificación según probabilidad	79,5%	20,5%	100,0%
		D	Recuento	900	255	1155
			% dentro de Calificación según probabilidad	77,9%	22,1%	100,0%
		E	Recuento	997	388	1385
			% dentro de Calificación según probabilidad	72,0%	28,0%	100,0%
	Total		Recuento	11627	1823	13450
			% dentro de Calificación según probabilidad	86,4%	13,6%	100,0%

Nota Elaboración propia en herramienta SPSS.

Con la Figura 38 confirmamos que el modelo predice similar al 70 % de la poblacion que se tomo como muestra de entrenamiento, asi como al 30 % de prueba. El modelo es estable. Graficamente podemos observar un comportamiento similar entre el 70 y el 30, en sus calificaciones tanto para los buenos y los malos pagadores.

Figura 40.

Gráfico calificación según probabilidad casos Comprobación y entrenamiento.



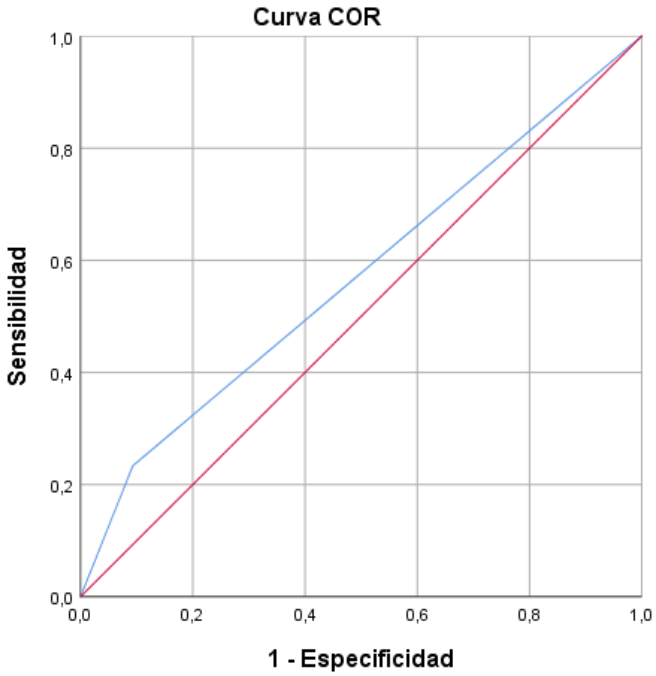
Nota Elaboración propia en herramienta SPSS.

11.4. Curva Cor

Cuantifica la capacidad de un indicador para discriminar entre buenos y malos pagadores, entre mayor sea el area bajo la curva el modelo tiene un poder de discriminacion alto. El resultado del modelo muestra una separacion entre la especifidad y sensibilidad obteniendo un area bajo la curva de 57%, se considera bueno para el modelo teniendo en cuenta que en la practica diaria o criterio experto para los modelos de scoring que se trabajan en la entidad financiera se considera aceptable un area bajo la curva del 40 %.

Figura 41.

Gráfico Curva COR.



Nota Elaboración propia en herramienta SPSS.

Área bajo la curva	
Variables de resultado de prueba:	DEFAULT_60
Área	
	0,570

Nota Elaboración propia en herramienta SPSS.

12. Conclusiones

El objetivo del presente trabajo es ofrecer a la cooperativa Financiera Comultrasan una herramienta cuantitativa que permita identificar los clientes que ingresan del segmento de créditos bajos montos (Menor a 2 SMMLV) como óptimos pagadores o pesimos pagadores.

Para tal fin se ajustaron dos modelos y se tuvo en cuenta el modelo 2 (Figura 34) que mostró variables significativas y se considera aceptable en el desempeño teniendo en cuenta las pruebas realizadas al mismo (Figura 37), el modelo presenta un buen rendimiento cuando de calcular la probabilidad de default de los asociados buenos se trata, pero no tan claro cuando se trata de los malos pagadores. Se elaboró una tabla de score distribution (Figura 35) con puntos de corte de probabilidad, en el cual se recomienda tener en cuenta clasificación AA y A aprobación en línea, (créditos aprobados automáticamente), clasificación B aprobación centro de crédito, (créditos que sean revisados por los analistas de crédito y den su criterio experto) a las personas que se encuentren en estos rangos se consideran buenos clientes dada la probabilidad mínima de incumplimiento en el pago de sus créditos. Dicha propuesta estará sujeta a la revisión por parte de la cooperativa para su implementación y el interés de atender un buen porcentaje de estos clientes asegurando un menor rango de pérdidas por el pago inoportuno de las obligaciones otorgadas.

La creación de este modelo consiste en crear un apoyo al área de crédito en la gestión diaria y búsqueda de nuevas plataformas de aprobación rápida para los asociados desde sus hogares, así como optimizar recursos físicos y humanos para la cooperativa, el complemento de este modelo junto con el análisis experto del analista lograra que el resultado final sea más óptimo y por lo tanto se tenga una mejor calidad de la cartera. Con el anterior diseño del modelo se podrá tener bases más sólidas para aprobar o negar una solicitud de crédito.

La implementación de este modelo de scoring para personas independientes sin experiencia crediticia y montos de aprobación hasta 2 SMMLV, muestra un avance en el área de crédito, como se mencionaba anteriormente, ayuda a mejorar tiempos de respuesta a los clientes y a cuantificar el riesgo mediante la probabilidad de incumplimiento.

Se observa que con la implementación del modelo se propone un indicador de mora del 9.33% menor al que se viene observando en la actualidad que es 13.52% para este tipo de segmento de créditos, atendiendo un 69% de la totalidad de los clientes bajo una aprobación automática y con revisión de algunos perfiles por parte de los analistas de crédito, sin embargo, esta decisión estará sujeto al alcance de mercado que se desee atender por parte de la cooperativa.

Referencias Bibliográficas

Amat Rodrigo, J. (Febrero de 2020). *Comparación de distribuciones: test Kolmogorov–Smirnov*. Obtenido de https://www.cienciadedatos.net/documentos/51_comparacion_distribuciones_kolmogorov%E2%80%93smirnov

Arango, L., & Restrepo, D. (2017). Diseño de un modelo de Scoring. *Diseño de un modelo de Scoring para el otorgamiento de crédito de consumo en una compañía de financiamiento colombiana.* Medellín, Colombia: Escuela de Economía y finanzas EAFIT.

Beltrán Martínez, B. (2014). *Minería de Datos.* Obtenido de [Figura]: http://bbeltran.cs.buap.mx/NotasMD.pdf

Building credit scorecards using SAS and Python. (28 de Febrero de 2016). *Building Credit Scorecards Using Credit Scoring for SAS® Enterprise MinerTM.* Obtenido de https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/building-credit-scorecards-using-credit-scoring-for-SAS-enterprise-miner-104182.pdf

Bulding and Implementing Better. (04 de Noviembre de 2016). Calificación crediticia inteligente. *crediticia inteligente : creación e implementación de mejores cuadros de mando de riesgo crediticio*. Wiley.

Camarero Rioja, L., Almazan Llorente, A., & Mañas Ramirez, B. (2018). *Regresión Logística Fundamentos y Aplicación a la investigación sociologica.* Obtenido de Analisis Multivariante: https://www2.uned.es/socioestadistica/Multivariante/Odd_Ratio_LogitV2.pdf

Comité de Supervisión Bancaria de Basilea. (Diciembre de 2017). *Resumen de las reformas de Basilea III.* Obtenido de https://www.bis.org/bcbs/publ/d424_hlsummary_es.pdf

Cuaderno Economico. (2013). METODOLOGÍA PARA UN SCORING DE CLIENTES SIN REFERENCIAS CREDITICIAS. En O. E. García, *METODOLOGÍA PARA UN SCORING DE CLIENTES SIN REFERENCIAS CREDITICIAS* (págs. 139 - 165). ISSN. Obtenido de https://revistas.unal.edu.co/index.php/ceconomia/article/view/38348/40677

Ealde. (20 de Marzo de 2018). *Los 4 tipos de Riesgo de Crédito.* Obtenido de Finanzas, Gestión de Riesgo: https://www.ealde.es/gestion-de-riesgos-de-credito/

Echemendía Tocabens, B. (2011). Definiciones acerca del riesgo y sus implicaciones. *Revista Cubana de Higiene y Epidemiología*, 49(3), 470-481.

Financiera Comultrasan. (Mayo de 2020). *Financiera Comultrasan*. Obtenido de Pagina oficial virtual: https://www.financieracomultrasan.com.co/es

Gutiérrez Girault, M. A. (11 de Diciembre de 2007). *Modelos de credit Scoring.* Obtenido de Modelos de credit Scoring, Qué, cómo, cuándo y para qué. Munich: Munich Personal RePEc Archive: https://mpra.ub.uni-muenchen.de/16377/

Hernández Cabrera, J. L. (03 de Enero de 2012). *Técnica de los árboles de causa y efecto para solución de problemas sociales*. Obtenido de https://www.gestiopolis.com/tecnica-arboles-causa-efecto-solucion-de-problemas-sociales/

Laboratorio e infectolodía. (27 de Febrero de 2012). *Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos.* Obtenido de [Figura]: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-10182012000200003

López, P. &. (2015). *Metodología de la investigación social cuantitativa.* Obtenido de https://ddd.uab.cat/record/129382

Ministerio de Hacienda. (Septiembre de 2013). *Sistema de Administración de Riesgo Operativo – SARO.* Obtenido de Gestión de Riesgo - Sistema de Administración de Riesgo Operativo – SARO: https://repository.ucc.edu.co/bitstream/20.500.12494/15617/2/2019_riesgo_organizaciones_cooperativas.pdf

Moine, J. (2013). *Metodología para el descubrimiento de conocimiento en bases de datos (tesis de grado).* La Plata: Universidad Nacional de la Plata.

Peiro Ucha, A. (Junio de 2018). *Economipedia.* Obtenido de Economipedia - Riesgo de Credito: https://economipedia.com/definiciones/riesgo-de-credito.html

Perez López, C. (2004). Técnicas de Análisis Multivariante de Datos Aplicaciones con SPSS. [Figura]. *Madrid, España: Pearson Prentice Hall.*

Perez López, C. (2004). *Técnicas de Análisis Multivariante de Datos Aplicaciones con SPSS.* Madrid, España: Pearson Prentice Hall.

Presidencia de la República. (02 de abril de 1993). Decreto 663 de 1993. *por medio del cual se actualiza el Estatuto Orgánico del Sistema Financiero y se modifica su titulación y numeración*. Bogotá, Colombia: Ministerio de Justicia . Obtenido de http://www.secretariasenado.gov.co/senado/basedoc/estatuto_organico_sistema_financiero.html

Rankia, & Trecet, J. C. (09 de Abril de 2020). *¿Qué es el SARC y para qué sirve?* Obtenido de ¿Qué es el SARC y para qué sirve?: https://www.rankia.co/blog/mejores-creditos-y-prestamos-colombia/4104690-que-sarc-para-sirve

Saavedra García, M. L., & Saavedra García, M. J. (2010). Modelos para medir el riesgo de crédito en la banca. En M. L. Saavedra García, & M. J. Saavedra García, *Modelos para medir el riesgo de crédito en la banca* (págs. 23(40), 295). Bogotá: Cuadernos de Administración.

Samaniego Medina, R. (Noviembre de 2008). El riesgo de crédito en el marco del Acuerdo de Basilea II. *El riesgo de crédito en el marco del Acuerdo de Basilea II*. Madrid, España: Delta Publicaciones Universitarias.

Superintendencia Financiera de Colombia. (1995). Circular básica, contable y financiera (circular externa 100 de 1995). Bogotá, Colombia: Ministerio de Hacienda.

Superintendencia Financiera de Colombia. (Abril de 2017). *Direcciones de RIesgo de Credito*. Obtenido de https://www.superfinanciera.gov.co/jsp/Publicaciones/publicaciones/loadContenidoPublicacion/id/60916/dPrint/1/c/00

Apéndice A. Árbol de Decisión