**Proyecto Final: Modelos de Clasificación Aplicados a la Predicción del Riesgo Crediticio**

Angélica Sepúlveda | Andrés Pedraza | Jefry Llerena

Candidatos a Magíster en Analítica de Datos, Universidad del Norte, Barranquilla

**RESUMEN / ABSTRACT**

La predicción del riesgo crediticio es un componente clave en la toma de decisiones financieras dentro de las entidades crediticias, ya que permite anticipar el comportamiento de pago de los solicitantes de crédito. En este contexto, los modelos de clasificación se utilizan para estimar la probabilidad de incumplimiento, categorizando a los clientes como solventes (TARGET = 0) o propensos al impago (TARGET = 1). El conjunto de datos [Home Credit Default Risk](https://www.kaggle.com/competitions/home-credit-default-risk/data), publicado en [Kaggle](https://www.kaggle.com/competitions/home-credit-default-risk/data), proporciona una base rica y compleja que incluye características socioeconómicas, laborales, demográficas y crediticias de más de 300 mil clientes, lo que permite aplicar y evaluar una variedad de algoritmos de clasificación supervisada (Flores, Malca et al., 2017).

Modelos de machine learning como Random Forest, XGBoost, Clasificación Bayesiana, K-vecinos más cercanos, redes neuronales, entre otros, son comúnmente utilizados para este tipo de tareas debido a su capacidad para manejar datos estructurados y variables tanto numéricas como categóricas. Estos algoritmos permiten no solo predecir el riesgo crediticio con alta precisión, sino también interpretar qué variables tienen mayor peso en la decisión, lo que es fundamental para la transparencia del modelo en entornos financieros regulados. A través de métricas como la AUC-ROC, el F1-score y la matriz de confusión, es posible evaluar el desempeño de los modelos y ajustarlos para minimizar el riesgo de clasificaciones erróneas, especialmente en contextos de datos desbalanceados como los presentes en este conjunto de datos (Kotsiantis et al., 2007).

Credit risk prediction is a key component in financial decision-making within lending institutions, as it helps anticipate the repayment behavior of loan applicants. In this context, classification models are used to estimate the probability of default, categorizing clients as either reliable (TARGET = 0) or likely to default (TARGET = 1). The [Home Credit Default Risk](https://www.kaggle.com/competitions/home-credit-default-risk/data) dataset, published on [Kaggle](https://www.kaggle.com/competitions/home-credit-default-risk/data), offers a rich and complex foundation that includes socioeconomic, employment, demographic, and credit-related features for over 300,000 clients, enabling the application and evaluation of a wide range of supervised classification algorithms (Flores, Malca et al., 2017).

Models such as Logistic Regression, Random Forest, XGBoost, and LightGBM are commonly used for this type of task due to their ability to handle structured data and both numerical and categorical variables. These algorithms not only enable high-accuracy credit risk prediction but also allow for interpretability, highlighting the most influential features a critical aspect in regulated financial environments. Through metrics such as AUC-ROC, F1-score, and the confusion matrix, the performance of the models can be assessed and fine-tuned to minimize the risk of misclassification, particularly in imbalanced datasets like the one presented in this challenge (Kotsiantis et al., 2007).

# PALABRAS CLAVE / KEY WORDS:

EDA, riesgo crediticio, clasificación, Home Credit, preprocesamiento / EDA, credit risk, classification, Home Credit, preprocessing.

**INTRODUCCIÓN**

El presente informe resume el proceso de análisis exploratorio de datos (EDA) aplicado al conjunto de datos [Home Credit Default Risk](https://www.kaggle.com/competitions/home-credit-default-risk/data). El objetivo principal es comprender la estructura y calidad de los datos, identificar patrones relevantes, evaluar la distribución de variables clave y preparar el conjunto de datos para tareas de modelado predictivo orientadas a la clasificación del riesgo de incumplimiento crediticio.

# REVISIÓN LITERARIA

# PREPROCESAMIENTO Y LIMPIEZA DE DATOS

## DESCRIPCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos original está compuesto por múltiples archivos que representan la información financiera, demográfica y conductual de los solicitantes de crédito, como se muestra en la siguiente tabla. El archivo principal applicationtrain.csv contiene 307.511 observaciones y 122 variables, incluyendo la variable objetivo TARGET, la cual indica si un cliente ha incumplido (TARGET = 1) o no (TARGET = 0) con el pago de su crédito.

|  |  |
| --- | --- |
| **Archivo CSV** | **Descripción** |
| application\_train.csv | Datos de clientes que ya recibieron un préstamo, incluyendo la variable TARGET. |
| application\_test.csv | Datos de nuevos clientes sin información de TARGET, para predicciones. |
| bureau.csv | Créditos anteriores reportados por otras instituciones financieras. |
| bureau\_balance.csv | Saldos mensuales asociados a los créditos anteriores (bureau). |
| previous\_application.csv | Historial de solicitudes de crédito anteriores a la actual. |
| POS\_CASH\_balance.csv | Saldos de préstamos tipo punto de venta o efectivo. |
| credit\_card\_balance.csv | Información mensual sobre tarjetas de crédito. |
| installments\_payments.csv | Registro de pagos realizados en cuotas por préstamos anteriores. |
| sample\_submission.csv | Archivo para cargar predicciones en [Kaggle](https://www.kaggle.com/competitions/home-credit-default-risk/data). |
| HomeCredit\_columns\_description.csv | Diccionario de variables con descripciones completas. |

**Tabla 1.** Fuentes de datos de Home Credit Default Risk

**Antes del filtrado**:

* **0:** 79.35% → Clientes sin problemas de pago.
* **1:** 6.97% → Clientes con incumplimientos.
* **NaN:** 13.68% → Eliminados posteriormente.

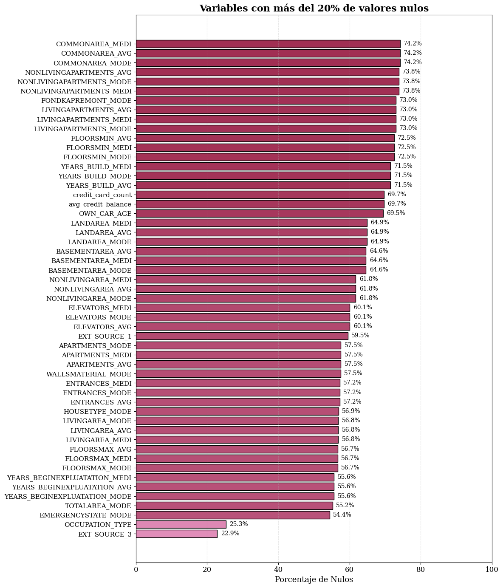
**Análisis**:

* La variable objetivo presenta un fuerte desbalance de clases.
* Es importante tener esto presente para futuras estrategias de muestreo o balanceo como SMOTE o undersampling.

## PREPROCESAMIENTO DE DATOS

### Tratamiento de valores nulos

* Se hace un filtrado del DataFrame original df para conservar solo las filas donde la columna TARGET es igual a 1.
* Lo que indica que hay 24,825 casos positivos y 135 casos negativos.
* Se calcula cuántos valores faltantes (NaN) hay por columna (na\_counts).
* Se calcula el porcentaje de nulos respecto al total de filas (na\_pct).
* Se crea un nuevo DataFrame na\_df con dos columnas:
* columna: nombre de la columna original.
* nulos: porcentaje de valores nulos.
* Filtra na\_df para conservar solo las columnas con más del 20% de valores nulos y las ordena.
* Finalmente se crea un gráfico de barras horizontal donde cada barra representa una columna con más del 20% de nulos:



**Ilustración 1.** Variables con valores nulos mayores al 20%

Con este diagnóstico, se procede a seleccionar los nombres de las columnas cuyo porcentaje de valores nulos supera el 40%, usando el análisis previo sobre el subconjunto df\_target\_1 y se eliminan esas 51 columnas con más del 40% de nulos en los casos TARGET =1 del DataFrame original df (no solo el subconjunto), ya que se consideran poco útiles debido a la gran cantidad de valores faltantes y esto ayuda a limpiar el conjunto sin perder demasiada información relevante para los positivos (los casos más escasos). Las nuevas dimensiones del dataset: 307.511 filas y 84 columnas.

### Reducción por dominancia y correlación

Se eliminaron variables con una sola categoría o donde más del 90 % de los registros compartían el mismo valor. También se removieron aquellas con alta multicolinealidad (VIF ¿10), como DAYS EMPLOYED y AMT CREDIT.

### Transformaciones adicionales

* Se eliminaron valores negativos en variables como total pagado.
* Se aplicaron transformaciones logarítmicas a montos financieros.
* Se agruparon y recodificaron categorías raras en variables categóricas.
* Se banalizaron variables de conteo para facilitar su análisis.
  1. **Preparación y limpieza**

Se cargaron y consolidaron múltiples fuentes de datos proporcionadas por la competencia, integrando información sobre solicitudes actuales, historiales crediticios, pagos, y consultas a buró. Posteriormente se aplicaron transformaciones como:

* Eliminación de variables con más del 40 % de valores nulos.
* Supresión de columnas con una sola categoría dominante (>90 %).
* Detección de colinealidad mediante VIF y eliminación de variables redundantes.
* Normalización y transformación logarítmica de montos financieros.

## Reducci´on por dominancia y correlacio´n

Se eliminaron variables con una sola categor´ıa o donde m´as del 90 % de los registros compart´ıan el mismo valor. Tambi´en se removieron aquellas con alta multicolinealidad (VIF ¿10), como DAYS EMPLOYED y AMT CREDIT.

## Transformaciones adicionales

Se eliminaron valores negativos en variables como total pagado.

Se aplicaron transformaciones logar´ıtmicas a montos financieros.

Se agruparon y recodificaron categor´ıas raras en variables categ´oricas.

Se binarizaron variables de conteo para facilitar su an´alisis.

* 1. **Análisis de variables**
     1. **Variables numéricas.** Se analizaron variables como ingreso total, cuota periódica y suma total crédito, identificando valores atípicos mediante IQR y distribuciones sesgadas que requirieron transformaciones logarítmicas.

* + 1. **Variables categóricas**. Se agruparon categorías poco frecuentes, transformaron a dummies o se recodificaron variables como tipo organización, ocupación y estado civil, mejorando la consistencia del conjunto de datos y reduciendo la dimensionalidad.
  1. **Gráficos y visualizaciones**

Se generaron histogramas, boxplots, mapas de calor de correlación y gráficos de barras para explorar:

* La distribución de la variable objetivo (TARGET).
* La relación entre características clave y la probabilidad de incumplimiento.
* La composición de variables categóricas y binarias.
  1. **Conclusión**

El EDA permitió refinar el conjunto de datos a 54 variables finales limpias, sin valores nulos, categorizadas adecuadamente y con estructura lista para modelado predictivo. Se detectó un fuerte desbalance de clases (91.9 % pagos vs. 8.1 % incumplimientos), lo que deberá considerarse al aplicar modelos de clasificación en etapas posteriores.

# ANÁLISIS EXPLORATORIO DE LOS DATOS