Inteligencia Artificial

Práctica 2

Análisis y Predicción de Datos

Realizado por:
Alberto González Fernández, Nº EXP 21719077

Índice

Contexto

Los accidentes de tráfico son la primera causa de fallecimientos de jóvenes en España y desde hace tiempo se encuentra en el top diez de mayor número de muertes causadas en todo el mundo. Desde que el uso del automóvil comenzó a extenderse, este problema no ha hecho nada más que aumentar, por lo que, para frenarlo, en España, se hicieron numerosos cambios a lo largo de los años, el más notable fue la introducción del carnet por puntos en 2006, que, debido al miedo de los conductores de perder puntos y potencialmente el permiso completamente, bajó mucho el número de víctimas mortales en los siguientes años hasta quedar justo por debajo de los dos mil anuales. Pero esto no es suficiente, ya que aun así este número puede ser reducido con más medidas.

Otro causante de la reducción del número de víctimas mortales es la organización europea EURONCAP, quien se dedica a hacer test de seguridad frente a distintos tipos de colisiones y dando una nota final del cero al cinco para informar a los posibles compradores si el vehículo es seguro en caso de accidente, produciendo un aumento en el concienciamiento de los fabricantes por la seguridad, y, aunque algunos modelos siguen suspendiendo el test, el número de casos en lo que esto pasa ni roza los número cuando se creó esta organización.

Por último, otro gran factor en la ayuda a la prevención es la mejora y el aumento de la infraestructura de carreteras, ya que hace años la gran mayoría de carreteras eran comarcales y secundarias, mientras que ahora, con el aumento del número de autopistas y autovías se ha aumentando la seguridad, y, aunque las carreteras secundarias son menos utilizadas en la actualidad, se siguen mejorando para evitar cualquier posible incidente debido a este motivo.

Objetivo

El objetivo de este estudio es, haciendo uso de una base de datos que recoge información de hasta dos millones de accidentes en los Estados Unidos, aplicar el análisis a realizar a España, ya que no hay este tipo de información pública en el país.
Se ha usado una base de datos de los estados unidos ya que aunque se encuentre en otro continente, es otro país occidental con un contexto muy equiparable al contexto de españa, por lo que con lo que se analize se podrán sacar conclusiones aplicables a cualquier país occidental.

Base de Datos

La base de datos recoge dos millones de datos de accidentes en Estados Unidos. Se divide en cuatro tipos de datos, que serán los que se analizan posteriormente.

El primero es la severidad, del uno al cuatro cuanto más grande el valor mayor impacto sobre el tráfico, siendo uno un accidente que no produce ningún tipo de retención, mientras que cuatro se trata de una retención de más de dos horas.
El segundo es la localización del accidente, ciudad, estado, etc.
El tercero son las condiciones meteorológicas, temperatura, humedad, clima, etc.
El último es la condición de la carretera, si el accidente se produce en un stop, un semáforo, un cruce, etc.

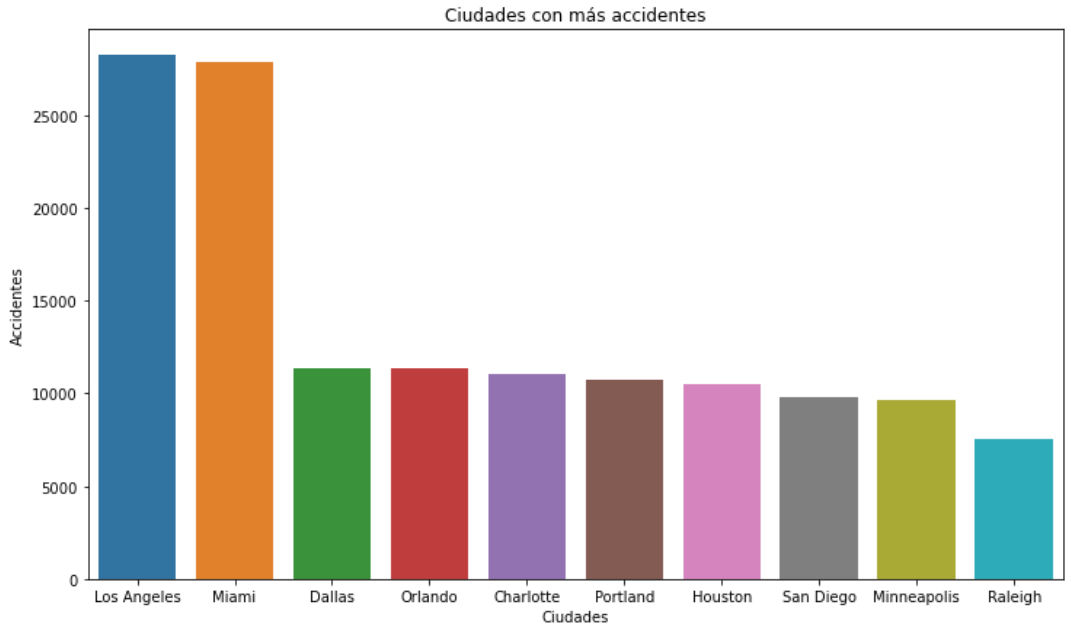
Cabe destacar que la base de datos solo almacena datos sobre circunstancias ajenas al conductor, ya que si se quisieran registrar datos sobre el coche se necesitaría la telemetría del propio vehículo, cosa que es muy difícil ya que es imposible tener acceso a los datos de todos los vehículos en circulación, por lo que la velocidad o el giro del coche no se tendrán en cuenta en este estudio. Por otra parte, los datos del propio conductor tampoco se pueden obtener, ya que se tendrían que hacer test a los conductores antes y después del accidente, otra vez algo imposible ya que hacer experimentos de este tipo es totalmente ilegal, si el conductor se despista o esté en estado en embriaguez son datos de los que no se disponen.
Por tanto, el estudio solo se realizará sobre circunstancias ajenas al conductor.

Estos datos han sido obtenidos desde la página web de Kaggle.

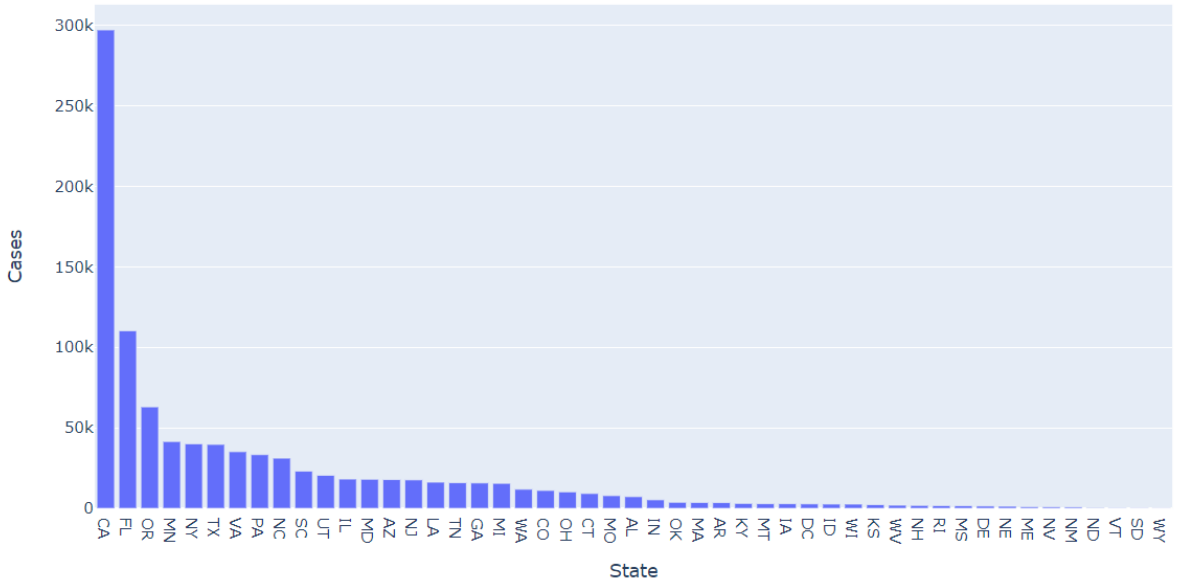
Análisis

Para seguir el análisis se recomienda tener cerca el código.
El primer paso para realizar el análisis es cargar el archivo con los datos haciendo uso de la librería *pandas*, ya que se hará uso de dataframes para la realización de la práctica. Se hace un drop de las columnas de las que no se van a hacer uso y se dropean las filas con algún dato vacío, dejando 977072 accidentes. Una vez hecho esto, ya se puede comenzar el análisis.

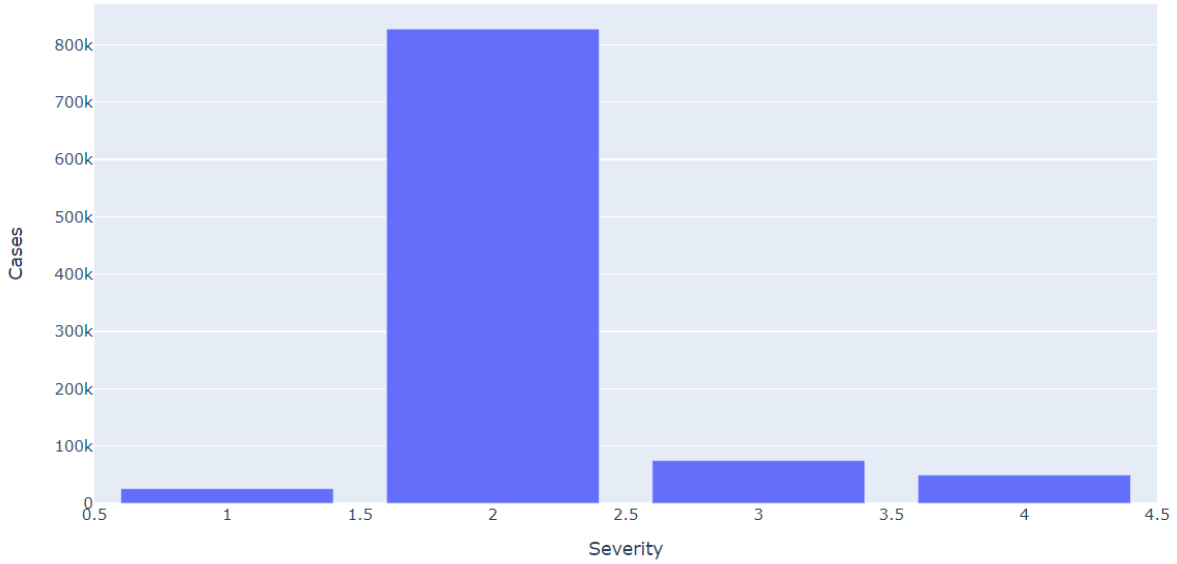
Lo primero que se hace es calcular el top diez de ciudades con mayor número de accidentes, información muy importante para poder focalizar posibles soluciones en el contexto de estas poblaciones. Para esto, se coje las ciudades y sus accidentes, se obtienen los diez primeros, y se añaden en un gráfico de barras de la librería *seaborn*.
Podemos observar que hay una gran diferencia del nuemro de accidentes de Los Angeles y Miami del resto.



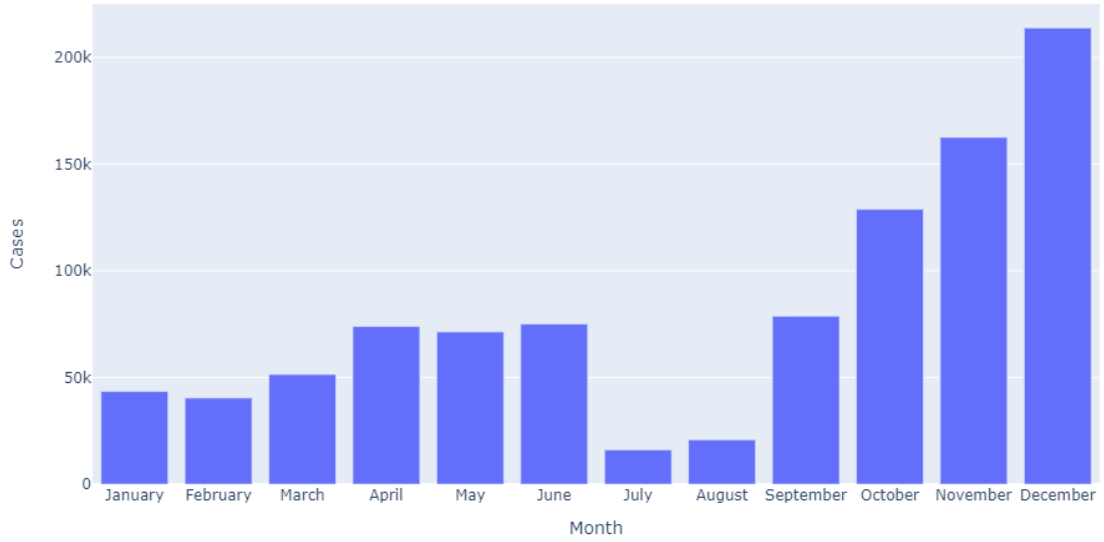
Para el resto de gráficos se ha usado *plotly express* dentro de la librería de *plotly*, por lo que el formato es diferente.
Se han obtenido el número de accidentes por estado.
Los estados más accidentados son California y Florida, cosa que tiene mucho sentido ya que sus capitales estado coronan el top diez de ciudades mas accidentadas.



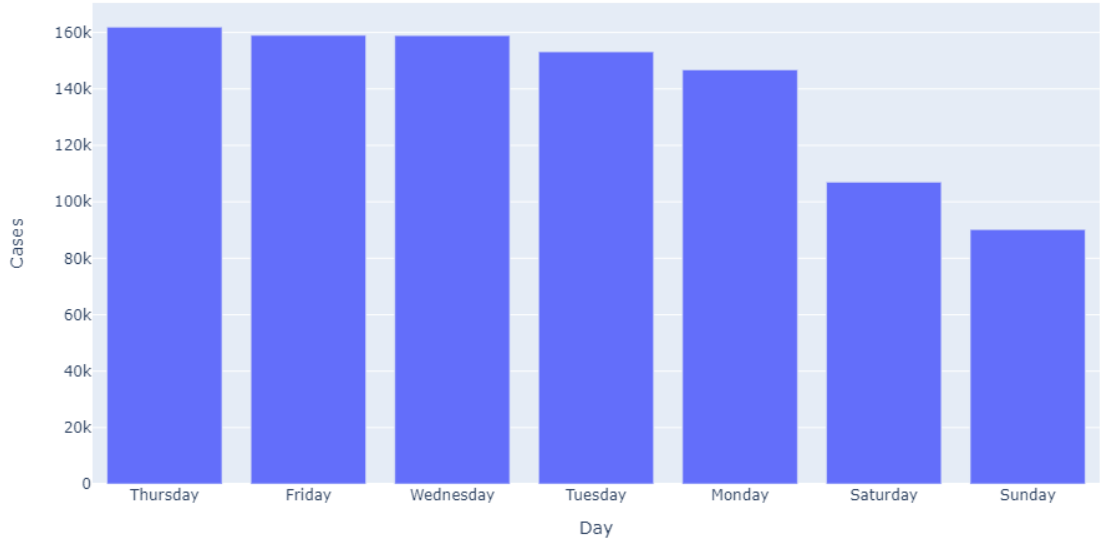
La distribución de casos por cada tipo de severidad.
La gran mayoría son de una severidad leve, retenciones de entre diez y treinta minutos, el mismo tipo de retención que se puede encontrar en cualquier carretera española.



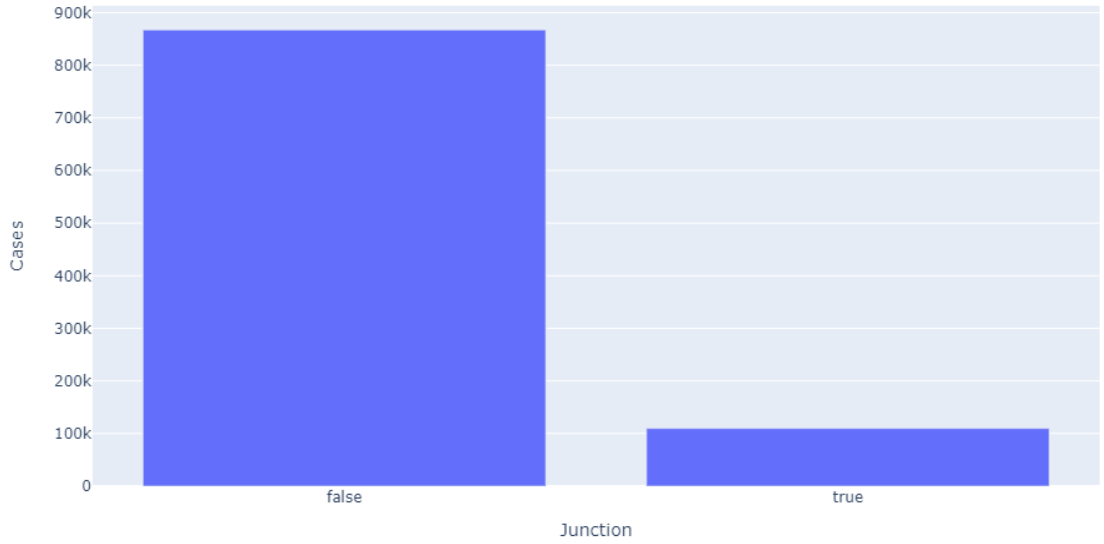
El número de casos por cada mes del año.
El último tercio del año es el más accidentado.

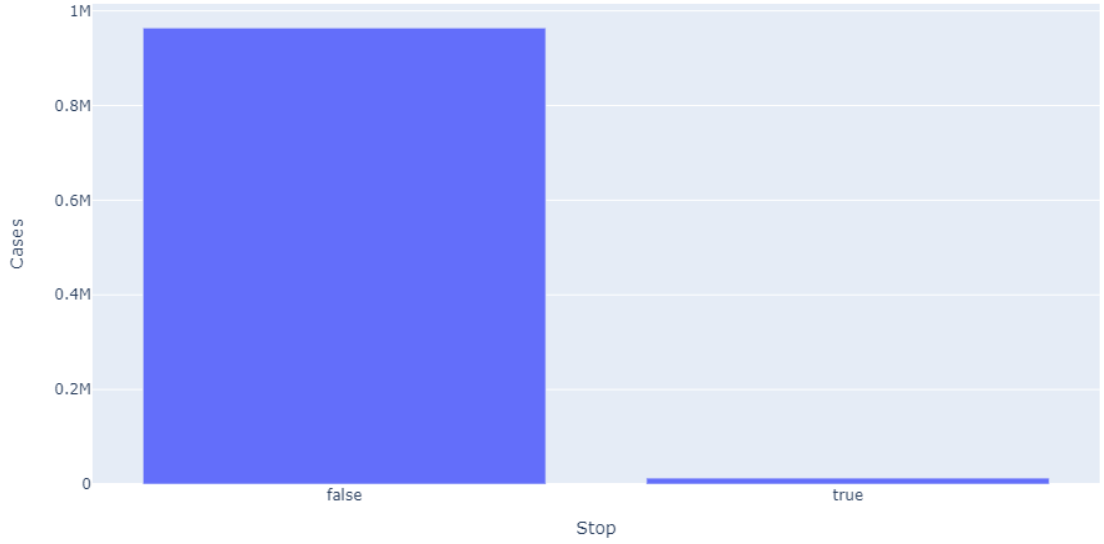


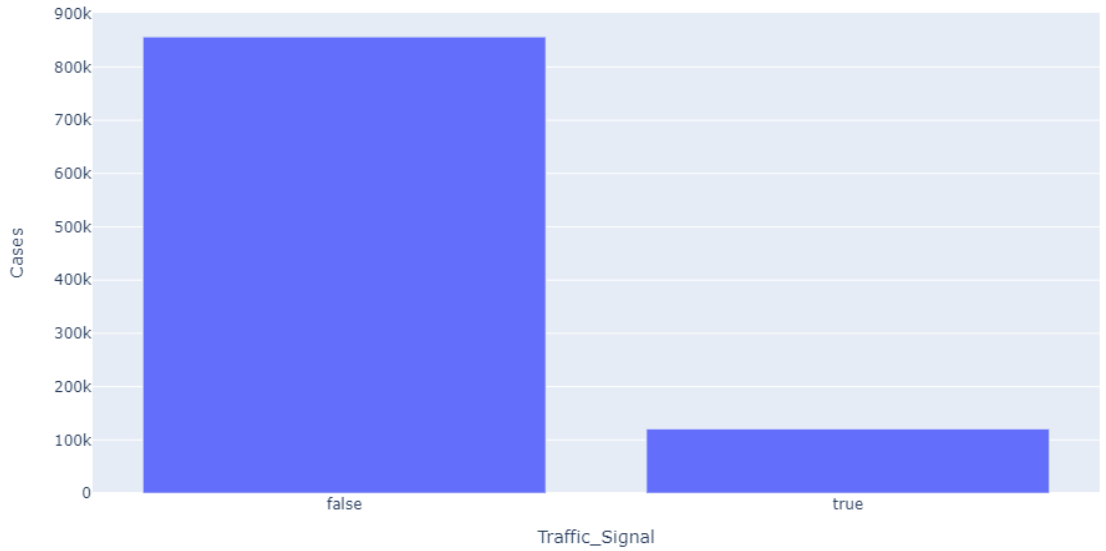
Accidentes cada día de la semana.
El mayor número de accidentes se produce en días laborables, ya que la gran mayoría de conductores coinciden a la misma hora en la carretera.

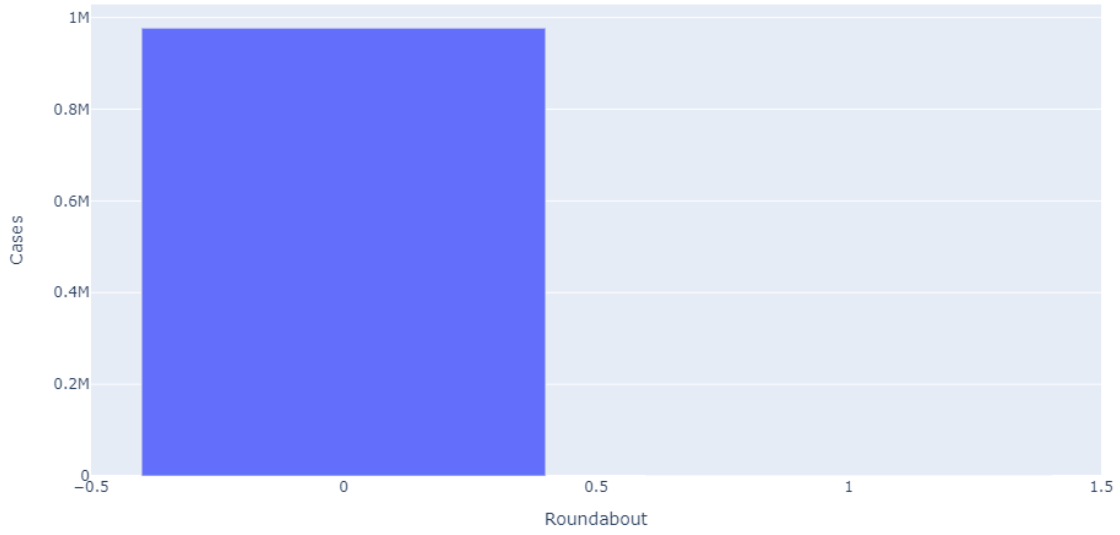


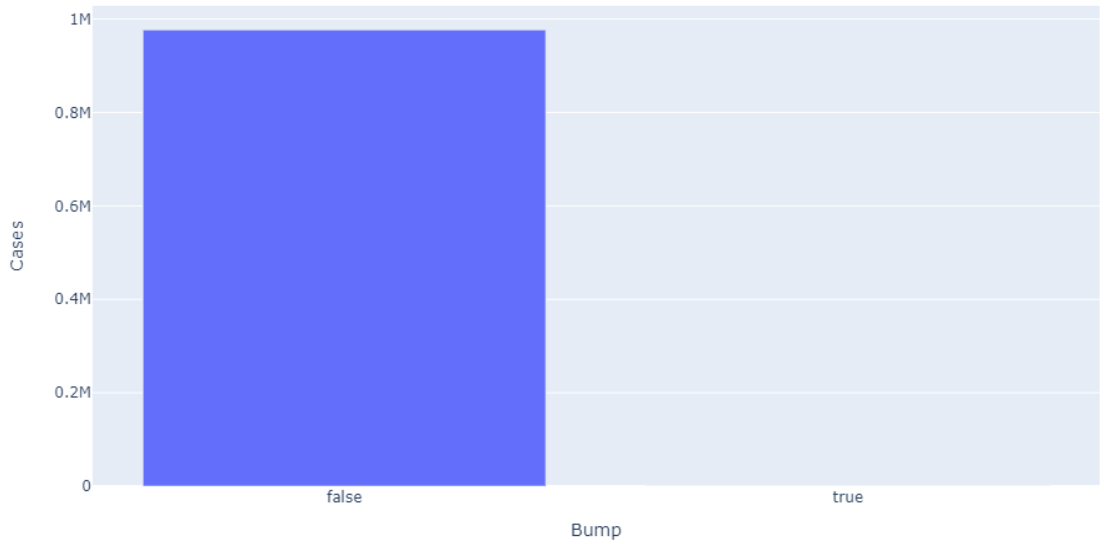
Una vez graficado esto, lo siguiente son las diferentes condiciones de la carretera.

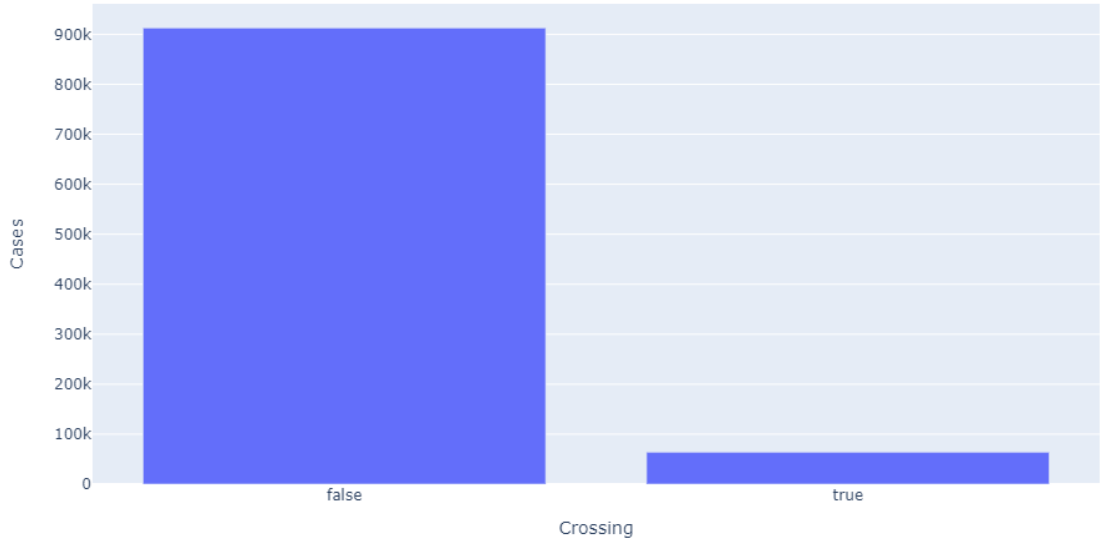

900k
800k
700k
600k
500k
400k
300k
200k
100k
0
Cases
false
true
Junction

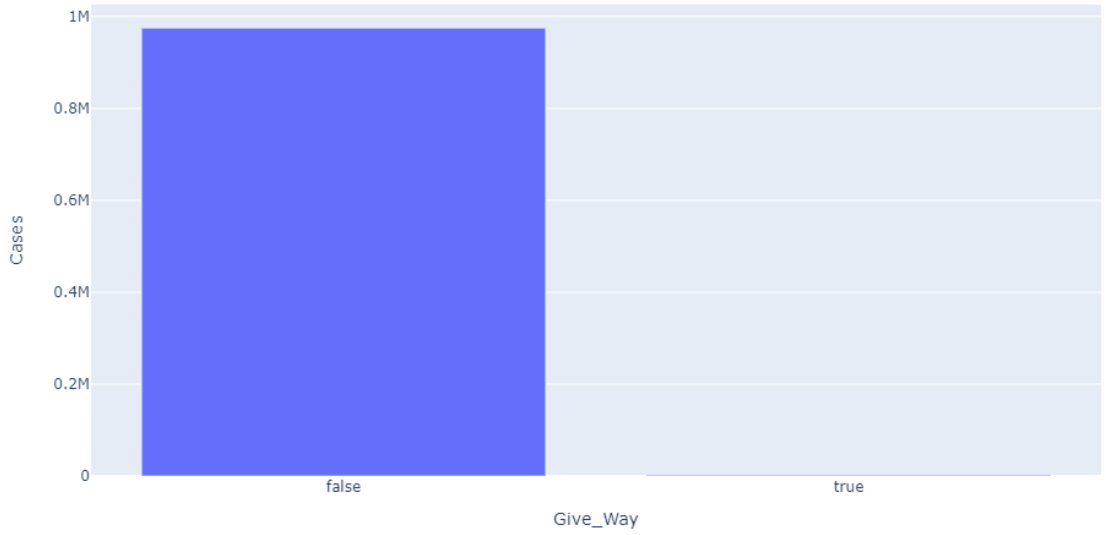

1M
0.8M
0.6M
0.4M
0.2M
0
Cases
false
true
Stop


900k
800k
700k
600k
500k
400k
300k
200k
100k
0
Cases
false
true
Traffic_Signal


1M
0.8M
0.6M
0.4M
0.2M
0
Cases
−0.5
0
0.5
1
1.5
Roundabout

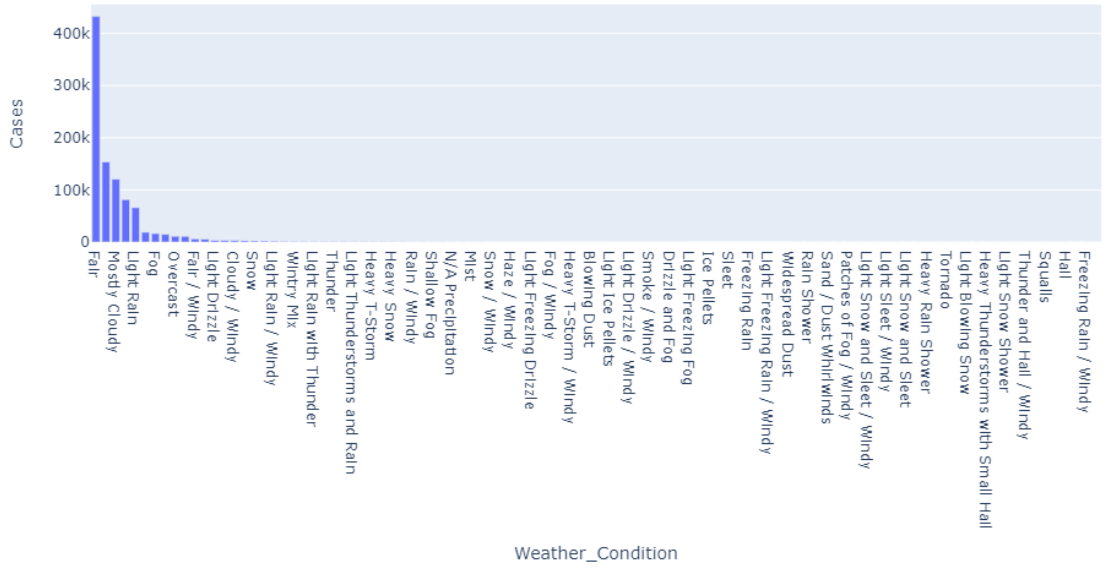

1M
0.8M
0.6M
0.4M
0.2M
0
Cases
false
true
Bump

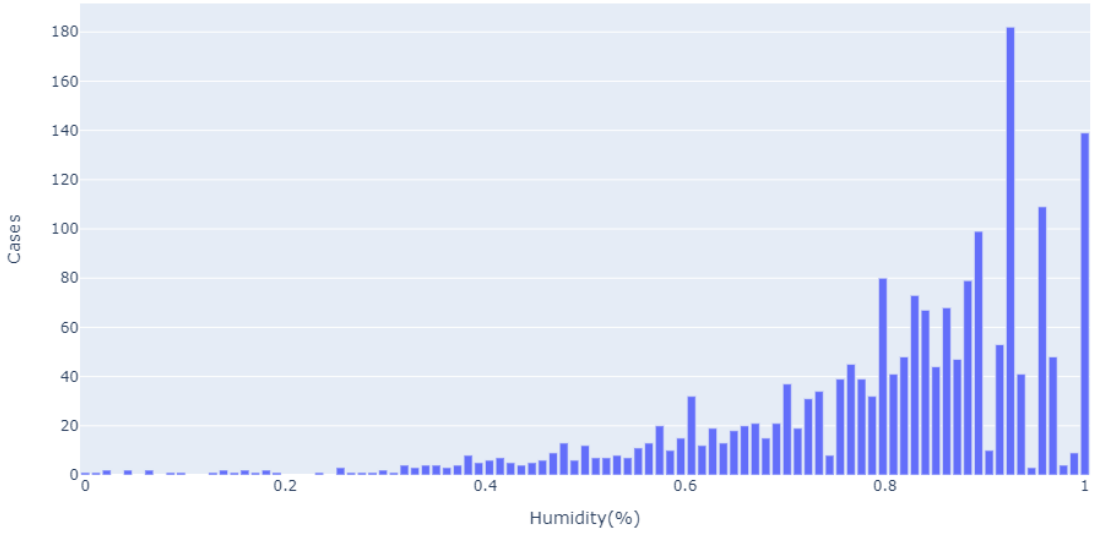




Como se puede observar, muchos de estos accidentes no se producen en ningún tipo de señal ni situación debido a la carretera, pero cabe destacar que en los semáforos y los stops hay un mayor porcentaje de incidentes en comparación con el resto.

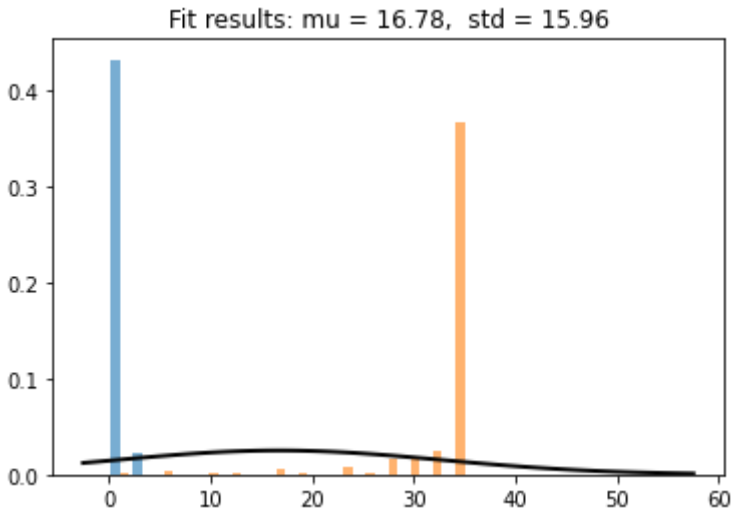
Por último, se analizan las situaciones meteorológicas.



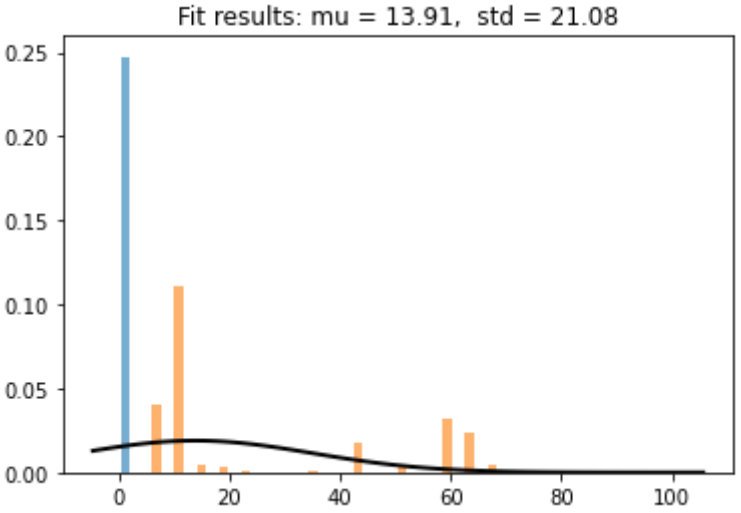


Como se puede observar, la condición meteorológica no tiene mucha relación con los accidentes, ya que la gran mayoría se producen con un clima soleado.
Respecto a la humedad, si tiene un factor en los accidentes, ya que puede producir que la carretera esté más resbaladiza, y por tanto, sea más propio a que se produzca algún incidente.
Cabe destacar que también se estudian la velocidad del viento y las precipitaciones, pero como hay registros de valor tan grande, la gráfica debe ser ampliada para observar los datos, sino la gráfica parece estar vacía.

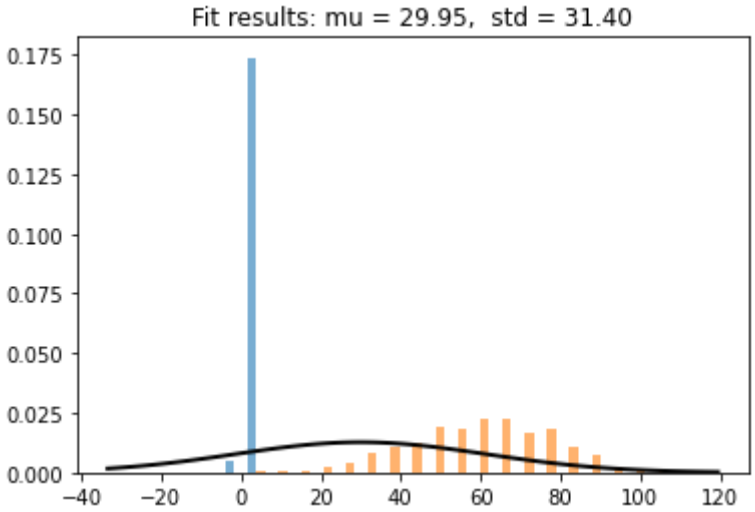
Una vez hechas todas las gráficas, se procede a realizar la comparación de todos estos posibles factores a la severidad, siendo las barras azules en la severidad y las naranjas el otro factor a comparar.



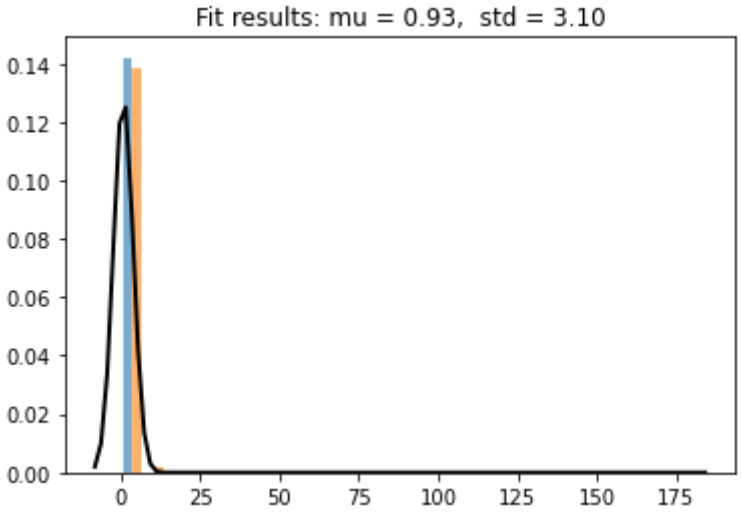
La visibilidad tiene muy poca relación con la severidad ya que la gran mayoria de los accidentes tienen una visibilidad perfecta de mas de treinta millas en adelante.



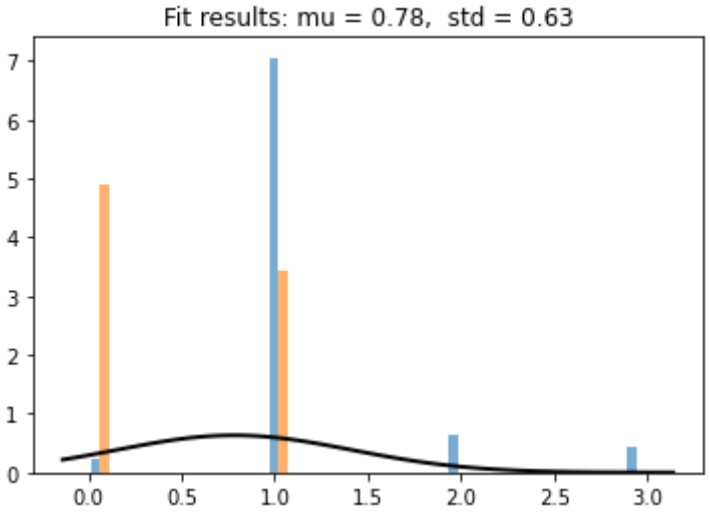
Como se ha visto anteriormente, el clima no tiene relación con la severidad, puesto que la gran mayoría de accidentes se han producido con un clima soleado.



La temperatura no tiene relación con la severidad, ya que la gran mayoría se encuentran en la media de entre cincuenta y noventa grados fahrenheit, ya que es una base de datos estadounidense.

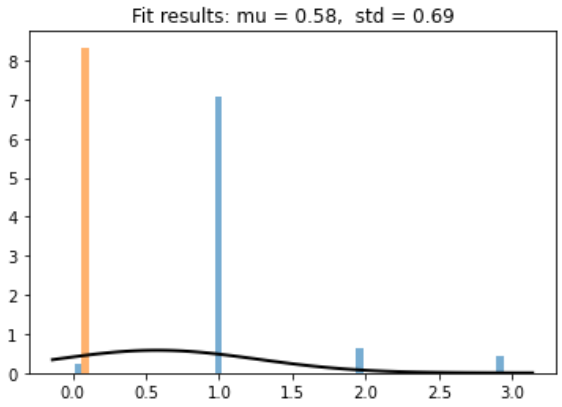


La gran mayoría de los accidentes se producen cuando no hay precipitaciones, por lo que tiene mucha relación la severidad con la falta de precipitaciones, algo que según el conocimiento común se contradice, ya que de normal se piensa que cuando llueve hay más posibilidades de accidente, cosa que la comparación nos dice que es totalmente falso.

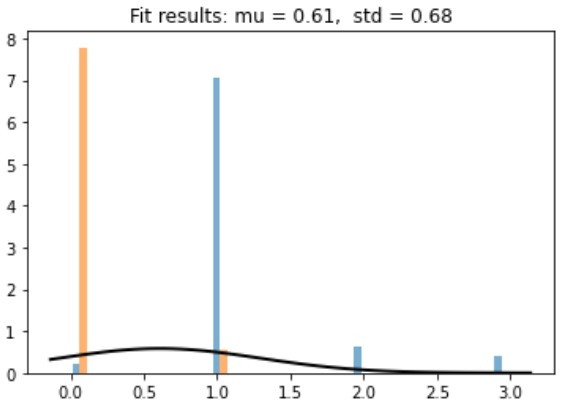


Respecto al momento del día hay dos posibilidades, cero siendo día y uno siendo noche. Como se puede observar, se producen más accidentes de día que de noche, ya que hay más gente en la carretera, y tiene relación con la severidad.

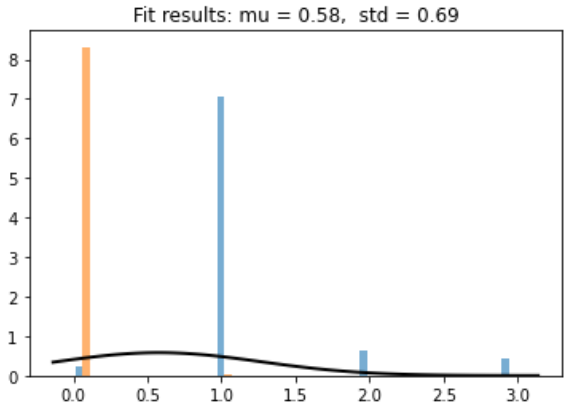
Para finalizar nuestro estudio, se realizará la comparación entre la severidad y las diferentes condiciones en la carretera.



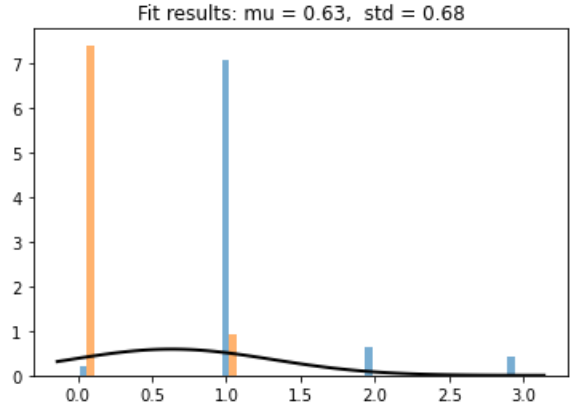
Bache



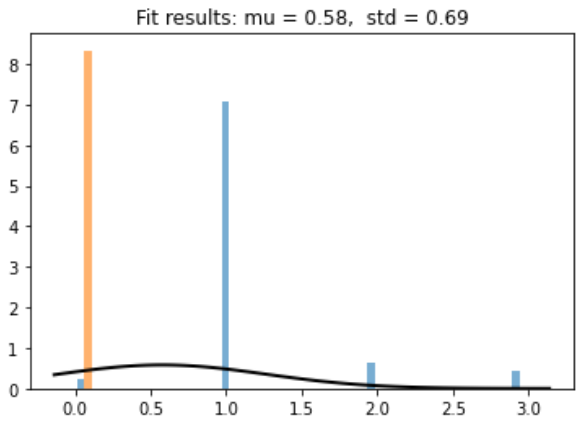
Cruce



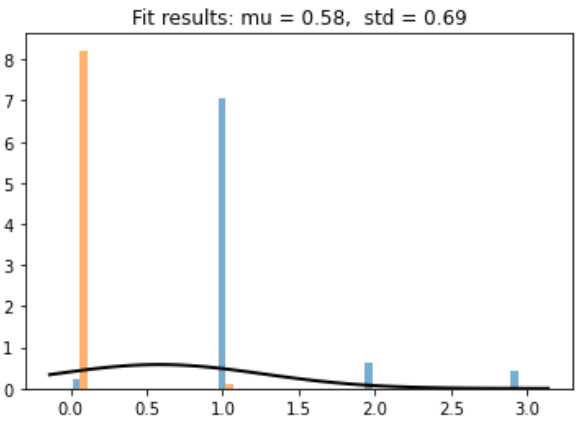
Ceda



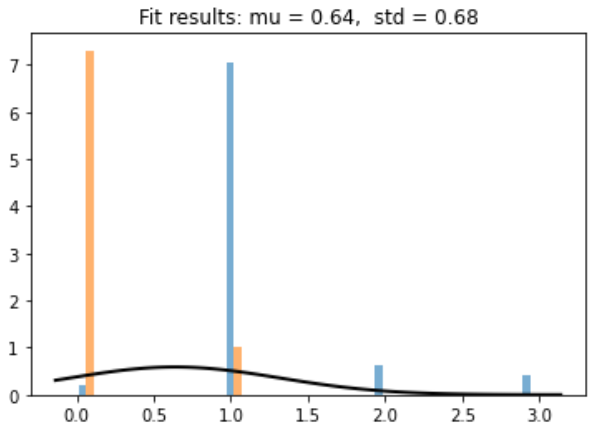
Salida de autopista



Rotonda



STOP



Semáforo

Una vez hecho todo el estudio, se puede concretar que de todas las posibilidades que un conductor se puede encontrar en la carretera mientras conduce, los momentos en los que se debe parar completamente el vehículo, ya sea un STOP o un semáforo, es donde más posibilidad hay que se produzca un accidente. Esto puede deberse a que el conductor se distrae o no frena por algún motivo.
También se puede observar que las salidas y entradas de autopistas tiene bastante posibilidad de accidente, ya uqe hay conductores que van a gran velocidad y que, para tomar una salida se espera hasta el último momento para salir y así adelantar a otros conductores, o a la hora de entrar en la autopista, algunos conductores pasar directamente al carril izquierdo, tratándose de esto una maniobra de gran peligrosidad.

Predicción

Para realizar la predicción, se hace un recorte de la base de datos, ya que al tratarse de dos millones de entradas, tarda en realizarse la tarea más de ocho horas, por lo que para realizar un ejemplo de predicción se usan 1935 entradas.
Antes de nada, se hace un cambio de tipo de variable de objeto o booleano a numeral, ya que de esta forma no da error.
Lo siguiente es buscar columnas que estén encaradas, en este caso se encuentra una, por lo que se hacen dummies para poder tratarla.
El próximo paso es realizar el preprocesado haciendo uso del *MinMax Scaler* de la librería *sklearn*. Se ha establecido un rango de cero a uno, por lo que todas las entradas pasan a tener de valor un número entre cero y uno.
Por último, haciendo uso de un *train*, cuatro variables, dos para el entrenamiento y dos para la predicción. La distribución será de un ochenta por ciento para el entrenamiento, y el veinte restante para realizar la predicción.
Cargaremos en un array todos los modelos que se utilizaran para compararlos. Se usarán los modelos más comunes, *Regresión Logística*, *Random Forrest*, *SVM*, *KNN*, *Decision Tree*, *Gaussian Naive Bayes*.
Para finalizar, se hará un proceso del array, haciendo entrenamiento y predicción sobre cada modelo contenido. Una vez acabado, se mostrará una tabla con el nombre de cada algoritmo, junto a su porcentaje de acierto y la desviación media. También hay otros dos valores contenidos en la tabla pero por algún error que no ha sido posible descubrir aparecen en la tabla como valores vacíos.

Una vez finalizada la predicción, podemos concretar que *Decision Tree* es el modelo más adecuado, con un cien por cien de acierto, cercanamente seguido de *Regresión Logística*, *Random Forest* y *Gaussian Naive Bayes*, con un porcentaje de acierto mayor al noventa y cinco por ciento. Los otros dos modelos *SVM* y *KNN* no son los modelos adecuados para esta predicción, ya que tienen un porcentaje de acierto demasiado bajo, siendo setenta y ochenta y siete por ciento respectivamente.

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
0	Logistic Regression	NaN	NaN	97.61	1.08
1	Random Forest	NaN	NaN	96.32	1.99
2	SVM	NaN	NaN	70.35	2.93
3	KNN	NaN	NaN	87.08	2.06
4	Decision Tree Classifier	NaN	NaN	100.00	0.00
5	Gaussian NB	NaN	NaN	95.80	1.01

Conclusiones

Tras realizar el estudio y análisis se pueden sacar como conclusiones que el mayor número de accidentes se producen en el desplazamiento al trabajo debido a la gran acumulación de conductores. Las conducciones ajenas al conductor no son tan influyentes y si se pudiesen obtener los datos del vehículo y conductor se podrá buscar una mejor solución al problemas, pero aun así, con los datos que se tienen, se podría buscar alguna medida para mejorar la situación actual.