Uso de modelos de clasificación para la detección de fraude de tarjetas bancarias

IT ACADEMY

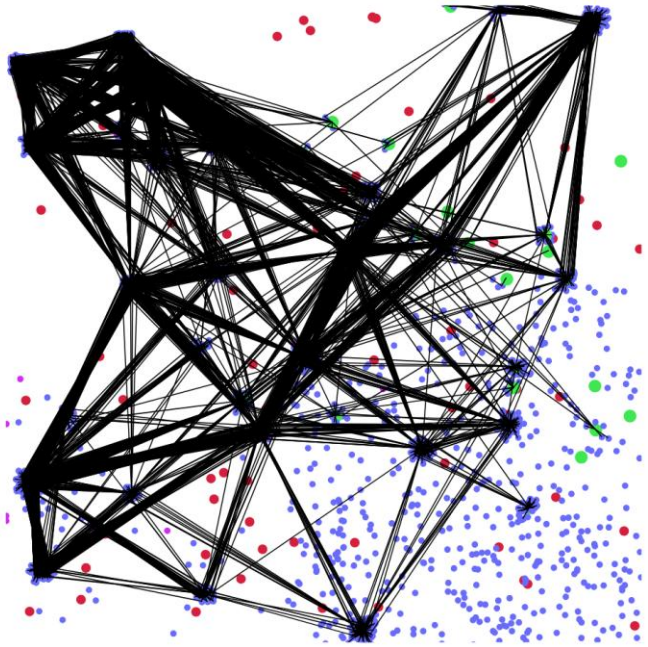
Bootcamp: Data Analytics
Nombre: Fernando Poblete Osses

www.linkedin.com/in/fernando-poblete-osses/

https://github.com/fernando-6561

Origen de los datos: modelados sintéticamente

Origen de Datos y Simulación



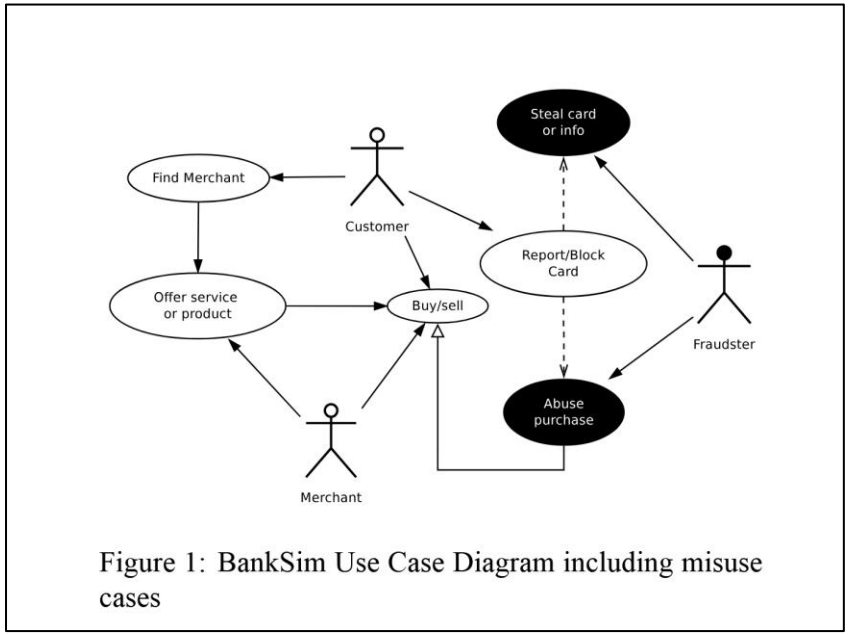


Figure 1: BankSim Use Case Diagram including misuse cases

Base de datos reales

Los datos sintéticos se derivan de un dataset real español del período 2012-2013, proporcionando una base sólida para simulaciones.

Simulación MABS

La simulación MABS incluye agentes como clientes y comercios, permitiendo escenarios realistas de interacción y transacción.

Privacidad y Control

El uso de datos sintéticos garantiza la privacidad, facilita el control de variables y permite una generación de datos segura y escalable.

Detección de fraude

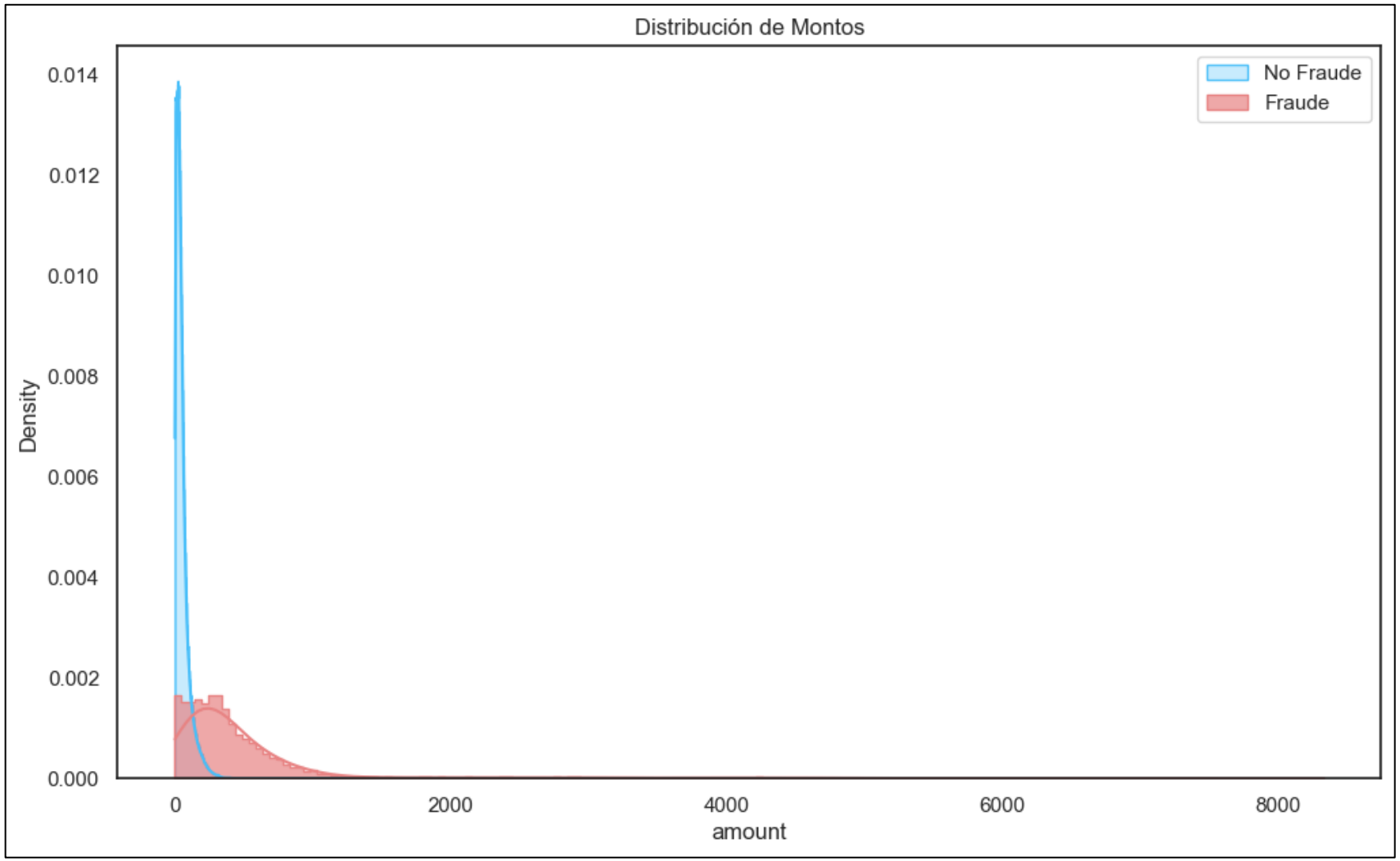
Este enfoque es ideal para crear datos controlados y seguros, especialmente útiles en la detección de fraude bancario.

Base de Datos

	step	customer	age	range	gender	merchant	category	amount	fraud
	Missing: 0% Distinct: <1%	Missing: 0 (0%) Disti... 4112 (<1%)	Missing: 0 (0%) Distinct: 8 (<1%)	Missing: 0 (0%) Distinct: 8 (<1%)	Missing: 0 (0%) Distinct: 4 (<1%)	Missing: 0 (0%) Distinct: 50 (<1%)	Missing: 0 (0%) Distinct: 15 (<1%)	Missing: 0 (0%) Distinct: 23767 (4%)	Missing: 0 (0%) Distinct: 2 (<1%)
	180 Distinct values	4112 Distinct values	2 31% 3 25% 4 18% Other 25%	26-35 31% 36-45 25% 46-55 18% Other 25%	F 55% M 45% E <1% Other <1%	M18230726... 50% M348934600 35% M85975013 4% Other 11%	es_transportation 85% es_food 4% es_health 3% Other 8%	Min 0.0 Max 8329.96	0 99% 1 1%
0	0	C1093826151	4	46-55	M	M348934600	es_transportation	4.55	0
1	0	C352968107	2	26-35	M	M348934600	es_transportation	39.68	0
2	0	C2054744914	4	46-55	F	M1823072687	es_transportation	26.89	0
3	0	C1760612790	3	36-45	M	M348934600	es_transportation	17.25	0
4	0	C757503768	5	56-65	M	M348934600	es_transportation	35.72	0
5	0	C1315400589	3	36-45	F	M348934600	es_transportation	25.81	0
6	0	C765155274	1	19-25	F	M348934600	es_transportation	9.1	0
7	0	C202531238	4	46-55	F	M348934600	es_transportation	21.17	0
8	0	C105845174	3	36-45	M	M348934600	es_transportation	32.4	0
9	0	C39858251	5	56-65	F	M348934600	es_transportation	35.4	0
10	0	C98707741	4	46-55	F	M348934600	es_transportation	14.95	0
11	0	C1551465414	1	19-25	M	M1823072687	es_transportation	1.51	0
12	0	C623601481	3	36-45	M	M50039827	es_health	68.79	0

Análisis descriptivo de los datos

Monto de transacciones



Distribución

Tanto para transacciones que son fraude y no lo son, las distribuciones no son normales

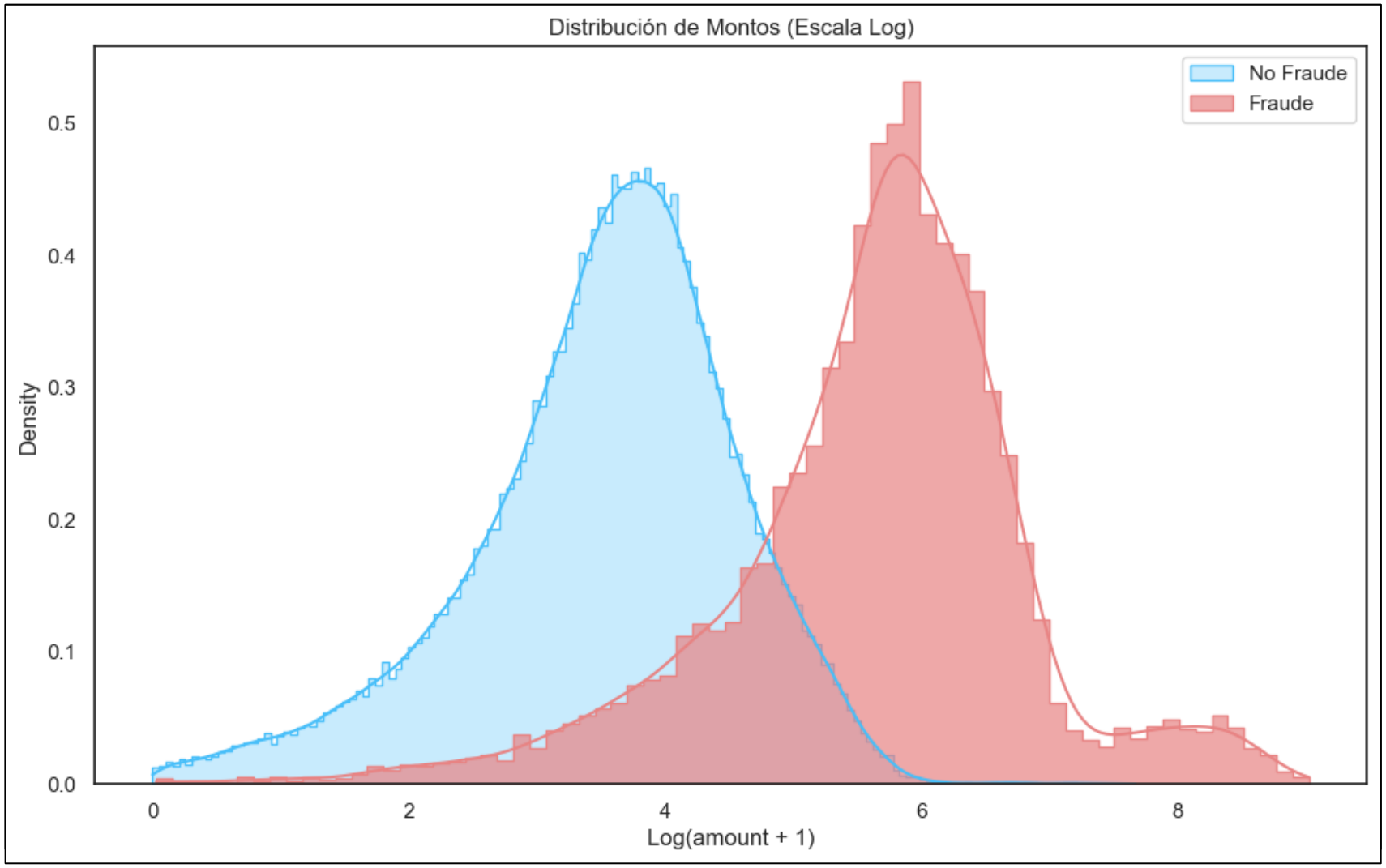
Fraudes

Tienen un rango más amplio de valores, teniendo una distribución de cola larga

No Fraudes

Se concentran en montos de menor valor

Monto de transacciones



Distribución

Tanto para transacciones que son fraude y no lo son, las distribuciones no son normales

Fraudes

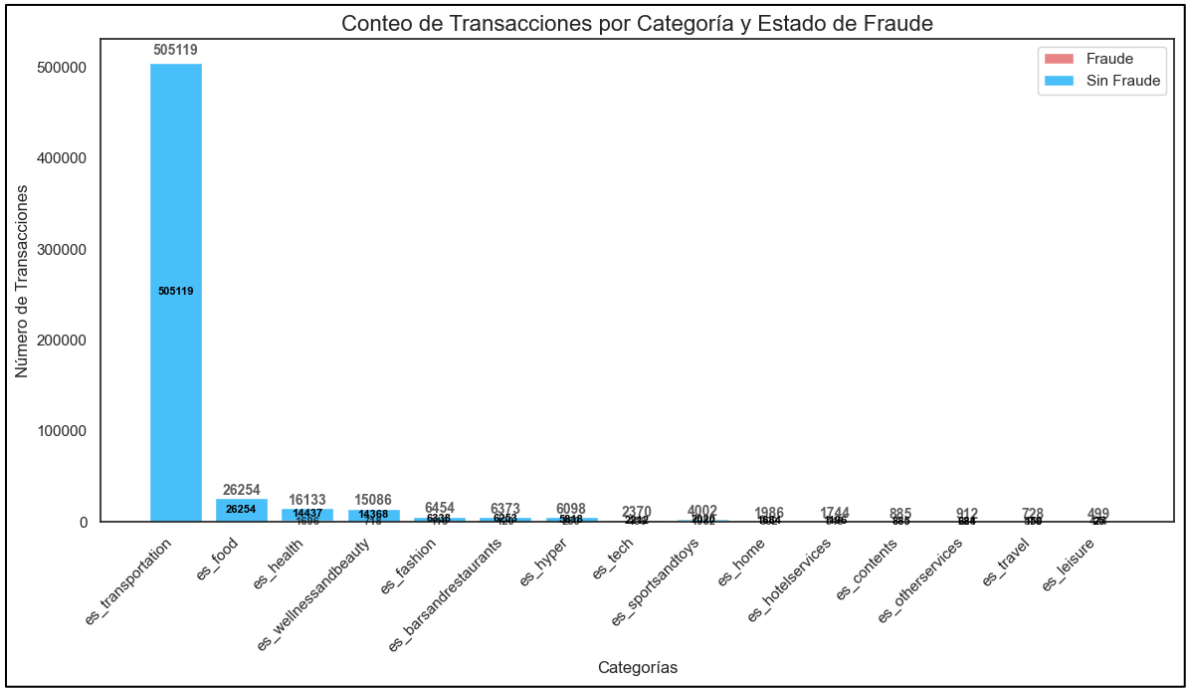
Tienen un rango más amplio de valores, teniendo una distribución de cola larga

No Fraudes

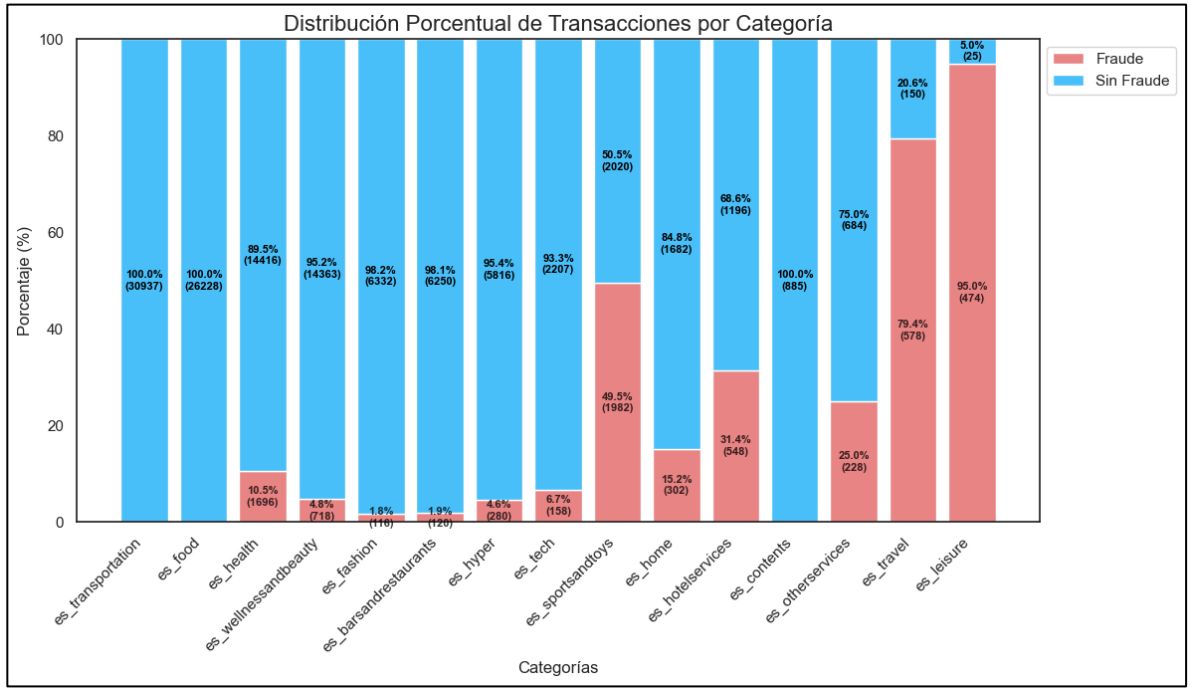
Se concentran en montos de menor valor

Segmentación por categorías de compra ✓

- **Hay un sobre muestreo de la categoría ‘transportation’**

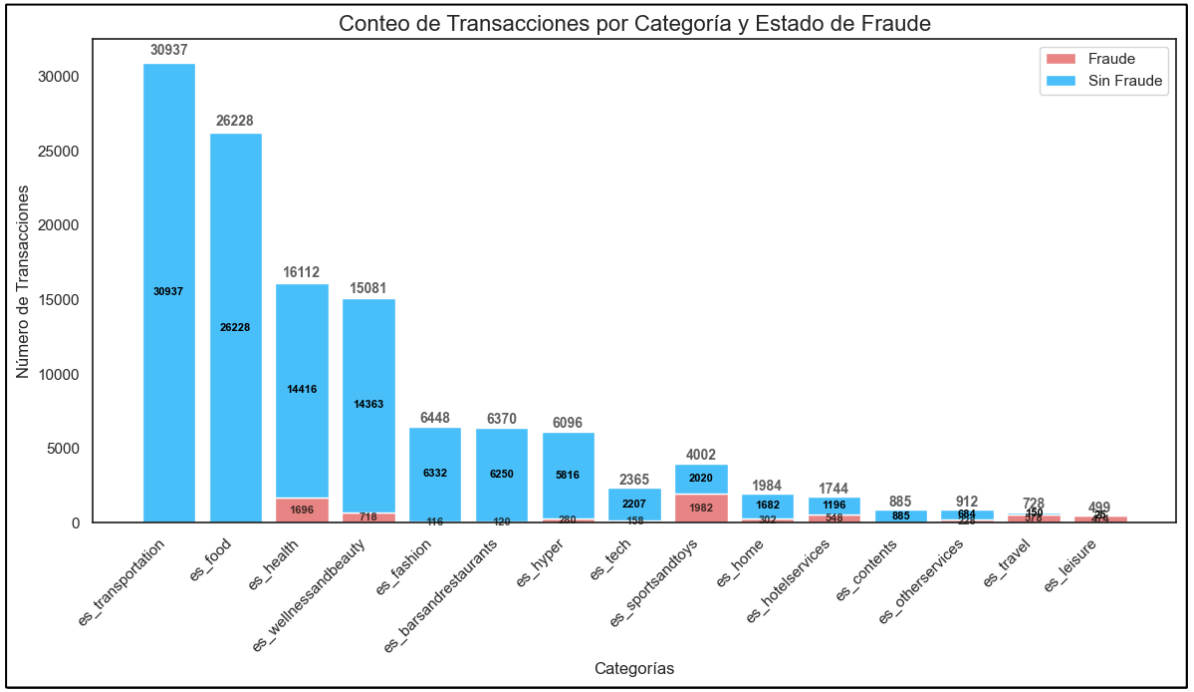


- **Existen categorías de productos que presentan mayores fraudes**

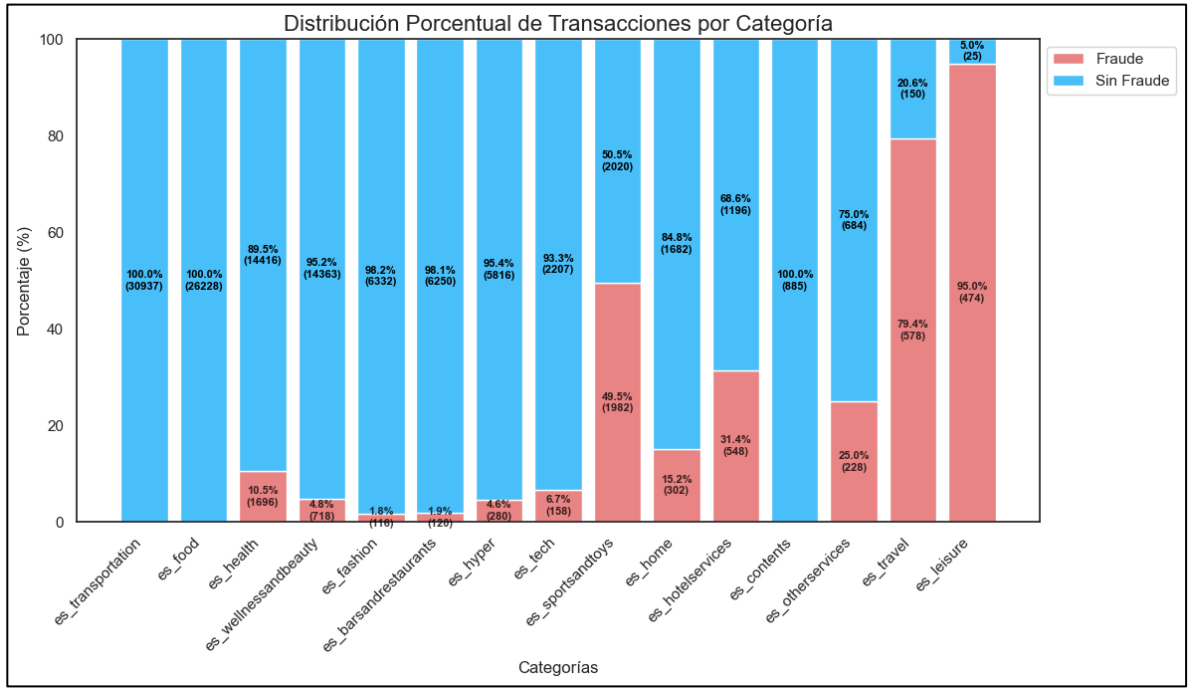


Segmentación por categorías de compra ✓

- **Hay un sobre muestreo de la categoría ‘transportation’**

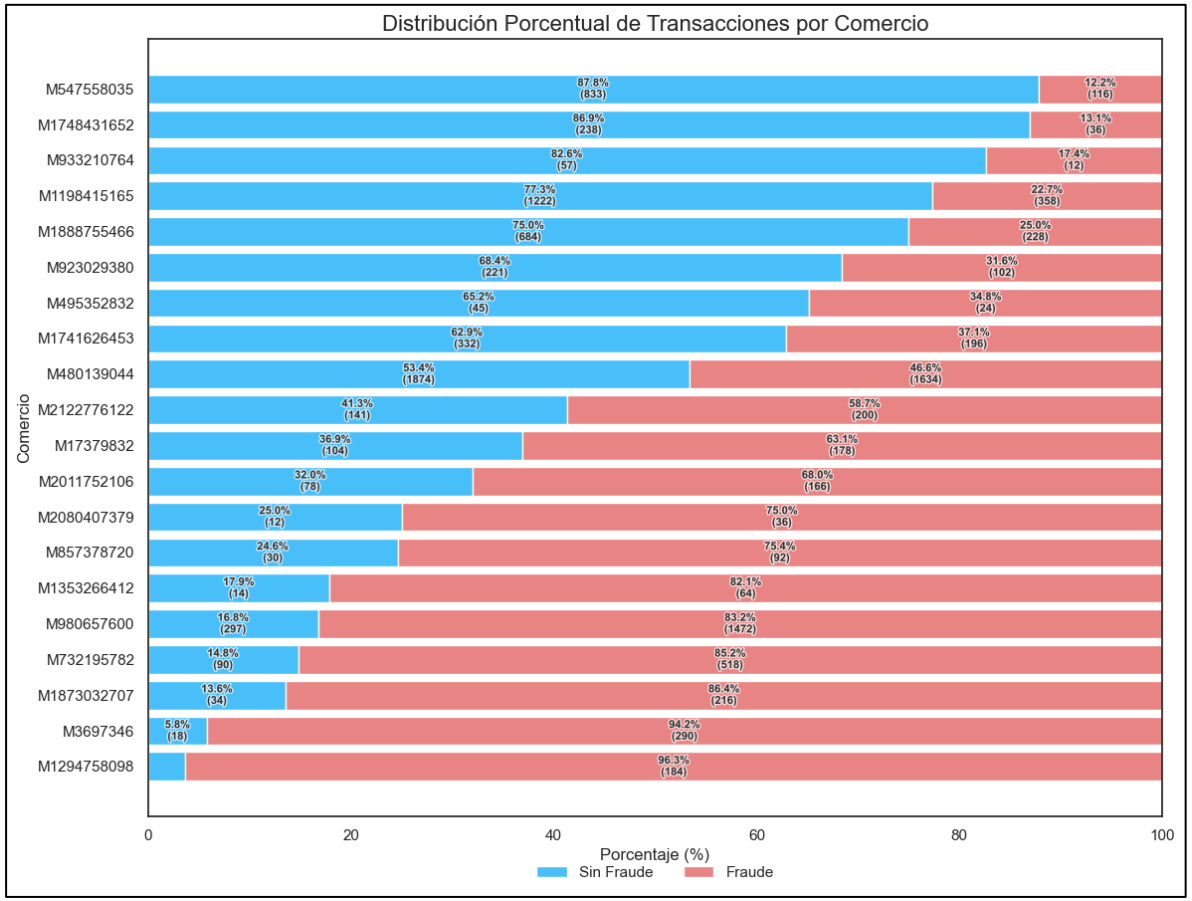


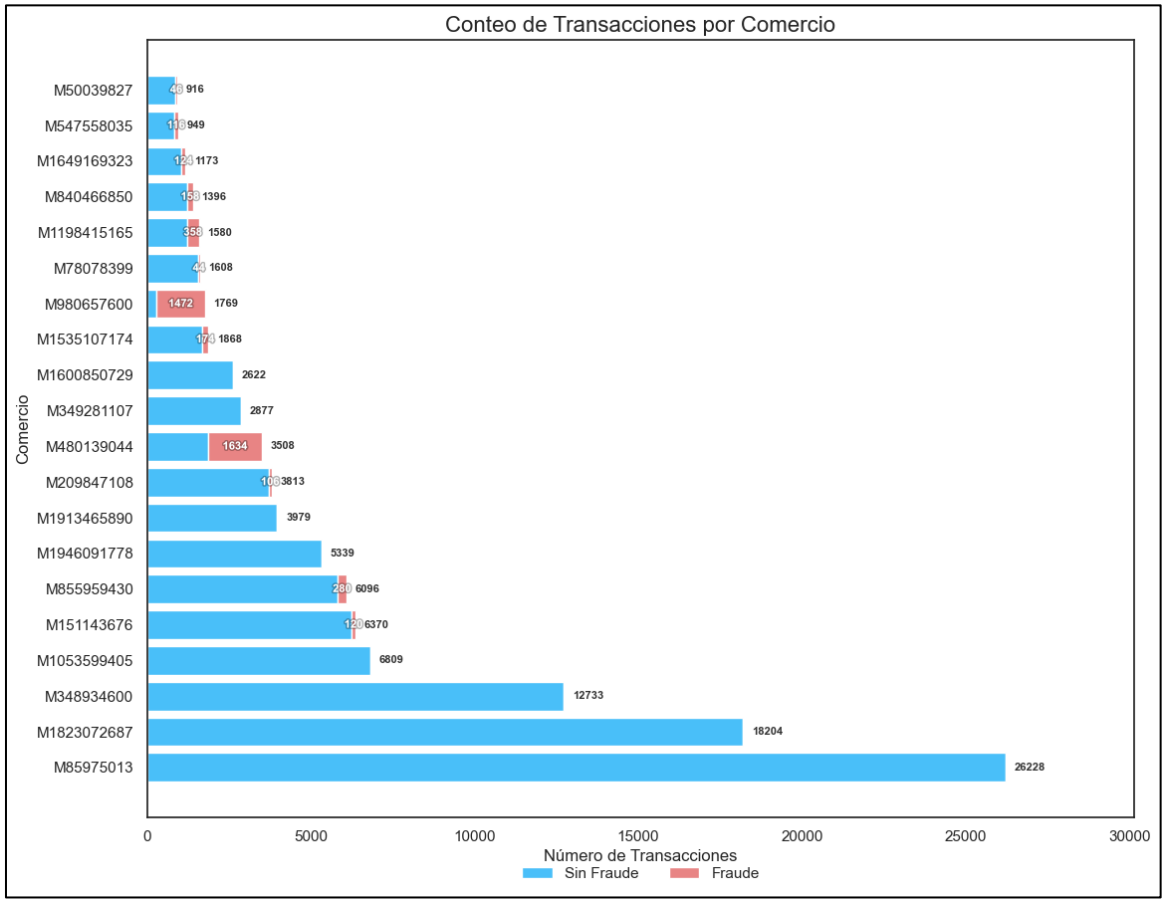
- **Existen categorías de productos que presentan mayores fraudes**



Segmentación por comercios ✓

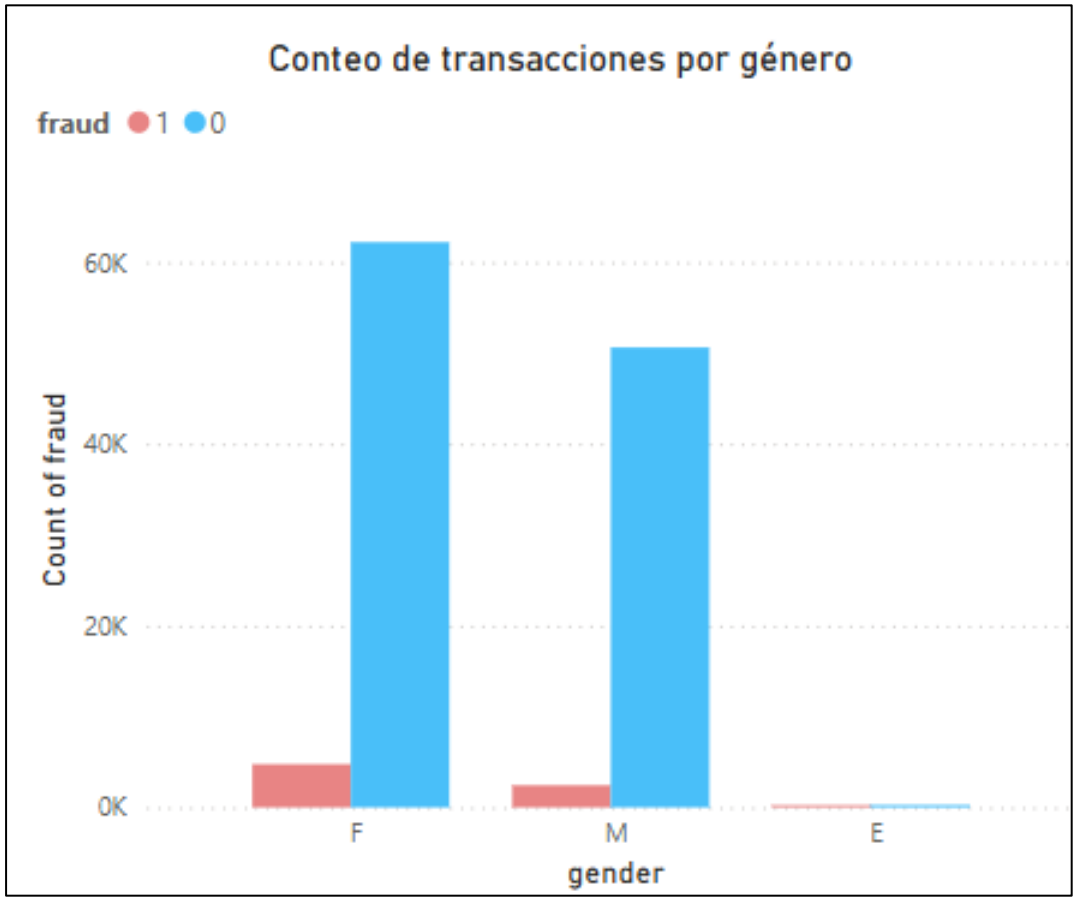
- **Transacciones y Fraudes concentradas por comercio**





Segmentación por género ✓

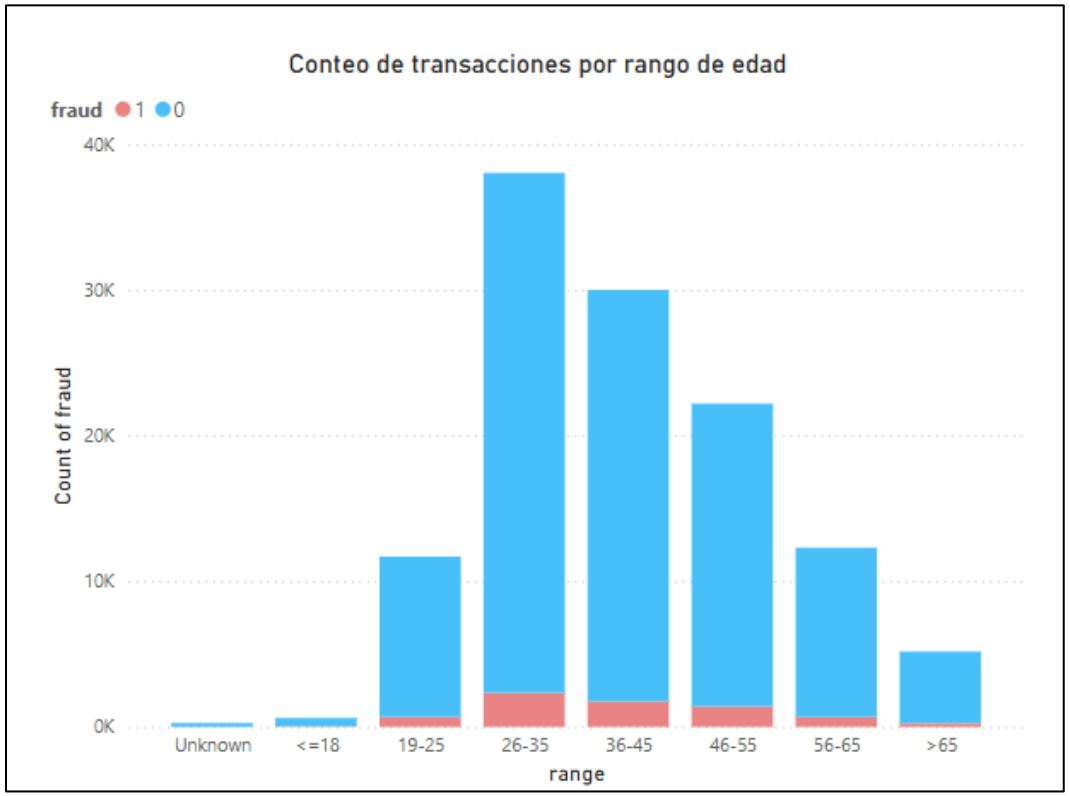
- **No se observan grandes diferencias de transacciones fraudulentas por género en la muestra**



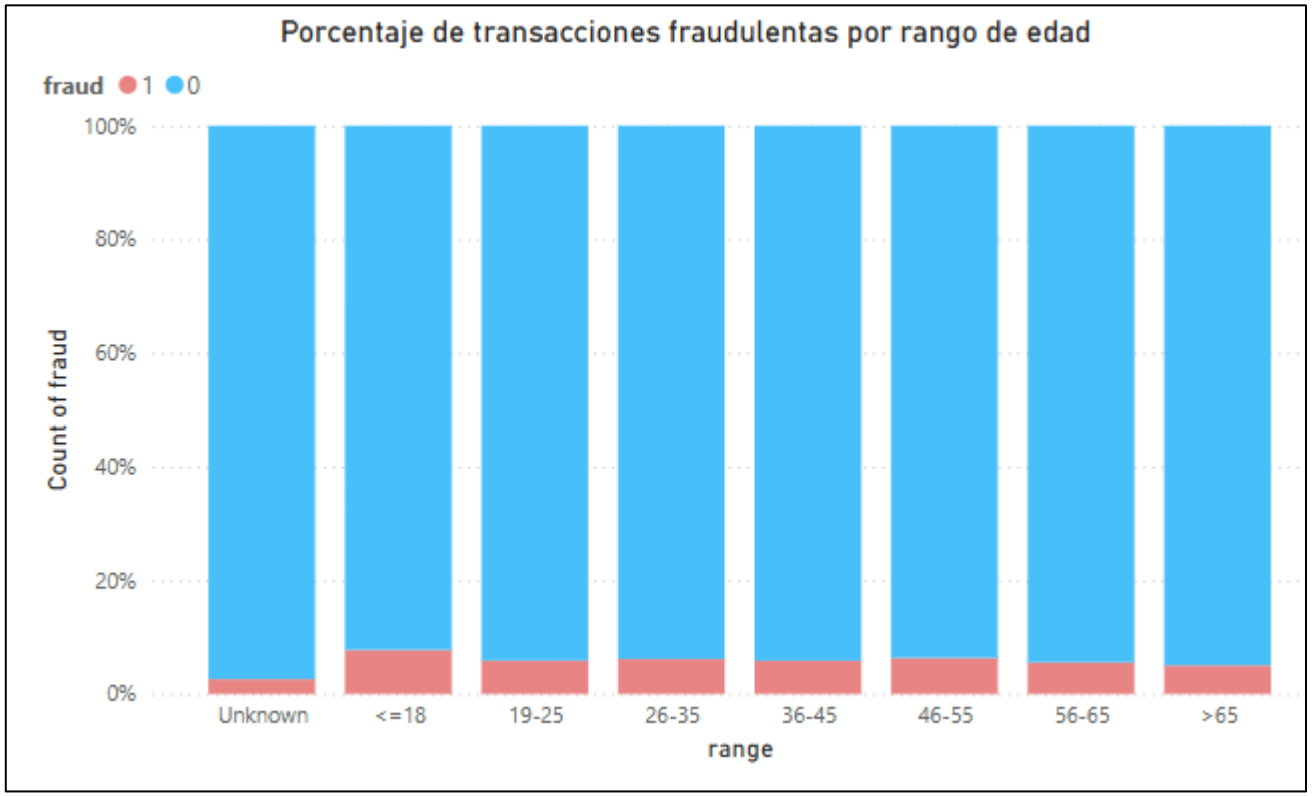
- **Se excluyó del análisis a las personas con género no identificado, ya que no contaban con ninguna transacción con fraude, debido al tamaño del grupo en la simulación**

Segmentación por rango de edades ✓

- El muestro es desbalanceado por rango de edad



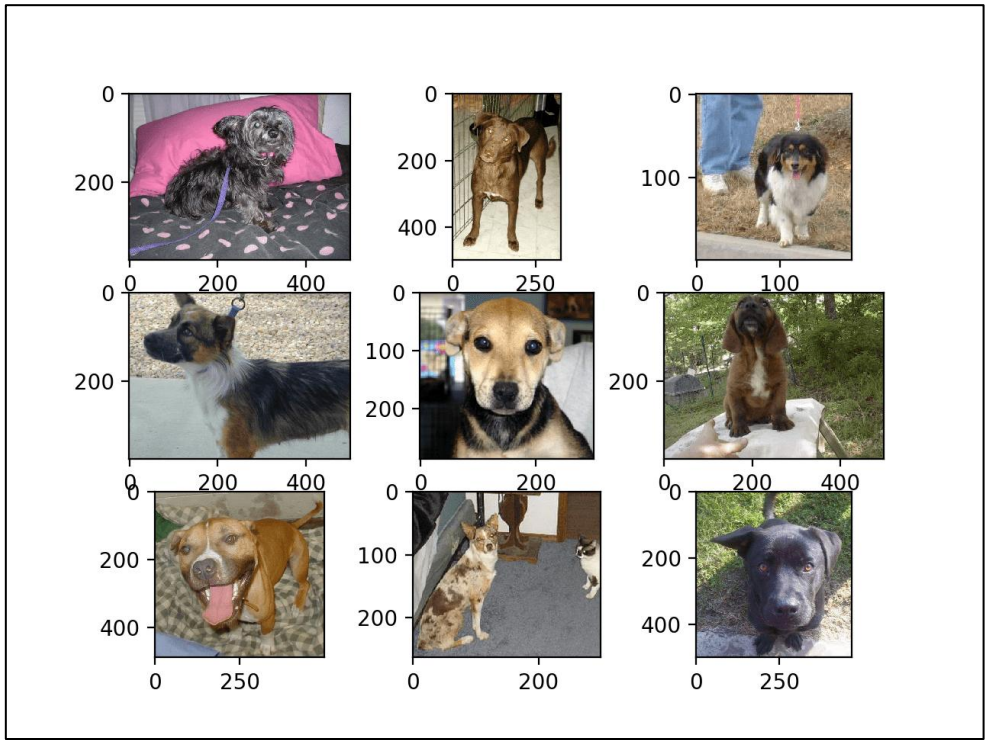
- No se observan grandes diferencias de porcentaje de transacciones fraudulentas por rango de edad

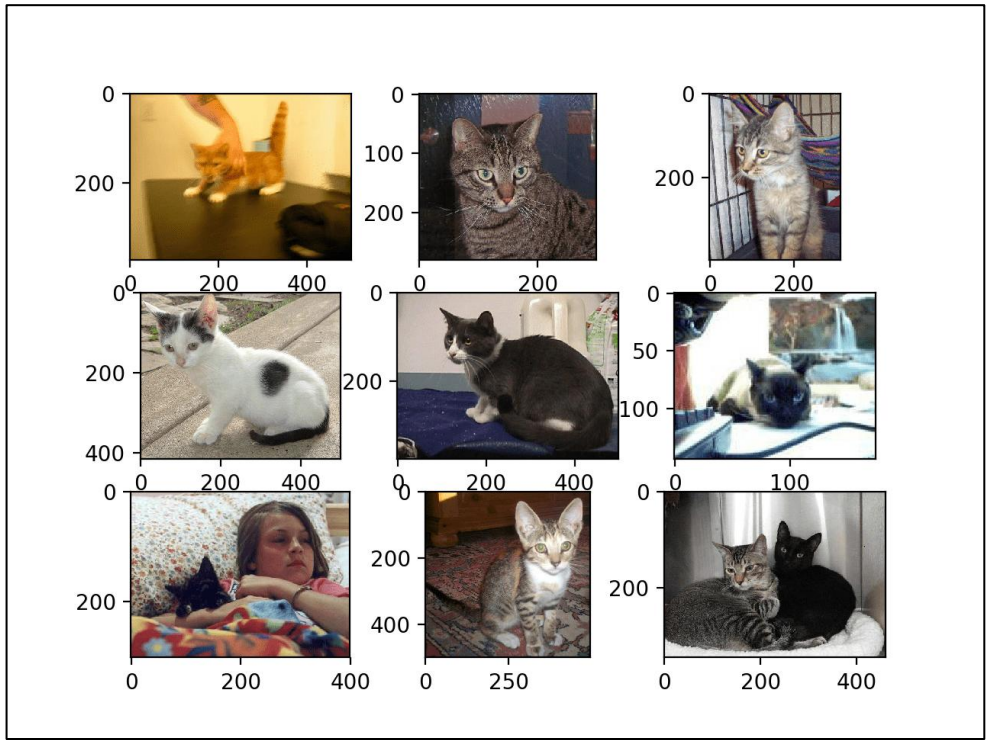


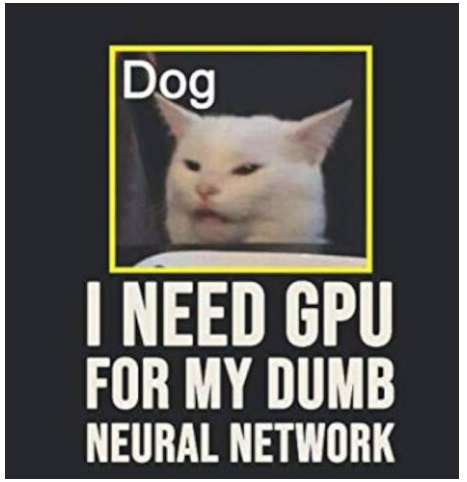
Creación y comparación de modelos de clasificación

¿Qué es un Modelo de Clasificación?

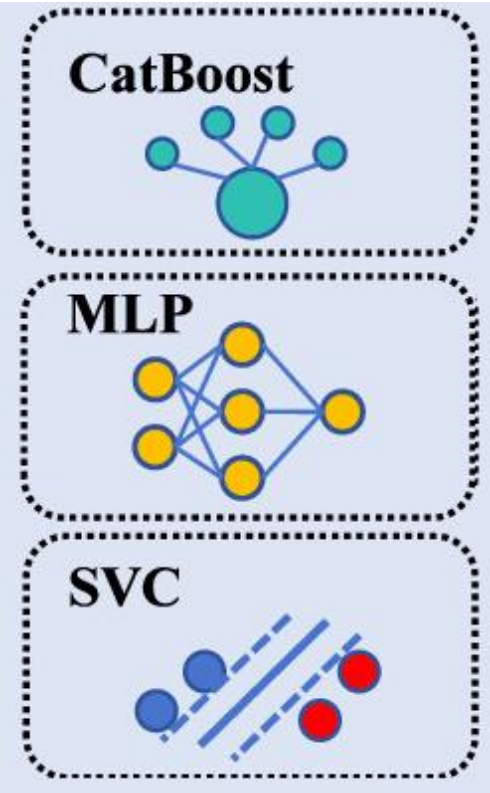
- **Datos etiquetados entrenan el modelo:** Se le dan ejemplos que ya se sabe qué grupo pertenecen
- **El modelo aprende patrones:** Descubre cómo son las características de cada grupo
- **Aprende a clasificar nuevos datos:** Luego puede decir a qué grupo pertenece información nueva que no ha visto







Selección de algoritmos de clasificación



CatBoost - Tipo: Gradient Boosting optimizado

▶ Automatiza variables categóricas (sin preprocesamiento manual).

▶ Rápido con GPU.

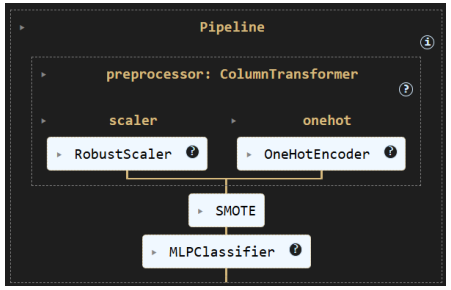
MLP (Multi-Layer Perceptron) - Tipo: Red Neuronal básica

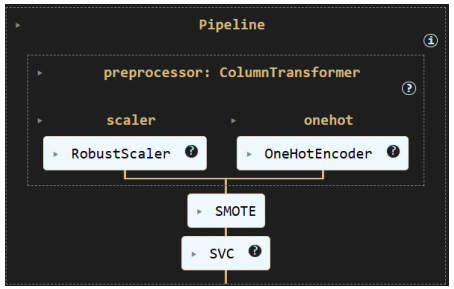
▶ Capas ocultas no lineales (aprende patrones complejos).

▶ Requiere muchos datos y ajuste fino.

SVC (Support Vector Classification) - Tipo: Clasificador de fronteras (SVM).

▶ Busca el hiperplano óptimo (maximiza margen entre clases).





¿Qué es una Matriz de Confusión?

Definición

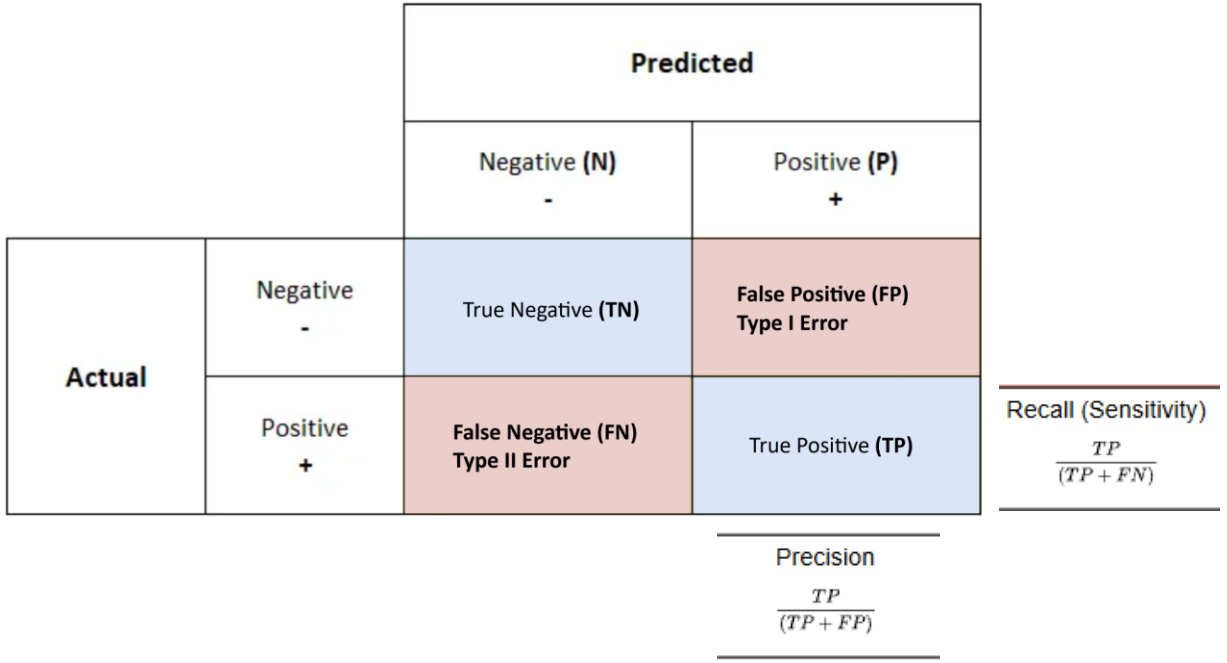
Una matriz de confusión es una herramienta visual usada para evaluar el rendimiento de un modelo de clasificación en aprendizaje automático.

Componentes Clave

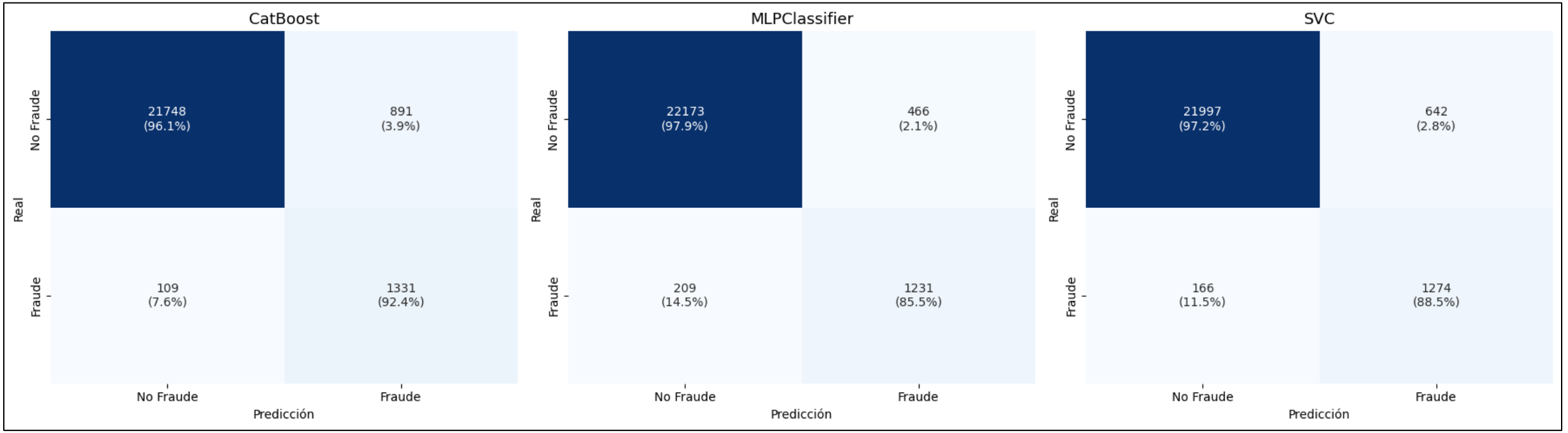
Contiene cuatro elementos: VP (Verdaderos Positivos), FP (Falsos Positivos), FN (Falsos Negativos) y VN (Verdaderos Negativos).

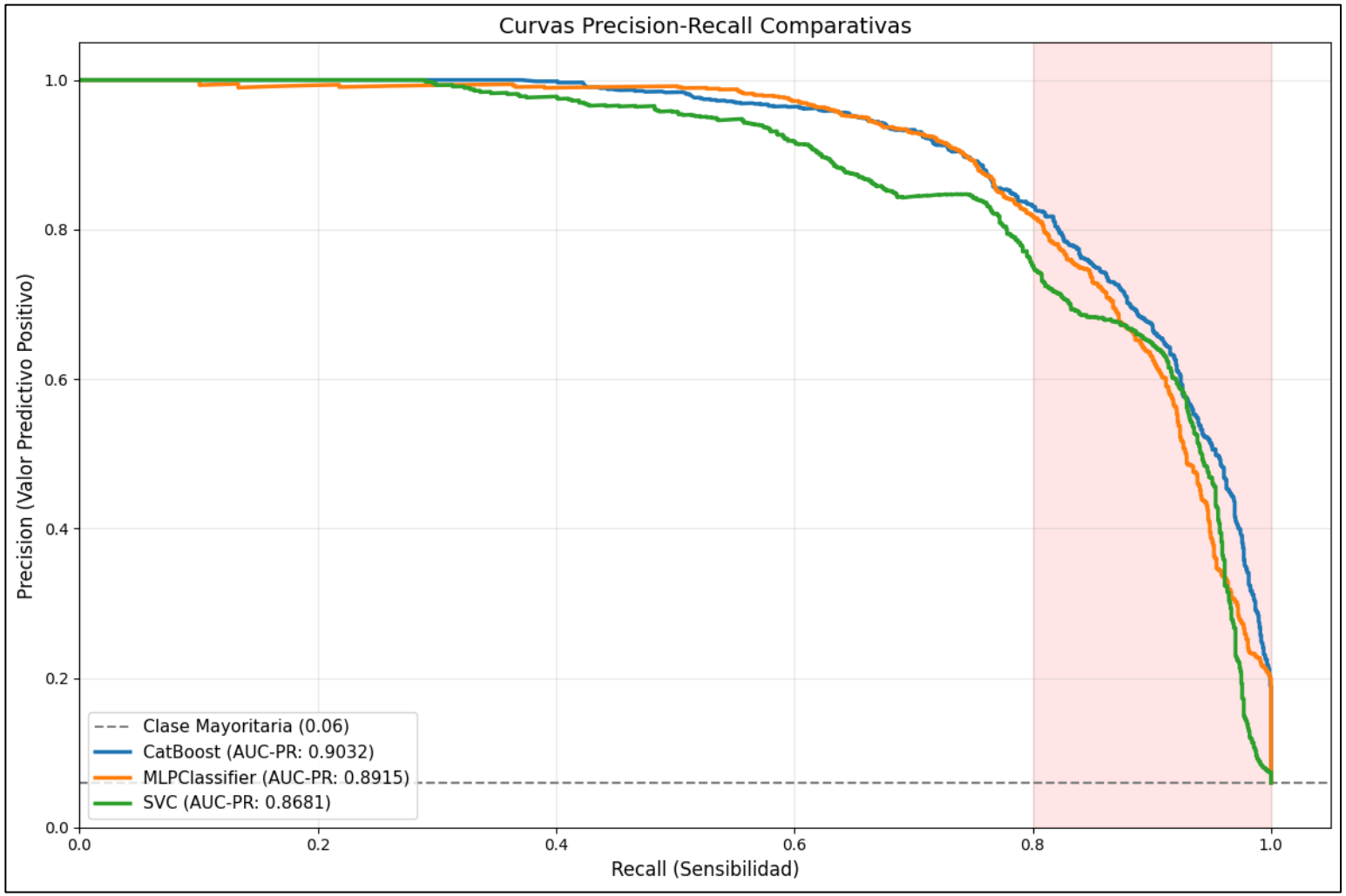
Importancia de FN

FN, que implica un fraude no detectado, puede tener un costo alto, subrayando la importancia de su minimización en modelos de clasificación.



Comparación de rendimiento y métricas





- **El umbral de clasificación** determina el equilibrio entre **precisión** y **recuperación (recall)**
- **Umbral más alto (más estricto):**
  - El modelo necesita **más confianza** para predecir la clase positiva.
  - **Aumenta la precisión:** Menos falsos positivos (menos casos negativos incorrectamente clasificados como positivos).
  - **Disminuye la recuperación:** Más falsos negativos (más casos positivos incorrectamente clasificados como negativos).
- **Relación fundamental:**
  - Existe una **contrapartida (trade-off)** entre precisión y recuperación al ajustar el umbral.
  - Esto se debe a que el umbral **modifica la rigurosidad** del modelo para clasificar un caso como positivo.

Conclusiones y recomendaciones

Principales hallazgos del análisis

- **Contexto del Problema**
- **Tarea:** Clasificación binaria (transacción legítima vs. fraudulenta)
- **Desafío:** Desbalanceo extremo de clases (muy pocos fraudes)

Criterio	CatBoost	MLP	SVC
Velocidad de entrenamiento	Muy rápida	Moderada	Muy lenta
Rendimiento general	Superior tras ajustes	Similar a otros	Similar a otros
Manejo variables categóricas	Nativo (sin preprocesamiento)	Requiere codificación	Requiere codificación

Conclusión Final

- **Modelo recomendado:** CatBoost (por su velocidad, rendimiento y eficiencia tras ajustes)
- **Ventaja decisiva:** Menor tiempo de entrenamiento/predicción + manejo óptimo de datos categóricos
- **Consideración práctica:** Trade-off entre precisión y recall debe alinearse con la estrategia antifraude (ej: priorizar seguridad vs. experiencia del usuario)

Preguntas

IT ACADEMY

Bootcamp: Data Analytics
Nombre: Fernando Poblete Osses

www.linkedin.com/in/fernando-poblete-osses/

https://github.com/fernando-6561