

# Detección de fraudes

¿Cuáles son los patrones que se combinan para que haya un fraude en transacciones con tarjetas de crédito?

Creado por: Bianca Asinari

# Contenido

1. Contexto y audiencia
2. Hipótesis/preguntas de interés
3. Análisis Exploratorio e insights
4. Algoritmo/Métricas
5. Conclusión



# Contexto, objetivo y audiencia

## Contexto

La detección de fraudes con tarjetas de crédito es un problema inminente para todas las empresas del sector bancario. El fraude se ha presentado en los últimos años como uno de los grandes males para las empresas, las economías y la sociedad en general.

En el caso que se analiza a continuación, la entidad bancaria convive en un ambiente de riesgos de actos fraudulentos, los cuales se maximizan con la diversificación de actividades, la globalización y las transacciones electrónicas.

## Objetivo

El objetivo de este estudio es poder determinar cuales son las distintas combinaciones de factores que tienen probabilidad de convertirse en fraude. A lo largo de este proyecto se quieren generar medidas que puede tomar el Banco para que esa cantidad de fraudes se minimicen y en el mejor de los casos se reduzcan a 0.

## Audiencia

Nuestro cliente es una entidad financiera que nos pide detectar y prevenir el fraude con tarjetas de crédito ya que es una necesidad fundamental para mejorar la eficiencia del negocio.

## Datos

Los registros obtenidos provienen de transacciones con tarjetas de crédito en los Estados Unidos entre los años 2019 y 2020.

# Primeras preguntas de interés

¿Cuáles son las combinaciones de factores que generan más probabilidad de fraudes?

**Algunas preguntas que ayudan a responder esta pregunta:**

1. ¿Hubo algún mes/día del mes/día de la semana donde aumento la detección de fraudes?
2. ¿Existe algún patrón marcado en el tipo de comercio donde se cometen mas fraudes?
3. ¿Existe alguna relación entre el género y el fraude?
4. ¿Existe algún patrón relacionado con la edad de las personas que cometen fraudes?
5. ¿Existe algún patrón de la transacción según el monto del gasto efectuado?
6. ¿Existe alguna relación entre la ciudad, la cantidad de habitantes y la cantidad de fraudes?

# Resumen Metadata

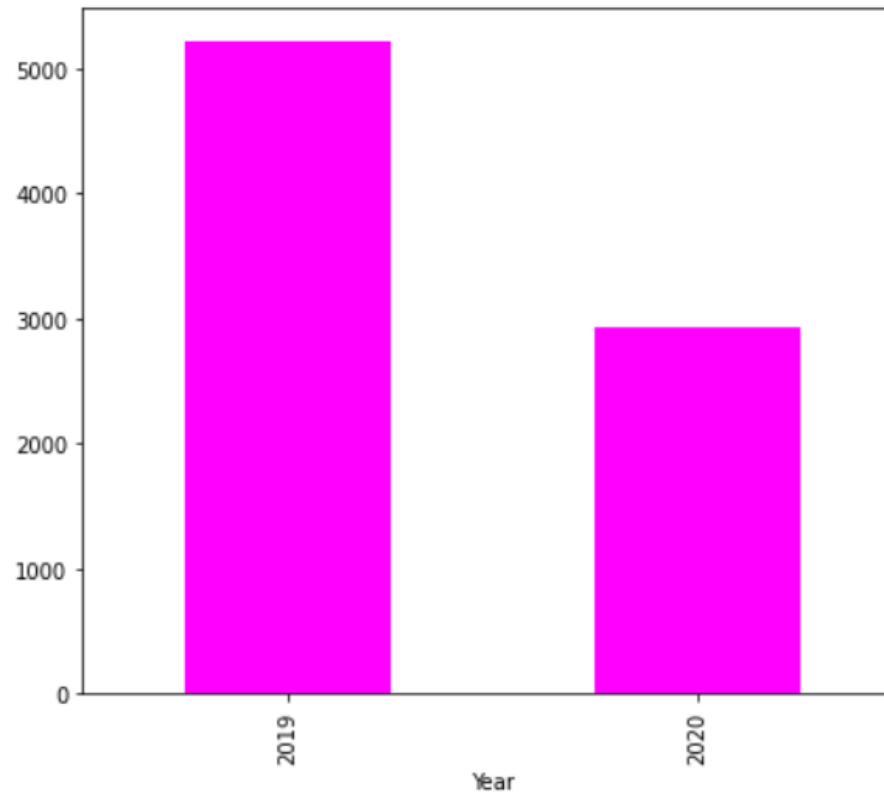
**1604294**

Total registros

**8151**

Registros Fraudulentos

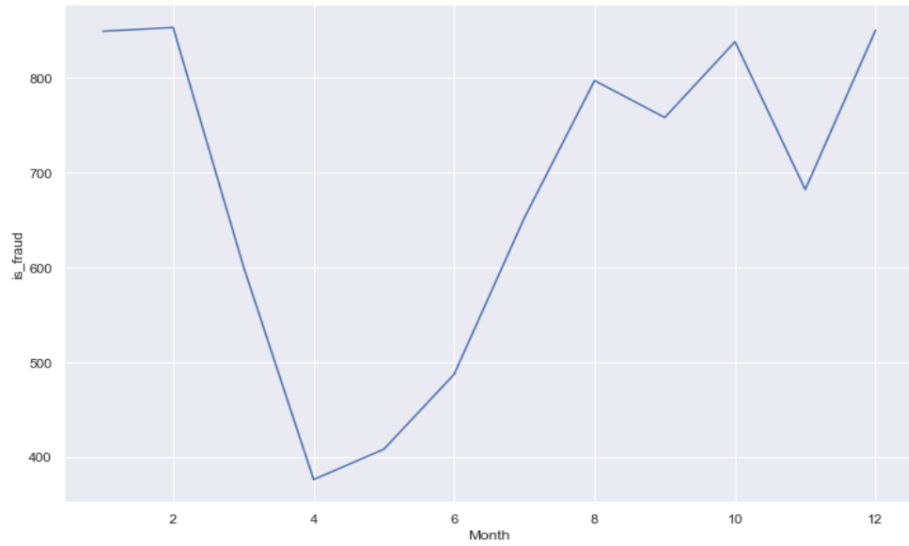
## Cantidad de fraudes por año



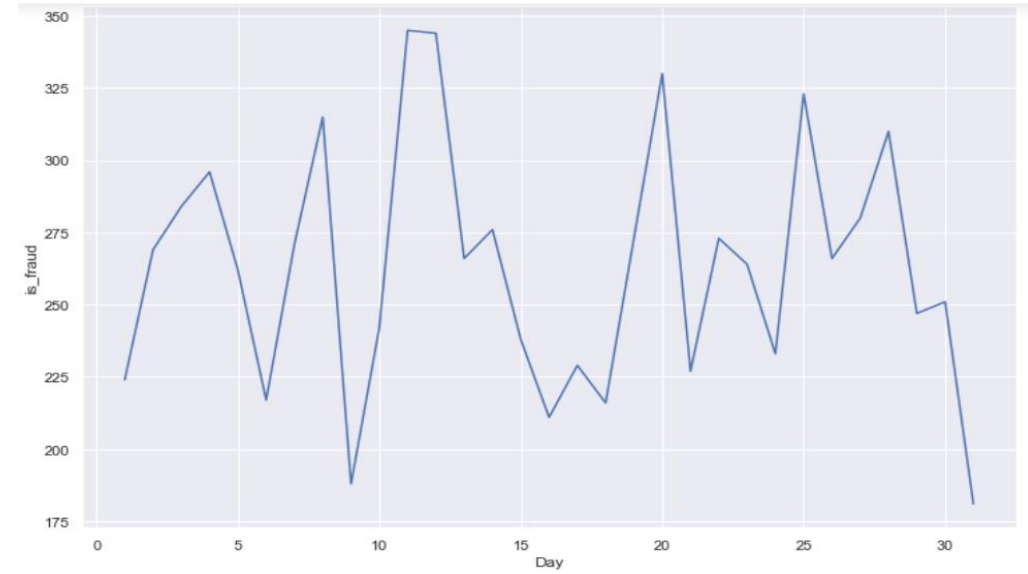
# ANÁLISIS EXPLORATORIO

# ¿Hubo algún mes/día del mes/día de la semana donde aumentó la detección de fraudes?

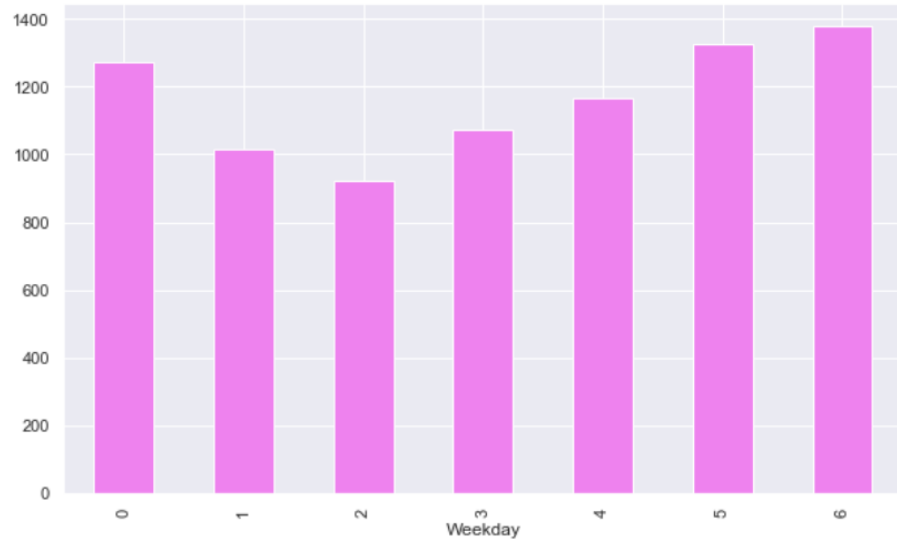
Cantidad de fraudes por mes



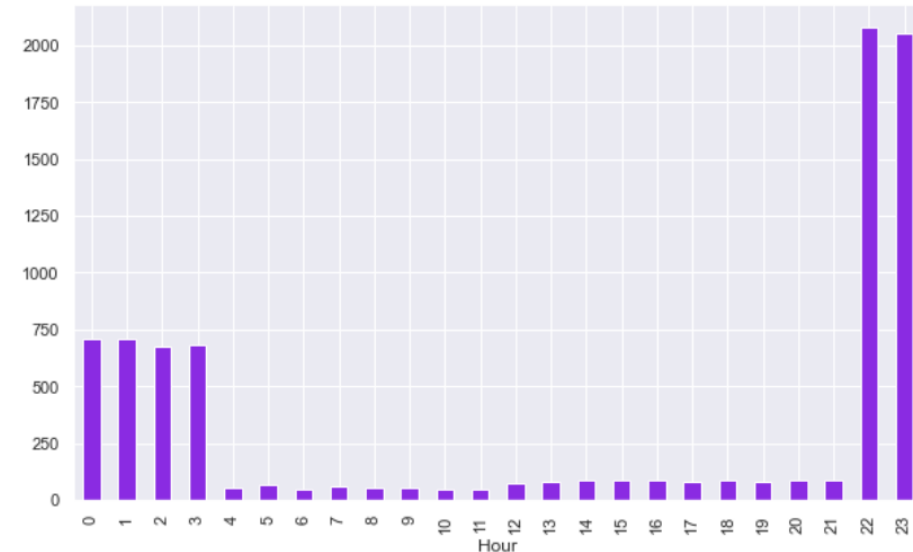
Cantidad de fraudes por día del mes



Cantidad de fraudes por día de la semana



Cantidad de fraudes por hora



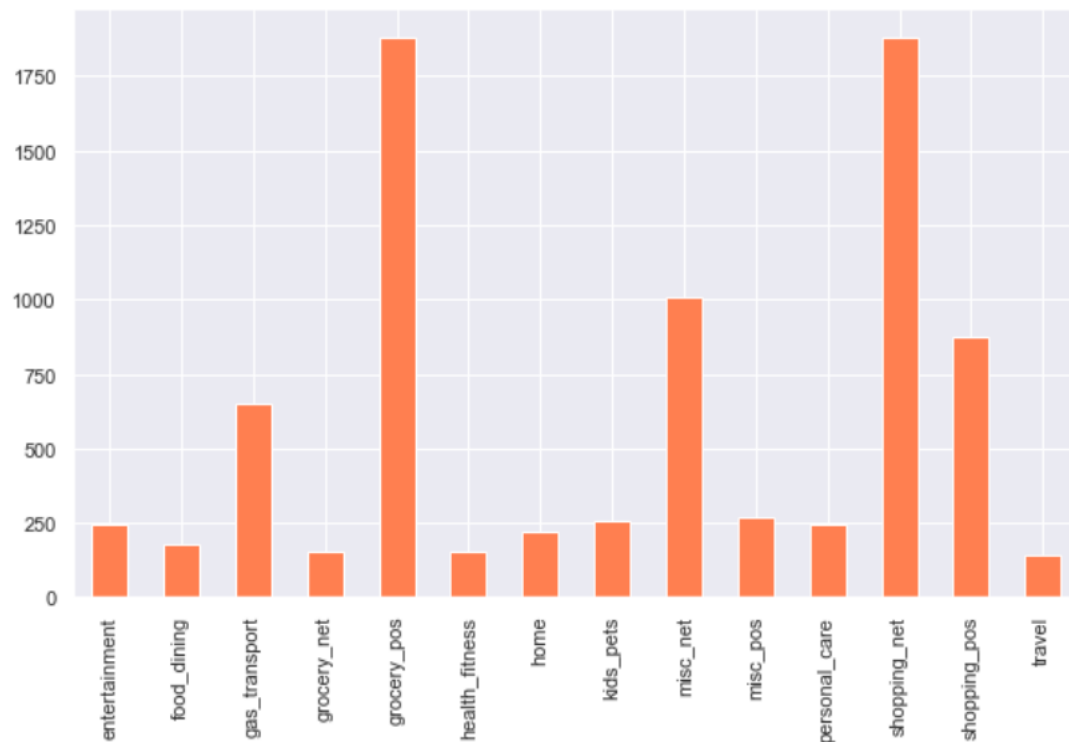
## ¿Hubo algún mes/día del mes/día de la semana donde aumentó la detección de fraudes?

- En lo que respecta a los meses, aquellos con **menor cantidad de fraudes** son abril, mayo y junio. Mientras que los que tienen más cantidad de fraudes son octubre, diciembre y enero.
- En cuanto a los días del mes, el pico máximo se da entre los días 10 y 12 y el pico mínimo a fin de mes.
- Respecto a los días de la semana se observa un **ascenso** de las transacciones fraudulentas para los días viernes, sábado y domingo.
- Se observa un marcado aumento de los fraudes en primer lugar entre las 22 y las 23 horas y en segundo lugar, entre las 0 y las 3 horas. Esto podría deberse ya que hay más fraudes a través de internet, por lo que las personas en el horario que está abierto un comercio no cometen fraudes pero si lo hacen a través de compras online.

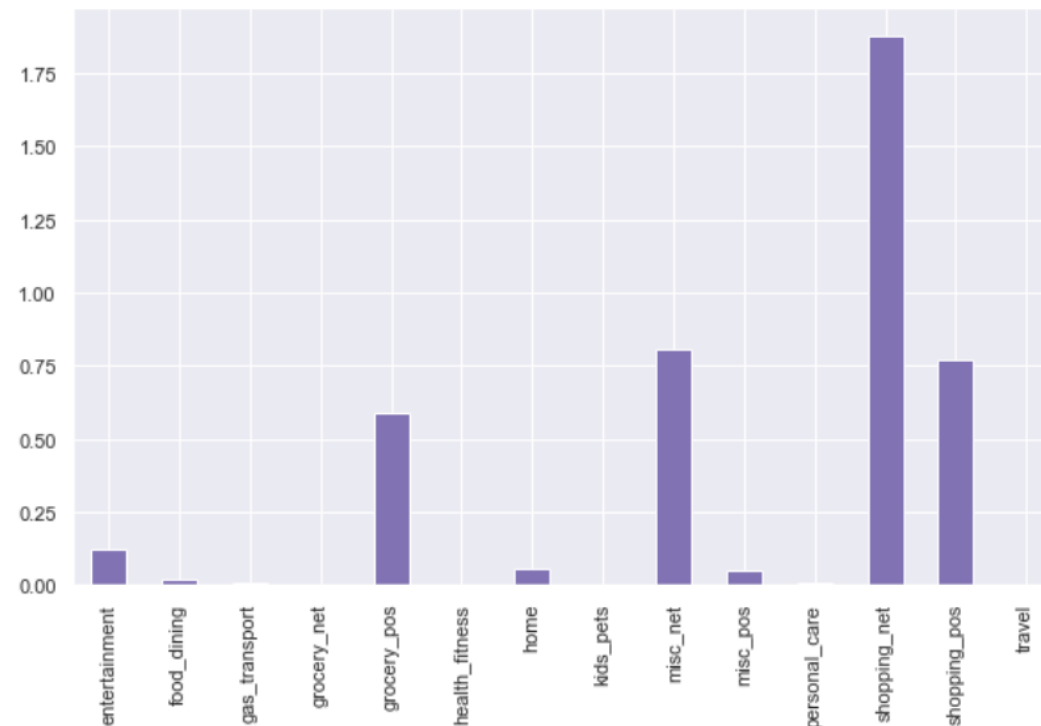


## ¿En cuáles comercios se cometen mas fraudes?

Fraudes por comercio



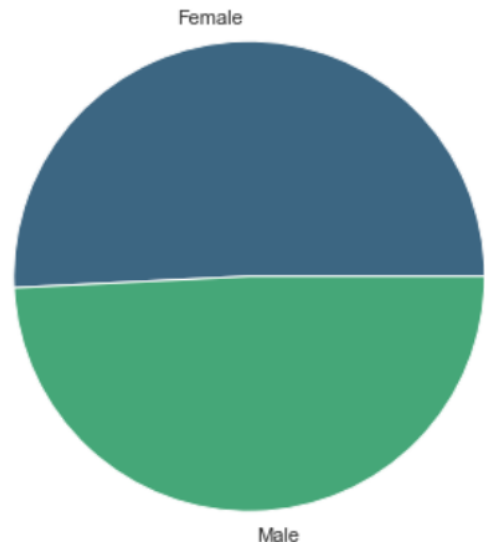
Monto de fraudes por comercio



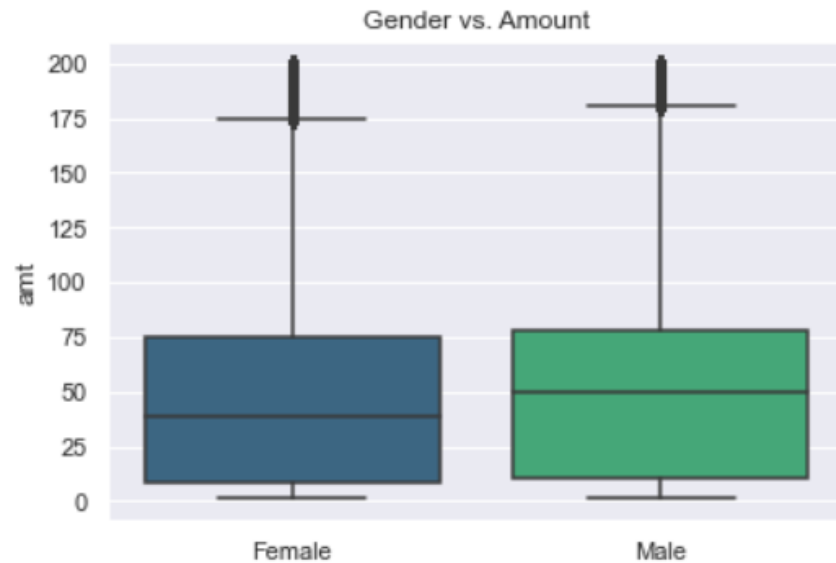
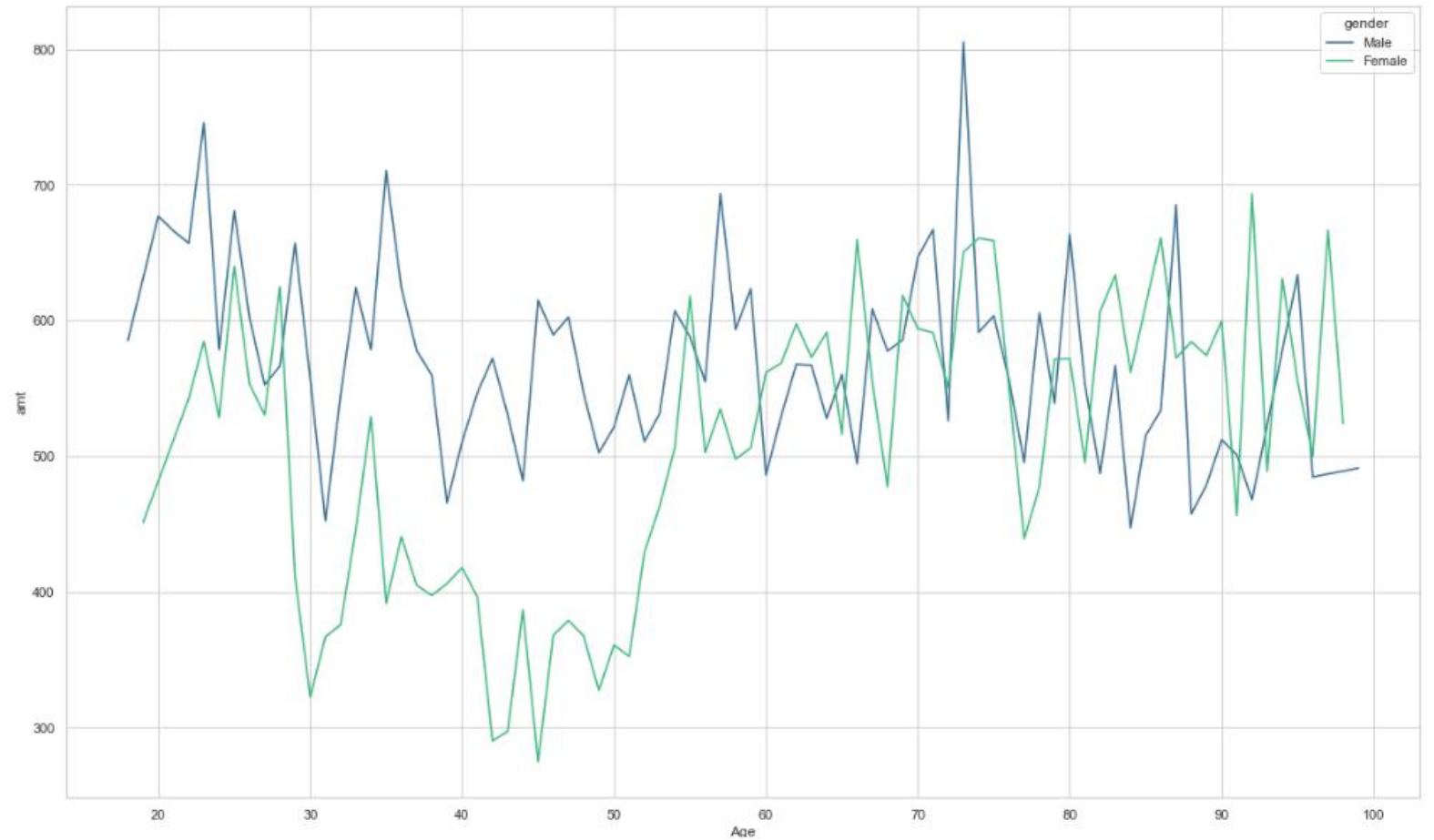
- Se observa que en determinados comercios hay más fraudes que otros. El fraude tiende a ocurrir con más frecuencia en 'Shopping\_net', 'Grocery\_pos' y 'misc\_net', mientras que 'home' y 'kids\_pets', entre otros, tienden a tener transacciones más normales que fraudulentas.

# ¿Existe alguna relación entre el género y el fraude?

Fraudes por género



Monto del fraude por edad y género

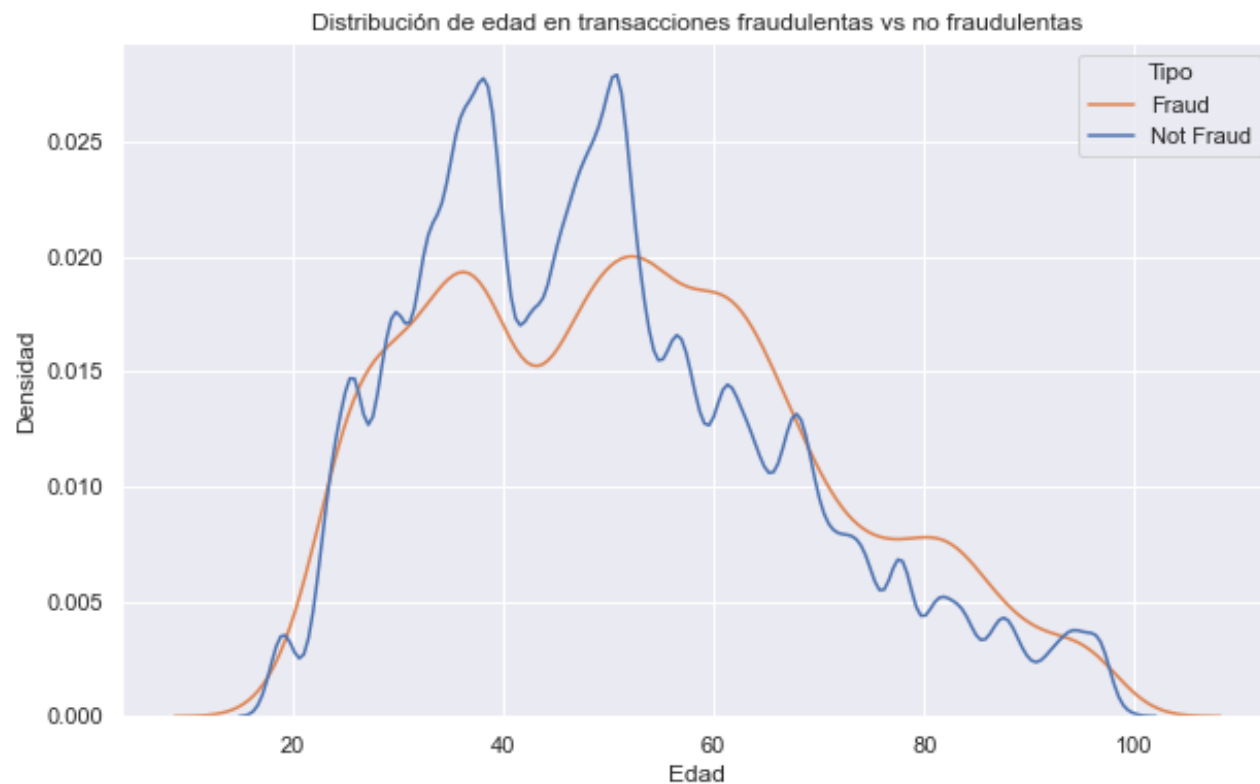


## ¿Existe alguna relación entre el género y el fraude?

- En azul se pueden ver los fraudes cometidos por mujeres y en verde los cometidos por hombres.
- Se observa una leve diferencia entre los fraudes causados por el sexo masculino y el sexo femenino. Según la revista New York Times, en Estados Unidos en promedio el sueldo de las mujeres es un 18% más bajo que el de los hombres, razón por la cual ellas pueden tener una mayor tendencia a recurrir al delito.
- La brecha salarial también puede comprobarse en la media del monto del gasto, donde la media del gasto de los hombres es mayor que el de las mujeres.
- A su vez, la media del monto de fraudes es mayor para el sexo masculino.
- En lo que respecta a la edad Los patrones más visibles pueden resumirse en que los montos fraudulentos más bajos se ven reflejados en mujeres entre los 35 y los 50 años, en el resto de las edades para las mujeres se mantiene relativamente constante. Mientras que para los hombres se tiene un pico de montos altos entre los 20 y los 25 años aproximadamente y otro pico aun mayor entre los 70 y 75 años.

# ¿Existe alguna relación entre la edad y el fraude?

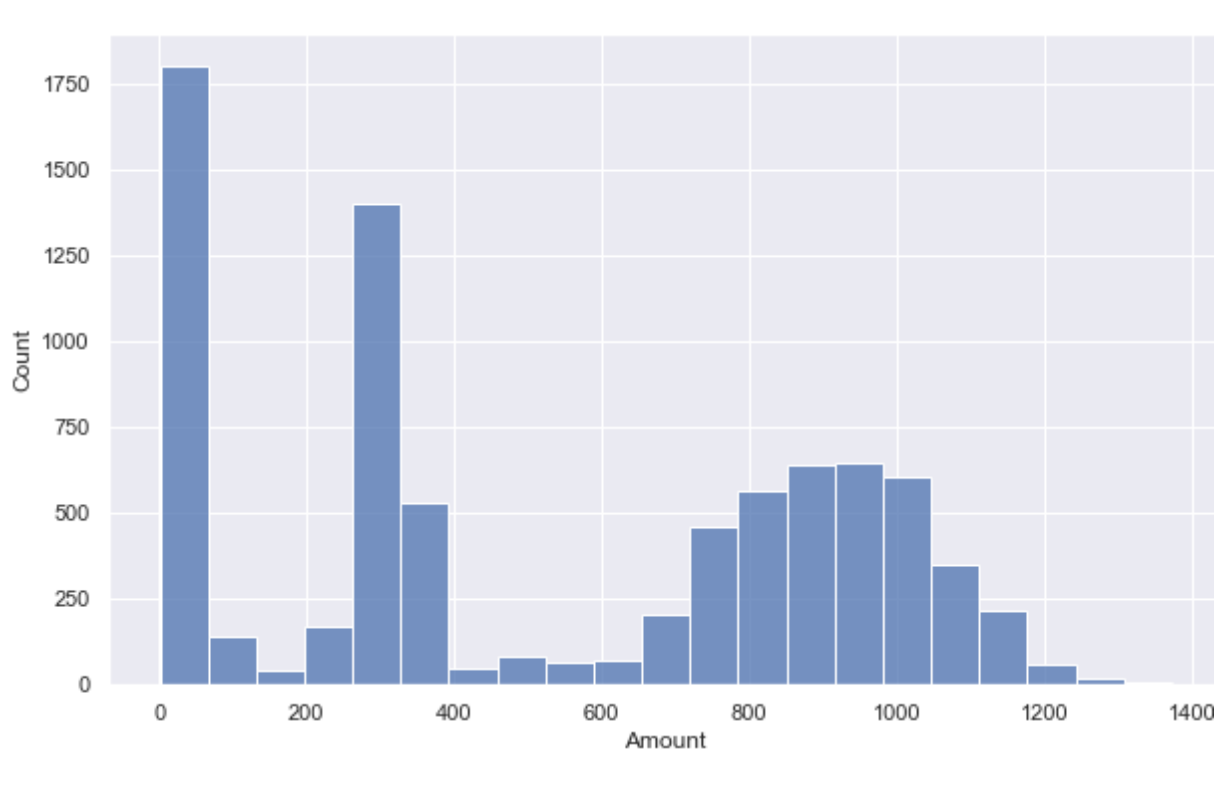
## Fraudes mostrados en forma porcentual según edad



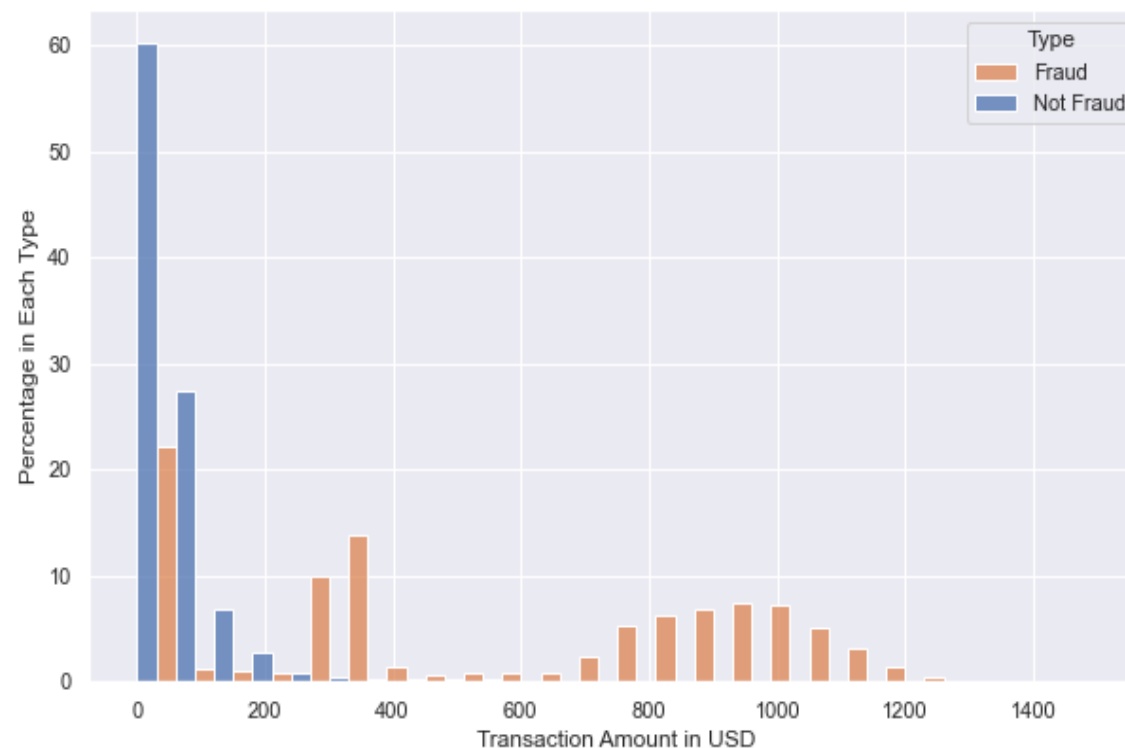
- En lo que respecta al primer gráfico, la distribución de edad es notablemente diferente entre 2 tipos de transacciones. En las transacciones normales, hay 2 picos en la edad de 37-38 y 49-50, mientras que en las transacciones fraudulentas, la distribución por edad es un poco más suave y el segundo pico incluye un grupo de edad más amplio de 50-65. Esto sugiere que las personas mayores son potencialmente más propensas al fraude.

## ¿Existe alguna relación entre el monto gastado y el fraude?

Monto de transacciones fraudulentas



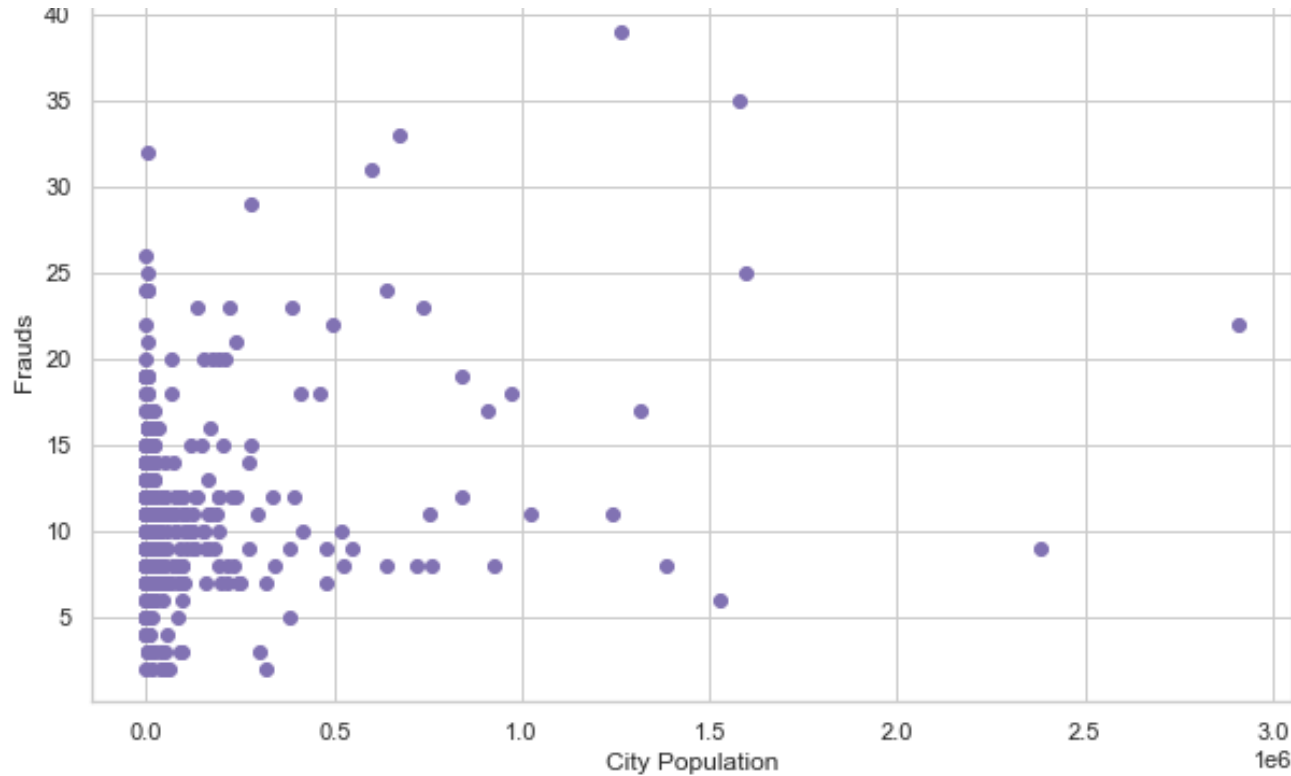
Comparación porcentual entre tipo de transacciones



Mientras que las transacciones normales tienden a rondar los 200 USD o menos, se observa que las transacciones fraudulentas alcanzan un máximo de 300 USD y luego en el rango de 800-1000 USD

## ¿Existe alguna relación entre la ciudad donde se cometen los fraudes y la cantidad de fraudes?

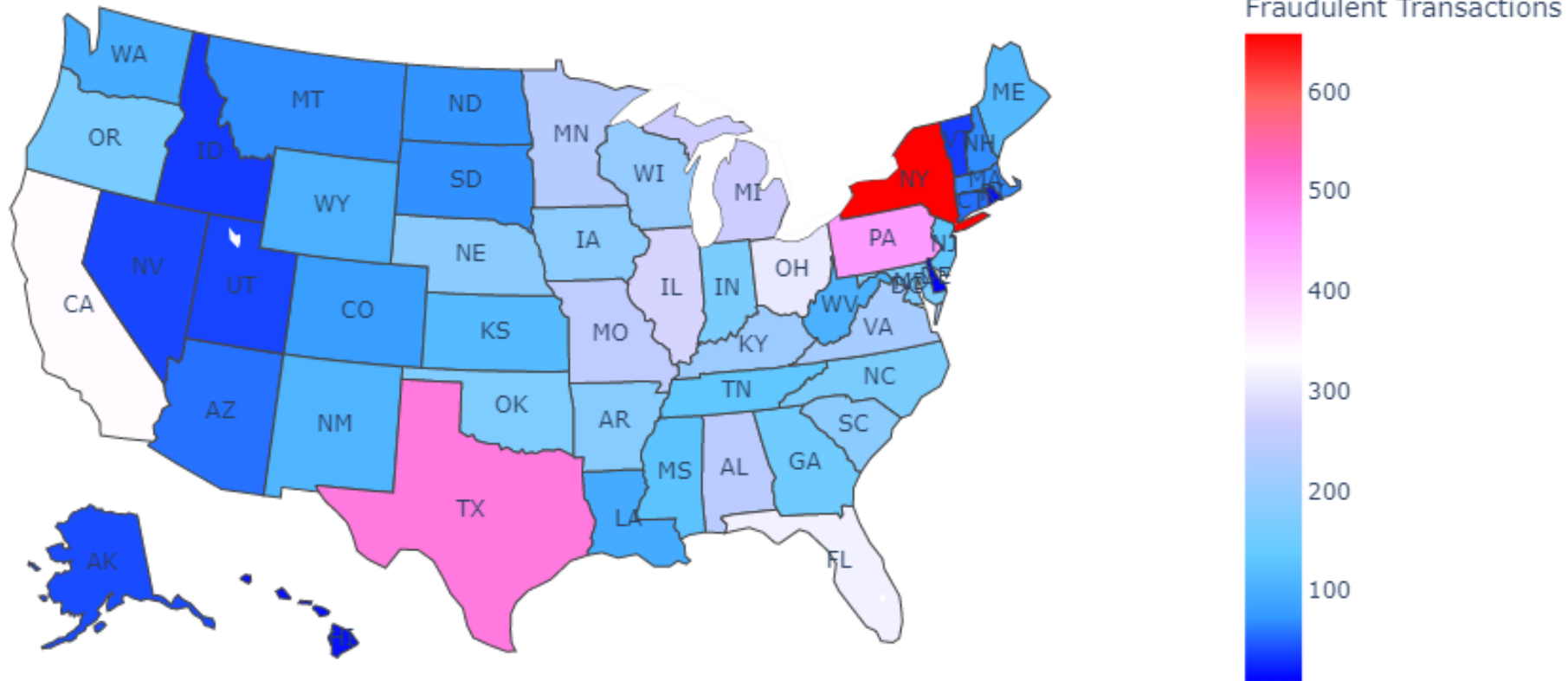
Población vs. fraudes



Se observa un patrón marcado de la existencia de fraudes en ciudades pequeñas de hasta medio millon de habitantes.

¿Existe alguna relación entre el Estado donde se cometen los fraudes y la cantidad de fraudes?

## Población por es vs. fraudes



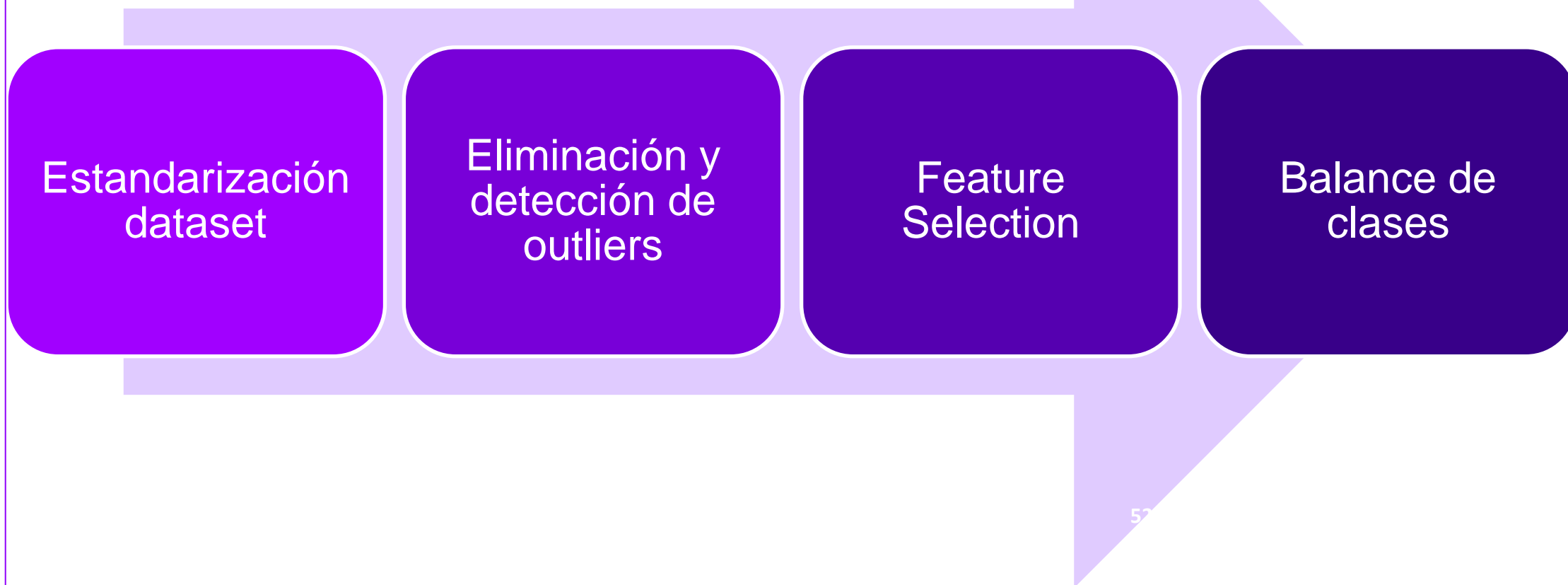
- En cuanto a los Estados, se establece que hay una leve diferencia entre los estados de la costa Oeste con los de la costa Este, siendo la cantidad de fraudes mayor para esta última.

# ALGORITMO SELECCIONADO Y MÉTRICAS



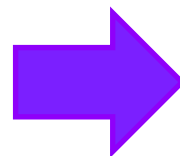
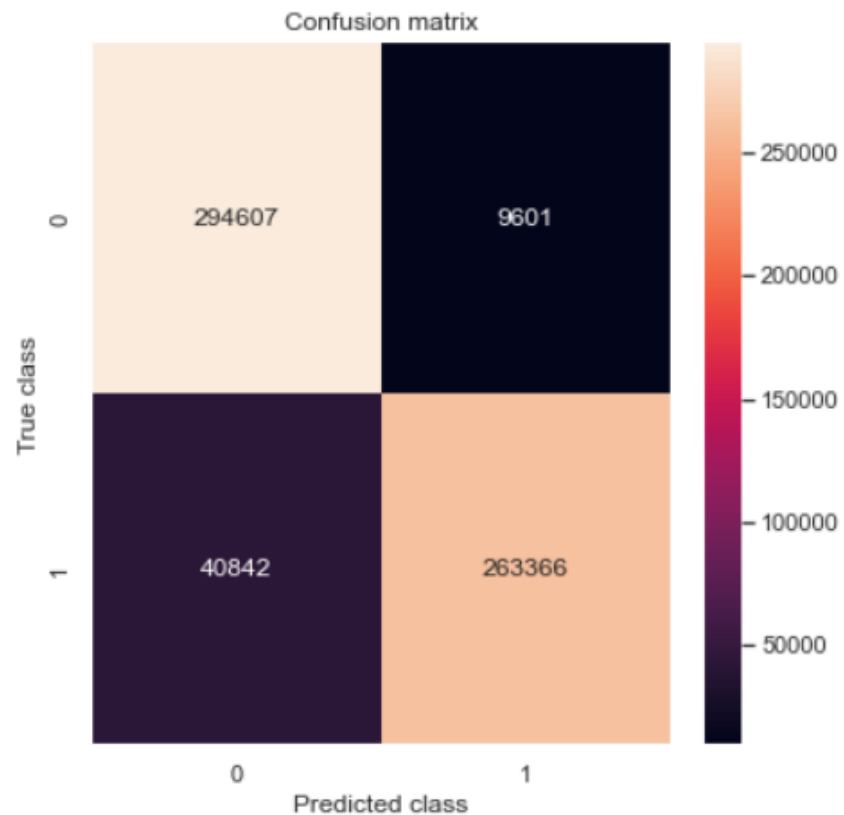
## Algoritmo Seleccionado

Pasos previos a la selección:



## Algoritmo Seleccionado

### Matriz de confusión



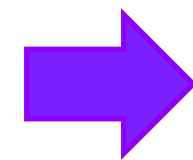
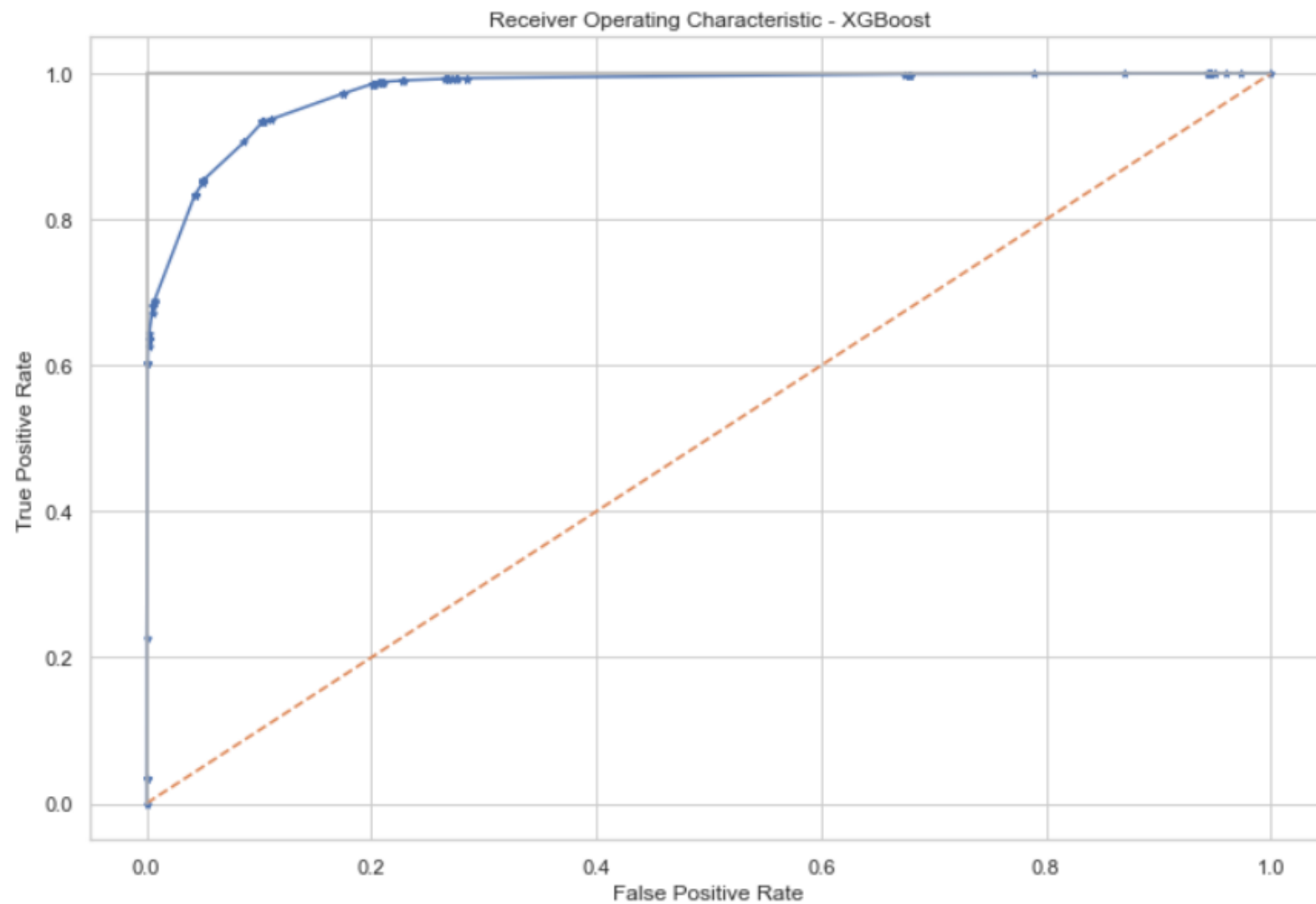
**Extreme Gradient Boosting (XGB)**



	precision	recall	f1-score	support
0	0.88	0.97	0.92	304208
1	0.96	0.87	0.91	304208
accuracy			0.92	608416
macro avg	0.92	0.92	0.92	608416
weighted avg	0.92	0.92	0.92	608416

## Algoritmo Seleccionado

ROC AUC Score → Área bajo la curva



**0.976**



---

F1-Score (Fraude) = 91%

Conclusión

F1-Score (NoFraude) = 92%

Area bajo la curva = 0.976

---

# ¡Objetivo cumplido!

Se establece que el objetivo planteado inicialmente de predecir el comportamiento de los clientes en cuanto a cometer un fraude o no a través del algoritmo de XGB fue cumplido.

El modelo XGB arrojó un valor de area bajo la curva de 0.976. Métrica que se eligió como la más relevante para el análisis.

A su vez, para clasificar gente que puede cometer un fraude, este mismo nos ofrece un f1-score de 91% y 92% para quienes no cometerán fraude por lo que se puede establecer que el balance y la combinación entre Precision y Recall son buenos.

Por otro lado, las demás métricas también están por encima del 86%.

Es decir, se obtuvo un algoritmo confiable a la hora de Clasificar clientes fraudulentos y no fraudulentos.