# Construcción de scores crediticios apoyado de matrices de transición para entidades financieras

Artículo presentado como requisito parcial para el curso
*Herramientas de Datos 2*

Profesor: M.Sc. Luis Alberto Juárez Potoy

Autores:

Joseph Romero Chinchilla - C37006
Sebastián Astúa Morales - C10763
Pablo Alonso Chaves Madrigal - B82150

Universidad de Costa Rica
Facultad de Ciencias Actuariales y Matemáticas Puras
Junio, 2025

# 1. Introducción

Los *scores* crediticios son herramientas fundamentales para las entidades financieras. Según MyFICO (s.f.), "un puntaje crediticio informa a los prestamistas sobre su solvencia, es decir, la probabilidad de que pague un préstamo según su historial crediticio. Se calcula utilizando la información de sus informes crediticios" (MyFICO, s.f.).

Estos puntajes se construyen a partir de variables como el historial de pagos, el nivel de utilización del crédito disponible, la antigüedad del historial crediticio, la diversidad de tipos de crédito y otros factores relevantes. Así, permiten segmentar a los clientes según su nivel de riesgo, facilitando decisiones como la aprobación o rechazo de solicitudes, así como el monto y condiciones del crédito otorgado.

Una aplicación valiosa de los *scores* crediticios es su uso en la construcción de matrices de transición, que permiten modelar la evolución de los estados crediticios de los individuos en el tiempo. Según Castro Ropero y Quintero Torres (2022), "la matriz de transición/incumplimiento crediticio es una herramienta clave para el análisis del riesgo crediticio, la cual muestra el deterioro o mejora de la calidad crediticia de un agente entre dos o más periodos" (Castro Ropero & Quintero Torres, 2022).

En este contexto, el presente trabajo se propone desarrollar un modelo actuarial basado en matrices de transición que permita analizar la dinámica de los estados crediticios y estimar la probabilidad de impago. Este enfoque actuarial busca aprovechar técnicas probabilísticas y estadísticas propias de las ciencias actuariales para cuantificar el riesgo asociado al comportamiento crediticio de los clientes a lo largo del tiempo. De este modo, se busca responder la siguiente pregunta de investigación:

> ¿Cómo cambian los estados de pago de los clientes en el tiempo y qué características están asociadas con una mayor probabilidad de impago, según un modelo actuarial basado en matrices de transición?

# 2. Exploración de Datos

Antes de hacer ejecutar los modelos, se realizó una exploración inicial del dataset para entender su estructura y características principales:

- **Dimensiones y tipos de datos:** El dataset cuenta con XX filas y YY columnas, donde se identificaron variables numéricas y categóricas.
- **Estadísticas descriptivas:** Se analizaron las medidas de tendencia central y dispersión de las variables numéricas, así como la frecuencia de categorías en las variables cualitativas.
- **Variables categóricas:** Se examinaron las variables categóricas, identificando la cantidad de valores únicos y su distribución.
- **Correlación:** Se calculó la matriz de correlación entre variables numéricas, y se visualizó mediante un mapa de calor, detectando relaciones significativas entre variables.
- **Visualización:** Se crearon histogramas para variables numéricas, mostrando su distribución, y gráficos de barras para variables categóricas, revelando patrones en los datos.

## 2.1. Resultados de la Exploración

- Se notaron variables con distribuciones sesgadas que podrían afectar el rendimiento del modelo.
- Las variables categóricas presentan distintas cardinalidades, lo cual se consideró para su codificación.
- La matriz de correlación indicó que algunas variables numéricas están fuertemente correlacionadas, lo que puede influir en la multicolinealidad.

## 2.2. Resultados de los Modelos

### 2.2.1. Random Forest

- **Mejores parámetros:** max_depth = None, max_features = None, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 200.

- **Accuracy en test:** 93.3 %
- **Matriz de confusión:**

$$\begin{bmatrix} 4430 & 33 \\ 348 & 890 \end{bmatrix}$$

- **Reporte de clasificación:**

| Clase | Precisión | Recall | F1-score | Soporte |
|---|---|---|---|---|
| 0 | 0.93 | 0.99 | 0.96 | 4463 |
| 1 | 0.96 | 0.72 | 0.82 | 1238 |
| Accuracy | | 0.93 | | |
| Macro promedio | 0.95 | 0.86 | 0.89 | 5701 |
| Promedio ponderado | 0.94 | 0.93 | 0.93 | 5701 |

#### 2.2.2. Gradient Boosting

- **Accuracy:** 93.2 %
- **Matriz de confusión:**

$$\begin{bmatrix} 4423 & 40 \\ 343 & 895 \end{bmatrix}$$

- **Reporte de clasificación:**

| Clase | Precisión | Recall | F1-score | Soporte |
|---|---|---|---|---|
| 0 | 0.93 | 0.99 | 0.96 | 4463 |
| 1 | 0.96 | 0.72 | 0.82 | 1238 |
| Accuracy | | 0.93 | | |
| Macro promedio | 0.94 | 0.86 | 0.89 | 5701 |
| Promedio ponderado | 0.93 | 0.93 | 0.93 | 5701 |

# 3. Metodología

Este estudio adopta un enfoque cuantitativo, utilizando técnicas de aprendizaje supervisado para modelar la probabilidad de impago de clientes en función de sus características demográficas y financieras. Se emplean los modelos de regresión logística, *Random Forest* y *Gradient boosting* debido a su eficacia. La implementación se realiza en Python, aprovechando su ecosistema de bibliotecas especializadas en análisis de datos y aprendizaje automático.

## 3.1. Herramientas y tecnologías

- **Python**: Lenguaje de programación de propósito general, ampliamente utilizado en análisis de datos y aprendizaje automático por su sintaxis clara y la disponibilidad de bibliotecas especializadas como `pandas` y `matplotlib`.(Rodríguez Rivas, 2022).

- **Jupyter Notebook**: Entorno interactivo que permite combinar código, visualizaciones y texto explicativo, facilitando la documentación y presentación del análisis.

- **GitHub**: Plataforma de control de versiones utilizada para almacenar y gestionar el código fuente del proyecto, promoviendo la colaboración y el seguimiento de cambios.

## 3.2. Descripción de los modelos

**Regresión Logística**.La regresión logística estima la probabilidad de que ocurra un evento, como votar o no votar, en función de un conjunto de datos determinado de variables independientes.Por ejemplo, para predecir la probabilidad de ocurrencia de un evento binario, como el impago de un préstamo. Este modelo estima la relación entre una variable dependiente binaria y una o más variables independientes, utilizando la función logística para garantizar que las predicciones se encuentren en el rango [0,1].(IBM, s.f.-a).

**Random Forest**. Este es un algoritmo de aprendizaje automático basado en la construcción de múltiples árboles de decisión, donde cada árbol es entrenado con una muestra aleatoria del conjunto de datos y una selección aleatoria de variables. La predicción final se obtiene mediante la agregación de las predicciones individuales de los árboles, generalmente por votación mayoritaria.(IBM, s.f.-b). Este enfoque reduce la varianza del modelo y mejora su capacidad de generalización, siendo especialmente útil en conjuntos de datos con alta dimensionalidad

y relaciones no lineales entre variables. Además, *Random Forest* proporciona medidas de importancia de variables, lo que facilita la identificación de los factores más relevantes en la predicción del impago.(Freire López, 2020).

**Gradient boosting**. Este es un algoritmo de aprendizaje por conjuntos que produce predicciones precisas combinando múltiples árboles de decisión en un solo modelo. Fue introducido por Jerome Friedman, y utiliza modelos base para aprovechar sus fortalezas, corrigiendo errores y mejorando la capacidad predictiva. Al capturar patrones complejos en los datos, el *Gradient boosting* destaca en diversas tareas de modelado predictivo. (Clark & Lee, 2025)

# 4. Implementación

El trabajo desarrollado, de manera momentánea, en este proyecto investigativo consistió en la implementación y entrenamiento de modelos para predecir el riesgo crediticio de clientes utilizando un conjunto de datos reales. Para ello, se emplearon dos técnicas de aprendizaje automático supervisado: **Random Forest** y **Gradient Boosting**.

La implementación se dividió en las siguientes etapas:

- **Carga y preparación de datos:** Se cargó el dataset `credit_risk_dataset.csv` y se realizó una limpieza básica, además de una transformación para adecuar los datos al modelo.
- **Entrenamiento:** Se entrenaron los modelos con los datos de entrenamiento, optimizando hiperparámetros mediante validación cruzada para mejorar el rendimiento.
- **Evaluación:** Se evaluaron los modelos usando el conjunto de prueba, calculando métricas como *accuracy*, matriz de confusión y reporte de clasificación con precisión, recall y F1-score.
- **Comparación:** Este apartado se reserva para la siguiente etapa.

# 5. Análisis FODA

## Fortalezas

- La regresión logística, *Random Forest* y *Gradient boosting* son modelos reconocidos por su eficacia en la predicción del riesgo crediticio.
- La implementación en Python, junto con Jupyter y GitHub, permite seguir buenas prácticas en ciencia de datos.
- El *dataset* contiene información clave como ingresos, empleo, historial crediticio, tasa de interés y monto del préstamo, lo cual permite construir modelos con buena capacidad explicativa.
- La metodología permite tanto la predicción como la interpretación de los factores que influyen en el impago, lo cual agrega valor práctico y analítico.

## Oportunidades

- La estructura del análisis puede ser replicada en *datasets* locales en caso de tener acceso a información de bancos nacionales o entes reguladores.
- La base desarrollada puede ser mejorada incluyendo otros algoritmos de aprendizaje automático o técnicas de explicación de modelos más avanzadas.

## Debilidades

- El conjunto de datos utilizado no representa las condiciones crediticias del sistema financiero costarricense, lo que limita la aplicabilidad directa de los resultados.
- Si hay muy pocos casos de impago, los modelos pueden sobreajustarse a la clase mayoritaria y presentar bajo desempeño en la detección de impagos.
- La ausencia de una validación con datos externos o históricos reales limita la capacidad del modelo para generalizar.

## Amenazas

- *Random Forest* puede sobreajustarse si no se controlan adecuadamente los hiperparámetros, especialmente en conjuntos de datos pequeños.
- En instituciones financieras, se requiere trazabilidad y explicación de las decisiones automatizadas, lo cual puede ser difícil de lograr con modelos complejos si no se acompañan de herramientas explicativas.
- Si el conjunto de datos no es lo suficientemente grande o no tiene variables derivadas útiles, la capacidad predictiva y la generalización del modelo pueden verse comprometidas.

# Referencias


Castro Ropero, J. A., & Quintero Torres, I. M. (2022). Propuestas metodológicas para el establecimiento de un score de crédito a partir de metodologías de componentes principales.

Clark, B., & Lee, F. (2025, abril). What is Gradient Boosting? https://www.ibm.com/think/topics/gradient-boosting

Freire López, J. (2020). *Modelo de clasificación de riesgo crediticio utilizando Random Forest* [Tesis de pregrado]. Universidad Internacional SEK. https://repositorio.uisek.edu.ec/handle/123456789/4256

IBM. (s.f.-a). ¿Qué es la regresión logística? [Consultado el 4 de junio de 2025]. https://www.ibm.com/mx-es/topics/logistic-regression

IBM. (s.f.-b). What is random forest? [Consultado el 4 de junio de 2025]. https://www.ibm.com/mx-es/topics/logistic-regression

MyFICO. (s.f.). What is a credit score? [Consultado el 4 de junio de 2025]. https://www.myfico.com/credit-education/credit-scores

Rodríguez Rivas, J. G. (2022). Uso de Python para el análisis de datos aplicado en la investigación. *Revista INCAING*, (Noviembre–Diciembre), 33-40. https://www.researchgate.net/publication/366157405_Uso_de_Python_para_el_analisis_de_datos_aplicado_en_la_investigacion