GOBIERNO DE MÉXICO | CONAHCYT CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS | INFOTEC

**BIBLIOTECA INFOTEC**
**VISTO BUENO DE TRABAJO TERMINAL**

Maestría en Ciencia de Datos e Información
(MCDI)

Ciudad de México, a 28 de junio de 2024

**UNIDAD DE POSGRADOS**
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Diseño de modelo para la recuperación de unidades en cartera vencida de tipo automotriz por medio de visitas domiciliarias"

Desarrollado por la alumna: **Yuridiana de Jesús Reyes Delgado** bajo la asesoría de la **Dra. Guadalupe O. Gutiérrez Esparza** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que la alumna no adeuda materiales de la biblioteca de INFOTEC.

**No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.**

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

**Mtro. Carlos Josué Lavandeira Portillo**
Director Adjunto de Innovación y Conocimiento

jah

CJLP/jah

C.c.p. Felipe Alfonso Delgado Castillo.- Gerente de Capital Humano.- Para su conocimiento.
Yuridiana de Jesús Reyes Delgado.- Alumna de la Maestría en Ciencia de Datos e Información.- Para su conocimiento.

MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO

GERENCIA DE CAPITAL HUMANO

POSGRADOS

# "Diseño de Modelo para la recuperación de unidades en cartera vencida de tipo automotriz por medio de visitas domiciliarias"

SOLUCIÓN ESTRATÉGICA
Que para obtener el grado de MAESTRA EN CIENCIAS DE DATOS E INFORMACIÓN

Presenta:

**Yuridiana de Jesús Reyes Delgado**

Asesora:

**Dra. Guadalupe O. Gutiérrez Esparza**

Ciudad de México, 06, 2024

CONACYT

INFOTEC POSGRADOS

# Agradecimientos

En este momento de culminación y reflexión, deseo expresar mi más profundo agradecimiento a todas aquellas personas que han sido pilares fundamentales en este viaje académico y personal.

A mi familia y amigos, a quienes amo profundamente y sé que me aman con la misma intensidad, les agradezco de corazón por estar siempre a mi lado, ofreciéndome su apoyo incondicional, su comprensión y su amor. Su presencia en mi vida ha sido una fuente constante de fortaleza y motivación.

De manera especial, deseo reconocer y agradecer a mi asesora, la Dra. Guadalupe Gutiérrez, por su invaluable contribución a mi formación y a este trabajo. Su guía experta, su dedicación y su paciencia han sido fundamentales para mi desarrollo académico y personal. Gracias, Dra. Guadalupe, por compartir su conocimiento, por inspirarme a superar cada desafío y por creer en mi capacidad para llevar a buen término este proyecto. Su apoyo ha sido un regalo invaluable que siempre llevaré conmigo.

A todos, gracias por formar parte de este viaje, por su fe inquebrantable en mí y por ayudarme a convertir este sueño en realidad.

# Tabla de contenido

# Índice de figuras

# Índice de cuadros

# Siglas y abreviaturas

(**IM**): Indice de Marginación

(**IMN**): Indice de Marginación Nominal

(**KM**): Kilómetro

(**KPI**): Key Performance Indicator

(**MySql**): My Structured Query Language

(**S. de R.L de C.V**): Sociedad de Responsabilidad Limitada de Asociación o Sociedad Civil

(**SEPOMEX**): Servicio Postal Méxicano

(**STE**): Saldo Total Exigible

(**SVM**): Support Vector Machine

# Introducción

Las empresas dedicadas a la gestión de cartera vencida surgieron en la década de los 90's, principalmente como resultado de la crisis bancaria en México y la necesidad de los bancos y otras instituciones financieras de recuperar activos en mora (Girón and Correa, 1997). A medida que aumentaba el número de préstamos y tarjetas de crédito otorgados, también crecía la proporción de cuentas en situación de impago, lo cual llevo a cabo la aparición de empresas especializadas en la recuperación de créditos vencidos (Arce Montaño, 2017).

Uno de los principales problemas a los que se enfrentan las empresas de cartera vencida en México es la alta tasa de morosidad (Conde Hernández et al., 2003). Los motivos detrás de la morosidad son diversos y pueden incluir dificultades económicas de los deudores, cambios en las circunstancias personales o laborales, falta de educación financiera y otros factores socioeconómicos. La recuperación de cartera vencida implica enfrentar una serie de desafíos, como el seguimiento de los deudores, la negociación de acuerdos de pago y, en algunos casos, la ejecución de acciones legales para recuperar los activos.

Para abordar estos desafíos, las empresas de cartera vencida han implementado diversas estrategias y enfoques que van desde el análisis de datos y segmentación, negociación y acuerdos de pago, colaboración con despachos legales e incluso la aplicación de algoritmos de aprendizaje automático (Bareño Amezquita, 2023). El aprendizaje automático ha tenido un impacto significativo en la recuperación de carteras vencidas al proporcionar herramientas y técnicas avanzadas para analizar datos, identificar patrones y predecir comportamientos futuros de los deudores. A través de modelos de aprendizaje automático, las empresas han mejorado la eficiencia y efectividad de sus estrategias de recuperación, optimizando la asignación de recursos y enfoconda su esfuerzo en las cuentas con mayor probabilidad de éxito (Buitrón et al., 2022; Wang et al., 2021).

Un estudio reciente llevado a cabo por (Singh et al., 2021) se centra en el uso de al-

goritmos de aprendizaje automático para predecir la aprobación de préstamos en el sector bancario. El objetivo principal del estudio es utilizar datos históricos de los candidatos para construir un modelo de aprendizaje automático que prediga si un nuevo solicitante recibirá o no un préstamo. En este trabajo se utilizaron algoritmos de clasificación, tales como regresión logística, random forest y support vector machines.

En otro estudio llevado a cabo por (Nazemi et al., 2022) se utilizan técnicas de aprendizaje automático para predecir la tasa de recaudo en 65,000 mil créditos de consumo en el sector de telecomunicaciones de una empresa Alemana dedicada a adquirir créditos en mora. Los algoritmos que utilizaron para la construcción del modelo predictor fue la máquina de vectores de soporte, regresión lineal y aprendizaje profundo.

Otro estudio realizado por (Tumuluru et al., 2022) aborda el desafío de estimar el riesgo involucrado en la aprobación de préstamos utilizando modelos de aprendizaje automático. Este análisis se enfoca en la aplicación de dichos modelos para identificar tendencias dentro de un conjunto de datos estándar de préstamos aprobados, con el objetivo de prever posibles incumplimientos futuros. Para lograr esto, se emplearon varios modelos de aprendizaje automático, incluyendo Random Forest, Support Vector Machine, K-Nearest Neighbor y regresión logística. Siendo Random Forest el modelo que alcanzó un mayor desempeño en comparación con los otros modelos utlizados.

Otro modelo de aprendizaje automático que también demostró su eficacia para el apoyo de la gestión de cobranza y recuperación de carteras vencidas fueron las redes neuronales, como lo indica Buitron (Buitrón et al., 2022). Buitron aplicó modelos de aprendizaje automático para anticipar la probabilidad de pago de los clientes a una agencia de cobranza durante los primeros tres meses después de acordar el pago. La implementación incluyó redes neuronales y regresión logística, identificando características (features) clave que influyerón en el cumplimiento del cliente. En este caso, los resultados resaltaron el rendimiento de las redes neuronales en términos de desempeño.

En este trabajo se considera una base de datos correspondiente a la empresa RK Representaciones Empresariales S. de R.L de C.V. fundada en 2008, especializada en la recuperación de cartera vencida de tarjetas de crédito, préstamos de consumo, prés-

tamos personales, préstamos automotrices y préstamos hipotecarios. La relevancia de esta investigación radica en el desarrollo de un modelo que mejore la eficacia en la recuperación de deudas pendientes, con un enfoque particular en la recuperación de vehículos. Por lo anterior, para lograr una recuperación efectiva, se considera fundamental contar con herramientas y estrategias que permitan identificar y clasificar las cuentas de los titulares en cuanto a la probabilidad de obtener pagos.

Mediante el análisis de la cartera vencida y la aplicación de modelos de aprendizaje automático, RK Representaciones Empresariales podrá optimizar sus esfuerzos de recuperación, enfocándose en aquellas cuentas que demuestren mayor probabilidad de pago. Esto no solo mejorará la eficiencia operativa de la empresa, sino que también contribuirá a fortalecer su posición en el mercado y aumentar su rentabilidad.

En este estudio, se analizarán los datos disponibles, se implementarán modelos de aprendizaje automático y se medirá la efectividad de los modelos desarrollados. Los resultados y conclusiones derivados de este análisis proporcionarán la base para una estrategia o plan de trabajo para recuperación de vehículos o cobro de deudas pendientes, la cual se centrará en optimizar el proceso de selección de visitas.

# Capítulo 1

# Identificación del problema

# 1 Capítulo 1

## 1.1. Planteamiento del problema

Las empresas que gestionan carteras vencidas en el área automotriz prefieren recuperar coches impagados, esto debido a la necesidad de éstas de proteger sus activos financieros y mantener la estabilidad económica. La recuperación de estos vehículos, que representan inversiones significativas, permite mitigar las pérdidas asociadas con los impagos al posibilitar su venta para recuperar fondos.

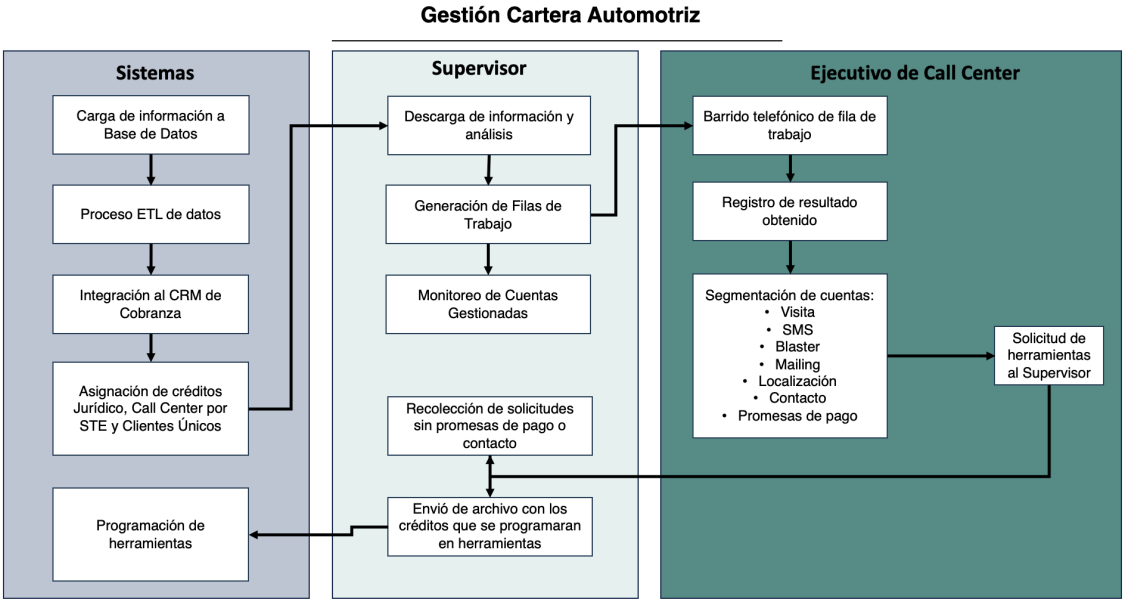
En este caso la empresa RK Representaciones Empresariales S. de R.L. de C.V. también ha mostrado la necesidad de recuperar la cartera vencida automotriz, con el objetivo específico de recuperar unidades morosas para su posterior venta. No obstante, a pesar de contar con años de experiencia y disponer de un historial detallado de las cuentas, la empresa no ha optimizado el potencial de la información histórica disponible ni la gestión de visitas domiciliarias.

Con base a lo anterior, la problemática de la empresa comienza con el registro de las cuentas deudoras en el sistema de cobranza, seguido por la asignación de éstas a ejecutivos que validan los números telefónicos y realizan investigaciones en redes sociales sobre los titulares. Las cuentas son clasificadas en categorías como promesas de pago, convenios, solicitudes de mensajes, llamadas automatizadas, correos electrónicos y visitas para notificaciones de pago. Algunas de éstas son asignadas a visitas, las cuales son gestionadas por ejecutivos domiciliarios, asimismo, en este caso es el Coordinador quien planifica la ruta más eficiente considerando estados y costos de traslado. Cabe destacar que el proceso de visitas se aplica únicamente en localidades cercanas a la Ciudad de México y sus alrededores.

En el caso de visitas que deban realizarse en el resto de los Estados o en áreas catalogadas como riesgosas debido a su ubicación, se lleva a cabo un análisis que considera tanto el monto del crédito como la viabilidad de recuperación del vehículo por parte

del equipo legal. Una vez se tiene identificado el número de cuentas a visitar, se genera un reporte por día, el cual se comparte con el coordinador, quien da seguimiento a cada caso de los titulares contactados. Cabe resaltar que este último proceso es realizado por cada ejecutivo de cobranza, sin tener en cuenta un análisis que calcule la probabilidad de pago. Las estrategias de cobro que implican visitas domiciliarias constituyen un gasto considerable para la empresa, y frecuentemente incurren en pérdidas debido a la falta de un análisis previo que identifique cuándo son aconsejables estas visitas y qué factores claves deben evaluarse para predecir la probabilidad de pago por parte del deudor. Este enfoque justifica la realización de la visita domiciliaria. Este procedimiento se ilustra en la Figura 1.1.

Figura 1.1: Diagrama. Gestión Cartera Automotriz




Fuente: Elaboración propia, 2023.

Adicionalmente, el cuadro 1.1 muestra un resumen sobre los datos de cartera vencida automotriz de la empresa. Ésta destaca que aproximadamente el 53 % de la cartera vencida está compuesta por saldos entre 100,001 a 250,000. Esta concentración temporal sugiere una oportunidad estratégica para la recuperación de estos vehículos, ya que son relativamente saldos medios que podrían tener un valor de mercado competitivo en comparación con modelos con saldos altos.

Cuadro 1.1: Rango de saldos

| **Rango Saldos** | **Clientes** | **Créditos** | **STE** | **Participación** |
|---|---|---|---|---|
| 1. <= 50,000 | 365 | 367 | 8,237,106.50 | 3% |
| 2. 50,001 - 100,000 | 332 | 338 | 24,770,463.35 | 8% |
| 3. 100,001 - 150,000 | 372 | 379 | 46,346,030.6 | 15% |
| 4. 150,001 - 200,000 | 300 | 311 | 52,871,768.94 | 17% |
| 5. 200,001 - 250,000 | 275 | 292 | 61,582,522.01 | 20% |
| 6. 250,001 - 300,000 | 112 | 124 | 30,419,901.96 | 10% |
| 7. 300,001 - 350,000 | 49 | 54 | 15,833,930.27 | 5% |
| 8. 350,001 - 400,000 | 25 | 31 | 9,380,368.20 | 3% |
| 9. 400,001 - 500,000 | 46 | 60 | 20,475,518.38 | 7% |
| 10. >= 500,000 | 45 | 81 | 32,513,343.73 | 11% |
| **Total** | **1,921** | **2,037** | **302,430,953.90** | **100%** |

Fuente: Elaboración propia, 2023.

En conclusión, el problema de la empresa RK Representaciones Empresariales S. de R.L. de C.V. radica en la gestión óptima de su cartera automotriz vencida. A pesar de la clara necesidad de recuperar unidades impagadas para su posterior venta, la empresa no ha aprovechado de manera efectiva la información histórica acumulada y carece de herramientas analíticas que permitan una evaluación precisa de la probabilidad de pago de los clientes. Este enfoque ineficiente se refleja en el proceso actual de asignación de visitas domiciliarias, que representa una inversión significativa pero no se traduce consistentemente en recuperaciones efectivas.

### 1.1.1. Objetivos

#### Objetivo general

Mejorar el proceso de recuperación de cartera vencida automotriz mediante la aplicación y adaptación de algoritmos de aprendizaje automático consolidados. Utilizando modelos específicos como árboles de decisión, bosques aleatorios, máquinas de soporte vectorial (SVM) y métodos de boosting de gradiente, adaptados para abordar las características de cartera vencida. Este enfoque se basará en la identificación y evaluación de atributos clave de los deudores con mayor probabilidad de cumplir con sus pagos y la optimización de la asignación de recursos en las actividades de cobranza.

**Objetivos específicos**

Objetivo específico no 1. Realizar una clasificación en el conjunto de datos para identificar cuentas saldadas y pendientes, con el fin de que el modelo utilice esta información como referencia y extraiga las características más relevantes. Esto permitirá un análisis que facilitará la optimización de la gestión de visitas domiciliarias, enfocándose en las cuentas con mayor probabilidad de ser recuperadas.

Objetivo específico no 2. Identificar los atributos más importantes en casos de éxito de las cuentas con pago.

Objetivo Específico 3: Diseñar un plan para optimizar la asignación de visitas domiciliarias con base a los resultados del modelo de aprendizaje automático con mejor desempeño.

Objetivo específico no 4. Evaluar el desempeño de los modelos de aprendizaje automático mediante métricas adecuadas para determinar la confiabilidad de éstos.

## 1.2. Límites y alcances

### 1.2.1. Límites

Los titulares de los créditos pueden tener diferentes perfiles,así como diversos motivos para no realizar la devolución de la unidad automotriz, lo que podría limitar la capacidad del modelo para predecir con precisión el comportamiento. Asimismo, la falta de datos completos, actualizados o precisos sobre las cuentas puede limitar la capacidad del modelo para identificar con precisión las características relevantes y clasificar las cuentas de manera efectiva.

Por otra parte, la implementación de una prueba piloto requiere de una autorización de recursos económicos y coordinación con las áreas involucradas para el seguimiento de las visitas, por lo cual no es posible implementar el modelo sin previa autorización por parte del área Directiva de la empresa.

De igual manera, la temporalidad puede afectar al rendimiento del modelo puesto que los patrones de comportamiento de los deudores pueden cambiar con el tiempo debido a factores económicos, legales o sociales.

### 1.2.2. Alcances

El proyecto se centrará en el desarrollo de un modelo de aprendizaje automático que apunte a alcanzar un equilibrio en la precisión, dentro del rango del 82 % al 85 %. Este modelo estará específicamente diseñado para optimizar la recuperación de unidades automtrices de carteras vencidas, mediante la implementación estratégica de visitas domiciliarias.

## 1.3. Justificación

El uso de modelos de aprendizaje automático diferentes sectores industriales resulta en numerosos beneficios. Estos incluyen la optimización de procesos (Flores de Valgas Williams et al., 2023; Santillán Veliz, 2022), lo cual mejora la eficiencia operativa, la reducción de costos operativos al identificar y eliminar ineficiencias, y el aumento de las ganancias al maximizar oportunidades y minimizar riesgos (Cabanillas Romero, 2022; Quintero Acuña, 2023). En este contexto, la información se convierte en un recurso crítico. Su análisis y aplicación adecuados son esenciales para tomar decisiones bien informadas y estratégicas que respalden estos beneficios.

En el dinámico entorno empresarial actual, la adopción de tecnologías tales como el aprendizaje automático se ha vuelto crucial para mantener la competitividad y eficiencia (Francés Monedero, 2020). Para una empresa enfocada en la recuperación de carteras vencidas, la integración de algoritmos de aprendizaje automático presenta una oportunidad para gestionar de manera más rápida y adecuada sus operaciones, especialmente en la recuperación de unidades automotrices y la optimización de visitas domiciliarias.

Con base a lo anterior, la presente investigación se considera factible, ya que se cuenta

con la autorización de la dirección, los datos necesarios y los recursos humanos y tecnológicos requeridos para llevar a cabo los procesos requeridos para el desarrollo del modelo predictivo.

La incorporación de un modelo predictivo para la selección de créditos con mayor probabilidad de recuperación beneficiará directamente a la dirección, al aumentar el número de ingresos obtenidos y reducir el tiempo en que los créditos son liquidados. Esto permitirá un mejor control sobre la cartera en la cual la empresa enfoca sus recursos, sin depender exclusivamente del conocimiento, la antigüedad o la intuición del personal operativo, como se ha hecho hasta ahora en la selección y gestión de visitas domiciliares para la cobranza y recuperación de vehículos, optimizando así el proceso actual. Este enfoque puede replicarse en la selección de créditos hipotecarios, tarjetas de crédito y préstamos personales, los cuales también son productos que ha adquirido la empresa en otras cuentas. De esta manera, la asignación de cartera para el personal operativo se basará en la probabilidad de recuperación mediante los resultados del modelo predictivo.

## 1.4. Contribución

Este trabajo contribuye a la mejora de procesos del área de sistemas, donde desempeño un papel clave, al introducir nuevas herramientas basadas en aprendizaje automático que buscan optimizar los resultados en los procesos de recuperación de cartera vencida automotriz. No obstante, la contribución esencial reside en una mejora significativa y estratégica para el área operativa de la empresa, especialmente en lo que respecta a la gestión de cartera vencida de vehículos. Mediante el análisis y aprovechamiento de los datos históricos acumulados mediante la implementación de modelos de aprendizaje automático. Esta integración tecnológica se propone para ofrecer facilitar la toma de decisiones estratégicas en la administración del crédito y en la programación eficiente de visitas domiciliarias.

Se propone el uso de modelos de aprendizaje automático en el proyecto con el objetivo de mejorar la eficiencia y reducir los costos asociados con las visitas domiciliarias, las

cuales a menudo resultan en pérdidas económicas para la empresa. Mediante la mejora de la selección de visitas, fundamentada en atributos significativos y/o patrones predictivos, se busca mejorar la tasa de éxito en la recuperación de vehículos.

# Capítulo 2

# Marco Teórico

# 2 Capítulo 2

## 2.1. Marco teórico

### 2.1.1. Administradora de Cartera

#### Cartera Vencida

La cartera vencida se refiere a aquellos créditos o préstamos que no han sido saldados en la fecha estipulada, de acuerdo con los términos y condiciones establecidos en el contrato. Este estado se alcanza después de diversos intentos de cobranza por parte de la entidad original. Subsecuentemente, los créditos ingresan a concursos mercantiles, donde son puestos a disposición de administradoras de cartera. Dichas entidades asumen la titularidad de los créditos, conforme a lo dispuesto en el articulo 27 bis de la ley que regula las sociedades de información crediticia (CNBV, 2002).

El proceso que se debe de llevar a cabo para notificar al cliente de acuerdo a la Ley para Regular las Sociedades de Información Crediticia publicado en el Diario de la Federación el 15 de enero del 2002 (CNBV, 2002), se describe en los siguientes puntos:

1. **Notificación de la Venta o Cesión**: El Usuario notifica al Cliente sobre la venta o cesión de la cartera de crédito de acuerdo con la legislación común.

2. **Información a las Sociedades de Información Crediticia**: Dentro de los veinte días hábiles siguientes a la notificación al Cliente, el Usuario informa a las Sociedades de Información Crediticia. Se proporciona información completa sobre el comprador o cesionario, incluyendo nombre, domicilio, Registro Federal de Contribuyentes y otros datos de identificación. Se menciona la fecha de la cesión o venta.

3. **Inclusión en Reportes de Crédito Especiales**: Las Sociedades de Información Crediticia incluyen una anotación en los Reportes de Crédito Especiales que in-

dica el nombre de la persona a la que se vendió o cedió alguno de los créditos.

4. **Actualización de Información**: La actualización de la información sobre los créditos vendidos o cedidos se realiza utilizando el mismo número asignado al crédito objeto de la venta o cesión.

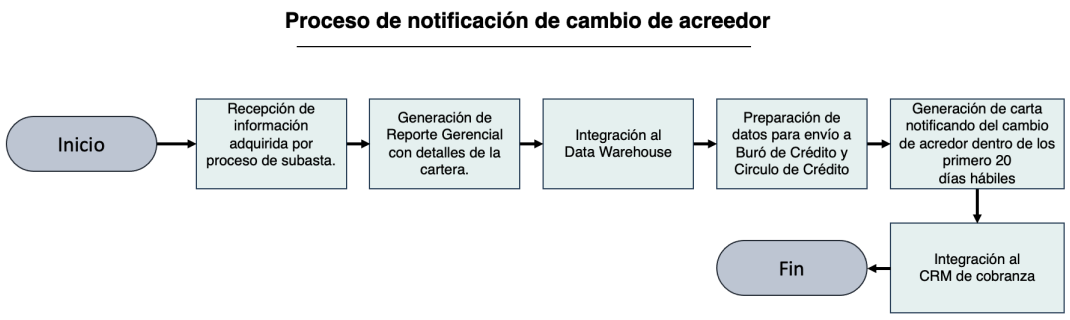
5. **Obligación del Comprador o Cesionario**: El comprador o cesionario de la cartera de crédito tiene la obligación de actualizar los registros de la Sociedad de Información Crediticia en relación con los créditos adquiridos.

6. **Eliminación de Información en Caso de Imposibilidad de Actualización**: En caso de que la venta o cesión se realice a personas que no sean Usuarios o si el vendedor o cedente deja de ser Usuario, las Sociedades de Información Crediticia incluyen una anotación que indica la imposibilidad de actualizar los registros. La información del crédito respectivo debe eliminarse del historial crediticio del Cliente en un plazo máximo de cuarenta y ocho meses.

7. **Atención de Reclamaciones**: La obligación de atender las reclamaciones de los Clientes en relación con la información crediticia recae en la persona responsable de actualizar la información del crédito vendido o cedido.

Internamente en el proceso que realiza la empresa tomando como base lo ya mencionado se muestra en la figura 2.1.

Figura 2.1: Diagrama. Proceso de notificación de cambio de acreedor




Fuente: Elaboración propia, 2023.

**Crédito Automotriz**

La financiación de vehículos se ofrece bajo dos modalidades principales:

- **Leasing**: Los créditos de tipo Leasing se refieren a la modalidad de pago por uso del automóvil. En este caso, el cliente realiza pagos de renta por un período definido, que generalmente oscila entre 24 y 36 meses. A diferencia de otros tipos de crédito, en el Leasing el cliente no adquiere la propiedad del automóvil, ya que solo paga por su uso durante el período acordado. Al finalizar el contrato, el cliente devuelve el automóvil a la entidad financiera. Una vez el crédito es vendido a una administradora de cartera, esta tiene el derecho de recuperar la unidad para realizar la venta.
- **Crédito**: Los créditos de tipo Credit ofrecen financiamiento por parte de una entidad crediticia para la compra de un automóvil nuevo o usado. El cliente realiza pagos mensuales durante un período que puede extenderse hasta 72 meses. A medida que se realizan los pagos, el cliente se convierte en propietario del automóvil. Esta modalidad es comúnmente utilizada por individuos que desean adquirir un vehículo para uso personal. En el cual, la administradora de cartera realiza un acuerdo de pago que puede ser finiquitar la cantidad adeudada o la devolución de la unidad para finiquitar el crédito, en cuyo caso será sometido a una revisión sobre el estado de la unidad y si este cubre el saldo pendiente (VW, 2023).

### 2.1.2. Gestión en la cartera automotriz

La empresa aplica las siguientes prácticas para la gestión de cobranza (Siappas, 2022):

1. Desarrollo de estrategias: Estas son evaluadas tomando como base el plazo de la deuda, los montos de los créditos y el número de créditos que tiene cada titular. Se puntúan para determinar cuáles serán los créditos gestionados en primera instancia.
2. Definición de KPIs: Haciendo uso de los datos disponibles en el sistema, se defi-

nen indicadores clave de rendimiento (KPIs) que pueden medirse posteriormente. Esto permite evaluar la cartera, la efectividad de las estrategias, análisis de costos, asignación y pagos.

3. Incentivos por recuperación: Estos se enfocan en determinar porcentajes de pago de comisiones por la cantidad recuperada o las unidades automotrices recuperadas durante el mes. Esto permite evaluar al personal identificando las habilidades de recuperación y ajustando la cantidad de créditos asignados.

4. Automatización de documentos: La generación de notificaciones de pago es sumamente importante, y dado que el número de solicitudes puede ser elevado, realizarlo de forma manual implica pérdida de tiempo y recursos.

5. Dar seguimiento: Es importante utilizar recordatorios de pago por medio de correo electrónico o mensaje de texto. Al establecer contacto con un cliente moroso, tener contacto en horarios específicos puede reforzar el sentido de seriedad y urgencia para el titular. Mantener contacto de esta manera resulta menos invasivo para ellos.

### 2.1.3. Aprendizaje automático: conceptos, métodos y métricas

#### Definición

Acorde con Samuel et al. (1959), "el aprendizaje automático es una rama de la inteligencia artificial que permite a los sistemas aprender y mejorar a partir de la experiencia". Éste se divide en dos categorías principales: aprendizaje supervisado, donde el modelo se entrena con datos etiquetados para predecir resultados (como clasificación y regresión), y aprendizaje no supervisado, que trabaja con datos sin etiquetar para descubrir patrones ocultos o agrupaciones (como en la detección de anomalías o la segmentación de clientes) (Helm et al., 2020).

En este estudio, se implementan algoritmos de aprendizaje supervisado, aplicados a un conjunto de datos procesado y clasificado previamente. Sin embargo, es importante reconocer que estos algoritmos presentan varios desafíos inherentes. Un aspecto crí-

tico es la dependencia de grandes volúmenes de datos etiquetados, cuya adquisición puede resultar en un proceso tanto costoso como complejo. Además, estos algoritmos son susceptibles al sobreajuste, es decir, pueden adaptarse excesivamente a los datos de entrenamiento, perdiendo la capacidad de generalizar a nuevos conjuntos de datos (Bilmes, 2020).

Otro problema de los algoritmos de aprendizaje automático es la selección de los atributos o característcas más importantes (features) (Cai et al., 2018) en problemas donde se considera que el número de atributos es alto (alta dimensionalidad). Este problema consiste en identificar, entre las variables disponibles, aquellas que están directamente relacionadas con la identificación de la clase (atributos relevantes). La cantidad y calidad de los atributos seleccionados tienen un impacto significativo en el rendimiento de los algoritmos utilizados.

**Métodos para selección de atributos**

Los conjuntos de datos pueden contener una mezcla de atributos malos y buenos. Los atributos malos son aquellas variables redundantes que generan lentitud e inexactitud en el desempeño del modelo. Las técnicas de selección de características permiten reducir la dimensionalidad de un conjunto de datos de manera que solo contenga atributos buenos que maximizen el rendimiento de los algoritmos y, por ende, posibiliten alcanzar una mayor precisión (Dash and Liu, 2000). Para la selección de atributos existen varios métodos de aprendizaje automático disponibles, que generalmente se clasifican en filtros (Hall, 1999), métodos envolventes (wrappers) (Kohavi and John, 1997) y métodos integrados (embedded) (Fu et al., 2009).

En este estudio, se implementó un enfoque de filtro para identificar las características más significativas, empleando tanto la correlación de Pearson como la importancia de variables derivada de un modelo Random Forest.

### Correlación de Pearson

El coeficiente de correlación de Pearson es una medida estadística utilizada para evaluar la fuerza y la dirección de la relación lineal entre dos variables. Desarrollado por Karl Pearson en 1895, este coeficiente se representa por el símbolo r y puede tomar valores entre -1 y 1. Un valor de 1 indica una relación lineal positiva perfecta, -1 una relación lineal negativa perfecta, y 0 indica que no hay una relación lineal entre las variables.

El coeficiente de correlación de Pearson se define matemáticamente como:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{2.1}$$

donde:

- $X_i$ y $Y_i$ son los valores individuales de las variables $X$ y $Y$.
- $\bar{X}$ y $\bar{Y}$ son los promedios de las variables $X$ y $Y$.
- La suma $\sum$ se realiza sobre todos los índices $i$ de las observaciones de las variables.

El numerador representa la covarianza de las dos variables, mientras que el denominador es el producto de las desviaciones estándar de X y Y. Esto hace que el coeficiente sea adimensional, facilitando su interpretación.(Hernández-Lalinde et al., 2018).

### Algoritmos de aprendizaje automático

### Arboles de decisión

Los árboles de decisión son modelos predictivos comúnmente utilizados para establecer sistemas de clasificación basados en múltiples covariables o para una variable objetivo. Son fáciles de interpretar, libres de ambigüedad y robustos incluso con valores faltantes. Tanto variables continuas como discretas se pueden emplear como variables objetivo o independientes(Song and Lu, 2015).

Un árbol de decisión se asemeja a un diagrama de flujo, donde los nodos internos representan atributos, las ramas indican decisiones y cada hoja representa un resultado. El nodo raíz es el punto de partida del árbol.

El árbol de decisión busca seleccionar el mejor atributo utilizando una medida de selección, convertir ese atributo en un nodo de decisión y dividir el conjunto de datos en subconjuntos más pequeños. Este proceso se repite recursivamente hasta que todas las instancias pertenecen a la misma categoría o no hay más atributos para dividir. Asimismo, en un árbol de decisión, medidas como la ganancia de información y el índice de Gini ayudan a dividir los datos al asignar puntuaciones a cada atributo (Gonzalez, 2019).

En este tipo de algoritmo no requiere un alto preprocesamiento de datos y es útil para la ingeniería de características y la selección de variables. No obstante, puede ser sensible al ruido y propenso al sobreajuste. Además, es importante equilibrar los conjuntos de datos para evitar sesgos (Rodrigo, 2020a).

Algunos de los párametros más importantes de los árboles de desición son los siguientes:

- **ccp_alpha**: Controla la poda basada en el costo de complejidad para prevenir el sobreajuste.
- **criterion**: Mide la calidad de una división; "gini"se usa para la impureza de Gini.
- **max_depth**: Profundidad máxima del árbol, limitando la complejidad del modelo para evitar el sobreajuste.
- **min_samples_leaf y min_samples_split**: Número mínimo de muestras requeridas para estar en un nodo hoja y para dividir un nodo, respectivamente, controlando el tamaño del árbol.

#### Random Forest

Random Forest introducido por Breiman y Adele Cutler (Breiman, 2001) es un algoritmo de aprendizaje supervisado que construye múltiples árboles de decisión durante

el entrenamiento y realiza predicciones mediante la agregación de los resultados de estos árboles. En este algoritmo, algunos de los parámetros más importantes son ntree y mtry. Ntree especifica la cantidad de árboles en el bosque, construyendo cada uno de manera independiente a partir de una muestra aleatoria del conjunto de datos. Aumentar el número de árboles puede incrementar la precisión, pero también eleva el costo computacional y puede alcanzar un punto donde los beneficios adicionales disminuyen. Por ello, es común realizar pruebas con diferentes cantidades de árboles para lograr un balance entre precisión y eficiencia. Por otro lado, mtry determina cuántas características se evalúan para dividir cada nodo en los árboles. En clasificación, generalmente es la raíz cuadrada del total de características, y en regresión, un tercio de éstas. Elegir correctamente el valor de mtry es fundamental, ya que un número bajo podría no captar todas las relaciones entre variables, mientras que uno alto podría causar sobreajuste en el modelo.

Otros párametros importantes son:

- **criterion**: 'entropy' se utiliza para la ganancia de información, determinando la calidad de las divisiones.
- **max_depth**: la profundidad máxima de los árboles, controlando la complejidad.
- **n_estimators**: Número de árboles en el bosque.
- **oob_score**: usa muestras fuera de bolsa para estimar la precisión general del modelo.

Este modelo también cuenta con una opción para obtener las variables más importantes. En este algoritmo la importancia de las variables se determina principalmente mediante dos métodos: la reducción del error y la permutación de características. Para la reducción del error, cada árbol se construye usando una muestra aleatoria del conjunto de datos, y en cada división, se selecciona un subconjunto aleatorio de características. El algoritmo evalúa la contribución de cada característica a la disminución del error de predicción, utilizando el índice Gini en clasificación o el error cuadrático medio en regresión. Por otro lado, la permutación de características mide el impacto en el error de predicción cuando los valores de una variable se alteran aleatoriamente; un

aumento significativo en el error indica la importancia de esa variable. La importancia se calcula para cada árbol individualmente y luego se promedia a través del bosque, proporcionando una medida robusta y generalizada de la relevancia de cada característica en el modelo.

#### Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vctorial (SVM) introducidas por (Cortes and Vapnik, 1995) buscan construir un hiperplano en un espacio multidimensional para separar diferentes clases, buscando iterativamente el hiperplano óptimo que minimice el error y maximice la separación entre clases.

Las SVM funcionan seleccionando vectores de soporte, que son los puntos de datos más cercanos al hiperplano. Estos vectores determinan la posición del hiperplano y el margen (la distancia entre las clases más cercanas). El objetivo es encontrar el hiperplano que maximice este margen, logrando la mejor separación posible entre las clases (Rodrigo, 2020c).

Las ventajas de las SVM incluyen su capacidad para manejar datos no lineales mediante el uso de funciones kernel y su efectividad en espacios de alta dimensión. No obstante, las SVM pueden ser sensibles al ruido en los datos y al sobreajuste si no se regula adecuadamente. Además, puede presentar sesgo en conjuntos de datos desequilibrados, por lo que es recomendable balancear los datos antes de su uso (Rodrigo, 2020c).

Algunos de los párametros importanes de las SVM se describen a continuación:

- **C**: parámetro de regularización, que controla el trade-off entre lograr un margen alto y clasificar todos los ejemplos de entrenamiento correctamente.
- **gamma**: parámetro de kernel para 'rbf', 'poly' y 'sigmoid'. Afecta la influencia de los puntos de datos individuales.
- **kernel**: tipo de función del kernel a utilizar en el algoritmo, aquí es 'linear', indicando un hiperplano lineal de separación.

### Gradient Boosting

Gradient Boosting fue desarrollado por Jerome H. Friedman (Friedman, 2002, 2001a) basándose en la idea de Leo Breiman(Breiman, 1997) . El modelo consiste en un conjunto de árboles de decisión individuales entrenados de forma secuencial, donde cada modelo busca corregir las debilidades de sus predecesores. La predicción de una nueva observación se logra combinando las predicciones de todos los árboles individuales, integrando efectivamente el algoritmo de descenso de gradiente y el método de boosting.

Las principales ventajas del modelo incluyen(IBM, 2020):

- **Facilidad de implementación**: Se puede utilizar con diversas configuraciones de hiperparámetros para mejorar el ajuste.
- **Reducción de sesgo**: Al combinar varios aprendices débiles de manera secuencial, se mejora interactivamente el modelo, reduciendo el sesgo presente en árboles de decisión poco profundos.
- **Eficiencia computacional**: Los algoritmos de boosting seleccionan características que aumentan su poder predictivo durante el entrenamiento, lo que ayuda a reducir la dimensionalidad y mejorar la eficiencia computacional.

Algunos desafíos potenciales incluyen:

- **Complejidad en la interpretación**: La combinación de múltiples árboles puede dificultar la interpretación en comparación con un solo árbol.
- **Pérdida de información con predictores continuos**: La categorización de predictores continuos durante la segmentación de nodos puede llevar a la pérdida de información.

Otros aspectos a considerar son el cálculo intensivo, ya que el entrenamiento secuencial puede ser difícil de escalar y lento. El modelo, al ser un conjunto (ensemble) y basado en boosting, combina múltiples modelos con el objetivo de equilibrar la desviación promedio de las predicciones y mejorar la precisión en diferentes datos de entrena-

miento. En boosting, se ajustan secuencialmente múltiples modelos sencillos, llamados '*Weak learners*', ya que cada modelo aprende del anterior(Rodrigo, 2020b).

A continuación se describen algunos de los párametros importantes de este modelo:

- **ccp_alpha**: parámetro de poda para controlar la complejidad y prevenir el sobreajuste.
- **learning_rate**: encoge la contribución de cada árbol para prevenir el sobreajuste.
- **n_estimators**: número de etapas de refuerzo para ejecutar, es decir, la cantidad de árboles en el modelo.
- **max_depth**: limita la profundidad de los árboles, afectando la complejidad del modelo.

#### Métricas de rendimiento para clasificación

#### Matriz de confusión

Una matriz de confusión es una tabla que muestra el rendimiento de un modelo predictivo. Esta tabla se deriva de la comparación del resultado de un modelo predictivo con los valores reales. Cada fila contiene los valores predichos por el modelo. Cada columna representa los valores reales.

Cuadro 2.1: Matriz de Confusión

| | **Predicción Positiva** | **Predicción Negativa** |
|---|---|---|
| **Condición Positiva** | VP | FN |
| **Condición Negativa** | FP | VN |

Fuente: Elaboración propia, 2023.

donde: $VP$ = Verdadero Positivo, $VN$ = Verdadero Negativo, $FP$ = Falso Positivo y $FN$ = Falso Negativo.

### Precisión

La precisión mide la proporción de verdaderos positivos sobre el total de positivos predichos. Indica cuán confiables son las predicciones positivas del modelo(Hastie et al., 2009).

$$\text{Precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{2.2}$$

### Precisión balanceada

La precisión balanceada es una mejor estimación del rendimiento de un clasificador cuando existe una distribución desigual de dos clases en un conjunto de datos.

$$\text{Precisión Balanceada} = \frac{1}{2}\left(\frac{\text{VP}}{\text{VP} + \text{FN}} + \frac{\text{VN}}{\text{VN} + \text{FP}}\right) \tag{2.3}$$

### Sensibilidad y Especificidad

En este trabajo se usó también la sensibilidad y la especificidad, las cuales son métricas para evaluar modelos de aprendizaje automático. La sensibilidad mide la capacidad del modelo para identificar correctamente las instancias positivas, mientras que la especificidad evalúa su capacidad para identificar correctamente las instancias negativas (Hastie et al., 2009). La sensibilidad se calcula como:

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.4}$$

y la especificidad se determina como:

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{2.5}$$

**Curva de ROC**

Una curva ROC (Característica Operativa del Receptor) mide el rendimiento de un clasificador basado en qué tan bien separa el grupo que está siendo evaluado en aquellos que pertenecen a una clase y a otra. Una curva ROC es una representación gráfica de la Tasa de Verdaderos Positivos (TPR) en el eje $y$, y la Tasa de Falsos Positivos (FPR) en el eje $x$.

$$\text{TPR} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.6}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{VN}} \tag{2.7}$$

El rendimiento se mide mediante el AUC. El AUC de una curva ROC varía entre [0,1], donde 1 representa una clasificación perfecta.

### 2.1.4. Trabajos relacionados

Algúnos de los trabajos que se relacionan con el objetivo del presente trabajo, se describen a continuación para ejemplificar como han sido abordados.

En este estudio (Nazemi et al., 2022) se evaluaron diversas técnicas de aprendizaje automático, como deep learning, boosting y regresión de vectores de soporte, para predecir la tasa de recuperación de más de 65,000 créditos de consumidores en mora del sector de las telecomunicaciones, adquiridos por una empresa alemana de terceros. Se definieron medidas de rendimiento ponderadas basadas en el valor de la exposición al incumplimiento para comparar modelos de tasa de recuperación. El enfoque propuesto resulta útil para que una empresa de terceros gestione el riesgo de su cartera de créditos en mora. La principal conclusión es que, de todos los métodos evaluados, el modelo de deep learning supera significativamente a los demás en términos de medidas de rendimiento ponderadas fuera de muestra.

La capacidad de prever las tasas de recuperación en deudas en mora se vuelve crucial dada la creciente deuda del consumidor a nivel mundial. El aumento en la deuda del consumidor, tanto en Europa como en los Estados Unidos, destaca la importancia de contar con modelos de tasa de recuperación confiables. Este estudio se enfoca en la industria de las telecomunicaciones y destaca que el deep learning podría ser una herramienta potencialmente mejoradora del rendimiento en la gestión del riesgo crediticio.

En este estudio (Sefik Ilkin Serengil and Koroglu, 2022), se llevó a cabo una comparación de varios algoritmos de aprendizaje automático para predecir la probabilidad de incumplimiento crediticio (NPL) en un conjunto de datos de carteras de clientes de un banco privado en Turquía. Se abordó el desafío de desequilibrio de clases y se utilizaron métricas de rendimiento como Precisión, Recuperación, Puntuación F1, Exactitud del desequilibrio, y Especificidad para evaluar los modelos.

Entre los algoritmos evaluados, LightGBM destacó como el mejor según estas métricas. Además, se abordó la interpretabilidad de los modelos mediante herramientas de Inteligencia Artificial Explicable (XAI) como SHAP y LIME.

Se identificaron características importantes para la predicción de NPL, como el historial de pagos, el comportamiento de las tarjetas de crédito, el riesgo, la antigüedad del cliente, entre otros. El estudio sugiere que estos modelos pueden ser útiles para prever el incumplimiento crediticio, con LightGBM como el algoritmo más eficaz en este contexto. Sin embargo, se enfatiza la importancia de la transparencia algorítmica, y se exploró la explicabilidad del modelo para una mejor comprensión de sus resultados.

En este estudio (Sefik Ilkin Serengil and Koroglu, 2022), se centra en la utilización de minería de datos para predecir préstamos no rendidos (NPL) en la industria bancaria, reconociendo la inevitabilidad de préstamos problemáticos y su impacto en la reducción del capital bancario. Se destaca la importancia de cuidar adecuadamente a los deudores con dificultades de pago. La investigación utiliza el historial de pago de los deudores para prever préstamos problemáticos, empleando técnicas de minería de datos. Se comparan varios algoritmos, destacando Random Forest con la mayor precisión

del 96.55 %. El estudio se basa en la metodología CRISP-DM y se realiza en cinco pasos, desde la comprensión del negocio hasta la evaluación. Se detallan herramientas como MS Excel y Rapid Miner, y se aborda el desafío de prever la calidad crediticia de los deudores. Además, se propone un método que involucra la identificación, selección y transformación de datos. Se concluye que la minería de datos, especialmente con Random Forest, es eficaz para predecir deudores problemáticos, y se sugiere la implementación en sistemas bancarios centrales para mejorar la gestión crediticia.

Los trabajos relacionados presentan enfoques diversos para abordar el objetivo del presente trabajo. Se destaca la aplicación de técnicas avanzadas de aprendizaje automático, como deep learning y algoritmos como LightGBM, para predecir la tasa de recuperación y la probabilidad de incumplimiento crediticio en diversas industrias, incluyendo las telecomunicaciones y la banca. Estos estudios subrayan la importancia de contar con modelos confiables y precisos.

# Capítulo 3

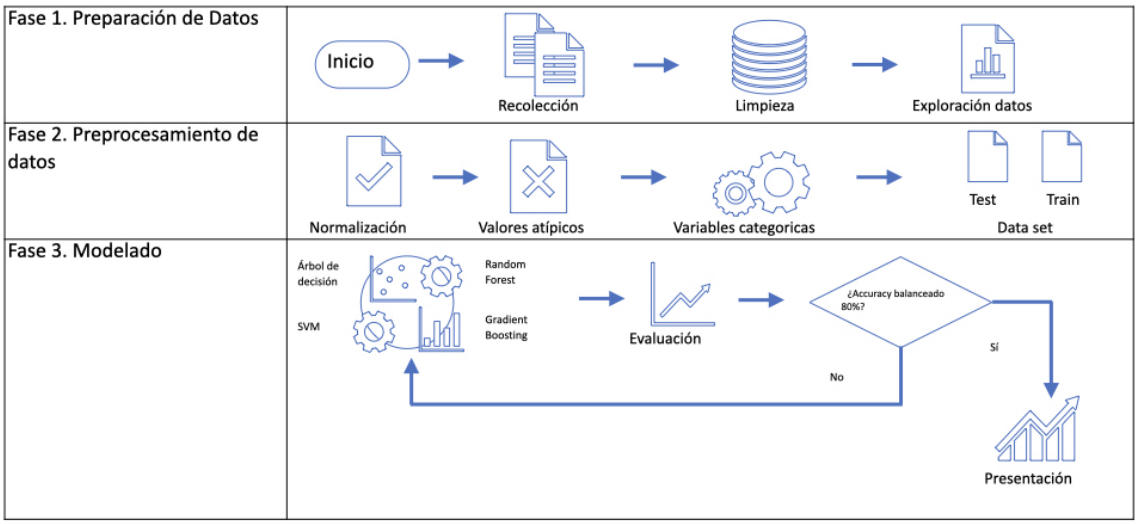
# Modelo a emplear

# 3 Capítulo 3

## 3.1. Marco metodológico

En esta sección se describe el marco metodológico, el cual se define bajo un esquema de trabajo que comprende la preparación de datos, el preprocesamiento de datos, la construcción del modelo, la evaluación y la presentación de propuestas, como se ilustra en la figura 3.1

Figura 3.1: Esquema de trabajo




Fuente: Elaboración propia, 2023.

### 3.1.1. Preparación de Datos

#### Segmentación de Cartera vencida

Una vez que la información ha sido proporcionada por el área Jurídica, esta es segmentada de la siguiente manera.

- **Créditos área Operativa**: Se asigna la cartera que no cuenta con los siguientes es-

tatus de calificación emitidos por la institución que originó el crédito: Demanda, Robo de identidad, fraude, Siniestro. Posteriormente, los créditos que no cuentan con dichas calificaciones son asignados al área operativa con el objetivo de gestionar el crédito para lograr su liquidación. El equipo de ejecutivos de cobranza se encarga de este portafolio durante un mes. Durante este período, se lleva a cabo una investigación que incluye la revisión de expedientes físicos, digitales y llamadas, utilizando la información proporcionada por la entidad que realizó la venta. Después de este período, los créditos se retiran y se asignan a otro gestor para que realice el mismo proceso.

- **Créditos área de Jurídico**: Los créditos que no se han asignado al área operativa pasan al área de Jurídico, en el cual son gestionados por la vía legal.

#### Créditos área Operativa

Los créditos cuentan con información que ha sido proporcionada por la entidad que originó el crédito, la cual indica la situación de endeudamiento actual del crédito. A continuación, se describen las fuentes de datos utilizadas:

- Creación de base de datos: Se crea una estructura de base de datos que incluye tablas y la configuración necesaria para el análisis histórico.
- Base de Inventario: Esta base de datos integra información sobre cada crédito, incluyendo datos personales que se utilizarán como variables, sin revelar información sensible. También incluye información sobre saldos pendientes y detalles del automóvil, entre otros.
- Base de Notificaciones: En esta base de datos se registran todas las notificaciones realizadas en cada ruta de visitas realizada. Contiene información como la localidad a la que se desplazaron los gestores domiciliarios.
- Base de Cobranza: Esta base de datos describe la fecha, el monto y el tipo de pago que los titulares han realizado.
- Base de Sepomex: Se utiliza esta base de datos de acceso público para revisar los

estados y direcciones que se incluyen en las notificaciones.

- Base de índice de marginación de CONAPO: Esta base de datos de acceso público, correspondiente al año 2022, proporciona información sobre el índice de marginación a nivel de municipio para todos los estados de la República.
- Base de Viaticos: Se utiliza un catálogo generado en la base de datos para identificar el costo promedio del traslado desde la Ciudad de México al destino en el cual se realizarán las visitas.

Estas fuentes de datos son utilizadas para analizar y mejorar el proceso de cobranza de unidades automotrices.

**Crédito Automotriz**

Para este trabajo, se usaran solo los créditos de tipo Credit que cuenten con automóvil a recuperar. Como se describe en el capitulo 2.

### 3.1.2. Preprocesamiento de Datos

En esta fase, se utiliza el conjunto de datos que se asigna al area operativa. Estos datos se utilizan para llevar a cabo un análisis y una estandarización de variables en un entorno de Notebook de Jupyter.

Durante este análisis exploratorio de datos, se examinan las diferentes variables y se realizan los procedimientos necesarios para garantizar su calidad y coherencia. Esto puede incluir la limpieza de datos, la identificación y manejo de valores atípicos, la imputación de datos faltantes y la normalización de variables si es necesario.

Una vez que los datos han sido preparados, se procede a la selección de variables que serán utilizadas en los modelos que se emplearán. Esta selección implica identificar las variables más relevantes y significativas para el análisis y descartar aquellas que no aportan información relevante o pueden introducir sesgos o ruido en el modelo.

### Variables Categóricas

Desde esta etapa, es crucial asegurar que la información sea compatible con los diversos modelos de estimación propuestos.

Se integro la cateogría de Catálogo de Saldos(SaldosCat) y Region, esto con el fin de agrupar la cantidad adeudada por los titulares y la ubicacion geografica en al que se encuentran en el pais de México. Estas a su vez pasaran a ser convertidas a variables *dummy* como se puede observar en la tabla 3.1 y 3.2

Cuadro 3.1: Catálogo de Regiones

| **Región** |
|---|
| Centrosur |
| Oriente |
| Noroeste |
| Oeste |
| Sureste |
| Centronorte |
| Noreste |

Fuente: Elaboración propia, 2023.

Cuadro 3.2: Catálogo de Saldos

| **SaldosCat** |
|---|
| 1. <= 50,000 |
| 2. 50,001 - 100,000 |
| 3. 100,001 - 150,000 |
| 4. 150,001 - 200,000 |
| 5. 200,001 - 250,000 |
| 6. 250,001 - 300,000 |
| 7. 300,001 - 350,000 |
| 8. 350,001 - 400,000 |
| 9. 400,001 - 500,000 |
| 10. >= 500,000 |

Fuente: Elaboración propia, 2023.

### Tratamiento de datos faltantes

Para el conjunto de datos, solo se encontraron tres variables, d_tipo_asenta, GM_2020, IMN_2020 de los cuales se imputaron los datos usando la técnica de valor constante, con el fín de que todos los valores que se encuentren en el set de datos cuenten con valor como se puede observar en la tabla 3.3.

### Estandarización de datos

StandardScaler es un método de normalización que se utiliza para estandarizar las características de los datos eliminando la media y escalando a varianza unitaria. Este pro-

Cuadro 3.3: Variables con datos faltantes

| **Variable** | **Valores Faltante** | **Valor Imputado** |
|---|---|---|
| d_tipo_asenta | 18 | COLONIA |
| GM_2020 | 18 | MUY BAJO |
| IMN_2020 | 18 | media |

Fuente: Elaboración propia, 2023.

ceso se realiza característica por característica. La fórmula utilizada para estandarizar los datos es la siguiente:

$$z = \frac{(x - u)}{s} \tag{3.1}$$

Donde:

- z: Es la puntuación estandarizada.
- x: Es el valor original de la característica.
- u: Es la media de la característica.
- s: Es la desviación estándar de la característica.

Este proceso asegura que cada característica tenga una media de 0 y una desviación estándar de 1, asegurando que los datos tengan una distribución normal y en escalas similares.

**Entrenamiento y Prueba**

La validación de modelos de aprendizaje automático es crucial para asegurar su efectividad. Dividir el conjunto de datos en subconjuntos de entrenamiento y prueba permite evaluar el rendimiento del modelo en un entorno controlado (Hastie et al., 2009).

En este trabajo se dividió el conjunto de datos original en dos subconjuntos independientes: el conjunto de entrenamiento y el conjunto de prueba. El conjunto de entrenamiento se utilizó para ajustar el modelo, mientras que el conjunto de prueba fue usado para evaluar el rendimiento de éste. En este estudio, se asignaron 2/3 de los datos al conjunto de entrenamiento y 1/3 al conjunto de prueba, siguiendo las recomendaciones de autores como Hastie 2009 y Friedman 2001b, entre otros.

Posteriormente se utilizó una validación cruzada de 10 iteraciones, la cual divide el conjunto de datos completo original en k nuevos conjuntos de datos independientes (en este caso 10). Luego, realiza k bucles en los que k - 1 particiones del conjunto de datos original se utilizan para el entrenamiento y el resto para la prueba. Para cada pliegue, se calcula y suma la medida de evaluación del modelo obtenida a partir de la matriz de confusión. Cuando todos los k bucles han terminado, se obtiene la precisión de la validación cruzada.

Mediante la implemantación del proceso anterior en las ejecuciones de los algoritmos también fue posible obtener los párametros más adecuados para cada modelo.

La tabla 3.4 muestra un conjunto de datos dividido en tres categorías: Base completa, Base entrenamiento y Base prueba. Para cada conjunto, se proporcionan las observaciones sin pago y su porcentaje correspondiente, así como las observaciones con pago y su respectivo porcentaje. La Base completa cuenta con 1,921 observaciones sin pago (70 %) y 578 con pago (30 %). La Base de entrenamiento incluye 940 observaciones sin pago (70 %) y 404 con pago (30 %), sumando 1,344 observaciones. Finalmente, la Base de prueba tiene 403 observaciones sin pago (70 %) y 174 con pago (30 %), con un total de 577 observaciones.

La conservación de la proporción de las clases entre los conjuntos de entrenamiento y prueba es crucial para garantizar que el modelo se entrene y evalúe en condiciones representativas del conjunto de datos completo.

Cuadro 3.4: División conjunto de datos

| **Datos** | **Obs. sin Pago** | **% sin Pago** | **Obs. con Pago** | **% con Pago** | **Total Obs.** |
|---|---|---|---|---|---|
| Base completa | 1,343 | 70 % | 578 | 30 % | 1,921 |
| Base entrenamiento | 940 | 70 % | 404 | 30 % | 1,344 |
| Base prueba | 403 | 70 % | 174 | 30 % | 577 |

Fuente: Elaboración propia, 2023.

**Base entrenamiento Smote**

Muestra el efecto de aplicar SMOTE (Synthetic Minority Over-sampling Technique) al conjunto de entrenamiento. SMOTE es una técnica de sobremuestreo que crea observaciones sintéticas de la clase minoritaria para equilibrar las clases. Después de aplicar

SMOTE, la cantidad de observaciones sin pago quedó en 940 y sin pago quedó en 940, llevando el total a 1,880 observaciones.

La inclusión de SMOTE en el proceso de preparación de datos mejora el rendimiento de los modelos de clasificación en conjuntos de datos desequilibrados, al permitir que el modelo aprenda de manera más efectiva las características de la clase minoritaria, como se puede observar en el cuadro 3.5

Cuadro 3.5: División conjunto de datos Smote

| **Datos** | **Obs. sin Pago** | **% sin Pago** | **Obs. con Pago** | **% con Pago** | **Total Obs.** |
|---|---|---|---|---|---|
| Base entrenamiento Smote | 940 | 50 % | 940 | 50 % | 1,880 |

Fuente: Elaboración propia, 2023.

### 3.1.3. Modelado

En este trabajo, se emplearon algoritmos de clasificación como Árboles de decisión, Random Forest, Support Vector Machine y Gradient Boosting, evaluados repetidamente para asegurar la fiabilidad de los modelos. Las métricas seleccionadas para la evaluación fueron la precisión balanceada, sensibilidad y especificidad, las cuales se consideran adecadas para trabajar con datasets desbalanceados (Alarcón-Narváez et al., 2021). Asimismo, a pesar de que el uso de SMOTE equilibró el dataset, la elección de estas métricas fue pertinente. La precisión balanceada es crucial porque, incluso con un dataset equilibrado artificialmente, ofrece una visión imparcial del rendimiento del modelo en todas las clases, evitando el sesgo hacia la clase mayoritaria que la precisión general podría inducir.

De igual manera, se incluyeron la sensibilidad y la especificidad, ya que proporcionan una visión clara de la capacidad del modelo para identificar correctamente cada clase. La sensibilidad mejora típicamente para la clase minoritaria post-SMOTE, un aspecto vital para evitar la negligencia de esta clase. Sin embargo, la especificidad debe ser monitoreada cuidadosamente debido al posible aumento de falsos positivos generado por los datos sintéticos de SMOTE.

Finalmente, se incorporó la curva ROC como una medida complementaria de rendimiento, debido a su capacidad para evaluar la discriminación del modelo independientemente del desbalance de las clases. Esta combinación de métricas asegura una evaluación completa del modelo, validando su efectividad y capacidad de generalización en condiciones equitativas.

# Capítulo 4

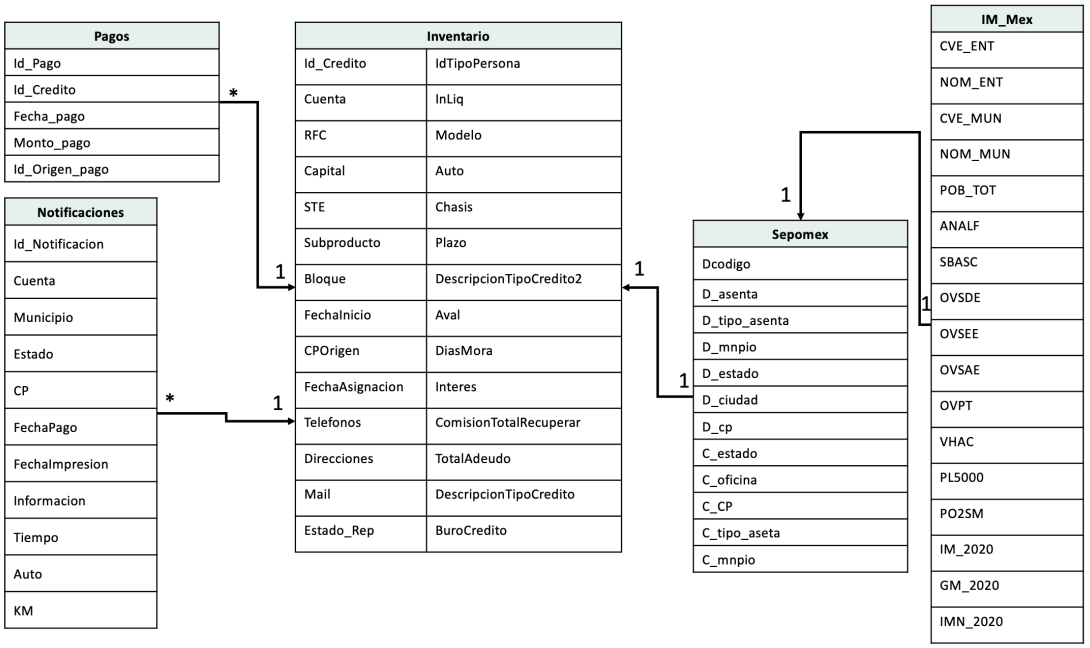
# Desarrollo de modelo a evaluar

# 4 Capítulo 4

## 4.1. Definición de Variables

La consolidación de las tablas de Inventario, Notificaciones, Pagos, Sepomex y IM Mex ha resultado en la obtención del conjunto de datos que se utilizará para el desarrollo del proyecto. Estas tablas se han transformado en MySQL para obtener el archivo final que será empleado en el proyecto, se puede observar la estructura de las tablas mostrado en la figura 4.1.

Figura 4.1: Diagrama. Gestión Cartera Automotriz




Fuente: Elaboración propia, 2023.

## 4.2. Análisis Exploratorio

Las variables obtenidas de la unión de las tablas, mostradas en la Figura 4.1, se han sometido a un proceso de limpieza de datos para asegurar su correcta manipulación

y análisis. Este proceso incluyó la eliminación de espacios en blanco al inicio y al final de los campos, así como de caracteres no deseados que podrían interferir con el procesamiento de los datos, tales como los caracteres de nueva línea, las tabulaciones, los retornos de carro y las comas. Estos elementos son comunes en los datos y pueden causar problemas en etapas de análisis al ser interpretados erróneamente como delimitadores o en formatos numéricos. Este meticuloso proceso de limpieza asegura que los datos estén estandarizados y libres de anomalías que puedan afectar las fases subsiguientes del análisis.

Variables como Cuenta, NoCreditos, y TotalSTE (suma total pendiente de pago) son fundamentales para el análisis relacionado con créditos. Estas variables detallan la cantidad de créditos y los montos asociados, así como aspectos demográficos y geográficos del titular del crédito como Estado_Rep, Edad, Genero, y Region. Esto permite una comprensión más profunda del perfil del cliente y su comportamiento de pago.

Variables como AgnosA (año de registro), AN2021 (notificaciones en 2021), AN2022, y AN2023 ofrecen una base para analizar tendencias y patrones a lo largo del tiempo de las notificaciones de pago realizadas.

Se detalla información específica del vehículo como el modelo (Vento, Jetta, etc.) y el estado de uso (Nuevo, Seminuevo, Usado), lo que es crucial para valorar el riesgo y la depreciación de los activos vinculados a los créditos.

La inclusión de variables operacionales como KM (kilómetros para cálculo de viáticos) y Tiempo (tiempo estimado de traslado) optimiza los procesos de logística en realización de rutas.

Variables como TotalAdeudo, Intereses y Pago (indicador de si el crédito ha sido liquidado) son esenciales para medir la salud financiera del portafolio y la efectividad de las estrategias de cobranza.

Variables como GM_2020 e IMN_2020, relacionadas con índices de marginación, ayudan a categorizar y normalizar los datos en términos de contexto socioeconómico, lo cual es vital para análisis regionales diferenciados y focalizados

Los resultados de la agrupación de la información realizada en la base de datos se pueden observar en la tabla 4.1 con un total de 1,921 registros (al 31 de julio de 2023).

Cuadro 4.1: Variables dataset

| **Variable** | **Descripción** |
|---|---|
| Cuenta | Identificador único del titular de uno o más créditos automotrices. |
| NoCreditos | Cantidad de créditos automotrices asociados al titular. |
| TotalSTE | Suma total pendiente de pago para saldar todos los créditos. |
| Agnos | Duración del crédito en años. |
| AgnosA | Año en que la cuenta fue registrada en el sistema. |
| MesA | Mes en que la cuenta fue registrada en el sistema. |
| DiaA | Día en que la cuenta fue registrada en el sistema. |
| Tel | Cantidad de créditos asociados a un número telefónico. |
| Dir_Trabajo | Indica si el titular ha proporcionado una dirección de trabajo. |
| Dir | Cantidad de créditos asociados a una dirección postal. |
| Mail | Cantidad de créditos asociados a un correo electrónico. |
| <=M2015 | Vehículos modelo 2015 o anteriores. |
| M2016 | Vehículos modelo año 2016. |
| M2017 | Vehículos modelo año 2017. |
| M2018 | Vehículos modelo año 2018. |
| M2019 | Vehículos modelo año 2019. |
| M2020 | Vehículos modelo año 2020. |
| >=M2021 | Vehículos modelo 2021 o posteriores. |
| Vento | Modelo de automóvil Vento. |
| Jetta | Modelo de automóvil Jetta. |
| Gol | Modelo de automóvil Gol. |
| Ibiza | Modelo de automóvil Ibiza. |
| Polo | Modelo de automóvil Polo. |
| Toledo | Modelo de automóvil Toledo. |
| Saveiro | Modelo de automóvil Saveiro. |

| | |
|---|---|
| Clasico | Modelo de automóvil Clásico. |
| Tiguan | Modelo de automóvil Tiguan. |
| Otros | Vehículos no clasificados por tipo. |
| Estado_Rep | Estado de la República del titular. |
| FNDia | Día de nacimiento del titular. |
| FNMes | Mes de nacimiento del titular. |
| FNAgno | Año de nacimiento del titular. |
| Edad | Edad actual del titular. |
| D_tipo_asenta | Clasificación del asentamiento según el catálogo de SEPOMEX. |
| Region | Región donde se ubica el estado del titular. |
| Genero | Género del titular. |
| Ocupacion | Profesión u ocupación registrada por el titular. |
| GM_2020 | Categoría dentro del índice de Marginación del 2020. |
| IMN_2020 | Índice de Marginación Normalizado del 2020. |
| SaldosCat | Clasificación de saldos según el total adeudado. |
| NoMunicipios | Cantidad de municipios notificados. |
| NoEstados | Cantidad de estados notificados. |
| NoCPs | Cantidad de códigos postales notificados. |
| NoFechasPagos | Cantidad de fechas de pago registradas. |
| NoCorrectas | Cantidad de direcciones correctas registradas en notificaciones. |
| NoIncorrectas | Cantidad de direcciones incorrectas registradas en notificaciones. |
| Visitas | Número de visitas realizadas al titular. |
| AN2021 | Notificaciones realizadas en el año 2021. |
| AN2022 | Notificaciones realizadas en el año 2022. |
| AN2023 | Notificaciones realizadas en el año 2023. |
| DM | Días de Mora hasta el momento de la venta. |
| Intereses | Intereses generados por concepto de cobranza. |
| CTR | Cuenta de Tasa Reducida por renegociación. |
| TotalAdeudo | Monto total adeudado hasta el momento de la venta del crédito. |

| | |
|---|---|
| Nuevo | Vehículos catalogados como nuevos. |
| Seminuevo | Vehículos catalogados como seminuevos. |
| Usado | Vehículos catalogados como usados. |
| OtrosE | Estado de uso del vehículo no especificado. |
| C_estado | Clave de estado según el catálogo de SEPOMEX. |
| C_cve_ciudad | Clave de ciudad según el catálogo de SEPOMEX. |
| C_muni | Clave de municipio según el catálogo de SEPOMEX. |
| KM | Kilómetros considerados para el cálculo de viáticos. |
| Tiempo | Tiempo estimado de traslado al lugar de visita. |
| Auto | Costo promedio de peajes por la ruta carretera utilizada. |
| Pago | Indicador de si el crédito ha sido liquidado o no. |

Fuente: Elaboración propia, 2023.

**Distribución geografica**

La región *Centrosur* tiene la mayor cantidad de deudas pendientes con 918 registros, lo que indica una concentración significativa de deudas automotrices en esta área. Le sigue la región *Oriente* con 257 registros y *Oeste* con 176 registros, lo que muestra una cantidad moderada de deudas pendientes. Las regiones *Centronorte*, *Noreste* y *Sureste* tienen una cantidad similar de registros, con un poco más de 150 cada una. Por otro lado, las regiones *Noroeste* tienen las cantidades más bajas de registros, con 83 mostrado en la figura 4.2.

En total, hay 1,921 registros, y la distribución por región puede indicar patrones económicos en diferentes partes del país.

Figura 4.2: Distribución Geografica por Region en México


Fuente: Elaboración propia, 2023.

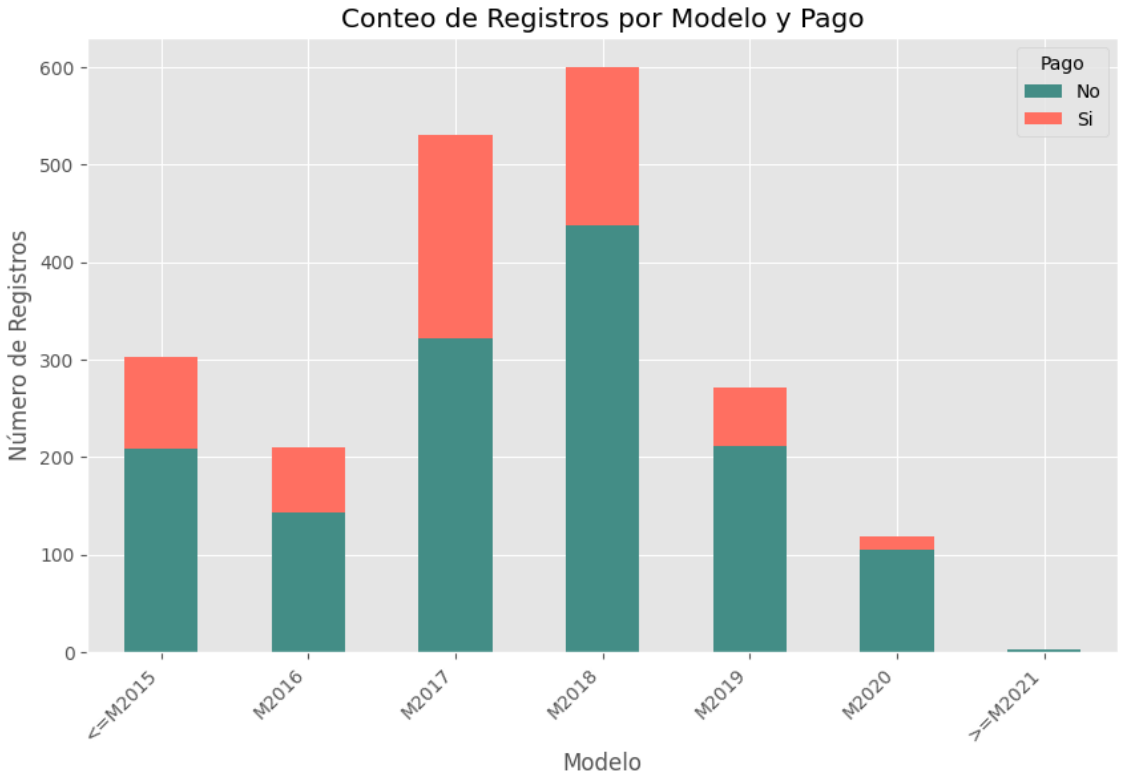
**Modelo y Sub Marca**

Como se muestrada en la figura 4.3, De los modelos más antiguos a los más recientes, los modelos de los años 2017 y 2018 tienen el mayor número de registros en total. Además, en todos los años, hay más registros de pagos no realizados (verde) que de pagos realizados (rojo). Esto es particularmente pronunciado en los modelos 2017 y 2018, donde la proporción de pagos no realizados es notablemente mayor que la de pagos realizados.

En el extremo más reciente del espectro, para los modelos del año 2021 y posteriores, hay una cantidad relativamente pequeña de registros.

El gráfico sugiere una tendencia en la que los propietarios de modelos de automóviles más antiguos están más al día con sus pagos que los propietarios de modelos más recientes.

Figura 4.3: Registro por Modelo y pago

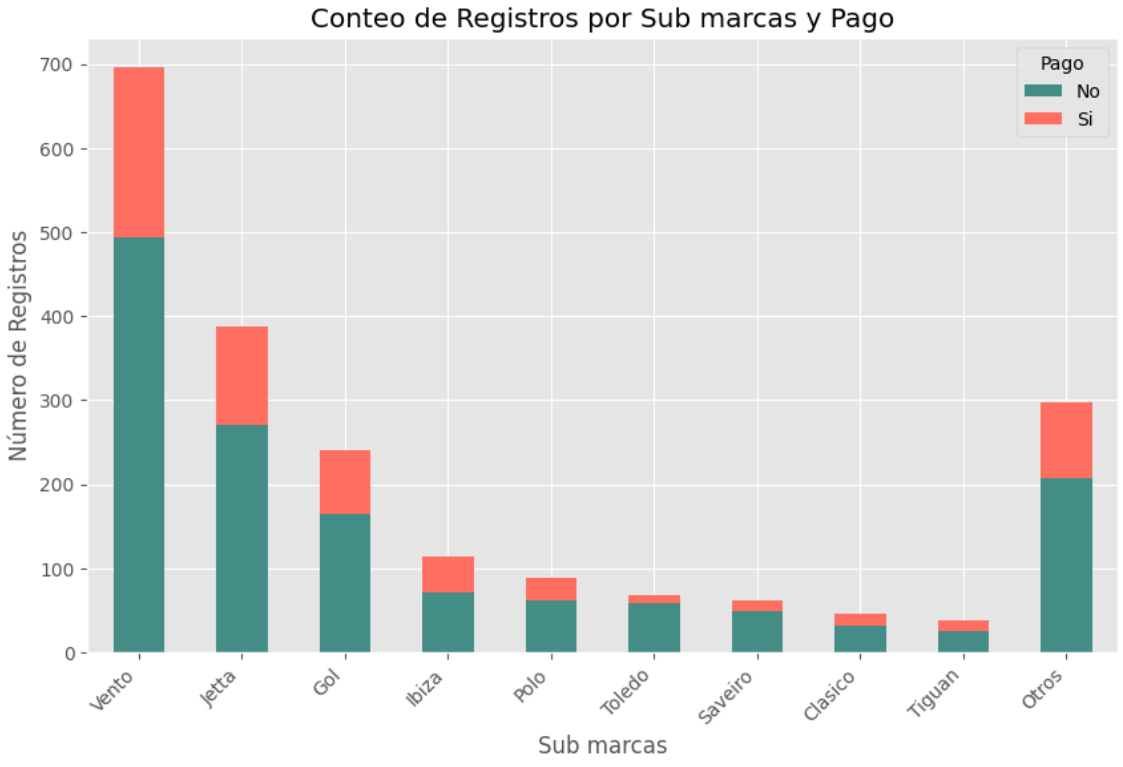



Fuente: Elaboración propia, 2023.

Mientras que la Figura 4.4 indica que la submarca *Vento, Jetta* y *Gol* tiene la mayor cantidad de registros sin pagos realizados. Las submarcas como *Ibiza, Polo, Toledo, Saveiro, Clásico y Tiguan* tienen menos registros en comparación, pero mantienen una proporción similar de pagos no realizados sobre los realizados.

Este gráfico podría sugerir que ciertas submarcas tienen tasas de incumplimiento más altas que otras. Por ejemplo, los propietarios de *Vento* pueden tener dificultades para realizar pagos más que los propietarios de otras marcas. Sin embargo, también es posible que *Vento* sea simplemente una submarca más popular y por lo tanto tenga más créditos otorgados.

Figura 4.4: Registro por SubMarca y pago




Fuente: Elaboración propia, 2023.

### 4.2.1. Análisis de Nulos

Se procede a analizar la cantidad de datos faltantes en las columnas, considerando la posibilidad de realizar imputaciones según la naturaleza y la información contenida en cada campo. A continuación, se presentan los resultados de los datos faltantes, como se observar en la tabla 4.4. Para las columnas d_tipo asenta y GM_2020, se opta por la imputación de los datos faltantes basándose en las categorías que se repiten con mayor frecuencia. En el caso de IMN_2020, se realiza el cálculo de la media para llenar los registros faltantes.

Cuadro 4.2: Variables con datos nulos

| **Variable** | **Observaciones** | **Valores Nulos** | **Porcentaje Nulos** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| IMN_2020 | 1,921 | 18 | 0.93 % |
| GM_2020 | 1,921 | 18 | 0.93 % |
| d_tipo_asenta | 1,921 | 18 | 0.93 % |

Fuente: Elaboración propia, 2023.

### 4.2.2. Variables Númericas

A continuación se integra una descripción de algúnas de las variables númericas con las que cuenta el dataset en el que se puede observar en la figura 4.5

- **Creditos**: En promedio, los titulares tienen aproximadamente 1 crédito cada uno, con la mayoría de ellos teniendo exactamente 1, como se refleja en la mediana.
- **TotalSTE** : La cantidad total que se debe por créditos tiene un promedio bastante alto y una desviación estándar grande, lo que indica una variación significativa en la cantidad que se debe entre los titulares de créditos. La diferencia entre el mínimo y el máximo es muy amplia, lo que sugiere que hay algunos titulares que deben cantidades extremadamente altas.
- **Agnos y AgnoA**: Estos se refieren al plazo de los créditos y al año de adquisición, respectivamente. El promedio de años es alrededor de 4, y el año promedio de adquisición es 2020, lo que indica que el conjunto de datos es bastante reciente.
- **Tel**: La mayoría de los titulares tienen al menos un teléfono asociado, lo cual es común en la era actual de la comunicación.
- **Modelos de Automóviles (Vento, Jetta, Gol, etc.)**: Hay una distribución variada de los modelos de automóviles. Algunos, como Vento y Jetta, parecen ser más comunes, mientras que otros son menos frecuentes.
- **Edad**: La edad promedio de los titulares es de aproximadamente 43 años, y la distribución de edades parece ser bastante normal, con un rango que va de jóvenes adultos a personas mayores.

- **IMN_2020**: El índice de marginación estandarizado promedio es bajo, lo que podría indicar que el conjunto de datos incluye a individuos de áreas con un nivel de vida relativamente alto.
- **TotalAdeudo**: Similar a TotalSTE, hay una amplia variabilidad en el monto total adeudado, con un promedio elevado.
- **Estado del Vehículo (Nuevo, Seminuevo, Usado)**: La mayoría de los vehículos son nuevos o seminuevos, lo que podría indicar una preferencia o capacidad de los titulares de crédito para adquirir vehículos más nuevos.
- **KM y Tiempo**: Estas variables tienen promedios y medianas altos, sugiriendo que podrían estar relacionadas con el cálculo de costos asociados al uso del automóvil, como el pago de viáticos para visitas de cobranza o seguimiento.

Figura 4.5: Muestra de Variables Númericas

| | count | mean | median | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|
| **NoCreditos** | 1921.0 | 1.060385 | 1.000000 | 0.303620 | 1.00000 | 1.000000 | 1.000000 | 1.000000 | 8.000000e+00 |
| **TotalSTE** | 1921.0 | 157434.124904 | 134649.880000 | 139015.543357 | 624.96000 | 67353.450000 | 134649.880000 | 212805.400000 | 1.675079e+06 |
| **Agnos** | 1921.0 | 3.980219 | 4.000000 | 0.981933 | 0.00000 | 3.000000 | 4.000000 | 5.000000 | 7.000000e+00 |
| **AgnoA** | 1921.0 | 2020.361791 | 2020.000000 | 0.839008 | 2019.00000 | 2020.000000 | 2020.000000 | 2021.000000 | 2.022000e+03 |
| **MesA** | 1921.0 | 7.282665 | 7.000000 | 3.584509 | 1.00000 | 4.000000 | 7.000000 | 11.000000 | 1.200000e+01 |
| **DiaA** | 1921.0 | 7.779282 | 3.000000 | 9.757220 | 1.00000 | 2.000000 | 3.000000 | 10.000000 | 3.100000e+01 |
| **Tel** | 1921.0 | 1.060906 | 1.000000 | 0.306079 | 0.00000 | 1.000000 | 1.000000 | 1.000000 | 8.000000e+00 |
| **Dir_Trabajo** | 1921.0 | 0.071317 | 0.000000 | 0.261436 | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 2.000000e+00 |
| **Dir** | 1921.0 | 1.060385 | 1.000000 | 0.307032 | 0.00000 | 1.000000 | 1.000000 | 1.000000 | 8.000000e+00 |
| **Mail** | 1921.0 | 0.930765 | 1.000000 | 0.466476 | 0.00000 | 1.000000 | 1.000000 | 1.000000 | 8.000000e+00 |
| **<=M2015** | 1921.0 | 0.157730 | 0.000000 | 0.386765 | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 4.000000e+00 |
| **M2016** | 1921.0 | 0.109318 | 0.000000 | 0.321976 | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 3.000000e+00 |
| **M2017** | 1921.0 | 0.276419 | 0.000000 | 0.456562 | 0.00000 | 0.000000 | 0.000000 | 1.000000 | 2.000000e+00 |
| **M2018** | 1921.0 | 0.312337 | 0.000000 | 0.482289 | 0.00000 | 0.000000 | 0.000000 | 1.000000 | 3.000000e+00 |
| **M2019** | 1921.0 | 0.141072 | 0.000000 | 0.359955 | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 3.000000e+00 |
| **M2020** | 1921.0 | 0.061947 | 0.000000 | 0.263814 | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 5.000000e+00 |
| **>=M2021** | 1921.0 | 0.001562 | 0.000000 | 0.039498 | 0.00000 | 0.000000 | 0.000000 | 0.000000 | 1.000000e+00 |

Fuente. Elaboración propia.

### 4.2.3. Variables Categóricas

A continuación se muestra una descripción de las variables categóricas con las que cuenta el dataset como se puede mostrar en la figura 4.6. La mayoría de los titulares posee solamente un crédito, al menos con un teléfono asociado. Sin embargo, la mayoría no tiene una dirección de trabajo, lo que sugiere posiblemente un sector de trabajadores independientes o una falta de requerimiento de información por parte de la entidad que originó el crédito. Los registros indican que la Ciudad de México (Distrito Federal) es la región con un mayor número de registros, y la mayoría de los titulares residen en zonas clasificadas como *Colonia* dentro de la región *Centrosur*, predominando el género masculino y siendo *Empleado* la ocupación más común. A pesar de una marginación catalogada como *Muy Bajo*, un número significativo de titulares tiene saldos altos, de 50,000 o más. Sin embargo, más de la mitad no ha liquidado su crédito, lo que podría reflejar desafíos en la capacidad de pago o en las estrategias de cobranza que la empresa ha aplicado.

Figura 4.6: Variables Categóricas

| | NoCreditos | Tel | Dir | Mail | Estado_Rep | d_tipo_asenta | region | genero | ocupacion | GM_2020 | SaldosCat | Pago |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 1921 | 1921 | 1921 | 1921 | 1921 | 1903 | 1921 | 1921 | 1921 | 1903 | 1921 | 1921 |
| **unique** | 5 | 6 | 6 | 6 | 32 | 15 | 7 | 2 | 12 | 5 | 10 | 2 |
| **top** | 1 | 1 | 1 | 1 | Distrito Federal | COLONIA | Centrosur | H | EMPLEADO | MUY BAJO | 3. 100,001 - 150,000 | No |
| **freq** | 1823 | 1820 | 1819 | 1580 | 485 | 1155 | 918 | 1236 | 855 | 1708 | 372 | 1343 |

Fuente. Elaboración propia.

### 4.2.4. Generación de insights

El 30 % de los titulares han realizado un pago como se puede observar en la tabla 4.3, de los cuales se analiza el perfil que se tiene hasta ahora mostrados en la figura 4.7 de la imagen proporcionan una visión integral sobre el perfil de los titulares de créditos que han efectuado pagos. Se destaca que la gran mayoría tiene un solo crédito y prácticamente todos contando con teléfono y correo electrónico. La dirección de trabajo no es común entre ellos. En cuanto a la distribución geográfica, Ciudad de México (Distrito Federal) domina en número de créditos pagados, y la mayoría de los pagos provienen

de áreas clasificadas como *Colonia*, lo cual sugiere que los sectores urbanos son clave en la actividad de pago. La región *Centrosur* sobresale como la más activa en pagos, lo que podría reflejar una economía regional más fuerte. Los hombres constituyen la mayoría de los que han realizado pagos, y el empleo formal aparece como la ocupación predominante, lo cual podría correlacionar estabilidad laboral con la capacidad de pago. Además, la categoría de marginación *Muy Bajo* es la más representada, sugiriendo que incluso en áreas con menos desventajas socioeconómicas, los desafíos para completar los pagos persisten. Los saldos altos categorizados como *1. >= 50,000* son los más frecuentes, indicando que los titulares con mayores deudas están cumpliendo con sus obligaciones de pago.

Cuadro 4.3: Créditos con pago

| **Pago** | **Porcentaje** |
|---|---|
| No | 69.91 % |
| Si | 30.09 % |

Fuente: Elaboración propia, 2023

Figura 4.7: Variables Categoricas de Titulares que han pagado

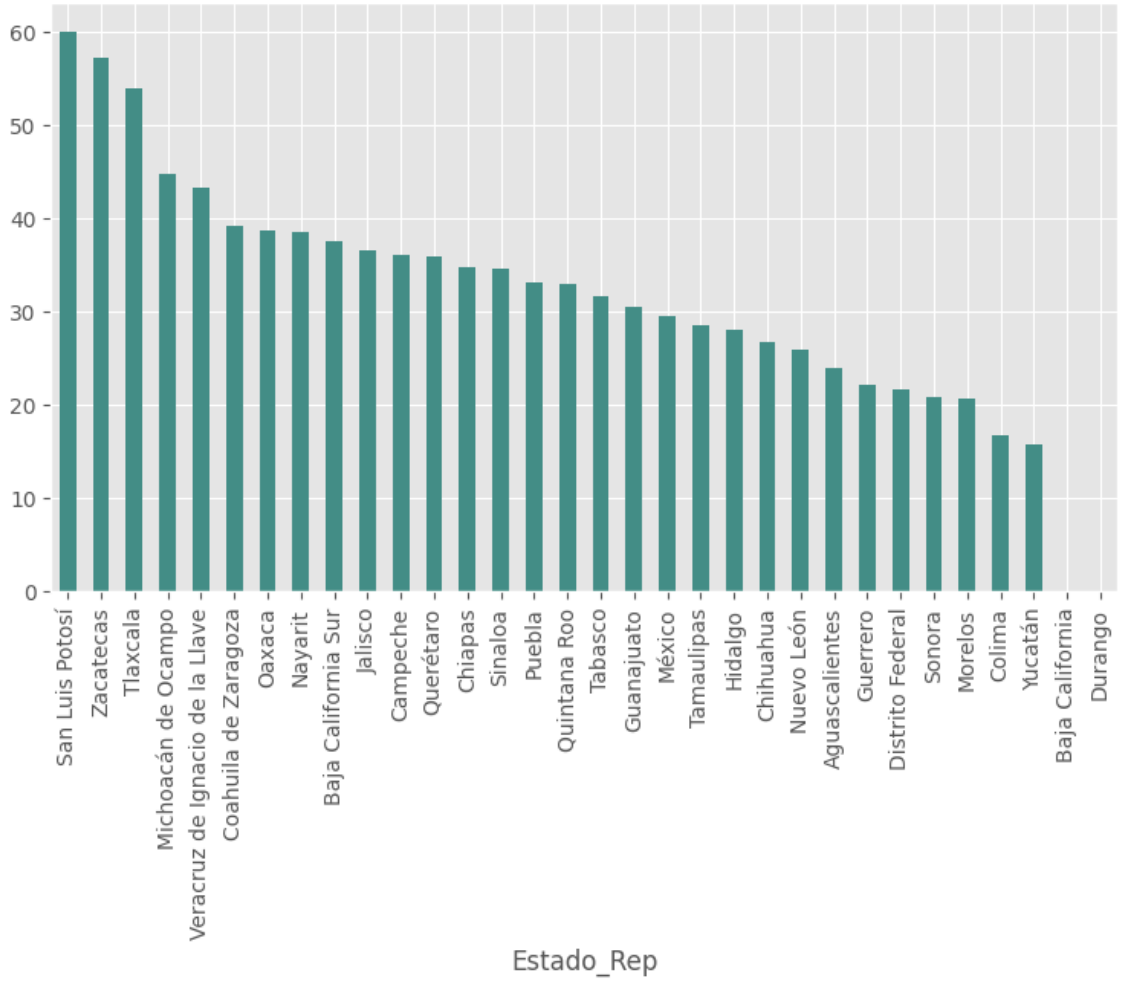

Fuente. Elaboración propia.

A continuación se muestra el cálculo de la media de pagos realizados en créditos, agrupada por las siguientes características: *Estado de la República*, *Categoría de saldos*, *Saldos* y *Edad*, como se muestra en la figura 4.8. Este cálculo representa la proporción de créditos automotrices que han sido saldados en cada estado de México, expresada en porcentaje.

Se observa que San Luis Potosí tiene el porcentaje más alto de pagos realizados, seguido por Zacatecas y Michoacán de Ocampo. En contraste, estados como Sonora, Yu-

catán y Baja California presentan los porcentajes más bajos de pagos realizados. Esto indica una variabilidad significativa en el comportamiento de pago entre los estados, lo cual podría estar influenciado por factores económicos regionales, diferencias culturales o políticas de crédito específicas de cada estado.

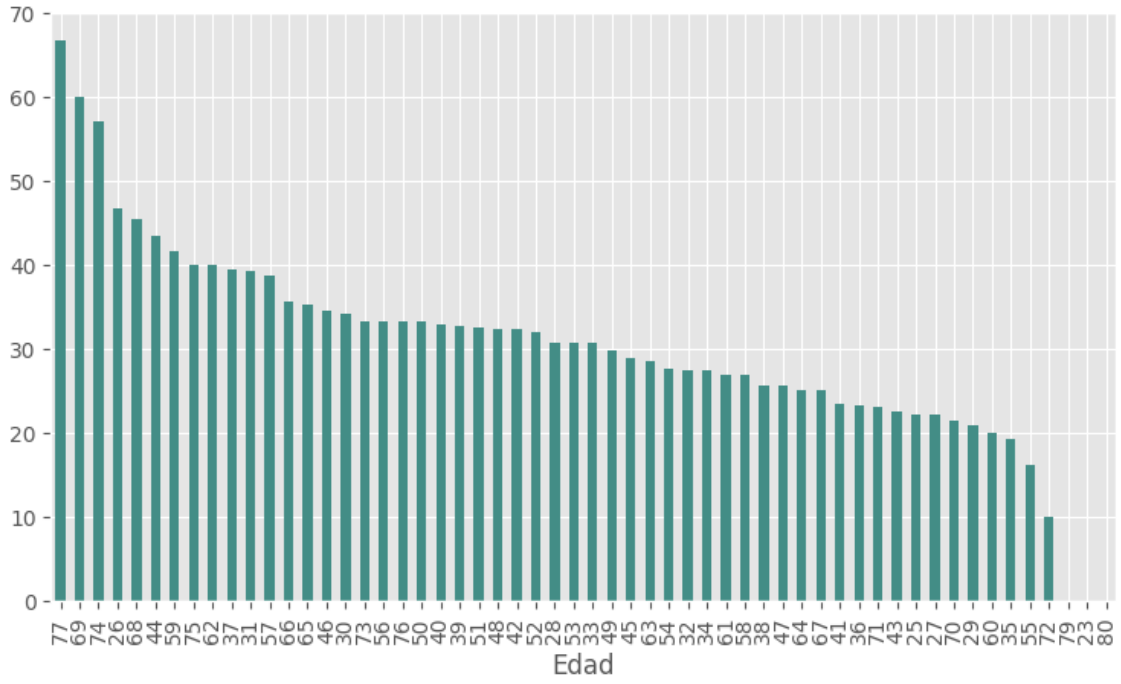
Figura 4.8: Pagos promedio por Estado de la República




Fuente. Elaboración propia.

La Figura 4.9 muestra que las personas más jóvenes tienden a estar al corriente con sus pagos de autos más a menudo que las personas de mayor edad. Las barras más grandes en el gráfico señalan que un número más alto de jóvenes ha pagado sus deudas. A medida que la gente envejece, el gráfico muestra que menos han terminado de pagar sus créditos de coches.

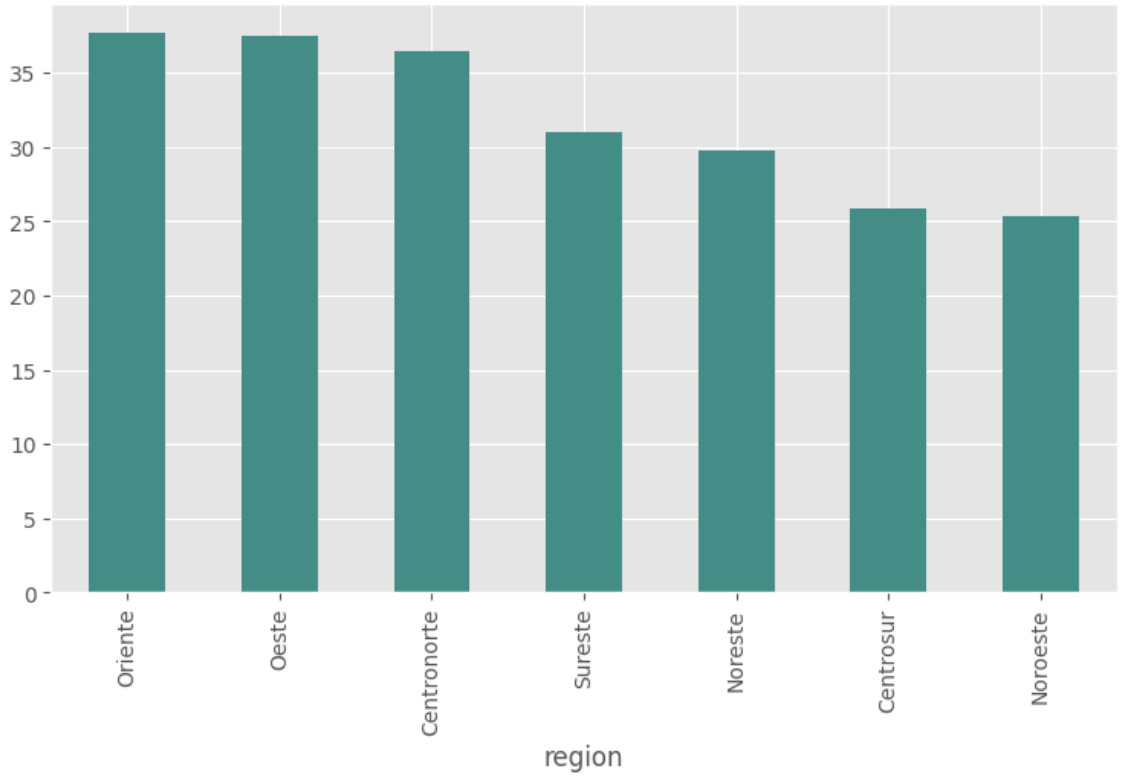
Figura 4.9: Pagos promedios por Edad




Fuente. Elaboración propia.

La Figura 4.10 muestra que las regiones con el porcentaje más altas corresponden a *Oriente* y *Oeste*, tienen porcentajes más altos de pagos realizados, mientras que regiones como *Centrosur y Noroeste* tienen porcentajes más bajos.

Figura 4.10: Pagos promedios por Región

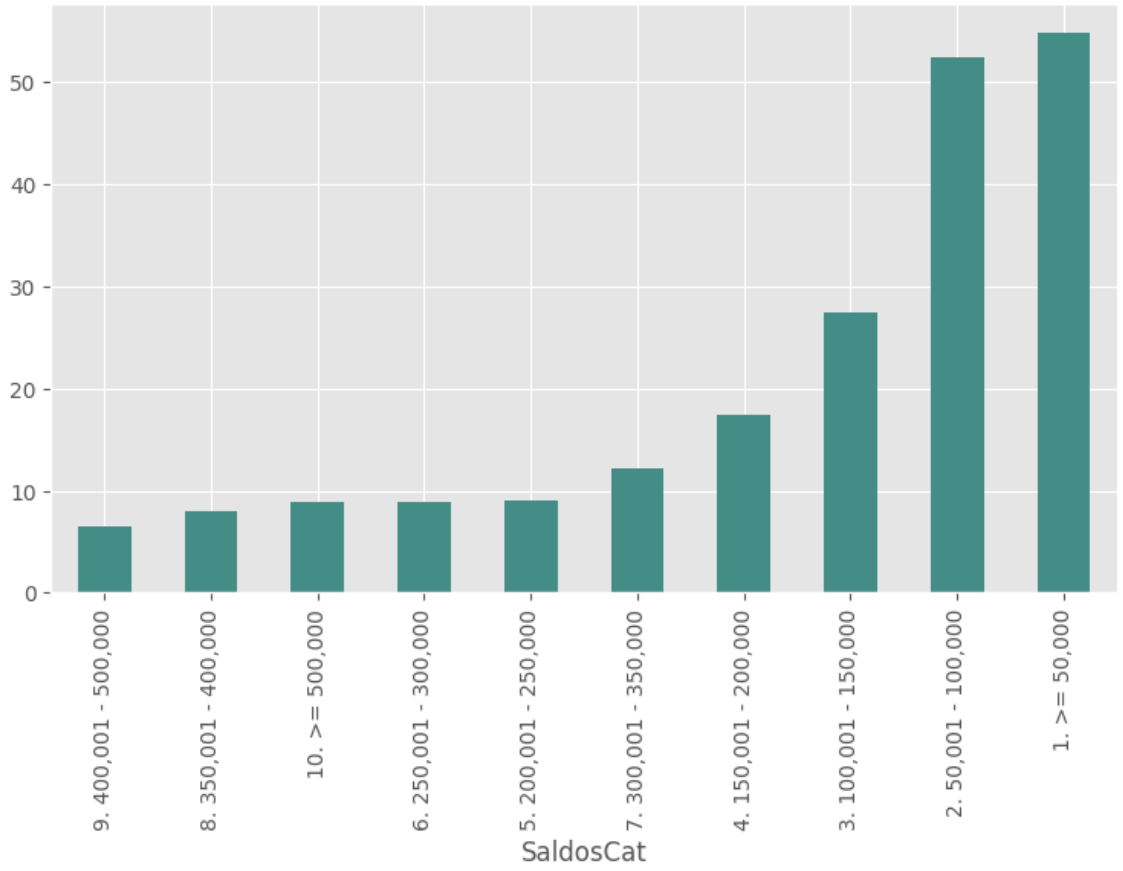



Fuente. Elaboración propia.

La Figura 4.11 ilustra los porcentajes de pago por cada categoría de saldo en la columna *SaldosCat*. Se observa que a menor cantidad de deuda, menor es el porcentaje de pagos realizados, mientras que las categorías con deudas más altas muestran un mayor porcentaje de pagos efectuados. Esto sugiere que los individuos con mayores deudas pueden tener una mayor propensión a cumplir con sus pagos o quizás tengan más recursos para hacerlo.

Es notable esta tendencia, ya que intuitivamente se esperaría que las deudas menores fuesen más fáciles de pagar; no obstante, la información presentada indica una realidad diferente.

Figura 4.11: Pagos promedios por Saldos




Fuente. Elaboración propia.

### 4.2.5. Resultados

Esta sección describe los resultados derivados de la implementación del análisis de correlación y la aplicación de algoritmos de aprendizaje automático.

#### Análisis de Correlación

La Figura 4.12 muestra una matriz de correlación para las variables del conjunto de datos. A partir de esta figura, se puede notar que las variables con mayor correlación se distinguen por valores próximos a 1 o $-1$, así como por tonalidades más saturadas en la escala de colores. Las variables que exhiben una correlación más fuerte son aquellas con colores más oscuros de rojo, lo que indica una correlación positiva, y las que tienen

colores más oscuros de azul para una correlación negativa.

- **TotalSTE y TotalAdeudo**: Tienen una correlación muy alta de 0.95, lo que sugiere que a medida que uno aumenta, el otro también tiende a aumentar. Esto tiene sentido si ambas variables están relacionadas con la cantidad que se debe en créditos.

- **Tiempo y KM**: Tambien tienen una alta correlación de 0.94, indicando que mayores distancias están asociadas con tiempos de viaje más largos, lo cual es lógico.

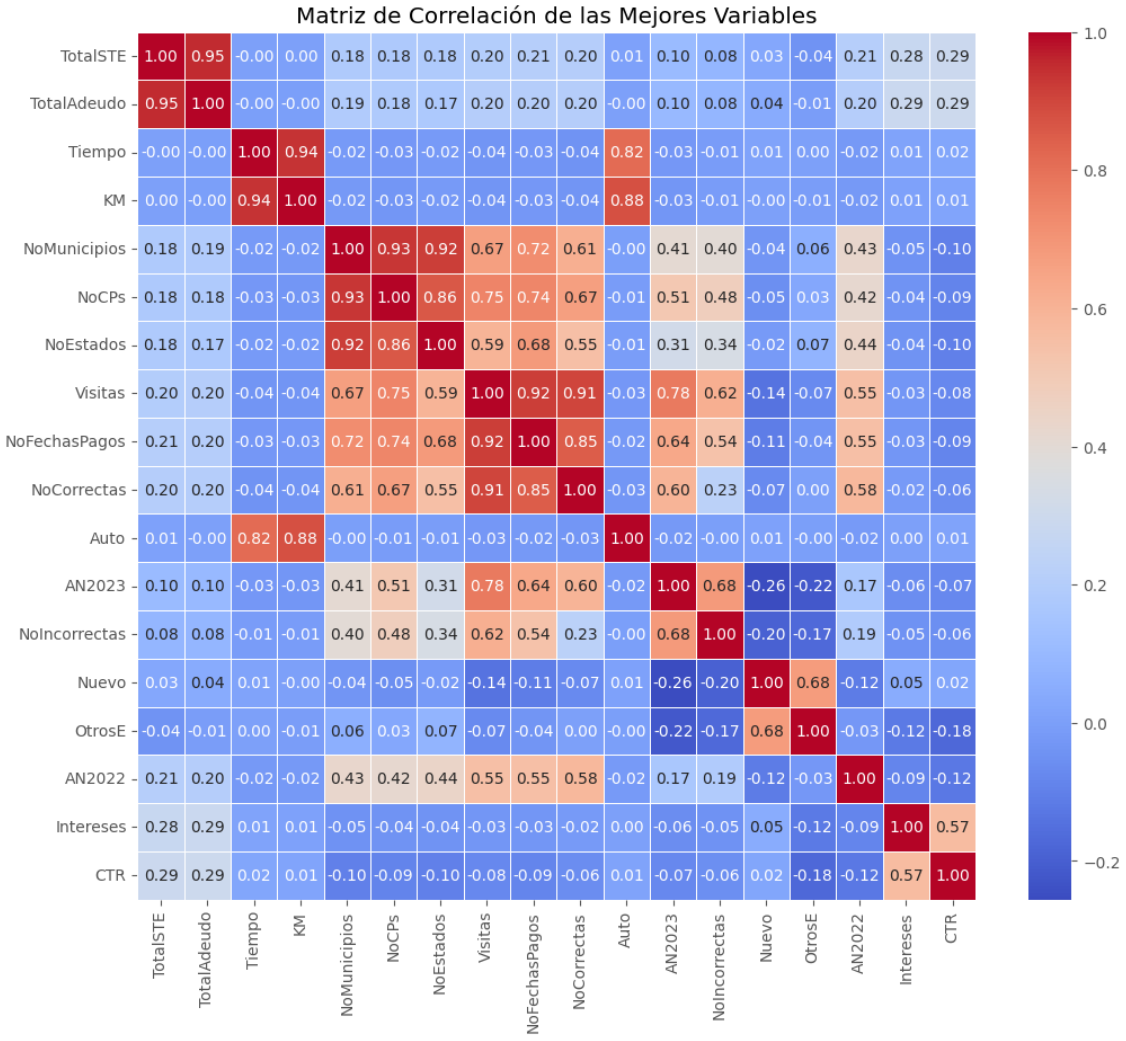
- **NoMunicipios, NoCPs, NoEstados, Visitas, NoFechasPagos y NoCorrectas**: Muestran correlaciones moderadas a altas entre sí, lo que sugiere que cuando hay más notificaciones en un área, hay una tendencia a que haya más fechas de pago, visitas y direcciones correctas asociadas.

- **AN2023**: Tiene correlaciones moderadas con Visitas, NoFechasPagos y NoCorrectas, lo que puede implicar que en el año 2023 se realizaron esfuerzos significativos de cobranza o seguimiento de los créditos.

- **AN2022, Intereses y CTR**: Tienen correlaciones bajas con la mayoría de las otras variables, lo que indica que estos factores no varían sistemáticamente con las características de los créditos o los esfuerzos de cobranza en este conjunto de datos.

  Las correlaciones altas pueden indicar redundancia de información entre las variables, mientras que las correlaciones bajas sugieren que las variables proporcionan información única.

Figura 4.12: Matriz de Correlación




Fuente: Elaboración propia, 2023.

Con base a lo anterior, la Figura 4.12 destaca dos pares de variables por su fuerte relación lineal positiva: TotalSITE y Totaldeudas con un coeficiente de 0.95, y NoMunicipios y NCPs con un coeficiente de 0.93.

### Modelos de aprendizaje automático

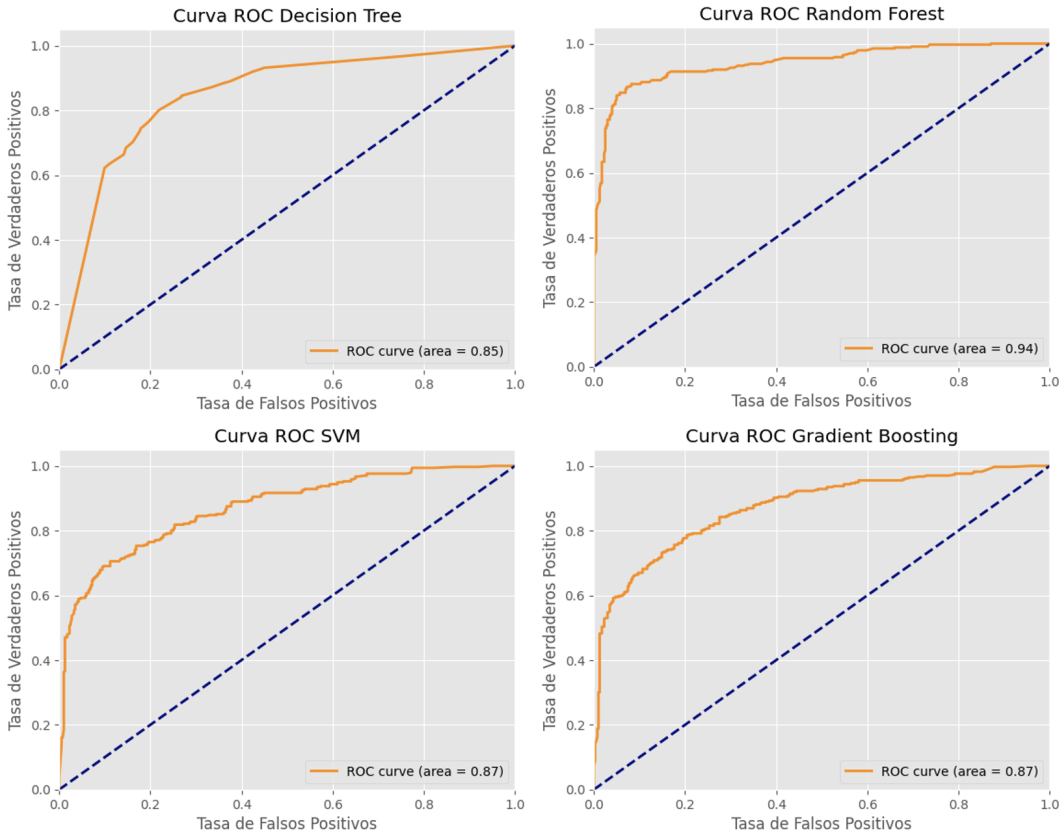
En cuanto a los resultados obtenidos por los modelos de aprendizaje automático aplicando SMOTE (ver Tabla 4.4) se destaca que el modelo que presentó el mejor desempeño fue Random Forest, logrando una precisión balanceada del 88%. Este modelo

también mostró una alta especificidad del 87 % y la mayor sensibilidad de los modelos evaluados con un 88 %. Los parámetros óptimos que se utilizaron para este modelo incluyen un número de estimadores de 100 y una profundidad máxima de 20. En comparación, el modelo Decision Tree mostró una sensibilidad del 85 % y una especificidad del 70 %, con un balanced accuracy del 77 % utilizando una profundidad máxima de 10 como parámetro óptimo. El modelo SVM, con un kernel linear y un parámetro C de 10, tuvo una sensibilidad del 75 %, una especificidad del 81 %, y un balanced accuracy del 78 %. Por último, el modelo Gradient Boosting, con 2000 estimadores y una profundidad máxima de 10, alcanzó una sensibilidad del 77 %, una especificidad del 81 % y un balanced accuracy también del 78 %.

Cuadro 4.4: Métricas de evaluación de los modelos (elaboración propia)

| **Modelo** | **Parametros Optimos** | **Sensibilidad** | **Especificidad** | **Balanced Accuracy** |
|---|---|---|---|---|
| **Decision Tree** | max_depth: 10 | 0.85 | 0.70 | 0.77 |
| **Random Forest** | n_estimators=100, max_depth: 20 | 0.88 | 0.87 | 0.88 |
| **SVM** | C: 10, kernel: 'linear' | 0.75 | 0.81 | 0.78 |
| **Gradient Boosting** | n_estimators=2000, max_depth: 10 | 0.77 | 0.81 | 0.78 |

Figura 4.13: Curva ROC de lo modelos



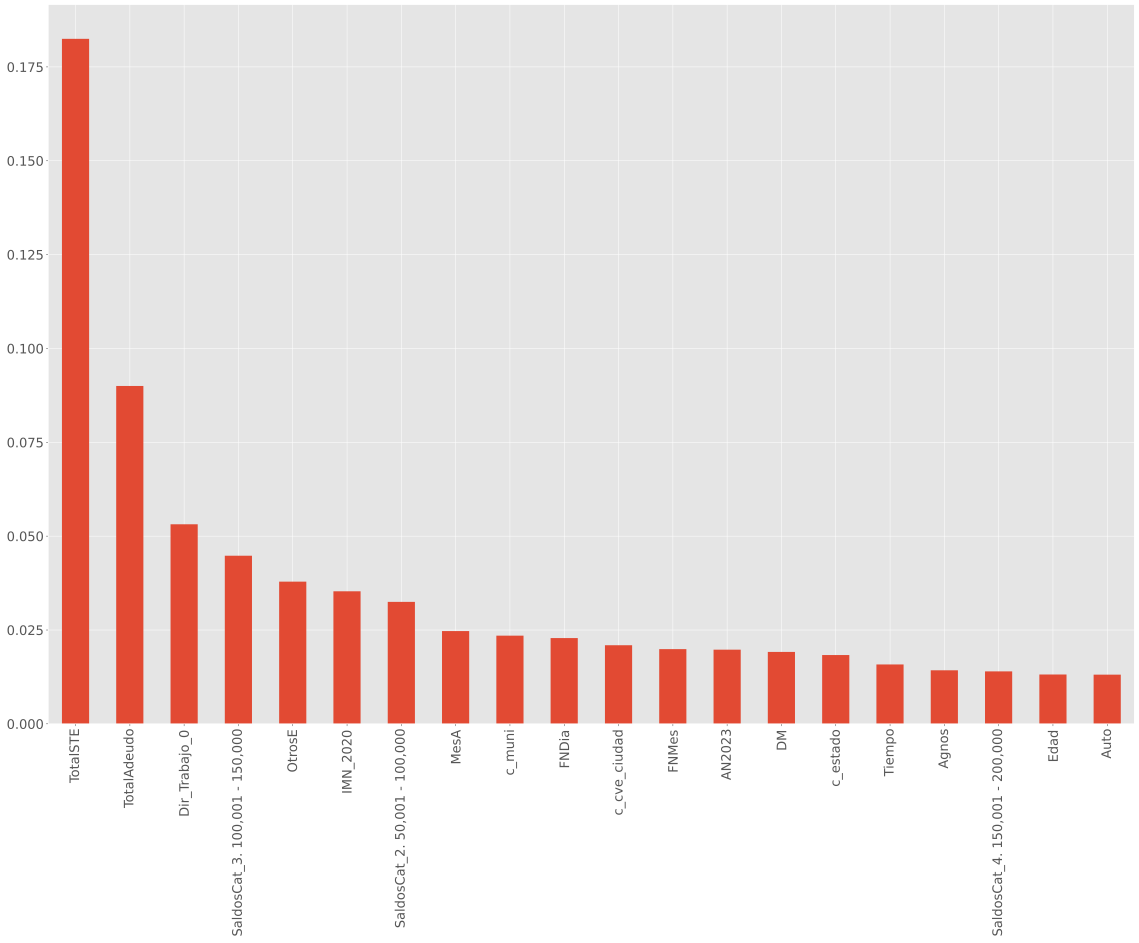

Fuente: Elaboración propia, 2024.

Con base en los resultados obtenidos de las gráficas de Curva ROC 4.13, se puede evaluar la efectividad de los cuatro modelos de machine learning: Árbol de Decisión, SVM, Bosque Aleatorio y Gradient Boosting, en clasificar correctamente casos positivos y negativos. El Bosque Aleatorio destaca como el modelo más competente, con un área bajo la curva (AUC) de 0.94. Esto indica una excelente capacidad para diferenciar entre las clases, demostrando un alto grado de sensibilidad y especificidad.

La implementación de técnicas de preprocesamiento de datos y balanceo de clases permitió abordar el problema de sesgo en los datos, lo que a menudo puede afectar negativamente la precisión del modelo. Estas técnicas están detalladas en la Sección 3.1.2, donde se explica su aplicación y efectividad. Este enfoque hacia la calidad de los datos subrayó la importancia de un buen fundamento de datos en la construcción de modelos predictivos confiables.

Además, el proceso de selección de variables fue clave para mejorar la interpretación del modelo y la velocidad de entrenamiento. Se identificaron y priorizaron las variables más influyentes, lo que permitió una comprensión más profunda de los factores que contribuyen al éxito en la recuperación de cuentas. Entre estas variables TotalSITE y Totaldeudas destacaron con una alta correlación, lo que implica una fuerte dependencia lineal entre éstas y la variable objetivo, como se puede observar en la figura mostrada en 4.14.

Figura 4.14: Caracteristicas más importantes




Fuente: Elaboración Propia, 2024

**Plan de optimización de la asignación de visitas domiciliarias**

Con base en los resultados del mejor modelo y la relevancia de las variables más importantes que incluyen el monto total para liquidar créditos, el total del adeudo generado,

la dirección laboral del deudor, el índice de marginación, los datos demográficos del deudor, la fecha de nacimiento y el mes de registro de la cuenta en el sistema, se presenta un plan estructurado para optimizar la asignación de visitas domiciliarias. El flujo de trabajo mostrado en la figura 4.15 representa un sistema integral para la optimización de la asignación de visitas domiciliarias, utilizando un enfoque basado en datos para mejorar la eficiencia y efectividad de las operaciones de recuperación de deudas. A continuación, se detallan los componentes clave del proceso:Este enfoque se fundamenta en una integración precisa de análisis predictivo y estrategias de segmentación, orientadas a maximizar la eficacia y eficiencia de las operaciones de recuperación de deudas.

**1. Integración y Análisis de Datos** Consolidar una base de datos actualizada que incluya las variables identificadas como las más importantes por el modelo Random Forest, como el monto total para liquidar créditos, el total del adeudo, y datos demográficos del deudor. Esto asegura que el análisis y las decisiones se basen en la información más actualizada y relevante.

**2. Modelo RF (Random Forest):** Utilizando técnicas de aprendizaje automático, el modelo RF procesa y analiza los datos integrados para identificar patrones. Basándose en estos patrones, el modelo hace predicciones sobre cuáles estrategias de cobro podrían ser más exitosas.

**3. Planificación de Visitas Domiciliarias**: Este paso implica la distribución y asignación eficiente de los recursos disponibles, asegurando que cada visita sea realizada a las cuentas de mayor prioridad.

- **Gestores Domiciliarios**: Con las predicciones y segmentaciones proporcionadas por el modelo RF, se asignan gestores domiciliarios. Estos recursos humanos son fundamentales para realizar las visitas domiciliarias y llevar a cabo las estrategias de recuperación.
- **Notificaciones y Viáticos**: Para cada visita planificada, se generan notificaciones para los gestores y se asignan viáticos. Esto garantiza que los gestores estén informados sobre sus asignaciones y cuenten con los recursos necesarios para

llevarlas a cabo.

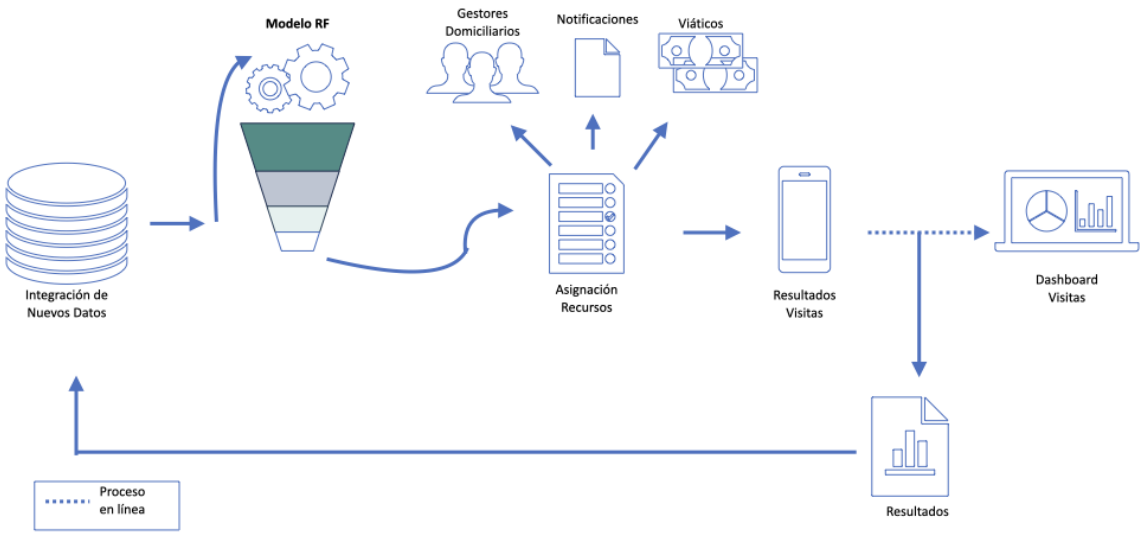
**4. Evaluación de Resultados y Ajustes**: Analizar los resultados de las visitas domiciliarias en términos de pagos recuperados y eficiencia de las rutas. Utilizar los datos recopilados para realizar ajustes continuos en el modelo predictivo y en la estrategia de segmentación y planificación de visitas. Esta información es crucial para evaluar el desempeño de las estrategias de cobro y la efectividad de los gestores.

**5. Dashboard Visitas**: Los resultados de las visitas se visualizan en un dashboard o panel de control. Esta herramienta analítica permite a los administradores evaluar la eficacia de las visitas y tomar decisiones informadas basadas en datos concretos.

**6. Retroalimentación y Mejora Continua**: Finalmente, el análisis de los datos recopilados a partir de las visitas y presentados en el dashboard permite realizar una evaluación general de las estrategias de recuperación. Con esta información, se pueden tomar decisiones estratégicas para mejorar futuras operaciones y ajustar el proceso según sea necesario.

Este flujo de trabajo optimizado busca maximizar la recuperación de deudas a través de un enfoque analítico y estratégico, asegurando que las visitas domiciliarias sean lo más productivas posible.

Figura 4.15: Diagrama. Optimización de Asignación de Visitas Domiciliarias




Fuente: Elaboración Propia, 2024

**Beneficios de la Propuesta**

- **Mayor Eficacia**: Al predecir la probabilidad de éxito y optimizar las rutas, se maximiza la recuperación de deudas.
- **Eficiencia Operativa**: Las rutas optimizadas y asignaciones en base a datos reducen el tiempo y el costo de las operaciones.
- **Adaptabilidad**: La metodología puede adaptarse rápidamente a cambios, gracias al análisis en tiempo real y al aprendizaje automático.
- **Mejora Continua**: El enfoque basado en datos garantiza una optimización y mejora continua del proceso.

### 4.2.6. Discusión

En la sección de resultados se observaron correlaciones significativas entre ciertas variables, como entre el total adeudado y el total de deuda con servicio, y entre el número de municipios y códigos postales, indicando relaciones lineales fuertes entre estas va-

riables.

El modelo de Random Forest demostró ser el más efectivo, logrando una precisión balanceada del 88 %, con alta especificidad y sensibilidad. Se destacó la importancia de un buen manejo de los datos y la selección de variables para mejorar la interpretación del modelo y la eficiencia en el entrenamiento, identificando las variables más influyentes en la recuperación exitosa de cuentas.

Se propuso un plan para optimizar la asignación de visitas domiciliarias, basado en los resultados del mejor modelo y las variables más relevantes. Este plan incluye la integración y análisis de datos, modelo rf (Random Forest), planificación de visitas domiciliarias, evaluación de resultados y ajustes, dashboard visitas, retroalimentación y mejora Continua.

A pesar de ciertas limitaciones temporales para la implementación completa del plan, se anticipa que este enfoque mejorará significativamente la eficiencia y eficacia de las operaciones de recuperación de deudas, al enfocar los esfuerzos en las cuentas con mayor probabilidad de pago y optimizar las rutas de visita.

# Conclusiones

# Conclusiones

En esta sección se presentan las conclusiones obtenidas del trabajo con base a los objetivos específicos planteados sección 1.1.1.

En este proyecto, se logró mejorar significativamente el proceso de recuperación de cartera vencida automotriz mediante la implementación y adaptación de diversos algoritmos de aprendizaje automático. Se utilizaron modelos específicos como árboles de decisión, bosques aleatorios, máquinas de soporte vectorial (SVM) y métodos de boosting de gradiente. Estos modelos fueron adaptados para abordar las características de los datos de cartera vencida, permitiendo una mejor identificación y evaluación de los atributos clave de los deudores con mayor probabilidad de cumplir con sus pagos.

En relación con el objetivo específico número 1, se logró realizar una clasificación detallada de las cuentas, identificando aquellas con mayor probabilidad de ser recuperadas. De esta forma, se categorizaron 578 registros como cuentas que resultaron en pago y 1,343 como cuentas sin pago. Sin embargo, para mejorar la precisión y fiabilidad del modelo predictivo, se implementó SMOTE, una técnica de balanceo de datos. Esta estrategia se empleó con el propósito de equilibrar la representación entre las clases de datos, mitigando así cualquier sesgo que pudiera estar presente debido a la desproporción inicial en la distribución de las clases.

En el marco del objetivo específico número 2, se llevó a cabo un análisis de correlación. Dentro de este análisis, se descubrió una correlación significativamente alta de 0.95 entre TotalSITE y Totaldeudas, lo que indicó una fuerte relación entre estas variables. De igual manera, se encontró una correlación de 0.93 entre No. de Municipios y NCPs, lo que demuestró otra fuerte conexión directa. Otra de las técnicas aplicadas para determinar los atributos más importantes en casos de éxito de las cuentas con pago fue Random Forest, en donde las variables más importantes fueron: monto total

a pagar para liquidar créditos (TotalSTE), total adeudo generado hasta la venta del crédito (TotalAdeudo), dirección de trabajo del deudor (Dir_Trabajo), índice de marginación (IMN_2020), día de la fecha de nacimiento (FNDia), municipio (C_muni), Ciudad (C_cve_ciudad), mes de la fecha de nacimiento (FNMes), año de la fecha de nacimiento (FNAgno) y mes de la carga de la cuenta en el sistema (MesA).

Con base al objetivo específico número 3, para la optimización de asignación de visitas domiciliarias se diseñó un plan que integró el resultado del Modelo para identificar y priorizar eficazmente las cuentas con mayor probabilidad de pago. Aunque el plan fue diseñado, abarcando desde la integración y análisis de datos, modelo RF (Random Forest), planificación de visitas domiciliarias, evaluación de resultados y ajustes, dashboard visitas, retroalimentación y mejora continua, no se logró implementar completamente dentro del plazo establecido para recoger resultados concretos y evaluar su impacto.

A pesar de esta limitación temporal, el plan propuesto se perfila como una mejora significativa para la empresa, prometiendo una gestión de cobranzas más eficiente y focalizada. Al enfocar los esfuerzos en las cuentas con mayor predisposición al pago y optimizar las rutas de visita, se espera no solo aumentar la tasa de recuperación de deudas sino también optimizar los recursos y reducir costos operativos.

Con respecto al objetivo específico número 4, la implementación delas métricas sensibilidad, especificidad y balanced accuracy permitió determinar el mejor modelo. En este análisis, se concluyó que el modelo Random Forest sobresalió como el más competente, evidenciado por un balanced accuracy del 88 %. Además, este modelo demostró un rendimiento notablemente equilibrado, alcanzando una sensibilidad de 88 % y una especificidad de 87 %. La consistencia entre la sensibilidad y la especificidad indica que Random Forest no solo identificó correctamente una alta proporción de casos exitosos sino que también determinó efectivamente los falsos positivos, reafirmando su utilidad como una herramienta fiable para la predicción en el ámbito de las cuentas con pago.

El diseño de un modelo predictivo implica mucho más que la creación de un algoritmo nuevo desde cero; también requiere una selección adecuada e implementación de algoritmos preexistentes adaptados a resolver problemas específicos. En este proyecto, el desafío radicó en la recuperación de cartera vencida automotriz y en evaluar la probabilidad de pago de los créditos. La elección de modelo y técnicas específicas se justifica por la naturaleza y cantidad de los datos que se tienen.

La incorporación de técnicas avanzadas de preprocesamiento y balanceo de datos fué esencial para aumentar la precisión y la efectividad del modelo predictivo. Estas estrategias permitieron que el modelo sea capaz de ofrecer mejor desempeño en predicciones, lo cual se alinea al objetivo de diseñar un modelo predictivo que mejore y optimice la gestión de la recuperación de cartera vencida automotriz.

En conclusión, los resultados alcanzados en este proyecto no solo satisfacen los objetivos específicos delineados, sino que también contribuyen significativamente al objetivo general de diseñar un modelo predictivo efectivo para la recuperación de cartera vencida automotriz. La identificación meticulosa de las cuentas con mayor probabilidad de recuperación y un análisis exhaustivo de las correlaciones y atributos clave han permitido desarrollar un modelo robusto, basado en algoritmos como Random Forest. Este modelo no solo predice con alta precisión la probabilidad de pago, sino que también optimiza los recursos al permitir una gestión más enfocada y eficiente de las cobranzas, un aspecto crucial para la mejora del proceso de recuperación de deudas en el sector automotriz.

# Bibliografía


Alarcón-Narváez, D., Hernández-Torruco, J., Hernández-Ocaña, B., Chávez-Bosquez, O., Marchi, J., and Méndez-Castillo, J. J. (2021). Toward a machine learning model for a primary diagnosis of guillain-barré syndrome subtypes. *Health Informatics Journal* 27, 14604582211021471

Arce Montaño, G. I. (2017). Medidas de prevención para la cartera vencida en instituciones financieras caso: Institución bancaria x

Bareño Amezquita, C. (2023). Modelo predictivo de clientes en mora

Bilmes, J. (2020). Underfitting and overfitting in machine learning. *UW ECE course notes* 5

Breiman, L. (1997). Arcing the edge

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

Buitrón, A., Rodríguez, C., Calisto, M. B., and Bonilla, S. (2022). Machine learning in finance: An application of predictive models to determine the payment probability of a client

Cabanillas Romero, J. C. (2022). Eficacia de modelos machine learning para el pronóstico del riesgo crediticio en la cartera consumo. coopac san josé cartavio

Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing* 300, 70–79

CNBV (2002). *LEY PARA REGULAR LAS SOCIEDADES DE INFORMACIÓN CREDITICIA*. Tech. rep.

Conde Hernández, R., Núñez Estrada, H. R., et al. (2003). Conceptualización y debate sobre la crisis bancaria en méxico

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning* 20, 273–297

Dash, M. and Liu, H. (2000). Feature selection for clustering. In *Pacific-Asia Conference on knowledge discovery and data mining* (Springer), 110–121

Flores de Valgas Williams, A. S., García Moreno, V. A., et al. (2023). Aplicación de machine learning para el control interno de la identificación de anomalías y proyección de mermas en la producción de papel higiénico: un enfoque de auditoría. *ESPOL. FCSH*

Francés Monedero, T. (2020). Impacto del machine learning en el sistema financiero

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics Data Analysis* 38, 367–378. doi:10.1016/S0167-9473(01)00065-2

Friedman, J. H. (2001a). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 1189–1232

Friedman, J. H. (2001b). Greedy function approximation: a gradient boosting machine. *Annals of statistics* , 1189–1232

Fu, H., Xiao, Z., Dellandréa, E., Dou, W., and Chen, L. (2009). Image categorization using esfs: a new embedded feature selection method based on sfs. In *Advanced Concepts for Intelligent Vision Systems: 11th International Conference, ACIVS 2009, Bordeaux, France, September 28–October 2, 2009. Proceedings 11* (Springer), 288–299

Girón, A. and Correa, E. (1997). *Crisis bancaria y carteras vencidas* (Universidad Nacional Autónoma de México, Instituto de Investigaciones . . . )

[Dataset] Gonzalez, L. (2019). Árboles de decisiÓn clasificaciÓn - teorÍa | 49 curso machine learning con python

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, vol. 2 (Springer)

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E.,

et al. (2020). Machine learning and artificial intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine* 13, 69–76

Hernández-Lalinde, J., Espinosa-Castro, J.-F., Peñaloza Tarazona, M., Rodríguez, J., Chacón, J., Carrillo Sierra, S., et al. (2018). Sobre el uso adecuado del coeficiente de correlación de pearson: definición, propiedades y suposiciones. *Archivos Venezolanos de Farmacologia y Terapeutica* 37, 587–595

[Dataset] IBM (2020). ¿qué es boosting? | ibm

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence* 97, 273–324

Nazemi, A., Rezazadeh, H., Fabozzi, F. J., and Höchstötter, M. (2022). Deep learning for modeling the collection rate for third-party buyers. *International Journal of Forecasting* 38, 240–252. doi:10.1016/j.ijforecast.2021

Quintero Acuña, L. K. (2023). Aplicación de machine learning a un modelo tradicional de prevención y detección de fraude en entidad financiera proyectado periodos trimestrales

[Dataset] Rodrigo, J. A. (2020a). Arboles de decision python

[Dataset] Rodrigo, J. A. (2020b). Gradient boosting con python

[Dataset] Rodrigo, J. A. (2020c). Máquinas de vector soporte (svm) con python

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development* 3, 210–229

Santillán Veliz, C. A. (2022). *El machine Learning como ventaja competitiva en el desarrollo de sistemas predictivos en el área de la inteligencia artificial.* B.S. thesis, Babahoyo: UTB-FAFI. 2022

[Dataset] Sefik Ilkin Serengil, U. G. T. E. B. B., Salih Imece and Koroglu, B. (2022). A comparative study of machine learning approaches for non performing loan prediction with explainability - volume 12 number 5 (sept. 2022) - international journal of machine learning (ijml)

[Dataset] Siappas, M. (2022). Best practices for successful debt recovery of auto loans

Singh, V., Yadav, A., Awasthi, R., and Partheeban, G. N. (2021). Prediction of modernized loan approval system based on machine learning approach. In *2021 International Conference on Intelligent Technologies (CONIT)*. 1–4. doi:10.1109/CONIT51480.2021.9498475

Song, Y. Y. and Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry* 27, 130–135. doi:10.11919/j.issn.1002-0829.215044

Tumuluru, P., Burra, L. R., Loukya, M., Bhavana, S., CSaiBaba, H., and Sunanda, N. (2022). Comparative analysis of customer loan approval prediction using machine learning algorithms. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. 349–353. doi:10.1109/ICAIS53314.2022.9742800

VW (2023). Tipos de financiamiento para comprar seminuevos VW

Wang, S., Yan, X., Zheng, B., Wang, H., Xu, W., Peng, N., et al. (2021). Risk and return prediction for pricing portfolios of non-performing consumer credit. In *Proceedings of the Second ACM International Conference on AI in Finance*. 1–9