First Submission - Final Project Big Data

Paula Ramos, Karen Uribe-Chaves y Juan D. Urquijo

July 17, 2022
Repository Link: *Github*

El objetivo de nuestro Final project será replicar el análisis del paper *Comprehensive Analysis of Traffic Accidents in Seoul: Major Factors and Types Affecting Injury Severity* de Jeong et al (2022) con datos de Bogotá.

1 Paper Selection

Jeong et al (2022) investigan los índices de accidentalidad y mortalidad por accidentes de tráfico en Seúl - Corea del Sur, así como identificar los principales factores que afectan la gravedad de los accidentes. Por medio de modelos de aprendizaje supervisado (XGBoost y Regresiones Logit) y aprendizaje no supervisado (DBSCAN), predicen la severidad o gravedad de los accidentes, por medio de ciertas variables independientes: factores humanos, viales y ambientales. Los autores utilizaron eXtreme Gradient Boosting (XGBoost) por su robustez para el sobreajuste, Regresión logística por su superioridad para manejar datos categóricos y DBSCAN por la libertad del número de clusters.

Metodos analíticos

- *XGBoost*: Un algoritmo de conjunto que combina múltiples árboles de decisión y se basa en la metodología de boosting. Según los autores a través de este metodo, se deben optimizar tres hiperparámetros para evitar el sobreajuste y aumentar la precisión, y obtiene los siguientes resultados: la tasa de aprendizaje es de 0,1, la profundidad del árbol es de 3, el número de aprendices débiles es de 200 y la mayor precisión, que es del 68,95%.

- *Regresión logística*: Se usa como remplazo de la regresión lineal cuando la variable dependiente es binaria, en este caso puntual los autores buscan escoger entre la regularización L1 y L2 y optimizar el valor del parametro de regularización C. Encuentran que el mejor modelo es L1, con parametro C=1, dado que las métricas RMSLE, RMSE, MAE y los hiperparámetros de Accuracy en el set de training y en el de test son los mejores.

- *Clustering - DBSCAN*: En esta metodología hay dos parámetros de entrada: (i) eps y (ii) minPts. En el cual MinPts se fija en aproximadamente el doble de la dimensionalidad, es decir, 30 para un espacio de 14 dimensiones. Con respecto a el valor de eps, los autores trazan la distancia al (MinPts-1) vecino más cercano para cada uno de los puntos muestreados, ordenados en la muestra, encontrando la distancia a un *elbow* de la curva. Como resultado, obtienen tres clusters principales que corresponden al 97,8 del conjunto de datos.

Los resultados encontrados por los autores, dan cuenta de la poca relevancia de las condiciones climáticas contrario a lo intuitivo. En el modelo de regresión logística se evidencia que la infracción a la ley es la variable más relacionada con los accidentes de tráfico graves. Los autores resaltan la importancia de variables relacionadas con peatones en las tres aproximaciones, lo cual brinda evidencia a favor de la importancia de sus predicciones en las decisiones de política pública relacionadas con seguridad vial en Seúl.

2 Data

Basandosé en los datos del Seoul Metropolitan Government's Traffic Accident Dataset (SMGTAD) y el análisis descrito por Jeong et al (2022), se utilizarán las bases del Anuario de Siniestralidad Vial de 2017 y 2018 con información de Bogotá, Colombia. Los datos se pueden encontrar en: *simur.gov.co*. Las variables y datos a utilizar según factor humano, vial y ambiental son:

Table 1: Comparativo entre las varibles de referencia y los datos disponibles para replicar

Factores de análisis	SMGTAD	Anuario Siniestralidad Vial
Human Factor	Accident type	Atropello, incendio, caída de ocupante, choque, volacamiento, otro.
	Violation of law	Con embriaguez, con velocidad.
	Perpetrator's gender	f, m, n/a
	Perpetrator's age	De 1-98 años
	Victim's gender	f, m, n/a
	Victim's age	De 1-119 años
Road Factor	Road surface	Con huecos
	Road type	Cicloruta, glorieta, intersección, lote o predio, paso a nivel, paso elevado, paso inferior, ponton, puente, tramo de via, via peatonal.
	Perpetrator's vehicle type	Automovil, bicicleta, bicitaxi, bus, buseta, camion o furgon, camioneta, campero, cuatrimoto, microbus, motocarro, motocicleta, motociclo, tracto-camión , volqueta.
Environmental Factor	Occurrence day	1 enero 2017 a 31 Diciembre 2019
	Occurrence time	00:00 a 23:59
	Day of the week	Lunes, martes, miércoles, jueves, viernes, sábado, domingo.
	Weather	Granizo, Lluvia, Lluvia/Lluvia, Lluvia/Normal, Niebla, Niebla/Normal, Normal, Normal/Viento, Viento, Viento/Lluvia, Viento/Normal

Para finalizar, debido a los puntos en común entre las variables utilizadas por Jeong et al(2022) y las variables del Anuario de Siniestralidad Vial de Bogotá, se concluye que es posible replicar este ejercicio con las metodologías de machine learning vistas en clase.

3 Referencias

Jeong, H.; Kim, I.; Han, K.; Kim, J. Comprehensive Analysis of Traffic Accidents in Seoul: Major Factors and Types Affecting Injury Severity. Appl. Sci. 2022, 12, 1790. https://doi.org/10.3390/app12041790