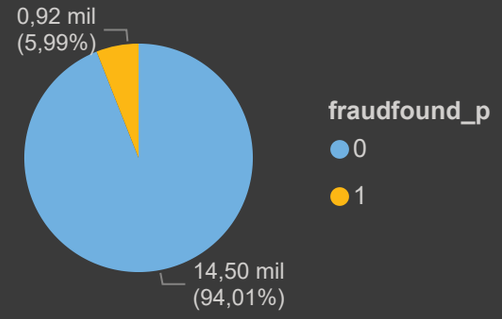
# Análisis de fraude en accidentes vehiculares

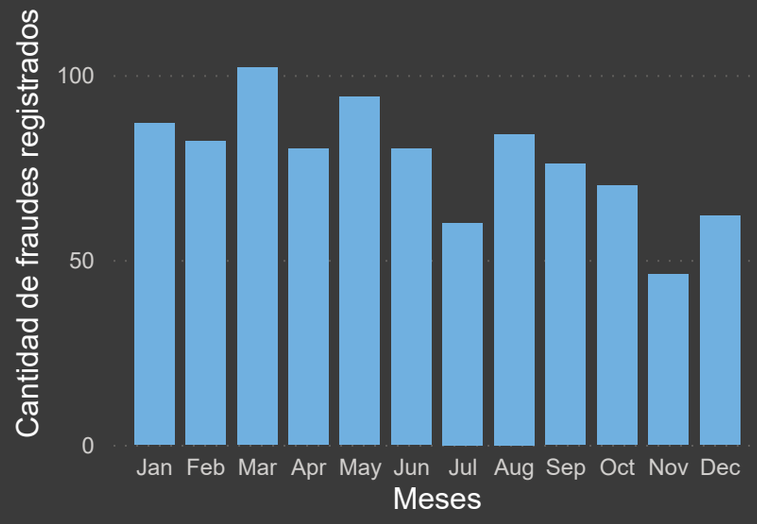
## Distribución de fraudes registrados




Las distribuciones de fraude por día y semana tienden a ser uniformes, sin embargo, para los meses se tiende a ver una disminución para meses mas tardíos, igual que en los años.

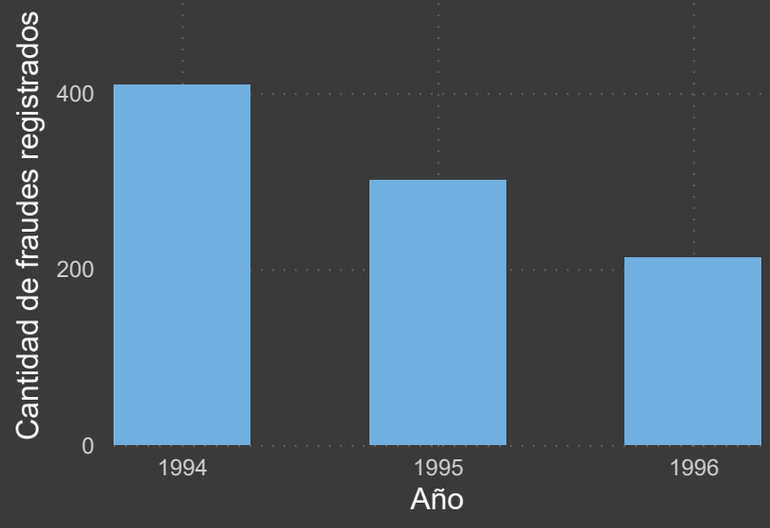
Estos cambios pueden deberse al efecto de las estaciones, para el caso de los meses. Para el caso de los años se ve una reducción notable de la cantidad de fraudes registrados.

## Cantidad de fraudes por mes

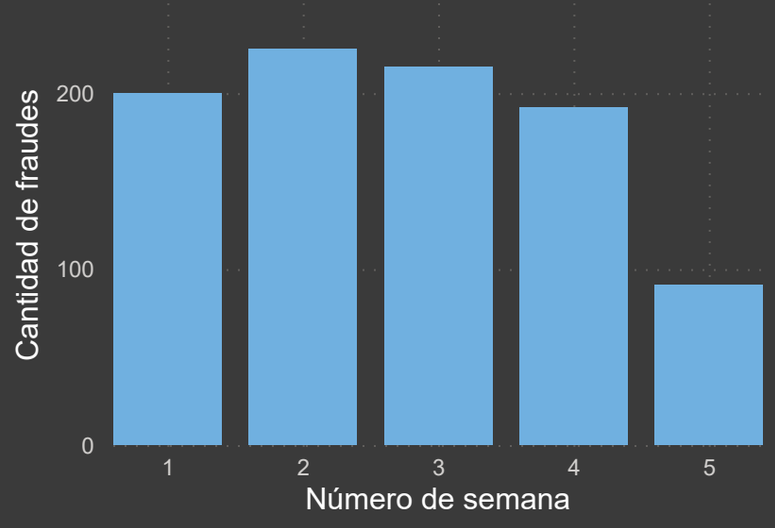



## Cantidad de fraudes por día de ocurrencia

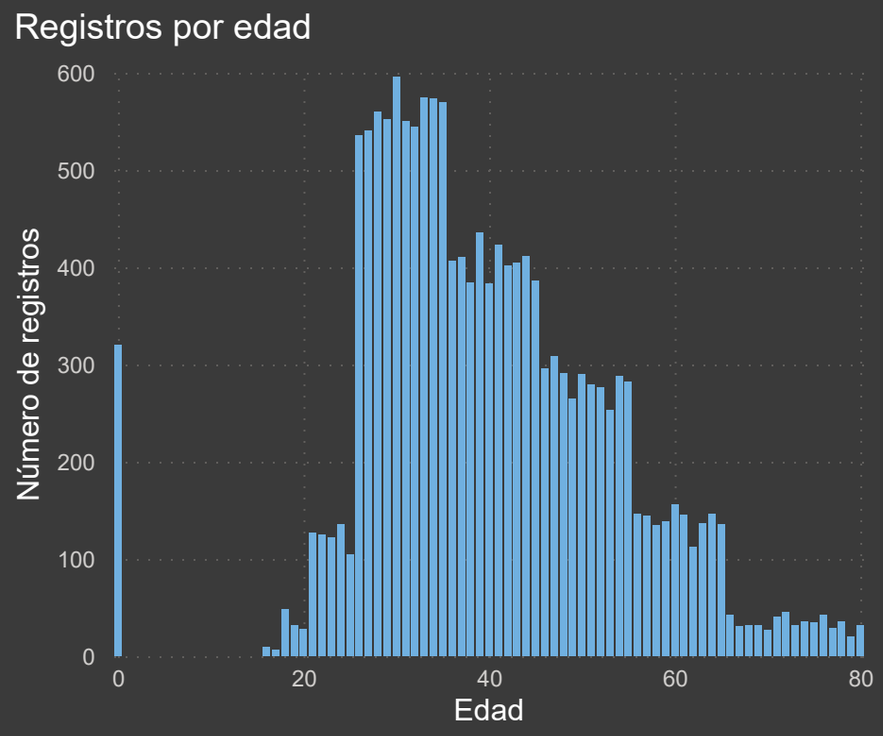

## Cantidad de fraudes por año




## Cantidad de fraudes semana de ocurrencia



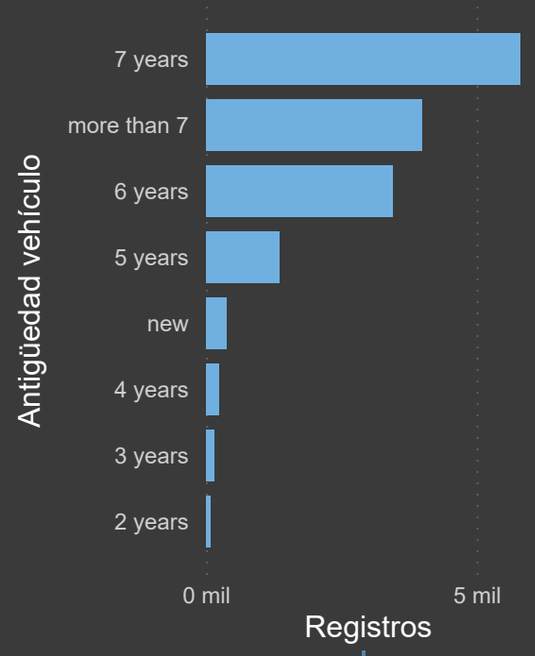
# Consistencia de la información registrada



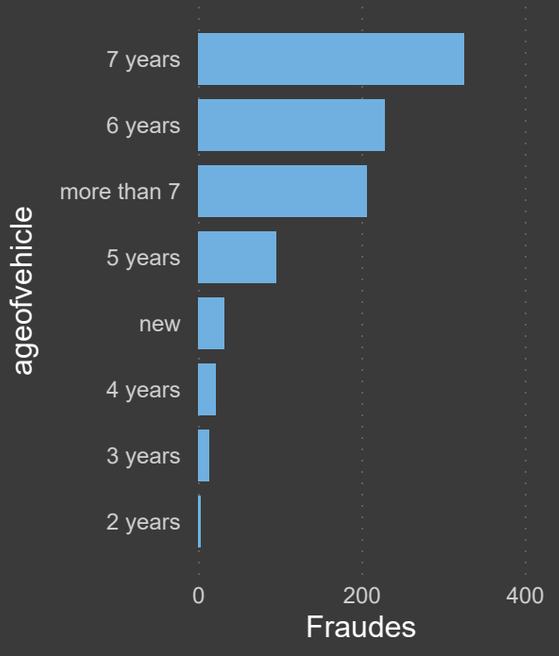

Como se puede observar, hay ciertas inserciones que no registran una edad, por lo que pueden considerarse como registros incompletos. Algo que se debe tener en cuenta a la hora de analizar las características que llevan a fraudes.

# Fraudes por característica del vehiculo

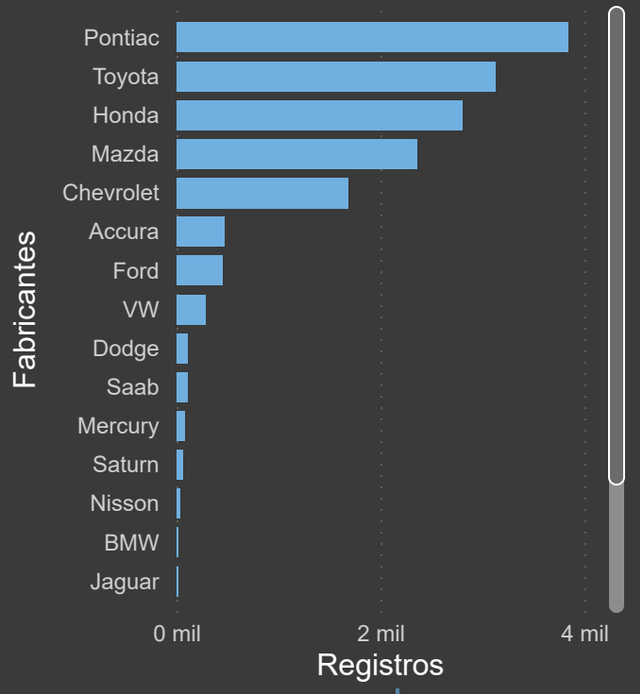
Registros por antigüedad de vehiculo



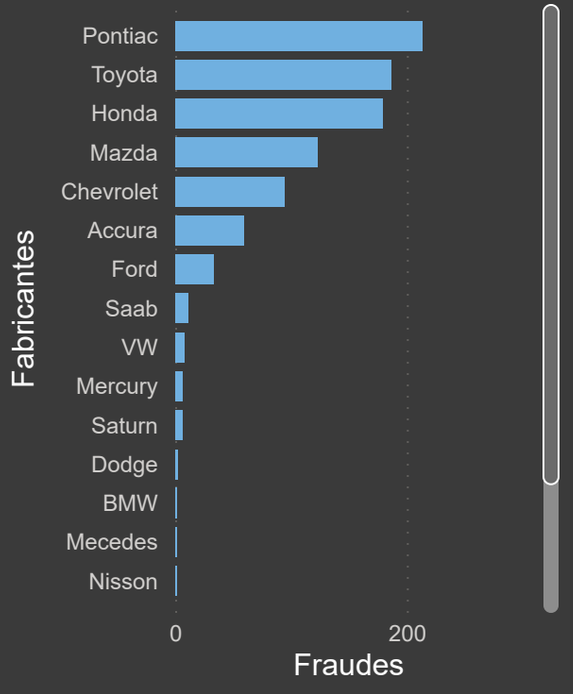

Fraudes por antigüedad de vehículo




Registros por fabricante



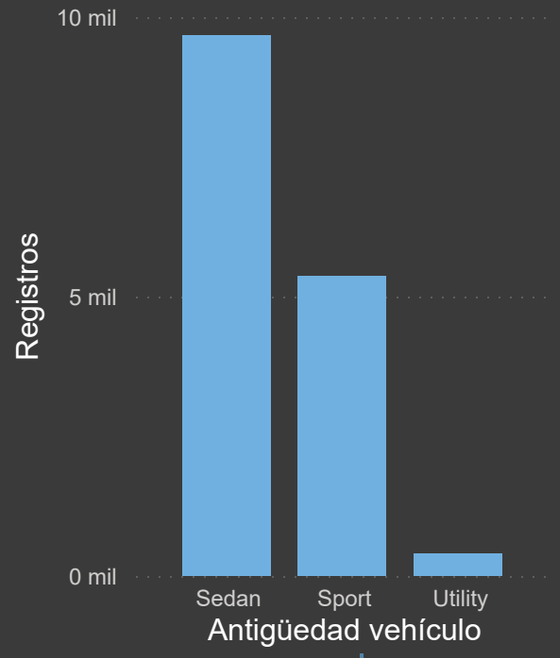

Fraudes por fabricante



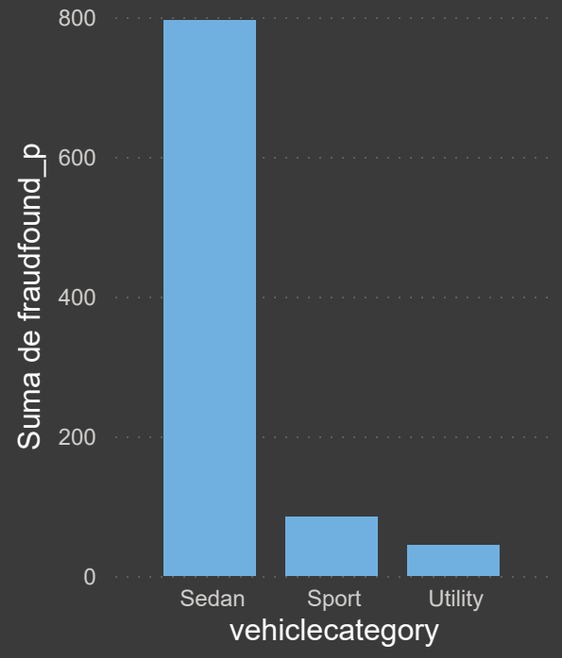

No se ven diferencias notables en la cantidad de fraudes al analizar antigüedad y fabricantes. Se mantiene la distribución

# Fraudes por característica del vehiculo

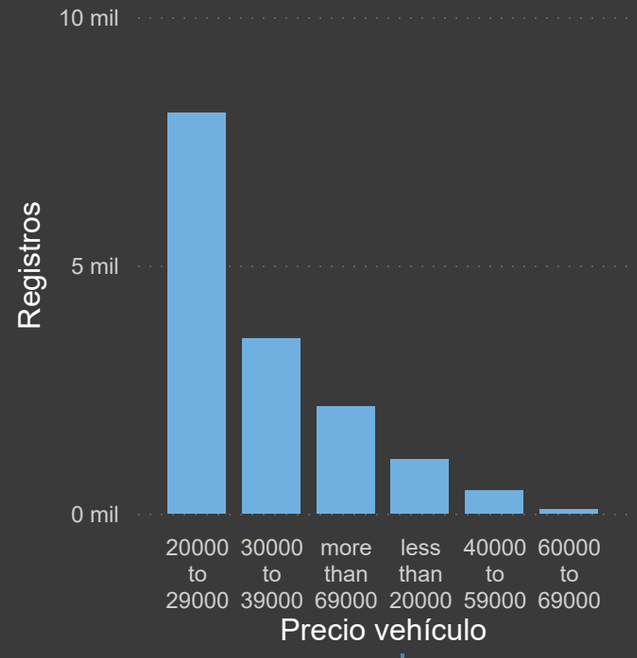
Registros por tipo de vehiculo



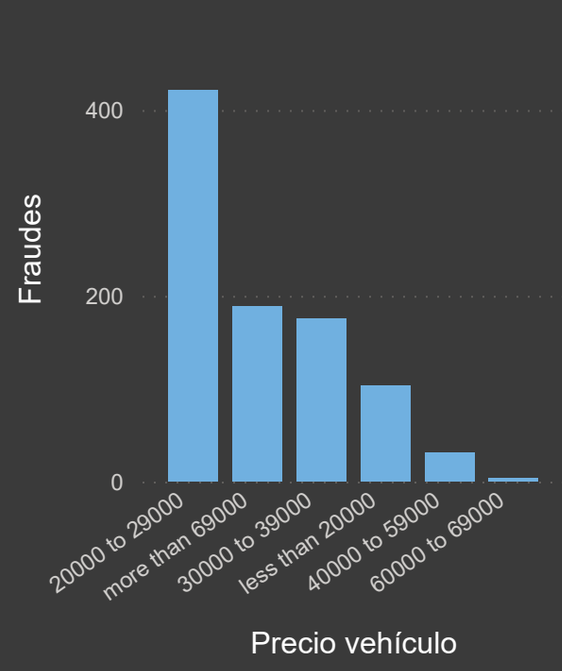

Fraudes por tipo de vehículo




Registros por precio



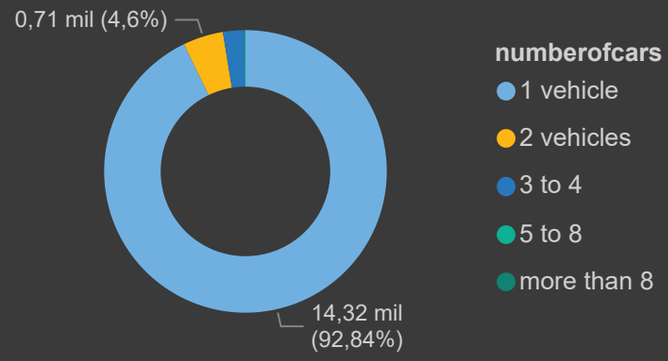

Fraudes por precio



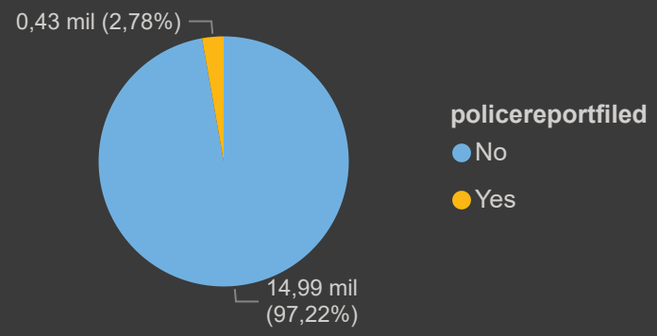

Se ve que cuando hay fraude, la gran mayoría de los vehículos involucrados son de tipo sedan. También tiende a haber más fraudes (en porcentaje) en vehículos con precios medios ($30.000 - $39.000).

# Variables con poca información

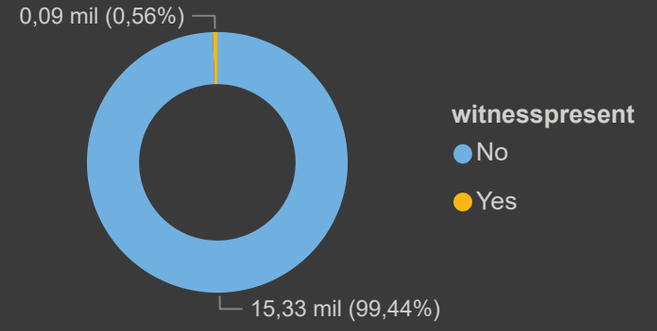
Número de carros involucrados en el siniestro




Indicación de si se llenó o no reporte policial



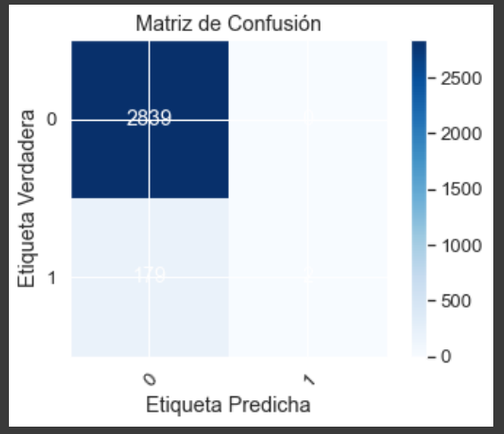

Indicación de presencia de testigos en el siniestro




Se evidencian ciertas variables de baja importancia en la categorización de fraudes o no debido a la predominancia que se ve en las distribuciones.

# Conclusiones

Con el análisis realizado y algunas modificaciones adicionales, se corrió un modelo de Machine Learning Para tratar de predecir, según las características disponibles, si un suceso puede ser fraudulento o no.




Presiona para acceder al repositorio

Los resultados del modelo muestran que se predice adecuadamente cuando no es fraude, pero se predice erróneamente cuando si existe fraude. El desbalance de clases hace que no haya suficiente información que ayude a determinar los patrones asociados a fraude.

De este modo, se recomienda hacer una limpieza de la información recolectada y balancear el conjunto por medio de técnicas como SMOTE.

Es necesario suprimir aquellas variables que hacen ruido, así como encontrar aquellas que contienen información valiosa.