**Bootcamp Data Science**

# Memoria Técnica - EDA Detección de Fraude en Transacciones con Tarjetas de Crédito

**Alumnos:**

María Jesús Sanchéz Pimienta

Carmen Gómez García-Atance

MADRID | ENERO 2026

## 1. Introducción y contexto

El fraude con tarjetas de crédito representa un problema relevante en el sector financiero debido a su impacto económico y reputacional, y a la necesidad de detectar operaciones fraudulentas en tiempo real. La detección temprana y precisa de transacciones fraudulentas es especialmente compleja por el fuerte desbalanceo entre operaciones legítimas y fraudulentas y por el comportamiento cambiante de los defraudadores.

El objetivo de este proyecto es realizar un **Análisis Exploratorio de Datos (EDA)** sobre un conjunto de transacciones reales con tarjetas de crédito, con el fin de comprender la estructura del dataset, identificar patrones relevantes, analizar el desbalanceo de la variable objeto, evaluar la complejidad del problema de detección de fraude y extraer conclusiones que puedan servir como base para futuros modelos de detección de fraude.

El análisis se centra en comprender los datos, no en entrenar modelos de Machine Learning, y sirve como base para justificar enfoques analíticos posteriores.

## 2. Elección de la temática y obtención de los datos

### Enfoque seleccionado

Se ha seguido el **Enfoque B: de datos a problema**. El dataset fue seleccionado desde la plataforma Kaggle tras explorar diferentes conjuntos de datos disponibles, y contiene transacciones realizadas con tarjetas de crédito por clientes europeos durante septiembre de 2013.

### Dataset

- **Nombre**: Credit Card Fraud Detection
- **Fuente**: Kaggle (mlg-ulb/creditcardfraud)
- **Periodo**: Septiembre de 2013 (dos días de transacciones)

El dataset contiene **284.807 transacciones**, de las cuales **492 corresponden a fraude**, lo que representa aproximadamente el **0,17%** del total. Así mismo, el dataset contiene 31 variables, donde la variable objetivo, Class, indica si una transacción es fraudulenta (1) o legítima (0).

Las variables V1–V28 son componentes principales obtenidas mediante PCA por motivos de confidencialidad. Las únicas variables no transformadas son Time y Amount. Adicionalmente, se incluyen las variables Time, que representa el tiempo transcurrido desde la primera transacción registrada, y Amount, que indica el importe de la transacción.

## 3. Definición de hipótesis

Se plantearon las siguientes hipótesis iniciales:

1. Las transacciones fraudulentas presentan patrones diferenciados respecto a las legítimas en algunas variables.
2. El importe de la transacción (Amount) está relacionado con la probabilidad de fraude.
3. El fraude no puede explicarse mediante una única variable, sino mediante la combinación de varias.

Estas hipótesis se contrastan a lo largo del análisis exploratorio.

## 4. Preprocesado de los datos

El dataset se descargó mediante la API oficial de Kaggle y se integró en un único DataFrame. No fue necesaria la integración de múltiples fuentes.

Las acciones de preprocesado incluyeron:

- Carga y revisión de la estructura del dataset.
- Verificación de tipos de datos.
- Creación de una variable transformada Amount_log mediante transformación logarítmica.

En esta primera etapa se realizó la carga del dataset y se hizo una inspección general de su estructura, para evaluar la calidad de los datos. Se comprobó que el conjunto de datos no contiene valores nulos ni inconsistencias evidentes en los tipos de datos. Además, todas las variables numéricas se encuentran correctamente definidas.

## 5. Limpieza de datos

### 5.1 Valores faltantes

No se detectaron valores nulos en ninguna de las variables.

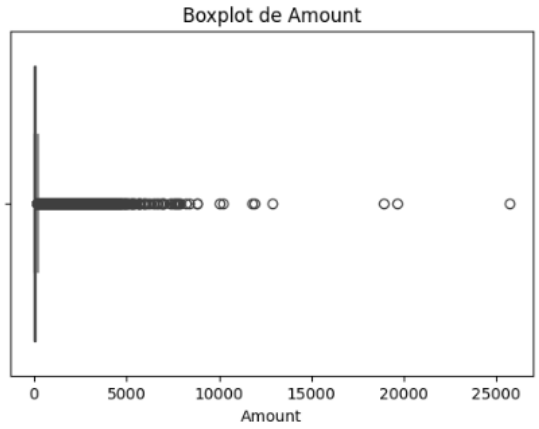
### 5.2 Duplicados

Se identificaron **1.081 filas duplicadas exactas**, correspondientes a múltiples grupos de transacciones idénticas. Dado que estos duplicados podían distorsionar el análisis, se procedió a su eliminación para evitar posibles sesgos en el análisis posterior y garantizar la integridad de los resultados.

### 5.3 Outliers

La variable Amount presenta una fuerte asimetría y numerosos valores extremos. El análisis mostró que los outliers presentan una mayor proporción de fraude que el conjunto total. Es decir, se observó que la proporción de transacciones fraudulentas entre los outliers **(≈ 0,27 %)** es superior a la proporción de fraude en el conjunto total del dataset **(≈ 0,17 %)**.

Este resultado sugiere que los valores extremos contienen información relevante para la detección de fraude. Por este motivo, se decidió **no eliminar los outliers**, ya que podrían aportar valor predictivo en etapas posteriores. En su lugar, se aplicó una **transformación logarítmica** para facilitar el análisis.



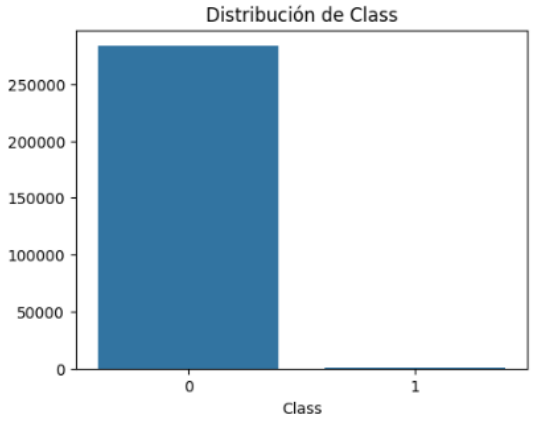

## 6. Análisis Exploratorio de Datos

### 6.1 Análisis univariante

El análisis univariante permitió estudiar la **distribución individual** de las variables:

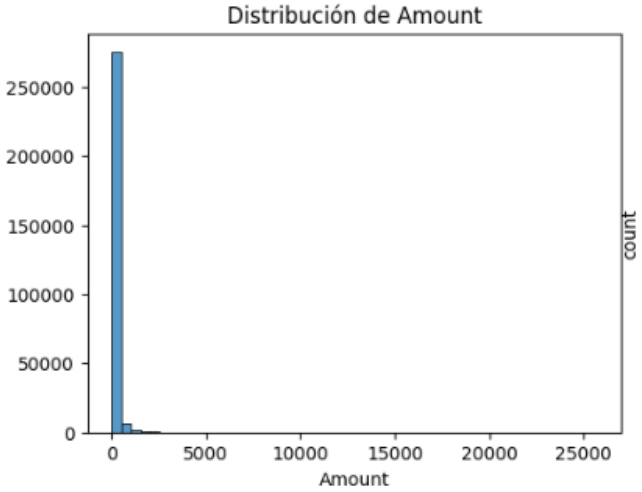
- **Time**: muestra una distribución no uniforme a lo largo del periodo analizado.
- **Amount**: presenta una distribución altamente asimétrica con cola derecha pronunciada.
- **Class**: evidencia un desbalanceo extremo, con una proporción muy reducida de transacciones fraudulentas.
- Las **variables V1–V28** presentan distribuciones centradas en torno a cero, coherentes con la aplicación de PCA.

Uno de los principales retos identificados en el dataset es el fuerte desbalanceo de la variable objetivo **Class**, ya que únicamente alrededor del **0,17 % de las transacciones corresponden a fraude**, mientras que el resto son transacciones legítimas. La escasa proporción de transacciones fraudulentas puede provocar que modelos predictivos obtengan altas métricas globales sin detectar adecuadamente el fraude.

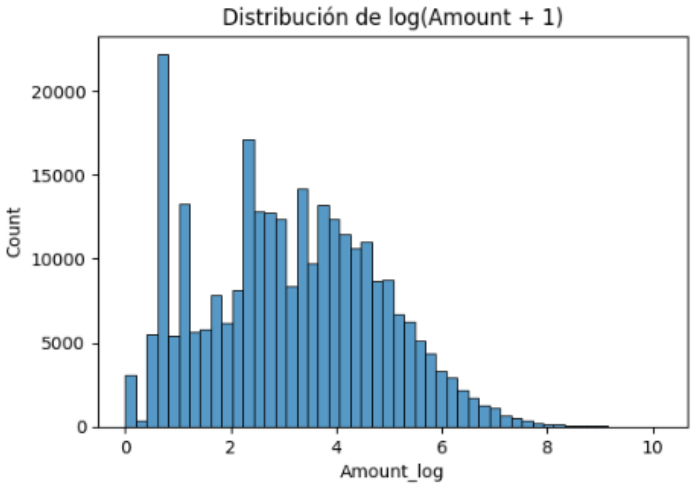



Este aspecto es especialmente relevante de cara a la construcción de modelos predictivos, y deberá ser tenido en cuenta en fases posteriores del proyecto, mediante el uso de métricas adecuadas y técnicas específicas para datos desbalanceados.

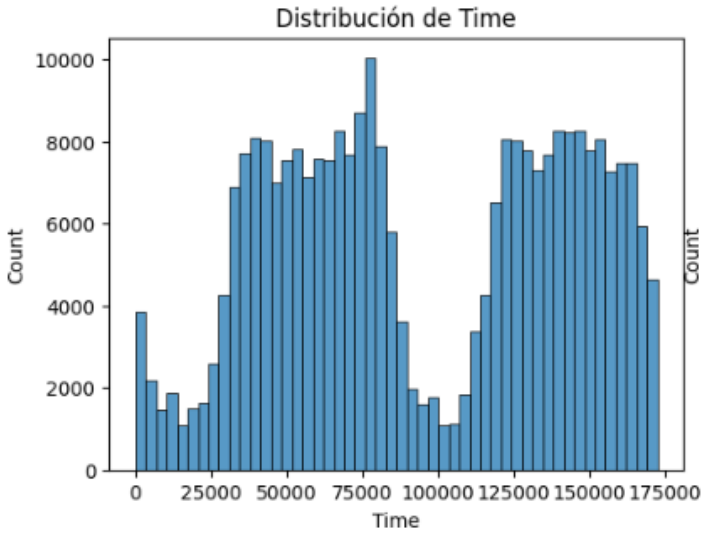
La variable **Amount** presenta una distribución altamente asimétrica, con una cola derecha muy pronunciada, lo que indica la presencia de valores extremos. Es decir, la mayoría de las transacciones se encuentran concentradas en importes bajos, junto a la presencia de valores extremos.



La transformación log(Amount + 1) redujo significativamente la asimetría de la variable Amount, facilitando su análisis.




La variable **Time** no sigue una distribución uniforme y muestra un patrón temporal coherente con la actividad diaria de las transacciones.

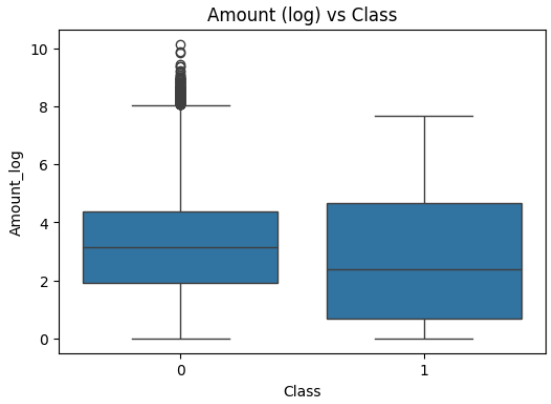



En cuanto a las **variables V1–V28**, al haber sido transformadas mediante PCA, presentan distribuciones aproximadamente centradas en cero y con varianza similar, lo cual es consistente con la naturaleza del preprocesamiento aplicado.

### 6.2 Análisis bivariante

Se analizaron relaciones entre pares de variables, prestando especial atención a la relación con la variable objetivo.

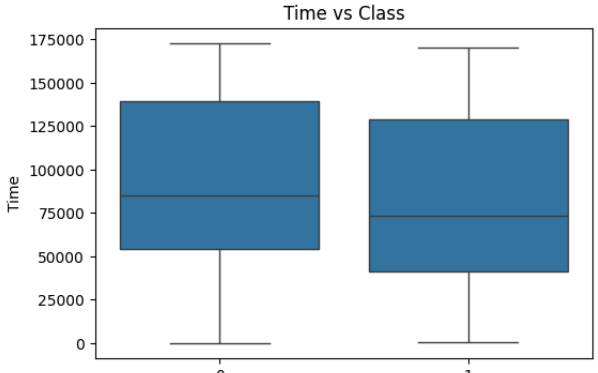
- **Amount_log vs Class:** *Las transacciones fraudulentas presentan una mediana de importe ligeramente inferior y mayor dispersión, lo que sugiere que el fraude no se limita a importes elevados.*




La comparación estadística por clase muestra que la mediana del importe en transacciones fraudulentas es de 9,82 €, frente a 22 € en transacciones legítimas. No obstante, la media del importe en fraude es superior, lo que indica una distribución más dispersa con presencia de valores extremos.

El análisis por percentiles refuerza este comportamiento. En los percentiles centrales (25% y 50%), las transacciones fraudulentas presentan importes inferiores a los de las transacciones legítimas. Sin embargo, en los percentiles superiores (90% y 95%), los importes asociados al fraude son significativamente mayores, lo que evidencia la coexistencia de fraudes de bajo importe con un reducido número de fraudes de alto impacto económico.

- **Time vs Class:** No se identificaron patrones temporales claros asociados al fraude, lo que indica que este ocurre a lo largo de todo el periodo analizado.



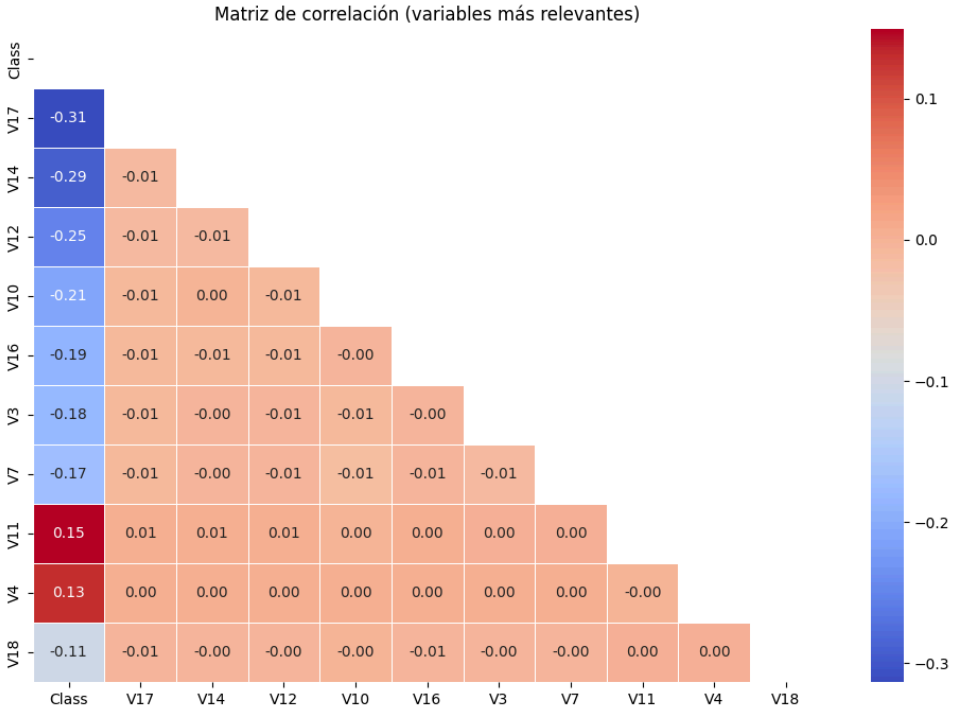
- **Correlaciones con Class:** Algunas componentes PCA (como V17, V14 o V12) presentan correlaciones moderadas con la variable objetivo, aunque ninguna variable individual muestra una relación fuerte.

La segmentación del importe en cuantiles muestra que el porcentaje de fraude no aumenta de forma monotónica con el valor de la transacción. Se observa una mayor proporción de fraude tanto en los importes más bajos como en los más elevados, mientras que los rangos intermedios presentan menores tasas de fraude. Este resultado descarta la existencia de un umbral de importe sencillo para la detección del fraude.

### 6.3 Análisis multivariante

El análisis multivariante se realizó mediante diferentes enfoques:

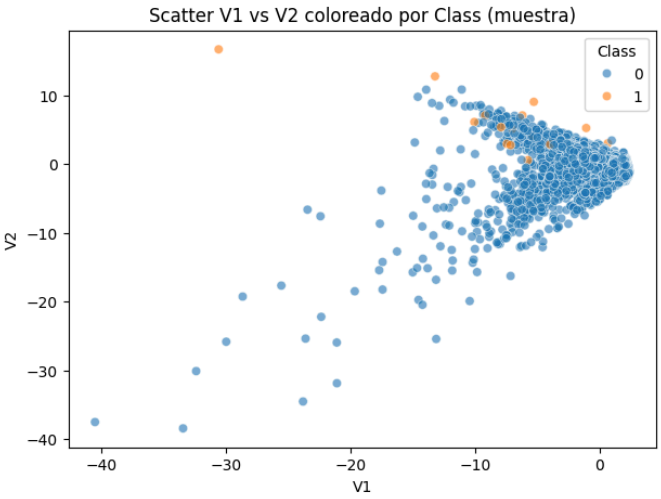
- **Matriz de correlación**: no se identificaron correlaciones lineales fuertes ni multicolinealidad significativa, lo que indica que las variables aportan información complementaria.

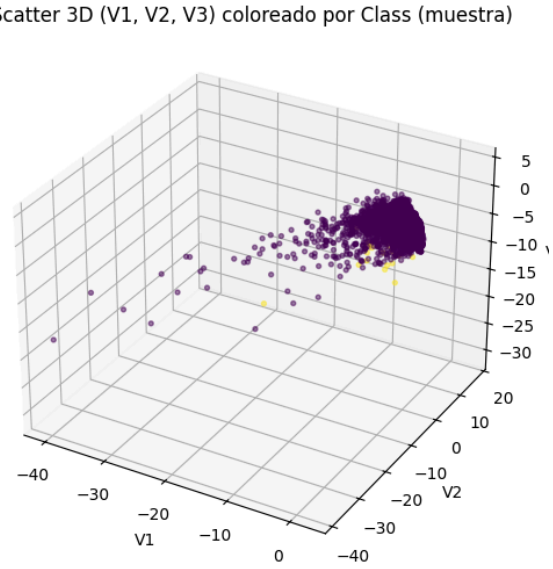



La distribución acumulada de la variable *Amount_log* por clase muestra que una mayor proporción de transacciones fraudulentas se concentra en importes bajos. A

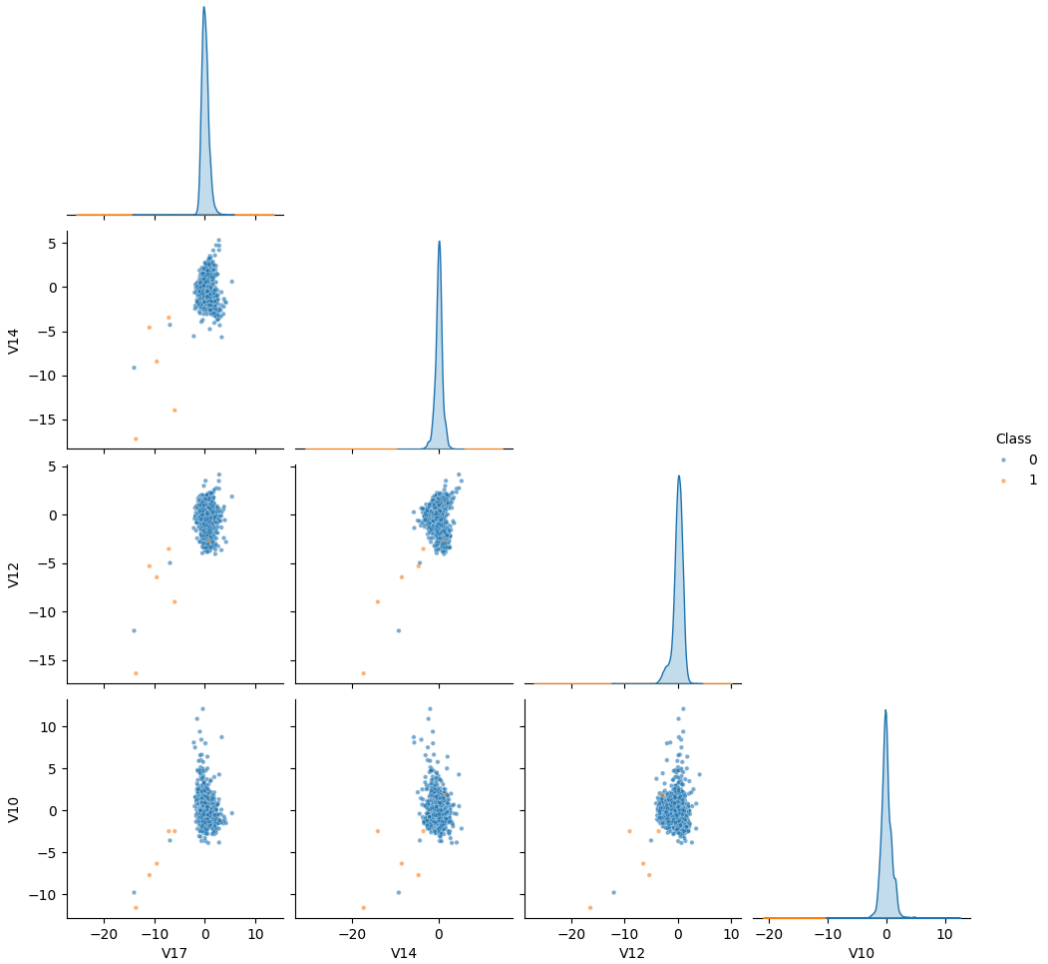
medida que aumenta el importe, las curvas de fraude y no fraude convergen, lo que refuerza la ausencia de un umbral claro y confirma la necesidad de considerar múltiples variables de forma conjunta.

- **Proyecciones PCA (2D y 3D)**: las transacciones fraudulentas aparecen solapadas con las legítimas, sin separación clara.




- **Pairplots de variables relevantes**: confirman la ausencia de patrones simples y la complejidad del problema.



Estos resultados indican que la detección del fraude requiere considerar múltiples variables de forma conjunta.

**7. Verificación de hipótesis**

- **Hipótesis 1**: Parcialmente cumplida. Existen patrones, pero no hay separaciones claras.
- **Hipótesis 2**: Refutada. El fraude no se asocia únicamente a importes elevados.
- **Hipótesis 3**: Cumplida. El problema es claramente multivariante.

**8. Conclusiones**

El análisis exploratorio muestra que el fraude con tarjetas de crédito es un fenómeno complejo, altamente desbalanceado y difícil de detectar mediante reglas simples. Ninguna variable individual permite identificar el fraude de forma clara, y no existen relaciones lineales fuertes que faciliten su detección.

El EDA justifica la necesidad de enfoques multivariantes y técnicas avanzadas en fases posteriores de modelado.

A modo de síntesis, la siguiente tabla resume los principales hallazgos cuantitativos del análisis exploratorio, incluyendo el nivel de desbalanceo del dataset, las diferencias de importe entre clases y las variables con mayor relación con el fraude.

| **Insight** | **Valor** |
|---|---|
| Porcentaje de fraude total | 0.17% |
| Mediana Amount No Fraude | 22.0 |
| Mediana Amount Fraude | 9.82 |
| Porcentaje de outliers en Amount | 11.17% |
| Variables más correlacionadas con Class | V17, V14, V12, V10 |

## 9. Recomendaciones y próximos pasos

Como continuación de este trabajo se recomienda:

- Aplicar técnicas de modelado multivariante sensibles al desbalanceo.
- Utilizar métricas adecuadas como Recall, F1-score o AUPRC.
- Evaluar técnicas de selección de variables y reducción de dimensionalidad.

Este análisis exploratorio proporciona una base sólida para el desarrollo de modelos de detección de fraude más avanzados.