Análisis de las intervenciones de Seguridad Vial en Ciudad de Buenos Aires

- Cátedra ClusterAI -

Abstract

El trabajo se compone de dos partes. En primer lugar, un análisis exploratorio de los datos y en segundo, se aplica un método de aprendizaje no supervisado con el objetivo de visualizar distintos grupos de intervenciones con características en común.

Keywords

Autopistas, AUSA, Gobierno de la Ciudad, Aprendizaje no supervisado, Clustering

1 INTRODUCCIÓN

El objetivo del trabajo es analizar las intervenciones de seguridad vial en las autopistas AUSA en la Ciudad Autónoma de Buenos Aires para luego formar clusters que nos sirvan de herramienta para la toma de decisiones preventivas.

2 DATASET

Seleccionamos un dataset del Gobierno de la Ciudad de Buenos Aires, el cual nos muestra las intervenciones de seguridad vial en las autopistas de AUSA entre el año 2014 y julio 2020. Se compone de 6639 samples y 15 features. Las features nos dan información sobre el periodo, la fecha, la autopista, el ramal, el km, las condiciones meteorológicas, la superficie de la vía, la cantidad de lesionados y fallecidos, el tipo de siniestro y los tipos de vehículos (moto, liviano, bus y camión) involucrados en las distintas samples. Es importante aclarar que cada sample será una intervención y que el dataset fue tomado el 04/09/2020.

https://data.buenosaires.gob.ar/dataset/seguridad-vial-autopistas-ausa/archivo/juqdkmgo-1872-resource.

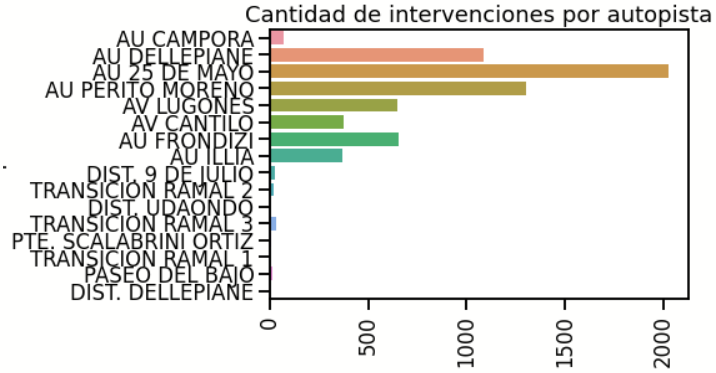
3 ANÁLISIS EXPLORATORIO DE DATOS

En primer lugar, se verificó que no existieran Nulls, se corrigieron errores de redacción en la entrada de datos y eliminamos una única fila que figuraba SD (Sin Datos).

Luego de tener el dataset listo confeccionamos un heatmap para ver la correlación entre autopistas por periodo (mes)

El resultado fue una baja correlación lineal, siendo la máxima 0.40 entre Au 25 de Mayo y Au Dellepiane. En el mismo se observaron espacios en color blanco entre algunas autopistas.

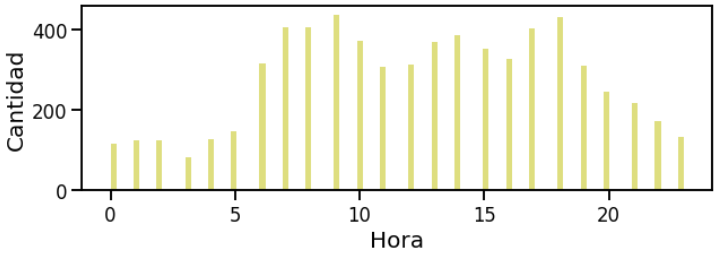
Suponemos que se debe a la poca cantidad de samples que contienen estos valores.



La suposición anterior se confirma luego de realizar un countplot con la cantidad de intervenciones por autopista. Además, obtuvimos que las tres autopistas con mayor cantidad de intervenciones son:

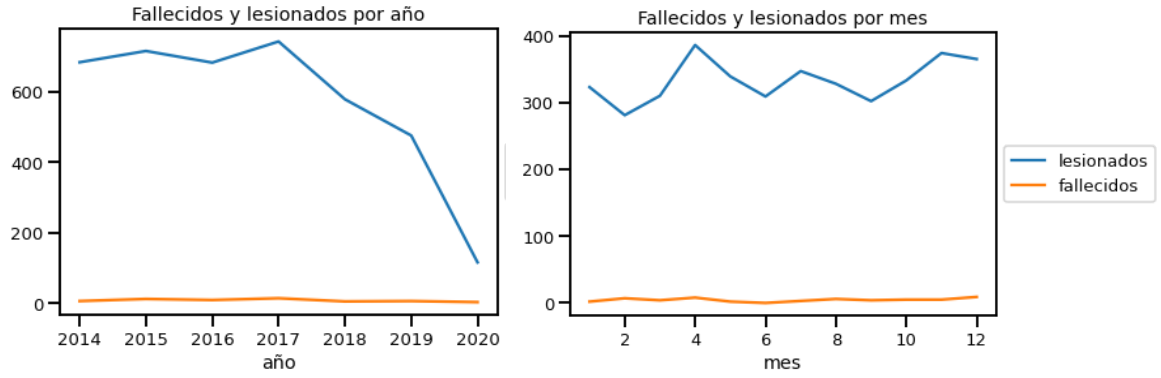
1. Autopista 25 de Mayo
2. Autopista Perito Moreno
3. Autopista Dellepiane

Profundizando más, se realizó un histograma con la cantidad de intervenciones y la hora de las mismas.



Se observa que a lo largo del día los picos de intervenciones coinciden con las horas con más flujo de vehículos (horas pico) de 9 a 10hs y de 18 a 19hs.

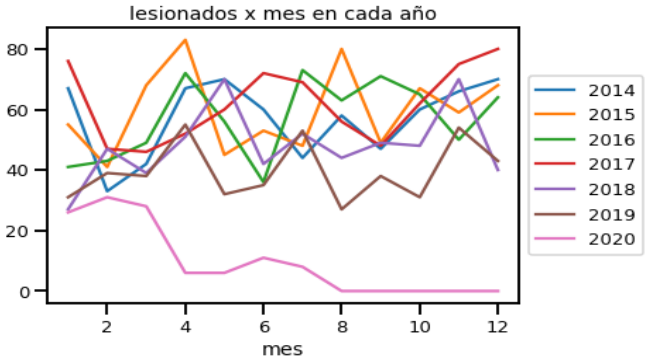
Realizamos un análisis de la cantidad de lesionados o fallecidos por año y por mes, con el fin de obtener una tendencia en los años del análisis.



Hay una proporción mucho mayor de lesionados que de fallecidos.

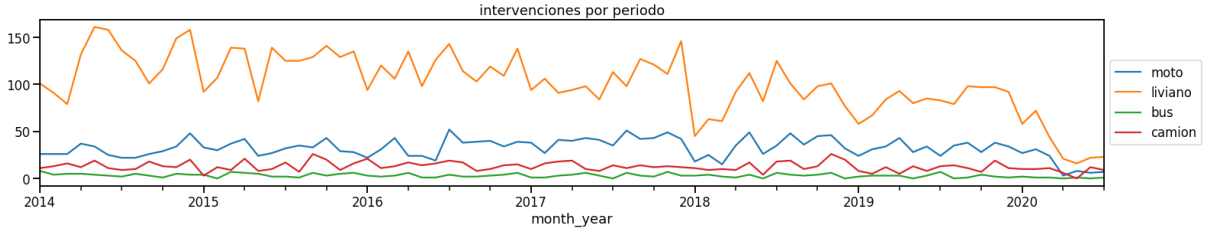
Si bien el año 2020 todavía no concluye, encontrándonos a mitad de año la tendencia indica que la cantidad de lesionados en los últimos tres años está disminuyendo.

Respecto a los meses, se observa un pico de lesionados en el mes de Abril. Para confirmar que la gráfica no esté sesgada por solo año con altos valores de lesionados en el mes hacemos el mismo análisis para cada año.



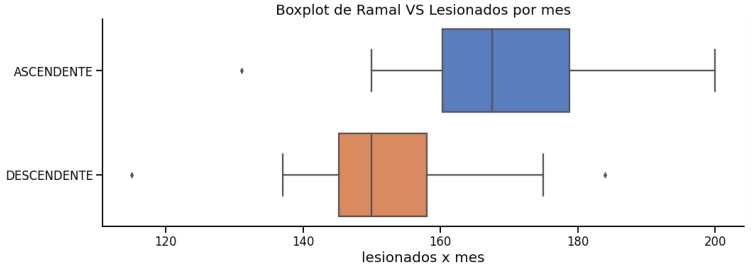
No es un hecho puntual de un año sino que se comprueba que en Abril siempre está entre los meses con mayor lesionados a excepción del 2017. Una posible causa es el fin de semana largo de Semana Santa donde hay un flujo extraordinario de vehículos.

Procedimos a hacer un análisis de la cantidad de intervenciones por tipo de vehículo.



La mayor cantidad de intervenciones involucran vehículos livianos (hasta 2700 kg) y también es importante recalcar la tendencia a la baja desde el 2014.

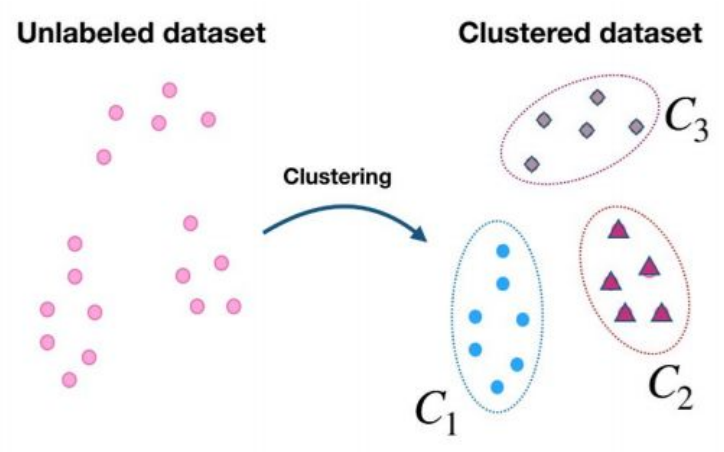
Siguiendo con el análisis, realizamos boxplots por tipo de ramal.



El ramal “ascendente” es el sentido de circulación hacia el km 0 en Plaza Mariano Moreno, Capital Federal y el “descendente” el sentido opuesto. El tercer cuartil de “descendente” no llega ni al primer cuartil de “ascendente”. Se concluye que hay mayor cantidad de lesionados por mes yendo hacia Capital Federal que alejándose.

4 MODELO DE APRENDIZAJE

Una vez finalizado el EDA, se procedió a realizar clustering sobre nuestros datos. Estamos hablando de un método de aprendizaje no supervisado que sirve para segmentar nuestros datos en subconjuntos o grupos.

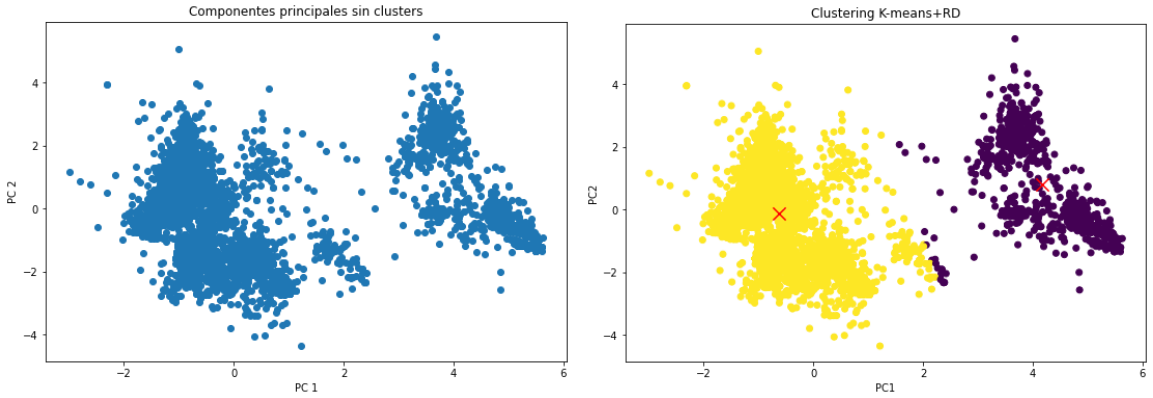


[1]

En primer lugar, creamos dummies con las features hora, autopista, condiciones_meteorologicas, superficie_de_la_via y tipo_de_siniestro, y las unimos con el dataset “intervenciones” filtrado previamente por las features útiles para este modelo: lesionados, fallecidos, moto, liviano, bus y camion.

En segundo lugar, realizamos una reducción de la dimensionalidad a través del método Principal Component Analysis (PCA) que crea nuevas features combinando linealmente las originales para poder explicar de mejor forma la variabilidad de nuestros datos. Así es como reducimos a dos componentes

En tercera instancia, aplicamos un modelo de Clustering conocido como K-Means[2] para medir la similaridad entre muestras. Cada cluster estará identificado por un centroide. El grupo que se le asigne a cada muestra va a depender de la distancia euclidiana cuadrática al centroide más próximo. El objetivo es generar la menor distancia posible intraclusters (dentro de un mismo grupo) y la mayor interclusters (entre grupos).



Para medir la calidad de los clusters se utilizó el Silhouette Index (S), el cual nos indica la similaridad intercluster (entre muestras de distintos grupos) e intracluster (muestras del mismo grupo). El objetivo es minimizar la primera y maximizar la segunda. El indicador toma valores entre -1 y 1. Cuanto más se aproxime a los valores extremos mejor serán nuestros clusters y cuando más cerca de 0 no serán de buena calidad. En nuestro caso, el mejor índice obtenido fue de 0,65 para dos clusters.

5 RESULTADOS

Cluster 1

- Compuesto por una amplia mayoría de las intervenciones (5769 registros).
- Las intervenciones se dieron en su gran mayoría con un buen clima (media de 0.999827)
- Los fallecidos y lesionados se distribuyen en proporciones similares y se observan outliers que hacen referencia a días particulares con muchos más lesionados que la media.
- Las intervenciones que involucran a motos son parte de este cluster.

Cluster 0

- Compuesto por la minoría (869 registros).
- Predomina el clima lluvioso en las intervenciones con una media de 0.966628 y por lo tanto, el estado de la ruta es mojada/húmeda.
- En el cluster 0 los mismos colisionan con obstáculos fijos.

Características comunes

- Las intervenciones que involucran vehículos livianos están distribuidos en ambos clusters, al igual que bus y camión.
- La gran mayoría de las intervenciones se dan en la autopista 25 de mayo. Esto refleja el resultado que obtuvimos con el countplot realizado en el EDA.
- Los tipos de accidentes que más se dan en ambos clusters son de colisión entre 2 o más vehículos.

6 CONCLUSIÓN

En un principio los datos parecían inseparables, pero luego de aplicar este modelo de aprendizaje no supervisado obtuvimos dos clusters con las características mencionadas anteriormente. El análisis será una herramienta más a la hora de tomar decisiones con el objetivo de disminuir la cantidad de lesionados y fallecidos en las autopistas. Después de todo, cuidar a las personas es lo más importante.

REFERENCIAS

[1] Imagen: Interpretación de Señales Genómicas Humanas utilizando técnicas de Machine Learning para mejorar el Diagnóstico Médico , Martín Palazzo.

[2] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, *24*(7), 881-892.

[3] https://github.com/clusterai/clusterai_2020