MAESTRÍA EN CIENCIA DE DATOS

Fecha de entrega:

Estudiantes: Ruth Alexandra Ruiz Rey

Luis Alejandro Rodríguez Daza

Director: Wilmer Edicson Garzón Alfonso

Codirector: Dante Conti

El presente documento avala la entrega del trabajo de grado por parte del director.

Documentos anexos: copia digital del Trabajo de Grado (1).

_____

Firma Director

_____

Firma Codirector

Ruth Alexandra Ruiz

_____

Firma Estudiante 1

Alejandro Rodriguez

_____

Firma Estudiante 2

Diseño, construcción e implementación de un visualizador basado en modelos de clasificación y predicción relacionado con las siniestralidades viales para algunas ciudades de Colombia

Luis Alejandro Rodriguez Daza
Ruth Alexandra Ruiz Rey

Trabajo de grado para optar al título de
Magíster en Ciencia de Datos

Director
Wilmer Edicson Garzón Alfonso
Doctor

Codirector
Dante Conti
Doctor

Maestría en Ciencia de Datos
Bogotá D.C., Colombia
2023

Publicado en 2023 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Reconocimiento o Agradecimientos

Queremos expresar nuestros más sinceros reconocimientos y profundos agradecimientos a todas las personas que contribuyeron de manera significativa a la realización de este documento de tesis de maestría. Es un honor poder dedicar esta sección para expresar nuestra gratitud hacia aquellos que han sido pilares fundamentales en nuestro camino académico.

En primer lugar, queremos agradecer a nuestro director de tesis por su guía experta, apoyo y constante motivación a lo largo de este arduo proceso. Sus conocimientos, experticia y dedicación han sido invaluables para alcanzar los objetivos propuestos en esta investigación. Su paciencia y disposición para brindarnos retroalimentaciones constructivas nos han permitido crecer como estudiantes y como profesionales.

Asimismo, nos gustaría extender nuestros reconocimientos a todos los profesores y profesoras que compartieron su sabiduría y conocimientos con nosotros a lo largo de nuestra carrera de maestría. Sus enseñanzas y mentoría han sido fundamentales en nuestra formación académica y personal. Agradecemos su compromiso y dedicación en impartir educación de calidad y su disposición para resolver las dudas y orientarnos en la dirección correcta.

No podemos dejar de mencionar a nuestros compañeros y compañeras de clase, quienes nos brindaron apoyo emocional y compartieron experiencias en este viaje académico. Nuestras discusiones y debates enriquecieron nuestras perspectivas y nos impulsaron a esforzarnos aún más. Sus palabras de aliento y amistad han sido un pilar fundamental para superar los momentos desafiantes y celebrar los logros alcanzados juntos.

Además, deseamos expresar nuestra gratitud a nuestras familias por su amor incondicional, comprensión y apoyo constante a lo largo de nuestra formación académica. Su confianza en nosotros y su sacrificio han sido la fuerza motriz detrás de nuestros logros. Agradecemos de corazón el respaldo que siempre me nos han brindado y que ha sido fundamental para el crecimiento y éxito.

Resumen

La accidentalidad vial es una constante de todas las ciudades del mundo. En Colombia, por ejemplo, se registraron 618 accidentes en el primer mes del año 2023, lo que agrava la tasa de mortalidad en el país, una tendencia que según la agencia nacional de seguridad vial va en aumento año tras año. Adicionalmente cada accidente genera repercusiones en la movilidad de la ciudad, y repercusiones económicas para el gobierno, los accidentados y las aseguradoras. Dada la situación se vuelve necesario analizar todas las variables sobre los accidentes para así encontrar los patrones más comunes y determinar las mejores acciones a seguir.

El proyecto aborda esta problemática mediante el tratamiento de datos utilizando la metodología KDD. A través de un dashboard se hace un seguimiento detallado de las causas de los accidentes viales, tanto los actores, características de los accidentes y dando así también un panorama general de las tendencias de los accidentes en las ciudades analizadas.

Adicionalmente, mediante técnicas de clusterización y análisis de minería de datos se obtienen resultados sobre las hipótesis sobre cómo se comportan los accidentes, para así hacer uso de modelos de aprendizaje automático para obtener las proyecciones de cuantos accidentes ocurren en cada una de las zonas. Esto genera herramientas con las que se podrían generar procesos de toma de decisiones sobre las zonas más accidentadas.

Abstract

Road accidents are a constant in all cities of the world, in Colombia for example, 618 accidents were generated in the first month of the year 2023, this becomes one more cause of mortality in the country since it increases year by year according to the national road safety agency, additionally each accident generates repercussions in the mobility of the city, and economic repercussions for the government, the injured and insurers. Given the situation, it becomes necessary to analyze all the variables on accidents in order to find the most common patterns and follow up on the best actions.

The project uses KDD methodology for data processing, and thus, through a dashboard, a detailed follow-up of the causes of road accidents, the actors, the characteristics of the accidents and also giving a general overview of where accidents occur in the cities analyzed.

Additionally, through clustering and data mining analysis techniques, results are obtained on the hypotheses about how accidents behave, in order to make use of machine learning models to obtain projections of how many accidents occur in each of the areas. This generates tools that could be used to generate decision-making processes on the most accident-prone areas.

Tabla de contenido

1 Introducción

1.1 Problemática

El estudio de la seguridad vial abarca una amplia gama de disciplinas, lo que refleja la complejidad y la interconexión de factores involucrados en la seguridad en las carreteras. Desde la ingeniería de tránsito hasta la epidemiología, pasando por la legislación, la educación y la tecnología, cada disciplina aporta una perspectiva única y fundamental para comprender y abordar este desafío. La minería de datos emerge como un enfoque especialmente valioso en este contexto, ya que permite analizar grandes conjuntos de datos para identificar patrones, tendencias y factores de riesgo que influyen en la seguridad vial.

Los datos recopilados y analizados no solo revelan ubicaciones críticas y condiciones climáticas asociadas con accidentes, sino que también proporcionan información crucial sobre factores humanos y vehiculares involucrados. Esta información es fundamental para que las autoridades tomen medidas preventivas antes de que ocurran accidentes, comprendan las causas subyacentes de los mismos y diseñen intervenciones específicas para abordarlas. Además, los datos demográficos y de comportamiento permiten una personalización más efectiva de las campañas de concienciación, al dirigirse a grupos específicos de población con necesidades y riesgos particulares.

La capacidad de realizar análisis en tiempo real proporciona una respuesta más rápida de los servicios de emergencia y la implementación de medidas correctivas eficaces. Además, los modelos predictivos de riesgo en carreteras, alimentados por datos, ayudan a planificar intervenciones preventivas y a asignar recursos de manera más eficiente. En conjunto, la combinación de la minería de datos con un enfoque centrado en la seguridad vial proporciona una visión más profunda de los desafíos y oportunidades para mejorar la seguridad en las diferentes tipologías de vías.

Adicionalmente, este estudio cobra una importancia aún mayor en el contexto colombiano, donde las cifras alarmantes de siniestralidad vial reflejan la urgente necesidad de acción. En el año 2022, se reportaron aproximadamente 41,147 siniestros viales, según datos de la Agencia Nacional de Seguridad Vial (ANSV), en los que fallecieron 8,348 personas y resultaron lesionadas 39,396, representando un incremento del 6.8% en víctimas fatales respecto al año 2021. Este aumento subraya la necesidad de abordar este problema de manera integral y urgente. Además, los Objetivos de Desarrollo Sostenible (ODS) establecidos en la Agenda 2030 enfatizan la importancia de mejorar la seguridad vial como parte de una agenda global de salud y desarrollo sostenible. En este sentido, las metas específicas relacionadas con la seguridad vial destacan la importancia de reducir las muertes y lesiones por accidentes de tráfico, así como promover sistemas de transporte sostenible que aborden las necesidades de grupos vulnerables de la sociedad. En resumen, el estudio de la seguridad vial y su aplicación práctica a través de la minería de datos no solo es fundamental para mejorar la seguridad en las carreteras, sino que también contribuye al logro de objetivos más amplios de desarrollo y bienestar público.

Se propone en este trabajo una metodología escalable, aplicable a cualquier ciudad que requiera ser analizada, en contraste con la práctica común de enfocarse únicamente en las ciudades principales. La flexibilidad de esta metodología permite adaptarse a las necesidades específicas de cada ciudad, aunque requiera el reentrenamiento de modelos según el caso. En líneas generales, los pasos seguidos en este estudio pueden servir como guía para investigar las características demográficas y circunstanciales de los accidentes en distintas ciudades.

De esta manera se eligieron tres ciudades que cuentan datos abiertos, las ciudades elegidas para el proyecto (Bogotá, Medellín y Barranquilla), con el fin de clasificar, predecir y dejar evidencias que permitan tomar decisiones para reducir los índices de siniestralidad vial. Con el objetivo de tomar decisiones informadas, se ha llevado a cabo un proceso de complementación de información y se han desarrollado métodos de clusterización para obtener un mayor conocimiento sobre las diversas zonas de cada ciudad. Además, se ha implementado un modelo de predicción que permite comprender en profundidad cada área, ya que se considera que existen patrones diferenciados según el día de la semana y el tipo de zona geográfica.

La clasificación se enfoca en agrupar áreas similares en función de sus características actuales, mientras que la predicción se centra en estimar valores futuros o tendencias en función de patrones históricos y factores específicos. Ambos enfoques son valiosos para la toma de decisiones informadas en la reducción de la siniestralidad vial.

El trabajo se limita a la información de los accidentes en general, puesto que los datos pluviométricos y en general los demás datos climáticos no se encuentran disponibles por zonas de la ciudad y en ciudades grandes no se tiene el mismo patrón del clima en todas las zonas geográficas. Adicionalmente no todas las ciudades cuentan con información geográfica de los accidentes, incluso en algunas solo se encuentran los recuentros de todos los accidentes sin características específicas de cada uno de estos.

1.2 Justificación

Los resultados de los análisis de accidentalidad vial presentados se deben ajustar a la realidad de los que se vive día a día, puesto que las condiciones por tipo de actores viales y patrones van cambiando año a año, por políticas gubernamentales, aumento de autos en las ciudades, deterioro de las vías u otros factores económicos. Por lo tanto, se presentará unos resultados que se pueden replicar año con año para que sean actualizables y escalables en diferentes lugares.

1.3 Planteamiento del problema

La siniestralidad vial es una de las problemáticas más relevantes y preocupantes en los países, y Colombia no es una excepción. La necesidad de ampliar los estudios existentes sobre accidentes de tránsito es crucial, ya que las cifras de siniestros viales tienden a aumentar constantemente en el país (Agencia Nacional de Seguridad Vial (ANSV), 2020). Para abordar esta problemática de manera efectiva, es necesario realizar un análisis profundo de diversos factores relacionados con la movilidad en cada ciudad.

Cada ciudad presenta características únicas que influyen en la ocurrencia de accidentes de tránsito, y, por lo tanto, es fundamental considerar factores como la demografía, la densidad vehicular, las infraestructuras viales y el comportamiento de los usuarios de las vías. Estos factores varían de una ciudad a otra y tienen un impacto significativo en la cantidad y gravedad de los siniestros.

Realizar un análisis exhaustivo de estos factores permitirá clasificar y predecir los siniestros de manera más precisa y eficiente. Al comprender cómo se interrelacionan estos factores y cómo afectan la seguridad vial, se pueden implementar medidas y políticas específicas para mitigar los riesgos y mejorar la seguridad en las vías.

Es importante destacar que la prevención y gestión de los accidentes de tránsito requieren un enfoque integral y multidisciplinario. No se trata solo de implementar medidas de control y vigilancia, sino también de promover una cultura de seguridad vial, mejorar la infraestructura vial, fomentar la educación y concienciación de los conductores y peatones, así como garantizar una respuesta efectiva en caso de siniestros.

1.4 Objetivos y Pregunta de Investigación

Objetivo general

Generar modelos de clasificación que permitan segmentar las zonas viales y modelos de predicción sobre la siniestralidad por zona, para tener sugerencias de vialidad teniendo en cuenta la recopilación histórica de accidentes de las principales ciudades de Colombia.

Objetivos específicos

- Construir una base de datos de siniestralidad a nivel nacional y con diferenciaciones entre ciudades apta para el proyecto.
- Establecer criterios de clasificación sobre los diferentes aspectos de los accidentes.
- Segmentación por nivel de accidentalidad en cada zona de cada ciudad.
- Generar un modelo que prediga la accidentalidad vial por periodos y por ciudad/zona.
- Seleccionar y comparar un grupo de modelos avanzados que permita mejorar la predicción de la siniestralidad vial en varias ciudades del país.
- Realizar un tablero de resultados.

1.5 Antecedentes o estado del arte

El presente trabajo busca pronosticar la tasa de siniestralidad vial de varias ciudades principales de Colombia. Para identificar patrones en relacionados en la accidentalidad y así agruparlo en características determinantes, por ejemplo, zonas, días de la semana, tipo de accidentes y/o agentes viales

No obstante, existen enfoques que ya se han realizado sobre la ciudad de Bogotá, como es en el caso de Gutiérrez-Osorio y Pedraza (2019) en el cual se hace una caracterización de los accidentes para los años 2016 y 2017 con diferentes algoritmos dando resultados de: análisis de distribución y clasificación de accidentes de tráfico. Nuestro estudio se diferencia del realizado por Gutiérrez-Osorio y Pedraza (2019) en varios aspectos:

Alcance y extensión: Mientras que el estudio de Gutiérrez-Osorio y Pedraza (2019) se centró exclusivamente en la ciudad de Bogotá y analizó los datos de accidentes para los años 2016 y 2017, nuestro proyecto tiene un alcance más amplio. Analizaremos las características actuales no solo de Bogotá, sino también de otras ciudades principales de Colombia, Medellín y Barranquilla.

Enfoque metodológico: Aunque ambos estudios emplean algoritmos para caracterizar los accidentes de tráfico, nuestra investigación busca ir más allá. Utilizaremos diferentes técnicas, incluyendo métodos de clusterización y modelos de predicción, con el objetivo de obtener una visión más profunda y detallada de

las zonas de cada ciudad y entender mejor los patrones relacionados con el día de la semana y la geografía.

Objetivos y conclusiones: Es probable que las preguntas de investigación y los objetivos de nuestro proyecto difieran de los planteados por Gutiérrez-Osorio y Pedraza (2019). En consecuencia, las conclusiones y recomendaciones que obtengamos pueden variar y contribuir con nuevos conocimientos a la problemática de la siniestralidad vial en las ciudades analizadas.

Conforme a ello, el presente marco conceptual presentado en esta investigación se basa en el modelo básico utilizado frecuentemente en ciencia de datos de 5 etapas, las cuales son: Recolección de datos, ETL, minería de datos, aprendizaje automático e interpretación y evaluación (Pérez, et al 2018), y la forma en la cual cada una de estas etapas son igualmente importantes para lograr el objetivo general del proyecto. Se tendría en cuenta una perspectiva que en muchas ocasiones se le presta muy poca atención y se refiere a la perspectiva del lado de los ciclistas. Para esto, se tendría en cuenta, trabajos anteriores en esta materia (Daraei et al, 2021) y se aplicaría a la movilidad colombiana. En donde se realizará un análisis exploratorio exhaustivo para conocer cuáles son los estímulos externos que afectan a las percepciones de los conductores en general.

Después de haber analizado la manera de enfocar el problema de (Sohn & Lee, 2003), donde se muestra una técnica interesante de clasificación, se encuentra que los algoritmos de fusión son más efectivos para la clasificación de datos que si lo hiciéramos con un solo clasificador. Los factores que se usan en ese estudio son: Tamaño de la vía, forma del carro, categoría del accidente, velocidad, mecanismos de protección, los cuales determinan la severidad del accidente. Estos resultados son relevantes, pero debido a que la tipología de los accidentes en Colombia difiere a otros países dado que las características de las vías son distintas él estudio también difiere en sus resultados siendo muy complejo incorporar estos algoritmos con la información disponible.

Un enfoque adicional direccionado al estudio de imágenes requiere acceso a datos en tiempo real, lo cual generaría un valor muy grande para la detección de accidentes y acción rápida sobre éstos. En las ciudades con el uso de GPS y con el uso de redes neuronales (Parsa et al., 2019) se desarrolla un estudio por medio de series de tiempo en el cual hace la clasificación de que se considera como un accidente y que se considera no accidente, las técnicas propuestas Long Short-Term Memory (LSTM) y Gated Recurrent Unit (GRU) permitían hacer una detección rápida de los accidentes.

Cabe agregar que los accidentes son inducidos por muchos factores, uno de estos factores son las condiciones climáticas, las cuales afectan tanto la visibilidad de los conductores como las condiciones en las que estos conducen (Novkovic et al., 2017) el artículo determina una alta correlación entre los accidentes de tráfico la lluvia, nieve y las altas temperaturas. En este estudio se quiso explorar de nuevo esta correlación debido a que Colombia se encuentra en una ubicación geográfica en el trópico.

Por otra parte, el método usado por (Yuan, Zhou, & Yang, 2018) muestra un tratamiento de datos interesante para usarlo en una ciudad como Bogotá puesto que se caracteriza el problema por medio de cuadriculas y nuevamente se usa una red neuronal Long Short-Term Memory (LSTM) que permite predecir el comportamiento del tráfico y por medio de dichas cuadriculas dar una solución escalable al problema.

Por último, en cuanto a la parte de los pronósticos, se cubriría una gran gama de metodologías y/o algoritmos básicos y avanzados de aprendizaje automático (Gutierrez-Osorio & Pedraza, 2020), que

guiaran las acciones y recomendaciones para el éxito del proyecto, entre ellas, tipo de vehículo más seguro, calles y/o carreras con menos siniestralidades viales, estimar un costo con más precisión, dependiendo del tipo de accidente de tráfico. En este sentido, el objeto de investigación recae en como los factores socio geográficos de las distintas ciudades a analizar estimulan a las actitudes y percepciones de los conductores y como ellos a su vez reaccionan por medio de las recomendaciones fruto de este trabajo.

1.6 Alcance y limitaciones

En este proyecto, se seleccionaron tres ciudades importantes de Colombia para su análisis en relación con la seguridad vial. Estas ciudades fueron elegidas debido a su relevancia y tamaño, lo que permite obtener una muestra representativa de las dinámicas de tránsito y accidentes en el país. Las ciudades seleccionadas fueron Bogotá, Medellín y Barranquilla.

Construir una base de datos de siniestralidad a nivel nacional y con diferenciaciones entre ciudades apta para el proyecto es uno de nuestros objetivos clave. Sin embargo, al trabajar con datos abiertos disponibles, surgen ciertas limitaciones en términos de la cantidad y calidad de variables disponibles en cada base de datos. Es importante tener en cuenta que la recopilación de datos de accidentes puede variar entre las diferentes ciudades y las fuentes de información utilizadas. Por lo tanto, es posible que las bases de datos disponibles para cada ciudad no contengan la misma cantidad de variables o que la información recolectada difiera en términos de su nivel de detalle o precisión.

Estas diferencias en las variables disponibles pueden afectar el análisis comparativo entre las ciudades y limitar la capacidad de realizar análisis más profundos o detallados. Por ejemplo, si una ciudad tiene datos más completos y detallados sobre el tipo de accidente, las condiciones climáticas o el estado de las carreteras, mientras que otra ciudad carece de esa información, puede haber limitaciones en la capacidad de realizar comparaciones precisas y exhaustivas entre ellas.

A pesar de estas limitaciones, se realizaron esfuerzos para maximizar la calidad y utilidad de los datos disponibles. Se llevó a cabo un proceso de limpieza y estandarización de los datos, se realizaron transformaciones y se generaron variables adicionales cuando fue posible. Además, se aplicaron técnicas de análisis estadístico y modelado para extraer información significativa de los datos disponibles y proporcionar una comprensión más profunda de las dinámicas de accidentes en cada ciudad.

2 Fundamentos Teóricos

Esta parte del proyecto de grado describe los elementos en los que se basa el estudio y el conjunto de modelos estadísticos para el procesamiento de la información que sustenta el procesamiento del estudio.

2.1 Minería de datos

En la actualidad, con el avance de la tecnología, se genera y recopila continuamente una gran cantidad de datos sin procesar heterogéneos. La minería de datos es el análisis de la recopilación, limpieza, procesamiento, análisis y adquisición de información valiosa mediante el descubrimiento de patrones y relaciones ocultas o implícitas en los datos (Aggarwal, 2015).

Las fases de recolección, preprocesamiento y análisis de datos conforman el proceso de minería de datos, donde el desafío de cada aplicación es diferente (Aggarwal, 2015). Como resultado, se utilizan varias metodologías, que incluyen *Knowledge Discovery in Databases* (KDD), *Cross Industry Standard Process for Data Mining* (CRISP-DM) y SEMMA (*Sample, Explore, Modify, Model, Assess*), para aplicar y llevar a cabo este proceso. Estas metodologías ofrecen una guía para llevar a cabo el proceso de manera sistemática y desafiante (Moine et al., 2011).

2.2 Aprendizaje Automático (Machine Learning)

Según Ertel (2017) y Géron (2017), el aprendizaje automático es la rama de la informática que permite a las computadoras aprender de los datos sin programación explícita mediante la ejecución de tareas repetidas. El objetivo es crear modelos que puedan predecir con precisión las etiquetas o valores numéricos de los registros que se desconocen, o modelos con buena capacidad de generalización. El conjunto de datos se divide en dos categorías en este sentido: el conjunto de entrenamiento, que contiene el conocimiento que el algoritmo extraerá y aprenderá, y el conjunto de prueba, que permite evaluar, a través de una métrica de desempeño, si el algoritmo tiene poder de generalización con nuevos datos (Ertel, 2017).

Según el nivel y el tipo de supervisión, estos algoritmos se dividen en cuatro grupos: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje por refuerzo. Los algoritmos que se discutirán a continuación están relacionados con el aprendizaje supervisado.

2.2.1 Regresión lineal múltiple

El modelo de regresión lineal simple es una técnica de modelado estadístico que tiene como propósito determinar la relación de dependencia, entre una variable dependiente o respuesta Y, con una variable independiente, explicativa, regresora o covariable X, (Heumann et al., 2016), representado como:

$$Y = \beta 0 + \beta 1 X \quad (1)$$

donde $\beta 0$ es el intercepto y $\beta 1$ el parámetro de pendiente que indica el cambio del valor de Y, cuando el valor de X cambia en una unidad. Si $\beta 1$ es positivo, significa que el valor de Y incrementa, si X incrementa y si $\beta 1$ es negativo, significa que el valor de Y decrece, si X decrece.

Los supuestos de la regresión lineal (Hoffmann, 2010) son:

• Homocedasticidad: La variación del error es constante para todas las combinaciones de las variables explicativas, es decir, $Var(\varepsilon i|Xi) = \sigma 2$. La prueba de homocedasticidad Breusch-Pagan, puede ser utilizada para comprobar este supuesto.

• Independencia: Los valores de Y son independientes entre sí, es decir, las observaciones son independientes. La prueba de independencia Durbin-Watson, puede ser utilizada para comprobar este supuesto.

• Normalidad: Los errores se distribuyen normalmente con media cero y varianza constante. Las pruebas de normalidad Shapiro-Wilk y Kolmogorov-Smirnov, pueden ser utilizadas para comprobar este supuesto.

2.2.2 Regresión Ridge

La regresión de Ridge, también conocida como la regularización de *Tikhonov*, es una forma regularizada de la regresión lineal, que utiliza el término de penalización L2

La regularización L2, también conocida como "Ridge", es una técnica fundamental en el campo de la ciencia de datos y el aprendizaje automático. Su objetivo principal es controlar la complejidad de los modelos y prevenir el sobreajuste al introducir una penalización en la función de costo del modelo. Esta penalización se basa en la suma de los cuadrados de los coeficientes de las características utilizadas en el modelo de regresión o clasificación.

La característica distintiva de la regularización L2 es que no fuerza los coeficientes de las características a ser exactamente cero como lo hace la regularización L1 (Lasso). En cambio, reduce los coeficientes hacia valores cercanos a cero, pero generalmente no los elimina por completo. Esto tiene el efecto de suavizar los coeficientes y evitar que tomen valores extremadamente grandes, lo que puede ser útil para evitar problemas de multicolinealidad y mejorar la estabilidad del modelo.

La regularización L2 es especialmente eficaz cuando se trabaja con conjuntos de datos que tienen muchas características (alta dimensionalidad) y cuando se busca un equilibrio entre la simplicidad del modelo y su capacidad de generalización. Al agregar la penalización L2 a la función de costo, se logra una reducción en la varianza del modelo, lo que puede llevar a un mejor rendimiento en la predicción de datos nuevos y no vistos.

Como resultado, el algoritmo no solo ajustará los datos, sino que también mantendrá los pesos del modelo lo más bajos posible (Géron, 2017), aunque el término de regularización reduce la interpretabilidad.

2.2.3 Regresión Lasso

Least Absolute Shrinkage and Selection Operator Regression, también conocida como regresión Lasso, es una forma regularizada de la regresión lineal, que utiliza el término de penalización L1

La regularización L1, también conocida como "Lasso" (*Least Absolute Shrinkage and Selection Operator*), es una técnica de regularización ampliamente utilizada en el campo de la ciencia de datos y el aprendizaje automático. Su objetivo principal es controlar la complejidad de los modelos y prevenir el sobreajuste al introducir una penalización en la función de costo del modelo. Esta penalización se basa en la suma de los

valores absolutos de los coeficientes de las características utilizadas en el modelo de regresión o clasificación.

La regularización L1 tiene la propiedad única de forzar algunos de los coeficientes de características a cero, lo que efectivamente realiza la selección automática de características. En otras palabras, L1 elimina o reduce la influencia de las características menos importantes, lo que simplifica el modelo y mejora la interpretabilidad al tiempo que mantiene un buen rendimiento predictivo. Esto la convierte en una herramienta valiosa para lidiar con conjuntos de datos de alta dimensionalidad y para identificar las variables más relevantes en un problema de modelado.

En contraste con la regresión Ridge, el algoritmo tiende a eliminar completamente los pesos de las características menos significativas, es decir, mantendrá los pesos del modelo lo más pequeños posible mientras pierde interpretabilidad al darles un valor cerca de cero (Géron, 2017).

2.2.4 Árbol de decisión

Él árbol de decisión es un método supervisado de aprendizaje automático, usado para clasificación y regresión, cuyo objetivo es construir un modelo con buena capacidad de generalización, mediante reglas de decisión derivadas de las variables explicativas. Algunos de los algoritmos empleados son CART y ID3, junto con las métricas para medir la calidad de la división que permite seleccionar la mejor (Ertel, 2017).

La construcción del árbol de decisión para regresión se realiza mediante un proceso iterativo, para encontrar la mayor reducción posible de las métricas mencionadas, de la siguiente forma (Breiman et al., 1984):

a. El proceso iterativo inicia en la parte alta del árbol, donde todas las observaciones están dentro de la misma región.

b. Se identifica los puntos de posibles umbrales para cada una de las variables explicativas ($X1$, $X2$, $\cdots$, Xp). A diferencia de las variables categóricas, en las variables continuas, se utiliza discretización binaria, que convierte a los atributos estableciendo un umbral, es decir, se ordenan de menor a mayor los valores y el punto intermedio entre cada par, se escoge como los umbrales. En el caso de variables categóricas, los umbrales son los niveles de cada una.

c. Luego, se selecciona una métrica y se calcula el valor para cada posible división, generada en el paso anterior.

d. Se selecciona el umbral que obtuvo la mejor división, de acuerdo con la métrica, y si existen dos o más divisiones con el mismo valor, la elección será aleatoria.

e. Finalmente, se repite de forma iterativa, hasta alcanzar el criterio de parada, como número mínimo de observaciones por región.

Para predecir un nuevo dato, se debe recorrer el árbol entrenado, evaluado en cada nodo la condición establecida, hasta llegar a una de las hojas y calcular el promedio de los valores de la variable respuesta que pertenecen a la hoja.

Las características del algoritmo son (Tan et al., 2005):

- Es un método no paramétrico, es decir, no requiere suposiciones previas de las distribuciones con respecto a las clases y los otros atributos.

- Hallar un árbol óptimo es un problema, entonces debe utilizarse un enfoque heurístico.

- Son computacionalmente económicos y aplicarlo para calcular una clasificación es rápido.

- Pueden generar árboles muy complejos que no generalizan los datos y son inestables, es decir, una variación puede generar un árbol completamente diferente.

- A diferencia de otros algoritmos, es fácilmente interpretable.

2.2.5 Random Forest

El algoritmo de *Random Forest*, está compuesto de árboles de decisión, por lo que puede tener mayor estabilidad y precisión, cuyos métodos más comunes de aprendizaje en conjunto son: *bagging, boosting y stacking* (Sullivan, 2017).

El *bagging* es un método que busca la reducción de la varianza entre las predicciones, al sumar las salidas de dos o más clasificaciones o predicciones que se entrenan con diferentes muestras del conjunto de datos, aplicando los siguientes pasos (Sullivan, 2017):

a. Crear múltiples conjuntos de datos: Se forman conjuntos de datos con muestreo con reemplazo.

b. Construir múltiples modelos: Estos se entrenan con cada uno de los conjuntos del paso anterior.

c. Combinar los modelos: Cada modelo genera predicciones individuales y son combinadas con una técnica de agregación, eligiendo la clasificación más votada o el promedio de las clasificaciones.

Las características del algoritmo son (Sullivan, 2017):

- Se pude emplear para clasificación y regresión.

- Es posible que maneja una gran cantidad de datos en alta dimensionalidad y puede estimar de manera efectiva datos faltantes.

- No es tan bueno en la clasificación que, en la regresión, ya que no puede hacer predicciones más allá de rango de los datos de entrenamiento.

- Si los datos tienen demasiado ruido, tienden a ajustarse demasiado.

2.2.6 K-Nearest Neighbors

Con el algoritmo *K-Nearest Neighbors*, el modelo se construye simultáneamente con la prueba de los datos en lugar de después de que los datos de entrenamiento se usen para crear un modelo de aprendizaje. Este proceso es iterativo y calcula la distancia o similitud entre cada ejemplo del conjunto de

prueba y cada ejemplo del conjunto de entrenamiento. Luego elige los k vecinos más cercanos y calcula el valor como el promedio de las predicciones o el rango más popular (Tan et al., 2005)

Las características del algoritmo son (Tan et al., 2005):

• Es parte de una técnica más general, conocida como aprendizaje basado en instancias, puesto que utiliza observaciones específicas del conjunto de entrenamiento para hacer predicciones.

• Es un algoritmo perezoso, debido a que no requiere la construcción de un modelo y por ello, puede ser bastante costoso computacionalmente.

• Puede producir predicciones incorrectas, sino se toman el tipo de distancia y los pasos de preprocesamiento adecuados.

2.3 Elementos teóricos adicionales que considerar

A continuación, se enuncia el método validación cruzada y las métricas utilizadas en el contexto propio del aprendizaje automático.

2.3.1 Validación cruzada

Debe probar un modelo de aprendizaje automático con datos nuevos para evaluar su rendimiento. Es posible saber si un modelo aún necesita ajustarse, se ha sobreajustado o está "bien generalizado" en función de cómo se desempeña cuando se le dan datos desconocidos.

La validación cruzada es uno de los métodos más populares para evaluar el rendimiento de un modelo de aprendizaje automático. Este enfoque incorpora una técnica de remuestreo que permite la evaluación del modelo incluso con datos escasos.

Es necesario separar primero una parte de los datos de la serie de datos de entrenamiento antes de realizar un "CV" (validación cruzada). Posteriormente, el modelo se probará y validará utilizando esos datos en lugar de para el entrenamiento.

En el aprendizaje automático, la validación cruzada se usa con frecuencia para evaluar varios modelos y elegir el mejor para un problema determinado. Hay menos sesgos con este método que con los otros, y también es simple de entender, aplicar y usarse para entrenar el modelo, esos datos se usarán para probarlo y validarlo más adelante.

2.3.2 Métricas

Las métricas de rendimiento permiten observar el progreso de los modelos en las tareas de aprendizaje automático. A continuación, algunas métricas utilizadas.

a. *Root Mean Square Error* (RMSE):

RMSE (Root Mean Square Error) es una medida comúnmente utilizada para evaluar la precisión de un modelo de regresión o pronóstico. Representa la raíz cuadrada del promedio de los errores cuadrados entre los valores pronosticados por el modelo y los valores reales observados.

El cálculo del *RMSE* implica los siguientes pasos:

- Para cada punto de datos en el conjunto de prueba, se calcula la diferencia entre el valor pronosticado por el modelo y el valor real observado.

- Estas diferencias se elevan al cuadrado para eliminar los signos negativos y resaltar el error absoluto.

- Se calcula el promedio de los errores cuadrados obtenidos.

- Finalmente, se toma la raíz cuadrada del promedio de los errores cuadrados para obtener el *RMSE*.

El RMSE proporciona una medida de la dispersión de los errores pronosticados en relación con los valores reales. Cuanto menor sea el valor del RMSE, mayor será la precisión del modelo en la predicción de los valores observados. Por lo tanto, se utiliza como una métrica de evaluación para comparar diferentes modelos o ajustar los parámetros del modelo con el objetivo de minimizar los errores pronosticados.

Es importante tener en cuenta que el RMSE utiliza los errores al cuadrado, lo que puede hacer que los valores atípicos tengan un impacto más significativo en la métrica en comparación con otras medidas de error, como el *MAE (Mean Absolute Error).*

b. Mean Absolute Error (MAE):

MAE (Mean Absolute Error) es una medida utilizada para evaluar la precisión de un modelo de regresión o pronóstico. Representa el promedio de las diferencias absolutas entre los valores pronosticados por el modelo y los valores reales observados.

El cálculo del MAE implica los siguientes pasos:

- Para cada punto de datos en el conjunto de prueba, se calcula la diferencia absoluta entre el valor pronosticado por el modelo y el valor real observado.

- Estas diferencias absolutas se suman y se calcula el promedio.

El MAE proporciona una medida de la magnitud promedio de los errores pronosticados en relación con los valores reales. A diferencia del RMSE, el MAE no eleva los errores al cuadrado y, por lo tanto, no otorga un mayor peso a los errores más grandes. Esto hace que el MAE sea menos sensible a valores atípicos y más interpretable en términos absolutos.

Una ventaja del MAE es que refleja directamente la diferencia promedio entre las predicciones y los valores observados, lo que facilita su interpretación. Sin embargo, el MAE no considera la

dirección del error, es decir, si el modelo subestima o sobreestima los valores. Si se desea penalizar los errores más grandes, el RMSE puede ser más adecuado.

c. *Mean Absolute Percentage Error* (MAPE):

MAPE (*Mean Absolute Percentage Error*) es una medida utilizada para evaluar la precisión de un modelo de pronóstico o regresión en términos de porcentaje de error absoluto promedio.

El cálculo del MAPE implica los siguientes pasos:

- Para cada punto de datos en el conjunto de prueba, se calcula el error absoluto como la diferencia absoluta entre el valor pronosticado por el modelo y el valor real observado.
- Luego, se calcula el porcentaje de error absoluto dividiendo el error absoluto por el valor real observado y multiplicándolo por 100.
- Estos porcentajes de error absoluto se suman y se calcula el promedio.

El MAPE proporciona una medida de la precisión relativa del modelo en términos de porcentaje de error promedio en relación con los valores reales. Es útil para evaluar modelos cuando se desea tener una idea de la magnitud de los errores en relación con el tamaño de los valores observados.

Sin embargo, es importante tener en cuenta que el MAPE puede ser sensible a divisiones por cero si existen valores reales cercanos a cero en el conjunto de datos. Además, el MAPE puede presentar problemas cuando se tienen valores reales muy pequeños o cercanos a cero, ya que los errores relativos pueden ser amplificados. En tales casos, se recomienda considerar otras métricas, como el MAE o el RMSE.

2.3.3 Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica estadística ampliamente utilizada en diversas áreas de investigación, reconocida por su capacidad para reducir la dimensionalidad de conjuntos de datos multivariados y encontrar estructuras subyacentes en ellos (James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013). Según Jolliffe (2002), el PCA es una técnica útil para analizar y reducir la dimensionalidad de conjuntos de datos complejos, permitiendo una mejor comprensión y visualización de las relaciones entre variables. Además, Izenman (2013) destaca que el PCA es una técnica ampliamente utilizada en el análisis de datos multivariados, permitiendo reducir la dimensionalidad de los datos y descubrir la estructura subyacente, facilitando la interpretación y visualización de los resultados.

2.3.4 *Análisis Factorial de Datos Mixtos* (FAMD)

El FAMD (Análisis Factorial de Datos Mixtos) es una técnica de análisis de datos que combina el Análisis Factorial con datos mixtos, es decir, conjuntos de datos que contienen variables de diferentes tipos, como variables categóricas y variables numéricas. Según Tenenhaus et al. (2005), el FAMD permite trabajar con datos mixtos y analizar la estructura subyacente de un conjunto de datos que incluye diferentes tipos de

variables. El FAMD se basa en la descomposición de la matriz de covarianza de los datos mixtos y utiliza técnicas de extracción de factores para identificar las dimensiones latentes o factores que explican la estructura de las variables (Le Roux & Rouanet, 2010). Esta técnica es útil para reducir la dimensionalidad de los datos mixtos, interpretar las relaciones entre las variables y realizar análisis adicionales como clustering o análisis discriminante (Tenenhaus et al., 2005).

3 Metodología

En este apartado se describe la forma en cómo se desarrolló el proyecto de investigación. El proyecto en general se consolida en los siguientes pasos:

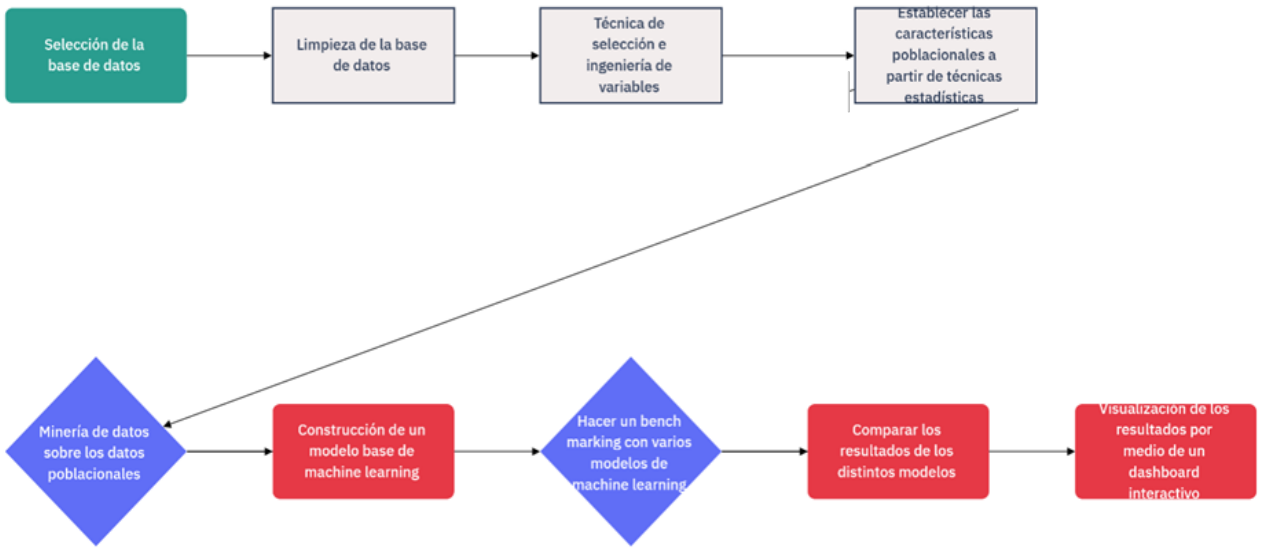


Figura 3-1 Metodología del proyecto

La anterior metodología es una adaptación de la metodología KDD.

Selección de la base de datos.

Para cada una de las ciudades involucradas en el proyecto, se procedió a utilizar las fuentes de datos abiertos de cada un formato e información distintas, para ellos se empezó buscando cuales eran los datos que tenían en común y las variables más relevantes. A continuación, se presenta el diccionario de datos del maestro de datos obtenido a partir de las bases de Bogotá, Medellín y Barranquilla.

Nombre de la variable	Tipo dato	Descripcion de la variable
CIUDAD	CATEGORICA	Ciudad en la que ocurrio el accidente
CODIGO_ACCIDENTE	CATEGORICA	Codigo del siniestro
CODIGO_ACC	CATEGORICA	Codigo unico del siniestro con la ciudad
DIRECCION	CATEGORICA	Direccion donde ocurrio el siniestro
GRAVEDAD	CATEGORICA	Gravedad del accidente
CLASE_ACC	CATEGORICA	Clase de accidente
BARRIO	CATEGORICA	Barrio o upz en la que ocurrio el siniestro
ZONA	CATEGORICA	Zona (Localidad o Comuna)
ANIO	NUMERICA	Año del siniestro
MES	CATEGORICA	Mes del siniestro
DIA	NUMERICA	Dia del siniestro
HORA	CATEGORICA	Hora del siniestro
DIA_SEMANA	CATEGORICA	Dia de la semana del siniestro
LATITUD	CATEGORICA	Latitud de la ubicación del siniestro
LONGITUD	CATEGORICA	Longitud de la ubicación del siniestro
DISENO	CATEGORICA	Diseño de la via
CANTIDAD_HERIDOS	NUMERICA	Cantidad de heridos en el accidente
CANTIDAD_ILESO	NUMERICA	Cantidad de ilesos en el accidente
CANTIDAD_MUERTOS	NUMERICA	Cantidad de muertos del siniestro
CANTIDAD_VICTIMAS	NUMERICA	Cantidad de victimas totales del siniestro
CANTIDAD_VEHICULOS	NUMERICA	Cantidad de vehiculos involucrados en el siniestro
FLAG_ILESOS	BINARIA	Variable de si son solo ilesos o no
EDAD PROMEDIO	NUMERICA	Edad promedio de los involucrados en el accidente
EDAD MEDIANA	NUMERICA	Edad mediana de los involucrados en el accidente
CANTIDAD HOMBRES	NUMERICA	Cantidad de hombres victimas del siniestro
CANTIDAD MUJERES	NUMERICA	Cantidad de mujeres victimas del siniestro

Tabla 3-2 Variables de la base de datos

Bogotá:

OBJECTID	FORMULA	CODIGO_A	FECHA_OC	ANO_OCU	DIRECCION	GRAVEDAD	CLASE_AC	LOCALIDA	FECHA_HO	LATITUD	LONGITUD	CIV	PK_CALZA	HORA	DESCR_CHOQUE
1	A000640275	4484660	2017/06/12 0	2017	AV AVENIDA E	SOLO DANOS	CHOQUE	ENGATIVA	2017/06/12 0	469,380,685	##########	10006772	221236	05:30:00	Vehiculo
2	A001233353	10533499	2020/11/19 0	2020	CL 26 S- KR 5C	CON HERIDO	OTRO	PUENTE ARA	2020/11/19 0	4,603	-74,121	16004560		02:05:00	#N/D
4	A001232786	10533629	2020/11/10 0	2020	KR 9 - CL 100	SOLO DANOS	CHOQUE	USAQUEN	2020/11/10 1	4,682	-74,042	30001107		13:30:00	Vehiculo
7	A000200705	4412699	2015/05/11 0	2015	CL 63A-KR 72	SOLO DANOS	CHOQUE	CIUDAD BOLI	2015/05/11 1	458,718,669	##########	19001483	136166	10:50:00	Vehiculo
8	A000402862	4447845	2016/06/08 0	2016	KR 27-CL 9 14	SOLO DANOS	CHOQUE	LOS MARTIRE	2016/06/08 2	460,764,758	##########	14000548	239719	21:30:00	Vehiculo
9	A001179874	10533587	2020/08/03 0	2020	AU NORTE - C	CON MUERTC	ATROPELLO	SUBA	2020/08/03 1	4,778	-74,042	1006455		14:05:00	#N/D
10	A000240105	4424883	2015/09/26 0	2015	KR 52A-CL 13	SOLO DANOS	CHOQUE	SUBA	2015/09/26 1	472,462,598	##########	11008301	31431	18:00:00	Objeto Fijo
12	A001233064	10533503	2020/11/23 0	2020	AU NORTE - C	SOLO DANOS	OTRO	USAQUEN	2020/11/23 1	4,796	-74,039			11:50:00	#N/D
13	A000551010	4468708	2016/12/27 0	2016	CL 69A-KR 89	CON HERIDO	CHOQUE	ENGATIVA	2016/12/27 1	469,357,846	##########	10006813	219627	19:00:00	Vehiculo
15	A000686495	4495519	2017/10/02 0	2017	AV AVENIDA (	CON HERIDO	CHOQUE	FONTIBON	2017/10/02 0	465,932,955	##########	50008531	272348	09:20:00	Vehiculo
17	A22245	4467629	2016/12/16 0	2016	KR 21-CL 127	SOLO DANOS	CHOQUE	USAQUEN	2016/12/16 1	470,653,939	##########	1004358	202250	15:00:00	Vehiculo
18	A001180302	10533662	2020/09/14 0	2020	KR 12 - CL 81	CON MUERTC	ATROPELLO	USME	2020/09/14 2	4,508	-74,113	5003431		21:30:00	#N/D
20	A000641417	4493355	2017/09/12 0	2017	KR 50-CL 26 0	SOLO DANOS	CHOQUE	TEUSAQUILLO	2017/09/12 0	464,165,433	##########	13001016	194626	07:53:00	Vehiculo
23	A001180467	10533606	2020/08/22 0	2020	KR 30 - CL 64	CON MUERTC	ATROPELLO	BARRIOS UNI	2020/08/22 1	4.66	-74,077	12002271		19:45:00	#N/D
25	A000815605	10448175	2018/05/19 0	2018	KR 18-CL 2 02	CON MUERTC	VOLCAMIENT	CIUDAD BOLI	2018/05/19 2	452,632,143	-741,408,926	50000366		20:00:00	#N/D
28	A000473024	4459693	2016/10/04 0	2016	KR 77-CL 19 0	SOLO DANOS	CHOQUE	FONTIBON	2016/10/04 1	465,483,949	##########	0		12:15:00	Vehiculo
29	A001236422	10530927	2020/12/24 0	2020	TV 13 BIS A - (	CON HERIDO	ATROPELLO	RAFAEL URIBI	2020/12/24 0	4,574	-74,113	18002038		08:55:00	#N/D
30	A001130325	10536544	2020/01/13 0	2020	AV AVENIDA (	CON HERIDO	ATROPELLO	RAFAEL URIBI	2020/01/13 0	4,579	-74,107	18001196		09:40:00	#N/D
31	A000403683	4448989	2016/06/20 0	2016	KR 11-CL 72 0	SOLO DANOS	CHOQUE	CHAPINERO	2016/06/20 1	465,714,915	##########	2001059	38023	11:50:00	Vehiculo
35	A000816158	10448832	2018/05/23 0	2018	CL 23-KR 104	SOLO DANOS	CHOQUE	FONTIBON	2018/05/23 0	468,186,145	##########	9001330	214724	06:50:00	Vehiculo
36	A001235699	10533544	2020/12/22 0	2020	TV 78 H - CL 4	CON MUERTC	CHOQUE	KENNEDY	2020/12/22 1	4,616	-74,158	8012380		15:30:00	Vehiculo
38	A001175820	10533489	2020/05/28 0	2020	KR 78 - CL 57	CON MUERTC	ATROPELLO	KENNEDY	2020/05/28 1	4,604	-74,173	30000651		12:45:00	#N/D
39	A001233294	10533579	2020/12/10 0	2020	AK 80 - CL 42	SOLO DANOS	CHOQUE	KENNEDY	2020/12/10 1	4,623	-74,163	8006166		11:45:00	Vehiculo
43	A001183932	10533860	2020/10/18 0	2020	CL 64 S- KR 21	SOLO DANOS	CHOQUE	CIUDAD BOLI	2020/10/18 2	4,566	-74,146	19005834		22:39:00	Vehiculo
45	A001185526	10533669	2020/11/10 0	2020	CL 26 - KR 11	SOLO DANOS	CHOQUE	FONTIBON	2020/11/10 1	4,695	-74.14	9004111		19:00:00	Vehiculo

Tabla 3-3 Variables reales base de Bogotá

Medellín:

NRO_R	ICA	GRAVE	CLASE_	DIRECC	SEXO	EDAD	CONDIC	MES	DIA_SE	DIA	HORA	ANO	LATITUD	LONGITUD	COMUNA
1470492	1	Heridos	Atropello	CL 106 B C	M	12	Peaton	Ene	Jue	1	00:00-00:59	2015	6.294617511	-75.54160996	01 - Popular
1470491	1	Heridos	Atropello	CL 48 A CR	F	62	Peaton	Ene	Jue	1	00:00-00:59	2015	6.258484108	-75.61037162	13 - San Javier
1470349	1	Heridos	Choque	CL 113 CR	F	30	Acompana	Ene	Jue	1	00:00-00:59	2015	6.309526171	-75.57044549	05 - Castilla
1470731	1	Heridos	Atropello	CR 50 C CL	M	66	Peaton	Ene	Jue	1	00:00-00:59	2015	6.272356955	-75.56150168	04 - Aranjuez
1470356	1	Heridos	Caida Ocu	CR 64 C CL	F	22	Acompana	Ene	Jue	1	01:00-01:59	2015	6.276299355	-75.5733641	05 - Castilla
1470445	1	Heridos	Choque	CR 48 CL 9	M	22	Acompana	Ene	Jue	1	01:00-01:59	2015	6.284498952	-75.5554293	04 - Aranjuez
1470742	1	Heridos	Atropello	CR 71 A CL	M	5	Peaton	Ene	Jue	1	01:00-01:59	2015	6.223813876	-75.59360484	16 - Belen
1470358	1	Heridos	Atropello	CL 17 CR 8	M	20	Peaton	Ene	Jue	1	02:00-02:59	2015	6.222828342	-75.60515535	16 - Belen
1470357	1	Heridos	Otro	CR 36 CL 9	M	45	Motociclis	Ene	Jue	1	02:00-02:59	2015	6.282663371	-75.54543119	03 - Manrique
1470487	1	Heridos	Atropello	CR 20 CL 6	M	19	Peaton	Ene	Jue	1	02:00-02:59	2015	Sin Inf	Sin Inf	Sin Inf

Tabla 3-3 Variables reales base de Medellín

Seguido a esto se empieza a buscar como unir la información para dejar un formato estándar para todas las ciudades, como primer paso se estandariza el formato de las fechas y se ponen nombres estándar para las localidades y las comunas a los cuales nombraremos zonas en la base de datos. Adicionalmente se estandariza para todas las ciudades los campos de día de la semana, día, mes, año y hora del accidente.

La variables de víctimas e información etaria y sexo se encuentran en bases separadas sin agregación alguna para las ciudades de Bogotá y Medellín, así que se procede a hacer agregaciones a la bases de datos para poder llevarla a la base final y así enriquecer la base de datos quedando como variables de información sobre las personas involucradas en el accidente así se construye las columnas cantidad de heridos, cantidad de ilesos, cantidad de muertos, cantidad de mujeres y hombres, edad mediana y edad promedio de las víctimas.

Se construye también una variable única de código de accidente para poder diferenciar los códigos que se parecen entre las ciudades y para Barranquilla se le asigna un código mediante una numeración artificial debido a que no cuenta con un código para cada uno de los siniestros

Las bases cuentan con bastantes caracteres que pueden generar problemas a la hora de subir la base al programa por lo tanto se decide quitar tildes y la letra ñ que no siempre se decodifican de manera correcta en Python. Adicionalmente se cambian las variables categóricas a mayúsculas debido a que algunas no se encuentran escritas de la misma forma, esto evita que más tarde se tomen como términos diferentes debido a su escritura.

En las variables de cantidad se llenan de ceros para los valores faltantes y se crea la variable FLAG_ILESOS que indica si hay solo ilesos en el caso que quede SI y que hay heridos o muertos en el caso de decir NO.

Las bases originalmente cuentan con datos del 2020 pero es un año que se obviara debido a la pandemia global Covid -19 que modifico los patrones de comportamiento de la población en general.

La base de Bogotá no cuenta con una variable de barrio o sector por lo tanto se procede a hallar en que unidad de planeación zonal o UPZ se encuentra el siniestro, debido a que las localidades son muy extensas y en Bogotá se toman las decisiones de la alcaldía por UPZ que es la unidad intermedia entre las manzanas y las Localidades.

Limpieza de la base de datos.

En el proceso de análisis de las bases de datos descargadas, se llevó a cabo una minuciosa revisión de los valores atípicos y los valores faltantes presentes en cada una de ellas. Este paso resulta fundamental para garantizar la integridad y la calidad de los datos utilizados en el análisis posterior. Durante este proceso de detección y evaluación de valores atípicos, se identificaron aquellos puntos de datos que se desviaban significativamente de la distribución esperada o que parecían ser errores o anomalías. Estos valores atípicos se examinaron cuidadosamente para determinar si eran legítimos o si requerían ser corregidos o eliminados del conjunto de datos. Por otro lado, también se detectaron valores faltantes, es decir, aquellos campos o variables que no tenían información registrada. Estos valores faltantes pueden ser problemáticos, ya que pueden afectar la capacidad de realizar análisis y obtener conclusiones significativas a partir de los datos. Por lo tanto, se realizó una evaluación exhaustiva para determinar qué datos podían ser llenados o imputados utilizando técnicas estadísticas o métodos de inferencia, y qué datos no eran recuperables y debían ser excluidos del análisis. Este proceso de limpieza y tratamiento de datos faltantes y valores atípicos es esencial para garantizar la confiabilidad y la validez de los resultados obtenidos a través del análisis de datos, y para asegurar que las conclusiones y decisiones tomadas se basen en información precisa y completa.

Técnica de selección e ingeniería de variables.

Se llevaron a cabo diversas transformaciones en los datos con el fin de facilitar y mejorar el análisis. Una de las transformaciones comunes fue la normalización de los datos. La normalización es un paso crucial que permite ajustar las variables a una escala comparable, eliminando las diferencias en las unidades de medida y los rangos de valores. Esto es especialmente importante cuando se trabaja con variables que tienen diferentes magnitudes, ya que la falta de normalización podría distorsionar los resultados del análisis de0 *clustering*. Al aplicar la normalización, se aseguró que todas las variables tuvieran una distribución similar y que ninguna de ellas tuviera un peso desproporcionado en el proceso de agrupación.

Además de la normalización, también se generaron variables adicionales que se consideraron relevantes para el análisis. Estas variables se crearon a partir de combinaciones o transformaciones de las variables existentes, con el objetivo de capturar características o patrones más específicos. Por ejemplo, podrían haberse creado variables de suma, diferencia o promedio a partir de variables relacionadas para obtener una medida agregada que refleje mejor la tendencia general. Asimismo, se podrían haber generado variables categóricas o indicadores binarios a partir de ciertos rangos de valores o umbrales específicos. Estas variables adicionales ayudaron a enriquecer el conjunto de datos y a capturar información adicional que podría haber sido relevante para la identificación de los grupos o *clusters*.

En resumen, la transformación de los datos para la clusterización implicó tanto la normalización de las variables como la generación de variables adicionales. Estas transformaciones permitieron obtener una representación más adecuada de los datos y mejorar la capacidad de discriminación entre los diferentes grupos o *clusters* identificados. Al aplicar estas técnicas, se buscó maximizar la eficacia y la calidad del análisis de *clustering*, proporcionando una visión más completa y precisa de la estructura subyacente de los datos y permitiendo una interpretación más sólida de los resultados obtenidos.

Establecer las características poblacionales a partir de técnicas estadísticas

Dado el enfoque de clusterización realizado a nivel de localidad en Bogotá y comuna en Medellín, se pudo obtener una visión más detallada de las características de la población en cada clúster identificado. Mediante el análisis de los grupos formados, fue posible identificar patrones y tendencias en términos de características demográficas, socioeconómicas y culturales de las personas que residen en cada área.

Minería de datos sobre los datos poblacionales.

En el análisis de los conjuntos de datos, se llevó a cabo un examen exhaustivo de la cantidad de hombres y mujeres involucrados en los accidentes registrados. Esta desagregación por género permitió obtener una visión clara de la distribución de los accidentes en función de la composición de género de los afectados. A través de esta información, fue posible identificar posibles disparidades o patrones específicos relacionados con la participación de hombres y mujeres en accidentes de tránsito.

Además de la desagregación por género, se realizó un análisis detallado de la media y la mediana de la edad de los afectados en los accidentes. Estas medidas estadísticas proporcionaron información clave sobre la distribución de edades de los involucrados y permitieron identificar tendencias o características específicas en relación con la edad de los afectados. Por ejemplo, al comparar la media y la mediana de edad, se podría determinar si existen diferencias significativas en cuanto a la distribución de edades, como una mayor concentración de accidentes en ciertos grupos de edad o si hay una dispersión amplia en la edad de los afectados.

Este análisis de la edad de los afectados en los accidentes también puede revelar información valiosa sobre la relación entre la edad y la incidencia de accidentes. Por ejemplo, se podría identificar si hay una mayor propensión a sufrir accidentes en ciertos rangos de edad, como los conductores jóvenes o los conductores de mayor edad. Además, este análisis permitiría evaluar si hay diferencias significativas en términos de edad entre hombres y mujeres involucrados en accidentes, lo que podría proporcionar información adicional sobre los factores subyacentes que contribuyen a la ocurrencia de los accidentes.

Construcción de un modelo base de aprendizaje automático.

Con el objetivo de realizar una predicción precisa de la cantidad de accidentes, se recurrió a la construcción de un modelo base de aprendizaje automático utilizando regresión lineal. La regresión lineal es un método ampliamente utilizado en el análisis predictivo que permite establecer una relación lineal entre una variable dependiente (en este caso, la cantidad de accidentes) y una o varias variables independientes.

Para construir el modelo de regresión lineal, se utilizaron los datos históricos disponibles, donde se registraba la cantidad de accidentes y las variables independientes correspondientes en un período determinado. Estos datos se dividieron en un conjunto de entrenamiento y un conjunto de prueba, con el fin de evaluar la capacidad predictiva del modelo. Durante el proceso de entrenamiento, el modelo ajustó los parámetros de la línea de regresión para minimizar la diferencia entre los valores reales y los valores predichos.

Una vez que el modelo de regresión lineal fue entrenado y validado, se utilizó para predecir la cantidad de accidentes en un período futuro. Se introdujeron los valores de las variables independientes correspondientes al período de predicción en el modelo y se obtuvieron las predicciones de la cantidad de accidentes esperada. Estas predicciones proporcionaron una estimación cuantitativa basada en las relaciones lineales identificadas en los datos de entrenamiento.

Es importante tener en cuenta que la construcción de un modelo base de regresión lineal es el primer paso en el análisis predictivo y que existen otros modelos más complejos y sofisticados que podrían mejorar la precisión de las predicciones. Sin embargo, la regresión lineal proporciona una base sólida y comprensible para comenzar a explorar las relaciones entre las variables y realizar predicciones iniciales. A medida que se disponga de más datos y se profundice en el análisis, se podrían considerar otros modelos más avanzados que tengan en cuenta la no linealidad de las relaciones o la interacción entre las variables, con el objetivo de mejorar aún más la capacidad de predicción de la cantidad de accidentes.

Hacer un benchmarking con varios modelos de aprendizaje automático.

Con el objetivo de evaluar y seleccionar el mejor modelo para este proyecto de predicción de accidentes, se construyeron varios modelos alternativos. La elección del modelo más adecuado se basó en la comparación de diversas métricas de evaluación, entre las cuales se consideraron el *Error Porcentual Absoluto Medio (MAPE), el Error Absoluto Medio (MAE) y Root Mean Square Error (RMSE).*

El MAPE es una medida que proporciona una estimación del error porcentual promedio entre las predicciones del modelo y los valores reales. Cuanto menor sea el MAPE, mayor será la precisión del modelo en términos de porcentaje de error.

El MAE es una métrica que calcula el promedio de las diferencias absolutas entre las predicciones y los valores reales. Esta métrica permite evaluar la magnitud promedio de los errores del modelo, sin tener en cuenta su dirección.

Por su parte, el RMSE es una medida que calcula la raíz cuadrada del promedio de los errores al cuadrado. Esta métrica penaliza más los errores grandes y proporciona una estimación de la dispersión de los errores.

Cada modelo alternativo fue evaluado utilizando estas métricas, y se compararon los resultados obtenidos. Se buscaba seleccionar aquel modelo que presentara los valores más bajos para el MAPE, el MAE y el RMSE, lo que indicaría una mejor capacidad de predicción y menor error en las estimaciones. Además de estas métricas, también se pueden considerar otras medidas de evaluación como el Coeficiente de Determinación (R^2), que indica la proporción de variabilidad explicada por el modelo, o realizar pruebas de validación cruzada para evaluar la capacidad de generalización del modelo a nuevos datos. Estas métricas proporcionaron una evaluación cuantitativa de los modelos y sirvieron como criterio objetivo para la selección del modelo más adecuado para este proyecto en particular.

Comparar los resultados de los distintos modelos.

Con base en el punto anterior, una etapa crucial en el análisis de los resultados de los modelos de predicción fue la evaluación de su consistencia y coherencia con la realidad del mundo. Aunque las métricas y los indicadores pueden proporcionar una medida cuantitativa de la calidad del modelo, es

igualmente importante analizar si los resultados obtenidos tienen sentido desde una perspectiva cualitativa y contextual.

Para lograr esto, se llevó a cabo un análisis detallado de las predicciones realizadas por los modelos y se compararon con el conocimiento experto o la intuición de los profesionales involucrados en el proyecto. Se examinaron aspectos como las tendencias generales identificadas, la dirección de los cambios esperados y la relación con variables conocidas o factores clave en el dominio de los accidentes de tránsito.

Por ejemplo, si el modelo predice un aumento significativo en la cantidad de accidentes en una zona específica debido a factores como el deterioro de las carreteras o el aumento del flujo vehicular, esto sería coherente con las expectativas y el conocimiento previo sobre cómo estos factores influyen en la ocurrencia de accidentes. Por otro lado, si las predicciones no se alinean con las condiciones o patrones conocidos en el mundo real, podría ser una señal de que el modelo necesita ajustes o que se deben considerar otras variables relevantes que no se tuvieron en cuenta inicialmente.

En última instancia, este análisis de consistencia se realizó para asegurarse de que los resultados obtenidos no fueran simplemente el resultado de una coincidencia estadística o de un ajuste inapropiado del modelo a los datos. Se buscó confirmar que los resultados tuvieran sentido desde el punto de vista lógico y práctico, y que pudieran proporcionar información valiosa y confiable para la toma de decisiones y la implementación de medidas preventivas en el ámbito de la seguridad vial.

Visualización de los resultados por medio de un dashboard interactivo.

Para facilitar la comprensión y el acceso a los resultados y análisis obtenidos, se desarrolló un dashboard interactivo que proporciona una visualización más amigable al usuario. Este dashboard fue diseñado con el objetivo de presentar de manera clara y concisa la información relevante derivada de los modelos de predicción y los análisis realizados.

El diseño del dashboard se basó en la identificación de las necesidades y requisitos de los usuarios finales, como gerentes, analistas o tomadores de decisiones en el ámbito de la seguridad vial. Se consideraron aspectos como la usabilidad, la capacidad de personalización y la presentación visual efectiva de los datos y resultados.

El dashboard incluye diferentes componentes y elementos interactivos, como gráficos, tablas y mapas, que permiten explorar los resultados desde diferentes perspectivas y niveles de detalle. Los gráficos ofrecen una representación visual clara de las tendencias, patrones y relaciones identificadas en los datos, lo que facilita la comprensión y el análisis de la información.

Además, el dashboard ofrece la posibilidad de filtrar y segmentar los datos según diferentes variables, lo que permite a los usuarios explorar y comparar los resultados de manera personalizada. Por ejemplo, se pueden filtrar los resultados por zona geográfica, tipo de accidente, período de tiempo o cualquier otra variable relevante para el análisis.

El aspecto interactivo del dashboard permite a los usuarios realizar exploraciones y consultas ad hoc, lo que implica una mayor autonomía y flexibilidad en la exploración de los resultados. Además, se incluyeron funciones de exportación y descarga de datos, lo que permite a los usuarios extraer la información necesaria para análisis posteriores o informes adicionales.

4 Resultados

En el marco de esta investigación sobre seguridad vial, se emprendió un proceso de recopilación y análisis de datos con el objetivo de profundizar en la comprensión de los accidentes de tráfico en distintas ciudades. La base de esta investigación radica en la recopilación de datos procedentes de fuentes de datos abiertos de múltiples localidades, los cuales fueron posteriormente unificados en una única base de datos. Un elemento fundamental que guio este proceso fue la selección cuidadosa de variables específicas que serían objeto de análisis exhaustivo.

Requisito Mínimo de Variables

El requisito mínimo de variables es un componente fundamental en nuestro enfoque de modelado, ya que establece los pilares sobre los cuales construimos nuestras predicciones y análisis. Este requisito no solo se basa en la inclusión de un número determinado de variables, sino que también implica una cuidadosa selección de aquellas que son más relevantes y significativas para el problema que estamos abordando.

Al considerar qué variables incluir en nuestro modelo, nos hemos asegurado de abarcar una amplia gama de factores que pueden influir en el fenómeno que estamos estudiando. Desde variables demográficas hasta indicadores económicos y ambientales, cada una de estas dimensiones ofrece una perspectiva única que enriquece nuestra comprensión del tema en cuestión.

La Tabla 4-1 proporciona una descripción detallada de las variables mínimas que deben estar presentes en cada ciudad que consideramos para nuestro análisis. Estas variables han sido cuidadosamente seleccionadas y evaluadas para garantizar que representen aspectos clave que impactan en el fenómeno en estudio, cada variable aporta su propia contribución al panorama general.

Además de la selección de variables, también hemos establecido un requisito mínimo de datos para cada ciudad. Entendemos que la cantidad y la calidad de los datos son fundamentales para el éxito de nuestro modelo. Por lo tanto, hemos fijado un umbral de 1000 registros por ciudad y año como el mínimo necesario para realizar un entrenamiento adecuado del modelo lo cual limitaría la inclusión de ciudades muy pequeñas puesto que no se lograría una predicción acertada. Esta cantidad de datos nos permite capturar una amplia variedad de escenarios y tendencias, lo que enriquece la capacidad predictiva de nuestro modelo.

Al cumplir con estos requisitos mínimos de variables y datos, no solo estamos garantizando la integridad de nuestro análisis, sino que también estamos sentando las bases para resultados más precisos y confiables. Nuestro enfoque riguroso nos permite confiar en que nuestras conclusiones están respaldadas por una base sólida de evidencia empírica, lo que a su vez fortalece la credibilidad y la utilidad de nuestro trabajo.

Geolocalización	**Dirección**
	Barrio
	Zona
Georeferencia	Latitud
	Longitud
Tiempo	Dia
	Mes
	Año
Tiempo_2	Hora
	Dia de la semana
Tipología	Gravedad
	Clase de accidente
	Diseño
Consecuencias	Cantidad heridos
	Cantidad ileso
	Cantidad muertos
	Cantidad victimas
	Cantidad vehiculos
Tipología demográfica	Cantidad hombres
	Cantidad mujeres
	Edad promedio
	Edad mediana

Tabla 4-1 Variables usadas para los modelos

En adición, se partió de la recopilación de datos de fuentes de datos abiertos de varias ciudades, consolidándolos en una única base de datos. La selección de variables específicas desempeñó un papel crucial en este proceso. A continuación, se presentarán los resultados de un análisis exhaustivo de seguridad vial en tres ciudades: Bogotá, Medellín y Barranquilla, con un enfoque en el periodo comprendido entre 2018 y febrero de 2020, año por año.

4.1 Análisis exploratorio

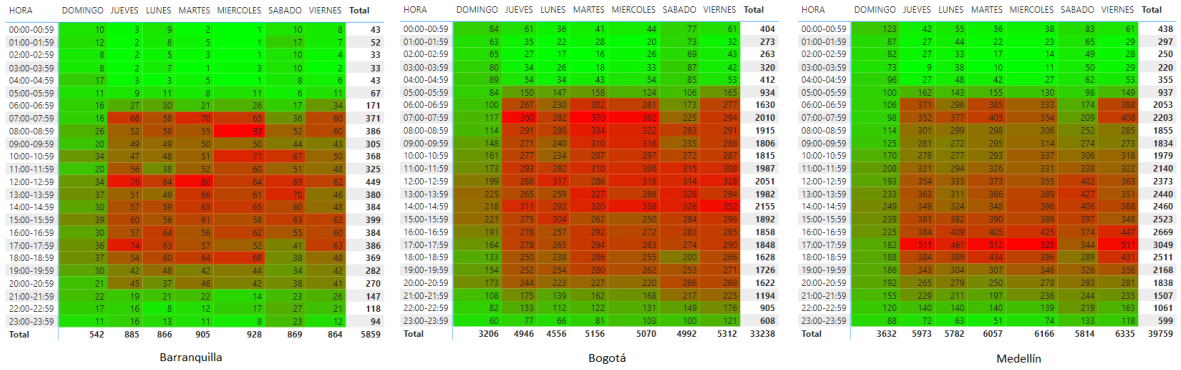
Análisis de Accidentes por Día y Hora:

Uno de los aspectos más destacados de esta investigación es el análisis de la temporalidad de los accidentes de tráfico. Durante el periodo evaluado, comprendido entre 2018 y febrero de 2020, año por año, se encontraron patrones intrigantes. Los domingos emergen como el día con la menor cantidad de accidentes en general. Asimismo, se observa que las horas de la madrugada registran el menor número de incidentes viales. Esta información tiene un valor sustancial, ya que puede influir en la formulación de políticas y estrategias de seguridad vial. Los gráficos generados a través de Power BI proporcionaron una representación visual clara y efectiva de estos patrones temporales, lo que facilita su comprensión y comunicación.

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	217	106	100	79	83	170	130	885
01:00-01:59	162	64	74	55	44	155	68	622
02:00-02:59	155	56	55	36	41	128	75	546
03:00-03:59	161	45	71	29	47	147	73	573
04:00-04:59	202	84	85	90	82	155	112	810
05:00-05:59	195	321	301	321	265	210	325	1938
06:00-06:59	222	665	556	708	640	364	699	3854
07:00-07:59	231	778	717	845	781	470	762	4584
08:00-08:59	254	644	637	687	711	587	636	4156
09:00-09:59	293	601	561	655	680	553	602	3945
10:00-10:59	365	602	559	631	705	645	655	4162
11:00-11:59	401	670	614	688	697	704	678	4452
12:00-12:59	426	718	714	739	738	785	753	4873
13:00-13:59	495	679	619	759	736	823	691	4802
14:00-14:59	497	717	675	733	797	792	788	4999
15:00-15:59	499	716	742	713	694	744	706	4814
16:00-16:59	446	719	730	753	759	712	792	4911
17:00-17:59	382	863	789	863	863	659	864	5283
18:00-18:59	358	688	687	784	719	527	745	4508
19:00-19:59	370	637	606	629	652	613	669	4176
20:00-20:59	386	554	539	523	540	597	591	3730
21:00-21:59	285	423	371	381	418	484	486	2848
22:00-22:59	219	289	260	274	287	395	360	2084
23:00-23:59	159	165	142	143	185	256	251	1301
Total	**7380**	**11804**	**11204**	**12118**	**12164**	**11675**	**12511**	**78856**

Figura 4-1 Concentración accidentes en los días de la semana por horas 2018

La figura 4-1 presenta un análisis detallado de la concentración de accidentes a lo largo de los días de la semana y las horas del día durante el año 2018. Resalta los momentos del día y de la semana en los que ocurren la mayoría de los accidentes, lo que puede ser fundamental para la planificación de medidas de seguridad y vigilancia vial. En el rango horario de las 5pm a las 6 pm se puede ver la mayor concentración de accidentes de lunes a viernes lo cual se podría prevenir con un cambio de los horarios de salida de las diversas empresas. Se puede observar como el Domingo ocurren menos accidentes que el resto de los días.



HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	10	3	9	2	1	10	8	43
01:00-01:59	12	2	8	5	1	17	7	52
02:00-02:59	8	2	5	3	1	10	4	33
03:00-03:59	8	2	7	1	3	10	2	33
04:00-04:59	17	3	3	5	1	8	6	43
05:00-05:59	11	9	11	8	11	6	11	67
06:00-06:59	16	27	30	21	26	17	34	171
07:00-07:59	16	66	58	70	65	36	60	371
08:00-08:59	26	52	58	55	83	52	60	386
09:00-09:59	20	49	49	50	50	44	43	305
10:00-10:59	34	47	48	51	71	67	50	368
11:00-11:59	20	56	38	52	60	51	48	325
12:00-12:59	34	76	64	80	64	69	62	449
13:00-13:59	37	51	49	66	61	70	46	380
14:00-14:59	30	57	59	65	65	60	48	384
15:00-15:59	39	60	56	61	58	63	62	399
16:00-16:59	30	57	64	56	62	55	60	384
17:00-17:59	36	74	63	57	52	41	63	386
18:00-18:59	37	54	60	64	68	38	48	369
19:00-19:59	30	42	48	42	44	34	42	282
20:00-20:59	21	45	37	46	42	38	41	270
21:00-21:59	22	19	21	22	14	23	26	147
22:00-22:59	17	16	8	12	17	27	21	118
23:00-23:59	11	16	13	11	8	23	12	94
Total	542	885	866	905	928	869	864	5859

Barranquilla

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	84	61	36	41	44	77	61	404
01:00-01:59	63	35	22	28	20	73	32	273
02:00-02:59	65	27	17	16	26	69	43	263
03:00-03:59	80	34	26	18	33	87	42	320
04:00-04:59	89	54	34	43	54	85	53	412
05:00-05:59	84	150	147	158	124	106	165	934
06:00-06:59	100	267	230	302	281	173	277	1630
07:00-07:59	117	360	282	370	362	225	294	2010
08:00-08:59	114	291	280	334	322	283	291	1915
09:00-09:59	148	271	240	310	316	235	286	1806
10:00-10:59	161	277	234	287	297	272	287	1815
11:00-11:59	173	293	282	310	306	315	308	1987
12:00-12:59	199	288	317	286	319	314	328	2051
13:00-13:59	225	265	259	327	286	326	294	1982
14:00-14:59	218	311	292	320	336	326	352	2155
15:00-15:59	221	275	304	262	250	284	296	1892
16:00-16:59	191	278	257	292	272	283	285	1858
17:00-17:59	164	278	265	294	283	274	290	1848
18:00-18:59	133	250	238	286	255	200	266	1628
19:00-19:59	154	252	254	280	262	253	271	1726
20:00-20:59	173	244	223	227	220	266	269	1622
21:00-21:59	108	175	139	162	168	217	225	1194
22:00-22:59	82	133	112	122	131	149	176	905
23:00-23:59	60	77	66	81	103	100	121	608
Total	3206	4946	4556	5156	5070	4992	5312	33238

Bogotá

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	123	42	55	36	38	83	61	438
01:00-01:59	87	27	44	22	23	65	29	297
02:00-02:59	82	27	33	17	14	49	28	250
03:00-03:59	73	9	38	10	11	50	29	220
04:00-04:59	96	27	48	42	27	62	53	355
05:00-05:59	100	162	143	155	130	98	149	937
06:00-06:59	106	371	296	385	333	174	388	2053
07:00-07:59	98	352	377	405	354	209	408	2203
08:00-08:59	114	301	299	298	306	252	285	1855
09:00-09:59	125	281	272	295	314	274	273	1834
10:00-10:59	170	278	277	293	337	306	318	1979
11:00-11:59	208	321	294	326	331	338	322	2140
12:00-12:59	193	354	333	373	355	402	363	2373
13:00-13:59	233	363	311	366	389	427	351	2440
14:00-14:59	249	349	324	348	396	406	388	2460
15:00-15:59	239	381	382	390	386	397	348	2523
16:00-16:59	225	384	409	405	425	374	447	2669
17:00-17:59	182	511	461	513	528	344	511	3049
18:00-18:59	188	384	389	434	396	289	431	2511
19:00-19:59	186	343	304	307	348	328	356	2168
20:00-20:59	192	265	279	250	278	293	281	1838
21:00-21:59	155	229	211	197	236	244	235	1507
22:00-22:59	120	140	140	140	139	219	163	1061
23:00-23:59	88	72	63	51	74	133	118	599
Total	3632	5973	5782	6057	6166	5814	6335	39759

Medellín

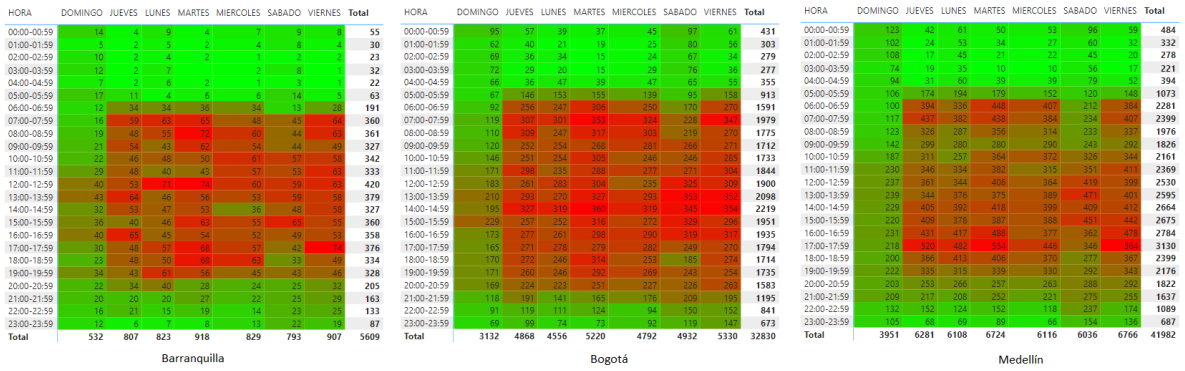
Figura 4-2 Concentración accidentes en los días de la semana por horas y ciudades año 2018

La figura 4-2 extiende el análisis anterior al desglosar los datos por ciudades, permitiendo una comparación entre Bogotá, Medellín y Barranquilla. Examina cómo varía la concentración de accidentes en diferentes horas del día y días de la semana en estas ciudades. De manera que se encuentran más concentrados los accidentes para la ciudad de Barranquilla entre las 7 am y las 6 pm, mientras que en Bogotá y Medellín la concentración empieza una hora antes y empieza a descender hasta las 10 pm. Por su parte las horas críticas se diferencian en las 3 ciudades siendo las 12m la hora crítica en Barranquilla, las 2pm para Bogotá y las 5pm en Medellín. Adicionalmente, por día de la semana se conserva menos accidentalidad el domingo y este patrón se conserva para las 3 ciudades.

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	232	103	109	91	105	202	128	970
01:00-01:59	169	66	79	55	56	148	92	665
02:00-02:59	187	55	83	38	47	114	56	580
03:00-03:59	158	50	62	25	41	140	54	530
04:00-04:59	167	69	113	80	87	147	108	771
05:00-05:59	190	331	351	340	297	229	311	2049
06:00-06:59	204	684	617	790	691	395	682	4063
07:00-07:59	252	803	746	856	756	507	818	4738
08:00-08:59	252	683	589	745	677	496	670	4112
09:00-09:59	283	605	577	610	625	553	612	3865
10:00-10:59	355	608	559	719	679	629	687	4236
11:00-11:59	430	692	609	713	649	675	778	4546
12:00-12:59	460	675	698	784	659	803	771	4850
13:00-13:59	492	701	692	758	735	883	811	5072
14:00-14:59	456	785	758	831	754	802	824	5210
15:00-15:59	485	706	676	766	715	845	793	4986
16:00-16:59	444	773	723	840	719	730	848	5077
17:00-17:59	413	839	817	901	785	637	908	5300
18:00-18:59	393	686	709	788	686	495	690	4447
19:00-19:59	427	638	622	687	644	578	643	4239
20:00-20:59	394	511	529	536	514	539	587	3610
21:00-21:59	347	428	369	444	419	509	479	2995
22:00-22:59	239	292	250	295	226	410	351	2063
23:00-23:59	186	173	150	170	171	295	302	1447
Total	**7615**	**11956**	**11487**	**12862**	**11737**	**11761**	**13003**	**80421**

Figura 4-3 Concentración accidentes en los días de la semana por horas año 2019

Continuando con el análisis temporal, la figura 4-3 se centra en los patrones de accidentes durante el año 2019. Compara la concentración de accidentes en diferentes días y horas, lo que puede revelar tendencias anuales. Donde se conservan los mismos patrones que para el análisis general.



Barranquilla

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	14	4	9	4	7	9	8	55
01:00-01:59	5	2	5	2	4	8	4	30
02:00-02:59	10	2	4	2	1	2	2	23
03:00-03:59	12	2	7	2	2	8	1	32
04:00-04:59	7	2	6	2	1	3	1	22
05:00-05:59	17	11	4	6	6	14	5	63
06:00-06:59	12	34	34	36	34	13	28	191
07:00-07:59	16	59	63	65	48	45	64	360
08:00-08:59	19	48	55	72	60	44	63	361
09:00-09:59	21	54	43	62	54	44	49	327
10:00-10:59	22	46	48	50	61	57	58	342
11:00-11:59	29	48	40	43	57	53	63	333
12:00-12:59	40	53	71	74	60	59	63	420
13:00-13:59	43	64	46	56	53	59	58	379
14:00-14:59	32	53	47	53	36	48	58	327
15:00-15:59	36	40	46	63	55	65	55	360
16:00-16:59	40	65	45	54	52	49	53	358
17:00-17:59	30	48	57	69	57	42	74	376
18:00-18:59	23	48	50	68	63	33	49	334
19:00-19:59	34	43	61	56	45	43	46	328
20:00-20:59	22	34	40	28	34	25	32	205
21:00-21:59	20	20	20	27	22	25	29	163
22:00-22:59	16	21	15	19	14	23	25	133
23:00-23:59	12	6	7	8	13	22	19	87
Total	532	807	823	918	829	793	907	5609

Bogotá

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	95	57	39	37	45	97	61	431
01:00-01:59	62	40	21	19	25	80	56	303
02:00-02:59	69	36	34	15	24	67	34	279
03:00-03:59	72	29	20	15	29	76	36	277
04:00-04:59	66	36	47	39	47	65	55	355
05:00-05:59	67	146	153	155	139	95	158	913
06:00-06:59	92	256	247	306	250	170	270	1591
07:00-07:59	119	307	301	353	324	228	347	1979
08:00-08:59	110	309	247	317	303	219	270	1775
09:00-09:59	120	252	254	268	281	266	271	1712
10:00-10:59	146	251	254	305	246	246	285	1733
11:00-11:59	171	298	235	288	277	271	304	1844
12:00-12:59	183	261	283	304	235	325	309	1900
13:00-13:59	210	293	270	327	293	353	352	2098
14:00-14:59	195	323	319	360	319	345	358	2219
15:00-15:59	229	257	252	316	272	329	296	1951
16:00-16:59	173	277	261	298	290	319	317	1935
17:00-17:59	165	271	278	279	282	249	270	1794
18:00-18:59	170	272	246	314	253	185	274	1714
19:00-19:59	171	260	246	292	269	243	254	1735
20:00-20:59	169	224	223	251	227	226	263	1583
21:00-21:59	118	191	141	165	176	209	195	1195
22:00-22:59	91	119	111	124	94	150	152	841
23:00-23:59	89	98	74	73	92	119	147	673
Total	3132	4868	4556	5220	4792	4932	5330	32830

Medellín

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	123	42	61	50	53	96	59	484
01:00-01:59	102	24	53	34	27	60	32	332
02:00-02:59	108	17	45	21	22	45	20	278
03:00-03:59	74	19	35	10	10	56	17	221
04:00-04:59	94	31	60	39	39	79	52	394
05:00-05:59	106	174	194	179	152	120	148	1073
06:00-06:59	100	394	336	448	407	212	384	2281
07:00-07:59	117	437	382	438	384	234	407	2399
08:00-08:59	123	326	287	356	314	233	337	1976
09:00-09:59	142	299	280	280	290	243	292	1826
10:00-10:59	187	311	257	364	372	326	344	2161
11:00-11:59	230	346	334	382	315	351	411	2369
12:00-12:59	237	361	344	406	364	419	399	2530
13:00-13:59	239	344	376	375	389	471	401	2595
14:00-14:59	229	405	392	418	399	409	412	2664
15:00-15:59	220	409	378	387	388	451	442	2675
16:00-16:59	231	431	417	488	377	362	478	2784
17:00-17:59	218	520	482	554	446	346	564	3130
18:00-18:59	200	368	413	406	370	277	367	2399
19:00-19:59	222	335	315	339	330	292	343	2176
20:00-20:59	203	253	266	257	263	288	292	1822
21:00-21:59	209	217	208	252	221	275	255	1637
22:00-22:59	132	152	124	152	118	237	174	1089
23:00-23:59	108	68	69	89	66	154	133	687
Total	3951	6281	6108	6724	6116	6036	6766	41982

Figura 4-4 Concentración accidentes en los días de la semana por horas y ciudades año 2019

La figura 4-4 amplía el análisis temporal para el año 2019, incluyendo la comparación entre las tres ciudades. Esto permite identificar diferencias en los patrones de accidentes en función de la ubicación geográfica y la temporalidad. Para 2019, se tiene que las horas donde más se concentra la accidentalidad para Barranquilla es de 6am a 9 pm, mientras que en Bogotá de 5am a 10pm y Medellín de 5am a 11pm puesto que en esta ciudad los accidentes se reparten de manera mas constante a lo largo de día.

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	132	50	68	46	51	97	67	511
01:00-01:59	111	34	39	21	27	63	34	329
02:00-02:59	89	22	32	15	32	66	30	286
03:00-03:59	78	20	42	25	33	54	38	290
04:00-04:59	87	64	54	55	51	88	56	455
05:00-05:59	102	183	173	171	215	173	158	1175
06:00-06:59	147	284	303	319	334	236	292	1915
07:00-07:59	156	395	325	395	387	289	354	2301
08:00-08:59	137	379	304	331	387	288	342	2168
09:00-09:59	148	383	310	340	353	287	342	2163
10:00-10:59	194	330	316	383	377	299	337	2236
11:00-11:59	221	382	378	397	340	376	359	2453
12:00-12:59	231	419	348	424	430	444	421	2717
13:00-13:59	253	442	408	465	430	469	454	2921
14:00-14:59	295	403	404	449	403	432	456	2842
15:00-15:59	241	420	388	439	429	462	422	2801
16:00-16:59	248	397	414	450	444	392	469	2814
17:00-17:59	233	454	443	429	469	347	467	2842
18:00-18:59	209	361	357	426	387	308	396	2444
19:00-19:59	218	360	320	337	343	314	348	2240
20:00-20:59	220	313	279	283	281	307	303	1986
21:00-21:59	198	208	173	214	200	245	278	1516
22:00-22:59	149	173	129	132	127	210	183	1103
23:00-23:59	68	99	80	72	76	148	151	694
Total	**4165**	**6575**	**6087**	**6618**	**6606**	**6394**	**6757**	**43202**

Figura 4-5 Concentración accidentes en los días de la semana por horas año 2020

Continúa el análisis temporal con la figura 3-5 explorando la concentración de accidentes durante el año 2020. En donde se observa los cambios en los patrones de accidentes en comparación con los años anteriores, se tiene que esta más concentrado entre las 12 y las 7pm.

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	7	2	8	3	2	4	1	27
01:00-01:59	8	2	2	2	2	3	1	20
02:00-02:59	8	2	4	1	2	4		21
03:00-03:59	4	1				1	1	7
04:00-04:59	7	2	3	2	2	6	1	23
05:00-05:59	3	6	7	6	3	7	7	39
06:00-06:59	10	15	13	17	20	15	17	107
07:00-07:59	14	32	31	38	30	20	30	195
08:00-08:59	10	41	29	37	33	33	26	209
09:00-09:59	7	29	39	38	31	21	35	200
10:00-10:59	9	28	33	35	26	26	28	185
11:00-11:59	22	33	26	39	32	25	29	206
12:00-12:59	20	35	25	47	42	48	45	262
13:00-13:59	12	36	35	36	31	37	34	221
14:00-14:59	14	34	37	31	32	36	39	223
15:00-15:59	23	34	36	33	41	38	26	231
16:00-16:59	17	28	37	35	40	30	38	225
17:00-17:59	11	34	34	29	33	32	34	207
18:00-18:59	14	38	28	40	35	21	26	202
19:00-19:59	17	21	25	20	25	16	20	144
20:00-20:59	17	20	18	19	21	13	11	119
21:00-21:59	9	13	8	14	13	19	17	93
22:00-22:59	7	12	9	7	2	4	6	47
23:00-23:59	2	5	1	6	6	6	4	30
Total	272	503	488	535	504	465	476	3243

Barranquilla

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	73	31	34	30	31	59	42	300
01:00-01:59	50	26	16	10	15	39	21	177
02:00-02:59	41	11	11	10	18	46	21	158
03:00-03:59	47	12	25	18	19	43	23	187
04:00-04:59	42	35	21	27	18	44	32	219
05:00-05:59	62	94	85	80	110	88	81	600
06:00-06:59	94	133	140	176	161	127	148	979
07:00-07:59	93	198	137	199	177	152	184	1140
08:00-08:59	75	196	163	172	201	155	192	1154
09:00-09:59	90	208	149	168	190	176	195	1176
10:00-10:59	110	174	157	200	184	147	176	1148
11:00-11:59	105	173	204	203	172	199	200	1256
12:00-12:59	116	216	165	210	235	189	204	1335
13:00-13:59	151	222	192	251	216	245	244	1521
14:00-14:59	159	205	191	235	225	237	254	1506
15:00-15:59	123	195	166	232	197	230	212	1355
16:00-16:59	141	176	180	209	214	209	219	1348
17:00-17:59	138	212	191	201	205	182	224	1353
18:00-18:59	110	158	173	204	195	171	193	1204
19:00-19:59	125	186	178	184	178	181	211	1243
20:00-20:59	113	163	158	147	151	183	179	1094
21:00-21:59	100	110	92	113	101	142	151	809
22:00-22:59	85	88	72	69	76	110	98	598
23:00-23:59	39	56	48	31	53	81	93	401
Total	2282	3278	2948	3379	3342	3435	3597	22261

Bogotá

HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
00:00-00:59	52	17	26	13	18	34	24	184
01:00-01:59	53	6	21	9	10	21	12	132
02:00-02:59	40	9	17	4	12	16	9	107
03:00-03:59	27	7	17	7	14	10	14	96
04:00-04:59	38	27	30	26	31	38	23	213
05:00-05:59	37	83	81	65	102	78	70	536
06:00-06:59	43	136	150	126	153	94	127	829
07:00-07:59	49	165	157	158	180	117	140	966
08:00-08:59	52	142	112	122	153	100	124	805
09:00-09:59	51	146	122	134	132	90	112	787
10:00-10:59	75	128	126	148	167	126	133	903
11:00-11:59	94	176	148	155	136	152	130	991
12:00-12:59	95	168	158	167	153	207	172	1120
13:00-13:59	90	184	181	178	183	187	176	1179
14:00-14:59	122	164	176	183	146	159	163	1113
15:00-15:59	95	191	186	174	191	194	184	1215
16:00-16:59	90	193	197	206	190	153	212	1241
17:00-17:59	84	208	218	199	231	133	209	1282
18:00-18:59	85	165	156	182	157	116	177	1038
19:00-19:59	76	153	117	133	140	117	117	853
20:00-20:59	90	130	103	117	109	111	113	773
21:00-21:59	89	85	73	87	86	84	110	614
22:00-22:59	57	73	48	56	49	96	79	458
23:00-23:59	27	38	31	35	17	61	54	263
Total	1611	2794	2651	2704	2760	2494	2684	17698

Medellín

Figura 4-6 Concentración accidentes en los días de la semana por horas y ciudades año 2020

Finaliza el análisis temporal para el año 2020 por medio de la figura 4-6, teniendo en cuenta las diferencias entre las tres ciudades. Estos datos pueden ayudar a comprender la influencia de eventos específicos en la concentración de accidentes y sus patrones no cambian con respecto a los años anteriores.

MES	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
1	496	766	818	990	1044	706	796	5616
2	566	865	948	954	978	933	1056	6300
3	600	1151	874	1023	892	1144	1114	6798
4	724	949	1254	957	947	898	1003	6732
5	606	1165	805	1170	1242	929	1032	6949
6	637	925	739	918	956	1148	1277	6600
7	648	925	1118	1150	909	867	854	6471
8	556	1123	888	855	1203	949	1280	6854
9	650	894	890	967	901	1165	1060	6527
10	525	917	1030	1211	1201	916	981	6781
11	581	1147	736	1040	926	962	1130	6522
12	791	976	1103	883	964	1058	928	6703
Total	**7380**	**11803**	**11203**	**12118**	**12163**	**11675**	**12511**	**78853**

Figura 4-7 Concentración de la accidentalidad por días en los meses del año 2018

Por otro lado, la figura 4-7 cambia el enfoque al análisis de la accidentalidad a lo largo de los meses del año 2018. Examina cómo varía la cantidad de accidentes en diferentes meses y días de la semana. Y no se encuentra un día de la semana que sea más crítico que los otros para todos los meses puesto que varían mes a mes.

MES	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
1	480	945	720	920	968	777	805	5615
2	559	911	912	922	908	849	982	6043
3	631	901	847	992	874	1042	1159	6446
4	558	826	1151	1088	929	779	850	6181
5	571	1208	912	938	1106	896	1252	6883
6	673	888	708	995	866	1149	941	6220
7	566	925	1103	1245	1095	899	1031	6864
8	645	1243	961	1068	891	1233	1394	7435
9	802	1056	1207	1127	986	1052	1118	7348
10	627	1241	874	1320	1227	967	1114	7370
11	608	804	826	1056	963	1195	1269	6721
12	895	1008	1266	1191	924	923	1088	7295
Total	**7615**	**11956**	**11487**	**12862**	**11737**	**11761**	**13003**	**80421**

Figura 4-8 Concentración de la accidentalidad por días en los meses del año 2019

La figura 4-8 extiende el análisis anterior al año 2019, permitiendo una comparación anual de la concentración de la accidentalidad en diferentes meses y días de la semana y para este caso, los viernes son los días más críticos en la mayoría de los meses.

MES	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES	Total
1	545	968	788	892	998	759	1087	6037
2	650	980	935	1062	952	1259	1042	6880
3	523	662	744	758	709	510	624	4530
4	133	271	250	243	284	201	202	1584
5	304	416	362	352	457	489	447	2827
6	327	501	505	669	485	544	591	3622
7	243	698	523	551	678	428	576	3697
8	337	572	657	542	587	573	562	3830
9	213	303	292	400	381	336	359	2284
10	271	439	280	329	317	469	489	2594
11	329	348	366	382	336	419	390	2570
12	290	417	385	438	422	407	388	2747
Total	**4165**	**6575**	**6087**	**6618**	**6606**	**6394**	**6757**	**43202**

Figura 4-9 Concentración de la accidentalidad por días en los meses del año 2020

Finaliza el análisis de la concentración de la accidentalidad en el año 2020 con la figura 4-9, proporcionando datos sobre la variación de accidentes en diferentes meses y días de la semana durante ese año, por su parte en 2020 hay una menor concentración de accidentes debido a la pandemia Covid-19.

Análisis de Accidentes por Zonas:

El análisis geográfico es otro pilar esencial de esta investigación. Se llevó a cabo un detallado análisis de la distribución espacial de accidentes en tres ciudades: Bogotá, Medellín y Barranquilla. De manera que se pueda ver la frecuencia de accidentalidad por cada una de las zonas y así identificar cuáles son los lugares críticos para cada ciudad, para así tomar medidas sobre estas zonas y minimizar la siniestralidad vial

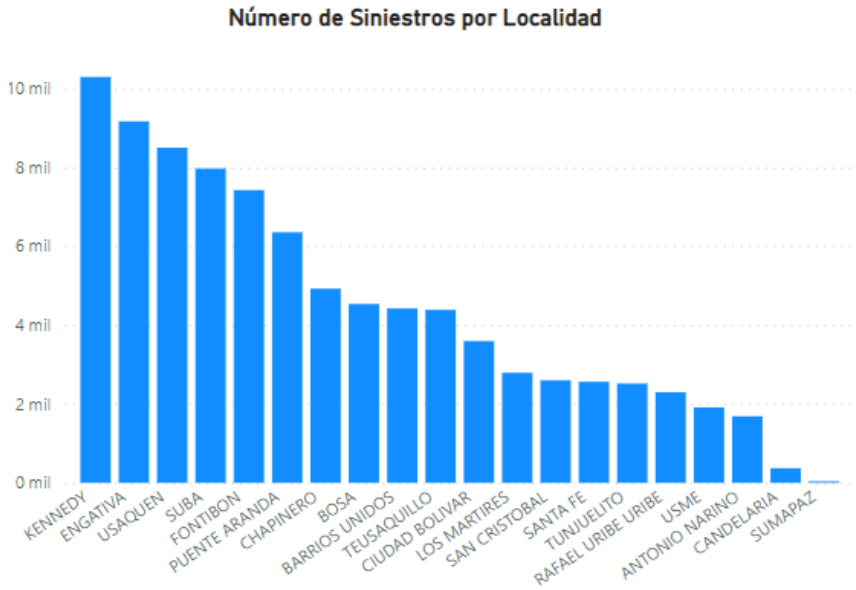


Figura 4-10 Número de siniestros por localidad en Bogotá

La figura 4-10 detalla el número de siniestros por localidad en Bogotá, resaltando las áreas con una alta concentración de accidentes. Por ejemplo, se identificaron cinco localidades que sobresalen por su alta concentración de accidentes, entre ellas Kennedy, Fontibón, Engativá, Suba y Usaquén. Lo cual proporciona información clave para la planificación de intervenciones de seguridad vial en la ciudad.

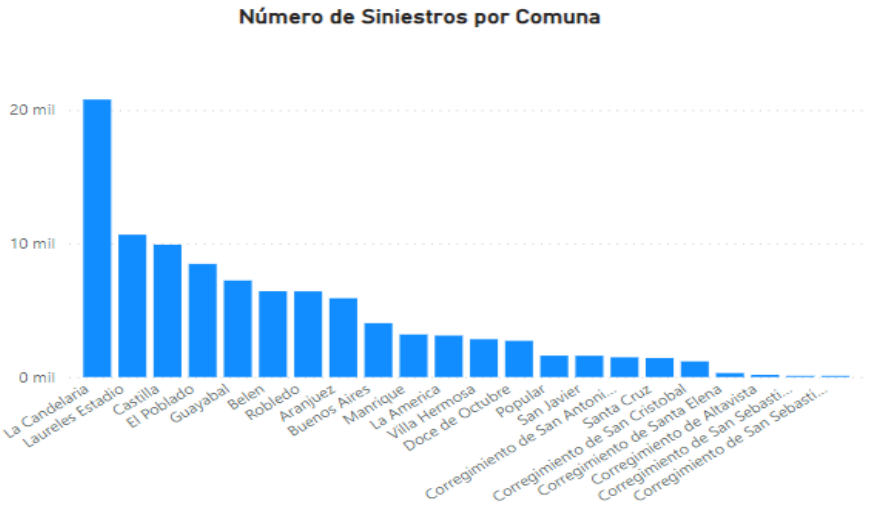


Figura 4-11 Número de siniestros por comuna (Medellín)

En el caso de Medellín, la figura 4-11 permite observar que hay una comuna (La Candelaria) en donde predomina un gran número de accidentes. Esto puede ser esencial para la asignación de recursos y estrategias de seguridad vial en Medellín

Matriz de Correlaciones:

La matriz de correlaciones se erigió como una herramienta esencial para explorar las relaciones entre las variables de estudio. Empleando herramientas como Jupyter dentro de Anaconda, se llevó a cabo un análisis profundo. A pesar de la exhaustiva revisión, es importante señalar que no se obtuvieron conclusiones significativas a partir de estas correlaciones. Este hallazgo podría indicar que las variables seleccionadas no presentan asociaciones claras entre sí o que se requiere una investigación más avanzada para desentrañar completamente las relaciones subyacentes.

```
import pandas as pd
import numpy as np
from pandas_profiling import ProfileReport
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Cargar tus datos
df = pd.read_excel("C:/Users/yerit/OneDrive - ESCUELA COLOMBIANA DE INGENIERIA JULIO GARAVITO/aticulo para proyecto grado/datos/E
```

Figura 4-12 Ejemplo librerías

La figura 4-12 presenta un ejemplo de las librerías y herramientas utilizadas para llevar a cabo el análisis de correlaciones y explorar las relaciones entre las variables de estudio.

```
corr_df = df.corr()
print(corr_df, "\n")
```

```
                  CODIGO_ACCIDENTE      ANIO       MES       DIA   LATITUD  \
CODIGO_ACCIDENTE          1.000000 -0.002752 -0.000582 -0.002277  0.004367
ANIO                     -0.002752  1.000000 -0.048514 -0.007913 -0.068190
MES                      -0.000582 -0.048514  1.000000 -0.008208 -0.033009
DIA                      -0.002277 -0.007913 -0.008208  1.000000 -0.007643
LATITUD                   0.004367 -0.068190 -0.033009 -0.007643  1.000000
LONGITUD                 -0.004345  0.066548  0.032247  0.007665 -0.996680
CANTIDAD_HERIDOS         -0.003449  0.019653 -0.009157  0.000209  0.113870
CANTIDAD_ILESO           -0.068203 -0.063498  0.001614  0.004389  0.121699
CANTIDAD_MUERTOS         -0.000436  0.000083  0.002645 -0.002050 -0.049096
CANTIDAD_VICTIMAS        -0.005545  0.053160  0.015865  0.006545 -0.686168
CANTIDAD_VEHICULOS        0.002108  0.011800  0.006237  0.002381  0.097000
EDAD PROMEDIO             0.159995  0.054626  0.008054  0.001776 -0.202136
EDAD MEDIANA              0.154357  0.053151  0.007656  0.001924 -0.195348
CANTIDAD HOMBRES         -0.005631  0.043275  0.015426  0.004998 -0.703513
CANTIDAD MUJERES         -0.002349  0.009915  0.004735  0.003717 -0.104968
```

Figura 4-13 Tabla de correlaciones

La figura 4-13 presenta unas tablas que resume las correlaciones entre las variables de estudio. Este es un componente central del análisis de correlaciones. Por ejemplo, de esta figura se puede ir deduciendo que la cantidad de hombres se relaciona con la latitud negativamente, esto quiere decir que a medida que se va más al norte hay menos hombres relacionados a los accidentes.

```
                   LONGITUD  CANTIDAD_HERIDOS  CANTIDAD_ILESO  \
CODIGO_ACCIDENTE  -0.004345         -0.003449       -0.068203
ANIO               0.066548          0.019653       -0.063498
MES                0.032247         -0.009157        0.001614
DIA                0.007665          0.000209        0.004389
LATITUD           -0.996680          0.113870        0.121699
LONGITUD           1.000000         -0.117973        0.119476
CANTIDAD_HERIDOS  -0.117973          1.000000       -0.580509
CANTIDAD_ILESO     0.119476         -0.580509        1.000000
CANTIDAD_MUERTOS   0.046663         -0.026535       -0.147066
CANTIDAD_VICTIMAS  0.687473          0.536829        0.143162
CANTIDAD_VEHICULOS 0.061630         -0.197192        0.697218
EDAD PROMEDIO      0.205972         -0.172542        0.145726
EDAD MEDIANA       0.199279         -0.182839        0.146843
CANTIDAD HOMBRES   0.703470          0.275601        0.272117
CANTIDAD MUJERES   0.108002          0.541373       -0.125277
```

Figura 4-14 Tabla de correlaciones 2

Se continúa presentando tablas de correlaciones adicionales para explorar las relaciones entre las variables de estudio desde diferentes perspectivas. De la figura 4-14 se puede concluir que la cantidad de víctimas involucrados en los accidentes aumenta hacia el occidente del país y a vez está relacionado de manera positiva con la cantidad de hombres involucrados. Y, por último, entre más vehículos involucrados hay más ilesos en los accidentes, esto quiere decir que no siempre son accidentes de gravedad

	CANTIDAD_MUERTOS	CANTIDAD_VICTIMAS	CANTIDAD_VEHICULOS	\
CODIGO_ACCIDENTE	-0.000436	-0.005545	0.002108	
ANIO	0.000083	0.053160	0.011800	
MES	0.002645	0.015865	0.006237	
DIA	-0.002050	0.006545	0.002381	
LATITUD	-0.049096	-0.686168	0.097000	
LONGITUD	0.046663	0.687473	0.061630	
CANTIDAD_HERIDOS	-0.026535	0.536829	-0.197192	
CANTIDAD_ILESO	-0.147066	0.143162	0.697218	
CANTIDAD_MUERTOS	1.000000	0.073927	-0.097264	
CANTIDAD_VICTIMAS	0.073927	1.000000	0.354060	
CANTIDAD_VEHICULOS	-0.097264	0.354060	1.000000	
EDAD PROMEDIO	0.042428	0.076160	0.020984	
EDAD MEDIANA	0.040179	0.059325	0.011225	
CANTIDAD HOMBRES	0.062287	0.833994	0.451912	
CANTIDAD MUJERES	0.015915	0.480004	-0.063277	

Figura 4-15 Tabla de correlaciones 3

La figura 4-15 proporciona otras deducciones, como, por ejemplo, que la edad ya sea promedio o mediana no está correlacionado ya sea con la cantidad de muertos, victimas o vehículos involucrados en los accidentes.

	EDAD PROMEDIO	EDAD MEDIANA	CANTIDAD HOMBRES	\
CODIGO_ACCIDENTE	0.159995	0.154357	-0.005631	
ANIO	0.054626	0.053151	0.043275	
MES	0.008054	0.007656	0.015426	
DIA	0.001776	0.001924	0.004998	
LATITUD	-0.202136	-0.195348	-0.703513	
LONGITUD	0.205972	0.199279	0.703470	
CANTIDAD_HERIDOS	-0.172542	-0.182839	0.275601	
CANTIDAD_ILESO	0.145726	0.146843	0.272117	
CANTIDAD_MUERTOS	0.042428	0.040179	0.062287	
CANTIDAD_VICTIMAS	0.076160	0.059325	0.833994	
CANTIDAD_VEHICULOS	0.020984	0.011225	0.451912	
EDAD PROMEDIO	1.000000	0.992645	0.090865	
EDAD MEDIANA	0.992645	1.000000	0.078831	
CANTIDAD HOMBRES	0.090865	0.078831	1.000000	
CANTIDAD MUJERES	-0.019663	-0.026846	-0.058115	

Figura 4-16 Tabla de correlaciones 4

La figura 4-16 permite reforzar la idea que las variables relacionadas con la edad presentan correlaciones débiles o nulas contra las demás variables.

	CANTIDAD MUJERES
CODIGO_ACCIDENTE	-0.002349
ANIO	0.009915
MES	0.004735
DIA	0.003717
LATITUD	-0.104968
LONGITUD	0.108002
CANTIDAD_HERIDOS	0.541373
CANTIDAD_ILESO	-0.125277
CANTIDAD_MUERTOS	0.015915
CANTIDAD_VICTIMAS	0.480004
CANTIDAD_VEHICULOS	-0.063277
EDAD PROMEDIO	-0.019663
EDAD MEDIANA	-0.026846
CANTIDAD HOMBRES	-0.058115
CANTIDAD MUJERES	1.000000

Figura 4-17 Tabla de correlaciones 5

La figura 4-17 finaliza la presentación de tablas de correlaciones, ofreciendo un enfoque completo para comprender las relaciones entre las variables de estudio.

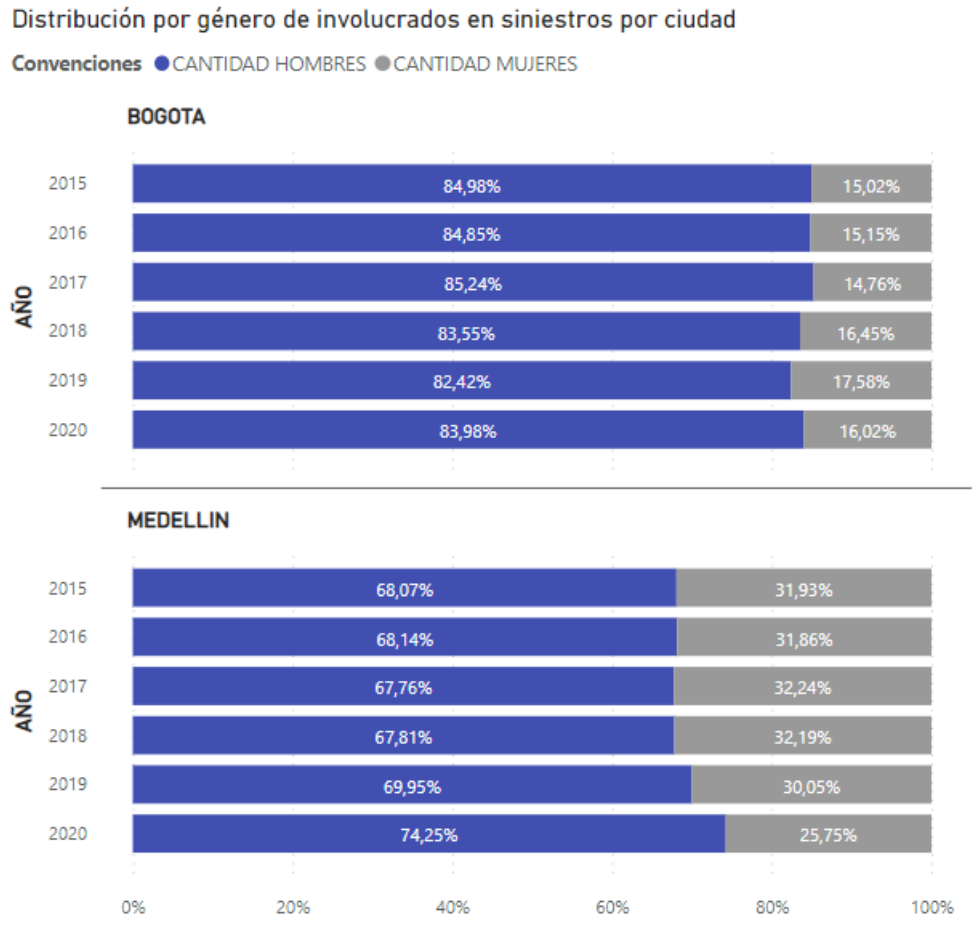


Figura 4-18 Distribución de siniestrados por genero

La figura 4-18 se enfoca en la distribución de siniestrados por género, durante los años observados se tiene que los hombres se accidentan en mayor proporción que las mujeres, adicionalmente, en Bogotá se accidentan en mayor proporción los hombres que las mujeres que en Medellín donde la proporción es menor en alrededor de 10 puntos porcentuales.

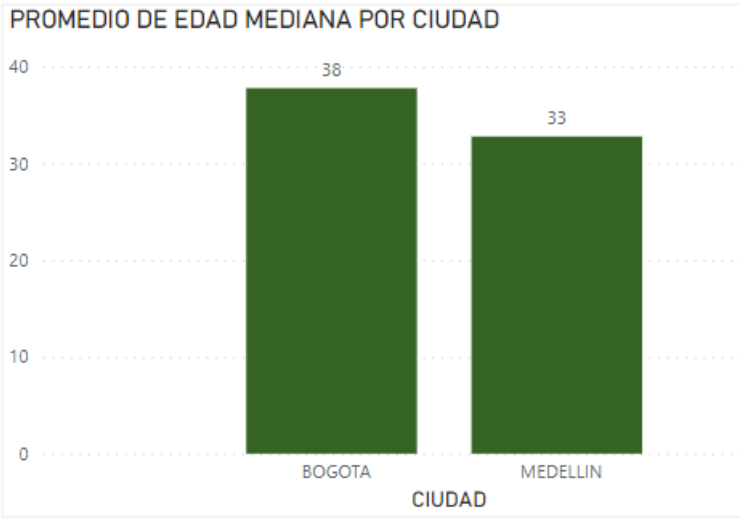


Figura 4-19 Promedio de la edad en siniestrados por ciudad

La figura 4-19 analiza el promedio de edad en los siniestrados, lo que puede ayudar a comprender la edad promedio de las personas involucradas en accidentes de tráfico en las ciudades estudiadas. Se observa que la edad promedio de los involucrados en los siniestrados es similar para las dos ciudades por lo tanto no es determinante este factor para las ciudades.

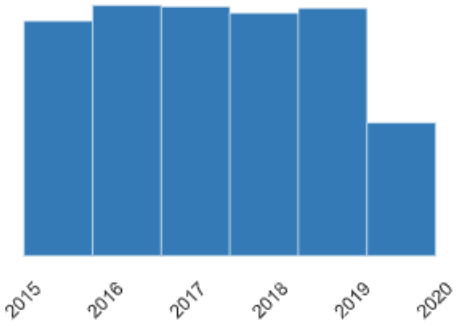


Ilustración 4-20 Cantidad de accidentes por año

Examina la cantidad de accidentes por año, enfocándose en las tendencias a lo largo del tiempo y revelando si un año en particular se destacó por su cantidad de siniestros. La cantidad de datos en el año 2020 confirma que ha sido un año atípico en cuanto a la cantidad de siniestros en efecto por el confinamiento debido al COVID 19

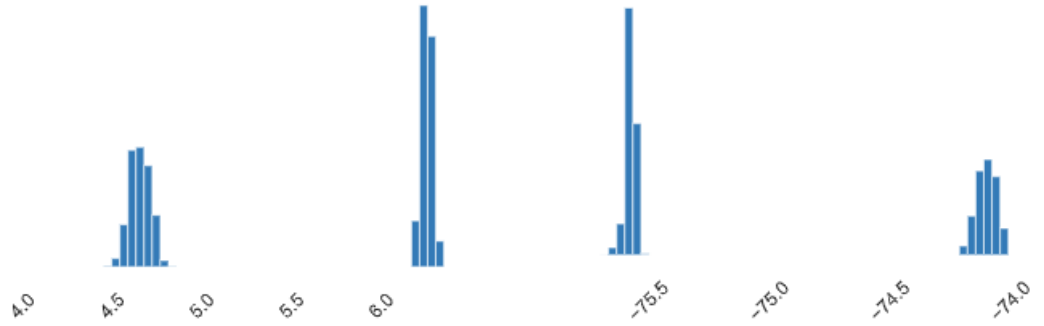


Figura 4-21. Distribución de la Latitud

Figura 4-22. Distribución de la longitud

La distribución de la latitud y la longitud también se analizó para revisar la consistencia de los datos, puesto que como solo se están analizando 3 ciudades no debería tener mucha varianza. Esto se ve reflejado en las figuras 4-21 y 4-22 en donde los datos se agrupan en dos grupos.

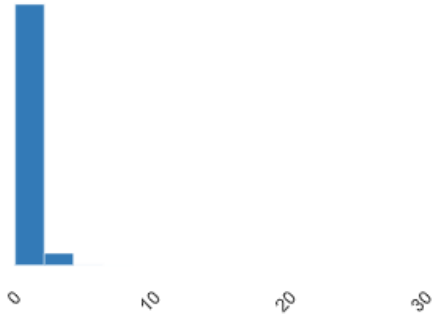


Figura 4-23. Distribución de los Ilesos

De otra manera se tiene que la cantidad de ilesos y de muertos tienen un comportamiento sesgado a la izquierda, y la cantidad de accidentados hombres y mujeres como es natural conservan el mismo comportamiento. Por su parte, la mayor cantidad de ilesos en un accidente fue de 30 y la mayor cantidad de muertes en un accidente fue de 9 personas. En adición, los ilesos son concentran alrededor del cero como se observa en la figura 4-23. Es importante recordar que esta variable implica para la ciudad de Medellín

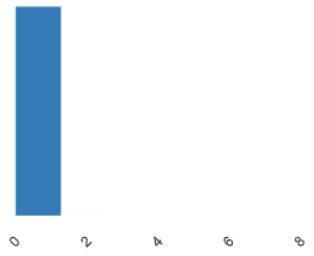


Figura 4-24. Distribución de los Fallecidos

La figura 4-24 muestra la distribución de los fallecidos en los accidentes, revelando cómo se agrupan en diferentes valores y proporcionando información sobre la gravedad de los incidentes. La mayoría de las siniestralidades viales presentan 0 o 1 fallecido.

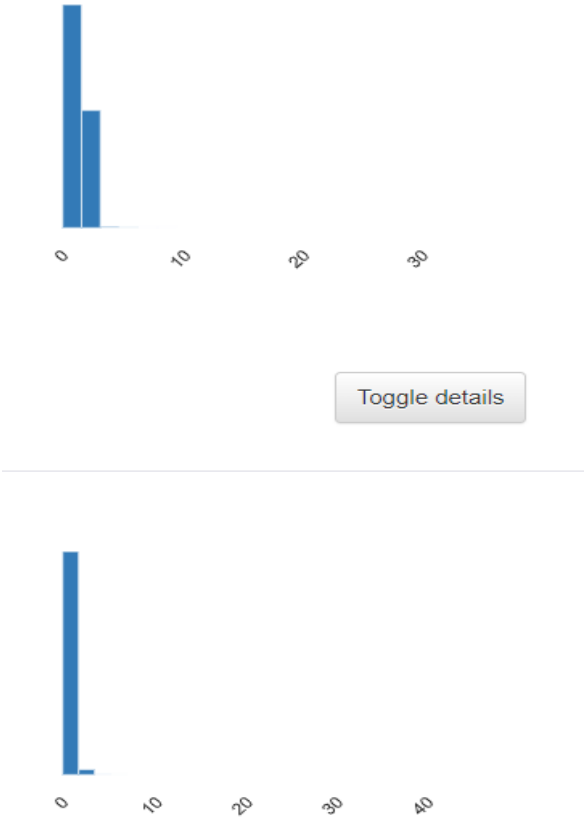


Figura 4-25. Distribución fallecidos de Hombres y Mujeres

Después de un minucioso análisis inicial de nuestros datos, decidimos llevar a cabo pruebas adicionales con el fin de profundizar y obtener una comprensión más completa de la estructura de nuestros datos. En esta etapa, optamos por utilizar la prueba FAMD (Análisis Factorial de Datos

Mixtos), una técnica poderosa que nos permite explorar las relaciones entre variables tanto cuantitativas como cualitativas.

Al examinar las variables cuantitativas, dedicamos especial atención a los detalles, presentando los resultados de manera exhaustiva en la Figura 4-26. A pesar de este enfoque detallado, lamentablemente no observamos contribuciones significativas adicionales con respecto a lo que ya habíamos descubierto a través del análisis de la matriz de correlaciones previo. Esto sugiere que las relaciones entre las variables cuantitativas ya han sido capturadas en gran medida por los métodos previos de análisis, lo que refuerza los resultados obtenidos anteriormente.

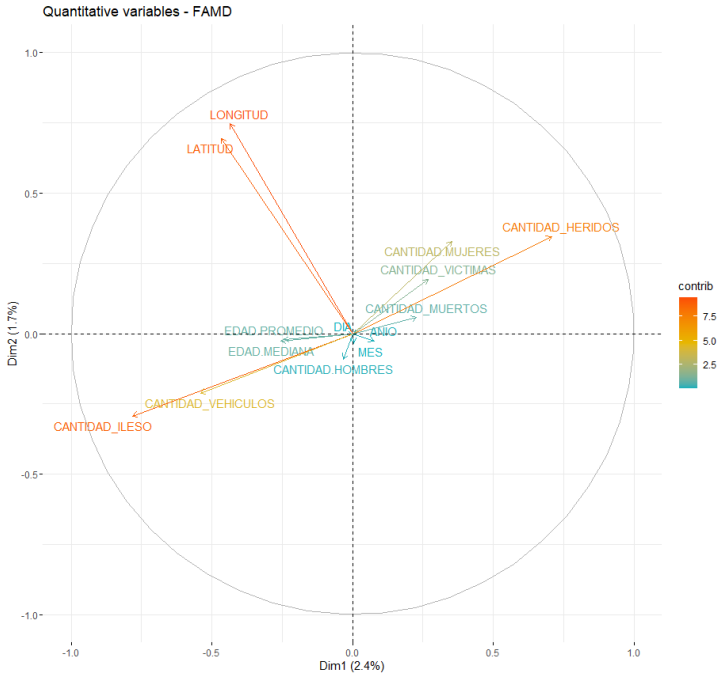


Figura 4-26. Distribución Contribución de las Variables Cuantitativas – FAMD

En cuanto a las variables cualitativas, nuestra exploración reveló un patrón similar. La mayoría de los valores se agruparon en las proximidades del origen en el espacio factorial, lo que indica una dispersión limitada de los datos en estas dimensiones. Este fenómeno se ilustra claramente en la Figura 4-27 donde se muestra las contribuciones de las variables en las dimensiones 1 y 2, donde se observa una concentración alrededor del punto de origen lo que indica que no hay una que contribuya más que las demás o que este muy relacionada con las otras. Como consecuencia de esta distribución, no encontramos nueva información sustancial que pudiera enriquecer nuestro entendimiento de las relaciones entre las variables cualitativas.

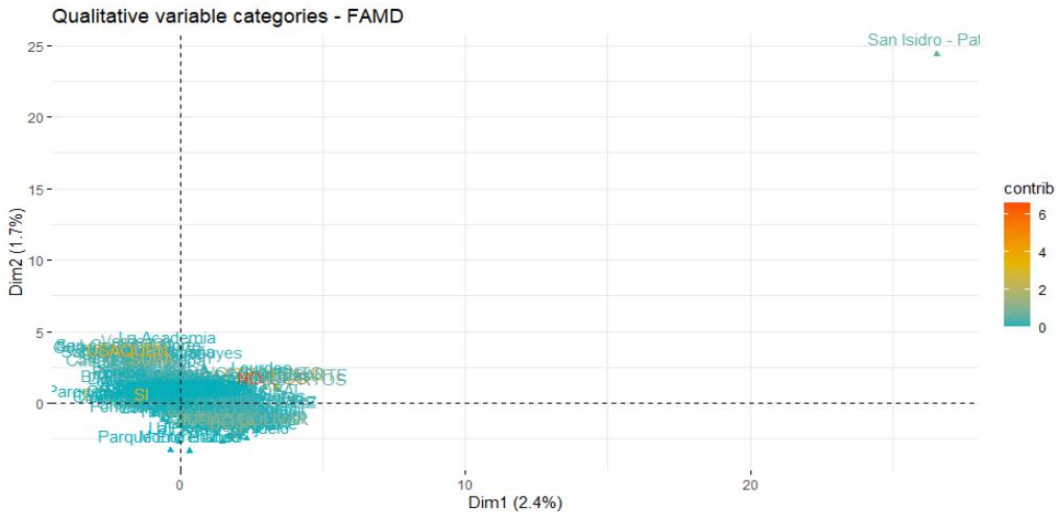


Figura 4-27. Distribución Contribución de las Variables Cuantitativas - FAMD

Es importante destacar que esta falta de novedades no debe interpretarse como una limitación de nuestro estudio, sino más bien como una indicación de la complejidad inherente de los datos y la robustez de los análisis previamente realizados puesto que las figuras 4-28 y 4-29 nos muestran las contribuciones de las variables en la dimensión 1 en la cual son significativas no obstante en la dimensión 2 se pierde explicabilidad de los datos pues tiene una contribución menor al 5% la mayoría de las variables lo cual no permite una reducción de dimensionalidad. En consecuencia, concluimos que reducir el número de variables no sería apropiado en este contexto, ya que los datos parecen requerir todas las dimensiones para ser explicados de manera adecuada. Este hallazgo resalta la riqueza y la profundidad de la información contenida en nuestro conjunto de datos completo, subrayando la importancia de mantener un enfoque holístico en nuestro análisis.

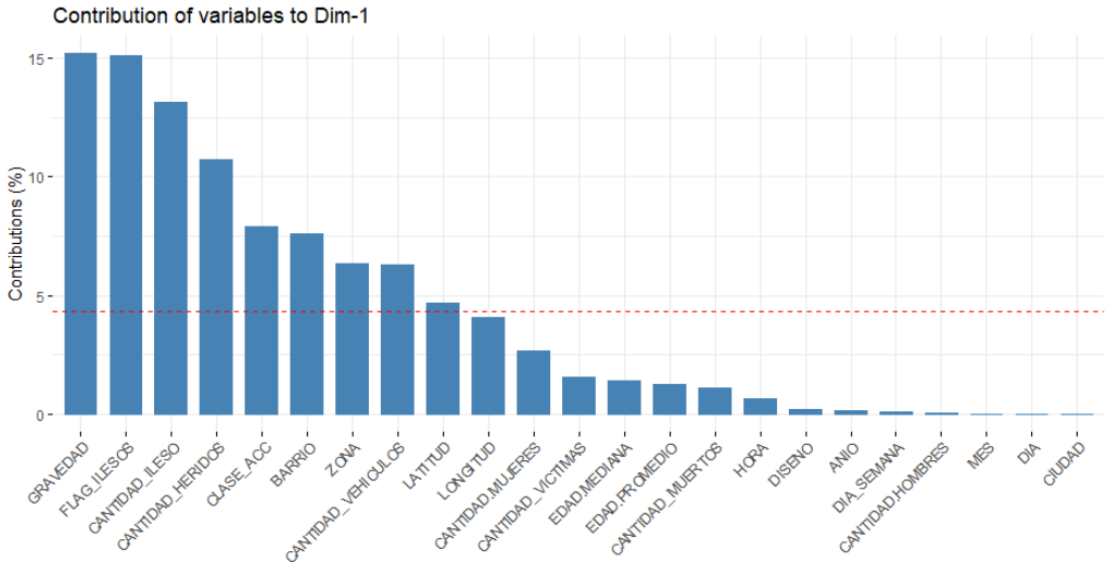


Figura 4-28. Distribución Contribución de las Variables en dimensión 1 - FAMD

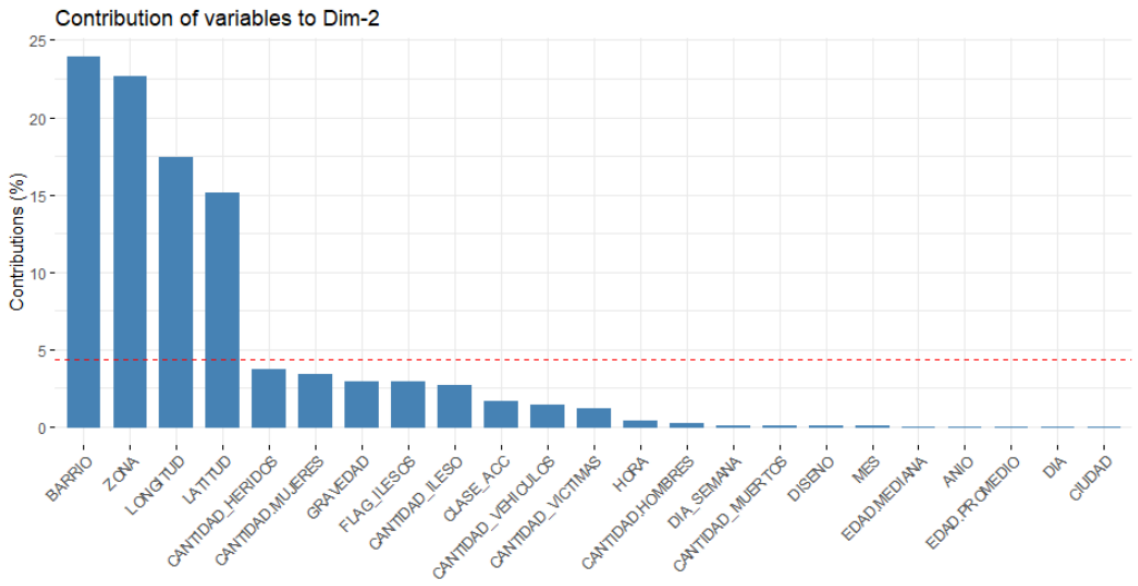


Figura 4-29. Distribución Contribución de las Variables en dimensión 2 - FAMD

4.2 Clasificación

La clasificación no supervisada a se realizó por zonas de manera que cada zona por ciudad puede tener comportamientos similares a otras mediante los valores de sus datos, sin necesidad de tener en cuenta la ubicación geográfica.

4.2.1 Variables

Para la clasificación no supervisada se hizo uso de las variables de tipo numérico usando una escala de normalización con la escala del mínimo y máximo. En adición, las variables usadas fueron las siguientes:

-CANTIDAD_ACCIDENTES
-CANTIDAD_HERIDOS
-CANTIDAD_ILESO
-CANTIDAD_MUERTOS
-CANTIDAD_VICTIMAS
-CANTIDAD_VEHICULOS
-EDAD MEDIANA
-CANTIDAD HOMBRES
-CANTIDAD MUJERES

4.2.1.1 Modelo

Por otro lado, la técnica del codo es comúnmente utilizada en el análisis de *clustering* o clasificación no supervisada (agrupamiento) para determinar el número óptimo de *clusters*, que se denota como "k". El objetivo es identificar el número de grupos o *clusters* en los que se pueden agrupar los datos

de manera que se maximice la cohesión (similitud dentro del mismo *cluster*) y se minimice la separación (diferencia entre *clusters*).

El proceso implica ajustar el algoritmo de *clustering*, como *k-means*, para diferentes valores de k y evaluar cómo varía la cohesión en función de k. Generalmente, a medida que aumenta k, la cohesión aumenta, ya que se pueden formar *clusters* más pequeños y específicos. Sin embargo, llega un punto en el que el aumento de k no proporciona una mejora significativa en la cohesión. Este punto, donde la mejora disminuye, se conoce como el "codo" en el gráfico de cohesión vs. k.

El valor de "k" óptimo es aquel en el que el gráfico de cohesión vs. k muestra un quiebre o codo, indicando que no es necesario tener más *clusters* para describir los datos de manera efectiva. El valor de "k" en el punto del codo se elige el número óptimo de *clusters* para el análisis de *clustering*.

Por ciudad, se procedió a realizar la técnica del codo para detectar el k óptimo, la visualización de los *clusters* y sus respectivos centroides. Los resultados son los siguientes:

Segmentación para Bogotá:

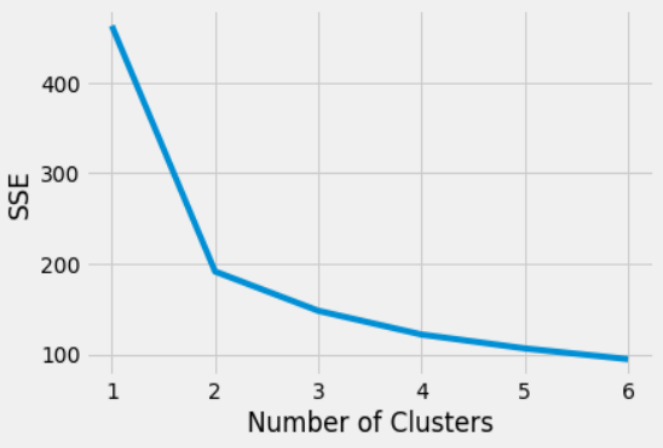


Figura 4-30. gráfico de cohesión vs. K – Bogotá

La figura 4-30, muestra que el valor de k óptimo sería de 2 pero se considera que puede llegar a ser un número insuficiente de grupos por lo que se consideró tomar un valor de k = 3.

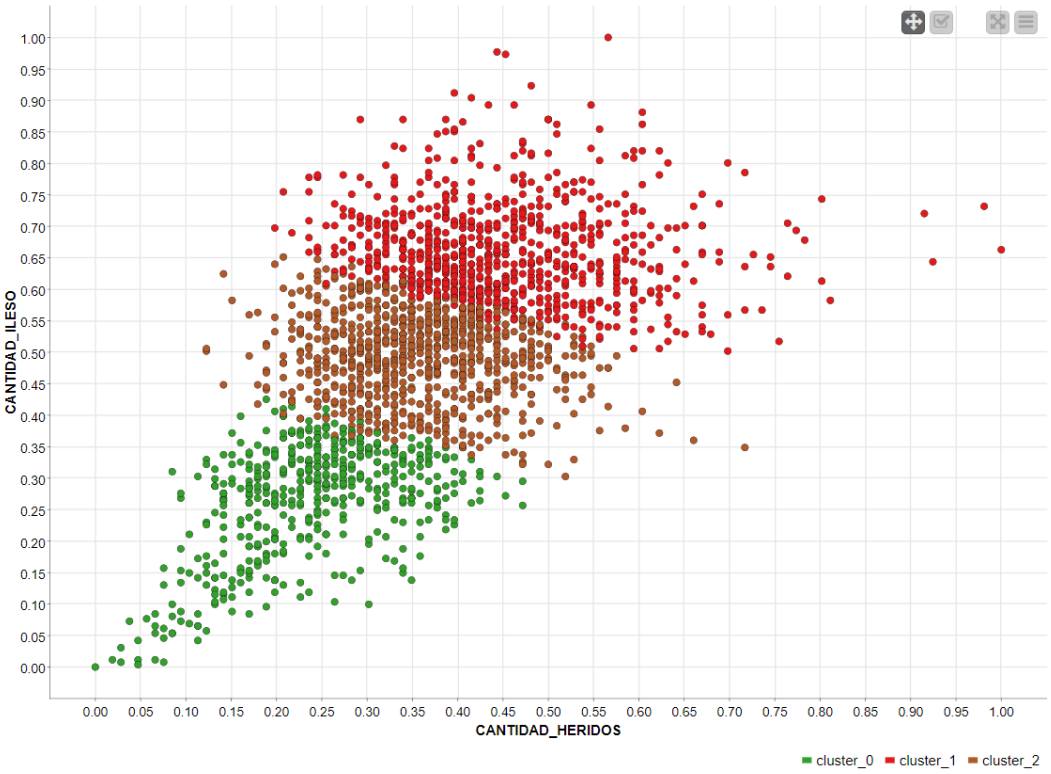


Figura 4-31. Distribución de la segmentación, en este caso, comparando las variables CANTIDAD_ILESO y CANTIDAD_HERIDOS

Se puede apreciar con base en la figura 4-31, que la segmentación es adecuada, ya que se puede diferenciar bien los diferentes grupos.

Por medio de un dendograma que se puede visualizar en la figura 4-32 también se pueden visualizar los clústers en dos o tres ideales:

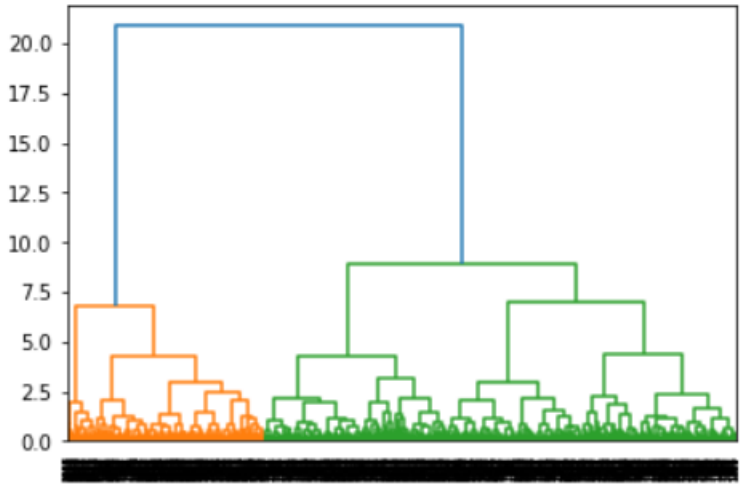


Figura 4-32. Dendograma - Bogotá

A continuación, con base en los centroides obtenidos, se puede definir 3 grupos bajo variables de interés claramente diferenciados por lo que se considera que los resultados de los 3 grupos son satisfactorios para poder describir la tendencia de cantidad de accidentes por zona.

Bogotá

Cluster\Variable	Frecencia accidentalidad	Nivel de víctimas	Nivel de Ilesos	Nivel de fallecidos	Nivel de vehículos involucrados	Nivel Hombres Involucrados	Nivel mujeres involucradas
0	Alta	Alta	Alta	Media - Baja	Alta	Alta	Media - Baja
1	Baja	Baja	Baja	Baja	Baja	Baja	Baja
2	Media-Baja	Media-Baja	Media-Baja	Baja	Media-Baja	Media-Baja	Baja

Tabla 4-33 Centroides de los clusters -Bogotá

El segmento 0 se caracteriza por tener un valor alto en la mayoría de las variables de interés a comparación del segmento 1 que cuenta un valor bajo en todas las variables de interés. Continuando con el análisis, para construir una adecuada visualización geográfica, se procedió a analizar y agrupar las localidades en algunas de los 3 segmentos con base en la cantidad de siniestros viales, el resultado es el siguiente:

Detalle por localidad

ZONA	cluster_0	cluster_1	cluster_2
KENNEDY	18387	2778	0
ENGATIVA	16085	5	2947
USAQUEN	13459	0	4271
SUBA	13209	0	4051
FONTIBON	11012	1487	2398
PUENTE ARANDA	6702	2425	3571
USME	0	3556	13
BOSA	926	7449	1
CIUDAD BOLIVAR	311	6807	1
TUNJUELITO	79	4698	1
RAFAEL URIBE URIBE	50	4629	2
SAN CRISTOBAL	48	4076	584
ANTONIO NARINO	2	3166	99
SUMAPAZ	0	5	1
LOS MARTIRES	255	548	4826
CHAPINERO	4650	0	6242
BARRIOS UNIDOS	2997	4	6338
TEUSAQUILLO	2807	10	6557
SANTA FE	104	57	4774
CANDELARIA	0	11	788

Tabla 4-34 Detalle clusters- Bogotá

Verde Zonas occidentales y norte ->Mayor accidentalidad

Morado: Centro de la ciudad -> Accidentalidad intermedia

Roja: Sur > Menor accidentalidad

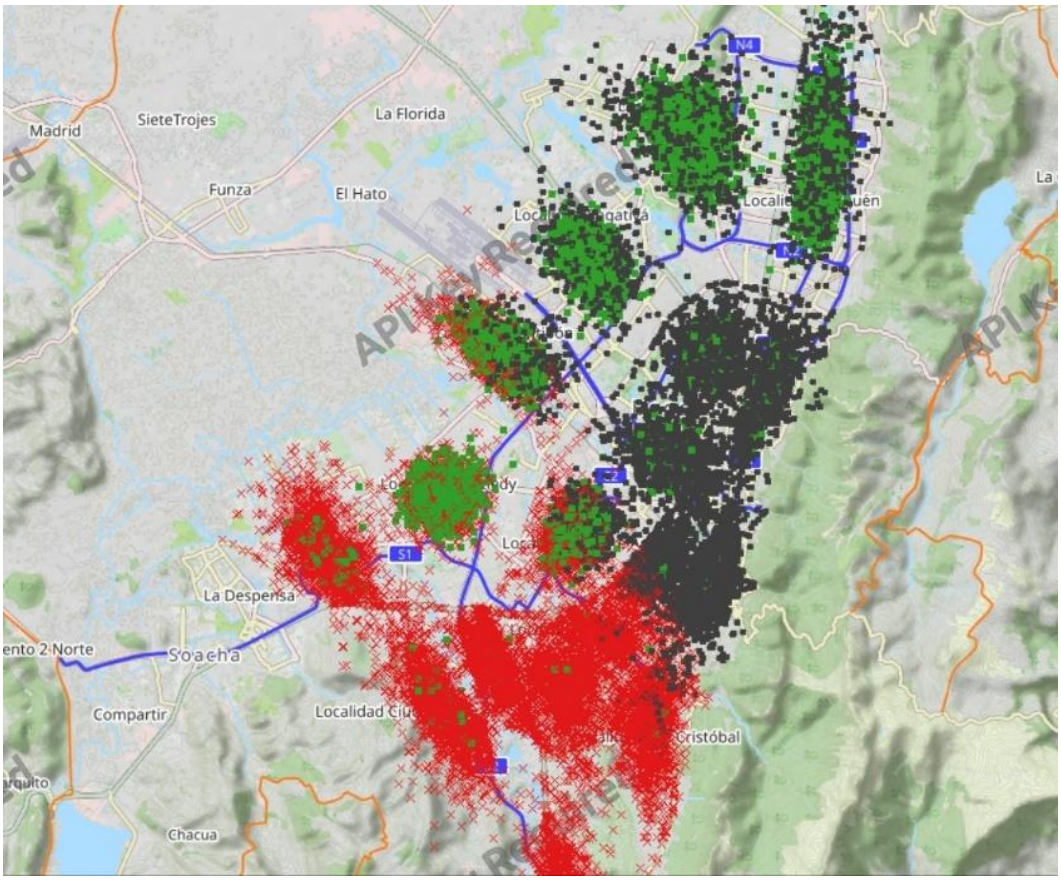


Figura 4-35. Clusterización accidentes - Bogotá

Esta segmentación nos permite ver donde se concentran los accidentes teniendo en cuenta la siniestralidad del sector. verde: zonas occidentales y norte, gris: centro de la ciudad y roja: Sur

Segmentación para Medellín:

Figura 4-36. gráfico de cohesión vs. K - Medellín

La figura 4-36, muestra que el valor de k óptimo sería de 2 pero se considera que puede llegar a ser un número insuficiente de grupos por lo que se consideró tomar un valor de k = 3.

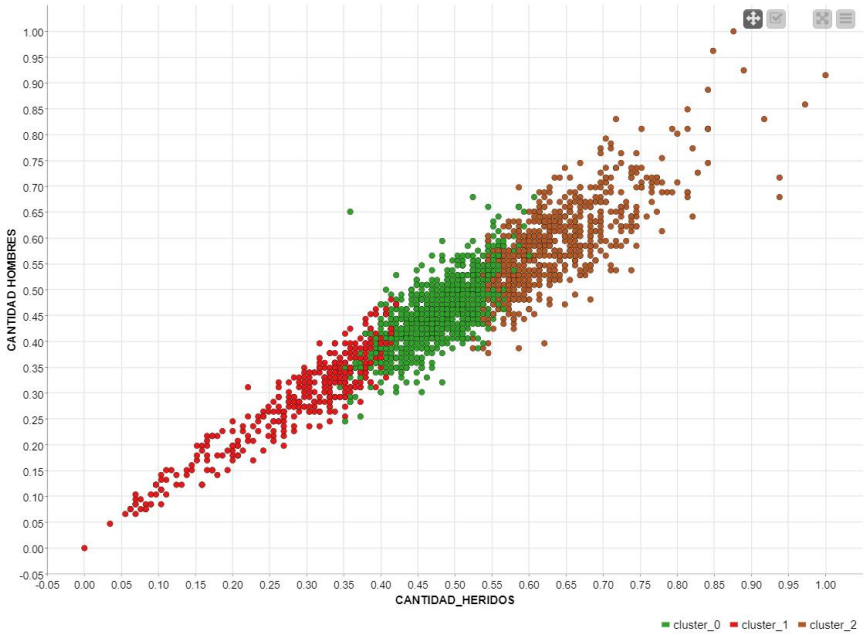


Figura 4-37. Distribución de la segmentación, en este caso, comparando las variables CANTIDAD_ILESO y CANTIDAD_HERIDOS - Medellín

Por medio de un dendograma que se puede visualizar en la figura 4-38 también se evidencian los clústers en dos o tres ideales

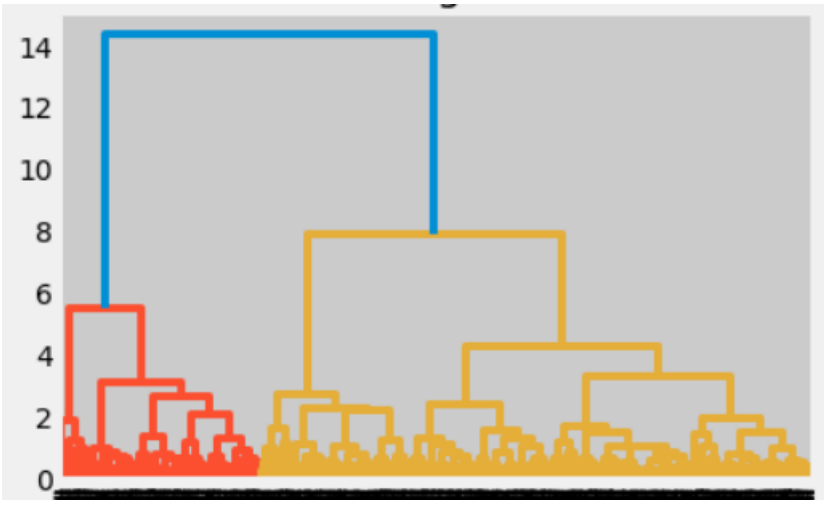


Figura 4-38. Dendograma - Medellín

A continuación, con base en los centroides obtenidos, se puede definir 3 grupos bajo variables de interés claramente diferenciados.

Medellín

Cluster\Variable	Frecencia accidentalidad	Nivel de víctimas	Nivel de fallecidos	Nivel Hombres Involucrados	Nivel mujeres involucradas
0	Alta	Alta	Media - Baja	Media - Alta	Media - Baja
1	Media-Baja	Media-Baja	Baja	Baja	Baja
2	Baja	Baja	Baja	Baja	Baja

Tabla 4-39 Centroides de los clusters - Medellín

El segmento 0 caracteriza por tener un valor alto en la mayoría de las variables de interés a comparación del segmento 2 que cuenta un valor bajo en todas las variables de interés. Continuando con el análisis, para construir una adecuada visualización geográfica, se procedió a analizar y agrupar las localidades en algunas de los 3 segmentos con base en la cantidad de siniestros viales, el resultado es el siguiente:

Detalle por comuna

ZONA	cluster_0	cluster_1	cluster_2
La Candelaria	45255	203	1830
Laureles Estadio	17283	1084	6350
Castilla	14154	0	8241
Buenos Aires	601	7316	951
El Poblado	2932	16831	0
Belen	2992	12253	0
Guayabal	3430	13269	0
Corregimiento de San Antonio de Prado	0	3573	0
Corregimiento de Altavista	0	446	7
Corregimiento de Santa Elena	0	406	134
Aranjuez	3405	1	9741
Robledo	5459	0	9400
Manrique	464	0	6547
Villa Hermosa	344	12	5710
La America	286	129	6887
Doce de Octubre	161	0	6100
San Javier	31	85	3797
Santa Cruz	14	0	3075
Popular	3	0	3373
Corregimiento de San Cristobal	0	5	2406
Corregimiento de San Sebastian de Palmitas	0	1	14

Tabla 4-40 Detalle clusters- Medellín

Verde: Centro ->Mayor accidentalidad

Rojo: Sur ->Accidentalidad intermedia

Morado: Norte ->Menor accidentalidad

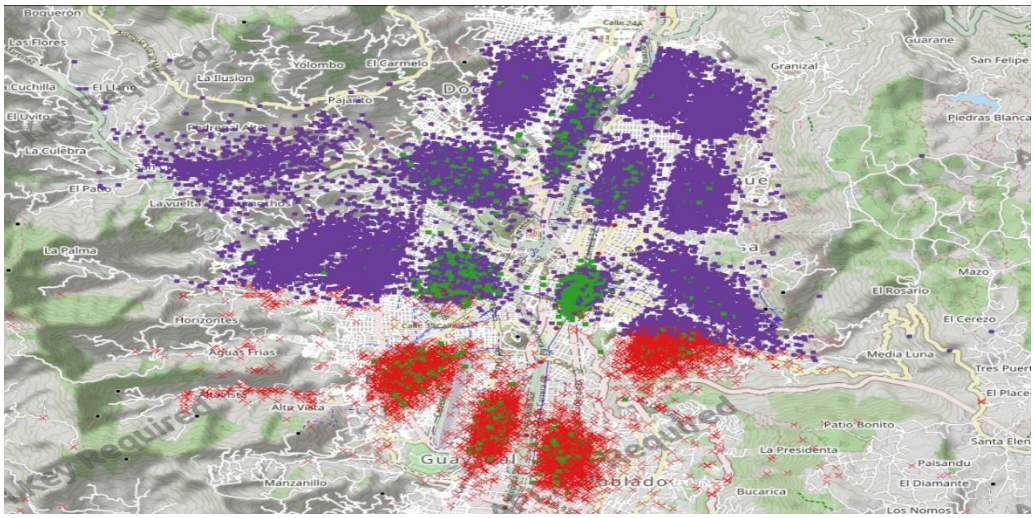


Figura 4-41. Clusterización accidentes - Bogotá

El comportamiento para Medellín se concentra en tres *clusters*: verde: Centro rojo: sur y morado: norte.

Segmentación para Barranquilla:

Para barranquilla, se quiere revisar si hay una posible caracterización de accidentes a pesar de su falta de información, puesto que para esta ciudad solo se tenían las siguientes variables:

CANTIDAD_HERIDOS
CANTIDAD_VICTIMAS

Por medio del método del codo se tiene que la Clusterización ideal esta entre dos y tres *clusters*:

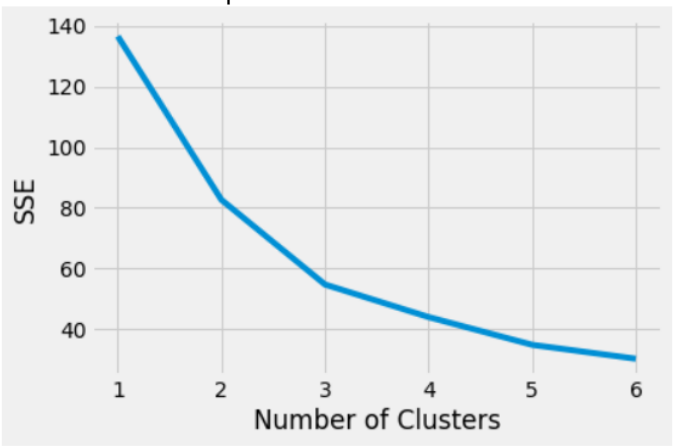


Figura 4-42. gráfico de cohesión vs. K - Barranquilla

Por medio de un dendograma que se puede visualizar en la figura 4-42 también se pueden visualizar los clústers en dos o tres ideales

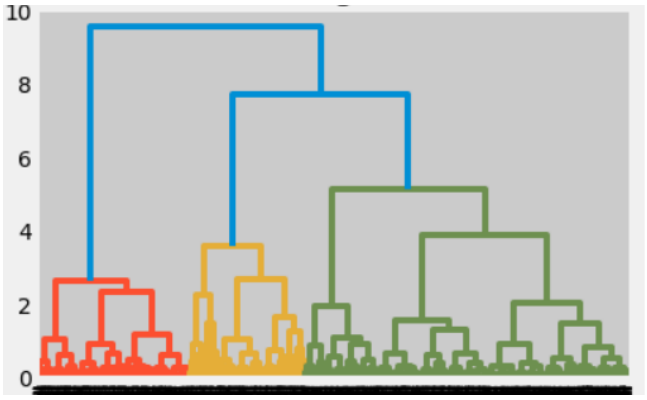


Figura 4-43. Dendograma – Barranquilla.

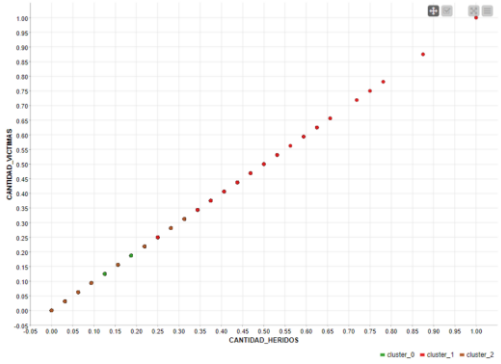


Figura 4-44. Distribución de la segmentación, en este caso, comparando las variables CANTIDAD_ILESO y CANTIDAD_HERIDOS - Barranquilla

Dado los resultados anteriores y debido a la escasa información disponible no es posible realizar y/o construir una segmentación adecuada para la ciudad de Barranquilla porque no está segmentada por zona. Por lo tanto, no se puede adaptar un *clustering* sobre una ciudad con tan poca información.

Resumiendo, esta sección del proyecto. El trabajo realizado para el proceso de segmentación por ciudad, utilizando la técnica del codo y otras herramientas de análisis de *clustering*, ha aportado contribuciones significativas en el campo de la seguridad vial y la gestión de accidentes de tráfico en Colombia. A través de este proceso, se ha logrado una comprensión más profunda de la siniestralidad vial en diferentes ciudades del país, lo que es fundamental para la toma de decisiones informadas por parte de las autoridades y planificadores urbanos.

Comenzando con el análisis de Bogotá, se observa que la técnica del codo sugirió inicialmente que un valor óptimo de "k" podría ser 2, lo que implicaría una segmentación en dos grupos. Sin embargo, se optó por un valor de 3 para obtener segmentos más detallados y representativos. Esta decisión

resultó en la identificación de tres grupos claramente diferenciados basados en variables críticas relacionadas con la accidentalidad vial. La caracterización geográfica de las localidades en Bogotá, que muestra zonas con alta, intermedia y baja siniestralidad, es una de las contribuciones más notables de este trabajo. Esta información es de gran utilidad para las autoridades locales, ya que les permite enfocar sus esfuerzos y recursos en áreas específicas que requieren una atención prioritaria en términos de seguridad vial.

En el caso de Medellín, el análisis de clustering siguió una estructura similar. Nuevamente, la técnica del codo sugirió un valor óptimo de "k" de 2, pero se optó por 3 *clusters* para una segmentación más detallada. Esto permitió la identificación de tres grupos claramente diferenciados en base a variables relacionadas con la siniestralidad vial. Además, al igual que en Bogotá, se realizó una caracterización geográfica de las comunas en Medellín, lo que proporciona información valiosa sobre las áreas con diferentes niveles de accidentalidad. Esta segmentación geográfica es esencial para la planificación de medidas específicas de seguridad vial en diferentes partes de la ciudad.

Sin embargo, en el caso de Barranquilla, la falta de información limitó la capacidad de realizar una segmentación adecuada. A pesar de los esfuerzos, la escasa disponibilidad de datos no permitió una agrupación efectiva de la siniestralidad vial en la ciudad. Esto subraya la importancia de contar con datos de calidad y cantidad suficiente para llevar a cabo análisis de clustering efectivos.

En cuanto a las contribuciones generales de este trabajo en comparación con el estado del arte, varias mejoras se destacan:

1. **Segmentación Detallada:** La segmentación por ciudad proporciona una visión detallada de la siniestralidad vial en diferentes zonas, lo que permite una comprensión más precisa de los patrones de accidentes en áreas específicas. Esta segmentación va más allá de los enfoques anteriores que podrían considerar las ciudades en su totalidad, lo que a menudo no permite la identificación de áreas críticas.
2. **Caracterización Geográfica:** La capacidad de caracterizar geográficamente las zonas con diferentes niveles de siniestralidad es una contribución importante. Esto facilita la identificación de áreas críticas que requieren intervención y planificación en materia de seguridad vial, lo que puede salvar vidas y reducir lesiones.
3. **Selección de K Óptimo:** La elección del número óptimo de *clusters* se basa en una técnica robusta como la del codo, lo que garantiza una segmentación apropiada y fundamentada en datos. Esto es una mejora significativa en comparación con enfoques anteriores que podrían seleccionar "k" de manera más arbitraria.
4. **Comparativo con el Estado del Arte:** Este trabajo se distingue por introducir mejoras sustanciales en el análisis de siniestralidad vial en ciudades colombianas (generalmente se concentran en una sola ciudad). En el estado del arte, es común encontrar enfoques que se centran en la descripción general de la siniestralidad en ciudades únicas. La contribución principal de este proyecto radica en la generación de una segmentación detallada y geográficamente relevante de varias ciudades en simultáneo, lo que permite una comprensión más profunda de la distribución de accidentes de tráfico en diferentes zonas y reduce tiempo de ejecución.

   Esta mejora en la segmentación geográfica permite una visión más específica de las áreas con diferentes niveles de siniestralidad, lo que es crucial para la planificación y la toma de

decisiones en seguridad vial. A diferencia de investigaciones previas que podrían abordar la seguridad vial de manera más generalizada, este trabajo identifica zonas críticas con alta siniestralidad y áreas con menor riesgo. Estos hallazgos son fundamentales para dirigir recursos y esfuerzos de manera más efectiva, implementando medidas preventivas y correctivas específicas donde más se necesitan.

4.3 Predicción

El papel crucial del modelo predictivo en la anticipación y gestión de la siniestralidad vial se convierte en un elemento central en la formulación de estrategias efectivas para mejorar la seguridad en nuestras calles. Este componente esencial de la investigación y gestión de la seguridad vial tiene como objetivo principal prever con precisión la cantidad de accidentes que podrían ocurrir en una ubicación específica en un período determinado.

Lo fascinante de estos modelos radica en su capacidad para aprender y adaptarse. Alimentados con datos históricos y la posibilidad de implementarlo a futuro en tiempo real, estos modelos pueden ajustarse de manera continua, lo que les permite mantenerse actualizados frente a los cambios en las condiciones del tráfico y del entorno. Esta adaptabilidad no solo mejora la precisión de las predicciones, sino que también proporciona una base sólida para la toma de decisiones informadas en materia de seguridad vial.

La práctica recomendada de llevar a cabo una calibración anual del modelo asegura su relevancia y precisión continuas. Al aprovechar la gran cantidad de datos recopilados a lo largo del tiempo, esta calibración garantiza que el modelo esté siempre actualizado y listo para ofrecer predicciones confiables. Este enfoque sistemático y continuo es fundamental para la implementación efectiva de políticas y medidas destinadas a proteger a los usuarios de las carreteras y minimizar los riesgos asociados con la movilidad.

Por último, en el contexto de la presente tesis, el análisis exhaustivo del papel y la efectividad de los modelos predictivos en la gestión de la seguridad vial proporciona una valiosa contribución al campo de estudio. Además, ofrece una base sólida para el desarrollo de recomendaciones prácticas y estrategias de intervención que tienen como objetivo mejorar la seguridad vial y reducir el número de accidentes en puntos estratégicos.

4.3.1 Variables para los modelos

Se generó la variable "gravedad_dia" esta variable clasifica los días según la incidencia de accidentes, diferenciando entre aquellos con muchos y pocos siniestros viales. En esta propuesta, "mucho" y "poco" son términos relativos que se utilizan para indicar una mayor o menor cantidad de accidentes en comparación con un punto de referencia. En este caso, se ha establecido un estándar para clasificar los días de la semana en términos de gravedad de accidentes. Los días de lunes a sábado se consideran "mucho" con un valor de 10, lo que implica que se espera una cantidad sustancial de accidentes en estos días en comparación con otros días de la semana. Por otro lado,

el domingo se clasifica como "poco" con un valor de 5, lo que indica que se espera que ocurran menos accidentes en los domingos en comparación con los días de lunes a sábado.

Esta categorización es valiosa porque simplifica la evaluación de la gravedad de los accidentes en función de los días de la semana. Permite una fácil clasificación y análisis de datos relacionados con la siniestralidad vial, lo que es esencial para la toma de decisiones en seguridad vial. Esta variable se basa en patrones de accidentes reales y proporciona información práctica que puede utilizarse para asignar recursos de manera más eficiente y tomar medidas preventivas.

Por otra parte, se crea la variable de *Gravedad zona*, esta variable clasifica las zonas según la incidencia de accidentes, diferenciando entre aquellos con muchos y pocos siniestros viales. La clasificación se fundamenta en una agrupación basada en percentiles, lo que permite asignar un grado de importancia a los incidentes según el día de la semana. En el caso de Medellín se clasifica de acuerdo en la figura 4-45:

ZONA	Total	Gravedad zona
Corregimiento de San Sebastian de Palmitas	15	0
Corregimiento de Altavista	453	0
Corregimiento de Santa Elena	540	0
Corregimiento de San Cristobal	2411	0
Santa Cruz	3089	0
Popular	3376	2
Corregimiento de San Antonio de Prado	3573	2
San Javier	3913	2
Villa Hermosa	6066	2
Doce de Octubre	6261	2
Manrique	7011	5
La America	7302	5
Buenos Aires	8868	5
Aranjuez	13147	5
Robledo	14859	5
Belen	15245	8
Guayabal	16699	8
El Poblado	19763	8
Castilla	22395	8
Laureles Estadio	24717	8
La Candelaria	47288	10

Figura 4-45. Distribución de zonas con base en la variable Gravedad zona - Medellín

Estos se seleccionaron de acuerdo con los percentiles de cómo se distribuye la variable de cantidad de accidentes por sector tomando valores de 0,2,5,8 y 10 que representan los niveles de accidental donde 0 es el más y 10 es el más alto.

Y para la ciudad de Bogotá se pusieron los pesos dependiendo al clúster que se categorizó en el capítulo anterior, se puede observar en la figura 4-46:

ZONA	Gravedad zona	Cluster
KENNEDY	10	Verde
ENGATIVA	10	Verde
USAQUEN	10	Verde
SUBA	10	Verde
FONTIBON	10	Verde
PUENTE ARAN	10	Verde
USME	5	Roja
BOSA	5	Roja
CIUDAD BOLI'	5	Roja
TUNJUELITO	5	Roja
RAFAEL URIBE	5	Roja
SAN CRISTOB.	5	Roja
ANTONIO NA	5	Roja
SUMAPAZ	5	Roja
LOS MARTIRE	2	Morada
CHAPINERO	2	Morada
BARRIOS UNI	2	Morada
TEUSAQUILLO	2	Morada
SANTA FE	2	Morada
CANDELARIA	2	Morada

Figura 4-46. Distribución de zonas con base en la variable Gravedad zona - Bogotá

Para el modelo predictivo se hizo uso de las variables las siguientes variables:

-CANTIDAD_ACCIDENTES
-CANTIDAD_HERIDOS
-CANTIDAD_ILESO
-CANTIDAD_MUERTOS
-CANTIDAD_VICTIMAS
-EDAD MEDIANA
-CANTIDAD HOMBRES
-CANTIDAD MUJERES
-GRAVEDAD_DIA
-GRAVEDAD_ZONA

Resultados de Modelos de Predicción:

Los modelos de predicción evaluados para la cantidad de accidentes por zona en Bogotá arrojaron resultados interesantes. Se utilizaron tres tipos de modelos: Regresión Lineal, *Random Forest* y *XGBoost*, y se evaluaron métricas clave para determinar su desempeño. Las métricas incluyen el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Porcentaje Absoluto Medio de Error (MAPE).

Ejemplo de los resultados:

BARRIO	ZONA	ANIO	MES	DIA	ACCIDENTES	Prediction
Niza	SUBA	2019	10	19	2	1
Niza	SUBA	2019	10	22	1	1
Niza	SUBA	2019	10	29	3	2
Niza	SUBA	2019	11	1	1	1
Niza	SUBA	2019	11	3	1	1
Niza	SUBA	2019	11	4	1	1
Niza	SUBA	2019	11	5	1	1
Niza	SUBA	2019	11	10	1	1
Niza	SUBA	2019	11	18	2	2
Niza	SUBA	2019	11	19	3	2

Figura 4-47. Resultado de las predicciones bajo modelo Random Forest - Bogotá

Bogotá

Modelo \ Métrica	MAE	MSE	RMSE	MAPE
Regresión Lineal	0.69	0.96	0.98	48%
Random Forest	0.5	0.88	0.94	21%
XGBoost	0.69	0.97	0.98	49%

Figura 4-48. Resultado de los modelos de predicción - Bogotá

De acuerdo con los resultados presentados en la figura 4-48, se tiene lo siguiente:

- **Regresión Lineal** presentó un MAE de 0.69, un MSE de 0.96, un RMSE de 0.98 y un MAPE del 48%.
- ***Random Forest*** destacó con un MAE de 0.5, un MSE de 0.88, un RMSE de 0.94 y un MAPE del 21%.
- ***XGBoost*** obtuvo un MAE de 0.69, un MSE de 0.97, un RMSE de 0.98 y un MAPE del 49%.

Basándonos en los resultados anteriores, se recomienda el modelo *Random Forest* para Bogotá, ya que obtuvo los valores más bajos en las métricas evaluadas. Es importante destacar que el objetivo era lograr un MAPE menor al 50%, un RMSE cercano a 1 y un MAE menor a 0.5. El modelo *Random Forest* cumplió con estos criterios y, por lo tanto, se considera la elección óptima para predecir la cantidad de accidentes por zona en Bogotá.

Para Medellín los resultados se presentan en la figura 4-49

Medellín

Modelo \ Métrica	MAE	MSE	RMSE	MAPE
Regresión Lineal	0.51	0.79	0.89	32%
Random Forest	0.35	0.60	0.77	14%

XGBoost	0.51	0.79	0.88	31%

Figura 4-49. Resultado de los modelos de predicción - Medellín

- **Regresión Lineal** arrojó un MAE de 0.51, un MSE de 0.79, un RMSE de 0.89 y un MAPE del 32%.
- ***Random Forest*** nuevamente se destacó con un MAE de 0.35, un MSE de 0.60, un RMSE de 0.77 y un MAPE del 14%.
- ***XGBoost*** obtuvo un MAE de 0.51, un MSE de 0.79, un RMSE de 0.88 y un MAPE del 31%.

Al igual que en Bogotá, el modelo *Random Forest* se destaca por su desempeño sobresaliente en Medellín. Cumplió con los criterios de un MAPE menor al 50%, un RMSE cercano a 1 y un MAE menor a 0.5, lo que lo convierte en la elección recomendada para predecir la cantidad de accidentes por zona en Medellín.

Los modelos no aciertan mucho en el caso de Barranquilla como se observa en la figura 4-50 puesto que no tenemos una distribución clara de donde ocurrieron los accidentes. Esto nos da una información muy escasa para poder hacer una predicción certera.

Barranquilla

Modelo \ Métrica	**MAE**	**MSE**	**RMSE**	**MAPE**
Regresión Lineal	4.51	31	5.60	48%
Random Forest	4.28	28	5.34	41%
XGBoost	4.85	36	6.01	56%

Figura 4-50. Resultado de los modelos de predicción - Barranquilla

En cuanto a los resultados de los modelos de predicción, se evaluaron tres tipos de modelos (Regresión Lineal, *Random Forest* y *XGBoost*) utilizando métricas clave (MAE, MSE, RMSE y MAPE). Estos modelos proporcionan una base sólida para la toma de decisiones informadas en seguridad vial. La elección del modelo óptimo para cada ciudad se basa en el rendimiento de estas métricas y se presenta claramente en la propuesta.

5 Visualizador

Finalmente, el proyecto culmina con el desarrollo de una herramienta de visualización que permite observar las diferentes etapas realizadas, incluyendo el análisis de datos y los resultados de los modelos implementados. El cual se puede encontrar en el siguiente vínculo:

Microsoft Power BI

Adentrémonos en la historia de seguridad vial en Bogotá, Medellín y Barranquilla, una narrativa que revela la importancia de las herramientas analíticas en la accidentalidad vial. Desde la identificación de patrones hasta la predicción de riesgos y la visualización de datos, estas herramientas ofrecen una vía prometedora para abordar los desafíos inherentes a la movilidad urbana. Al integrar la analítica en los procesos de toma de decisiones, las autoridades pueden mejorar significativamente la seguridad de los ciudadanos en las principales ciudades colombianas.

Imagínate en un viaje visual donde las cifras y los gráficos cobran vida, pintando un retrato vívido de la realidad de nuestras calles. Comenzamos este emocionante recorrido con un detallado análisis de las estadísticas de accidentes en Bogotá y Medellín. Aquí exploramos la cantidad de incidentes por año, la edad promedio de los afectados y la distribución de género en cada ciudad como, por ejemplo, el porcentaje de mujeres en accidentes en Bogotá esta entre 15% y 18% y para Medellín entre 30% y 33% lo cual es una gran diferencia entre los comportamientos de las dos ciudades y también lo semejante es que la mayor proporción de accidentes se encuentra en los hombres. Se puede observar en la figura 5.1 los siniestros que se han presentado en los últimos años y comparar sus características con la visualización cantidad de accidentes, los promedios de edad y cantidad de accidentes por ciudad, con la gráfica de cantidad de siniestros por mes donde se ve como para las dos ciudades en agosto y octubre se muestran picos de accidentalidad en todos los años es el comportamiento de cantidad de accidentes que muestra ciertas semejanzas en todos los años.

A través de esta lente analítica, descubrimos sorprendentes similitudes entre ciudades que, a primera vista, podrían parecer distintas. Revelamos cómo los patrones de accidentes transcurren de manera similar, incluso en entornos geográficos y culturales diferentes.

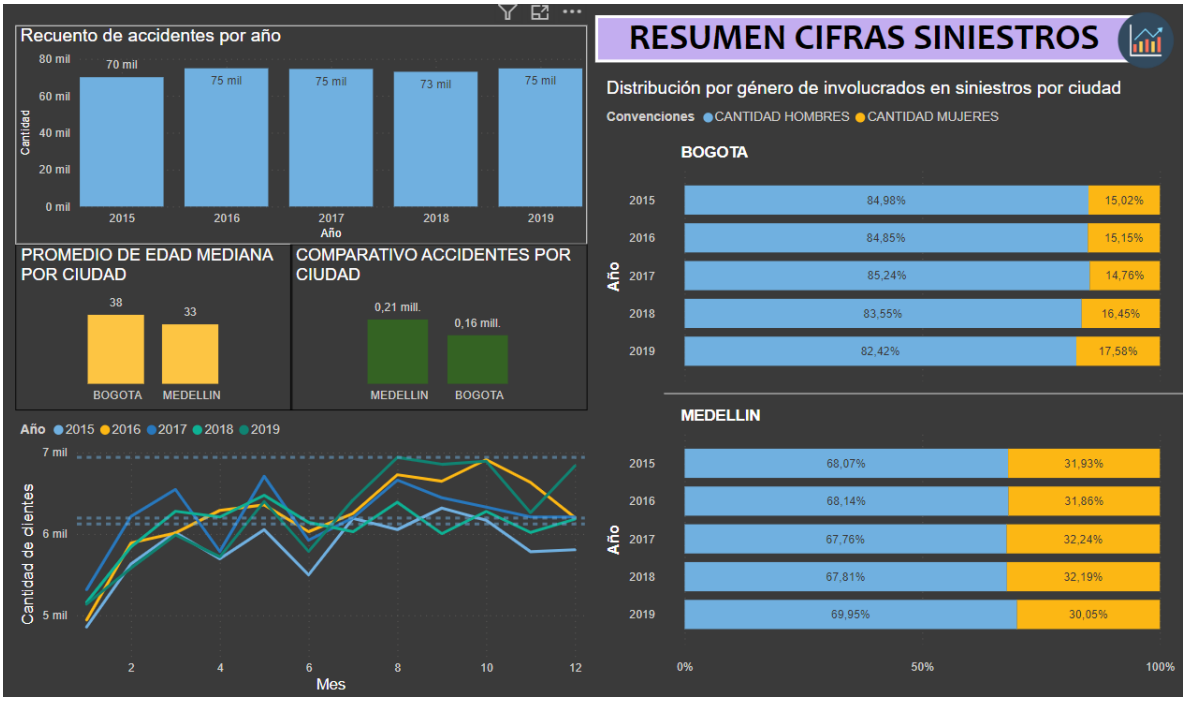
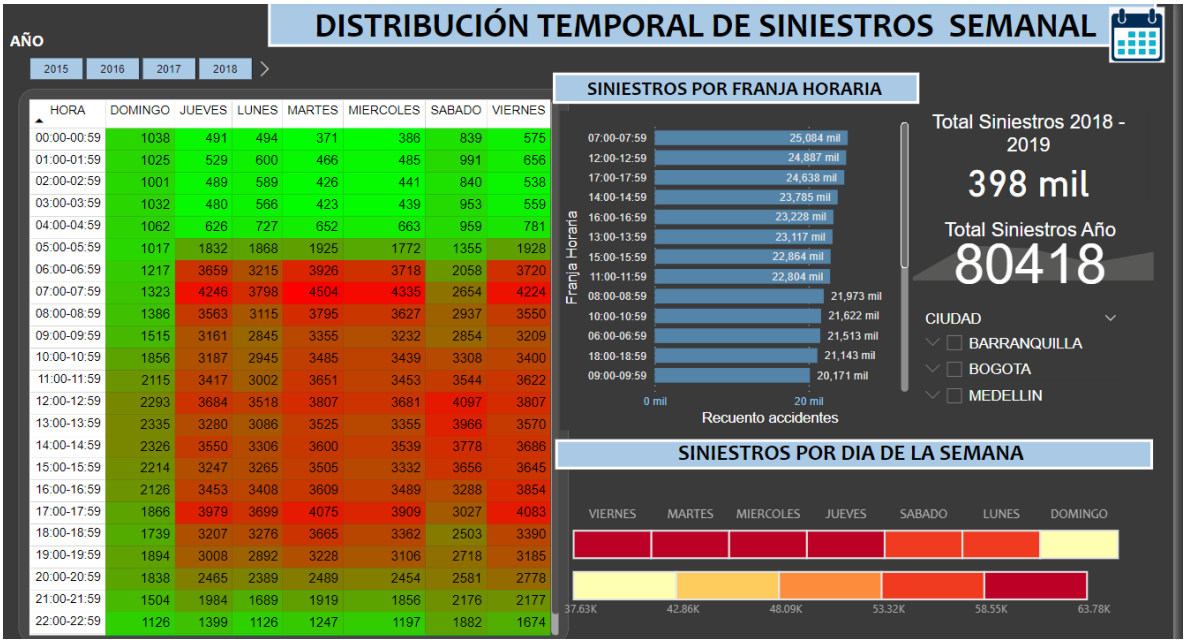


Figura 5-1. Resumen dashboard

Nuestro viaje continúa adentrándonos en la frecuencia semanal de los accidentes. Descubrimos en la figura 5.2 la hora óptima para evitar los peligros de la carretera y analizamos las tendencias a lo largo de los años, detectando cambios sutiles pero significativos en la seguridad vial, es de interés poder ver año a año si hubo algún cambio de tendencia y actualizarlo anualmente nos permite ver año a año para poder analizar estos cambios de tendencias, podría hasta detectarse faltantes de información o cambios gubernamentales en las vías, puesto que no tendría sentido que cambiaran mucho los datos de un año a otro si nada cambiara, esto se complementa con una grafica de frecuencia del día de la semana que nos cuenta en este caso el domingo es el día de color más claro es decir cuando ocurren menos accidentes en la semana, así que podría suponerse que si sales ese día de casa estas menos propenso a sufrir algún accidente, también existe una cifra global de cuantos datos ingresaron en los últimos años y la cantidad total de siniestros



HORA	DOMINGO	JUEVES	LUNES	MARTES	MIERCOLES	SABADO	VIERNES
00:00-00:59	1038	491	494	371	386	839	575
01:00-01:59	1025	529	600	466	485	991	656
02:00-02:59	1001	489	589	426	441	840	538
03:00-03:59	1032	480	566	423	439	953	559
04:00-04:59	1062	626	727	652	663	959	781
05:00-05:59	1017	1832	1868	1925	1772	1355	1928
06:00-06:59	1217	3659	3215	3926	3718	2058	3720
07:00-07:59	1323	4246	3798	4504	4335	2654	4224
08:00-08:59	1386	3563	3115	3795	3627	2937	3550
09:00-09:59	1515	3161	2845	3355	3232	2854	3209
10:00-10:59	1856	3187	2945	3485	3439	3308	3400
11:00-11:59	2115	3417	3002	3651	3453	3544	3622
12:00-12:59	2293	3684	3518	3807	3681	4097	3807
13:00-13:59	2335	3280	3086	3525	3355	3966	3570
14:00-14:59	2326	3550	3306	3600	3539	3778	3686
15:00-15:59	2214	3247	3265	3505	3332	3656	3645
16:00-16:59	2126	3453	3408	3609	3489	3288	3854
17:00-17:59	1866	3979	3699	4075	3909	3027	4083
18:00-18:59	1739	3207	3276	3665	3362	2503	3390
19:00-19:59	1894	3008	2892	3228	3106	2718	3185
20:00-20:59	1838	2465	2389	2489	2454	2581	2778
21:00-21:59	1504	1984	1689	1919	1856	2176	2177
22:00-22:59	1126	1399	1126	1247	1197	1882	1674

Figura 5-2. Distribución temporal de siniestros dashboard

Avanzamos hacia una exploración demográfica en la figura 5-3 más específica donde de manera porcentual se puede comparar mes a mes la cantidad de accidentes en esta claramente se puede diferenciar el mes con mas accidentes y el que se reportaron menos accidentes.

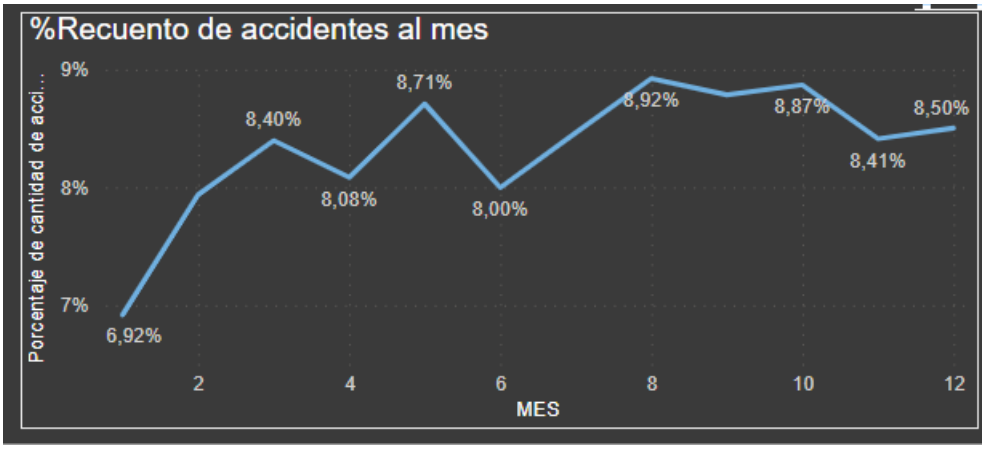


Figura 5-3. Porcentaje de accidentes por mes dashboard

También fue posible contrastar los tipos de accidentes como es la gravedad de los daños sucedidos en estos a pesar de que en barranquilla no hay muchos reportes se puede ver que la tendencia con Bogotá y Medellín se conserva en la figura 5-4.

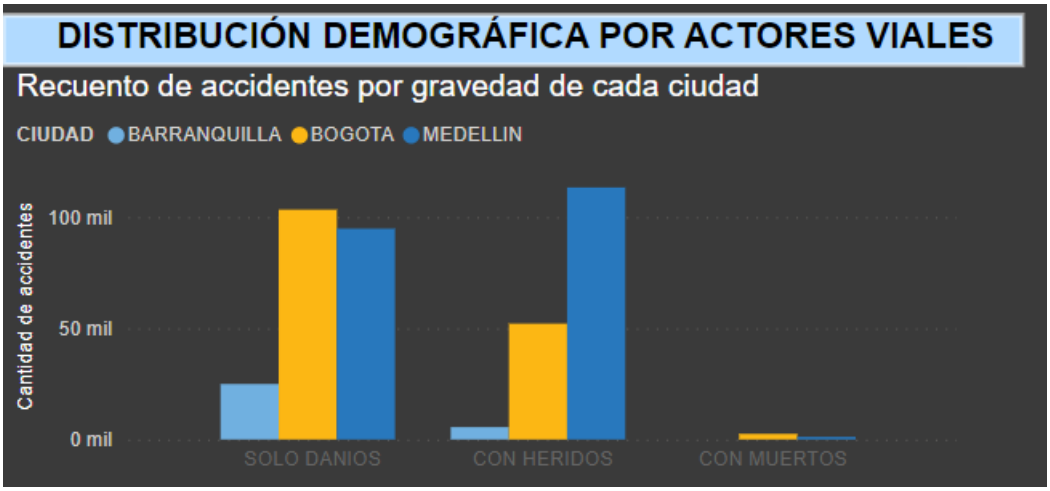


Figura 5-4. Distribución demográfica por tipo de accidente dashboard

Adicionalmente, igual que en el resumen se ve que la cantidad de mujeres es mucho menor en proporción que la cantidad de hombres, aun así, la información adicional que otorga la gráfica en la figura 5-5 es sobre la cantidad de personas implicadas en los accidentes y finalmente las proporciones de cuantos accidentes sucedieron en cada una de las ciudades donde a pesar de que Bogotá es una ciudad mas grande los reportes de accidentes de la ciudad de Medellín es mayor.

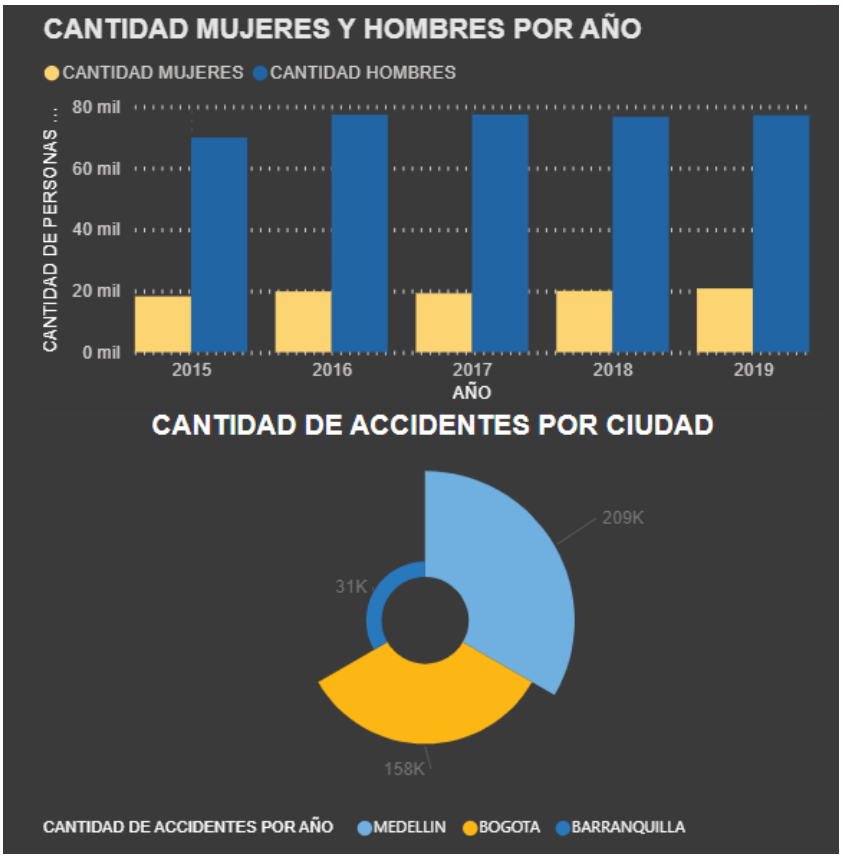


Figura 5-5. Distribución por cantidad de hombres y mujeres y comparación de cantidad de siniestros dashboard

Ahora en las características se puede filtrar por ciudad, en este análisis, se encuentran valiosas pistas que permiten entender mejor la dinámica de los accidentes de tránsito en las ciudades. Al filtrar los datos por ciudad, se ven detalles fascinantes. Por ejemplo, en la Figura 5-6 se observa como en Medellín, la Candelaria emerge como un punto crítico de incidentes, mientras que, en Bogotá, el barrio de Kennedy se destaca por su alta incidencia. Además, se observa que la mayoría de los accidentes ocurren en tramos de vía urbana, sugiriendo la necesidad de intervenciones específicas en áreas de alto tráfico y congestión. Sorprendentemente, muchos de estos accidentes no resultan en heridos, predominando los choques. Estos hallazgos nos brindan una oportunidad única para dirigir campañas de seguridad vial y fortalecer mecanismos de emergencia, enfocándolos en donde más se necesita. Pero esto es solo el comienzo de nuestra aventura en el vasto mundo de la ciencia de datos aplicada a la seguridad vial.

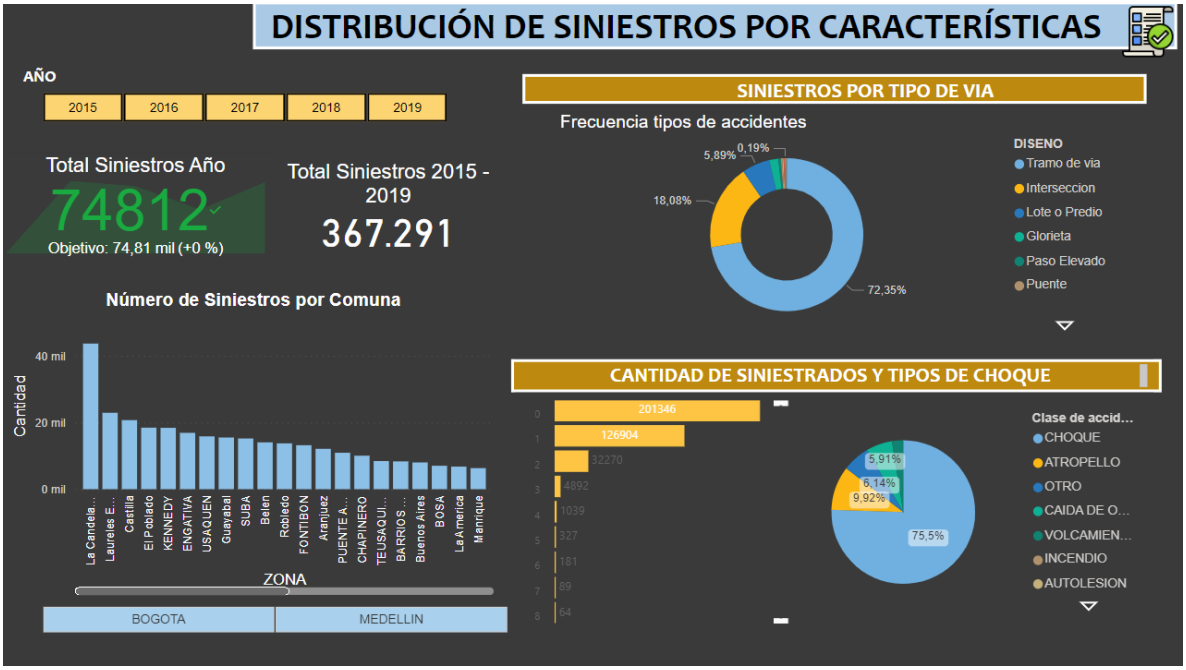


Figura 5-6. Distribución de siniestros por características dashboard

Seguido a lo anterior ahora se ve geográficamente en la figura 5-7 la distribución de cantidad de accidentes por localidades en el caso de Bogotá y las comunas en el caso de Medellín donde el amarillo son los puntos mas críticos y el rojo son puntos donde hay alta accidentalidad y el azul baja frecuencia de accidentalidad y en la parte inferior una animación donde ser dinámica se van visualizando los accidentes mes a mes como fueron reportados. Por ejemplo, para Medellín en la figura 5-8 se ve bien marcado como alta accidentalidad en el barrio candelaria, mientras que en Bogotá se ve mas marcado en puntos específicos.

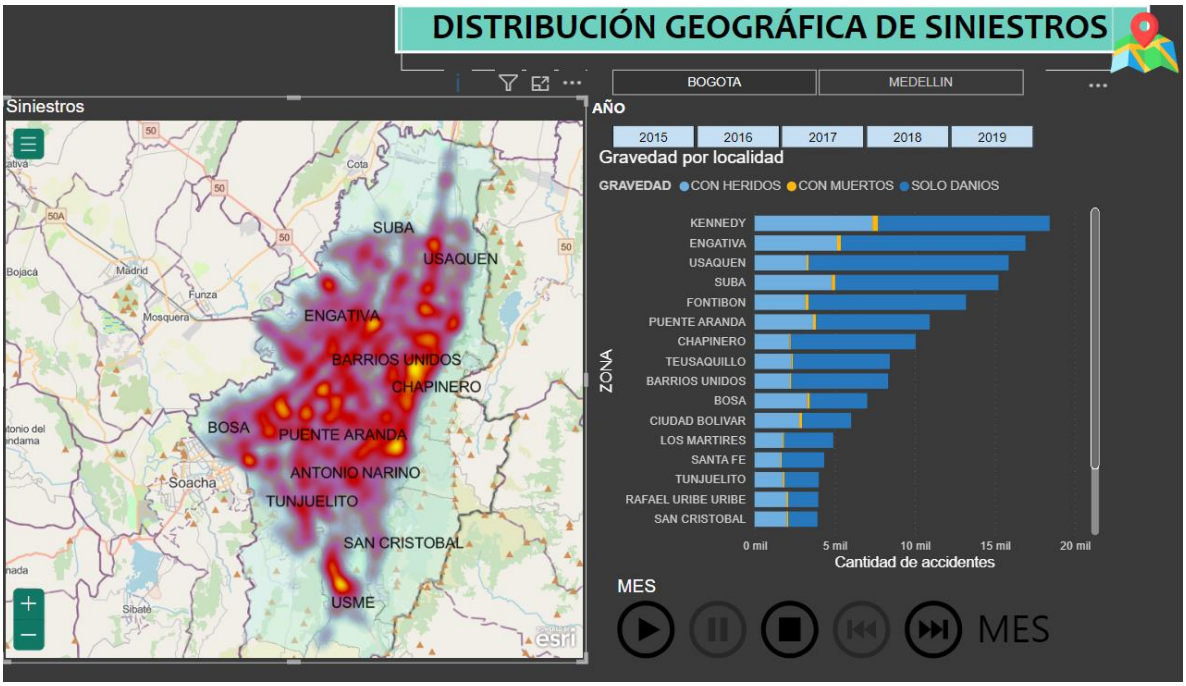


Figura 5-7. Distribución de geográfica de siniestros dashboard Bogotá

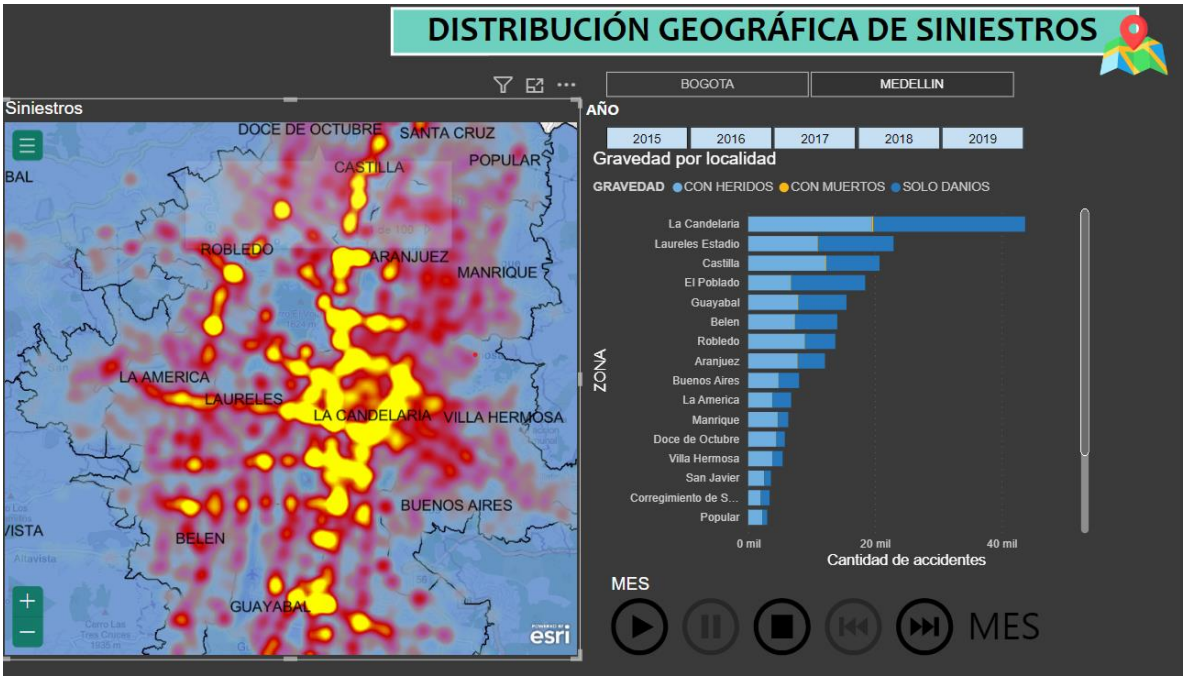


Figura 5-8. Distribución de geográfica de siniestros dashboard Medellín

Finalizando nuestro viaje en la figura 5-9, podemos encontrar las gráficas que se derivan de la predicción de los datos. En estas se encuentran las distribuciones con los accidentes que se predicen en el modelo como queda distribuido por varios factores, en un periodo determinado y la clasificación que queda a partir de los modelos en el mapa geográfico, donde es posible verlo para Bogotá y Medellín, al final no se ven los datos de Barranquilla debido a que no eran suficientes para los modelos. Con estos datos se pueden ver los patrones y conclusiones del estudio. En adición, para cada una de las ciudades se puede observar adicionalmente la distribución de los accidentes de acuerdo con el cluster asignado por él modelo de clusterización.

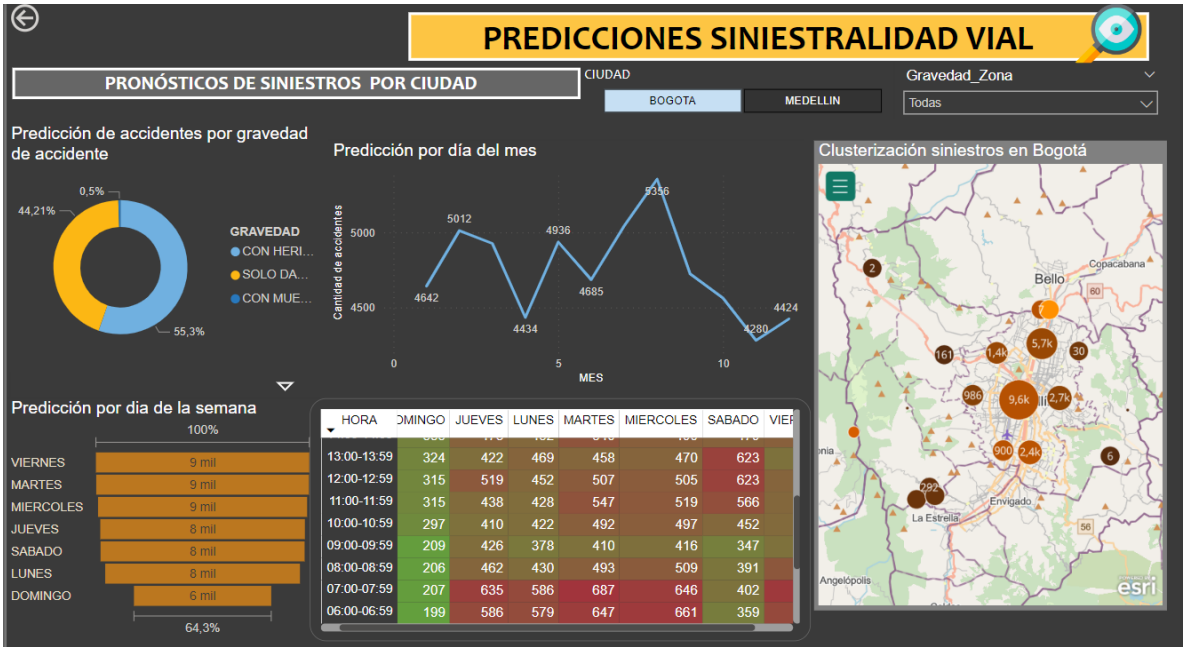


Figura 5-9. Resultados predicción dashboard

6 Conclusiones

- La construcción de una base de datos nacional de siniestralidad, con desglose por ciudades, establece un sólido fundamento para el éxito del proyecto. Esta plataforma integral facilita un análisis detallado y respalda decisiones informadas, contribuyendo significativamente a la comprensión y gestión efectiva de la siniestralidad a nivel nacional.

- La definición de criterios de clasificación se basó en una revisión exhaustiva de la literatura científica relacionada con la seguridad vial y la investigación de accidentes. Los criterios se seleccionaron cuidadosamente después de considerar las mejores prácticas y enfoques previamente utilizados en el campo. Esto asegura que nuestros criterios de clasificación estén respaldados por una sólida base teórica y científica, lo que aumenta su fiabilidad y utilidad para abordar cuestiones de seguridad vial.

- Aunque existían algunas investigaciones previas que abordaron la segmentación de la accidentalidad en zonas de ciudades, nuestro enfoque se destacó por su enfoque más integral y detallado. Hemos considerado factores adicionales y datos más actualizados, lo que nos permitió realizar una clasificación mucho más precisa y valiosa. Esto ayudará a las autoridades y planificadores urbanos a tomar decisiones informadas para mejorar la seguridad vial en áreas específicas de las ciudades.

- La implementación de un modelo predictivo basado en regresión lineal para la accidentalidad vial por periodos y ubicación geográfica emerge como un recurso esencial. Este enfoque no solo proporciona una herramienta efectiva para anticipar y gestionar riesgos, sino que también contribuye a la planificación estratégica de medidas preventivas.

- Nuestro trabajo incluyó una evaluación exhaustiva de modelos avanzados de predicción de siniestralidad vial. A través de esta comparación, pudimos demostrar de manera sólida que los modelos seleccionados superaron significativamente a los modelos que se han venido usando en términos de precisión y capacidad predictiva. Esta contribución destaca la eficacia de nuestra metodología y proporciona a las ciudades de Colombia una base sólida para mejorar sus estrategias de seguridad vial

- Nuestro tablero de resultados, que se encontrará disponible una vez se apruebe este proyecto, es una herramienta interactiva que facilita la visualización y comprensión de los resultados de nuestra investigación. Esto garantiza que no solo los investigadores, sino también los encargados de la toma de decisiones y el público en general puedan acceder y entender fácilmente los resultados de nuestro trabajo. Esto contribuye a la transparencia y a la difusión efectiva de los hallazgos de nuestra investigación.

7 Líneas Futuras

- Con ayuda de la base de datos construida en las ciudades observadas, se pueden identificar patrones relacionados con la cantidad de accidentes que ocurren en cada zona. Estos patrones pueden ser analizados y utilizados como base para implementar acciones gubernamentales enfocadas en mejorar la seguridad vial. Al tener en cuenta el estado de las carreteras y el flujo vehicular con el que cuenta la ciudad, es posible desarrollar estrategias que reduzcan la incidencia de accidentes y protejan la vida de los ciudadanos. Por ejemplo, si se detecta que ciertas zonas de la ciudad tienen una alta concentración de accidentes debido a la presencia de baches o calles en mal estado, se podrían destinar recursos para mejorar la infraestructura vial en esas áreas específicas. Además, al considerar el flujo vehicular, se pueden implementar medidas como la instalación de semáforos, señalización adecuada o la creación de vías exclusivas para reducir la congestión y prevenir colisiones. Estas acciones gubernamentales, respaldadas por el análisis de los patrones de accidentes, buscan garantizar la seguridad de los conductores, peatones y ciclistas, fomentando una convivencia vial más segura y eficiente en las ciudades.

- Los patrones de accidentes en Bogotá y Medellín son importantes para las autoridades gubernamentales, ya que permiten direccionar estrategias específicas de prevención y mejorar la seguridad vial en las áreas identificadas como más propensas a la ocurrencia de accidentes. Por lo cual se hace necesario contactar a las autoridades viales de cada ciudad con el fin de presentar los resultados del proyecto.

- Agregar variables relacionadas con las predicciones meteorológicas de acuerdo con la zona de las ciudades a los modelos de predicción y decidir si impacta o no en el comportamiento de siniestralidad vial.

- La utilización efectiva del Big Data será un recurso invaluable para la gestión de la seguridad vial en la ciudad en el futuro próximo. El punto de partida será la recolección diaria de datos en tiempo real, tarea que se llevará a cabo gracias a la recopilación activa que realizará la alcaldía. Esta información, capturada en el flujo continuo de la vida urbana, abarcará una amplia gama de variables que serán de gran utilidad para comprender y abordar los desafíos de seguridad en las carreteras. Para aprovechar al máximo esta valiosa fuente de datos, se propondrá la implementación de un proceso de Extracción, Transformación y Carga (ETL) automatizado. Este proceso automatizado garantizará la eficiencia en la recopilación y transformación de datos, preparándolos de manera óptima para su integración en el modelo de predicción. Será esencial que este modelo se ajuste de manera anual, permitiendo así su adaptación continua a medida que evolucionen las condiciones y tendencias en la ciudad. Esta práctica asegurará que las predicciones generadas reflejen con precisión la realidad dinámica de la seguridad vial. Además de mantener el modelo actualizado, se propondrá la creación de gráficos automáticos que proporcionarán una representación visual clara de la distribución de datos y las tendencias emergentes. Estos

gráficos no solo servirán como herramientas de análisis, sino que también actuarán como alertas tempranas para identificar cambios drásticos en las tendencias, lo que permitirá una respuesta rápida y efectiva ante posibles problemas de seguridad vial en áreas específicas de la ciudad.

8 Bibliografía

- Abdi, H., & Williams, L. J. (2010). *Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.*
- Agencia Nacional de Seguridad Vial (ANSV). Informe Anual de Siniestralidad Vial. Bogotá, Colombia: ANSV, 2020.
- Aggarwal, C. C. (2015). *Data Mining: The Textbook. Springer.*
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees. CRC Press.*
- Daraei, A., et al. (2021). *Factors influencing the bicycle safety perception: A case study of Tehran, Iran. International Journal of Sustainable Transportation, 15(1), 63-78.*
- Departamento Nacional de Planeación. (2018). Plan Nacional de Seguridad Vial 2018-2022 (p. 127). Bogotá, Colombia: Editorial Nacional.
- Ertel, W. (2017). *Introduction to Artificial Intelligence: Methods and Applications. Springer.*
- Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media.*
- Gutierrez-Osorio, C., & Pedraza, F. (2019). *Análisis de distribución y clasificación de accidentes de tráfico: caso Bogotá. Enfoque UTE, 10(2), 119-135.*
- Gutierrez-Osorio, C., & Pedraza, F. (2020). *Predictive modelling in road safety: a case study of Bogotá, Colombia. International Journal of Computational Science and Engineering, 22(1), 107-116.*
- Heumann, C., Schomaker, L., & Shalabh. (2016). *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R. Springer.*
- Hoffmann, H. (2010). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. Springer.*
- Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques. Springer.*
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
- Jolliffe, I. T. (2002). *Principal Component Analysis. Springer*
- Le Roux, B., & Rouanet, H. (2010). *Multiple Correspondence Analysis. SAGE Publications.*
- Moine, F. R., Letham, B., & Liaw, A. (2011). *Practical Data Science with R. Manning Publications.*
- Novkovic, M., Arsenovic, M., Sladojevic, S., Anderla, A., & Stefanovic, D. (2017). Data science applied to extract insights from data-weather data influence on traffic accidents. INFOTEH-JAHORINA, 16, 387-392.

- Parsa, H., Chauhan, K., Taghipour, S., Derrible, S., & Mohammadian, A. (2019). *Deep learning-based real-time accident detection in smart cities. Journal of Transport Geography, 79, 102466.*

- Pérez, R., et al. (2018). *Modelo de ciencia de datos aplicado a la evaluación del desempeño de estudiantes de educación superior en carreras de tecnología. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología, 26, 99-112.*

- Sohn, S. Y., & Lee, J. (2003). *A new approach to traffic accident analysis: The application of the Bayesian networks. Accident Analysis & Prevention, 35(3), 403-415.*

- Sullivan, J. M. (2017). *Introduction to Uncertainty Quantification. Springer.*

- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining. Pearson.*

- Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., & Lauro, C. (2005). *PLS Path Modeling. Computational Statistics & Data Analysis, 48(1), 159-205.*

- Yuan, Z., Zhou, J., & Yang, L. (2018). *Traffic flow prediction based on deep learning: A survey. IEEE Transactions on Intelligent Transportation Systems, 19(11), 3686-3704.*

9 Anexos.

Anexo 1: Código construido para realizar la asignación de los códigos de accidente por localidad /comuna.

```
Código para enriquecer la información en caso que la localidad esta vacia
```

```
import pandas as pd
import numpy as np
import math
from matplotlib import pyplot as plt
import geopandas as gpd
import time
import matplotlib as mpl
import matplotlib.dates as mpdates
from pandas import DataFrame

from shapely.geometry import shape,Point

path ='C:/Users/yerit/OneDrive/Documentos/uni/aticulo para proyecto grado/datos/'
base=pd.read_excel('C:/Users/yerit/OneDrive/Documentos/uni/aticulo para proyecto grado/datos/BASE_CONSOLIDADA.xlsx',sheet_name='Base')

%matplotlib inline
```

```
poligonos=gpd.read_file(path+'unidad-de-planeamiento13.geojson',driver='GeoJSON')
```

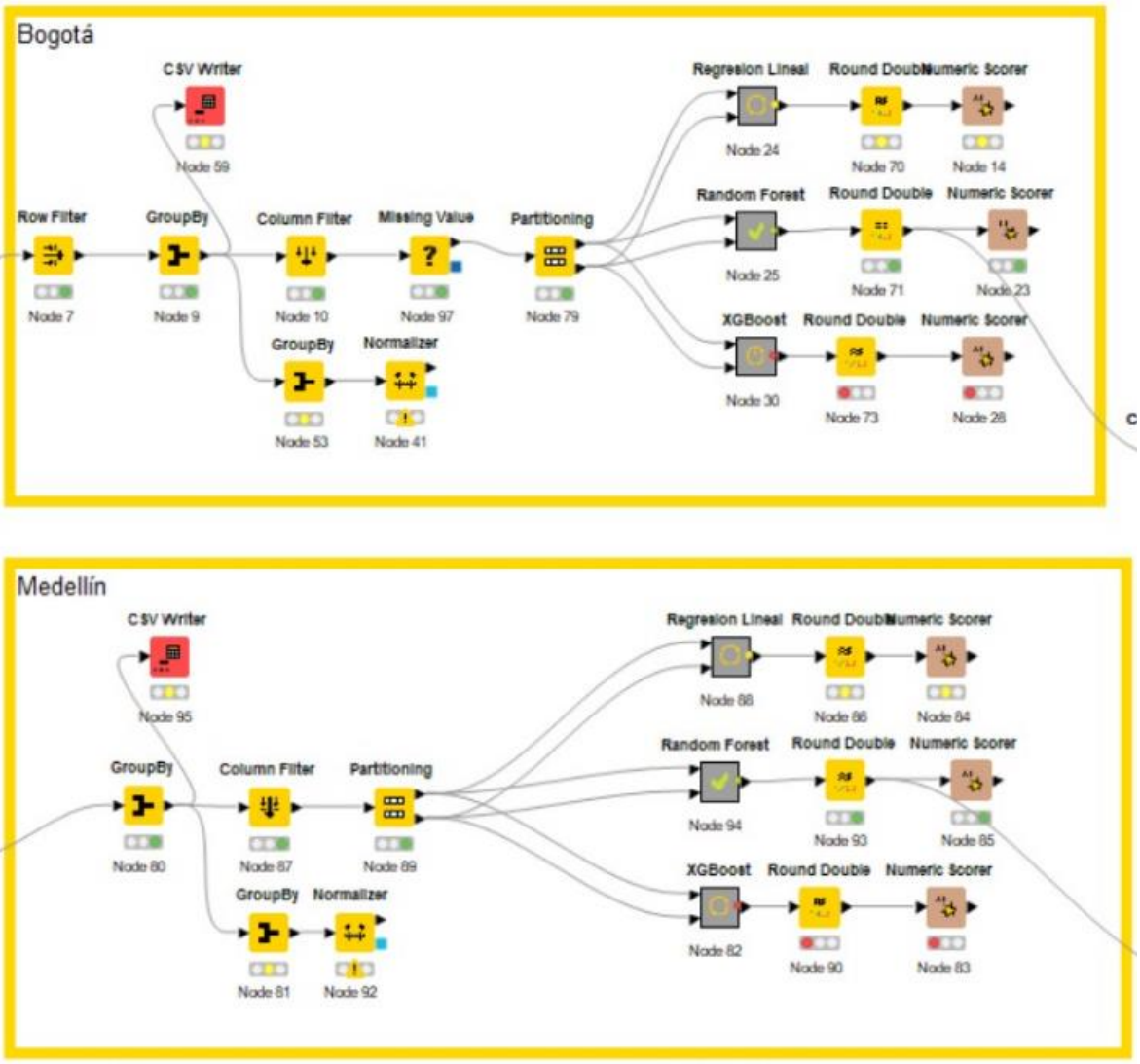
```
POLIGONO1=poligonos['geometry'][1]
if POLIGONO1.contains(Point(-74.090924,-74.090924)):
    print("Si")
else:
    print("No")
```

```
base2=pd.DataFrame(base[base["CIUDAD"]=="BOGOTA"])
base2.count()

len(base2)
```

```
##Aregla las variables iniciales para cada punto

x_1=base2["LATITUD"]
y_1=base2["LONGITUD"]
z_1=[[["CODIGO_ACCIDENTE"],["Zona"]]]
type(z_1)
j=0

while j<len(base2)  :
    point=Point(y_1[j],x_1[j])
    a=0
    u=0
    #Verifica cada poligono para ver si contiene el punto
    while u<len(poligonos['uplnombre']) and a==0 :
    #Crear la forma del polígono
        polygon =shape(poligonos['geometry'][u])
        #Condicion: Si el punto esta dentro del polígono se agrega el nombre de la localidad
        if polygon.contains(point):
            a=1
            z_1.append([base2['CODIGO_ACCIDENTE'][j],poligonos['uplnombre'][u]])
        else:
            a=0
        u=u+1
    j=j+1

#revisar un punto
print(z_1[0])
```

```
df_sector=DataFrame(z_1[1:],columns=['CODIGO_ACCIDENTE','Zona'])
df_sector.to_csv('sector.csv')
```

Anexo 2: Flujo construido en KNIME para la comparación de los distintos modelos.



Anexo 3: Vista previa del dashboard construido.