



PREDICCIÓN DE

MODELO DE DETECCIÓN DE FRAUDE CON TARJETAS DE CRÉDITO

FRAUDE

Resumen



Proyecto de Machine Learning

01 Presentación del problema

02 Planteamiento técnico del problema

03 Descripción de los datos

04 Análisis

05 Modelo utilizado

06 Resultados

Presentación del problema



Descripción

Este proyecto se enfoca en analizar las transacciones realizadas con tarjetas de crédito de una entidad bancaria específica.

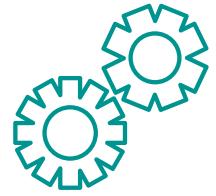
Propósito

El objetivo principal de este proyecto es desarrollar un modelo de machine learning capaz de predecir qué transacciones realizadas con tarjeta de crédito son fraudulentas.

Contexto

Los datos utilizados en este proyecto provienen de usuarios de tarjetas de crédito estadounidenses.

Planteamiento del problema técnico

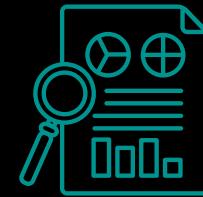


Este problema se presenta debido a la creciente sofisticación de los métodos fraudulentos, generando un importante desafío para los sistemas de seguridad.

La detección temprana de transacciones fraudulentas ayuda a prevenir pérdidas financieras significativas tanto para el banco como para los clientes, cargos no autorizados o robos de identidad.

Buscamos desarrollar un modelo de Machine Learning que pueda analizar grandes volúmenes de datos de transacciones y, a partir de patrones y anomalías identificadas, clasificar de manera precisa las transacciones como legítimas o fraudulentas.

Descripción del dataset original



555.719	22	851	50	2020
Instancias	Features	Ciudades	Estados de EE.UU.	6 meses
<hr/>				
919	46	14	\$ 67.35	694
Usuarios	Promedio de edad	Categorías de gasto	Gasto promedio	Comercios

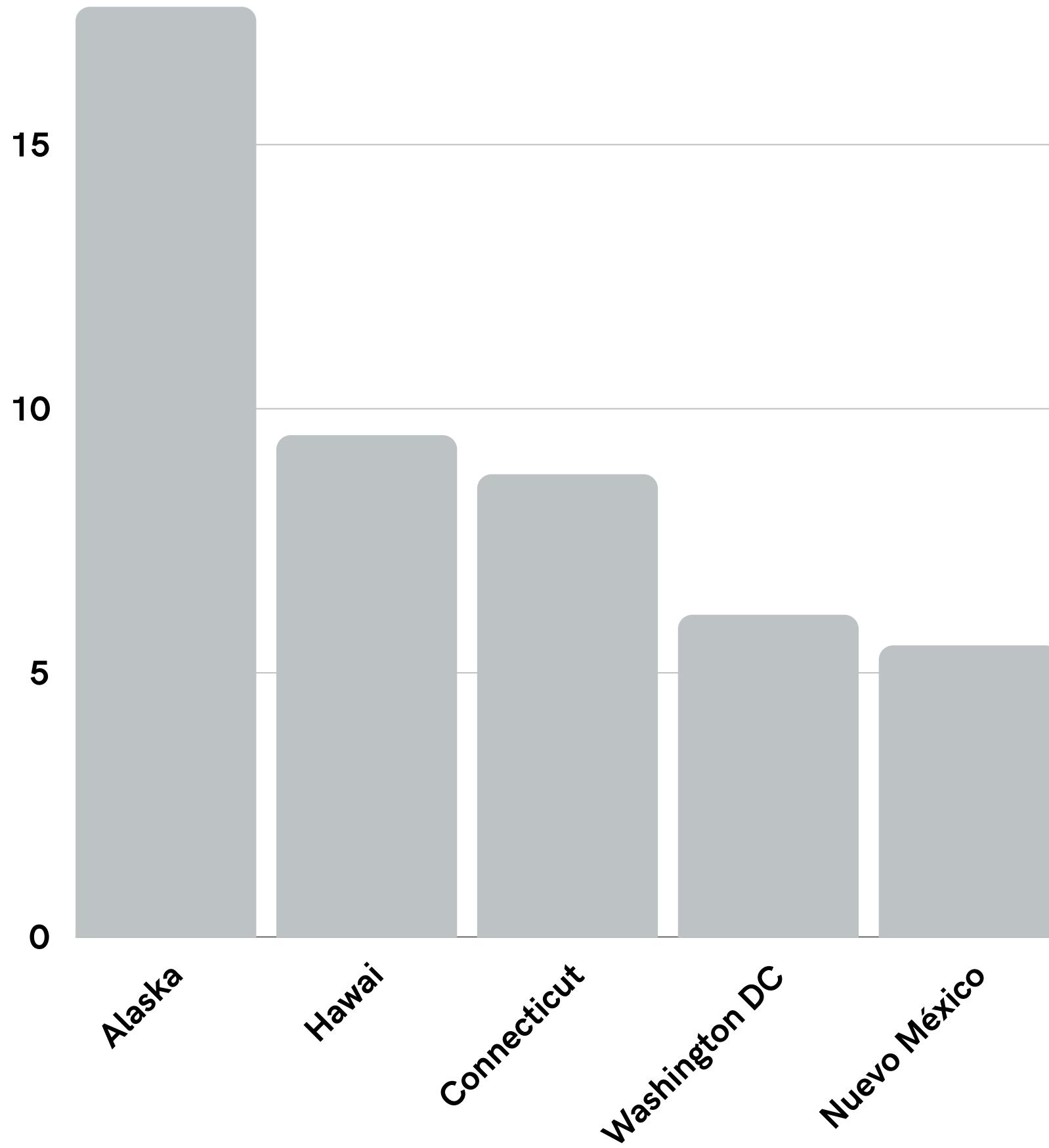
Análisis y principales estadísticas obtenidas

Transacciones fraudulentas



- Cantidad total defraudada: \$1,133,324.68
- Porcentaje cantidad defraudada: 2.9%
- Importe medio defraudado: \$410





Fraude

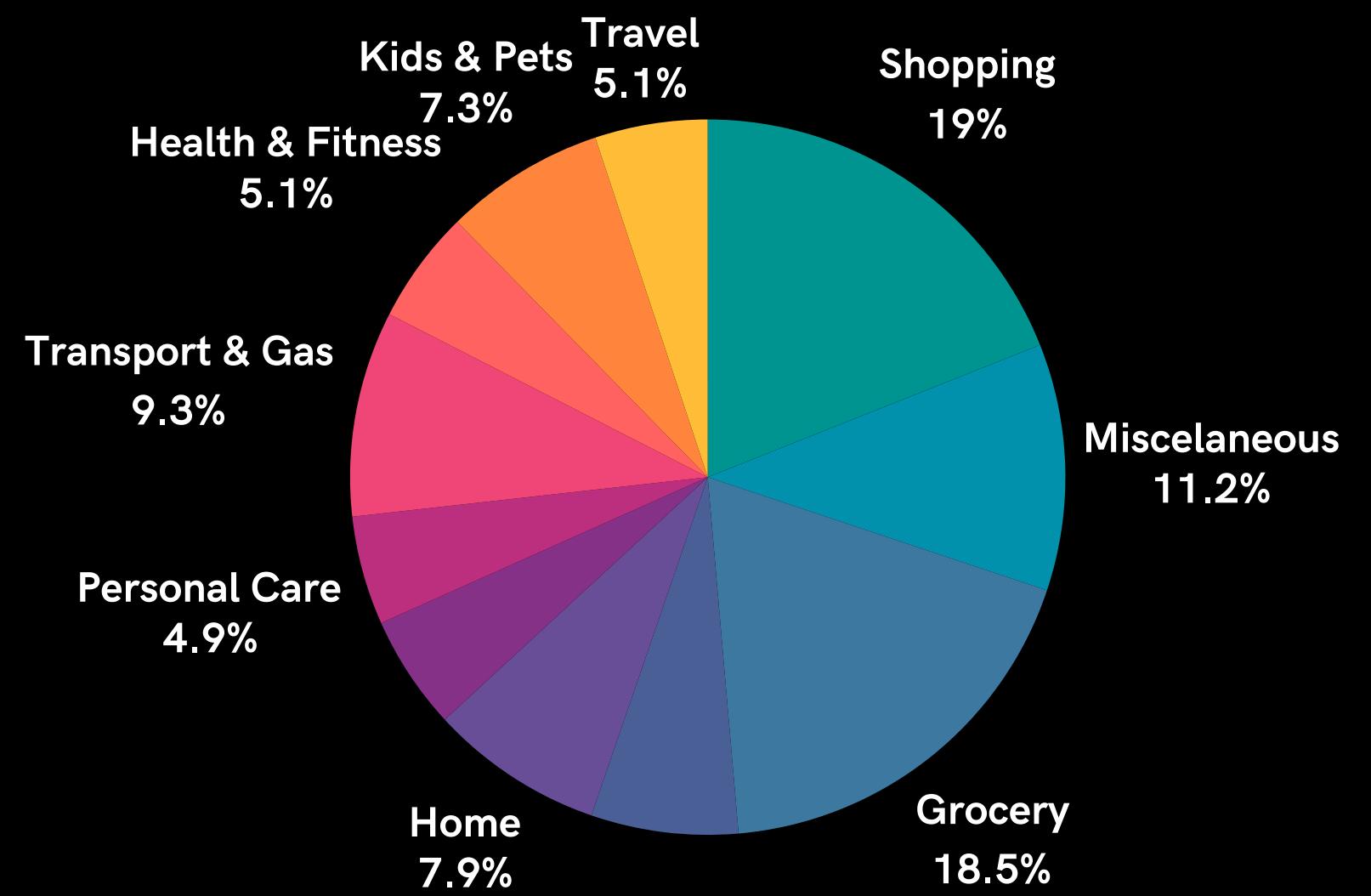
Porcentaje por Estado

17.61%

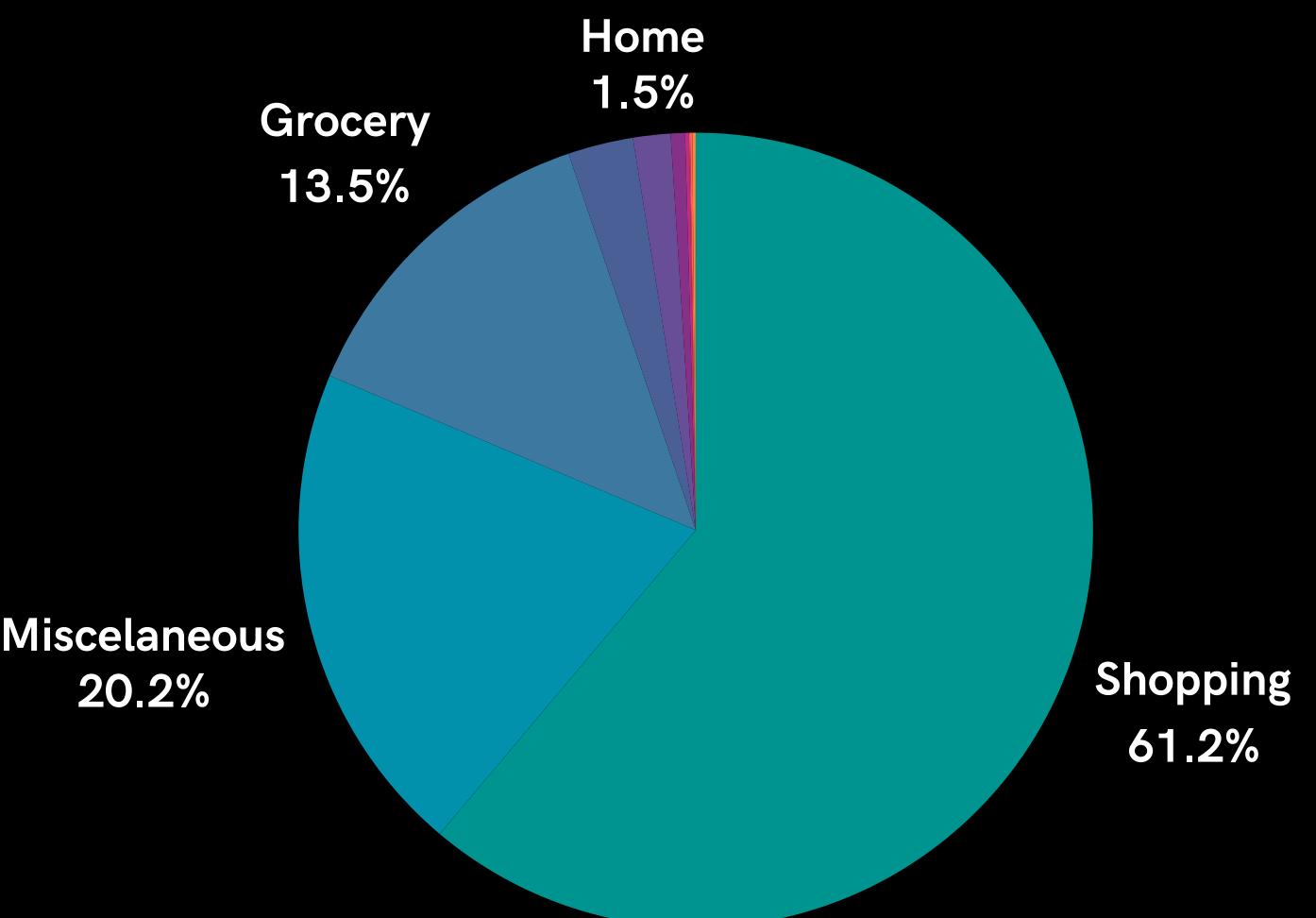
Alaska es, con diferencia, el Estado con mayor tasa de fraude entre los usuarios de tarjetas de crédito de esta entidad bancaria.

El promedio de los restantes Estados es inferior al 2.5%.

Análisis y principales estadísticas obtenidas



Gasto



Fraude

Feature Engineering



Modificaciones en el dataset

Creación de nuevas features

- Distancia entre el usuario y el comercio
- Categorización de los gastos
- Features temporales (horas, días de la semana, etc.)

Transformación de características

- Get dummies
- K-Means
- Normalización

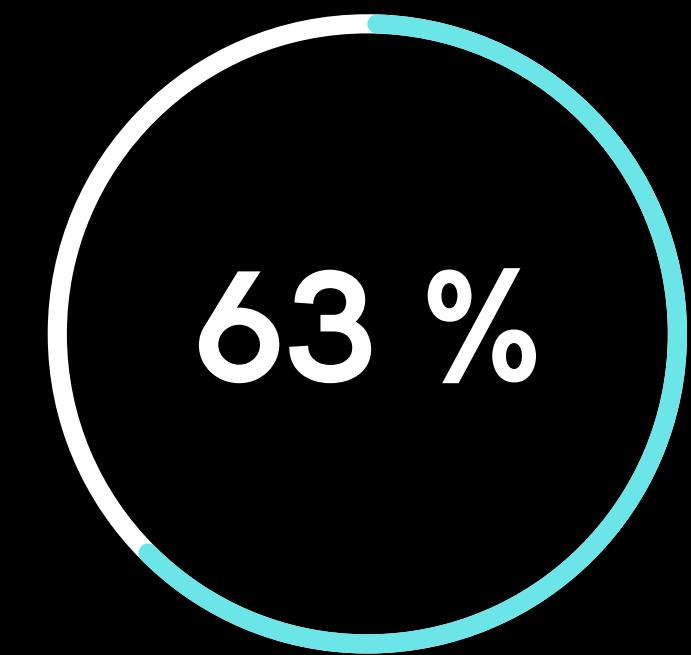
Tratamiento outliers y selección de features

- Clustering
- Eliminación de features poco relevantes o altamente correlacionadas

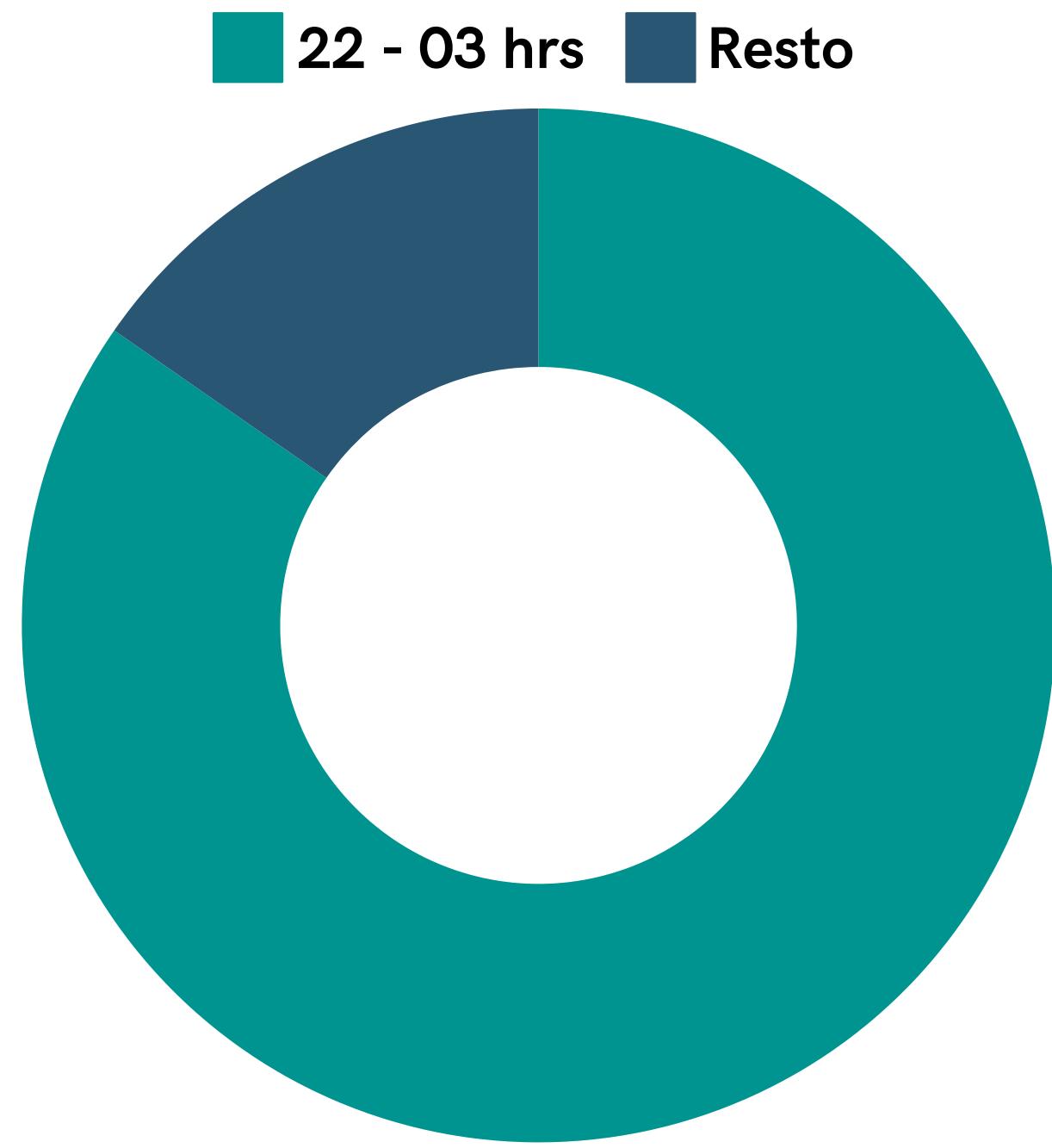


Principales datos obtenidos

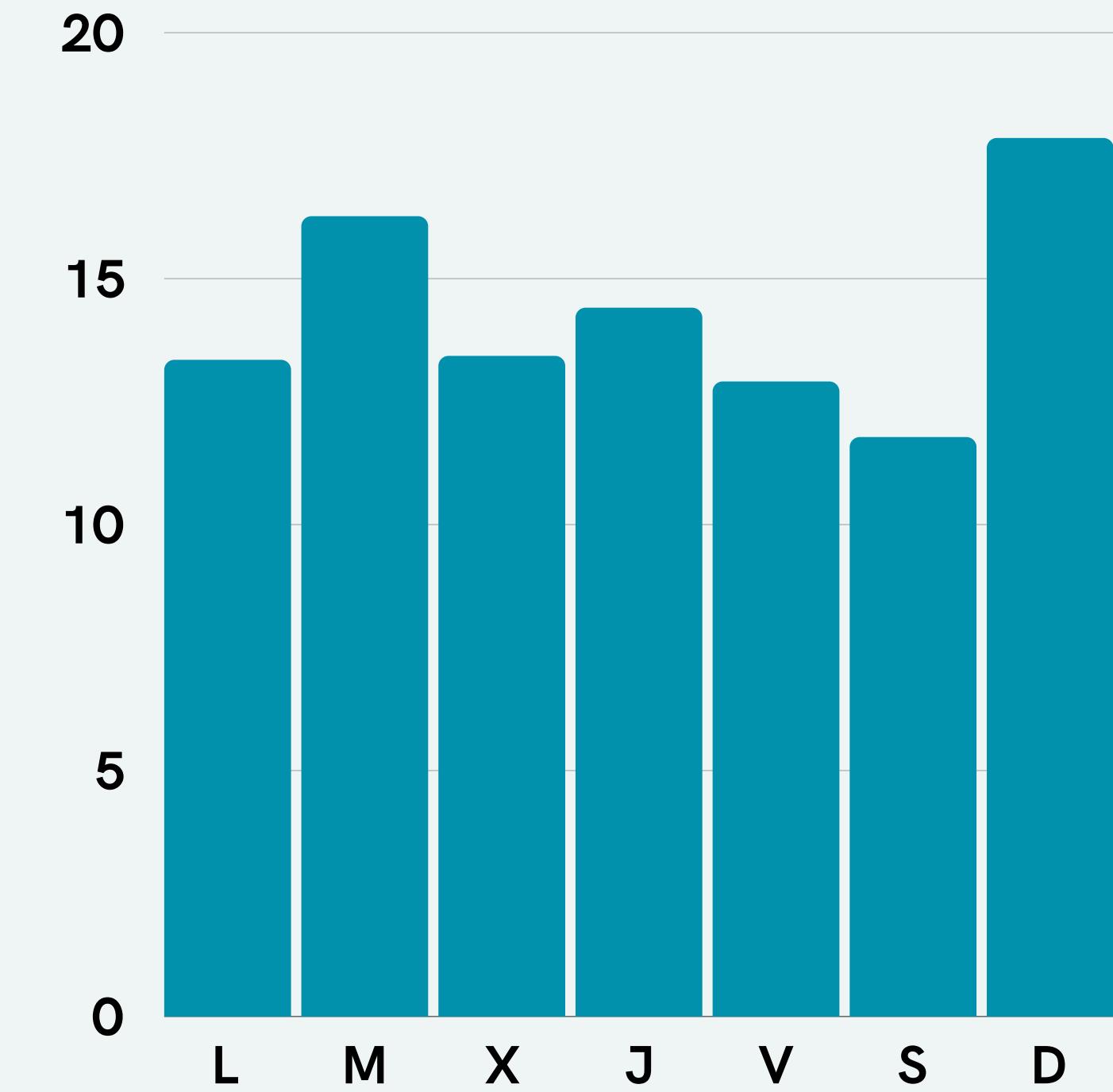
Transacciones fraudulentas son realizadas a través de internet



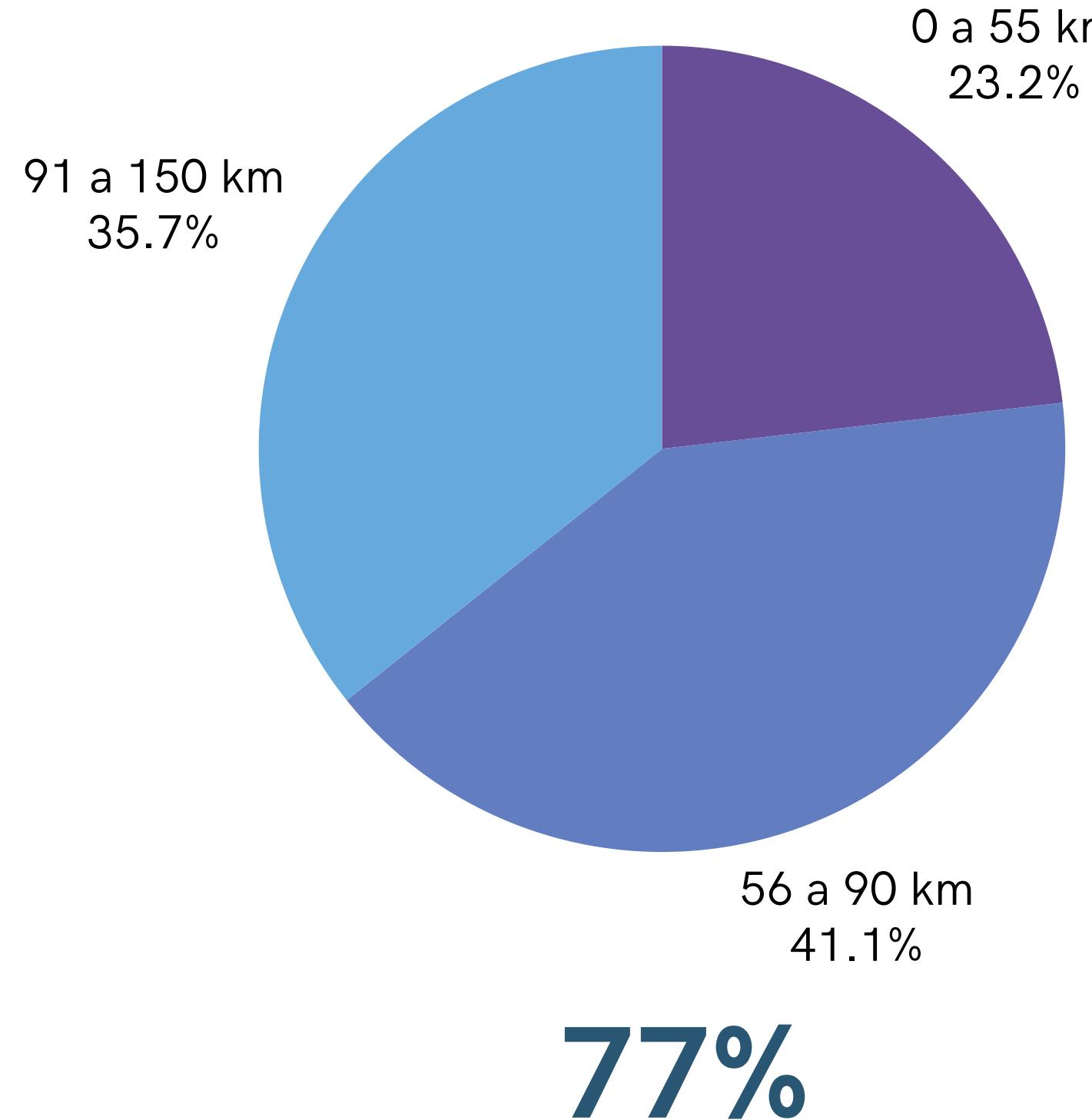
La mayoría de las transacciones fraudulentas en internet se destinan a la compra de productos diversos, excluyendo alimentos



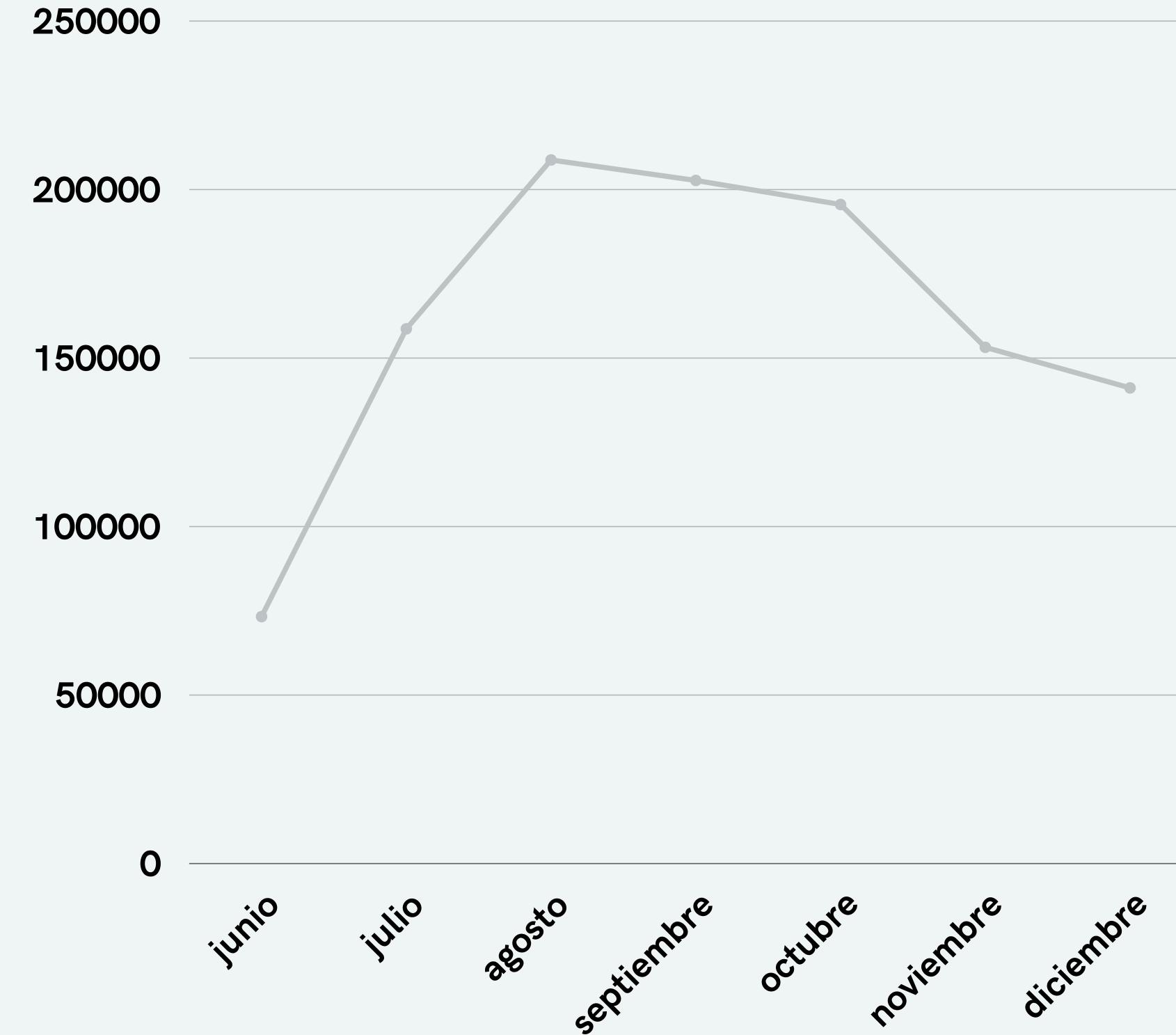
85%
de los fraudes se comete entre las
22:00 y las 03:00 horas.



30%
de los fraudes se comete el fin de
semana



Recorre de 56 a 150 km
para cometer el fraude
(en comercio físico)



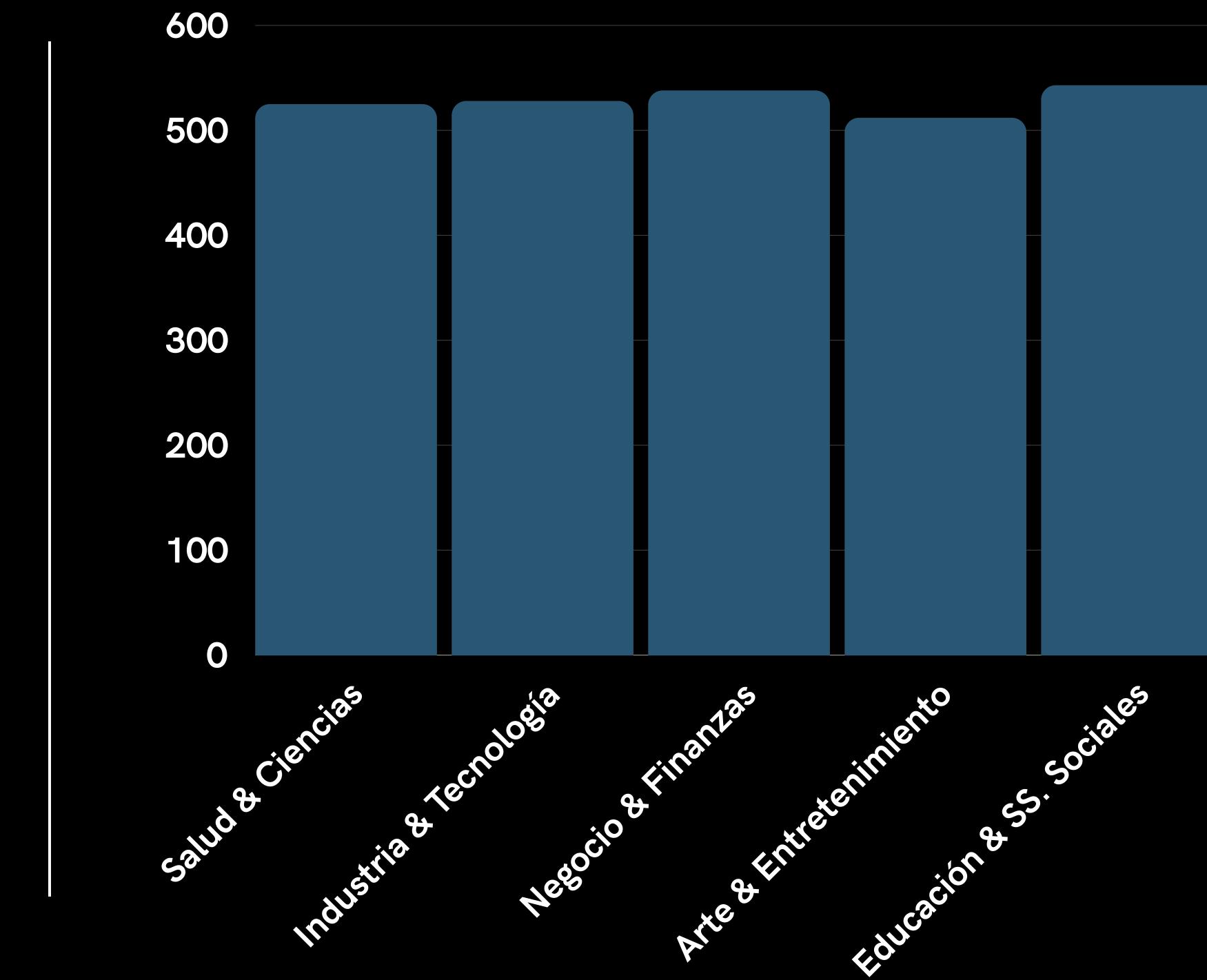
186%
Variación porcentual de
junio a agosto

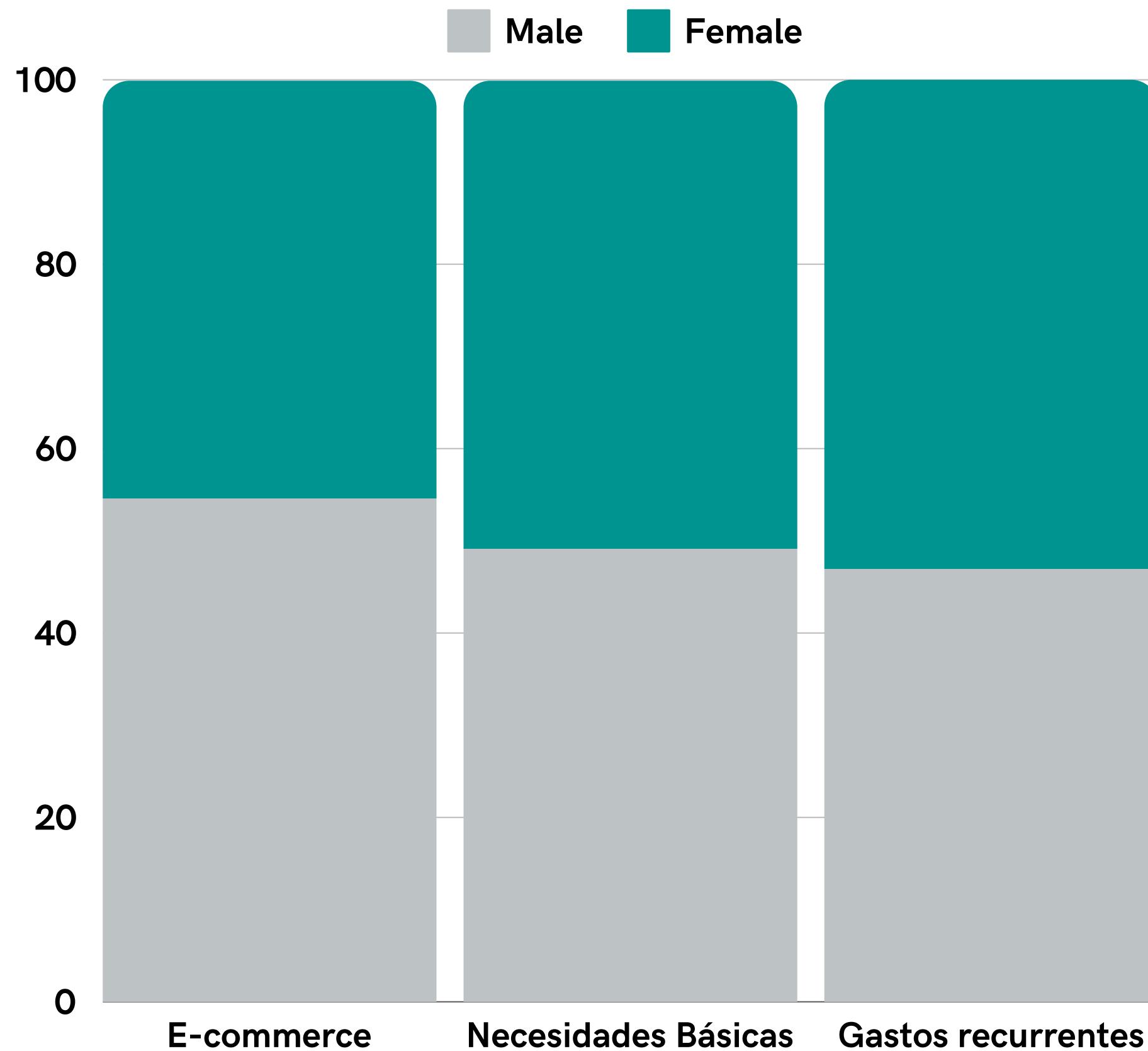
Fraude por tipo de empleo

\$530

**Importe medio defraudado
por tipo de trabajo**

No se observan diferencias
relevantes en función del tipo
de empleo del defraudador



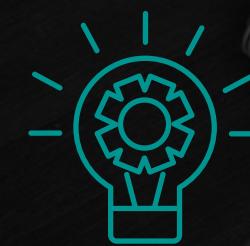


Quién defrauda?

Y en qué categoría?

- Los hombres defraudan más en compras hechas a través de Internet
- Las mujeres defraudan más en bienes básicos (alimento, transporte) y gastos recurrentes (gimnasios, mascotas, cuidados, etc.)

Modelo utilizado



XGBoost

F1-Score

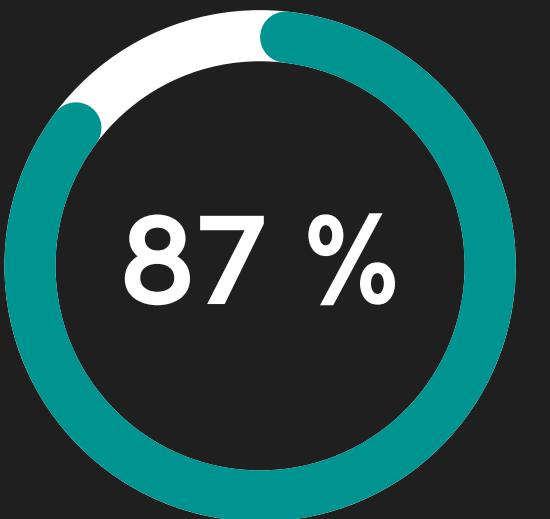
Tamaño del dataset
(550000, 90)

Target muy
desbalanceado

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	110718
1	0.95	0.81	0.87	426
accuracy			1.00	111144
macro avg	0.97	0.90	0.94	111144
weighted avg	1.00	1.00	1.00	111144

F1-Score



Resultados obtenidos

Característica	Importancia
amt	487.0
age	442.0
city_pop	382.0
hour_trans	236.0
distance_kms	226.0
gender_binary	80.0
category_misc_pos	51.0
category_travel	47.0
category_food_dining	45.0
basic_needs	44.0

Feature Importance



Gracias!

Julio Gómez Expósito