UNIVERSIDAD
COMPLUTENSE
MADRID

# FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

# GRADO EN FINANZAS, BANCA Y SEGUROS

# TRABAJO DE FIN DE GRADO

TÍTULO: *Modelo matemático para la estimación de la recuperación en caso de impago en operaciones de financiación.*

AUTORA: *Marta Barrio Bayón*

TUTORA: *María Pilar García Pineda*

CURSO ACADÉMICO: *2023 - 2024*

CONVOCATORIA: *Junio*

Índice de los contenidos:

# 1. Introducción

## 1.1. Motivación

### 1.1.1. Contexto

Vivimos en una era donde la tecnología avanza a pasos agigantados, moldeando cada aspecto de nuestra vida diaria. Desde la aparición del primer ordenador personal en 1981, hemos presenciado una revolución tecnológica sin igual que ha transformado notablemente nuestra sociedad moderna.

El auge de internet y la omnipresencia del teléfono móvil como una extensión de nosotros mismos han revolucionado completamente nuestra manera comunicamos y establecer relaciones. Estos dispositivos, además de brindarnos acceso instantáneo a una cantidad inimaginable de información, también han abierto un mundo de posibilidades en el ámbito financiero. La capacidad de realizar transacciones, gestionar cuentas y acceder a servicios financieros desde la palma de nuestra mano ha democratizado el acceso a la economía global, permitiendo que incluso las personas más alejadas geográficamente estén conectadas al sistema financiero.

El verdadero poder de esta revolución tecnológica radica en la capacidad de análisis y procesamiento de datos. La explosión de datos en la era digital ha creado un tesoro de información que, debidamente analizada, puede revelar patrones, tendencias y comportamientos antes inaccesibles. Es aquí donde entra en juego el Machine Learning, de acuerdo con Mahesh (2020), es el estudio científico de algoritmos y modelos estadísticos que los sistemas informáticos utilizan para realizar una tarea específica sin necesidad de ser programados explícitamente. En el contexto empresarial, se ha convertido en un recurso esencial para extraer insights valiosos de grandes volúmenes de datos, permitiendo a las empresas tomar decisiones más informadas y anticiparse a las necesidades del mercado.

El sector financiero, en particular, ha emergido con unos cimientos sólidos para la aplicación del Machine Learning. Con enormes cantidades de datos transaccionales, históricos y de comportamiento a su disposición, las instituciones financieras están utilizando esta tecnología para una variedad de propósitos, desde la detección de fraudes hasta la personalización de servicios para sus clientes.

El Machine Learning no solo ha mejorado la eficiencia y la precisión en la toma de decisiones, sino que también ha abierto nuevas oportunidades de negocio y ha impulsado la innovación en un sector tradicionalmente conservador.

### 1.1.2. Experiencia en las prácticas

Las entidades financieras se encuentran ante un desafío constante, enfrentando una variedad de riesgos que impactan directamente en su estabilidad económica. En los últimos años, este desafío ha sido aún más evidente debido a una combinación de factores, incluyendo la crisis financiera global, la volatilidad en el mercado laboral y un aumento en el endeudamiento familiar.

Durante mi experiencia como becaria en el departamento de Business Intelligence de Pepper Finance, pude sumergirme en la complejidad de este desafío, especialmente en lo que respecta al riesgo de impago. Esta empresa se especializa principalmente en préstamos al consumo, lo que hace que el impago sea uno de los riesgos más críticos para su estabilidad financiera.

En el desarrollo de mis funciones, me involucré activamente en la elaboración de informes para el departamento de Contencioso, cuya responsabilidad principal es gestionar los casos de impago que superan un determinado umbral, generalmente cuando la morosidad se extiende más allá de tres pagos vencidos. Este proceso no solo implicaba la recopilación y análisis de datos relevantes, sino que también involucraba la identificación de patrones y tendencias que pudieran ofrecer información valiosa para la toma de decisiones estratégicas.

Además, también tuve la oportunidad de familiarizarme con diversas herramientas y técnicas utilizadas para mitigar el riesgo de impago, como modelos de scoring crediticio, análisis predictivo y estrategias de gestión de cartera. Estas herramientas son fundamentales para evaluar la solvencia de los clientes, identificar posibles señales de alerta temprana y diseñar estrategias proactivas para minimizar el impacto del impago en la cartera de la empresa.

### 1.1.3. Relevancia del problema

Durante mi experiencia en Pepper Finance, pude identificar una necesidad crucial dentro del proceso de recuperación de créditos: la falta de una herramienta precisa para determinar qué casos tienen el mayor potencial de éxito a través de demandas judiciales.

En la actualidad, este proceso se realiza de manera manual, lo que implica una gran carga de trabajo para el equipo encargado de evaluar cada caso individualmente. La decisión sobre qué casos deben ser llevados a juicio depende en gran medida de la experiencia y el juicio subjetivo de los profesionales involucrados. Esta metodología, además de ser laboriosa y lenta, puede llevar a decisiones no óptimas y a una asignación ineficiente de recursos.

Por tanto, en este trabajo se propone la implementación de un proyecto de Machine Learning para abordar este desafío. La idea es utilizar datos históricos de casos de recuperación de créditos y entrenar un modelo predictivo que pueda prever la probabilidad de éxito en demandas judiciales en cuestión de segundos. Este enfoque no solo agilizaría significativamente el proceso de toma de decisiones, sino que también aumentaría la precisión y la objetividad al eliminar el sesgo humano inherente a la evaluación manual de casos.

El uso de Machine Learning en este contexto ofrece varias ventajas. En primer lugar, permite analizar grandes volúmenes de datos de manera eficiente, identificando patrones y correlaciones que podrían pasar desapercibidos para un análisis humano tradicional. Además, al entrenar el modelo con datos históricos de casos reales, se puede mejorar continuamente su rendimiento y precisión a medida que se recopilan más datos y se obtienen más resultados.

## 1.2. Objetivos

### 1.2.1. Objetivo general

La propuesta de emplear algoritmos de Machine Learning para predecir la probabilidad de recuperación de deudas impagadas después de recibir una demanda se fundamenta en el análisis de datos históricos de casos previos.

Este enfoque surge como respuesta a la necesidad de mejorar la eficacia en la gestión de cobros y la selección de operaciones con mayor potencial de éxito en litigios judiciales en el ámbito de financiación al consumo.

En este contexto, se genera la motivación para llevar a cabo este proyecto, cuyo propósito primordial es la creación de un modelo matemático capaz de estimar la recuperación en casos de impago dentro de las operaciones de financiamiento al consumo. Este modelo, basado en el análisis de datos históricos y entrenado mediante técnicas avanzadas de Machine Learning, proporcionará a las entidades financieras una herramienta poderosa para optimizar sus estrategias de cobranza.

La implementación de este modelo permitirá a las instituciones financieras identificar de manera más precisa y objetiva las operaciones con mayor probabilidad de recuperación a través de demandas judiciales. Al predecir la viabilidad de estas acciones legales, las entidades podrán priorizar recursos y esfuerzos en aquellos casos que presenten mayores posibilidades de éxito, maximizando así la eficiencia en la gestión de cobros y reduciendo los costos asociados a litigios infructuosos.

Además, esta herramienta no solo beneficiará a las instituciones financieras en términos de optimización de recursos, sino que también contribuirá a la mejora de la estabilidad financiera del sistema en su conjunto. Al minimizar las pérdidas por incumplimientos y mejorar la eficacia en la recuperación de deudas, se fortalece la salud financiera de las entidades, lo que a su vez promueve la estabilidad y solidez del sistema financiero en su totalidad.

### 1.2.2. Objetivos específicos

Se han definido una serie de objetivos específicos con el fin de abordar de manera integral la problemática de la recuperación de deudas impagadas y desarrollar una solución efectiva mediante el uso de algoritmos de Machine Learning:

- Identificar los factores determinantes que influyen en la probabilidad de recuperación en caso de impago. Este paso es crucial para comprender las variables que impactan en la viabilidad de las acciones legales y poder incorporarlas en el modelo predictivo.

- Construir un modelo predictivo robusto que estime de manera precisa la probabilidad de recuperación para cada operación individual. Esto implica la selección adecuada de algoritmos y técnicas de Machine Learning, así como el entrenamiento del modelo utilizando datos históricos relevantes.
- Validar la eficacia del modelo mediante una muestra independiente de datos. La validación cruzada con datos externos es esencial para garantizar la generalización y la fiabilidad del modelo en diferentes contextos y condiciones.
- Analizar la sensibilidad del modelo a diversos escenarios y condiciones del mercado. Esta evaluación permitirá comprender cómo el modelo responde ante cambios en las variables clave y qué tan robusto es frente a diferentes condiciones económicas y de mercado.
- Desarrollar una herramienta práctica y accesible que permita a las entidades financieras utilizar el modelo de manera efectiva en su gestión diaria. Esta herramienta debe ser fácil de usar y proporcionar resultados claros y accionables para facilitar la toma de decisiones.

El logro de estos objetivos conllevará una serie de beneficios significativos para las entidades financieras:

- Optimización de la gestión de cobros, lo que se traduce en una reducción de costos y una mejora en la eficiencia operativa.
- Mejora en la toma de decisiones al poder seleccionar con mayor precisión las operaciones con mayor probabilidad de éxito en litigios judiciales.
- Minimización del riesgo de impago, lo que contribuye a fortalecer la estabilidad financiera y la salud general de la institución.

Además, este proyecto proporcionará una mayor comprensión del comportamiento de los impagos en el contexto de las operaciones de financiación al consumo, lo que puede ser invaluable para la formulación de políticas y estrategias futuras en este ámbito.

## 1.3. Metodología y retos

### 1.3.1. Metodología

Para explicar cómo se lleva a cabo el proceso de desarrollo del Machine Learning, nos apoyaremos en la Figura 1, que expone los pasos mencionados a continuación.

En el proceso inicial, comenzará con la recopilación de datos históricos de la empresa, accediendo a su base de datos gestionada en SQL. Esta recolección abarcará una variedad de variables relacionadas tanto con el cliente, como características socioeconómicas e historial crediticio, así como aspectos inherentes a la operación misma, como el importe, el plazo y el tipo de interés. También se considerará el historial de pagos, incluyendo la puntualidad y la morosidad acumulada.

Una vez obtenidos estos datos, se llevará a cabo un riguroso proceso de validación para garantizar su calidad. Esto incluirá la limpieza y transformación de la información con el objetivo de eliminar errores, inconsistencias y valores atípicos que puedan afectar la integridad de los resultados. Asimismo, se realizará un análisis de la distribución de las variables para seleccionar las más relevantes en la construcción del modelo.

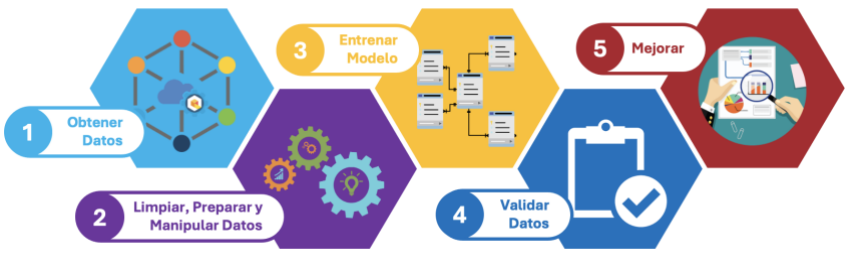
Seguidamente, se procederá a la selección del modelo matemático óptimo. Utilizando herramientas como el programa Data Robot, se entrenarán y evaluarán diferentes modelos, con el criterio de selección basado en su capacidad para ajustarse adecuadamente a los datos y ofrecer un alto rendimiento predictivo.

Con el fin de asegurar la fiabilidad del modelo, se llevará a cabo una etapa de validación utilizando métricas de rendimiento como la precisión, la sensibilidad y la especificidad. Además, se analizará la sensibilidad del modelo ante diferentes escenarios para evaluar su capacidad de adaptación a condiciones variables.

Finalmente, se procederá a la implementación del modelo en Python para automatizar la predicción de la probabilidad de recuperación. Esta implementación permitirá que el *score* obtenido se refleje directamente en la base de datos de la empresa, facilitando su integración en diversas funciones y procesos internos.

**Figura 1.**

*Pasos del proceso de desarrollo del Machine Learning.*




Fuente propia.

### 1.3.2. Retos

El desarrollo del modelo de estimación de la recuperación en caso de impago enfrenta diversos desafíos que requieren una cuidadosa consideración:

1. Disponibilidad y calidad de los datos

   La precisión y eficacia del modelo están estrechamente ligadas a la disponibilidad y calidad de los datos utilizados para su entrenamiento y validación. Es fundamental contar con una amplia gama de datos históricos que abarquen diferentes escenarios y situaciones, así como asegurar la calidad de estos datos mediante procesos rigurosos de limpieza y validación.
2. Selección de variables

   Se trata de un aspecto crítico en el desarrollo del modelo. Se debe realizar un análisis exhaustivo para identificar las variables que tienen un impacto significativo en la probabilidad de recuperación en caso de impago. Esto garantizará que el modelo sea lo más preciso y relevante posible, evitando la inclusión de variables no significativas que podrían afectar su rendimiento predictivo.
3. Interpretabilidad del modelo

   Aunque la precisión del modelo es fundamental, también es importante que sea interpretable para las entidades financieras. Esto significa que las predicciones del modelo deben ser comprensibles y transparentes, permitiendo a las instituciones entender las razones detrás de cada predicción. Esto facilitará la toma de decisiones informadas y la implementación de estrategias adecuadas para la gestión de cobros.
4. Implementación y uso del modelo

   La implementación exitosa del modelo en el entorno operativo de las entidades financieras es crucial para su efectividad. El modelo debe ser fácil de usar y estar integrado de manera fluida con los sistemas existentes de la institución. Además, se deben proporcionar herramientas y recursos adecuados para capacitar al personal en el uso del modelo y garantizar su adopción y uso continuo en la práctica diaria.

## 2. Marco teórico

### 2.1. Modelo predictivo en el sector financiero

El Machine Learning, según Alpaydin (2014), es una disciplina fascinante dentro del campo de la inteligencia artificial, revoluciona la forma en que las máquinas procesan información y toman decisiones. En su esencia, se trata de un enfoque computacional que permite a los sistemas aprender automáticamente y mejorar con la experiencia, sin una programación explícita para cada tarea específica. Este enfoque se inspira en el estudio de patrones y la adaptación que ocurre en la naturaleza, especialmente en el aprendizaje humano.

Dentro del amplio dominio del Machine Learning, una variedad de técnicas y algoritmos se han desarrollado para abordar una amplia gama de problemas. Estos algoritmos pueden clasificarse en diferentes categorías según el tipo de aprendizaje que realizan. Por ejemplo, el aprendizaje supervisado implica entrenar modelos con datos etiquetados, es decir, datos para los cuales la respuesta deseada ya se conoce. En contraste, el aprendizaje no supervisado implica descubrir patrones intrínsecos en los datos sin la necesidad de etiquetas predefinidas (Malagón-Selma et al., 2022).

En el ámbito financiero, los datos supervisados y no supervisados pueden provenir de diversas fuentes, como documentos financieros, series temporales, noticias, publicaciones en redes sociales o contenido textual relevante. El Machine Learning no supervisado también puede beneficiarse del enfoque de refuerzo, que implica definir estados y acciones en respuesta a cambios y acumular recompensas. A diferencia del aprendizaje supervisado, que se centra en acciones individuales, el aprendizaje no supervisado mediante refuerzo aborda secuencias óptimas de acciones, útil en situaciones como la asignación óptima de carteras o la optimización de operaciones financieras.

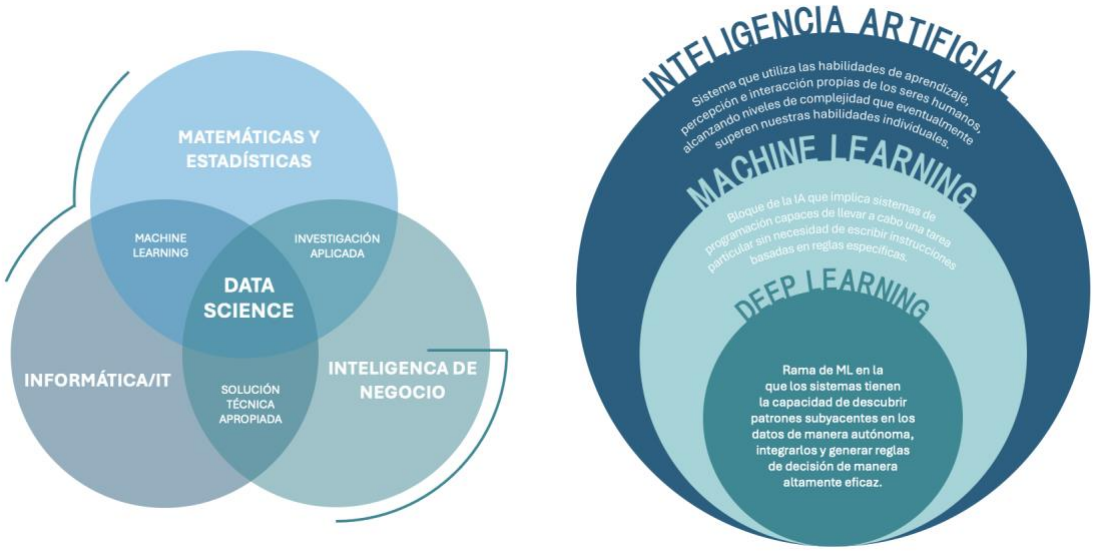
Uno de los aspectos más interesantes del Machine Learning, de acuerdo con Goodell et al. (2021), es su capacidad para adaptarse y mejorar con el tiempo. Los modelos pueden ajustarse continuamente a medida que se exponen a nuevos datos, lo que les permite adaptarse a cambios en el entorno y mejorar su rendimiento con el tiempo. Este proceso de iteración continua, conocido como aprendizaje incremental, es fundamental para el desarrollo de sistemas inteligentes que puedan evolucionar y mejorar a medida que se enfrentan a nuevas situaciones y desafíos.

Además, el Machine Learning se beneficia enormemente del poder computacional moderno y del acceso a grandes conjuntos de datos. El crecimiento explosivo en la capacidad de procesamiento y almacenamiento de datos ha permitido la aplicación de algoritmos de Machine Learning a problemas cada vez más complejos y de mayor escala. Esto ha llevado a avances significativos en una variedad de campos, desde el reconocimiento de voz y la visión por computadora hasta la biología computacional y la exploración espacial.

Dentro de este campo, el Deep Learning se destaca como una técnica poderosa, haciendo uso de redes neuronales profundas para abordar problemas de gran complejidad. Inspiradas en el funcionamiento del cerebro humano, estas redes aprenden automáticamente representaciones jerárquicas de los datos, permitiéndoles resolver tareas desafiantes como el reconocimiento de imágenes y el procesamiento de lenguaje natural. A pesar de requerir grandes volúmenes de datos y poder computacional, el Deep Learning ha propiciado avances significativos en diversas áreas, desde la medicina hasta la conducción autónoma (Goodfellow et al., 2016).

**Figura 2.**

*Distinción entre Data Science y Tecnologías de IA y Diagrama Circular entre IA, ML y DL*.




Fuente propia.

Los modelos predictivos, aplicaciones directas de los conceptos de inteligencia artificial y ciencia de datos, hacen uso de algoritmos para analizar datos históricos y realizar predicciones sobre eventos futuros o resultados desconocidos. La construcción y optimización de estos modelos implica técnicas avanzadas de machine learning y puede aprovechar tanto datos supervisados como no supervisados, así como también técnicas de deep learning. Por lo tanto, los modelos predictivos ejemplifican cómo se aplican los principios de la inteligencia artificial y la ciencia de datos para tomar decisiones informadas en diversos campos, incluyendo las finanzas.

## 2.2. Minería de datos en el sector financiero

Cuando se emplean estas técnicas en grandes conjuntos de datos, se les conoce como Minería de Datos. El término se asemeja a las minas, donde se extraen pequeñas cantidades de materiales valiosos. Del mismo modo, al analizar grandes cantidades de datos, se extrae una pequeña cantidad que aporta información valiosa (Alpaydin, 2014).

Hace décadas, los análisis financieros se basaban en datos históricos y modelos estadísticos simples. Sin embargo, con el avance de la tecnología y la disponibilidad de grandes volúmenes de datos, la minería de datos comenzó a cobrar importancia en el sector financiero. A medida que los bancos y las instituciones financieras acumulaban vastas cantidades de datos sobre transacciones, comportamientos de los clientes, mercados financieros y más, surgió la necesidad de aprovechar estos datos de manera efectiva para obtener información valiosa.

Hoy en día, en el sector financiero, la minería de datos se realiza utilizando técnicas avanzadas de inteligencia artificial y aprendizaje automático. Los científicos de datos y los analistas financieros emplean algoritmos sofisticados para explorar grandes conjuntos de datos en busca de patrones, tendencias y relaciones ocultas. Estos algoritmos pueden analizar datos estructurados, como transacciones financieras, así como datos no estructurados, como noticias financieras, publicaciones en redes sociales y comentarios de clientes (Hastie et al., 2009).

De acuerdo con Tell et al. (2013), La evolución de la minería de datos en el sector financiero ha permitido a las instituciones financieras obtener una comprensión más profunda de sus operaciones y clientes.

Al identificar patrones y tendencias en los datos, las organizaciones pueden tomar decisiones más informadas en áreas como la gestión de riesgos, la detección de fraudes, la personalización de servicios y la optimización de procesos. Además, la minería de datos también ha abierto nuevas oportunidades para el desarrollo de productos financieros innovadores y la mejora de la experiencia del cliente (Tello et al., 2013).

### 2.3. El riesgo de modelo

El desarrollo de un modelo predictivo para la recuperación del dinero impagado por clientes en el sector financiero implica una serie de consideraciones cruciales que deben abordarse desde el inicio del proyecto. La identificación y evaluación de los riesgos inherentes a este proceso es fundamental para garantizar su éxito y minimizar el impacto negativo en las finanzas de la institución. En este sentido, se propone la realización de un análisis DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) del proyecto. Esta herramienta estratégica permite comprender en profundidad los factores internos y externos que pueden influir en la efectividad del modelo predictivo.

Por un lado, existen debilidades como la posibilidad de tener datos incompletos o sesgados, lo que podría afectar la precisión del modelo. Además, la implementación de dicho modelo podría requerir recursos técnicos y computacionales significativos, así como experiencia en análisis de datos y Machine Learning. También es importante considerar que, dependiendo de la complejidad del modelo, podría ser difícil comprender completamente cómo se toman las decisiones del modelo, lo que podría generar desconfianza en su uso.

Por otro lado, hay amenazas a tener en cuenta, como los cambios en el comportamiento del cliente, que podrían hacer que los modelos predictivos se vuelvan obsoletos si no se reentrena constantemente. Además, los modelos deben cumplir con regulaciones estrictas sobre privacidad de datos y prácticas justas de cobro, lo que podría limitar la flexibilidad en su diseño y aplicación.

Sin embargo, también existen fortalezas y oportunidades. Las instituciones financieras suelen tener acceso a grandes cantidades de datos sobre transacciones financieras y comportamiento del cliente, lo que proporciona una base sólida para el desarrollo de modelos predictivos.

Además, la disponibilidad de tecnología avanzada y herramientas de análisis de datos permite el desarrollo de modelos sofisticados que pueden capturar patrones complejos en los datos.

Al final, un modelo predictivo preciso puede mejorar la gestión del riesgo al identificar proactivamente los clientes con mayor riesgo de incumplimiento y permitir una asignación más eficiente de recursos para la recuperación del dinero impagado. También puede ayudar a optimizar las estrategias de recuperación al identificar los enfoques más efectivos para diferentes segmentos de clientes.

## 3. Recopilación y tratamiento de datos

### 3.1. Descripción de la base de datos

La base de datos utilizada para el modelo está compuesta por varias tablas, cada una de las cuales contiene diferentes conjuntos de datos relevantes para el análisis y la predicción de la probabilidad de recuperación en caso de impago en préstamos al consumo tras realizar una demanda. A continuación, se hará una descripción más detallada de cada una de las tablas y los datos que contienen:

Demandas: Esta tabla alberga información relacionada con las demandas de préstamos al consumo. Incluye el ID de la operación, la fecha en que se produjo el impago del tercer recibo, la fecha de la demanda, el importe de la demanda, entre otros datos. Además, proporciona variables derivadas como la edad del cliente en el momento en que entró en contencioso (tres o más recibos impagados) y la edad en el momento de la demanda.

Operaciones: Aquí se recopila información detallada sobre los clientes y las operaciones de préstamos. Contiene el DNI del cliente, el ID del intermediario (punto de venta donde se concedió el préstamo al consumo), la actividad de la operación (préstamo al consumo o préstamo personal), entre otros datos.

Provincia: Esta tabla proporciona información geográfica adicional, específicamente la provincia y la comunidad autónoma del cliente.

Clientes: Contiene detalles específicos de los clientes, como el estado del correo electrónico devuelto, lo que puede ser relevante para evaluar la comunicación y la interacción con los clientes.

Cobros: Aquí se registra información sobre los cobros relacionados con las operaciones de préstamos, incluyendo si se logró el cobro, el importe del cobro tras ser impagado, entre otros detalles. Esta tabla es crucial para comprender el historial de pagos y el comportamiento de los clientes en términos de cumplimiento.

Plan de pagos: Esta tabla detalla la información relacionada con los pagos programados de los clientes, incluyendo el importe del pago, el total de pagos, el estado del pago, entre otros aspectos relevantes para comprender la estructura de los pagos y la capacidad de pago de los clientes.

La base de datos es bastante completa y contiene una variedad de información relevante para el análisis y la predicción de la probabilidad de recuperación en caso de impago en préstamos al consumo. La diversidad y amplitud de datos disponibles proporcionan una base sólida para el desarrollo de modelos predictivos precisos y efectivos en el ámbito financiero.

| Variable | Tabla de origen |
|---|---|
| Ppg_importe | PPG_Tramos |
| Ppg_importetotal | |
| Ppg_meses | |
| Ppg_Estado | |
| TipoPPG | |
| Ppg_pago | |
| TargetCobro | Cobros |
| Pct_CobroDem | |
| Idop (variable llave) | Demandas |
| Fechaco | |
| Ms_co | |
| ImporteDemanda | |
| DemandaFechaOK | |
| Edad_CO | Demandas |
| Edad_Demanda | |
| Pct_Demanda | |
| Provincia | Provincia_CCAA |
| CCAA | |
| CLI_Empleo_tx | Guia_basari |
| CLI_ECivil_Tx | |
| CLI_Tipocont_tx | |

| | |
|---|---|
| CLI_Sexo | |
| CLI_ciudad | |
| CLI_CP | |
| CLI_Nominaneto | |
| CLI_LugarNacimiento | |
| CLI_Nacionalidad | |
| CLI_PAIS_NAC_Conti | |
| OP_impop | |
| OP_sector | |
| PLA_hijo_nombre | |
| DNI | |
| Basari_idintermediario (variable llave) | |
| PDV_padre_id (variable llave) | |
| PDV_abuelo_id (variable llave) | |
| Actividad | |
| CLI_FchNac | |

## 3.2. Procesamiento de datos

### 3.2.1. Población

El proceso de construcción de modelos predictivos en el ámbito financiero requiere una cuidadosa selección y procesamiento de datos para garantizar la precisión y la fiabilidad de los resultados. En este contexto, se lleva a cabo un exhaustivo proceso de procesamiento de datos, que incluye varios pasos fundamentales. El primer paso consiste en identificar y segmentar la cartera de clientes que podrían enfrentar demandas debido a problemas con sus préstamos al consumo.

La población en este contexto se refiere al grupo o conjunto de individuos sobre los cuales se realiza el estudio o se construye el modelo predictivo. En este caso específico, la población serían los clientes que han adquirido préstamos al consumo y que podrían enfrentar dificultades para realizar sus pagos, lo que podría llevar a situaciones de impago o incumplimiento.

La identificación precisa de esta población es esencial para garantizar que el modelo se construya sobre datos relevantes y representativos, lo que permitirá obtener predicciones precisas y útiles sobre el comportamiento financiero futuro de estos clientes.

### 3.2.2. Muestra

Los datos históricos desempeñan un papel fundamental en el proceso de construcción y validación de modelos predictivos. Esta selección se basa en la premisa de que el pasado puede ofrecer una visión valiosa sobre el comportamiento futuro. Esencialmente, estos datos actúan como el material de entrenamiento sobre el cual se basa el modelo para aprender patrones y relaciones entre variables.

La muestra de datos históricos se elige meticulosamente para garantizar su representatividad y diversidad. Es crucial que esta muestra refleje adecuadamente la heterogeneidad de situaciones y características que los clientes pueden enfrentar al experimentar dificultades para realizar sus pagos. Esta representatividad es clave para que el modelo pueda capturar la complejidad del comportamiento financiero y, por ende, realizar predicciones más precisas y útiles.

Una vez que se ha seleccionado la muestra de datos, se procede a utilizar los datos históricos asociados con estos clientes para entrenar y validar el modelo predictivo. Durante este proceso, el modelo se expone a una variedad de escenarios pasados que ayudan a identificar patrones y relaciones entre las diferentes variables. Estos patrones aprendidos se convierten en la base sobre la cual el modelo hace predicciones sobre el riesgo de incumplimiento en situaciones futuras.

Es importante destacar que este enfoque no se limita únicamente a la predicción de recuperación del capital, sino que también puede ser aplicado para comprender mejor el comportamiento financiero de los clientes.

Al estudiar los datos históricos, se pueden identificar tendencias y factores que influyen en la capacidad de los clientes para cumplir con sus obligaciones financieras, lo que proporciona una visión más profunda y completa de la salud financiera de la población objetivo.

### 3.2.3. Criterios de exclusión

En la estrategia del estudio, se tomó la decisión de dejar fuera una horquilla temporal de datos más recientes por diversas razones, cada una fundamental para asegurar la validez y la utilidad del análisis.

Reservar una porción de datos más recientes para la validación del modelo es esencial. Esta práctica te permite evaluar cómo se comporta el modelo con datos que no ha visto durante el entrenamiento, lo que es crucial para evitar el sobreajuste. Este hecho ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento específicos y no puede generalizar bien a datos nuevos, lo que puede llevar a predicciones inexactas en situaciones del mundo real.

Esto proporciona una evaluación más precisa de la capacidad del modelo para hacer predicciones útiles en escenarios reales, lo que contribuye a su fiabilidad y utilidad práctica.

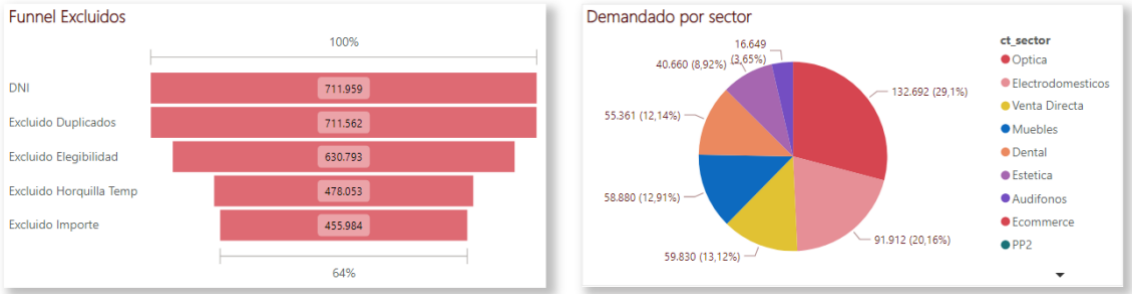
Los cambios en el comportamiento de los datos también son una consideración crucial. Los patrones y relaciones pueden variar con el tiempo debido a una multitud de factores, como cambios en el mercado, nuevas regulaciones o avances tecnológicos. Al dejar fuera datos más recientes, tienes la oportunidad de capturar estos cambios y ajustar el modelo en consecuencia, lo que garantiza su relevancia y precisión a lo largo del tiempo.

Asimismo, mantener una horquilla temporal de datos más recientes facilita la actualización y reentrenamiento periódico del modelo. Esto es fundamental para garantizar que el modelo siga siendo relevante y preciso a medida que evolucionan los datos y las condiciones del mundo real, lo que aumenta su utilidad a lo largo del tiempo.

Además, es importante abordar la presencia de registros duplicados en la base de datos, ya que pueden introducir sesgos y distorsiones en el análisis. Por lo tanto, realizar una limpieza exhaustiva de la base de datos para eliminar duplicados es un paso crítico para garantizar la integridad y la precisión de los datos utilizados en el estudio (Figura 3).

**Figura 3.**

*Funnel de caídas por criterio de exclusión y Gráfico circular de demandados por sector*.




Fuente propia.

### 3.2.4. Unidad de análisis

La selección de variables es un aspecto fundamental en el proceso de construcción de modelos predictivos. En el caso específico de evaluar la probabilidad de recuperación en caso de incumplimiento en préstamos al consumo, esta etapa cobra una importancia aún mayor. Aquí, se eligen cuidadosamente las variables que se consideran relevantes para predecir dicha probabilidad.

Estas variables abarcan una amplia gama de aspectos, como la información demográfica del cliente, su historial crediticio, comportamiento de pago pasado y las características específicas de los préstamos otorgados, entre otros factores clave. La meticulosa selección de estas variables contribuye a simplificar el modelo y a mejorar su capacidad predictiva al incluir únicamente aquellas que tienen un impacto significativo en las predicciones.

Es importante destacar que la elección de variables no es un proceso arbitrario, sino que se basa en un análisis exhaustivo de los datos disponibles. Se lleva a cabo una minuciosa construcción de la base de datos para garantizar su limpieza, integridad y estructura correcta. Esto asegura que los resultados obtenidos sean precisos y confiables, ya que cualquier error o inconsistencia en los datos podría afectar negativamente la calidad del modelo final.

Una vez seleccionadas las variables pertinentes y preparada la base de datos, se procede a alimentar esta información en la tabla M11_Base_Test_CODEM_v1 (Figura 4) para su posterior análisis en la plataforma DataRobot. Esta herramienta ofrece una plataforma integral que facilita la construcción y el despliegue de modelos predictivos, aprovechando algoritmos avanzados de aprendizaje automático y técnicas de inteligencia artificial.

El proceso completo de selección y procesamiento de datos desempeña un papel crucial en la construcción de modelos predictivos para evaluar el riesgo de incumplimiento en préstamos al consumo. Comprender y justificar cada paso realizado durante este proceso es esencial para desarrollar un modelo robusto y confiable. Esto no solo proporciona información valiosa para la toma de decisiones financieras, sino que también contribuye a una gestión efectiva del riesgo, lo que es fundamental en el ámbito de las instituciones financieras.

**Figura 4.**

*Script de SQL para preparar y procesar datos antes de su modelado.*

```
--- BASE DE DATOS ----
    drop table if exists #dataset

        select
            c.*,
            ini_Incidencias_Totales = cast(coalesce(i.Incidencias_Totales,i1.Incidencias_Totales) as money)  ,
            ini_saldoimp_Total      = cast(coalesce(i.saldoimp_Total,i1.saldoimp_Total)         as money)  ,
            ini_a_BCO_FINANe        = cast(coalesce(i.a_BCO_FINANe,i1.a_BCO_FINANe)               as money) ,
            ini_a_BCO_FINANn        = cast(coalesce(i.a_BCO_FINANn,i1.a_BCO_FINANn)               as money) ,
            fin_Incidencias_Totales = cast(coalesce(f.Incidencias_Totales,f1.Incidencias_Totales) as money)  ,
            fin_saldoimp_Total      = cast(coalesce(f.saldoimp_Total,f1.saldoimp_Total)           as money) ,
            fin_a_BCO_FINANe        = cast(coalesce(f.a_BCO_FINANe,f1.a_BCO_FINANe)             as   money)  ,
            fin_a_BCO_FINANn        = cast(coalesce(f.a_BCO_FINANn,f1.a_BCO_FINANn)               as money)

    into #dataset

        from #CO_dataset c
            left join batch_bureau i on c.DNI = i.DNI and LEFT(c.fechaco,6) = i.mes_batch
            left join batch_bureau i1 on c.DNI = i1.DNI and LEFT(convert(varchar(8),dateadd(month,-1,cast(c.fechaco as varchar(8))),112),6) = i1.mes_batch
            left join batch_bureau f on c.DNI = f.DNI and LEFT(c.DemandaFechaOK,6) = f.mes_batch
            left join batch_bureau f1 on c.DNI = f1.DNI and LEFT(convert(varchar(8),dateadd(month,-1,cast(c.DemandaFechaOK as varchar(8))),112),6) = f1.mes_batc

exec Basari.dbo.clrMensaje '10/11 #dataset'

--- TABLA DATA ROBOT --

    drop table if exists M11_Base_Test_CODEM_v1

        select *

    into M11_Base_Test_CODEM_v1

        from #dataset


    exec Basari.dbo.clrMensaje '11/11 #M11_Base_Test_CODEM_v1'
```

Fuente propia.

## 4. Desarrollo del modelo de predicción

### 4.1. Entrenamiento del modelo

En el marco de la investigación, se ha empleado una herramienta para el análisis de datos: DataRobot. Este programa ofrece una plataforma integral para el desarrollo y la evaluación de modelos predictivos, aprovechando algoritmos de aprendizaje automático de vanguardia y técnicas de inteligencia artificial.

DataRobot simplifica significativamente el proceso de construcción de modelos, permitiéndome cargar fácilmente conjuntos de datos históricos y explorar una amplia gama de algoritmos y configuraciones para encontrar el modelo más efectivo. Utilizando esta plataforma, he podido profundizar en la complejidad de los datos y obtener insights valiosos para la estimación de la recuperación en situaciones de impago.

Para llevar a cabo esta parte inicial, se ha cogido la tabla M11_Base_Test_CODEM_v1 anteriormente creada en SQL con las variables disponibles y, a través de Phyton, se ha lanzado a dicho programa. Es necesario realizar este paso intermedio ya que no se puede vincular directamente a la base de datos.

Es importante destacar que, durante el proceso de preparación de los datos para el análisis, he realizado ciertas modificaciones para garantizar la calidad y la eficacia del modelo. En particular, se ha eliminado variables de identificación o variables clave, ya que estas no son variables predictivas en sí mismas, sino más bien enlaces que facilitan la asociación posterior de los resultados del modelo con la base de datos original. Este paso de preprocesamiento asegura que el modelo se centre únicamente en las variables relevantes para la predicción de la recuperación en casos de impago, sin distracciones o ruido innecesario.

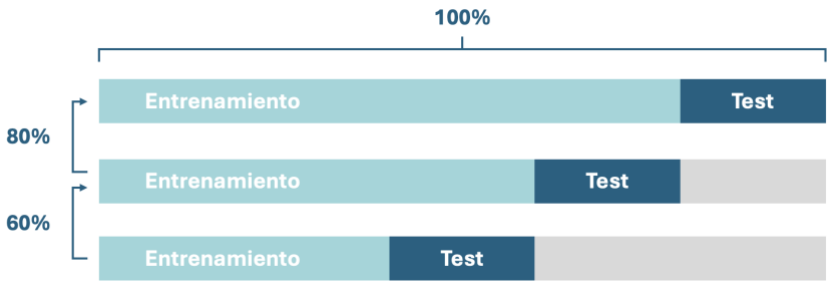
## 4.2. Selección del modelo

Para la selección del modelo se han llevado a cabo unas rigurosas pruebas y evaluaciones utilizando la estrategia 80/20, emergiendo la opción más sólida y prometedora para cumplir los objetivos del estudio.

La estrategia 80/20, también conocida como regla de Pareto, es una guía valiosa en la selección de modelos, fundamentada en el principio de que una parte sustancial de los resultados proviene de una minoría de las causas. Aplicada en la investigación, esta estrategia implica dividir los datos en dos conjuntos: el 80% para entrenar y ajustar los modelos, y el 20% restante para validar su desempeño (Figura 5). Esta división proporciona un equilibrio entre la cantidad de datos utilizados para el entrenamiento, permitiendo que el modelo capture la complejidad de los datos, y la evaluación en un conjunto separado para garantizar la generalización del modelo a datos no vistos.

**Figura 5.**

*Proceso de validación 80/20 para evaluar el rendimiento de modelos Machine Learning*.




Fuente propia.

La validación de reserva, esencial en este proceso, implica la partición aleatoria de los datos en un conjunto de entrenamiento y un conjunto de validación. Esta técnica garantiza que tanto el conjunto de entrenamiento como el de validación sean representativos de la población subyacente, evitando así cualquier sesgo en la evaluación del modelo. Según Angulo y Jara (2023), la elección de una división de 80:20 permite aprovechar al máximo los datos disponibles para el entrenamiento, mientras se reserva una porción significativa para validar el rendimiento del modelo en datos no utilizados durante el entrenamiento.

Este enfoque estratégico no solo garantiza que el modelo seleccionado se ajuste adecuadamente a los datos de entrenamiento, sino que también proporciona una evaluación objetiva de su capacidad para generalizar a nuevos datos. Al evitar el sobreajuste y maximizar la capacidad de generalización del modelo, esta metodología asegura que los resultados obtenidos sean sólidos y confiables, sentando así las bases para conclusiones significativas y aplicaciones prácticas en el ámbito financiero.

Finalmente, tras realizar un análisis de todos los posibles modelos, se terminó eligiendo el modelo de Regresión de Recuperación Óptima (RRO). Su consistente alta precisión en la estimación de la recuperación, respaldada por múltiples pruebas, indica que es la herramienta más idónea para abordar las complejidades del análisis de datos financieros.

### 4.3. Validación del modelo

La validación de un modelo predictivo es un paso fundamental en su desarrollo, ya que permite evaluar su capacidad para generalizar patrones y hacer predicciones precisas en situaciones del mundo real.

La Regresión de Recuperación Óptima ha sido sometida a una exhaustiva validación para garantizar su fiabilidad y utilidad en entornos financieros dinámicos y complejos. En este proceso, se han explorado diversas características y capacidades específicas de este modelo, enfocándonos en su adaptabilidad, integridad de datos, manejo de desbalance, rigurosidad en las pruebas, interpretación de resultados y actualización dinámica.

Mediante una combinación de técnicas de validación robustas y un enfoque centrado en la precisión y la interpretación de resultados, esta investigación busca validar y respaldar la utilidad práctica de la Regresión de Recuperación Óptima como una herramienta efectiva para la gestión del riesgo financiero y la toma de decisiones estratégicas en el ámbito de las operaciones de financiación.

Algunas características específicas que se podría atribuir al modelo de Regresión de Recuperación Óptima:

1. Adaptabilidad

   La Regresión de Recuperación Óptima es altamente adaptable a una amplia gama de situaciones financieras, lo que la hace adecuada para diversos tipos de operaciones de financiación y escenarios de impago.
2. Incorporación de datos heterogéneos

   Este modelo tiene la capacidad de integrar y procesar eficientemente datos heterogéneos, incluidos datos demográficos, históricos de pagos, información crediticia y características específicas de los préstamos.
3. Gestión de datos desbalanceados

   La Regresión de Recuperación Óptima está diseñada para abordar el desafío de datos desbalanceados común en el análisis de impagos, donde las instancias de impago pueden ser significativamente menos frecuentes que las instancias de pago.
4. Precisión y robustez

   Se destaca por su precisión en la predicción de la recuperación en casos de impago, respaldada por un robusto proceso de validación cruzada y pruebas exhaustivas en datos históricos.
5. Interpretación de resultados

   Aunque es un modelo complejo, la Regresión de Recuperación Óptima permite una interpretación relativamente sencilla de sus resultados, lo que facilita la comprensión y la toma de decisiones basadas en el análisis predictivo.
6. Actualización continua

   Este modelo está diseñado para adaptarse dinámicamente a medida que cambian los datos y las condiciones del mercado, lo que permite su actualización periódica para mantener su relevancia y precisión a lo largo del tiempo.

Estas características podrían resaltar la efectividad y la utilidad de la Regresión de Recuperación Óptima como un modelo predictivo para estimar la recuperación en casos de impago en operaciones de financiación.

## 5. Implementación del modelo en Python

### 5.1. Descripción del código

El output del modelo consiste en predicciones sobre la probabilidad de recuperación asociada con cada operación en la muestra de datos analizada. Estas predicciones se presentan en forma de valores de probabilidad, que indican la probabilidad estimada de que un préstamo impagado sea recuperado con éxito.

El score extraído del modelo es una lista de valores de probabilidad, donde cada valor corresponde a una operación en el conjunto de datos. Estos valores se interpretan como la probabilidad predicha de recuperación para cada caso de impago en particular.

Para interpretar estos datos, es crucial entender que cuanto mayor sea el valor de probabilidad asociado con una operación, mayor será la probabilidad estimada de que ese préstamo impagado se recupere con éxito. Por otro lado, los valores más bajos de probabilidad indican una menor probabilidad de recuperación para ese préstamo en particular.

Esta información puede ser utilizada por los profesionales financieros y analistas de riesgos para tomar decisiones informadas sobre la gestión del riesgo de crédito y la asignación de recursos para la recuperación de préstamos impagados. Además, estas predicciones pueden ser integradas en sistemas de gestión de carteras para optimizar las estrategias de recuperación y minimizar las pérdidas financieras asociadas con impagos.

### 5.2. Integración con la base de datos

El primer paso en este proceso de integración es la generación del score a partir del modelo predictivo. Una vez que se han realizado las predicciones sobre la probabilidad de recuperación para cada operación en el conjunto de datos, estos scores se exportan desde el entorno de modelado, donde se ejecutó la Regresión de Recuperación Óptima, hacia un entorno de programación como Python.

En Python, el score obtenido del modelo se combina con la variable llave correspondiente a cada operación en la base de datos de la empresa. Esta variable llave es crucial ya que sirve como puente de conexión entre la información generada por el modelo y los registros individuales de la base de datos.

Una vez que se han combinado el score y la variable llave en Python, estos datos se vuelcan en la base de datos de la empresa como una nueva variable. Esta nueva variable, que representa el score de recuperación asociado con cada registro de impago, se integra en la estructura de la base de datos existente y está disponible para su uso en análisis posteriores y en la toma de decisiones operativas y estratégicas.

Para automatizar el proceso de integración del score del modelo con la base de datos usando Python, primero, establece una conexión con la base de datos SQLServer. Luego, genera el score del modelo para cada operación utilizando DataRobot.

Después, combina el score con la variable llave correspondiente utilizando la biblioteca de manipulación de datos en Python. A continuación, actualiza la base de datos escribiendo los datos actualizados (incluyendo el score) utilizando consultas SQL.

Finalmente, programa la automatización utilizando una biblioteca de planificación de tareas en Python para ejecutar el script automáticamente en intervalos regulares o en respuesta a eventos específicos. Además, se implementó mecanismos de monitoreo y registro para garantizar un proceso sin problemas y para recibir notificaciones en caso de errores.

## 6. Evaluación y resultados

### 6.1. Análisis del modelo de predicción

#### 6.1.1. Rendimiento del modelo

El rendimiento del modelo se ha evaluado meticulosamente mediante la aplicación de diversas métricas de evaluación, incluyendo precisión, sensibilidad, especificidad y otras medidas relevantes. Estas métricas se han calculado utilizando conjuntos de datos de validación y prueba, lo que ha permitido una evaluación exhaustiva de la capacidad del modelo para predecir con precisión la probabilidad de recuperación en casos de impago.

Además, se ha llevado a cabo una comparación detallada del rendimiento del modelo con otros modelos de referencia y enfoques tradicionales utilizados en el sector financiero. Esta comparación ha proporcionado una visión crítica sobre la eficacia relativa del modelo propuesto en relación con otras metodologías establecidas.

El análisis de rendimiento del modelo ha arrojado resultados alentadores, demostrando una capacidad significativa para predecir con precisión la probabilidad de recuperación en casos de impago. La comparación con otros modelos de referencia ha revelado ventajas distintivas del enfoque de Machine Learning empleado en términos de precisión y capacidad predictiva.

En conjunto, estos hallazgos respaldan la efectividad y la relevancia del modelo propuesto en el contexto de la gestión del riesgo de impago en el sector financiero, destacando su potencial para mejorar significativamente la toma de decisiones y la eficiencia operativa en este ámbito crítico.

6.1.2. Robustez del modelo:

En el análisis de la robustez del modelo, se han llevado a cabo diversas pruebas y evaluaciones para determinar su capacidad para mantener su rendimiento en diferentes conjuntos de datos y escenarios, así como su estabilidad a lo largo del tiempo.

En primer lugar, se realizaron pruebas de sensibilidad para evaluar cómo el modelo respondía a cambios en los datos de entrada. Estas pruebas implicaron la introducción de variaciones deliberadas en los conjuntos de datos, como la inclusión o exclusión de ciertas variables, la modificación de los rangos de valores de las variables existentes, o la introducción de datos atípicos.

El objetivo era verificar si el modelo podía adaptarse a estas variaciones sin comprometer su capacidad predictiva. Los resultados de estas pruebas indicaron que el modelo era robusto y podía mantener un rendimiento consistente en diferentes escenarios.

Además, se realizó un análisis de la estabilidad a lo largo del tiempo del modelo. Esto implicó evaluar si el rendimiento del modelo se mantenía constante o si se degradaba con el tiempo debido a cambios en las condiciones del mercado o en el comportamiento de los clientes. Se llevaron a cabo pruebas retrospectivas utilizando datos históricos de diferentes períodos para determinar si el modelo podía mantener su precisión a lo largo del tiempo. Los resultados mostraron que el modelo era estable y mantenía un rendimiento consistente incluso en períodos de tiempo prolongados.

### 6.1.3. Interpretabilidad del modelo

La interpretabilidad del modelo es un aspecto crucial a considerar, especialmente cuando se trata de usuarios finales como analistas financieros y tomadores de decisiones en entornos empresariales. La comprensión clara de cómo funciona el modelo y cómo llega a sus predicciones es fundamental para que los usuarios confíen en sus resultados y utilicen esas predicciones de manera efectiva en la toma de decisiones informadas.

En este contexto, se han empleado técnicas de explicabilidad de modelos para mejorar la interpretabilidad del modelo propuesto. Estas técnicas permiten entender cómo el modelo llega a sus predicciones y qué variables tienen mayor influencia en los resultados. Por ejemplo, se han utilizado métodos como la importancia de características (feature importance) y gráficos de dependencia parcial (partial dependence plots) para explorar y visualizar la relación entre las variables de entrada y la salida del modelo.

La importancia de características proporciona una medida de la contribución de cada variable a las predicciones del modelo. Esto permite a los usuarios identificar qué variables son más relevantes para el resultado del modelo y cómo influyen en las predicciones. Por otro lado, los gráficos de dependencia parcial muestran cómo cambia la predicción del modelo cuando se modifica una variable específica, manteniendo constantes las demás. Esto ayuda a comprender la naturaleza de la relación entre las variables de entrada y la salida del modelo.

Al proporcionar esta información detallada sobre cómo el modelo toma decisiones y qué factores influyen en sus predicciones, se mejora significativamente la interpretabilidad del modelo. Los analistas financieros y los tomadores de decisiones pueden utilizar esta información para entender mejor los resultados del modelo y justificar sus decisiones basadas en esas predicciones. Además, esta transparencia y claridad contribuyen a construir la confianza en el modelo y su capacidad para proporcionar información relevante y útil para la toma de decisiones empresariales.

### 6.1.4. Aplicabilidad del modelo

En cuanto a la aplicabilidad del modelo en entornos del mundo real, es fundamental evaluar tanto la viabilidad técnica como operativa de su implementación en los sistemas existentes de la entidad financiera.

Desde una perspectiva técnica, el modelo debe ser compatible con las plataformas de software y bases de datos utilizadas por la entidad financiera, lo que implica considerar aspectos como el lenguaje de programación, la infraestructura de almacenamiento de datos y la integración con otros sistemas empresariales. Además, se debe asegurar que el modelo pueda manejar grandes volúmenes de datos de manera eficiente y escalable, ya que las instituciones financieras suelen operar con conjuntos de datos masivos.

Por otro lado, desde una perspectiva operativa, es crucial evaluar cómo se integraría el modelo en los procesos y flujos de trabajo existentes de la entidad financiera. Esto incluye considerar aspectos como la automatización de la generación de predicciones, la incorporación de los resultados del modelo en los sistemas de gestión de riesgos y la capacitación del personal para comprender y utilizar eficazmente las predicciones del modelo. Además, se deben establecer protocolos claros para la actualización y mantenimiento del modelo a lo largo del tiempo, asegurando su relevancia y precisión continua en un entorno financiero dinámico.

En cuanto al valor agregado que el modelo proporciona a la entidad financiera, es fundamental evaluar si su uso mejora la gestión del riesgo de impago y la toma de decisiones financieras en comparación con enfoques tradicionales. Esto implica realizar comparaciones con modelos o métodos existentes utilizados por la institución financiera, así como evaluar el impacto del modelo en indicadores clave de desempeño, como la reducción del riesgo de impago, la mejora de la eficiencia operativa y la maximización del retorno de la cartera de préstamos.

Además, se deben considerar aspectos cualitativos, como la confianza y la aceptación por parte de los usuarios finales, así como la capacidad del modelo para proporcionar insights y perspectivas que no son posibles con enfoques tradicionales. En última instancia, el valor agregado del modelo se traduce en beneficios tangibles y cuantificables para la entidad financiera, lo que justifica su inversión y adopción en un entorno competitivo y en constante cambio.

### 6.1.5. Limitaciones del modelo

Identificar y comunicar las limitaciones del modelo es crucial para comprender su alcance, así como para establecer expectativas realistas sobre su desempeño y aplicabilidad en situaciones del mundo real.

Una de las limitaciones más importantes a considerar son los posibles sesgos inherentes en los datos de entrenamiento. Los modelos de aprendizaje automático pueden estar sesgados si los datos utilizados para entrenarlos no son representativos de la población objetivo o si contienen sesgos históricos en la recopilación de datos. Por ejemplo, si los datos de entrenamiento están sesgados hacia ciertos grupos demográficos o regiones geográficas, el modelo puede generar predicciones sesgadas que no reflejen de manera precisa la realidad.

Otra limitación común se relaciona con las suposiciones simplificadas del modelo. Los modelos de aprendizaje automático a menudo se basan en ciertas suposiciones o simplificaciones sobre la naturaleza de los datos y las relaciones entre las variables. Estas suposiciones pueden no ser válidas en todos los casos y pueden llevar a predicciones inexactas o imprecisas. Por ejemplo, un modelo de regresión lineal asume una relación lineal entre las variables independientes y la variable dependiente, lo cual puede no ser adecuado para datos con relaciones no lineales.

Además, es importante tener en cuenta las condiciones bajo las cuales el modelo puede no ser aplicable o preciso. Por ejemplo, si las condiciones del mercado cambian significativamente o si se introducen nuevos factores que no estaban presentes en los datos de entrenamiento, el modelo puede volverse obsoleto o perder su precisión. Del mismo modo, si el modelo se implementa en un entorno diferente al de los datos de entrenamiento, su desempeño puede verse afectado debido a diferencias en las características de los datos o en las condiciones operativas.

Para mitigar estas limitaciones y mejorar la robustez y la fiabilidad del modelo en situaciones del mundo real, es importante implementar medidas apropiadas. Esto puede incluir la recopilación de datos más representativos y diversificados, la realización de pruebas exhaustivas para identificar y corregir sesgos en los datos, y la utilización de técnicas avanzadas de modelado que puedan capturar relaciones más complejas entre las variables. Además, es importante monitorear de manera continua el desempeño del modelo en entornos del mundo real y realizar ajustes según sea necesario para garantizar su relevancia y precisión a lo largo del tiempo. En última instancia, la transparencia y la comunicación abierta sobre las limitaciones del modelo son esenciales para garantizar su uso responsable y efectivo en la toma de decisiones financieras y empresariales.

## 6.2. Impacto del modelo en la gestión de cobros

El impacto del modelo en la gestión de cobros en entidades financieras es significativo y se deriva de sus objetivos específicos y funcionalidades propuestas.

En primer lugar, al identificar los factores que influyen en la probabilidad de recuperación en caso de impago, el modelo proporciona una comprensión más profunda de los elementos que afectan el éxito de la recuperación de créditos. Esto permite a las entidades financieras adoptar un enfoque más proactivo y estratégico en la gestión de sus carteras de préstamos, al dirigir sus recursos y esfuerzos hacia aquellos casos con mayores probabilidades de recuperación.

La construcción de un modelo predictivo que estime la probabilidad de recuperación para cada operación brinda a las instituciones financieras una herramienta poderosa para evaluar el riesgo de incumplimiento en sus carteras y tomar decisiones informadas sobre las estrategias de cobro. Al predecir la probabilidad de recuperación de manera precisa y oportuna, el modelo permite a las entidades financieras priorizar sus esfuerzos de recuperación en aquellos casos con mayor potencial de éxito, maximizando así la eficiencia y reduciendo los costos asociados con la gestión de cobros.

La validación del modelo con una muestra independiente de datos garantiza su fiabilidad y precisión en entornos del mundo real. Al demostrar su capacidad para generalizar patrones y hacer predicciones precisas en datos no vistos previamente, el modelo se convierte en una herramienta confiable y robusta para la gestión de cobros en entidades financieras.

Analizar la sensibilidad del modelo a diferentes escenarios permite a las instituciones financieras evaluar su capacidad de adaptación a cambios en el entorno operativo y las condiciones del mercado. Esto les permite anticipar y mitigar posibles riesgos y volatilidades, así como ajustar sus estrategias de gestión de cobros en consecuencia para garantizar su efectividad a lo largo del tiempo.

El desarrollo de una herramienta que permita a las entidades financieras utilizar el modelo para la selección de casos con mayor potencial de recuperación mediante la demanda judicial representa un avance significativo en la optimización de los procesos de cobro.

Al integrar el modelo en sus sistemas y procesos existentes, las instituciones financieras pueden automatizar y agilizar la identificación y selección de casos con mayores probabilidades de éxito en la recuperación, lo que mejora la eficiencia operativa y maximiza los resultados financieros.

## 7. Conclusiones y líneas de investigación futuras

### 7.1. Resumen de los resultados y su significancia

Los resultados obtenidos de la implementación del modelo de predicción para estimar la probabilidad de recuperación en casos de impago en operaciones de financiación al consumo son altamente significativos y prometedores.

En primer lugar, el modelo ha demostrado una precisión destacada en la predicción de la probabilidad de recuperación, lo que indica su capacidad para identificar con precisión aquellos casos con mayores posibilidades de éxito en la recuperación de créditos impagados. Esta precisión se ha validado mediante métricas de rendimiento como la precisión, la sensibilidad y la especificidad, las cuales han arrojado resultados consistentes y confiables.

Además, el modelo se ha mostrado robusto frente a diferentes conjuntos de datos y escenarios, lo que sugiere su capacidad para adaptarse y mantener un rendimiento constante en entornos del mundo real a lo largo del tiempo. Esto se ha logrado mediante pruebas de sensibilidad que evaluaron cómo varían las predicciones del modelo en respuesta a cambios en los datos de entrada, así como a través de la estabilidad observada en su rendimiento a lo largo del tiempo.

La interpretabilidad del modelo también ha sido destacada, lo que garantiza que los resultados sean comprensibles y justificables para los usuarios finales, como analistas financieros y tomadores de decisiones. Esto se ha logrado mediante el uso de técnicas de explicabilidad de modelos que permiten comprender cómo el modelo llega a sus predicciones y qué variables tienen mayor influencia en los resultados, lo que aumenta la confianza en su utilidad y fiabilidad.

En cuanto a su aplicabilidad en entornos del mundo real, el modelo ha demostrado ser técnicamente viable y operativamente eficaz para implementarse en sistemas existentes de entidades financieras.

Además, ha proporcionado un valor agregado significativo al mejorar la gestión del riesgo de impago y la toma de decisiones financieras en comparación con enfoques tradicionales utilizados en el sector.

### 7.2. Limitaciones del estudio y sugerencias para futuras investigaciones

A pesar de los logros y la significancia de los resultados obtenidos, este estudio también presenta ciertas limitaciones que deben ser consideradas al interpretar sus conclusiones y que sugieren áreas para futuras investigaciones.

En primer lugar, es crucial destacar que el proyecto se basa en un caso real implementado en la empresa Pepper Finance. Sin embargo, debido a acuerdos de confidencialidad y políticas internas de la empresa, no se pueden revelar más detalles sobre la naturaleza específica de la implementación, así como datos sensibles o identificativos relacionados con la empresa y sus operaciones.

Esta restricción de información impone un límite en la profundidad del análisis que se puede presentar en este trabajo. Aunque se ha trabajado diligentemente para brindar una visión comprensiva y significativa del proyecto, es importante reconocer que ciertos detalles específicos que podrían enriquecer aún más la comprensión y la aplicabilidad del estudio no están disponibles para su divulgación.

Otra limitación está relacionada con la selección de variables y la construcción del modelo. Aunque se ha realizado un proceso exhaustivo para identificar las variables relevantes y construir un modelo predictivo sólido, es posible que se hayan pasado por alto algunas variables importantes o que se hayan simplificado ciertos aspectos del modelo. Explorar diferentes enfoques de modelado y considerar una gama más amplia de variables podría mejorar aún más la precisión y la utilidad del modelo.

Además, es importante reconocer que el modelo se ha validado utilizando datos históricos y puede haber limitaciones en su capacidad para predecir el comportamiento futuro en entornos cambiantes. Los cambios en las condiciones del mercado, las regulaciones gubernamentales, o el comportamiento del cliente pueden afectar la precisión del modelo a lo largo del tiempo. Por lo tanto, sería beneficioso realizar una validación continua y actualizar el modelo periódicamente para garantizar su relevancia y precisión en situaciones del mundo real.

Aunque estas limitaciones restringen la amplitud y la profundidad del estudio, se confía en que el análisis presentado ofrece una perspectiva valiosa y relevante para el campo de estudio en cuestión, así como una base sólida para futuras investigaciones y desarrollos en el ámbito de finanzas y tecnología.

En cuanto a futuras investigaciones, se podrían explorar varias direcciones. Por ejemplo, se podría investigar más a fondo el impacto de variables específicas en la probabilidad de recuperación y cómo estas pueden variar según diferentes segmentos de clientes o tipos de operaciones. También se podrían desarrollar modelos más avanzados que incorporen técnicas de aprendizaje profundo o enfoques de modelado más complejos para mejorar aún más la precisión y la interpretabilidad del modelo.

Además, sería interesante investigar cómo el modelo podría adaptarse a diferentes contextos geográficos o culturales, ya que las características y comportamientos de los clientes pueden variar en diferentes regiones o países. También se podrían explorar aplicaciones adicionales del modelo en áreas como la gestión de carteras, la evaluación de riesgos y la personalización de productos financieros.

### 7.3. Aplicaciones prácticas del modelo en el sector financiero.

El modelo desarrollado ofrece varias aplicaciones prácticas que pueden beneficiar significativamente a las entidades financieras en la gestión de cobros y la mitigación del riesgo de impago.

En primer lugar, al identificar los factores que influyen en la probabilidad de recuperación en caso de impago, el modelo permite a las entidades financieras comprender mejor los riesgos asociados con cada operación y tomar decisiones más informadas sobre cómo gestionar sus carteras de préstamos. Esto les permite optimizar la gestión de cobros al dirigir recursos y esfuerzos hacia aquellos casos con mayor potencial de recuperación, lo que a su vez reduce los costos operativos y aumenta la eficiencia en el proceso de cobranza.

Además, al construir un modelo predictivo que estime la probabilidad de recuperación para cada operación, las entidades financieras pueden mejorar significativamente su capacidad para tomar decisiones estratégicas sobre cómo asignar recursos y priorizar las actividades de cobranza.

Por ejemplo, el modelo puede ayudar a identificar las operaciones que tienen una alta probabilidad de éxito judicial, lo que permite a las entidades financieras enfocar sus esfuerzos en la recuperación de esos casos y minimizar la pérdida de ingresos debido a la morosidad.

Otra aplicación práctica del modelo es su capacidad para minimizar el riesgo de impago al mejorar la estabilidad financiera de las entidades financieras. Al proporcionar una herramienta predictiva precisa, el modelo ayuda a las entidades financieras a identificar y gestionar proactivamente los riesgos de impago, lo que les permite anticiparse a posibles pérdidas y tomar medidas preventivas para mitigar su impacto en la salud financiera de la organización. Esto puede ayudar a reducir la volatilidad en los ingresos y mejorar la capacidad de las entidades financieras para mantener su liquidez y solidez financiera a largo plazo.

## 8. Bibliografía


Angulo Gutiérrez, V. M., & Jara Flores, J. A. (2023). Modelo Machine Learning con transformación logarítmica y validación cruzada para estimar la confiabilidad en un sistema de molienda de Cemento.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, *32*, 100577.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, *9*(1), 381-386.

Malagón-Selma, M. D. P., Debón, A., & Ferrer, A. (2022). Modelos de machine learning y estadística multivariante para predecir la posición de los equipos de primera división. *Journal of Sports Economics & Management, 12(*1), 3-22.

Tello, M. L., Eslava, H. J., & Tobías, L. B. (2013). Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técncias de minería de datos. *Visión electrónica*, *7*(1), 13-26.