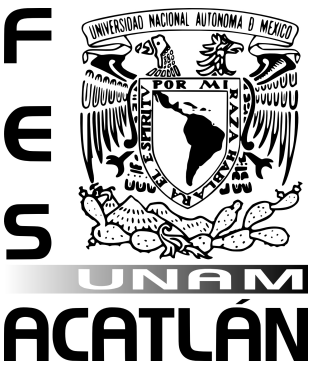


Universidad Nacional Autónoma de México
Facultad de Estudios Superiores Acatlán
Diplomado en Ciencia de Datos

Impacto de los accidentes viales en el flujo de tránsito - Estados Unidos

- Proyecto -

Autor:
Colmenares Apodaca Flavio Alberto

Profesores:

Pineda Rodríguez Lorena
Acosta González Oscar Daniel

Fecha,
Mayo 2021

Índice

Índice de figuras

Índice de cuadros

1. Introducción

Los accidentes viales son parte de la cotidianidad de cualquier conductor o transeúnte, un hecho que tiene impacto en aspectos como defunciones, daños a terceros, retraso en el fujo del transporte privado y público.

Además de tener un gran influencia económica generando millones de pérdidas cada año, surgiendo la necesidad de programas preventivos para disminuir los incidentes de tránsito buscando mejorar las condiciones en las vialidades, por ello los modelos de predicción de accidentes juegan un papel importante en implementar acciones de manera eficiente destinando recursos humanos y económicos en zonas de alto riesgo.

Figura 1: Accidentes Mensuales

2. Planteamiento

Los accidentes viales son particularmente interesantes, debido a que se pueden desarrollar diversos estudios entre los que destacan:

- Detectar zonas con alto índice de accidentes
- Determinar causas y efectos en los accidentes
- Predicción de accidentes en tiempo real

El presente trabajo busca analizar el impacto posterior al accidente, como es afectado el flujo del tránsito, así como, de ser posible determinar patrones como pueden ser zonas, distancia recorrida, condiciones atmosféricas, y como influyen en los percances viales.

2.1. Objetivos

Para este proyecto nos centraremos en los siguientes objetivos:

- Estructurar y visualizar los datos de tal forma que nos apoye con la comprensión de la tabla, para poder profundizar en el análisis posterior
- Encontrar las zonas y condiciones que causan mayor severidad[1] en accidentes automovilísticos
- Encontrar patrones en los datos, los cuales, nos permitan segmentar a nuestra población según las características en cada incidente de tránsito

[1]Severidad será considerado como el impacto que tuvo el accidente, en el retraso del tránsito vehicular.

3. Conjunto de Datos

Para este análisis tenemos una tabla que cubre accidentes automovilísticos en 49 estados pertenecientes a Estados Unidos, en un periodo que abarca desde Febrero 2016 hasta Junio 2020.

Estos fueron extraidos a partir de API's que nos proveen Bing y MapQuest.

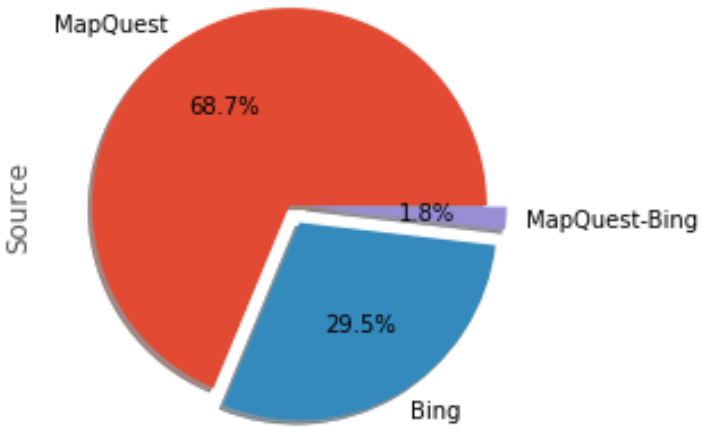


Figura 2: Datos obtenidos por API

3.1. Descripción y Visualización

Nuestra tabla consta de 3,513,617 registros y 49 variables, las cuales podemos seccionar de la siguiente manera:

- Variable Objetivo
- Variables Continuas
- Variables Discretas o Categóricas

- Variable de Texto
- Variables Geográficas

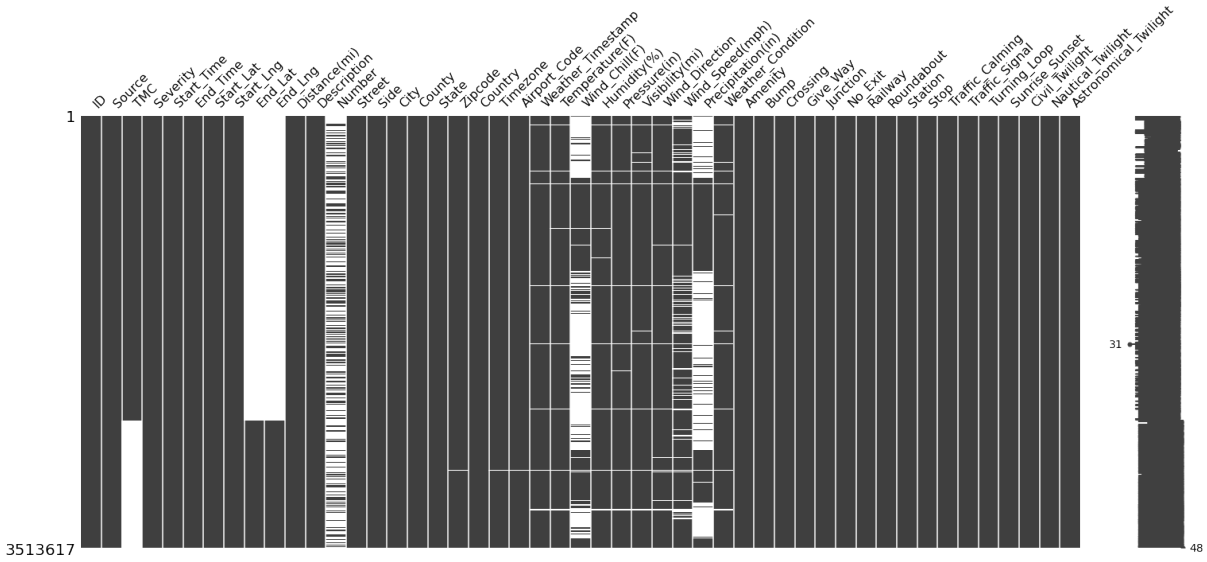


Figura 3: Variables y valores faltantes

3.1.1. Variable Objetivo

Nuestra Variable objetivo Severity nos dice el retraso en el tránsito consta de 4 niveles [1,2,3,4]

- 1: la menor afectación
- 4: la mayor afectación

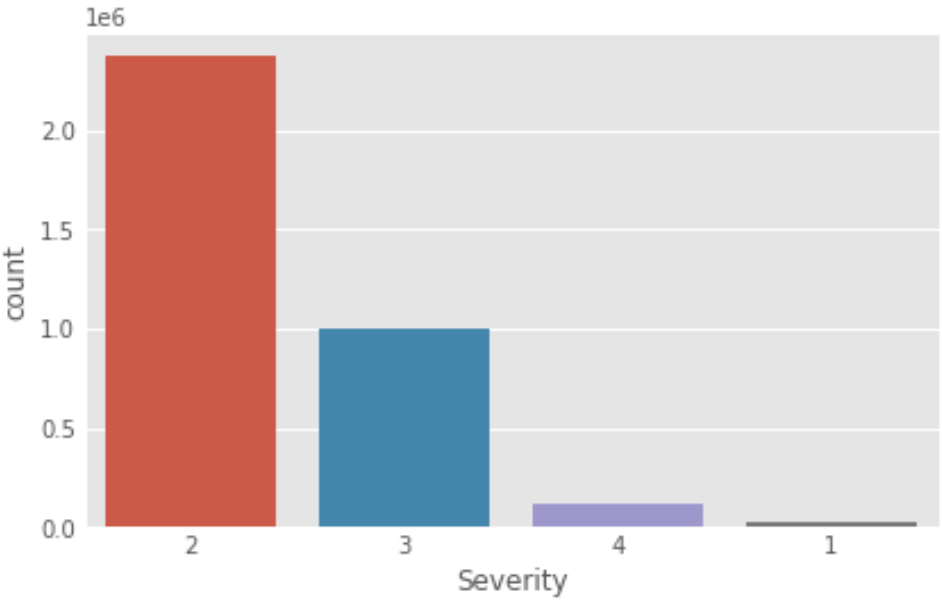


Figura 4: Distribución de la Severidad

3.1.2. Variables Continuas

Las variables continuas constan en su mayoría de condiciones atmosféricas:

- Temperature(F): Temperatura en Farenheit
- WindChill(F): Temperatura del aire en Farnheit

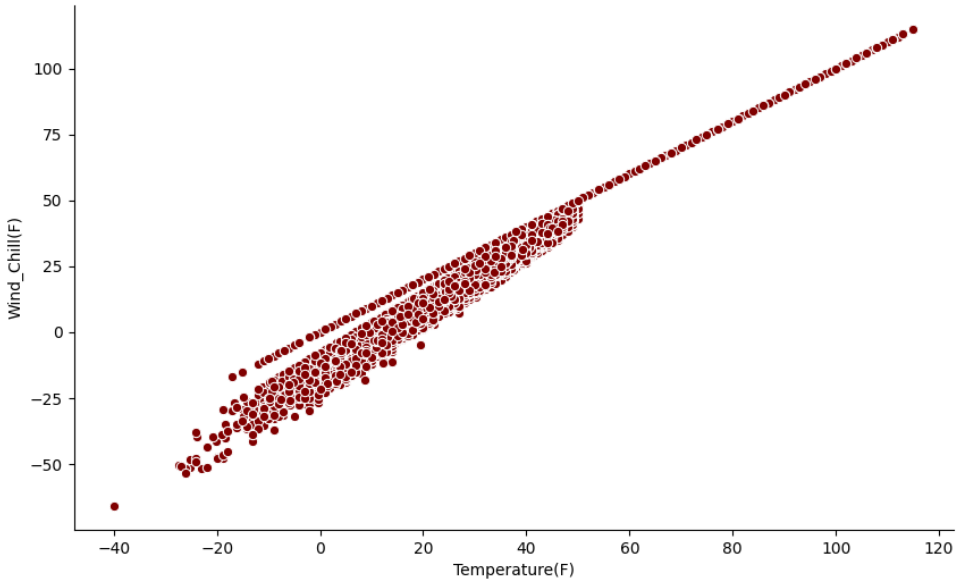


Figura 5: WindChill vs Temperature

- Humidity: Porcentaje de humedad
- Preassure(in): Presión atmosférica en pulgadas
- Visibility(mi): Visibilidad en millas
- Windspeed(mph): Velocidad del viento en millas por hora
- Precipitation(in): Precipitación en pulgadas

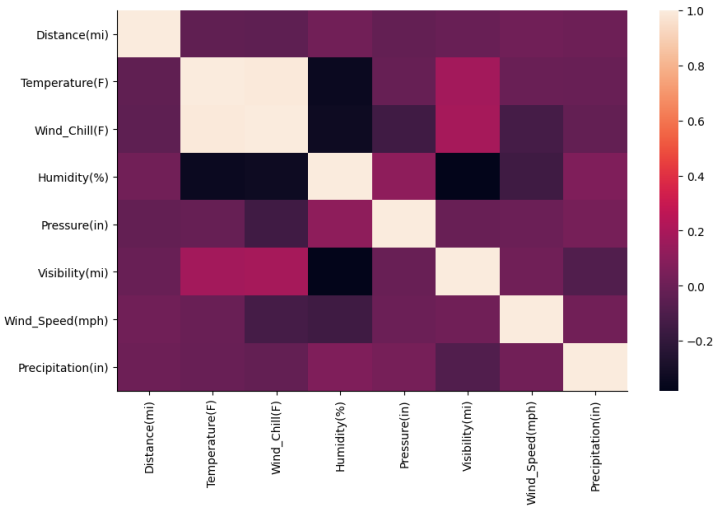


Figura 6: Correlaciones variables continuas

- Distance(mi): Espacio afectado por el retraso de tránsito (millas)

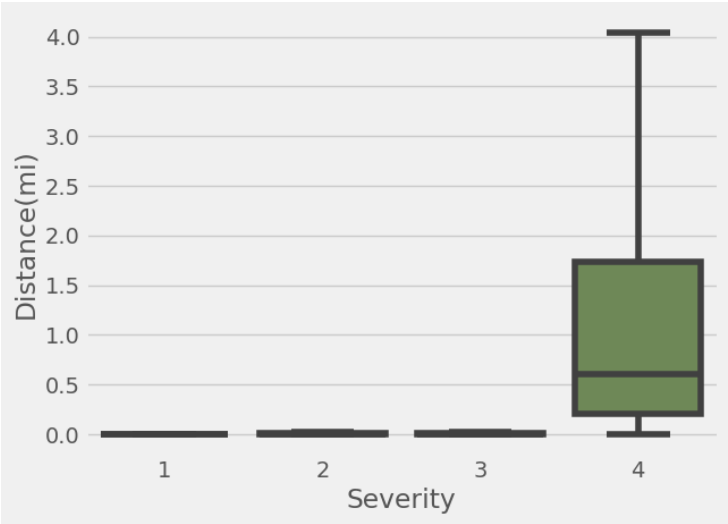


Figura 7: Distance vs Severity

3.1.3. Variables Discretas o Categóricas

En está parte contamos con 31 variables categóricas por lo que, de momento, nos enfocaremos en 3 de estas.

Variable	Descripción
State	Estado del registro de dirección
WeatherCondition	Tipo de Clima
Timezone	Zona horaria respecto a la locación

Cuadro 1: Variables Discretas

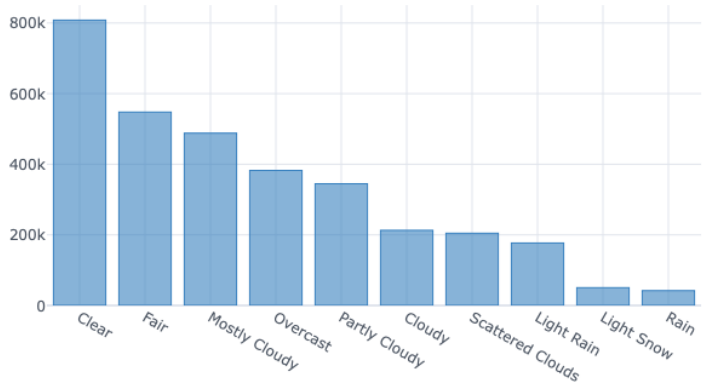


Figura 8: Top 10 tipos de clima en accidentes

3.1.4. Variable de Texto

En esta tabla contamos con la variable Description con la cual podemos extraer como fue reportado cada accidente.

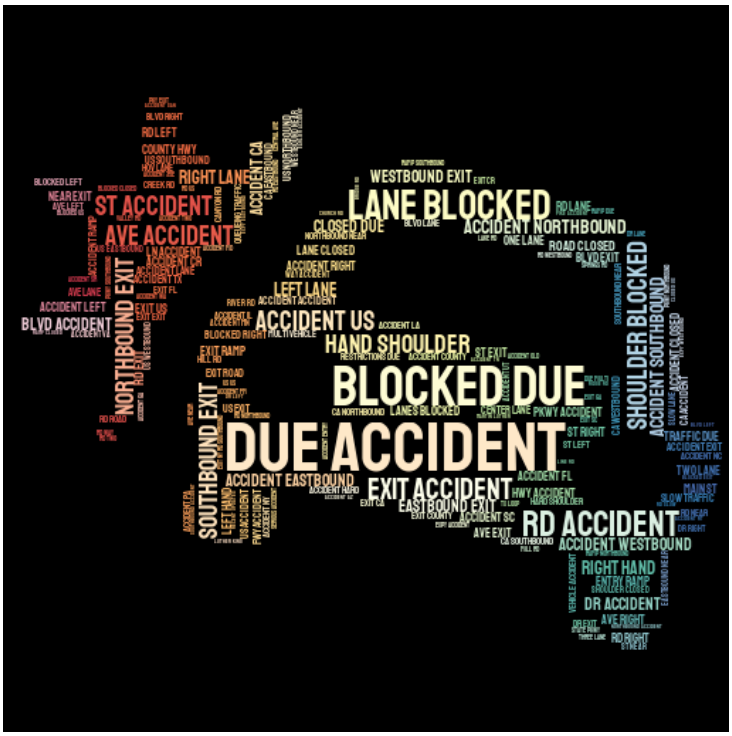


Figura 9: Frecuencias de la variable Description

3.1.5. Variables Geográficas

Con las siguientes variables se pretende detectar zonas con alto índice de accidentes y el impacto que tienen sobre la Severidad de los incidentes:

- StartLat: Latitud del inicio de la afectación de tránsito
- StartLng: Longitud del inicio de la afectación de tránsito
- EndLat: Latitud del final de la afectación de tránsito
- EndLng: Longitud del final de la afectación de tránsito

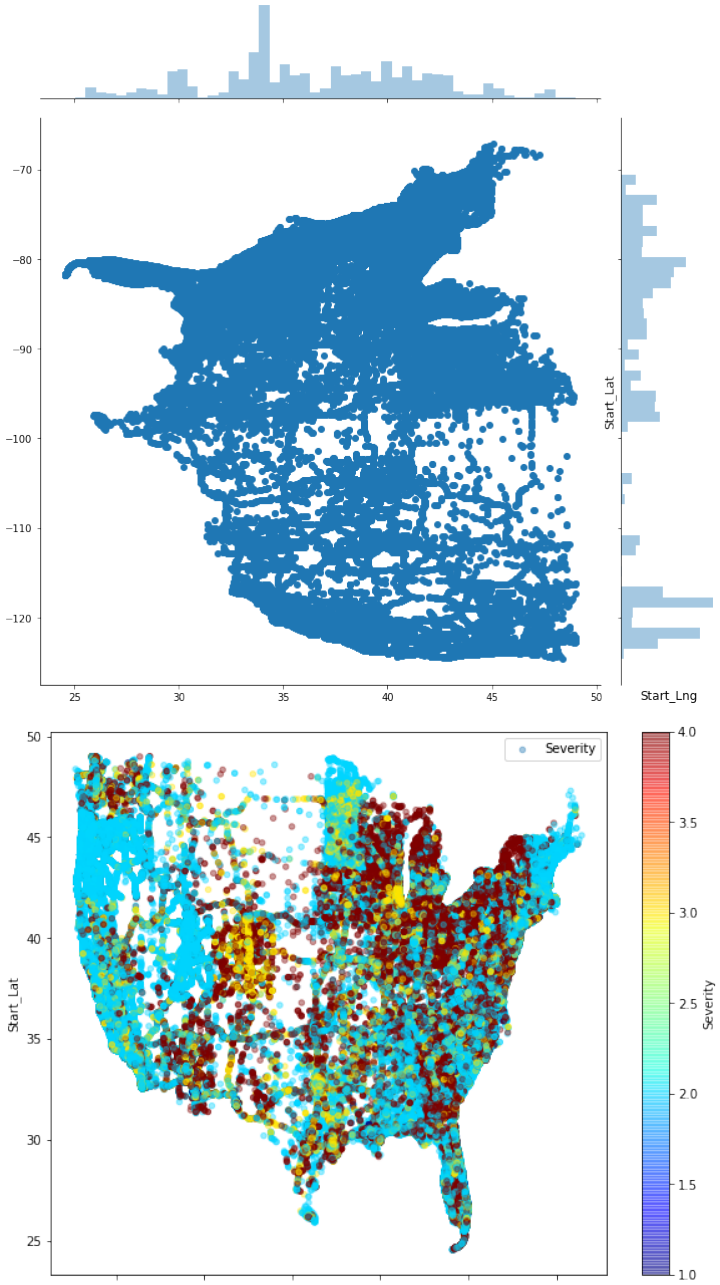


Figura 10: Distribución por latitudes y longitudes

4. Ingeniería de variables

Como vimos en el curso, el principal objetivo de la ingeniería de variables es aumentar la eficacia predictiva de nuestro modelo creando nuevas variables a partir de las anteriores y enriqueciendo nuestra tabla tomando nuevas fuentes de datos para este propósito.

Variable	Descripción
TimeMinutes	Tiempo que duró la obstrucción vial en minutos
Hour	Hora en que inició la obstrucción vial
Weekday	Día de la semana
Month	Mes del incidente
CountyFreq	Accidentes ocurridos por Condado
StateFreq	Accidentes ocurridos por Estado
CityFreq	Accidentes ocurridos por Ciudad
AirportFreq	Accidentes ocurridos por Código de Aeropuerto

Cuadro 2: Variables Propuestas

4.1. Unidad Muestral

Para este proyecto la unidad muestral serían los accidentes automovilísticos reportados por MapQuest, tomando como elemento de nuestro estudio la severidad del accidente en el restraso del tránsito.

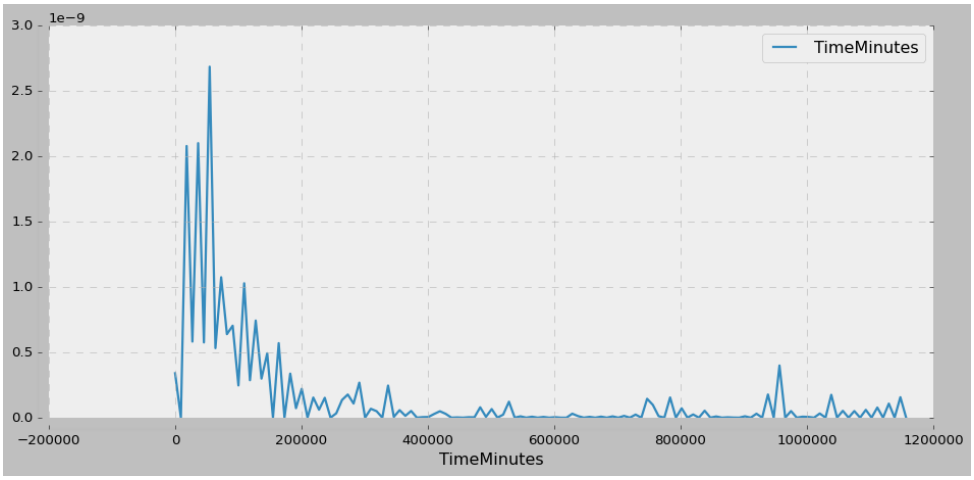


Figura 11: Tiempo hasta el final de la afectación en minutos

5. Tratamiento de Valores Atípicos

Un outlier es una observación anormal y extrema en los datos que puede afectar potencialmente a la estimación de los parámetros del mismo. Por lo que serán removidos por distintos métodos.

Variable	Distance(mi)	Visibility(mi)	WindSpeed(mph)	TimeMinutes
Mínimo	0	0	0	-35
Mediana	0	1.000000e+01	7	4.4
Máximo	3.336300e+02	1.400000e+02	9.840000e+02	1.421955e+06

Cuadro 3: Atípicos Detectados

5.1. Método Intuitivo

Este método solo lo ultizaremos para la variable TimeMinutes, con la cual nos dimos cuenta que en la tabla hay tiempos de inicio que son posteriores a los tiempos de termino de la afectación del tránsito por lo que se removerán los tiempos negativos de nuestra tabla.

5.2. Método Rango Intercuartil (IQR)

Este método se utilizó sobre las variables WindSpeed(mph) y Visibility(mi), ya que nos apoya con los outliers en otras variables y genera menos pérdida de información teniendo actualmente el .85 de nuestros datos.

Variable	Limite inferior	Limite superior
TimeMinutes	0	No aplica
WindSpeed(mph)	0	21.25
Visibility(mi)	0	10

Cuadro 4: Limites para remover outliers en el orden que fueron retirados

6. Análisis exploratorio

6.1. Limpieza

Para esta sección nos enfocamos en dos variables principalmente, Wind_Direction que limpiamos con ayuda de un diccionario que se muestra a continuación reduciendo los niveles de la misma.

Original	Diccionario
South	S
West	W
SSW	SW
WSW	SW
WNW	NW
NNW	NW

Cuadro 5: Fragmento diccionario de datos Wind_Direction

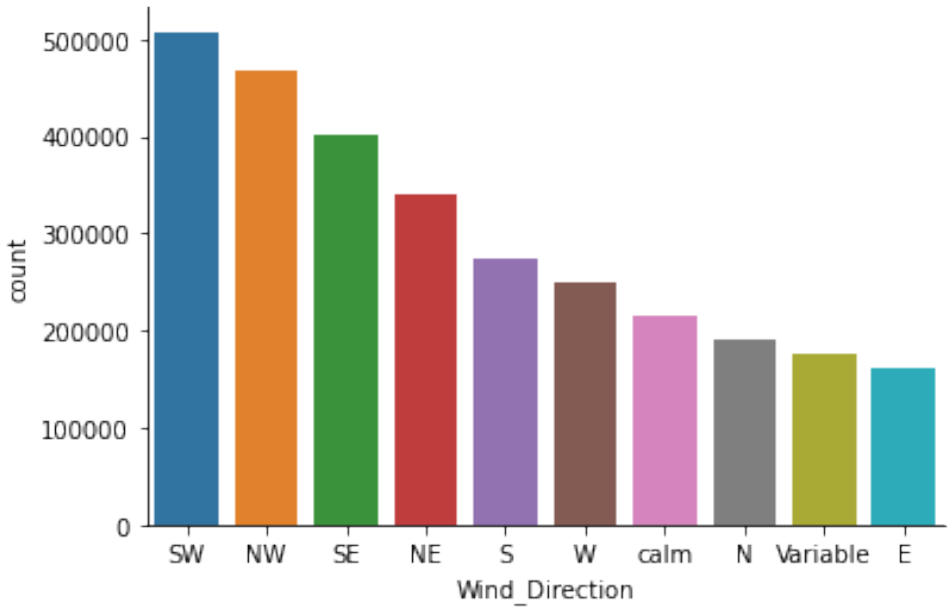


Figura 12: Frecuencias Wind_Direction

Para la variable Weather_Condition se agruparon por los climas que contenía cada registro, ej. si contenía la palabra cloud se agrupaba en cloudy.

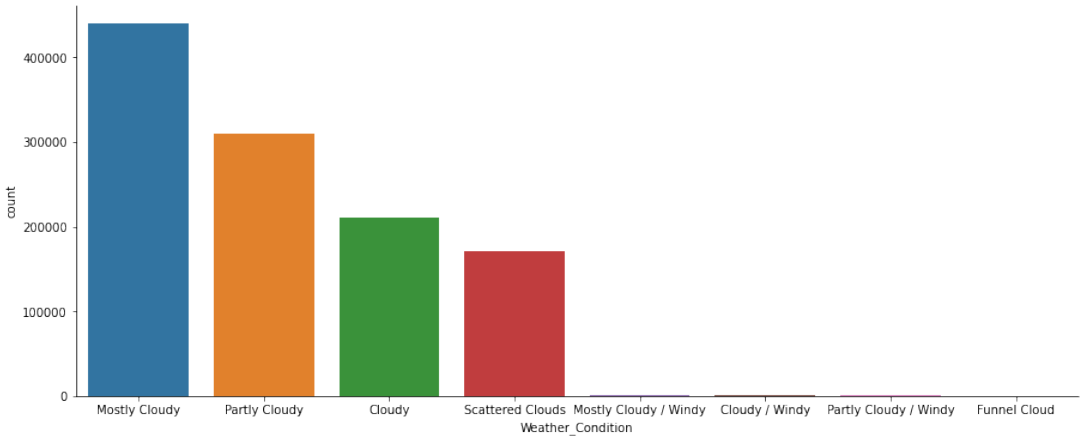


Figura 13: Registros agrupados en Cloudy

Esto debido a que teníamos 117 niveles en esta variable y gracias a esto fuimos capaces de agruparla a 10 niveles.

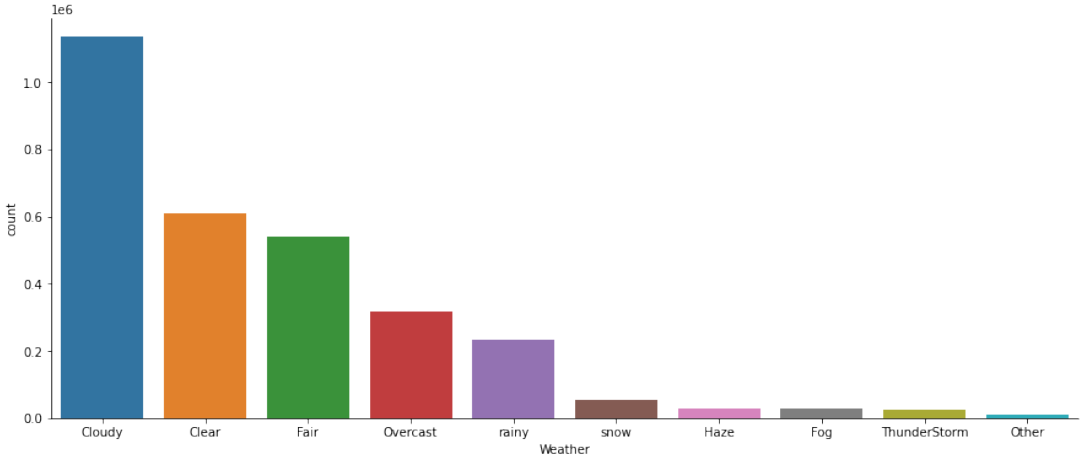


Figura 14: Frecuencia Climas agrupados

6.2. Escalas

Para la parte de las escalas, ocupamos el StandardScaler para que nuestras variables queden normalizadas para análisis posteriores.

Variable
Temperature(F)
Humidity(%)
Pressure(in)
Visibility(mi)
Wind_Speed(mph)

Cuadro 6: Variables Escaladas

6.3. Valores Ausentes

6.3.1. Univariado

Para realizar la imputación de las variables categóricas Side, Wind_Direction y Weather decidimos imputar por moda. Con lo cual nos quedó lo siguiente.

Variable	descripción	Moda
Side	El lado relativo donde ocurrió el accidente (L,R)	R
Wind_Direction	Dirreción en que soplaba el viento	SW
Weather	Clima del accidente	Cloudy

Cuadro 7: Imputación por moda

Figura 15: Frecuencias variable Side

6.3.2. Multivariado

Para realizar la imputación de las variables Humidity(%), Pressure(in) y Temperature(F). Todas previamente escaladas, optamos por usar el KNNImputer el cual mantuvo las distribuciones de nuestras variables.

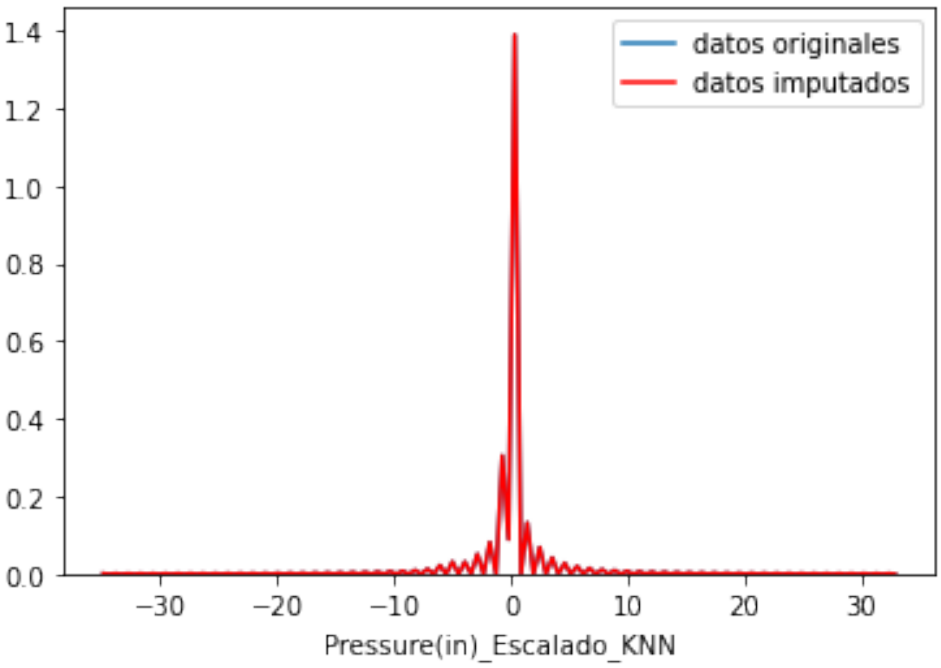


Figura 16: Valores Imputados vs Originales Pressure(in)

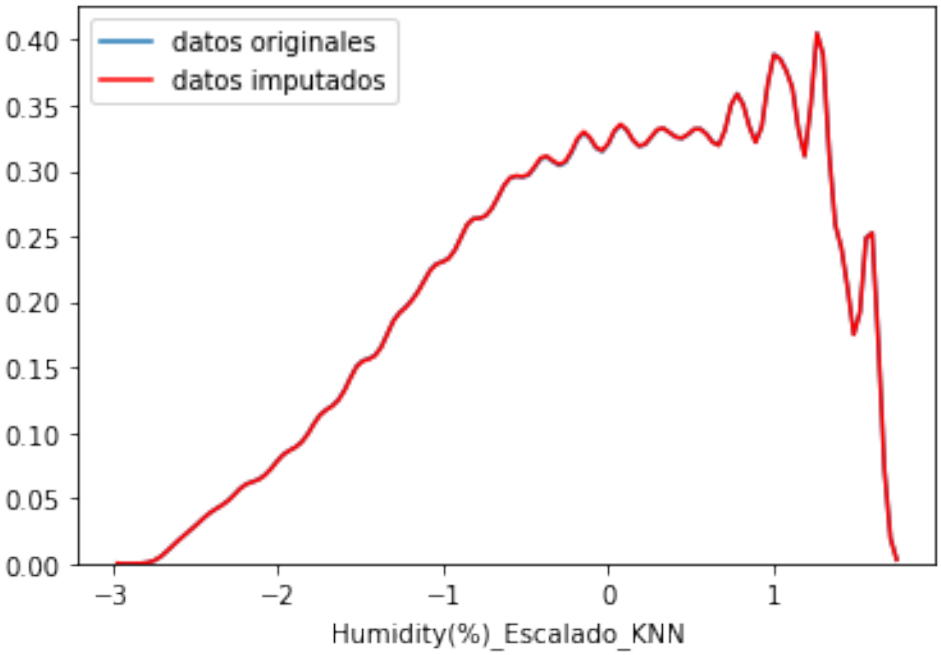


Figura 17: Valores Imputados vs Originales Humidity(%)

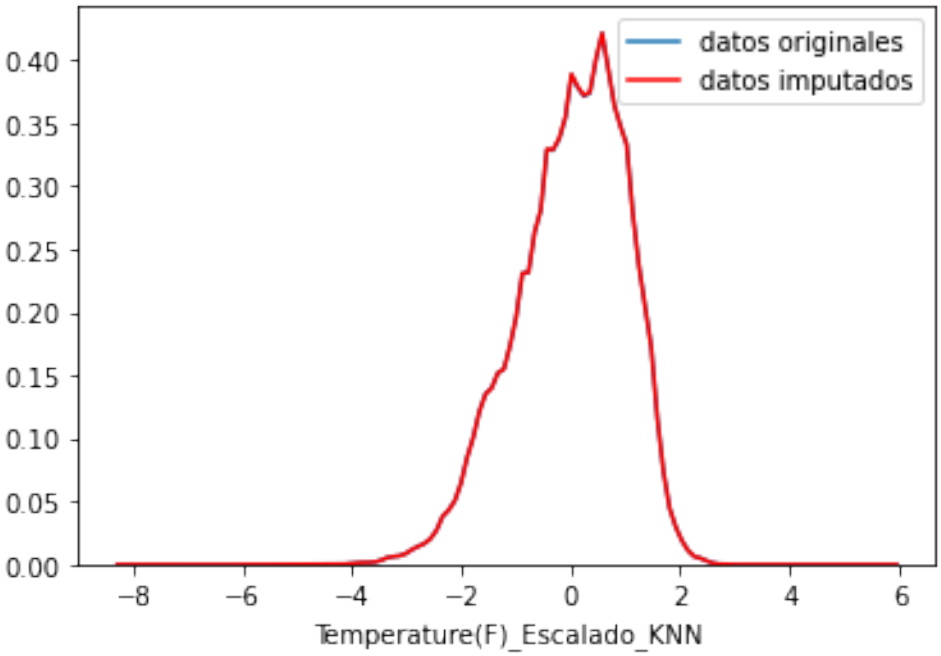


Figura 18: Valores Imputados vs Originales Temperature(F)

6.4. Reducción de dimensiones

6.4.1. Variables Unarias

Las siguientes variables descartadas por estar concentradas en un nivel.

Variable	Tasa de Concentración
Turning_Loop	1.0
Bump	0.999828
Give_Way	0.997344
No_Exit	0.998738
Railway	0.991251
Roundabout	0.999947
Traffic_Calming	0.999608

Cuadro 8: Variables concentradas en un nivel

6.4.2. Variables con Missings

Las siguientes variables serán descartadas por su alta tasa de Missings ya que sería peligroso intentar imputarlas.

Variable	Tasa de Missings
End_Lat	0.698702
End_Lng	0.698702
Number	0.643073
Wind_Chill(F)	0.463360
Precipitation(in)	0.528233

Cuadro 9: Variables descartadas por Missings

6.4.3. Análisis Componentes Principales (PCA)

Visualizamos nuestra variable objetivo con ayuda de los componentes principales, los cuales decidimos no ocupar para la reducción porque no nos explican suficiente varianza.

Variable estandarizadas
Temperature(F)
Humidity(%)
Pressure(in)
Visibility(mi)
Wind_Speed(mph)

Cuadro 10: Variables utilizadas para PCA

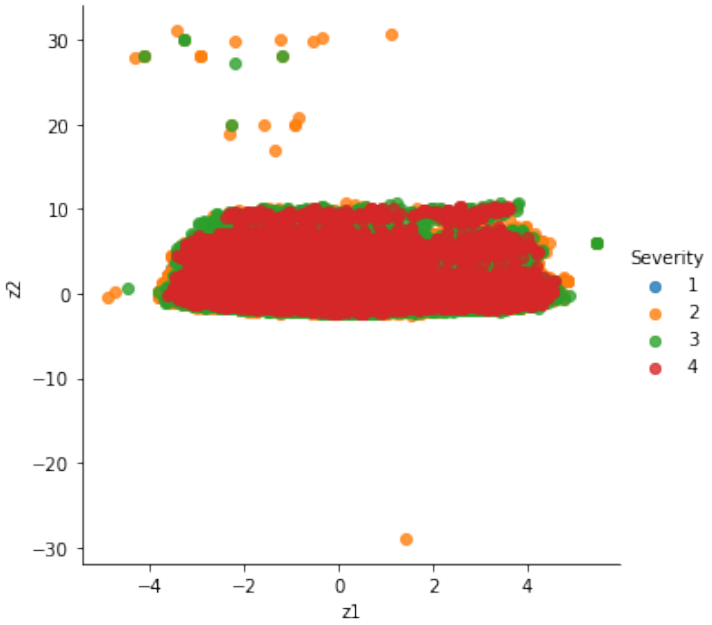


Figura 19: PCA con 2 Componentes

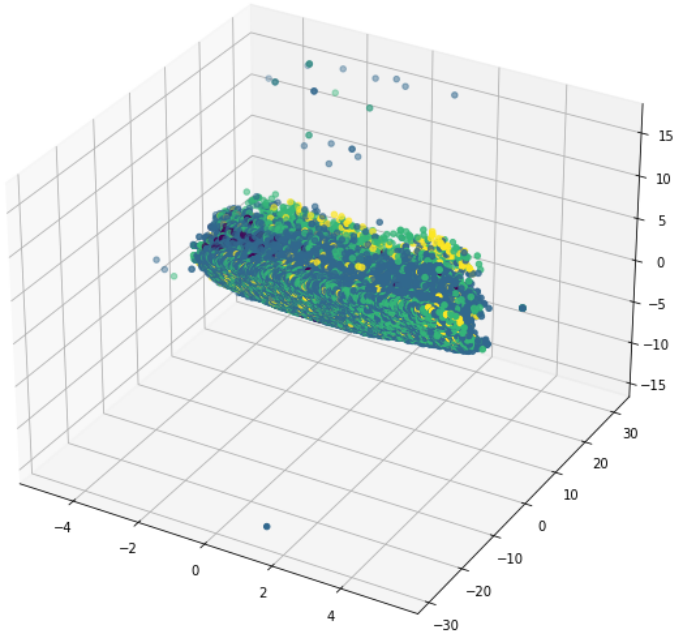


Figura 20: PCA con 3 Componentes

6.4.4. Análisis Discriminante Lineal (LDA)

Realizamos una visualización parecida de nuestra variable objetivo pero con ayuda de el Análisis de Discriminante, usando las mismas variables del PCA.

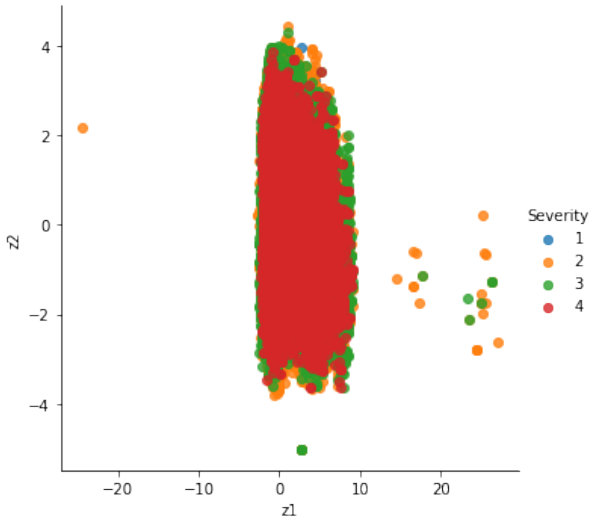


Figura 21: LDA con 2 Componentes

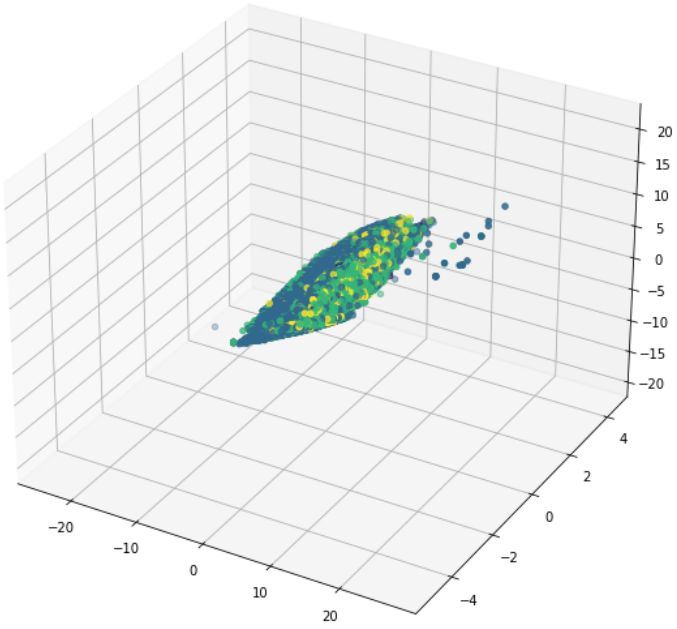


Figura 22: LDA con 3 Componentes

7. Modelos Supervisados

En los modelos supervisados buscamos predecir el valor correspondiente a una matriz de entrada.

$$y = f(x) \tag{1}$$

Siendo el resultado de esta estimación parte de un modelo regresivo o de clasificación, buscando generalizar situaciones no vistas, partiendo de la entrada presentada previamente.

7.1. Remuestreo

Una de las técnicas utilizadas para combatir el problema de "imbalance data", fue el remuestreo, aplicando el oversampling u undersampling en el entrenamiento sobre la minoría de nuestra variable objetivo para ayudar al modelo a generalizar en las predicciones.

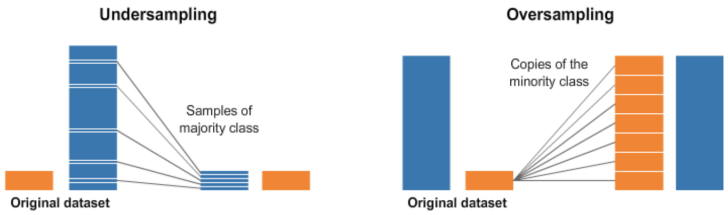


Figura 23: Ejemplo Oversampling y Undersampling

7.2. Modelos lineales

Un modelo lineal es aquel cuya interpretación se puede realizar fácilmente, es decir son altamente explicables y entendibles debido a la sencillez de su aplicación. Aunque muchas de las problemáticas actuales no son lineales, lo tomaremos como una primera aproximación para la solución de este problema que es la Severidad de los accidentes automovilísticos.

7.2.1. Severidad

Para esta primera estimación ocuparemos una Regresión Logística para buscar la causalidad de nuestro fenómeno, en esta ocasión optamos por convertir nuestra variable objetivo a dicotómica principalmente debido a la distribución inicial de la misma, quedando con las siguientes proporciones.

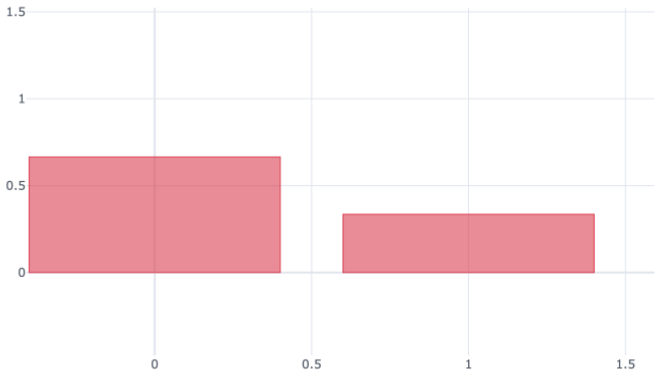


Figura 24: Severidad agrupada a pares

El modelo aplicado cuenta con un total de 30 variables involucradas, en este apartado realizaremos la lectura general para el estudio y se encontrarán más a detalle en el anexo.

Variable	Coeficiente
Stop	-2.29
Traffic Signal	-1.59
Amenity	-1.24
Rhode Island	.84
Missouri	1.50
Side	2.28

Cuadro 11: Estadígrafos Regresión Logística

Para este punto podemos destacar varios aspectos:

- La presencia de señales de tránsito, cruces peatonales y áreas de interés común (centros comerciales, aeropuertos, estaciones de transporte público), tienen un efecto positivo disminuyendo la severidad de los accidentes
- Se econtraron estados que son propensos a tener accidentes de mayor severidad Missouri, Rhode Island, Georgia, Connecticut, entre otros
- La cantidad de accidentes reportados en las Api, no tiene mayor influencia en la severidad
- Las horas 7 y 8 am tambien tienen un efecto positivo disminuyendo la severidad
- Los accidentes son de mayor seriedad cuando el auto queda del lado derecho relativo a la calle, es decir, cuando invade la acera peatonal

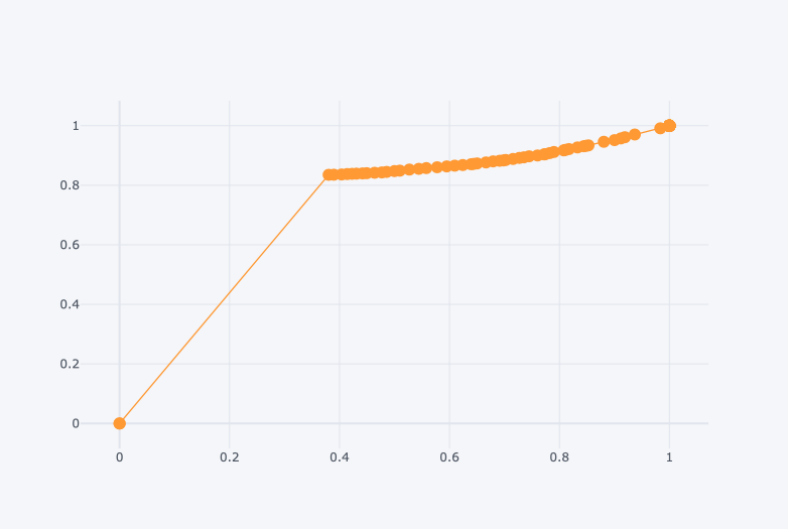


Figura 25: ROC .727

Con esto tenemos nuestro primer acercamiento, en encontrar las condiciones que causan mayor severidad, que era uno de nuestros objetivos iniciales, pero ¿Que aplicaciones podemos sacar de esto?

- Al detectar zonas de alto riesgo, es factible tomar medidas apropiadas para contrarrestarlo, y destinar de manera óptima los recursos encargados de solventar esta situación

7.2.2. TimeMinutes

Como análisis auxiliar a nuestra variable objetivo tenemos el tiempo en minutos que duró el incidente vial, esto con el fin de saber que características generan mayor pérdida de tiempo al momento de ocurrir el incidente vial.

Para este apartado el modelo seleccionado fue una Regresión Cresta, tomando en cuenta 127 variables de las cuales destacan:

Variable	Coeficiente
Colorado	-15.54
Clima Nublado	-9.81
Clima Despejado	-7.71
Clima Neblina	5.61
Vermont	11.76
Oregon	15.30

Cuadro 12: Estadígrafos Regresión Cresta

- A diferencia de la severidad, el tipo de clima es de los factores más importantes para la duración de los incidentes de tránsito
- Se econtraron horas específicas en la que los accidentes aumtentan su duración
- De igual forma la cantidad de accidentes reportados en las Api no tiene mayor influencia
- Tenemos meses que los meses tambien son un factor importante en la duración, siendo los que la aumentan Marzo, Abril, Mayo, Junio y Julio
- Los accidentes de mayor duración se encuentran en Oregon y Vermont

Gracias a este análisis auxiliar podemos entender más la variable objetivo, que es una combinación de la distancia afectada en millas, y la duracción del accidente.

- A partir de las condiciones meteorológicas y geográficas podemos conocer un estimado de la duración de la afectación de tránsito

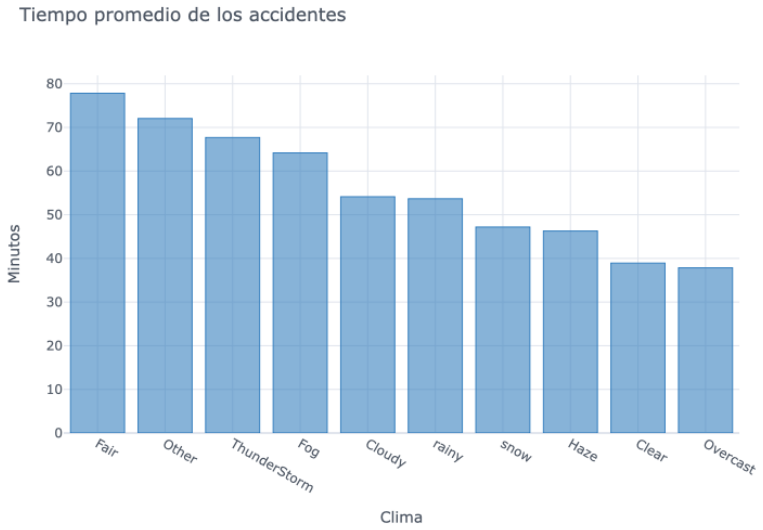


Figura 26: Accidentes en minutos por tipo de clima

7.2.3. Credit Scoring

Para este apartado, daremos un puntaje a los incidentes de tránsito y obtendremos los puntos de corte para determinar la severidad del mismo tomando la variable objetivo de forma dicotómica igual que en la sección anterior.

Discretizamos nuestras variables para obtener el information value (IV), y nuestro threshold para que una variable sea considerada en el modelo es que su IV sea mayor a .02, con lo cual obtenemos la siguiente tabla:

Variable	IV
Side	0.65
Traffic_Signal	0.46
State	0.35
Crossing	0.24
Weekday	0.09
Junction	0.07
Stop	0.06
Hour	0.049
Timezone	0.045
Amenity	0.038
Station	0.035
Start_Lng	0.028
Weather	0.024

Cuadro 13: Information Value

Con los cuales tras aplicar el modelo podemos obtener el rango de los scores, y encontrar los puentos de corte.

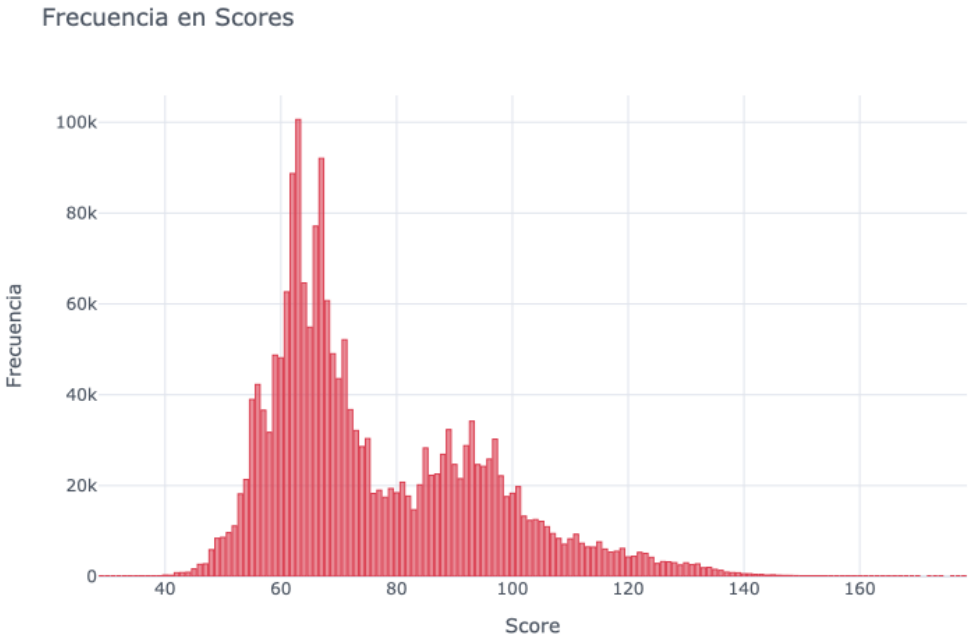


Figura 27: Distribución de Scores

Para este modelo el scoring tiene un rango entre [29,178] y el punto de corte empieza en 71 puntos.

Cuadro 14: Lado relativo del accidente

Side	Score
Izquierda	28
Derecha	2

Cuadro 15: Presencia señal de tráfico

Traffic_signal	Score
No	3
Si	21

Cuadro 16: Presencia de un cruce

Crossing	Score
No	5
Si	16

Cuadro 17: Presencia intersecciones cercanas

Junction	Score
No	6
Si	0

Cuadro 18: Presencia de un Alto

Stop	Score
No	5
Si	32

Cuadro 19: Presencia de áreas de confort

Amenity	Score
No	5
Si	20

Cuadro 20: Presencia de Estaciones

Station	Score
No	5
Si	14

Cuadro 21: Zona horaria

Timezone	Score
US/Central	4
US/Eastern	6
US/Mountain	4
US/Pacific	8

Cuadro 22: Clima

Weather	Score
Clear	5
Cloudy	5
Fair	9
Fog	10
Haze	4
Other	6
Overcast	5
Thunderstorm	2
Rainy	3
Snow	1

Cuadro 23: Día de la semana

Weekday	Score
Fri	6
Mon	7
Sat	-1
Sun	-1
Thu	6
Tue	7
Wed	7

Donde notamos que toman relevancia variables como el clima, el día de la semana y los estados donde ocurren los accidentes.

7.3. Modelos no lineales

Debido a la complejidad y no linealidad de las problemáticas actuales debemos recurrir a modelos robustos que nos ayuden a detectar patrones con mayor presición que los modelos del apartado anterior.

Para está sección introduciremos dos conceptos importantes:

- Modelos Caja Blanca: Son aquellos algoritmos de aprendizaje automático que tienen un alto grado de explicabilidad e interpretabilidad
- Modelos Caja Negra: Son aquellos algoritmos de aprendizaje automático, en los cuales complejidad de la función utilizada para modelar, no nos permite realizar una interpretación de la misma

7.3.1. TimeMinutes

Para la continuaciòn del análisis de la variable TimeMinutes utilizaremos un DecisionTreeRegressor, de lo cual rescatamos lo siguiente:

- De los 52 estados pertenecientes a EUA, 22 no fueron relevantes para la regresión
- La presión atmosférica fue la variable más importante para la regresión siendo el corte en 29.66 pulgadas
- La población por estado obtenida con ayuda de un estimado del año 2013 no resultó significativo

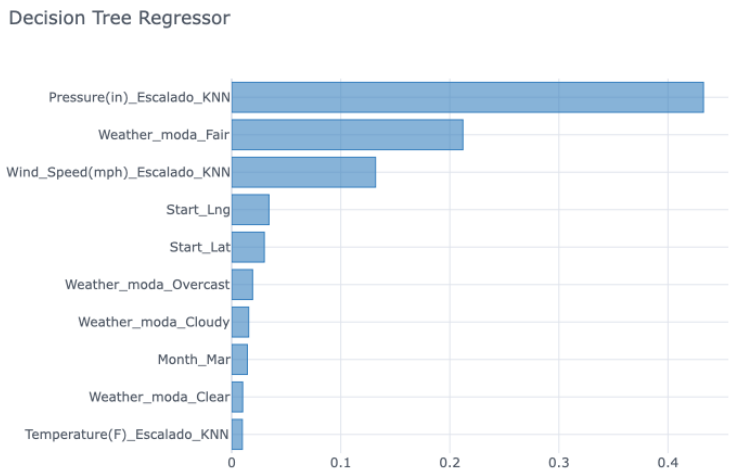


Figura 28: Las 10 variables más importantes del modelo de regressiòn

Tambien podemos notar que en contraste con el modelo lineal, se descartaron los estados como las variables con mayor predictibilidad tomando su lugar la latitud y la longitud.

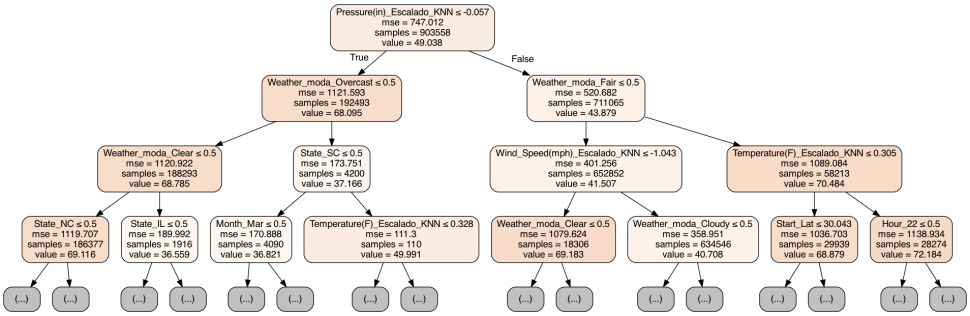


Figura 29: Árbol de regresión

Además notamos que la duración de la afectación de tránsito es mayor cuando el clima está despejado.

7.3.2. Severidad

El primero de dos modelos no lineales que utilizaremos para la Severidad, es el DecisionTreeClassifier, que por tener la bondad de ser un modelo caja blanca nos permite una mayor interpretación de los resultados

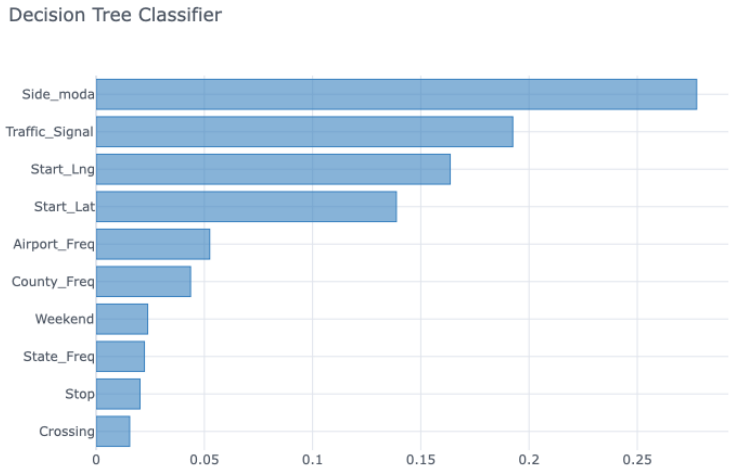


Figura 30: Las 10 variables más importantes del modelo de clasificación

- Notamos que la variable Side, permanece como la más explicativa hasta ahora
- La ausencia de señales de tránsito aumenta la severidad de los accidentes, como habíamos notado anteriormente
- La cantidad de accidentes reportados por estado, va relacionado a la severidad de los mismos

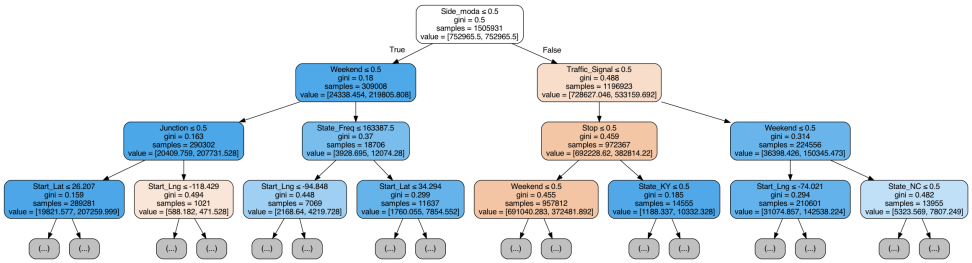


Figura 31: Árbol de clasificación

En general el modelo realiza los cortes para clasificar con base en condiciones geográficas y tomando en cuenta las señales de tránsito. Con lo cual es posible identificar hotspot de accidentes viales con severidad alta.

7.3.3. XGBoost Classifier

Para el segundo modelo no lineal que ocuparemos para la Severidad, tenemos al XGBoost Classifier, una ventaja que tenemos en este caso es que a pesar de ser un modelo

de caja negra, es decir, de dificil interpretación. Tenemos que el feature importance fue parecido al del árbol de clasificación, por lo que nos apoyaremos en el primero para la explicabilidad del modelo.

- Con el XGboost tuvimos una mejora del 16 % respecto al modelo lineal y del 9 % respecto del árbol de decisión
- El score final del XGboost fue del 0.844
- Debido a las pocas observaciones de los valores extremos optamos, además de poder hacer la comparación de todos los modelos. se optó por dejar la variable objetivo como dicotómica

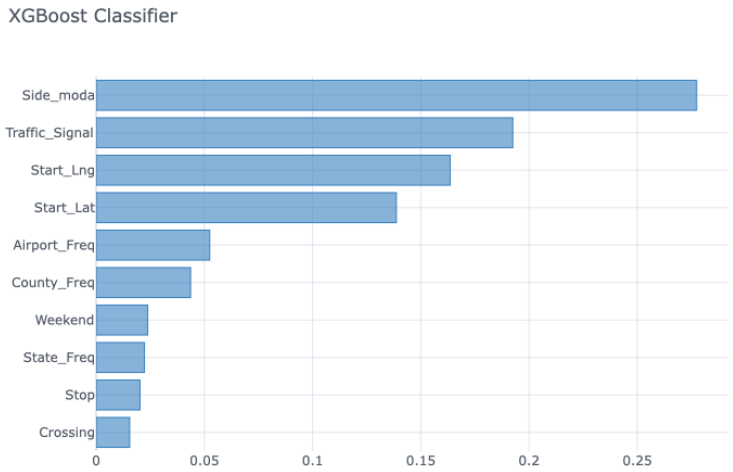


Figura 32: Las 10 variables más importantes del XGBoost

8. Modelos No Supervisados

En los modelos no supervisados buscamos encontrar patrones ocultos en los datos a partir de criterios de similitud.

$$S_n = \{X^{(i)}, i = 1, 2, ..., n\} \tag{2}$$

Siendo el resultado de los criterios el conjunto de los clusters, para está sección tenemos tres opciones:

- Clustering jerárquico
- Clustering de optimización
- Clustering de densidad

8.1. Clustering

8.1.1. Gaussian Mixtures

El modelo que ocuparemos para está sección es Gaussian Mixtures, siendo este una extensión probabilística de k-means, con el cual describimos los clusters a partir de Gaussianas.

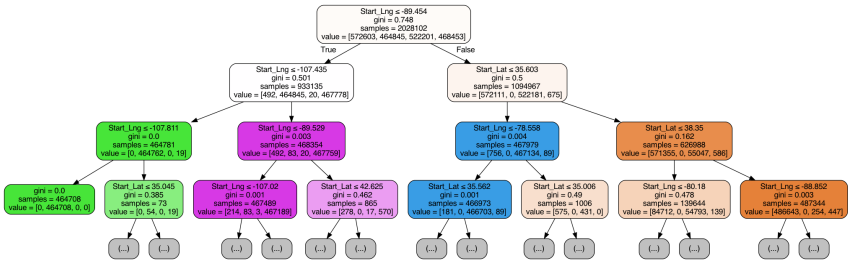


Figura 33: Árbol de decisión para perfilamiento

8.1.2. Perfilamiento

Para nuestro perfilamiento tenemos cuatro grupos, los cuales se comportan de la siguiente manera:

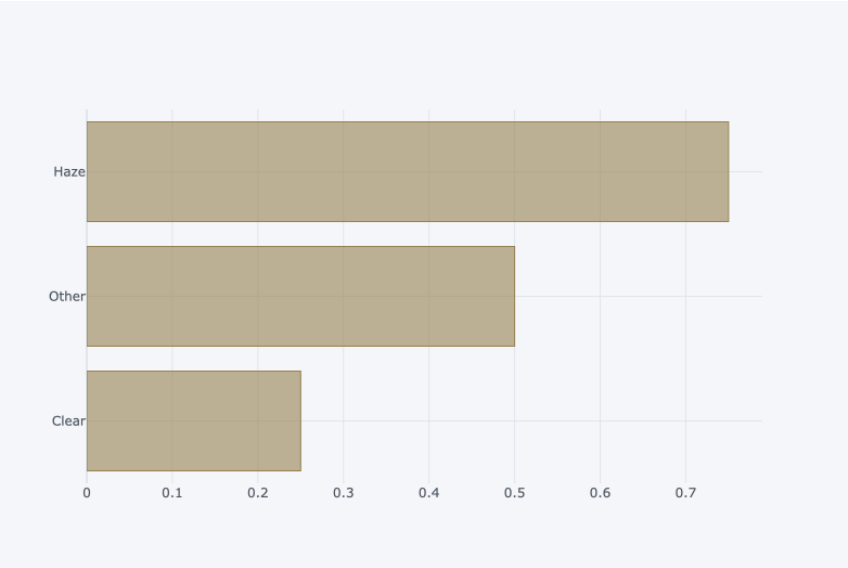
Accidentes por eventos sociales.

- Ocurren de noche o madrugada (8pm-3am)
- En fines de semana
- Severidad ★

¿En donde ocurren este tipo de accidentes?
Principalmente se encuentran estados de la zona horaria Pacífico o Montaña, por ejemplo:

- Arizona, California, Idaho

¿En que climas ocurren estos accidentes?



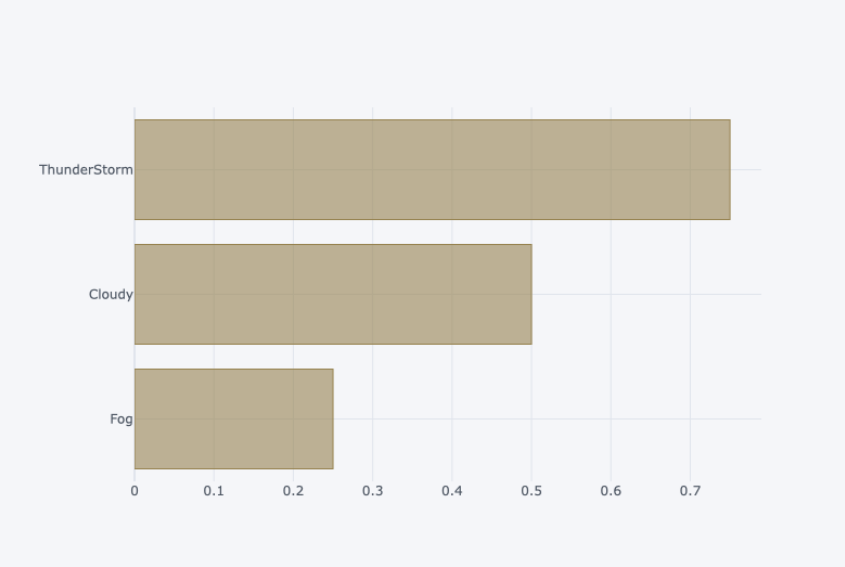
Accidentes menores.

- Alrededor de las 9am
- De lunes a domingo
- Severidad ★★

¿En donde ocurren este tipo de accidentes?
Principalmente se encuentran estados de la zona horaria Central, por ejemplo:

- Arkansas, Iowa, Kansas

¿En que climas ocurren estos accidentes?



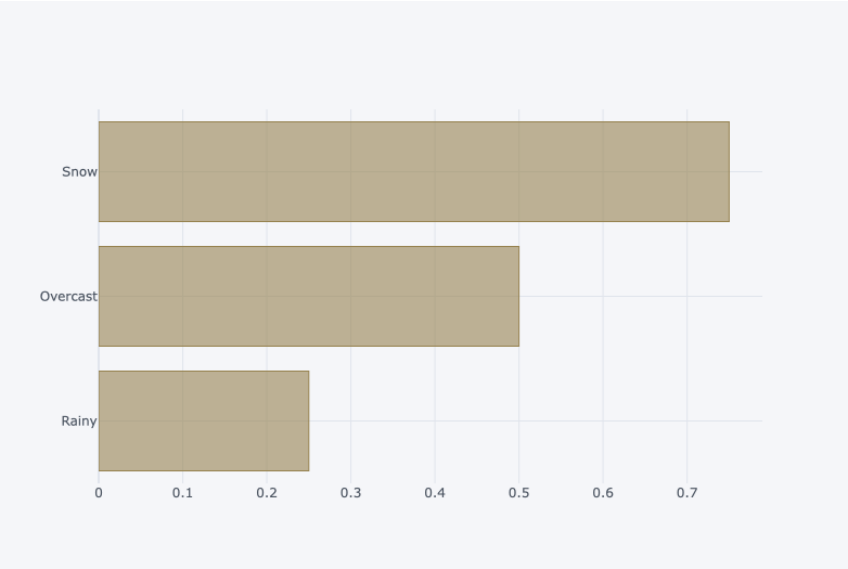
Accidentes por trabajo.

- Ocurren en la mañana (4-11am)
- De lunes a viernes
- Severidad ★★★

¿En donde ocurren este tipo de accidentes?
Principalmente se encuentran estados de la zona horaria del Este, por ejemplo:

- Connecticut, Delaware, Massachusetts

¿En que climas ocurren estos accidentes?



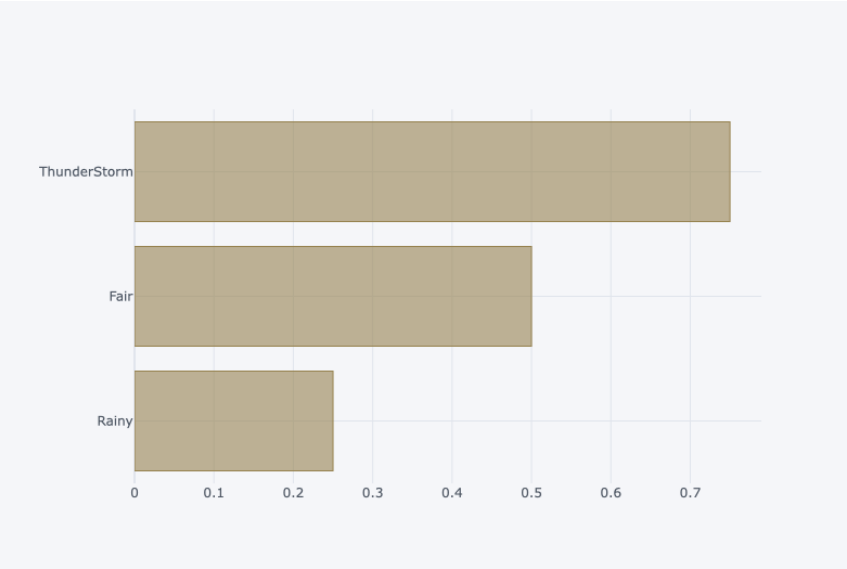
Accidentes graves.

- Alrededor de las 3pm
- De lunes a viernes
- Severidad ★★★★

¿En donde ocurren este tipo de accidentes?
Principalmente se encuentran estados de la zona horaria del Este, por ejemplo:

- Georgia, Florida, Kentucky

¿En que climas ocurren estos accidentes?



9. Resultados

9.1. Modelos Supervisados

9.1.1. Primeras observaciones

En general con los resultados de los modelos podemos notar que los accidentes de tránsito siguen un patrón que es posible identificar, y con base en la técnica utilizada puede verse afectada la relevancia de las variables en los análisis.

Con lo cual en primera instancia podemos remarcar lo siguiente:

- La relevancia que toman los aspectos geográficos a pesar hablar del mismo país, nos indica que hay una diferencia definida en cultura vial y prevención de accidentes
- En primera instancia, se pensaba que el número de accidentes sería un factor clave para la predicción de la severidad de los mismos, dado que no fue el caso posteriormente enriqueciendo la tabla a partir de otras fuentes se analizará si la densidad poblacional puede ser un factor relevante para el mismo
- Estos modelos están enfocados a prevención de accidentes, al ser varias de las características de origen geográfico es factible pensar en reforzar campañas de prevención en los estados que son propensos a tener accidentes de mayor gravedad
- El análisis de la variable TimeMinutes resulta importante e interesante, ya que definiendo las condiciones climáticas/geográficas que generan mayor afectación al momento de ocurrir un accidente, es posible generar un estimado de la duración del mismo. Dado que uno de los negocios con mayor revenue son las compañias que proveen vehículos de transporte con conductor, tener conocimiento de esto es de gran importancia, como método de calidad de servicio dado que puede marcar la diferencia entre competidores

9.1.2. Observaciones complementarias

Notamos que abordando nuestro problema a partir de distintos modelos, gran parte de las variables relevantes para la clasificación o regresión se mantuvieron, con lo cual para terminar está sección tenemos lo siguiente.

- ¿Como resuelve el modelo la problemática de severidad? Como se ha mencionado anteriormente, confirmamos que existen hotspots de manera matemática con apoyo de nuestros modelos, y no solo eso tambien tenemos que las señales de tránsito tienen gran impacto en la severidad de los mismos, con lo cual una propuesta para disminuir la severidad sería señalizar las áreas mayormente afectadas en términos de severidad
- ¿Como resuelve el modelo la problemática de la duración del accidente? En este caso, viendolo desde un enfoque distinto a la prevención, lo que si podemos hacer es una vez ocurrido el accidente determinar en que condiciones ocurrió y así tener una estimación de la duración de la afectación y en caso de ser considerada una afectación grave, abrir paso a vías externas para evitar la aglomeración vehicular
- ¿A que sectores les beneficia este estudio? Al ser sobre la afectación vial, empresas que proveen vehículos de transporte con conductor, paquetería e incluso una institución gubernamental dedicada al transporte podrían sacar partido de este estudio

9.2. Modelos No Supervisados

9.2.1. Observaciones

En general con las agrupaciones realizadas a partir de los modelos, podemos rectificar o identificar lo siguiente:

- La idea de hotspots de accidentes con ayuda del día de la semana, la hora y el lugar
- Podemos crear escenarios de los tipos de accidentes, apoyandonos en las agrupaciones
- Una mayor claridad de las condiciones en que ocurren los accidentes según su severidad
- Y finalmente llegar a una mayor comprensión del fenómeno que modelamos en secciones anteriores

¿Como nos apoya esto en nuestro estudio? Dando las condiciones de cada accidente podemos realizar story telling basandonos en resultados determinando los escenarios en los cuales surgen cada accidente con su respectiva severidad.

10. Conclusiones

10.1. Personales y Siguientes Pasos

En general, me pareció muy interesante como factores geoespaciales nos ayudan a clasificar la severidad de los accidentes de tránsito con una precisión a mi parecer aceptable, ampliando el panorama respecto a los accidentes y como el entorno influye en ellos dejando tanto dudas como respuestas.

- Notamos hotspots en términos de los accidentes y como están dados por factores como la zona ya sea el estado o incluso si ocurrió cerca de una zona de interés común como una plaza, horario y clima del accidente. pero yendo de forma más refinada ¿Como afectán variables como el ancho de la avenida? o ¿Cual era el estado de la avenida previo al accidente?

Tambien es de mi interés ver como se comportaría este fenómeno añadiendo datos del conductor, e incluso del tipo de vehículo registrado en el accidente para identificar si hay grupos prominentes respecto a la severidad de los accidentes. por ejemplo ¿si el vehículo está registrado para transportar pasajeros como influye esto en la gravedad del accidente? o si en caso contrario es un vehículo privado ¿Es de relevancia?

Para concluir este apartado, y remitiendome a los objetivos inicialmente planteados:

- ¿Pudimos estructurar y visualizar los datos para comprender el fenómeno? Si, el análisis previo jugó un papel crucial incluso modificando nuestra unidad muestral, realizando la limpieza apropiada pudimos realizar inferencias.

- ¿Cuales fueron las condiciones que causan mayor severidad? En general zonas sin señales de tránsito o sin zonas recreativas, en horarios con gran movilización de tránsito como hora de entrada al trabajo o salida de escuelas generan grandes aglomeraciones. A lo cual podemos agregar que las condiciones climáticas desfavorables como lluvias o nevadas

- ¿Encontramos patrones definidos para poder segmentar los accidentes según sus características? A pesar de que si fue posible segmentar los datos rectificando los hallazgos anteriores, pienso que hubiera sido más interesante tener información de los conductores para poder generar sus perfiles y revisar los accidentes desde una perspectiva diferente

¿Cuales serían los siguientes pasos?

- A mediano plazo sería recopilar más información para complementar el estudio, incluso buscar información de historiales de choque por persona para saber como se ve afectado este hecho por el tiempo

- A largo plazo, pensaría en un modelo para predecir accidentes en tiempo real una vez se tenga mayor conocimiento del tema. Tanto de negocio como técnico

11. Anexo

11.1. Complemento modelos lineales

11.1.1. Arquitecturas

Regresión Logística

- LogisticRegression(C=0.5, penalty='l1', solver='liblinear', tol=1e-05)

Variable	Coeficiente
Start_Lat	-0.0201
Amenity	-1.240
Crossing	-0.946
Junction	0.603
Weekday	0.09
Station	-0.753
Stop	-2.295
Traffic_Signal	-1.593
State_Freq	-1.053e-06
Airport_Freq	-3.77e-07
Weekend	0.762
Side_moda	2.288
State_AZ	-0.526
State_CA	0.267
State_CT	0.669
State_GA	0.681
State_IL	-0.325
State_LA	-0.988
State_MN	0.532
State_MO	1.509
State_NC	-1.025
State_OK	-1.249
State_RI	0.834
State_SC	-0.857
State_TX	-0.129
State_VA	0.348
Timezone_US/Central	-0.187
Timezone_US/Pacific	0.252
Hour_7	-0.368
Hour_8	-0.371
Weather_moda_Fair	-0.334

Regresión Cresta

- Ridge(alpha=19, tol=0.01)
- Coeficientes omitidos por volumen

Credit Scoring

- LogisticRegression(C=0.05, solver='saga', tol=0.01)

Variable	Coeficiente
Side_moda	-0.903
Traffic_Signal	-0.783
State	-0.786
Crossing	-0.430
Weekday	-0.726
Junction	-0.436
Stop	-0.881
Hour	-0.833
Timezone	0.530
Amenity	-0.576
Station	-0.469
Start_Lng	-0.014
Weather_moda	-0.858

11.2. Complemento modelos no lineales

11.2.1. Arquitecturas

DecisionTreeRegressor

- DecisionTreeRegressor(criterion='mse', max_depth=12, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')

DecisionTreeClassifier

- DecisionTreeClassifier(class_weight='balanced', criterion='gini', max_depth=15, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')

XGBClassifier

- XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1, importance_type='gain', interaction_constraints='', learning_rate=0.300000012, max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan, monotone_constraints='()', n_estimators=100, n_jobs=0, num_parallel_tree=1, objective='binary:logistic', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact', validate_parameters=1, verbosity=None)