Dibujo de un animal con la boca abierta

Descripción generada automáticamente con confianza baja

**UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS**

**FACULTAD DE INGENIERÍA**

**METODOLOGÍA PARA DATA SCIENCE**

**TB1**:

COMPRENSIÓN DE NEGOCIO Y DATOS

**PROPUESTA:**

DETECCIÓN DE TRANSACCIONES FRAUDULENTAS EN TARJETAS DE CRÉDITO

**INTEGRANTES**

Poma Espinoza, Gustavo Arturo (u20221c138)

Diaz Villanueva, Jeffrey Ulises (u202124269)

Pastor Salazar, Tomas Alonso (u201916314)

**PROFESORA**

Ubaldo Gamarra, Victoria Alejandra

2025 - 02

**ÍNDICE**

[**1. Introducción 3**](#_e5oyiy1c0rn5)

[**2. Objetivo 4**](#_3i98bj7lsatu)

[**3. Dataset 4**](#_vev6xnw6oje)

[**4. Roles 5**](#_izzmm0q0xt46)

[4.1. Gustavo Arturo Poma Espinoza – Data Engineer 5](#_jf93h170exsc)

[4.2. Jeffrey Ulises Díaz Villanueva – Data Analyst 5](#_aea6m59r30sy)

[4.3. Tomás Alonso Pastor Salazar – Data Scientist 5](#_7jftntpewucw)

[**5. Visualizaciones 6**](#_4ohclaa7ci5h)

[5.1. Conteo y proporción 6](#_tc05p46lg17m)

[5.2. Histograma de Amount 7](#_4yce3gbgq1vl)

[5.3. Matriz de correlación 7](#_488lbdevb47q)

[5.4. Dispersión índice–v alor con límites IQR 8](#_jmtto59hd3wg)

[5.5. Link del Github 8](#_a7glnpb2mzmw)

[**6. Análisis EDA 9**](#_8p8nto66114v)

[6.1 Valores nulos y duplicados 9](#_1fxv46p7eoc7)

[6.2. Separar numéricas y objetivo 9](#_k14lku47o6ed)

[**7. Bibliografía 10**](#_ykvyjetdl5z7)

# **1. Introducción**

El fraude con tarjetas de crédito constituye uno de los principales desafíos en el ámbito financiero contemporáneo, afectando tanto a instituciones bancarias como a millones de usuarios en todo el mundo. Este fenómeno no solo genera pérdidas económicas considerables, sino que también debilita la confianza en los sistemas digitales de pago, cuya adopción se ha acelerado con el auge del comercio electrónico y los servicios financieros en línea.

De acuerdo con el *Nilson Report* (2024), las pérdidas globales por fraude con tarjetas superaron los 35 mil millones de dólares, y se estima que esta cifra podría duplicarse hacia 2030 si no se implementan mecanismos de detección temprana más eficientes. En América Latina, la situación es especialmente preocupante: según la *Asociación de Bancos de México* (ABM, 2023), más del 40 % de los reclamos financieros están relacionados con transacciones no reconocidas o fraudulentas.

Frente a este panorama, la ciencia de datos emerge como una herramienta esencial para combatir el fraude financiero. Mediante el análisis de grandes volúmenes de datos y la aplicación de modelos de aprendizaje automático, es posible identificar patrones anómalos que diferencien transacciones legítimas de fraudulentas en tiempo real. En este contexto, el presente trabajo se centra en la exploración del dataset *creditcard.csv*, con el propósito de comprender sus características, detectar desequilibrios de clases y sentar las bases para el desarrollo de un modelo predictivo que contribuya a la seguridad de las operaciones financieras.

Ante esta problemática, cobra especial relevancia la detección temprana de transacciones fraudulentas mediante el análisis automatizado de los datos. Implementar estrategias preventivas basadas en inteligencia artificial y aprendizaje automático resulta esencial para reducir las pérdidas financieras y fortalecer la confianza de los usuarios en los sistemas digitales de pago. En este escenario, los modelos de clasificación se han consolidado como herramientas clave, ya que permiten procesar grandes volúmenes de información transaccional en tiempo real, identificar comportamientos atípicos y anticipar posibles fraudes antes de que se concreten, contribuyendo así a la seguridad y sostenibilidad del ecosistema financiero.

# **2. Objetivo**

Desarrollar un modelo predictivo de clasificación binaria utilizando el algoritmo Random Forest, con la capacidad de identificar si una transacción es legítima o fraudulenta, a partir del análisis de las variables del dataset *creditcard.csv*. El modelo buscará alcanzar un F1-score igual o superior a 0.70, con el fin de ofrecer una herramienta confiable para la detección temprana de fraudes financieros y la protección de los usuarios y las instituciones bancarias.

# **3. Dataset**

El dataset seleccionado, titulado “Credit Card Fraud Detection”, contiene información sobre transacciones realizadas con tarjetas de crédito por diferentes usuarios en un periodo de dos días. Cada registro representa una transacción individual e incluye variables numéricas derivadas de un proceso de reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA), lo cual permite preservar la privacidad de los datos originales.

Este conjunto de datos tiene como objetivo facilitar el estudio de patrones de comportamiento financiero que permitan distinguir entre operaciones legítimas y fraudulentas, constituyendo una base ideal para el desarrollo de modelos de clasificación supervisada en el contexto de la detección temprana de fraude.

* **Tamaño**: 284,807 registros y 31 variables.
* **Variable objetivo**: Class, que toma el valor 0 para transacciones legítimas y 1 para transacciones fraudulentas.

| **Nombre de la columna** | **Descripción breve** |
| --- | --- |
| **Time** | Tiempo transcurrido en segundos desde la primera transacción registrada en el conjunto de datos (numérica). |
| **V1 a V28** | Variables anónimas resultantes de la transformación PCA que representan características financieras no reveladas (numéricas). |
| **Amount** | Monto de la transacción realizada (numérica). |
| **Class** | Variable objetivo que indica si la transacción es fraudulenta (1) o legítima (0). |

# **4. Roles**

Se asignaron roles a cada uno de los integrantes del grupo para hacer diferentes tareas en el transcurso del proyecto:

## **4.1. Gustavo Arturo Poma Espinoza – Data Engineer**

Encargado del procesamiento y preparación de los datos. Su labor incluye la limpieza, integración y optimización del dataset *creditcard.csv*, garantizando su calidad y consistencia para el análisis posterior. Además, diseña los pipelines de datos y gestiona la estructura del repositorio del proyecto.

## **4.2. Jeffrey Ulises Díaz Villanueva – Data Analyst**

Responsable del análisis exploratorio de datos (EDA) y de la interpretación de patrones en las variables. Desarrolla visualizaciones que permiten comprender la distribución de las transacciones y el comportamiento de las clases, aportando información clave para la toma de decisiones y la selección de variables relevantes.

## **4.3. Tomás Alonso Pastor Salazar – Data Scientist**

Lidera el desarrollo del modelo predictivo de detección de fraude. Se encarga de implementar, entrenar y validar el algoritmo Random Forest, optimizando sus hiperparámetros y evaluando su desempeño mediante métricas como precisión, recall y F1-score. Además, documenta los resultados y conclusiones del análisis.

# **5. Visualizaciones**

## **5.1. Conteo y proporción**

A graph with a blue bar

AI-generated content may be incorrect.

El gráfico muestra la distribución de la variable Class, donde 0 representa transacciones legítimas y 1 fraudulentas. Se evidencia un fuerte desbalance de clases, lo que hace necesario aplicar técnicas de balanceo o ajuste de pesos para evitar sesgos del modelo hacia la clase mayoritaria.

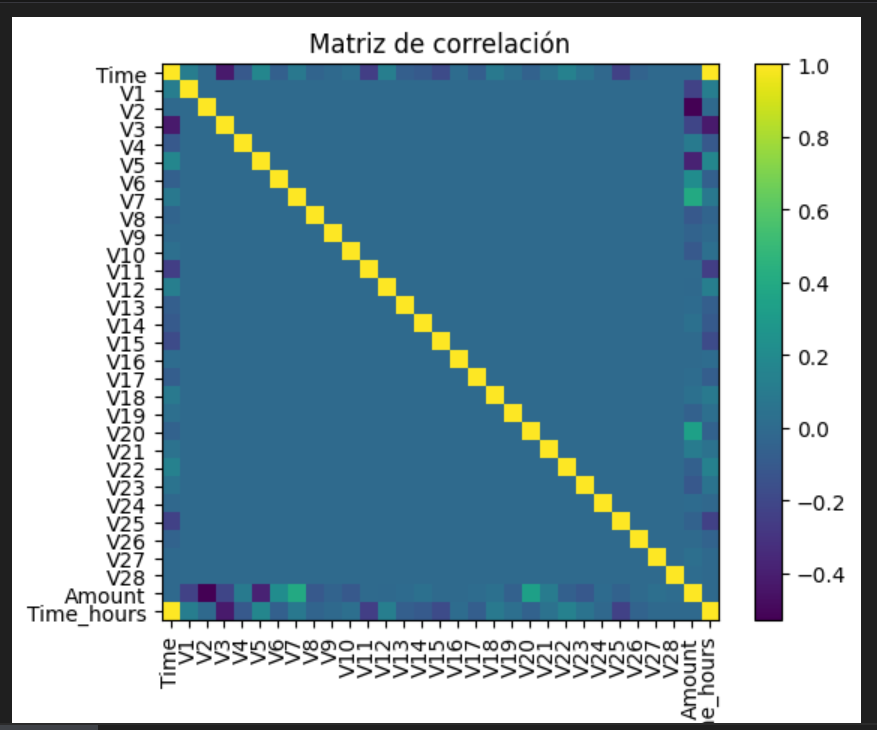
## **5.2. Histograma de Amount**

A graph with numbers and a line

AI-generated content may be incorrect.

El histograma de Amount muestra que la mayoría de las transacciones se concentran en montos bajos, mientras que los valores altos son poco frecuentes. Esto indica una distribución sesgada a la derecha, característica común en datos financieros.

## **5.3. Matriz de correlación**



La matriz de correlación muestra una baja relación entre la mayoría de variables, lo que indica independencia entre las componentes transformadas por PCA. Solo se observan correlaciones moderadas en algunas variables cercanas, lo que sugiere una adecuada reducción de redundancia en los datos.

## **5.4. Dispersión índice–v** **alor con límites IQR**

A graph of blue dots

AI-generated content may be incorrect.A blue dotted line with numbers

AI-generated content may be incorrect.A graph of blue dots

AI-generated content may be incorrect.

Los gráficos de dispersión con límites IQR evidencian la presencia de outliers en variables como *Amount* y *V27*. Aunque la mayoría de los datos se concentran cerca del rango intercuartílico, existen valores extremos que podrían influir en el desempeño del modelo y requerir un tratamiento previo.

Esta visualización demuestra que los estudiantes con deudas presentan una proporción más alta a la deserción respecto a los que se mantienen al día con sus pagos.

## **5.5. Link del Github**

<https://github.com/magic-tps/TB1-Comprensi-n-de-los-Datos>

# 

# **6. Análisis EDA**

## **6.1 Valores nulos y duplicados**

A screen shot of a computer code

AI-generated content may be incorrect.

## **6.2. Separar numéricas y objetivo**

A screen shot of a computer code

AI-generated content may be incorrect.

# 

# **7. Bibliografía**

Afriyie, J. K., et al. (2023). A supervised machine-learning algorithm for detecting and preventing credit card fraud. *Journal of Big Data & Analytics in Healthcare*. [https://www.sciencedirect.com/science/article/pii/S2772662223000036](https://www.sciencedirect.com/science/article/pii/S2772662223000036?utm_source=chatgpt.com)

Sundaravadivel, P., et al. (2025). Optimizing credit card fraud detection with random forests and SMOTE. *Scientific Reports*. [https://www.nature.com/articles/s41598-025-00873-y](https://www.nature.com/articles/s41598-025-00873-y?utm_source=chatgpt.com)

Xuetong, N., Wang, L., & Yang, X. (2019). A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv*. [https://arxiv.org/abs/1904.10604](https://arxiv.org/abs/1904.10604?utm_source=chatgpt.com)