

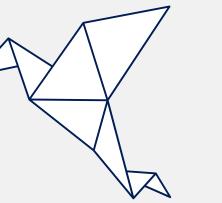


Fraude en e-commerce

Proyecto: Paula Mariana Gilio

CoderHouse

índice



Objetivo y audiencia



Hipótesis y preguntas planteadas



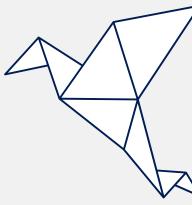
Análisis exploratorio y visualizaciones



Algoritmos y mejoras



Objetivo principal

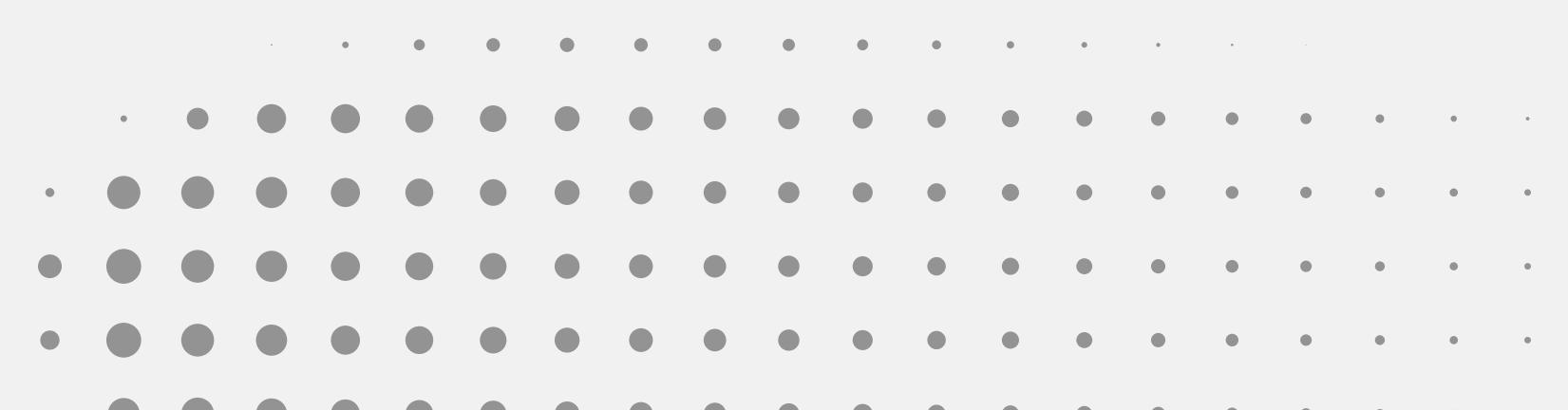


El objetivo es el de desarrollar un modelo predictivo robusto y eficiente que permita detectar y prevenir transacciones fraudulentas en plataformas de comercio electrónico.

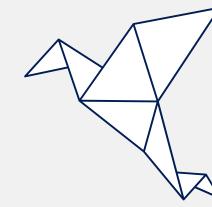
El proyecto tiene como meta reducir el número de fraudes mediante la identificación temprana de patrones y comportamientos anómalos en las transacciones.

Al implementar este sistema, se busca proteger tanto a las empresas como a los consumidores de pérdidas financieras, mejorar la seguridad de las plataformas, y mantener la confianza en el comercio digital.

Tras estos objetivos indiscutibles, se necesita determinar: si se podría mejorar la seguridad, optimizar la gestión de recursos y prevenir futuros ataques.



Audien cia



Este análisis intentará responder y entender los diferentes comportamientos maliciosos para servir, posteriormente, a la toma de decisiones.

Por lo qué, esta dirigido a aquellas pequeñas o grandes empresas que operen con plataformas de comercio electrónico, entidades que procesen pagos y gestionen cuentas de clientes y pretendan minimizar el riesgo de fraude.



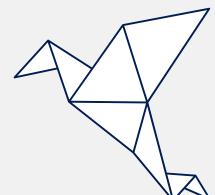
Hipótesis y preguntas planteadas



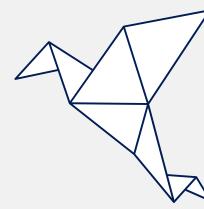
Con el fin de implementar estrategias proactivas, se planteó una serie de preguntas que pudieran ser respondidas mediante gráficos de rápida visualización que puedan dar contexto y mejor entendimiento.



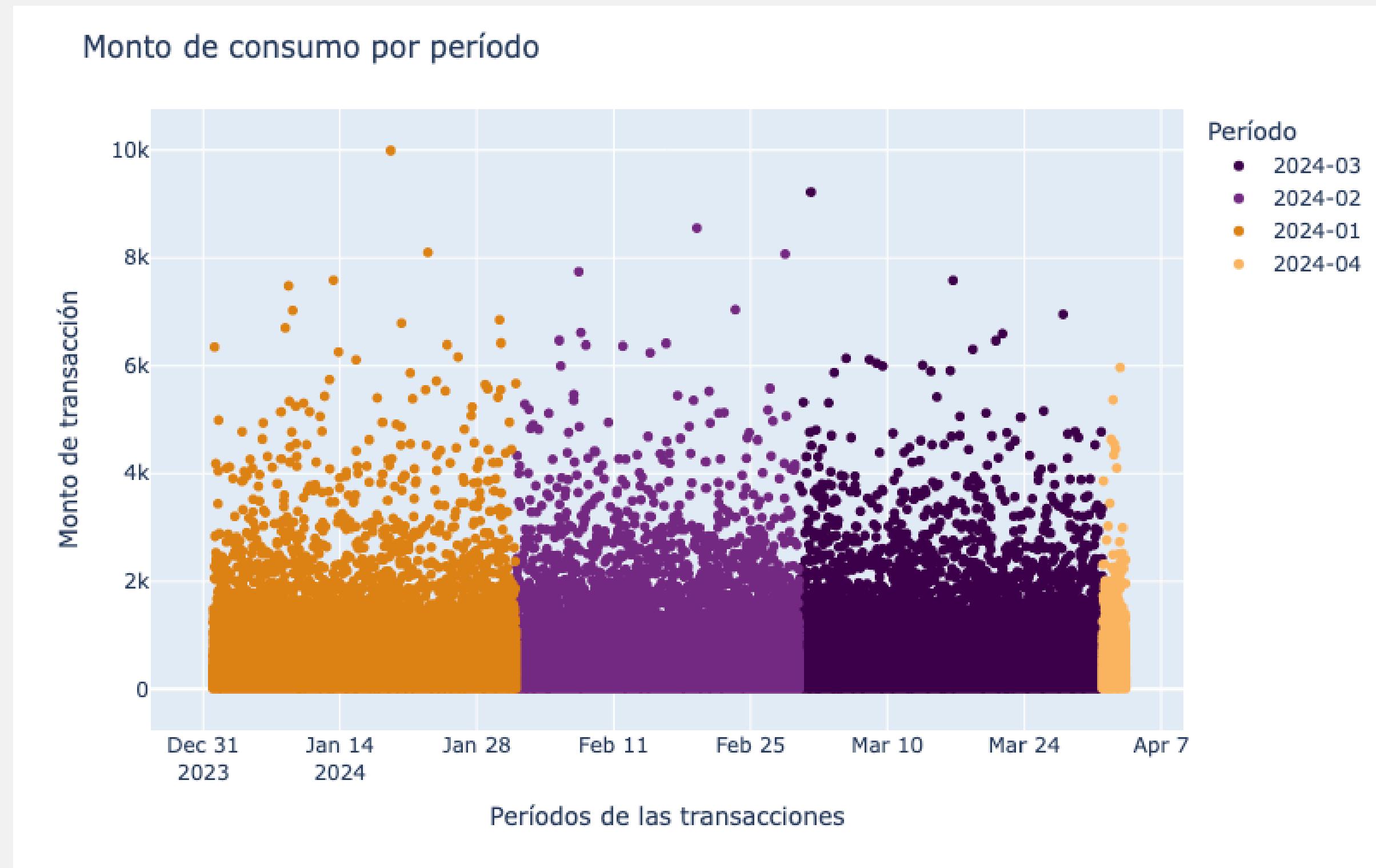
HIPÓTESIS: El modelo será capaz de predecir la ocurrencia de ataques con una precisión mayor o igual al 90%.



- A.** ¿Existen compras irregulares según los diferentes períodos del año?
- B.** ¿Cierto rango etario es más propenso al consumo?
- C.** ¿Hay un dispositivo elegido para realizarlas?, ¿La edad puede influir en esto?
- D.** ¿Existen transacciones de montos muy elevados con respecto a los días de apertura de una cuenta?
- E.** ¿Las transacciones realizadas en horas no laborables podrían tener una mayor probabilidad de ser fraudulentas?
- F.** ¿En qué medida existe mayor o menor fraude según el medio de pago utilizado?
- G.** ¿Es probable la apertura de una cuenta para cometer fraude?
- H.** ¿Existe alguna categoría de producto más consumida? ¿Es probable que alguna de ellas proporcione en mayor cantidad fraude?



A. ¿Existen compras irregulares según los diferentes períodos del año?

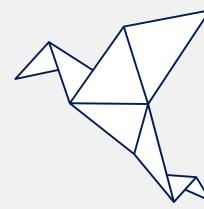


En todos los períodos se observan varias transacciones realizadas dentro de los dos mil dólares, los cuales parecerían ser valores normales.

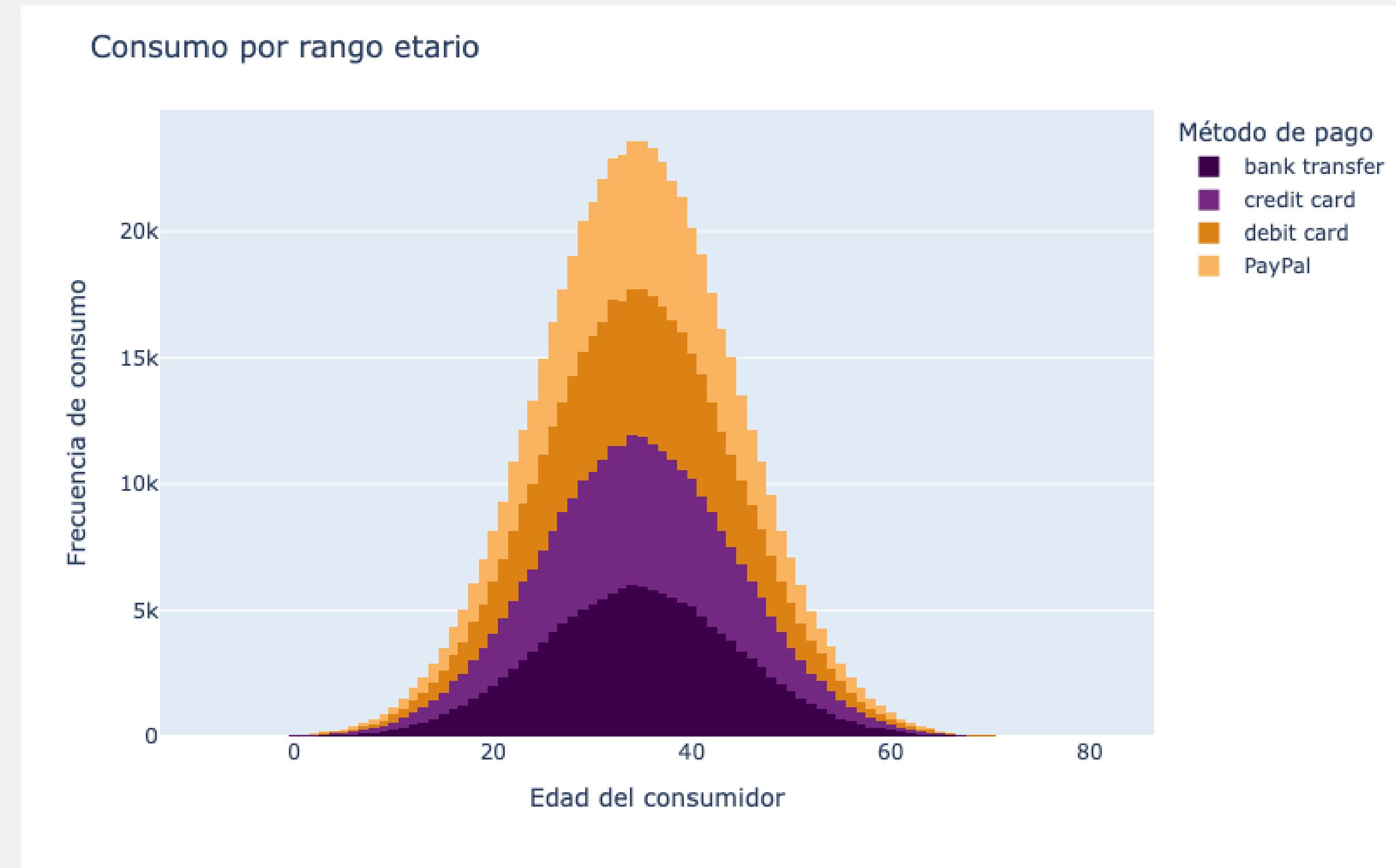
Existen varios valores por sobre los cuatro mil dólares distribuidos en todos los períodos, siendo enero uno de los meses con transacciones de consumos atípicos y más altos. Por otro lado se observan varios consumos entre los seis mil y ocho mil, identificando este mes con más actividad de consumo.

Los meses de febrero y marzo están equitativos sus consumos, incluyendo transacciones altas de entre seis mil a ocho mil dólares y unas muy pocas mayores a los ocho mil dólares.

El período de abril es el menos representativo ya que sus datos abarcan una semana únicamente.



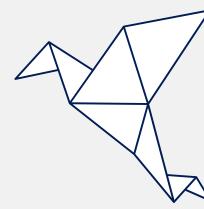
B. ¿Ciertos rango etario es más propenso al consumo?



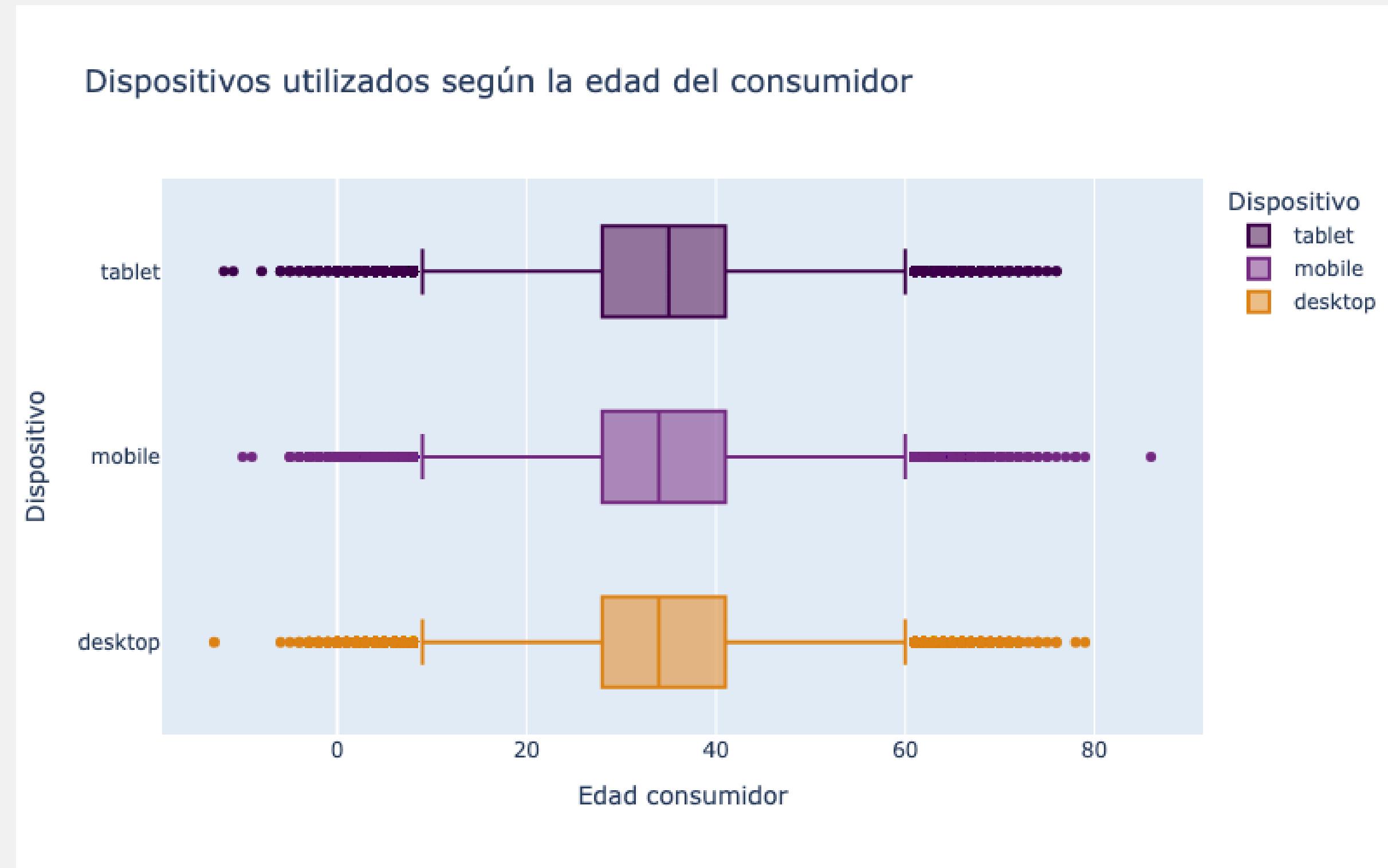
Se puede observar en el gráfico la asimetría que representa el consumo con respecto a la edad de los consumidores, determinando que la edad promedio de consumo son entre los 33 a 36 años.

Se observa que a partir de los 60 años disminuye notablemente el consumo, quedando pocos casos de consumidores que realizan transacciones, los cuales son aproximadamente entre dos mil y tres mil.

Respecto a los medios de pago utilizados, puede observarse que son equitativos, tanto en métodos como en edades de los consumidores, liderando PayPal el medio más representativo.



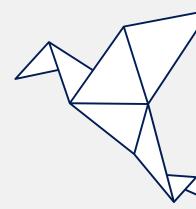
C. ¿Hay un dispositivo elegido para realizarlas?, ¿La edad puede influir en esto?



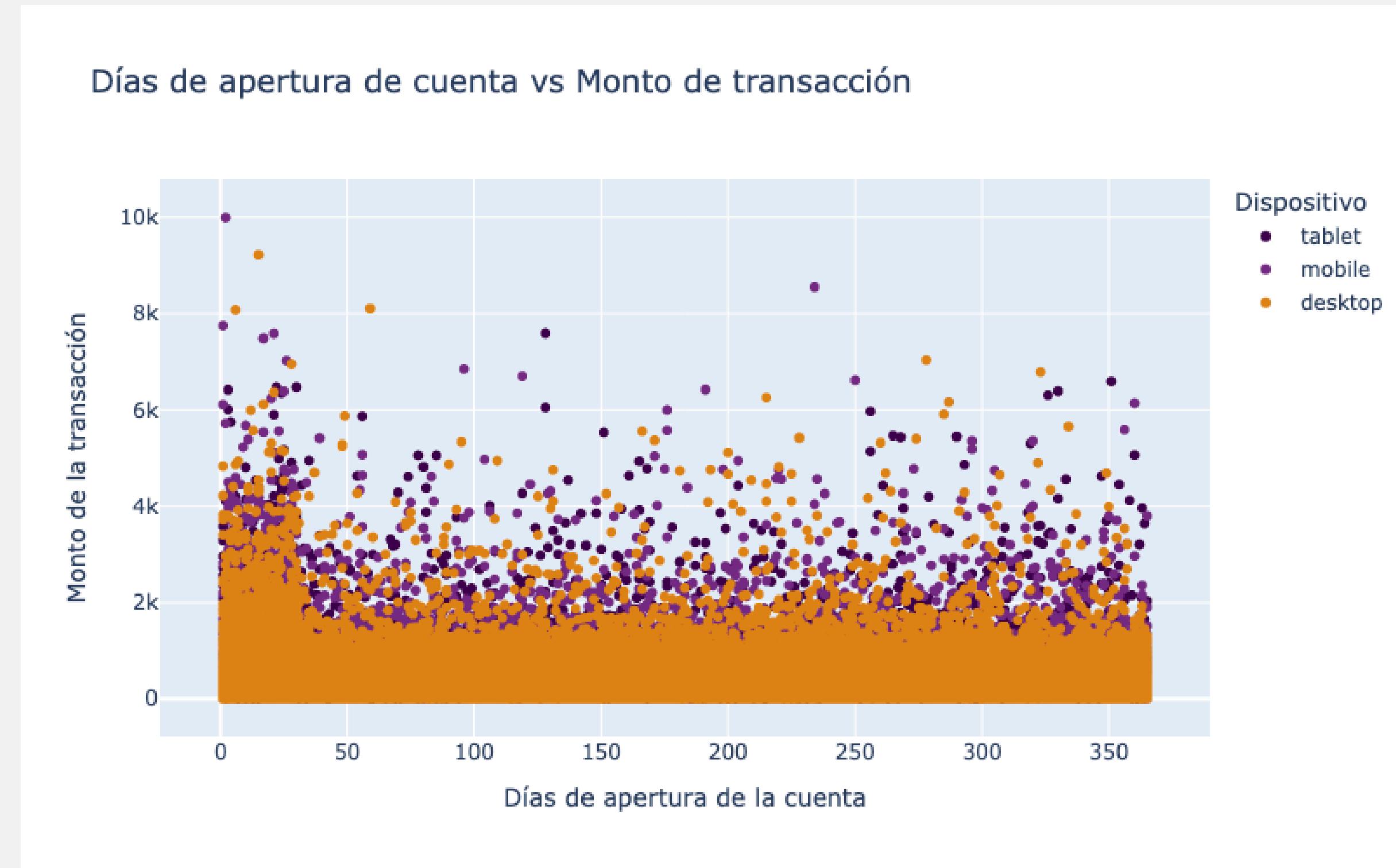
El gráfico muestra que el equipo electrónico utilizado para realizar las transferencias, es indistinto y representan números muy similares. Así mismo que la distribución entre cuartiles es simétrica sin importar el equipo.

Por otro lado, la media de mobile y desktop es de 34 años y las tablets es de 35 años. También se identificó un valor irregular en la utilización del dispositivo móvil de un usuario de 86 años, el cual podría o no ser fraudulento.

La evidencia de outliers en edades menores a 0 años hasta 17 años, posteriormente será tratada ya que, no es lógico. El rango etario mayor a 60, si bien aparece como valor atípico, podría o no indicar fraude.



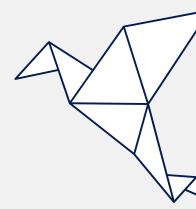
D. ¿Existen transacciones de montos muy elevados con respecto a los días de apertura de una cuenta?



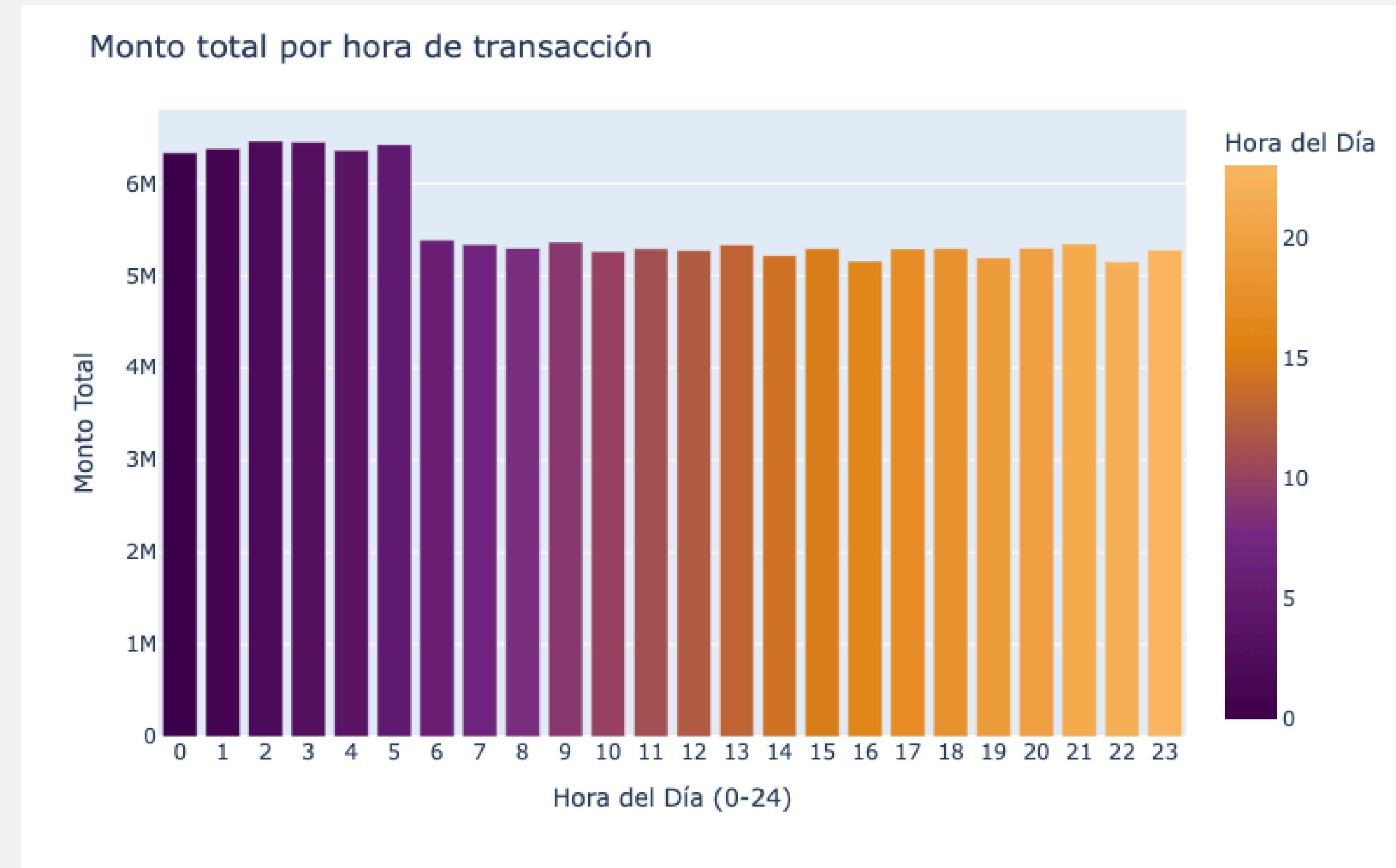
Se observa en el gráfico que, las transacciones realizadas, contienen valores de 1500 dólares las cuales son considerados normales.

Se observa que entre el período 0 y 30 días de apertura de la cuenta, hay varios valores con transacciones muy elevadas, las cuales superan los 2500 dólares y llegan hasta los cinco mil dólares, existen monto más elevados los cuales no superan los **15 días desde el día de apertura**.

Entre los 50 y los 365 días las transacciones más frecuentes se encuentran entre los 2500 dólares. También se representan consumos por encima de los 2500 dólares, pero se reducen notablemente en cantidad.



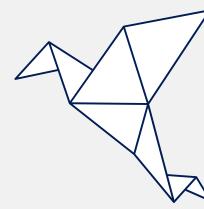
E. ¿Las transacciones realizadas en horas no laborables podrían tener una mayor probabilidad de ser fraudulentas?



Existe una pequeña varianza en los montos de transacciones realizadas por horas nocturnas, donde los montos más altos llegan a los 6500 millones de dólares, en el horario de las 0 am a las 5am.

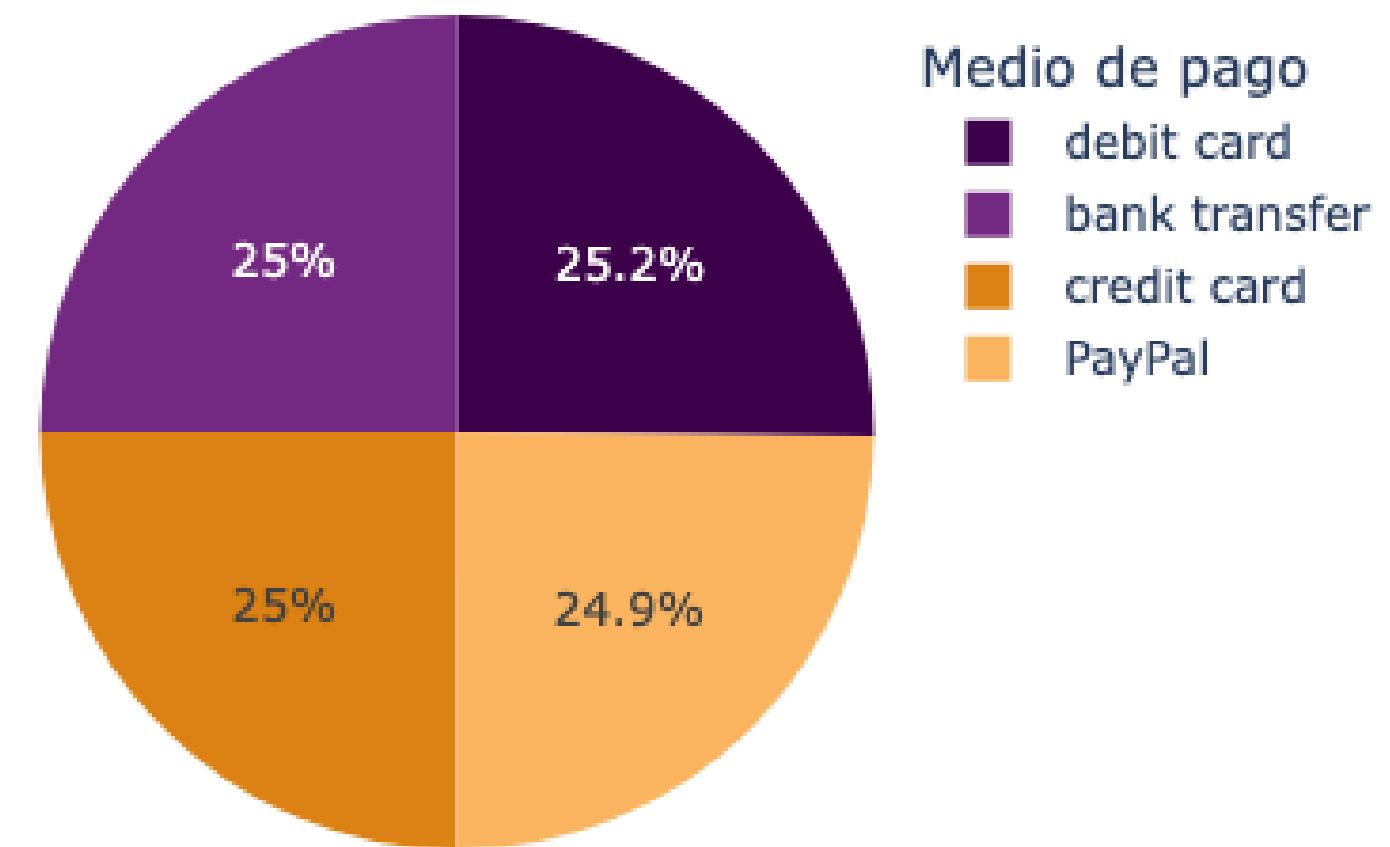
A partir de las 6 am y hasta las 23pm las transacciones fluctúan entre los 6 y 5 millones.

Lo cual podría ser un indicativo que en las horas más representativas del consumo, posiblemente fraudulento, podrían ser las horas de la madrugada.
Y no existe algún indicativo que por fuera del horario laboral exista mayores transacciones, las cuales impliquen fraude.



F. ¿En qué medida existe mayor o menor fraude según el medio de pago utilizado?

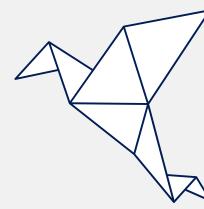
Fraude según método de pago



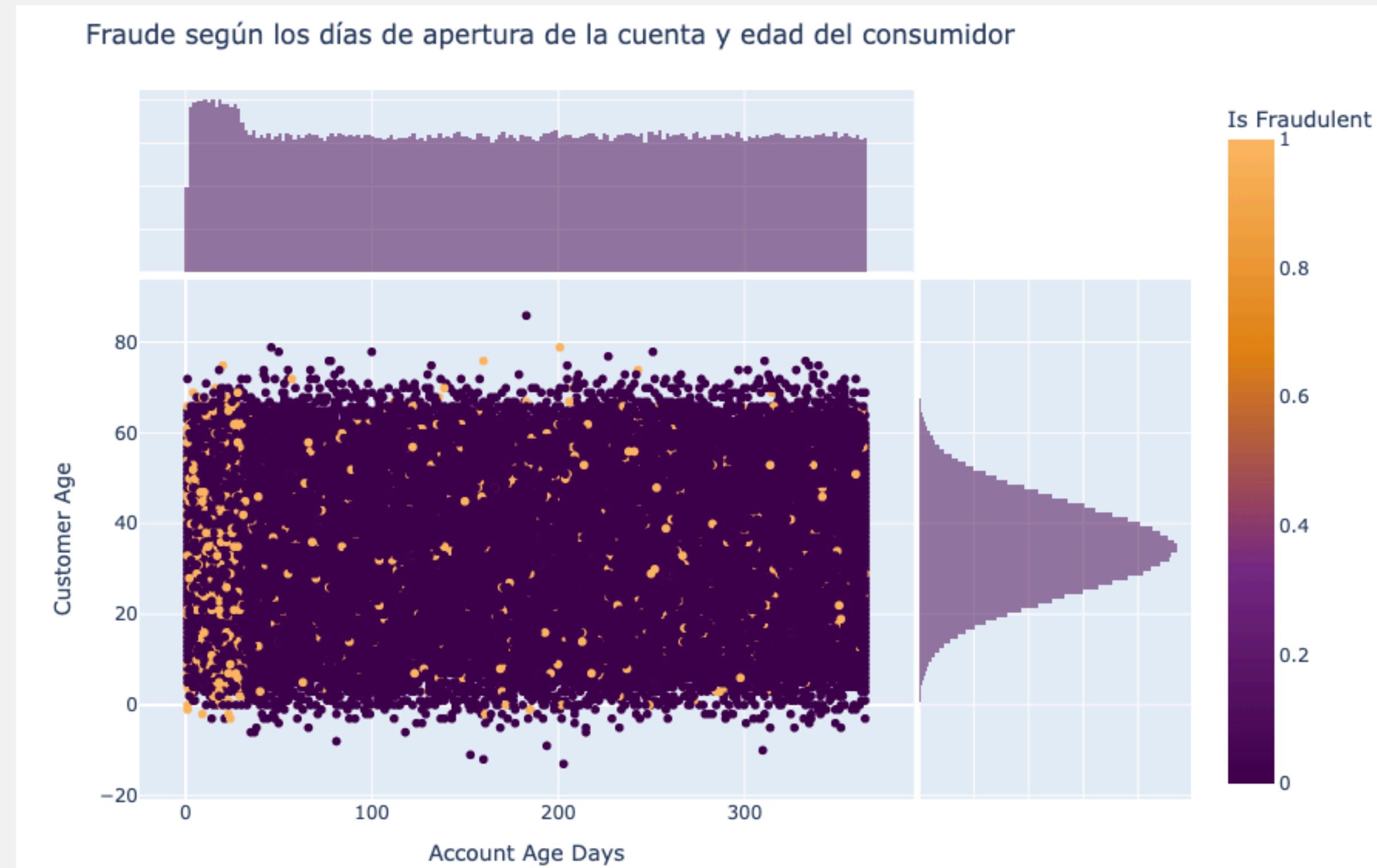
Se puede observar en el gráfico que los medios utilizados para realizar las transacciones son perfectamente equitativos.

Aún así podemos determinar que el recurso más utilizado es:

- Transferencia bancaria, de los cuales 7377 transacciones son fraudulentas.
- Tarjeta de débito, con 7418 transacciones fraudulentas.
- Tarjeta de crédito y PayPal, en último lugar con 7330 y 7364 transacciones fraudulentas.

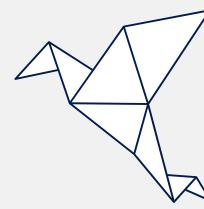


G. ¿Es probable la apertura de una cuenta para cometer fraude?

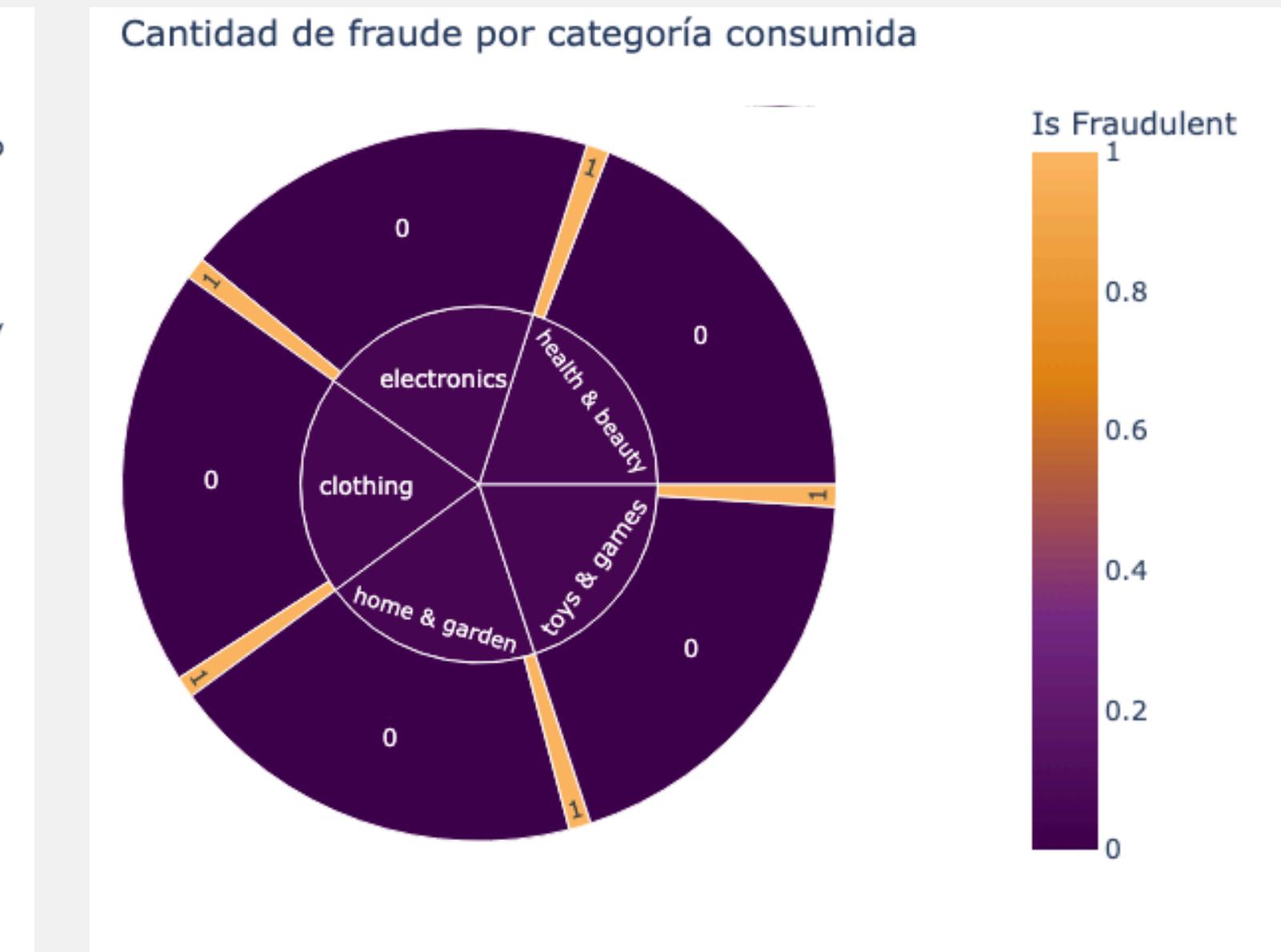
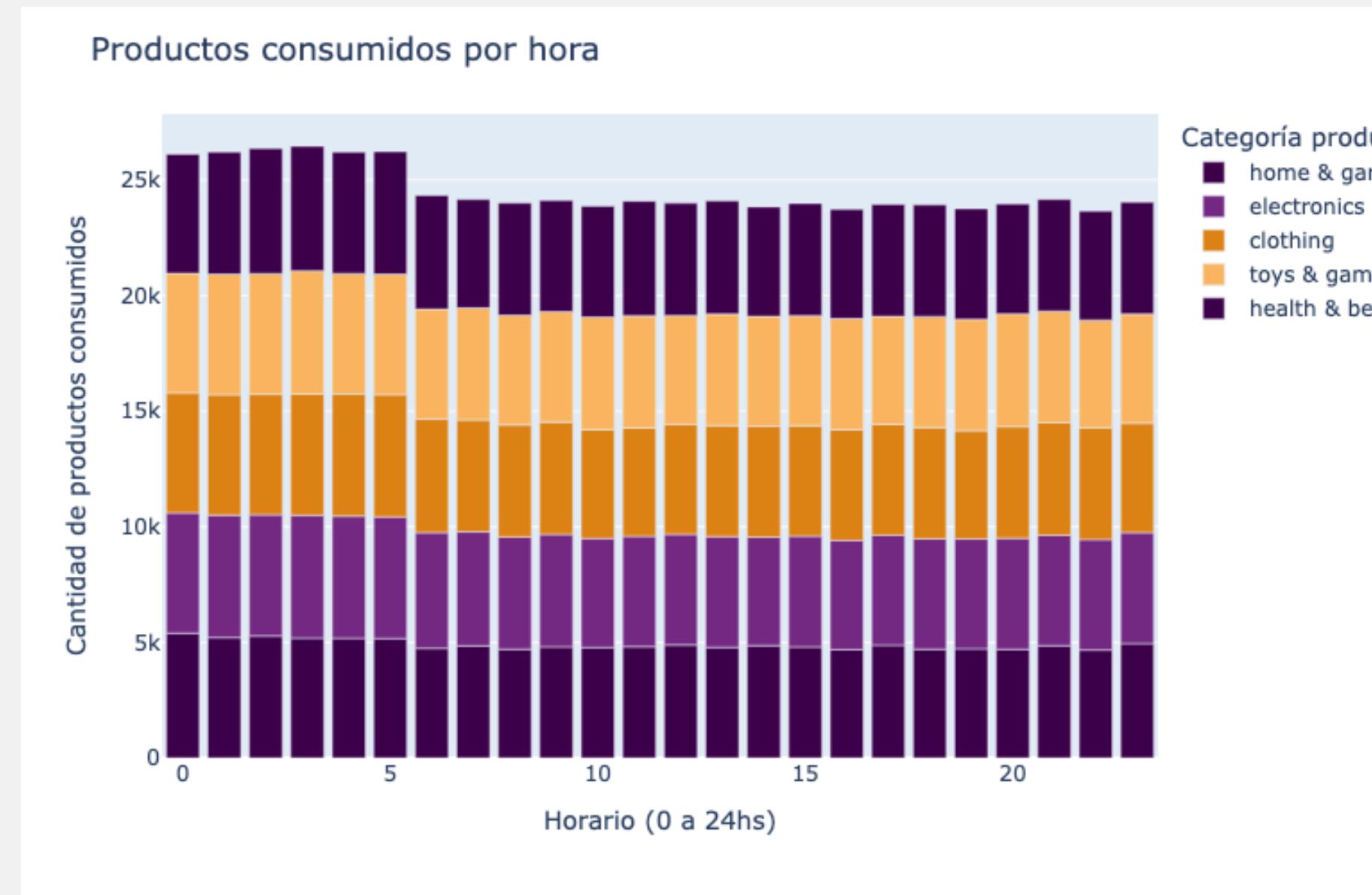


Se aprecian en tono naranja los usuarios considerados fraude y podemos destacar que se encuentran, en gran parte, los primeros días de apertura de una cuenta, aproximadamente hasta los 40 días y que son en edades variadas.

Posteriormente encontramos fraude, en el transcurso de los días de apertura de las cuentas, en menor cantidad y distribución. Pero alguno de ellos representados en usuarios menores a los 20 años y mayores a 70 años.

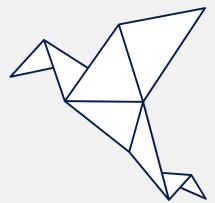


H. ¿Existe alguna categoría de producto más consumida? , ¿Es probable que alguna de ellas proporcione en mayor cantidad fraude?



El gráfico nos representa pequeñas variaciones de las categorías consumidas en los diferentes momentos del día, siendo así los horarios de la madrugada en los que el consumo es mayor.

También podemos ver el que el pico máximo de consumo es a las 3 am. Y tener en cuenta que **no existe diferencias significativas** entre las diferentes categorías.



Algoritmos y mejoras

En este proyecto se analizará un conjunto de datos, el cual cuenta con 1.472.952 registros, sobre la detección de fraude en el comercio electrónico.

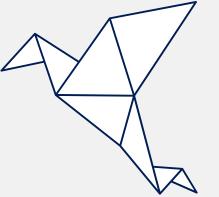
Incluye un conjunto de transacciones que se han etiquetado como fraudulentas o legítimas, permitiendo analizar patrones y características comunes en las transacciones fraudulentas.

Para posteriormente facilitar la investigación y desarrollo de modelos de detección de fraude. El dataset contiene múltiples variables relevantes como, *la edad del consumidor, información sobre el método de pago y el tipo de dispositivo utilizado, la ubicación geográfica, el horario de la transacción, y la cantidad y segmento del producto.*

En este proyecto se utilizaran técnicas de machine learning de aprendizaje supervisado, con el objetivo de entrenar modelos que puedan identificar y prevenir actividades fraudulentas.



Observaciones



Se implementaron diferentes algoritmos de clasificación para la obtención de las diferentes métricas necesarias para el conjunto de datos.

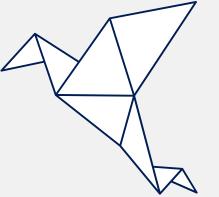
Posteriormente, se realizarán una serie de múltiples ajustes para garantizar la mejora.

Al calcular métricas se detectaron falsos positivos. Lo cual significa que nivel empresarial es primordial reducirlos, ya que el modelo predice que hay usuarios que no cometieron fraude pero realmente si lo hicieron.

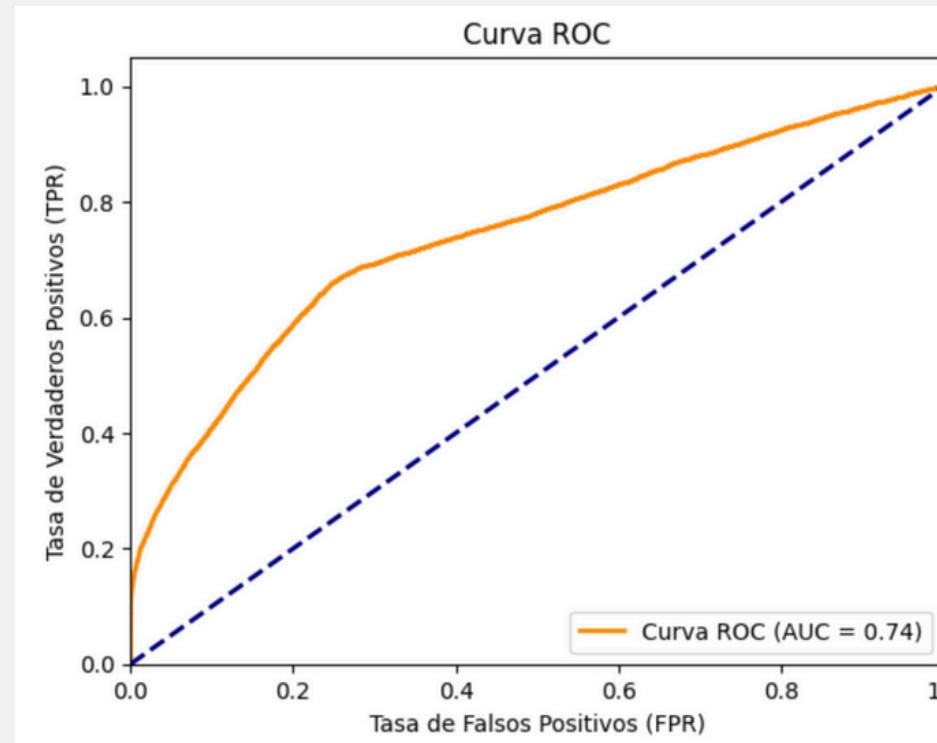
| | Score | Accuracy | Precision | Recall | F1 Score | \ |
|------------------------|----------|----------|-----------|----------|----------|---|
| Logistic Regression | 0.954441 | 0.955014 | 0.915052 | 0.104814 | 0.582475 | |
| Random Forest | 0.875677 | 0.876295 | 0.218251 | 0.576477 | 0.624310 | |
| KNeighborsClassifier | 0.957176 | 0.951747 | 0.553350 | 0.152270 | 0.606953 | |
| DecisionTreeClassifier | 0.955043 | 0.955506 | 0.819979 | 0.134517 | 0.604105 | |
| Gradiant Boosting | 0.954954 | 0.955345 | 0.841743 | 0.125299 | 0.597572 | |
| XGBoost | 0.955158 | 0.955404 | 0.823151 | 0.131103 | 0.601611 | |
| | AUC | | | | | |
| Logistic Regression | 0.741067 | | | | | |
| Random Forest | 0.770078 | | | | | |
| KNeighborsClassifier | 0.667873 | | | | | |
| DecisionTreeClassifier | 0.753618 | | | | | |
| Gradiant Boosting | 0.770124 | | | | | |
| XGBoost | 0.770513 | | | | | |



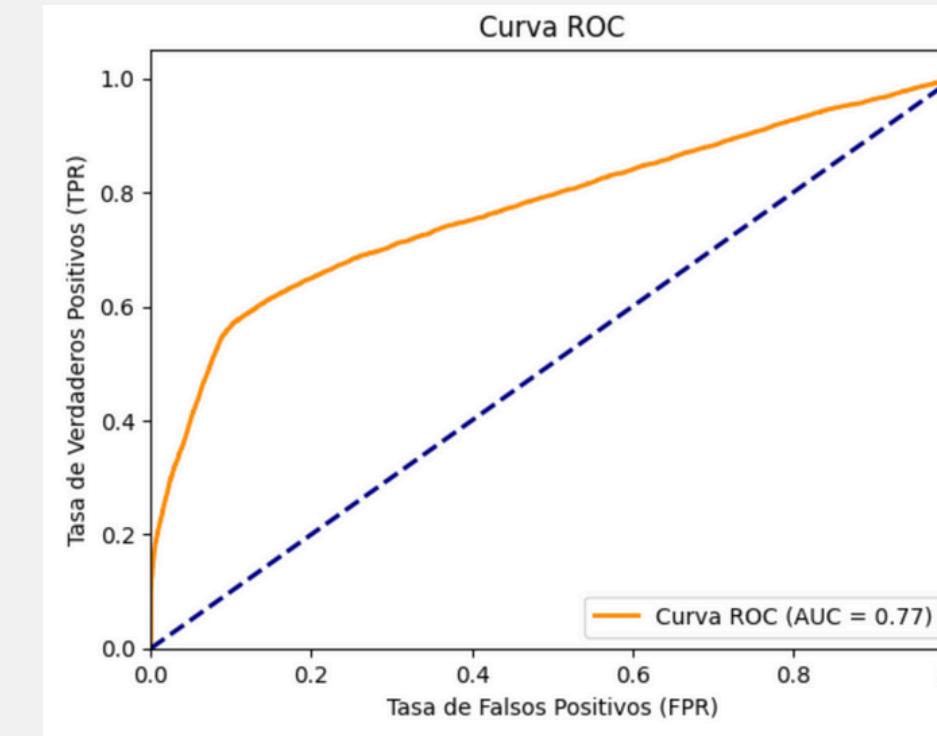
Curva ROC y AUC



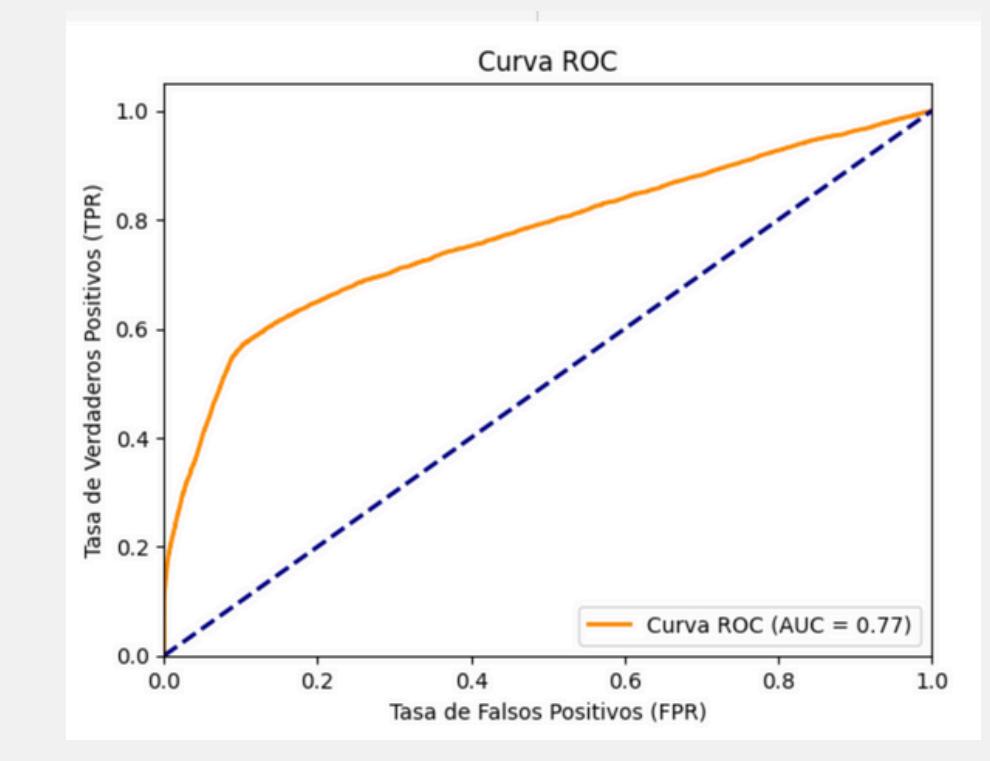
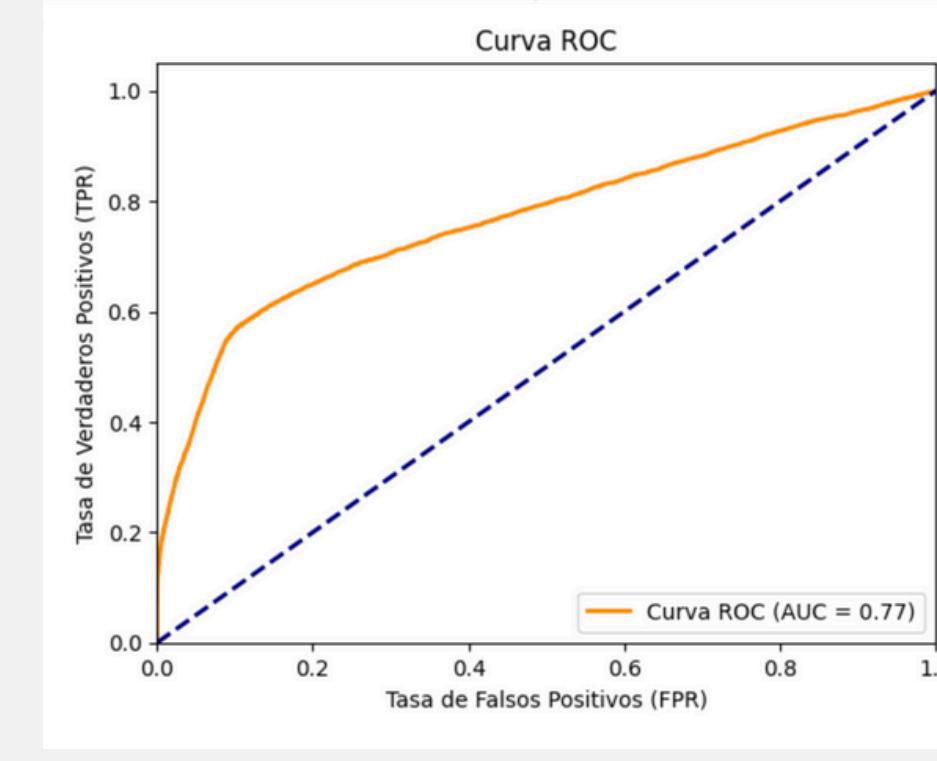
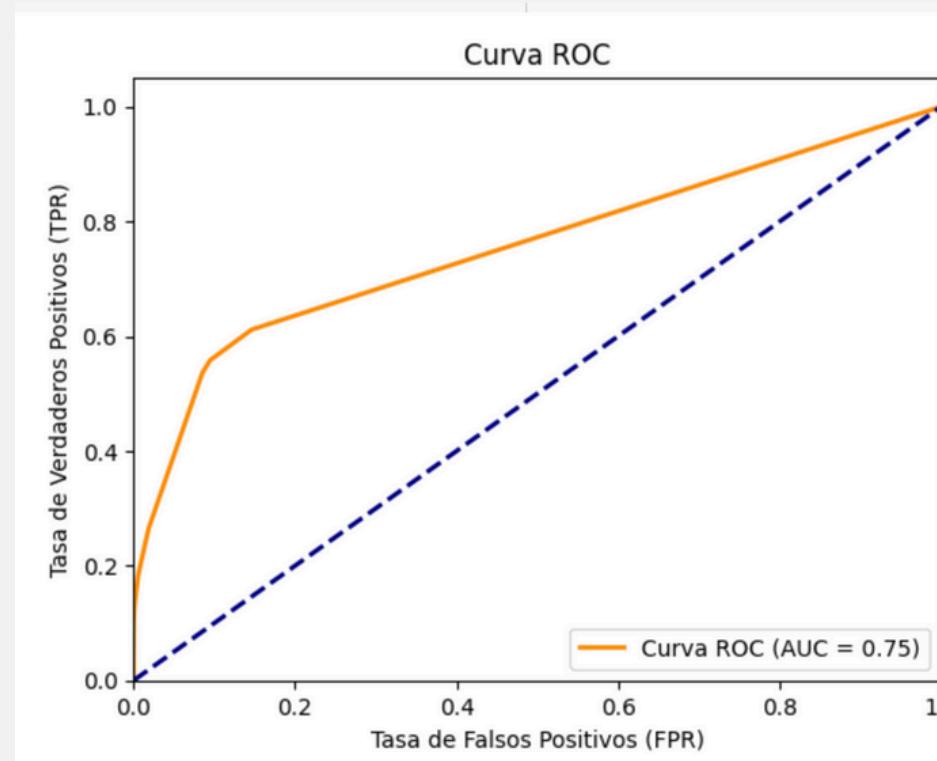
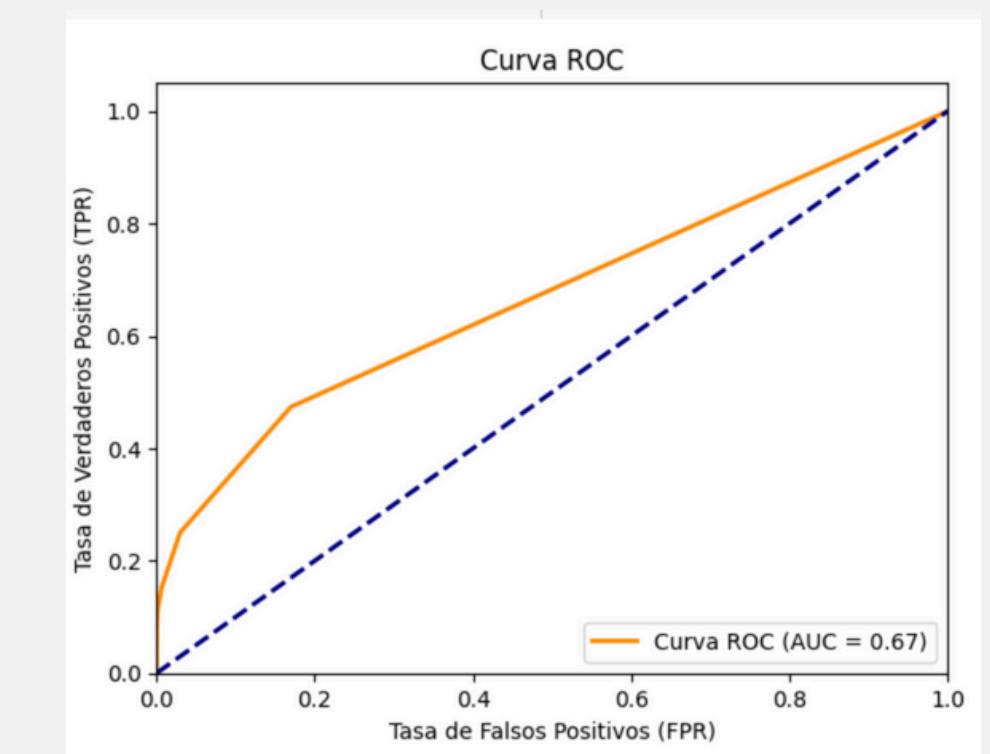
Regresión logística



Random Forest



KNeighbors



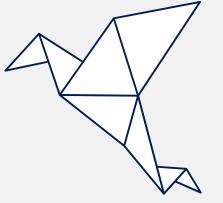
Árbol de decisión

XG Boosting

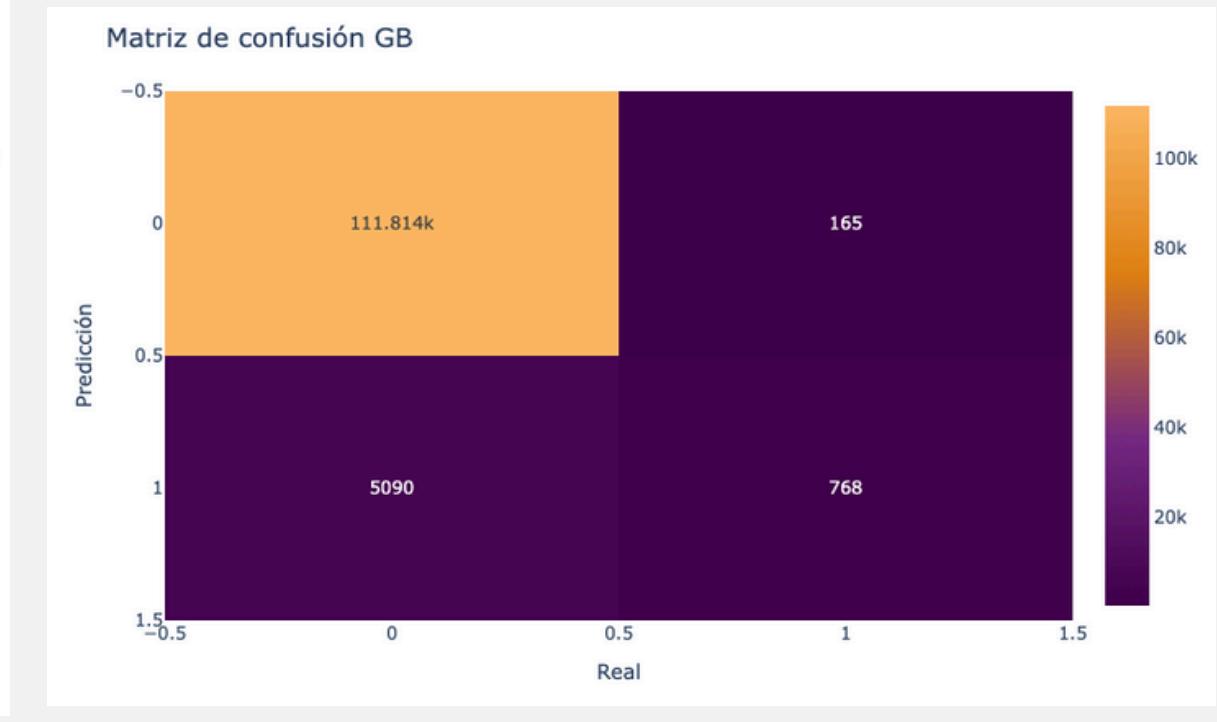
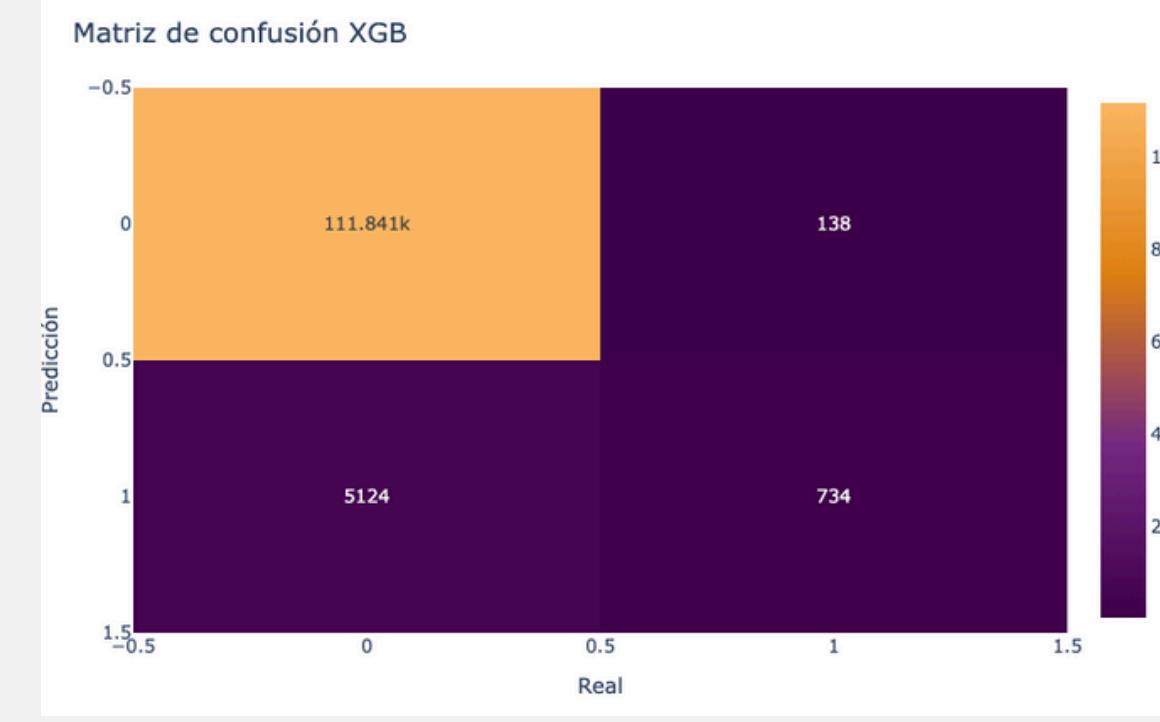
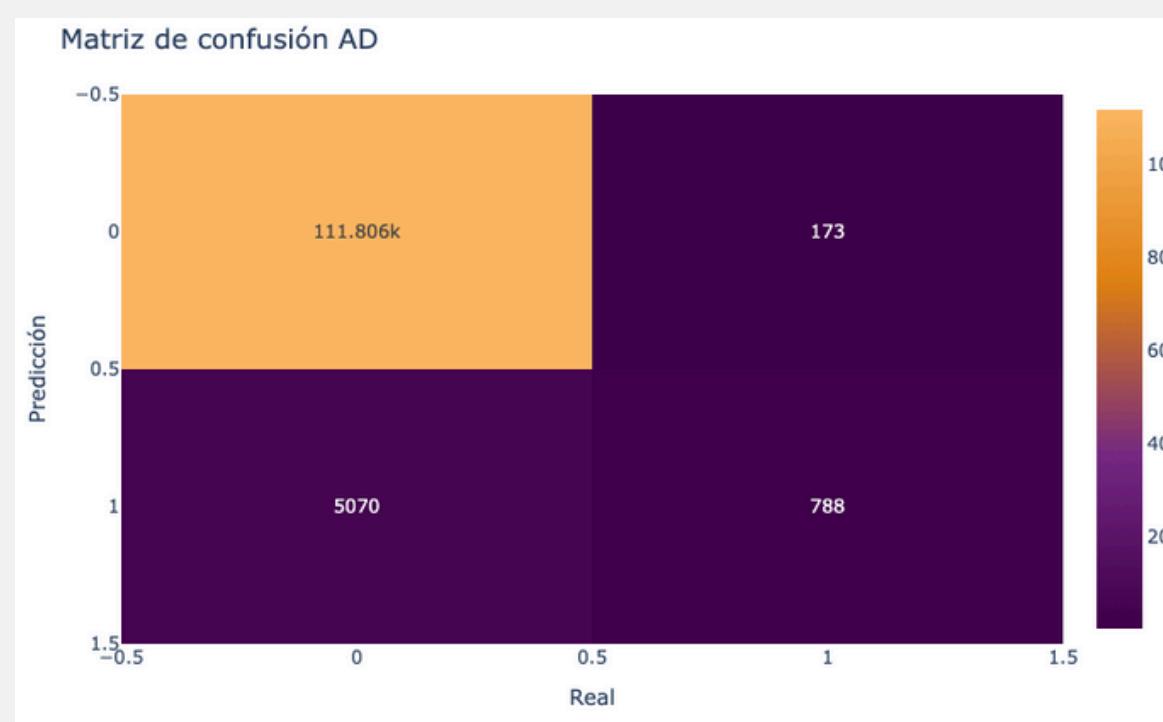
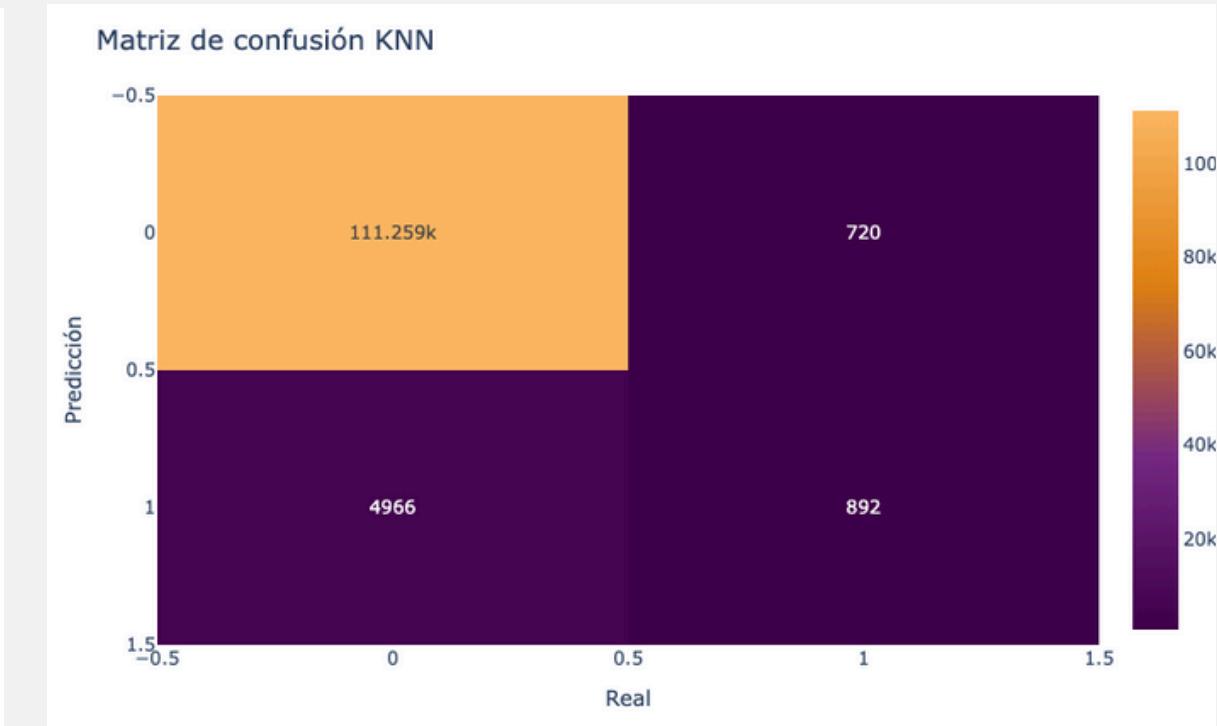
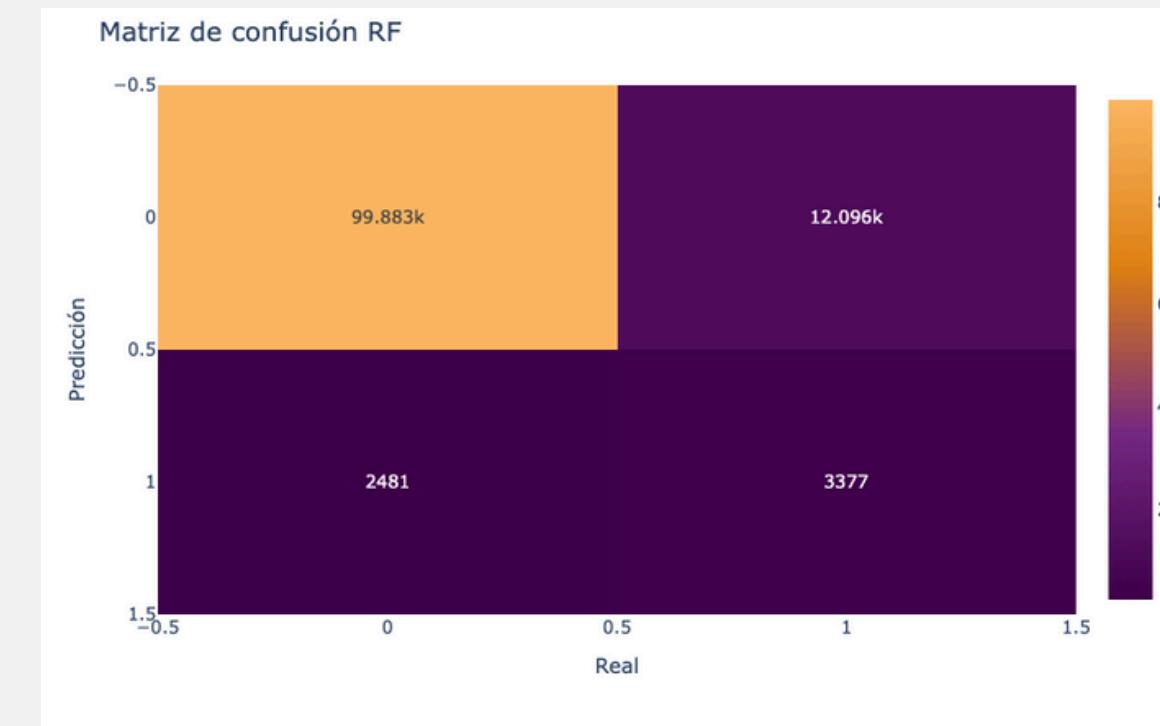
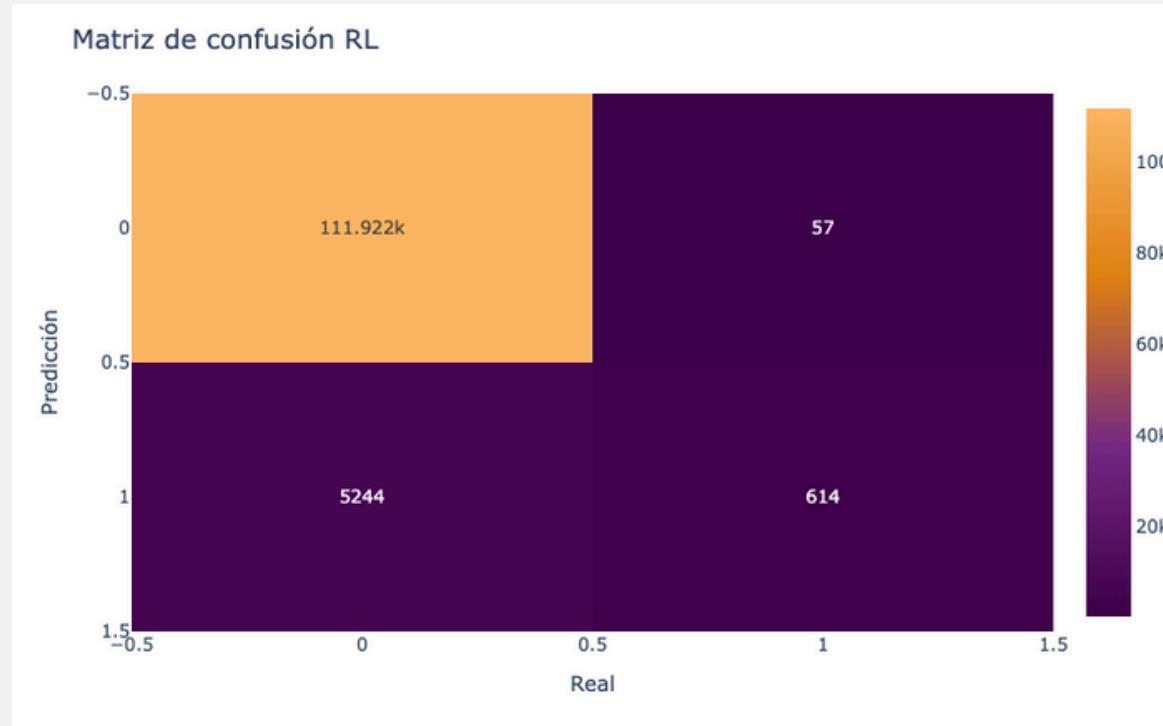
Gradient Boosting



Matriz de confusión



Regresión logística



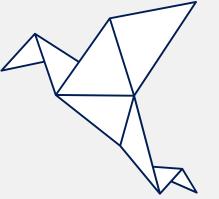
Árbol de decisión

XG Boosting

Gradient Boosting



Stratified K-Fold



La estrategia principal era reducir el F1 Score, ya que el conjunto de datos se encontraba compartiendo datos con una clase minoritaria y otras mayoritarias.

El F1 Score es la media armónica entre la precisión y la sensibilidad.

Representa un balance entre ambos, pudiendo visualizar como captura el modelo a la clase positiva, que es sensible a la clase minoritaria, mientras evita los falsos positivos. Esto lo convierte en una métrica especialmente valiosa ya que precisión y sensibilidad son importantes.

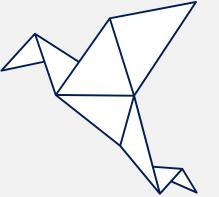
En el gráfico se representa las métricas obtenidas de los diferentes modelos a través del método Stratified K-fold

```
F1 Score promedio (Stratified K-Fold) for Logistic Regression: 0.19
F1 Score promedio (Stratified K-Fold) for Random Forest: 0.32
F1 Score promedio (Stratified K-Fold) for KNeighborsClassifier: 0.24
F1 Score promedio (Stratified K-Fold) for DecisionTreeClassifier: 0.23
F1 Score promedio (Stratified K-Fold) for Gradiant Boosting: 0.21
F1 Score promedio (Stratified K-Fold) for XGBoost: 0.23
```

Parecería ser que el modelo con mejor performance es el Random Forest. Las métricas de los demás modelos que rondan los 0,20 y 0,24 son aceptables.



K-Fold y Grid Search



K-fold provee una estimación más robusta del rendimiento del modelo, en modelos basados en árboles ayuda a reducir la sobreoptimización, o por ejemplo en algoritmos de KNN los datos pueden variar mucho dependiendo de la distribución de los datos, por lo que K-Fold ayuda a obtener una evaluación estable.

Por otro lado Grid Search se basa en probar de forma exhaustiva todas las combinaciones posibles de hiperparámetros, que mejor balancea precisión y generalización.

K-Fold CV – Accuracy promedio for XGBoost: 0.9550

K-Fold CV – Desviación estándar for XGBoost: 0.0005

CPU times: user 11min 50s, sys: 4.63 s, total: 11min 55s

Wall time: 11min 42s

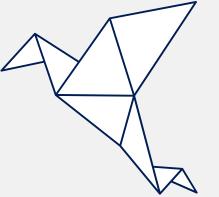
```
[107] # mejores hiperparámetros GridSearchCV
      print(f'Mejores hiperparámetros for {nombre}: {grid_search.best_params_}')
→ Mejores hiperparámetros for XGBoost: {'max_depth': 3, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 150}
```

```
[108] # mejor puntuación obtenida
      print(f'Mejores F1-score obtenido for {nombre}: {grid_search.best_params_}')
→ Mejores F1-score obtenido for XGBoost: {'max_depth': 3, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 150}
```

Ambas técnicas de optimización concluyen que el modelo que mejor performance es el XG Boosting.



Randomized Search



Randomized es otra técnica de optimización que explora de forma aleatoria a través de los hiperparámetros ajustables.

En este gráfico podemos apreciar cuales serian los mejores hiperparámetros según el modelo a utilizar.

```
Mejores Parámetros (RandomizedSearch) for Logistic Regression: {'penalty': 'l2', 'C': 0.1}
Accuracy (RandomizedSearch) for Logistic Regression: 0.95501
Mejores Parámetros (RandomizedSearch) for Random Forest: {'min_samples_split': 5, 'max_depth': 3}
Accuracy (RandomizedSearch) for Random Forest: 0.88
Mejores Parámetros (RandomizedSearch) for KNeighborsClassifier: {'weights': 'uniform', 'n_neighbors': 3}
Accuracy (RandomizedSearch) for KNeighborsClassifier: 0.94727
Mejores Parámetros (RandomizedSearch) for DecisionTreeClassifier: {'min_samples_split': 2, 'max_depth': 7}
Accuracy (RandomizedSearch) for DecisionTreeClassifier: 0.95534
Mejores Parámetros (RandomizedSearch) for Gradiant Boosting: {'n_estimators': 300, 'max_depth': 7, 'learning_rate': 0.2}
Accuracy (RandomizedSearch) for Gradiant Boosting: 0.95479
Mejores Parámetros (RandomizedSearch) for XGBoost: {'min_samples_split': 5, 'max_depth': 7}
Accuracy (RandomizedSearch) for XGBoost: 0.95524
CPU times: user 1h 37min 49s, sys: 18.4 s, total: 1h 38min 7s
Wall time: 1h 35min 58s
```

```
# Mejor estimador entrenado

print(f'Mejores hiperparámetros for {nombre}: {random_search.best_params_}')
# mejor puntuación obtenida
print(f'Mejores F1-score obtenido for {nombre}: {random_search.best_score_}')

Mejores hiperparámetros for XGBoost: {'min_samples_split': 5, 'max_depth': 7}
Mejores F1-score obtenido for XGBoost: 0.2293547733058229
```

Randomized concluye que el modelo que mejor performa es el XG Boosting.



Conclusiones

- Basándonos en el análisis de métricas y el ajuste de hiperparámetros, el modelo XGBoost es la mejor opción para este conjunto de datos si se prioriza la reducción de falsos positivos, logrando un buen equilibrio con una alta sensibilidad y un AUC sólido.
- Gradient Boosting es una alternativa sólida que, comprende el mismo rendimiento en el AUC, proporciona menos falsos positivos pero es bajo en F1 score, de igual manera proporciona resultados consistentes y comprensibles.
- Finalmente, el Árbol de decisiones representa una buena opción de modelo balanceado que, si bien supera al XGBoost en sensibilidad, su precisión es menor. Aun así sigue siendo robusto y adecuado para aplicaciones donde ambos tipos de error falsos positivos y falsos negativos tienen costos similares.

En resumen, con las optimizaciones todo recae sobre el Trade-Off por lo que, aumentar la precisión tiende a disminuir la sensibilidad y viceversa.

Se puede concluir que para aplicaciones en las que minimizar los falsos positivos es crucial, XGBoost con ajuste de hiperparámetros es la elección más efectiva, ya que busco que el modelo no solo tenga un buen rendimiento en un conjunto específico, sino que generalice bien a diferentes subconjuntos de datos, proporcionando una detención robusta de la clase minoritaria sin generar demasiados falsos negativos.