Título del Trabajo de Grado: IMPLEMENTACIÓN DE UN MODELO DE RIESGO DE CRÉDITO PARA EL OTORGAMIENTO Y LA RENOVACIÓN ÁGIL DE MICROCRÉDITOS PARA MIPYMES APLICANDO TÉCNICAS DE MACHINE LEARNING.

Autor(es): Julian Ernesto Díaz Arboleda, Jorge Gonzalez Rivera, Miguel Eduardo Rodríguez Vivas

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Luis Eduardo Girón Cruz
Director

Isabel Cristina Garcia Arboleda
Jurado

David Arango Londoño
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 26 de Junio de 2023

Acta de Correcciones al Documento de Trabajo de Grado

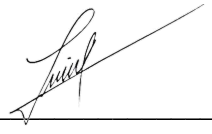
Santiago de Cali, 26 de junio de 2023

Autor(es):
Julian Ernesto Díaz Arboleda, Jorge Gonzalez Rivera, Miguel Eduardo Rodríguez Vivas
ID: 8971666, 0015508, 0036232

Título del Trabajo de Grado: "IMPLEMENTACIÓN DE UN MODELO DE RIESGO DE CRÉDITO PARA EL OTORGAMIENTO Y LA RENOVACIÓN ÁGIL DE MICROCRÉDITOS PARA MIPYMES APLICANDO TÉCNICAS DE MACHINE LEARNING"

Director: Luis Eduardo Girón Cruz

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



Firma del Director del Trabajo de Grado

Santiago de Cali, 29 de mayo de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado "Implementación de un modelo de riesgo de crédito para el otorgamiento y la renovación ágil de microcréditos para MiPymes aplicando técnicas de Machine Learning", el cual será realizado por los estudiantes Julian Ernesto Díaz Arboleda con código 8971666, Jorge Gonzalez Rivera con código 0015508 y Miguel Eduardo Rodríguez Vivas con código 0036232, bajo la dirección del profesor Luis Eduardo Girón Cruz.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Estudiantes:

_____	_____	_____
Julián E. Díaz Arboleda	Jorge González Rivera	Miguel E. Rodríguez Vivas
CC. 1.047.369.681	CC. 16.929.308	CC. 94.536.402

Director:

_____
Luis Eduardo Girón Cruz
CC. 16.632.200

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

Santiago de Cali, 29 de mayo de 2023

Doctora
Gloría Inés Alvarez V.
Directora Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "Implementación de un modelo de riesgo de crédito para el otorgamiento y la renovación ágil de microcréditos para MiPymes aplicando técnicas de Machine Learning" el cual fue realizado por los estudiantes Julian Ernesto Díaz Arboleda con código 8971666, Jorge Gonzalez Rivera con código 0015508 y Miguel Eduardo Rodríguez Vivas con código 0036232 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de Luis Eduardo Girón Cruz.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Estudiantes:

Julián E. Díaz Arboleda	Jorge González Rivera	Miguel E. Rodríguez Vivas
CC. 1.047.369.681	CC. 16.929.308	CC. 94.536.402

Director:

Luis Eduardo Girón Cruz
CC. 16.632.200

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

Fundación
Santo Domingo

Barranquilla, Mayo 29 de 2023

Señores
Julian Ernesto Díaz Arboleda
Jorge Gonzalez Rivera
Miguel Eduardo Rodríguez Vivas
Estudiantes de la Maestría en Ciencia de Datos de la Pontificia Universidad Javeriana Cali

La Fundación Santo Domingo otorga su aval y apoyo para la realización de la investigación propuesta en colaboración con ustedes. Su proyecto, que busca desarrollar un modelo de evaluación de riesgo de crédito basado en técnicas de aprendizaje automático para las MiPymes en los departamentos de Atlántico y Bolívar, es de gran relevancia para promover el desarrollo empresarial y la inclusión financiera en la región.

Autorizamos el uso de la información proporcionada por la Fundación Santo Domingo con fines exclusivamente académicos para uso exclusivo del desarrollo de la investigación anteriormente mencionada, para lo cual deben asegurar que los datos serán tratados de forma confidencial y se garantizará la privacidad y protección de estos, de acuerdo con las leyes y regulaciones aplicables en materia de protección de datos.

Les instamos a llevar a cabo la investigación de manera ética y responsable, respetando los principios de integridad académica. Estamos dispuestos a brindarles el apoyo necesario durante el desarrollo de la investigación y esperamos que esta colaboración sea exitosa y beneficie a ambas partes.

Atentamente,

Jose Alberto Bedoya Ramos
Director Unidad de Financiación y Desarrollo Empresarial
Fundación Santo Domingo
jbedoya@fundacionsantodomingo.org
605 3710707 - Ext 48043

FICHA RESUMEN
PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

TÍTULO: **IMPLEMENTACIÓN DE UN MODELO DE RIESGO DE CRÉDITO PARA EL OTORGAMIENTO Y LA RENOVACIÓN ÁGIL DE MICROCRÉDITOS PARA MIPYMES APLICANDO TÉCNICAS DE MACHINE LEARNING.**

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Ingeniería, Ciencias Económicas y Administrativas, Finanzas, microcréditos.
4. ESTUDIANTE(S): Julián Ernesto Díaz Arboleda con código 8971666, Jorge González Rivera con código 0015508 y Miguel Eduardo Rodríguez Vivas con código 0036232
5. CORREO ELECTRÓNICO: juliandiazec@javerianacali.edu.co, jorge.gonzalez@javerianacali.edu.co, miguelrodriguez@javerianacali.edu.co
6. DIRECCIÓN Y TELEFONO: Calle 18 # 118-250, 3015493966
7. DIRECTOR: Luis Eduardo Girón Cruz
8. VINCULACIÓN DEL DIRECTOR: Planta, Departamento de Economía
9. CORREO ELECTRÓNICO DEL DIRECTOR: legiron@javerianacali.edu.co
10. CODIRECTOR(ES): N/A
11. GRUPO O EMPRESA QUE LO AVALA: Fundación Santo Domingo - Dirección de Financiamiento y Desarrollo empresarial
12. OTROS GRUPOS O EMPRESAS: N/A
13. PALABRAS CLAVE (al menos 5): Scoring de crédito, microcrédito, modelo de riesgo, riesgo de crédito, Machine Learning.
14. ODS QUE APLICA EL PROYECTO (Agenda 2023): 8 Trabajo decente y Crecimiento Económico, 1 Fin de la Pobreza, 9 Industria, innovación e Infraestructura.
15. FECHA DE INICIO: Julio de 2022
16. DURACIÓN ESTIMADA (En meses): 8
17. RESUMEN:

El riesgo de crédito para las micro, pequeñas y medianas empresas (MiPymes) en Colombia representa un desafío significativo para las entidades financieras y las propias empresas, ya que un mal manejo de la concesión de créditos puede generar incumplimientos y pérdidas económicas considerables. En su mayoría, las instituciones financiadoras recurren a modelos de evaluación de riesgo basados en métodos tradicionales basados en la consulta en centrales de riesgo, en donde en gran proporción las MiPymes pueden no estar registradas o peor aún, estar mal calificadas por incumplimientos pasados o falta de historial crediticio, generándose así una autoexclusión de las MiPymes en el sistema financiero. La Fundación Santo Domingo (FSD) y su Dirección de Financiamiento y Desarrollo Empresarial, ofrecen servicios financieros y no financieros para apoyar el desarrollo empresarial y la creación de empleo en Colombia. Su objetivo principal es fomentar la inclusión financiera y el acceso al crédito para MiPymes en el país. Como alternativa a los modelos tradicionales de valoración del riesgo de crédito hoy en día son cada vez más utilizados aquellos que incorporan el procesamiento de los datos con técnicas de Machine Learning (ML), bajo este contexto, en el presente proyecto se presenta la implementación de un modelo de riesgo de crédito basado en técnicas de ML para la FSD, que le permita la concesión y renovación de microcréditos a MiPymes del departamento de Atlántico y Bolivar. Para lograr este objetivo, el proyecto incorpora la revisión del estado del arte relacionado con el problema, la caracterización y análisis de los datos históricos de préstamos, la limpieza y preparación de los datos, la selección de características relevantes, la reducción de la dimensionalidad y la implementación del algoritmo de ML para crear el modelo predictivo. Se espera que la implementación de este modelo permita a la FDS tomar decisiones de préstamo más precisas y efectivas, lo que a su vez puede aumentar la tasa de aprobación de préstamos y reducirá la tasa de incumplimiento de pagos.

IMPLEMENTACIÓN DE UN MODELO DE RIESGO DE CRÉDITO PARA EL OTORGAMIENTO Y LA RENOVACIÓN ÁGIL DE MICROCRÉDITOS PARA MIPYMES APLICANDO TÉCNICAS DE MACHINE LEARNING.

Julián Ernesto Díaz Arboleda
Jorge González Rivera
Miguel Eduardo Rodríguez Vivas

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director:
Luis Eduardo Girón Cruz

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MARZO DE 2023

ABSTRACT

Credit risk for micro, small, and medium-sized enterprises (MSMEs) in Colombia represents a significant challenge for financial institutions and the companies themselves, as mismanagement of credit granting can generate considerable defaults and economic losses. Mostly, financial institutions resort to risk assessment models based on traditional methods relying on credit bureau queries, where a large proportion of MSMEs may not be registered or, worse yet, poorly rated due to past defaults or lack of credit history, generating self-exclusion of MSMEs from the financial system. The Santo Domingo Foundation (FSD) and its Financing and Business Development Directorate offer financial and non-financial services to support business development and job creation in Colombia. Its main objective is to promote financial inclusion and access to credit for MSMEs in the country. As an alternative to traditional credit risk valuation models, those incorporating data processing with Machine Learning (ML) techniques are increasingly being used. In this context, this project presents the implementation of an ML-based credit risk model for the FSD, enabling the granting and renewal of microcredits to MSMEs in the departments of Atlántico and Bolivar. To achieve this goal, the project includes the review of the state of the art related to the problem, characterization and analysis of historical loan data, data cleaning and preparation, selection of relevant features, dimensionality reduction, and implementation of the ML algorithm to create the predictive model. It is expected that the implementation of this model will enable the FSD to make more accurate and effective loan decisions, which in turn can increase the loan approval rate and reduce the payment default rate.

Keywords: credit scoring, microloans, risk model, credit risk, Machine Learning.

RESUMEN

El riesgo de crédito para las micro, pequeñas y medianas empresas (MiPymes) en Colombia representa un desafío significativo para las entidades financieras y las propias empresas, ya que un mal manejo de la concesión de créditos puede generar incumplimientos y pérdidas económicas considerables. En su mayoría, las instituciones financiadoras recurren a modelos de evaluación de riesgo basados en métodos tradicionales basados en la consulta en centrales de riesgo, en donde en gran proporción las MiPymes pueden no estar registradas o peor aún, estar mal calificadas por incumplimientos pasados o falta de historial crediticio, generándose así una autoexclusión de las MiPymes en el sistema financiero. La Fundación Santo Domingo (FSD) y su Dirección de Financiamiento y Desarrollo Empresarial, ofrecen servicios financieros y no financieros para apoyar el desarrollo empresarial y la creación de empleo en Colombia. Su objetivo principal es fomentar la inclusión financiera y el acceso al crédito para MiPymes en el país. Como alternativa a los modelos tradicionales de valoración del riesgo de crédito hoy en día son cada vez más utilizados aquellos que incorporan el procesamiento de los datos con técnicas de Machine Learning (ML), bajo este contexto, en el presente proyecto se presenta la implementación de un modelo de riesgo de crédito basado en técnicas de ML para la FSD, que le permita la concesión y renovación de microcréditos a MiPymes del departamento de Atlántico y Bolívar. Para lograr este objetivo, el proyecto incorpora la revisión del estado del arte relacionado con el problema, la caracterización y análisis de los datos históricos de préstamos, la limpieza y preparación de los datos, la selección de características relevantes, la reducción de la dimensionalidad y la implementación del algoritmo de ML para crear el modelo predictivo. Se espera que la implementación de este modelo permita a la FDS tomar decisiones de préstamo más precisas y efectivas, lo que a su vez puede aumentar la tasa de aprobación de préstamos y reducirá la tasa de incumplimiento de pagos.

Palabras Claves: puntaje de crédito, microcrédito, modelo de riesgo, riesgo de crédito, Machine Learning.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

1. INTRODUCCIÓN

En Colombia, las micro, pequeñas y medianas empresas (MiPymes) son una parte fundamental de la economía del país [1]. Sin embargo, muchas de estas empresas tienen dificultades para obtener financiamiento por falta de historial crediticio y falta de información financiera confiable. Esto ha llevado a un alto riesgo de crédito para las instituciones financieras que otorgan préstamos a estas empresas. Además, cuando los clientes incumplen sus compromisos y las condiciones pactadas se generan perdidas, por lo que se requiere que las instituciones tengan mecanismos que permitan identificar la probabilidad de incumplimiento de los clientes desde que solicitan los créditos, para que la decisión de aprobación se pueda tomar mediante la medición de la exposición de riesgo inherente a cada crédito por otorgar.

Con el propósito de contribuir al impulso de las estrategias de inclusión financiera en los departamentos de Atlántico y Bolívar, la Fundación Santo Domingo (FSD) y su Dirección de Financiamiento y Desarrollo Empresarial, busca ampliar sus coberturas para beneficiar a muchos más microempresarios y acelerar el crecimiento productivo y sostenible de sus negocios. La FSD contempla actualmente un modelo de evaluación de riesgo basado en los métodos tradicionales de consulta en las centrales de riesgo, el comportamiento del microempresario en su relación histórica con la entidad y las condiciones del entorno económico nacional y sectorial.

No obstante, el uso de técnicas de aprendizaje automático para el desarrollo de modelos de riesgo de crédito se ha convertido en una solución cada vez más popular en la industria financiera [2]. Estos modelos pueden ayudar a las instituciones financieras a tomar decisiones más informadas y precisas al evaluar la capacidad de las MiPymes para cumplir sus obligaciones financieras.

En este contexto, con este trabajo se presenta en primer lugar abordar el marco conceptual, en el cual se enuncia el panorama de las MiPymes en Colombia, se contextualiza sobre el riesgo de crédito y lo que implica para la entidad financiadora, qué es un modelo de riesgo y para qué sirve, así como los métodos de solución basados en aprendizaje automático. Posteriormente se abordan las fases de implementación de un modelo de riesgo de crédito para el otorgamiento y la renovación de microcréditos para MiPymes aplicando técnicas de aprendizaje automático. Para lo cual, se utilizaron técnicas de procesamiento de datos y análisis estadístico avanzado para analizar diferentes variables financieras y no financieras que influyen en el riesgo crediticio, se presentan los resultados del modelamiento y evaluación de un modelo predictivo para la Fundación Santo Domingo capaz de clasificar a las MiPymes según su nivel de riesgo crediticio y que ayuda a tomar decisiones más acertadas y justas en cuanto a la concesión y renovación de créditos, finalmente se entregan las conclusiones del trabajo.

2. DEFINICIÓN DEL PROBLEMA

2.1. PLANTEAMIENTO DEL PROBLEMA

Según el RUES (Registro Único Empresarial y Social), el 99% del tejido empresarial en Colombia se debe a micro, pequeñas y medianas empresas, MiPymes. Estas empresas son parte importante de la activación económica y representan un motor frente al desarrollo económico y social ya que generan aproximadamente el 79% del empleo y aportan 40% al Producto Interno Bruto (PIB) del país [1]. A pesar de su relevancia para el desarrollo económico del país, este segmento de empresas enfrenta múltiples retos para su supervivencia, sólo 34 de cada de 100 nuevas empresas, siguen funcionando luego de 5 años y dentro de los retos identificados para las MiPymes se encuentran prácticas gerenciales adecuadas, acceso y mantenimiento de capital humano adecuado, innovación, uso de tecnologías de información y alianzas. [3] Sin embargo, cada una de estas dimensiones requiere de inversiones por parte del empresario y acceso a capital, siendo este último aspecto clave dadas las actuales condiciones del país ya que cerca del 38% de los créditos negados a microempresarios obedecen a falta de historial crediticio o reportes negativos [4], generando un proceso de autoexclusión al sistema financiero por parte de los microempresarios.

Esta situación obliga a muchos microempresarios a recurrir a fuentes de financiación sin respaldo legal (como el “gota a gota”) o utilizar recursos propios que ponen en riesgo la capacidad de crecimiento de la empresa, y en algunos casos, el patrimonio familiar [5]. La informalidad de los procesos contables, financieros y operacionales, se constituyen, también, en una razón relevante para que las MiPymes tengan impedimento de acceso al crédito según las condiciones del sector financiero tradicional, puesto que sin esa información es difícil medir la capacidad de pago de estas empresas y valorar de forma adecuada el riesgo de otorgar un crédito.

Este problema no es ajeno para la Fundación Santo Domingo (FSD) y su Dirección de Financiamiento y Desarrollo Empresarial, la cual gestiona e impulsa servicios de inclusión financiera para que familias de bajos ingresos impulsen y aceleren el crecimiento productivo y sostenible de sus negocios, en los departamentos de Atlántico y Bolívar. Para valorar el riesgo de cada crédito la Fundación utiliza la información reportada por la central de riesgo, DataCrédito [6], la cual otorga un puntaje a partir de la evaluación de 300 variables distribuidas en 5 dimensiones: historial de pagos (35%), monto adeudado al sistema financiero (30%), antigüedad historial crediticio (15%), tipo de cuentas (10%) y actividad crediticia reciente (10%), información que es bastante útil para tomar la decisión de aprobación, sin embargo, debido al impacto de la pandemia por COVID -19, muchas empresas han visto afectado su puntaje de crédito, debido a impagos, retrasos y disminución de la actividad productiva, lo cual ha llevado a que a partir del año 2022 siete de diez clientes potenciales consultados tengan un reporte negativo o carezcan de historia crediticia al ser nuevas empresas, haciendo que se tengan que implementar mecanismos subjetivos para complementar la evaluación, y poder determinar la posibilidad del otorgamiento del crédito, dentro del nivel o apetito de riesgo definido por la organización, con el objetivo de cumplir las metas de colocación de cartera.

Este proceso de evaluación personalizado, si bien está permitiendo alcanzar una mayor tasa de colocación

en comparación a sí sólo se tomara la decisión basada en el puntaje de DataCrédito no ha logrado medir de forma acertada el riesgo para la Fundación, puesto que a marzo del 2022 el indicador de cartera vencida (ICV) se ubicaba en el 27.2% [7], frente al promedio del 7.5% de la cartera de microfinanzas de los diferentes actores del sector.

Estos niveles de cartera tienen un impacto negativo en las utilidades y en el capital de trabajo disponible, debido a la provisión de cartera que debe realizarse, lo cual obliga a buscar nuevos mecanismos de fondeo, probablemente con un mayor costo a los utilizados de manera regular. A partir de esta situación la Fundación ha requerido de eficientes mecanismos que reduzcan los índices del no pago, valorar y medir el riesgo de cada nuevo crédito o renovación, bajo las condiciones del contexto económico y sectorial actual, en armonía con un crecimiento sostenible y rentable que permita generar utilidades y apalancar su crecimiento.

Por su parte, el uso de algoritmos de aprendizaje automático permite una evaluación más precisa y eficiente del riesgo crediticio de los solicitantes de crédito [8]. Los modelos de evaluación de riesgo de crédito basados en aprendizaje automático pueden analizar grandes cantidades de datos, incluyendo datos no estructurados, y detectar patrones y relaciones que podrían ser difíciles de identificar por medios convencionales. Además, los modelos de aprendizaje automático pueden ser entrenados en datos históricos para aprender de los patrones de comportamiento de los usuarios, lo que les permite mejorar su precisión con el tiempo y adaptarse a diferentes contextos y situaciones. Otro aspecto importante es que los modelos de aprendizaje automático pueden ser más justos y menos sesgados que los modelos tradicionales de scoring de crédito, ya que utilizan una amplia variedad de datos para evaluar el riesgo crediticio en lugar de depender de unas pocas variables o características. Esto puede ayudar a reducir la discriminación y mejorar la inclusión financiera para las personas que han sido históricamente marginadas o excluidas del sistema crediticio.

2.2. FORMULACIÓN DEL PROBLEMA

¿Cómo puede la Fundación Santo Domingo construir e implementar un modelo de evaluación de riesgo de crédito basado en técnicas de Machine Learning que le permita una disminución de la tasa de incumplimiento?

3. OBJETIVOS DEL PROYECTO

3.1. OBJETIVO GENERAL

Desarrollar un modelo de riesgo de crédito basado en técnicas de Machine Learning para la Fundación Santo domingo, que le permita la concesión y renovación de microcréditos a MiPymes del departamento de Atlántico y Bolivar.

3.2. OBJETIVOS ESPECÍFICOS

- Caracterizar y explorar las fuentes de datos suministradas por la Fundación Santo Domingo.
- Efectuar el entrenamiento de los modelos en Machine Learning que apliquen para la evaluación de concesión o renovación de microcréditos.
- Realizar un análisis comparativo de la precisión de los modelos utilizados en la concesión o renovación de microcréditos.
- Evaluar la efectividad y precisión del modelo seleccionado, en términos de la toma de decisiones sobre la concesión y renovación de microcréditos.

4. MARCO TEÓRICO

4.1. Contexto de las MiPymes en Colombia

En Colombia, las MiPymes (Micro, Pequeñas y Medianas Empresas) están definidas por la Ley 905 de 2004 [9], que establece que las MiPymes son aquellas empresas que cumplen con uno o varios de los siguientes criterios:

- Microempresa: empresa con un máximo de 10 trabajadores y ventas anuales o activos totales que no superen los 501 salarios mínimos legales mensuales vigentes (SMMLV).
- Pequeña empresa: empresa con un máximo de 50 trabajadores y ventas anuales o activos totales que no superen los 5.000 SMMLV.
- Mediana empresa: empresa con un máximo de 200 trabajadores y ventas anuales o activos totales que no superen los 30.000 SMMLV.
- Estos límites los actualiza cada año el gobierno colombiano para ajustarse a la inflación y a las necesidades del mercado.

Este sector, según cifras del RUES (Registro Único Empresarial y Social) en 2022, corresponde al 99% del tejido empresarial en Colombia, donde el 73,8% corresponde a personas naturales y el 26,2% a sociedades. Estas generan aproximadamente el 79% del empleo y aportan 40% al Producto Interno Bruto (PIB) del país [1]. Por tal razón, se convierten en jugadores indiscutibles en el desarrollo económico del país, en la generación de empleo, en la productividad y en la competitividad.

A pesar de su importancia, las MiPymes en Colombia enfrentan varios desafíos, incluyendo:

- Acceso limitado a financiamiento: Las MiPymes tienen mayores dificultades para obtener crédito que las grandes empresas debido a su falta de historial crediticio, garantías y recursos financieros limitados.
- Falta de innovación: La mayoría de las MiPymes en Colombia se dedican a sectores tradicionales y tienen poca capacidad de innovación tecnológica o de productos.
- Bajo nivel de productividad: Las MiPymes en Colombia tienen una baja productividad en comparación con las empresas más grandes debido a una menor inversión en tecnología, capacitación y procesos de gestión.
- Competencia desigual: Las MiPymes enfrentan competencia desleal de empresas informales que no cumplen con las regulaciones y están fuera del alcance de la fiscalización.

Para abordar estos desafíos, el gobierno colombiano [10] ha implementado programas y políticas para apoyar el desarrollo de las MiPymes, incluyendo medidas de financiamiento, capacitación empresarial y apoyo para la innovación.

4.2. Contexto de la Fundación Santo Domingo - Unidad de Financiamiento y Desarrollo Empresarial

La Fundación Santo Domingo (FSD) es una organización sin ánimo de lucro establecida en Colombia en 1956 por los empresarios Julio Mario Santo Domingo y Beatrice Dávila. La fundación tiene como objetivo

principal contribuir al desarrollo social y económico del país a través de la inversión en proyectos de educación, salud, cultura, medio ambiente, desarrollo económico y comunitario. [11]

La FSD ha financiado y liderado proyectos en diversas áreas en todo el país, incluyendo la construcción de hospitales, escuelas y bibliotecas, la promoción de iniciativas empresariales y la preservación del patrimonio cultural y natural. La organización también es propietaria de empresas en varios sectores, como la industria cervecera, la cadena de tiendas de grandes superficies y la compañía de medios de comunicación Caracol Televisión. Los ingresos generados por estas empresas son destinados a financiar los proyectos sociales y comunitarios de la Fundación Santo Domingo.

Por su parte, la Unidad de Financiamiento y Desarrollo Empresarial (UFDE) [12] de la FDS ofrece servicios financieros y no financieros para apoyar el desarrollo empresarial y la creación de empleo en Colombia. Su objetivo principal es fomentar la inclusión financiera y el acceso al crédito para MiPymes en el país.

La UFDE ofrece una variedad de productos y servicios financieros, como préstamos, líneas de crédito y garantías, con tasas de interés y plazos competitivos. Además, también brinda asesoramiento y capacitación empresarial para ayudar a los emprendedores y empresarios a mejorar sus habilidades en áreas como finanzas, marketing y gestión de recursos humanos. Y trabaja en colaboración con otras entidades y organizaciones para apoyar la creación y el fortalecimiento de las MiPymes en Colombia, y contribuir así al desarrollo económico y social del país.

La Fundación cuenta con su propia segmentación o clasificación de las MiPymes a partir de su estrategia y modelo de negocio en sus territorios de operación (Cartagena, Barranquilla, y municipios aledaños a estas dos ciudades). Estos tres segmentos de clientes se determinan a partir de la antigüedad del negocio, nivel de ventas o activos, historial crediticio y actividades económicas. En cuanto a esta última variable, se cuentan con unas actividades económicas objetivo de esfuerzo comercial por cada segmento, sin embargo, esto no es excluyente para el otorgamiento de crédito a personas o empresas con negocios. Esta clasificación interna de las MiPymes difiere a la establecida en la Ley 905 de 2004 [9], al incluir elementos adicionales como la antigüedad de la unidad productiva, historial crediticio y actividad económica. Así mismo, excluye la variable de número de empleos generados. Las características de cada segmento se describen en la Tabla *1*.

Para atender estos clientes, se cuenta con 5 líneas de crédito disponible, en donde a partir del segmento en el que se encuentre cada cliente, se cuentan con unas condiciones para montos de desembolso y plazos. Las líneas de crédito se describen a continuación:

- **Capital de Trabajo:** Dirigida para atender las necesidades a corto plazo de liquidez como compra de materias primas y mercancías, pagos a proveedores y empleados. Esta línea permite consolidar el crecimiento de las empresas clientes.
- **Pago a Proveedores:** Esta línea de crédito permite al empresario un manejo optimo y oportuno a sus compromisos de pago con su red de proveedores, aprovechando descuentos por pronto pago, volúmenes de compra. Esta línea se gira directamente a los proveedores.

- **Negociación de Facturas:** Facilita al empresario un manejo optimo y oportuno a sus compromisos de pago con su red de proveedores, aprovechando descuentos por pronto pago, volúmenes de compra. Esta línea se gira directamente a los proveedores.
- **Adquisición de activos Fijos:** Es un crédito diseñado para financiar los proyectos de expansión y compra en activos fijos de las empresas. Dirigida a financiar necesidades de largo plazo
- **Adecuaciones Locativas:** Línea destinada al respaldar a nuestros empresarios en la financiación de sus proyectos de mejoramiento o adecuación de sus negocios. No requiere hipoteca de bienes inmuebles.

Tabla 1 - Características segmentos de clientes Unidad de Desarrollo Empresarial

Variable	Emprendedor independiente	Emprendedor en desarrollo	Comercial
Descripción	Persona natural con unidad básica de negocio, que busca financiarse para sostener su actividad y proveer con esta a su unidad familiar	Persona natural con unidades de negocio vinculada a cámara de comercio, que busca el crecimiento de su actividad a través de financiación y acompañamiento empresarial	Persona natural o jurídica con mínimo 3 años de antigüedad, con información financiera formal y necesidades de financiación y escalonamiento
Antigüedad del negocio (mínimo)	1 año	6 meses en cámara de comercio	>= 3 Años micro >3 años emprendedor etapa temprana
Ingresos mensuales ventas	Hasta 8 SMMLV	Mayor a 8 SMMLV	Activos superiores a 300 SMMLV
Historial crédito	Con o sin experiencia crediticia	Con o sin experiencia crediticia	Con o sin experiencia crediticia
Actividades económicas foco	○ Comercio de ropa ○ Confecciones ○ Misceláneas ○ Restaurantes ○ Venta de comida rápida en casa ○ Ferreterías	○ Confecciones ○ Comercio de ropa ○ Restaurantes y gastronomía ○ Alquiler y arrendamiento ○ Ferreterías	○ Construcción ○ Industria alimentos ○ Transporte ○ Automotriz ○ Ferreterías

Fuente: elaboración propia

4.3. Microcrédito

El microcrédito es un tipo de préstamo de pequeña cantidad de dinero [13] que se otorga a personas de bajos ingresos o a pequeñas empresas que no tienen acceso a los servicios financieros tradicionales. El objetivo del microcrédito es apoyar el desarrollo económico y social de comunidades y regiones con altos índices de pobreza, a través del financiamiento de proyectos productivos.

Los microcréditos suelen ser otorgados por organizaciones no gubernamentales (ONG), cooperativas de crédito, bancos y otras entidades financieras especializadas. Estas organizaciones suelen tener requisitos

de crédito menos rigurosos que los bancos tradicionales y pueden ofrecer tasas de interés más bajas y plazos de pago flexibles.

El microcrédito se utiliza comúnmente para financiar pequeñas empresas o proyectos productivos, como la compra de maquinaria o materias primas, la expansión de un negocio existente, o la creación de un nuevo negocio. El objetivo final es permitir a los prestatarios mejorar sus ingresos y su calidad de vida, y contribuir al desarrollo económico de sus comunidades.

4.4. Riesgo de crédito

El riesgo de crédito se define como la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que un deudor o contraparte incumpla sus obligaciones [14]. El Riesgo de crédito se encuentra regulado en Colombia por la Superintendencia Financiera y está estipulado a través de la Circular Externa 100 de 1995 y la Circular Externa de 2002, donde todas las entidades vigiladas deben cumplir con el establecimiento de un Sistema de Administración de Riesgo, con el fin de evitar poner en peligro la situación de solvencia y liquidez. Por tanto, las entidades financieras deben cumplir con un adecuado sistema de administración de riesgo de crédito y definir unas políticas claras al respecto. Dentro de todo este sistema, se incluyen las diferentes fuentes de financiación, las garantías que se deben exigir, las provisiones que deben tener, el tipo de créditos a financiar, las diferentes variables que van a ser estudiadas por cada segmento que vaya a ser atendido (personas naturales, pymes, empresas grandes) y el desarrollo de modelos para el otorgamiento de créditos.

En Colombia, se pueden prestar servicios financieros sin estar sujeto a regulación, dado que la regulación se aplica para aquellas entidades que captan recursos del público y no a la colocación de crédito. Este arreglo normativo permite que las nacientes Fintech u organizaciones del sector social como ONG puedan otorgar créditos sin estar obligados al cumplimiento de indicadores de Solvencia o Liquidez requeridas por esta Superintendencia. El no estar regulada, no la exime de cumplir algunas disposiciones o normativas dictadas por la Superintendencia Financiera u otras entidades del estado para el ejercicio de su actividad, tales como el Decreto 1072 de 2015 [15] el cual establece entre otras "*disposiciones para las operaciones de crédito, disposiciones relacionadas con operaciones de crédito otorgadas por personas naturales o jurídicas cuyo control y vigilancia sobre su actividad crediticia no haya sido asignada a alguna autoridad administrativa en particular*".

A pesar de no estar regulada, la fundación cuenta con un Gobierno Corporativo que realiza control y seguimiento a las métricas de negocio y al cumplimiento de los diferentes requisitos normativos. Para esto, la organización cuenta con un área de auditoría interna la cual se apoya en una firma de auditoría externa. La actividad de microcrédito está incluida de forma permanente en el plan de auditoria anual, en el comité de auditoría. Así mismo, en las siguientes instancias del Gobierno Corporativa se realiza seguimiento a la operación y sus métricas de desempeño: Comité directivo, Comité de Auditoría, Comité Asesor de Financiación y Desarrollo Empresarial, Junta directiva y el Consejo de Familia.

4.5. Machine Learning

El Machine Learning (Aprendizaje automático en español), es una rama de la inteligencia artificial que se centra en desarrollar algoritmos y modelos capaces de aprender y mejorar automáticamente a partir de los datos [15]. Estos modelos se dividen en dos categorías principales: supervisados y no supervisados.

Los modelos supervisados se utilizan cuando se dispone de un conjunto de datos etiquetados, es decir, se conoce la variable objetivo que se quiere predecir. Dentro de los modelos supervisados se encuentran los modelos de clasificación y los modelos de regresión. Los modelos de clasificación se utilizan cuando la variable objetivo es categórica y se busca asignar una clase a cada instancia del conjunto de datos. Por otro lado, los modelos de regresión se emplean cuando la variable objetivo es numérica y se busca predecir un valor continuo.

Los modelos no supervisados son utilizados cuando no se dispone de datos etiquetados o de una variable objetivo-específica. Estos modelos se centran en encontrar patrones, estructuras o agrupaciones inherentes en los datos sin la necesidad de una guía explícita. Otra aplicación de los modelos no supervisados es que incluyen la reducción de dimensionalidad, como el análisis de componentes principales (PCA) y el análisis de conglomerados, que permiten resumir la información en un espacio de menor dimensionalidad mientras se conserva la mayor cantidad de variabilidad posible.

La fase de entrenamiento de un modelo de machine learning consiste en utilizar un conjunto de datos de entrenamiento para ajustar los parámetros del modelo y lograr que se adapte a los patrones y relaciones presentes en los datos. Una vez que el modelo ha sido entrenado, se pasa a la fase de prueba o test, donde se utiliza un conjunto de datos separado y no visto previamente por el modelo para evaluar su rendimiento y medir su capacidad de generalización.

Para validar y evaluar la calidad de los modelos, se emplean diferentes técnicas. Una de ellas es la validación cruzada, que consiste en dividir el conjunto de datos en varias partes y realizar múltiples entrenamientos y pruebas, alternando qué partes se usan para cada fase. Otra técnica común es la separación del conjunto de datos en conjunto de entrenamiento, conjunto de validación y conjunto de prueba, donde el conjunto de validación se utiliza para ajustar los hiperparámetros del modelo y el conjunto de prueba se utiliza para evaluar su rendimiento final.

En cuanto a las métricas de calidad, estas varían dependiendo del tipo de modelo y del objetivo específico. Para modelos de clasificación, se suelen utilizar métricas como la precisión, el recall, la F1-score y la matriz de confusión. En el caso de los modelos de regresión, las métricas más comunes son el error medio cuadrático (MSE), el error medio absoluto (MAE) y el coeficiente de determinación (R^2).

4.6. Machine Learning y riesgo de crédito

En el otorgamiento de créditos, el Machine Learning se utiliza para analizar los datos crediticios de los solicitantes y predecir su capacidad de pago y riesgo crediticio. El Aprendizaje Automático permite

construir modelos predictivos precisos a partir de grandes cantidades de datos históricos, lo que permite identificar patrones y tendencias en los datos de crédito. Estos modelos son capaces de aprender de forma autónoma y mejorar su capacidad de predicción con el tiempo. Algunas de las técnicas de Machine Learning que se utilizan en el otorgamiento de créditos son el análisis discriminante, las máquinas de soporte vectorial, la regresión logística, los árboles de decisión y las redes neuronales artificiales. [16]

4.7. Modelos de aprobación de crédito

Los modelos de clasificación de aprobación de créditos son algoritmos de aprendizaje automático diseñados para evaluar la elegibilidad de los solicitantes de crédito y tomar decisiones sobre la aprobación o rechazo de préstamos. Estos modelos se entrenan utilizando datos históricos que contienen información relevante sobre los solicitantes, como su historial crediticio, ingresos, deudas, empleo, entre otros factores. A través de técnicas de aprendizaje automático, los modelos analizan y extraen patrones y relaciones en los datos para predecir la probabilidad de que un solicitante cumpla con sus obligaciones crediticias.

Algunos de los modelos de clasificación comúnmente utilizados en la aprobación de créditos incluyen:

- Modelo Logit simple: también conocido como regresión logística. Utiliza la función logística para modelar la probabilidad de que una muestra pertenezca a una clase específica en función de sus variables predictoras. Este modelo permite interpretar los coeficientes asociados a cada variable para entender su impacto en la clasificación.
- Árboles de Decisión: Estos modelos utilizan una estructura de árbol para tomar decisiones basadas en una serie de condiciones y características del solicitante.
- Random Forest: Es un ensamblaje de múltiples árboles de decisión. Cada árbol se entrena con una muestra aleatoria de los datos y vota para la clasificación final.
- Gradient Boosting: Este modelo construye una secuencia de árboles de decisión, donde cada árbol se enfoca en corregir los errores cometidos por el árbol anterior, mejorando gradualmente la precisión de las predicciones.
- Naive Bayes: Es un modelo probabilístico que asume la independencia condicional de las características y utiliza el teorema de Bayes para calcular la probabilidad de pertenencia a una clase determinada.
- K-Nearest Neighbors (KNN): Este modelo clasifica a los solicitantes en función de la similitud con los vecinos más cercanos en el espacio de características.
- Extra Trees Classifier: Este modelo es una variante del algoritmo Random Forest que se basa en la técnica de ensamblaje de árboles de decisión.
- Máquinas de soporte vectorial: Estas buscan encontrar un hiperplano óptimo que separe las diferentes clases en un espacio multidimensional, maximizando el margen entre las muestras más cercanas de cada clase.

Estos modelos de clasificación utilizan diferentes enfoques y técnicas para evaluar y predecir la probabilidad de riesgo crediticio de los solicitantes. Su elección depende de las características del conjunto de datos y las necesidades específicas de la institución financiera.

5. METODOLOGÍA

A continuación, se presenta la ruta metodológica que se siguió para el desarrollo de la solución, partiendo de un marco teórico previamente establecido, que tiene como base todo el músculo cognitivo de las técnicas de machine Learning para el desarrollo de modelos de evaluación de riesgo de crédito, lo cual permite implementar y aplicar los siguientes pasos:

Definición del problema: El primer paso es definir el problema de negocio y el objetivo del modelo de Machine Learning. En este caso, el objetivo es determinar el otorgamiento o renovación de créditos a clientes potenciales.

Selección de datos: Selección y recolección los datos necesarios para el modelo, como información financiera, historial crediticio, información personal, etc. Para lo cual la base de datos fue suministrada por la Fundación Santo Domingo y fue preprocesada y segmentada para los periodos de análisis.

Análisis exploratorio de datos: se realiza un análisis exploratorio de los datos para identificar patrones y relaciones en los datos y comprender mejor la distribución de los datos. Esto incluye gráficos, estadísticas descriptivas para comprender mejor el problema de negocio.

Preprocesamiento de datos: Se realizar la limpieza y transformación de los datos para prepararlos para el modelado. Esto incluye la imputación de valores faltantes, la normalización de los datos y la selección de características relevantes.

Selección del modelo: Seleccionar el modelo de Machine Learning que mejor se adapte al problema y a los datos. Esta fase se alinea con la revisión de la literatura frente a los diferentes modelos que pueden usarse para problemas de crédito.

Entrenamiento del modelo: Entrenar los modelos seleccionados con los datos preparados. El objetivo es ajustar los parámetros del modelo para obtener la mejor precisión posible.

Validación del modelo: Evaluar la calidad de los modelos entrenados mediante técnicas de validación, como la validación cruzada, y seleccionar el mejor modelo en función de las métricas de evaluación.

Implementación y pruebas: Se confirman los parámetros para modelo seleccionado y se realizar pruebas para asegurarse de que el modelo funciona correctamente.

6. RESULTADOS

En esta sección se presentan los resultados de la caracterización y exploración de las fuentes de datos suministradas por la Fundación Santo Domingo, así como los resultados del entrenamiento de los modelos de Machine Learning aplicados para la evaluación de concesión o renovación de microcréditos. También se incluye un análisis comparativo de la precisión de los modelos utilizados, evaluando su efectividad y precisión en la toma de decisiones sobre la concesión y renovación de microcréditos. Los resultados obtenidos proporcionan una visión detallada del desempeño del modelo desarrollado y su capacidad para mejorar el proceso de evaluación de riesgo de crédito de la Fundación Santo Domingo.

6.1. Exploración y Caracterización de los Datos

Se trabajó con una base de datos de 3.507 créditos generados desde el 01 de enero de 2021 hasta el 15 de julio de 2022 con comportamiento de pago de al menos 4 meses, periodo crítico identificado por la entidad en la que la mayor cantidad de clientes caen en default; variable dependiente, aquellos clientes que al cierre del mes tenían 30 días o más de mora. La base de datos inicial se encontraba conformada por 27 variables explicativas que se detallan a continuación.

El análisis exploratorio se realizó a partir de la segmentación propia de MiPymes definida por la Fundación para sus clientes, la cual se realiza con unas variables diferentes establecidas en la Ley 905 de 2004 [9] de acuerdo con su modelo de negocio y estrategia. Debido a las pocas variables disponibles para los clientes pertenecientes al segmento comercial y la decisión de la organización de enfocar su operación en los clientes de tipo persona natural, se omite del ejercicio los microempresarios de este segmento.

6.1.1. Variables explicativas no categóricas

- **Edad:** Años de vida del cliente, puede tomar un valor máximo de 86. El 10% de las observaciones no se reportan, por lo cual se decide imputar con la media.
- **Personas a cargo:** Número de personas económicamente a cargo del cliente. Puede tomar valores del 0 hasta el 9.
- **Activos:** Valor de los activos declarados de la unidad productiva**.** Puede tomar valores del 0 hasta un máximo de $1.972.222.112.
- **Pasivos:** Valor de los pasivos declarados de la unidad productiva**.** Puede tomar valores del 0 hasta un máximo de $1.357.172.874.
- **Ingresos:** Valor de los ingresos mensuales declarados de la unidad productiva**.** Puede tomar valores del 0 hasta un máximo de $1.257.290.342.
- **Gastos:** Valor de los gastos mensuales declarados de la unidad productiva**.** Puede tomar valores del 0 hasta un máximo de $1.252.744.104.
- **Monto crédito:** Valor del crédito desembolsado al cliente**.** Puede tomar valores desde $200.000 hasta $160.273.533.

- **Cobertura garantía:** Indica que porcentaje del crédito es cubierto por algún tipo de garantía bancaria. Puede tomar valores desde 0 (créditos sin garantía) hasta del 0.9. Menos del 1% de las observaciones no se reportan, por lo cual se decide imputar con la media.
- **Tasa crédito:** Indica la tasa efectiva anual a la que fue entregada el crédito. Puede tomar valores desde 12 hasta del 47.
- **No cuotas:** Indica el número de cuotas en las que se pagaría el crédito. Puede tomar valores desde 1 hasta del 153.
- **Valor cuota:** Indica el valor de la cuota mensual a la que fue entregada el crédito. Puede tomar valores desde $30.000 hasta $12.500.000.
- **Empleos permanentes:** Indica el número de empleos permanentes reportados por el cliente. Puede tomar valores desde 0 hasta 416. Menos del 1% de las observaciones no se reportan, por lo cual se decide imputar con la media.
- **Mes default:** Indica el número de meses transcurridos antes de que el cliente incurra en 30 días de mora. Puede tomar valores desde 0 hasta 21.

En la Tabla *2* se realiza un análisis descriptivo de cada una de las 13 variables no categóricas.

Tabla 2 - Caracterización de variables no categóricas

Variable	Número de observaciones	Datos faltantes	Promedio	Mínimo	Máximo
Edad	3507	259	45.1	18	86
Personas a cargo	3507	0	0.32	0	9
Activos	3507	0	$ 26.312.818	$ -	$ 1.972.222.112
Pasivos	3507	0	$ 6.458.602	$ -	$ 1.357.172.874
Ingresos	3507	0	$ 8.879.397	$ -	$ 1.257.290.342
Gastos	3507	0	$ 6.432.129	$ -	$ 1.252.744.104
Monto crédito	3507	0	$ 4.904.549	$ 200.000	$ 160.273.533
Cobertura garantía	3507	11	0.37	0	0,9
Tasa	3507	0	31.1	12%	47%
No cuotas	3507	0	21.5	1	153
Valor cuota	3507	0	$ 402.580	$ 31.294	$ 12.500.000
Empleos permanentes	3507	1	1.95	0	416
Default	3507	0	No aplica	0	1
Mes default	3507	0	2.37	0	21

Fuente: elaboración propia

6.1.2. Variables categóricas

- **Género:** Puede tomar los valores de femenino o masculino.
- **Estrato:** Puede tomar valores del 1 al 6 de acuerdo con la estratificación socioeconómica del cliente.
- **Barrio:** Nombres de los barrios de ubicación de la unidad productiva. Existen 408 categorías, se pretende reducir su dimensionalidad agrupándola por localidades o IPM de localidades. Debido a no existir aún definidos los criterios de agrupamiento, la variable no se incluye en el modelo final.

- **Ciudad:** Municipio de ubicación de la unidad productiva. Existen 27 ciudades**.**
- **Profesión:** Profesión u oficio declarado del microempresario.
- **Nivel de estudios:** Variable que capta el máximo nivel de estudios alcanzado por el cliente y cuenta con 10 categorías.
- **Estado civil.** Variable que capta el estado civil del cliente al momento de la solicitud y cuenta con 7 categorías.
- **Tipo de vivienda.** Variable que capta la propiedad de la vivienda del cliente al momento de la solicitud. Cuenta con 6 categorías.
- **Actividad económica:** A partir del código de actividad económica del CIIU se disminuye la dimensionalidad de esta variable, agrupando el CIIU a dos dígitos con lo cual se tendrían 65 categorías.
- **Teléfono fijo:** Se crea una variable dicotómica para capturar si el cliente reporta tener un teléfono fijo o no. Toma el valor de 0 cuando no lo tiene y el valor de 1 al reportar tenerlo.
- **Teléfono celular:** Se crea una variable dicotómica para capturar si el cliente reporta tener un teléfono celular o no. Toma el valor de 0 cuando no lo tiene y el valor de 1 al reportar tenerlo.
- **Correo electrónico:** Se crea una variable dicotómica para capturar si el cliente reporta tener un correo electrónico o no. Toma el valor de 0 cuando no lo tiene y el valor de 1 al reportar tenerlo.
- **Tipo cliente:** identifica si el cliente es nuevo para la institución o corresponde a una renovación de una obligación ya adquirida.
- **Garantía**: identifica si el préstamo del cliente tiene alguna garantía para la entidad, que funciona como un seguro para la entidad financiera en caso de default del cliente. No corresponde a una garantía real sobre algún bien del cliente.

6.1.3. Análisis Exploratorio de los datos

Se realizó un análisis descriptivo de la información con tablas cruzadas en función de conocer las características de los clientes y posibles relaciones de estas con el comportamiento de pago del cliente (default). En la Tabla *3* se presenta la caracterización de cada perfil de cliente de acuerdo con variables financieras claves del negocio tales como activos, pasivos, ingresos entre otros. Posteriormente, en la Tabla 4 se analiza la participación de clientes y cartera por segmento. A partir de estas dos tablas se describe a continuación el perfil y comportamiento por cada segmento de clientes.

Tabla 3 - Caracterización de los segmentos de los clientes

Características segmentos clientes					
Segmento	**Activos promedio**	**Ingresos promedio**	**Pasivos promedio**	**Empleos promedio**	**Ticket crédito**
Emprendedor en desarrollo	$67	$16	$18	1,4	$9
Emprendedor independiente	$12	$6	$2	1,1	$3

Fuente: elaboración propia. Cifras en millones de pesos

Tabla 4 - Caracterización de la cartera por segmento

Cartera por segmento		
Segmento	**% clientes**	**% cartera colocada**
Emprendedor en desarrollo	28,25%	52,03%
Emprendedor independiente	71,75%	47,97%
Total	**100%**	**100%**

Fuente: Elaboración Propia.

Emprendedor independiente: El 71,75 % de los créditos otorgados en el periodo corresponden a este segmento de unidades productivas con un ingreso promedio inferior a los 7 millones de pesos mensuales y autoempleo. Este segmento representa el 47,97% de la cartera total colocada en personas naturales.

Emprendedor en desarrollo: Este segmento se caracteriza por tener un ingreso promedio de 16 millones, activos por 67 millones y un promedio 1.4 empleos por unidad productiva. Este segmento representa cerca del 28% del total de clientes y el 52,03% de la cartera colocada.

A partir de esta información se concluye que todos los clientes clasificados por la Fundación en los segmentos de Emprendedor independiente y Emprendedor en Desarrollo correspondientes al 92,74 % de clientes de la Fundación son Microempresas, según la clasificación establecida por Ley 905 de 2004 [9], considerando que no tienen más de 10 trabajadores y las ventas anuales o activos totales no superan los 501 salarios mínimos legales vigentes (SMMLV).

Tabla 5 - Top 10 cartera colocada y número de clientes por actividad económica

Actividad económica	% cartera	% clientes
Comercio al por menor, excepto el de vehículos automotores y motocicletas	31%	49%
Comercio al por mayor y en comisión o por contrata	10%	9%
Actividades de servicios de comidas y bebidas	9%	7%
Obras de ingeniería civil	5%	6%
Elaboración de productos alimenticios	5%	5%
Almacenamiento y actividades complementarias al transporte	3%	4%
Actividades especializadas para la construcción de edificios y obras de ingeniería civil	3%	2%
Otras actividades de servicios personales	3%	2%
Construcción de edificios	3%	1%
Transporte terrestre; transporte por tuberías	2%	1%

Fuente: Elaboración Propia.

Al analizar las actividades económicas a las que se dedican los clientes y que se detallan en la Tabla 5, se encuentra que los empresarios dedicados al comercio al por menor representan el 49% de los clientes y el 31% de los recursos colocados, ocupando el primer lugar en estas variables analizadas. En segundo lugar, aparecen las unidades productivas dedicadas al comercio al por mayor, lo cuales representan el 9% de los clientes y el 10% de los recursos colocados. En el top 10 de actividades económicas se destacan actividades

asociadas la construcción, transporte, restaurantes y fabricación de productos alimenticios. En estas actividades se concentra el 73% de los recursos colocados y el 85% de los clientes (Anexo 1).

En la Tabla *6* se detalla la distribución de clientes y cartera por cada uno de los segmentos y ciudades de ubicación de los clientes. El 91 % de los clientes se concentran en Barranquilla, Cartagena y Soledad, representando el 93 % de la cartera, donde Barranquilla tiene el mayor nivel de participación con el 50 % de los recursos colocados. Por segmentos, se encuentra que para el segmento *Emprendedor Independiente* la mayor cantidad de clientes (65%) y recursos (65%) se encuentran en Barranquilla; para el caso del segmento *Emprendedor en Desarrollo* el 44% de los clientes y el 41% de la cartera se ubica en Cartagena.

Tabla 6 - Clientes y cartera por segmentos y ciudades

Ciudad	Independiente		Desarrollo		Total	
	% clientes	% cartera	% clientes	% cartera	% clientes	% cartera
Barranquilla	64%	65%	36%	37%	41%	50%
Cartagena	30%	30%	44%	41%	41%	36%
Soledad	3%	2%	12%	13%	9%	7%
Otros	4%	3%	8%	9%	9%	7%
Total	100%	100%	100%	100%	100%	100%

Fuente: Elaboración Propia.

Se complementó el análisis descriptivo inicial, con información descriptiva de los clientes por género, nivel estudios, tipo de vivienda entre otros. Los clientes son mayoritariamente mujeres con el 63% de participación, donde el 51,4% de ellas tienen estudios superiores, el 67,9% se encuentran casadas o en unión libre y el 57,93% residen en una vivienda propia o familiar. En el caso de los hombres, 37% del total de clientes, 50,7% cuentan con estudios superiores, el 60,8% están casados o en unión libre y el 62,7% viven en una vivienda propia o familiar (Anexo 2 y Anexo 3). El 72 % de los clientes mujeres pertenecen al segmento de emprendedor independiente, frente al 66 % que están en este segmento. En el segmento emprendedor en desarrollo se ubican el 26% de mujeres y 30% de los hombres respectivamente (Anexo 4).

Default

En esta sección se presenta el resultado del análisis exploratorio y descriptivo en función del default, es decir, aquellos clientes que incurren en 30 días o más de mora durante los primeros 12 meses del crédito. De acuerdo con lo observado en la Tabla *7* el 28,78% de la muestra de clientes analizados cayeron en default equivalente al 22,3% del total de la cartera colocada. El segmento con mayor incidencia del default es el *Emprendedor Independiente,* donde en promedio 3 de cada 10 clientes incurren en una mora superior a 30 días, equivalente al 28,1% de los recursos colocados en este segmento. En cuanto al segmento *Emprendedor en Desarrollo* 2 de cada 10 clientes caen en default representando el 16,9% de los recursos colocados en el segmento.

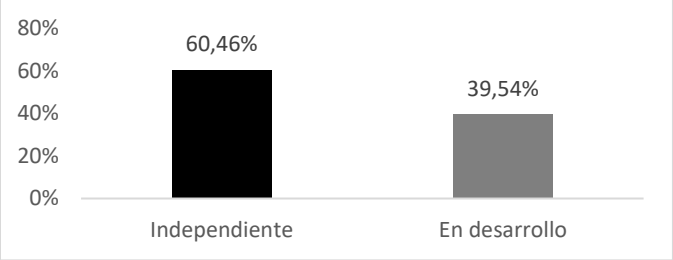
Tabla 7 - Default por segmento - % clientes y cartera

Segmento	% Clientes default del segmento	% cartera default del segmento
Emprendedor en desarrollo	19,94%	16,96%
Emprendedor independiente	32,25%	28,13%
Total	**28,78%**	**22,32%**

Fuente: Elaboración Propia.

Los clientes del segmento *emprendedor independiente* representan el 60,46% del total de recursos colocados en clientes en default, mientras los clientes del segmento *independiente representan* el 39,5% restante como se observa en la Figura *1*.

Figura 1 - Distribución de la cartera en default por cada segmento



Fuente: Elaboración Propia.

Una variable que podría incidir en el nivel de default es la actividad económica en el que se encuentre la unidad productiva receptora del crédito. En la Tabla *8* se observa el default de las 10 actividades económicas de mayor incidencia de default participación por número de clientes. En esta tabla se observa que el 43% de los clientes y el 32% de la cartera en default se encuentran en las actividades relacionadas con el comercio al por menor, en un segundo lugar, con el 11% en clientes y cartera en default los clientes con unidades productivas destinadas a servicios de comidas y bebidas. Se destacan las unidades dedicadas al Comercio al por mayor y en comisión o por contrata, quienes representan el 3% de los clientes, pero el 11% de la cartera total en default. En estas actividades se concentra el 78% y el 85% de los recursos y clientes en default respectivamente.

Tabla 8 - Default por actividad económica según número de clientes y cartera colocada (%)

Actividad económica	% clientes	% cartera
Comercio al por menor, excepto el de vehículos automotores y motocicletas	43%	32%
Actividades de servicios de comidas y bebidas	11%	11%
Elaboración de productos alimenticios	8%	6%
Otras actividades de servicios personales	7%	4%
Confección de prendas de vestir	4%	2%
Comercio al por mayor y en comisión o por contrata	3%	11%
Actividades administrativas y de apoyo de oficina y otras actividades de apoyo a las empresas	3%	2%
Comercio, mantenimiento y reparación de vehículos automotores y motocicletas	2%	3%
Otras actividades profesionales, científicas y técnicas	2%	2%
Actividades especializadas para la construcción de edificios y obras de ingeniería civil	2%	6%

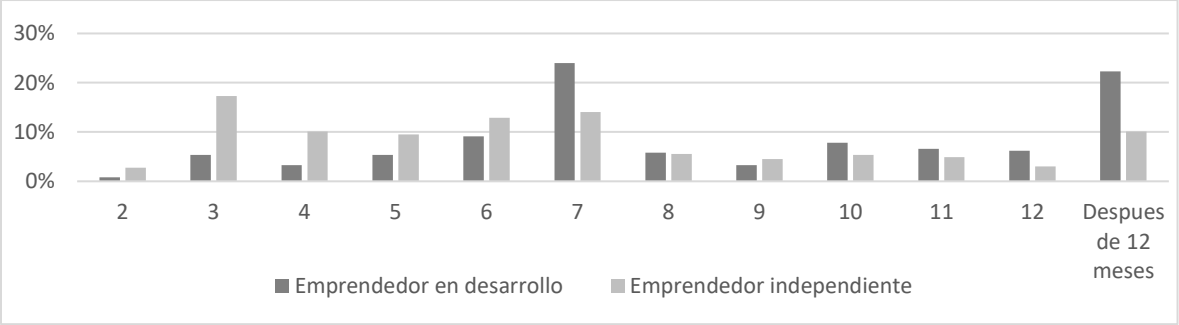
Fuente: Elaboración Propia.

Este comportamiento es muy similar a nivel de segmentos, donde las 2 principales actividades económicas concentran el total de clientes y cartera en default, al observar que los recursos en default de las 3 principales actividades económicas corresponden al 63% del segmento *Independiente* y el 66% de los *Emprendedores en Desarrollo* (Anexo 5 y Anexo 6).

Una variable que permite identificar la calidad del análisis de riesgo de colocación es el tiempo que un cliente "malo" tarda en caer en default, entre más tiempo pase entre la entrega del crédito y la mora, se entiende que obedece a nuevas condiciones o cambios en la situación del empresario en comparación al momento del crédito y entre más cercano se encuentre puede ser un indicativo de problemas en la evaluación crediticia del cliente. El análisis evidencia que los clientes caen en default entre el sexto (6) y séptimo (7) mes de vigencia del crédito, pero presenta un comportamiento diferencial por segmento.

En los 4 primeros meses del crédito, en el segmento de emprendedor en desarrollo cayeron en default el 9,5% de los clientes y el 30,1% del segmento independiente. Ampliando el periodo a 7 meses se observa que en el segmento de emprendedor en desarrollo cayeron en default el 47,9% del segmento y el 66,5% del segmento independiente. El 22,3% del segmento emprendedor en desarrollo incurren en una mora de más de 30 días luego de 12 meses de vigencia del crédito en comparación al 10% del segmento independiente. Como se observa, el comportamiento de pago de este último segmento difiere tiene un peor desempeño al otro analizado y la cartera se deteriora más rápido como se muestra en la Figura *2*.

Figura 2 - Mes de ocurrencia default por segmentos

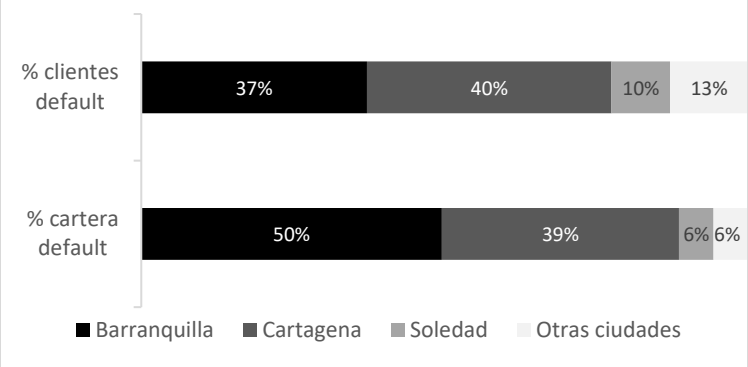


Fuente: Elaboración Propia.

Según esta información, para el segmento de emprendedor independiente se tiene más dificultad para evaluar el riesgo, pues tras 4 meses la tasa de default es superior a la del otro segmento.

Las condiciones económicas de las diferentes ciudades también pueden llegar a incidir en la calidad de la cartera. En la Figura *3* se observa que en la ciudad de Barranquilla se ubican el 37% de las obligaciones correspondiente al 50% del total de recursos mientras en la ciudad de Cartagena tiene una participación mayor en número de obligaciones (40%) pero esto representa el 39% del total de la cartera. Esto se explica por la diferencia en el ticket promedio, mientras en Barranquilla este asciende a $12.3 millones en Cartagena se ubica en $7.9 millones.

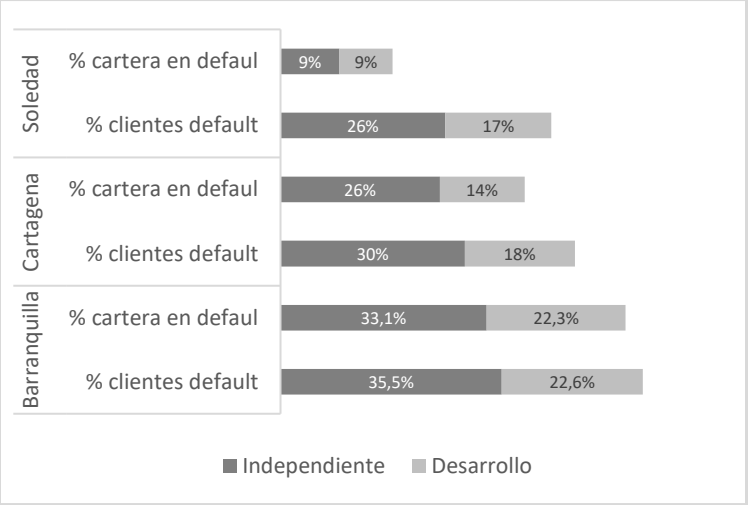
Figura 3 – Default por clientes y monto cartera por ciudades (%)



Fuente: Elaboración Propia.

En materia de segmentos por ciudades, la Figura *4* muestra que en Barranquilla el 35% de los clientes del segmento independiente se encuentran en default, en comparación al 22% del otro segmento. En el caso de Cartagena, el segmento independiente también presenta una mayor incidencia (30%) de default en comparación a lo observado en el otro segmento.

Figura 4 – Default por clientes y monto cartera por ciudades y segmentos (%)



Fuente: Elaboración Propia.

Finalmente, la Tabla *9* presenta un resumen del default presentado de acuerdo con las principales características sociodemográficas de este tipo de clientes. No hay diferencias entre la proporción de mujeres que reciben un crédito que caiga en default en comparación con hombres en la misma condición, pues presentan tasas de incidencia similares (26,7% mujeres contra 27,7% hombres), pero equivalen a proporciones diferentes del monto en default (16% mujeres contra 5% hombres). Por nivel de escolaridad, se encuentra que, a medida que aumenta el nivel educativo, aumenta la proporción de clientes en morosidad, donde más del 30% de los con estudios superiores a nivel tecnológico, universitario o postgrado están en default frente a los clientes con menor nivel de estudios. Otra característica que llama la atención, son los clientes con vivienda propia, donde el 24% de estos clientes caen en default, el valor más bajo para los diferentes tipos de tenencia de la vivienda, posiblemente asociado a una mayor estabilidad económica y flujo de caja disponible.

Tabla 9 – Default por características sociodemográficas de los clientes (%)

Variable	Categoría	%Clientes default	% Cartera en default
Genero	Femenino	27%	16%
	Masculino	28%	5%
Nivel de estudios	Analfabetismo	0%	0%
	No escolarizado	25%	0%
	Primaria	21%	1%
	Secundaria	28%	11%
	Técnica	27%	6%
	Tecnológica	34%	1%
	Universitaria	30%	3%
	Especialización	30%	0%
	Magister	50%	0%
Estado Civil	Divorciado - Separado	30%	25%
	Otro	31%	8%
	Soltero	35%	31%
	Viudo	30%	24%
	Unión libre	29%	25%
	Casado	21%	17%
Tipo de vivienda	Otra	38%	37%
	Sin vivienda	28%	17%
	Arriendo	26%	21%
	Familiar	33%	23%
	Inmueble con Hipoteca	0%	0%
	Propia	24%	22%
Personas a cargo	0	27%	20%
	1	34%	28%
	2	25%	15%
	3	16%	12%
	4	13%	23%
	5	0%	0%
	6	0%	0%
	9	0%	0%
Estrato	1	24%	19%
	2	29%	22%
	3	30%	26%
	4	20%	25%
	5	27%	29%
	6	0%	0%

Fuente: Elaboración Propia

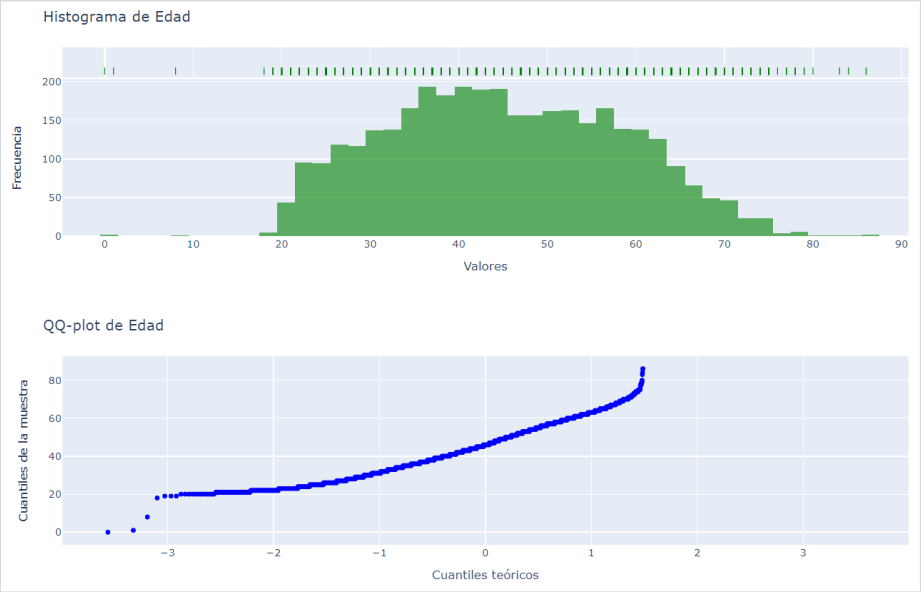
6.2. Tratamiento y selección de variables

6.2.1. Tratamiento de datos faltantes.

Los datos faltantes pueden corresponder a la ausencia de información o a dificultades en la captura del dato; dicha situación puede convertirse en un inconveniente, sobre todo en variables con una tasa significativa de valores nulos (p.e. *"Estrato"* con el 25.8% y *"Nivel_Estudios"* con el 12.5%).

Ante esta situación y la imposibilidad de ignorar esos registros, se definieron dos técnicas de imputación de datos: la estrategia de valor constante para variables categóricas y la mayoría de las variables numéricas, con excepción de la variable *"Edad"* que por su porcentaje destacable de valores nulos (6.8%) se emplea la estrategia basada en la media de los datos [18].

Figura 5 - Histograma y un gráfico QQ-plot de variable "Edad" (Distribución normal).



Fuente: Elaboración Propia.

La imputación por la media (o promedio) es una técnica que consiste en reemplazar los valores nulos de una variable numérica por el valor promedio de los valores no nulos de la misma variable en todo el conjunto de datos. Por otro lado, la imputación por la mediana es similar, pero en lugar de utilizar el valor promedio, se utiliza la mediana de los valores no nulos de la variable.

La elección la media, entre estas dos técnicas de imputación, se debe al resultado del análisis de la distribución de los datos en la variable numérica *"Edad"*. Según la Figura 5, los datos están distribuidos en el histograma de forma simétrica; además se observa en el gráfico QQ-plot, que compara los cuantiles de la muestra con los cuantiles teóricos de una distribución normal, que los puntos se ajustan bien a una línea casi recta, lo que indica que la variable sigue una distribución normal.

En lo referente a las variables categóricas, se realizó un reemplazo, creando la categoría *"Sin información"*

para la mayoría y reusando la categoría *"DESCONOCIDA"* para el caso de la variable *"Profesion"*. Esto para no perder información, sumada a que ciertos clientes solicitantes de créditos de bajo monto pueden ser renuentes a compartir ciertos datos y recoger ese comportamiento puede resultar relevante. Para las demás variables numéricas con un número marginal de valores nulos, se hacen los reemplazos con el valor cero (0), que tiene dentro del caso de uso, un significado similar de información no disponible.

6.2.2. Tratamiento a valores atípicos en variables numéricas.

La existencia de valores atípicos en el conjunto de datos puede tener un impacto negativo en la precisión del modelo. Los valores atípicos pueden influir en los parámetros del modelo y en los umbrales de decisión, lo que puede afectar negativamente la capacidad del modelo para generalizar a nuevos datos [19].

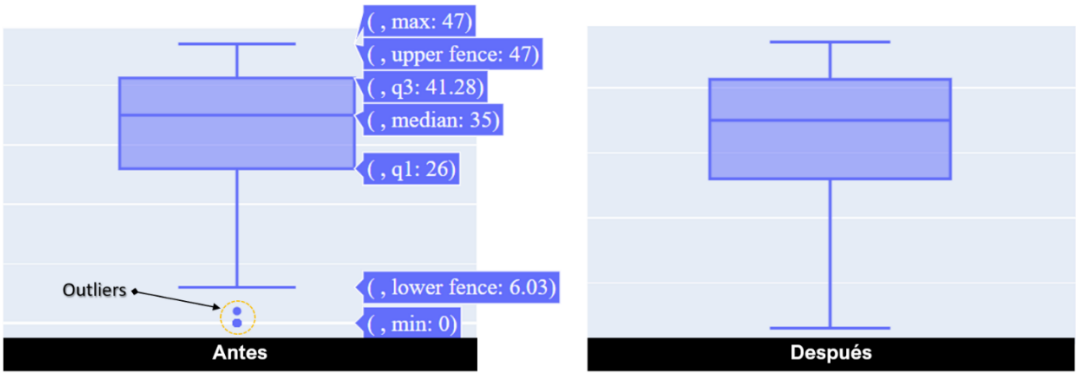
Figura 6 - Detección automática (AutoViz) de valores atípicos en variables numéricas.

	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
Monto_credito	303	int64	0	0.000000	8.007400	0	
Tasa	50	float64	0	0.000000	1.321353	0	skewed: cap or drop outliers
Nro_Cuotas	42	int64	0	0.000000	1.109937	0	
Empleos_Permanentes	39	int64	0	0.000000	1.030655	0	
Mes_default	21	int64	0	0.000000	0.554968	0	
Cobertura_garantia	8	float64	0	0.000000	0.211416	0	

Fuente: Elaboración Propia.

Hay varias técnicas para identificar y eliminar valores atípicos, que incluyen métodos gráficos, estadísticos y modelos de detección de anomalías. Usando una combinación de herramientas gráficas como *AutoViz* y *Plotly* se evidencia que tan sólo la variable *"Tasa"* presenta una fracción mínima de valores atípicos como se muestran en la Figura *6* y la Figura *7*.

Figura 7 - Antes y después de aplicar la técnica IQR a valores atípicos en la variable Tasa de crédito.



Fuente: Elaboración Propia.

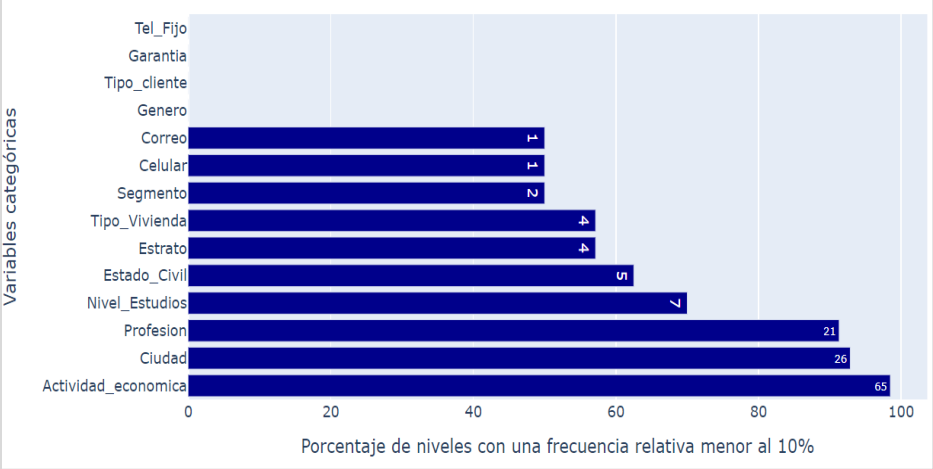
Para remover los valores atípicos se emplea la técnica IQR que utiliza una medida estadística simple, el rango intercuartílico. El rango intercuartílico se define como la diferencia entre el tercer cuartil (*Q3=75%*) y el primer cuartil (*Q1=25%*) de un conjunto de datos. Los valores atípicos se identifican y remueven utilizando un rango de valores que se encuentra fuera de este rango.

Específicamente, cualquier valor que se encuentre por debajo del primer cuartil menos 1,5 veces el rango intercuartílico o por encima del tercer cuartil más 1,5 veces el rango intercuartílico se considera un valor atípico y se elimina del conjunto de datos [20].

6.2.3. Tratamiento a niveles raros en variables categóricas.

Una casuística para revisar en los conjuntos de datos y relacionada con algunas variables categóricas tiene que ver con la cantidad elevada de niveles, es decir, categorías con alta cardinalidad. Si una variable categórica con niveles de alta cardinalidad se codifica a valores numéricos, entonces la matriz resultante es una matriz dispersa. Esto no solo hace que el experimento sea lento debido al tamaño del conjunto de datos, sino que también introduce ruido en el experimento.

Figura 8 - Frecuencia de niveles raros por variable categórica.



Fuente: Elaboración Propia.

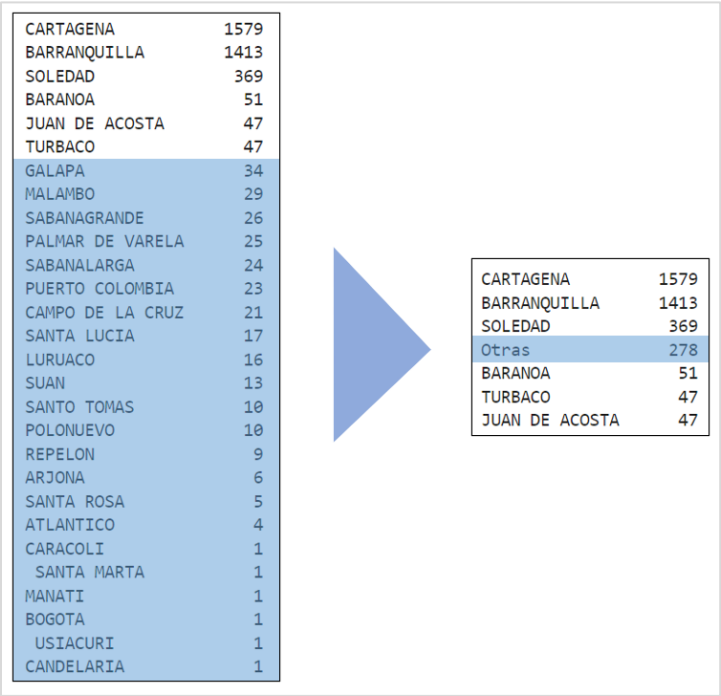
La matriz dispersa se puede evitar combinando los "niveles raros" entre aquellas categorías cuya frecuencia relativa sea marginal en relación con las demás. En este caso, se agrupan todos los niveles en aquellas variables categóricas cuyo porcentaje relativo sea menor al *10%* del total de las observaciones en una nueva categoría a la que llamaremos *"Otras"*, sólo en aquellas variables categóricas que tengan más de 20 "niveles raros" y correspondan a más del *70%* del total de valores únicos.

De acuerdo con la Figura *8*, sólo las siguientes variables categóricas cumplen con el criterio para identificación de "niveles raros":

- **Ciudad:** tiene 26 niveles (92.86%)
- **Actividad económica*:*** tiene 65 niveles (98.48%)
- **Profesión:** tiene 21 niveles (91.3%)

En la Figura *9* se presenta a manera de ejemplo la combinación de valores para niveles raros en la variable categórica ciudad, para la cual, después de aplicar la transformación se agrupan las categorías con menor frecuencia en la categoría "otras".

Figura 9 - Ejemplo de aplicación de la combinación de niveles raros.



Fuente: Elaboración Propia.

6.2.4. Transformación de variables

A partir de características nominales existentes en el conjunto de datos como los activos, pasivos, ingresos, gastos, personas a cargo y valor de la cuota, que por su naturaleza tienen datos muy dispersos y no son candidatas para alimentar al modelo, se harán cálculos de indicadores financieros para medir la capacidad de asumir una responsabilidad crediticia, su solvencia económica y posición financiera.

Para evaluar que una persona tiene capacidad para asumir una nueva deuda se calculará el *Nivel de Endeudamiento*, indicador que mide la proporción de los pasivos totales de una persona en relación con sus activos totales.

- *Nivel de endeudamiento* = Pasivos totales / Activos totales

Para evaluar la solvencia económica de una persona se empleará el indicador financiero *Capacidad de endeudamiento*, que permite establecer la cantidad o monto máximo por la que una persona puede endeudarse sin arriesgar su solvencia económica. Este indicador financiero sugiere entonces que, con los ingresos totales que recibe una persona periódicamente, debe ser capaz de cubrir tanto los diferentes gastos obligatorios como el pago de cuota mensual del nuevo crédito por asumir, sin perjudicar la solvencia financiera.

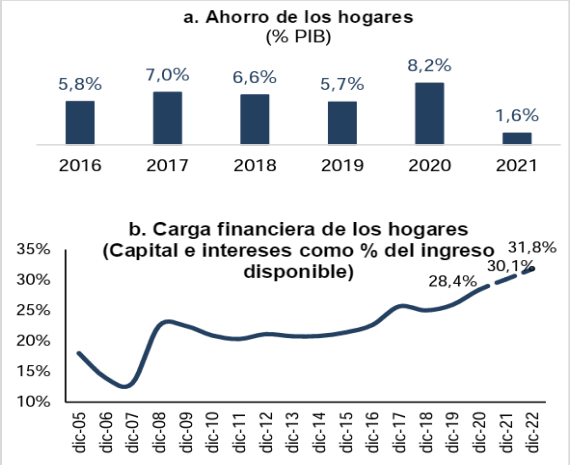
Dentro del cálculo de la *Capacidad de Endeudamiento* se tiene en cuenta un límite que va entre el 30% y el 40% de los ingresos netos mensuales, correspondiente al índice de carga financiera sugerido por Asobancaria en su informe de 2022 [21]. El ingreso neto es aquel monto de dinero que resulta de la diferencia entre los ingresos y gastos totales. No es recomendable exceder el límite recomendado porque, de hacerlo, se pondría en riesgo la estabilidad y salud financiera de la persona, y por ende aumentaría el riesgo de default.

- *Ingresos netos* = Ingresos mensuales – Gastos mensuales
- *Capacidad de endeudamiento* = Ingresos netos x 0,3
- *Capacidad de pago* = (Capacidad de endeudamiento - Valor cuota crédito) / Valor cuota crédito

Niveles de capacidad de pago:

- Ninguna: Capacidad de endeudamiento < Valor cuota crédito
- Débil: Capacidad de endeudamiento => Valor cuota crédito < Valor cuota crédito * 1,10
- Fuerte: Capacidad de endeudamiento => Valor cuota crédito * 1,10

Figura 10 - Ahorro y carga financiera de los hogares colombianos.



Fuente: Asobancaria, 2022.

Para medir la posición financiera de la persona se empleará el indicador de clase social conforme a lo definido por el DANE en su informe *"Caracterización pobreza monetaria y resultados clases sociales"* de 2020 [22].

A partir de la actualización de los valores de corte de la *metodología de López Calva y Ortiz Juárez (2011)* [23], en el caso de Colombia se obtienen los siguientes umbrales:

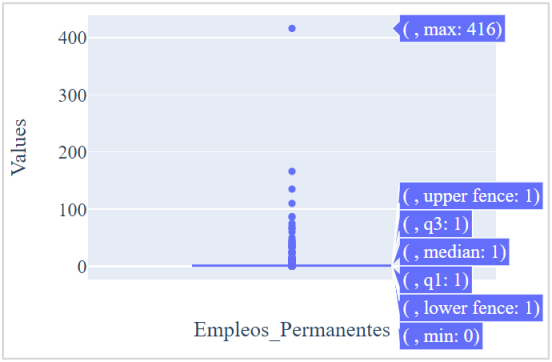
- Los *pobres* se definen como aquellos con un ingreso per cápita inferior a la línea de pobreza monetaria (25 líneas de pobreza diferenciadas)
- Los *vulnerables* corresponden con ingreso per cápita entre la línea de pobreza y $653.781 mensuales.

- La *clase media* está compuesta por aquéllos a quienes corresponde como ingreso per cápita al interior del hogar entre $653.781 y $3.520.360 al mes.
- La *clase alta* está conformada por personas cuyo ingreso per cápita al interior del hogar corresponde con más de $3.520.360 mensuales.

6.2.5. Variables irrelevantes.

Del conjunto de datos se eliminan las variables originales utilizadas para obtener las variables transformadas y agregadas al conjunto de datos: Capacidad de pago, el Nivel de endeudamiento y la Clase social.

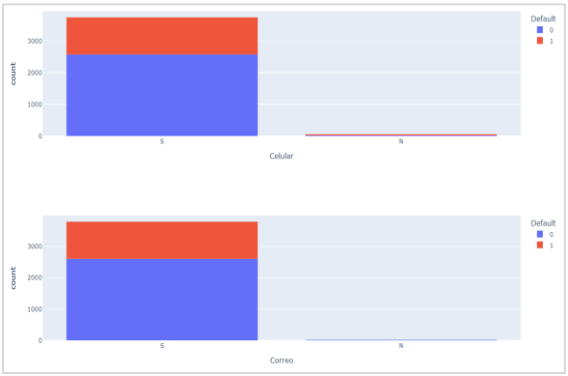
Figura 11 - Gráfico de caja y bigotes de variable numérica Empleos permanentes.



Fuente: Elaboración propia.

También se elimina del conjunto de datos la variable *"Empleos_Permanentes"*, por contener un número elevando de valores atípicos que de tratarse podría ocasionar un sobreajuste al modelo, además que evidencia una alta concentración de datos alrededor de la media, como se puede observar en el análisis de cuartiles de la Figura 11.

Figura 12 - Gráfico bivariado de frecuencias Tiene celular y Tiene correo vs. Default.



Fuente: Elaboración propia.

Por último, también se excluye del conjunto de datos las variables categóricas *"Celular"* y *"Correo"* porque la distribución de frecuencias es muy amplia y marcada entre los niveles de sí-contiene" y no-contiene, como se muestra en la Figura 12.

6.2.6. Codificación de valores ordinales y nominales.

Las variables categóricas en el conjunto de datos contienen valores de etiqueta (ordinales o nominales) en lugar de números continuos. La mayoría de los algoritmos de aprendizaje automático no pueden tratar directamente con variables categóricas y deben transformarse en valores numéricos antes de entrenar un modelo. El tipo más común de codificación categórica es la codificación One-Hot, donde cada nivel categórico se convierte en una característica separada en el conjunto de datos que contiene valores binarios (1 o 0).

Esto es ideal para funciones que tienen datos categóricos nominales, es decir, que los datos no se pueden ordenar. En otros escenarios diferentes, se deben utilizar otros métodos de codificación. Por ejemplo, cuando los datos son ordinales, es decir, los datos tienen niveles intrínsecos, se debe usar la codificación ordinal.

Cuando las variables categóricas en el conjunto de datos contienen niveles con un orden natural intrínseco como Bajo, Medio y Alto, estas deben codificarse de manera diferente a las variables categóricas nominales (donde no hay un orden intrínseco para, por ejemplo, Hombre o Mujer), de tal manera que se aplique un orden creciente de menor a mayor a los niveles, según el sentido al tratarse de valores ordinales.

Teniendo en cuenta lo anterior y para la adecuada codificación de valores categóricos se emplea el método *"OneHotEncoding"* y sólo para valores ordinales se emplea el método *"OrdinalEncoding"* con el siguiente agrupamiento de niveles en orden ascendente:

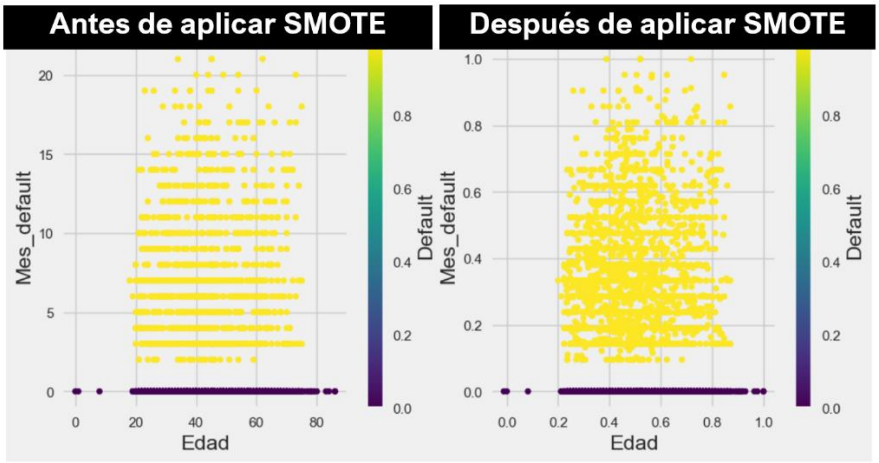
- **Capacidad pago:** 'Ninguna', 'Débil', 'Fuerte'
- **Nivel endeudamiento:** 'Bajo', 'Medio', 'Alto'
- **Clase social:** 'Pobre', 'Vulnerable', 'Media', 'Alta'

6.2.7. Balanceo de la variable objetivo.

En problemas de clasificación ocurre, en algunas ocasiones, que las clases, sean binarias o múltiples, no están representadas proporcionalmente y por ende una clase predomina sobre las otras o bien una clase tiene poca representación dentro del conjunto de datos. Si se intenta construir un clasificador sin tener en cuenta este desbalance, entonces el clasificador resultante podría predecir la clase mayoritaria, dado que esa acción minimiza el error cometido al acertar la mayoría de las veces, y clasificará erróneamente de forma sistemática los elementos de la clase minoritaria.

Para el caso de estudio, la clase es binaria (toma valores de 0 o 1) y está representada en la variable objetivo *Default*, que según las observaciones evidencia un desequilibrio donde la clase minoritaria uno (1: Susceptible a caer en mora/default) es precisamente la de mayor interés.

Figura 13 - Gráfico con el antes y después de balancear usando SMOTE.



Fuente: Elaboración propia.

Para resolver esto se aplicó la técnica SMOTE, acrónimo de Synthetic Minority Over-sampling Technique, basada en generar de forma sintética nuevos elementos de la clase minoritaria usando como referencia los elementos de dicha clase ya presentes en el conjunto de datos. Para ello, a partir de un elemento de la clase minoritaria elegido al azar, se escoge un cierto número de vecinos más cercanos y se genera un nuevo elemento combinándolos linealmente de forma ponderada introduciendo un factor aleatorio para generar elementos parecidos. En la Figura 13 se muestra el resultado de aplicar SMOTE al conjunto de datos.

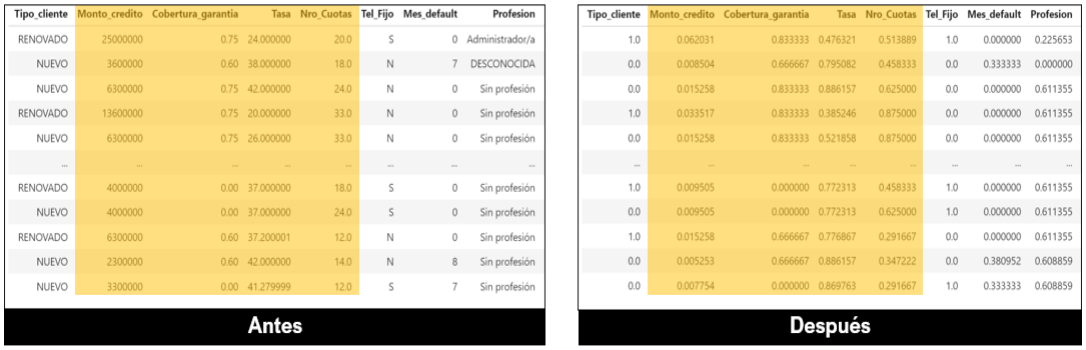
6.2.8. Normalización y re-escalamiento de valores numéricos.

La normalización es una técnica que se aplica a menudo como parte de la preparación de datos para el aprendizaje automático. El objetivo de la normalización es cambiar la escala de los valores de las variables numéricas dentro del conjunto de datos sin distorsionar las diferencias en los rangos de valores ni perder información.

Hay varios métodos disponibles para la normalización, para el caso de estudio se aplicó *Min-Max Scaler* basados en dos siguientes criterios:

- Escala de los datos: Min-Max es un método de normalización adecuado cuando se aplica a diferentes escalas de los datos. Puede ser útil para normalizar datos continuos como el Monto del crédito, en cuyo caso los datos tienen valores muy dispares en términos de magnitud.
- Distribución de los datos: El método de normalización debe manejar diferentes tipos de distribución de los datos. Para el caso de estudio, no todas las variables numéricas contienen datos que siguen una distribución normal, por eso el método Min-Max Scaler puede ser una mejor elección a otros como el método Z-Score o Abs-Max.

Figura 14 - Antes y después de normalizar usando Min-Max Scaler.



Antes

Tipo_cliente	Monto_credito	Cobertura_garantia	Tasa	Nro_Cuotas	Tel_Fijo	Mes_default	Profesion
RENOVADO	25000000	0.75	24.000000	20.0	S	0	Administrador/a
NUEVO	3600000	0.60	38.000000	18.0	N	7	DESCONOCIDA
NUEVO	6300000	0.75	42.000000	24.0	N	0	Sin profesión
RENOVADO	13600000	0.75	20.000000	33.0	N	0	Sin profesión
NUEVO	6300000	0.75	26.000000	33.0	N	0	Sin profesión
...	...	...	...	...	...	...	...
RENOVADO	4000000	0.00	37.000000	18.0	S	0	Sin profesión
NUEVO	4000000	0.00	37.000000	24.0	S	0	Sin profesión
RENOVADO	6300000	0.60	37.200001	12.0	N	0	Sin profesión
NUEVO	2300000	0.60	42.000000	14.0	N	8	Sin profesión
NUEVO	3300000	0.00	41.279999	12.0	S	7	Sin profesión

Después

Tipo_cliente	Monto_credito	Cobertura_garantia	Tasa	Nro_Cuotas	Tel_Fijo	Mes_default	Profesion
1.0	0.062031	0.833333	0.476321	0.513889	1.0	0.000000	0.225653
0.0	0.008504	0.666667	0.795082	0.458333	0.0	0.333333	0.000000
0.0	0.015258	0.833333	0.886157	0.625000	0.0	0.000000	0.611355
1.0	0.033517	0.833333	0.385246	0.875000	0.0	0.000000	0.611355
0.0	0.015258	0.833333	0.521858	0.875000	0.0	0.000000	0.611355
...	...	...	...	...	...	...	...
1.0	0.009505	0.000000	0.772313	0.458333	1.0	0.000000	0.611355
0.0	0.009505	0.000000	0.772313	0.625000	1.0	0.000000	0.611355
1.0	0.015258	0.666667	0.776867	0.291667	0.0	0.000000	0.611355
0.0	0.005253	0.666667	0.886157	0.347222	0.0	0.380952	0.608859
0.0	0.007754	0.000000	0.869763	0.291667	1.0	0.333333	0.608859

Fuente: Elaboración propia.

6.2.9. Selección de variables y reducción de dimensionalidad

La reducción de la dimensionalidad se refiere a técnicas y métodos utilizados para disminuir el número de variables o características en un conjunto de datos, mientras se trata de preservar la mayor cantidad posible de información útil [17]. El objetivo de la reducción de la dimensionalidad es simplificar la estructura de datos y reducir la complejidad computacional en el análisis posterior. Esto es de particular interés para la entidad, ya que no tiene un proceso digital automatizado para evaluar el crédito tras consultar en las centrales de riesgo. Un modelo simple podría facilitar su adopción, así como la construcción de un proceso de negocio y tecnológico simple. Existen varias técnicas y métodos para reducir la dimensionalidad, incluyendo la selección de características, la extracción de características y la proyección de datos.

Para esto se compararán diversas técnicas y sus resultados bajo una regresión logística econométrica para identificar su poder estadístico bajo un enfoque tradicional. Las 4 técnicas seleccionadas corresponden a métodos usados en la industria para desarrollar score de créditos, describiéndose a continuación, *i) Regresión Lasso,* técnica de regularización enfocada en encontrar un equilibrio entre la simplicidad y precisión del modelo, para esto añade una penalización al modelo de regresión lineal obligando a algunos de los coeficientes a ser cero, permitiéndole descartar variables redundantes o irrelevantes, *ii) Information value (IV),* a partir de un valor numérico que cuantifica el poder predictivo de una variable continua independiente x en la captura de la variable dependiente binaria, a partir del análisis de cada variable independiente individual, iii) *Combinación information value y regresión logística,* en enfoques econométricos se utiliza el IV como un método previo para reducir la dimensionalidad y seleccionar variables predictoras de una regresión logística binaria, donde posteriormente se descartan las variables no significativas [26], *iv) Extra Tree Classifier (ET),* método basado en el árbol de decisiones el cual reduce el espacio de características al seleccionar el subconjunto de características óptimo, a partir de una puntuación otorgada a cada variable. Por lo tanto, mejora la precisión de la predicción y reduce la complejidad del modelo [27].

Tabla 10 - Resultado selección de variables por técnicas utilizadas

Variables independientes	Lasso	Information Value (IV)	IV - Regresión Logit	Extra Tree Classifier	No veces seleccionada
Genero	Si	No	No	No	1
Edad	Si	Si	No	Si	3
Estrato	Si	Si	Si	Si	4
Ciudad	Si	Si	Si	Si	4
Nivel Estudios	Si	No	No	Si	2
Estado Civil	Si	Si	Si	Si	4
Tipo Vivienda	Si	Si	Si	Si	4
Actividad económica	Si	Si	No	Si	3
Tipo cliente	Si	Si	Si	Si	4
Monto crédito	Si	Si	Si	Si	4
Segmento	Si	Si	Si	No	3
Garantía	Si	Si	Si	Si	4
Cobertura garantía	Si	Si	Si	Si	4
Tasa	Si	Si	Si	SI	4
Nro. Cuotas	Si	Si	No	Si	3
Tel Fijo	Si	No	No	No	1
Profesión	Si	Si	Si	No	3
Nivel endeudamiento	Si	No	No	No	1
Capacidad pago	Si	Si	No	No	2
Clase social	Si	No	No	No	1
Total	20	15	11	13	

Fuente: Elaboración Propia

Los resultados de la selección se muestran en la Tabla *10*, donde se observa que bajo la regresión Lasso no se descartan variables, bajo la técnica del *Information Value* (IV) se descartan 5 variables, ET descarta 7 variables y bajo el método combinado del *Information Value* (IV) y regresión logística se descartan 9 variables. Para comparar los resultados se realiza una regresión logística con las variables seleccionadas para cada técnica, con excepción de la regresión Lasso, y se evalúan sus métricas de desempeño, lo cual se muestra en la Tabla *11*, evidenciando que las variables seleccionadas bajo el método combinado del *Information Value* (IV) y Regresión Logística cuenta con el mejor desempeño, permitiendo seleccionar así las once (11) variables con las cuales se realizará la selección del modelo bajo machine learning.

Tabla 11 - Métricas Modelo Logit por técnica de selección utilizadas

Métrica	Information Value (IV)	IV - Regresión Logit	Extra Tree Classifier
Precision	0,4521	0,6994	0,4437
Recall	0,5739	0,7154	0,5478
F1 Score	0,5057	0,7073	0,4903
Exactitud	0,6592	0,7040	0,6539
AUC	0,6352	0,7040	0,6792

Fuente: Elaboración Propia

De las variables seleccionadas, cuatro (4) corresponden a características socioeconómicas de tipo personal de los clientes como el *"estrato civil", "Tipo de vivienda", "Estado civil" y "Profesión";* cuatro (4) corresponden a información del crédito: *"Monto crédito", "Garantía", "Cobertura Garantía", y "Tasa",* mientras las tres (3) restantes*: "Ciudad", "Segmento" y "Tipo de cliente",* a otro tipo de condiciones. Las

variables seleccionadas resultado pueden estar explicadas por las características de la clase y segmento de los clientes objeto de análisis, empresarios que actúan como persona natural ubicados en los segmentos de *emprendedor independiente* y *emprendedor en desarrollo* en donde el nivel de desarrollo productivo promedio es bajo, dado los niveles de ventas, activos y número de empleados registrados, lo cual lleva a que estas variables no permitan diferenciar los clientes *"buenos"* de los *"malos"* y sean variables excluidas del modelo, a pesar de la transformación realizada.

6.3. Evaluación de modelos

En la industria se han usado de forma frecuente las técnicas de análisis discriminante lineal, regresiones logísticas, modelos de regresión logística, árbol de decisiones entre otros [25]. Para seleccionar el modelo de Machine Learning que mejor se adapte al problema y a los datos disponibles, se evalúa el rendimiento de las distintas técnicas a través de una validación cruzada. Se incluyen las siguientes métricas de calidad:

- Accuracy (Exactitud): Es la métrica que mide la proporción de predicciones correctas realizadas por un modelo en relación con el total de predicciones. Se calcula dividiendo el número de predicciones correctas entre el número total de predicciones.
- AUC (Area Under the Curve): Es una métrica utilizada para evaluar la capacidad de discriminación de un modelo de clasificación. Representa el área bajo la curva ROC (Receiver Operating Characteristic) y cuantifica la capacidad del modelo para distinguir entre las clases positiva y negativa. Valores más cercanos a 1 indican un mejor rendimiento del modelo.
- Recall (Recuperación): También conocido como sensibilidad o tasa de verdaderos positivos, mide la proporción de casos positivos que el modelo es capaz de identificar correctamente. Se calcula dividiendo los verdaderos positivos entre la suma de los verdaderos positivos y los falsos negativos.
- Prec (Precisión): Es la proporción de casos positivos predichos correctamente en relación con el total de casos clasificados como positivos por el modelo. Se calcula dividiendo los verdaderos positivos entre la suma de los verdaderos positivos y los falsos positivos.
- F1 (F1-score): Es una métrica que combina la precisión y el recall en un solo valor. Es útil cuando se desea encontrar un equilibrio entre ambas métricas. Se calcula como la media armónica entre la precisión y el recall.
- Kappa (Cohen's Kappa): Es una métrica que mide la concordancia entre las predicciones de un modelo y las etiquetas reales corrigiendo la concordancia aleatoria. Proporciona una medida de la precisión del modelo más allá de la precisión que se esperaría por casualidad.
- MCC (Matthews Correlation Coefficient): Es una métrica que mide la calidad de las clasificaciones binarias, teniendo en cuenta tanto los verdaderos positivos como los verdaderos negativos, así como los falsos positivos y falsos negativos. Proporciona una medida equilibrada incluso cuando las clases están desequilibradas.
- Matriz de confusión: herramienta que muestra el desempeño de un algoritmo de clasificación, describiendo cómo se distribuyen los valores reales y las predicciones del modelo, permitiendo

analizar su desempeño y evidenciar que tipos de aciertos y errores está teniendo el modelo analizado.

En este proyecto, se utilizó la biblioteca PyCaret para el entrenamiento y validación cruzada de los modelos de aprendizaje automático. PyCaret nos permitió realizar una validación cruzada con kfold = 10, lo que implica dividir nuestros datos en 10 conjuntos o "pliegues" de igual tamaño. En cada iteración del proceso de validación cruzada, se selecciona un pliegue diferente como conjunto de prueba y los otros nueve pliegues se utilizan como conjunto de entrenamiento. Esto se repite 10 veces para asegurarnos de que cada pliegue se haya utilizado una vez como conjunto de prueba. Al final de cada iteración, se calculan las métricas de evaluación del modelo, como la precisión o el área bajo la curva (AUC), y se obtiene un promedio de estas métricas para evaluar el rendimiento general del modelo. La validación cruzada con kfold = 10 nos permite obtener una estimación más precisa del rendimiento del modelo y reducir el sesgo introducido por una única división de los datos en conjuntos de entrenamiento y prueba.

El resultado es una cuadrícula de puntuación con las principales métricas de evaluación, en este caso ordenadas por AUC de mayor a menor, como se muestra en la Tabla *12*.

Se observa que, para los datos disponibles, tres (3) de las cinco (5) técnicas con mejor evaluación (*Light Gradient Boosting Machine, Extreme Gradient Boosting, Gradient Boosting Classifier*) corresponden a métodos de conjunto o Boosting, en donde en un proceso iterativo se ajusta un modelo inicial y se va mejorando de forma secuencial, intentando compensar las debilidades del anterior. Son usados de forma frecuente para abordar problemas de riesgo de crédito [29]. En cuanto a las otras dos técnicas en el top 5, por un lado se encuentra el algoritmo *Extra Trees Classifier*, el cual ha demostrado rendimientos superiores en la evaluación de riesgo de crédito frente a otras técnicas como *Random Forest, Extreme Gradient Boosting, Naive Bayes y KNearest Neighbor (KNN)* [26],y consiste en generar una serie de árboles de decisión aleatorios en varias submuestras del conjunto de datos, entrenados de manera individual, se fusionan bajo un proceso estándar buscando reducir la varianza y controlar el sobreajuste. Finalmente, se tiene el *Random Forest*, un enfoque primario de Machine Learning basado en reglas, frecuentemente usado para abordar problemas de crédito [31], y se basa en la construcción de un sinfín de diferentes árboles de decisión, creando un “bosque” que es luego agregado y en el cual se llega a un único resultado.

Tabla 12 – Cuadrícula de puntuaciones medias validadas de forma cruzada para selección del modelo.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Light Gradient Boosting Machine	0.754	0.844	0.740	0.762	0.750	0.509	0.509
Random Forest Classifier	0.758	0.842	0.754	0.760	0.757	0.516	0.517
Extreme Gradient Boosting	0.757	0.839	0.749	0.761	0.755	0.515	0.516
Extra Trees Classifier	0.748	0.827	0.760	0.743	0.751	0.496	0.497
Gradient Boosting Classifier	0.745	0.827	0.741	0.747	0.743	0.490	0.490
Ada Boost Classifier	0.725	0.816	0.737	0.721	0.728	0.451	0.451
Logistic Regression	0.703	0.775	0.685	0.711	0.697	0.407	0.409
Linear Discriminant Analysis	0.699	0.774	0.691	0.701	0.695	0.398	0.399
Quadratic Discriminant	0.661	0.756	0.835	0.620	0.711	0.322	0.344

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Analysis							
Naive Bayes	0.676	0.751	0.814	0.638	0.715	0.352	0.367
K Neighbors Classifier	0.650	0.705	0.701	0.637	0.667	0.301	0.303
Decision Tree Classifier	0.686	0.687	0.696	0.682	0.689	0.373	0.373
Dummy Classifier	0.499	0.500	0.100	0.049	0.066	0.000	0.000
SVM - Linear Kernel	0.578	0.000	0.668	0.642	0.548	0.156	0.240
Ridge Classifier	0.699	0.000	0.691	0.702	0.696	0.398	0.399

Fuente: Elaboración Propia

6.4. Entrenamiento del modelo

Teniendo en cuenta que no existen diferencias significativas en las métricas de mayor AUC y precisión para las 5 técnicas con mejor desempeño se decide entrenar y evalúa el rendimiento de un estimador determinado mediante validación cruzada. El resultado es una cuadrícula de puntuaciones con la secuencia de diez (10) iteraciones de validación cruzada y la media calculada. En la Tabla *13*, se muestra un resumen de las medias de estas iteraciones para las 5 técnicas anteriormente mencionadas.

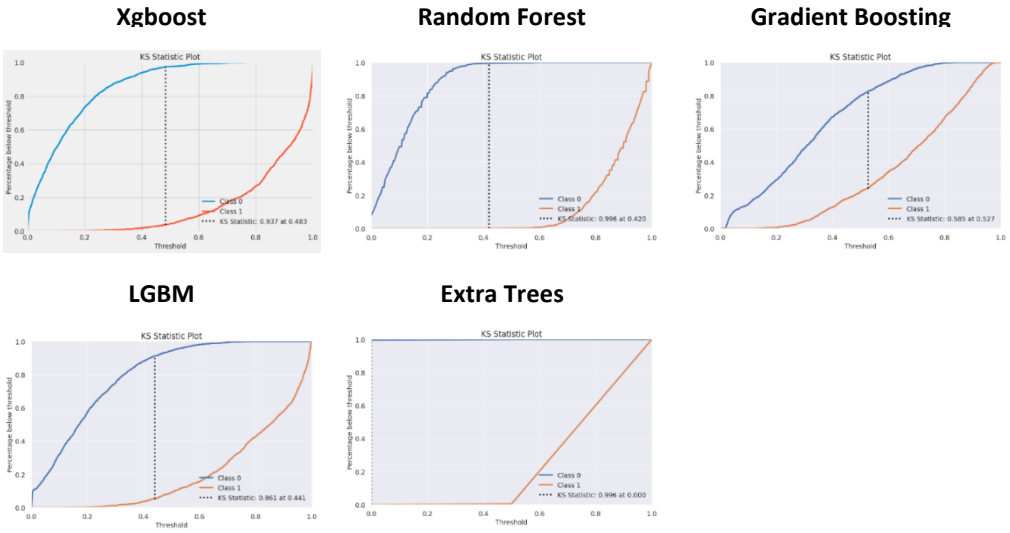
Tabla 13 - Cuadrícula de puntuaciones medias validadas de forma cruzada para modelos entrenados

Model	Accuracy	AUC	Recall	Prec	F1	Kappa	MCC
Light Gradient Boosting Machine	0.755	0.848	0.741	0.762	0.751	0.510	0.511
Random Forest Classifier	0.756	0.843	0.741	0.765	0.752	0.513	0.514
Extreme Gradient Boosting	0.753	0.839	0.737	0.762	0.748	0.506	0.507
Gradient Boosting Classifier	0.737	0.835	0.732	0.742	0.736	0.475	0.476
Extra Trees Classifier	0.747	0.829	0.746	0.748	0.747	0.495	0.496

Fuente: Elaboración Propia

Estas métricas se complementan con el análisis gráfico del área bajo la curva (AUC) y la matriz de confusión, donde se observa que no existen mayores diferencias entre los modelos y cada uno tiene una buena capacidad de predicción de clientes buenos y malos o en default (Anexo 7 y Anexo 8). Sin embargo, si se encuentran diferencias en otras métricas de evaluación. En el caso del estadístico Kolmogorov Smirnov, utilizado para medir la capacidad del modelo de diferenciar las clases, en este caso entre clientes buenos y malos, donde en un puntaje de 0 a 100, valores cercanos a 100 indican la capacidad del modelo para clasificar de forma correcta y cercanos a 0 su baja o nula capacidad. En la Figura 15, se observa como la técnica de Extra Trees tiene un estadístico de 0, es decir no posee una buena capacidad para diferenciar. Por otro lado, bajo la técnica de Gradient Boosting Classifier y Extreme Gradient Boosting se identifican con los valores más alto para esta métrica.

Figura 15 - Kolmogorov Smirnov para las técnicas evaluadas

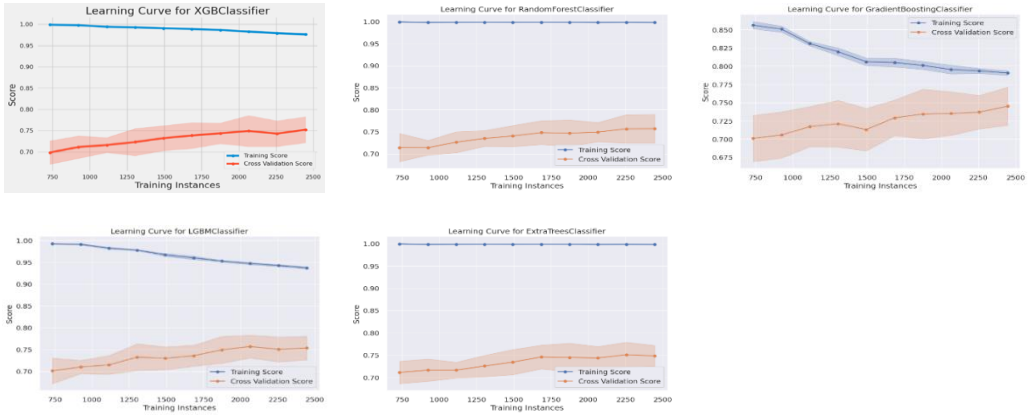


Fuente: Elaboración Propia

Finalmente, se evalúa la curva de aprendizaje, la cual muestra cómo el rendimiento del modelo varía a medida que se incrementa la cantidad de datos de entrenamiento utilizados. Permite evaluar como mejora el rendimiento del entrenamiento a partir de una muestra más grande y a su vez como el conjunto de prueba mejora a medida que se aumenta la cantidad de datos de entrenamiento y no está sobre ajustando los datos de entrenamiento.

En la Figura *16*, se observa que para las técnicas de *Extreme Gradient Boosting, Extra Trees Classifier, Random Forest Classifier,* indican un potencial sobre ajuste de estas técnicas dado que su puntaje es casi perfecto, donde además la brecha con el puntaje de la curva de validación cruzada o de prueba es muy alta. En el caso de la técnica de *Gradient Boosting Classifier* se observa un puntaje aceptable alejado del sobreajuste y una tendencia hacia la convergencia de las dos curvas, con mejoras de la curva de prueba sin grandes pérdidas de rendimiento en el conjunto de entrenamiento a medida que aumenta la cantidad de datos.

Figura 16 - Curva de entrenamiento para las técnicas evaluadas



Fuente: Elaboración Propia

De acuerdo con lo anterior, el modelo desarrollado bajo la técnica *Gradient Boosting Classifier comienza a perfilarse como el potencial a modelo a seleccionar*, al presentar un desempeño superior en las últimas 2 métricas analizadas en comparación a las demás. Se procederá a realizar el testing de cada modelo con los datos de prueba y se evaluaran las métricas resultantes para proceder a seleccionar el modelo definitivo.

6.5. Testing y selección del modelo

Se realiza el testing de cada uno de los modelos creados en la fase anterior y se generan las respectivas métricas de evaluación. En la Tabla *14* se observa que con los datos de prueba el modelo desarrollado bajo la técnica *de Gradient Boosting Classifier,* tiene un rendimiento ligeramente superior a las demás en especial en la precisión y el área bajo la curva. Esto se confirma con el reporte de clasificación de la Tabla *15*, en la cual se observa que es la técnica con mayor capacidad de predicción de los clientes buenos y de la categoría objetivo: el default. Esta técnica logra clasificar de forma correcta el 55% de los clientes malos como tal y el 76,5% de los clientes buenos reales (Anexo 9). Las demás técnicas logran tener desempeños aceptables y parejos en la clasificación de clientes "buenos", pero no ocurre lo mismo para los clientes en default.

Tabla 14 - Métricas del testing de las técnicas seleccionadas

Métrica	Rf	Xgboost	LGBM	ET	GBC
Accuracy	0,6681	0,6510	0,6681	0,6524	0,6923
Precision	0,5180	0,4907	0,5171	0,4935	0,5519
Recall	0,4772	0,4357	0,5021	0,4730	0,5519
F1	0,4968	0,4615	0,5095	0,4831	0,5519
MSE	0.3176	0.3361	0.3404	0.3518	0,3148
AUC	0,6225	0,5996	0,6285	0,6096	0,6588

Rf: Random Forest, XgBoost: Extreme Gradient Boosting, LGBM: Light Gradient Boosting Machine, ET: Extra Trees, GBC: Gradient Boosting Classifier. **Fuente:** Elaboración Propia

Tabla 15 - Reporte de clasificación del testing de las técnicas seleccionadas

Precision			Recall			F1		
Técnica	0	1	Técnica	0	1	Técnica	0	1
Xgboost	0,72	0,49	Xgboost	0,76	0,44	Xgboost	0,74	0,46
Rf	0,74	0,52	Rf	0,77	0,48	Rf	0,75	0,5
LGBM	0,74	0,52	LGBM	0,75	0,5	LGBM	0,75	0,51
ET	0,73	0,49	ET	0,75	0,47	ET	0,74	0,48
GBC	0,77	0,55	GBC	0,77	0,55	GBC	0,77	0,55

Rf: Random Forest, XgBoost: Extreme Gradient Boosting, LGBM: Light Gradient Boosting Machine, ET: Extra Trees, GBC: Gradient Boosting Classifier. **Fuente:** Elaboración Propia

A partir de los resultados anteriores se selecciona el modelo desarrollado bajo la técnica de *Gradient Boosting Classifie*r, el cual presenta mejor capacidad para detectar los clientes buenos y malos, mayor precisión general, sin presentar indicios de sobreajuste o alta varianza.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

La caracterización de datos suministradas por la Fundación Santo Domingo proporcionó una comprensión profunda de la información disponible para la evaluación de riesgo de crédito. Esto permitió identificar variables relevantes y garantizar la calidad de los datos utilizados en el modelo de Machine Learning.

La codificación de variables para normalizarlas fue fundamental para garantizar la comparabilidad y la consistencia en el análisis de riesgo de crédito. La aplicación de técnicas como la codificación ordinal, la codificación one-hot o la normalización z-score, permitió estandarizar las variables en un rango común, facilitando la interpretación, el análisis de los resultados del modelo y el uso de diferentes algoritmos.

Un aspecto clave del trabajo y resultados del modelo se asocia al proceso realizado para la selección de variables y reducción de la dimensionalidad. Al emplear diferentes técnicas como la Regresión LASSO, Information Value (IV) y Extra Tree Classifier, se logró evaluar el impacto de la técnica para eliminar la redundancia y la correlación entre las variables, mejorando así la calidad de los resultados y reduciendo el tiempo de procesamiento. Sus resultados permitieron seleccionar las variables más representativas para las unidades productivas de personas naturales, las cuales representan más del 90% de los clientes de la Fundación y su potencial uso practico.

El entrenamiento de los modelos de Machine Learning aplicados en la evaluación de concesión o renovación de microcréditos demostró ser efectivo y prometedor. Los algoritmos utilizados, en especial el Gradient Boosting Classifier, mostró niveles de precisión del 76,5% en la predicción del riesgo crediticio de los clientes foco de la organización. Adicionalmente, el análisis comparativo de la precisión de los modelos utilizados reveló que el Gradient Boosting Classifier superó a otros algoritmos en términos de rendimiento y capacidad predictiva. Su capacidad para manejar grandes conjuntos de datos y su capacidad para controlar el sobreajuste lo convierten en una opción sólida para la evaluación de riesgo de crédito. Su proceso de elección demuestra la importancia y utilidad de comparar diversas métricas de rendimiento que permitan encontrar la técnica con mejor rendimiento, pertinencia y ajuste a los datos disponibles.

Se resalta la utilidad que tuvieron las pruebas de Kolmogórov-Smirnov y la curva de aprendizaje, las cuales permitieron evidenciar que las técnicas con mejor AUC presentaban problemas de sobreajuste u overfitting lo que desencadenaría problemas en el futuro, pues los modelos no podrían considerar nuevos datos de entrada validos al salirse del patrón establecido. Lo anterior permite confirmar que no existen técnicas ideales para resolver un problema, sino que la modelación es un proceso iterativo e integral en el que deben evaluarse de forma individual diversas pruebas y ajustarse diferentes parámetros para encontrar la mejor combinación con los datos disponibles sin incurrir en desarrollas modelos “cuasi perfectos” o incapaces de mostrar la realidad y adaptarse a la realidad del negocio, con sus imperfecciones o limitaciones asociadas.

La evaluación de la efectividad y precisión del modelo seleccionado demostró que el modelo de riesgo de

crédito basado en técnicas de Machine Learning en general fue capaz de tomar decisiones precisas y confiables sobre la concesión y renovación de microcréditos, en especial para clasificar de forma adecuada los clientes buenos, para los cuales el modelo desarrollado podría permitir la implementación de un proceso ágil de preaprobación o aprobación para estos clientes. Existe una oportunidad de mejora del modelo para clasificar en una proporción más alta los potenciales clientes malos.

En general, el desarrollo e implementación de un modelo de riesgo de crédito basado en técnicas de Machine Learning ha demostrado ser una herramienta prometedora para la Fundación Santo Domingo, permitiéndole la implementación de una herramienta objetiva y automatizada para evaluar solicitudes de crédito, haciendo uso de datos históricos y la identificación de patrones de comportamiento para mejorar el proceso de aprobación de créditos.

7.2. Trabajos Futuros

Para la organización el desarrollo de las diferentes etapas del trabajo ha generado oportunidades de trabajos a futuro en materia de analítica y ciencia de datos. En un primer aspecto se identificaron oportunidades de redefinir los segmentos de clientes a través de un ejercicio de clústeres, permitiendo ajustar la estrategia comercial y las políticas de riesgo a partir de entender a profundidad las características de cada grupo en función de su comportamiento de pago.

El ejercicio no contempla la modelación del default para las unidades productivas jurídicas, dado el bajo nivel de representación de este segmento en la muestra de créditos y la disponibilidad de variables, por lo que el modelo resultante solo debe aplicarse para los clientes de los segmentos de Emprendedor en Desarrollo e Independiente. Por tal razón, se debe trabajar en la estructuración de un modelo para el segmento *Comercial*, enfocado en las características de la unidad productiva, al ser empresas de mayor consolidación y desarrollo, donde se puedan incluir variables financieras, pertinentes para evaluar este tipo de MiPymes.

Finalmente, para el modelo desarrollado se puede evaluar la posibilidad de incluir un aspecto geográfico del barrio tal como una agrupación superior a nivel de localidades o una variable proxy de las condiciones socioeconómicas del barrio como el Índice de pobreza multidimensional, datos disponibles para Cartagena, Barranquilla y Soledad donde se concentra la operación de la Fundación. En este mismo sentido, se podrían incluir indicadores económicos regionales, datos demográficos o información del sector específico al que pertenecen las microempresas. La inclusión de variables financieras que permitieran capturar con mayor precisión y diferenciar entre si a las unidades productivas como el flujo de caja disponible, rentabilidad, podría ayudar a mejorar la sensibilidad del modelo y aumentar la identificación de “malos” clientes.

A nivel académico, un aspecto interesante es la comparación de datos a nivel regional para evaluar el desempeño y ajuste de diferentes técnicas de Machine Learning a las características particulares de cada región y de los microempresarios de dichas regiones.

REFERENCIAS BIBLIOGRÁFICAS

[1] ConfeCámaras - Red de Cámaras de Comercio, «Informe de dinámica de creación de empresas,» ConfeCámaras, Bogotá D.C., 2022.

[2] S. Guo, J. Ma y H. Liao, «The applications of machine learning in credit scoring: A systematic review,» *Journal of Ambient Intelligence and Humanized Computing,* 2021.

[3] ConfeCámaras - Red de Cámaras de Comercio, «Fortaleza del tejido empresarial colombiano - Recuperación post pandemia - 2022,» Bogotá D.C, 2022.

[4] Banca de las Oportunidades; Superintendencia Financiera de Colombia; CAF, «Estudio de demanda de inclusión financiera,» 2017.

[5] G. Ariza-Arteaga, J. Cely-Gómez y A. Plazas-Támara, «La economía criminal en Colombia: el caso del préstamo 'gota a gota'.,» *Revista Científica General José María Córdova,* 2021.

[6] DataCredito, «Datacrédito Experian Colombia: ¿Qué es Datacrédito?,» 2022. [En línea]. Available: https://www.datacredito.com.co/nosotros.

[7] Fundación Santo Domingo, «Proyecciones financieras Unidad de Financiamiento y Desarrollo Empresarial,» 2022.

[8] A. Ampountolas, T. Nyarko Nde, P. Date y C. Constantinescu, «A Machine Learning Approach for Micro-Credit Scoring,» *Risks,* 2021.

[9] Congreso de la República de Colombia, «Ley 905 de 2004, por la cual se establecen medidas para promover el empleo y se dictan otras disposiciones.,» 26 agosto 2004. [En línea]. Available: http://www.secretariasenado.gov.co/senado/basedoc/ley_0905_2004.html.

[10] Ministerio de Comercio, Industria y Turismo, «MI PYMES,» 2023. [En línea]. Available: https://www.mipymes.gov.co/.

[11] Fundación Santo Domingo, «Somos Fundación,» 2023. [En línea]. Available: https://www.fundacionsantodomingo.org/somos/.

[12] Fundación Santo Domingo, «Desarrollo Empresarial - Unidad de Financiamiento y Desarollo Empresarial,» 2023. [En línea]. Available: https://www.fundacionsantodomingo.org/desarrollo-empresarial/.

[13] Encyclopædia Britannica, «Microcredit,» 2023. [En línea]. Available: https://www.britannica.com/topic/microcredit.

[14] Super Intendencia Financiera de Colombia, *Circular Externa 100 de 1995,* Bogotá D.C., 1995.

[15] I. y. T. Ministerio de Comercio, «Decreto 1072 de 2015,» 2015. [En línea]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=62953.

[16] D. J. Hand, H. Mannila y P. Smyth, «Principles of data mining,» *MIT press,* 2001.

[17] R. Araújo et al., «Machine Learning Techniques Applied to Credit Scoring: A Systematic Literature Review,» *Journal of Intelligent & Fuzzy Systems,* vol. 40, nº 3, 2021.

[18] L. Useche y D. Mesa, «Una introducción a la imputación de valores perdidos,» *Terra Nueva Etapa,* 2006.

[19] D. M. Hawkins, «Identification of outliers.,» *Chapman and Hall/CRC,* 1980.

[20] M. Krstajic, J. Smith, R. G. Garrett y R. A. Hardie, «Effective exploratory data analysis: A review of methods and their practical application.,» *Journal of Advertising Research,* 2021.

[21] Asobancaria, «Banca & Economía,» Bogotá D.C., 2022.

[22] DANE, «Caracterización pobreza monetaria y resultados clases sociales 2020,» Bogotá D.C., 2020.

[23] L. López-Calva y E. Ortiz-Juárez, «A Vulnerability Approach to the Definition of the Middle Class,» Mimeo, World Bank, 2011.

[24] H. Abdi y L. J. Williams, «Principal component analysis. Wiley Interdisciplinary Reviews,» *Computational Statistics,* 2010.

[25] B. Lund y D. Brotherton, 2013. [En línea]. Available: https://www.lexjansen.com/mwsug/2013/AA/MWSUG-2013-AA14.pdf.

[26] V. Jaramillo y W. Ossa, «Machine Learning para la estimación del riesgo de crédito en una cartera de consumo,» 2021. [En línea]. Available: https://repository.eafit.edu.co/bitstream/handle/10784/29589/Wbeimar_OssaGiraldo_Veronica_JaramilloMarin_2021.pdf?sequence=2&isAllowed=y.

[27] Y. Lean, Z. Xiaoming y Y. Hang, «An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity,» *Expert Systems with Applications,* vol. 202, 2022.

[28] F. Medeiros Assef y M. Arns Steiner, «Ten-year evolution on credit risk research: a Systematic Literature Review approach and discussion,» *Ingenieria e investigación,* vol. 40, nº 2, pp. 50-71, 2020.

[29] S. Trishita, K. Souvik, K. Saroj y S. Saptarsi, «Credit Risk Prediction using Extra Trees Ensemble,» *ternational Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON),* 2023.

[30] G. Loterman, I. Brown, D. Martens, C. Mues y B. Baesens, «Benchmarking regression algorithms for loss given default modeling,» *International Journal of forecasting ,* vol. 28, pp. 161-170, 2018.

[31] M. Khan, M. A. Hossain, M. I. Jantan, «Credit Scoring Models using Random Forest and Extra Trees Classifier.,» *International Journal of Computer Science and Network Security (IJCSNS),* vol. 21, nº 4, pp. 171-179, 2021.

8. ANEXOS

Anexo 1 - Monto y Plazos de Desembolsos por Segmento de Clientes

Variable	Emprendedor independiente	Emprendedor en desarrollo	Comercial
Líneas	Todas	Todas	Todas
Monto máximo crédito	De acuerdo con niveles de ventas reportados ○ Hasta 4 SMMLV: 4 SMMLV. ○ Entre 4 y 8: SMMLV equivalente a un mes de ventas	120 SMMLV	○ Hasta $330 Pequeña empresas ○ Hasta $150 Micro, ○ Hasta $150 Emprendedor etapa temprana
Plazo	Hasta 12 meses para ventas <= 4 SMMLV. Hasta 24 meses para ventas >4 y <= 8 SMMLV.	○ Capital de trabajo: Hasta 24 meses, ○ Activos fijos: hasta 48 meses ○ Otros: Hasta 36 meses.	○ Activo fijo: Hasta 60 meses ○ Otras líneas: Hasta 48 meses

Fuente: Elaboración Propia

Anexo 2 - Distribución de clientes naturales por género

Variable	Categoría	%
Género	Femenino	63,0%
	Masculino	37,0%

Fuente: Elaboración Propia

Anexo 3 - Características Sociodemográficas de las Personas Naturales por Género

Variable	Categoría	Mujeres	Hombres	Total
Nivel de estudios	Analfabetismo	0%	0%	0%
	No escolarizado	0%	0%	0%
	Primaria	3%	4%	4%
	Secundaria	45%	44%	45%
	Técnica	32%	29%	31%
	Tecnológica	5%	5%	5%
	Universitaria	14%	17%	15%
	Especialización	0%	1%	0%
	Magister	0%	0%	0%
Estado Civil	Divorciado	1%	0%	1%

Variable	Categoría	Mujeres	Hombres	Total
	Otro	0%	1%	0%
	Separado	1%	0%	1%
	Soltero	30%	38%	33%
	Viudo	0%	0%	0%
	Unión libre	42%	38%	41%
	Casado	26%	22%	24%
Tipo de vivienda	Otra	8%	12%	9%
	Sin vivienda	25%	10%	19%
	Arriendo	10%	16%	12%
	Familiar	21%	28%	23%
	Inmueble con Hipoteca	0%	0%	0%
	Propia	37%	35%	36%
Personas a cargo	0	84%	78%	82%
	1	10%	13%	11%
	2	6%	6%	6%
	3	1%	2%	1%
	4	0%	0%	0%
	5	0%	0%	0%
	6	0%	0%	0%
	9	0%	0%	0%
Estrato	1	34%	39%	36%
	2	45%	44%	44%
	3	14%	13%	14%
	4	4%	3%	4%
	5	2%	1%	2%
	6	0%	0%	0%

Fuente: Elaboración Propia

Anexo 4 - Distribución de Clientes Personas Naturales por Género y Segmento

Segmento	Mujeres	Hombres	Total
Emprendedor independiente	72%	66%	70%
Emprendedor en Desarrollo	28%	34%	30%
Total	100%	100%	100%

Fuente: Elaboración Propia

Anexo 5 – Porcentaje de Clientes en default por Segmento y Actividad Económica

Actividad económica	Independiente	En Desarrollo
Comercio al por menor	44%	45%
Actividades de servicios de comidas y bebidas	11%	13%
Elaboración de productos alimenticios	8%	9%
Otras actividades de servicios personales	8%	8%
Confección de prendas de vestir	5%	2%
Comercio al por mayor y en comisión o por contrata	2%	4%
Actividades administrativas y de apoyo de oficina y otras actividades de apoyo a las empresas	3%	1%
Comercio, mantenimiento y reparación de vehículos automotores y motocicletas	2%	4%
Otras actividades profesionales, científicas y técnicas	2%	2%

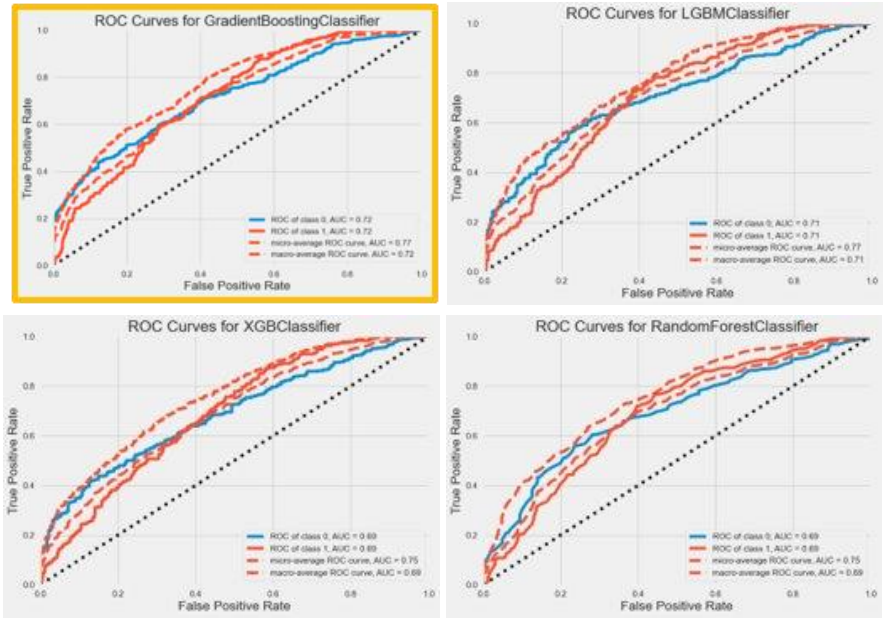
Fuente: Elaboración Propia.

Anexo 6 – Porcentaje de Recursos (monto) en default por segmento y actividad económica

Actividad económica	Independiente	En Desarrollo
Comercio al por menor, excepto el de vehículos automotores y motocicletas	44%	44%
Actividades de servicios de comidas y bebidas	9%	13%
Comercio al por mayor y en comisión o por contrata	2%	6%
Actividades especializadas para la construcción de edificios y obras de ingeniería civil	1%	1%
Elaboración de productos alimenticios	7%	6%
Construcción de edificios	0%	0%
Otras actividades de servicios personales	7%	7%
Comercio, mantenimiento y reparación de vehículos automotores y motocicletas	2%	5%
Confección de prendas de vestir	4%	1%
Transformación de la madera y fabricación de productos de madera y de corcho, excepto muebles	1%	0%

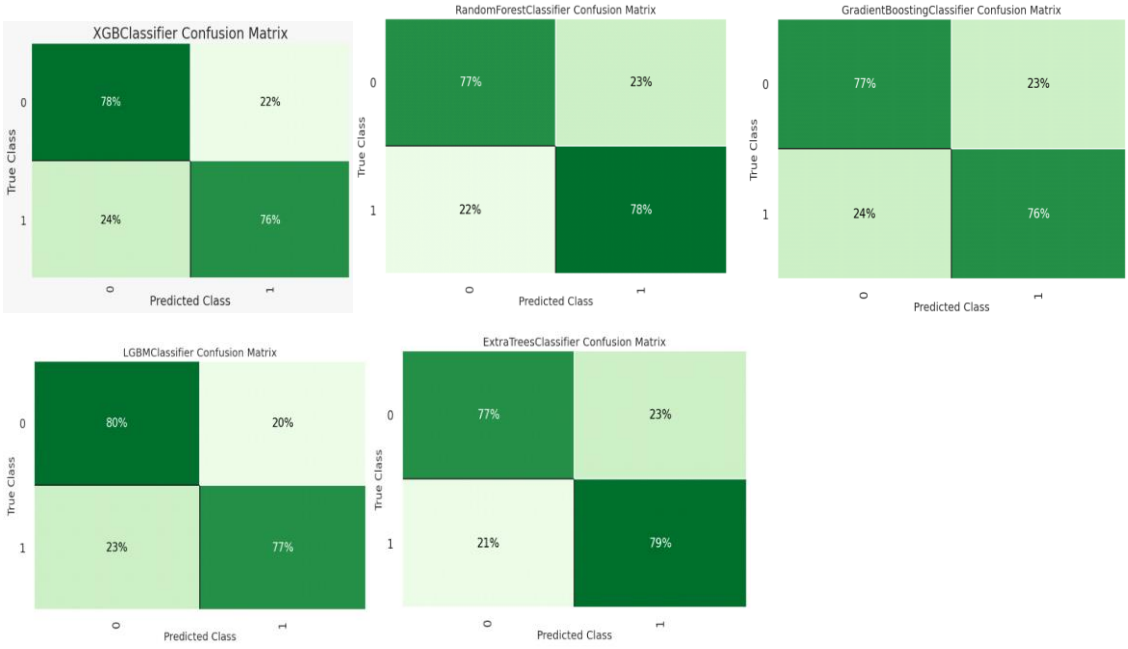
Fuente: Elaboración Propia

Anexo 7 - Área Bajo la Curva (AUC) de los Modelos Entrenados



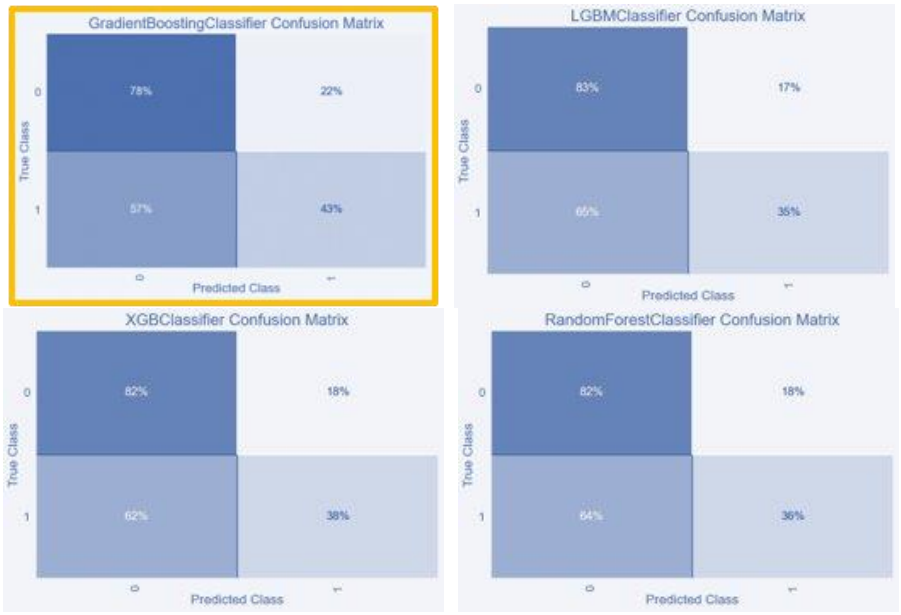
Fuente: Elaboración Propia

Anexo 8 - Matriz de Confusión de los Modelos Entrenados



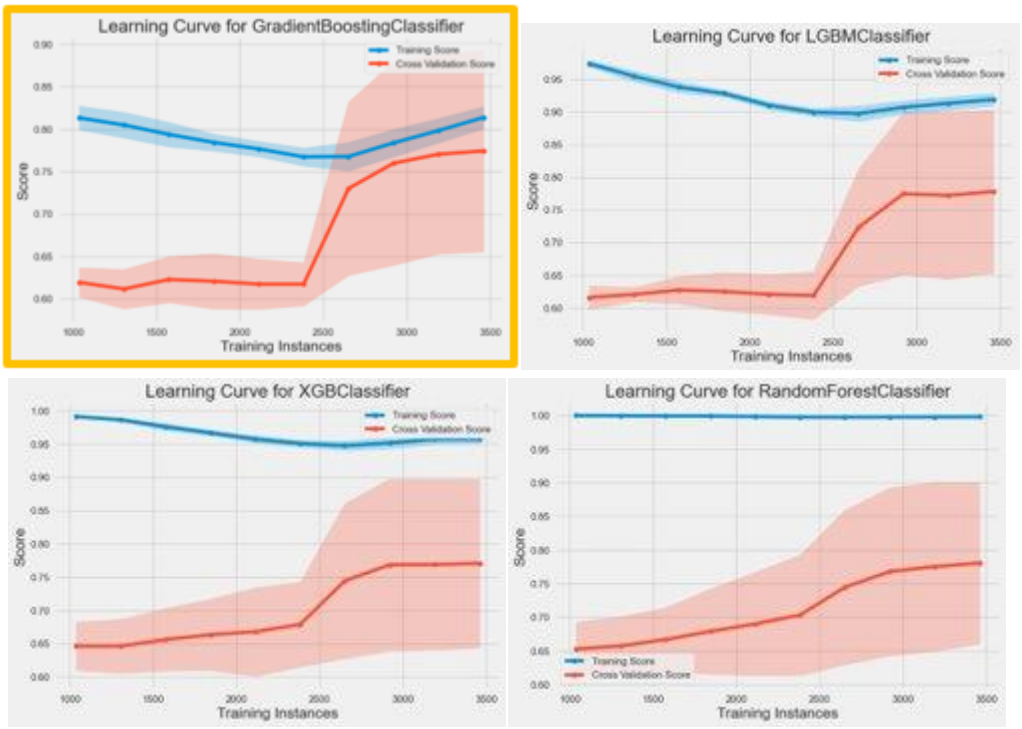
Fuente: Elaboración Propia

Anexo 9 - Matriz de Confusión: Testing de las técnicas seleccionadas



Fuente: Elaboración Propia

Anexo 10 - Curva de aprendizaje de las técnicas seleccionadas



Fuente: Elaboración Propia