Universidad Don Bosco

Escuela de Computación

Facultad de Ingeniería

Datawarehouse y minería de datos

Desafío practico 3

Grupo 01L

Integrantes:

Rivas González, Cesar Josué RG180141

Docente:

Alexander Alberto Sigüenza Campos

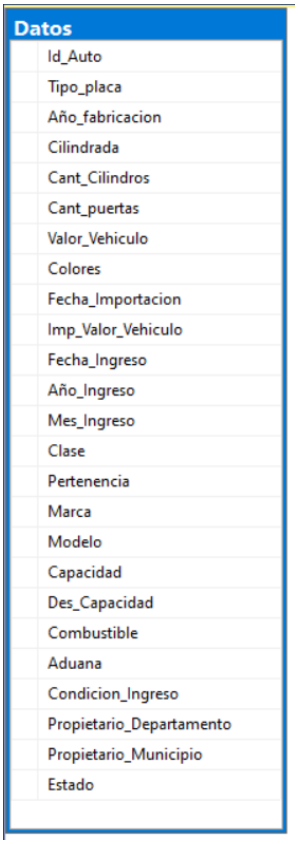
Fecha:

12 de noviembre de 2020

Para el análisis de los datos mediante métodos de minería, se utilizaron las herramientas de visual studio de SQL Analysis services.

Parte I: Análisis de Parque Vehicular

Para poder utilizar las herramientas de visual studio, primero se realizo un modelo de base de datos para poder utilizarlo como data source, para cargar los datos del .csv a la base de datos, se hizo por medio de un etl simple. La tabla en la base de datos quedo de la siguiente manera:



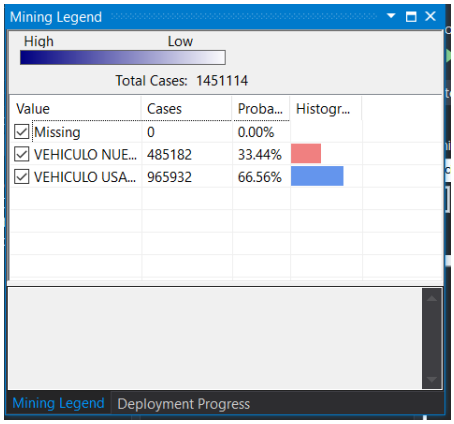
Posteriormente, se decidió analizar estos datos por medio de los algoritmos de árboles de decisión y de reglas de asociación.

1. **Arboles de decisión**

   Para la creación del árbol de decisión, solo se creó un nuevo proyecto multidimensional en visual studio, y se agrego como data source el modelo de base de datos antes mostrado. Se utilizo como columna key una columna Id_auto que fue generada automáticamente en el etl, para la columna de predicción, se tomó la condición de ingreso del auto, si este es nuevo o usado, y para las columnas de input se tomaron el año de fabricación, la aduana por la que ingresó, la clase de auto, el año de ingreso, el tipo de placa y la pertenencia.

   Cabe destacar que se utilizo el 30% de los datos para realizar las predicciones.

Una vez creada la estructura de minería, la procesamos y comenzamos el análisis. Según el modelo de minería, podemos determinar que la mayoría de los autos que ingresan son usados:



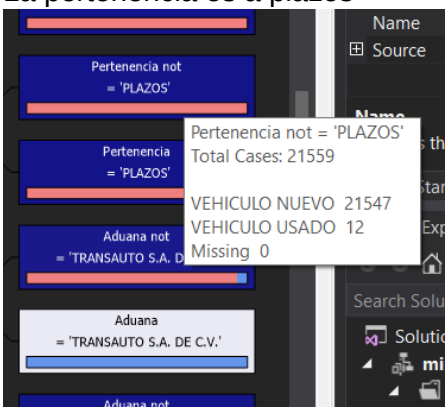
Probabilidad de vehículos nuevos

Por cuestiones de comodidad no es posible mostrar todo el árbol de decisión ya que por la cantidad de información se hizo muy extenso, pero a continuación se listan los parámetros que mas afectan a que un auto sea nuevo según lo estudiado:

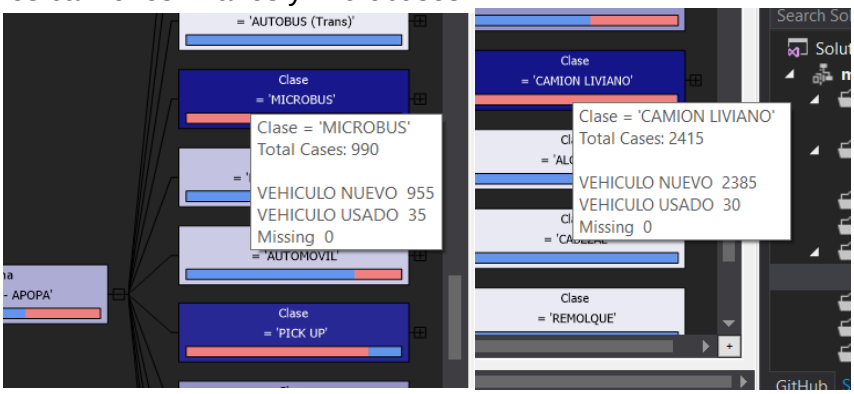
a) El tipo de placa es de motocicleta

b) El año de ingreso es de 2010 para arriba (2010 hacia abajo la probabilidad de que sea nuevo baja mucho)

c) La pertenencia es a plazos



d) En caso de no ser motocicleta, los mas comunes que se presentan nuevos, son los camiones livianos y microbuses.

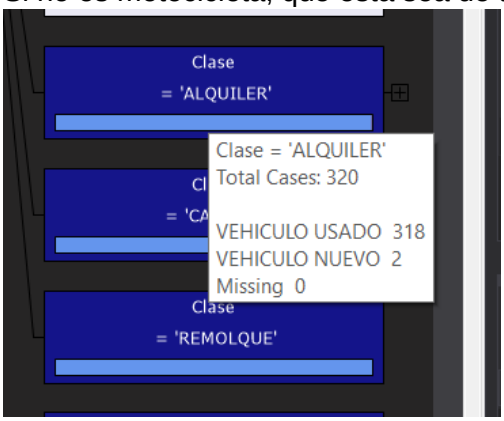


Probabilidad de vehículos usados

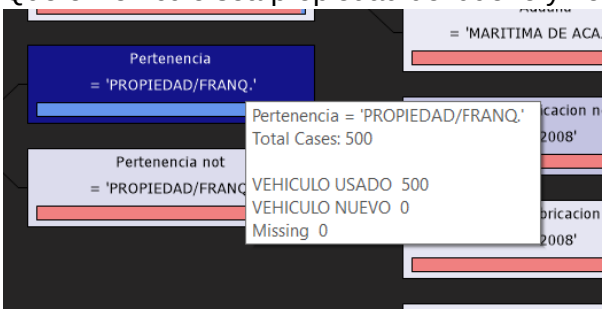
Como se mencionó anteriormente, los vehículos usados son mas comunes, y los parámetros que mas afectan a que un vehículo sea usado son los siguientes:

a) Que sea de tipo motocicleta, pero esta haya sido ingresada entre 1994 y 2000

b) Si no es motocicleta, que esta sea de tipo cabezal, alquiler o remolque

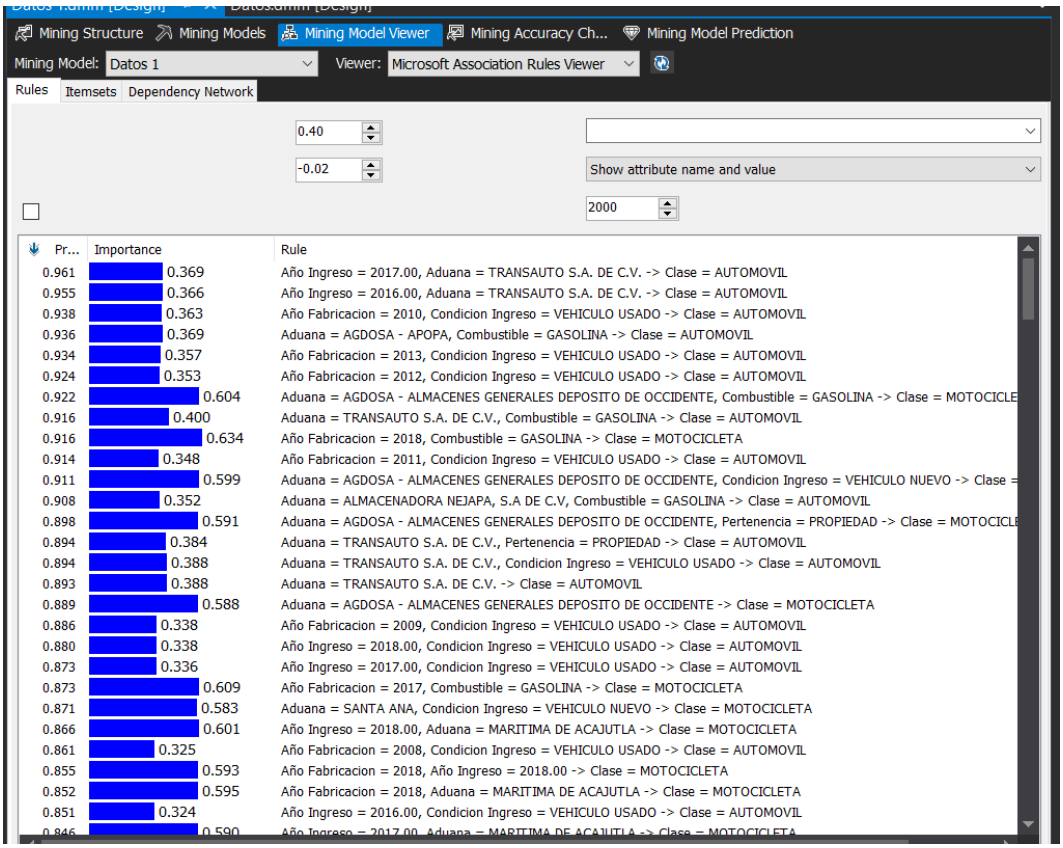


c) Que el vehículo sea propiedad del dueño y no lo esté pagando aun



Reglas de asociación

Para el análisis por medio de reglas de asociación, se hizo uso del mismo data source, pero esta vez se utilizo como columna de predicción, la de la clase del auto, una vez procesados los datos podemos proceder al análisis.



Aquí se muestran las asociaciones que se hicieron que tienen una mayor probabilidad e importancia. De esto podemos apreciar lo siguiente:

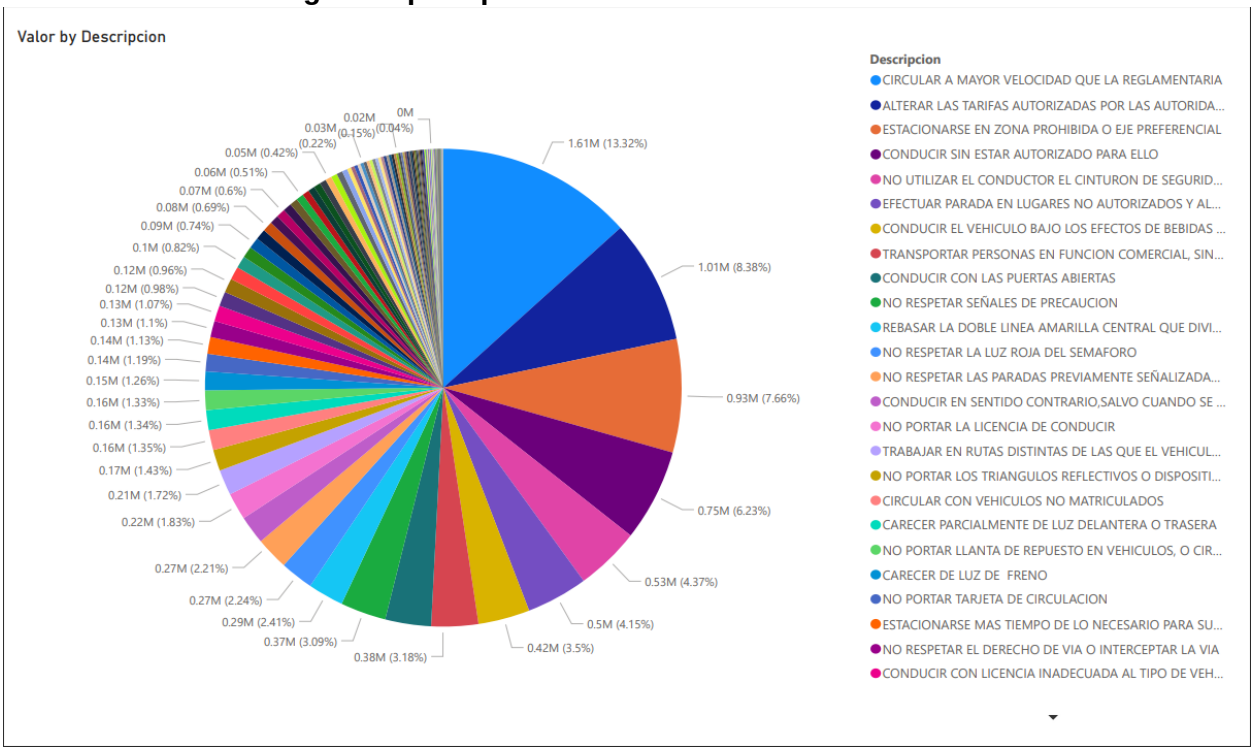
- Hay una gran probabilidad que si es de tipo automóvil, este haya ingresado en 2017 o 2016 y en la aduana de TRANSAUTO S.A DE C.V.
- Hay una gran probabilidad de que si es un automóvil, este sea fabricado en el 2010,2012 o 2013 y sea un vehículo usado.
- Si el vehículo es una motocicleta, es muy probable que haya sido fabricada en 2018 (es decir, que sea nueva ya que los datos son del 2018)
- Si es una motocicleta ingresada en la aduana de Santa Ana, es muy probable que sea nueva
- En general, es muy probable que una motocicleta que ingrese a la aduana sea nueva
- En cambio, si es una automóvil, lo mas probable es que sea usado

Parte II: Análisis de esquelas

Para el análisis de el otro documento de esquelas, también se hizo uso de un etl para cargar los datos a una base de datos de SQL server y poder utilizarlos con visual studio. En este caso, se hizo uso de las herramientas de power BI para hacer el análisis, además de la técnica de agrupamiento por cluster o k-means.

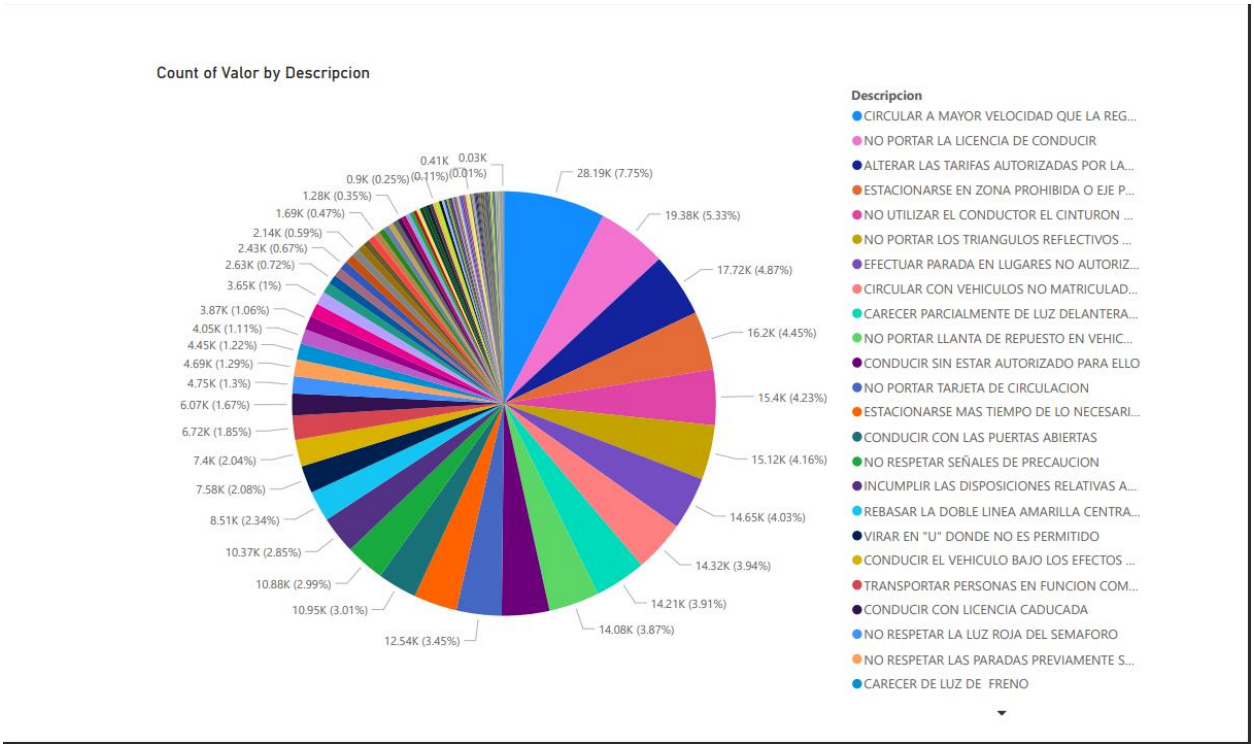
Analisis por medio de power BI

a) Grafico de valor de ingresos por tipo de infracción



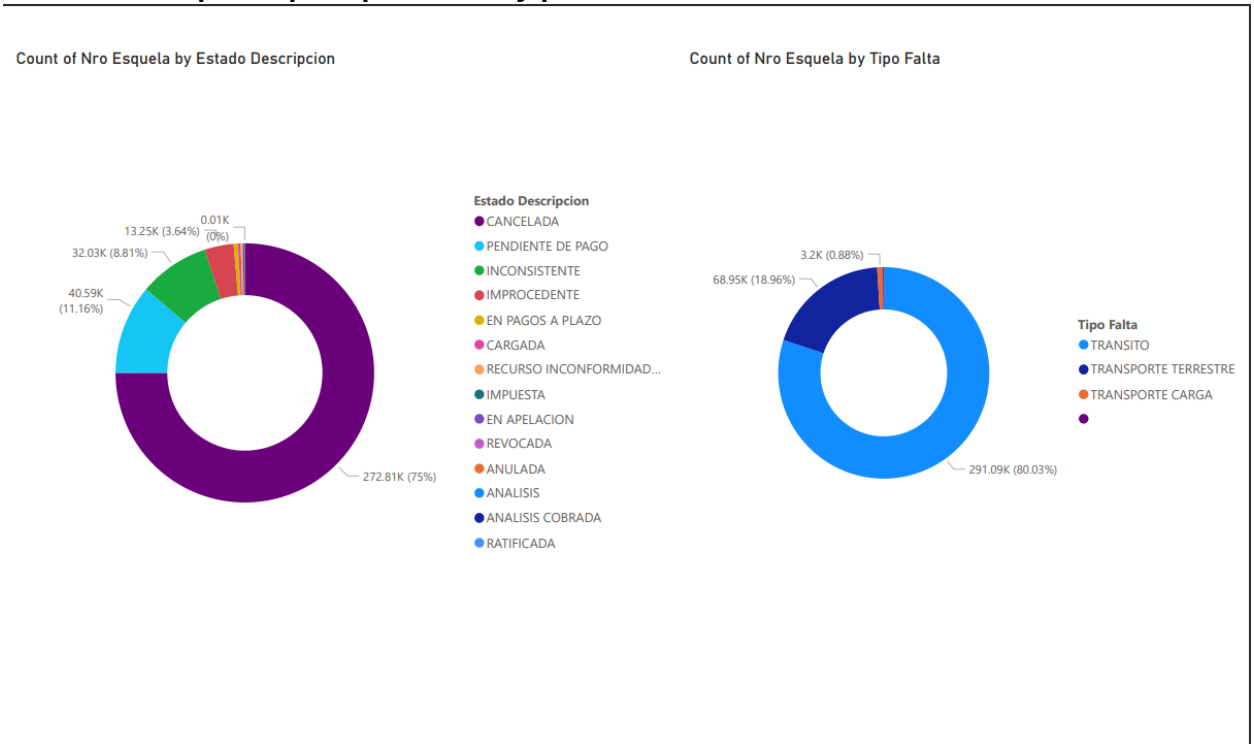
Hay un gran numero de infracciones como se puede apreciar en el gráfico, pero como podemos ver las infracciones que más genera dinero mediante multas, son las de circular a exceso de velocidad, alterar las tarifas impuestas, y estacionarse en zonas prohibidas, por mencionar solo las mayores tres.

b) Conteo del numero de infracciones por cada tipo



Por otro lado, podemos ver que en el caso del conteo, el tipo de infracciones con mayor numero de faltas son: circular sobre el limite de velocidad, no portar la licencia de conducir, y alterar las tarifas impuestas.

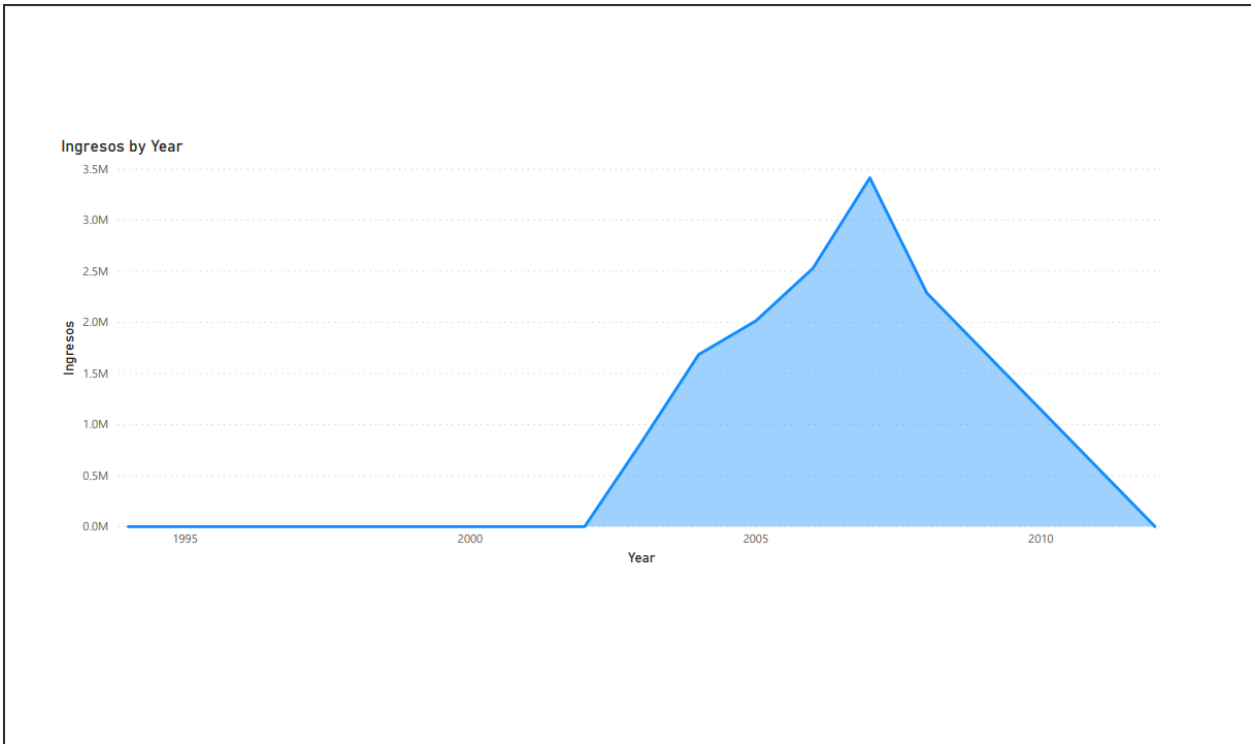
c) Conteo de esquelas por tipo de falta y por su estado actual



En el grafico de la izquierda podemos apreciar que alrededor de un 75% de las infracciones que fueron emitidas han sido canceladas por los infractores. Por otro lado, también podemos ver que un 11% de ellas están pendientes y otro buen porcentaje es inconsistente.

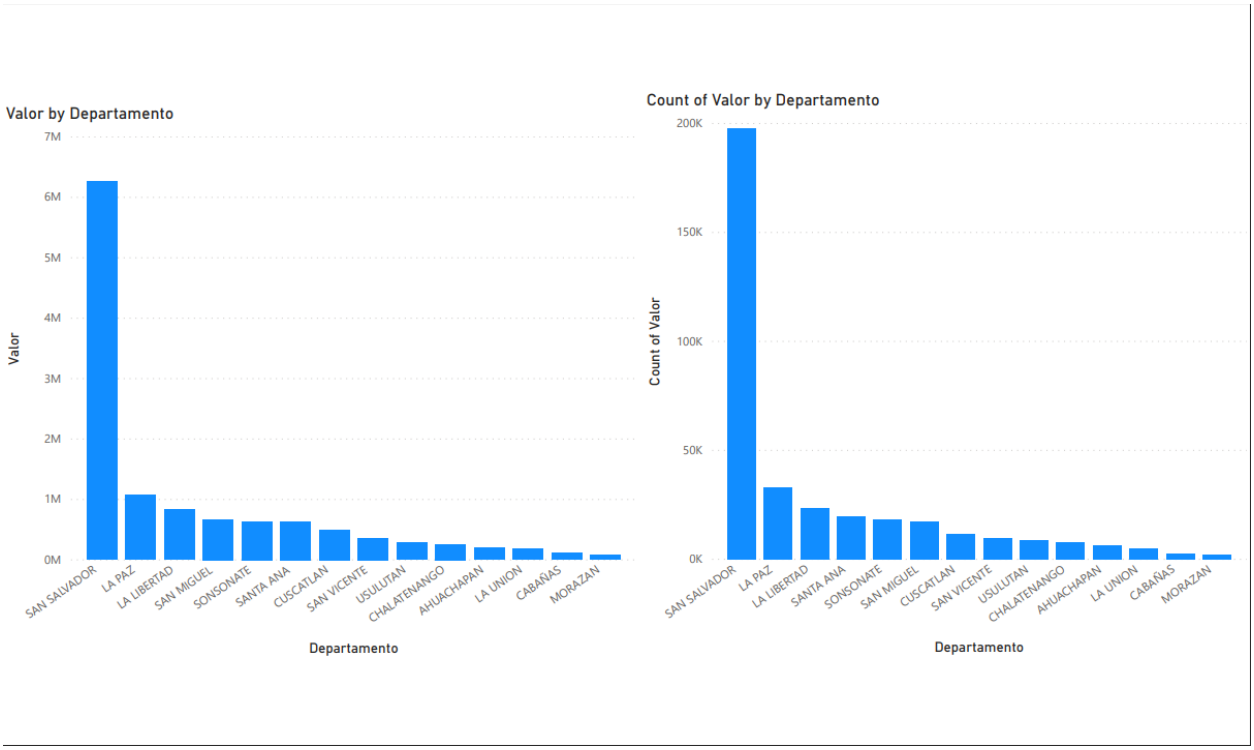
En el grafico de la derecha podemos ver que la gran mayoría de las infracciones emitidas son a vehículos particulares, mientras que otro porcentaje considerable es medios de transporte publico como buses, microbuses, mientras que una pequeña parte de las infracciones son emitidas a transporte de mercancía.

d) Ingresos por año



En este grafico se muestran los ingresos que se tuvieron por multas cada año desde 2002 hasta 2012. Como podemos ver los ingresos fueron incrementando cada año, hasta que llegaron a alrededor de 3.5 millones en el año 2007 y desde entonces han ido disminuyendo.

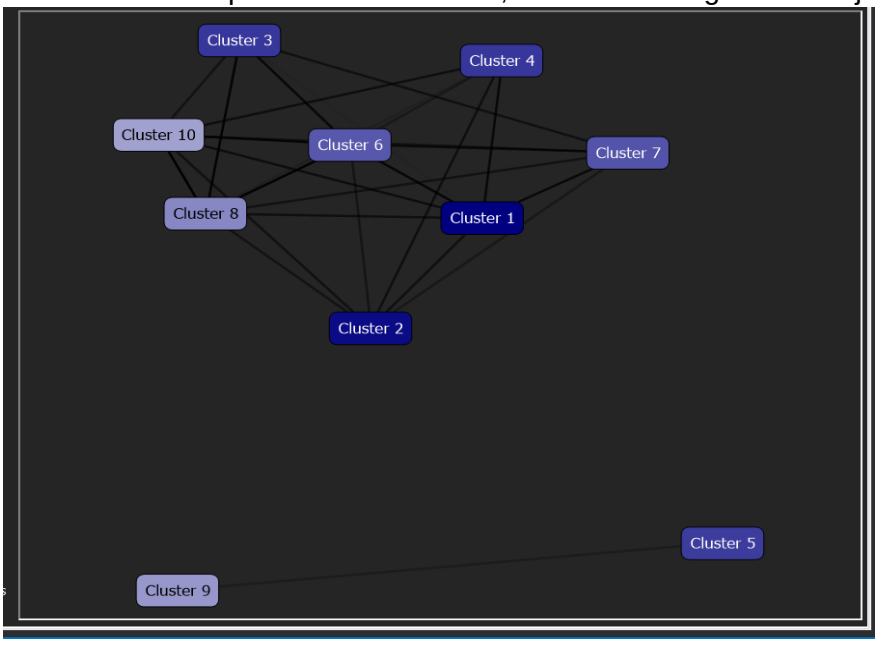
e) Ingresos por departamento y conteo de multas por departamento



En ambos gráficos se puede apreciar que el departamento que tiene mas multas e ingreso por infracciones es la capital, y que luego le sigue el departamento de la paz y la libertad.

Análisis por medio de k-means

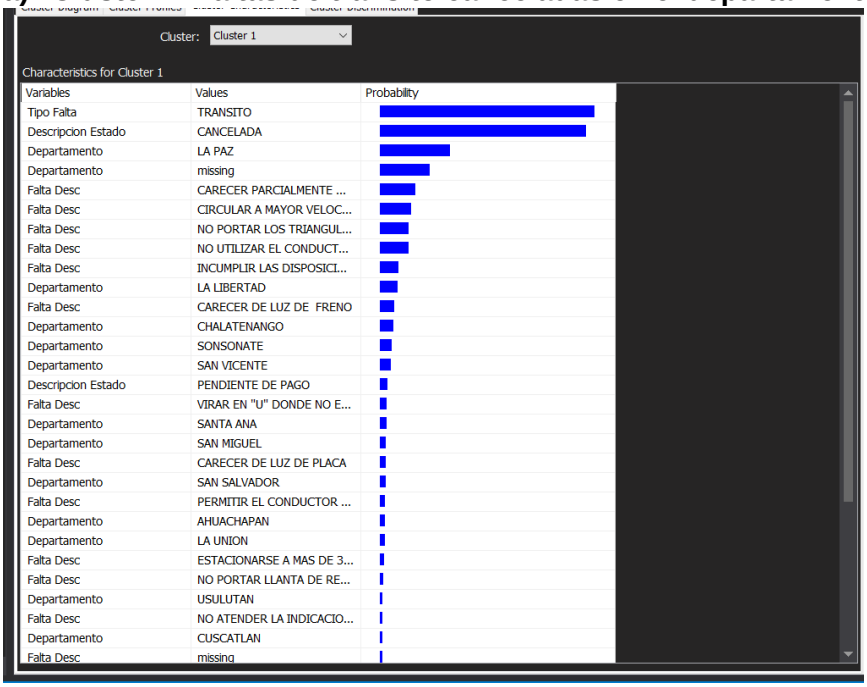
Para hacer el análisis por medio de k-means se utilizó la herramienta de SQL analysis service. Una vez procesado el modelo, se obtuvo el siguiente conjunto de clústeres:



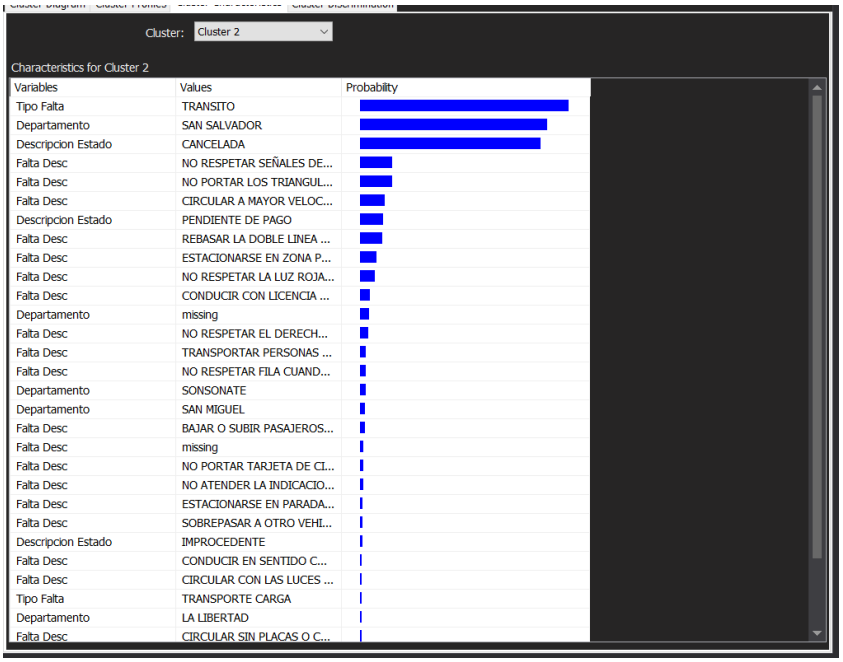
Se obtuvo un gran grupo de clústeres y otro grupo mas pequeño, lo que los diferencia es el tipo de vehículo al que se le emitió la multa, el grupo grande de clúster son multas a vehículos particulares, mientras que el grupo pequeño son a vehículos de transporte.
En dicho diagrama, un color mas oscuro significa un clúster con mas densidad de población.

A continuación se muestra la población de cada clúster detalladamente:

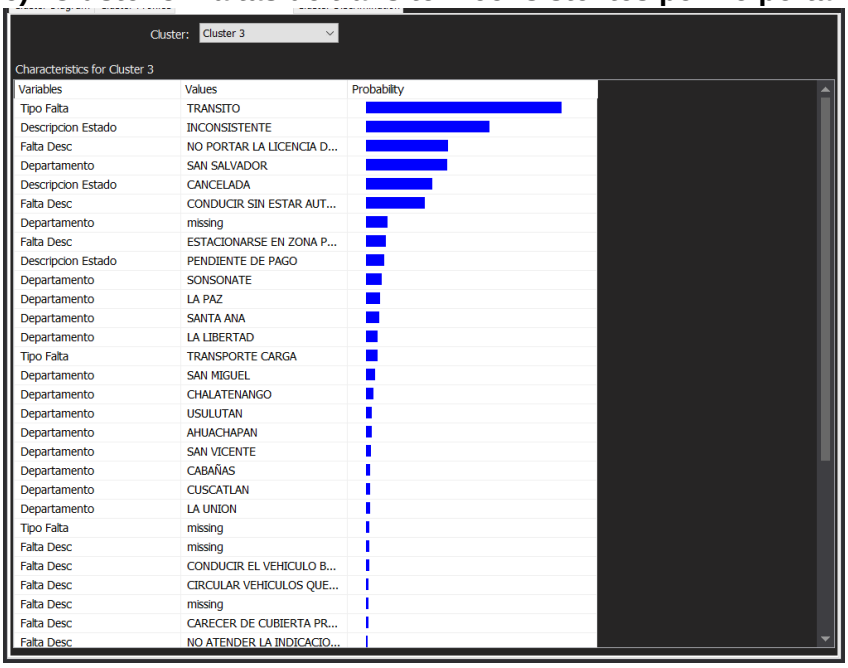
a) Clúster 1: Faltas de transito canceladas en el departamento de la paz



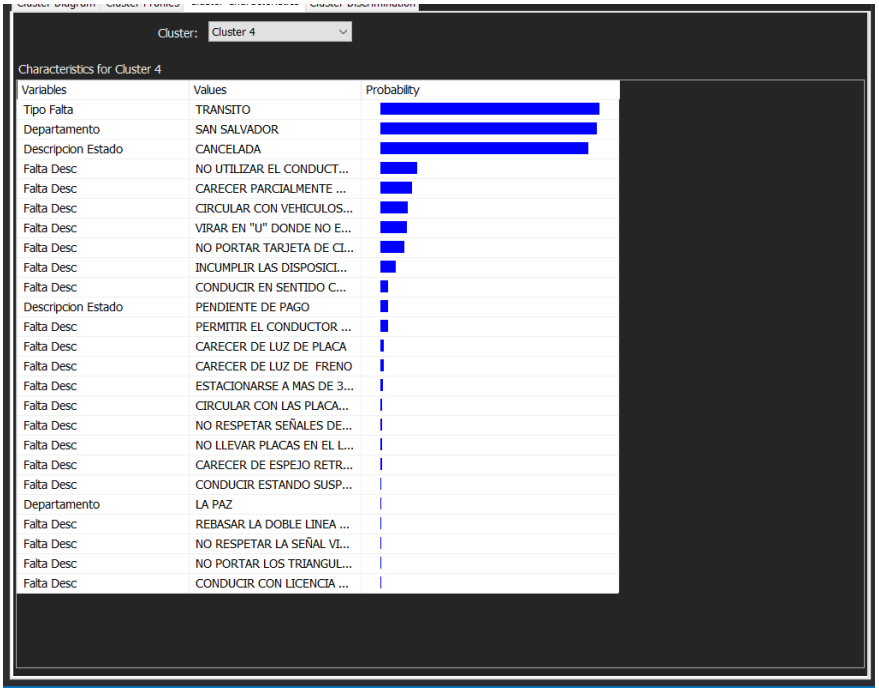
b) Clúster 2: Faltas de transito canceladas en San Salvador por no respetar señales de transito



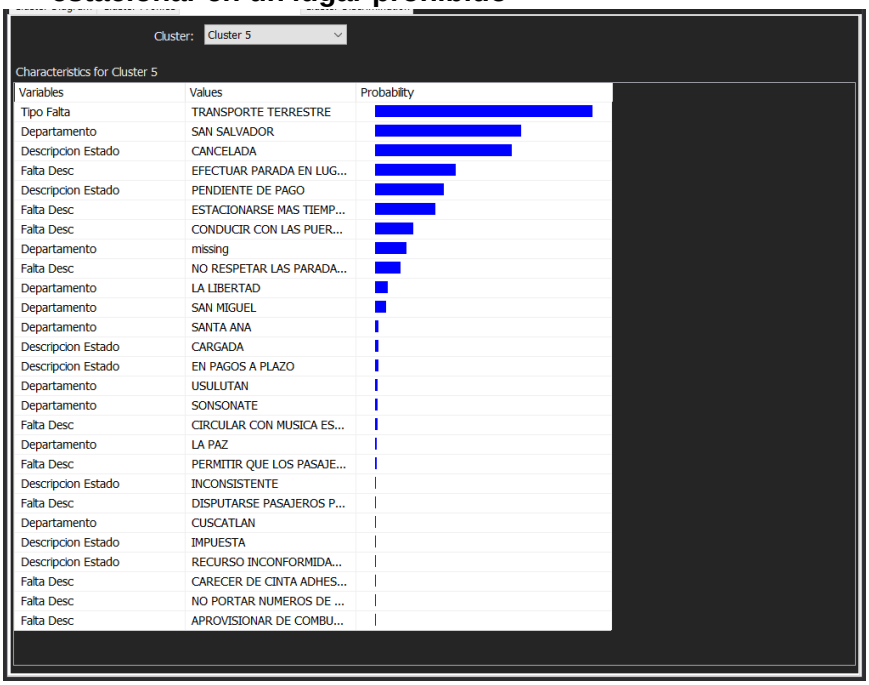
c) Clúster 3: Faltas de transito inconsistentes por no portar licencia



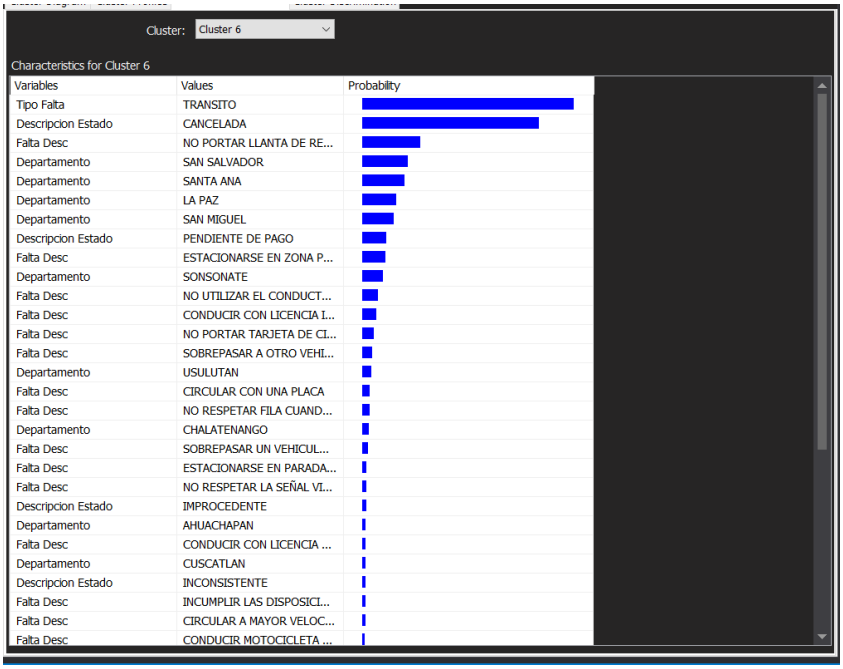
d) Clúster 4: Faltas de transito canceladas en San Salvador por no utilizar cinturón de seguridad



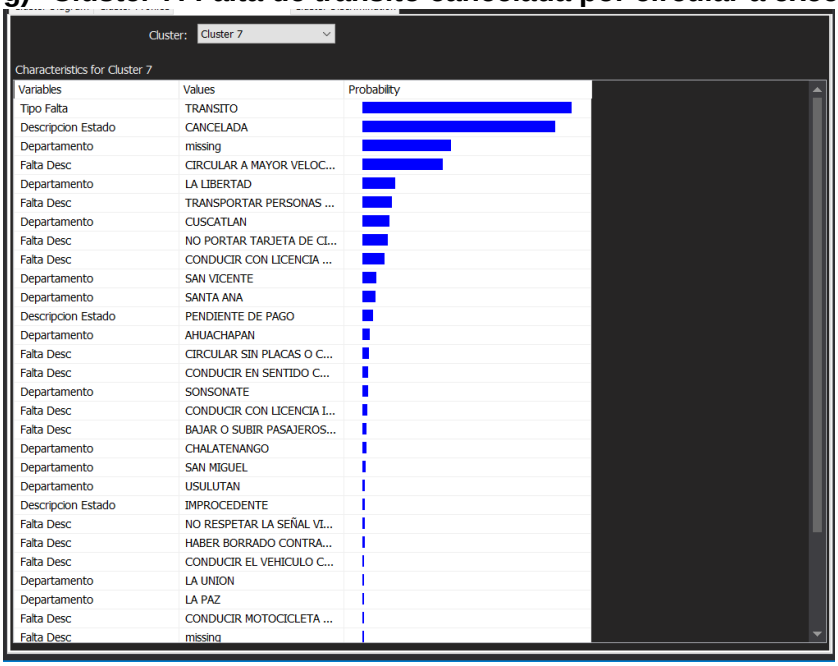
e) Clúster 5: Faltas de transporte terrestre canceladas en San Salvador, por estacionar en un lugar prohibido



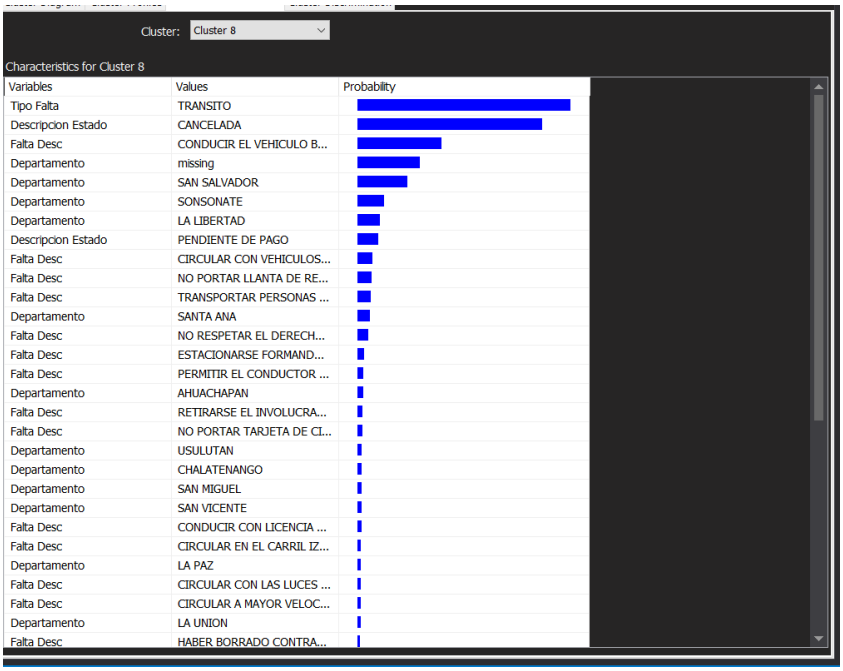
f) Clúster 6: Falta de transito cancelada por no portar llanta de repuesto



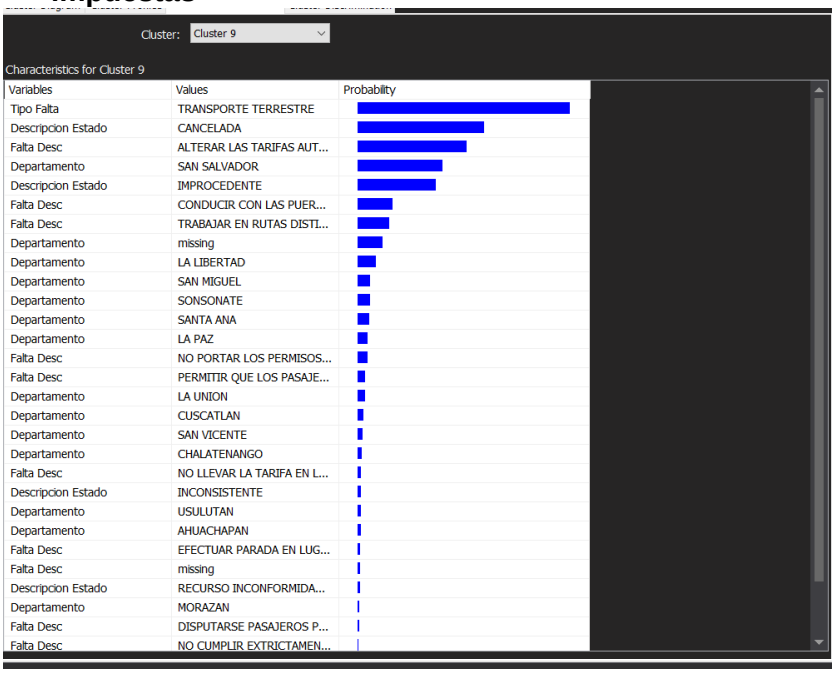
g) Clúster 7: Falta de transito cancelada por circular a exceso de velocidad



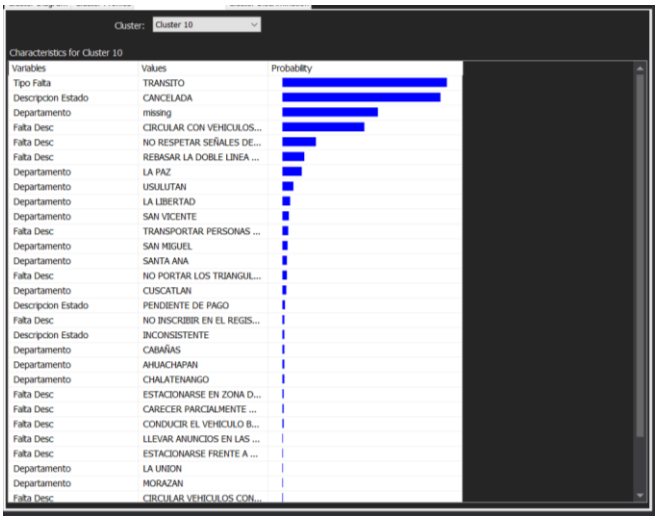
h) Clúster 8: Falta de transito cancelada por conducir bajo los efectos del alcohol



i) Clúster 9: Falta de transporte terrestre cancelada por alterar las tarifas impuestas



j) Clúster 10: Falta de transito cancelada por circular con vehículos no matriculados.


Cluster:
Cluster 10
Characteristics for Cluster 10
Variables
Values
Probability
Tipo Falta
TRANSITO
Descripcion Estado
CANCELADA
Departamento
missing
Falta Desc
CIRCULAR CON VEHICULOS...
Falta Desc
NO RESPETAR SEÑALES DE...
Falta Desc
REBASAR LA DOBLE LINEA ...
Departamento
LA PAZ
Departamento
USULUTAN
Departamento
LA LIBERTAD
Departamento
SAN VICENTE
Falta Desc
TRANSPORTAR PERSONAS ...
Departamento
SAN MIGUEL
Departamento
SANTA ANA
Falta Desc
NO PORTAR LOS TRIANGUL...
Departamento
CUSCATLAN
Descripcion Estado
PENDIENTE DE PAGO
Falta Desc
NO INSCRIBIR EN EL REGIS...
Descripcion Estado
INCONSISTENTE
Departamento
CABAÑAS
Departamento
AHUACHAPAN
Departamento
CHALATENANGO
Falta Desc
ESTACIONARSE EN ZONA D...
Falta Desc
CARECER PARCIALMENTE ...
Falta Desc
CONDUCIR EL VEHICULO B...
Falta Desc
LLEVAR ANUNCIOS EN LAS ...
Falta Desc
ESTACIONARSE FRENTE A ...
Departamento
LA UNION
Departamento
MORAZAN
Falta Desc
CIRCULAR VEHICULOS CON...