ANÁLISIS DE ACCIDENTES DE TRÁNSITO EN AUTOPISTAS AUSA

Alumnos: Juan Carlos Escaray Lara leg: 1504186

Burgos Ontiveros Victoria leg: 1630740

RESUMEN

La ciencia de datos es un campo interdisciplinario que involucra métodos, procesos y sistemas científicos que extraen conocimiento o comprenden mejor los datos en diferentes formas (ya sean estructuradas o no estructuradas).

Nuestro proyecto tiene como objetivo explicar el comportamiento de diferentes eventos mediante el establecimiento de métodos de Machine Learning y averiguar si existe una relación entre diferentes variables en nuestro data set.

Realizamos la búsqueda de nuestro data set teniendo en cuenta que necesitábamos de samples y features que nos permitieran aplicar alguno de los modelos explicados durante la cursada.

Elegimos un data set del Gobierno de la Ciudad de Buenos Aires ya que consideramos que el análisis de accidentes es un tema de interés para toda la comunidad, en este caso de la Ciudad de Buenos Aires.

Se utilizó el data set descargado de la pagìna de la Ciudad, donde se registran las incidencias de accidentes sobre distintas autopistas, con fecha, tipo de evento y vehículos involucrados.

A continuación el link del data set.

https://data.buenosaires.gob.ar/dataset/seguridad-vial-autopistas-ausa

OBJETIVO

Según el tipo de accidente y las condiciones climáticas se busca predecir que autopista tuvo mayor fallecidos y lesionados. (según tipo de vehiculo)

Este objetivo es planteado desde la premisa que sólo conseguimos datos de accidentes, no encontramos un data set que contenga los datos de las condiciones de la autopista o condiciones meteorológicas de días que no haya habido accidentes. Por lo tanto el enfoque está puesto en la predicción de personas lesionadas o fallecidas.

EDA

Limpieza del data set y acondicionamiento de la tabla.

Verificamos que no haya filas con valores null. Resultado: Ninguna sample con valores nulos.

Verificamos que no haya filas con valores nan. Resultado: Ninguna sample con valores nan.

Análisis del tipo de columna. Para poder realizar el análisis por fecha tuvimos que modificar el formato de la columna “fecha”

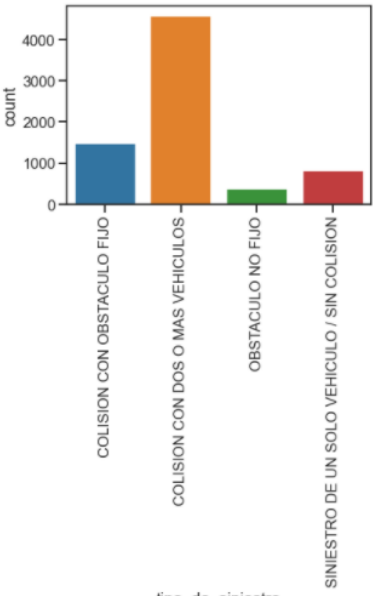
Se verifica que se tenia datos futuros a la fecha de descarga del data set y se procedió a eliminarlo del data set

Verificamos que en las columnas no haya valores repetidos. Tuvimos que unificar algunos criterios ya que considerábamos que se trataban de lo mismo. Por ejemplo, BRUMA O HUMO u OTRO TIPO DE SINIESTRO con OTRO.

Descartamos la columna “PK”, “Banda y/o ramal” y “superficie de la vía”. Si bien parece ser un indicador del km de la autopista en donde se produjo el accidente, vemos cuando mostramos los valores únicos que también contiene samples con datos en forma de str. Este tipo de configuración dificulta el análisis; Se decidió borrar la columna “superficie de la vía” ya que se

evidencia que tiene correlación con “condiciones metereológicas” por lo tanto no sumaba al análisis.

ANÁLISIS EXPLORATORIO DE DATOS



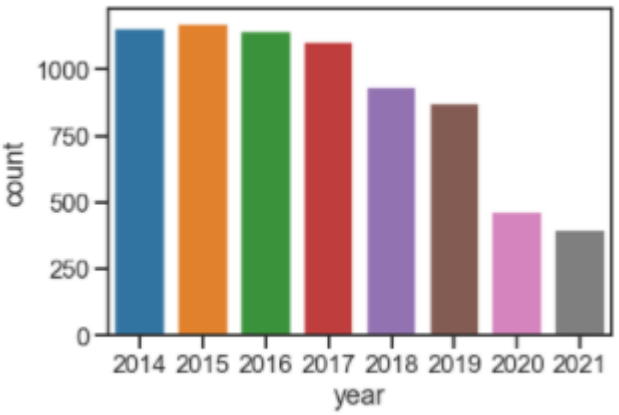
Accidentes por tipo de colisión

Como nos interesa saber la cantidad de accidentes por tipo de colisión procedemos a realizar las siguientes acciones:

Contamos la frecuencia por tipo de colisión y graficamos los resultados.

Se puede observar que la gran mayoría de los accidentes son por COLISIÓN CON DOS O MÁS VEHÍCULOS.

Accidentes por año



Como nos interesa saber la cantidad de accidentes por año procedemos a realizar las siguientes acciones:

- Creamos la columna 'year' con el año indicado en la columna 'fecha'
- Contamos frecuencia por año y graficamos los resultados

Los datos de nuestro data set van de enero 2014 hasta septiembre del 2021. Podemos observar con este análisis que en los años 2014 al 2017 la cantidad de accidentes son similares. En el año 2020 2021 se puede evidenciar una caída significativa de un aproximadamente 60% debido al contexto covid-19.

Accidentes por hora según el día de la semana

Como nos interesa saber la cantidad de accidentes por hora de cada día de la semana procedemos a realizar las siguientes acciones:

En este análisis se observa que de lunes a viernes el histograma por hora se mantiene.

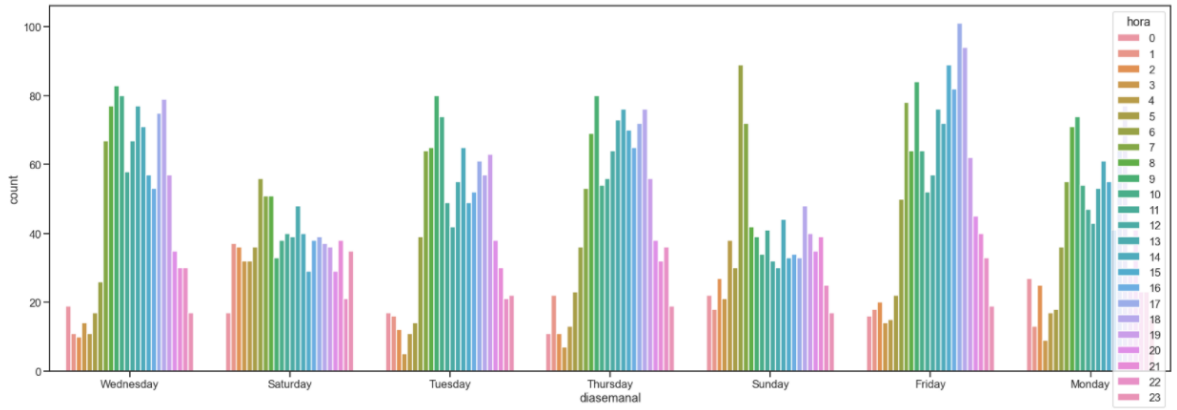
- Creamos la columna 'diasemanal' con el día de la semana indicado en la columna 'fecha'

Sábados y domingos el histograma es más amesetado, con un claro pico en la mañana de los domingos. La causa supuesta es las condiciones de los conductores a la hora de volver de salidas nocturnas de los días sábados por la noche.

- Contamos la frecuencia por hora de cada día de la semana y graficamos los resultados

Deducimos que los picos de 7a 9 hs y de 16 a 18 hs se debe a que son los horarios de entrada y salida de la jornada laboral.

Se evidencia que el día con mayor incidencia de accidentes es el día viernes, esto puede deberse al inicio del fin de semana y la ansiedad de llegar a casa (segùn el pico del gráfico 18hs)

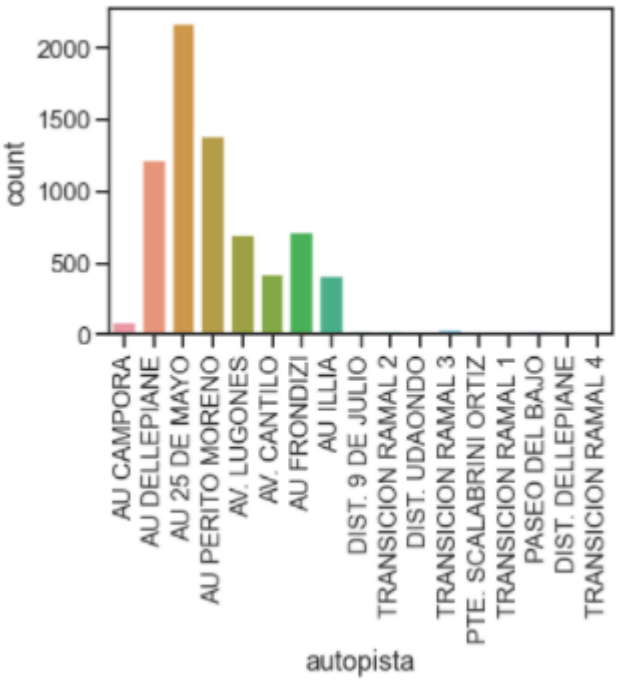


Accidentes por autopista

Debido a que nos interesa saber la cantidad de accidentes por autopista procedemos a realizar las siguientes acciones:

- Contamos la frecuencia por autopista (principales ramales y/o transiciones de vía) y graficamos los resultados.

Se observa que la mayor cantidad de accidentes se presenta en las principales accesos a la Ciudad de Buenos Aires como la AU 25 DE MAYO, AU PERITO MORENO y AU DELLEPIANE



DUMMIES

Realizamos dummies en las columnas y eliminamos las columnas categóricas.

Se modificó la columna lesionados y fallecidos en binario para poder analizar si habrá o no fallecidos utilizando el valor 1 como la ocurrencia de lesionados y/o fallecidos y como 0 la no ocurrencia.

Para esto "dummificamos" la columna creada como 'lesionados_fallecidos', usando la función loc. Con esto se busca predecir el impacto que genera heridos/fallecidos en el accidente.

Por lo tanto, hicimos un proceso de DUMMIES de manera manual.

APRENDIZAJE SUPERVISADO

Se comenzó a utilizar el aprendizaje supervisado para poder analizar si había una relación entre las condiciones meteorológicas, tipo de siniestro y el tipo de vehiculo con la cantidad de lesionados y fallecidos.

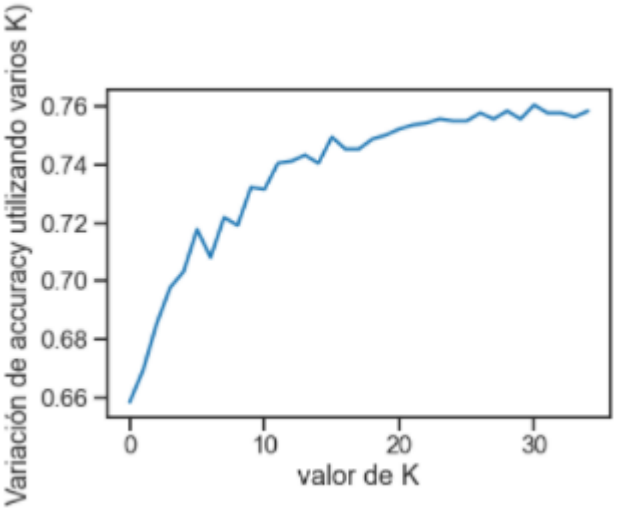
La feature que en nuestro modelo se uso como función dependiente (y). Entonces se sumó las columnas Lesionados y Fallecidos para predecirlas en una sola categoría. Luego se eliminó las columnas originales ya que eran valores categóricas y entorpecian el análisis

MODELOS

Knn

Se inicia la etapa de predicción intentando modelizar con KNN. En principio se obtiene un accuracy de 0.72 ingresando como parámetro la cantidad de vecinos = 5. Luego, via un Gridsearch se obtiene el valor óptimo de vecinos para este modelo: 30.

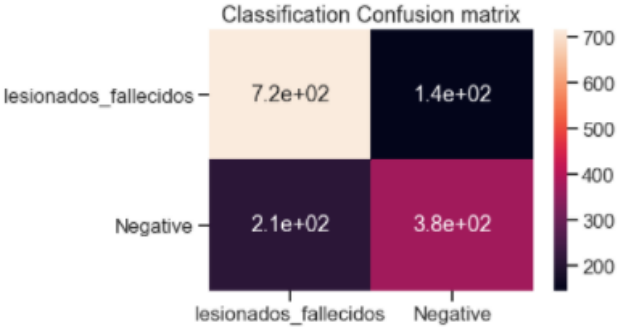
Con este valor, el accuracy sube hasta 0.72. Es en esta instancia cuando se decide recurrir a los metadatos implícitos en las features generando nueva información para alimentar al modelo. Repitiendo el Gridsearch se obtiene la nueva curva de accuracy en función de la cantidad de vecinos asumida.



El máximo valor que logramos alcanzar con este método es de **0.754 de Accuracy**. El indicador AUC trepa hasta 0.735.

Matriz de confusión

Nos permite visualizar los True Positive y True Negative de los resultados del modelo KNN.

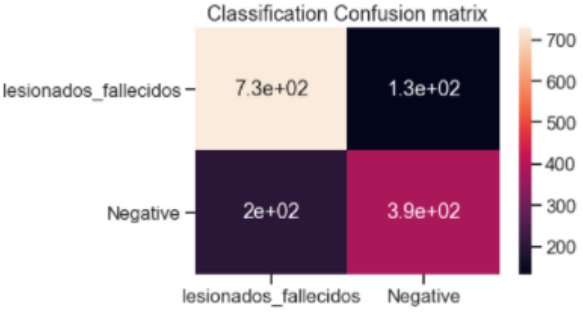


Svm

Una vez terminado el KNN se continúa por evaluar el clasificador de SVM. Se alcanzan valores similares al KNN por lo que se recurre nuevamente a un Gridsearch para optimizar el método sin grandes diferencias en los resultados. **El Accuracy es 0.770** y el AUC 0.752.

Matriz de confusión

Nos permite visualizar los True Positive y True Negative de los resultados del modelo SVM.



RESULTADOS

Usando modelos KNN el máximo valor que logramos alcanzar de Accuracy es 0.**754**

Por otro lado, con SVM el resultado final del proyecto es de **Accuracy = 0.770, AUC 0.752.**

CONCLUSIONES

Se analiza un caso particular como la seguridad vial con fines académicos desde la página del gobierno de BA para poner en práctica los modelos aprendidos y el impacto que generan, el análisis de estos datos evaluando otros modelos puede generar un impacto con la posibilidad de salvar vidas. La aplicación práctica tiene incidencia sobre los servicios de salud, seguridad, control vehicular.

Libros: Machine Learning & Pattern Recognition, The Art of Data Science" de Roger D. Peng y Elizabeth MatsuiData Science Handbook