



Grupo 6

INSURANCE COMPANY

Análisis de Fraude en
Seguros de Automóviles





PROBLEMA



Los fraudes generan pérdidas económicas significativas para las aseguradoras, por lo que es importante desarrollar modelos que analicen patrones en los reclamos para diferenciarlos de aquellos legítimos.

SOLUCIÓN

Desarrollar un modelo predictivo con técnicas de Machine Learning.

Analizar los datos históricos de reclamos, identificando patrones asociados al fraude. A través de la ingeniería de características y el uso de algoritmos como DecisionTreeClassifier, se optimiza la predicción de reclamos fraudulentos.

Emplear técnicas de interpretabilidad como SHAP para garantizar la transparencia y comprensión del modelo por parte de los usuarios.



METODOLOGÍA

ANÁLISIS EXPLORATORIO

15,420 mil

Cientes

19

Modelos Vehículos

VehicleCategory	Valor	Accidents	Female	Male
Sedan	62,72%	Urban	2227	11595
Sport	34,75%	Rural	193	1405
Utility	2,54%	Total	2420	13000

Variable: Make	Count	Percentage
Make		
Pontiac	3837	24.9
Toyota	3121	20.2
Honda	2800	18.2
Mazda	2354	15.3
Chevrolet	1681	10.9
Accura	472	3.1
Ford	450	2.9
VW	283	1.8
Dodge	109	0.7
Saab	108	0.7
Mercury	83	0.5
Saturn	58	0.4
Nisson	30	0.2
BMW	15	0.1
Jaguar	6	0.0
Porche	5	0.0
Mecedes	4	0.0
Ferrari	2	0.0
Lexus	1	0.0

Variable: MonthClaimed	Count	Percentage
MonthClaimed		
Jan	1446	9.4
May	1411	9.2
Mar	1348	8.7
Oct	1339	8.7
Jun	1293	8.4
Feb	1287	8.3
Nov	1285	8.3
Apr	1271	8.2
Sep	1242	8.1
Jul	1225	7.9
Dec	1146	7.4
Aug	1126	7.3

AgeOfVehicle	Sedan	Sport	Utility	Total
7 years	22,56%	14,12%	0,99%	37,66%
more than 7	16,83%	7,91%	1,08%	25,82%
6 years	14,36%	7,69%	0,31%	22,36%
5 years	5,73%	2,98%	0,09%	8,80%
new	1,26%	1,10%	0,06%	2,42%
4 years	1,05%	0,43%	0,01%	1,49%
3 years	0,62%	0,36%		0,99%
2 years	0,31%	0,16%		0,47%
Total	62,72%	34,75%	2,54%	100,00%

Variable: MaritalStatus	Count	Percentage
MaritalStatus		
Married	10625	68.9
Single	4683	30.4
Divorced	76	0.5
Widow	35	0.2

LIMPIEZA DE DATOS

No hay nulos.

Elimina el registro 0 para la variable **DayOfWeekClaimed**.

DATOS

No existen duplicados

BasePolicy

Collision
Liability
All Perils

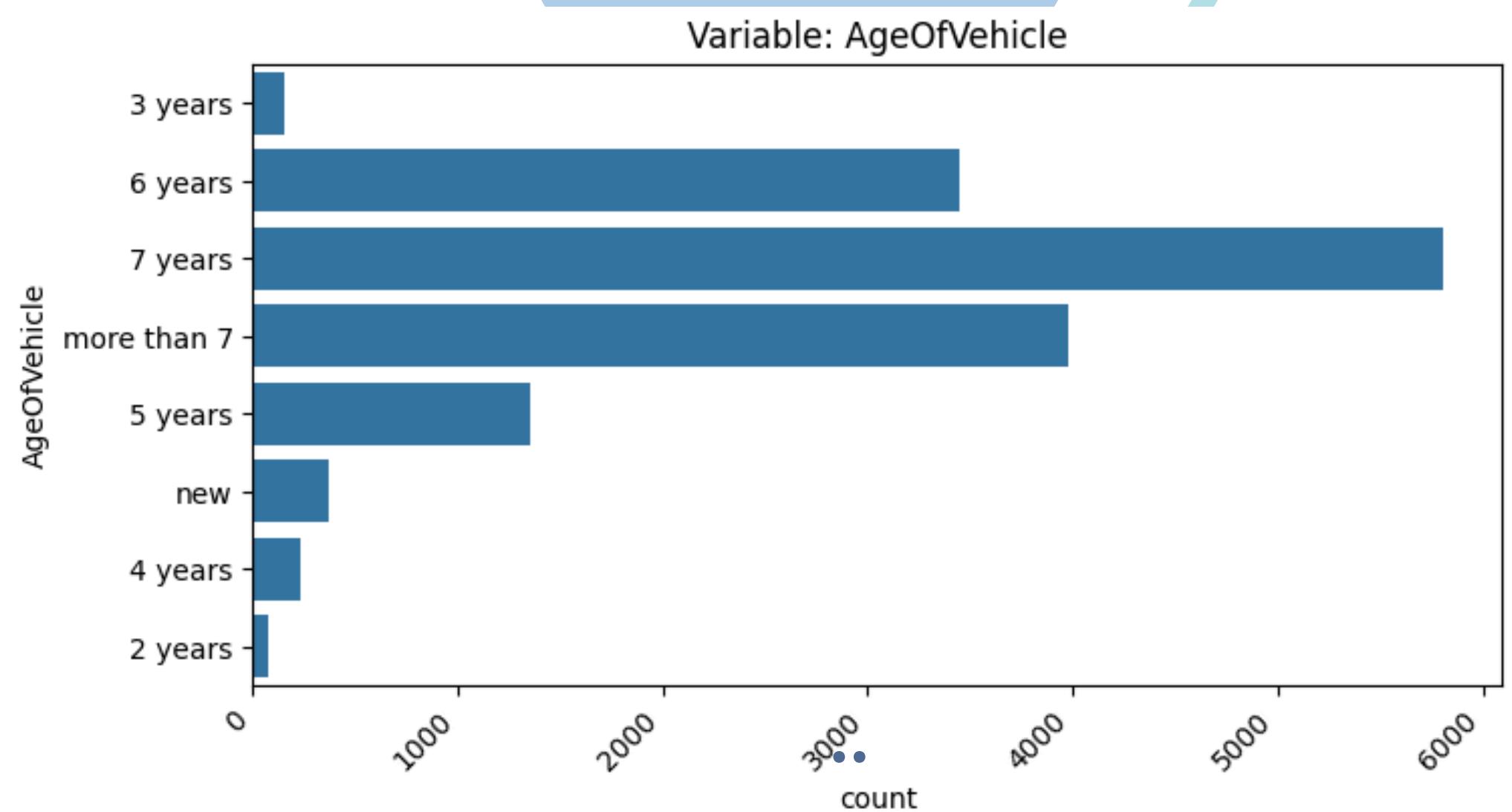
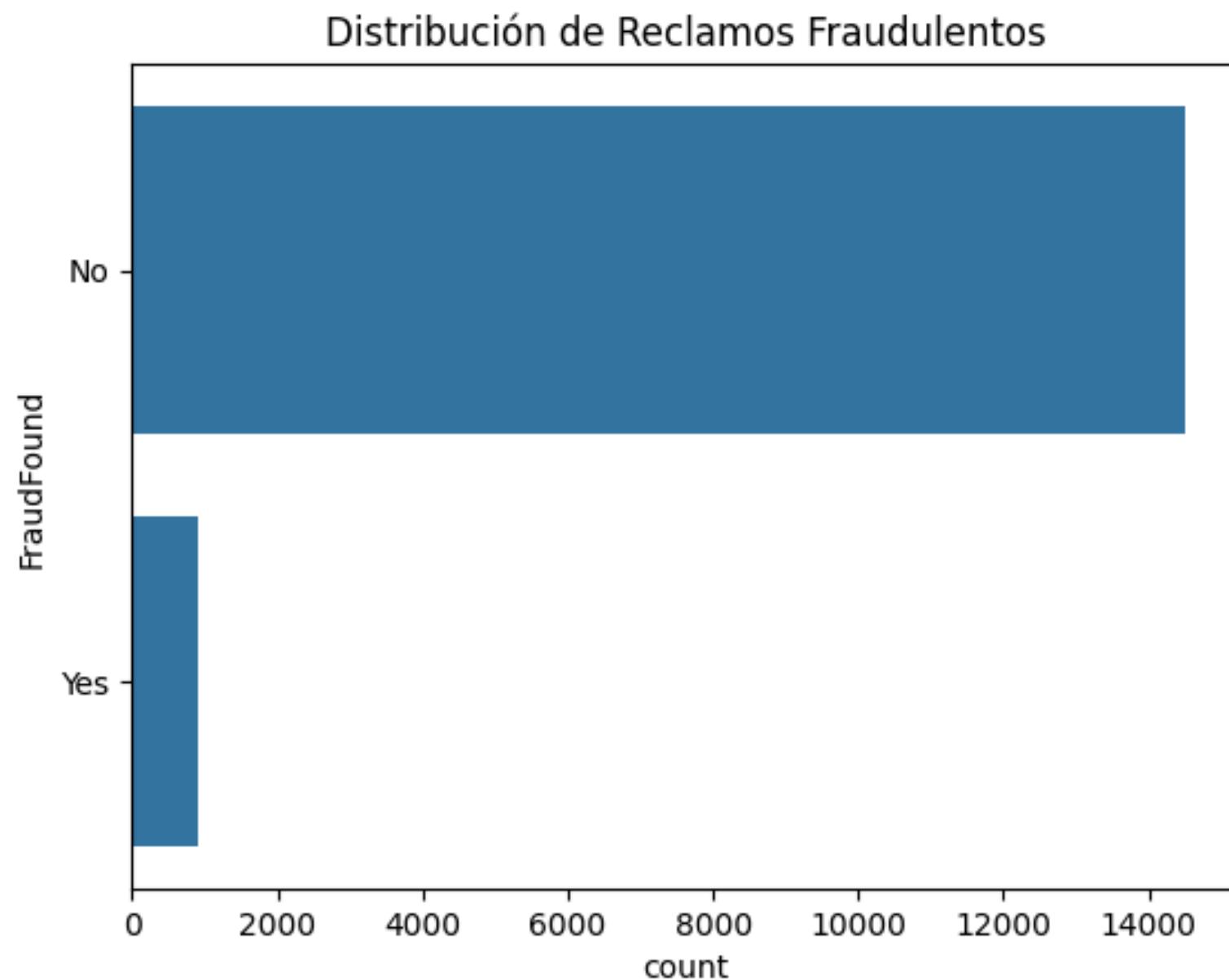
TARGET

Variable:	FraudFound	Count	Percentage
FraudFound	No	14496	94.0
	Yes	923	6.0

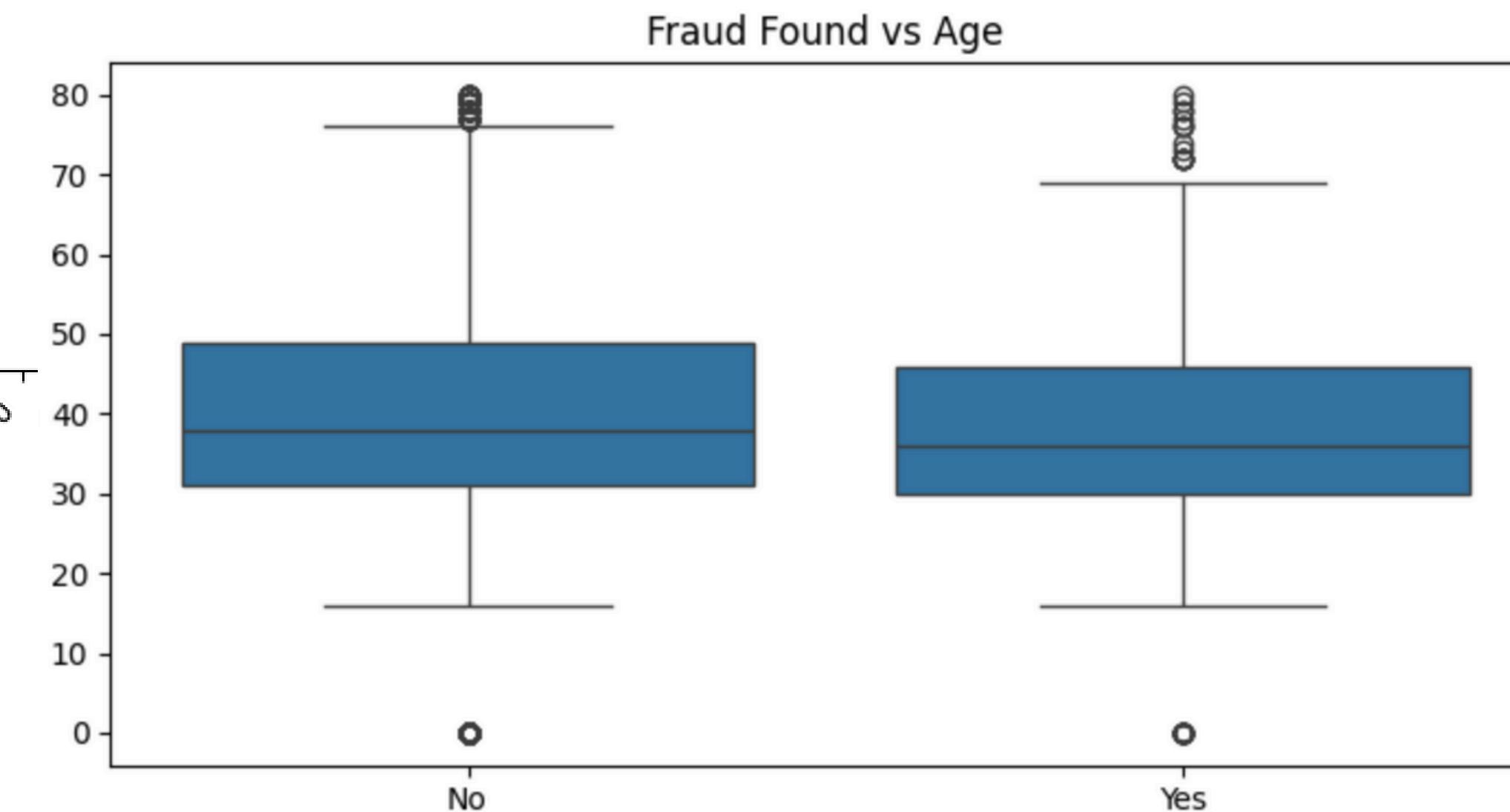
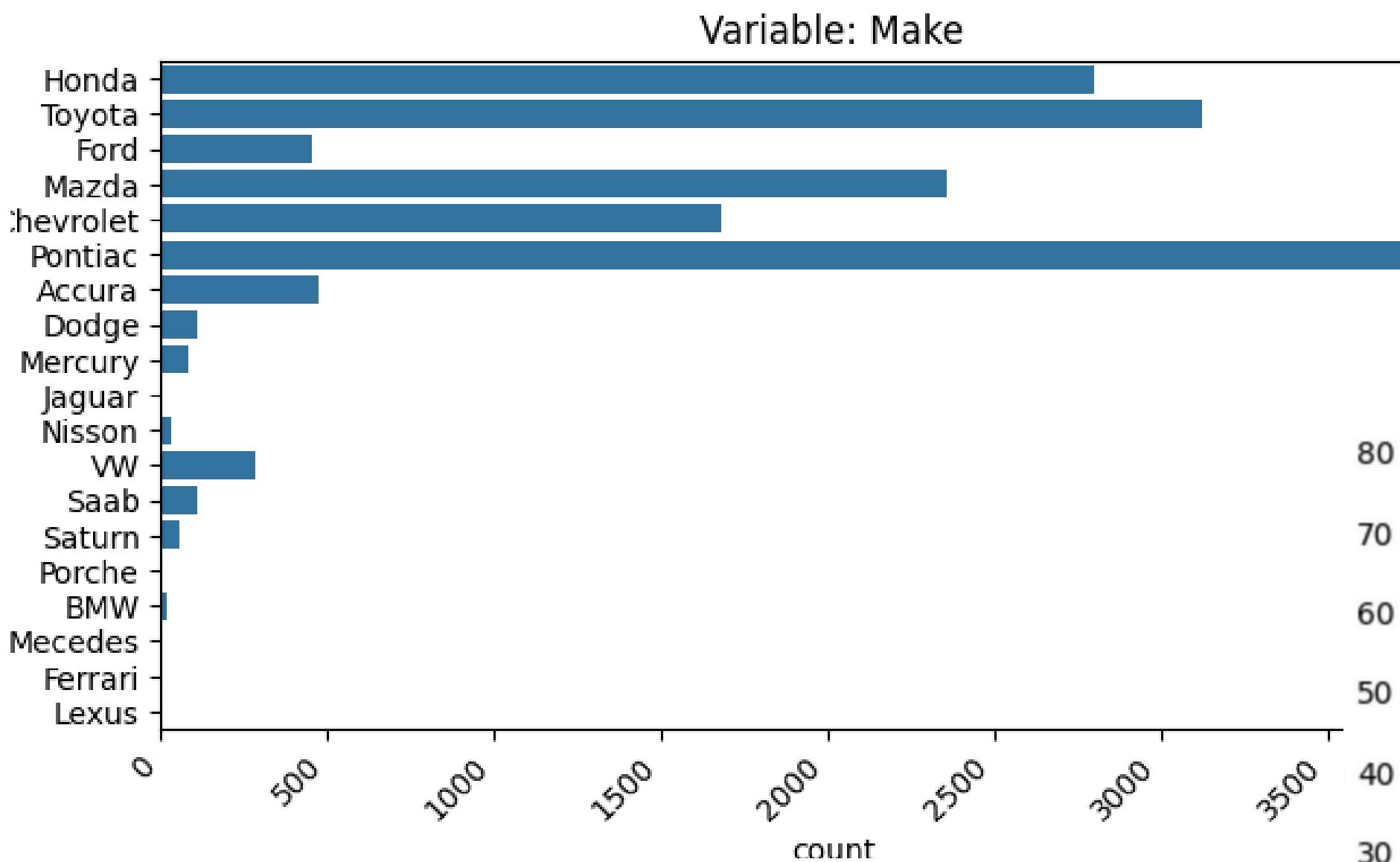
LIMPIEZA DE DATOS

Se elimina PolicyNumber, dado que es un número único que identifica la póliza de seguro.

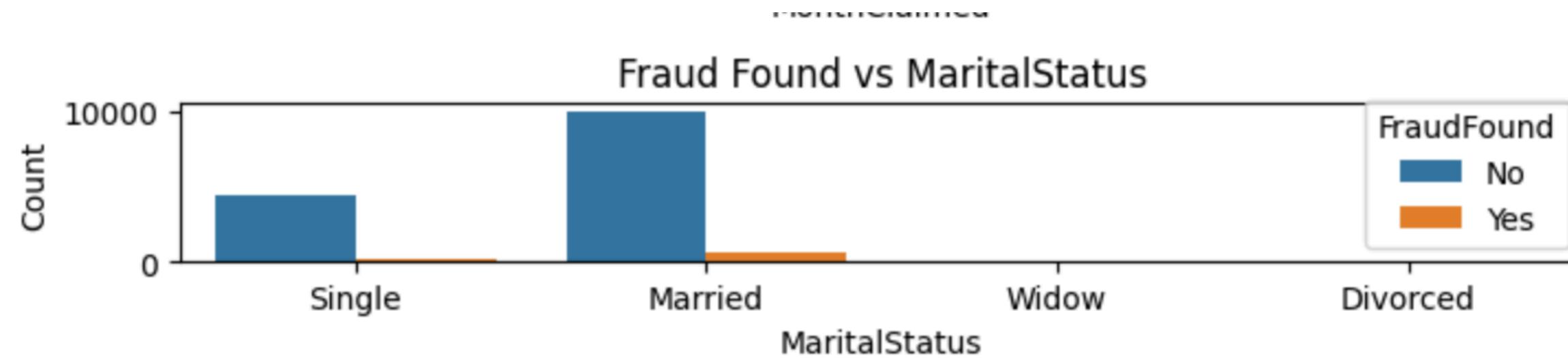
VISUALIZACION DE LOS DATOS



VISUALIZACION DE LOS DATOS



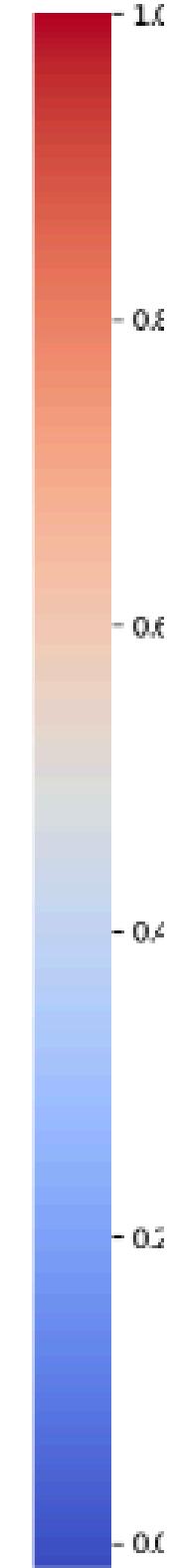
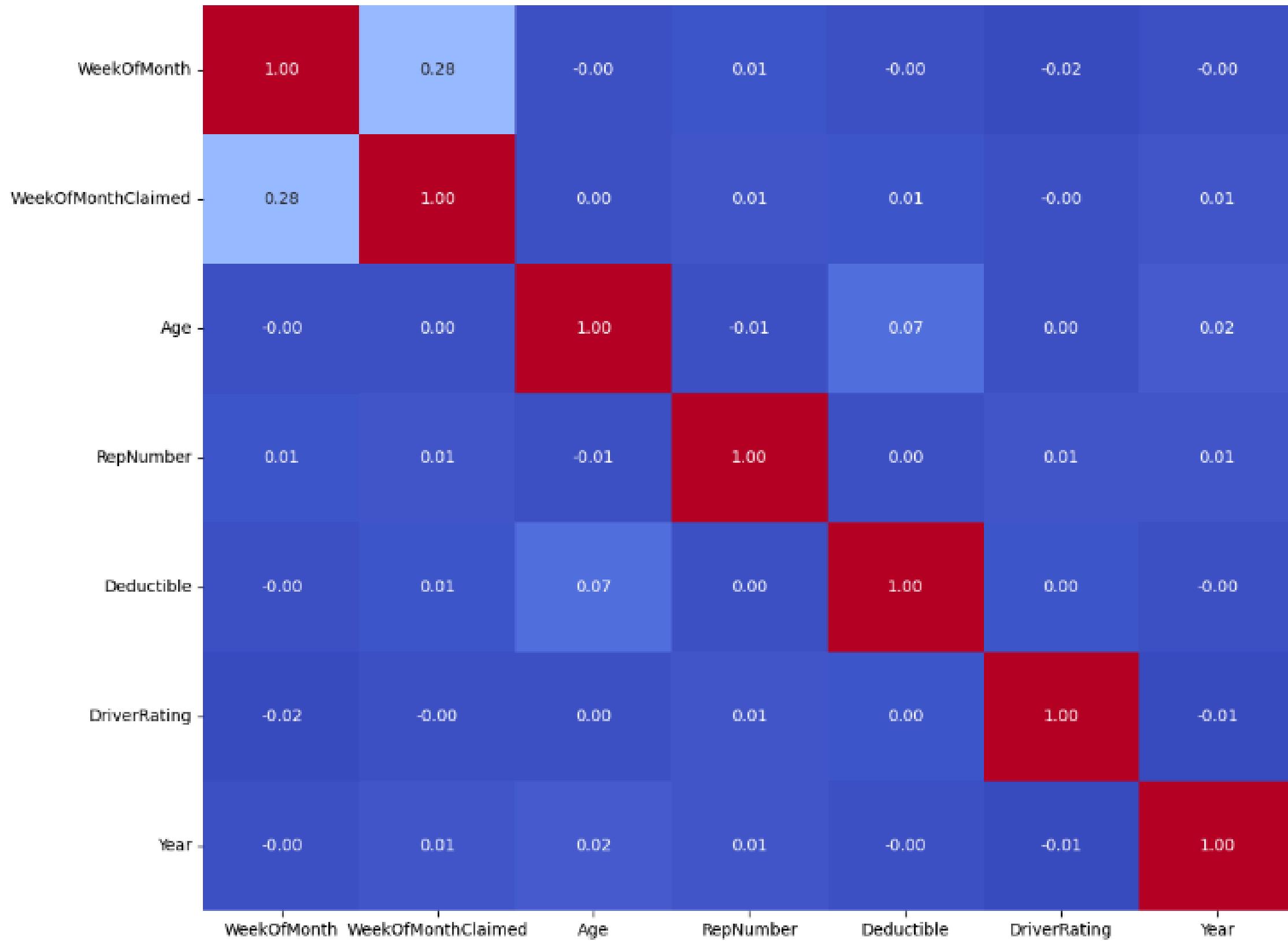
VISUALIZACION DE LOS DATOS



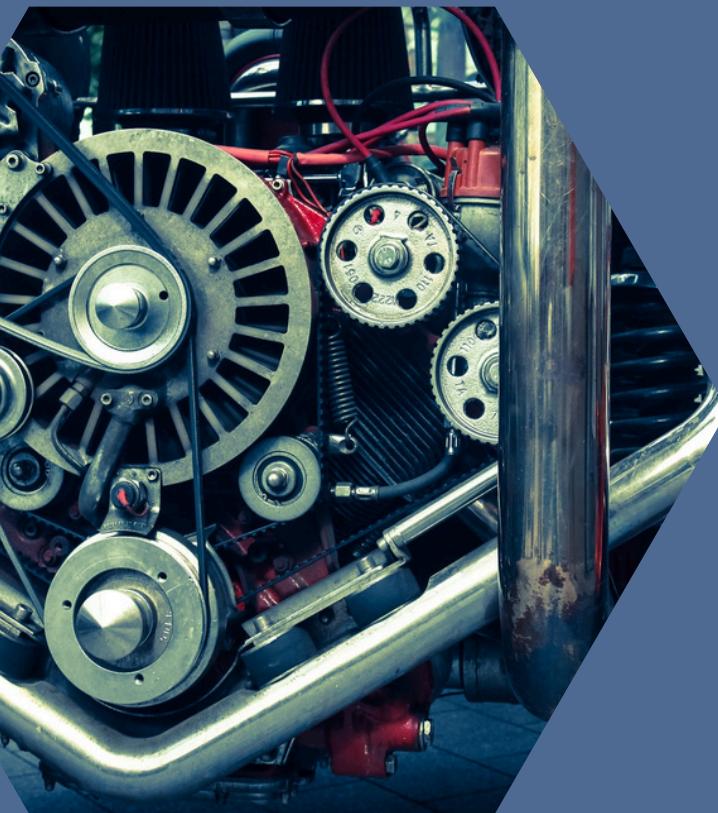
VISUALIZACION DE LOS DATOS



Matriz de Correlación de Variables Numéricas



RESAMPLING



Fraud Found

Desbalance significativo "No": 14496 vs 'Yes' :923)

Undersampling

+

Oversampling

'No': 4320 vs 'Yes':1384)



SPLITTING

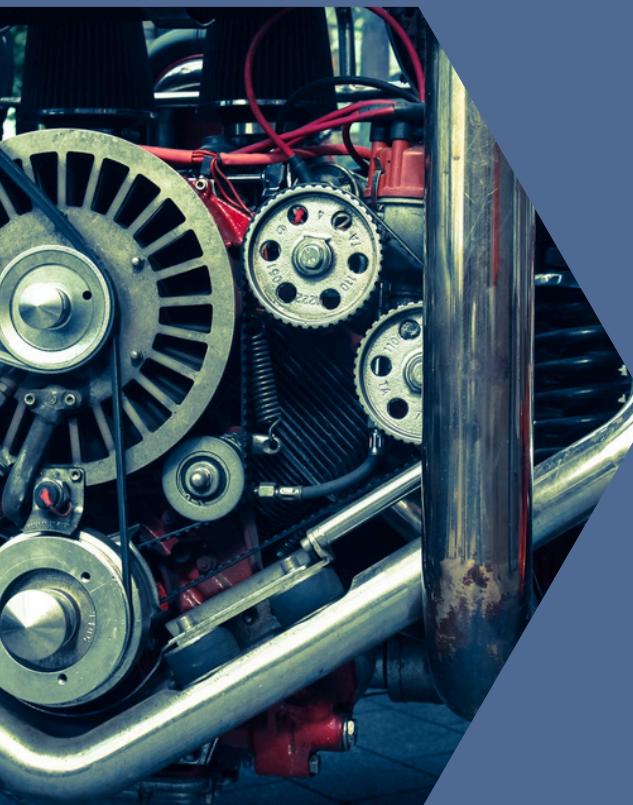
Del target: "Fraud found" se dividió los datos en:

- 70% entrenamiento
- 15% validación
- 15% test

FraudFound	proportion
No	0.714286
Yes	0.285714

PROCESAMIENTO DE DATOS

FEATURING INGENIERING



Duración entre accidente y reclamo

Month	MonthClaimed	AccidentToClaimDuration
Dec	Jan	1
Jan	Jan	0
Oct	Nov	1
Jun	Jul	1
Jan	Feb	1

Riesgo basado en la edad

Age	RiskLevel
21	Alto
34	Medio
47	Medio
65	Bajo
27	Alto

Combinación de variables categóricas
Marca y Categoría

Make	VehicleCategory	Make_VehicleCategory
Honda	Sport	Honda_Sport
Honda	Sport	Honda_Sport
Honda	Sport	Honda_Sport
Toyota	Sport	Toyota_Sport
Honda	Sport	Honda_Sport

PROCESAMIENTO DE DATOS

FEATURING INGENIERING

Marital Status
Nuevas clasificación

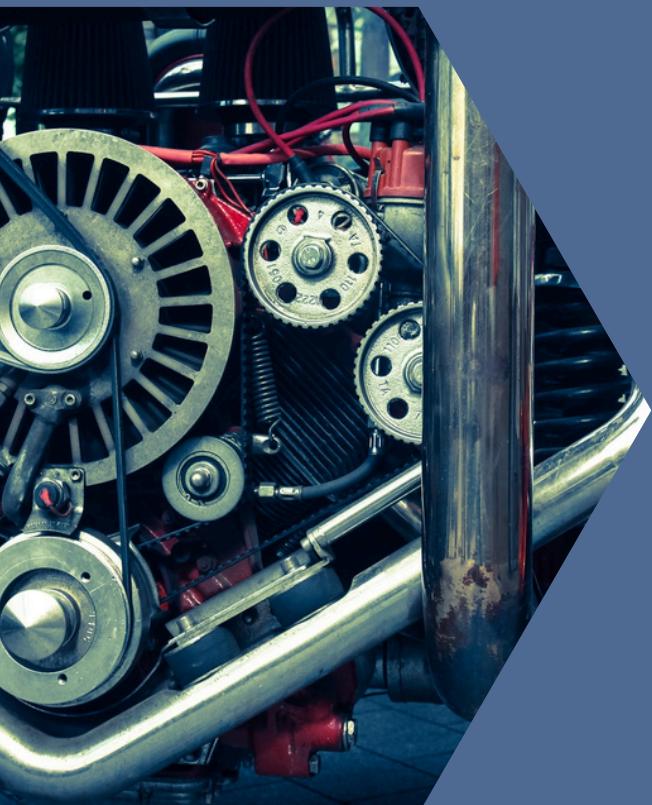
MaritalStatus
Single
Single
Married
Married
Single

Number of suppliments
Nuevas clasificación

NumberOfSuppliments
Bajo
Bajo
Bajo
Alto
Bajo

Marital Status
Nuevas clasificación

AddressChange-Claim
Reciente
No cambia



ENCODING

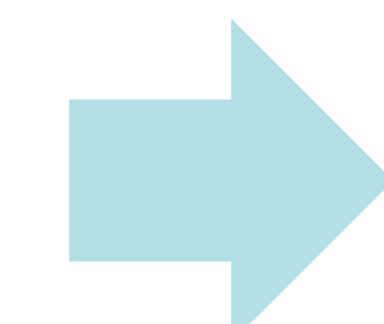
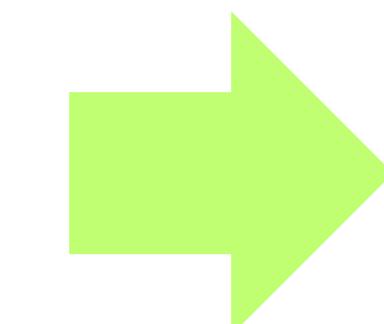
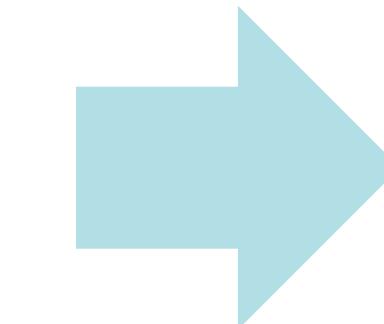
◆ variables categóricas ◆ Técnica aplicada

```
[ 'Month',
  'DayOfWeek',
  'Make',
  'AccidentArea',
  'DayOfWeekClaimed',
  'MonthClaimed',
  'Sex',
  'MaritalStatus',
  'Fault',
  'PolicyType',
  'VehicleCategory',
  'VehiclePrice',
  'Days:Policy-Accident',
  'Days:Policy-Claim',
  'AgeOfVehicle',
  'PoliceReportFiled',
  'WitnessPresent',
  'AgentType',
  'NumberOfSupplements',
  'AddressChange-Claim',
  'NumberOfCars',
  'BasePolicy',
  'FraudFound',
  'Make_Vehicle_Category',
  'RiskLevel',
  'Make_VehicleCategory']
```

Label encoder

Onehot encoder

Ordinal encoder

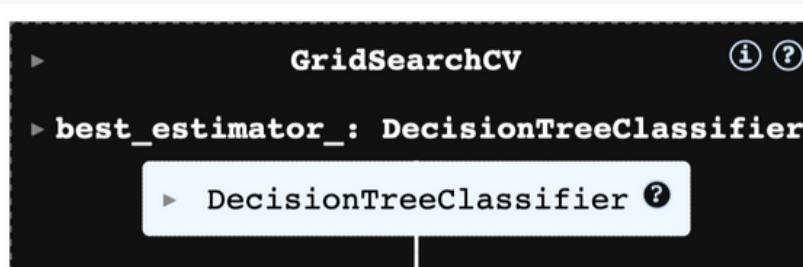


'AccidentArea', 'Sex',
'MaritalStatus', 'Fault'
'PoliceReportFiled',
'WitnessPresent', 'AgentType',
'AddressChange-Claim'

'VehiclePrice',
'PastNumberOfClaims',
'AgeOfVehicle', 'Days:Policy-
Accident', 'Days:Policy-Claim',
'NumberOfSupplements'

'VehicleCategory',
'NumberOfCars',
'BasePolicy'

ENTRENAMIENTO Y TUNEO DE HIPERPARAMETROS



max_depth	min_samples_split	min_samples_leaf
profundidad del árbol	dividir un nodo	debe contener una hoja
3, 5 y 10.	2, 5 y 10	1, 2 y 4

El mejor modelo encontrado es un árbol de decisión

max_depth=5: El árbol se ajustó a una profundidad máxima de 5.

min_samples_leaf=2: Cada hoja del árbol tiene al menos 2 muestra.

min_samples_split=10: Un nodo se divide si tiene al menos 10 muestras.

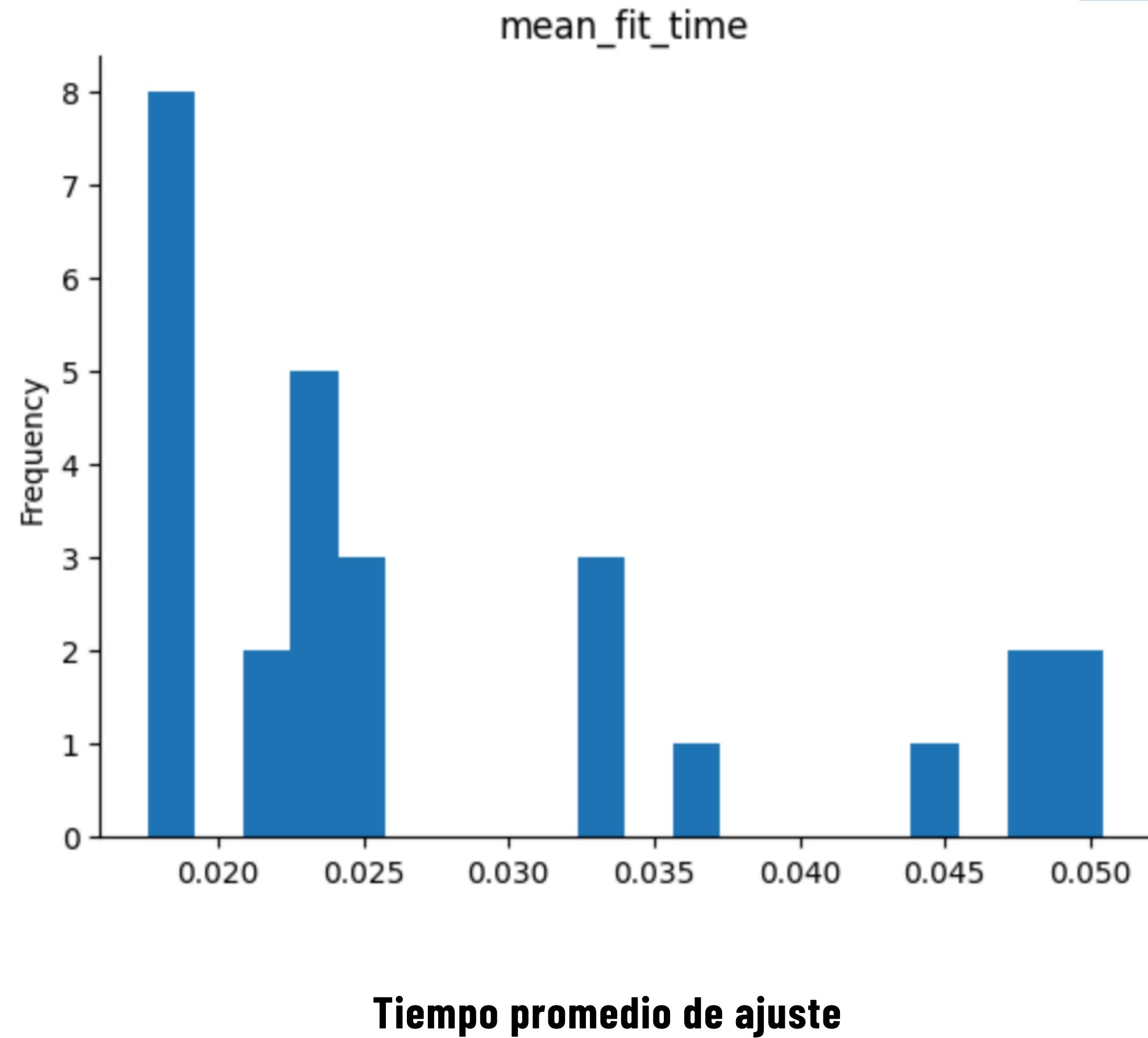
class_weight='balanced': El peso de cada clase en función de su frecuencia en los datos



Metricas

La mayoría de las combinaciones de hiperparámetros tienen tiempos de ajuste que caen alrededor de 0.02 segundos

La complejidad de las configuraciones de hiperparámetros no varía de manera drástica



RESULTADO



RENDIMIENTO

Accuracy en el conjunto de validación para Decision Tree 1: 0.70

Accuracy en el conjunto de validación para Decision Tree 2: 0.79

Informe de clasificación para Decision Tree 1:

	precision	recall	f1-score	support
No	0.97	0.66	0.78	692
Yes	0.45	0.93	0.68	208
accuracy			0.72	900
macro avg	0.71	0.79	0.69	900
weighted avg	0.85	0.72	0.74	900

Clase 0 (No Fraude):
Indica un buen balance,
aunque la sensibilidad (recall)
es algo baja.

Clase 1 (Fraude):
Indica un compromiso entre
precisión y recall, aunque la
precisión para identificar el fraude
es baja, el modelo compensa con
una alta tasa de recall.

RECALL DE UNOS: 0.93

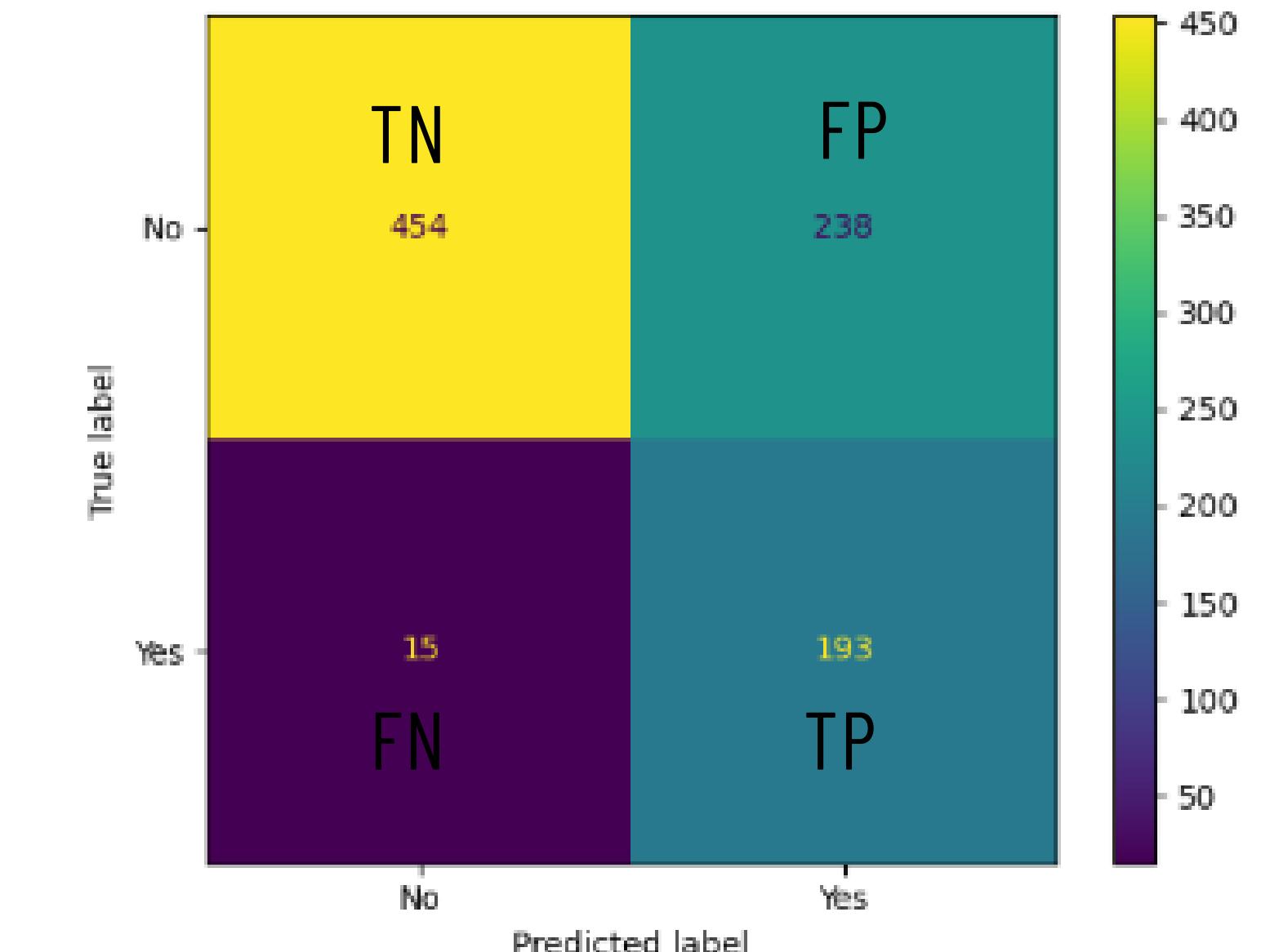
Total costo por falsos positivos (FP): \$71.400

Total costo por falsos negativos (FN): \$965.000

Total ahorro por verdaderos positivos (TP): \$2'270.000

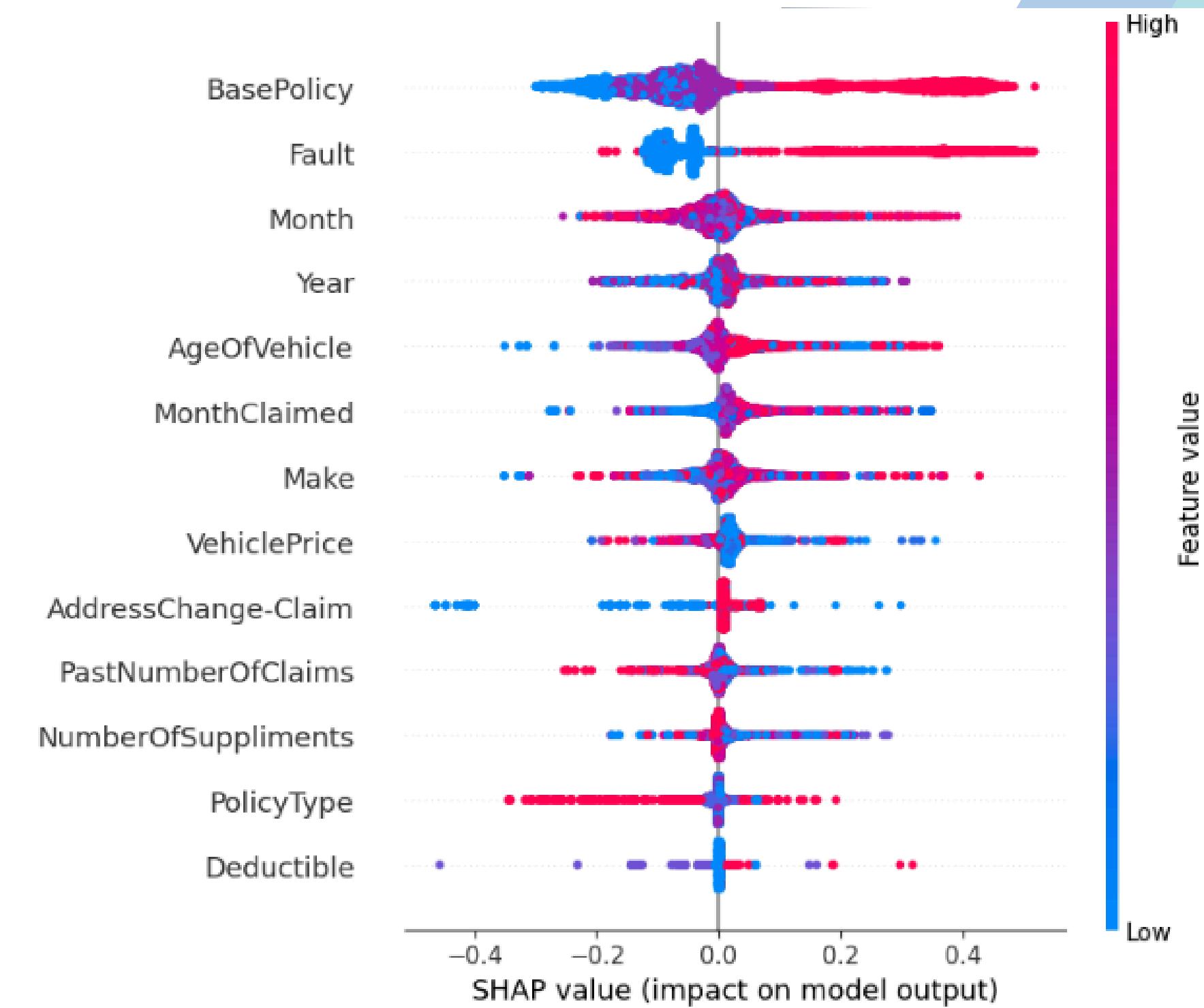
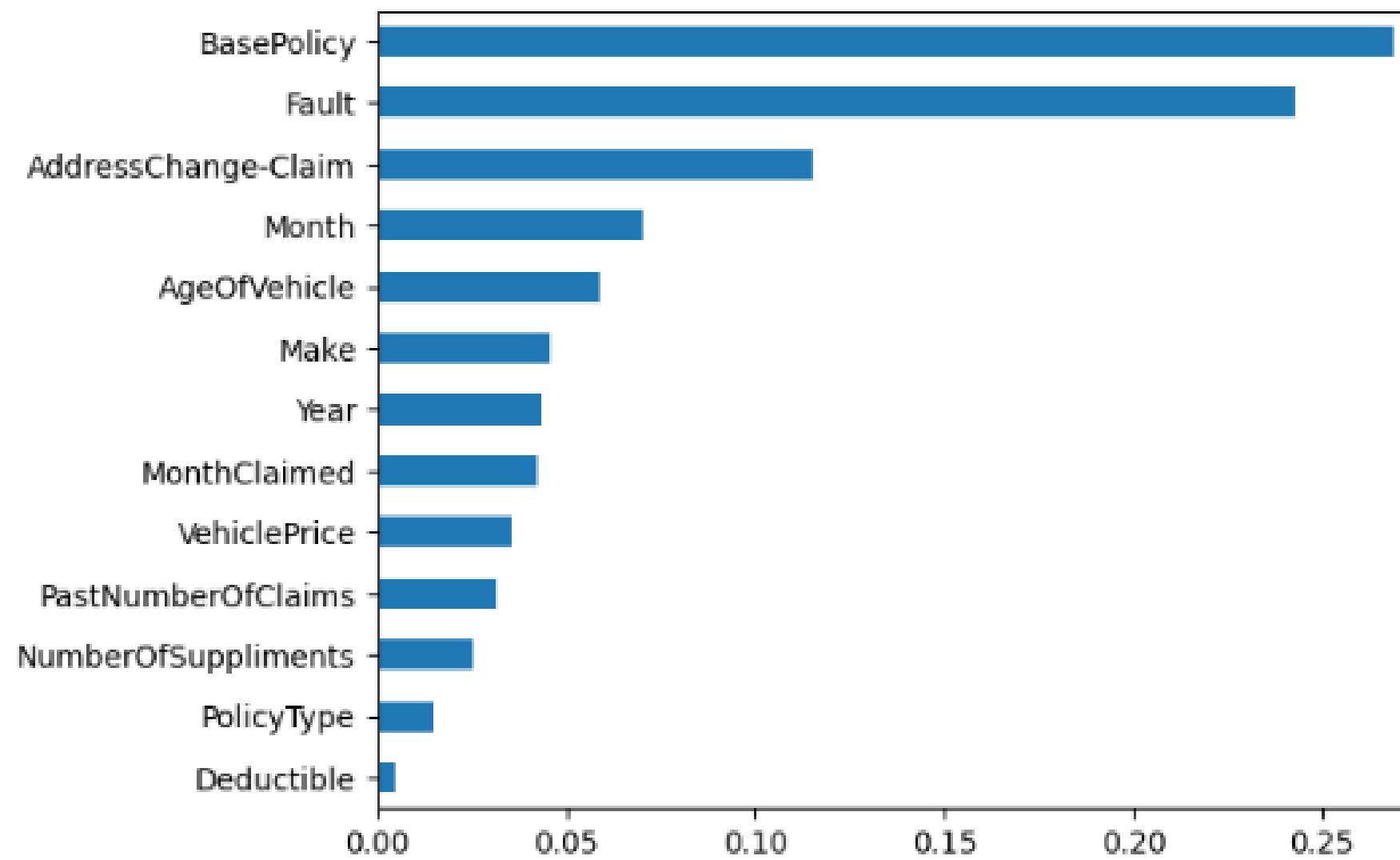
Impacto económico neto: \$1'233.600

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7baab8951c30>



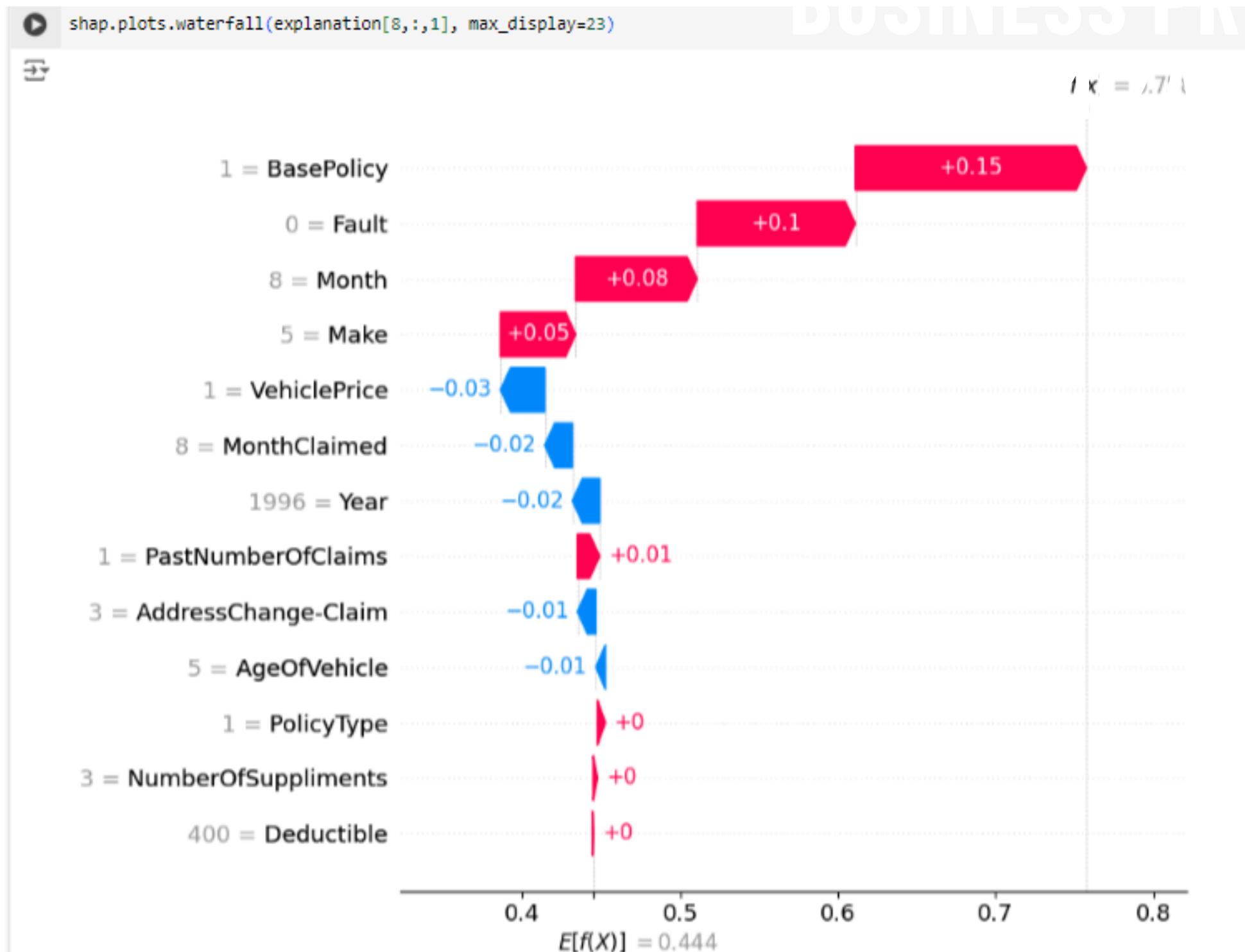
INTERPRETABILIDAD / EXPLICABILIDAD DEL MODELO

- BasePolicy - más importante, influye más en las predicciones del modelo.
- Fault y AddressChange - Claim - alta importancia, el modelo utiliza significativamente para tomar decisiones



- BasePolicy tiene un fuerte impacto positivo y negativo en las predicciones, siendo una de las variables más influyentes

INTERPRETABILIDAD / EXPLICABILIDAD DEL MODELO



- BasePolicy es la más importante y tiene un gran impacto positivo.
- Fault también es bastante influyente.
- Month aporta una contribución significativa al modelo.

MODELO ADICIONAL

Top 10 + Target

- Preparación del Conjunto de Datos
- División del Conjunto de Datos
- Entrenamiento Modelos Alternativos
- Evaluación de Modelos
- Ajuste de Hiperparámetros

▼ RandomForestClassifier ⓘ ?

```
RandomForestClassifier(random_state=42)
```

	precision	recall	f1-score	support
0	0.97	0.96	0.96	1312
1	0.89	0.92	0.90	497
accuracy			0.95	1809
macro avg	0.93	0.94	0.93	1809
weighted avg	0.95	0.95	0.95	1809
	[[1254 58]			
	[40 457]]			

El modelo parece ser robusto y tiene un buen equilibrio entre precisión y recall, aunque hay margen para mejorar la identificación de casos fraudulentos.

Mejorar Potencial: Aplicar técnicas de oversampling para mejorar la detección de la clase positiva (fraude).

Visualización: Considera visualizar la matriz de confusión y las métricas de evaluación para una mejor interpretación.



IMPLEMENTACIÓN EN EL NEGOCIO

**Participación e integración de
Stakeholders
para que el modelo se alinee con las
necesidades del negocio**

MAPEA CÓMO EL
MODELO ENCAJA
EN LOS PROCESOS
DE DETECCIÓN DE
FRAUDE
EXISTENTE

AUTOMATIZACIÓN
PARA OBTENER UNA
PREDICCIÓN EN
TIEMPO REAL

ACTUALIZACIÓN
PERIÓDICA Y
FEEDBACK DE
RESULTADOS
CON FP Y FN

ENTRENAMIENTO
EN DATOS
ACTUALES Y
MONITOREO

LIMITACIONES



PRODUCCIÓN

TIEMPO DE IMPLEMENTACIÓN

DATOS Y SU CALIDAD

EVOLUCIÓN DEL FRAUDE

OVERFITTING

COSTOS Y RECURSOS



CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

El desempeño sólido del modelo ha mostrado un alto nivel de precisión (accuracy del 95%), lo que significa que puede detectar la gran mayoría de los casos de fraude de manera efectiva.

RECOMENDACIONES

Aunque el rendimiento inicial es excelente, es fundamental que el modelo sea monitoreado regularmente para asegurarse de que mantenga su eficacia.

Capacitación del personal, ayudará a mejorar la interacción entre las revisiones manuales y automáticas.



GRACIAS

GRUPO 6
SAMIRA CARRILLO
CRISTINA CEDEÑO
GENESIS RODRIGUEZ