DETECCIÓN DEL FRAUDE EN TRANSACCIONES DE TARJETAS DE CRÉDITO

ANÁLISIS EXPLORATORIO DE DATOS (EDA).

BOOTCAMPO DE DATA SCIENCE EN THE BRIDGE. REALIZADO POR CARMEN GÓMEZ GARCÍA-ATANCE Y MARÍA JESÚS SÁNCHEZ PIMIENTA. 2025/2026.

ÍNDICE

- Introducción y contexto
- Elección de la temática y obtención de los datos
- Definición de hipótesis
- Preprocesado de los datos
- Limpieza de datos
- Análisis Exploratorio de Datos
- Verificación de hipótesis
- Conclusiones
- Recomendaciones y próximos pasos

1. Introducción y contexto

¿Por qué es un problema relevante?

- Impacto económico y reputacional
- Necesidad de detección en tiempo real
- Fraude = eventos raros

El gran reto: datos desbalanceados

Solo el 0,17 % de las transacciones son fraude

- Riesgo de conclusiones engañosas
- Dificultad para detectar patrones

1. Introducción y contexto

OBJETIVOS

- Comprender la estructura del dataset
- Analizar el desbalanceo de la variable objetivo
- Identificar patrones relevantes
- Sentar las bases para modelado futuro

ALCANCE DEL PROYECTO

INCLUYE	NO INCLUYE
▪ Análisis Exploratorio de Datos ▪ Estadística descriptiva ▪ Visualización	▪ Entrenamiento de modelos ▪ Evaluación predictiva

2. Elección de la temática y obtención de los datos

¿Cómo se eligieron los datos?

- Enfoque B
- Exploración previa de datasets
- Selección por relevancia real

DATASET

- 284.807 transacciones
  - 492 fraudes
- Septiembre de 2013
  - Dos días de transacciones
- 31 variables

31 variables

- V1–V28
  - Componentes PCA (anonimizadas)
- Time
- Amount
- Class
  - Variable objetivo

3. Definición de hipótesis

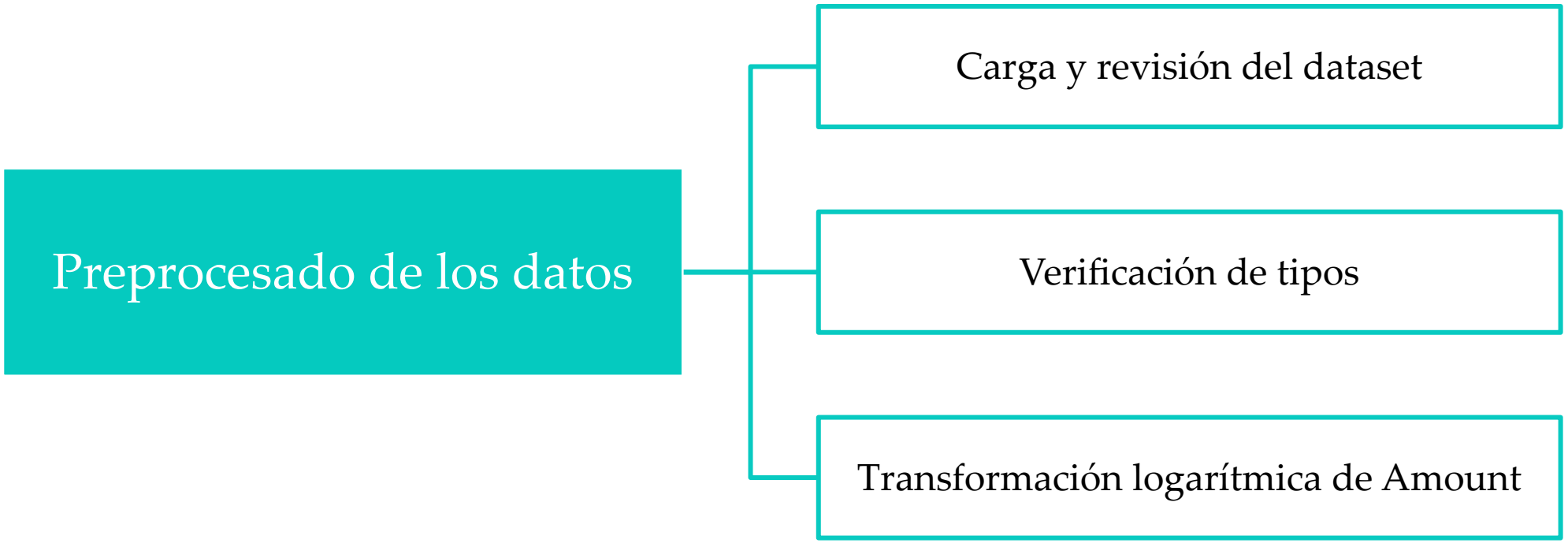
1. Existen patrones diferenciados
2. El importe está relacionado con el fraude
3. El fraude es multivariante

3. Definición de hipótesis

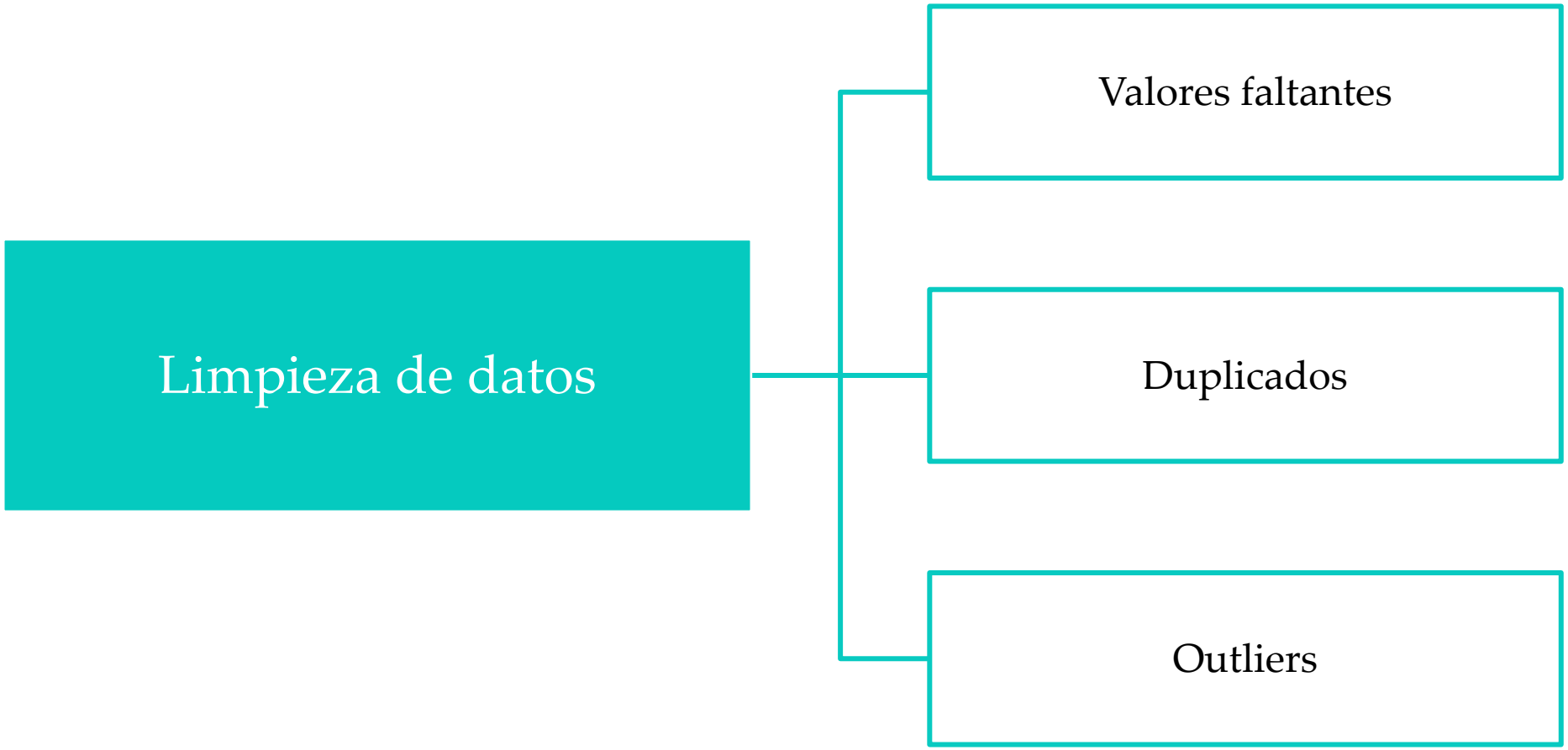
HIPÓTESIS INICIALES

1. Existen patrones diferenciados
2. El importe está relacionado con el fraude
3. El fraude es multivariante

4. Preprocesado de los datos



5. Limpieza de datos



5. Limpieza de datos

Valores faltantes → No se detectaron en ninguna de las variables

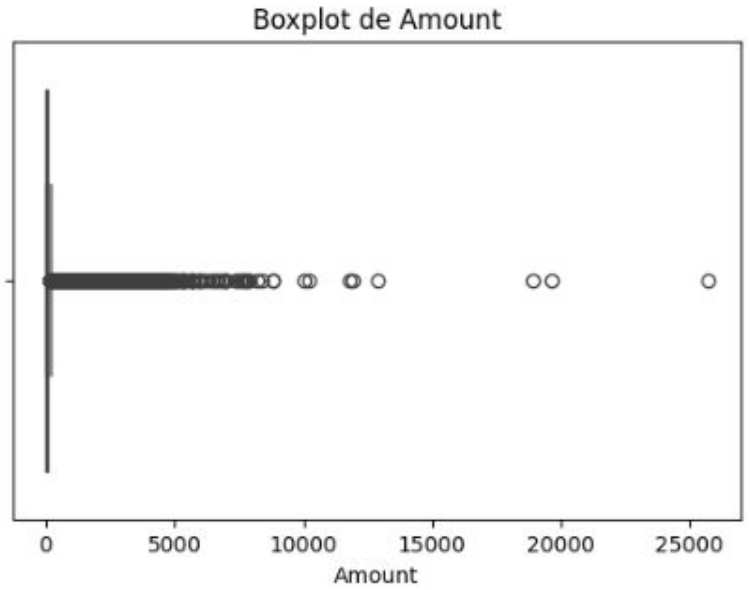
Duplicados → 1.081 filas duplicadas eliminadas

- Prevención de sesgos
- Mejora de la calidad del análisis

5. Limpieza de datos

Outliers

- Más fraude entre outliers
- No se eliminan
- Se transforman

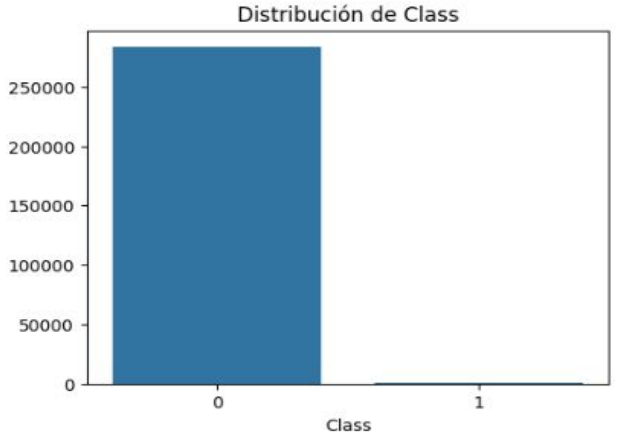


6. Análisis Exploratorio de Datos

6. 1. Análisis Univariante

Class

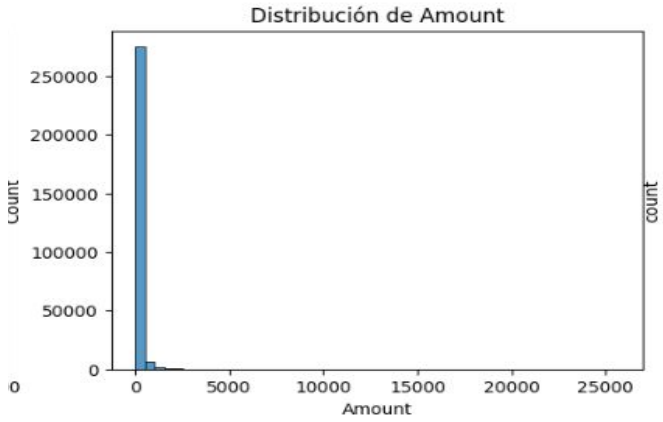
Evidencia un desbalanceo extremo



El 99,83 % de las transacciones son legítimas

Amount

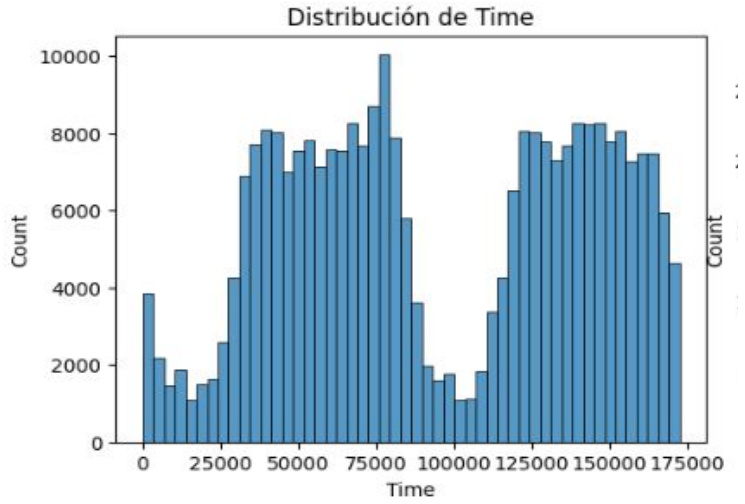
Distribución altamente asimétrica con cola derecha pronunciada



6. 1. Análisis Univariante

Time

Distribución no uniforme

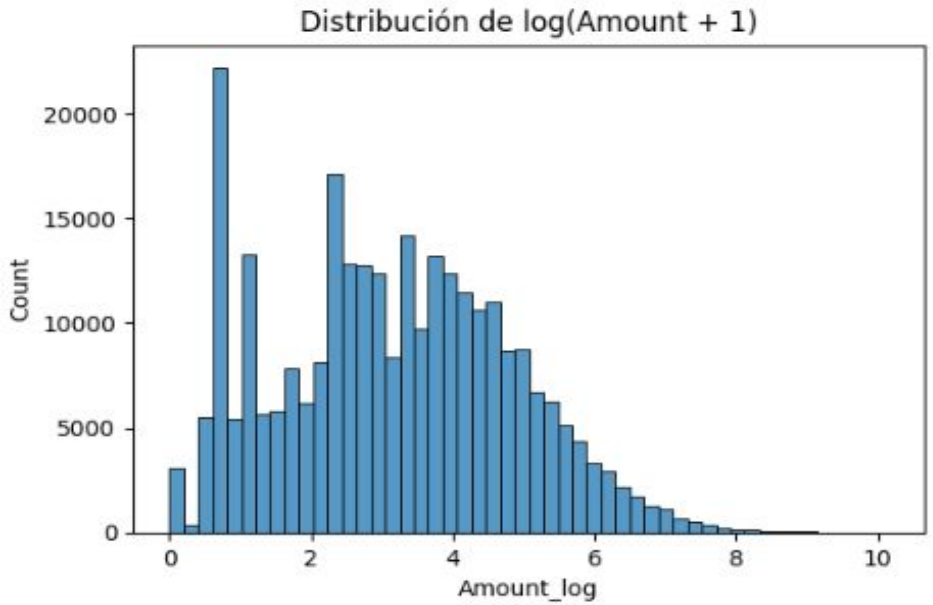


V1-V28

Distribuciones centradas en torno a cero, coherentes con la aplicación de PCA

6. 1. Análisis Univariante

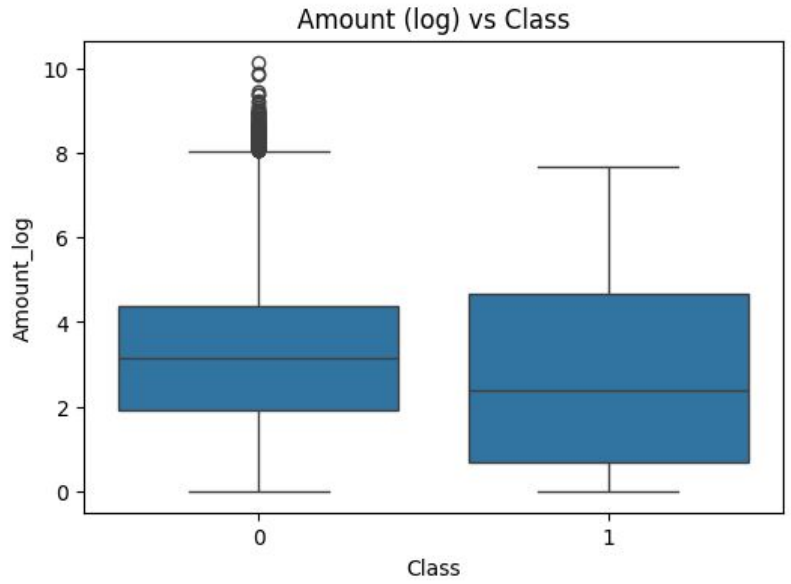
Amount → La transformación logarítmica de esta variable redujo significativamente su asimetría, facilitando su análisis.



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

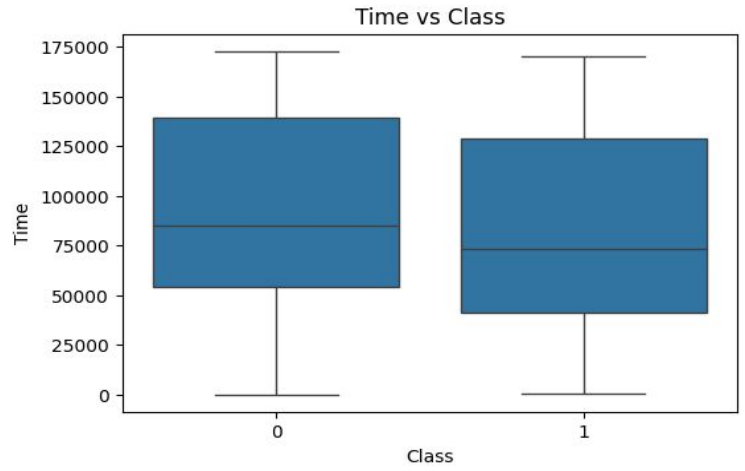


El fraude no se limita a importes altos

6. 2. Análisis Bivariante

Time vs Class

No hay patrón temporal claro



Correlaciones con Class

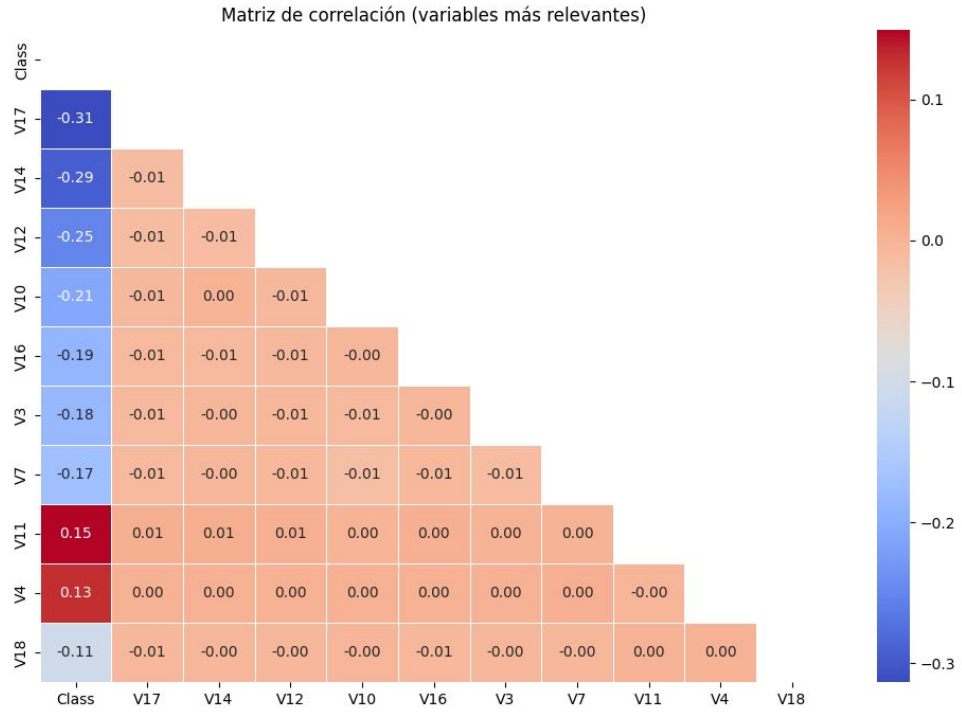
Ninguna variable PCA explica el fraude por sí sola

6.3. Análisis Multivariante

Matriz de correlación

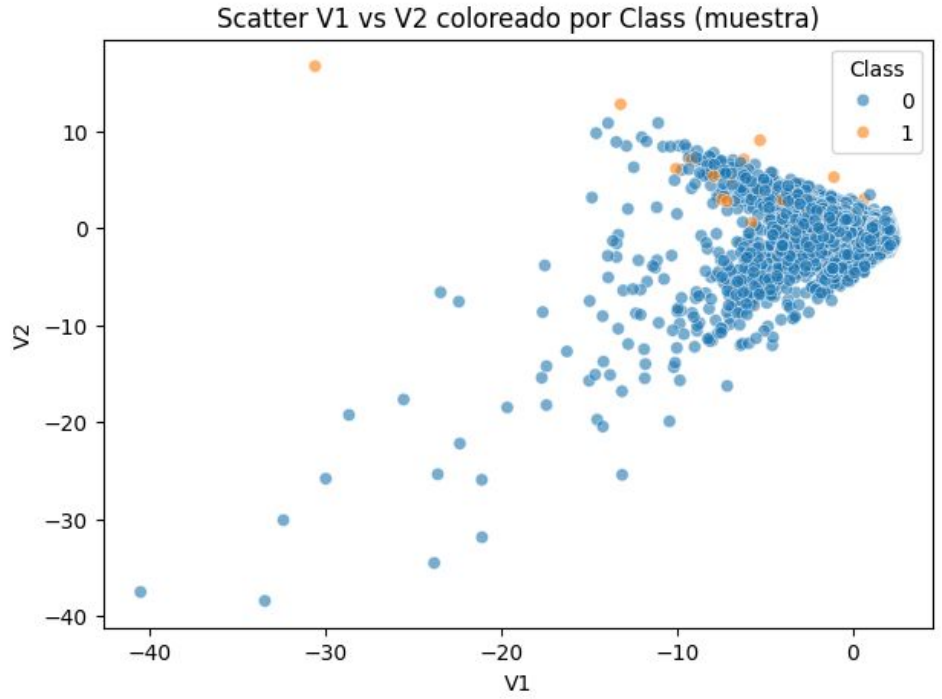
No se identificaron correlaciones lineales fuertes ni multicolinealidad significativa

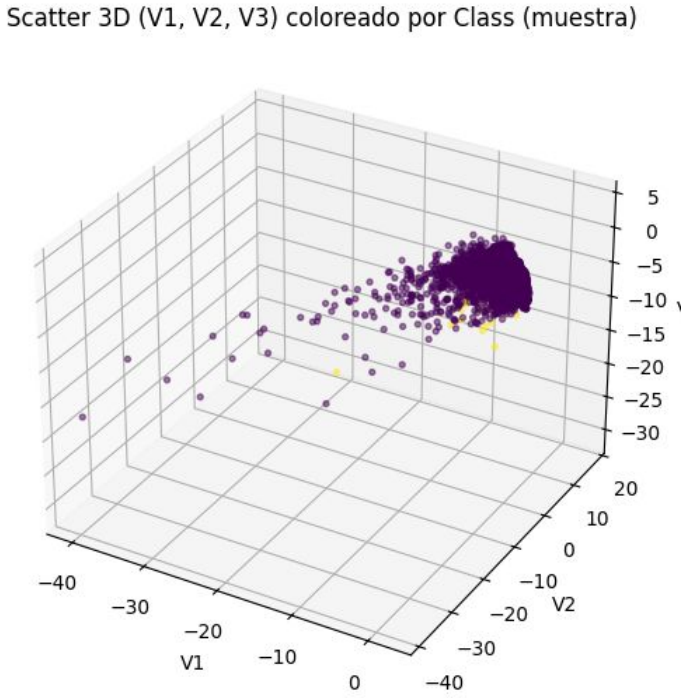
Las variables aportan información complementaria



6.3. Análisis Multivariante

Proyecciones PCA (2D y 3D) → Las transacciones fraudulentas aparecen solapadas con las legítimas

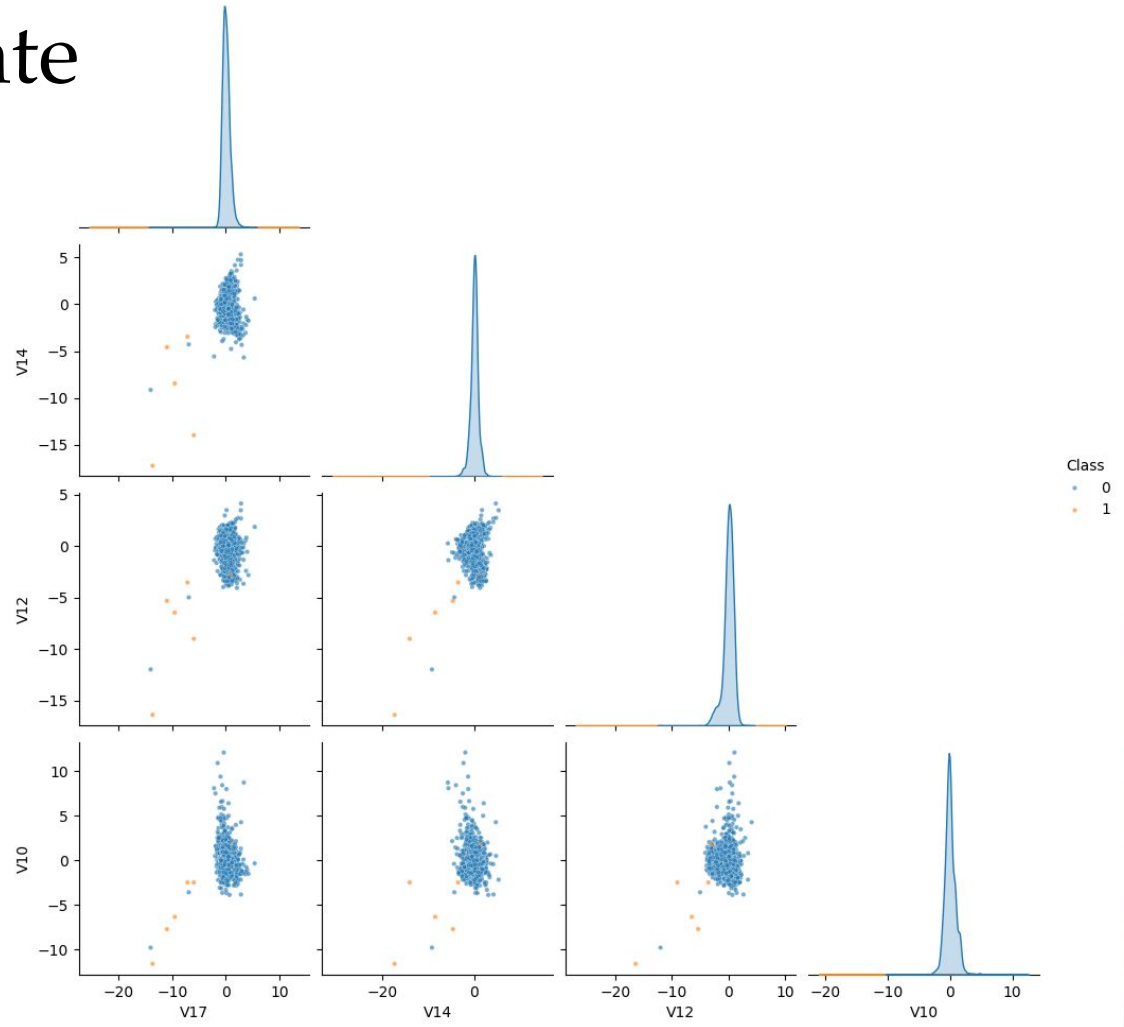




6.3. Análisis Multivariante

Pairplots de variables relevantes

Confirman la ausencia de patrones simples y la complejidad del problema.



7. Verificación de hipótesis

Hipótesis 1 → Las transacciones fraudulentas presentan patrones diferenciados respecto a las legítimas en algunas variables.

Parcialmente cumplida

Hipótesis 2 → El importe de la transacción está relacionado con la probabilidad de fraude.

Refutada

Hipótesis 3 → El fraude no puede explicarse mediante una única variable, sino mediante la combinación de varias

Confirmada

8. Conclusiones

Fraude raro y complejo

No hay reglas simples

Necesidad de enfoque multivariante

8. Conclusiones

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

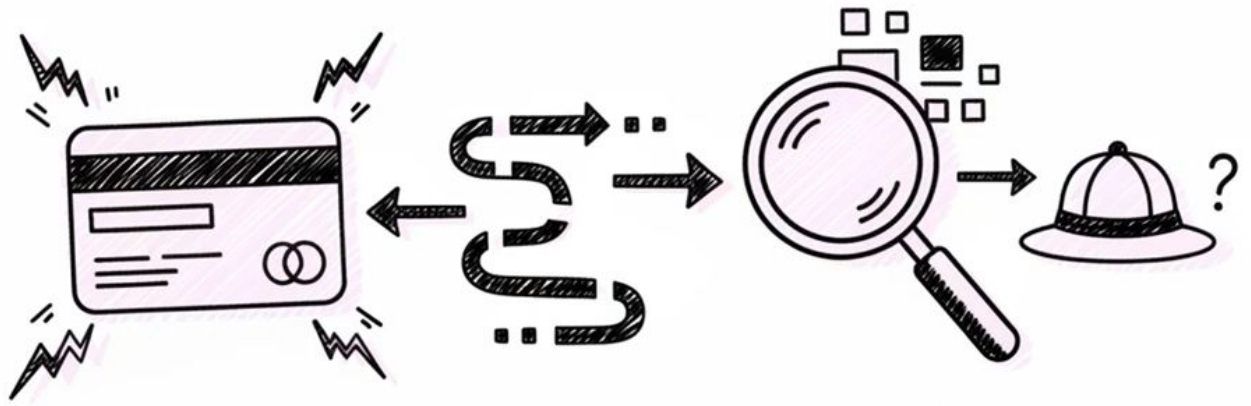
El EDA sirve como base para una detección de fraude robusta

9. Recomendaciones y próximos pasos

Se recomienda:

- Modelos sensibles al desbalanceo
- Métricas adecuadas
- Selección de variables

Desenmascarando el Fraude



FIN

¡Esperamos que os haya gustado!