

# Data Science Project



**financial  
fraud  
detection**



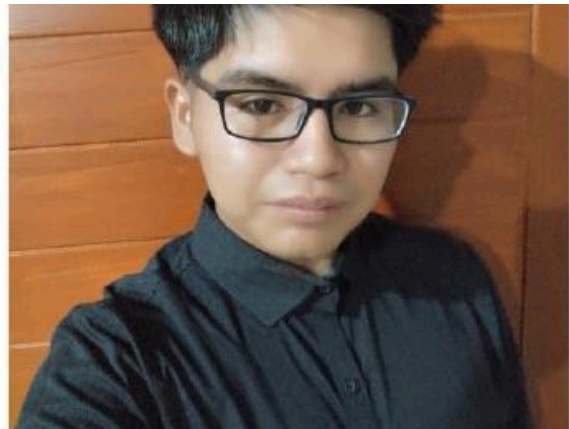
Jhonatan Rodriguez  
Jorge Perez  
Esteban Ferraz





# Científicos de Datos

Nuestro equipo se especializa en análisis de datos y desarrollo de modelos de machine learning para detectar y prevenir transacciones fraudulentas.



Jhonatan Rodriguez



Jorge Perez



Esteban Ferraz

GitHub

<https://github.com/JhonatanRC03>

Linkedin

<https://www.linkedin.com/in/jrc03>

GitHub

<https://github.com/JorgePere27>

Linkedin

<https://www.linkedin.com/in/jorge-perez-1b3621232/>

GitHub

<https://github.com/estebanferraz1>

Linkedin

<https://www.linkedin.com/in/esteban-ferraz/>

# Fraude por Pago en Línea

2023  
38MM USD

2028  
91MM USD

2023 - 2028  
362MM USD

Un reporte de Juniper Research estima que las pérdidas comerciales por fraudes en pagos en línea aumentarán de 38 mil millones de dólares en 2023 a 91 mil millones de dólares en 2028.



# Problema de negocio



La urgencia por detectar fraudes en transacciones móviles de dinero ha llevado a una empresa del segmento Fintech a buscar soluciones innovadoras.

Nuestro objetivo es desarrollar un modelo de machine learning que pueda distinguir de manera precisa entre transacciones legítimas y fraudulentas, estableciendo así un estándar de seguridad en el sector financiero global.

# Preprocesamiento de Datos: La Base del Éxito

```
1 #Cargamos la base de datos en un dataframe de pandas y visualizamos una muestra
2 df_banco = pd.read_csv("fraud_detect.csv")
3 df_banco.head()
```

	step	type	amount	nameOrig	oldbalanceOrig	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.00	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.00	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.00	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.00	0.00	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.00	0.00	0	0

```
1 #verificamos la información general de nuestros datos
2 df_banco.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6362620 entries, 0 to 6362619
Data columns (total 11 columns):
 #   Column          Dtype
---  -
0   step            int64
1   type            object
2   amount          float64
3   nameOrig        object
4   oldbalanceOrig  float64
5   newbalanceOrig  float64
6   nameDest        object
7   oldbalanceDest  float64
8   newbalanceDest  float64
9   isFraud         int64
10  isFlaggedFraud  int64
dtypes: float64(5), int64(3), object(3)
memory usage: 534.0+ MB
```

- El preprocesamiento de datos es esencial para el rendimiento de modelos de machine learning, abarcando desde la limpieza y manejo de valores faltantes hasta la normalización. Se trabajó con conjuntos de datos de 100,000 y 6 millones de entradas para garantizar robustez y escalabilidad.

# Exploración de Datos: Descubriendo Patrones Ocultos

Overview Alerts 10 Reproduction	
Dataset statistics	
Number of variables	12
Number of observations	6355023
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	630.3 MiB
Average record size in memory	104.0 B

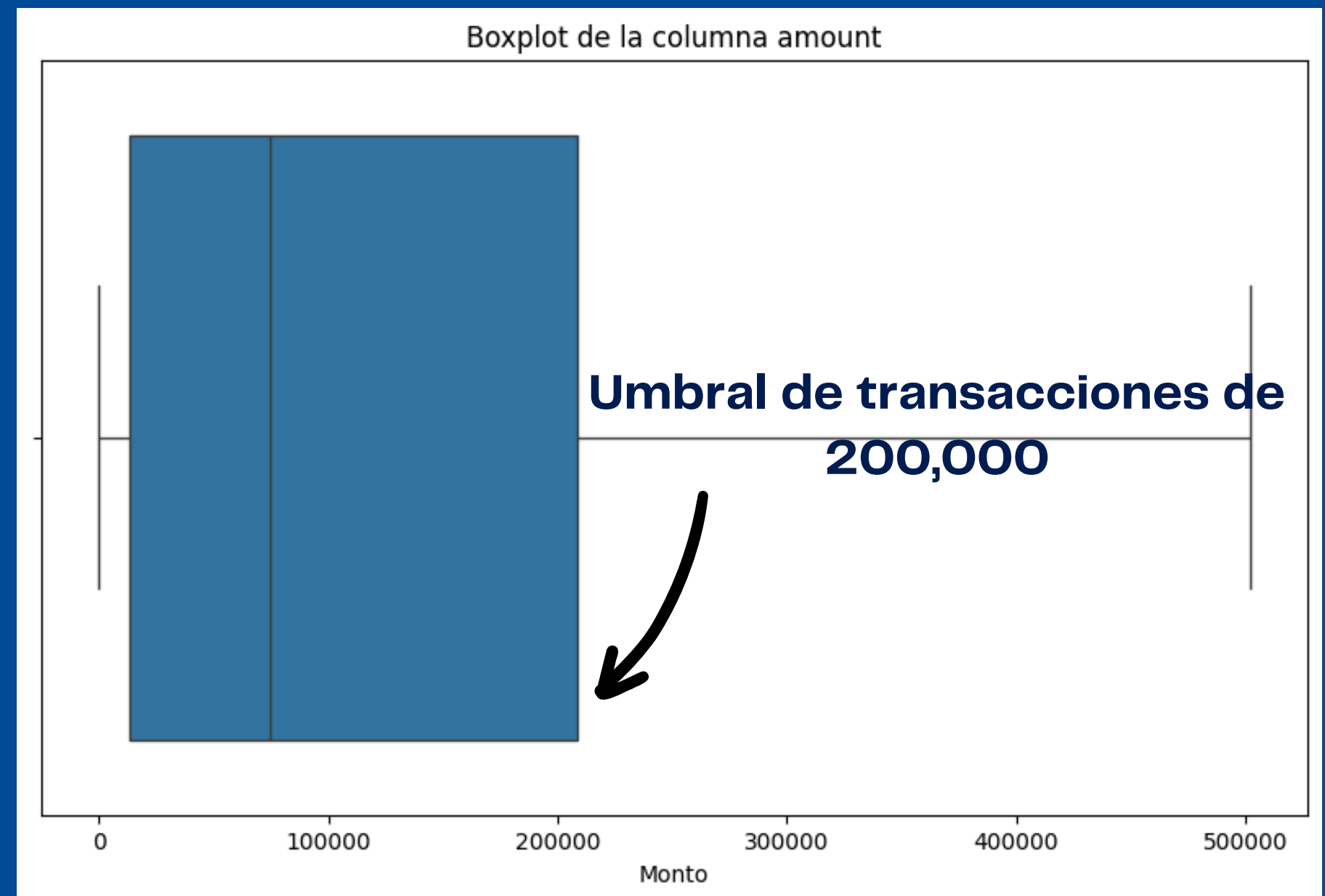
- Análisis exhaustivo para identificar variables clave y relaciones en la detección de fraudes. Se aborda el desafío del desbalanceo de clases de datos con técnicas de reequilibrio para mejorar la eficacia del modelo.

# Datos interesantes

**1.05%**

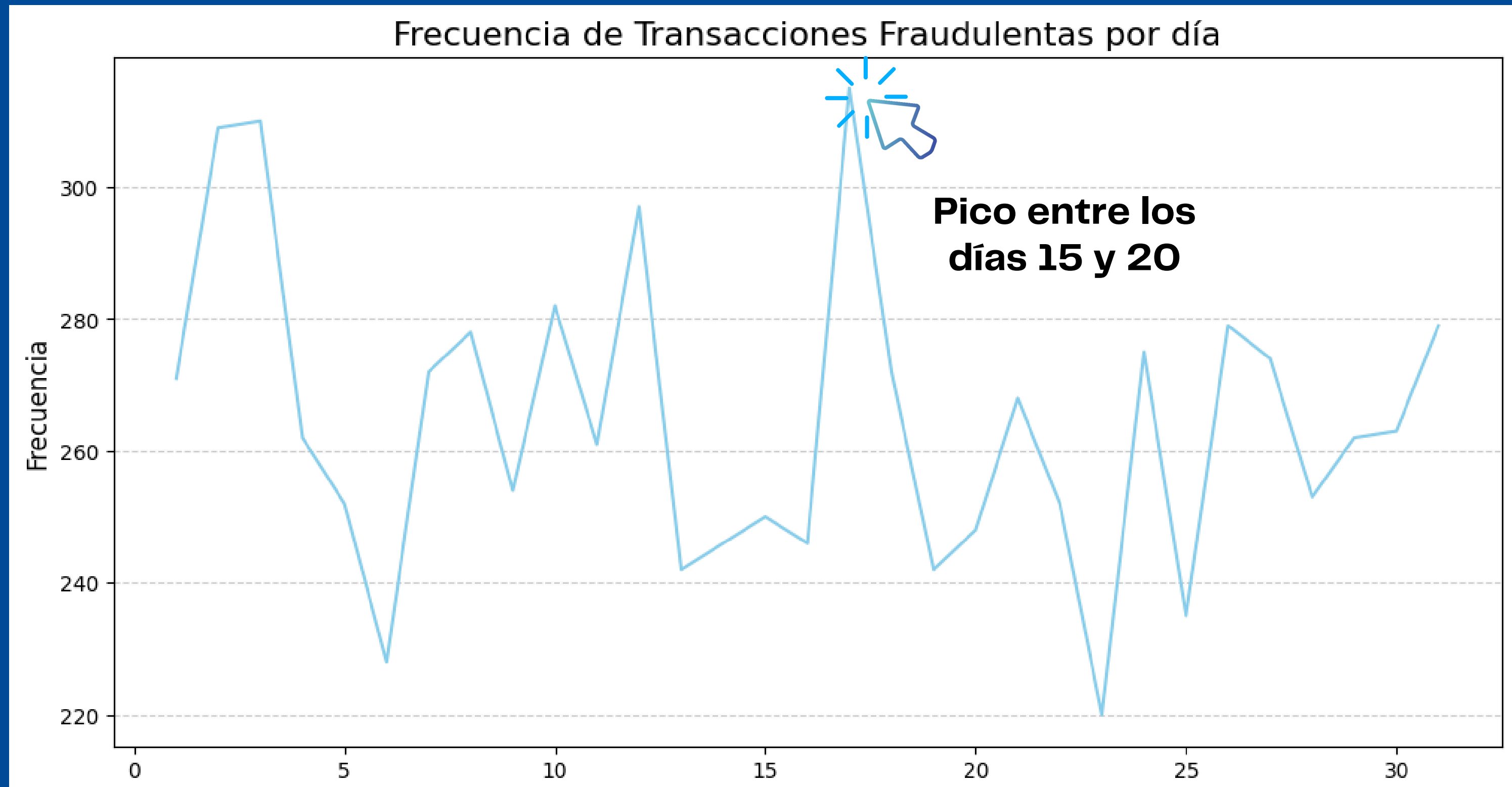
**12,056,415,427.84**

Transacciones fraudulentas



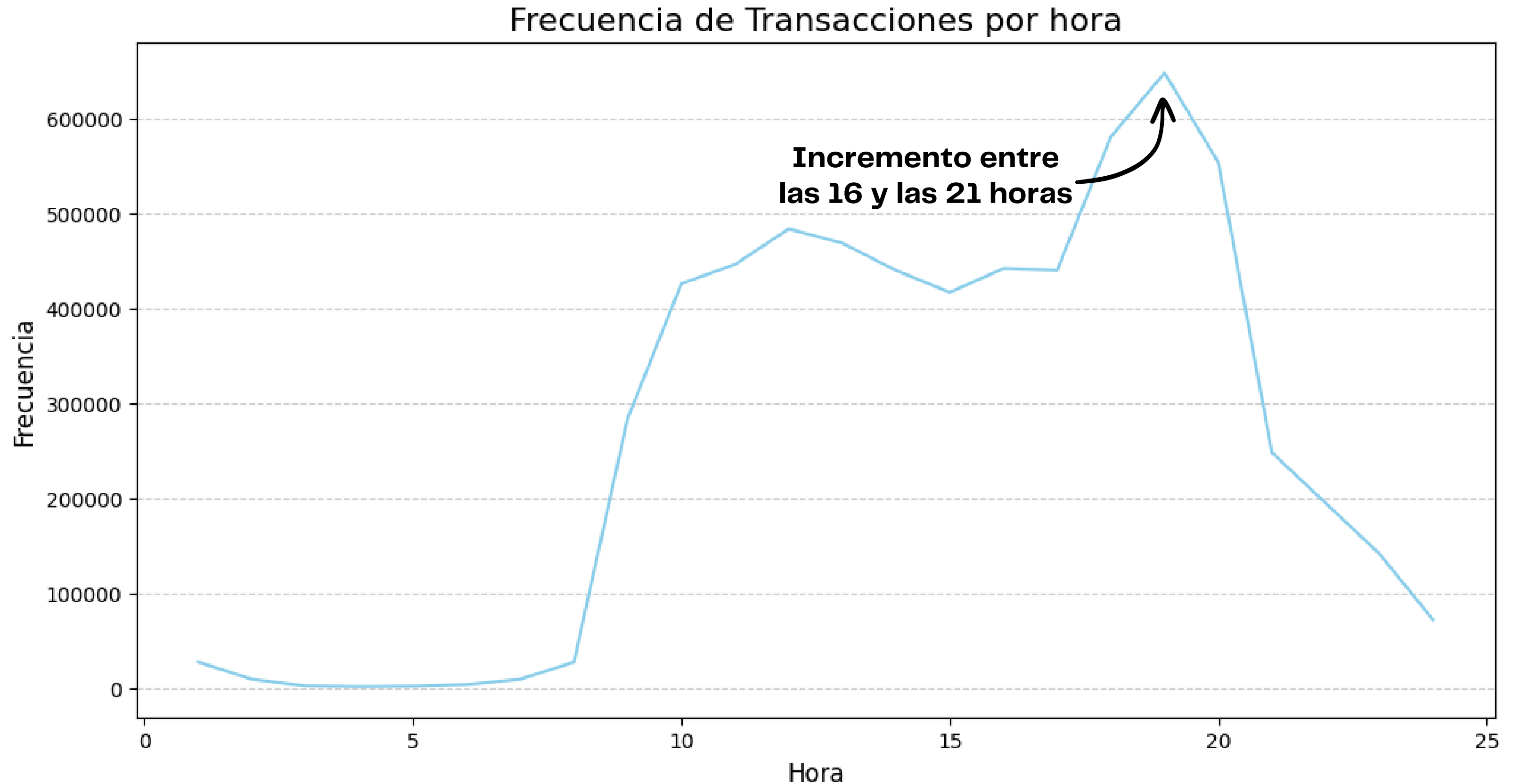
Monto transacciones

# Datos interesantes



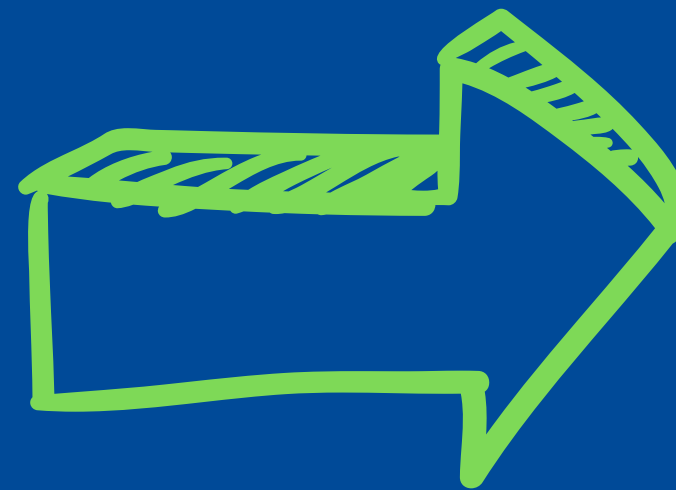
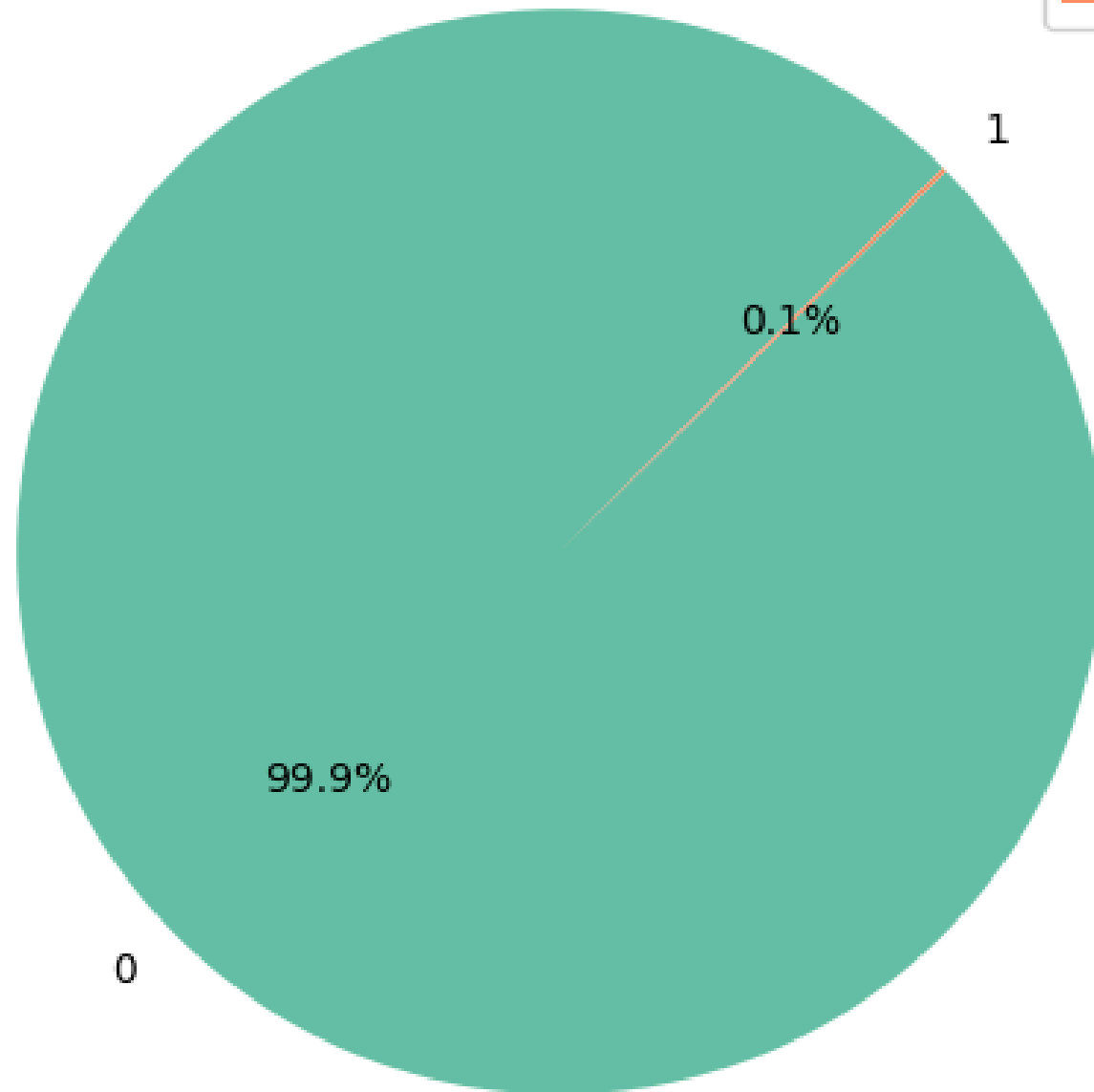
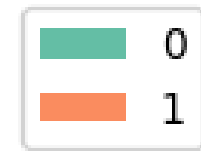


# Datos interesantes

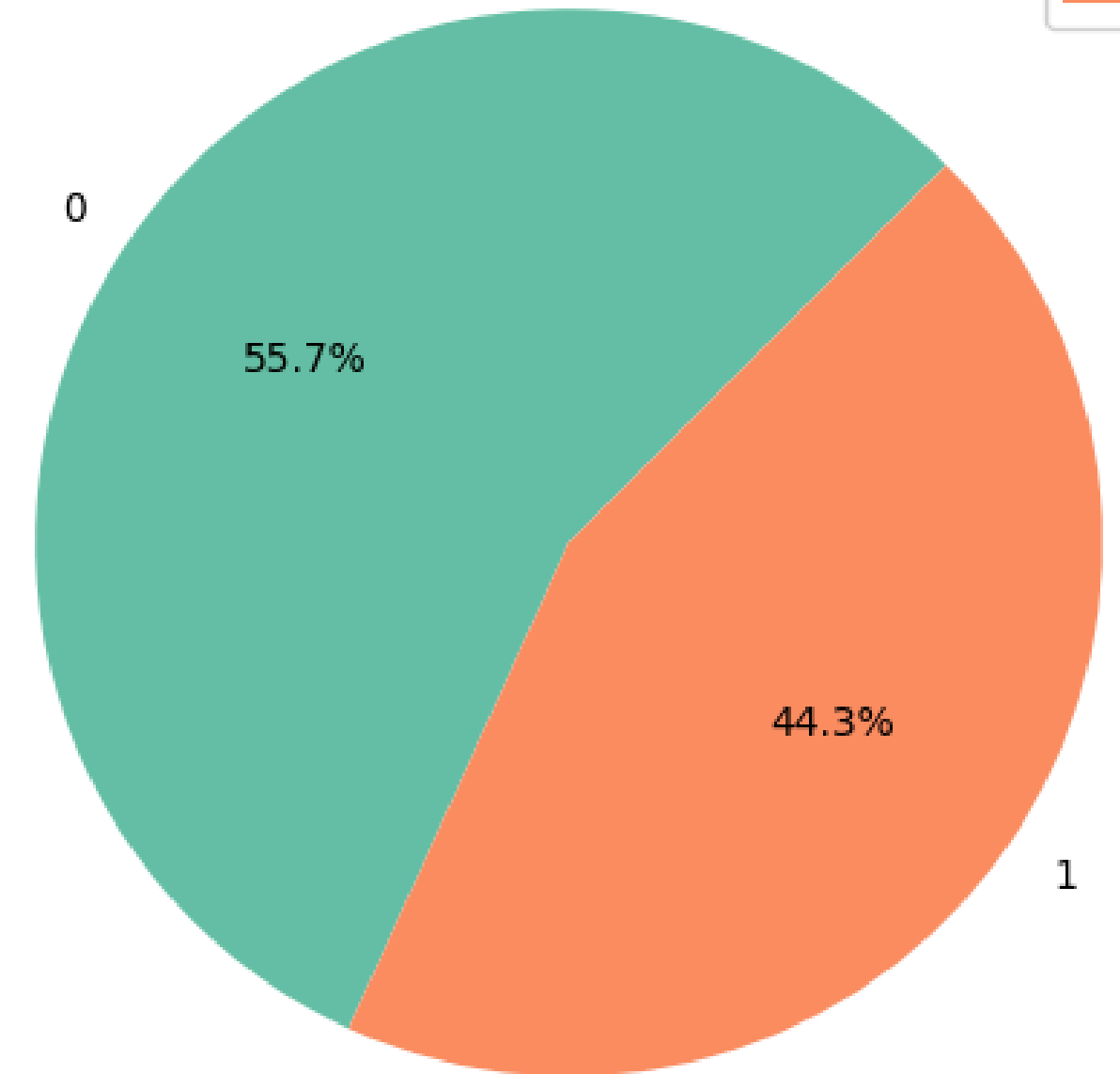
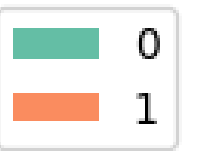


# Balanceo de datos

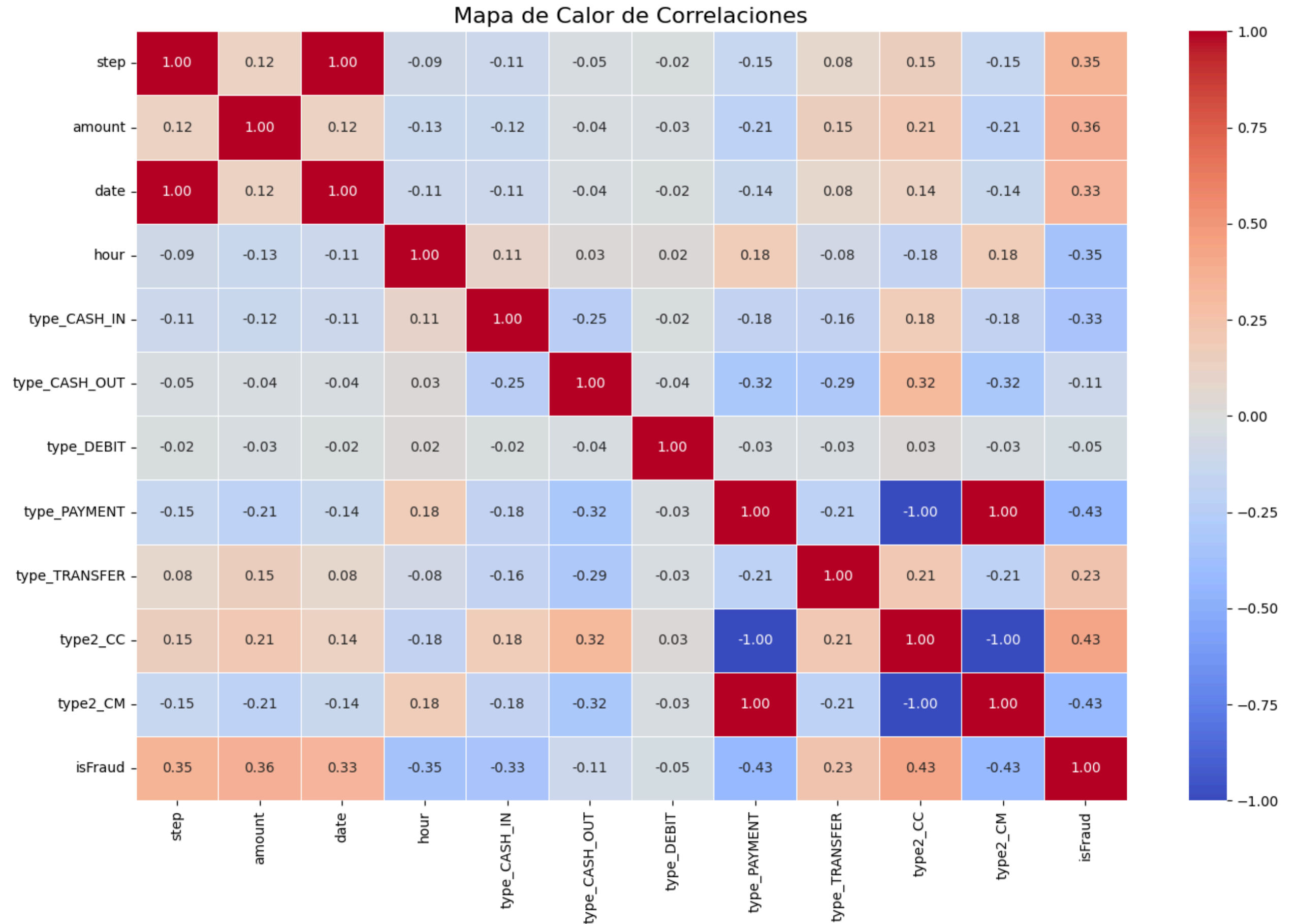
Grafico de pizza de la potabilidad



CGrafico de pizza de la potabilidad

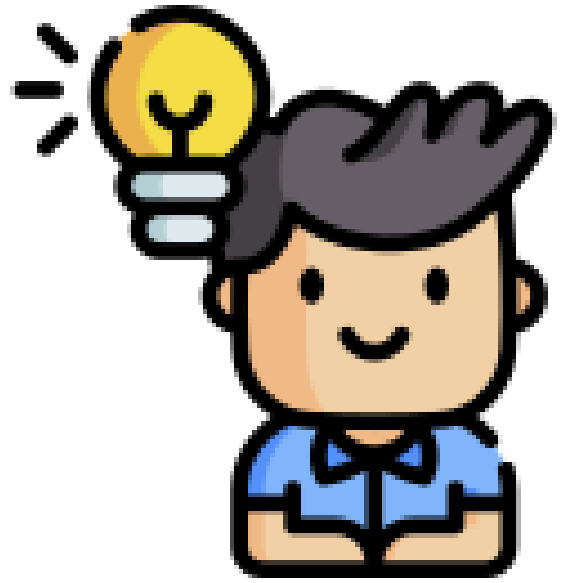


# Matriz de correlación



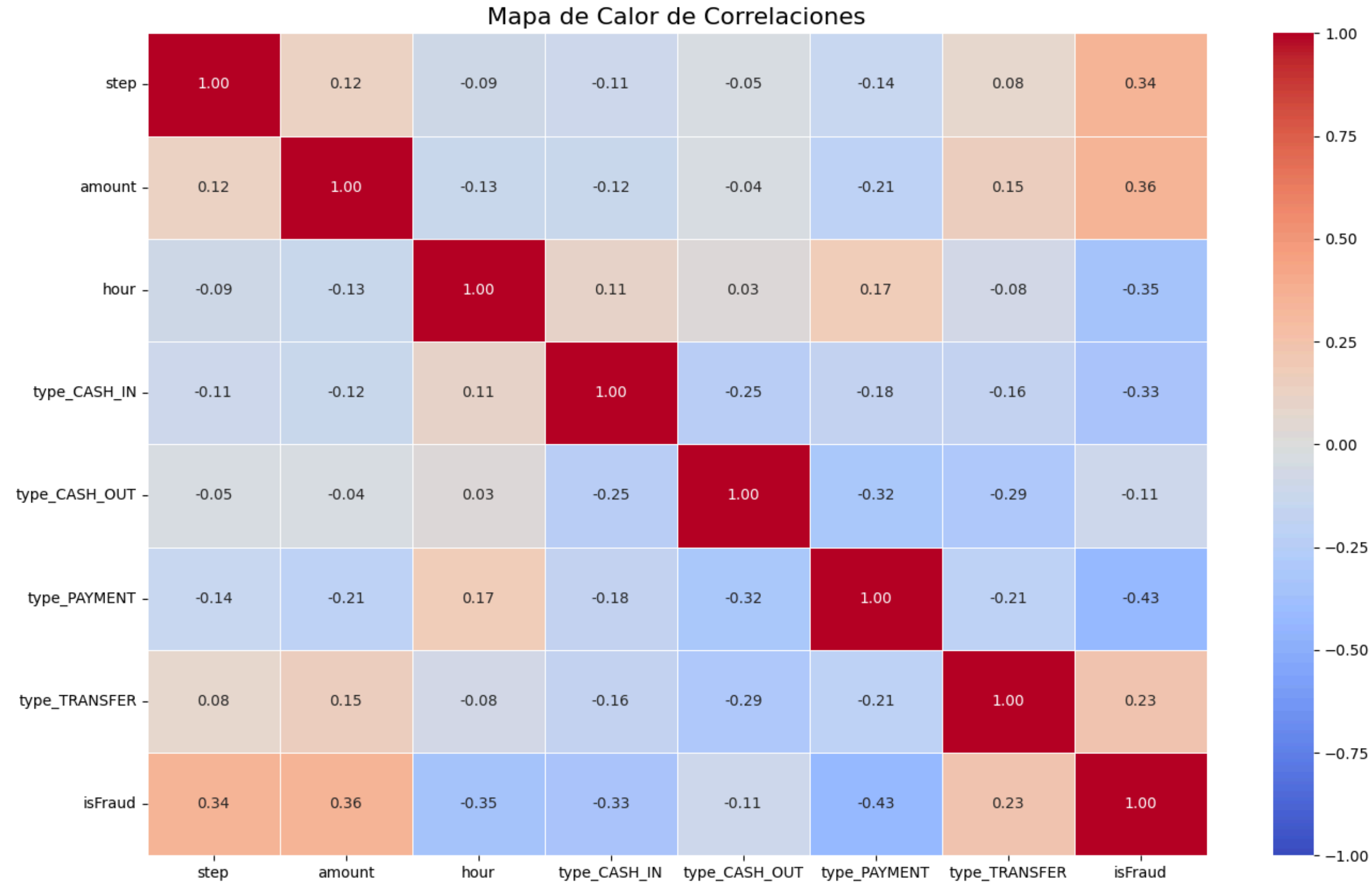


# Matriz de correlación



## Características Eliminadas

- type\_DEBIT
- type2\_CC
- type2\_CM
- Date

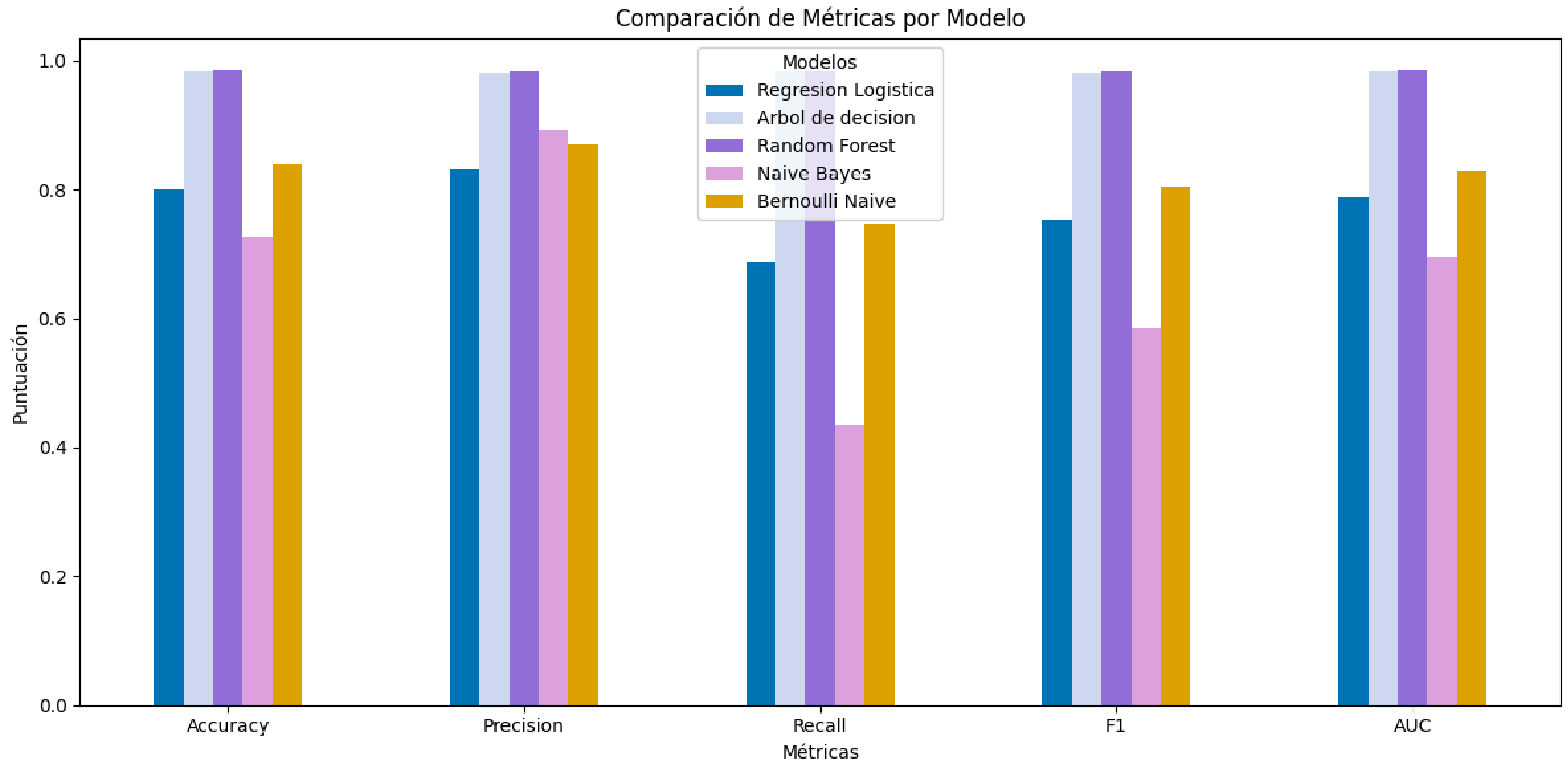
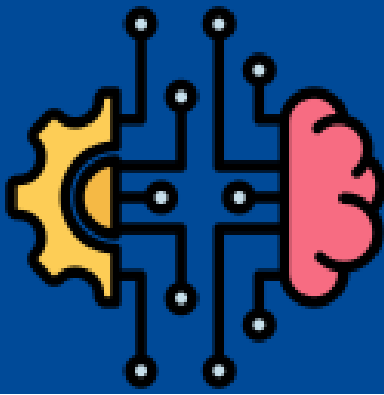


# Construcción y Evaluación de Modelos



- Experimentación con Algoritmos: Probamos una variedad de algoritmos de machine learning, incluyendo regresión logística, árbol de decisión, random forest, naive bayes y bernoulli bayes, para detectar transacciones fraudulentas.
- Evaluación Rigurosa: Evaluamos el rendimiento de cada modelo utilizando métricas clave como precisión, recall, F1-score y área bajo la curva ROC, asegurándonos de seleccionar el modelo más efectivo para la detección de fraudes.

# Evaluación de modelos de machine learning





# Conclusiones y recomendaciones



El modelo de Random Forest ha demostrado ser altamente efectivo, logrando una precisión y recall superiores al 98%. Esto sugiere que el modelo es capaz de identificar correctamente la mayoría de las transacciones fraudulentas sin muchos falsos positivos.

La alta precisión y recall indican que el modelo no solo es bueno para identificar fraudes, sino que también comete pocos errores al clasificar transacciones no fraudulentas como fraudulentas. Esto es crucial en un sistema de detección de fraudes, donde los falsos positivos pueden resultar en una mala experiencia para los clientes legítimos.

El AUC-ROC cercano a 1 muestra que el modelo tiene una excelente capacidad para discriminar entre transacciones fraudulentas y no fraudulentas. Este es un indicador fuerte de un buen desempeño del modelo en escenarios reales.



# Data Science Project



**financial  
fraud  
detection**



Jhonatan Rodriguez  
Jorge Perez  
Esteban Ferraz

