Análisis de la Siniestralidad Vial en Bogotá: Factores que afectan la gravedad de los siniestros

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

August 2, 2022

Abstract

El objetivo de este trabajo es replicar el paper *Comprehensive Analysis of Traffic Accidents in Seoul: Major Factors and Types Affecting Injury Severity* de Jeong et al (2022) con datos del *Anuario de Siniestralidad Vial* 2017 y 2018 para Bogotá. Así, se utilizaron modelos de aprendizaje supervisado (XGBoost y Regresiones Logit) y aprendizaje no supervisado (DBSCAN) para predecir la severidad o gravedad de los siniestros viales en la ciudad. Como resultado de nuestro análisis se encuentra que los factores críticos que afectan la severidad de los accidentes dependen principalmente de factores humanos y de la vía, y en menor relevancia los factores ambientales. En particular para Bogotá, los vehículos de dos ruedas (motos y bicis), así como los peatones y los choques, son relavantes a la hora de predecir la severidad de los siniestros viales en la ciudad; resultado similar a lo encontrado por Jeong et al (2022) en Seúl.

Palabras clave: siniestros viales, víctimas, Machine learning, Logit, XGBoost, DBSCAN.

1 Introduction

El análisis de los siniestros viales se hace relevante para los expertos en movilidad y seguridad vial en las principales ciudades. En primer lugar, la siniestralidad vial se considera un tema de gran relevancia en la salud pública a nivel mundial. Según cifras de la Organización Mundial de la Salud (OMS) aproximadamente 1.3 millones de persona mueren al año como consecuencia de un siniestro vial. En Bogotá, cifras del Instituto Nacional de Medicina Legal y Ciencias Forenses indican que los siniestros viales son una de las principales causas de muerte en población joven por encima de enfermedades graves como el EPOC, Cáncer y Diabetes. Del mismo modo, si se analizan las muertes violentas en Bogotá, la siniestralidad causa más víctimas fatales que los suicidios (Anuario de Siniestralidad Vial de Bogotá 2019).

Por otro lado, la siniestralidad vial también representa un costo económico para el país. A pesar de la reducción en las víctimas fatales en siniestros viales en Bogotá durante 2021, el Plan Nacional de Desarrollo 2018-2022 muestra que la siniestralidad vial implica un costo del 3.6% del PIB del país año tras año, siendo Bogotá la ciudad que más fallecidos por esta causa presenta en el país. Según el Observatorio Nacional de Seguridad Vial, en Colombia durante 2021 se observó un incremento de la siniestralidad del 35.3% respecto a 2020 y 12% frente a 2019, lo que implica 7,720 personas fallecidas en siniestros viales.

Dada la relevancia del análisis de siniestralidad vial y tomando como referencia el estudio de Jeong et al (2022) para la ciudad de Seúl, surge el interés de comprender las principales causas y variables que generan siniestros viales graves en Bogotá, con el fin de identificar factores claves que permitan formular soluciones que reduzcan los siniestros fatales en la capital del país. Este trabajo aporta a la literatura reciente, como estudio de los datos de siniestralidad en Bogotá para descubrir elementos relacionados con los siniestros mortales, y aportar a políticas de seguridad vial enfocadas en aquellos elementos de mayor relevancia.

A partir del uso de modelos de Machine Learning de aprendizaje supervisado (Logit y XGBoost) y no supervisado (DBSCAN), identificamos para los años 2017 y 2018 las principales variables que generarían un siniestro vial con víctimas fatales en la ciudad. La idea clave de esta investigación es examinar factores correlacionados en la gravedad de los siniestros viales mediante la aplicación de estos modelos. Se encuentra que los factores ambientales no tienen incidencia en la predicción, sin embargo el tipo de accidente, la edad y el género como factores humanos, así como el tipo de vehículo son relevantes para el análisis de la siniestralidad vial; en línea con lo encontrado por Jeong et al (2022).

El documento se encuentra organizado de la siguiente manera: en la sección 2 se explican los datos utilizados y una breve descripción de los mismos. En la sección 3, se explica la construcción y desarrollo de los tres modelos mencionados y sus resultados. Finalmente, las conclusiones de este trabajo se presenta en la sección 4.

2 Data

Los datos que se usaron para predecir la severidad de los siniestros en Bogotá, correspoden a la información del *Anuario de Siniestralidad Vial* para los años de 2017 y 2018, compuesta por 3 bases:

1. *Siniestros*: Conjunto de datos sobre las características de los 72,123 siniestros ocurridos en los dos años, tales como: fecha, tipo de siniestro, tipo de vía, gravedad del siniestro, situación de embriaguez o de velocidad, entre otras.

2.*Conductores*: Datos acerca de los 137,012 conductores involucrados, como el vehículo en el que conducían, la edad, el sexo, si llevaba cinturón, chaleco y/o casco.

3. *Víctimas*: Información sobre las 18,468 víctimas del siniestros, tales como edad, sexo, si eran peatones, si llevaban cinturón, chaleco y/o casco, vehículo en el que viajaban.

Los autores del paper se enfocan en la variable de severidad con dos categorías "Severo" o "Leve", en la base se crea la variable en función de la gravedad (Con muertos, con heridos y con daños). Además, en el estudio identifican tres factores que agrupan varias variables:

- *Human factor*, que contiene el tipo de accidente, la violación a la ley, las características del cupable y las víctimas.

- *Road Factor*, en esta tienen en cuenta el tipo de vía, el tipo de vehículo del culpable y las víctimas.

- *Environmental Factor*, en la cual se encuentrasn la estación, el día de la semana, tiempo y clima.

Es importante mencionar que no es posible determinar el culpable del accidente, por lo tanto, se toma al conductor como proxy. Además, en algunos accidentes hay más de un conductor involucrado, de modo que se generan variables a nivel de siniestro, como el número de vehículos involucrados, la edad promedio de conductor, la edad promedio de la víctima, el número de mujeres y hombres involucrados, los grupos etareos conductores y víctimas. Adicionalmente, se crea la variable tiempo (Día, noche y amanecer) y se recategoriza el Clima (Soleado, lluvioso y niebla).

Se divide la base de manera aleatoria, con el fin de entrenar los modelos con train y probar su eficiencia en la base test. Posterior a ello, se aplica el enfoque de remuestreo *Up-sampling* para balancear la muestra.

3 Models and Results

3.1 Logit

Para determinar la probabilidad de que un siniestro sea severo, se especifica el siguiente modelo:

$$Gravedad = \beta X + u \tag{1}$$

Donde *Gravedad* es la variable dependiente de severidad del siniestro, con las categorías 1 - Severo y 0 - Leve. Por su parte, X representa la matriz de covariables: *TipoAccidente, Estado de mbriaguez, Infracción por velocidad, Otro tipo de infracción, Número de hombres y Número de mujeres conductores y víctimas, Categorías de edad conductores y víctimas, Número de autos, de vehiculos de transporte público, de motos, de biciletas, de otros vehiculos y de peatones* (solo para víctimas) [Factor humano]; *Tipo de vía, Vía con huecos* [Factor vial]; *Tiempo, Día y Clima* [Factor ambiental].

El mejor modelo es un logit lasso con α de 0, parámetro λ de 0.0245 y un porcentaje de precisión de 92%. Los demás hiperparámetros del modelo en la base train son los siguientes:

Table 1: Hiperparámetros del modelo logit

	lambda	ROC	Sens	Spec	Accuracy	Kappa
60	0.02	0.97	0.95	0.90	0.92	0.85

Una vez seleccionado el modelo, se realizan las predicciones en la base test y se clasifican los accidentes como Severo o Leve tomando como punto de corte 0.5 y 0.52 según el threshold. Para ambos puntos de corte se obtienen resultados satisfactorios, sin embargo, se toma el cutoff de 0.5, para reducir el error tipo II, es decir, predecir la mayor cantidad de los siniestros severos. En la predicción de la base test se observa un porcentaje de falsos positivos de 2.7% y de falsos negativos 16.9%.

La importancia de las variables se encuentra en la Figura A.3 del anexo de este trabajo. Se resalta la incidencia de los vehículos tipo Motocicleta y Bici (Dos ruedas) de los conductores involucrados, así como el género de las víctimas y los peatones como variables más relevantes en la predicción realizada.

3.2 XGBoost

Las variables usadas para estimar este modelo fueron las mismas usadas en el modelo Logit anterior. Se usó una grilla estandar con rango de rondas 250 a 500. Una profundidad del árbol que ajusta el modelo de 4, 6 y 8 nodos. Un rango de la tasa de aprendizaje del modelo entre 0.01, 0.3 y 0.5. Un rango de observaciones en la región final del arbol de entre 10, 25 y 50. Dando como resultado

- *ROC, Sens y Accuracy*: El árbol escogido fue el que mejor desempeño tuvo en el parametro Sens de 0,8841659. El ROC fue de 0,9781584 y el Accuracy fue de 0,9259324.
- *Hiper-parámetros del mejor árbol estimado*: Tasa de aprendizaje del modelo de 0,3; profundidad del árbol 6 nodos; penalización por particiones del árbol 1; porcentaje de subsampleo de columnas para el árbol 0,7; número de observaciones en la región final del árbol de 50; y porcentaje de subsampleo de observaciones para el árbol 0,6. El número de rondas de árboles fue 250.
- *Importancia Variables*: En el árbol estimado A.2, se encuentra que las variables más importantes según la estimación son: i) Número de motos conductor, ii) Tipo de accidente = Choque, iii) Número de Bicis conductor y iv) Número de peatones víctimas. El ranking completo se encuentra en la Figura A.3 del anexo de este trabajo.

3.3 DBSCAN

DBSCAN es un algoritmo de aprendizaje automático no supervisado que forma las agrupaciones según la densidad de los puntos de datos o qué tan cerca están los datos, y aquellos puntos que se encuetran por fuera de las regiones con alta densidad se consideran ruido o *outliers*. En esta metodología tiene un enfoque paramétrico que funciona con dos parámetros: (i) eps y (ii) minPts.

Table 2: Caracteristicas de los Clusters - DBSCAN

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Tiempo	Día 67,82%	Día 0,42%	Día 0,50%	Día 0,07%
Día	Viernes 15,83%	Lunes 0,11%	Sábado 0,15%	Miércoles 0,02%
Tipo accidente	Choque 84,70%	Choque 0,40%	Choque 0,60%	Choque 0,13%
Huecos	No 98,08%	Si 0,69%	No 0,83%	No 0,13%
Clima	Soleado 94,91%	Soleado 0,66%	Soleado 0,82%	Soleado 0,13%
Tipo de vía	Tramo de vía 78,91%	Tramo de vía 0,63%	Tramo de vía 0,70%	Tramo de vía 0,10%
Conductores Moto	0 74,43% 1 22,20%	0 0,37% 1 0,32%	0 0,70% 1 0,13%	0 0,10% 1 0,03%
Conductores Hombre	2 61,61%	1 0,38%	1 0,46%	1 0,04%
Edad Víctima	Joven 86,25%	Joven 0,59%	Joven 0,68%	Joven 0,04%
Peatón Víctima	0 87,61% 1 9,69%	0 0,67% 1 0,02%	0 0,60% 1 0,23%	0 0,13% 1 0,003%
Víctima Mujeres	0 87,36% 1 9,34%	0 0,56% 1 0,11%	0 0,70% 1 0,12%	0 0,04% 1 0,08%

- *Epsilon (eps)*: Es el radio de los vecinos alrededor del punto p de la muestra. En el caso de nuestro modelo, el parámetro se fijó en 10.
- *minPts*: Es el número mínimo de puntos de datos en la vecindad de un punto particular para definir un clúster. Este parámetro se fijó en 62 siguiendo la metodología de Jeong et al (2022) por el doble de la dimensionalidad.

El resultado del análisis arrojó cuatro (4) clusters y 150 puntos de ruido, como se evidencia en la Tabla 2. Dada la alta frecuencia de Choques, siniestros en Tramo de la Vía y Víctimas Jóvenes, el clúster 1 es el que más observaciones contiene de toda la muestra. Sin embargo, los resultados no son concluyentes dado que los 4 clusters encontrados únicamente explican el 17% de la varianza muestra según lo explicado en los componentes principales 1 y 2, a pesar de que el cluster 1 recoge el 98% de las observaciones. La distribución de los clusters se evidencia en la Figura A.4 del anexo a este trabajo.

No obstante, consideramos relevante mencionar en comparación con los resultados de Jeong et al (2022), que en efecto los factores ambientales no parecen tener incidencia en la severidad de los siniestros viales; sin embargo el género, la edad y el tipo de vehículo moto en Bogotá si juegan un papel importante en la predicción.

4 Conclusions and recommendations

A partir de los datos de siniestralidad vial para Bogotá en los años 2017 y 2018, se identificaron los principales factores que afectan la gravedad de la siniestralidad vial usando modelos de Logit Lasso, XGboost y DBSCAN.

Para el caso de Logit- Lasso, se encontró que las variables más relevantes son *Conductores en moto, Conductores en bici, Conductores en automóvil, Mujeres como víctimas y Choques.* Para XGBoost, los resultados fueron similares arrojando como variables más relevantes *Conductores en moto, Choques, Conductores de bici, Peatones víctimas, Conductores en automóvil y Vícima joven*,

siendo consistente con las encontradas por la regresión logística. Por su parte, en DBSCAN los resultados únicamente explican el 17% de la variabilidad de los datos a pesar de incluir el 98% de las observaciones en los clusters encontrados, sin embargo resaltan el género, la edad y el tipo de vehículo moto como relevantes en el cluster 1, que es el que recoge el mayor número de observaciones.

Los resultados encontrados coinciden con lo reportado por Jeong et al (2022), quienes indican que los factores ambientales no tienen incidencia en la predicción, como el clima. Sin embargo, algunos modelos si arrojaron relevante la hora del día (Madrugada) y el día de la semana (Viernes). Por su parte, también coinciden los resultados en resaltar que los peatones involucrados en un siniestro vial son relavantes a la hora de predecir la gravedad, ya que en su mayoría son víctimas fatales. Finalmente, los vehículos como motocicletas o bicicletas, también resultan relevantes en la severidad de los siniestros, brindando evidencia a favor de los resultados encontrados por los autores ya mencionados.

Para futuros estudios, se recomienda que los datos utilizados puedan identificar al culpable del siniestro con el fin de agregar otro escalón al análisis ya presentado. En términos de relevancia para la formulación de políticas públicas de seguridad vial, nos unimos a la recomendación de implementar medidas preventivas para peatones así como consciencia en los conductores, ya que son el actor vial más vulnerable. Por otra parte, y en particular para Bogotá, que tiene un parque automotor en crecimiento, especialmente en motocicletas, recomendamos medidas más contundentes respecto a la expedición de licencias de conducción y el respeto por las normas de tránsito.

5 Data and Code availability

El desarrollo de este trabajo, código y datos pueden encontrarse en:

- Repository Link: *Github*
- Data: Los datos del Anuario de Siniestralidad Vial para 2017 y 2018 se pueden encontrar en *simur.gov.co*.

6 References

- ITF (2021), Road Safety Annual Report 2021: The Impact of Covid-19, OECD Publishing, Paris.
- Organización Mundial de Salud - OMS (2020). Global Status Report on Road Safety 2019.
- Jeong, H., Kim, I., Han, K., Kim, J. (2022). Comprehensive Analysis of Traffic Accidents in Seoul: Major Factors and Types Affecting Injury Severity. Applied Sciences, 12(4), 1790.
- Secretaría de Movilidad (2020). Anuario de Siniestralidad Vial 2019. Visión Cero BOG. Alcaldía Mayor de Bogotá.

A *Appendix*

Figure A.1: Importancia de las Variables Logit-Lasso Up Sampling

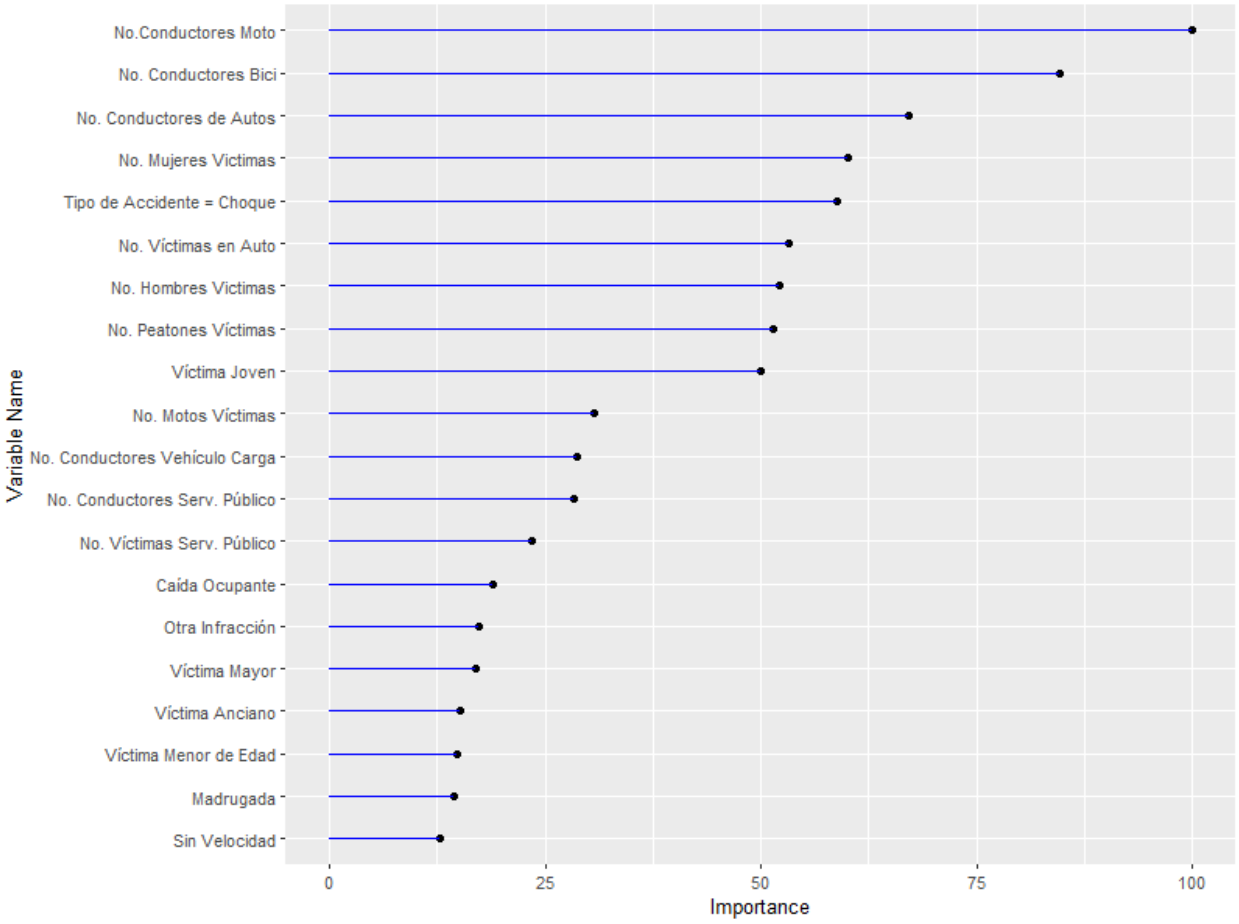

Figure A.2: Árbol estimado por XGBoost

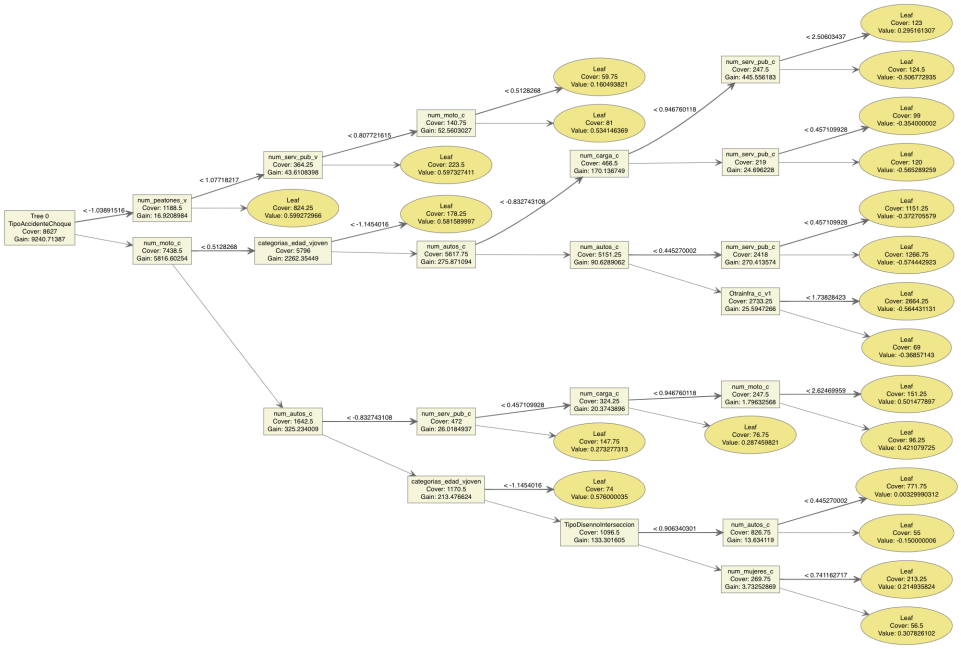

Figure A.3: Importancia de las Variables XGBoost

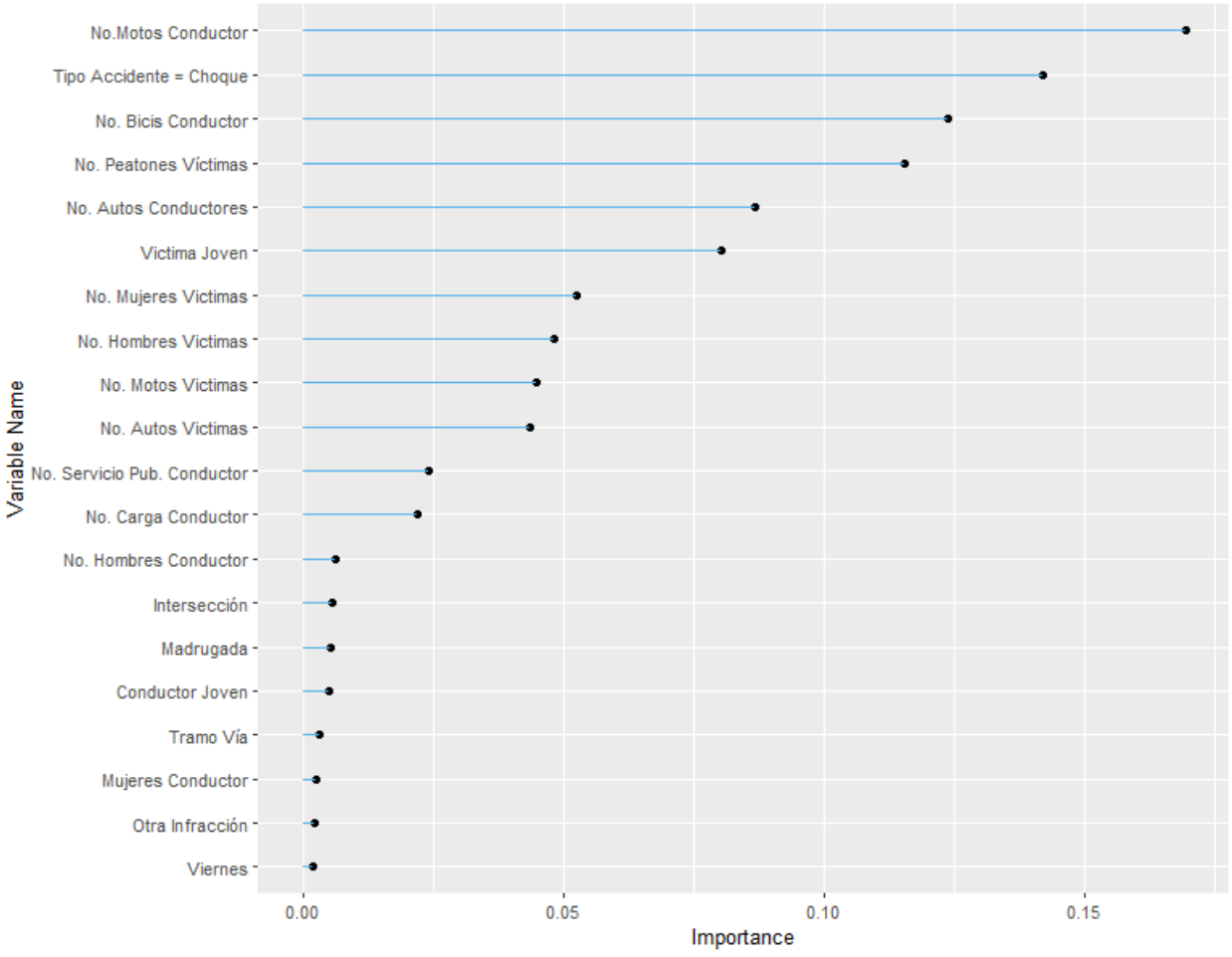

Figure A.4: Clusters DBSCAN