Clustering de intervenciones viales AUSA

Hugo Britez, Matias Tamborini

UTN.BA INGENIERÍA INDUSTRIAL

clusterAI

UTN.BA UNIVERSIDAD TECNOLÓGICA NACIONAL FACULTAD REGIONAL BUENOS AIRES

Introducción

El trabajo desarrollado es en base a las intervenciones de la seguridad vial en las autopistas AUSA. Los registros obtenidos son desde inicios del 2014 hasta julio del 2020 y fueron proporcionados por el Gobierno de la Ciudad Autónoma de Buenos Aires. El objetivo es analizar las intervenciones para luego formar clusters que nos sirvan de herramienta para la toma de decisiones preventivas.

Dataset

El dataset del Gobierno de la Ciudad de Buenos Aires se compone de 6639 samples y 15 features. Cada sample es una intervención de seguridad. Las features nos describen cada intervención y son las siguientes:

- Periodo
- Fecha
- Autopista
- Ramal
- Kilómetro
- Condición meteorológicas
- Superficie de la vía
- Cantidad de lesionados
- Cantidad de fallecidos
- Tipo de siniestro
- Tipo de vehículos (moto, liviano, bus y camión)

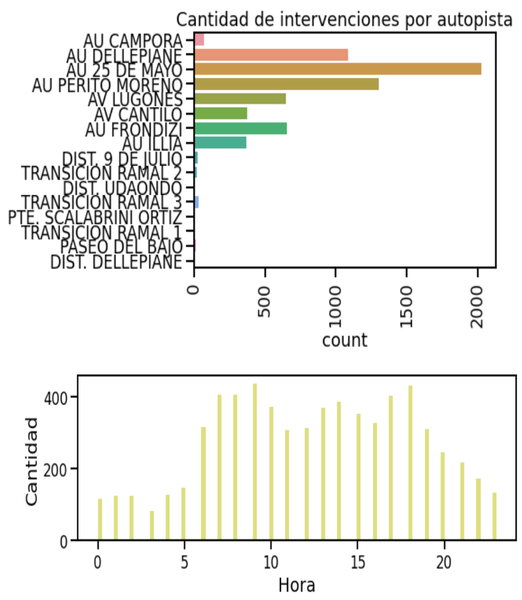
Análisis Exploratorio de Datos

Realizamos un análisis cuantitativo de las intervenciones vs autopistas de AUSA.

Podemos observar que la mayor cantidad de intervenciones se da en Au. 25 de mayo, Au. Dellepiane y Au. Perito Moreno.

A continuación, realizamos un histograma que nos muestre la distribución de las intervenciones por hora.

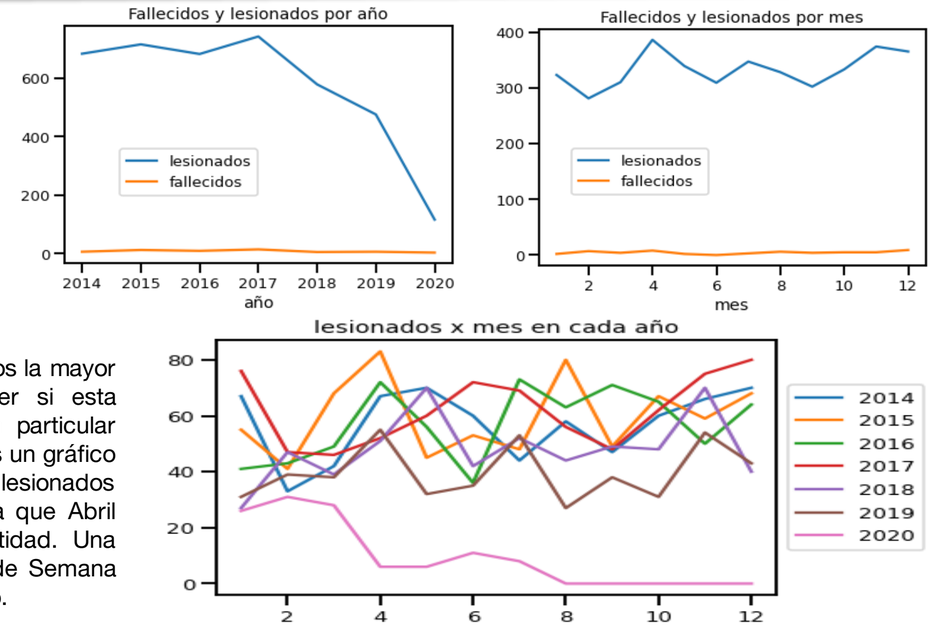
Cabe destacar, que los picos de intervenciones coinciden con las horas pico o también conocidas como las horas con mayor flujo vehicular.



A la derecha se pueden observar dos gráfico de fallecidos y lesionados. Uno nos permite ver la cantidad por año y el otro por meses.

En referencia a los años podemos observar, incluso teniendo en cuenta que el año 2020 todavía no acabó, una tendencia a disminuir las cifras de lesionados.

Respecto a los números por mes, observamos la mayor cantidad de lesionados en Abril. Para ver si esta cantidad no esta sesgada por un año en particular donde haya habido muchas lesiones creamos un gráfico que nos muestre año a año la cantidad de lesionados por mes. A excepción del 2017 se observa que Abril siempre es de los meses con mayor cantidad. Una posible causa sea el fin de semana largo de Semana Santa donde el flujo vehicular es muy elevado.



Aprendizaje no supervisado

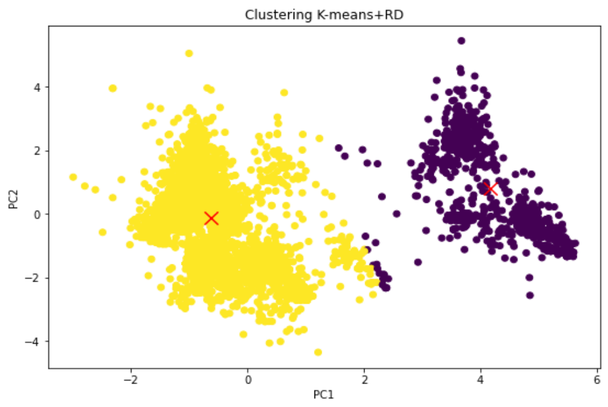
Se optó usar un modelado de datos a través de **clustering**, un método de aprendizaje no supervisado que nos va a permitir busca segmentar datos en subconjuntos o grupos.

En primer lugar reducimos 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

Luego usamos un modelo llamado K-Means 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). Obtuvimos un S = 0.65

Obtuvimos dos clusters, en amarillo el 1 y en violeta el 0



Resultados/Conclusiones

Cluster 1

•Compuesto por una amplia mayoría. 5769 intervenciones.

•Las intervenciones se dieron en su gran mayoría con un buen clima. La media es de 0.999827.

•Las intervenciones que involucran a motos son parte de este cluster.

Cluster 0

•Compuesto por la minoría. 869 intervenciones.

•Predomina el clima lluvioso en las intervenciones con una media de 0.966628. Esto se relaciona al estado de la ruta es mojada/húmeda en gran porcentaje.

•Colisiones con obstáculos fijos.