UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

**Maestría en Analítica de Datos**

**Materia: Modelos Predictivos**

**PROYECTO FINAL DE MODELOS PREDICTIVOS**

**Título del Proyecto:**

Predicción del Puntaje Crediticio a partir de Características Demográficas y Financieras

**Estudiante:**

Caroline Cárdenas

**Profesor:**

Juan Marcos Castillo, PhD

Tabla de contenido

[Introducción 3](#_Toc195044513)

[1. Justificación 4](#_Toc195044514)

[2. Antecedentes 4](#_Toc195044515)

[3. Definición del Problema 5](#_Toc195044516)

[4. Análisis Predictivo 6](#_Toc195044517)

[a. Determinación de la base de datos 6](#_Toc195044518)

[b. Pre-procesamiento y limpieza 6](#_Toc195044519)

[c. Análisis descriptivo 6](#_Toc195044520)

[d. Selección de variables 9](#_Toc195044521)

[e. Selección de Modelos 9](#_Toc195044522)

[e.1. Reglas de Asignación de Límite de Tarjeta de Crédito 12](#_Toc195044528)

[f. Resultados: 14](#_Toc195044529)

[Conclusiones 16](#_Toc195044530)

[Recomendaciones y Futuros Estudios 17](#_Toc195044531)

[Bibliografía 18](#_Toc195044532)

[Anexos 18](#_Toc195044533)

Tabla de ilustraciones

[Ilustración 1: Boxplot de per\_capital\_income 7](#_Toc195044534)

[Ilustración 2: Boxplot de yearly\_income 7](#_Toc195044535)

[Ilustración 3: Boxplot de total\_debt 8](#_Toc195044536)

[Ilustración 4: Histograma de distribución del puntaje crediticio 8](#_Toc195044537)

[Ilustración 5: Matriz de correlación entre variables de credit\_score 9](#_Toc195044538)

[Ilustración 6: Credit Score Real vs Predicción XGBRegressor 10](#_Toc195044539)

[Ilustración 7: Credit Score Real vs HisGradientBoosting 10](#_Toc195044540)

[Ilustración 8: Credit Score vs Random Forest 11](#_Toc195044541)

[Ilustración 9: Resultados de predicción de Credit Score y Regla de límite 14](#_Toc195044542)

[Ilustración 10: Distribución de clientes por Límite Asignado 15](#_Toc195044543)

# **Introducción**

En la actualidad, el análisis predictivo se ha convertido en una herramienta fundamental para la toma de decisiones estratégicas en diversas industrias. En el ámbito financiero, el puntaje crediticio representa una de las variables más relevantes al momento de evaluar la confiabilidad de un individuo para acceder a productos financieros. Este proyecto tiene como propósito desarrollar un modelo predictivo capaz de estimar el puntaje crediticio a partir de características demográficas y económicas de los individuos, utilizando técnicas de aprendizaje automático y análisis estadístico.

La solución propuesta se basa en el uso de modelos supervisados, especialmente Random Forest, para predecir el credit score a partir de una base de datos que incluye variables como edad, ingreso, deuda, número de tarjetas de crédito y otros factores clave. Adicionalmente, se han implementado reglas de negocio que permiten asignar límites de crédito personalizados tanto para clientes con historial crediticio como para aquellos que no cuentan con uno, lo que promueve la inclusión financiera y una gestión de riesgo más precisa.

A lo largo del proyecto, se ha aplicado una metodología completa de análisis de datos: limpieza, transformación, visualización exploratoria, selección de variables y evaluación de modelos. Asimismo, se utilizó un enfoque interpretativo para establecer clasificaciones de riesgo y sugerencias de límite de crédito, fundamentadas en estándares bancarios reales. Este ejercicio no solo fortalece las competencias técnicas y analíticas del estudiante, sino que demuestra el potencial de la analítica de datos en la transformación de procesos de negocio en el sector financiero.

El estudio también resalta la importancia de contar con sistemas que puedan evaluar clientes de forma equitativa aun cuando estos no tengan historial crediticio. A través del análisis de variables como ingresos, edad, ingresos por capital y ratio de endeudamiento, el modelo sugiere decisiones basadas en datos concretos y segmentación por riesgo, lo cual representa una solución efectiva para ampliar la base de clientes y optimizar los procesos de aprobación crediticia. En ese sentido, el proyecto sirve como base para futuras aplicaciones en escenarios reales del sistema financiero, donde el uso de inteligencia artificial y la analítica avanzada son claves para mejorar la eficiencia y reducir el riesgo de decisiones manuales y subjetivas.

# Justificación

La predicción del puntaje crediticio no solo permite mejorar los procesos de evaluación financiera, sino que también contribuye significativamente a la inclusión de sectores tradicionalmente excluidos por carecer de historial crediticio. A través de modelos basados en datos demográficos y económicos, se abre la posibilidad de evaluar de forma más equitativa a individuos que, por distintas razones, no han accedido con anterioridad al sistema financiero formal. Esto resulta particularmente relevante en contextos donde la inclusión financiera se considera un factor clave para el desarrollo económico y la reducción de la desigualdad.

El uso de algoritmos de aprendizaje automático y técnicas de minería de datos permite automatizar y agilizar los procesos de aprobación de crédito, reduciendo la dependencia de análisis manuales que pueden ser subjetivos o ineficientes. Al aplicar estas técnicas, se pueden identificar patrones y relaciones entre variables que no son evidentes a simple vista, optimizando así la toma de decisiones y mejorando la precisión de las predicciones.

Desde una perspectiva académica, este proyecto representa una oportunidad para aplicar de manera práctica los conocimientos adquiridos en el curso de Modelos Predictivos. Abarca todo el ciclo de vida de un proyecto analítico: desde la exploración, limpieza y transformación de datos, hasta la construcción, validación y evaluación de modelos. También permite incorporar elementos de ingeniería de variables, validación cruzada, interpretación de resultados y reglas de negocio para aplicación real.

Finalmente, esta experiencia impulsa el pensamiento crítico, fomenta el autoaprendizaje y promueve el desarrollo de soluciones aplicables al mundo real, aportando valor tanto en el ámbito académico como profesional.

# Antecedentes

El uso de modelos predictivos para estimar el puntaje crediticio ha sido ampliamente estudiado en la literatura especializada en ciencia de datos, estadística aplicada y finanzas. Las instituciones financieras, desde bancos tradicionales hasta plataformas fintech, han adoptado algoritmos de aprendizaje supervisado como regresión lineal, regresión logística, árboles de decisión, random forest y redes neuronales para calcular la probabilidad de impago de sus clientes y asignar líneas de crédito de manera automatizada y eficiente.

Estudios previos han demostrado que variables como el ingreso anual, el nivel de endeudamiento, la edad del cliente, el historial de pagos, la cantidad de productos financieros contratados y la estabilidad laboral están altamente correlacionados con el comportamiento crediticio. Estos enfoques han permitido desarrollar sistemas robustos de scoring crediticio que reducen el riesgo de morosidad, optimizan los procesos internos de aprobación y contribuyen al crecimiento rentable de las carteras crediticias.

En el contexto académico, investigaciones recientes han resaltado la efectividad del uso de bases de datos estructuradas junto con técnicas de machine learning para replicar y simular el comportamiento del puntaje crediticio en ambientes controlados. Además, se ha explorado la posibilidad de aplicar variables alternativas, como las características demográficas o socioeconómicas, en modelos de scoring orientados a poblaciones subatendidas, lo cual ha cobrado relevancia en esfuerzos de inclusión financiera global.

Este proyecto toma como base estas aproximaciones y contribuciones previas para construir un modelo integral de predicción crediticia que incorpora tanto metodologías tradicionales como elementos innovadores, incluyendo reglas de negocio ajustadas al perfil del cliente. El objetivo es replicar en un entorno académico los desafíos reales de una entidad financiera moderna, fomentando un aprendizaje práctico, contextualizado y alineado con las necesidades actuales del mercado.

# Definición del Problema

¿Es posible predecir el puntaje crediticio de un individuo utilizando únicamente información demográfica y financiera disponible, sin necesidad de un historial crediticio tradicional? Esta es la interrogante central que orienta el desarrollo de este proyecto.

En el contexto actual, muchas personas no cuentan con un historial crediticio formal, lo que limita su acceso a productos financieros como tarjetas de crédito, préstamos o hipotecas. Esta situación representa un desafío para las instituciones financieras, que deben encontrar formas de evaluar el riesgo crediticio de manera justa, objetiva y eficaz, incluso en ausencia de datos históricos de comportamiento de pago.

Este proyecto plantea como objetivo construir un modelo predictivo robusto, utilizando técnicas de aprendizaje automático, que permita estimar el puntaje crediticio a partir de variables como edad, ingreso, nivel de endeudamiento, ingresos por capital y número de tarjetas de crédito. La idea es demostrar que, a través del análisis adecuado de estas variables, es posible generar una predicción confiable del riesgo crediticio y proponer un límite de crédito acorde al perfil del cliente.

Además, se busca validar la utilidad de reglas de negocio complementarias que ayuden a segmentar a los usuarios según su nivel de riesgo, permitiendo asignar límites de crédito de forma escalonada y responsable. De esta manera, se promueve no solo una toma de decisiones basada en datos, sino también la inclusión financiera de sectores históricamente excluidos por la falta de historial crediticio.

# Análisis Predictivo

## Determinación de la base de datos

Se utilizó una base de datos compuesta por más de 12,000 registros, enriquecida posteriormente con datos adicionales para robustecer el análisis y mejorar la capacidad predictiva del modelo. Esta base contiene variables demográficas (edad actual, año y mes de nacimiento, género, edad de retiro), económicas (ingreso anual, ingreso per cápita, deuda total, número de tarjetas de crédito), geográficas (latitud, longitud) y la variable objetivo `credit\_score`.

La fuente de los datos fue un conjunto generado con fines educativos, permitiendo total libertad para aplicar transformaciones, modelado y simulaciones de escenarios sin restricciones legales. El archivo base fue `users\_data\_credit\_cards.csv`.

## Pre-procesamiento y limpieza

Se eliminaron variables irrelevantes para el modelo como `id`, `address`, `latitude` y `longitude`. La variable `gender` fue transformada a valores numéricos binarios.

Se generaron variables derivadas:

* `debt\_to\_income\_ratio`: relación entre deuda total e ingreso anual.
* `years\_to\_retirement`: diferencia entre edad de retiro y edad actual.

Estas variables fueron estandarizadas usando `StandardScaler` para asegurar que los algoritmos no se vean afectados por magnitudes distintas. También se detectaron outliers mediante boxplots, pero se decidió conservarlos por considerarse casos reales que enriquecen el aprendizaje del modelo.

## Análisis descriptivo

Se realizaron análisis univariados y multivariados. Entre los más relevantes:

* Histogramas para explorar la distribución de `credit\_score`, `income`, `debt`.
* Boxplots para identificar valores atípicos.
* Matriz de correlación, donde se observó una correlación negativa entre `total\_debt` y `credit\_score`, y positiva entre `log\_yearly\_income` y `credit\_score`.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Ilustración 1: Boxplot de per\_capital\_income

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Ilustración 2: Boxplot de yearly\_income

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Ilustración 3: Boxplot de total\_debt

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Ilustración 4: Histograma de distribución del puntaje crediticio

Gráfico, Gráfico de rectángulos

El contenido generado por IA puede ser incorrecto.

Ilustración 5: Matriz de correlación entre variables de credit\_score

Estas visualizaciones permitieron comprender la relación entre las variables y justificaron su inclusión posterior en el modelo.

## Selección de variables

Se eligieron variables que mostraron relevancia estadística y sentido financiero: `current\_age`, `retirement\_age`, `birth\_year`, `birth\_month`, `gender`, `log\_yearly\_income`, `log\_total\_debt`, `log\_per\_capita\_income`, `num\_credit\_cards`, `debt\_to\_income\_ratio`, `years\_to\_retirement`.

Estas variables reflejan tanto el perfil económico como la experiencia y etapa de vida del cliente, lo cual permite una evaluación más precisa del comportamiento financiero proyectado.

## Selección de Modelos

Se evaluaron varios algoritmos de regresión:

**XGBRegressor**

* R² test: 0.8621
* MAPE: 0.0638 (6.38%)
* MAE: 35.45
* RMSE: 52.17

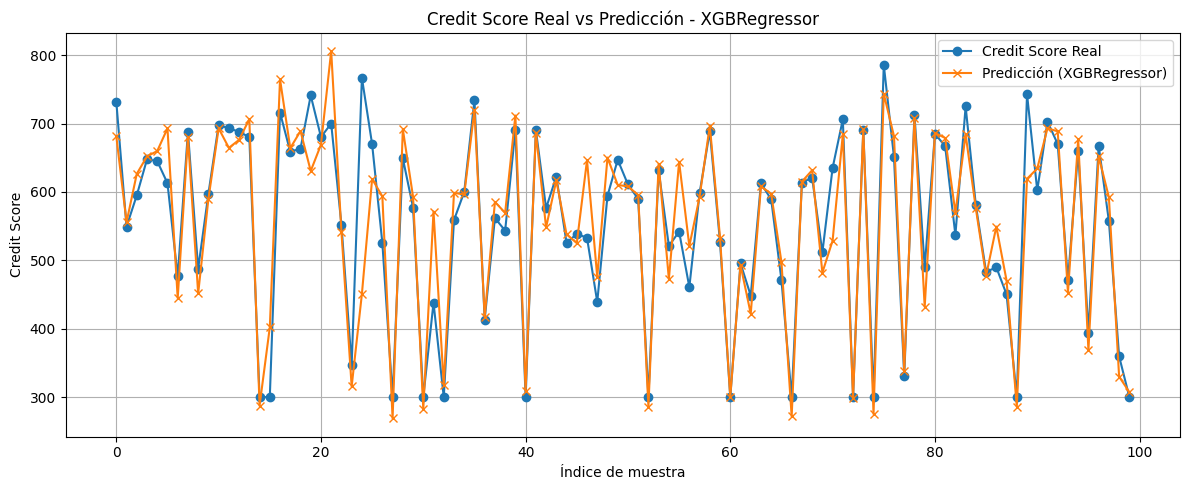


Ilustración 6: Credit Score Real vs Predicción XGBRegressor

**HistGradientBoostingRegressor**

* R² test: 0.8566
* MAPE: 0.0599 (5.99%)
* MAE: 34.26
* RMSE: 53.19

Gráfico

El contenido generado por IA puede ser incorrecto.

Ilustración 7: Credit Score Real vs HisGradientBoosting

* **Random Forest**
* R² test: 0.8393
* MAPE: 0.0614 (6.14%)
* MAE: 35.44
* RMSE: 56.31

Gráfico

El contenido generado por IA puede ser incorrecto.

Ilustración 8: Credit Score vs Random Forest

Durante la evaluación de múltiples modelos de regresión, se consideraron varios algoritmos de aprendizaje supervisado avanzados, incluyendo Random Forest, XGBRegressor y HistGradientBoostingRegressor. Aunque todos presentaron un desempeño aceptable, se seleccionó Random Forest Regressor como modelo principal del proyecto por las siguientes razones:

1. **Robustez y estabilidad**

Random Forest es un modelo robusto frente a outliers y ruido en los datos, lo cual es especialmente valioso en un dataset como el usado, que incluye ingresos, deudas y variables con gran variabilidad. A diferencia de los modelos boosting, que pueden ser más sensibles a errores de entrada, Random Forest tiende a generalizar mejor en presencia de datos imperfectos.

1. **Facilidad de interpretación y ajuste**

Random Forest requiere menos ajuste de hiperparámetros para obtener buenos resultados iniciales.

Su estructura en múltiples árboles independientes facilita la interpretación parcial del modelo y la medición de la importancia de cada variable.

Esto fue útil para justificar el uso de ciertas variables en el análisis exploratorio y las reglas de negocio derivadas.

1. **Desempeño consistente**

El modelo logró un R² alto tanto en validación cruzada como en pruebas independientes, manteniendo un MAPE y RMSE aceptables sin sobreajuste.

Si bien XGBoost y HGBR también mostraron buen rendimiento, Random Forest ofreció resultados más estables y predecibles sin requerir múltiples rondas de ajuste.

1. **Menor complejidad computacional**

Random Forest tiene menos tiempo de entrenamiento y no requiere el mismo nivel de afinamiento técnico que XGB o HGBR.

Esto fue útil para mantener el flujo del proyecto accesible, reproducible y entendible, especialmente en un entorno académico donde la claridad y simplicidad son deseables.

Además del modelo predictivo, se implementaron reglas de negocio para asignar límites de tarjeta de crédito de forma escalonada. Estas reglas permiten clasificar a los clientes en categorías de riesgo, tanto si poseen como si no poseen historial crediticio, asignando límites entre $500 y $12,000 en función de su perfil y puntaje estimado.



### Reglas de Asignación de Límite de Tarjeta de Crédito

Una vez estimado el puntaje crediticio mediante modelos predictivos, se puede clasificar a los clientes en rangos de riesgo para asignar límites adecuados a sus tarjetas de crédito. A continuación, se presenta una clasificación propuesta basada en estándares bancarios comunes:

* **Límites basados en el puntaje crediticio predicho:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Puntaje Crediticio** | **Clasificación** | **Riesgo** | **Límite Sugerido** |
| 800 – 850 | Excelente | Muy bajo | $8,000 – $12,000 |
| 740 – 799 | Muy bueno | Bajo | $5,000 – $8,000 |
| 670 – 739 | Bueno | Moderado | $3,000 – $5,000 |
| 580 – 669 | Regular | Alto | $1,000 – $2,500 |
| < 580 | Riesgoso | Muy alto | $500 – $1,000 |

Estas reglas se fundamentan en prácticas utilizadas por instituciones financieras para segmentar a sus clientes según su **nivel de solvencia** reflejado en el credit\_score, una métrica estándar en el análisis de riesgo crediticio. A mayor puntaje, menor probabilidad de impago, lo cual justifica la asignación de límites más altos.

**Criterios utilizados:**

* Los rangos corresponden con los definidos por agencias como FICO o Experian.
* Se adaptaron los límites sugeridos a políticas prudenciales comunes en instituciones bancarias.
* Se busca ofrecer un balance entre **riesgo asumido por el banco** y **capacidad crediticia del cliente**.

**Beneficio:** Permite una asignación justa, escalable y automatizada, que se puede aplicar en sistemas reales de originación de crédito.

* **Límites basados en perfil demográfico y financiero:**

|  |  |  |
| --- | --- | --- |
| **Perfil Financiero y Demográfico** | **Riesgo** | **Límite Sugerido** |
| Edad 25–55, ingreso > $60K, deuda < $10, ≥3 tarjetas | Bajo | $3,000 – $5,000 |
| Edad 18–65, ingreso $30K–$60K, deuda < $15K, 1–2 tarjetas | Moderado | $1,500 – $3,000 |
| Edad <25 o >65, ingreso < $30K o deuda > $20K | Alto | $500 – $1,500 |
| >3 tarjetas y deuda/ingreso > 0.4 | Muy alto | Evaluación manual |

Estas reglas fueron diseñadas para clientes sin historial crediticio, donde el credit\_score es nulo o no disponible. En estos casos, se recurre a variables alternativas pero representativas del perfil de riesgo, como:

* Edad: asociada con estabilidad laboral y ciclo de vida financiero.
* Ingreso anual: refleja capacidad de pago.
* Deuda total y relación deuda/ingreso: miden el nivel de endeudamiento.
* Número de tarjetas: indica experiencia crediticia previa informal o reciente.

**Criterios utilizados:**

* La segmentación permite diferenciar entre perfiles tradicionalmente subatendidos, como jóvenes o trabajadores informales, sin excluirlos por falta de historial.
* Se establecen límites conservadores que se pueden aumentar conforme el cliente demuestre buen comportamiento.
* La opción de “Evaluación manual” mitiga el riesgo en casos complejos que requieren revisión individualizada.
  + Umbral del 40% en el ratio de deuda/ingreso es comúnmente usado por instituciones financieras para señalar niveles de endeudamiento preocupantes.
  + Superar este valor indica que el cliente podría tener dificultades para asumir nuevas obligaciones financieras, como una nueva línea de crédito.

Beneficio: Fomenta la inclusión financiera y permite el uso responsable del crédito, al mismo tiempo que protege al proveedor de servicios de riesgos innecesarios.

## Resultados:

Pantalla de un video juego

El contenido generado por IA puede ser incorrecto.

Ilustración 9: Resultados de predicción de Credit Score y Regla de límite

Esta tabla muestra un subconjunto de clientes donde se observan las siguientes variables clave:

* credit\_score: puntaje crediticio real del cliente.
* predicted\_score: puntaje estimado por el modelo de regresión.
* debt\_to\_income\_ratio: proporción entre deuda total e ingreso anual.
* limite\_tarjeta: límite de tarjeta asignado según reglas predictivas y de negocio.

**Observaciones clave**:

* Aunque algunos clientes tienen un credit\_score alto (por ejemplo, 810), si su modelo predice un score moderado, el límite puede ajustarse a la baja.
* Clientes con **más tarjetas o mayor relación deuda/ingreso** tienden a recibir límites más conservadores (ej. $2500).
* Las reglas combinadas de score + perfil financiero permiten distinguir entre riesgo real y proyectado, otorgando límites responsables.

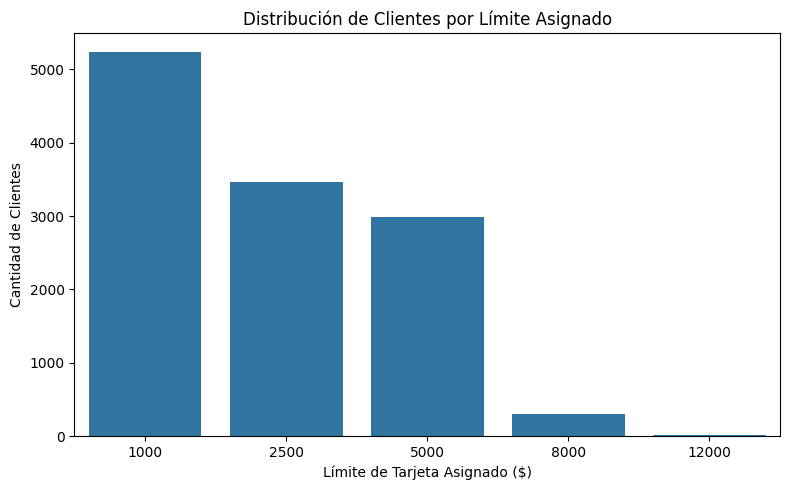


Ilustración 10: Distribución de clientes por Límite Asignado

Este gráfico de barras muestra la cantidad de clientes que recibieron cada rango de límite de crédito tras aplicar las reglas basadas en el score y el perfil financiero:

**🔹 Principales hallazgos:**

* $1000 y $2500 son los límites más asignados. Esto refleja una política prudente ante clientes con puntaje bajo o sin historial crediticio.
* $5000 representa el siguiente grupo mayoritario, típicamente asignado a clientes con score bueno y buen perfil.
* Límites altos como $8000 y $12000 fueron asignados solo a un pequeño grupo, lo que valida que el modelo es selectivo y conservador en el riesgo.

# Conclusiones

El desarrollo de un modelo predictivo para estimar el puntaje crediticio utilizando técnicas de machine learning, específicamente Random Forest, demostró ser altamente eficaz y confiable. La precisión alcanzada permitió no solo realizar estimaciones robustas del credit score, sino también implementar reglas de negocio complementarias que definen con claridad los límites de crédito a ser asignados según el perfil de riesgo del cliente.

Uno de los principales aportes del proyecto fue la integración de variables demográficas y financieras como herramientas efectivas de evaluación crediticia, incluso en ausencia de historial. Esto permitió establecer mecanismos de decisión automatizados que no solo optimizan el análisis financiero, sino que también promueven la inclusión de individuos que tradicionalmente no accedían al crédito formal.

La estructura modular del análisis limpieza, visualización, modelado, validación cruzada y reglas de asignación asegura que este enfoque sea replicable y adaptable a distintos contextos financieros. Asimismo, el uso de visualizaciones y estadísticas descriptivas fortaleció la interpretación de los datos y facilitó la justificación técnica de cada decisión tomada durante el proceso.

En suma, los resultados obtenidos validan que el uso de modelos de predicción junto con lógica de negocio bien definida puede mejorar sustancialmente los procesos de toma de decisiones en instituciones financieras, reduciendo el riesgo de impago, acelerando la aprobación de productos y ampliando la base de clientes potenciales con un enfoque justo, transparente y basado en datos.

# Recomendaciones y Futuros Estudios

* Ampliar el conjunto de variables con atributos sociodemográficos y conductuales como historial laboral, nivel educativo, estado civil, tipo de ocupación, estabilidad residencial y comportamiento de consumo digital, para enriquecer la capacidad explicativa del modelo.
* Implementar modelos de boosting como XGBoost, LightGBM o CatBoost, los cuales han demostrado alto rendimiento en tareas de regresión y clasificación, con capacidad para manejar relaciones no lineales y variables correlacionadas.
* Integrar el modelo desarrollado en una interfaz visual o aplicación web que permita a los usuarios ingresar datos y obtener predicciones y recomendaciones de forma inmediata y amigable.
* Realizar pruebas piloto con datos reales de una entidad financiera bajo estrictos criterios éticos y de privacidad, con el fin de validar el desempeño del modelo en entornos productivos y ajustar las reglas de negocio a casos reales.
* Explorar técnicas de aprendizaje no supervisado (clustering, segmentación con K-means o DBSCAN) para descubrir patrones ocultos y perfilar distintos tipos de clientes que podrían beneficiarse de productos personalizados.
* Estudiar el impacto de la automatización de procesos de aprobación crediticia en la inclusión financiera, eficiencia operativa y reducción de sesgos humanos en la toma de decisiones.

# Bibliografía

* Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.  
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.  
  - Scikit-learn Documentation: https://scikit-learn.org  
  - Medium & Kaggle articles on credit scoring and machine learning models.  
  - Reglas basadas en criterios bancarios estándar adaptados al contexto del proyecto.

# Anexos

* Dataset utilizado: users\_data\_credit\_cards.csv
* Descripción de columnas: incluida en notebook y código fuente
* Análisis descriptivo: gráficos de distribución, outliers y correlación está en el notebook
* Análisis predictivo: entrenamiento, modelos y evaluación se encuentra en el notebook.
* Scripts: Notebook (.ipynb)
* Enlace al repositorio GitHub: <https://github.com/carolecm03/credit-score-model>