



BOOTCAMP XPERIENCE

ESTRATEGIAS PARA LA PREVENCIÓN DEL FRAUDE DIGITAL



Mayo 2024



"FRAUDE 0 NO EXISTE"



DESCUBRIENDO PATRONES DE FRAUDE

La urgencia por detectar fraudes en transacciones móviles de dinero requiere soluciones innovadoras.

Los estafadores siempre están buscando la próxima oportunidad para encontrar un agujero en el proceso, y los comerciantes siempre tendrán más trabajo por hacer y nuevos problemas que enfrentar cuando se trata de prevenir el fraude.

LA TRANSFORMACIÓN DIGITAL ESTÁ CAMBIANDO EL PANORAMA DEL FRAUDE

+6%
CRECIMIENTO DE TRANSACCIONES DIGITALES
2023
+14%
CRECIMIENTO DEL FRAUDE DIGITAL+
+133%
CRECIMIENTO DEL FRAUDE DIGITAL

+70%
DE LA POBLACIÓN EN LATAM UTILIZA LOS SERVICIOS EN LÍNEA



+5%
TRANSACCIONES DIGITALES A NIVEL MUNDIAL FUERON SOSPECHOSAS DE FRAUDE DIGITAL
+13.5%
DE TODAS LAS TRANSACCIONES GLOBALES PARA LA CREACIÓN DE CUENTAS DIGITALES EN 2023 FUERON SOSPECHOSAS DE FRAUDE DIGITAL.

2.4 % ANUAL
TASA DE CRECIMIENTO DE ADOPCIÓN DE INTERNET



NUESTRO OBJETIVO ES EL SIMPLIFICAR TU LUCHA
CONTRA EL FRAUDE DIGITAL

COMBATE EL FRAUDE EN TUS CANALES DIGITALES
MINIMIZANDO LA FRICCIÓN PARA TUS CLIENTES
Y REDUCIENDO LA CARGA OPERATIVA PARA TU
ORGANIZACIÓN.

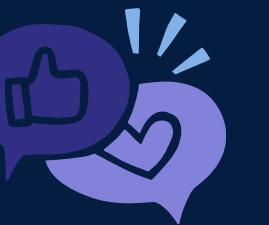
EN TIEMPO REAL



PRECISO

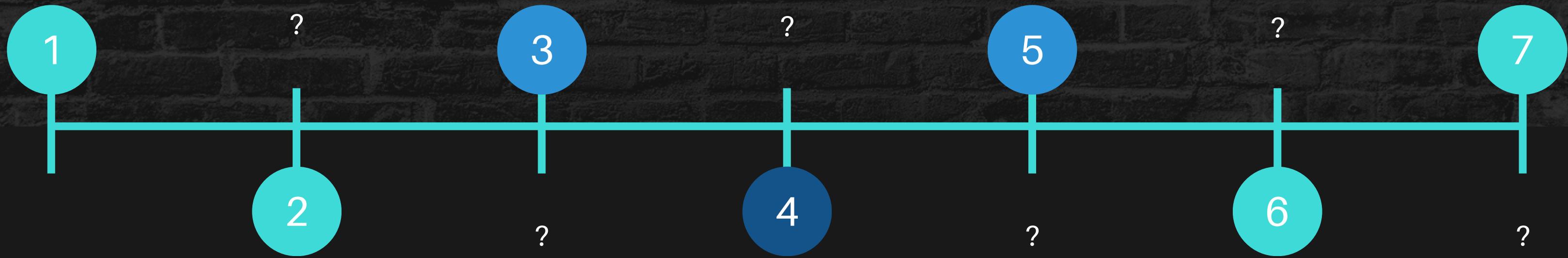


EFECTIVO

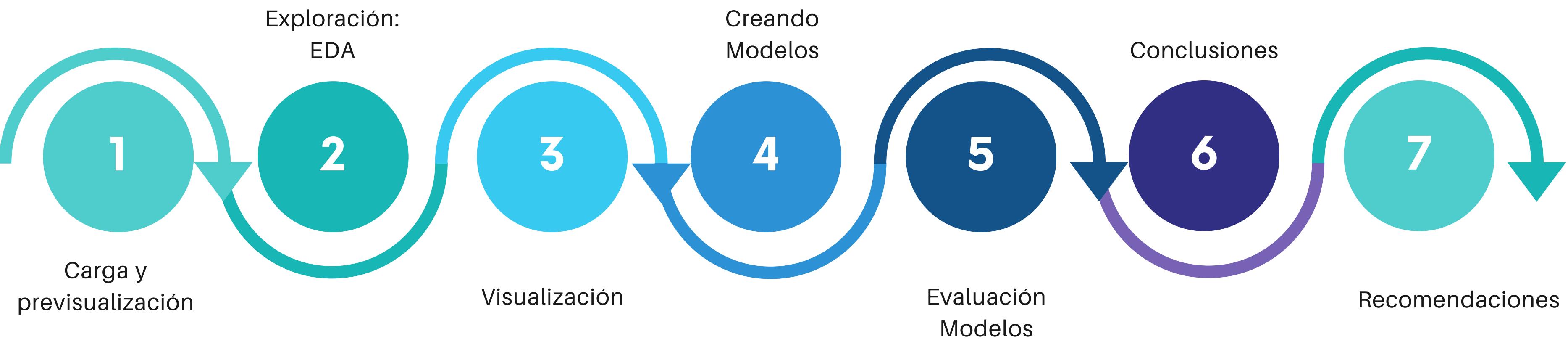


Conociendo la base de datos...

que tipo de información tenemos disponible?



ETAPAS DEL ANÁLISIS



DESCRIPCIÓN DEL DATASET

Data de cuentas de usuarios

```
df.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   oldbalanceOrg 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
```

Tipos de datos,
descripción, valores

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYOUT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYOUT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYOUT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
count	6362620.0	6362620.0	6362620.0	6362620.0	6362620.0	6362620.0	6362620.0	6362620.0
mean	243.4	179861.9	833883.1	855113.7	1100701.7	1224996.4	0.0	0.0
std	142.3	603858.2	2888242.7	2924048.5	3399180.1	3674128.9	0.0	0.0
min	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	156.0	13389.6	0.0	0.0	0.0	0.0	0.0	0.0
50%	239.0	74871.9	14208.0	0.0	132705.7	214661.4	0.0	0.0
75%	335.0	208721.5	107315.2	144258.4	943036.7	1111909.2	0.0	0.0
max	743.0	92445516.6	59585040.4	49585040.4	356015889.4	356179278.9	1.0	1.0

Identificación de los datos



Step: Mapea una unidad de tiempo en el mundo real. En este caso, 1 paso es 1 hora de tiempo.

Amount : Importe de la transacción en moneda local.

OldbalanceOrg: Saldo inicial antes de la transacción

NameDest: Cliente que es el destinatario de la transacción

Type: Cash-in, Cash-out, Debit, Payment y Transfer.

NameOrig: Cliente que inició la transacción

NewbalanceOrig : Nuevo saldo después de la transacción

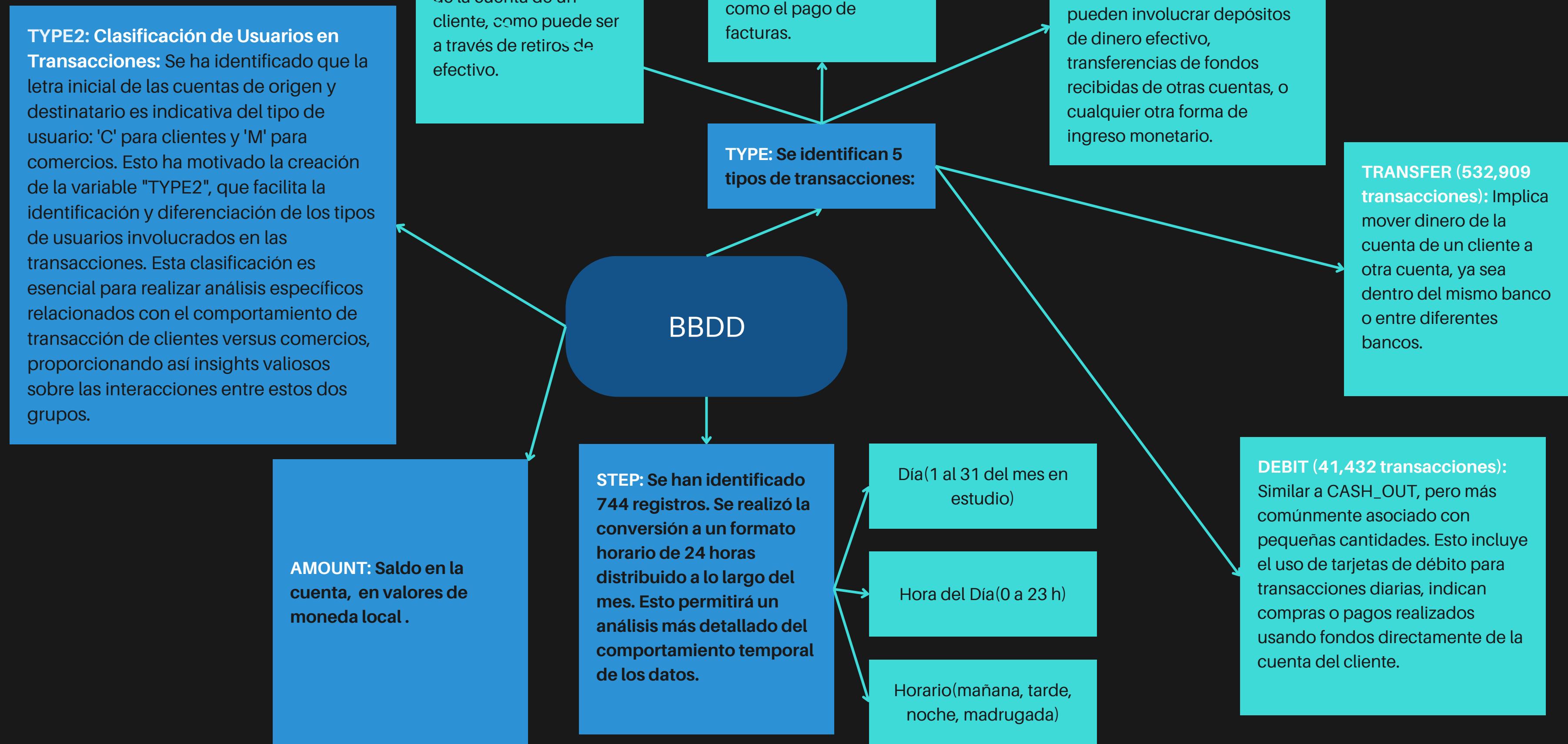
OldbalanceDest: Destinatario del saldo inicial antes de la transacción. Sin información para los clientes que comienzan con M (comerciantes).

NewbalanceDest: Destinatario del nuevo saldo después de la transacción. Sin información para los clientes que comienzan con M (comerciantes).

IsFlaggedFraud: El modelo de negocio tiene como objetivo controlar las transferencias masivas de una cuenta a otra y señala los intentos ilegales. Un intento ilegal en este conjunto de datos es un intento de transferir más de 200.000 en una sola transacción

IsFraud: Son transacciones realizadas por los agentes fraudulentos dentro de la simulación. El comportamiento fraudulento de los agentes tiene como objetivo obtener ganancias tomando el control de las cuentas de los clientes e intentar vaciar los fondos transfiriéndolos a otra cuenta y luego retirándolos del sistema.

Definición de Variables



Featuring_Engineering

Featuring Engineering

Variable Rango_fondo

Categorizar los valores de **oldbalanceDest** según los rangos:

- **Sin saldo:** Cuando oldbalanceDest es igual a 0.
- **Saldo :** Cuando oldbalanceDest es mayor que 0

Variable intento_vaciar

Esta variable ayuda a identificar si se intentó vaciar la cuenta de origen, una práctica comúnmente observada en incidentes de fraude. Para adaptarse a esta evolución en el comportamiento del fraude, hemos establecido un margen que considera un saldo remanente del 5%. Si el monto de la transacción representa el 95% o más del saldo original, se considera un intento de vaciar la cuenta, capturando así tanto los intentos completos como los casi completos de extracción de fondos.

* 1 = Si intentó vaciar cuenta

* 0 = No intentó vaciar cuenta

Variable hora de dia :

Hora del dia se usara el operador % para transformar cada step en una hora del dia de **0 a 23**.

Variable DIA

La columna se creó para determinar el día en el rango de 1 a 31 días.

Variable horario

Clasifica las horas del día en categorías como "**Madrugada**", "**Mañana**", "**Tarde**" y "**Noche**".

Evaluando Transacciones Fraudulentas

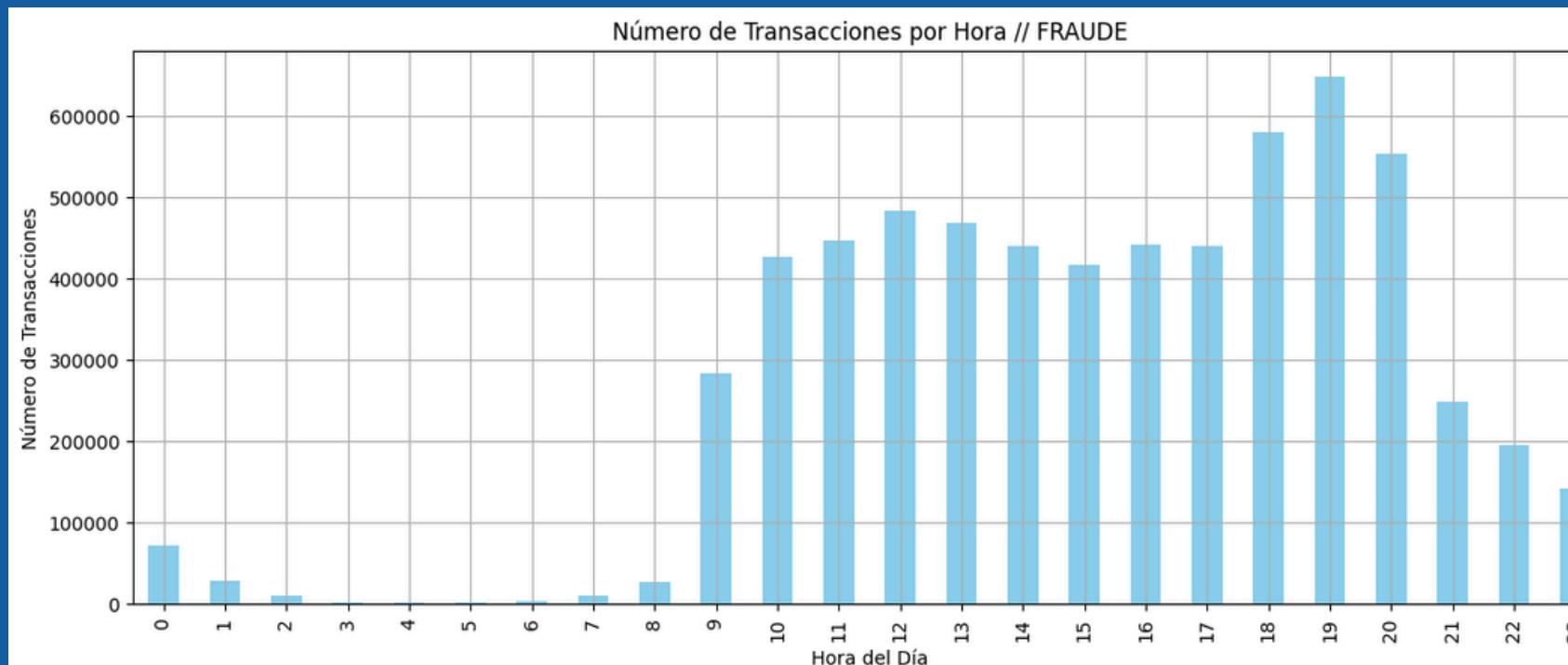
IS FRAUD = 1



IS FRAUD = 1

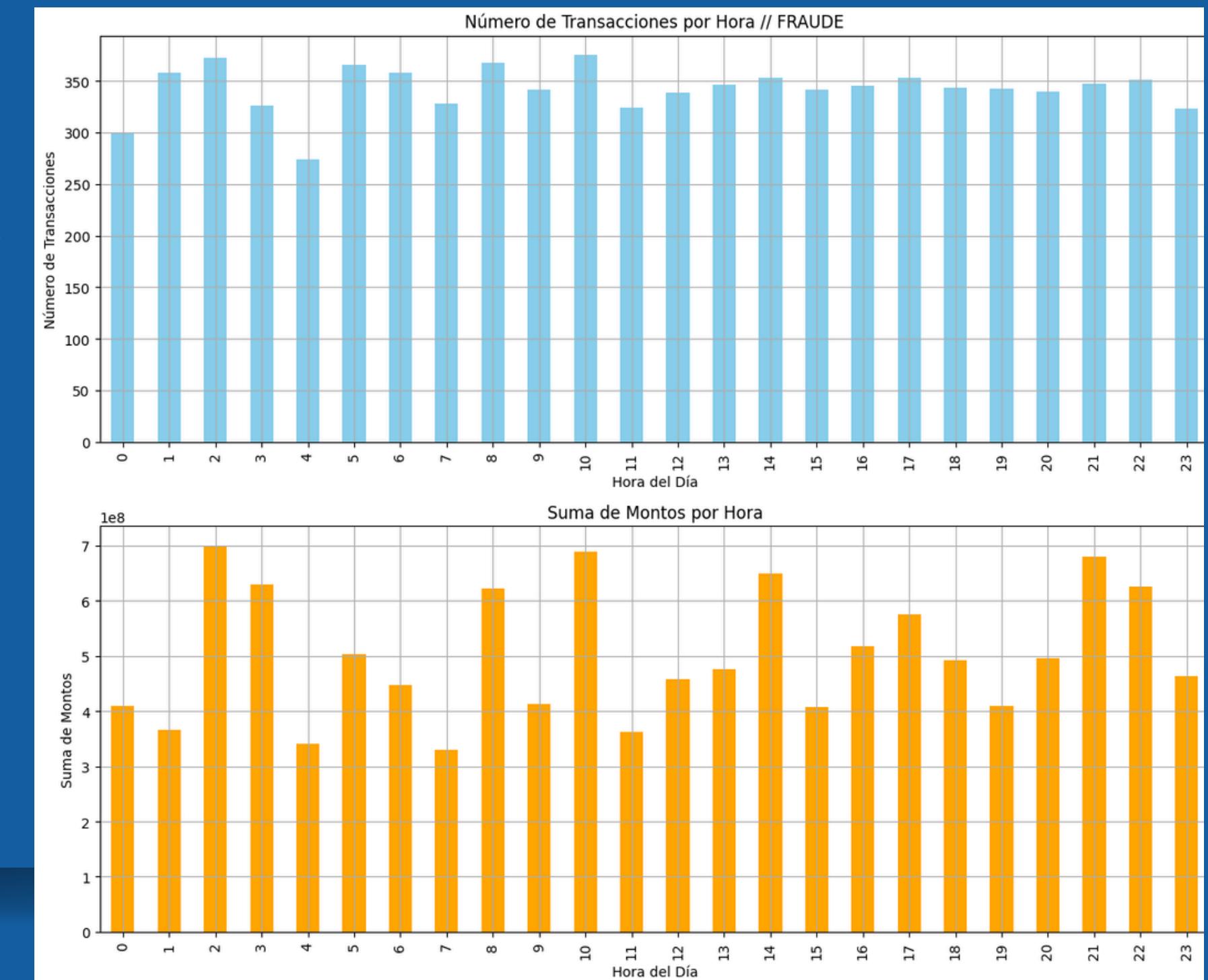
TRANSACCIONES POR HORA

Fraude en base de datos Total



VS

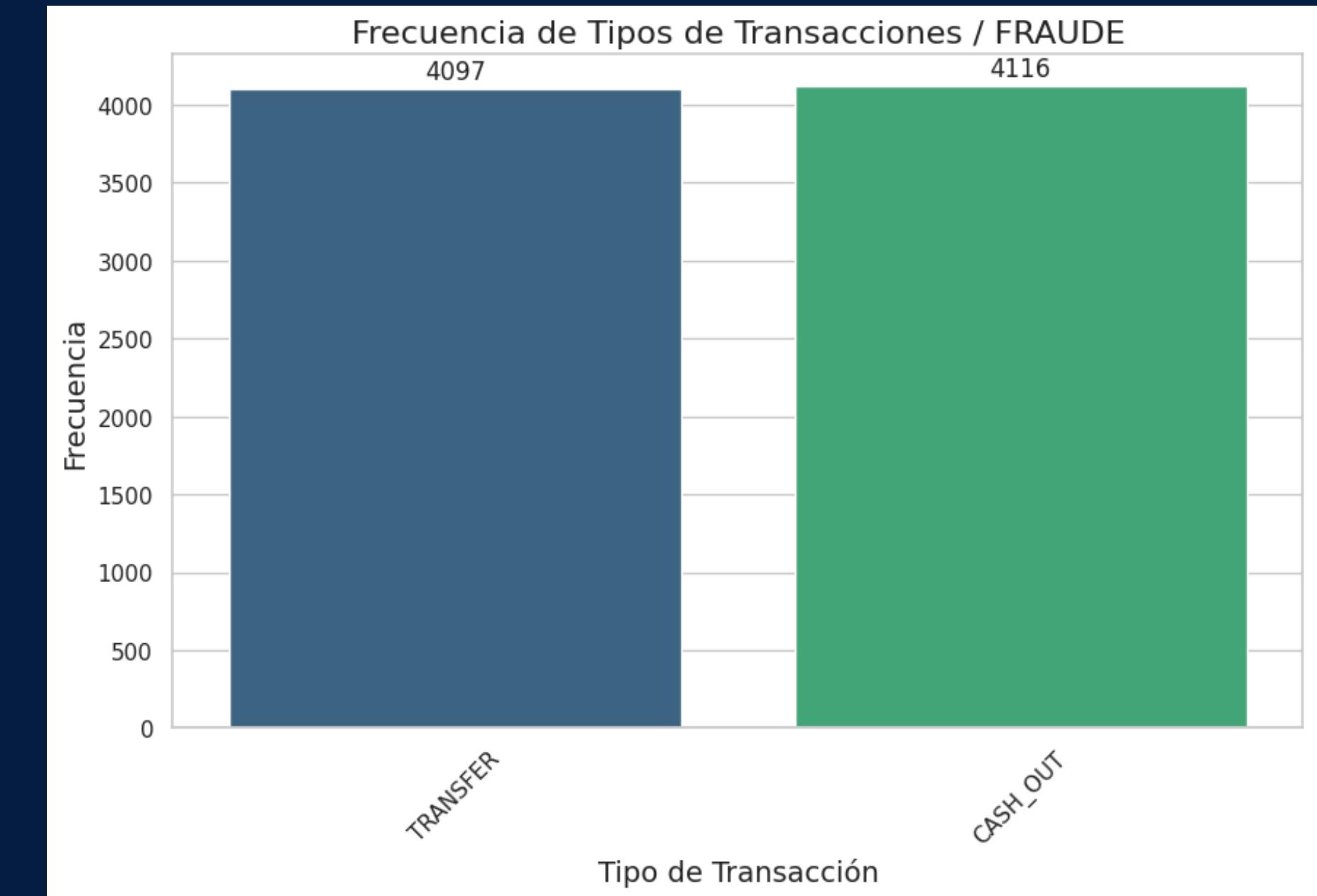
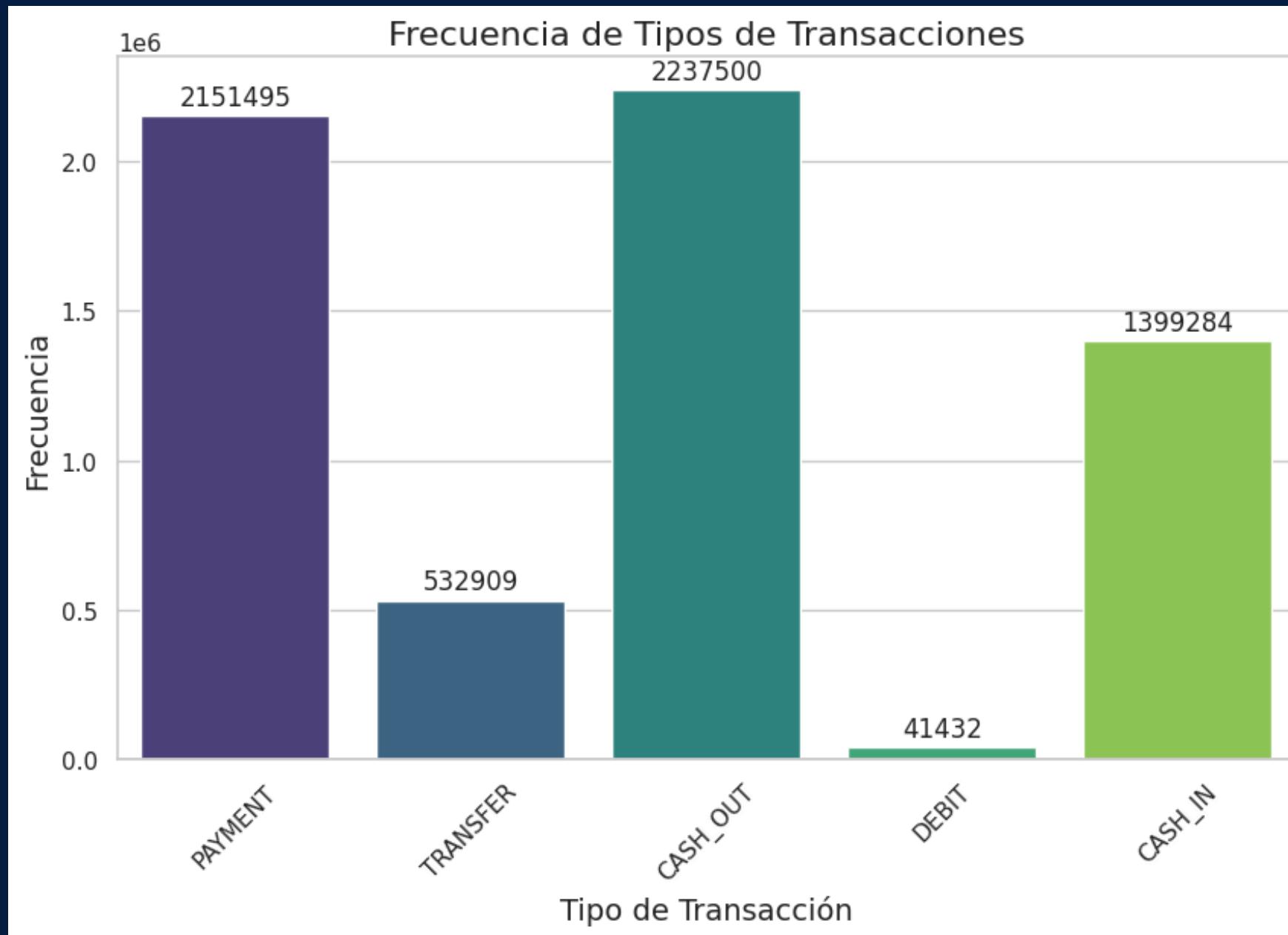
Base de datos de registros fraudulentos únicamente



TRANSACCIONES

Por
TIPO

Fraude identificado sólo en 2 tipos

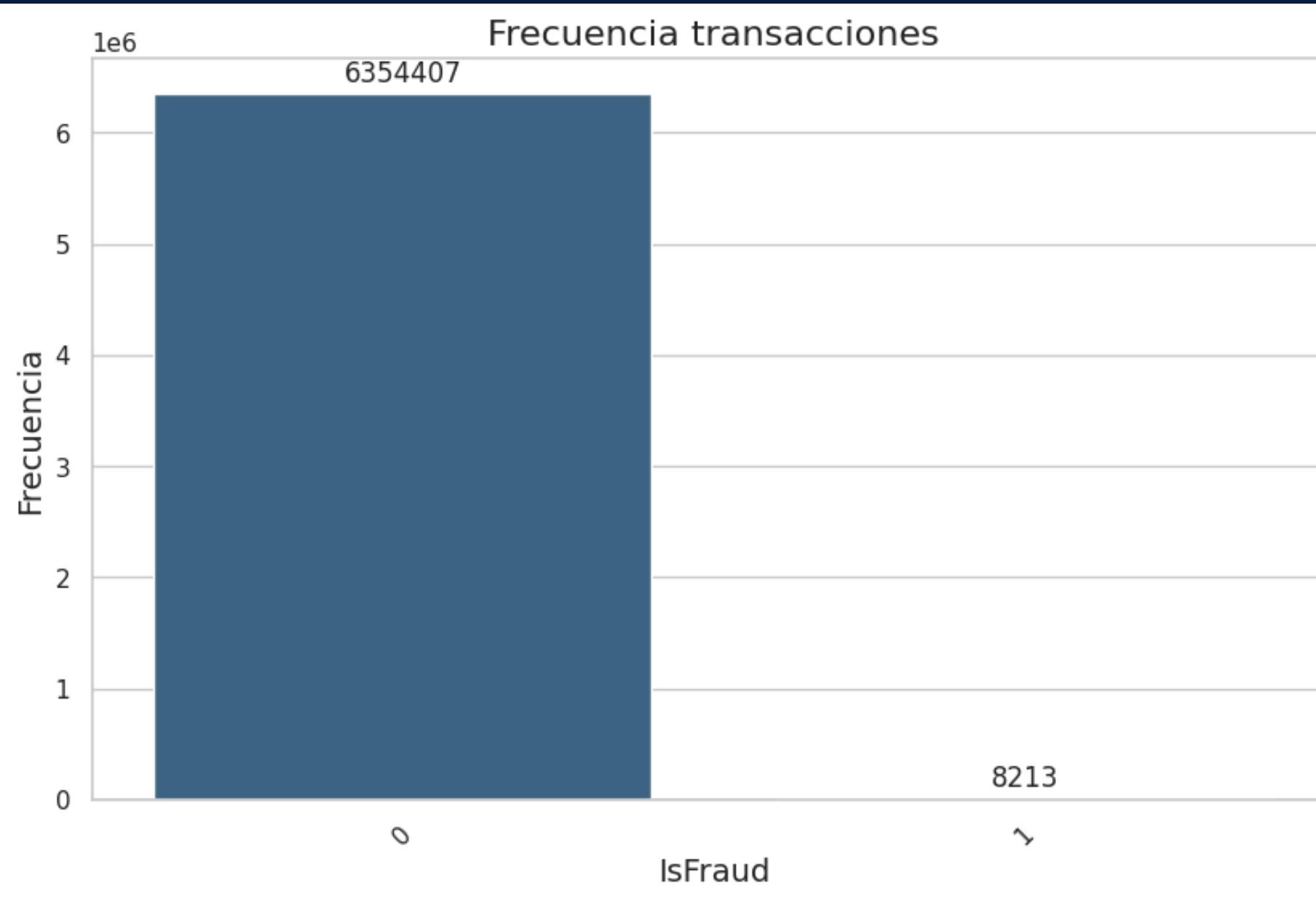


TRANSACCIONES

Por
FRAUDE

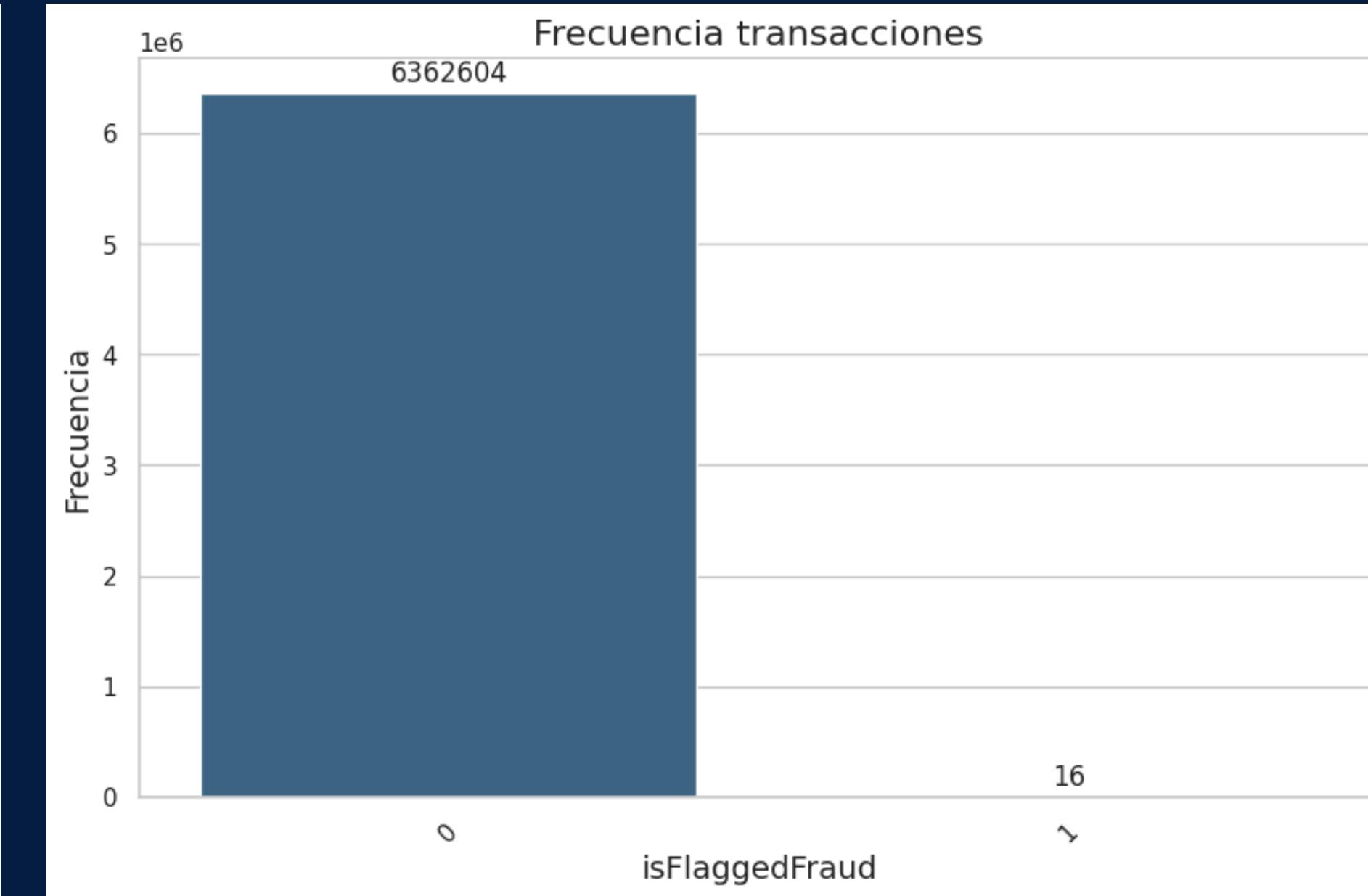
no=0

si=1



Masivo
si=1

no=0



TRANSACCIONES

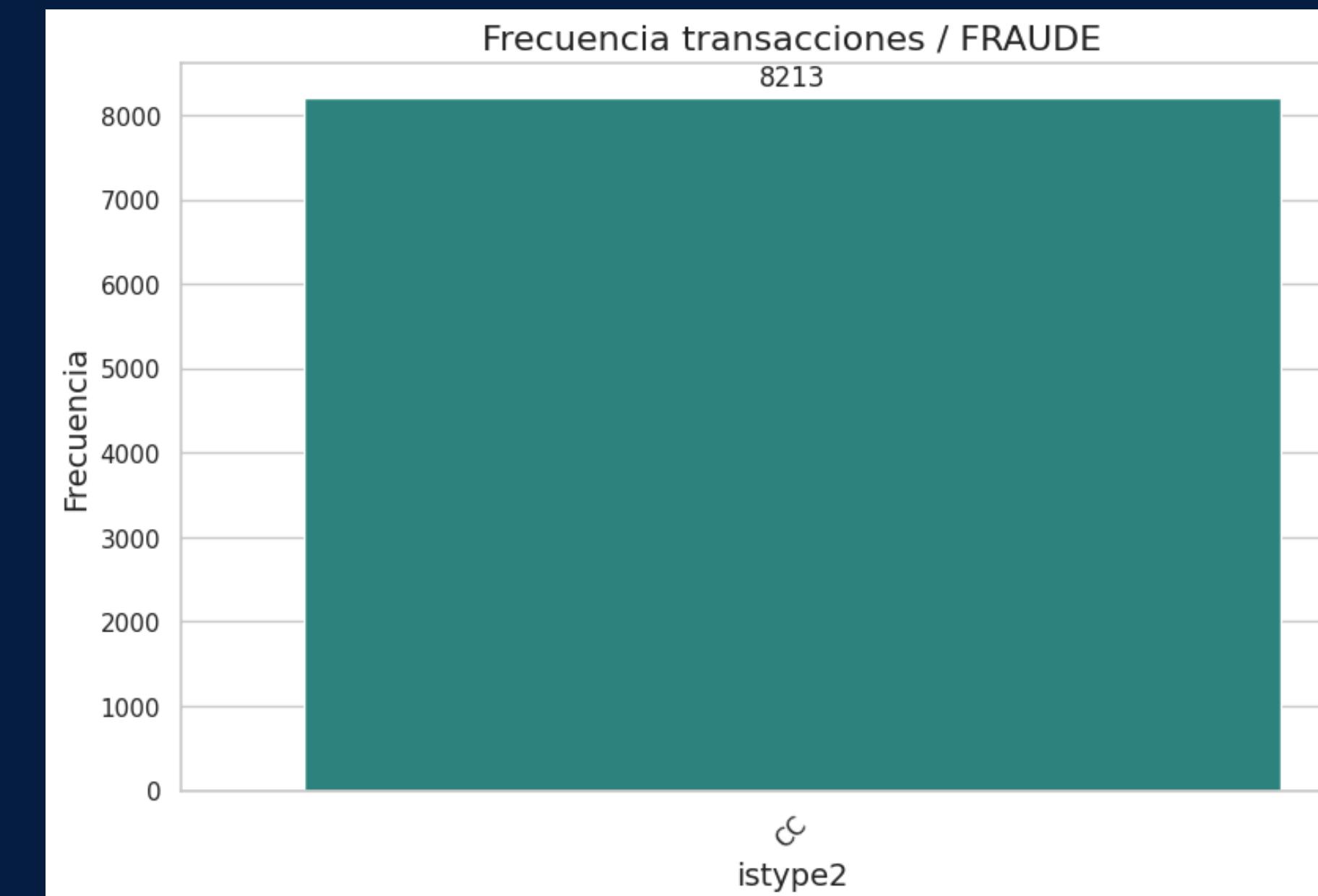
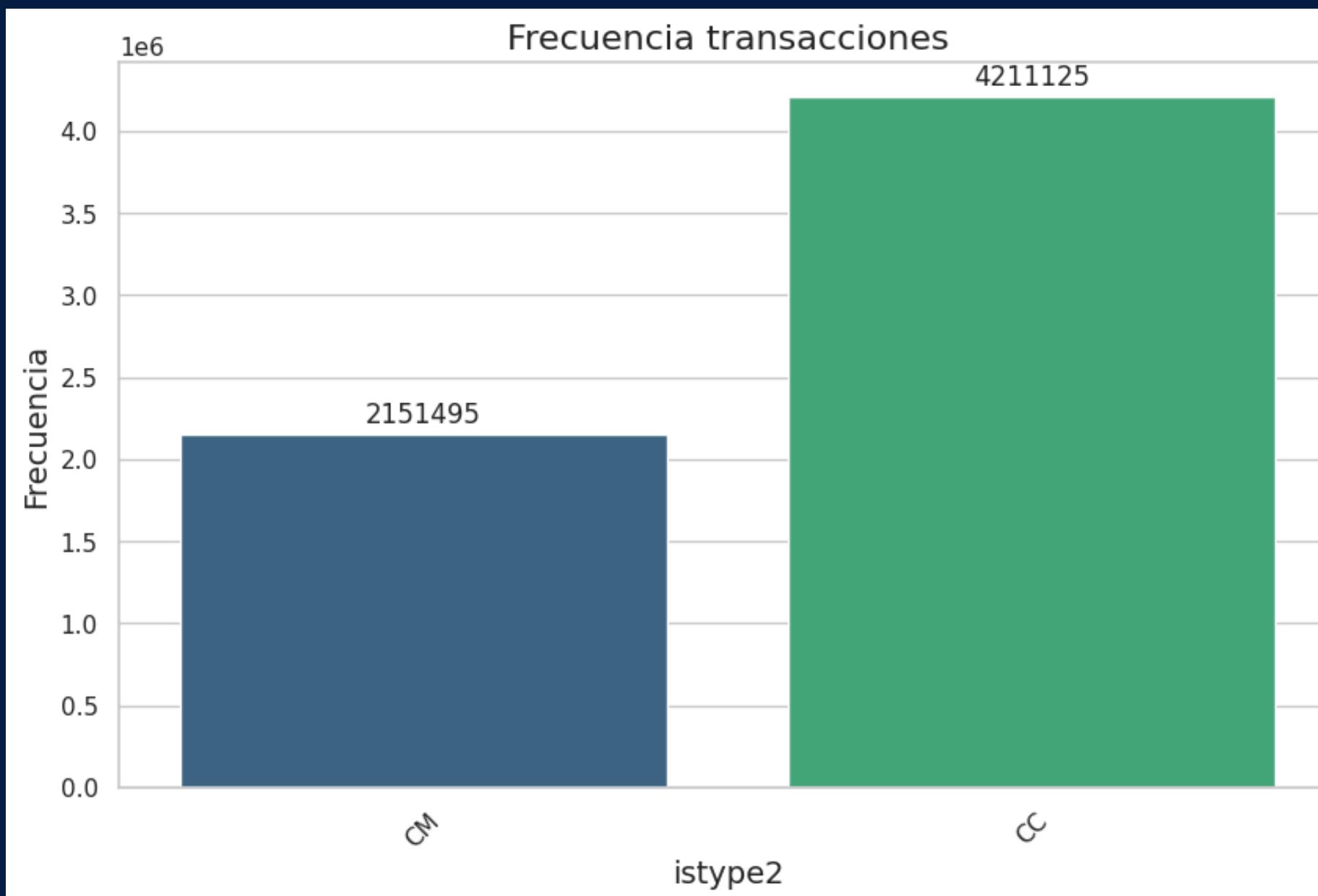
Por
TIPO
USUARIO

C=cliente

M=comerciante



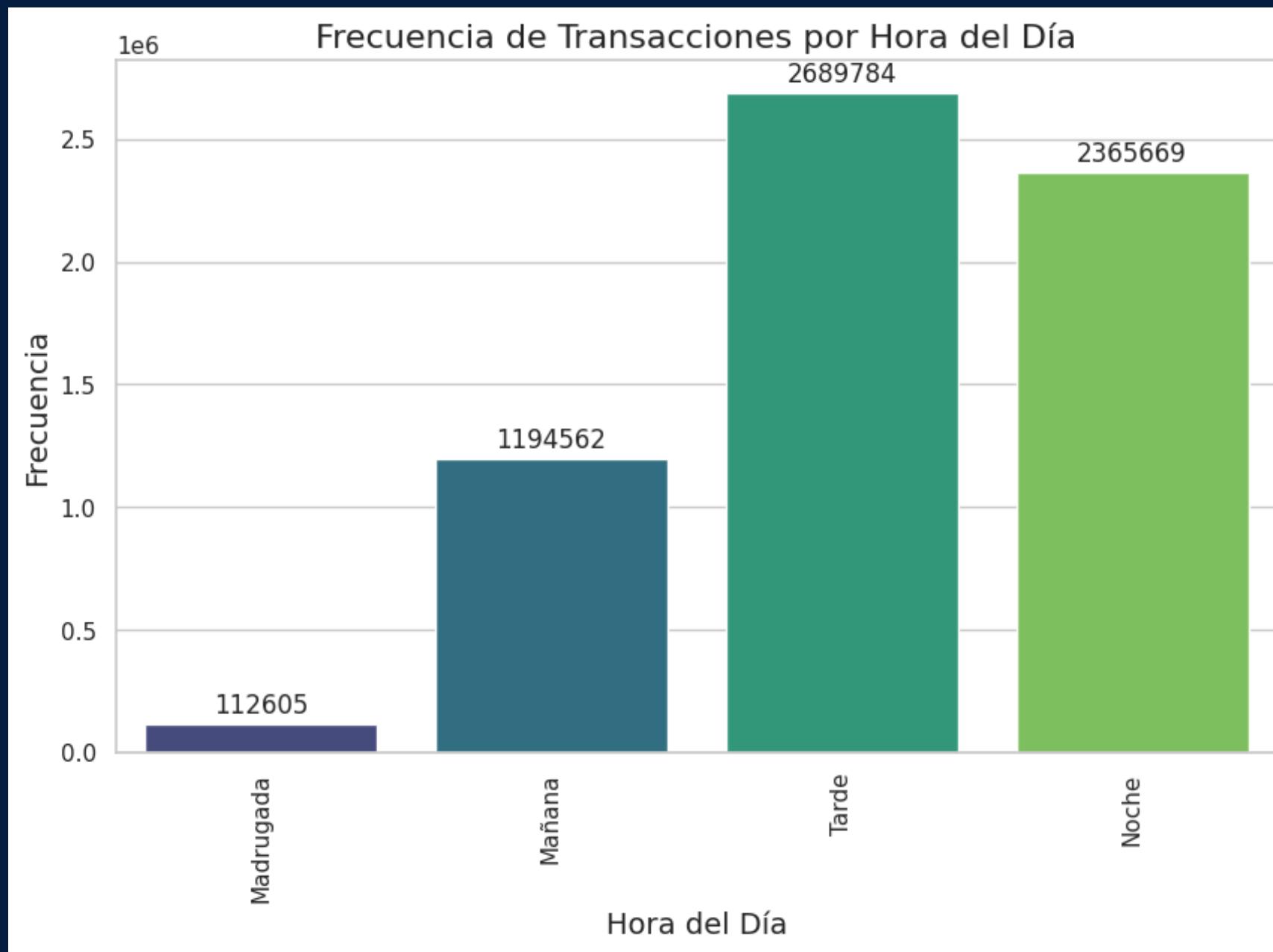
Fraude identificado sólo en tipo CC



TRANSACCIONES

Por
HORA DEL DÍA

En base de datos Total



Base de datos de registros fraudulentos únicamente

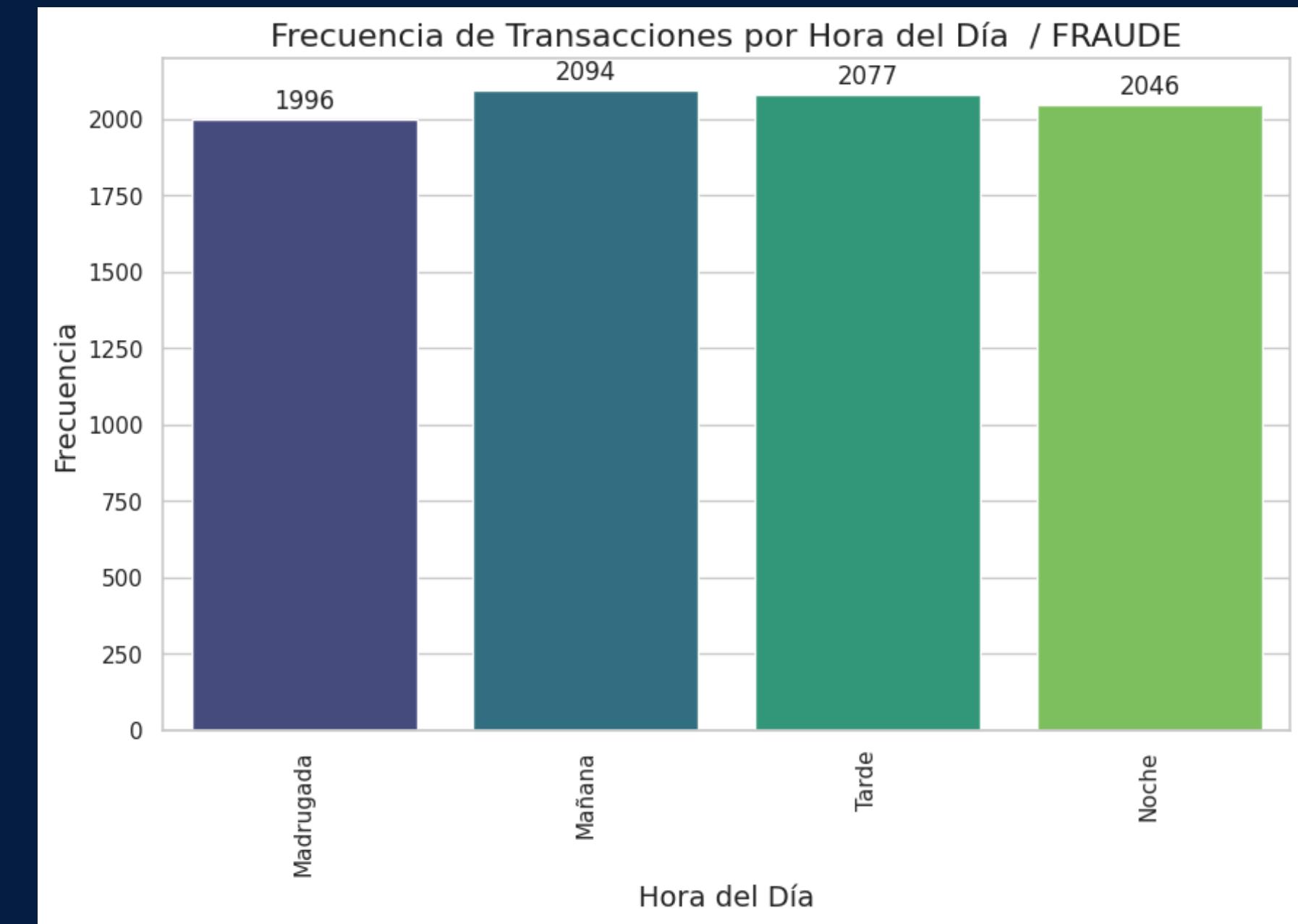


Gráfico de variable Rango_Fondo

Base de datos de registros fraudulentos únicamente

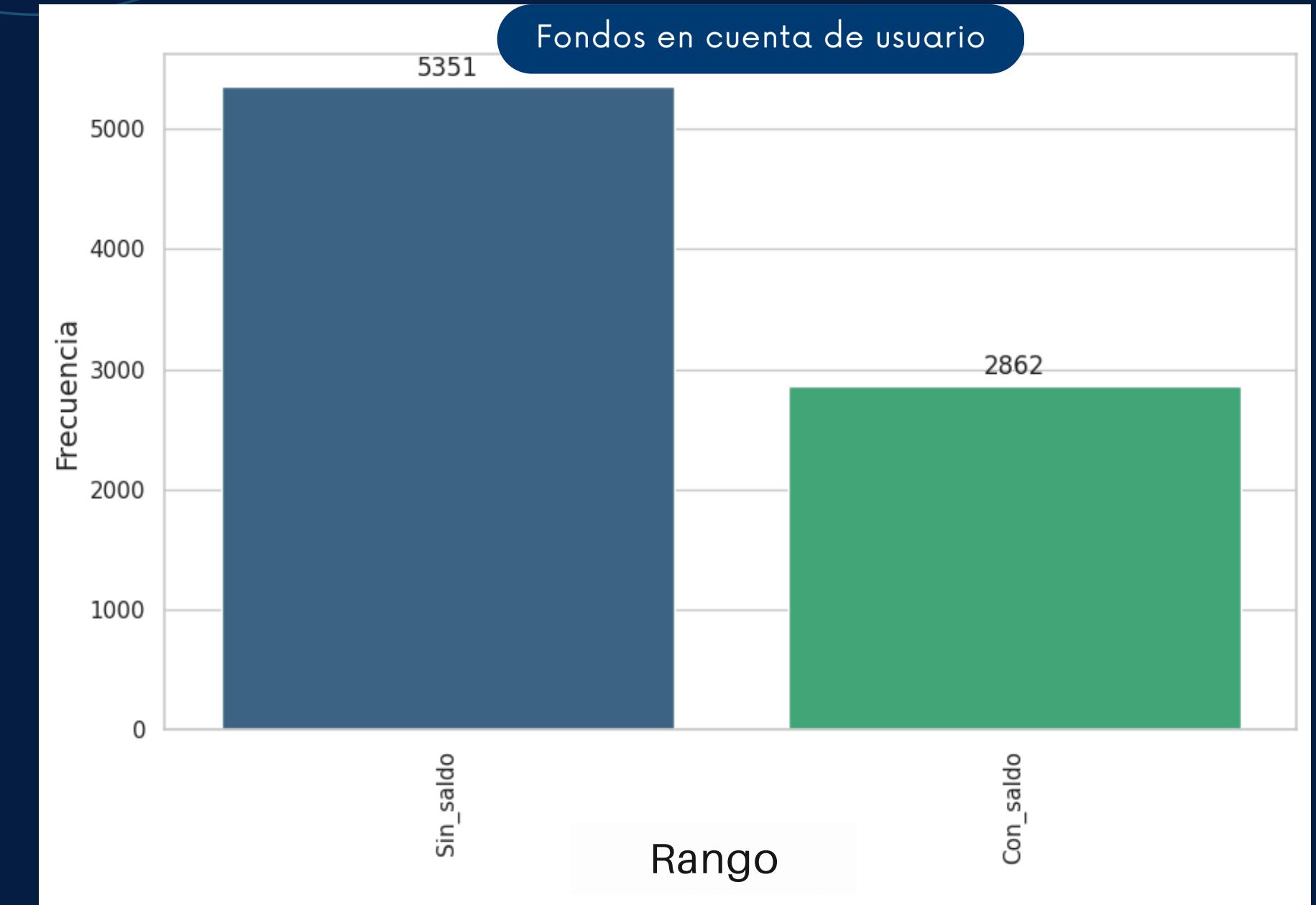
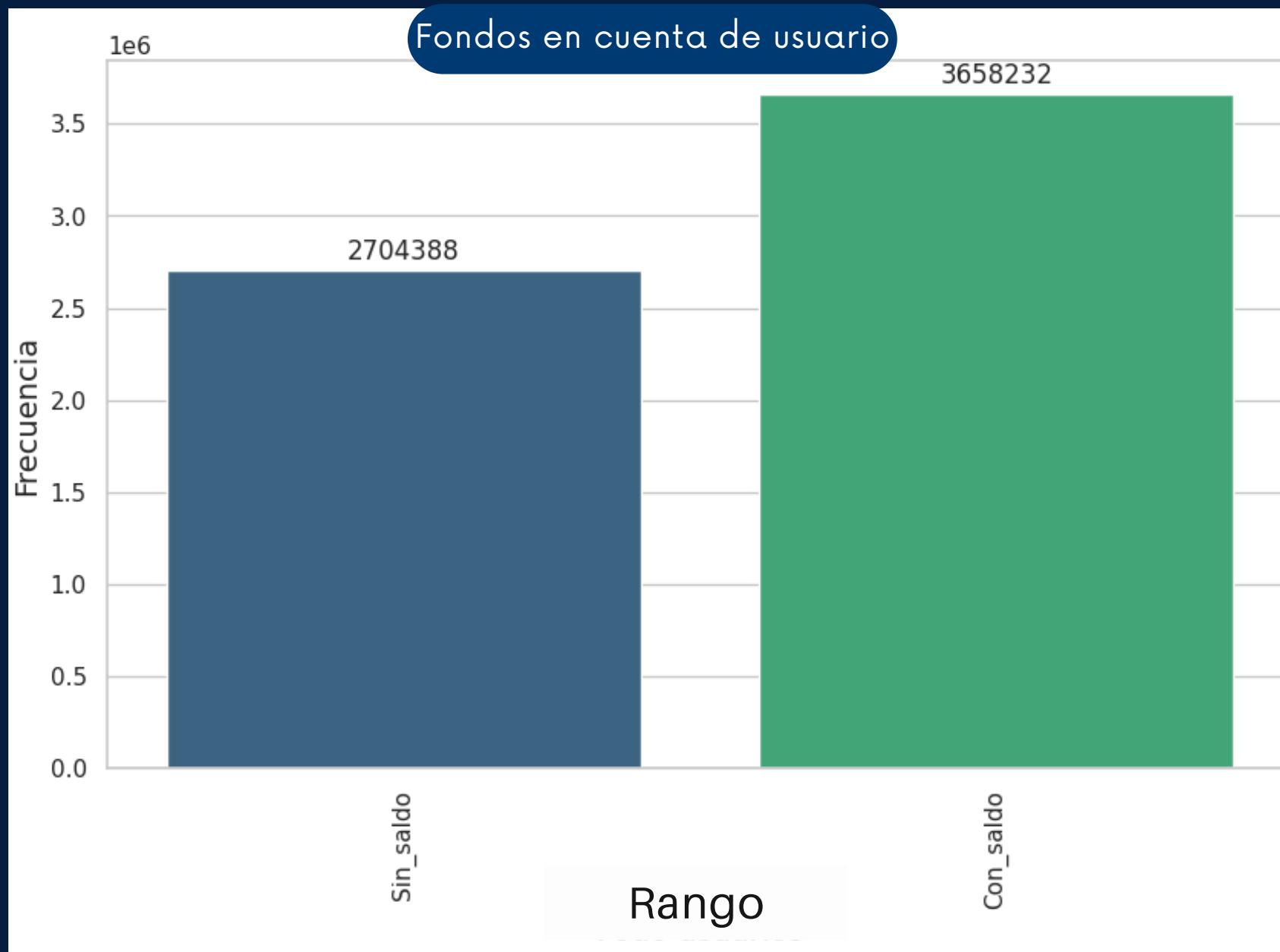
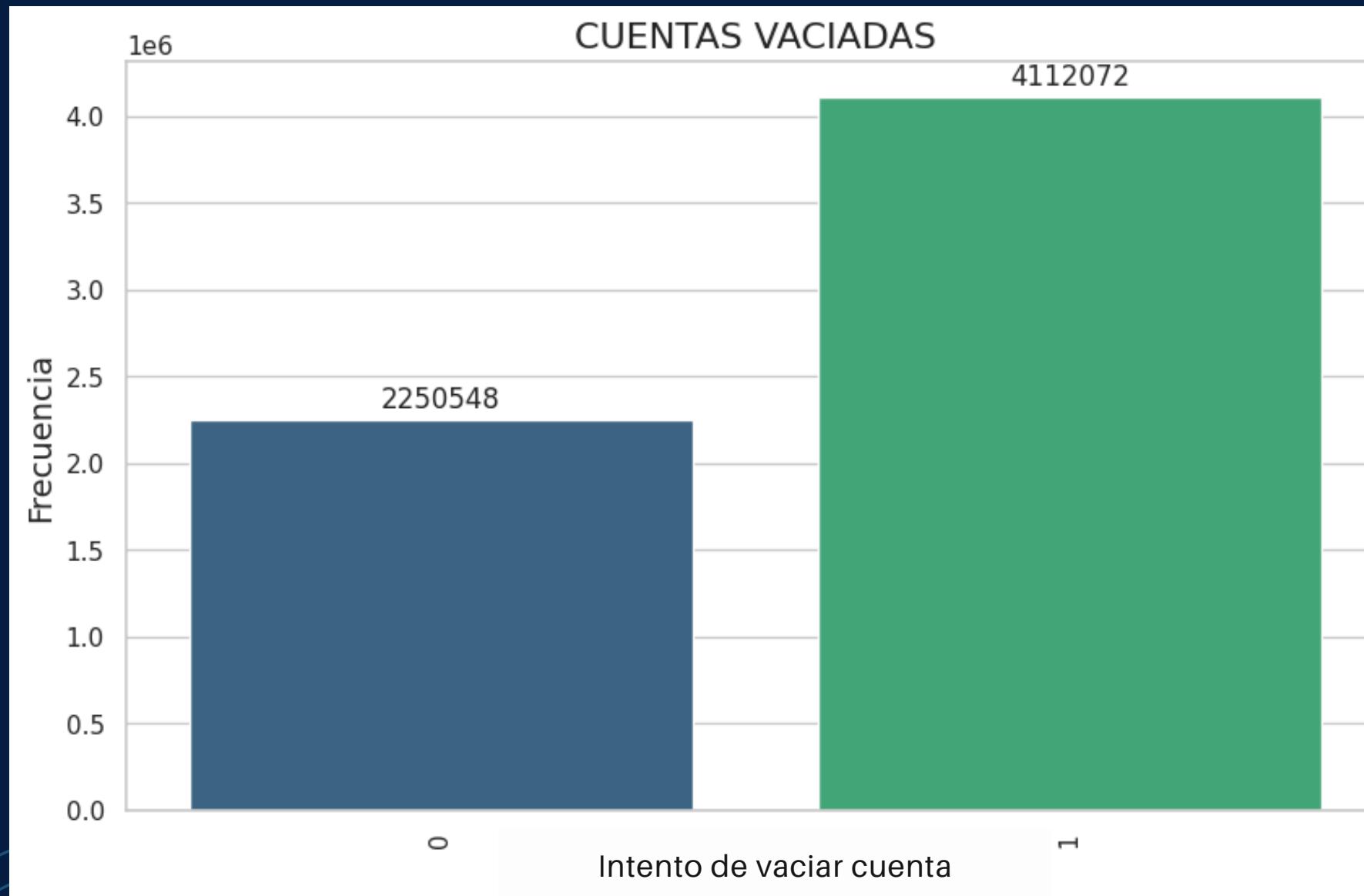


Gráfico de variable Cuentas_vaciadas

No intentó vaciar cuenta=0

Si intentó vaciar cuenta=1



Los defraudadores a menudo modifican sus estrategias cuando detectan que un patrón de fraude ha sido identificado, adaptando sus métodos para evitar la detección. En lugar de vaciar completamente las cuentas, pueden optar por dejar un pequeño saldo remanente.



PROPUESTA DE VALOR



PREMISAS

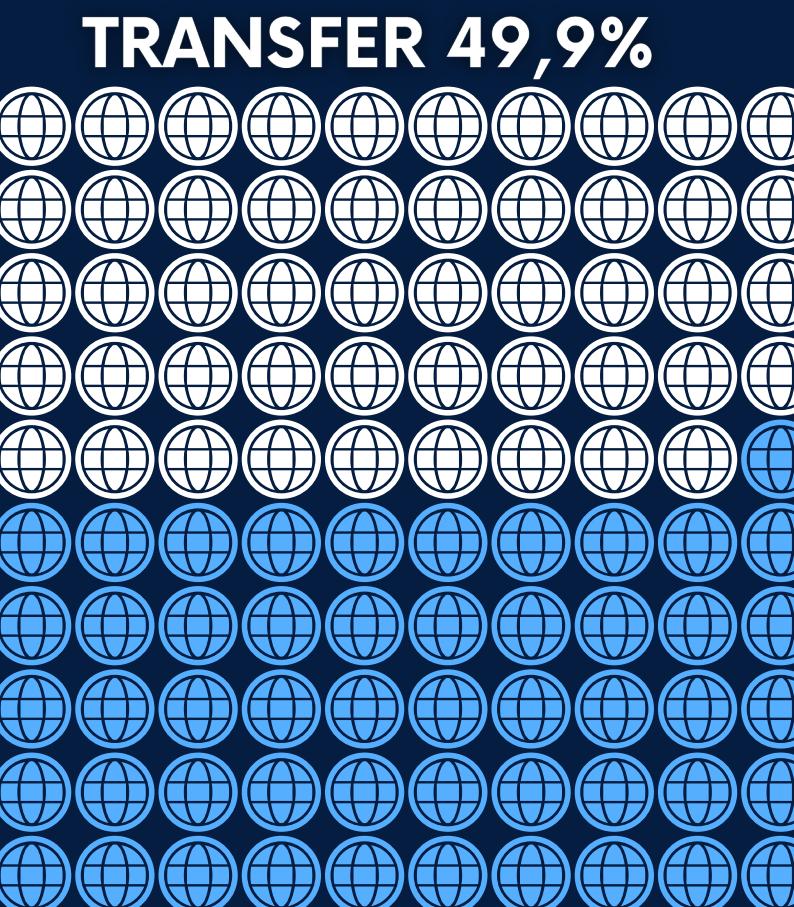


Se identifica que el fraude se aglutina solo en dos tipos de canales :

Cash Out y Transfer

```
<class 'pandas.core.frame.DataFrame'>
Index: 2770409 entries, 2 to 6362619
Data columns (total 17 columns):
 #   Column           Dtype  
--- 
 0   step             int64  
 1   type             object  
 2   amount            float64 
 3   nameOrig          object  
 4   oldbalanceOrg     float64 
 5   newbalanceOrig    float64 
 6   nameDest           object  
 7   oldbalanceDest     float64 
 8   newbalanceDest    float64 
 9   isFraud            int64  
 10  isFlaggedFraud    int64  
 11  type2             object  
 12  hora_del_dia      int64  
 13  dia               int64  
 14  horario            object  
 15  rango_fondo       object  
 16  intento_vaciar    int64  
dtypes: float64(5), int64(6), object(6)
memory usage: 380.5+ MB
```

**Total transacciones
2,77 M**

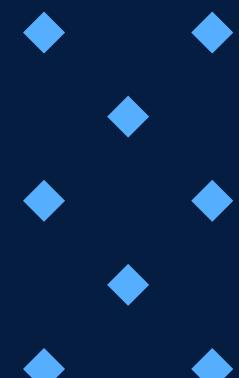


Estos canales han mostrado una incidencia significativamente mayor de actividades fraudulentas en comparación con otros métodos.

Nuestra propuesta de valor esta enfocada en el monitoreo y prevención exclusivamente de estos dos tipos de transacciones.

Esta estrategia podría reducir la cantidad de falsos positivos y optimizar la eficiencia de nuestros sistemas de detección de fraude.

Además, no se considera prudente incluir otros canales en el análisis debido a sus características inherentes, las cuales podrían no representar un riesgo potencial.



BALANCEO DE CLASES

SMOTE

Manejo del Desequilibrio de Clases:

SMOTE:

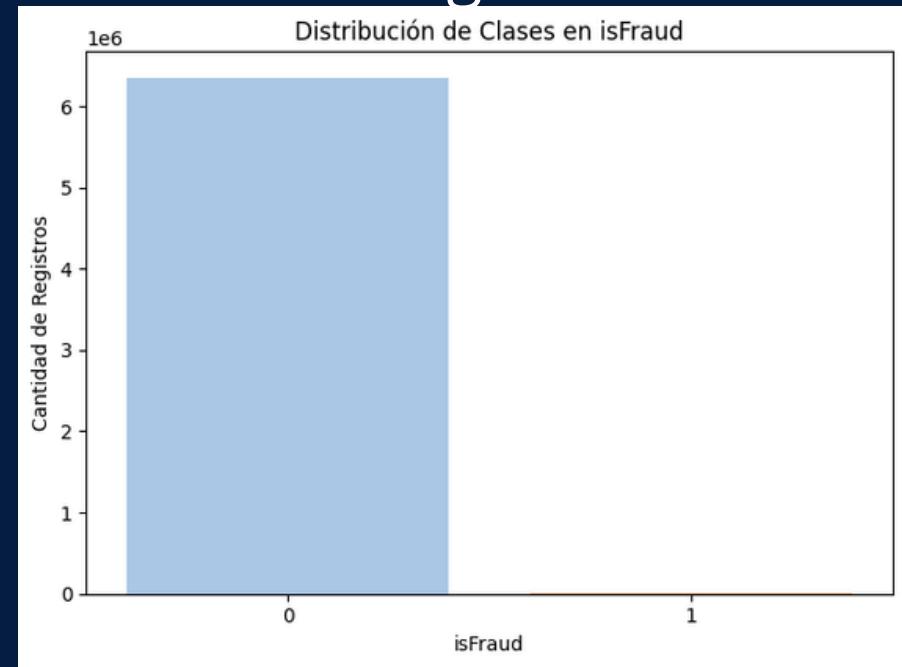
Ventajas:

- No modifica directamente los datos, lo que puede ser útil si el tamaño del conjunto de datos es limitado.
- Puede evitar problemas de sobreajuste que pueden surgir con técnicas de muestreo.

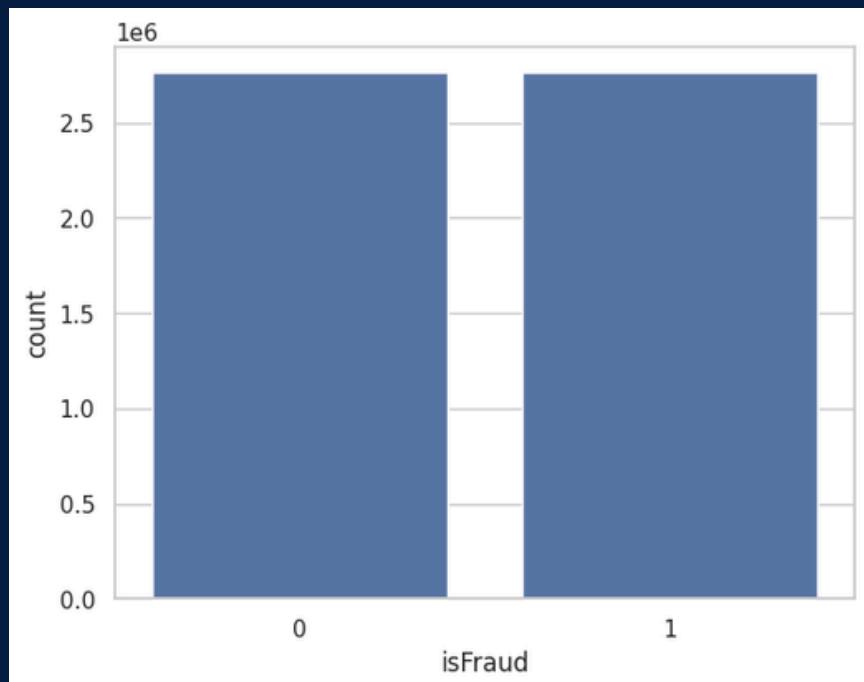
Desventajas:

- Puede no ser suficiente para abordar desequilibrios extremos.
- Dependiendo del modelo y la implementación, puede requerir ajustes adicionales en la configuración de pesos.

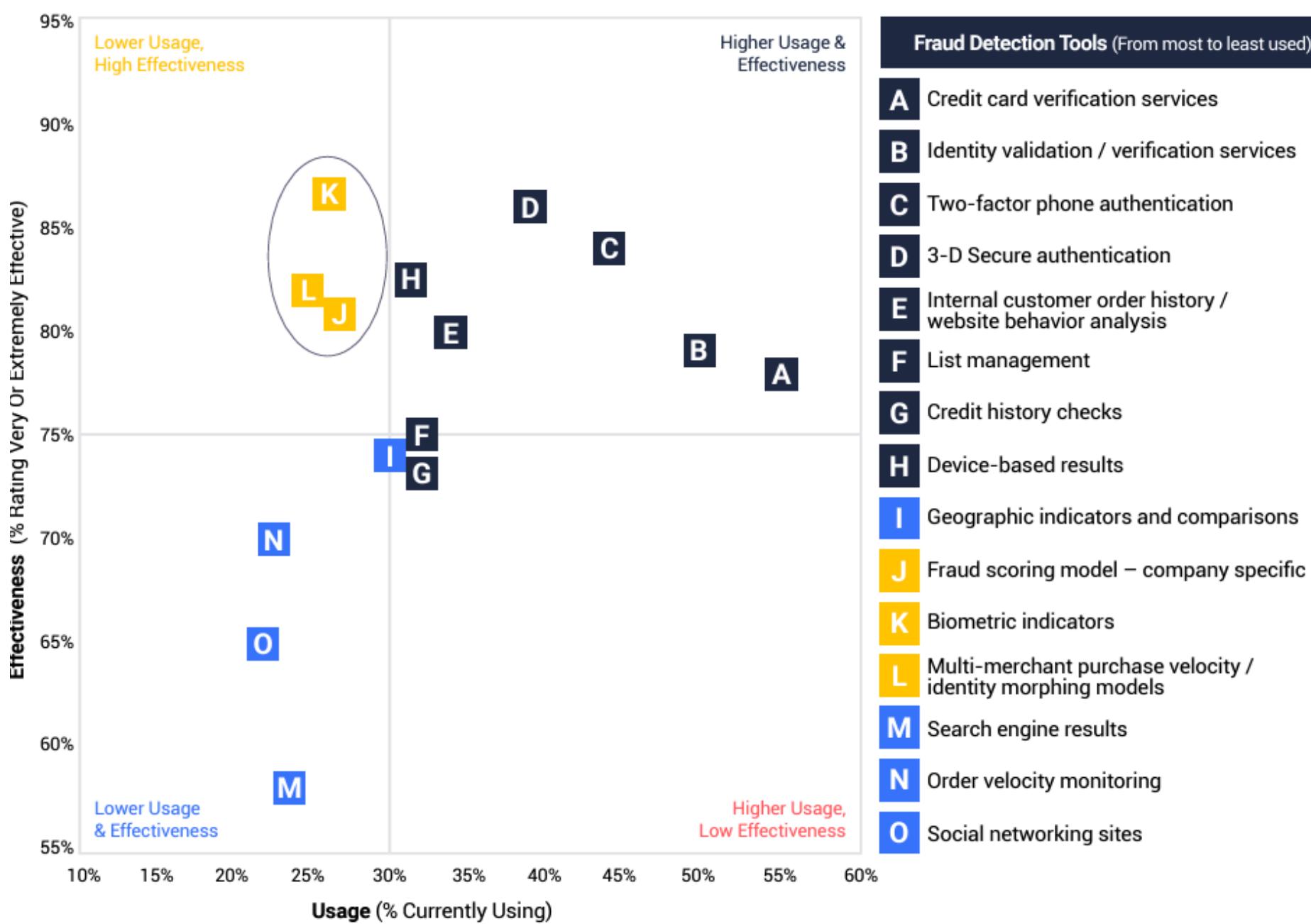
Data set original



Data set tratado y balanceado



Uso vs. Efectividad de las Herramientas de Detección de Fraudes



TIPOS DE FRAUDE EXPERIMENTADOS POR LOS COMERCIANTES -

Rankings de los últimos 3 años e Incidencia global (2023)

	2021 Rank	2022 Rank	2023 Rank	Global % Experiencing (2023)
Phishing / pharming / whaling	3	1	1	43% 
First-Party Misuse (i.e., friendly / chargeback fraud)	1	4	2 	34%
Card testing	2	2	3 	33%
Identity theft	4	3	4 	33%
Coupon / discount / refund abuse	5	7	5 	30%
Account takeover	7	5	6 	27%
Loyalty fraud	6	6	7 	22%
Affiliate fraud	8	8	8	22%
Re-shipping	12	11	9 	20% 
Botnets	10	9	10 	19%
Triangulation schemes	9	10	11 	17%
Money laundering	11	12	12	15%
AVG. # of attacks experienced	3	3	3	3

 Increased Ranking  Decreased Ranking  = Sig. Higher vs. 2022

Hay cuatro principales ataques de fraude, de más a menos prevalentes:
phishing / pharming / whaling, mal uso de primera parte (también conocido como "fraude amistoso"), prueba de tarjeta y robo de identidad .



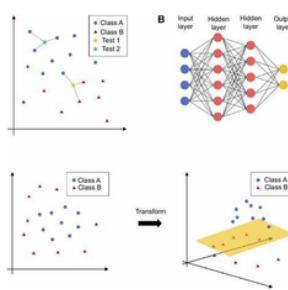
Consistente con los años anteriores del estudio, los comerciantes se vieron afectados por tres tipos diferentes de fraude, en promedio, con las pequeñas y medianas empresas alrededor de dos y las grandes empresas más propensas a enfrentarse con cuatro o más.



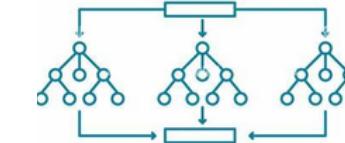
MACHINE LEARNING

Creación y evaluación de modelos

ML- CLASIFICACIÓN



KNN



RANDOM FOREST

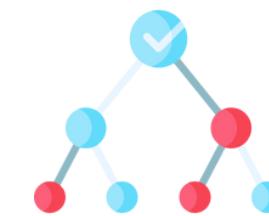
Probability B Will Happen Given Evidence A Has Already Happened $P(B|A)$
Probability A Will Happen Given Evidence B Has Already Happened $P(A|B)$
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability A Will Happen Given Evidence B Will Happen $P(A)$
Probability B Will Happen $P(B)$

NAIVE BAYES

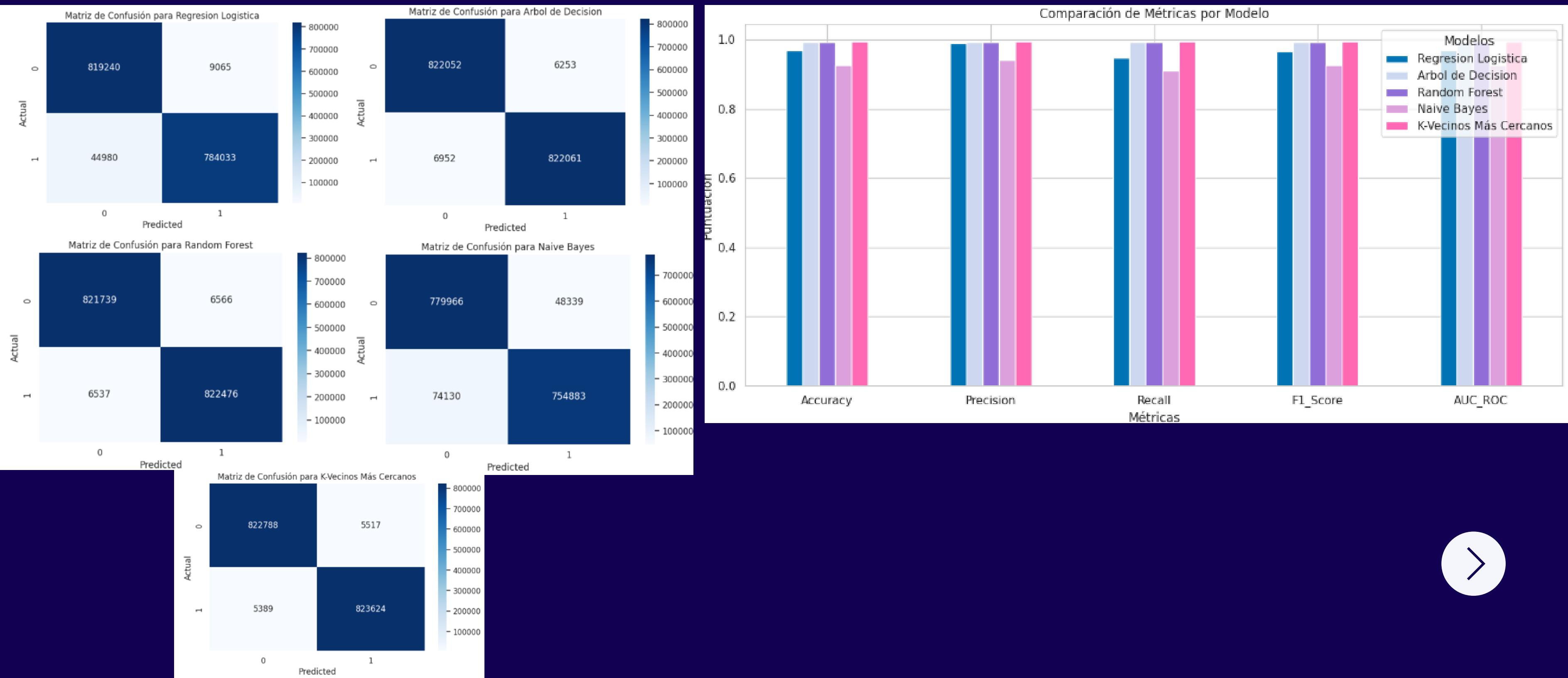


REGRESIÓN LOGÍSTICA



ÁRBOL DE DECISIÓN

Matriz de confusión / Métricas



Justificación de los Modelos Seleccionados

1. Modelo KNN (K-Nearest Neighbors)

- Eficiencia: KNN es rápido de implementar y puede proporcionar resultados precisos sin requerir complejas configuraciones. Esto significa que podemos obtener resultados rápidos y precisos sin invertir mucho tiempo en ajustes iniciales.
- Alta Precisión: El modelo KNN alcanzó una precisión del 99.34%, lo que lo hace altamente confiable para identificar fraudes. Esto es crucial para minimizar las pérdidas financieras y proteger los activos de la empresa.
- Adaptabilidad: No asume ninguna distribución específica de los datos, lo que le permite manejar diferentes tipos de transacciones y detectar patrones de fraude en diversas circunstancias.

Conclusión: KNN es adecuado para la detección de fraudes debido a su alta precisión y flexibilidad, ayudando a minimizar pérdidas financieras y proteger los activos de la empresa.

Justificación de los Modelos Seleccionados

2. Modelo Bernoulli de Naive Bayes

- Rapidez y Eficiencia: Este modelo es extremadamente rápido de entrenar y ejecutar, lo que permite una detección casi en tiempo real. Es ideal para sistemas que requieren respuestas inmediatas.
- Claridad de Resultados: Proporciona probabilidades claras de fraude, ayudando a priorizar las investigaciones de manera efectiva.
- Simplicidad: Con una precisión del 84.21%, es una excelente línea base y es muy útil para obtener resultados rápidos y comparativos con modelos más complejos.

Conclusión: Bernoulli Naive Bayes es una opción rápida y eficiente para la detección de fraudes, proporcionando resultados probabilísticos claros. Aunque tiene una precisión moderada, es útil para obtener respuestas rápidas y puede ser combinado con otros modelos para mejorar la efectividad.

Justificación de los Modelos Seleccionados

3. Modelo Árbol de Decisión (Decision Tree)

- Facilidad de Interpretación: Los árboles de decisión son fáciles de entender y visualizar, permitiendo a los equipos de negocio ver exactamente cómo se toman las decisiones de detección de fraude. Esto es esencial para la transparencia y la confianza en el sistema de detección.
- Alta Precisión: Con una precisión del 99.24%, este modelo es casi tan efectivo como KNN, lo que proporciona una segunda opción confiable con la ventaja añadida de ser fácilmente explicable.
- Flexibilidad: Puede manejar tanto datos categóricos como continuos, y es capaz de identificar relaciones complejas entre las características, mejorando la detección de patrones de fraude.

Conclusión: El Árbol de Decisión es altamente efectivo para la detección de fraudes debido a su alta precisión y facilidad de interpretación. Su transparencia en la toma de decisiones facilita la comprensión y confianza en el sistema, mejorando la toma de decisiones empresariales y ayudando a reducir las pérdidas financieras.



CONCLUSIONES

1. Eficiencia en la Detección de Fraudes:

- Modelos Altamente Precisos: Implementamos y evaluamos varios modelos de machine learning, destacándose el K-Nearest Neighbors (KNN) y el Árbol de Decisión por su alta precisión (99.34% y 99.24% respectivamente). Estos modelos demostraron una capacidad excelente para identificar transacciones fraudulentas, lo cual es esencial para proteger los activos de la empresa y reducir pérdidas.





CONCLUSIONES

2. Manejo Efectivo del Desequilibrio de Clases:

- Técnicas de Balanceo de Datos: La aplicación de técnicas como SMOTE fue fundamental para abordar el problema del desequilibrio en los datos (99% de transacciones legítimas frente a 1% de fraudulentas). Esto mejoró significativamente la capacidad del modelo para detectar fraudes, asegurando una protección más robusta.

3. Transparencia y Facilidad de Interpretación:

- Árbol de Decisión: Este modelo ofrece una clara visualización de cómo se toman las decisiones, lo que facilita la interpretación por parte de los equipos no técnicos. Esta transparencia es vital para generar confianza en el sistema de detección de fraudes.





CONCLUSIONES

4. Implementación Rápida y Eficiente:

- KNN y Naive Bayes: Estos modelos son rápidos de implementar y proporcionan resultados casi en tiempo real, permitiendo la detección inmediata de actividades fraudulentas. Esta capacidad de respuesta es crucial para prevenir fraudes de manera proactiva.

5. Impacto en la Seguridad Financiera:

- Reducción de Pérdidas Financieras: La implementación de estos modelos puede reducir significativamente las pérdidas financieras debidas a fraudes, mejorar la seguridad de las transacciones bancarias y aumentar la confianza de los clientes en el sistema financiero móvil.





CONCLUSION FINAL

La implementación de modelos de machine learning como KNN y Árbol de Decisión ha demostrado ser una solución altamente efectiva para la detección de fraudes en transacciones móviles de dinero. Estos modelos no solo proporcionan una alta precisión y rapidez en la detección, sino que también ofrecen transparencia y facilidad de interpretación, lo cual es esencial para generar confianza en el sistema. Al seguir las recomendaciones propuestas, la empresa puede establecer un estándar de seguridad en el sector financiero móvil, protegiendo sus activos y mejorando la confianza de los clientes.





RECOMENDACIONES

1. Despliegue en Producción:

- Utilización de Google Cloud Platform (GCP): Recomendamos desplegar los modelos en GCP para aprovechar su infraestructura escalable y capacidades de predicción en tiempo real. Esto permitirá una integración fluida con los sistemas existentes y una detección inmediata de fraudes.

2. Monitoreo y Actualización Continua:

- **Sistema de Monitoreo Continuo:** Establecer un sistema para monitorear el rendimiento del modelo en producción y realizar ajustes periódicos. Esto incluye la actualización del modelo con datos nuevos para mantener su efectividad frente a las nuevas tácticas de fraude.





RECOMENDACIONES

- 3. Combinación de Modelos para Mayor Robustez:
- Enfoque Ensamblado: Considerar la combinación de predicciones de KNN, Naive Bayes y Árbol de Decisión para mejorar la precisión y robustez de la detección de fraudes. Esto puede proporcionar una capa adicional de seguridad.

- 4. Gestión Adecuada de Datos:
- **Políticas de Privacidad y Seguridad:** Implementar políticas robustas para la gestión de datos, asegurando la calidad y actualización regular de los datos, así como la protección de la información sensible de los clientes.
-



NUESTRO EQUIPO DATA-DEFEND

DATA SCIENTIST



Christian
Fernandez



[Github](#)



[Linkedin](#)

DATA SCIENTIST



Giovanni
Roncancio



[Github](#)



[Linkedin](#)

DATA SCIENTIST



Leopoldo
Flores



[Github](#)



[Linkedin](#)

**"El primero y peor de todos los
fraudes es engañarse a si mismo."**

Philip James Bailey. Poeta Inglés

