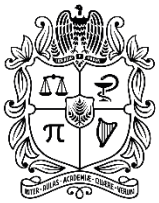


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Aprendizaje supervisado en la construcción de un modelo de *Credit Scoring* para cooperativas de ahorro y crédito en Colombia

Jonathan Cano Bedoya

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2021

Aprendizaje supervisado en la construcción de un modelo de *Credit Scoring* para cooperativas de ahorro y crédito en Colombia

Jonathan Cano Bedoya

Tesis de investigación presentada como requisito parcial para optar al título de:

Magister en Ingeniería Analítica

Director (a):
PhD. Claudia Stella Jiménez Ramírez

Codirector (a):
MSc. Luz Estela Sánchez Herrera

Línea de Investigación:
Aprendizaje supervisado

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión
Medellín, Colombia
2021

A mi Dios

A mi mamá que en paz descansa

A mis asesoras Claudia y Luz Estela

“El investigador sufre las decepciones, los largos meses pasados en una dirección equivocada, los fracasos. Pero los fracasos son también útiles, porque, bien analizados, pueden conducir al éxito. Y para el investigador no existe alegría comparable a la de un descubrimiento, por pequeño que sea...”

Sir. Alexander Fleming

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Nombre

Fecha 25/08/2021

Agradecimientos

Agradezco a la Universidad Nacional de Colombia por este tiempo en el que me he formado con los aportes y saberes de los docentes del área de Ciencias de la Computación y la Decisión de la Facultad de Minas, de los cuales he aprendido otras formas de abordar el conocimiento, la investigación y seguir aprendiendo.

En especial doy agradecimientos a mi directora de tesis la profesora Claudia Jiménez R., docente de la Facultad de Minas, quien tiene una amplia trayectoria y experiencia en docencia e investigación, aproveché su aporte en las asignaturas que me impartió y en su dirección para orientarme a culminar este proceso de formación. Gracias por estar siempre dispuesta.

Doy gracias a mi codirectora la profesora Luz Estela S., docente del Instituto de Matemáticas de la Universidad de Antioquia, de quien en mi formación como estadístico he recibo aportes importantes y por estar dispuesta para apoyarme.

Gracias al profesor Francisco Javier Díaz, docente de la Facultad de Minas quien me ha dado la oportunidad de acompañarlo como monitor en los cursos optimización, simulación e investigación de operaciones, sus consejos y aportes me han sido de mucho aprendizaje.

Resumen

Aprendizaje supervisado en la construcción de un modelo de *Credit Scoring* para cooperativas de ahorro y crédito en Colombia

Las entidades financieras del sector solidario brindan créditos a sus asociados, exponiéndose a un riesgo derivado de esta actividad. Este trabajo de investigación tiene como objetivo la propuesta metodológica de un modelo de riesgo de crédito, que asigne un puntaje de crédito a los deudores con base en la información que estos proporcionan a las entidades financieras. Se emplean técnicas analíticas de aprendizaje supervisado para la construcción del modelo con la metodología ASUM-DM.

Se toman los datos de una cartera de crédito de una compañía del sector solidario en Colombia para ilustrar el desarrollo metodológico en la construcción del modelo de *Scoring* para el *score* de comportamiento de los asociados. Se utilizan métodos Biplot, análisis de correspondencias y componentes principales para la reducción de dimensionalidad. También se utilizan las técnicas de árboles de decisión, modelos de regresión probabilísticos, K-Vecinos más cercanos, máquinas de soporte vectorial y redes neuronales. Los métodos anteriores son utilizados para el modelo de *Credit Scoring* en la asignación del puntaje de crédito y la selección de las variables que son significativas en el estudio. El modelo de regresión logística es empleado para el *Score de crédito* y se compara con las demás técnicas supervisadas mediante métricas de rendimiento.

Palabras clave: aprendizaje supervisado, reducción de dimensionalidad, clasificación, regresión logística, puntaje de crédito.

Abstract

Supervised learning in the construction of a Credit Scoring model for savings and credit cooperatives in Colombia

The financial companies of the solidarity sector provide credits to their associates, exposing themselves to a risk derived from this activity. The objective of this research work is the methodological proposal of a credit risk model, which assigns a credit score to credit debtors based on the information they provide to financial companies. Analytical techniques of supervised learning are used for the construction of the model accompanied by the ASUM-DM methodology.

Data are taken from a credit portfolio of a company in the solidarity sector in Colombia to illustrate the methodological development in the construction of the Scoring model for the behavior score of the associates. The Biplot methods, correspondence analysis and principal components are used for dimensionality reduction. Techniques of decision tree, probabilistic regression models, K-Nearest Neighbors, vector support machines and neural networks are also used. The previous elements are used for the Credit Scoring model in assigning the credit score and selecting the variables that are significant in the study. The logistic regression model is used for the credit score and is compared with the other techniques supervised by performance metrics.

Keywords: supervised learning, dimensionality reduction, classification, logistic regression, Credit Scoring.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Lista de abreviaturas

Abreviaturas

Abreviatura	Término
ACM	Análisis de Correspondencias Múltiple
ACP	Análisis de Componentes Principales
AFC	Análisis Factorial de Correspondencias
ANN	Redes Neuronales Artificiales
ASUM-DM	Analytics Solucionas Unified Method
AUC	Área bajo la curva ROC
Bagging	Bootstrapping agreggation
CART	Árbol de clasificación y regresión
CHAID	Detección automática ji-cuadrado
Credit Scoring	Puntaje de crédito
CRISP-DM	*Cross Standard Process for Data Mining*
DVS	Descomposición en valores singulares
Fogacoop	Fondo de Garantías del Sector Cooperativo
IV	Valor de la información
KDD	*Knowledge Discovery in Databases*
KNN	K-Vecinos más cercanos
MRE	Minimización del Riesgo Estructural
MSE	*Mean Square Error*
NIIF	Norma Internacional de Información Financiera
OOB	*Out of Bagging*
PDO	Puntos para Duplicar el Odd
PSM	Proporción de Saldo Máximo
Random Forest	Bosques Aleatorios
RMSE	*Root Mean Square Error*
ROC	Receiver Operating Characteristic
RyF	Recencia y Frecuencia
Supersolidaria	Superintendencia de la Economía Solidaria
SVMC	Máquinas de soporte vectorial para clasificación
WOE	Peso de la Evidencia

Introducción

El riesgo de crédito se presenta por el incumplimiento de una deuda que surge cuando un prestatario no cumple con los pagos requeridos del dinero que se le prestó [1]. Trata entonces del incumplimiento o *default* en las obligaciones contraídas en una operación de crédito [2]. Este no se debe confundir con *Credit Scoring* que es la asignación de un puntaje o valoración del crédito de los prestatarios. Uno de los propósitos de medir el riesgo crediticio es apoyar la toma de decisiones de los prestamistas. Las palabras *Credit Scoring* se mantendrán escritas así en este documento porque se han venido adoptando para referirse a la asignación del puntaje de crédito, aunque también se pueden usar el término *Credit score* o puntaje de crédito. Desde el aprendizaje supervisado, la estadística y la inteligencia artificial, el concepto se aborda como un conjunto de modelos, métodos y técnicas para clasificar clientes de crédito y estimar la probabilidad de incumplimiento [3].

Los prestamistas financieros tienen la posibilidad de aceptar o rechazar una solicitud crédito, al parecer dos opciones simples. La toma de decisiones se ha basado en la experiencia, juicios subjetivos, e incluye la evaluación de técnicas analíticas por el capital que los prestamistas ponen en riesgo [4]. El riesgo de crédito es un problema que los alerta, porque afecta a sus organizaciones y la economía en general. Este se analiza de forma multidimensional porque son varios los eventos que pueden ser determinantes del cumplimiento o no del pago del crédito. Esto implica que se aborde el problema con herramientas que sean eficaces para modelarlo y gestionarlo.

El riesgo de crédito considera tres aspectos: exposición al incumplimiento, probabilidad de incumplimiento y tasa de recuperación [5]. Bajo el marco de Basilea [6] este riesgo se debe identificar, evaluar y gestionar. La norma internacional de contabilidad NIIF-9 estará implementada en mayor porcentaje para el año 2022. Esta brindará las pautas para el tratamiento del incumplimiento en la determinación del riesgo de crédito de un instrumento financiero. También incluirá las indicaciones de las tasas de provisiones, que se deben calcular dependiendo de los días de mora y la información que se debe incluir de los prestatarios.

A razón de esto, las pérdidas crediticias deben ser estimadas, lo que implica la innovación de la gestión del riesgo de crédito con técnicas estadísticas y analíticas con información financiera y contable para mejorar las condiciones de los prestatarios, prestamistas y la salud financiera global.

En función de la gestión del riesgo se han venido presentando un gran número de investigaciones alrededor del mundo sobre el tema, exponiendo nuevos modelos que mejorarían la predicción del cumplimiento, escenario que no es general por las diferencias culturales, prácticas empresariales y normas gubernamentales. Los juicios de expertos, la estadística y las técnicas de analítica han considerado varios enfoques sistemáticos con la propuesta de modelos de calificación crediticia en cuanto a la asignación de un puntaje. El proceso en la última década con un punto de partida desde la crisis financiera Sub-Prime ha puesto de manifiesto el esfuerzo de muchos investigadores para detectar el riesgo y mitigarlo, pero no ha sido fácil porque aún se encuentran problemas transversales, como el tratamiento de datos estructurados y no estructurados, información desequilibrada, pocos datos abiertos para modelar, alta dimensionalidad, falta de precisión e interpretabilidad, y el tamaño de las bases de datos, entre otros. Es un tema que continúa preocupando a la economía global a los grandes y pequeños prestamistas [7].

El número amplio de modelos y la falta de datos abiertos para el análisis es uno de los retos en el trabajo investigativo. A pesar de que se han planteado diversas medidas para evaluar y comparar el desempeño de los modelos, aún se presentan vacíos que al ser identificados podrían mejorar el resultado predictivo. No necesariamente técnicas exhaustivas o de gran complejidad son las que podrían generar buenos resultados, es el tratamiento de la información y la gestión de los datos disponibles.

El objetivo de este trabajo está en proponer un modelo de *Credit Scoring* para cooperativas financieras de ahorro y crédito, implementando modelos de aprendizaje supervisado, analizando la información mediante el proceso de descubrimiento de conocimiento KDD (*Knowledge Discovery in Databases*) para el tratamiento de la información que es el insumo principal del modelamiento. Se analiza el tratamiento de la información cualitativa y cuantitativa en los modelos de aprendizaje supervisado. Gran parte de la información histórica y disponible se encuentra en datos de esta naturaleza.

Los modelos descritos en la literatura han implementado la información de las variables en forma cuantitativa y cualitativa para la gestión del riesgo en cuanto a descripción, clasificación y predicción.

En este trabajo se describe lo que ha venido ocurriendo en la última década sobre el tema y se comparará con el desempeño de algunos de los modelos de aprendizaje supervisado que más se han utilizado. El análisis de *Credit Scoring* ha comprendido el uso de un número amplío de variables para su estimación. [8].

Los modelos más empleados desde el enfoque del aprendizaje supervisado y estadístico [9, 10, 11] han sido la regresión lineal y logística, análisis discriminante, algoritmos de programación lineal, redes neuronales, algoritmos genéticos, modelos híbridos, árboles de clasificación y métodos de clúster.

El trabajo de investigación se presenta en este documento en seis capítulos. En el primero describe el problema de negocio y analítico en relación con *Credit Scoring*. El segundo describe el estado del arte y conceptos teóricos que se utilizan en el desarrollo. En el tercer capítulo se presenta parte de las fases de la metodología ASUM-DM y algunas técnicas analíticas y estadísticas para el análisis de los datos. Se utiliza un conjunto de datos de una entidad financiera y con éste se describe todo el proceso investigativo. En el cuarto capítulo se realiza un proceso de modelamiento para clasificación y obtención de la probabilidad de incumplimiento de los asociados. El quinto capítulo está dedicado a la construcción y evaluación del modelo de *Scoring* para la asignación del puntaje de crédito. En el último capítulo se presentan conclusiones y trabajo futuro. Los capítulos se organizan de acuerdo con el alcance los objetivos propuestos.

1. Conocimiento del problema de negocio y analítico

El problema de negocio está orientado a cooperativas financieras, las cuales tienen unas particularidades diferentes a otras entidades como las del sector bancario, por ejemplo, en tema regulatorio.

1.1 Planteamiento del problema

En el mercado de microcrédito, las cooperativas financieras son entidades que prestan dinero a personas naturales y jurídicas, empresas o negocios pequeños. Las condiciones del crédito las estipula cada entidad. Por ejemplo, las garantías por el dinero prestado, nivel de estudios de la persona, actividad económica que desarrolla la persona, entre otras. En Colombia el cooperativismo está vigilado por la Superintendencia de Economía Solidaria, las normas y características atenientes a este sector están documentadas en la Ley 454 de 1998 [13], a su vez que se describe las operaciones que se pueden realizar.

Las operaciones de crédito se dividen en carteras, para el propósito que el prestatario destine el dinero prestado. Entre ellas, la cartera de consumo, la cartera comercial, de vivienda y microcrédito. Lo que preocupa a estas entidades es la aprobación de una solicitud de crédito por el capital que está en riesgo. Aunque se aplican metodologías y técnicas como el análisis de experto para la toma de decisiones. Por lo cual, estas entidades requieren un modelo que les permita decidir si aceptar una solicitud de crédito para evitar pérdidas, que no se dan solo por el dinero que se presta, sino que a esto se le suman los costos asociados al proceso de recuperación de la cartera.

En el sistema financiero colombiano se implementa el Sistema de Administración del Riesgo Crediticio (SARC) documentado en la Circular Básica y Contable y Financiera [14]. Sistema implementado desde el año 2006 que se encuentra bajo la supervisión de la Superintendencia Financiera de Colombia y la Superintendencia de Economía Solidaria. Las cooperativas de ahorro y crédito hacen parte de esta última y si bien no están obligadas a seguir las directrices de Basilea [1], si están obligadas a mantener el sistema SARC. Este contiene varias componentes que se muestran a continuación.

- Políticas de administración del riesgo de crédito
- Procesos de administración del riesgo de crédito
- Modelos internos o de referencia para la estimación o cuantificación de pérdidas esperadas
- Sistema de provisiones para cubrir el riesgo de crédito
- Procesos de control interno

Dentro de los modelos internos, las entidades vigiladas tienen libertad de implementar estas componentes para la estimación de las probabilidades de incumplimiento de los prestatarios así provisionar la cartera. El riesgo de crédito es el concepto principal, pero este trabajo se centra en el análisis de *Credit Scoring* que hace parte de uno de los aspectos de este riesgo, en cuanto a los modelos y técnicas analíticas a implementar para estimar la probabilidad de incumplimiento y su clasificación. Las entidades supervisadas pueden implementar un modelo analítico basado en las bases de datos que tengan en sus registros, además son responsables de reportar los modelos a las entidades regulatorias para verificar la eficacia de éstos en cuanto al control del riesgo crediticio. También deben identificar las variables críticas del incumplimiento por parte de los prestatarios y de las que son importantes y significativas para la discriminación de los sujetos de crédito. Esto se hace de manera particular para cada entidad.

La implementación de los modelos utiliza la probabilidad de impago y el *Score* de crédito para estimar las provisiones de cartera por el capital en riesgo. Acorde con SARC, esto hace parte del seguimiento y control exigido por la Superintendencia Financiera. Variables como el tipo de cartera, el tiempo que durará el crédito, la calificación de la cartera en la escala *A*, *B*, *C*, *D* y *E* que se explicarán más adelante son parte del estudio analítico a implementar.

El modelo de aprendizaje supervisado a proponer de *Credit Scoring* para las cooperativas de ahorro y crédito, se evalúa en una entidad financiera que tiene su operación principal en la ciudad de Medellín. Los datos se mantendrán en la confidencialidad por lo cual no serán publicados. Sin embargo, se describe la información.

El análisis de *Credit Scoring* será el eje temático de negocio que esta investigación busca abordar, para dar un aporte a estas entidades ante la necesidad de un modelo eficaz de control del riesgo de crédito que se pueda valorar de forma automatizada [15]. El problema analítico está en el diseño de un modelo de control que incluya el análisis de los datos históricos de los asociados que las entidades poseen en su información. La manera de abordar los análisis se hará en varias fases

hasta obtener un modelo adecuado que deba alienarse a los objetivos de la organización (control de riesgo de crédito).

Esta investigación busca responder las siguientes preguntas:

Q1: ¿Cuáles son las variables determinantes de *Credit Scoring* en los datos estudiados?

Q2: ¿Cuáles son los modelos de aprendizaje supervisado con mejores resultados para la medición de *Credit Scoring* en las cooperativas financieras?

Q3: ¿Cuál metodología de análisis debe emplear una cooperativa financiera de ahorro y crédito al aplicar los modelos de aprendizaje supervisado?

1.2 Justificación del proyecto de tesis

Credit Scoring es una herramienta de análisis para el riesgo crédito al que están sometidas las entidades financieras que prestan dinero. Esto incluye entidades de macrocrédito como bancos centrales, bancos de segundo piso, bancos comerciales, y de microcrédito como las cooperativas financieras, fondos de empleados. La calificación crediticia o asignación de un puntaje de crédito es uno de los temas importantes en la toma de decisiones de estas entidades [15]. Las características o atributos que se han venido recogiendo en el tiempo son la información disponible, que se debe evaluar porque no todas son útiles para los objetivos de la aplicación de modelos de aprendizaje supervisado en el modelamiento del riesgo de crédito.

La evaluación crediticia es un proceso que incluye recopilar, analizar y clasificar diferentes elementos y variables crediticias para evaluar las decisiones de aprobar solicitudes de crédito [17]. El tema está vigente y continúa siendo objetivo por una gran cantidad de investigadores en todo el mundo [18]. La puntuación de riesgo de crédito se ha venido desarrollando principalmente por modelos de aprendizaje estadístico, aprendizaje automático con modelos supervisados e inteligencia artificial.

Los modelos son útiles porque pueden mejorar la eficacia de estas entidades financieras en la asignación de créditos a los prestatarios. Esto es porque se busca incrementar la cartera de préstamos reduciendo el número de solicitudes injustificadas o solicitudes que podrían tener alta probabilidad de incumplimiento. Es así, que en el problema de negocio y analítico, se deben analizar y plantear soluciones para mejorar la precisión en la valoración del prestatario y minimizar el riesgo de incumplimiento. Esto llevaría a reducir los incumplimientos y fraudes, también aceleraría la valoración y descripción del perfil de los asociados.

En una cooperativa, el modelo que tenga buen acierto predictivo ayudaría en la planificación y gestión de las provisiones por pérdidas ante el *default* (incumplimiento). Se evalúa objetivamente la información cualitativa y cuantitativa disponible, mejorando la calidad de los análisis por la selección de variables y procedimientos que son relevantes, es decir, que generan resultados contundentes y positivos para las entidades de crédito.

El comité de supervisión de Basilea exige en el análisis de riesgo de crédito, que las entidades financieras tengan un sistema de puntuación para ayudar a tomar las decisiones en el manejo del riesgo y la asignación de capital [1]. Todas las entidades así no estén sometidas a la supervisión de Basilea como lo son las cooperativas financieras, pueden considerar medir el riesgo de crédito por beneficio propio, de la población y de la economía. Los servicios de crédito se brindan para diferentes sectores [18], entre ellos el comercio en general, construcción, servicios públicos, industrial, financiero, minería, agricultura, turismo, transporte y otros. Es por esto, que medir el riesgo evita inestabilidades en una economía y mejora las condiciones financieras. Los préstamos ayudan a mejorar la economía [19].

Este trabajo se realiza porque desde el lado de las entidades del sector solidario hay una necesidad en modelar el riesgo de crédito, a pesar de que el interés de encontrar el modelo robusto se va perdiendo en los investigadores por la gran cantidad de modelos y técnicas empleadas para el tratamiento de este problema [20]. El interés está en fortalecer los modelos convencionales.

Este trabajo de investigación propone utilizar modelos de aprendizaje automático para evaluar el análisis de *Credit Scoring*. La propuesta aporta en la investigación porque desde la revisión de literatura se evidenció que gran parte de las investigaciones alrededor del mundo se centran en *Credit Scoring* para el sector bancario y no tanto para el sector de las microfinanzas. Además de

que no se ha presentado un modelo único para la gestión del riesgo de crédito, son varios los modelos implementados, pero en diferentes contextos como el caso de Australia, Alemania, China, Japón, Estados Unidos, Brasil y otros países donde los resultados de los mismos modelos han mostrado diferencias y se debe a la particularidad de cada contexto.

Los aportes que busca esta investigación están enmarcados en proponer un modelo de puntaje de crédito con técnicas de aprendizaje supervisado que sea útil para las cooperativas de ahorro y crédito en la gestión del riesgo de crédito, que apoye la toma de decisiones en este aspecto, minimizando las pérdidas asociadas al dinero que está en riesgo, a los costos de transacción y procesos para tratar de recuperar la cartera de crédito [21].

1.3 Objetivos

Los objetivos propuestos para el desarrollo de la investigación están en la misma línea de trabajo de las preguntas de investigación formuladas.

1.3.1 Objetivo general

Proponer un modelo de *Credit Scoring* con información de créditos de cooperativas financieras de ahorro y crédito que pertenecen al sector de la economía solidaria de Colombia, basado en modelos de aprendizaje supervisado para el análisis de clientes con riesgo de incumplimiento.

1.3.2 Objetivos específicos

- Analizar la información digital recopilada sobre los créditos en una cooperativa financiera.
- Seleccionar aquellas variables que pueden ser empleadas para la implementación del modelo de *Credit Scoring*.
- Ajustar modelos de aprendizaje supervisado para el análisis de *Credit Scoring* en las cooperativas financieras de ahorro y crédito.
- Evaluar el desempeño de los modelos de aprendizaje supervisado implementados comparando los resultados con métricas de análisis.

1.4 Metodología

En el trabajo se implementa la metodología ASUM-DM (*Analytics Solutions Unified Method*) de IBM la cual está basada en la metodología CRISP-DM (*Cross Standard Process for Data Mining*) para orientar el proceso de minería de datos donde se analizará el problema empresarial como un problema analítico [10].

Las fases de la metodología ASUM-DM son: la comprensión del negocio, enfoque analítico, requisitos y recolección de los datos, entendimiento de los datos, preparación de datos, modelamiento, evaluación, implementación y retroalimentación. Estas implicarán determinar los objetivos de la investigación, recopilar, describir, explorar y verificar la calidad de los datos para seleccionar un conjunto de información para el proceso de limpieza, seguido de la selección de las técnicas de modelado, y posteriormente evaluar y retroalimentar los modelos ajustados.

2. Estado del Arte y Marco Teórico

2.1 Estado del Arte

La revisión de literatura es extensa referente a *Credit Scoring*, los trabajos presentados a lo largo de la última década han estado creciendo, lo que hace que sea necesario mencionar los trabajos más relacionados con el trabajo de investigación propuesto. Hacer esto permite a los investigadores tomar los trabajos de revisión y seguir adelante con sus investigaciones. Aunque en esta investigación se han encontrado varios trabajos importantes de revisión de literatura [11, 12, 17, 18]. Se decidió consolidarlos para describir lo que presentan. El tema principal es *Credit Scoring* y de este se han desagregado varios tópicos que son importantes de revisar. Se encontraron trabajos que presentan sus análisis para un solo tópico y otros que tratan varios de ellos al tiempo, por ejemplo, algunos tratan la reducción de dimensionalidad y también modelos híbridos, otros trabajan máquinas de soporte vectorial o modelos de regresión y otros trabajan árboles de decisión y *Boosting*. A continuación, se presentará los aportes de la revisión de

literatura describiendo en las investigaciones seleccionadas los objetivos que han perseguido los autores para *Credit Scoring* y los aportes que han realizado.

El *Credit Scoring* es un tema que se ha documentado desde 1950. En el transcurso del tiempo se ha basado en un modelo de aprendizaje automático de predicción para estimar la probabilidad de *default* asociada a una operación de crédito. Es un análisis estadístico realizado por prestamistas e instituciones financieras para acceder a la solvencia de una persona. Los prestatarios utilizan la calificación crediticia, para tomar una decisión de asignar o extender un crédito. La necesidad latente ha estado en la implementación de un sistema de evaluación automática para las solicitudes de crédito de los prestatarios [11].

El *Credit Scoring* se ha manejado entonces como la asignación de un puntaje de crédito para calificar la cartera de consumo, vivienda, comercial o microcrédito. En Colombia la clasificación se da en una escala cualitativa:

> "Tratándose de la calificación de cartera, tradicionalmente se ha realizado el análisis teniendo en cuenta la atención del crédito por parte del deudor y factores tales como la solvencia del cliente, sus antecedentes comerciales con el resto del sistema y las garantías otorgadas. Con base en estos criterios se suele calificar la cartera en cinco categorías: (A) Normal, (B) Subnormal, (C) Deficiente, (D) Difícil recuperación, (E) Irrecuperable". [23]

La calificación *A*, son los que presentan un crédito al día, *B* son créditos con mora entre uno y dos meses, *C* para créditos con vencimientos entre dos y tres meses, *D* para créditos con mora entre tres y seis meses y *E* para vencimientos mayores a 6 meses. Esta calificación se da por tiempo en días del vencimiento del crédito.

Estas escalas pueden ser diferentes en cada país, se dan por la regulación que cada uno presenta, lo cual puede ser diferente en China, Taiwán, Indonesia, Estados Unidos, Colombia, México y así sucesivamente. Los modelos aplicados para evaluar esta asignación de puntaje se pueden traducir a probabilidades, donde la cartera menos riesgosa se traduce a una probabilidad de incumplimiento muy baja y en este mismo sentido, la cartera *E* de mayor mora y riesgo, tendría una probabilidad mayor de incumplimiento [17, 24]. Es por ese cambio de escala que el análisis de riesgo de crédito se

puede analizar en la dimensión de la probabilidad de incumplimiento. Sin embargo, hay una variable que se puede asignar y que es la variable más utilizada para el aprendizaje supervisado, es cuando se define que un crédito está en uno de dos posibles estados, “cumplimiento” o “no cumplimiento” o finalmente se califica es al prestatario “bueno” o “malo”, “buen pagador” o “mal pagador” [25]. Es así como se han implementado gran cantidad de modelos de riesgo de crédito para pronosticar la probabilidad de incumplimiento, o clasificando al cliente de buen o mal pagador. La tabla 2-1 muestra la relación de la escala cualitativa con las probabilidades asignadas.

En el caso de los bancos colombianos se ha implementado el modelo de regresión logística, lineal probabilístico y Probit para el análisis de *Credit Scoring* [26, 27]. La regresión logística también hace parte de los modelos de aprendizaje supervisado. Desde la revisión de literatura los modelos que más se han utilizado con fines de clasificación y hallar la probabilidad de incumplimiento se presentan en la tabla 2-3.

Estos modelos y técnicas acorde con los aportes de la estadística, las líneas de investigación, modelos de aprendizaje supervisado y la inteligencia artificial se pueden clasificar en las categorías de la tabla 2-2 y 2-3.

Tabla 2-1 Categorías de riesgo por probabilidad de incumplimiento (en términos porcentuales)

Categoría de reporte	Categoría agrupada	Comercial	Consumo	Vivienda	Microcrédito
AA	A	0-3.11	0-3	0-2	0-3
A	B	> 3.11-6.54	> 3-5	> 2-9	> 3-5
BB	B	> 6.54-11.15	> 5-28	> 9-17	> 5-28
B	C	> 11.15-18.26	>28-40	>17-28	>28-40
CC	C	> 18.26-40.96	>40-53	>28-41	>40-53
C	C	> 40.96- 72.75	>53-70	>41-78	>53-70
D	D	> 72.75-89.89	>70-82	>78-91	>70-82
E	E	>89.89-100	>82-100	>91-100	>82-100

Fuente: Superintendencia Financiera de Colombia [13]

La tabla 2-3 muestra los principales modelos implementados, no obstante, son múltiples las variaciones a estos modelos. Por ejemplo, de las redes neuronales y las máquinas de soporte vectorial se desprenden un número más amplio de modelos, que los investigadores han propuesto con la finalidad de adaptarlos a las situaciones particulares en la búsqueda de mejorar el poder de

predicción y clasificación. Los problemas identificados en el análisis de *Credit Scoring* se documentan a continuación:

- Reducción de dimensionalidad y selección de características o atributos: en esta parte a pesar de que los análisis que se han documentado en la literatura se han hecho con pequeñas y grandes bases de datos [16] con un registro mínimo de 220 observaciones, un máximo de 65524 y un promedio de 5831, con implicaciones de mínimo 4 variables, un máximo de 102 y un promedio de 23. Diversos autores manifestaron que la selección de variables importantes es un desafío en los modelos analizados [24], porque se busca encontrar aquellas variables con mayor poder predictivo y de clasificación. Algunas técnicas como el análisis de componentes principales, análisis de correlación son empleadas para tal fin.

Tabla 2-2 Líneas de investigación, inteligencia artificial implementadas en el período 1994-2020

Agrupamiento de las técnicas por categorías	**Agrupamiento de los modelos y técnicas de aprendizaje automático**
• Análisis univariado y multivariado • Variable dependiente limitada • Métodos probabilísticos • Regresión lineal y no lineal • Estadística no paramétrica • Análisis discriminante • Toma de decisiones con múltiples criterios	❖ Aprendizaje automático basado en reglas ❖ Técnicas de *Boosting* ❖ Inteligencia Artificial

Fuente: [18]

Tabla 2-3 Modelos de aprendizaje automático, técnicas de *Boosting*, inteligencia artificial implementados en el periodo 1994-2020

Modelos empleados en el periodo 1994-2020	Referencias
Análisis discriminante (DA), análisis discriminante cuadrático (QDA), regresión logística (LR), modelo lineal probabilístico (MLP), programación lineal (LP), máquinas de soporte vectorial (SVM), redes neuronales artificiales (ANN), Naive Bayes (NB), Naive Bayes aumentado (TAN), árboles de decisión (DT), C4.5, C5, redes bayesianas (B-Net), CART, bosques aleatorios (RF), modelos de ensamble, ensamble selectivo o híbridos, k-medias, vecino más cercano 10 (KNN10) y 100 (KNN100), perceptrón multicapa (PM), aprendizaje de cuantificación vectorial (LVQ), modelos difusos (FM), algoritmos evolutivos (EA), algoritmos genéticos (GA), optimización de colonia de hormigas (ACO), optimización de enjambre de partículas (PSO), búsqueda Tabú (TS), Optimización de colonias de abejas artificiales (ABCO), búsqueda de armonía (HS), regresión cuantil (QR), evolución diferencial (DE), modelo de cópulas, análisis de datos envolventes (DEA), algoritmo genético híbrido en modelo de puntuación dual, tablas de decisión (DT), detección automática de interacción *Chi Square* (CHAID), *Bagging*, Mapas autoorganizados (SOM), técnica de clasificación para la preferencia de orden por similitud con la solución ideal (CTOPSIS). Criterio de optimización multicriterio (MCOC), análisis factorial (FA).	[11] [12] [13] [14] [15] [16] [17]

- El tratamiento de los datos cuantitativos y cualitativos: gran parte de las variables implementadas en los modelos de aprendizaje automático toman una escala ordinal o nominal, a pesar de que la variable sea cuantitativa o cualitativa [20, 21, 22]. Al tener una variable cuantitativa y agruparla para asignarle una escala ordinal, hace que el número de dimensiones sea más amplio para los análisis, de alguna manera la naturaleza cuantitativa de la variable se deja en manos de la escala ordinal. El número de categorías óptimas para una variable no fue encontrado, aunque se observó entre 2 y 6 categorías.
- Las métricas de desempeño: la valoración del desempeño de los modelos es uno de los resultados principales en los estudios, ya que los resultados de la clasificación, por ejemplo, inciden directamente en los objetivos de las entidades financieras, clasificar mal genera un costo monetario operativo para las entidades de crédito. Las medidas de desempeño más utilizadas fueron error tipo I y II, coeficiente de GINI, medida F, curva ROC, área bajo la curva ROC (AUC), error cuadrático medio y su raíz cuadrada (MSE - RMSE), sensibilidad, especificidad, y la matriz de confusión que indica los porcentajes de buenos y malos clasificados [17, 30]. Se encontró que en algunos trabajos era necesario tener unas métricas que permitan comparar los múltiples modelos que se pueden aplicar.

- Entre otras problemáticas y trabajos futuros mencionados se resaltó la precisión y la interpretación de los modelos, la comparación entre modelos con métricas e hipótesis que permitan contrastar los resultados. Las bases de datos desequilibradas donde el mayor número de observaciones son de clientes que tienen bajo riesgo, es decir, donde el evento de interés que son los malos pagadores tiene poca frecuencia. Esto también tiene en cuenta el uso de técnicas de muestreo y remuestreo que permitan trabajar con la información disponible. La inferencia del rechazo que es el aprovechamiento de la información de clientes que ya fueron rechazados. Pocos datos abiertos para que los investigadores puedan entrenar los modelos.

Lo anterior da cuenta de los retos que enfrenta la implementación de modelos de *Credit Scoring* realizados por las técnicas de aprendizaje supervisado, técnicas de *Boosting* e inteligencia artificial. Las mejores técnicas [21] a lo largo de estos años han sido los árboles de clasificación, regresión logística, redes neuronales, máquinas de soporte vectorial, algoritmos genéticos, modelos difusos y de ensamble.

La regresión logística, los árboles de clasificación y decisión, las máquinas de soporte vectorial y las redes neuronales han sido los modelos de base para *Credit Scoring* los cuales han permitido comparar resultados con los nuevos modelos [17, 29, 30].

El análisis de los datos ha implicado para el estudio de bancos, y de entidades de microfinanzas el uso de variables, algunas de estas son: género, edad, estado civil, nivel educativo, ocupación, tipo de vivienda actual y tiempo de permanencia, tenencia de número de teléfono, número y tipos de productos bancarios, tenencia de vehículo, propósito del crédito, garantías, destino del crédito, monto, tiempo y tasa de interés, información del conyugue, número de personas a cargo, número de veces que la persona ha caído en mora, ingresos, sistema de pensión, gastos, calificación de cartera, nacionalidad, raza y variables macroeconómicas. Estas se pueden agrupar en varias categorías como los son [31]:

- Información demográfica
- Capacidad de reembolso
- Intención de reembolso
- Garantía conjunta
- Ambiente macroeconómico

2.2 Marco Teórico

El marco referencial presenta los materiales y métodos que se emplearán en el trabajo de investigación para el ajuste de los modelos de aprendizaje supervisado en el análisis de *Credit Scoring*.

El Modelo Lineal Probabilístico (MLP) es un modelo para medir una variable binaria, es decir, una variable que se caracteriza por tener distribución Bernoulli, donde la respuesta son dos posibles valores, 0 o 1, los eventos son mutuamente excluyentes, sus probabilidades son $P(Y=1)=p$ y $P(Y=0)=1-p$.

Se considera que cuando $Y=1$ se habla del éxito y cuando $X=0$ del fracaso, que son características de la distribución de probabilidad de esta variable. Sin embargo, el éxito y el fracaso depende del analista, por ejemplo, se desea estudiar la morosidad de un asociado en una cooperativa financiera. Si $Y=1$ indica que la persona está en mora y $Y=0$ que no lo está.

La forma del modelo está en la ecuación (1) que corresponde a un modelo lineal [32]. La interpretación del modelo debe realizarse en términos de la probabilidad de que ocurra el éxito, es decir, $P[Y_i=1|\boldsymbol{X}]=\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\cdots+\hat{\beta}_KX_{ik}$, donde el conjunto de variables $\boldsymbol{X}$ independientes pueden ser tanto cuantitativas como cualitativas.

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_kX_{ik}+\varepsilon_i \qquad (1)$$

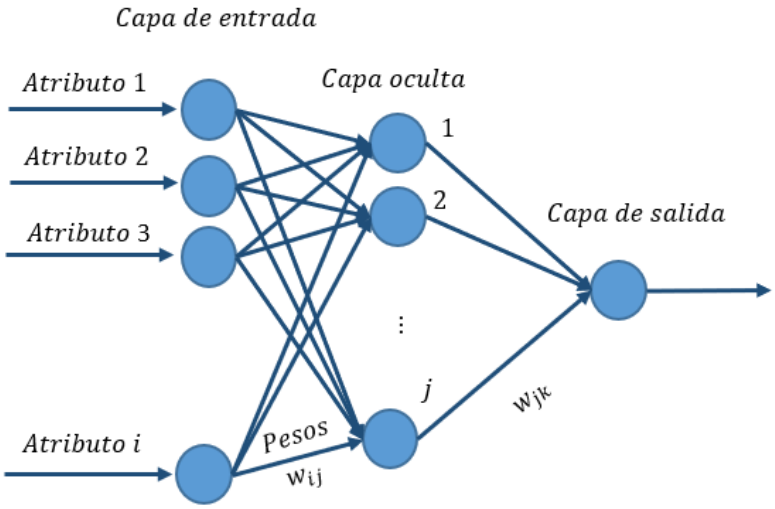
La regresión logística (modelo Logit) es útil para la predicción de una variable dependiente binaria en términos de las variables $\boldsymbol{X}$ que pueden ser también cuantitativas o cualitativas. Este modelo presenta una ventaja frente al modelo MLP debido a que las probabilidades de ocurrencia del éxito se encuentran en el intervalo $[0,1]$ dejando una interpretación más sencilla de la predicción de la ocurrencia del éxito. El modelo toma la forma de la ecuación (2) [32].

$$p_i=P[Y_i=1|\boldsymbol{X}]=\frac{e^{\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_kX_{ik}}}{1+e^{\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_kX_{ik}}} \qquad (2)$$

Por su lado, las redes neuronales son técnicas matemáticas motivadas por las operaciones del cerebro humano. Emplean variables lineales o no lineales en un proceso de entrenamiento a través de las capas de la red neuronal, hasta obtener un mejor resultado en clasificación y predicción. Las

entradas concurrentes de información a la red generan una salida. Un modelo inicial de red neuronal es el perceptrón multicapa, el cual tiene una capa de entrada, una o más capas ocultas y finalmente genera una salida, la topología de la red utiliza una propagación hacia delante o hacia atrás alimentando de información las capas. Un esquema de red neuronal puede ser como se muestra en la figura 2-1, es una forma general de considerar una estructura de red neural.

Figura 2-1 Topología de una red neuronal



Fuente: [19]

Las máquinas de soporte vectorial (SVM) son empleadas como clasificadores y también para el reconocimiento de patrones. SVM busca un hiperplano óptimo con un margen máximo que actúe como límite de decisión, para separar las dos clases diferentes. Dado un conjunto de entrenamiento con pares de instancias etiquetadas (x_i, y_i), donde i es el número de instancias $i = 1, 2, 3, \ldots, m$, $x_i \in \mathbb{R}^n\ (n - dimensional)$ y $y_i \in \{-1, +1\}$. El límite de decisión para separar dos clases diferentes en SVM generalmente se expresa como $wx + b = 0$. El hiperplano de separación óptimo es el que tiene un margen máximo. El problema de optimización convexo se puede definir como se muestra en la ecuación (3). El hiperplano óptimo es equivalente al problema de optimización de una función cuadrática, donde la función de Lagrange se utiliza para encontrar el máximo global. $\in_i$ es

la variable de holgura introducida para tener en cuenta la clasificación errónea, con C como el costo de penalización. [12].

$$min\ \emptyset(w,b) = \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{m} \in_i \quad (3)$$

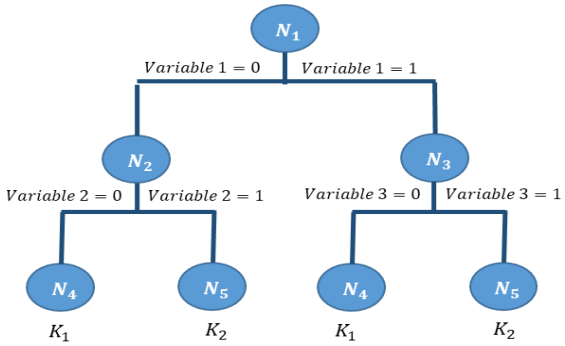
Sujeto a $y_i(wx_i + b) \geq 1$

Los árboles de decisión son también técnicas que se emplean para la calificación crediticia, se conocen como árboles de regresión, clasificación o participación recursiva. Las particiones o nodos del árbol se hacen basadas en reglas para definir si un nodo del árbol será intermedio o terminal. Estos algoritmos tienen en cuenta: (i) establecimiento del número de divisiones admisibles para cada nodo, (ii) definir una regla para declarar un nodo como intermedio o terminal, (iii) asignar a un nodo terminal cada uno de los K grupos con $K = 1,2,3,..,n$ que se definan, (iv) en cada paso de la división escogida permitirá que el nodo sea lo más puro posible, es decir, que no contenga más que los individuos de un solo grupo, (v) definir la regla de pertenencia de un nodo a un grupo lo cual se puede hacer con la regla de Bayes, (vi) se construye un árbol padre, del cual se desprendan otro subárboles anidados, calculando la tasa de mal clasificados en cada árbol anidado y así seleccionar el mejor, esto se puede hacer con validación cruzada [31]. La figura 2-2 muestra un esquema de árbol con un nodo raíz, dos intermedios y 4 terminales.

En las medidas de desempeño para los resultados de la clasificación y predicción se utiliza la matriz de confusión como uno de los criterios para evaluar la clasificación de un modelo, midiendo la proporción de casos clasificados correctamente como buen o mal crédito para un conjunto de datos, existen otras medidas como el coeficiente de GINI, KS, WOE, entre otras.

La curva de características operativas del receptor (ROC) representa en un gráfico bidimensional (plano) la proporción de casos malos clasificados como malos (sensibilidad) frente a la proporción de casos buenos clasificados como malos (1-especificidad). La sensibilidad es la unidad menos la tasa de error tipo II y la especificidad es igual a la unidad menos la tasa de error tipo I. En la tabla 2-4 se muestra la matriz de confusión.

Figura 2-2 Esquema de construcción de un árbol



Fuente: [33]

Tabla 2-4 Matriz de confusión

	Predicción positiva	**Predicción negativa**
Clase positiva	Verdadero positivo (VP)	Falso Negativo (FN)
Clase negativa	Falso positivo (FP)	Verdadero negativo (VN)

Fuente [30]

La proporción de predicciones verdaderas $(PPV = VP/(VP + FP))$ indica la predicción correcta de los créditos buenos frente al total de predicciones positivas, y la proporción de predicciones falsas $(PPV = VN/(VN + FN))$ indica la predicción correcta de los créditos malos frente al total de la predicción negativa y la precisión total $PT = (VP + VN)/(VP + FP + +FN + VN)$ indica la proporción correcta de todo el modelo.

Gran parte de las técnicas de base del aprendizaje supervisado como las mencionadas se han centrado en clasificación, algunas en regresión también [25]. Actividades adjuntas a estas tareas han implementado métodos de agrupamiento, entre ellos el K-vecino más cercano, ampliamente utilizado para apoyar estas técnicas, al ser un método de agrupamiento, implementa una medida de distancia, esto es en general para los métodos de agrupamiento como métodos jerárquicos. Estos métodos realizan fusiones sucesivas de los subconjuntos de un conjunto, es decir, métodos aglomerativos o ascendientes, o por divisiones sucesivas, es decir, métodos divisivos o descendientes. La ecuación (4) [33] muestra una familia de distancias conocida como Minkowski,

donde x_i^j es el valor del atributo j para la observación i, esta familia se construye para $r \geq 1$, si $r = 1$ se habla de la distancia *city block*, si $r = 2$ se refiere a la distancia euclidiana clásica.

$$d(i,i') = \sqrt[r]{\sum_{j=1}^{p} |x_i^j - x_{i'}^j|^r} \quad \text{con } j = 1, \dots, p \; atributos \text{ y r} \geq 1 \qquad (4)$$

La dimensionalidad en un conjunto de datos se refiere tanto al número de variables o atributos y al número de observaciones. Cuando el número de variables y de observaciones es grande se habla de datos con alta dimensionalidad. No necesariamente tener mucha información es lo necesario, posiblemente se encuentre información redundante que es necesario analizarla. La alta dimensionalidad genera problemas de interpretación de los fenómenos o resultados generados en un análisis de datos. Por ejemplo, al realizar un método de agrupamiento con la distancia euclidiana con alto número de variable y de observaciones, es posible realizar los cálculos en un software de apoyo, aunque el costo computacional puede ser bastante alto, y la interpretación será cada vez más compleja. En el aprendizaje supervisado se pueden aplicar varios métodos para reducir dimensionalidad, entre ellos el análisis de correlación, el análisis de componentes principales, los métodos Biplot, el análisis de correspondencias múltiples, entre otros. [22, 24, 33].

3. Caso de estudio

3.1 Preparación y análisis de datos

El sector solidario en Colombia está conformado por varias organizaciones, entre ellas cooperativas financieras, fondos de empleados, asociaciones mutuales, entre otras. El sector se ha venido desarrollando y estructurando desde el año 1931. La Superintendencia de la Economía Solidaria (Supersolidaria) es la entidad estatal que actualmente vigila y regula la actividad de estas organizaciones, otras entidades como el Fondo de Garantías del Sector Cooperativo (Fogacoop) también apoyan la actividad solidaria. La confianza del sector solidario es reflejada en el número de asociados a las entidades que lo conforman. El fortalecimiento de esta actividad se da cuando la dinámica de las operaciones que se desarrollan alcanza sus objetivos a nivel organizacional.

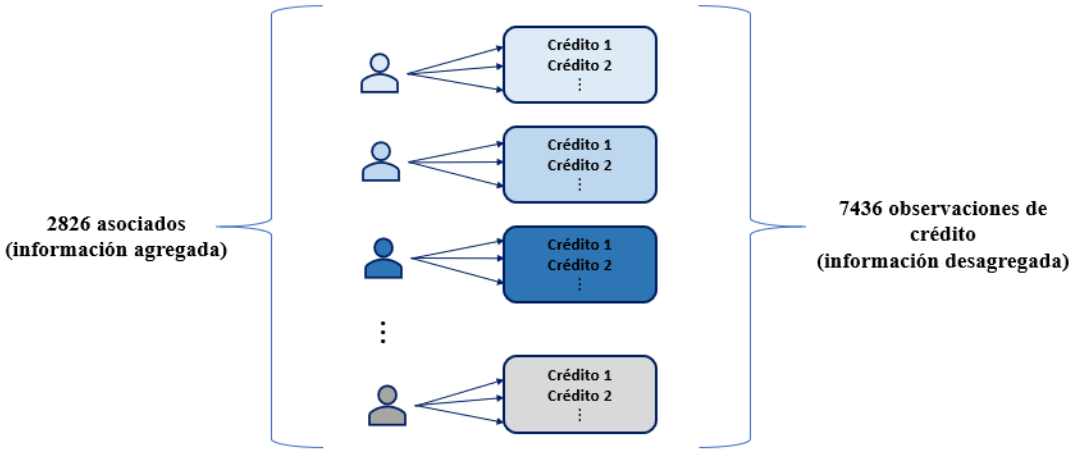
La Supersolidaria realiza un seguimiento al sector para detallar el comportamiento de la cartera de crédito, la distribución de éstos al igual que otras cuestiones como los ingresos, aportes, créditos y ahorros de los asociados en las entidades. La gestión del riesgo de crédito es parte de los requisitos que la Supersolidaria implementa. Las diferentes organizaciones bajo los parámetros y directrices que se les brindan desarrollan e implementan modelos de supervisión para la administración de este riesgo. El desarrollo de modelos analíticos que permitan monitorear el riesgo de crédito dará apoyo a las entidades del sector solidario para fortalecer su actividad operacional y mantener un crecimiento continuo.

Dentro de los análisis que se deben considerar, se trabaja el análisis de la clasificación de la cartera de crédito, la escala está en 5 categorías de riesgo, *A*, *B*, *C*, *D* y *E*, los cuales califican los registros de crédito de las entidades en los niveles normal, aceptable, apreciable, significativo y de incobrabilidad, respectivamente. La categoría *A* valora una situación de bajo o poco riesgo, y en ese sentido la categoría *E* valora una situación con alto riesgo, incluso incobrable. Categorizar en estas medidas se da acorde a los días de mora de créditos que están vencidos o atrasados entre 0 y 180 días. Esto implica que la información que se analice debe estar actualizada con relación a la vigencia y vencimiento de los créditos para poderlos valorar y asignar un puntaje de crédito.

El conjunto de datos que se utiliza para el caso de análisis contiene 20 atributos. La información está consolidada y actualizada para los asociados que se encuentran activos y con productos de crédito de la entidad financiera. Los atributos dan información del historial de los créditos y del estado de cuenta actual.

La información se divide en dos subconjuntos. El primero tiene en cuenta lo que se llamará “información desagregada” y hace referencia a un total de 7436 registros o solicitudes de crédito aprobadas que se encuentran activas. El segundo subconjunto será la “información agregada”, es la información de los 2826 asociados que tienen activa al menos una operación de crédito. Es decir, que los 7436 créditos que están distribuidos en los 2826 asociados. Esta información es necesaria trabajarla por separada, porque permitirá describir el comportamiento de los créditos y de los asociados para posteriormente dar un puntaje de crédito a cada asociado. A continuación, se describen los atributos que permitan obtener los metadatos de este conjunto de información:

Figura 3-1 Información agregada y desagregada del conjunto de datos



Fuente: Elaboración propia

3.2 Comprensión de los datos

- **Sexo**: Atributo cualitativo nominal con dos categorías, sexo masculino y femenino.
- **Edad**: Edad en años del asociado. La edad admisible es igual o superior a los 18 años, valores menores se limpian y superiores a 100 años se evaluar.
- **Estado civil**: Atributo cualitativo con 7 categorías. Los datos extraños son aquellos que no reportan el estado civil, y se clasifican en “ninguno”.
- **Profesión**: Profesión u ocupación actual del asociado. Se tienen 13 categorías que son complejas de combinar porque no se logra encontrar más afinidad entre las profesiones.
- **Antigüedad en la entidad**: Este atributo se mide en años y da cuenta de los asociados que llevan más tiempo en la entidad.
- **Salario**: Atributo numérico que muestra los ingresos mensuales recibidos por el deudor (asociado). Permite inferir su capacidad de pago. Valores muy extremos son posibles.
- **Aportes sociales**: Recaudo que la entidad recibe por cada asociado, valor que se acumula con el tiempo. El único caso en que se admita un valor de cero en el aporte es a razón de que el asociado

no tiene una fuente de ingreso que le permita aportar o porque se ha retirado de la entidad, pero su compromiso de crédito permanece.

- **Tipo de recaudo**: Atributo cualitativo que indica el medio de pago de las obligaciones de deuda. Puede ser recaudo por nomina (libranza) o por caja.
- **Línea de crédito**: Indica el nombre del destino del crédito. Este atributo cuenta con 11 categorías. En la información desagregada un asociado puede tener varias líneas de crédito. Es uno de los atributos quedan información relevante a la entidad porque permite llevar un control del destino la cartera de crédito de sus asociados para el diseño de estrategias de colocación de dinero y también de gestión y control interno ante las entidades regulatorias.
- **Cantidad de líneas de crédito**: En la información desagregada de las líneas de crédito, un asociado puede aparecer con varios créditos y con la posibilidad de que cada uno sea de una línea de crédito sea diferente. Se propuso la creación de este atributo de forma cualitativa, con la categoría “una” o “varias” líneas de crédito, y también de forma cuantitativa para conocer el total y así para poder agregar la información de los asociados.
- **Periodicidad de pago**: Es un atributo cualitativo con dos categorías la cual puede ser mensual o quincenal. Es admisible que un asociado con varios créditos pueda tener periodicidades de pago diferentes. Permite tener control de los pagos, hacer recaudos oportunos y gestionar la liquidez de la entidad financiera.
- **Cuota**: Valor que se paga quincenal o mensualmente. Los únicos créditos con cuota cero son los créditos de sobreprima, debido a que se pagan en un solo momento con las primas.
- **Monto inicial del crédito**: Es un atributo numérico que indica el valor del monto por el cual fue solicitado el crédito. A diferencia de la línea de crédito de sobreprima, los demás créditos no deben tener valores extraños de las cuotas, por ejemplo, cuotas en cero
- **Saldo de capital**: Es el valor que cada asociado adeuda al fondo por cada registro. puede estar en la base de datos repetidas veces por la tenencia de varios créditos, y la suma del saldo del capital se puede observar en la base de datos agregada.
- **Porcentaje de participación**: Un asociado puede tener varios créditos con montos iguales o diferentes, consecuente con esto, se totaliza todos los montos. El monto inicial del cada crédito se divide sobre este total y así se obtiene el porcentaje de participación.
- **Tasa de interés de ponderación**: Cada crédito tiene una tasa de interés efectiva anual. El producto de la tasa efectiva y el porcentaje de participación genera una tasa de interés ponderada para el crédito. Este atributo se crea con interés de estudiar su incidencia en el riesgo de crédito.

- **Costo promedio ponderado de capital**: Es la suma de las tasas de interés ponderación de los créditos de cada asociado.
- **Plazo en meses**: Hace referencia al tiempo en meses concedido a cada asociado para el pago de sus obligaciones por cada crédito.
- **Tasa efectiva anual**: Es la tasa de interés que se aplica a cada crédito. Varía por los acuerdos que se dan según la línea de crédito, el tiempo, el monto, el nivel de riesgo, entre otras.
- **Capital inicial total**: Este atributo totaliza la suma de todos los montos iniciales de los créditos de cada asociado.
- **Saldo de capital total**: Totaliza el valor del saldo total que adeuda cada asociado por todos los créditos que tenga activos en la entidad financiera.
- **Días en mora**: En la información desagregada cada crédito tiene sus días de mora y en la agregada cada asociado tiene la suma de los días en mora de todos los créditos que posee.
- **Cuotas en mora**: Número de cuotas atrasadas de los pagos que tiene cada asociado por cada crédito.
- **Total días y cuotas en mora**: Totalizan los días y cuotas en mora de cada asociado por todos los créditos que tenga en la entidad financiera.
- **Garantía**: Hace referencia al nombre de la garantía con la cual se está respaldando el crédito en caso de que el asociado incumpla con el pago de la obligación. Este atributo es cualitativo con 7 categorías. Un asociado puede tener varias categorías que respalden sus créditos, pero para cada uno se le asocia solo una garantía.
- **Cantidad Garantías**: Hay diferentes tipos de garantía que puede tener un solo asociado, por lo que, en la base de datos agregada, se crea el atributo que cuenta el número total de garantías para cada asociado, se construye de forma cualitativa, es decir, se indica “una” o “varias”, también se construyó de forma cuantitativa.
- **Número de créditos**: Se calcula el número de créditos de cada asociado, atributo que se puede ver en la información desagregada pero su utilidad se da más en la información agregada.
- **Proporción de deuda:** Es la razón entre el saldo de capital total y el capital inicial total.
- **Calificación de cartera**: Indica la calificación de riesgo de cada asociado según su comportamiento de pago crediticio, cartera por días en mora: *A* de 0-30, *B* de 31-60, *C* de 61-90, *D* de 91-180 y *E* más de 180.

En las categorías *B*, *C*, *D* y *E* se encuentran créditos con más de 30 días en mora. Esto se debe a la regla de arrastre descrita en la Circular Básica Contable y Financiera No. 004 de 2008 que dice que para poder constituir la respectiva provisión de cartera cuando el fondo califique cualquiera de los créditos de un deudor en categoría *B*, *C*, *D* y *E* deberá llevar a la categoría de mayor riesgo los demás créditos así se encuentren al día (regla de arrastre). Todos los créditos de un asociado deben tener la misma calificación de cartera.

- **Deudor riesgo:** Atributo que identifica con el valor de 0 si el asociado que tiene deuda no presenta riesgo, es decir, que está en la escala *A* de la calificación de cartera, y si está en la escala *B*, *C*, *D* o *E* entonces, el atributo valdrá 1. Este atributo se crea con fines de clasificación y predicción.

3.3 Organización y análisis de los datos

Previo al proceso de modelamiento se realiza una preparación del conjunto de datos. Éste se utiliza para la construcción de la información que será procesada y utilizada para el modelamiento. Las actividades para desarrollar son:

- Eliminación de inconsistencias y errores de información.
- Análisis de valores faltantes para determinar datos nulos.
- Implementación de estrategias para recuperar información faltante.
- Construcción de nuevos atributos.

3.3.1 Limpieza de datos

- La construcción de los metadatos y del análisis de los atributos se hizo en conjunto con el personal de la entidad financiera quienes indicaban los detalles para construir la descripción. En el sexo los valores faltantes se imputaron hallando los asociados que aparecían varias veces, y los que tenían ambos sexos se verificaron con información interna de la empresa. Así también se hizo con otras variables en casos faltantes y con valores diferentes.
- El estado civil “madre soltera” eran pocos registros y se decidió unificar con la categoría “Soltero”.

- En la profesión, la categoría “Sin profesión” no se logró imputar por la alta cantidad de registros, se trabajará como una categoría. La categoría de “Terapeuta” se unificó con “Salud” pues solo tenía 3 registros de los 7436. Así quedaron 13 de 18 categorías.
- Un asociado tenía el valor de los aportes negativos, se corrigió este error. En el tipo de recaudo 213 registros tenían el recaudo en blanco, solo 47 de ellos se lograron recuperar porque se identificó al asociado, los demás fueron actualizados nuevamente por la entidad.
- En la periodicidad de pago, aparecía una categoría con el nombre “catorcenal” se tomó la decisión con la entidad de unificarla con la categoría “quincenal” pues la diferencia en el pago era de un día.
- La línea de crédito “Vivienda” e “Inversión inmobiliaria” se agruparon en la categoría “Inmobiliaria”. Las líneas de “Emergencia”, “Ex asociados”, “Transición”, “Fomento” y “Accionario” se agruparon en la categoría “Otras Líneas”. Los agrupamientos se hicieron porque estas líneas tenían pocas observaciones.
- Para validar la cuota y el ingreso del asociado, se calculó la razón entre la cuota y el ingreso, valor que debería ser menor que 1, solo dos asociados tenían esta razón mayor a 1. Se validó con la entidad financiera y este registro se dio en condiciones especiales.
- En el monto inicial del crédito aparecían créditos que la cuota tenía valores extraños, es decir, valores que no alcanzaban a pagar el crédito en los plazos determinados, a pesar de que la información de la cuota se podía calcular porque se conoce la tasa de interés, el monto y el plazo, la recomendación por parte de la persona que facilitó la información era no contar con esos registros que en total fueron 8.
- Un registro en el monto inicial del crédito tenía un valor de 5.600.000 cuando en realidad era de 56.000.000, se detectó debido a que el cociente entre el saldo que se adeudaba sobre el monto inicial era mayor, y esto era una inconsistencia.
- Se construye el atributo de “porcentaje de participación”, “tasa de interés de ponderación” y “Costo promedio ponderado de capital” los cuales son de carácter financiero para analizar el riesgo de crédito. Esto se da a razón de incluir variables que puedan ser útiles y recomendadas por las entidades regulatorias.
- En el atributo de plazo en meses, se tomó la decisión de retirar los créditos con valores inferiores a 1 mes, pues se desea analizar aquellos que tienen un plazo mayor a 30 días para pagar.

- Se crearon los atributos “Capital inicial total”, “Saldo de capital total”, “Total días en mora” y “Total cuotas en mora”, para agregar la información por asociado.
  - En la información desagregada se encontraron 16 registros con días de mora entre 0 y 30, a los cuales se le actualizó a la categoría de cartera a la *A*, 12 registros a cartera *B*, 21 a cartera *C*, 24 a cartera *D* y 11 registros a cartera *E*.
  - En el atributo de garantías se observó inicialmente 10 categorías, se agruparon las que tenían como denominación un “codeudor” en una sola categoría, quedando así 7, no se decidió agrupar más debido a que éstos brindan información relevante para la entidad financiera.
  - Se creó el atributo “número de garantías” para contar cuántas respaldan los créditos de cada asociado. Por ejemplo, un asociado con 6 créditos lo respaldan 4 garantías (Aportes, Fondo solidario de garantías, prenda de vehículo y sin garantía). También se creó de manera cualitativa “Una” o “Varias categorías”.
  - En la creación del atributo de la cantidad de créditos por asociado, se observó que varios de los asociados se encontraban duplicados, se hizo un proceso de validación y se eliminaron registros duplicados.
  - La calificación crediticia estaba desactualizada con respecto a algunos asociados, es decir, si un asociado tiene la calificación más alta en “C”, entonces todos sus créditos deben estar en esa categoría.

3.3.2 Construcción de nuevos datos

La información inicial contaba con 20 atributos (variables), en el proceso de limpieza fue necesario crear los atributos que se describieron, porque permiten validar que la información sea consistente y, por otra parte, y acorde al negocio son necesario para el análisis del puntaje de crédito.

Los atributos iniciales y los que se construyeron están incluidos en la lista de la información que se debe considerar acorde con la Circular Externa No. 007 de la Superintendencia de la Economía Solidaria para la rendición de cuentas relacionadas con la actividad crediticia. Una de las variables a las que no se logró tener acceso fue el valor de las garantías diferentes a la de aportes sociales.

La clasificación previa de los atributos permite entender el contexto de donde provienen. Las variables sexo, edad, estado civil, profesión y salario son consideradas demográficas, las demás como el monto inicial del crédito, tasa de interés efectiva anual, plazo, número de cuotas del crédito y cuotas en mora, etc., son variables financieras y crediticias.

En la tabla 3-1 se muestran los atributos cualitativos todos en escala nominal, a excepción de la calificación de cartera que es ordinal. Una lectura de esta tabla indica para el caso del sexo que hay 7436 solicitudes de crédito, 4769 de éstas son de personas masculinas las cuales representan 1852 asociados de los 2826 que tienen de crédito.

En la tabla 3-2 se tiene el listado de los atributos o variables cuantitativas. Una de las columnas identifica el rango en el que se observó el registro de las 7436 solicitudes de crédito una vez realizado el proceso de limpieza de datos y creación de nuevos atributos. La descripción de las variables y la información de la tabla 3-1 y 3-2 son parte de los metadatos. La tabla 3-3 muestra la clasificación que se le dio a los atributos para el estudio, y se indica si éste será considerado para la exploración de los datos en forma desagregado o agregado.

3.3.3 Prueba ji-cuadrado

Se emplea la prueba ji-cuadro para contrastar la hipótesis de independencia entre dos variables categóricas (hipótesis nula). Esta prueba utiliza las frecuencias observadas de la muestra en el cruce de las categorías del par de variables. Si el valor del estadístico de prueba χ_0^2 es pequeño, más probable será que las variables sean estadísticamente independientes (o lo mismo que el *p*-valor sea alto), es decir, no se rechazaría la hipótesis de independencia, en caso contrario se rechazaría.

El número de combinaciones de tamaño 2 de las 11 variables cualitativas genera 55 tablas de contingencia bidimensionales, esto para el caso de la información desagregada. La tabla 3-4 ilustra el caso de la variable “línea de crédito” y “calificación de cartera” para la información desagregada, el gran total son los 7436 de registros de solicitudes de créditos. En la información agregada se construyó el atributo “número de líneas de crédito” en dos categorías (una o varias) y la tabla 3-5 muestra la distribución de frecuencias, el gran total es 2826 que representa el número de asociados. Nótese que en las 11 categorías desagregadas de la línea de crédito hay 3751 solicitudes de créditos en la categoría *A* de la calificación de cartera, esta cantidad de solicitudes corresponde a 1588 asociados. El número de tablas de contingencia bidimensionales para la información agregada son 36 resultado de la combinatoria de las 9 variables cualitativas como se muestra en la tabla 3-3.

Tabla 3-1 Variables cualitativas del conjunto de datos

Variable	Categorías	Etiqueta	Desagregado (Por créditos)	Agregado (Por asociado)
Sexo	Masculino	S_M	4769	1852
	Femenino	S_F	2667	974
Estado_Civil	Casado	E_C_C	3870	1466
	Divorciado	E_C_D	197	73
	Mujer Cabeza Familia	E_C_M	17	10
	Separado	E_C_Se	119	42
	Soltero	E_C_So	2229	859
	Union Libre	E_C_UL	866	323
	Viudo	E_C_V	138	53
Profesión	Administrador	PR_Ad	501	174
	Analista	PR_An	93	31
	Asistente	PR_As	180	67
	Auxiliar	PR_Au	256	80
	Contador	PR_Co	330	122
	Especialista	PR_Es	113	40
	Ingeniero	PR_In	2509	998
	Operario	PR_Op	97	33
	Profesional	PR_Pr	686	256
	Salud	PR_Sa	55	17
	Sin profesion	PR_SP	1948	765
	Tecnico	PR_Tco	57	21
	Tecnologo	PR_Tgo	611	222
Tipo_Recaudo	Caja	TR_C	1292	571
	Nomina	TR_N	6144	2255
Linea_Credito	Compra cartera	LC_CC	1119	839
	Diamante libre inversion	LC_DLI	451	305
	Dirigido	LC_D	464	403
	Educativo	LC_E	609	504
	Inmobiliaria	LC_I	1042	859
	Libre inversion	LC_LI	1535	1108
	Otras_lineas	LC_O	74	59
	Rotativo	LC_R	1472	656
	Sobreprima	LC_S	208	167
	Vacacional	LC_Va	50	50
	Vehiculo	LC_Ve	412	393
Cantidad_lineas_credito	Una	LC_U	2022	1275
	Varias	LC_V	5414	1551
Periodicidad_de_Pago	Quincenal	PE_Q	1574	2139
	Mensual	PE_M	5862	687
Garantia	Aportes	G_A	1090	765
	Codeudor	G_C	405	312
	Fondo Solidario de Garantias	G_FS	2455	1382
	Hipoteca Primer Grado	G_HP	806	523
	Hipoteca Segundo Grado	G_HS	690	487
	Prenda de Vehiculo	G_PV	320	286
	Sin Garantia	G_SG	1670	793
Cantidad_garantias	Una	G_U	2578	1464
	Varias	G_V	4858	1362
Calificacion_Cartera	A	CC_A	3751	1588
	B	CC_B	480	230
	C	CC_C	418	180
	D	CC_D	2638	780
	E	CC_E	149	48
Deudor_en_riesgo	No	0	3751	1588
	Si	1	3685	1238

Fuente: Elaboración propia

Tabla 3-2 Variables cuantitativas del conjunto de datos

Variable	Rango	Etiqueta	Desagregado (Por créditos)	Agregado (Por asociado)
Edad	22-93	Edad	7436	2826
Antigüedad entidad	0.31 - 48.90	Antiguedad	7436	2826
Salario	391000 - 95000000	Salario	7436	2826
Aportes Sociales	0 - 57792806	A_Social	7436	2826
Cantidad lineas credito num	1-6	C_Lineas	7436	2826
Cuota	0 -9481137	Cuota	7436	2826
Monto inicial del credito	100 - 626062320	Monto_I	7436	2826
Saldo Capital	1 - 626062320	Saldo_C	7436	2826
Porcentaje Participacion	0.00034-1	P_Participacion	7436	2826
Tasa de interés de ponderación	0.00005 - 0.21150	T_I_Ponderacion	7436	2826
Costo promedio ponderado	0.03230-0.21150	C_P_Ponderado	7436	2826
Plazo en meses	2-184	Plazo	7436	2826
Tasa EA	0,02-0,2172	Tasa_EA	7436	2826
Capital inicial total	120000-817665372	C_I_Total	7436	2826
Saldo capital total	1-748059509	S_C_Total	7436	2826
Dias mora	0 - 450	Dias_M	7436	2826
Cuotas mora	0 - 31	Cuotas_M	7436	2826
Total dias mora	0 - 9550	T_Dias_M	7436	2826
Total cuotas mora	0 - 201	T_Cuotas_M	7436	2826
Cantidad garantias num	1-4	Cantidad_G	7436	2826
Cantidad creditos	1-7	Cantidad_C	7436	2826
Proporción de deuda	0-1	P_Deuda	7436	2826

Fuente: Elaboración propia

Tabla 3-3 Clasificación variables del conjunto de datos

Atributo	Tipo	Categorías	Clasificación	Información desagregada	Información agregada
Sexo	Cualitativo	2	Demográfica	X	X
Estado_Civil	Cualitativo	7	Demográfica	X	X
Profesion	Cualitativo	13	Demográfica	X	X
Tipo_Recaudo	Cualitativo	2	Financiera y crediticia	X	X
Linea_Credito	Cualitativo	11	Financiera y crediticia	X	
Cantidad_lineas_credito	Cualitativo	2	Financiera y crediticia	X	X
Periodicidad_de_Pago	Cualitativo	2	Financiera y crediticia	X	X
Garantia	Cualitativo	7	Financiera y crediticia	X	
Cantidad_garantias	Cualitativo	2	Financiera y crediticia	X	X
Calificacion_Cartera	Cualitativo	5	Financiera y crediticia	X	X
Deudor_riesgoso	Cualitativo	2	Financiera y crediticia	X	X
Edad	Cuantitativo	No aplica	Demográfica	X	X
Antiguedad_entidad	Cuantitativo	No aplica	Financiera y crediticia	X	X
Salario	Cuantitativo	No aplica	Demográfica	X	X
Aportes_Sociales	Cuantitativo	No aplica	Financiera y crediticia	X	X
Cantidad_lineas_credito_num	Cuantitativo	No aplica	Financiera y crediticia	X	X
Cuota	Cuantitativo	No aplica	Financiera y crediticia	X	
Monto_inicial_del_credito	Cuantitativo	No aplica	Financiera y crediticia	X	
Saldo_Capital	Cuantitativo	No aplica	Financiera y crediticia	X	
Porcentaje_Participacion	Cuantitativo	No aplica	Financiera y crediticia	X	
Tasa_Int_Ponderacion	Cuantitativo	No aplica	Financiera y crediticia	X	
Costo_promedio_ponderado	Cuantitativo	No aplica	Financiera y crediticia	X	X
Plazo_en_meses	Cuantitativo	No aplica	Financiera y crediticia	X	
Tasa_EA	Cuantitativo	No aplica	Financiera y crediticia	X	
Capital_inicial_total	Cuantitativo	No aplica	Financiera y crediticia	X	X
Saldo_capital_total	Cuantitativo	No aplica	Financiera y crediticia	X	X
Dias_mora	Cuantitativo	No aplica	Financiera y crediticia	X	
Cuotas_mora	Cuantitativo	No aplica	Financiera y crediticia	X	
Total_dias_mora	Cuantitativo	No aplica	Financiera y crediticia	X	X
Total_cuotas_mora	Cuantitativo	No aplica	Financiera y crediticia	X	X
Cantidad_garantias_num	Cuantitativo	No aplica	Financiera y crediticia	X	X
Cantidad_creditos	Cuantitativo	No aplica	Financiera y crediticia	X	X
Proporcion_deuda	Cuantitativo	No aplica	Financiera y crediticia	X	X

Fuente: Elaboración propia

En los resultados de las 55 pruebas ji-cuadrado de la información desagregada (combinatoria de 11 tomados de 2). La variable que mostró independencia (donde no se rechazó la hipótesis nula)

fue el sexo con el tipo de recaudo, cantidad de líneas de crédito, periodicidad de pago, cantidad de garantías, calificación de cartera y deudor riesgoso, donde los *p*-valores fueron mayores a 0.05.

La figura 3-2 muestra que la cartera *A* y *D* son las que tienen mayor proporción de los asociados, aun así, en las diferentes categorías de la calificación de cartera se mantiene una mayor proporción de masculinos que de femeninos y esto hace que no se rechace la hipótesis nula de independencia. El mismo resultado se observó con el sexo y la variable deudor riesgoso.

Tabla 3-4 Tabla de contingencia de la información desagregada del conjunto de datos para la línea de crédito y calificación de cartera

	Calificación de cartera					
	A	**B**	**C**	**D**	**E**	Total
Compra cartera	616	73	55	368	7	1119
Diamante libre inversion	320	45	21	64	1	451
Dirigido	170	19	34	238	3	464
Educativo	371	41	34	156	7	609
Inmobiliaria	534	60	64	373	11	1042
Libre inversion	644	126	111	623	31	1535
Otras_lineas	43	7	10	11	3	74
Rotativo	798	52	44	498	80	1472
Sobreprima	81	17	10	98	2	208
Vacacional	23	5	4	18	0	50
Vehiculo	151	35	31	191	4	412
Total	3751	480	418	2638	149	7436

Fuente: Elaboración propia

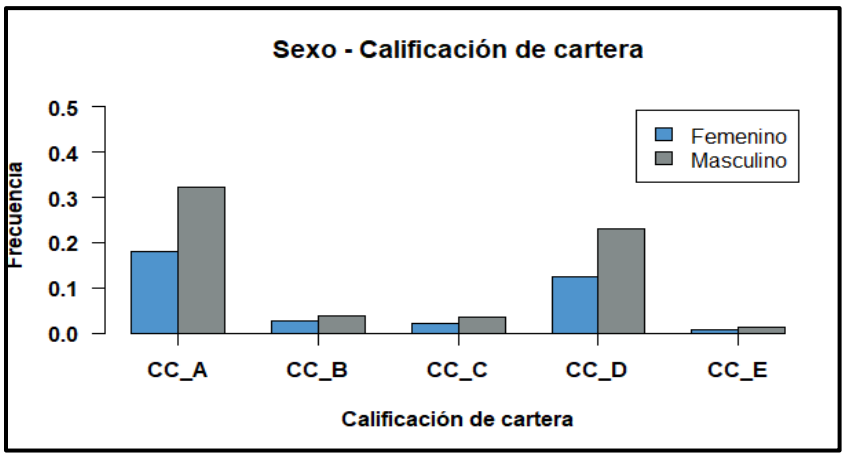
Tabla 3-5 Tabla de contingencia de la información agregada del conjunto de datos para la cantidad de líneas de crédito y calificación de cartera

	Calificación de cartera					
	A	**B**	**C**	**D**	**E**	Total
Cantidad de líneas de crédito	847	138	85	173	32	1275
	741	92	95	607	16	1551
Total	1588	230	180	780	48	2826

Fuente: Elaboración propia

La variable calificación de cartera y deudor riesgo mostraron un rechazo de la hipótesis nula con todos los atributos cualitativos a excepción del sexo. Recordando también que en la calificación de cartera *A* se calificaron como no riesgosos y cartera *B*, *C*, *D* y *E* son los riesgosos. Esto implica que para este estudio la variable sexo no está mostrando diferencias o asociaciones con el riesgo.

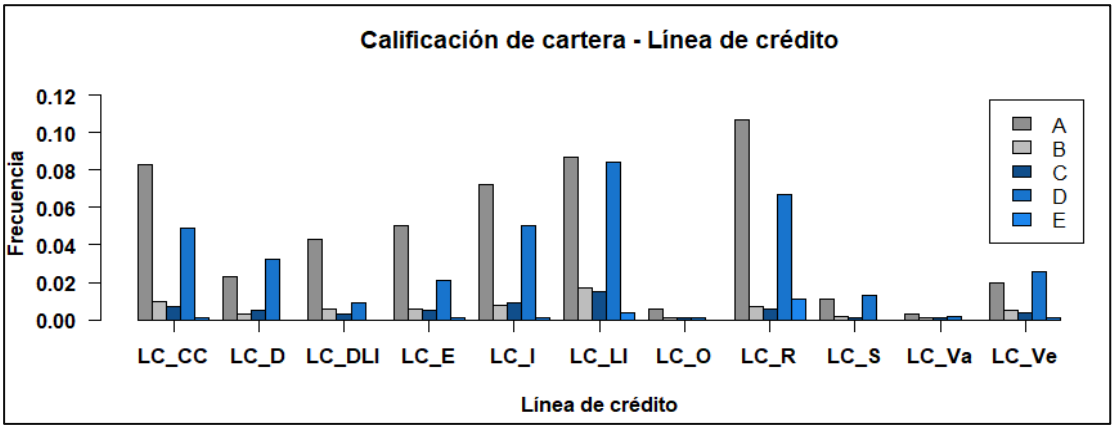
Figura 3-2 Perfil del sexo en la calificación de cartera del conjunto de datos



Fuente: Elaboración propia

Se rechazó la independencia entre la línea de crédito y la calificación de cartera, es decir, que hay una asociación entre estas variables y sus categorías. La línea de crédito con once categorías de compra de cartera (LC_CC), diamante libre inversión (LC_DLI), dirigido (LC_D), educativo (LC_E), inmobiliaria (LC_I), libre inversión (LC_LI), etc.) La variable calificación de cartera mostró dependencia estadística con las demás variables cualitativas a excepción del sexo.

Figura 3-3 Perfil de la línea de crédito en la calificación de cartera del conjunto de datos



Fuente: Elaboración propia

La cartera *A* está muy asociada con la compra de cartera, con diamante libre inversión, educación, inmobiliaria, créditos rotativos. La cartera *B* con diamante libre inversión, libre inversión, créditos de sobre prima, vacacionales y de vehículo. La cartera *C* con créditos dirigidos, inmobiliaria, libre inversión, vacacional y de vehículos. La cartera *D* con compra de cartera, créditos dirigidos, inmobiliaria, libre inversión, de sobreprima y vehículo. Por su parte la cartera *E* muestra asociación con créditos dirigidos, de libre inversión y rotativos. Acorde a la prueba de independencia, tiene que la línea de crédito muestra diferencias entre las proporciones de sus categorías y las de cartera de crédito, es decir, que hay una asociación entre ellas.

A este punto se podría continuar al proceso de técnicas de aprendizaje supervisado sin la variable sexo por la independencia que mostró con la mayoría de las variables del conjunto de datos y en especial con la variable cartera de crédito y deudor riesgoso, sin embargo, se considerará trabajar con ella para un análisis más exhaustivo de su incidencia en el puntaje de crédito.

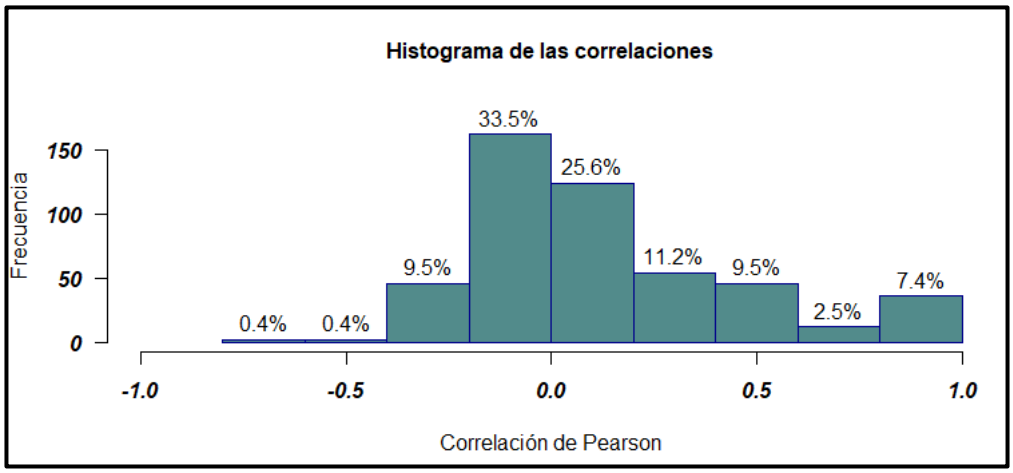
3.3.4 Análisis de correlación

El coeficiente de correlación de Pearson (ρ_{xy}) es una medida del grado de asociación lineal entre un par de variables cuantitativas. En el conjunto de datos se tienen 22 atributos, el total de

correlaciones serían 231 ($(22 \times 22 - 22)/2$). Un valor alto de ρ_{xy} en valor absoluto indica que hay una relación lineal fuerte entre las variables.

En la figura 3-4 se muestra que cerca del 10% de las correlaciones son superiores a 0.5, es decir, que gran parte de las variables no muestran una correlación lineal. Por otra parte, la tabla 3-6 indica la correlación de los once pares de variables que mayor correlación tenían. A mayor capital inicial adeudado se tenía también un saldo de capital total mayor, es decir, que todavía hay asociados que solicitan montos de dinero y que aún los tienen pendientes por pagar. A mayor salario mayor es el aporte del asociado a la entidad solidaria.

Figura 3-4 Histograma de las correlaciones de los atributos



Fuente: Elaboración propia

La tasa de interés de un crédito se pondera acorde con todos los créditos que tenga un asociado en la entidad, y ésta se calcula a partir de la participación por eso son atributos con una correlación alta. Es consistente que el total de días y cuotas en mora tengan correlación alta, pues una variable está en función de la otra. En el caso de las dos correlaciones negativas se da en el sentido de que, si un asociado tiene por ejemplo un crédito, su porcentaje de participación es del 100%, en cambio sí tiene dos créditos el porcentaje de participación de cada uno depende del monto, y ahí ya no sería 100 porciento, por eso estas correlaciones deben ser negativas.

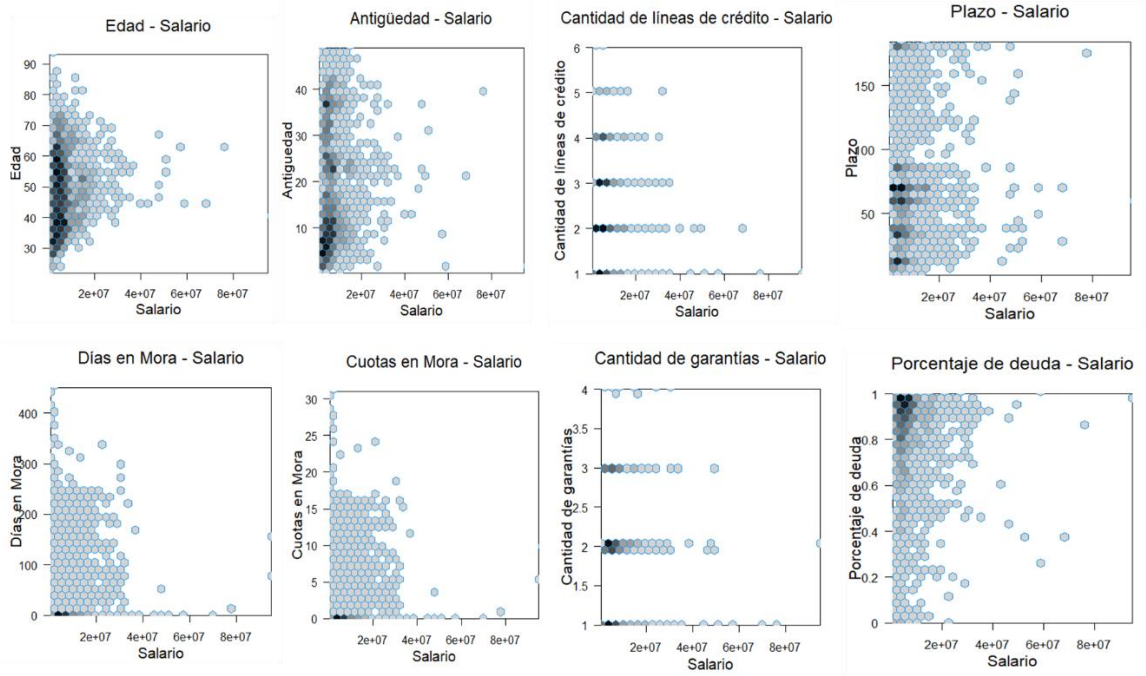
Tabla 3-6 Correlaciones significativas

Variable x	Variable y	Correlación ρ_{xy}
Capital inicial total	Saldo de capital total	0,97
Saldo de capital	Monto inicial	0,97
Cantidad de días en mora	Cantidad de cuotas en mora	0,96
Tasa de ponderación	Porcentaje de participación	0,94
Total cuotas en mora	Total días en Mora	0,83
Edad	Antigüedad	0,82
Monto Inicial	Cuota	0,81
Aportes	Salario	0,78
Saldo de capital	Cuota	0,75
Cantidad de créditos	Tasa de interés de ponderación	-0,57
Cantidad de créditos	Porcentaje de participación	-0,61

Fuente: Elaboración propia

Variables como el salario con la edad, antigüedad, cantidad de líneas de crédito, plazo de los créditos, días de mora, no mostraron una asociación lineal significativa. La figura 3-5 muestra que no hay una relación lineal entre el ingreso y las variables que se han puesto de ejemplo. Sin embargo, hay relaciones que se pueden considerar. Gran parte de los asociados son personas entre los 30 y 60 años con ingresos bajos, éstos muestran tener mayor concentración en tener hasta tres líneas de crédito. Esto permite entender que además de una asociación lineal (correlación) entre las variables es posible detectar otros patrones de comportamiento en los datos para las variables que tendrán incidencia en el puntaje de crédito para un asociado.

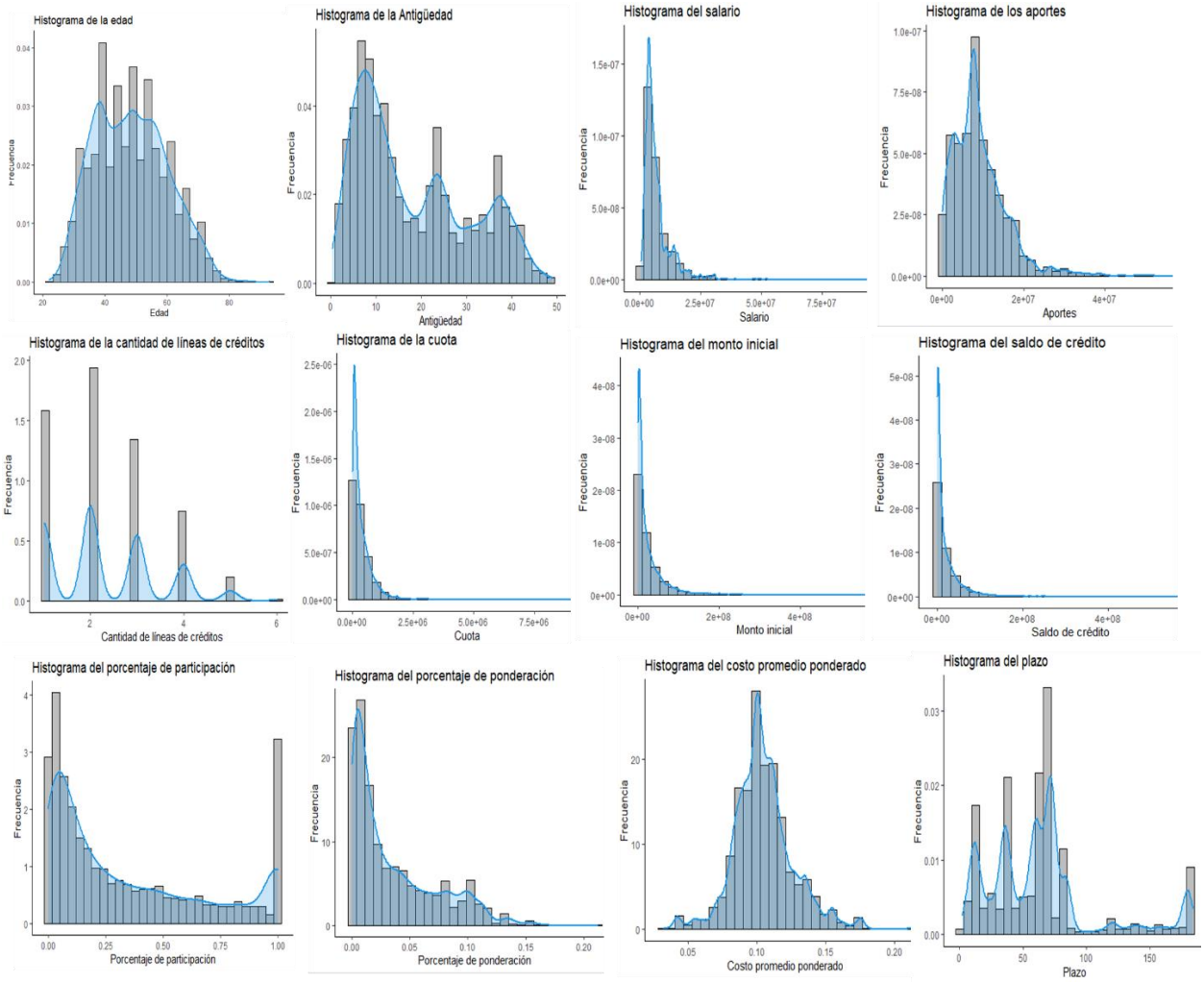
Figura 3-5 Dispersión entre el salario y otras variables cuantitativas



Fuente: Elaboración propia

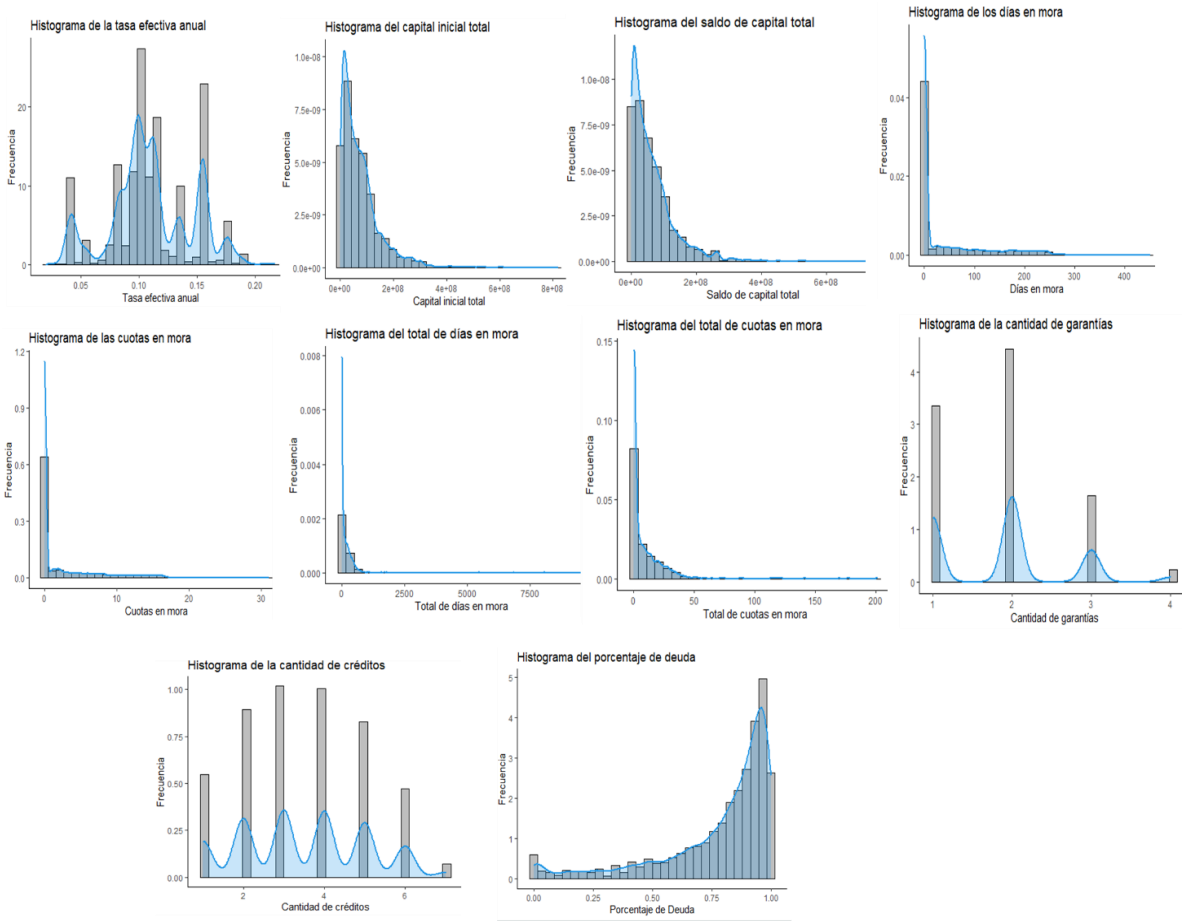
La figura 3-6 y 3-7 muestran la distribución de las variables continuas mediante el histograma de frecuencias. Algunas variables como el ingreso, los aportes, el saldo de capital total tienen formas asimétricas positivas. Lo que se observa con los ingresos que hay muchos asociados con ingresos medios o bajos y pocos con ingresos altos. El porcentaje de deuda muestra una asimetría negativa, pues muchos tienen porcentajes de deuda altos (saldos de capital por pagar). La variable que mostró una forma más simétrica fue el costo promedio ponderado.

Figura 3-6 Histograma de la distribución de variables cuantitativas (a)



Fuente: Elaboración propia

Figura 3-7 Histograma de la distribución de variables cuantitativas (b)



Fuente: Elaboración propia

3.4 Exploración y evaluación de los datos

3.4.1 Análisis de variables cualitativas

En las variables cualitativas se calcula un índice de variación ($IVar$) el cual considera calcular la frecuencia de la moda (n_{moda}) y dividirla sobre el total de datos de la muestra (n), esto indica la proporción de datos que están en la categoría con mayor frecuencia.

Por otra parte, las demás categorías que no están en la moda se llevan el resto de la proporción, este se calcula $IVar = 1 - n_{moda}/n$. Si el $IVar$ es cercano a 0 es porque la moda tiene una frecuencia casi igual al total de los datos, dejando una menor frecuencia a las demás categorías de la variable. Por el contrario, si el $IVar$ es cercano a 1, significa que la moda es pequeña en comparación con el total de datos [35]. En general indica el grado en el que las frecuencias de las categorías de una variable no coinciden con el de la moda, a menor valor del índice de variación menor número de categorías tiene la variable (menor variación).

La tabla 3-7 indica que la variable cualitativa "Línea de crédito" tiene un índice de variación $IVar$ de 0.794 es decir, que hay una alta variabilidad o dispersión respecto de las demás categorías comparado con la moda de esta variable. Además, varias de las variables tienen solo dos categorías, aun así, se calcula el $IVar$, y en general, se ha notado que hay poca variación entre las categorías de cada variable.

Tabla 3-7 Índice de variación de las variables cualitativas del conjunto de datos

Nombre variable	Numero categorías	Índice de variación
T_Recaudo	2	0.174
Periodicidad_P	2	0.212
Cantidad_L_C	2	0.272
Cantidad_G	2	0.347
Sexo	2	0.359
E_Civil	7	0.48
C_Cartera	5	0.496
Deudor_R	2	0.496
Profesión	13	0.663
Garantía	7	0.67
Linea_C	11	0.794

Fuente: Elaboración propia

En el análisis de las variables cualitativas el índice de variación ($IVar$) permite mirar que tan dispersas están las frecuencias en las categorías para cada variable. En la tabla 3-7 se observa que hay mayor dispersión en la variable, línea de créditos, garantía y profesión, esto es porque no hay una concentración de una frecuencia alta en alguna categoría de las variables. Por otro lado, en las

11 variables cualitativas se analizó el estadístico ji-cuadro, concluyendo que la variable sexo no mostró dependencia estadística con varias variables incluida la calificación de cartera y deudor riesgoso.

3.4.2 Análisis de correspondencias

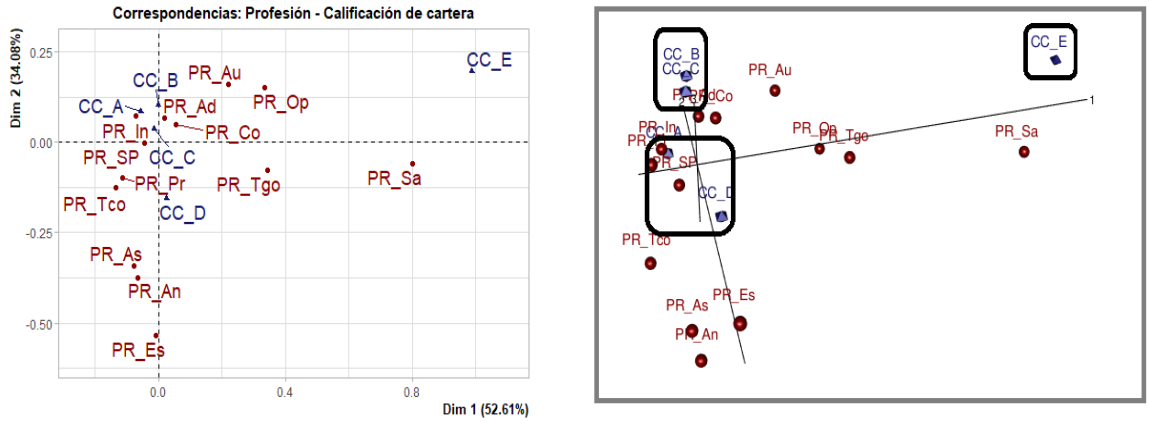
En el Análisis Factorial de Correspondencias (AFC) se busca examinar la asociación o relación estadística entre dos variables cualitativas a través de sus categorías. La extensión de este método a más de dos variables categóricas se conoce como el Análisis de Correspondencia Múltiple (ACM). En ambos tipos de análisis se obtiene la inercia, que es una medida de la asociación en variables cualitativas [35].

El análisis de correspondencias representa las variables, las categorías y las observaciones (solicitudes de crédito) en ejes coordenados ayudando analizar patrones en los datos.

Análisis Factorial de Correspondencias (AFC)

La variable calificación de cartera se codificó en la variable deudor riesgoso (Si=1 y No=0). Se comparó la calificación de cartera con las demás variables cualitativas y se notó que hay dependencia con estas. Obsérvese que en la figura 3-8 se ilustra un análisis factorial de correspondencias en el cruce de la variable cartera de crédito y profesión.

Figura 3-8 Análisis Factorial de Correspondencias entre la profesión y la cartera de crédito



(a) Vista en un plano (Dimensiones 1 y 2) (b) Vista en 3D (Dimensiones 1, 2, y 3)

Fuente: Elaboración propia

En la figura 3-8 se muestra mayor asociación de la cartera de crédito *E* con los profesionales de la salud (cercanía en la representación gráfica). Por su parte la cartera *D* se asocia más con los ingenieros, profesionales y personas sin profesión. La cartera *C* con administradores, ingenieros, sin profesión. La cartera *B* con ingenieros, sin profesión, auxiliares, contadores y administradores y la cartera *A* con ingenieros, administradores, sin profesión y auxiliares.

Con el Análisis Factorial de Correspondencias (AFC) como técnica analítica propia de estas variables, se analiza aquellas categorías que dan señales de patrones en los datos, propiamente aquellas que muestran mayor inercia. En la tabla 3-8 se analiza la calificación de cartera con la variable sexo (primera fila), se indicó que no hay diferencia en la inercia o asociación por la categoría masculino o femenino (no se indica el valor de la inercia solo se describe el resultado). Con la variable tipo de recaudo (fila 4) la categoría recaudo por caja indicó mayor asociación con las categorías de la calificación de cartera. De la misma manera, se hizo con las demás variables.

Tabla 3-8 AFC - Categorías con mayor contribución a la inercia con la calificación de cartera

Variable cualitativa	Categorías con mayor inercia
Sexo	Las correspondencias son similares para ambos sexos en las diferentes categorías de la cartera de crédito
Estado_Civil	Casados, solteros y en unión libre
Profesión	Salud, auxiliar, tecnólogo, analista, asistente, especialista, ingeniero
Tipo_Recaudo	Caja
Linea_Credito	Dirigido, diamante libre inversión, rotativo, compra de cartera
Cantidad_lineas_credito	Varias
Periodicidad_de_Pago	Mensual
Garantía	Sin garantía, aportes, fondo solidario, hipoteca de primer y segundo grado y codeudor
Cantidad_garantias	Varias

Fuente: Elaboración propia

Análisis de Correspondencias Múltiples (ACM)

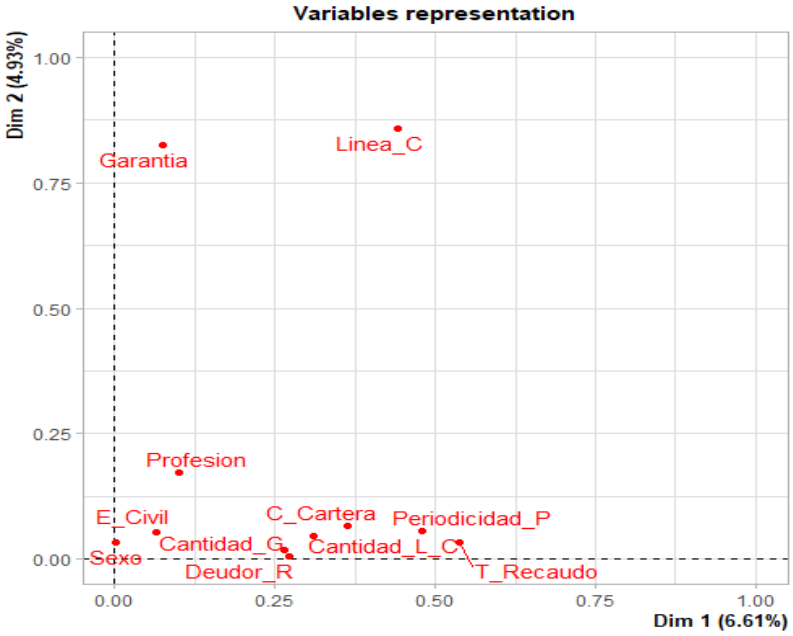
La inercia explicada por los primeros ejes seguirá siendo la máxima, sin embargo, al aplicar un ACM con muchas variables y muchas categorías la inercia explicada es pequeña, a razón de que se requieren varias dimensiones para representar toda esta información.

Se realiza un análisis de correspondencia múltiple con las 11 variables cualitativas. La implicación en utilizar todas, es que la inercia explicada sería pequeña. Sin embargo, las primeras dimensiones logran absorber la mayor inercia (asociación) y es ahí donde se puede lograr una mejor interpretación de las correspondencias.

En ACM se analiza la nube de variables (representación de variables), nube de las categorías de las variables y nube de individuos (solicitudes de crédito de los prestatarios). En la figura 3-9 se muestra la nube de variables, el primer plano representa las dos primeras dimensiones las cuales explican un 11.54% de la inercia total. Se observa que la variable garantía y línea de crédito están asociadas a la dimensión 2 y variables como la calificación de cartera, periodicidad de pago, cantidad de línea de crédito, y tipo de recaudo con la dimensión 1. Para una mejor interpretación se analiza la nube de categorías.

En las 11 variables hay 55 categorías, la interpretación gráfica de estas puede ser compleja (figura 3-10). Se analiza las categorías que sean más representativas (que estén alejadas del origen del plano).

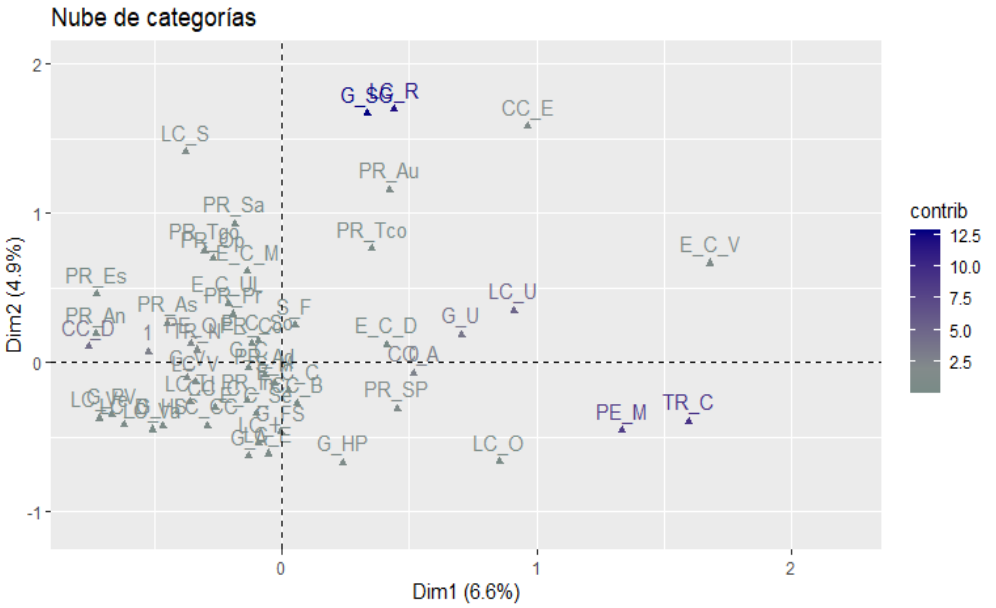
Figura 3-9 Nube de variables cualitativas comparativas del conjunto de datos



Fuente: Elaboración propia

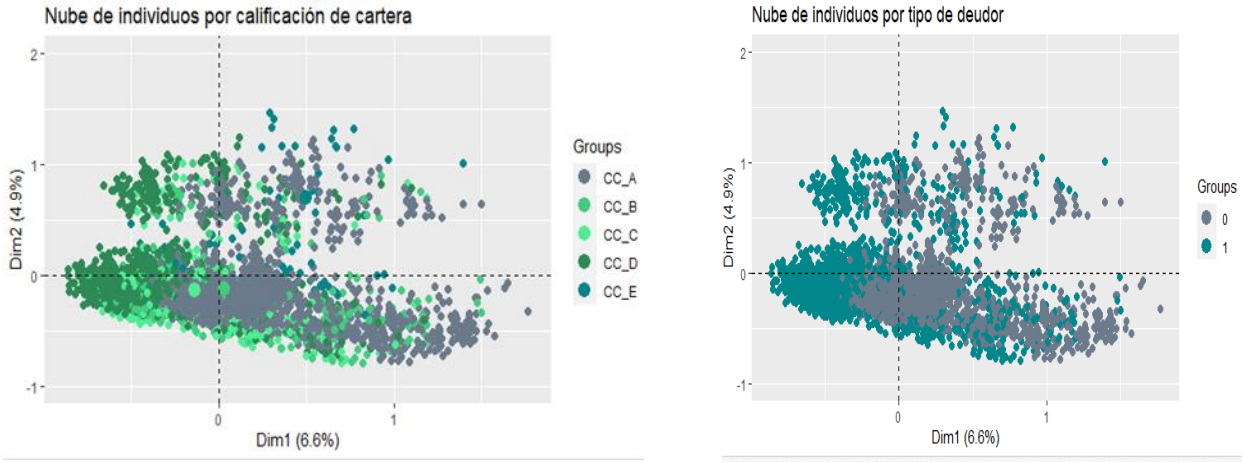
En la figura 3-11 se grafica los individuos, los que están cercanos es porque tienen atributos muy similares en las variables que los describen. Al igual que, dos categorías están cercanas (se corresponden) cuando los mismos individuos que toman valores en una categoría lo hacen también en la otra, esto en categorías de variables diferentes (figura 3-10). En el caso de las categorías de la misma variable su cercanía se da porque las frecuencias suelen ser similares, pero son excluyentes porque un individuo se identifica solamente en una. Las categorías que toman frecuencias bajas se encuentran alejadas del origen del plano, además representan un valor alto de inercia y a su vez son influyentes en el ACM.

Figura 3-10 Nube de categorías de las variables cualitativas



Fuente: Elaboración propia

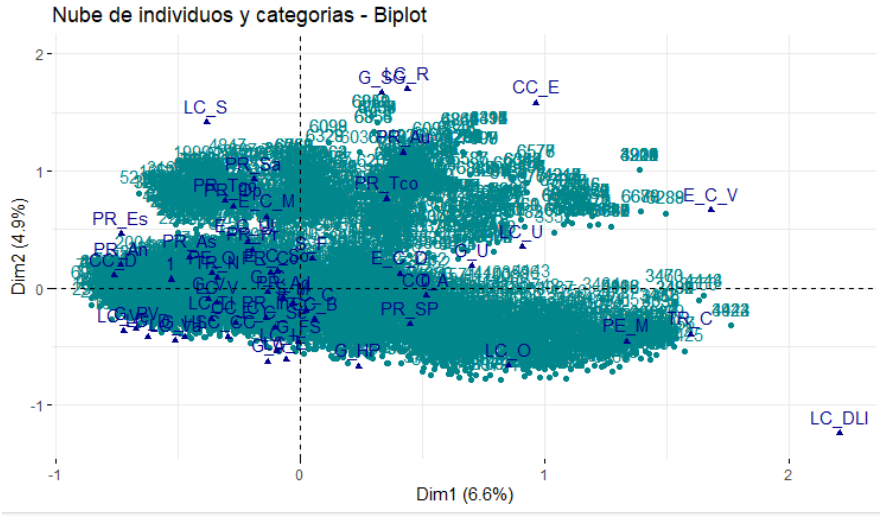
Figura 3-11 Nube de individuos de las variables cualitativas



(a) Resaltando la calificación de cartera (b) Resaltando por tipo de deudor

Fuente: Elaboración propia

Figura 3-12 Nube de categorías e individuos de las variables cualitativas



Fuente: Elaboración propia

La tabla 3-9 describe lo que ocurre en las figuras 3-9 a 3-12. Es una manera de interpretar la forma como se asociaron las categorías de las variables cualitativas del estudio y la forma como se correspondieron los individuos (solicitudes de crédito). Los cuadrantes están indicados en el sentido de las manecillas del reloj.

Del ACM se concluye que las categorías de las variables tipo de recaudo, cantidad de líneas de crédito y periodicidad de pago tienen una asociación o correspondencia entre sus categorías y así mismo, la línea de crédito y las garantías.

Tabla 3-9 Descripción del ACM

	Categorías	Individuos	Descripción
Cuadrante 1	10	1263	Femenino, divorciados, viudos, auxiliares, técnicos, con crédito rotativos, con una línea de crédito, sin garantía o con una garantía y cartera de crédito *E*
Cuadrante 2	16	1054	Mujer cabeza de familia, soltero, en unión libre, analista, asistente, contador, especialista, Operario, profesional, salud, tecnólogo, recaudo por nomina, línea de crédito de sobreprima, recaudo quincenal, cartera de crédito *D*, si son deudores.
Cuadrante 3	19	3205	Masculino, separado, administrador, ingeniero, con línea de crédito para compra de cartera, crédito dirigido, inmobiliaria, libre inversión, vacacional y de vehículo, es decir, lo que tienen varias líneas de crédito, con garantía de los aportes sociales, codeudor, fondo solidario, hipoteca de segundo grado, prenda de vehículo, es decir, con varias garantías y con cartera *C*
Cuadrante 4	10	1914	Casado, sin profesión, con recaudo por caja, cartera de diamante libre inversión y otras líneas, con recaudo mensual, con garantía de hipoteca en primer grado, con cartera *A* y *B* y no riesgosos
Total	55	7436	

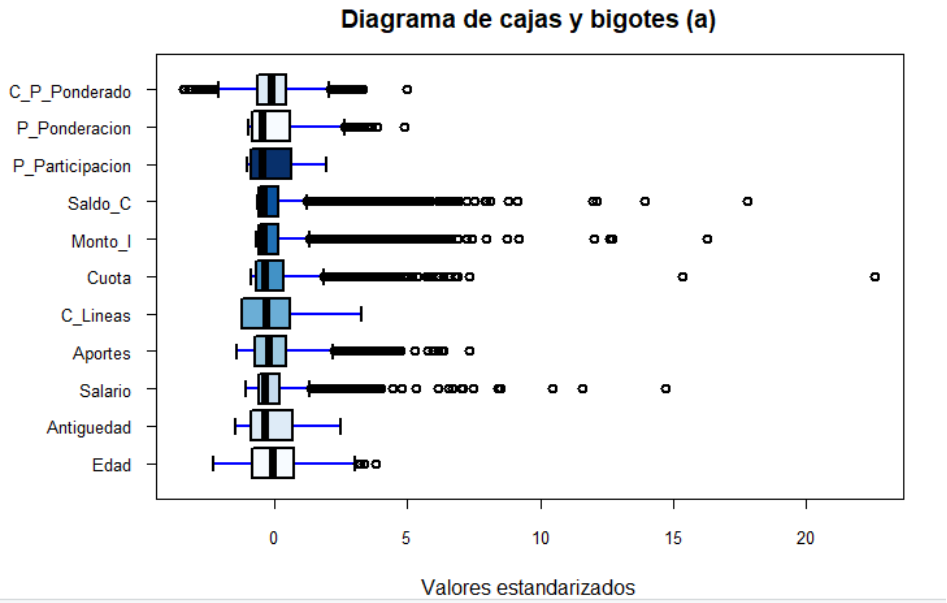
Fuente: Elaboración propia

3.4.3 Análisis de diagrama de cajas y bigotes

Los atributos cuantitativos del conjunto de datos se pueden explorar mediante diagrama de cajas y bigotes. Para comparar entre atributos en un solo gráfico, los datos deben estar estandarizados por las diferencias en las unidades o escalas de medida. Los diagramas de cajas y bigotes se muestran en las figuras 3-13 y 3-14, las variables que no presentaron datos atípicos son el porcentaje de participación, cantidad de líneas de crédito, antigüedad, cantidad de créditos. La edad, cantidad de garantías, y la tasa efectiva anual son las variables que muestran pocos datos atípicos y no muy alejados de los bigotes, y las demás variables con un número amplio de datos atípicos. Estos datos no son errores de la información porque en el proceso de limpieza quedaron justificados, obedece entonces a que así se comporta la distribución de la información.

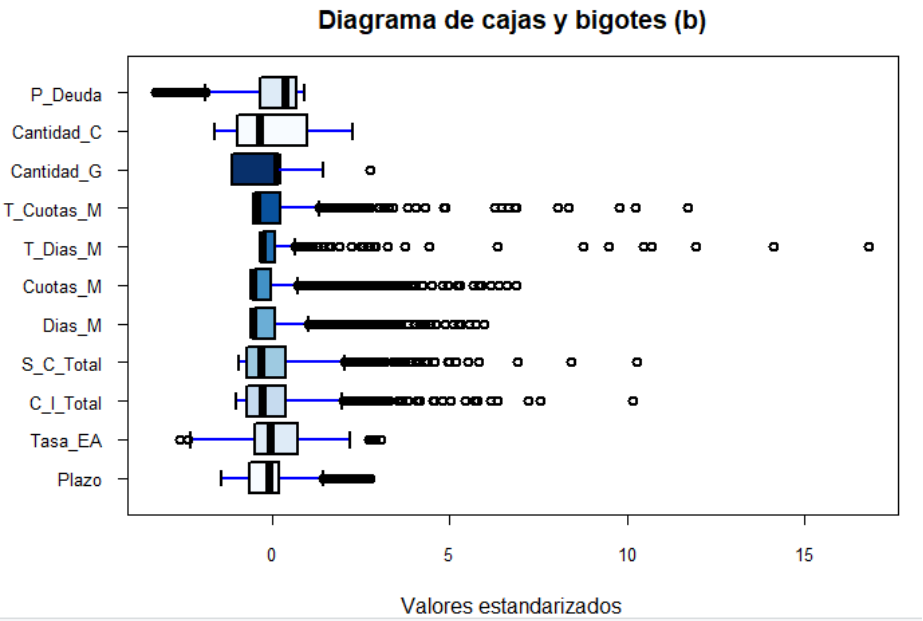
La cantidad de líneas de crédito, porcentaje de participación y ponderación, antigüedad, edad, cantidad de créditos y garantías, el saldo del capital inicial y saldo de cuenta total, la tasa efectiva anual y el plazo son variables que muestran niveles de variabilidad amplios. Se nota que la variable del costo promedio ponderado es la única que tiende a tener una distribución simétrica, las demás con asimetrías.

Figura 3-13 Diagrama de cajas y bigotes de las variables cuantitativas (a)



Fuente: Elaboración propia

Figura 3-14 Diagrama de cajas y bigotes de las variables cuantitativas (b)



Fuente: Elaboración propia

3.4.4 Métodos Biplot

Los métodos Biplot son una herramienta para comprender la composición de grandes volúmenes de datos. Son una representación en un mismo plano de individuos y variables, el objetivo es presentar la información contenida en una matriz de datos, en un espacio de menor dimensión con la menor pérdida de información, captando patrones entre conjuntos de individuos y variables, correlación entre variables y formación de grupos de individuos. El Biplot tiene la característica de reproducir aproximadamente el dato.

Por la alta dimensión de la información, estos métodos se vuelven importantes y se acompañan de una medida de ajuste global o medida de bondad de ajuste (en términos de dispersión – calidad de la representación) multidimensional que explica la aproximación de la alta dimensión en un espacio de menor dimensión.

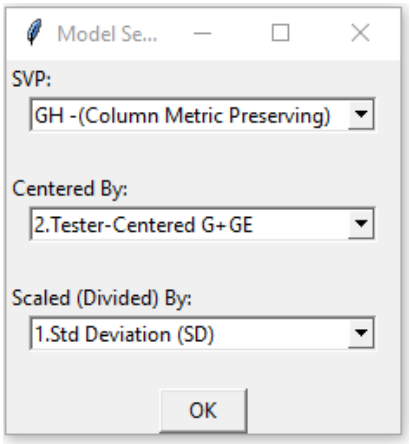
En las matrices de datos, las columnas de datos representan las variables y las filas las observaciones o individuos. En el caso de estudio, las columnas son todas las variables

cuantitativas como el salario, la edad, los días en mora, etc., y las filas representan cada una de las solicitudes de crédito. El Biplot de una matriz de datos $\boldsymbol{X}$ es una factorización $\boldsymbol{X = AB}$ siendo $\boldsymbol{A}$ los marcadores fila **y** $\boldsymbol{B}$ los marcadores columna. Los marcadores se hallan con la descomposición en valores singulares de $\boldsymbol{X}$**.**

Se distinguen tres tipos de Biplot clásicos: El *Column Metric Preserving* o GH-Biplot que conserva la correlación entre variables y es útil para cuando el interés son las variables. El *Row Metric Preserving* o JK-Biplot que es útil para centrarse en el estudio de los individuos. Una concentración en el estudio de las variables y los individuos al mismo tiempo se hace con el HJ-Biplot.

Se utiliza el paquete GGEBiplotGUI del software R-Project (véase la figura 3-15) el cual realiza el estudio de los diferentes Biplots. La figura 3-16 muestra el porcentaje de variación explicada por las diferentes dimensiones de la información, las dos primeras dimensiones explican el 98.82%, eso quiere decir que, en un plano formado por las dos primeras dimensiones se obtiene la mayor representación de los individuos y variables. El coseno del ángulo que se forma entre los vectores (marcadores columna) representa la correlación, y la longitud del vector es una aproximación de su desviación estándar. El GH Biplot se muestra en la figura 3-16, el primer plano factorial explica 41.47% de la varianza.

Figura 3-15 Paquete GGEBiplotGUI para Biplot



Fuente: Elaboración propia

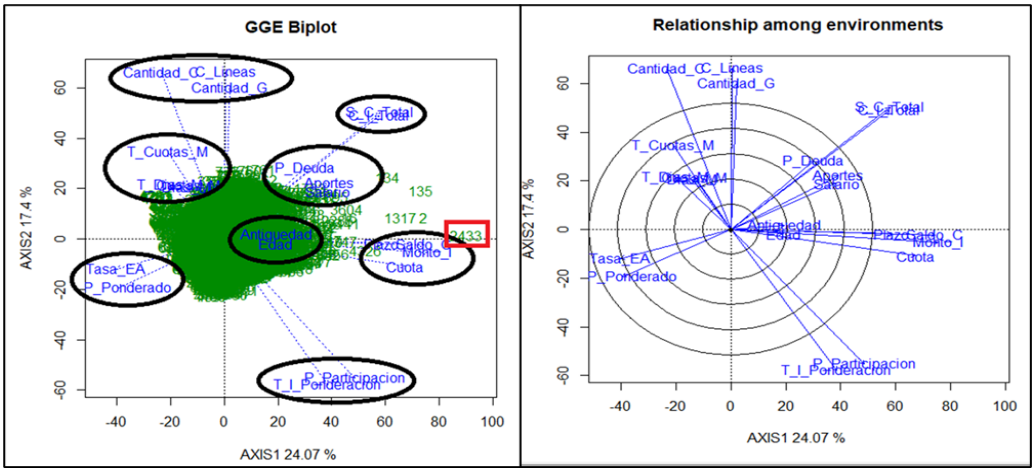
Las variables que se encuentran alejadas del centro del plano están bien representadas, la edad y antigüedad son las que posiblemente están mejor representadas en otras dimensiones, pero en

general se observa que con las dos primeras componentes se obtiene un gran porcentaje de varianza explicada. La tabla 3-10 muestra las varianzas y desviaciones estándar de las variables, se puede notar que algunas como el capital inicial total (C_I_Total) y el saldo de capital total (S_C_Total) tienen mayor varianza, y variables como las tasas de interés, tasa efectiva anual y costo promedio ponderado (Tasa_EA y C_P_Ponderado) lo contrario, esto se da por la unidad de medida de cada atributo, esto es influyente a la hora de aplicar métodos Biplot, por eso, se normalizan los atributos para que permanezcan en una misma escala y que eso ayude a interpretar bien el gráfico.

Otros de los resultados encontrados, es que el coseno del ángulo entre un par de variables representa su correlación, variables como el total de días en mora y la cantidad de días en mora tienen una correlación de 0,8313 y por eso en el Biplot están cercanas, otras como la cantidad de líneas de crédito (C_Lineas), cantidad de garantías (Cantidad_G) y cantidad de créditos (Cantidad_C) tienen una correlación considerable. Esto permite reducir aún más la dimensionalidad, porque se podría trabajar con la variable del número de días en mora o del número de cuotas en mora, pues ambos dan un nivel de información muy similar.

En la figura 8-17 se resaltan 8 posibles grupos de variables que tienden a tener alta correlación, y a efectos de aplicar modelos de aprendizaje supervisado, se podría implementar una variable de cada grupo, decisión que se tomará más adelante. Téngase también en cuenta que la mayoría de las observaciones (nube de puntos en color verde) está muy concentrada al origen, algunas observaciones, por ejemplo, la 2433 (de las 7436) se resalta en un recuadro rojo, está indica que toma valores altos en esas variables, por ejemplo, la observación es de las que tiene una cuota de crédito más altas. Para el caso de los individuos se analiza mediante agrupamiento posibles grupos, pero a este punto se desea aplicar el GH-Biplot para representar los variables.

Figura 3-16 Dimensión 1 y 2 del GH-Biplot



Fuente: Elaboración propia

Tabla 3-10 Varianza y desviación estándar de las variables cuantitativas

Atributo	Varianza	Desviación estándar
C_I_Total	5.348.060.978.878.930	73.130.438.114.912
S_C_Total	4.453.425.340.535.310	66.733.989.394.725
Monto_I	1.363.739.993.722.410	36.928.850.425.141
Saldo_C	1.155.748.422.301.570	33.996.300.126.655
Aportes	44.004.776.520.127	6.633.609.614.691
Salario	35.872.979.347.475	5.989.405.592.166
Cuota	163.238.761.846.922	404.028.169.621
T_Dias_M	311.457.090.356	558.083.408
Dias_M	4.789.353.004	69.205.152
Plazo	1.896.535.518	43.549.231
T_Cuotas_M	271.498.091	16.477.199
Antiguedad	151.804.861	12.320.912
Edad	135.293.585	11.631.577
Cuotas_M	17.712.534	4.208.626
Cantidad_C	2.406.276	1.551.218
C_Lineas	1.245.188	111.588
Cantidad_G	0,598136	0,7734
P_Participacion	0,1148	0,3389
P_Deuda	0,0578	0,2404
T_I_Ponderacion	0,0013	0,0363
Tasa_EA	0,0012	0,0348
C_P_Ponderado	0,0005	0,0216

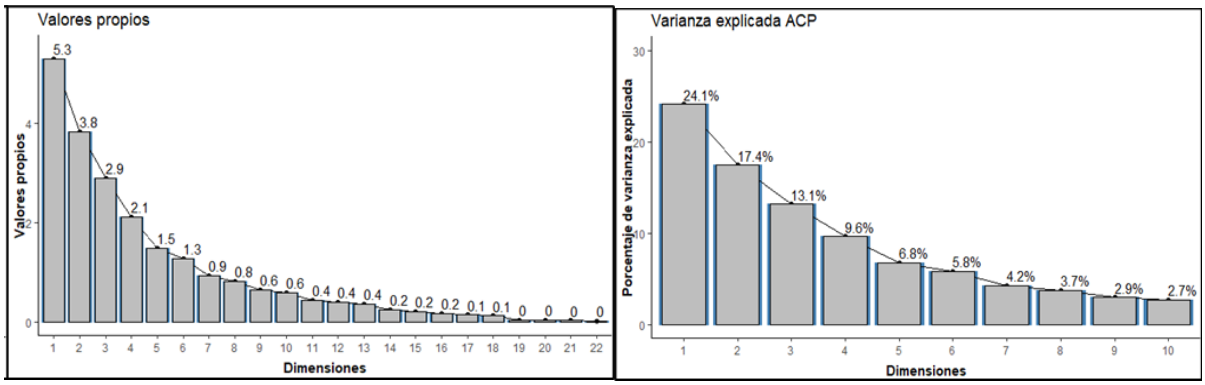
Fuente: Elaboración propia

3.4.5 Análisis de componentes principales

El método de análisis de componentes principales (ACP) tiene como objetivo la reducción de dimensionalidad, explicar la máxima variabilidad del conjunto de datos en un número reducido de factores con la mínima pérdida de información. Es un método lineal para el análisis de correlaciones en alta dimensión. La interpretación geométrica busca representar la información original en nuevas variables (componentes principales) que no están correlacionadas, a las cuales se les puede asociar grupos de variables, también permiten identificar grupos de individuos. Los resultados son muy congruentes con el de los métodos Biplot. El ACP se ha realizado con las variables escaladas, para que la varianza de cada una no afecte los resultados.

En la figura 3-17 se muestra que de las 22 componentes principales (porque hay 22 variables cuantitativas en el análisis) las dos primeras recogen un 41.5% (mismo porcentaje que el GH-Biplot) y las tres primeras dimensiones recogen el 54.61%. Note que la suma de los valores propios suma el número de variables, y el porcentaje de varianza explicada por cada componente principal es su valor propio sobre la suma de todos los valores propios.

Figura 3-17 Varianza explicada y valores propios

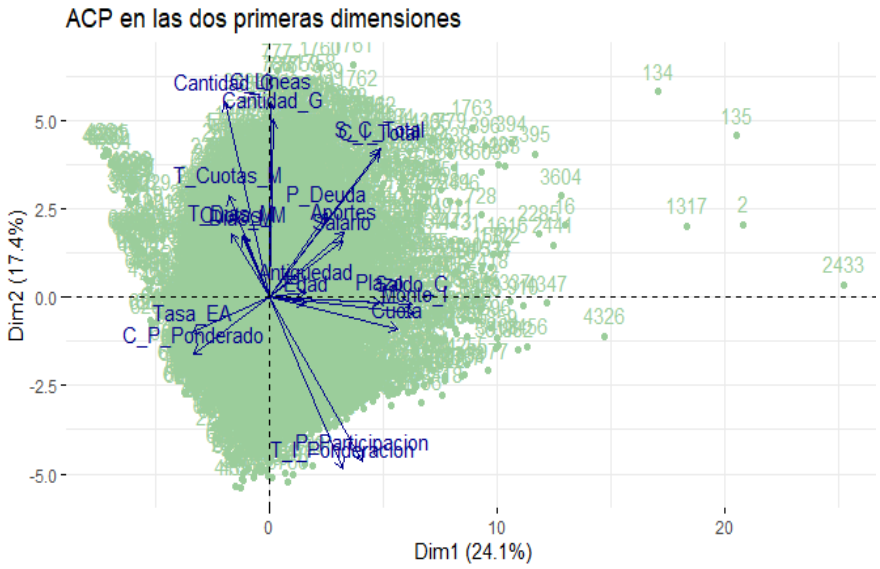


Fuente: Elaboración propia

La figura 3-18 muestra el gráfico de componentes principales en forma de Biplot (individuos y variables) y tiene los mismos resultados que el GH-Biplot. Lo que se aprecia del ACP es la selección de aquellas variables que tienen mayor contribución a la formación o explicación de las

primeras dimensiones, porque éstas podrían ser las seleccionadas para un proceso de modelación y así reducir más la dimensionalidad.

Figura 3-18 Componentes principales



Fuente: Elaboración propia

4. Modelamiento

4.1 Aspectos de la modelación

Una de las características de la modelación está en ordenar y clasificar a los asociados (prospectos) en niveles de riesgo desde el más bajo al más alto. Los modelos de *Scoring* se basan en información histórica. Esta herramienta tiene en cuenta variables demográficas, financieras y variables que puedan dar una mejor segmentación de la cartera y del grupo objetivo que se quiere gestionar, por ejemplo, indicadores de desempeño de pagos. Principalmente para emisión de nuevos créditos, crecimiento de la cartera de crédito enfocada en los asociados existentes y acciones de cobranza para que el negocio pueda mejorar el flujo de efectivo.

El conjunto de datos tiene un enfoque de análisis de cierre, es decir, la ventana de tiempo toma la información de cierre de periodo. En la variable días de mora, el máximo es de 450, es porque la ventana de corte se hizo a un tiempo de un año y cuatro meses. La evaluación del cierre se puede hacer mensual, bimestral, trimestral, cuatrimestral, semestral, anual o en un periodo más prolongado. Esto se hace para el análisis de cosechas o de créditos que ya fueron desembolsados y que incidirá en la nueva asignación de créditos.

El conjunto de datos estudiado tiene un proceso de limpieza y de selección de la información para el análisis de la población objetivo, la información recopilada no incluye exclusiones del negocio como temas regulatorios, ciclos económicos, procesos de cobro, asociados de alto riesgo, inactivos, con poca información histórica o de desempeño.

En los diferentes modelos analíticos de aprendizaje supervisado a emplear del conjunto de datos se decidió tomar un 70% para entrenamiento (*training*) y un 30% para prueba (*testing*). Se tuvo en cuenta que la distribución de malos y buenos tanto en el conjunto de entramiento como en el de prueba fuera muy similar.

4.1.1 Componentes del análisis predictivo

Los componentes que se utilizan en el análisis predictivo son:

- **Ventana de tiempo**: análisis de cierre de cartera con corte a 4 cuatrimestres.
- **Población objetivo**: asociados de la entidad financiera que tienen créditos activos, es decir, que el dinero desembolsado no ha sido cancelado, también define el perfil de asociado que la entidad desea estudiar, en el caso de estudio se consideran todos los asociados.
- **Variable objetivo**: define el incumplimiento para aquellos créditos que se encuentran con más de 90 días de mora, en el conjunto de datos se incluyó con el atributo “Deudor_en_Riesgo”.
- **Variables asociadas al objetivo**: son las covariables que se utilizan en los modelos de aprendizaje supervisado para determinar un ordenamiento o clasificación de los asociados, es decir, aquellas que explican o dan impacto en la variable objetivo.
- **Algoritmo o métrica**: se define la metodología de aprendizaje supervisado para el modelamiento. El modelo de *Scoring* es un modelo de clasificación porque la variable objetivo es categórica.

4.1.2 Selección de variables asociadas al objetivo

En el capítulo anterior se indicó 33 variables, 11 cualitativas y 22 cuantitativas. Al aplicar los métodos Biplot, el análisis de componentes principales y de correspondencias múltiples con la finalidad de reducir dimensionalidad, se determinó dejar todas las variables cualitativas. En cuanto a las variables cuantitativas, por la correlación que hay entre ellas y a efecto de poder utilizar la información agregada se utiliza el plazo, saldo de capital, cuota, total cuotas en mora, cantidad de líneas de crédito, aportes, salario, saldo de capital total, antigüedad, edad, cantidad de garantías, cantidad de créditos y costo promedio ponderado de los créditos.

Nótese que la variable cualitativa líneas de crédito tiene 11 categorías, en la información agregada un asociado que tenga más de un crédito activo podrá tener todos en la misma o diferentes categorías. Al agregar la información se creó la variable cuantitativa, que indica el número de líneas de crédito que tiene activo un asociado y este fue un procedimiento para recuperar esa información a nivel agregado, pero también se ha propuesto agregarla con una variable cualitativa que indicará en dos modalidades si el asociado posee “una” o “varias” líneas de crédito. Esto implica que para el proceso de modelamiento se pruebe a nivel agregado la variable que mejor resultados tendría si agregarla de forma cuantitativa o cualitativa. Con la variable plazo, saldo de capital y cuota ocurre algo similar, a nivel agregado se deberá trabajar con el plazo, saldo y cuota máxima que un asociado presente.

4.1.3 Interpretación de variables asociadas al objetivo

El análisis para el modelo de *Scoring* debe mantener aquellas variables que apuntan al objetivo del negocio, pues la dimensionalidad del modelo debe ser baja y evitando la redundancia en la información. Por eso se tiene una propuesta de analizar algunos indicadores de desempeño de los pagos, estos son:

- **Proporción de saldo máximo**: éste hará que para el caso del conjunto de datos agregado no se emplee la variable saldo de capital máximo y saldo total de capital sino el índice PSM

$$PSM = \frac{Saldo\ de\ capital\ máximo}{Saldo\ total}$$

- **Recencia y frecuencia**: la recencia indica el tiempo transcurrido desde el último pago que hizo el asociado, este valor se encuentra en la variable días en mora, pero el hecho de un cliente tener varios créditos con mayor o menor mora, se considera el número máximo de días en mora. La frecuencia es el número de créditos que el asociado tiene en la entidad, este valor está en la variable de cantidad de créditos.

$$RyF = \frac{Días\ en\ mora\ máximo}{Cantidad\ de\ créditos}$$

Si RyF es bajo indica que la recencia es baja y la frecuencia alta o lo mismo que el número de días es alto en comparación con la cantidad de créditos. Con este indicador se pasa de utilizar dos variables a emplear una sola.

En la tabla 4-1 se indica las variables que a nivel agregado se utilizan para el modelamiento. Se pasa de tener 33 variables a tener 21, aun faltando analizar si la variable de cantidad de garantías y líneas de crédito se deben trabajar como cuantitativas o cualitativas categorizadas en dos modalidades.

Tabla 4-1 Variables seleccionadas para el modelamiento

Cuantitativas agregado	Cualitativas agregado
Plazo máximo	Sexo
Proporción de saldo máximo	Estado civil
Recencia y frecuencia	Profesión
Cuota máxima	Periodicidad de pago
Total de cuotas en mora	Tipo de recaudo
Aportes	Cantidad garantías
Salario	Cantidad líneas de crédito
Antigüedad	Calificación de Cartera
Edad	Deudor riesgoso
Cantidad de garantías	
Cantidad de líneas de crédito	
Costo promedio ponderado	

Fuente: Elaboración propia

4.2 Árboles de decisión

Los árboles de decisión se aplican para detectar posibles incumplidores de pago en el proceso de aceptación o rechazo de solicitudes de crédito por las entidades financieras del sector solidario.

4.2.1 CART

El algoritmo CART (árbol de clasificación y regresión) se utiliza a modo de clasificación debido a que la variable de estudio es **deudor riesgoso** con dos categorías codificado con 1 si el asociado es riesgoso y 0 en caso contrario.

El nodo raíz indica el número de asociados en la categoría 0 y 1 (figura 4-1). Se comienza a dividir en nodos intermedios con particiones binarias. Se utilizan las diferentes variables predictoras en el análisis, para ello, se ha tomado de ejemplo la edad. Las categorías en las que se dividió son 4. Las divisiones binarias del nodo raíz depende de la variable predictora inicial.

Si una variable agrupada tiene 4 modalidades, tendrá $2^{4-1}-1=7$ divisiones binarias (la tabla 4-2 tiene las categorías de ambas variables), en general $2^{m-1}-1$ siendo m el número de categorías agrupadas como se muestra en la tabla 4-3. Solo una de estas 7 divisiones binarias será la seleccionada para empezar hacer las particiones sucesivas con carácter de clasificación descendente y divisivo. La división binaria seleccionada será la que presente mayor reducción de impureza (entropía). La figura 4-1 muestra el nodo raíz (nodo 0) y el primer nodo con la división binaria 1 y 5.

Tabla 4-2 Variable edad agrupada en 4 categorías en la información agregada

	Deudor riesgoso		
Edad	0	1	**Total**
[22,46)	640	578	**1218**
[46,55)	342	320	**662**
[55,65)	370	232	**602**
[65,93]	236	108	**344**
Total	**1588**	**1238**	**2826**

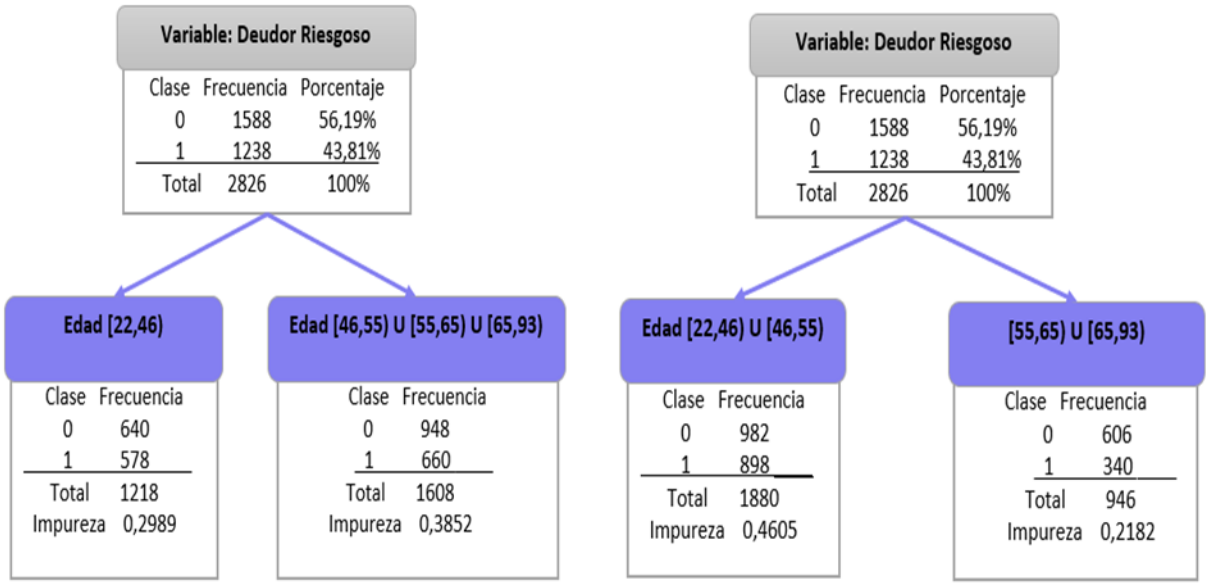
Fuente: Elaboración propia

Tabla 4-3 Divisiones binarias de una variable agrupada por categorías

División binaria	Nodo 1	Nodo 2
1	[22,46)	[46,55) ∪ [55,65) ∪ [65,93]
2	[46,55)	[22,46) ∪ [55,65) ∪ [65,93]
3	[55,65)	[22,46) ∪ [46,55) ∪ [65,93]
4	[65,93]	[22,46) ∪ [46,55) ∪ [55,65)
5	[22,46) ∪ [46,55)	[55,65) ∪ [65,93]
6	[22,46) ∪ [55,65)	[46,55) ∪ [65,93]
7	[22,46) ∪ [65,93]	[46,55) ∪ [55,65)

Fuente: Elaboración propia

Figura 4-1 Partición por el método CART división binaria 1 y 5 con la edad



(a) División binaria 1 (b) División binaria 5

Fuente: Elaboración propia

El método CART tiene como objetivo maximizar la homogeneidad interna de los nodos, la no homogeneidad es un indicador de impureza, medida de que un elemento elegido al azar del conjunto de datos sería etiquetado incorrectamente si fue etiquetada de manera aleatoria de

acuerdo con la distribución de las etiquetas. Uno de los indicadores de impureza es el índice de Shannon calculado como $-\sum_{i=1}^{k} P(a|r)\ln\left(P(a|r)\right)$ siendo $P(a|r) = n_r(a)/n_r$, es decir, la proporción de individuos del conjunto de datos en el nodo a que pertenecen a la clase o categoría r.

En el caso del nodo raíz la impureza se calcula como:

$$Impureza\ nodo\ raíz = -(p_1 \ln(p_1) + p_2 \ln(p_2)) = 0.65855$$

$$p_1 = proporción\ de\ la\ clase\ 0 = \frac{total\ de\ ceros}{total\ de\ observaciones} = \frac{1588}{2826} \approx 0.5619$$

$$p_2 = proporción\ de\ la\ clase\ 1 = \frac{total\ de\ unos}{total\ de\ observaciones} = \frac{1238}{2826} \approx 0.4381$$

En cada división binaria también se calcula la impureza de la misma manera, ponderada por la frecuencia de la subdivisión respecto al nodo raíz. Se calcula la impureza de la división binaria 1:

$$Impureza = -\left[\left(\frac{640}{1218}\ln\left(\frac{640}{1218}\right) + \frac{570}{1218}\ln\left(\frac{570}{1218}\right)\right)\right]\left(\frac{1218}{2826}\right) -$$

$$\left[\left(\frac{948}{1608}\ln\left(\frac{948}{1608}\right) + \frac{660}{1608}\ln\left(\frac{660}{1608}\right)\right)\right]\left(\frac{1608}{2826}\right)$$

$$Impureza = 0.2892 + 0.3852 = 0.6834$$

Ahora nótese que la variación de impureza del nodo 0 con la partición binaria 1 es $0.6855 - 0.6834 = 0.020$. Para las 7 posibles divisiones binarias se indica la variación de impureza como se muestra en la tabla 4-4. La mayor reducción de impureza la tiene la división binaria 5.

Tabla 4-4 Impureza y variación

División binaria	Impureza	Variación de impureza
Nodo raíz	0,6855	----------
1	0,6834	0,0020
2	0,6842	0,0013
3	0,6839	0,0015
4	0,6810	0,0045
5	0,6791	0,0064
6	0,6853	0,0002
7	0,6855	0,0000

Fuente: Elaboración propia

4.2.2 CHAID

El árbol de decisión basado en el algoritmo CHAID (Detección automática ji-cuadrado) examina las tablas cruzadas de las categorías de la variable **Deudor_Riesgoso** con las agrupaciones binarias que se realizan de las variables predictoras. Continuando con la variable edad, la división binaria muestra en la tabla 4-5 las frecuencias observadas y esperadas (si la variable Deudor_Riesgo y Edad fueran independientes).

Tabla 4-5 Frecuencias observadas y esperadas de la división binaria 1

Frecuencias observadas		
	Edad	
Deudor riesgoso	[22,46)	[46,55) U [55,65) U [65,93)
0	640	948
1	578	660
Total	1218	1608

Frecuencias esperadas		
	Edad	
Deudor riesgoso	[22,46)	[46,55) U [55,65) U [65,93)
0	684	904
1	534	704
Total	1218	1608

Fuente: Elaboración propia

En el nodo raíz la proporción de la clase 0 para el deudor riesgoso es de 0.5619, para la división binaria 0 en la edad de $[22,46)$ se esperan $1218 \times 0.5619 \approx 684$ y la agrupación

$[46{,}55) \cup [55{,}65) \cup [65{,}93]$ se esperan $1608 \times 0.5619 \approx 904$. El valor del estadístico ji-cuadrado con un grado de liberta es la suma de los errores así:

$$\chi_1^2 = \frac{(640-684)^2}{684} + \frac{(578-534)^2}{534} + \frac{(948-904)^2}{904} + \frac{(660-704)^2}{704} = 11.57$$

La probabilidad asociada a $\chi_1^2 = 11.57$ es 0.00067, lo cual indica a un nivel de significancia del 0.05 que se rechaza la hipótesis de independencia entre la variable deudor riesgoso y la variable edad (resultado que ya se había descrito pero el procedimiento en el método CHAID es diferente). La tabla 4-6 muestra el resultado de la prueba ji-cuadrado para las 7 divisiones binarias de la variable edad. Nótese que el valor más alto del estadístico ji-cuadrado lo tiene la división binaria 5 (también le corresponde el menor valor p).

Acorde con los resultados, el método CHAID ahorra tiempo al analista evitando el análisis de un gran número de tablas cruzadas bivariadas e identificando de manera eficiente la relación entre las variables, indicando aquella división binaria que muestra una mejor clasificación, que corresponde a la que tiene mayor valor del estadístico ji-cuadrado o menor valor p, que para el caso en estudio sería la división binaria 5 de la variable edad.

Tabla 4-6 Pruebas ji-cuadrado para las divisiones binarias de la variable deudor riesgoso con la edad

División binaria	Valor cuadrado	Valor-p
1	11,57	0,00067
2	7,21	0,0073
3	8,63	0,0033
4	24,51	0,0000
5	35,75	0,0000
6	18,53	0,0000
7	0,02	0,8952

Fuente: Elaboración propia

4.2.3 CTREE

CTREE (Árbol de inferencia condicional) es un algoritmo que utiliza la prueba t para contrastar la diferencia de proporciones de dos muestras. En el caso de la información agregada se tienen 2826 de asociados, y se conoce si es deudor riesgoso (1) o no (0). La tabla 4-7 ilustra la forma como se dispone la información para la prueba t. En la primera columna se identifica a cada asociado, en la segunda se pone el valor que toma en la variable deudor riesgoso. En la columna 3 se ilustra para la división binaria 1, en la cual se particionó la variable edad en intervalo 1 (categoría 1) [22,46) para formar un nodo y las tres restantes en el otro nodo, es decir, [46,55)∪[55,65)∪[65,93]. Debajo de la categoría [46,55) se escribe para cada muestra que está en ese rango de edad, el valor que toma en la variable deudor riesgoso. Por ejemplo. El asociado 6 no está en el rango de edad de [22,46) por lo tanto queda vacío, pero si está en el rango de edad [46,55)∪[55,65)∪[65,93] y debajo de esta columna se marca un 1 porque es un deudor riesgoso.

Tabla 4-7 Muestras de la variable deudor riesgoso y la edad en la información agregada

Muestra	Valor de la edad	Deudor Riesgoso	Nodo 1	Nodo 2
			$[22,46)$	$[46,55) \cup [55,65) \cup [65,93]$
1	[22,46)	1	1	
2	[55,65)	0		0
3	[55,65)	0		0
4	[55,65)	0		0
5	[46,55)	1		1
6	[55,65)	1		1
7	[22,46)	0	0	
8	[22,46)	1	1	
⋮	⋮	⋮	⋮	⋮
2823	[22,46)	1	1	
2824	[46,55)	0		0
2825	[22,46)	1	1	
2826	[55,65)	0		0

Fuente: Elaboración propia

La prueba t contrasta entonces las proporciones de la columna $[22,46)$ con la columna $[46,55) \cup [55,65) \cup [65,93]$. El estadístico t arrojó un valor de 3.41 con 2824 grados de libertad y la

probabilidad asociada a este valor o valor-p es del 0.0007, entonces se rechaza la hipótesis nula de que en la división [22,46) y [46,55)∪[55,65)∪[65,93] la proporción de deudores riesgosos sea la misma. Rechazar la hipótesis es un indicador de buena clasificación. Sin embargo, entre las 7 posibles divisiones binarias se escoge la que tenga mayor valor t o lo mismo que un p-valor más pequeño. La tabla 4-8 indica el resultado para las 7 divisiones donde se encuentra que la división binaria 5 con un valor t de 6.02 (el más alto) es la que mejor discrimina.

El árbol CHAID y CTREE llegan a una conclusión similar, ambos algoritmos soportan su procedimiento con una prueba estadística, a diferencia de CART que utiliza el criterio matemático de la impureza (o variación de entropía). No siempre se llega a indicar el mismo árbol en los tres algoritmos, en el caso en estudio ha coincidido con la división binaria 5 como aquella que mejor poder de clasificación tiene. Una de las ventajas de usar árboles, es que se pueden emplear tanto, variables predictoras cualitativas y cuantitativas y que tampoco requieren estandarizar o escalar los datos.

Tabla 4-8 Resultados prueba t para las divisiones binarias de deudor riesgoso y la edad

División binaria	Valor t-student	Valor-p
1	3,41	0,0007
2	2,69	0,0072
3	2,94	0,0033
4	4,97	0,0000
5	6,02	0,0000
6	1,01	0,3146
7	0,13	0,8953

Fuente: Elaboración propia

Tabla 4-9 Comparación árbol CART, CHAID y CTREE

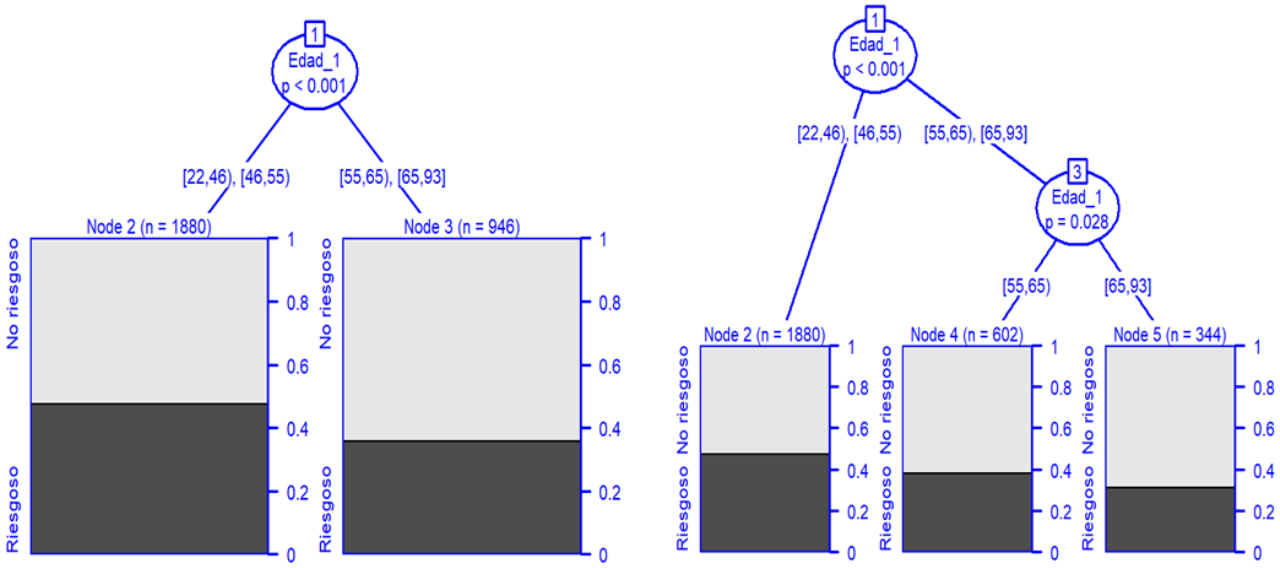
Algoritmo	Árbol final	Observación
CART (RPART)	División binaria 5	Se escoge este árbol porque tiene la mayor variación de entropía
CHAID	División binaria 5	Se escoge este árbol porque tiene mayor discriminación de la clase 0 y 1, se rechaza la hipótesis independencia entre las muestras del nodo 1 y 2, a través de la prueba ji-cuadrado. El umbral de rechazo es un p-valor menor a 0.05
CTREE	División binaria 5	Se escoge este árbol porque tiene mayor discriminación de la clase 0 y 1, se rechaza la hipótesis no igualdad de proporciones entre las muestras del nodo 1 y 2, a través de la prueba t-student. El umbral de rechazo es un p-valor menor a 0.05

Fuente: Elaboración propia

En la figura 4-2 se evidencia la participación con el árbol condicional para un nivel de profundidad del árbol (izquierda) y para dos niveles (derecha) se nota que se desagrega más del nodo que contiene los mayores de 55 años. Lo que se ha mencionado es que en el estudio de los árboles de decisión otras variables pueden tener mayor poder predictivo que la edad, que incluso no necesitaría una profundidad amplia del árbol, y este es el caso de la variable “calificación de cartera” la cual es directa, pues la cartera *A* son todos los no riesgosos y la *B*, *C*, *D*, y *E* son los riesgosos.

Al implementar estos algoritmos con todas las variables predictoras y solicitando una profundidad pequeña, la variable que mayor resultado muestra es la calificación de cartera, pues clasifica perfectamente la muestra (ver figura 4-3), esto indica que esta variable no es necesaria en el análisis pues la variable *target* se definió como riesgoso y no riesgoso y ahora se desprende de la calificación de cartera. En este caso, los analistas de riesgo de crédito en el caso de entidades del sector solidario, al definir su variable objetivo o a predecir solo deberán tomarla como riesgoso o no riesgoso, buen pagador o mal pagador, en vez de utilizar la calificación de cartera, porque pasan de utilizar una variable cualitativa de 5 categorías a una que tiene dos y esto es posible dado el problema de negocio.

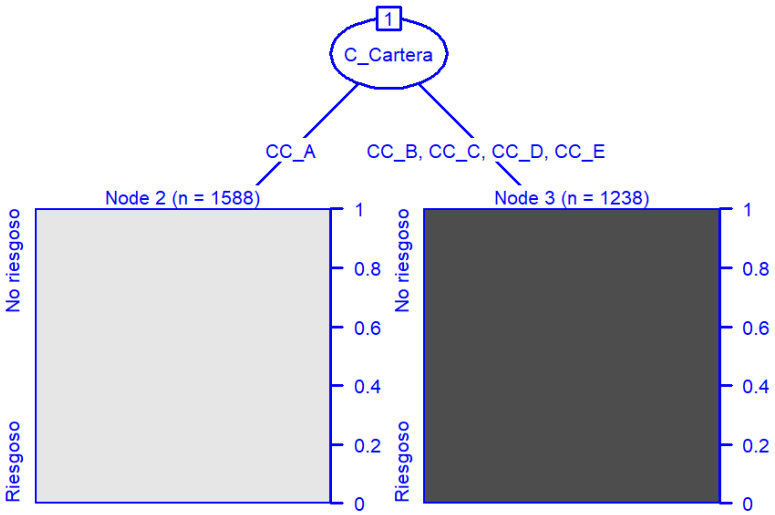
Figura 4-2 Partición por el algoritmo CTREE con uno y dos niveles de profundidad



(a) Con una profundidad (b) Con dos profundidades

Fuente: Elaboración propia

Figura 4-3 Partición por el algoritmo CART con las variables predictoras y un nivel de profundidad



Fuente: Elaboración propia

Una de las ventajas que tiene el uso de los árboles de decisión, es que son interpretables, pues en el árbol (b) de la figura 4-2 se muestra como logra particionar la edad en rangos.

4.2.4 Random Forest

Los árboles de decisión son sencillos de implementar e interpretar y permiten identificar los puntos de corte de las variables para establecer un perfil de asociados, lo cual es útil para la entidad financiera. Las divisiones sucesivas es un indicador de discriminación de la muestra en cada nivel del árbol.

Uno de los objetivos del uso de los árboles es reducir el error de clasificación de los individuos de la muestra. Un modelo de clasificación conlleva un error en la pertenencia al grupo, así se realizan varios modelos con la intención de encontrar aquel que indique un nivel de error menor (el más pequeño posible).

El ajuste de varios modelos es lo que se conoce como algoritmo de bosques aleatorios o bosques de decisión (*Random Forest*). Éste se basa en el submuestreo (*Bootstrapping*), técnica estadística que consiste en extraer submuestras con reemplazamiento, del mismo tamaño del conjunto de datos de entrenamiento.

La agregación de *Bootstrap* (*Bagging=Bootstrapping agreggation*) es un algoritmo de aprendizaje automático (*machine learning*) que emplea la combinación de modelos (*ensemble models – ensemble methods*) para obtener una mejor precisión y estabilidad de la predicción. Este ayuda a evitar el sobre ajuste, y se obtiene un mejor rendimiento de la predicción del algoritmo empleado (árboles de decisión) logrando reducir varianza, pero se va perdiendo la interpretabilidad. En cada submuestra se realiza un modelo, y para todos los modelos estimados se calcula un promedio de las predicciones.

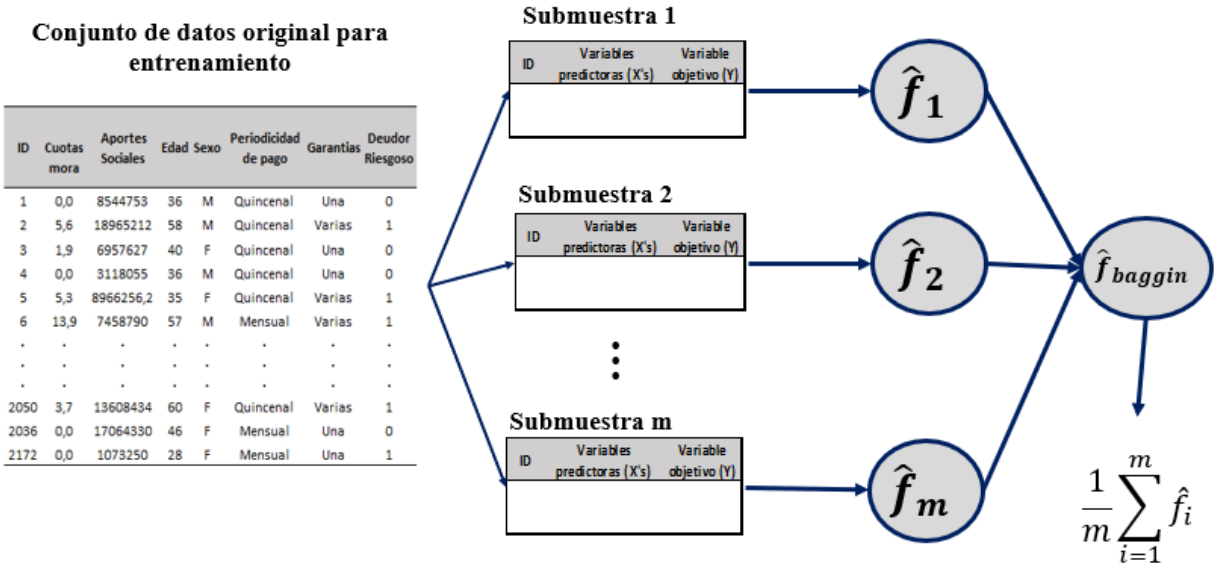
En las muestras repetidas se estima un valor particular de la población. Se puede emplear 2/3 del conjunto original para realizar submuestreos y la tercera parte restante se utiliza para hacer el testeo y validación. La tercera parte suele llamarse fuera de la bolsa (OOB- *Out of Bagging*). El bosque

aleatorio es entonces un algoritmo de aprendizaje formado por algoritmos más simples como los árboles de decisión incluyendo la metodología de Bagging.

Del conjunto de entrenamiento se extraen m muestras de tamaño fijo como se muestra en la figura 4-4, se estima para cada submuestra un modelo y se obtienen sus predicciones $\hat{f}_i$ (árbol de decisión), se ensamblan en el algoritmo de Bagging ($\hat{f}_{bagging}$) y con el promedio final se obtiene una estimación de la predicción de la variable objetivo (deudor riesgoso). Por ejemplo, si el asociado 2 que es un deudor riesgoso, y en 470 de los $m = 500$ árboles de decisión (predictores débil) que se realicen queda clasificado como riesgoso, en el promedio (470/500) quedará clasificado como riesgoso (predictor robusto) por mayoría de votos.

El desarrollo de la técnica supervisada de Random Forest permite aplicar múltiples predictores individuales (árboles), en este también se suele hacer un muestreo de las variables para que cada árbol no retenga siempre las mismas. A su vez se tiene en cuenta que las predicciones con este algoritmo superan los resultados de aplicar solo un árbol, la desventaja está en que no se obtienen las reglas de clasificación tal como se observan en árboles individuales.

Figura 4-4 Funcionamiento de *Random Forest*



Fuente: Elaboración propia

Random Forest también considera la independencia en el entrenamiento, pues no se requieren los resultados del entrenamiento de un árbol A para entrenar un árbol B (independencia). A esto, se le suma que utilizar un ensamble ayuda a reducir el problema de varianza (en árboles grandes) y sesgo (en árboles pequeños). Haciendo uso del bootstrapping en Bagging para Random Forest, se puede reducir el error de clasificación (OOB -*Out Of Bag*-) sin recurrir a la validación cruzada.

La desventaja que tiene usar *Random Forest* es que los árboles no se ven, tampoco se pueden interpretar y su ventaja está en la mejora del poder la predicción. Es diferente a cuando se usa 1, 2, 3, 4, …, 10 árboles, con diferentes niveles de profundidad, seleccionando variables que entran y salen e interpretándolas, sean cualitativas o cuantitativas e indiferente del algoritmo de entrenamiento (tabla 4-9), ver estos árboles sería algo como se muestra en la figura 4-5.

Figura 4-5 Esquema de muchos árboles de decisión entrenados

Fuente: Elaboración propia

A medida que el número de variables que se utilizan en *Random Forest* es mayor, no necesariamente el error de clasificación disminuye, este se da en la combinación todos los hiperparámetros: el número de variables, el número de nodos o profundidad del árbol y número de árboles a entrenar. Al aplicar una validación cruzada (particiones del conjunto de datos para entrenamiento y validación) a la combinación de hiperparámetros se obtuvieron varios modelos, de los cuales 11 con mejor ajuste, es

decir, con un porcentaje bajo de error de clasificación (o alto en el porcentaje de precisión) con una profundidad entre 3 y 5 nodos del árbol y con un número de variables predictoras entre 5 y 7 (ver tabla 4-10).

El aplicar el 70% de los datos para entrenamiento y 30% para prueba, el modelo final con *Random Forest* y validación cruzada de los hiperparámetros haciendo selección de la mejor combinación de éstos, se obtiene un nivel de precisión *(accuracy*) de 0.992. En la figura 4-6 se evidencia que a mayor número de árboles el error de clasificación disminuía estabilizándose cerca de 0.008 y utilizando validación cruzada con k=5 analizando el número de variables para mirar cuál era la disminución del error (ver figura 4-7).

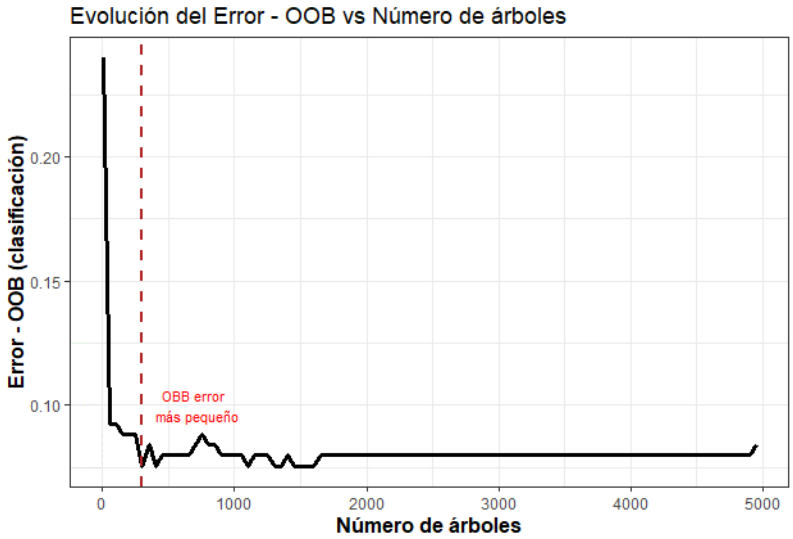
Tabla 4-10 Error de predicción en la búsqueda de combinación de hiperparámetros en *Random Forest*

Errore de clasificación								
		Número de variables						
Número de árboles /		**2**	**3**	**5**	**7**	**9**	**11**	**19**
200								
	1	0,571	0,439	0,261	0,159	0,159	0,164	0,164
	3	0,263	0,140	0,139	0,139	0,139	0,139	0,139
	5	0,110	0,119	0,108	0,105	0,090	0,078	0,093
	10	0,103	0,095	0,090	0,095	0,081	0,081	0,081
500								
	1	0,566	0,440	0,269	0,161	0,164	0,165	0,164
	3	0,177	0,140	0,137	0,139	0,139	0,139	0,140
	5	0,115	0,121	0,105	0,098	0,087	0,081	0,101
	10	0,101	0,098	0,087	0,081	0,071	0,081	0,084
1000								
	1	0,559	0,457	0,284	0,157	0,164	0,165	0,164
	3	0,164	0,142	0,137	0,139	0,139	0,137	0,139
	5	0,115	0,117	0,108	0,093	0,078	0,078	0,093
	10	0,095	0,093	0,087	0,081	0,067	0,071	0,084
2000								
	1	0,565	0,449	0,352	0,156	0,164	0,165	0,164
	3	0,205	0,140	0,137	0,139	0,139	0,137	0,139
	5	0,108	0,117	0,101	0,095	0,090	0,075	0,093
	10	0,098	0,098	0,084	0,081	0,071	0,064	0,081
5000								
	1	0,570	0,451	0,372	0,152	0,162	0,165	0,164
	3	0,159	0,133	0,137	0,139	0,139	0,137	0,139
	5	0,108	0,117	0,103	0,095	0,087	0,071	0,093
	10	0,098	0,101	0,087	0,078	0,071	0,067	0,081

Fuente: Elaboración propia

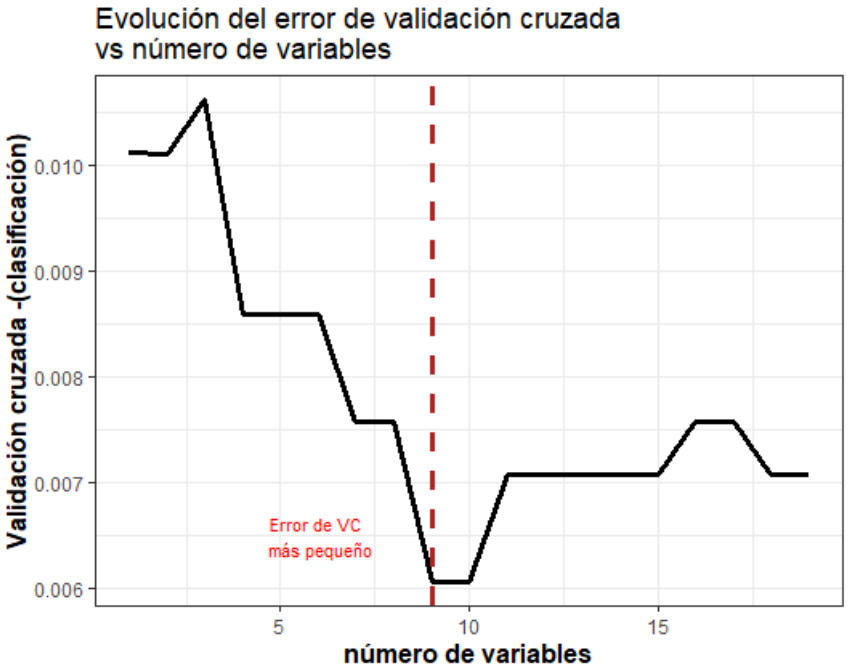
En el proceso de entrenamiento de los hiperparámetros para tratar de minimizar el error de clasificación, surge la necesidad de indicar cuáles variables tienen mayor nivel de importancia. El algoritmo empleado en el software R-Project da esta información. Las dos variables predictoras con mayor importancia en el modelo son la recencia y frecuencia (RyF) y el total de cuotas en mora. Ésta última está en total consistencia, pues al conocer las cuotas en mora es posible indicar en cuál calificación de cartera está y es consecuente con la clasificación. Por lo tanto, el modelo se puede emplear sin la variable total cuotas en mora, aunque se debe tener en cuenta, que esta es una variable necesaria para el estudio de un modelo de *Scoring* de comportamiento (ver la figura 4-8).

Figura 4-6 Evolución del error – OOB vs el número de árboles



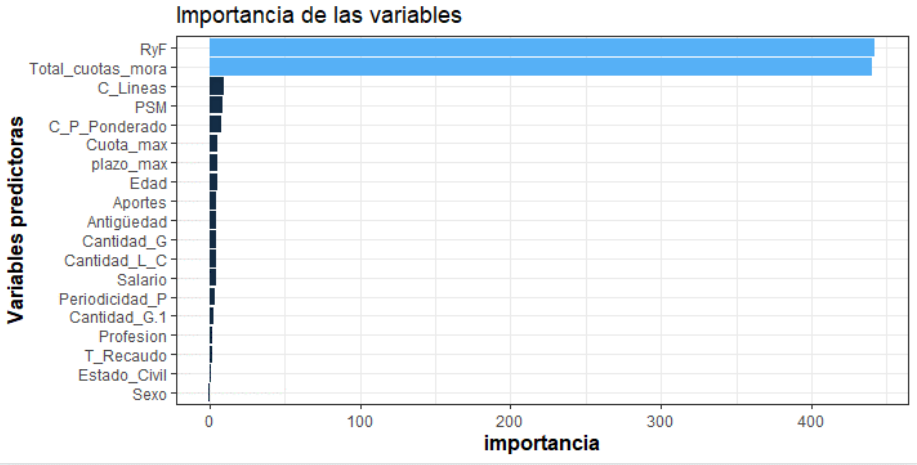
Fuente: Elaboración propia

Figura 4-7 Evolución del error de validación cruzada vs el número de variables



Fuente: Elaboración propia

Figura 4-8 Importancia de las variables con el algoritmo de Random Forest



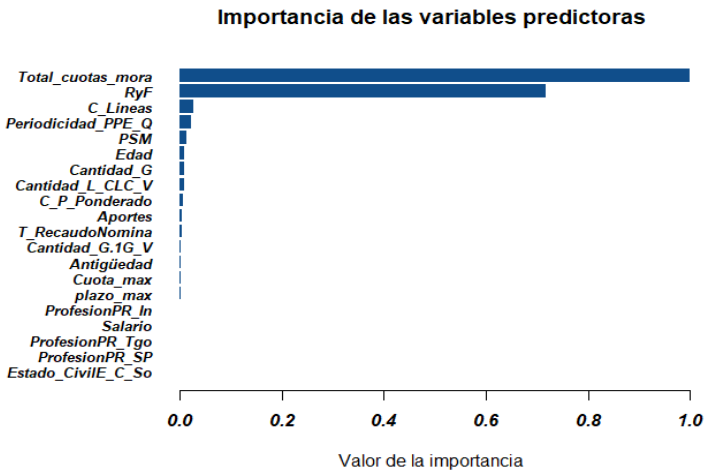
Fuente: Elaboración propia

4.2.5 XGBoost

XGBoost es un algoritmo de árboles de decisión de evolución del *Bagging, Random Forest, Boosting y Gradient Boosting* [36]. Utiliza la técnica del árbol CART agregando los árboles secuencialmente, aprendiendo de los errores de los árboles anteriores y continuando con la minimización del error de clasificación. En XGBoost a diferencia de *Random Forest* es que los árboles crecen hasta su máxima extensión, y en el procesamiento se pueden podar los árboles y evita el sobreajuste del modelo, además que trata con grandes volúmenes de datos. Se ha utilizado el algoritmo de aprendizaje supervisado XGBoost con el 30% de los datos de entrenamiento. Se obtiene una precisión del 0,996, con las variables total cuotas en mora, RyF, cantidad de líneas, PSM, edad, cantidad de garantías, cantidad de líneas de crédito, costo promedio ponderado, aportes y tipo de recuadro como las variables de mayor importancia.

El entrenamiento de la información con XGBoost se realizó teniendo como parámetros ya calibrados la profundidad de los árboles que fue de 3. Una inclusión del 40% de las variables predictoras y 5000 árboles simulados, una significancia del 5% para la selección de las ramas, y también con el factor de importancia del 0,0012 de los errores que se tenían en los modelos que se fueron iterando. Se calibraron en un proceso de optimización cruzada como en el caso del Random Forest. La figura 4-9 muestra la importancia de las variables.

Figura 4-9 Importancia de las variables predictoras con XGBoost



Fuente: Elaboración propia

4.2.6 Álgoritmo C5.0

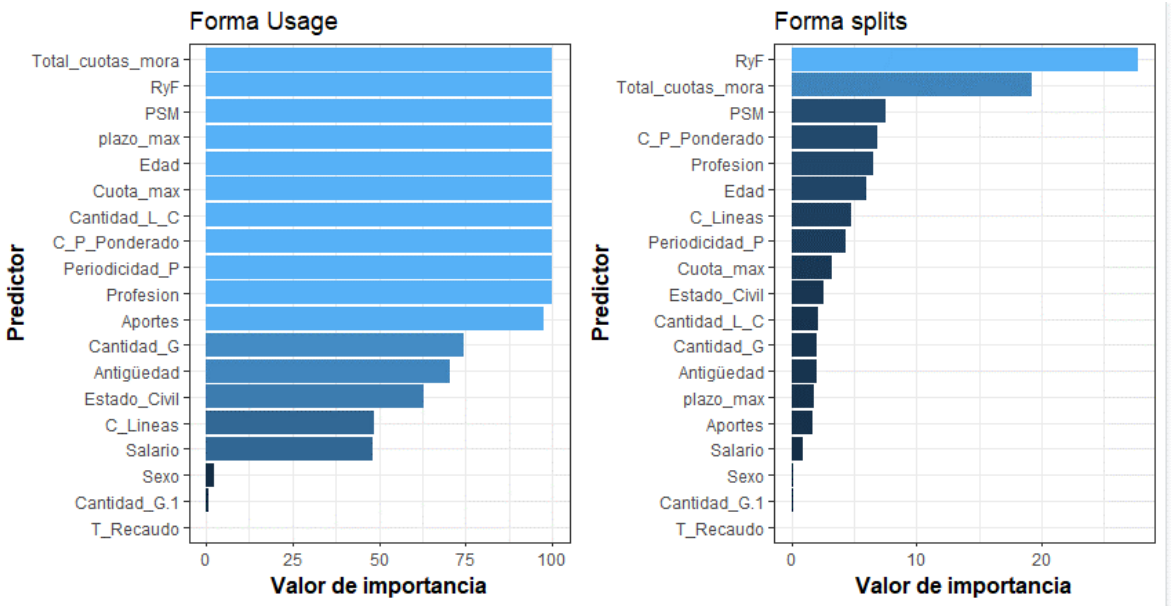
C5.0 es un algoritmo basado en árboles de decisión ensamblados, utiliza la entropía para las divisiones sucesivas de los árboles y el *Boosting* para formar predictores fuertes en el caso del problema de clasificación [37]. Una característica de este algoritmo es que se le puede asignar un peso a los errores de la clasificación, en el contexto de los datos. Para la entidad financiera es muy costoso asignarle un cupo de crédito a un asociado que no lo va a pagar en comparación con no darle cupo de crédito, así C5.0 puede recibir este costo y ser impulsado a mejorar la clasificación.

En la clasificación binaria se requiere indicar 4 valores para el costo, uno para la clase que es 0 y que se predice como 0, otro para la clase que es 0 y se predice como 1 y así para las otras dos. En la modelación se le asignó un peso de cero a los que quedan bien clasificados y diferente de cero a los que no (es decir, al error tipo I y II).

Se emplea el algoritmo de aprendizaje C5.0 el cual tiene un desarrollo muy similar al XGBoost, y es útil para abordar problemas de clasificación. Se ensamblaron 5000 modelos individuales y con ellos se prueba el conjunto de entrenamiento, se logró una precisión de 0,992.

Este algoritmo emplea dos formas de analizar la importancia de las variables predictoras, la primera es el ***Usage*** el cual toma los nodos en los que participa un predictor, y calcula la proporción de los datos de entrenamiento que están en ese nodo y al realizar un proceso de ensamble, se promedia esta proporción. La segunda forma es el ***Split*** que calcula la proporción de las divisiones en las que participa un predictor. Ambas métricas son diferentes lo cual implica que la importancia o influencia de una variable puede variar dependiendo la métrica. Los resultados implican que el árbol C5.0 tiene una buena capacidad de predicción y en cuanto la importancia de las variables el resultado sigue siendo el mismo.

Figura 4-10 Importancia de las variables con C5.0



Fuente: Elaboración propia

4.3 Métricas de discriminación

Para evaluar el poder de discriminación de una variable predictora respecto a la variable que se va a clasificar, se utilizan métricas. También aplican para evaluar la discriminación en modelos analíticos de aprendizaje supervisado.

4.3.1 Métrica del Valor de la Información (IV)

En la columna 1 de la tabla 4-11 se muestra las categorías en que se dividió la variable **total de cuotas en mora**. En la 4 y 5 se muestra la distribución de frecuencias de la variable objetivo (Deudor Riesgoso). En la 6 y 7 la distribución de las frecuencias relativas para estas dos categorías. La diferencia de las frecuencias relativas entre clases se calcula como la entropía o índice de Shannon [35], así para los que tienen cuotas en mora en el intervalo [0,1) la entropía sería:

$$(0.937 - 0.001)\left[\ln\left(\frac{0.937}{0.001}\right)\right] = 660.6\%$$

El cociente de las probabilidades $0.937/0.001$ se conoce como el *odd ratio*.

Tabla 4-11 Distribución de frecuencias para la variable total cuotas en mora

Total cuotas en mora	Frecuencia absoluta	Frecuencia Relativa	Frecuencia No riesgoso	Frecuencia Riesgoso	Frec Rela. No riesgoso	Frec Rela. Riesgoso	Diferencias de las frecuencias relativas entre clases
[0, 1)	1489	52,7%	1488	1	93,7%	0,1%	660,6%
[1, 8)	724	25,6%	100	624	6,3%	50,4%	91,7%
[9, 15)	270	9,6%	0	270	0,0%	21,8%	0,0%
[16, 25)	184	6,5%	0	184	0,0%	14,9%	0,0%
[26, 201]	159	5,6%	0	159	0,0%	12,8%	0,0%
	2826	100,00%	1588	1238	100%	100%	752,4%

Fuente: Elaboración propia

Al sumar las todas las diferencias de la columna 8 de la tabla 4-11 se obtiene el valor de la métrica IV, que para el ejemplo es de 752.4% (660.6% + 91.7% + 0% + 0% + 0%). Un umbral superior al 30% para IV indicaría que la variable predictora (en este caso total cuotas en mora), es una variable con alto poder de predicción.

4.3.2 Métrica de Kolmogorov Smirnov (KS)

A la tabla anterior se le añaden las columnas que calculan las frecuencias relativas acumuladas entre la clase de No riesgoso y Riesgoso, y una columna más que calcula las diferencias entre estas dos frecuencias (ver tabla 4-12 y figura 4-11), el valor de KS es el máximo de las diferencias de las frecuencias relativas acumuladas de estas dos clases [38].

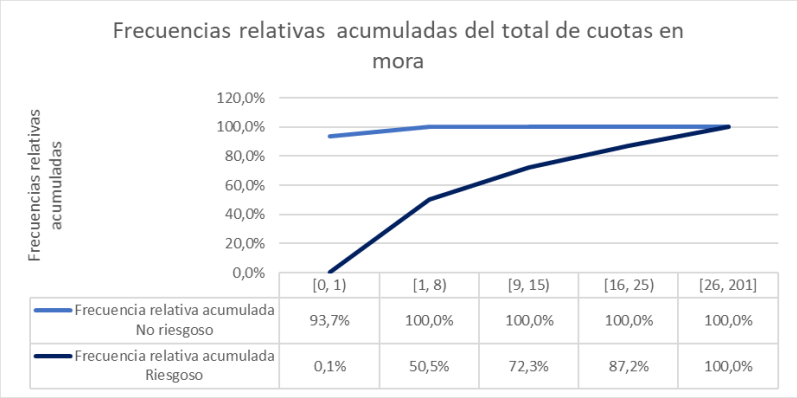
Tabla 4-12 Diferencias de las frecuencias relativas acumuladas entre clases

Total cuotas en mora	Frecuencia No riesgoso	Frecuencia Riesgoso	Frec Rela. No riesgoso	Frec Rela. Riesgoso	Diferencias de las frecuencias relativas entre clases	Frec Rela. Acum. No riesgoso	Frec Rela. Acum. Riesgoso	Diferencias de las frecuencias relativas acumuladas entre clases
[0, 1)	1488	1	93,7%	0,1%	660,6%	93,7%	0,1%	93,6%
[1, 8)	100	624	6,3%	50,4%	91,7%	100,0%	50,5%	49,5%
[9, 15)	0	270	0,0%	21,8%	0,0%	100,0%	72,3%	27,7%
[16, 25)	0	184	0,0%	14,9%	0,0%	100,0%	87,2%	12,8%
[26, 201]	0	159	0,0%	12,8%	0,0%	100,0%	100,0%	0,0%
	1588	1238	100%	100%	752,4%			

Fuente: Elaboración propia

La métrica KS se utiliza para determinar si la muestra de Riesgosos y No Riesgosos difieren significativamente. Valores mayores del 20% indican que la variable predictora tiene poder discriminatorio, pero ocurre en conjuntos de datos que no se presenten valores de KS muy altos, esto por la información que se utiliza para medir el *Score*, por ejemplo, la edad, estado civil, sexo del cliente, estrato socioeconómico que son variables estáticas (posiblemente desactualizadas). Un modelo para el *Score* de comportamiento con variables así suele tener un KS menor a 20%, porque es difícil predecir el comportamiento de un asociado que no muestre movimientos en variables asociadas a la toma de créditos. La figura 4-11 muestra el área que se forma entre las curvas de las dos clases.

Figura 4-11 Frecuencias relativas acumuladas del total de cuotas en mora



	[0, 1)	[1, 8)	[9, 15)	[16, 25)	[26, 201]
Frecuencia relativa acumulada No riesgoso	93,7%	100,0%	100,0%	100,0%	100,0%
Frecuencia relativa acumulada Riesgoso	0,1%	50,5%	72,3%	87,2%	100,0%

Fuente: Elaboración propia

4.3.3 Métrica del área bajo la curva (AUC)

El área bajo la curva ROC (AUC) es también una métrica empleada para el análisis de la discriminación. Se puede aplicar para evaluar el efecto de una sola variable respecto a la variable binaria que se trata en la clasificación (en este caso a la variable deudor). En el análisis de las dos clases de la variable deudor, el área bajo la curva (AUC) es otro indicador que se utiliza en la estimación de la capacidad que tiene una variable predictora para clasificar al deudor. Al ser un área bajo la curva, se puede integrar la función ROC, y entre las propiedades del AUC es que el área bajo la curva es 1. Mientras mayor sea el AUC mayor es el poder de discriminación (deudores

riesgosos totalmente clasificados como riesgosos y lo mismo con los no riesgosos). Un valor del AUC de 0.5 (50%) indica que es igualmente probable clasificar un deudor riesgoso como riesgoso o no, lo cual no es lo que se pretende [38].

Una de las formas de calcular el área debajo de la curva mayor (en este caso del deudor no riesgoso) es mediante el método trapezoidal, teniendo en cuenta que sería una aproximación del AUC.

En la columna 6 de la tabla 4-13 se calculan las sumas sucesivas para la clase **deudor no riesgoso**, nótese que se acumulan son los valores de la columna 4, y en la columna 7 se hacen diferencias en vez de sumas de la clase **deudor riesgoso**. El área bajo de la curva que se aproxima está en referencia a la figura 4-11 (estima la concentración de una clase en la clasificación binaria), y una aproximación de ésta se logra con la regla trapezoidal, así el AUC se calcularía como:

$$AUC = \frac{(193.7\% \times 50.4\%) + (200\% \times 21.8\%) + (200\% \times 14.9\%) + (200\% \times 12.8\%)}{2}$$

$$AUC = 98.33\%$$

Tabla 4-13 Aproximación del AUC

Total cuotas en mora	Frec Rela. No riesgoso	Frec Rela. Riesgoso	Frec Rela. Acum. No riesgoso	Frec Rela. Acum. Riesgoso	Sumas sucesivas de la clase No riesgoso	Diferencias sucesivas de la clase Riesgoso
[0, 1)	93,7%	0,1%	93,7%	0,1%	---------	---------
[1, 8)	6,3%	50,4%	100,0%	50,5%	193,7%	50,4%
[9, 15)	0,0%	21,8%	100,0%	72,3%	200,0%	21,8%
[16, 25)	0,0%	14,9%	100,0%	87,2%	200,0%	14,9%
[26, 201]	0,0%	12,8%	100,0%	100,0%	200,0%	12,8%
	100%	100%				

Fuente: Elaboración propia

4.3.1 Métrica del coeficiente de GINI

La métrica del coeficiente de GINI indica el poder discriminante que tiene una variable. Dos veces el AUC menos la unidad es el coeficiente de GINI. Esta métrica también se utiliza como una alternativa al AUC y mientras más grande sea el GINI entre una variable predictora y la que se va a predecir, más alto es el poder de clasificación. Hay casos en que el coeficiente GINI sea negativo y es por la razón de que el AUC es menor a 0.5, es decir, que se ubica debajo de la recta $y = x$ o la diagonal principal, en caso contrario siempre estará entre 0 y 1. En el caso de la variable total de cuotas en mora, se calculó el AUC= 98.33%, el índice de GINI es:

$$GINI = 2AUC - 1 = 2(0.9833) - 1$$

$$GINI = 96.66\%$$

4.3.1 Métrica del Peso de la Evidencia (WOE)

El peso de la evidencia (WOE- *Weight of Evidence*) permite analizar cómo se podría recategorizar de manera óptima una variable predictora. En el caso de emplear árboles de decisión, es óptimo tener las variables categorizadas previamente, con la finalidad de que el árbol también pueda encontrar divisiones sucesivas que tengan alto poder discriminante [12].

Al calcular el cociente entre el número de riesgosos y no riesgosos en cada categoría de la variable predictora se indica el número de veces que es el uno del otro. Si las variables que son cuantitativas o cualitativas nominales muestran un ordenamiento en el nivel de riesgo, es posible que la variable tenga capacidad discriminatoria (en el caso de cualitativas nominales se podría agrupar las categorías). Se busca entonces una manera de recategorizar teniendo cuidado de que se mantenga un orden de riesgo (porque es el objetivo del *Credit Scoring*), que el número de categorías agrupadas o agregadas sean continas, que no haya categorías en medio, y que los porcentajes en los que se van a agrupar no pierdan el ordenamiento.

El WOE analiza la relación entre una clase y otra en una escala logarítmica. Se calcula como el logaritmo natural del cociente de las proporciones de cada categoría de la variable predictora frente a las categorías de la variable a predecir (que solo tiene dos clases). Nótese que el WOE es posible cuando estas proporciones no son nulas. Cuando la tasa de riesgosos es pequeña, le corresponde un WOE alto (ver tabla 4-14), es decir, que está alineado al orden del riesgo, y ésta métrica puede ser negativa o positiva). Cuando hay discriminación pura no es posible calcular el WOE (como en el caso de la variable de calificación de cartera figura 4-3).

Fijemos el análisis en una variable cualitativa donde el WOE va de mayor a menor (la tasa de riesgosos va de menor a mayor), y si el WOE en sus valores de mayor a menor pasa de positivos a negativos, se podría recategorizar juntando categorías con WOE positivo y categorías con WOE negativos (por ejemplo, la variable cuota máxima). El investigador que construya el modelo se apoya en el WOE para recategorizar las variables, y acá entra una parte subjetiva en la cual tome decisiones acerca de esto pero que el WOE sea la métrica de apoyo para alinear y ordenar el riesgo. El WOE es un factor de distribución para darle un peso a cada categoría de la variable predictora y aplica para los algoritmos en los cuales la variable se debe discretizar y también en los que el analista decida en vez de usarla en la escala continua, utilizarla discreta.

En la tabla 14 en la columna 6 se calcula la razón de riesgosos sobre no riesgosos (81/55 = 1,47. para la categoría $[0.167\,,0.375)$), nótese que a mayor razón de riesgo el WOE es menor. La frecuencia de **No riesgoso** es de 3.46% (55/1588) y se **Riesgoso** de 6.54% (81/1238), el WOE es de -63.6% ($\ln(3.46\%/6.54\%)$). Se puede observar que el WOE está ordenado, lo que muestra que la organización de las categorías si muestra un ordenamiento y posibilita un poder de clasificación alto, sin embargo, como las tres primeras categorías tienen un WOE negativo, podría juntar la categoría 2 y 3 en una sola.

Tabla 4-14 Peso de la Evidencia para la variable PSM

PSM	Frecuencia Absoluta	Frecuencia Relativa	No riesgoso	Riesgoso	Razón de riesgosos sobre no riesgosos	Frec Rela. No riesgoso	Frec Rela. Riesgoso	WOE
[0.167, 0.375)	136	4,85%	55	81	1,47	3,46%	6,54%	-63,6%
[0.385, 0.594)	696	24,35%	334	362	1,08	21,03%	29,24%	-32,9%
[0.604, 0.812)	641	22,58%	348	293	0,84	21,91%	23,67%	-7,7%
[0.822, 1]	1353	48,23%	851	502	0,59	53,59%	40,55%	27,9%
	2826	100,00%	1588	1238		100,00%	100,00%	

Fuente: Elaboración propia

Tabla 4-15 Métricas de discriminación para el conjunto de datos

Variable	Número de categorías	Information Value	Kolmogorov Smirnoff	AUC	GINI
plazo_max	4	3,18%	5,36%	51,39%	2,78%
PSM	4	8,43%	13,04%	42,16%	-15,67%
RyF	4	640,77%	90,16%	98,92%	97,84%
Cuota_max	4	2,85%	3,11%	38,79%	-22,42%
Total_cuotas_mora	5	752,36%	93,62%	98,33%	96,66%
Aportes	5	0,77%	2,99%	42,51%	-14,97%
Salario	5	0,36%	1,30%	45,45%	-9,11%
Antigüedad	4	3,78%	8,78%	35,33%	-29,34%
Edad	4	5,99%	10,70%	35,05%	-29,90%
Cantidad_G	3	18,71%	17,73%	48,11%	-3,79%
C_Lineas	4	21,69%	18,77%	52,77%	5,53%
C_P_Ponderado	3	1,56%	3,26%	14,65%	-70,70%
Sexo	2	0,02%	0,62%	43,74%	-12,52%
Estado civil	7	1,36%	5,75%	39,52%	-20,96%
Profesión	12	3,93%	3,82%	50,88%	1,76%
Periodicidad de pago	2	1,82%	5,74%	50,04%	0,09%
Tipo de recuado	2	2,11%	5,77%	50,96%	1,91%
Cantidad de garantías	2	12,71%	17,73%	46,40%	-7,20%
Cantidad de líneas de crédito	2	14,48%	18,77%	50,16%	0,33%
Calificación de cartera	5	--------	100,00%	100,00%	100,00%

Fuente: Elaboración propia

En la tabla 4-15 se muestra que la variable PSM y RyF tienen los mejores resultados en todas las métricas presentadas, así mismo se acerca la variable de cantidad de garantías y cantidad de líneas de crédito. El analista podría ser flexible en aceptar un umbral para aceptar la forma como se vayan a emplear las variables en la modelación de las técnicas de aprendizaje supervisado. Esta tabla forma también una línea de base para la selección de variables para la modelación.

Tabla 4-16 Umbral de análisis de las métricas de validación

Umbral de evaluación de la métrica			
Information Value	**Kolmogorov Smirnoff**	**AUC**	**GINI**
• <2% Predictor no útil la relación entre el odd ratio de no riesgosos y riesgosos es débil. • [2%, 10%) la relación entre el odd ratio de no riesgosos y riesgosos es débil. • [10%, 30%) la relación entre el odd ratio de no riesgosos y riesgosos es medianamente fuerte. • [30%, 50%) la relación entre el odd ratio de no riesgosos y riesgosos es muy fuerte. • ≥50% Analizar la relación porque el predictor tiene una mayor relación.	• <20% Discriminación baja. • >=20% Discriminación aceptable.	• [50%, 70%) bajo poder discriminante • [70%,80%) poder discriminate aceptable • ≥80% alto poder discriminante (revisar)	• < 40% bajo poder discriminante • ≥ 40% poder discriminante aceptable

Fuente: Elaboración propia

4.4 Modelos de regresión

Las métricas de discriminación como se muestra en la tabla 4-15 son un paso inicial para el análisis de la relación entre la variable a predecir y las variables independientes. A continuación, se muestra el uso de los modelos de regresión que tienen un enfoque probabilístico para la clasificación binaria del tipo de deudor.

4.4.1 Modelo Lineal Probabilístico

El modelo lineal probabilístico conocido como MLP toma en cuenta una regresión lineal, sin embargo, como se indicó en el marco teórico, éste puede arrojar valores negativos para la variable predictora e incluso superiores a 1. Para solucionar este inconveniente se asigna 0 a los negativos y a los que son mayores que 1 se ajustan a la unidad. Así, se ajustan valores entre 0 y 1, considerados como la probabilidad de pertenecer a la clase, un valor superior indica que el asociado se clasificaba como riesgoso.

En la tabla 4-17 se resaltan las variables que mostraron mayor significancia estadística (variables seleccionadas). También se realizó una validación de los supuestos del modelo una vez se excluyeron las variables no seleccionadas y posterior se volvió a estimar la probabilidad.

Tabla 4-17 Variables de mejor rendimiento en el modelo MLP

Variable	*p*-valor	significancia
(Intercept)	0.003310	**
plazo_max	0.029832	*
PSM	0.000927	***
RyF	0.0000000000000002	***
Cuota_max	0.012975	*
Total_cuotas_mora	0.0000000000000002	***
Aportes	0.012139	*
Antigüedad	0.032382	*
Cantidad_G	0.0000000574490296	***
C_Lineas	0.0000000000000229	***
Estado_CivilE_C_D	0.397146	
Estado_CivilE_C_M	0.506424	
Estado_CivilE_C_Se	0.638352	
Estado_CivilE_C_So	0.967394	
Estado_CivilE_C_UL	0.705756	
Estado_CivilE_C_V	0.018637	*
ProfesionPR_An	0.254320	
ProfesionPR_As	0.551131	
ProfesionPR_Au	0.760636	
ProfesionPR_Co	0.679004	
ProfesionPR_Es	0.319677	
ProfesionPR_In	0.059984	.
ProfesionPR_Pr	0.900071	
ProfesionPR_Sa	0.991824	
ProfesionPR_SP	0.254945	
ProfesionPR_Tco	0.653998	
ProfesionPR_Tgo	0.200938	
Periodicidad_PPE_Q	0.0000464958522717	***
T_RecaudoNomina	0.038537	*

Fuente: Elaboración propia

4.4.2 Modelo Probit

Una alternativa al modelo MLP es el modelo de regresión Probit que sigue una forma similar mejorando el problema de tener un coeficiente de terminación bajo. También hace que las probabilidades de clasificación estén en el intervalo [0,1]. Tiene en cuenta que la relación entre la variable predicha y las variables productoras no es totalmente lineal.

La variable **deudor riesgoso** es dicotómica porque toma el valor de 0 si el deudor no es riesgoso y 1 si lo es. La distribución acumulada de la probabilidad puede tener la forma de una curva logística acumulativa, aunque también la normal ha funcionado, caso en el cual se habla de la regresión Probit o normal. Los errores del modelo tendrán distribución normal y la estimación se hace con mínimos cuadrados ponderados o máxima verosimilitud. El coeficiente de determinación en este modelo puede calcularse como el cociente entre el número de predicciones correctas sobre el total de observaciones, es lo mismo que la predicción total (*accuracy*). La regla para clasificar como riesgoso es si la probabilidad es mayor a 0.5. La tabla 4-18 resalta las variables de mejor capacidad predictiva (variables seleccionadas), las probabilidades se reestiman nuevamente una vez se retira las variables no seleccionadas.

Tabla 4-18 Variables de mejor rendimiento en el modelo Probit

Variable	Valor-p	Significancia
(Intercept)	0.0000008720	***
PSM	0.0000000473	***
RyF	< 0.0000000000000002	***
Total_cuotas_mora	0.0000000892	***
Antigüedad	0.06185	.
Edad	0.06455	.
Cantidad_G	0.01515	*
C_Lineas	0.73384	
C_P_Ponderado	0.00964	**
Estado_CivilE_C_D	0.08039	.
Estado_CivilE_C_M	0.90034	
Estado_CivilE_C_Se	0.47733	
Estado_CivilE_C_So	0.32742	
Estado_CivilE_C_UL	0.74896	
Estado_CivilE_C_V	0.73431	
Periodicidad_PPE_Q	0.02409	*
Cantidad_L_CLC_V	0.0000003752	***

Fuente: Elaboración propia

4.4.3 Modelo Logit

El modelo Logit es también una alternativa al modelo MLP y Probit, el cual mejor el problema de tener un coeficiente de terminación bajo y mantener las probabilidades en el intervalo [0,1]. Tiene en cuenta que la relación entre la variable predicha y las variables predictoras no es totalmente lineal. La distribución acumulada de la probabilidad puede tener la forma de una curva logística, caso en el cual se habla de la regresión Logit (o logística). Es decir, los errores del modelo tendrán distribución normal y la estimación se hace con mínimos cuadrados ponderados o máxima verosimilitud. El coeficiente de determinación en este modelo puede calcularse como el cociente entre el número de predicciones correctas sobre el total de observaciones, es lo mismo que la predicción total (*accuracy*). Si la probabilidad de deudor es mayor a 0,5 se clasifica como riesgoso en caso contrario como no riesgoso.

Este modelo puede emplear variables predictoras cualitativas y cuantitativas. Se realizó un proceso incluyendo todas las variables que se tenían para el conjunto de datos del caso de estudio. La siguiente figura muestra el modelo con las variables finales empleadas para estimar la probabilidad de ser un deudor riesgoso. Nótese que el tipo de recaudo, la periodicidad de pago, la edad y el costo promedio ponderado fueron ingresados como variables categorizadas.

Figura 4-12 Variables de mejor rendimiento en el modelo Logit

```
Call:
glm(formula = Deudor_R ~ PSM + RyF + Total_cuotas_mora + Cantidad_G +
    C_Lineas + plazo_max + t_recaudo_num + periodicidad_p_num +
    Edad_cat + C_P_Ponderado_cat, family = "binomial", data = df_model)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8061   0.0000   0.0027   0.0167   3.9472

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)        7.846675   1.550018   5.062 0.00000041420224610 ***
PSM                5.930341   1.332724   4.450 0.00000859543522561 ***
RyF               -0.338602   0.043177  -7.842 0.00000000000000443 ***
Total_cuotas_mora -1.641785   0.245337  -6.692 0.00000000002201968 ***
Cantidad_G        -1.069430   0.373121  -2.866             0.004155 **
C_Lineas          -0.981161   0.286385  -3.426             0.000612 ***
plazo_max          0.006709   0.004479   1.498             0.134172
t_recaudo_num1    -1.794822   0.884762  -2.029             0.042500 *
periodicidad_p_num1  2.205917   0.683093   3.229             0.001241 **
Edad_cat1         -0.281905   0.543156  -0.519             0.603752
Edad_cat2         -2.413981   0.843612  -2.861             0.004217 **
C_P_Ponderado_cat1 -0.499298   0.463958  -1.076             0.281851
C_P_Ponderado_cat2 -1.416250   0.833727  -1.699             0.089376 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3874.21  on 2825  degrees of freedom
Residual deviance:  189.84  on 2813  degrees of freedom
AIC: 215.84

Number of Fisher Scoring iterations: 12
```

Fuente: Elaboración propia

En los tres modelos de regresión probabilística indicados se indicó las variables que se mostraron significativas para explicar la variable **deudor riesgoso**, en este punto la intención no indicar el modelo matemático y sus parámetros, esto se hará más adelante con el modelo logístico. La regresión logística es muy útil para elaborar un modelo de Score de crédito. En el siguiente capítulo se ilustra una manera más analítica y comprensiva para el aprovechamiento de este modelo.

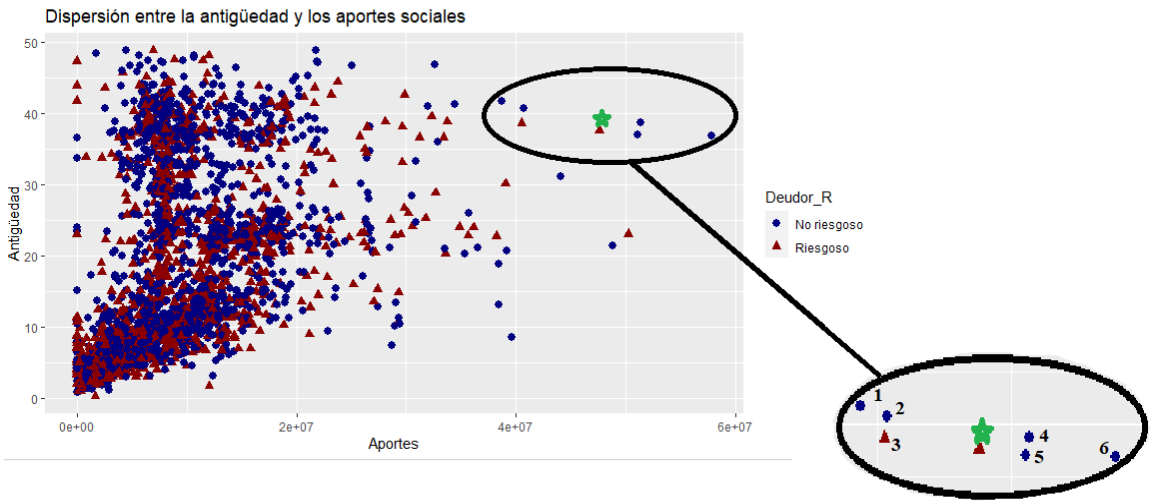
4.5 K-Vecinos más cercanos (KNN)

KNN es un algoritmo de aprendizaje supervisado para la clasificación, emplea distancias como la euclidiana, Mahalanobis, Manhattan entre otras (ecuación 4). Este algoritmo ha venido ganando importancia en la aplicación de diversas áreas de trabajo. La semejanza entre individuos depende de la métrica seleccionada. La predicción del individuo tiene en cuenta el número k de vecinos más cercanos. Este parámetro es subjetivo para el investigador, pero se puede medir para varios niveles de k el poder de predicción total bajo este algoritmo. Se clasifica en la clase donde la mayoría de sus k vecinos pertenezca (mayor al 50%).

Se utilizaron solamente las variables predictoras con mayor influencia identificadas desde los modelos de árboles decisión, disminuyendo la dimensionalidad del problema en concordancia con los resultados de las componentes principales y los métodos Biplot. Los datos para aplicar esta técnica fueron normalizados. El número de vecinos más cercanos se trabajó con varios niveles de k, empleando la distancia euclidiana en un vecindario pequeño, mediano y grande. En el análisis de dispersión por pares en las variables cuantitativas no se dio la oportunidad de remover datos atípicos que pudieran afectar la implementación de KNN.

Obsérvese en la figura 4-13 que el asociado señalado con la estrella es riesgoso, al considerar 6 vecinos cercanos, 5 de ellos (el 83.33%) son no riesgosos, al implementar este algoritmo este deudor quedaría mal clasificado (el objetivo pronosticado no sería el correcto). El deudor que se va a clasificar quedaría en el conjunto de datos de prueba y no en el de entrenamiento.

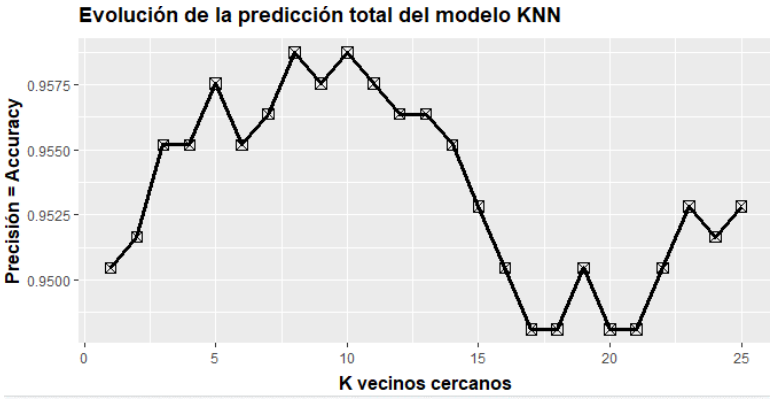
Figura 4-13 Dispersión entre la antigüedad y los aportes



Fuente: Elaboración propia

Con KNN no se hace una estimación de la probabilidad de pertenecer a una clase, solo importa la ubicación y las observaciones cercanas a esta. Se pronostica en la clase de la mayoría de sus vecinos. Téngase en cuenta que la complejidad del algoritmo se da en la elección de k, pues cuando k=1 se deben calcular N medias, en el otro extremo, cuando k=N solo se calcula una media, es decir, que la complejidad es inversa al número de vecinos cercanos para el caso de la regresión. La figura 4-14 indicó que con 5 vecinos cercanos se llegó un nivel de clasificación correcta del 95.75%.

Figura 4-14 Número de vecinos cercanos versus la predicción correcta en la clasificación



Fuente: Elaboración propia

4.6 Máquinas de soporte vectorial para clasificación (SVMC)

En la SVMC (*Support Vector Machine for Classification*) se construye un hiperplano que separa las clases, considerándose la clasificación binaria de riesgoso y no riesgoso. Las cuales están etiquetadas con $y \in \{-1,1\}$, de forma que la distancia entre el hiperplano óptimo y el patrón de entrenamiento más cercano sea máxima (es decir, se maximiza el margen que es la distancia entre ambos) para que no se presente un sesgo hacia alguna de las clases a separar.

En la expansión de las SVMC para funciones no lineales se utiliza el espacio de entrada $\mathrm{X} \subseteq \mathbb{R}^n$, a un espacio de mayor dimensión $\mathcal{H}$, denominado espacio de características. Se realiza un producto interno mediante una función no lineal. Es decir, $h(\mathbf{x}, \omega) = \langle \omega, \mathbf{x} \rangle_{\mathcal{H}} + b = k(\mathbf{x}, \omega) + b$. $\langle \omega, \mathbf{x} \rangle_{\mathcal{H}}$ puede escribirse como $\boldsymbol{\omega}\mathbf{x}^T = \sum_{i=1}^{n} \omega_i \, x_i$. El producto punto está representado por $k(\mathbf{x}, \omega)$ y es considerado un núcleo de Hilbert cuando satisface la condición de Mercer [39]. El hiperplano óptimo busca solucionar el problema de clases solapadas o puntos que contengan ruido, los puntos u observaciones que están cercanos entre una clase y otra se convierten en soportes. Se calcula la recta que los separa, perpendicular a ésta se traza una frontera que la divide (hiperplano óptimo) y que maximiza la distancia entre los soportes y el hiperplano óptimo, es decir, el margen de separación.

El hiperplano separador de las clases puede ser escrito como $h(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i \, k\big(x_i, x_j\big) + b$ siendo $\boldsymbol{\alpha}$ un vector de l elementos, los cuales son los puntos de soporte (conocidos como multiplicadores de Lagrange) que viven en un intervalo $(0, C)$. En la solución de la búsqueda del hiperplano óptimo, es donde se genera un hiperparámetro C que es regularizado para ajustar el hiperplano sobre aquellos puntos que son difíciles de clasificar (por el solapamiento). Así C, es un hiperparámetro libre de ser ajustado en la SVMC (el analista lo elegí en el entrenamiento de la SVM en función del error de clasificación). La frontera de clasificación se puede refinar regulando a C.

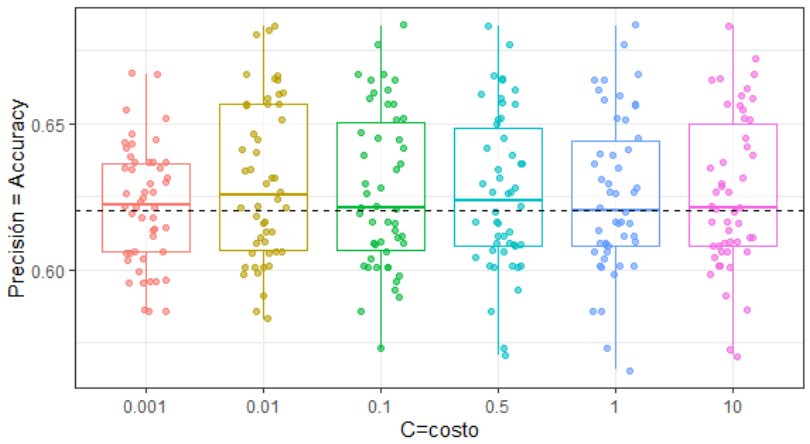
En SVMC se minimiza el margen bajo el principio de Minimización del Riesgo Estructural (MRE) y la selección de la función k (kernel) que calcula el producto punto de los puntos de entrada en el espacio de características $\mathcal{H}$ para la transformación del espacio de los datos. La función kernel cumple con dos características, es simétrica $k(\mathbf{x}, \omega) = k(\omega, \mathbf{x})$ y cumple la desigualdad de Cauchy-Schwarz [40]. Este enfoque es muy útil en el aprendizaje supervisado para determinar la frontera de

decisión (hiperplano) en el problema de clasificación para indicar si un asociado a una cooperativa financiera será bueno o malo.

En los datos de riesgo de crédito se notó en el análisis descriptivo que las variables no muestran relaciones lineales, por eso la función kernel toma los datos de entrada que no son separables linealmente y los mapea aun espacio de mayor dimensión mediante una función no lineal para hallar ese hiperplano que logra separar linealmente las clases. Se puede usar un kernel polinomial de grado superior a 1 para construir un clasificador de SVM o usando funciones gaussianas (funciones de base radial) o Laplacianas.

Una validación cruzada con 10 particiones y 5 repeticiones implica ajustar y evaluar el modelo $10 \times 5 = 50$ veces, cada vez con una partición distinta, más un último ajuste con todos los datos de entrenamiento para crear el modelo final. Se emplea por defecto la precisión total (*accuracy*). Además, se indicó un remuestreo. Promediando los 50 valores del *accuracy* obtenidos para cada valor del hiperparámetro, se identifica cual es el mejor. Finalmente, se reajusta el modelo empleando todas las observaciones de entrenamiento y el mejor valor del hiperparámetro. En este caso, entre los valores de C probados, $C = 0.01$ consigue los mejores resultados con un *accuracy* de 0.63. Si varios valores consiguen el mismo resultado, se selecciona el que da lugar al modelo más sencillo.

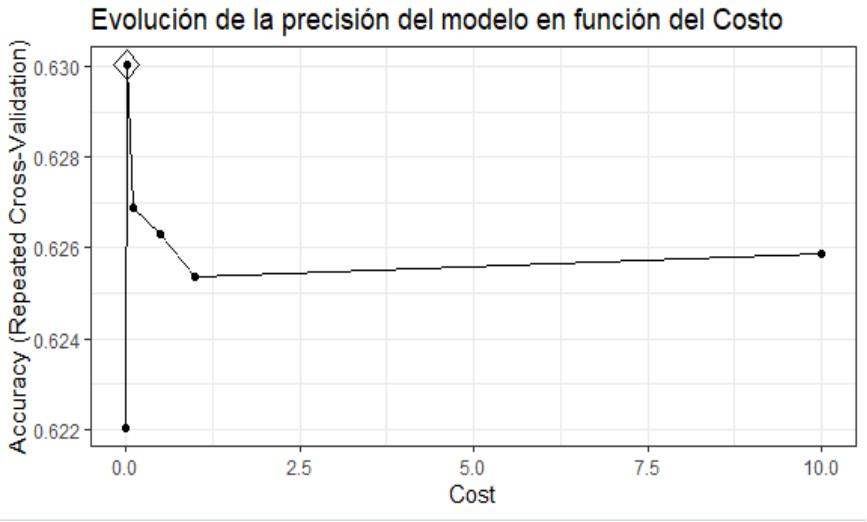
Figura 4-15 Valor del costo para SVMC versus la precisión total de la clasificación



Fuente: Elaboración propia

La figura 4-15 muestra los resultados para varios niveles del costo y en la figura 4-16 se muestra la evolución de la precisión en función del costo. Para el caso de los datos en estudio, la SVMC no mostró un nivel alto de precisión en la clasificación a pesar de que se redujo la dimensionalidad incluyendo solo las variables PSM, edad, cantidad de líneas, cantidad de garantías, plazo máximo, costo promedio, antigüedad, profesión y tipo de recaudo. Aun así, sigue siendo una alternativa supervisada para estimar las probabilidades de pertenecer a la clase de riesgoso y no riesgoso siempre y cuando la precisión sea alta, se podría fijar un umbral mayor al 95%.

Figura 4-16 Evolución de la precisión de la SVMC en función del Costo



Fuente: Elaboración propia

4.7 Redes neuronales artificiales (ANN)

La estimación de la probabilidad de que un asociado pertenezca a la clase de riesgoso o no riesgoso con las redes neuronales, requiere de indicar aquellos elementos como el número de capas, la función de costo, el tipo de variables que se van a incluir.

En la tabla 4-19 se indican todos los elementos empleados, el entrenamiento se hizo en el software RapidMiner. Se entrenó un perceptrón multicapa con conexión completa entre neuronas y capas (*Dad Fully-Connected Layer*).

Tabla 4-19 Entrenamiento de la red neuronal

Entrenamiento de la red neuronal	
Selección de variables	Se ingresan las variables que predicen la respuesta y que no covarían entre ellas.
Tipos de variables	Variable cuantitativa discreta Variable cuantitativa continua Variable cualitativa binaria Variable cualitativa politómica
Tratamiento de los datos	Normalización min-max.
Método de aprendizaje	Backpropagation=retropropagación Gradiente descendente =Backpropagation error
Neuronas en la capa de entrada	Igual al número de variables cuantitativas más el número de categorías de las variables cualitativas.
Neuronas en la capa intermedia	2k+1 por lo menos, siendo K el número de neuronas de la capa de entrada.
Neuronas en la capa de salida	Dos neuronas que identifican la salida binaria de la variable dependiente para el riesgo de crédito.
Función de coste	Cross- Entropy por ser un problema de clasificación.
Capas	Monocapa (una capa de entrada y una de salida). Bicapa (una capa de entrada, una capa oculta y una de salida). Multicapa (una capa de entrada, dos o mas ocultas.
Tipo de conexión	Hacia adelante (Feedforward) = la información va en una única dirección, desde la entrada hasta la salida
Función de activación	Sigmoidal logística / Tangencial hiperbólica/ Relu
Epocas	Número de iteracciones, 100, 120, 150
Tasa de aprendizaje	Acelera el proceso de aprendizaje, un valor pequeño constante inicial para alcanzar el mínimo

Fuente: Elaboración propia

En la tabla 4-20 se muestra la estructura de las 6 redes neuronales que mejor precisión total de buenos clasificados obtuvieron, se trató de analizar la sensibilidad del número de capas y del

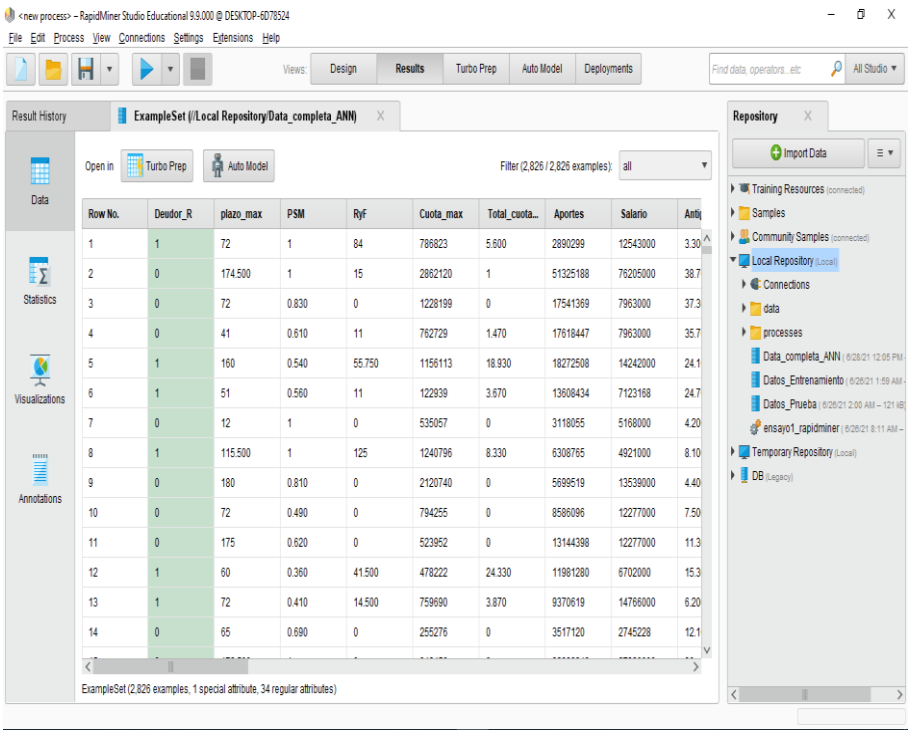
número de neuronas por capas, Nótese que en la capa de entrada hay ocho neuronas, una por cada variable predictora y en la capa de salida hay 2 neuronas que representan si es deudor riesgoso o no, también se indica la función de activación empleada en cada capa. El mejor resultado lo obtuvo una red neuronal con una capa oculta y un *cross-entropy* de 0.075.

Tabla 4-20 Estructura de las redes neuronales entrenadas

Red neuronal	Tasa de aprendizaje	Capas ocultas	Número de neuronas por capa	Funciones de activación	Error de clasificación (%)	Precisión (%)	*Cross-Entropy*
1	0,01	1	8; 33; 2	todas las capas con sigmoide	1,30	98,7	0,075
2	0,02	1	8; 80; 2	todas las capas con sigmoide	1,89	98,11	0,101
3	0,03	1	8; 201; 2	todas las capas con sigmoide	2,48	97,52	0,201
4	0,04	2	8; 201; 201; 2	sigmoide, sigmoide, Relu, simoide	1,65	98,35	0,172
5	0,06	2	8; 201; 121; 2	todas las capas con sigmoide	1,30	98,7	0,105
6	0,06	2	8; 201; 121; 2	sigmoide, Relu, Relu, sigmoide	1,30	98,7	0,13

Fuente: Elaboración propia

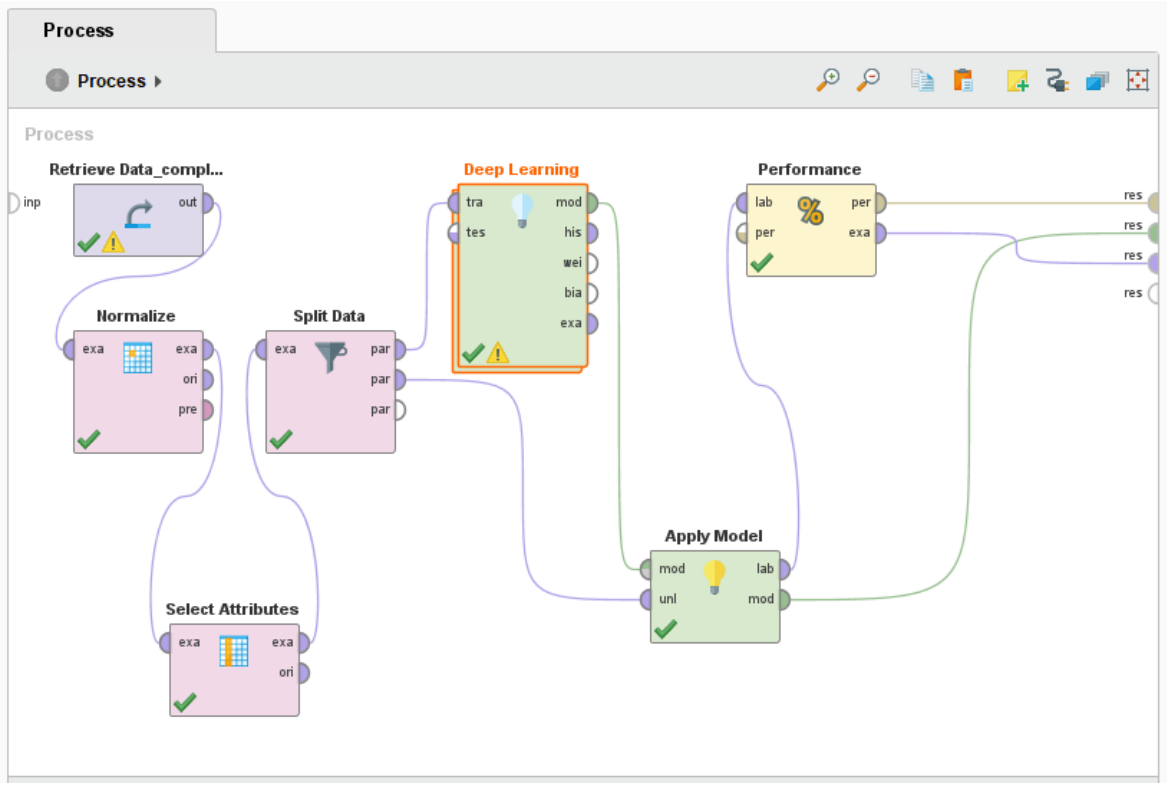
Figura 4-17 Entrada de las variables a RapidMiner



Fuente: Elaboración propia

La figura 4-17 y 4-18 ilustran el ingreso de los datos a RapidMiner y también el proceso de modelación con un esquema desde la normalización de los datos, la selección de atributos, partición en entrenamiento y prueba aplicación de la red neuronal (*Deep Learning*) análisis con datos de prueba y métricas de evaluación del desempeño del modelo. Las redes neuronales mostraron alto poder de clasificación.

Figura 4-18 Proceso de entrenamiento de las redes neuronales



Fuente: Elaboración propia

Compilando los resultados de las técnicas supervisadas como se muestra en la tabla 4-21 empleadas para obtener una clasificación de un deudor riesgoso o no en función de las covariables, y bajo un trabajo de entramiento y prueba se presentan los indicadores de precisión de la predicción obtenidos de las matrices de confusión.

Tabla 4-21 Comparación de las predicciones correctas en los modelos supervisados

Predicciones correctas	MLP	Probit	Logit	KNN	SVMC	ANN
Clase no riesgoso	86,57%	98,81%	98,74%	96,62%	62,46%	99,37%
Clase riesgoso	93,20%	99,03%	98,54%	94,59%	60,87%	97,84%
Predicción total (accuracy)	89,10%	98,90%	98,66%	95,73%	62,03%	98,70%

Fuente: Elaboración propia

5. Construcción del modelo de *Score*

Los modelos de aprendizaje supervisado presentados en la tabla 4-21 pueden ser empleados para la construcción del modelo de Score, por la razón que estiman las probabilidades de que una observación (asociado) pertenezca a la clase deudor riesgoso o no riesgoso, a excepción de KNN que indica la pertenencia de una observación basado en la cantidad de vecinos cercanos.

En este trabajo se consideró llamar las dos categorías de la variable a clasificar como “Deudor Riesgoso = 1” y “Deudor no riesgoso =0” en vez de “Malo =1 “ y “Bueno =0”; se utilizó riesgoso y no riesgoso y acá se debe entender que siempre hay riesgo por pequeña que sea la probabilidad de pertenecer a la clase de “Riesgoso”. En la construcción del modelo de Score se puede dar un paso más firme utilizando la probabilidad estimada para dar el puntaje de crédito a un asociado, y en este capítulo se van a cambiar un poco las condiciones anteriores. Ahora la variable a predecir tomará los valores “Deudor Riesgoso = 0” y “Deudor no riesgoso =1” lo cual se argumenta a continuación.

5.1 Modelo de regresión logística

Si bien, en el capítulo 4 se aplicó una regresión logística, en este se muestra como tomar mayor provecho de ella para el modelo de *Score*. Dos asociados con un buen *Score*, indica que tienen una baja probabilidad de incumplimiento (*default*). Pero analizar las variables se hace en el sentido de tener indicadores que marquen la diferencia, por ejemplo, diferencia entre un asociado con un ingreso alto de uno con ingresos medios o bajos así sea que tengan un score alto, porque aparte de que tienen menor riesgo, él de mayor ingreso tiene un signo de riqueza más alto. Esto se hace con

el fin de analizar un mejoramiento en la línea de productos, ampliación del monto de crédito, menor tasa de interés, entre otras ofertas que se le pueden aplicar mejor.

Finalmente se decidió agrupar todas las variables y realizar una regresión logística, se tuvo el apoyo de los árboles de decisión, las métricas de discriminación junto con las técnicas de reducción de dimensionalidad para indicar la categorización de las variables y realizar la mejor selección de estas. En la base del modelo de regresión logística está el deudor riesgoso y para las demás variables la idea es poner en la base a quienes son riesgos de ser deudores, así un análisis inicial para las 8 variables finales indica que:

- A mayor PSM menor riesgo
- A menor RyF menor riesgo
- A menor número garantías menor riesgo
- A menor cantidad de líneas hay menor riesgo
- A menor plazo menor riesgo
- A mayor costo menor es el riesgo
- La frecuencia de pago mensual mostró que es menos riesgosa
- A una edad mayor de 66 hay menor riesgo

Todas las variables sean cualitativas o cuantitativas se categorizaron, éstas pueden tener 2, 3, 4 o más categorías, lo intencional es que discriminen (para eso se utiliza el WOE, KS, GINI).

Hay una medida que se conoce como *odd* o razón de probabilidades, si p es la probabilidad de que el asociado sea no riesgoso, entonces $1-p$ es la probabilidad de que sea riesgoso (que incumpla), así el cociente $p/(1-p)$ es el *odd* que indica qué tantas veces es la probabilidad de **riesgoso** de **no riesgos.** Sí $p = 0.75$ entonces el *odd* $p/(1-p) = 0.75/0.25 = 3$ esto indica que la probabilidad de ser **no riesgoso** es tres veces de ser *riesgoso*. Un *odd alto* indica que hay mayor propensión de que el asociado sea no riesgoso. En la ecuación (5) se muestra el *odd* y en (6) el modelo de regresión logística.

$$odd = \frac{Probabilidad\ de\ no\ ser\ riesgoso}{probabilidad\ de\ ser\ riesgoso} = \frac{p}{1-p} \tag{5}$$

$$y(P) = Logit(P) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{6}$$

Con

- $y(P) = \begin{cases} 0, & deudor\ riesgoso \\ 1, & deudor\ no\ riesgoso \end{cases}$, $P(y=1)=p$ y $P(y=0)=1-p$
- $\beta_0, \beta_1, , \dots, \beta_k$ son los parámetros a estimar en el modelo Logit
- $X_1, \dots, X_k$ son las k variables explicativas del modelo

Nótese que se está etiquetando con $y=1$ al asociado que es no riesgoso (que es bueno) y $P(y=1)=p$ como la probabilidad ser no riesgoso, por eso si p es alta, indica que hay mayor propensión de que el asociado sea no riesgoso. El modelo de *Score* de crédito es una ponderación del modelo Logit que puede escribirse como $Score = aLogit + b$ donde a es un **factor de ponderación** y b un **factor de corrección.** Estos son los parámetros para estimar para hallar una escala o intervalo del *Score* [41], se puede escribir como se muestra en (7)

$$Score = a\ln\left(\frac{p}{1-p}\right) + b = a\ln(odd) + b \qquad (7)$$

Ahora considérese que se desea tener un *Score* en una escala de 0 a 1000 para dar un puntaje a una cartera de asociados. Se consideran dos escenarios:

Primer escenario (*Score* base = un *Score* de 700 y un *odd* de 50): Es un escenario base con un *Score* de 700 que es muy alto, y debe estar en la consistencia de que la probabilidad de ser un asociado no riesgoso deba ser también alta. Utilizando la ecuación (5) se despeja la probabilidad, y como se tiene un *odd* de 50, entonces queda que:

$$p = \frac{odd}{1+odd} = \frac{50}{1+50} = 0.9804$$

La tabla 5-1 indica para varios valores del *odd*, cuál sería el valor correspondiente de p, se ha resaltado el escenario base que es un *odd* de 50, y también se ha resaltado el *odd* de 100 con la razón de ver la probabilidad cuando el *odd* se duplica. De igual manera se observa que a mayor *odd*, mayor es la probabilidad de que el asociado sea un cliente no riesgoso (asociado que cumple con sus obligaciones) pero nótese que la probabilidad no se duplica.

En el primer escenario y consecuente con la ecuación (7), el *Score* base se puede escribir como:

$$700 = a\ln(50) + b$$

Este escore de 700 requiere que el *odd* que se incluya tenga una probabilidad de ser un deudor no riesgoso cercana a 1, pues está muy cerca a los 1000 puntos (valor máximo), además porque a mayor puntaje, mejor es la calificación crediticia, lo que indicaría que el deudor es menos riesgoso.

Tabla 5-1 *Odd* y su correspondiente probabilidad

odd	**p = probabilidad de ser bueno**
100	0,9901
99	0,9900
98	0,9899
97	0,9898
.	.
.	.
.	.
51	0,9808
50	0,9804
49	0,9800
.	.
.	.
.	.
3	0,7500
2	0,6667
1	0,5000
0,5	0,3333
0,4	0,2857
0,3	0,2308
0,2	0,1667
0,1	0,0909
0	0,0000

Fuente: Elaboración propia

Segundo escenario (*Score* base = un ***Score* de 700 + aumento y un *odd* de 100):** Es un escenario alternativo en el que se considera que el *odd* se duplica, es decir, que la probabilidad de ser bueno aumentará y estará más cercana a 1 (ver tabla 5-1). Supongamos que el *Score* al duplicar el *odd* aumenta 30 puntos porque si aumenta la probabilidad, aumenta el *Score*, pero este valor es solo

demostrativo. Consecuente con la ecuación 7 se podría escribir como se muestra en la ecuación (8) para describir los puntos que se requiere aumentar el *Score* base para que el *odd* se duplique:

$$Score + aumento = a\ln(2odd) + b \qquad (8)$$

Al consolidar los dos escenarios se tiene que:

- Primer escenario (base): $odd = \frac{p}{1-p} \quad \longrightarrow p = \frac{50}{50+1} = 0.9804$
- Segundo escenario (alternativo): $odd = \frac{p}{1-p} \quad \longrightarrow p = \frac{100}{100+1} = 0.9901$

Con este proceso se está determinando la equivalencia que va a tener el *Score* dentro de la escala que se quiere plantear (0-1000 puntos). Este proceso de alineamiento incluye 3 parámetros:

- Parámetro 1: *Score* de 700 puntos
- Parámetro 2: *odd* base de 50 puntos
- Parámetro 3: Puntos para duplicar el *odd* (PDO) igual a 30

Se debe encontrar los valores de a (factor de ponderación) y b (factor de corrección), para ellos se tiene que el $Score = a\ln(odd) + b$, y las dos ecuaciones bajo los dos escenarios planteados son:

Escenario 1 (base): $Score = a\ln(odd\ base) + b$

$$700 = a\ln(50) + b$$

$$700 = 3.9120a + b \quad (9)$$

y

Escenario 2 (alternativo): $Score + PDO = a\ln(2 \times odd\ base) + b$

$$700 + 30 = a\ln(100) + b$$

$$730 = 4.6052a + b \quad (10)$$

Así se obtiene la ecuación (9) y (10) de las cuales se genera un sistema de dos ecuaciones con dos incógnitas. Al usar un método como el de igualación, sustitución o reducción se encuentra que el valor del factor de **ponderación** es como se muestra en la ecuación (11), y el valor del factor de **corrección** se muestra en la ecuación (12).

$$a = \frac{PDO}{\ln(2)} \tag{11}$$

$$b = Score - aln(odd\ base) \tag{12}$$

Con los datos anteriores $a \approx 43.2808$ y $b \approx 530.6843$

Ahora se retoma la regresión logística, aplicada a la variable binaria que toma el valor de 1 si el asociado deudor no es riesgoso con probabilidad p y con el valor de 0 si es riesgoso con probabilidad $1 - p$. En la regresión logística se acostumbra a que la categoría que se etiqueta como 0 es la que está en la base, y los resultados del modelo permiten comparar la otra categoría con la que esta. La categoría de la base cumple para el caso de estudio de ser un deudor riesgoso, y además de esto, las 8 covariables seleccionadas (variables explicativas) se categorizaron marcando cada categoría como 0 y 1 para el caso de tener dos, o 0, 1 y 2 para el caso de tener tres, y así sucesivamente se podría hacer cuando se tengan más categorías (el número de categorías de las variables explicativas depende del proceso de analítico que se realice con las métricas de discriminación y analíticas como los árboles de decisión).

Figura 5-1 Resultado de la regresión logística final

```
> summary(glm_score)

Call:
glm(formula = Deudor_R ~ PSM_cat + RyF_cat + Cuotas_mora_cat +
    C_Lineas_cat + Edad_cat + plazo_max_cat + C_P_Ponderado_cat +
    periodicidad_p_num, family = "binomial", data = df_model)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0243  -0.1261   0.0008   0.0039   2.7977

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.2705     0.8107  -7.735 1.04e-14 ***
PSM_cat1             1.2256     0.4435   2.764  0.00572 **
RyF_cat1             7.5296     0.8310   9.061  < 2e-16 ***
Cuotas_mora_cat1     8.8416     1.3751   6.430 1.28e-10 ***
C_Lineas_cat1        2.5209     0.7877   3.200  0.00137 **
C_Lineas_cat2        3.5611     0.7993   4.455 8.38e-06 ***
Edad_cat1           -2.8491     1.2613  -2.259  0.02390 *
plazo_max_cat1      -1.1291     0.5781  -1.953  0.05080 .
C_P_Ponderado_cat1  -0.9366     0.3334  -2.810  0.00496 **
periodicidad_p_num1 -1.1841     0.4928  -2.403  0.01628 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fuente: Elaboración propia

En la figura 5-1 se muestra el modelo final ajustado, el cual cumplió con seleccionar las variables significativas (ya se habían seleccionado en todo el proceso analítico supervisado aplicado). Esta tabla no muestra los parámetros estimados para las variables que no están en la base. Se sabe que la variable que se encuentra en la base es el deudor riesgoso ($y = 0$). En el caso de la variable PSM se categorizó en dos, tomando el valor de 0 si $PSM \leq 0.69$ y 1 si $PSM > 0.69$, esto se hizo porque para esta variable valores pequeños indican que el deudor es riesgoso y por eso se va ordenando igual que se indicó con la variable a clasificar. En el caso de la variable de cantidad de líneas de crédito que registra un asociado, se organizó en 3 categorías, si es mayor a 2 se etiquetó con 0, si es 2 se etiquetó con 1, y si es menor a 2 se etiquetó con 2, se hace esto porque con ayuda de los árboles de decisión se identificó que los asociados con más de dos líneas de crédito se mostraban como riesgoso.

Lo descrito permite entender que la regresión logística se alinea con el modelo de *Credit Scoring* siempre y cuando esta se trabaje ordenada, es decir, que permita ordenar las variables, para que el *Score* sea consistente con la probabilidad de incumplimiento, para que corresponda a un *Score* alto una probabilidad alta de que el deudor no será riesgoso.

En la tabla 5-2 se muestra el ordenamiento del *Score*, la primera columna indica el nombre de la variable (incluye el intercepto), la columna 2 muestra la manera como se ordenó cada covariable en consistencia a que en la etiqueta 0 se organizaba los más riesgosos y se podría leer algo como, un deudor riesgoso tiene un PSM menor o igual a 0.69, un RyF mayor a 8.5, tener más de 1 cuota atrasa, tener préstamos activos en más de dos líneas de crédito y así sucesivamente. Es necesario hacer énfasis que el conocimiento del negocio es necesario para tratar el ordenamiento, porque se pueden presentar variables diferentes a PSM y RyF en las cuales no se sabe indicar muy bien para que valores hay mayor riesgo, por ejemplo, la edad, la periodicidad de pago, y el analista puede usar las herramientas analíticas para indicar como categorizar la variable con la finalidad de darle ordenamiento y esto permitirá una mejor interpretación del modelo.

Interpretación

En el modelo de regresión logística estimado, por el signo que tiene los coeficientes, se identifica que tener un PSM mayor a 0.69 aumenta la probabilidad de ser no riesgoso mientras los demás factores se mantengan constante. Un RyF inferior o igual a 8.5 también aumenta la probabilidad de no ser riesgoso. En cambio, solicitar créditos a plazos menores a 36 meses reduce la probabilidad de no ser riesgoso, realizar pagos mensuales también reduce la probabilidad, tener un costo promedio ponderado de los créditos superior a 11% disminuye esa probabilidad. Esta es la interpretación del signo de los parámetros estimados, sin embargo, no se desea indicar los efectos medios de cada variable sobre la probabilidad, porque con el valor del parámetro estimado se puede valorar el *Score* para cada variable.

Tabla 5-2 *Score* para las variables de la regresión logística

Variable	Categorías	Parámetro	*SCORE*
Deudor_R	Riesgoso (0)		
	No_Riesgoso (1)		
Intercepto		-6,2705	259,292
PSM	<= 0.69 (0)	0	0
	> 0.69 (1)	1,2256	**53,045**
RyF	> 8.5 (0)	0	0
	<= 8.5 (1)	7,5296	**325,887**
Total_cuotas_mora	> 1 (0)	0	0
	<= 1 (1)	8,8416	**382,672**
C_Lineas	>2 (0)	0	0
	= 2 (1)	2,5209	109,107
	<= 1 (2)	3,5611	154,127
plazo_max	> 36 (0)	0	0
	<= 36 (1)	-1,1291	-48,868
C_P_Ponderado	<= 0.11 (0)	0	0
	> 0.11 (1)	-0,9366	-40,537
periodicidad_p_num	Quincenal (0)	0	0
	Mensual (1)	-1,1841	-51,249
Edad_cat	<=66 (0)	0	0
	> 66 (1)	-2,8491	-123,311

Fuente: Elaboración propia

5.2 Asignación del puntaje

En la tabla 5-2 se muestra a aparte del ordenamiento, una columna que indica el parámetro estimado con la regresión logística, teniendo presente que si la variable explicativa tiene dos categorías solo se estima un parámetro, y este es para la categoría que no está en la base, porque la categoría de la base tiene parámetro cero. La otra columna de la tabla muestra el *Score* no para cada asociado sino para cada variable, y es que éstas también se les indica un *Score*. Los *Score* se calculan con la ayuda del factor de ponderación (a) y corrección (b) indicados en la ecuación (7). El *Score* se

calcula con ayuda de los parámetros estimados a razón del modelo de que el Logit es una combinación lineal de las variables como se muestra en la ecuación (13).

$$Logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + .. + \beta_k X_k \qquad (13)$$

El alineamiento se da teniendo en cuenta los tres parámetros del *odd* base, el *Score* base y los puntos para duplicar el *odd*, con los cuales se calcularon el factor de ponderación y corrección para construir la escala del *Score*. Entonces los parámetros (betas) para que estén en la misma escala del *Score* se multiplican por el factor (a) y al intercepto se le multiplica por este factor y se le suma el factor de corrección, así que para el intercepto queda:

$$\beta_0(a) + b = -6.2705(43.2808) + 530.6843 = 259.292$$

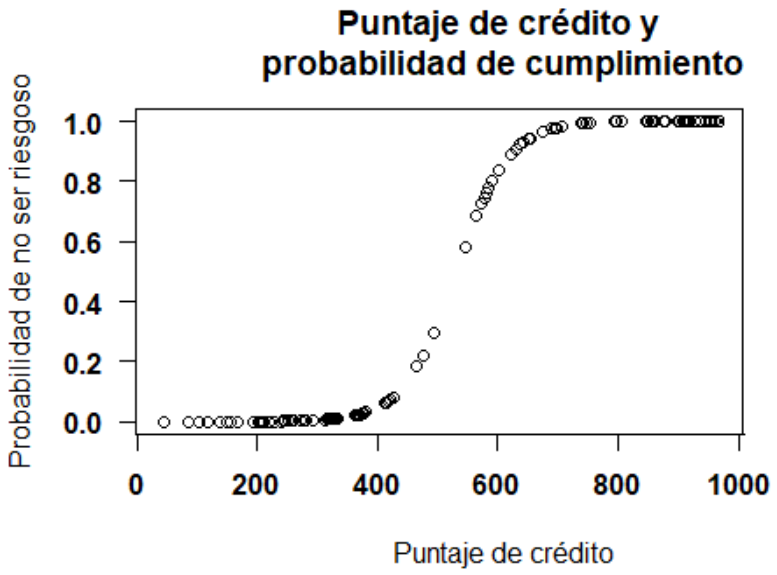
En el caso de alguna de las categorías de la variable, por ejemplo, para categoría que representa un PSM mayor a 0.69, el *Score* queda:

$$\beta_1(a) = 1.2256(43.2808) = 53.045$$

La tabla 5-2 da más información, la columna del *Score* indica el puntaje que le da cada categoría de cada variable a aun asociado, es decir que, para un asociado que tenga un PSM mayor a 0.69 tendría un puntaje de *Score* de 53.045 solo por PSM, pero si teniendo este PSM tiene menos de una línea de crédito, se le sumaría un puntaje de 154.27, pero si su frecuencia de pago es mensual, entonces se le restaría 40.537 puntos y así sucesivamente, algunas categorías suman puntos y otras restan punto, esto se da por el ordenamiento del *Score,* pero puede suceder que todas sumen punto solo que unas más que otras premiando con mayor puntaje a un deudor de menor riesgo.

Si se considera reemplazar los datos de cada asociado (valor que toma el asociado en las categorías) con el modelo estimado, entonces basta con multiplicar cada dato por el *Score* asignado a la variable. La figura 5-2 se muestra el comportamiento del *Score* asignado a cada asociado de los datos de prueba, se evidencia que el Score es consistente, pues a medida que la probabilidad de ser un deudor no riesgoso es mayor, le corresponde un *Score* mayor.

Figura 5-2 *Score* asignado a los asociados de la cartera de crédito



Fuente: Elaboración propia

También se les asignó el *Score* a los asociados de los datos de entrenamiento y la tabla 5-3 se indica los resultados de una muestra, recordando que los que tenían calificación de cartera *A* son los deudores no riesgosos y los demás presenten un nivel de riesgo de menor a mayor, siendo *B* menos riesgoso que *C* y así sucesivamente. La columna del puntaje muestra que a pesar de estar en la misma calificación de cartera el puntaje no es el mismo y se debe a que los asociados no todos presentan la misma información, pero el *Score* sigue estando ordenado en consistencia con el tipo de deudor, la calificación de cartera y la probabilidad de no ser riesgoso.

Tabla 5-3 *Score* de crédito y probabilidad de no ser riesgoso

Tipo de Deudor	Probabilidad de no ser riesgoso	Calificación de cartera	Puntaje de crédito
No Riesgoso	0,999990	CC_A	905
No Riesgoso	0,999985	CC_A	889
Riesgoso	0,184862	CC_B	428
Riesgoso	0,074192	CC_B	399
Riesgoso	0,027938	CC_C	370
Riesgoso	0,022986	CC_C	361

Riesgoso	0,001888	CC_D	266
Riesgoso	0,000741	CC_D	227
Riesgoso	0,000227	CC_E	180
Riesgoso	0,000149	CC_E	161

Fuente: Elaboración propia

5.3 Evaluación del modelo

La tabla 5-3 se indica el ordenamiento del *Score* para los datos de prueba, el modelo también asigna puntajes bajos del *Score* a clientes no riesgosos (en el primer quintil a 5 deudores no riesgosos) y es permisible un margen de error, sin embargo, del quintil 10 en adelante se muestra que se les asigna mayor *Score* a los deudores no riesgosos (a los buenos). Una de las métricas que se puede utilizar para evaluar si el *Score* quedó objetivamente bien asignado es el KS que se describió En el capítulo 4.

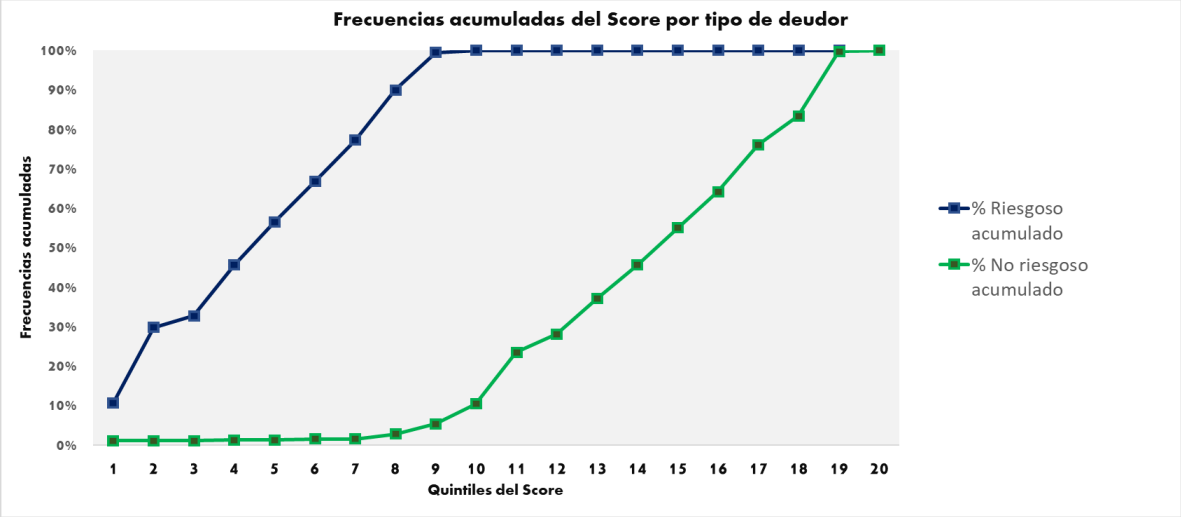
La tabla 5-5 muestra los valores que se indicaron para la construcción del KS, el valor máximo obtenido es de 94.1% un valor superior al 20% lo cual consta que el modelo presentó una buena discriminación.

Tabla 5-4 Quintiles del Score en los datos de prueba

Score categorizado por quinteles	Freccuencia Riesgoso	Freccuencia No Riesgoso
138 - 247	40	5
251 - 259	73	0
261 - 272	11	0
277 - 312	49	1
313 - 328	41	0
330 - 368	39	1
370 - 413	40	0
415 - 426	48	6
466 - 466	36	12
467 - 591	2	24
593 - 642	0	61
643 - 662	0	22
663 - 700	0	42
702 - 745	0	40
747 - 756	0	44
757 - 796	0	43
797 - 804	0	56
805 - 809	0	34
810 - 849	0	77
850 - 1000	0	1
Total	379	469

Fuente: Elaboración propia

Figura 5-3 Frecuencias acumuladas del *Score* de crédito por tipo de deudor



Fuente: Elaboración propia

Tabla 5-5 Evaluación del modelo de *Score* con KS

Rendimiento del KS					
Score	**Riesgoso acumulado**	**No riesgoso acumulado**	**% Riesgoso acumulado**	**% No riesgoso acumulado**	**Dif - Acumuladas**
138 - 247	40	5	10,55%	1,07%	9,49%
251 - 259	113	5	29,82%	1,07%	28,75%
261 - 272	124	5	32,72%	1,07%	31,65%
277 - 312	173	6	45,65%	1,28%	44,37%
313 - 328	214	6	56,46%	1,28%	55,19%
330 - 368	253	7	66,75%	1,49%	65,26%
370 - 413	293	7	77,31%	1,49%	75,82%
415 - 426	341	13	89,97%	2,77%	87,20%
466 - 466	377	25	99,47%	5,33%	94,14%
467 - 591	379	49	100,00%	10,45%	89,55%
593 - 642	379	110	100,00%	23,45%	76,55%
643 - 662	379	132	100,00%	28,14%	71,86%
663 - 700	379	174	100,00%	37,10%	62,90%
702 - 745	379	214	100,00%	45,63%	54,37%
747 - 756	379	258	100,00%	55,01%	44,99%
757 - 796	379	301	100,00%	64,18%	35,82%
797 - 804	379	357	100,00%	76,12%	23,88%
805 - 809	379	391	100,00%	83,37%	16,63%
810 - 849	379	468	100,00%	99,79%	0,21%
850 - 1000	379	469	100,00%	100,00%	0,00%
				KS	**94,14%**

Fuente: Elaboración propia

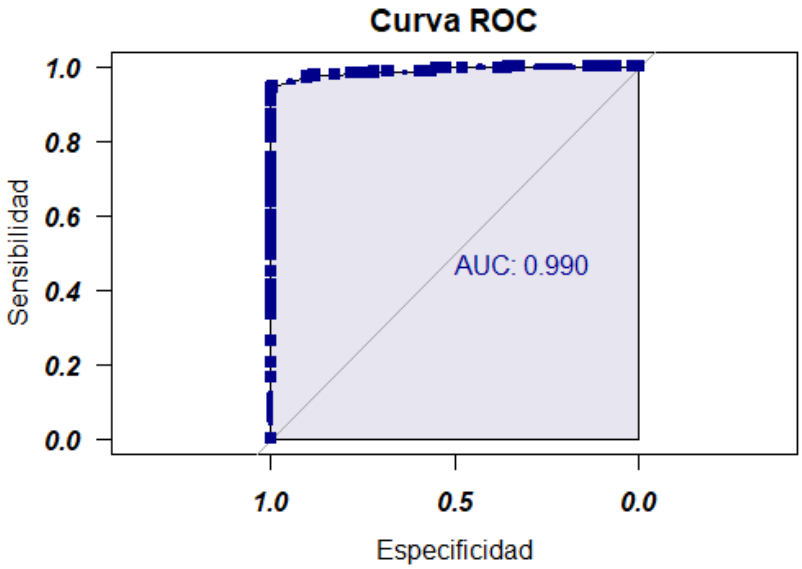
La curva ROC es otro indicador de valuación, para la regresión logística implementada en el modelo de *Credit Scoring*. La tabla 5-6 muestra la matriz de confusión obtenida y la figura 5-4 la curva ROC respectiva.

Tabla 5-6 Matriz de confusión modelo Logit con los datos de prueba

Modelo Logit		
Clase / Predicción	**No riesgoso**	**Riesgoso**
No riesgoso	440	29
Riesgoso	0	379

Fuente: Elaboración propia

Figura 5-4 Curva ROC para la regresión logística final



Fuente: Elaboración propia

La clasificación correcta de no riesgosos es 1 $(440/(440+0))$ y de riesgoso es de 0.929 $(379/(379+29))$ y la clasificación correcta completa (*accuracy*) de riesgosos y no riesgosos es de 0.966 $(((440+379)/(440+0+379+29)))$. El área bajo la cura muy cercana a 1. Ser no riesgoso y clasificado como riesgoso fueron 29 (Error tipo I), ser riesgoso y clasificado como no riesgoso fueron 0 (Error tipo II). La sensibilidad y especificidad fueron:

$$Sensibilidad = \frac{No\ riesgoso\ clasificado\ como\ no\ riesgoso}{total\ de\ no\ riesgosos} = \frac{440}{440+29} = 0.9382$$

$$Especificidad = \frac{Riesgoso\ clasificado\ como\ \ riesgoso}{total\ de\ \ riesgosos} = \frac{379}{379+0} = 1$$

Para una curva ROC se deben tener muchas matrices de confusión para tener valores de sensibilidad y especificidad, porque al estimar el modelo solo de tiene un valor de la especificidad y otro para la sensibilidad. Para generar diferentes matrices de confusión para el mismo resultado, se genera una infinidad de puntos de corte de la probabilidad, no siendo ya 0,5 la referencia del corte sino desplazando este punto de corte.

Tabla 5-7 Puntos de la corte de la probabilidad para la curva ROC

Puntos de corte	
P(No riesgoso) > 0.30	→ Se pronóstica como No riesgoso
P(Riesgoso) <= 0.30	→ Se pronóstica como Riesgoso
P(No riesgoso) > 0.35	→ Se pronóstica como No riesgoso
P(Riesgoso) <= 0.35	→ Se pronóstica como Riesgoso
P(No riesgoso) > 0.4	→ Se pronóstica como No riesgoso
P(Riesgoso) <= 0.40	→ Se pronóstica como Riesgoso
.	
.	
.	
P(No riesgoso) > 0.7	→ Se pronóstica como No riesgoso
P(Riesgoso) <= 0.7	→ Se pronóstica como Riesgoso

Fuente: Elaboración propia

Observaciones

- A este punto se realizó todo el proceso propuesto en los objetivos, pero antes de terminar este capítulo deseo indicar que todo el tema de procesado de los datos aplicando la metodología ASUM-DM se realizó en el software -R y RapidMiner.
- El modelamiento se puede aplicar en varios software, los recomendables son el software R-Project, Python, Matlab y RapidMiner, los cuales implementan en su mayoría los modelos vistos. En R-Project y Python se puede hacer un paralelo porque ambos cuentan con buenos recursos en libros, y en la web para el manejo de las librerías que contienen las funciones de aprendizaje automático supervisado, estos son dos programas *Open Source*. Matlab es un software licenciado que hoy en día cuenta con varios paquetes de funciones para realizar analítica avanzada incluyendo modelos de aprendizaje supervisado y requiere de programar código. Por su parte RapidMiner también es un software licenciado donde los modelos de aprendizaje se entrenan en un entorno de procesamiento que no requiere de elaborar código de programación, pero sí de actualizar los paquetes, además contiene funciones actualizadas para aplicar métodos analíticos.
- Los modelos entrenados en este trabajo se realizaron en el software R-Project y RapidMiner

6. Conclusiones y trabajo futuro

6.1 Conclusiones

Los algoritmos de aprendizaje supervisado ayudan en la modelación estableciendo reglas para que el modelo sea generalizable, para realizar réplicas en periodos siguientes con nuevos datos y obtener un puntaje de crédito y una probabilidad de incumplimiento nueva.

Es importante tener en cuenta la diferencia entre la visión que tiene el aprendizaje supervisado de lo real. Lo que el modelo aprende de la realidad está basado en los datos de entrenamiento y la visión que adquiere viene de ahí. Por ejemplo, si en los datos de entrenamiento un asociado con ingresos bajos es riesgoso, el modelo en predicción tendrá una precisión alta de la probabilidad de que no va a pagar. Ahora bien, si en una nueva predicción un asociado tiene bajos ingresos el modelo tratará de predecirlo como riesgoso, pero no es del todo así porque asociados de bajos ingresos no todos son riesgosos, esto depende de que en los datos de entrenamiento se hayan incluido casos así para que el modelo pueda aprender. Teniendo en cuenta esto, hay que considerar las probabilidades generadas por el modelo, como la seguridad que tiene este, desde su visión limitada, al realizar las predicciones, pero no como la probabilidad en el mundo real de que así lo sea.

Las métricas de evaluación de la discriminación son útiles para evaluar que tan bueno será el modelo para emplear un modelo de *Credit Scoring*, es un paso en el que debe intervenir el analista usando las técnicas de aprendizaje supervisado, las métricas de discriminación y técnicas de reducción de dimensionalidad. Los diferente modelos aplicados del aprendiza supervisado tiene sus ventajas y desventajas, aun así, los que sean seleccionados deben estar también muy alineados al objetivo que la entidad quiere con el riesgo de crédito, porque posiblemente sea solo evaluar la probabilidad de incumplimiento o construir el puntaje de crédito, todas las herramientas son útiles pero dependen en cierta forma del contexto del negocio, no es lo mismo un *Score* para clientes de una cooperativa financiera que para clientes bancarizados por las características que las diferencian.

En el modelo logístico que se estimó en el capítulo 4 no tiene los mismos resultados que el realizado en este capítulo 5, debido a que este último presentó un ordenamiento de las o variables predictoras, ordenamiento que se dio a través de categorías que, para el conjunto de datos en estudio, permitieron identificar puntos de corte en los cuales se encontraban los deudores riesgosos y no riesgosos. El analista que conoce bien el negocio puede también emprender un camino con ayuda de las diferentes técnicas empleadas en el capítulo de modelamiento para afianzar las variables que serán utilizadas en la modelación final para la construcción del modelo de *Scoring*.

Una vez se obtiene el *Score* de crédito para cada asociado de la cartera de créditos, es necesario iniciar un proceso de evaluación del modelo final, el cual debe estar en consistencia a que a mayor *Score* mayor debe ser la probabilidad de que sea un asociado poco riesgoso para la entidad financiera. El *Score de* Crédito debe tener un periodo de referencia en el cual se pueda actualizar para que se adapte a la dinámica de la economía. Por ejemplo, al tener dos asociados con un *Score* alto de crédito, ambos pueden acceder a créditos, pero una situación como la ocasionada por la Covid-19 ha puesto de manifiesto que el tipo de ocupación laboral de una persona puede afectar su cumplimiento crediticio, dado esto, el *Score* de crédito deberá actualizarse para apoyar la toma de decisiones en las entidades financieras.

Tabla 6-1 Comparación de los modelos de aprendizaje supervisado utilizados

Algoritmo	Análisis
Árboles de decisión y algoritmos de boosting (CART, CTREE, CHAID, Random Forest y C5.0)	Permitieron detectar puntos de corte en las divisiones binarias sucesivas que dan señales de una discriminación objetiva en variables cualitativas y cuantitativas. Su mejora se da cuando se utilizan algoritmos de boosting y validación cruzada aunque se pierde la interpretación de los árboles, se gana en una mejora de la estimación de la probabilidad de pertenecer a una clase. Estos algoritmos permiten indicar las variables con mayor poder de discriminación.
Modelos de regresión probabilística (MLP, LOGIT, PROBIT)	Las variables se codificaron desde un inicio para ser ingresadas al modelo. Al MLP se le hace el ajuste porque la probabilidad estimada no está en el rango [0, 1]. El modelo Probit mejora lo anterior y el modelo Logit permite obtener un ordenamiento de las variables. El analista tiene mayor control del modelo porque organiza la información para mantener un control del ordenamiento del puntaje de crédito. Los tres modelos permitieron indicar las variables con mayor poder de discriminación, además de que se puede tener estimado el efecto del aumento de cada variable en la probabilidad de ser un deudor riesgoso, que es una de las ventajas que tiene frente a otros algoritmos de aprendizaje supervisado.
k-vecinos más cercanos	Este algoritmo ayuda a indicar la pertenencia de un asociado a la clase de riesgoso o no riesgoso, en él se incluyeron un número considerable de variables predictoras que previamente se seleccionaron, pero no estima las probabilidades de pertenecer a una clase, solo la mide de acuerdo con el número de vecinos más cercanos.
Máquinas de soporte vectorial	Es útil en el escenario en el que el problema de credit scoring se analiza con un número amplio de variables, permitió estimar las probabilidades de pertenecer a una clase, y es útil en el caso en que separar las clases se hace cada vez más difícil por la alta dimensionalidad y el alto volumen de datos. Sin embargo, en el caso de estudio fue el modelo con el poder de calificación más pequeño. Ayudó en la estimación de las probabilidades de pertenecer a la clase de riesgoso y a indicar variables con alto poder de discriminación.
Redes neuronales	El algoritmo basado en redes neuronales para clasificación mostró alto poder de predicción y bajo error de clasificación, estimó las probabilidades de pertenecer a la clase de riesgosos y es manejable en cuanto a los parámetros que se le debieron indicar, en general, para el caso en estudio se utilizaron pocas capas de neuronas. A pesar de estimar la probabilidad, el algoritmo no permite la interpretación de los resultados.

Fuente: Elaboración propia

6.2 Trabajo futuro

Si bien en este trabajo se utilizó el modelo de regresión logística alineando los parámetros de este modelo con el *Score* de crédito que se deseó en una escala de 1 a 1000 y empleado el *odd* para asignar también un *Score* a cada asociado. Parte fundamental fue estimar la probabilidad de que el deudor sea no riesgoso (buen pagador). Los demás modelos del aprendizaje supervisado como

ANN, SVMC, CTREE, CART, CHAID, MLP, Probit, Random Forest, XGBoost, C5.0 entre otros, son un apoyo para estimar la probabilidad de pertenecer a una de las dos clases, y el trabajo está en utilizar estos modelos generando un esquema para asignar el *Score* de crédito buscando también interpretar los resultados de manera que se siga manteniendo un ordenamiento del *Score*.

Otra posibilidad está en que se indique el fin del ordenamiento del *Score*, en este trabajo se utilizó con fines de analizar un *Score de* comportamiento, pero se puede generar un esquema de un *S*core para generar nuevos créditos, para segmentación de productos de crédito.

También se debe tener una exhaustividad de las variables a nivel demográfico, crediticio y con posibilidades de incluir variables que den cuenta de la posición geográfica y de información micro y macroeconómica, pues desde la revisión de literatura se encontró que el *Scoring* depende en gran manera del contexto cultural, social y económico de la cartera de asociados que se analice.

Mejorar el proceso de automatización de la información, pues el *Score* debe estar actualizado constantemente lo que permite tomar mejores decisiones en planes de créditos y monitorear el riesgo asociado al incumplimiento por parte de los asociados, además de que genera valor para las entidades financieras porque les permitirá ser competentes en el mercado de crédito.

Bibliografía

[1] Basel Committee on Banking Supervision. “Credit Ratings and Complementary Sources of Credit Quality Information”. *Bank for International Settlements, Basel Committee on Banking Supervision*, pp. 1-183. (2000). Disponible en: www.bis.org/publ/bcbs72a.pdf

[2] K. Brown, y P. Moles. “Credit risk management”. *Credit Risk Management*. 16. (2014).

[3] L. Yu, X. Li, L. Tang, Z. Zhang y G, Kou. “Social credit: A comprehensive literature review”. *Financial Innovation*. Vol. 1(1). (2015). DOI:10.1186/s40854-015-0005-6.

[4] M. Schreiner. “Can credit scoring help attract profit-minded investors to microcredit?” *New partnerships for innovation in microfinance*, pp. 207-231. (2009). Doi:10.1007/978-3-540-76641-4_12.

[5] D. van Thiel, W.F.F, van Raaij. “Artificial Intelligent Credit Risk Prediction: An Empirical Study of Analytical Artificial Intelligence Tools for Credit Risk Prediction in a Digital Era”. *Journal of Risk Management in Financial Institutions*. Vol. 19(8). (2019).

[6] Basel Committee on Banking Supervision. “Basel Committee – 2010”. *Bank for International Settlements, Basel Committee on Banking Supervision*. (2010). Disponible en: https://www.bis.org/list/bcbs/spp_12/from_01012010/index.htm

[7] D. Huang, J. Zhou y H. Wang. “RFMS Method for Credit Scoring Based on Bank Card”. *Statistica Sinica*. Vol. 28(4), pp. 2903-2919. (2018). DOI: 10.5705/ss.202017.0043

[8] V. S. Ha, D. N. Lu, G.S. Choi, H.-N. Nguyen y B. Yoon. “Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Feature selection with Deep learning”. *International Conference on Advanced Communication Technology, ICACT*. Vol. 29, pp. 511-515. (2019). DOI: 10.23919/ICACT.2019.8701943

[9] S.O. Yaroshchuk, N.N. Shapovalova, A.M. Striuk, O.H. Rybalchenko, I.O. Dotsenko y S.V.B ilashenko. “Credit scoring model for microfinance organizations”. *CEUR Workshop Proceedings*. Vol. 2546, pp. 115-127. (2019).

[10] CAOBA. “Reporte técnico: Perfil Alianza Caoba”. Universidad de los Andes. (2017).

[11] S. Lessmann, B. Baesens, H. Seow y l. Thomas. “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research”. *European Journal of Operational Research*. Vol. 247(1), pp. 124-136. (2015). DOI: 10.1016/j.ejor.2015.05.030

[12] Y. Goh y L. S. Lee. “Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches”. *Advances in Operations Research. Hindawi Limited*. (2019). DOI: https://doi.org/10.1155/2019/1974794

[13] El Congreso de Colombia. “Ley 454 de 1998”. (1998). Disponible en: http://www.secretariasenado.gov.co/senado/basedoc/ley_0454_1998.html

[14] Superintendencia Financiera de Colombia. "Circular Básica Contable y Financiera (Circular Externa 100 de 1995)". *Superintendencia Financiera de Colombia*. (1995). Disponible en https://www.superfinanciera.gov.co/inicio/normativa/normativa-general/circular-basica-contable-y-financiera-circular-externa--de---15466

[15] A, Rodan, y H. Faris. “Credit risk evaluation using cycle reservoir neural networks with support vector machines readout”. *Lecture Notes in Computer Science*. Vol. 9621, pp. 595-604. (2016). DOI:10.1007/978-3-662-49381-6_57

[16] V.S. Ha y H.N. Nguyen. “FRFE: Fast recursive feature elimination for credit scoring”. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*. Vol. 168, pp. 133-142. DOI: 10.1007/978-3-319-46909-6_13. (2016).

[17] H. Abdou y J. Pointon. “Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature”. *Int. Syst. in Accounting, Finance and Management*. Vol. 18, pp. 59-88. (2011). DOI: 10.1002/isaf.325.

[18] F. A. Medeiro y M. T. S. Arns. “Ten-year evolution on credit risk research: a Systematic Literature Review approach and discussion”. *Ingeniería e investigación*. Vol. 40 (2), pp. 1-21. (2020). DOI: 10.15446/ing.investig.v40n2.78649

[19] M. Aláraj, M.A. bbod y M. Radi. “The applicability of credit scoring models in emerging economies: an evidence from Jordan”. *International Journal of Islamic and Middle Eastern Finance and Management*. Vol. 11(4), pp. 608-630. (2018). DOI: 10.1108/IMEFM-02-2017-0048

[20] U. Rahmani, Sukono, Riaman, Subiyanto y A.T. Bon. “Analysis of Credit Scoring Using Particle Swarm Optimization Algorithm under Logistic Regression Mode*l”. Proceedings of the International Conference on Industrial Engineering and Operations Management*, pp. 1201-1210. (2019).

[21] G. Chornous y I. Nikolskyi. “Business-Oriented Feature Selection for Hybrid Classification Model of Credit Scoring”. *Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018 1 October 2018*. Art. 8478534, pp. 397-401. (2018). DOI:10.1109/DSMP.2018.8478534

[22] J. Kalina. “High-dimensional data in economics and their (robust) analysis”. *Serbian Journal of Management*. Vol. 12. (2017). DOI: 10.5937/sjm12-10778.

[23] Superintendencia Financiera de Colombia. "Evaluación de cartera Concepto No. 1999039821-2. Octubre 14 de 1999". *Superintendencia Financiera de Colombia*. (1999). Disponible en https://www.superfinanciera.gov.co/jsp/Publicaciones/publicaciones/loadContenidoPublicacion/id/18387/dPrint/1/c/0

[24] A. S. Kumar, S. Ramesh, y S. Rahul. “A Technology on Credit Score System Assessing Public Perception in Bengaluru city”. *International Journal of Innovative Technology and Exploring Engineering*. Vol. 8(12), pp.1929-1934. (2019). DOI: 10.1016/j.eswa.2008.06.071

[25] G. Arutjothi y C. Senthamarai. “Prediction of loan status in commercial bank using machine learning classifier”. *International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, pp. 416-419. (2017). DOI: 10.1109/ISS1.2017.8389442.

[26] O. S. Pardo. "Perfil de riesgo de crédito para una cooperativa en Villavivencio a partir de un modelo Logit". Universidad y empresa. Vol. 22(38), pp. 237-256. (2020). DOI: 10.12804/revistas.urosario.edu.co/empresa/a.8266

[27] N. R. Ayala. y K. C. Delgado. "Aplicación del modelo SARC (Sistema de Administración del Riesgo Crediticio) a la cooperativa de servicios múltiples de VILLANUEVA LTDA - COOPVILLANUEVA LTDA” DEPARTAMENTO DE SANTANDER". Universidad de SANTANDER. (2014).

[28] A. Dimitris, M. Doumpos, P. M. Pardalos. y C. Zopounidis. “Computational approaches and data analytics in financial services: a literature review”. Journal of the Operational Research Society. Vol. 70(10), pp. 1581-1599. (2019) DOI: https://doi.org/10.1080/01605682.2019.1595193

[29] A, Lahsasna, R.N. Ainon y T.Y. Wah, “Credit Scoring Models Using Soft Computing Methods: A Survey”. *The International Arab Journal of Information Technology*. Vol. 7(2), pp. 115-124. (2010).

[30] V. M, A. García y J. S. Salvador. “An insight into the experimental design for credit risk and corporate bankruptcy prediction systems”. Journal of Intelligent Information Systems. Vol. 44(1), pp.1-31. (2014). DOI: 10.1007/s10844-014-0333-4

[31] Y. Zhang, G. Chi y Z. Zhang. “Decision tree for credit scoring and discovery of significant features: an empirical analysis based on Chinese microfinance for farmers”. *Filomat*. Vol. 32 (5), pp.1513–1521. (2018). DOI: 10.2298/FIL1805513Z

[32] D. E. Rodríguez y G. J. González. "Principios de Econometría". Fondo editorial ITM. Colombia. (2017).

[33] C. Lagrand, y L. M. Pinzón. "Análisis de datos, métodos y ejemplos". Editorial Escuela Colombiana de Ingeniería. Colombia. (2009).

[34] A. Villasante., S. Vignote y R. Blanco. “Análisis estadístico de los nombres comerciales de maderas en un país (España)”. Madera y bosques, vol. 20 (2), pp. 59-70. (2014).

[35] L. G. Díaz Monroy y M. A. Morales Rivera. “Análisis estadístico de datos categóricos”. *Universidad Nacional de Colombia.* (2009).

[36] J. J. Espinosa. “Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de crédito” Ingeniería Investigación y Tecnología, pp. 1-16. (2020) DOI: https://doi.org/10.22201/fi.25940732e.2020.21.3.022

[37] A. G. Cantú. “Árboles de decisión y su aplicación en el síndrome metabólico”. CIMAT, Centro de Investigación en matemáticas. México. (2019). Disponible en: https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/1013/1/MTY%20TE%203.pdf

[38] C. O. Montalván. “Credit Scoring aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito”. Universidad Andina Simón Bolívar. Ecuador. (2019). Disponible en: https://repositorio.uasb.edu.ec/bitstream/10644/6872/1/T2962-MGFARF-Montalvan-Credit.pdf

[39] G. A. Betancourt. “Las máquinas de soporte vectorial”. Scientia Et Technica, vol. XI, no. 27, pp. 67-72. Colombia. (2005).

[40] J. A. Resendiz. “Las máquinas de vectores de soporte para identificación en línea”. Tesis. Centro de Investigación y Estudios Avanzados del Politécnico Nacional. Departamento de Control Automático. México. (2006).

[41] I. Cañate Quian. “Scoring No Clientes (Modelización con información Big Data)”. *Universidad de Santiago de Compostela*. España. (2018).