RUTEO TSP PARA TIENDAS MINORISTAS UTILIZANDO MACHINE LEARNING BASADA EN LA CARACTERIZACIÓN Y PREDICCIÓN DE SEVERIDAD DE ACCIDENTES VIALES EN BOGOTÁ-COLOMBIA

LIZETH NATALIA MAYORQUIN CUESTAS

DIRECTOR:

JAIRO RAFAEL MONTOYA TORRES

UNIVERSIDAD DE LA SABANA

FACULTAD DE INGENIERÍA

MAESTRÍA EN ANALÍTICA APLICADA

NOVIEMBRE DE 2023

PÁGINA DE ACEPTACIÓN

Dr. Carolina Castañeda Pérez
Decide Soluciones

Dr. Alfonso Tullio Sarmiento Vásquez
Universidad de La Sabana

Gonzalo Enrique Mejía Delgadillo
Director de la Maestría en Analítica Aplicada

DEDICATORIA Y AGRADECIMIENTOS

A mi amada familia que ha sido mi mayor fuente de inspiración, agradecimiento y amor incondicional.

A mi amado papá que desde el cielo ilumina mi camino y me impulsa a no rendirme jamás.

A Jairo Montoya, mi tutor quien es un pilar fundamental, por guiarme con sabiduría y darme la oportunidad de trabajar a su lado en este proyecto de la mano con la Universidad de La Sabana

Tabla De Contenido

Tabla De Ilustraciones

Tablas

1. RESUMEN

El tráfico, como uno de los desafíos cotidianos a los que se enfrentan a diario tanto gobiernos como ciudadanos, se ve agravado por la accidentalidad vial; convirtiendo ésta, en una de las principales causas de congestión vehicular. Este problema tiene consecuencias adversas en los sistemas de costos empresariales, especialmente en metrópolis como Bogotá D.C., clasificada como la cuarta con mayor tráfico a nivel mundial según el Traffic Index (Becerra, 2022). Por este motivo, el presente estudio toma el caso de la ciudad de Bogotá y analiza registros cualitativos/categóricos sobre accidentes viales ocurridos entre el año 2015 y el año 2022, con miras a desarrollar tres enfoques complementarios entre sí ante el problema planteado, para finalmente entregar un modelo de ruta de transporte urbano para la distribución de una muestra de tiendas de una empresa. A partir de técnicas estadísticas descriptivas e inferenciales, el primer enfoque determinó que las variables de tiempo como día de la semana, mes y horario, junto con variables de geografía como localidad y dirección generan un sesgo o mayor influencia en la frecuencia, la probabilidad y la severidad de sufrir un accidente consecuente. El segundo enfoque utiliza un modelo de asignación de Centros de Distribución (CEDI's) a los puntos de entrega seleccionados, estandariza los nodos de las posibles rutas y utiliza técnicas de Machine Learning para predecir la severidad de accidentes en una hora de congestión vial masiva, basándose en las variables de mayor representación encontradas previamente. Estas dos fases convergieron en la última fase referente al diseño de una solución estática de ruteo tipo Traveling Salesman Problem (TSP) combinada con una ponderación probabilística de ocurrencia de un accidente. Algunas de las técnicas seleccionadas se fundamentan con un análisis de la literatura a partir de las propuestas disponibles en la base de datos de Scopus según las coincidencias de las necesidades del presente estudio. Con las bases de este análisis y los resultados obtenidos, se proveen elementos y recomendaciones para el ruteo de alguna muestra de tiendas de una empresa tipo Hard Discount, junto con algunas

sugerencias de trabajo en futuras investigaciones.

Palabras Clave: Severity Traffic Accident Prediction, Traffic Accident Factors, Route Design, Traveling Salesman Problem, Multiple Correspondence Analysis.

2. RESUMEN GRÁFICO

Ilustración 1.Resumen gráfico

Para el proceso investigativo se implementó la metodología ágil conocida como CRISP-DM, caracterizada por su enfoque iterativo y colaborativo. Este marco de trabajo estructurado facilitó la exploración, modelado y evaluación de datos en mejora continua, asegurando así la eficiencia y precisión en cada etapa del análisis. Un resumen visual de esta metodología se presenta a continuación en el siguiente gráfico.

3. INTRODUCCIÓN

Los tiempos empleados entre trayectos, los tumultos y la inseguridad, son algunos de los grandes desafíos de movilidad que tienen las metrópolis del mundo (Portafolio, 2019). La capital colombiana (Bogotá D.C.), no es una excepción, según Becerra (2022) para 2021 Bogotá se posicionó como la cuarta ciudad con mayor tráfico vehicular a nivel mundial, donde un trayecto puede durar hasta 55% más del tiempo normal. La justificación de este acontecimiento surge por varios factores influyentes, en los que destaca el clima, la mala infraestructura, la persistencia progresiva de las ventas de automóviles y motocicletas; a lo que se suma un nuevo plan consolidado de ordenamiento territorial indicado por Vásquez (2021), que prevalecerá durante 10 años más y posiblemente continuará aumentando los niveles de tráfico actuales en la ciudad, según lo indica el diario DNP (2020) y Semana (2022). En consecuencia, no solo se verá impactada la población y el estado colombiano, sino que también las empresas experimentarán repercusiones negativas en sus estructuras financieras, como el incremento de los costos laborales debido a que, ante un aumento en el tiempo de trabajo por la espera logística en el tráfico, las horas extras del personal aumentarán de manera proporcional; de igual forma, como indica Bernal (2022), este tiempo también afecta el bienestar de sus empleados junto con su productividad.

En adición, la OMS[1] (2022) interpreta los accidentes de tránsito como uno de los componentes principales que repercuten negativamente en la movilidad metropolitana, en la vida de los usuarios y por ende afecta el ODS[2] #3 que busca garantizar una vida sana y de bienestar. Ante este panorama, se seleccionó la accidentalidad vial como tema determinante de la congestión a lo largo del presente documento.

A nivel académico existen numerosos estudios que han evaluado

[1] OMS: Organización Mundial de la Salud
[2] ODS: Objetivo de Desarrollo Sostenible

independientemente factores influyentes en la accidentalidad vial, tanto internacionalmente como nacionalmente. Sin embargo, resulta inquietante observar que en Colombia hay una cantidad de artículos de investigación limitados en temas de accidentalidad vial para los últimos 5 años, a pesar de que la cantidad de periódicos en línea que hablan sobre estos temas son relativamente considerables (La República, Semana, GOV.CO, El Tiempo, etc.). Esto sin mencionar que, en el último año, el diario El Tiempo (2023) indicó que un estudio llevado a cabo por la Universidad Nacional Colombiana determinó patrones vitales de accidentalidad vial en Bogotá relacionados a tiempo y geografía.

Por otro lado, la revisión literaria indica la existencia de investigaciones internacionales que abordan técnicas de predicción de accidentes y estrategias de optimización de rutas basadas en factores como el tráfico, las distancias o los costos. Sin embargo, se destaca la ausencia de documentos que apliquen estas técnicas dentro de un modelo de ruteo considerando la probabilidad de severidad de accidentes predicha mediante Machine Learning a un tramo de ruta propuesto. Por lo tanto, no se ha identificado ningún documento a nivel global o local que pueda considerarse fácilmente replicable. Ahora bien, a pesar de que actualmente existen herramientas como Google Maps, Waze y aplicaciones similares que sugieren rutas alternas según las condiciones de tráfico, ninguna de ellas ofrece una entrada dinámica en el orden de los puntos a visitar. Es decir, si se proporcionan tres puntos en un orden específico, estas aplicaciones seguirán esa secuencia sin reorganizar los puntos según la menor distancia o tiempo de recorrido.

Considerando lo anteriormente expuesto, el estudio se ejecutó con información de tipo Open Access almacenada hasta el año 2022 de 2 bases de datos complementarias entre sí, extraídas de la página de la Secretaría Distrital de Movilidad (SDM, 2022) (Accidentes (221.909 registros) y Lesiones (114.347 registros)). Para ello se ejecutaron las 3 etapas que caracterizan la analítica de datos (Descripción, Predicción y Prescripción) teniendo presente la carencia de

artículos o modelos replicables de investigación, dando como resultado final del estudio el diseño de un algoritmo de rutas de distribución de productos tipo TSP con base en la caracterización y predicción del nivel de severidad de la accidentalidad vial urbana. El modelo TSP está planteado en un escenario ideal que no contempla restricciones de movilidad de camiones o aspectos similares.

Con dicho modelo, el objetivo es estimar mejor el tiempo y distancia de ruta considerando únicamente accidentalidad vial, evitando así costos extra por los retrasos en la entrega de mercancía hacia una muestra de tiendas al minorista en la ciudad capital.

4. PREGUNTA DE INVESTIGACIÓN

¿Cómo la aplicación de la Machine Learning (ML) para la predicción del nivel de severidad en accidentes viales puede integrarse con el diseño de rutas de distribución urbana de mercancías considerando la caracterización de accidentalidad vial de la capital colombiana?

Esta pregunta de investigación conduce a tres sub-preguntas:

1. ¿Cuáles son los factores que estadísticamente afectaron la accidentalidad de la ciudad de Bogotá entre 2015 y 2022?
2. ¿Cómo se pueden predecir los niveles de severidad de accidentalidad vial empleando técnicas de Machine Learning?
3. ¿Cuál es el impacto de la probabilidad de ocurrencia y la predicción del nivel de severidad en accidentalidad vial para un algoritmo de diseño de rutas?

5. MARCO CONCEPTUAL

Para la revisión del estado del arte y la selección de investigaciones del presente estudio se consideraron criterios rigurosos de limitación sistemática para no

obstaculizar su futura replicabilidad. La exploración se ejecutó en la base de datos Scopus teniendo en cuenta que esta fuente es una de las más enriquecidas de información global. La búsqueda se limitó a artículos de revisión e investigación en Journals desde el año 2018 hasta 2022, excluyendo al sector de la salud como medicina, neurología, psicología, entre otros. Teniendo en cuenta el amplio alcance del estudio, variedad de cadenas de búsqueda fueron utilizadas implementando palabras clave como Traffic congestion, Traffic Road, Road accident prediction, Traffic accident prediction, Accidents-Bogotá(Colombia), Design Routing, Traffic Flow Prediction, TSP, Accident Analysis and Prevention, entre otras similares, teniendo en cuenta que para la fase predictiva o prescriptiva los modelos de Inteligencia Artificial (IA) tienen mayor precisión de predicción/clasificación (Coursera, 2022). Se realizaron varias iteraciones de estudios en Scival para identificar pilares informativos, como las palabras clave que están en auge, los países donde se han llevado más investigaciones del tema, la distribución del clúster temático por años, el cuartil del Journal y el número de citas.

5.1. MARCO TEÓRICO

El tránsito vehicular es una problemática que aqueja a todas las ciudades y a la sociedad en general, ya que trae consigo varios impactos negativos. De primera mano, la circulación de vehículos contribuye a la emisión de gases y partículas contaminantes en el ambiente, así como también la congestión vehicular representa una amenaza potencial para la seguridad de las personas debido a la elevada probabilidad de accidentes viales asociados. Es importante anotar que un accidente vial se entiende como aquel siniestro que atenta contra la vida de la persona y puede generar una lesión o incluso la muerte. Gran cantidad de políticos e investigadores se han esforzado por encontrar soluciones o estrategias ante este problema, pero el crecimiento acelerado de la población en las ciudades y a su vez el crecimiento del parque automotor vuelve dichas soluciones prácticamente obsoletas, más aún para el caso de Bogotá que está ejecutando

proyectos de movilidad sostenible con una ventana de tiempo de 10 años aproximadamente. Dado que la accidentalidad vial es una de las principales teorías que sustentan el tráfico vial, es necesario abordar grosso modo los aportes científicos, que en principio describen algunos factores y variables influyentes en la frecuencia de accidentes de tráfico.

En el último año, según informes de El Tiempo (Mercado, 2023), un estudio realizado por la Universidad Nacional de Colombia determinó patrones vitales de accidentalidad vial en Bogotá; el análisis, respaldado por herramientas de inteligencia artificial y basado en datos previos al 2019, reveló información crucial sobre los días, horas y lugares con mayor incidencia de accidentes. Los resultados señalan tres franjas horarias críticas: de 6 am a 8 am, de 12 del mediodía a 3 de la tarde y de 5 pm a 8pm. Además, se identificaron los miércoles y viernes como los días de la semana con más reportes de choques. En términos de ubicaciones peligrosas o con mayor frecuencia de accidentes, las intersecciones señaladas por los investigadores de la Universidad Nacional de Colombia PlaS[3] son: “carrera 72 con calle 6.ª; autopista Sur con calle 68 sur; carrera 72 y calle 17; autopista Norte con calle 100; calle 9.ª con carrera 50; carrera 51 con calle 56A sur y la calle 80 sobre la salida hacia el río Bogotá.”.

Ahora bien, resulta indiscutible ver el factor humano como uno de los principales causantes de los accidentes en vía; algunos autores (Aarón et al., 2019; AlKheder et al., 2020; Das, S. et al., 2018; Klinjun et al., 2021b; Leonavičienė et al., 2020b; Ospina-Mateus, Quintana Jiménez, Lopez-Valdes, Berrio Garcia, et al., 2021; Xie et al., 2020b; K. Zhang et al., 2018) convergen en esta idea respecto a sus investigaciones y mencionan que los humanos podrían evitar este tipo de acontecimientos, por ejemplo, si tomaran las prevenciones que emite la educación vial o si fueran más conscientes de las graves consecuencias que pueden causar. La mayoría de estos autores indican las malas condiciones de seguridad vial (alta velocidad, ausencia del cinturón de seguridad, embriaguez, conducción en sentido

[3] PlaS: Programming Languages and Systems

contrario) y otros atributos o criterios psicosociales del conductor como algunos de los principales factores humanos que incrementan la frecuencia de accidentes. De igual forma, determinan que existen factores demográficos asociados incrementalmente a la frecuencia y gravedad de accidentes como el género (se accidentan más los hombres) y la edad (más accidentes relacionados con las personas de edad avanzada). En la misma línea las condiciones meteorológicas y el tiempo han sido factores independientemente evaluados por investigadores (Aarón et al., 2019; Cela & Montoya-Torres, 2022; Ospina-Mateus, Quintana Jiménez, Lopez-Valdes, Berrio Garcia, et al., 2021; Shiran et al., 2021; Xie et al., 2020a; K. Zhang et al., 2018), donde se concluye que horas pico, días lluviosos y días festivos o inicios de fines de semana, son parámetros que incrementan la frecuencia y/o gravedad de los accidentes.

Otros autores involucran en sus estudios factores como el vehículo (Ahmadi et al., 2020; Ospina-Mateus, Quintana Jiménez, Lopez-Valdes, Berrio Garcia, et al., 2021; Sattar et al., 2022; Xie et al., 2020a) y el entorno/infraestructura de las calles (Ahmadi et al., 2020; Arévalo-Támara et al., 2020; Chaparro et al., 2018; Klinjun et al., 2021a; Ospina-Mateus, Quintana Jiménez, Lopez-Valdes, & Sana, 2021b; Siamidoudaran & Iscioglu, 2019; Yang et al., 2022; K. Zhang et al., 2018). En la primera categoría se identifica alta frecuencia siniestral según la tipología de vehículo como: las motocicletas, bicicletas y autos de gran tamaño. Mientras que, en la segunda, se revela como resultado que las zonas con menos curvas, presencia de peatones, longitud del tramo, menor ancho del carril, ausencia de iluminación, condiciones en la carretera, localidades del oeste, suroeste y sureste de la ciudad, tienen mayor representación en la frecuencia y gravedad de los siniestros viales.

Es imperante observar que existe una gran variedad de investigaciones asociadas a la predicción de accidentes de tráfico, principalmente en términos de frecuencia o gravedad (Abadi et al., 2015; Ahmadi et al., 2020; Angarita-Zapata et al., 2021; Basso et al., 2021; Bustos et al., 2021; Guerra et al., 2022; Leonavičienė et al.,

2020a; D. Li et al., 2021; P. Li et al., 2020; Ma et al., 2021; Medina-Salgado et al., 2022; Ospina-Mateus, Quintana Jiménez, Lopez-Valdes, & Sana, 2021b; Ospina-Mateus, Quintana Jiménez, Lopez-Valdes, Berrio Garcia, et al., 2021; Park et al., 2018; Perafan-Villota et al., 2022; Rashidi et al., 2022; Santos et al., 2022; Sattar et al., 2022; Shiran et al., 2021; Siamidoudaran & Iscioglu, 2019; Terán et al., 2020; Xie et al., 2020a; Yang et al., 2022; J. Zhang et al., 2018), sin embargo ninguna está asociada a temáticas de diseño de rutas. El uso de La Inteligencia Artificial (IA), Big Data, Data Mining y la estadística son los métodos de mayor representación en los estudios analizados genéricamente. La implementación de la inteligencia artificial es la más común debido a su adaptación a situaciones personalizadas y su potencial disruptivo en las aplicaciones convencionales. En las investigaciones se observa que IA tiene diferentes disciplinas para la aplicación como son Machine Learning (ML), Deep Learning (DL), Deep Machine Learning (DML), Artificial Neuronal Networks (ANN), Spiking Neural Networks (SNN), Convolutional Neural Network (CNN) o Mixed Language Programming (MLP). La disciplina de inteligencia artificial más utilizado en el 80% de los artículos revisados para predicción de accidentes y su severidad, ha sido ML; luego de realizar un análisis del enfoque implementado, se observó que las técnicas de mayor uso han sido modelos supervisados como árboles de decisión, Support Vector Machine (SMV), Regresión Logística (RL), Random Forest (RF) o K-Nearest Neighbor (KNN), que se utilizan para dar solución a problemas de regresión y/o clasificación; por lo que considerando las características de los datos y la variable a predecir (severidad), algunas de estas técnicas podrían ajustarse al objetivo del presente documento.

Ante este contexto, Ghaemi et al. (2021), Giripunje et al. (2022), Jiang et al. (2022) y Liao et al. (2022) han investigado y ejecutado tácticas con ML para mitigar el tráfico vial mediante problemas del camino más corto, ruteo multicamino, Safe Route Mapping (SRM) y redes VANET (Vehicular Ad-hoc Networks) en tiempo real utilizando la interconexión de vehículos, lugares e incluso personas, pero por el

tipo de datos numéricos tampoco parecen ser 100% compatibles o aplicables para el ruteo del presente que usa datos categóricos.

Como consecuencia de los antecedentes mencionados, se planea implementar algunos métodos aprendidos durante el posgrado para las tres fases de los objetivos a desarrollar, dichos métodos serán mencionados en la sección del estado del arte.

5.2. ESTADO DEL ARTE

Según la revisión previa de las soluciones propuestas, existen diferentes formas de abarcar el problema según los tipos de datos con los que se cuente. No obstante, debido a la naturaleza del ejercicio, el presente documento implementa técnicas diversas considerando que el análisis se desarrolló en 2 software independientes durante tiempos distintos.

Para el caso de R Studio que buscaba tener una mejor comprensión de las variables y su caracterización, se implementaron técnicas complementarias tradicionales basadas en estadística descriptiva e inferencial. En lo que se encuentra una representación de los datos mediante la visualización gráfica, tablas de contingencia, frecuencias y porcentajes; pruebas de asociación entre variables con el estadístico para pruebas de hipótesis chi2 y finalmente un análisis de asociación e interpretación de categorías y variables mediante el análisis de correspondencia múltiple (MCA). En ello, se ejecutaron librerías como dplyr (gramática para la manipulación y operaciones con data frames), plotly- ggplot2 (sistema organizado de visualización de datos), lubridate (formatos en series de tiempo), treemapify (gráficos de mapas de arbol), showtext (formatos de fuente).

La prueba de hipótesis con la distribución chi-cuadrado como lo indica Alisha & Surjeet (2023), es un método sólido usado comúnmente para la selección y descarte de las variables categóricas que pueden influenciar o no modelos de aprendizaje automático, de predicción o clasificación. Por su parte, Das et.al. (2018) rectifica en su investigación de comportamientos en accidentes de tráfico,

que un método que determina las asociaciones clave entre las variables y las categorías es el Análisis de Correspondencia Múltiple, el cual también permite reducir las dimensionalidades, agrupar y visualizar las variables categorizadas en un mapa de mínima dimensionalidad.

Por otro lado, con base en las investigaciones anteriores que contemplan datos categóricos y cuantitativos (Angarita-Zapata et al., 2021; D. Li et al., 2021; Guerra A et al., 2020; Kenny S et al., 2020; Ospina-Mateus et al.,2021; Shiran et al., 2021), los modelos de clasificación seleccionados para la predicción de severidad, cuyas descripciones están basadas en el libro de Data Mining y Machine Learning (Mohammed, et al., 2020), fueron:

- **Decision Tree:** Es un modelo que toma decisiones para clasificar basado en características de los datos, cada nodo va representando un nuevo camino o una nueva pregunta donde sus respuestas están en cada ramificación. Este modelo no requiere normalizar los datos, pueden usar datos cuantitativos y cualitativos; además hace comparaciones de características individuales.
- **Random Forest:** Es un algoritmo de clasificación que se basa en crear y apilar variedad de árboles de decisión, disminuyendo así las posibilidades de sobreajuste de los datos, por ende, su forma de generalizar el modelo es relativamente mejor, además tampoco es sensible a la escala de los datos y no requiere normalización.
- **KNN (K-Nearest Neighbor):** Es un modelo de clasificación predice una clase basado en las clases (puntos) más cercanas, es decir, realiza una clusterización según la distancia a un punto, con ciertas características específicas; es sensible a la escala de las características.
- **Logistic Regression:** Es un algoritmo que modela por medio de la probabilidad (función logística), para decidir si una instancia pertenece a una clase particular; es sensible a la escala de características.

Con el fin de evaluar el rendimiento del modelo se seleccionaron cuatro métricas

de evaluación/validación de la predicción del nivel de severidad:

- **Accuracy:** Es una métrica general de exactitud del rendimiento que determina la proporción de predicciones correctas versus el total de predicciones, es 100% confiable cuando los datos están equilibrados
- **Precision:** Esta métrica calcula la proporción de predicciones positivas correctas entre todas las instancias predichas como positivas.
- **Recall:** La métrica de recuperación o sensibilidad, se usa para saber cuántos valores positivos son correctamente clasificados (proporción de instancias positivas predichas correctamente entre todas las instancias realmente positivas)
- **F1-Score:** Es la media harmónica de precisión (precision) y recuperación (recall). Perfecto para predicciones multiclase y datos desbalanceados.

Para finiquitar, Ismail, Dzulkifli (2021) mencionan en su investigación la selección del modelo TSP como “un método de optimización que busca circular por cada uno de los puntos con la mejor ruta que recorra la distancia mínima de viaje”, que es precisamente lo que busca el presente manuscrito. Para el diseño de ruteo TSP ejecutado en Python, se utilizó la librería Pulp que ayuda a modelar problemas de optimización mediante programación lineal para modelos de asignación junto con la heurística Greedy en python(Universidad de Granada, 2020). Dicha Heurística codificada trae consigo una variable binaria, la cual indica si se visita o no cada nodo, con la restricción de que inicie el recorrido en un nodo específico (en este caso el CEDI) y que los siguientes nodos (tiendas) se visiten una única vez; la toma de decisiones de la heurística busca minimzar el peso del tiempo total de recorrido. También se hizo uso de la documentación de la librería Networkx para el diseño de la ruta dirigido TSP; los parámetros principales para este modelo se importaron como matrices creadas con Google Maps desde archivos Excel tipo xlsx. La librería Networkx (2023) permite analizar las conexiones de las rutas, KPI’s estadísticos, e incluso representa el grafo de diferentes maneras.

6. OBJETIVOS

6.1. OBJETIVO GENERAL

Diseñar un modelo híbrido de ruteo para la distribución urbana, integrando técnicas de Machine Learning que consideren como elemento clave la caracterización de los accidentes viales en la ciudad de Bogotá.

6.2. OBJETIVOS ESPECÍFICOS

- Caracterizar las variables que han intervenido en la frecuencia de accidentalidad vial en la ciudad de Bogotá durante el periodo de 2015 a 2022, a través técnicas de estadística descriptiva e inferencial.
- Implementar un modelo de clasificación basado en técnicas de ML para la predicción de la severidad de accidentes viales en la ciudad de Bogotá, fundamentado en las variables de mayor impacto estudiadas anteriormente.
- Proponer un algoritmo estático para el diseño de rutas de distribución que tome en cuenta las predicciones de severidad de accidentalidad vial.

7. METODOLOGÍA

El ejercicio del presente documento se ejecutó en 2 software diferentes (R y Python), todos los scripts se corrieron en un computador con procesador Core i5, con memoria RAM de 8 GB y sistema operativo de 64 Bits con Windows 10. Para el caso de Python, la versión utilizada fue 3.10.8.

La metodología en este proyecto investigativo se divide en 3 secciones: Descripción (Ejecutado en R Studio), Prescripción y Predicción (Ejecutados en un Notebook de Python) con el fin de cumplir con cada objetivo específico planteado en la sección previa.

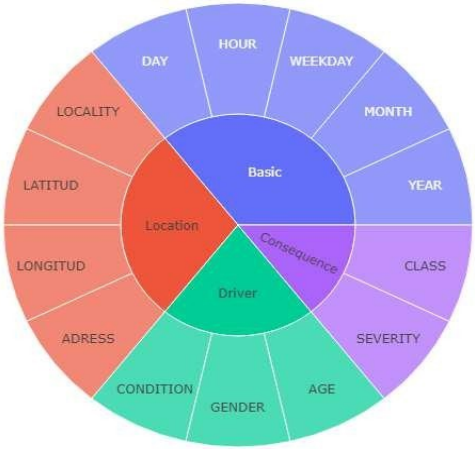
7.1. METODOLOGÍA DESCRIPTIVA

Inicialmente se utilizaron estadísticas descriptivas para dar una generalización a partir de los registros de la Secretaría Distrital de Movilidad (2022) referentes a accidentes de tránsito ocurridos durante el año 2015 y 2022 de 2 bases de datos complementarias entre sí. Dicho análisis contiene el diagnóstico actual del conglomerado de datos, limpieza y depuración de datos, estadística descriptiva e inferencial de los datos (que explica la relación entre variables y categorías), ya que los datos suministrados no son numéricos de tipo contable. La información contenía diferentes tipos de datos como fechas, texto, números ordinales y factores. Adicionalmente, en la limpieza de datos, se observó que algunas de las variables/detalles de registro en los conjuntos de datos no eran aprovechables o estaban incompletos, por tal motivo se seleccionaron las variables a analizar que se detallan en la Tabla 1. Cabe resaltar que allí existen variables asociadas directamente a información de tiempo (Basic), consecuencia del accidente (Consequence), información del pasajero (Driver) y espacio del siniestro (Location). Luego traducir las variables, la visualización de la categorización de las variables seleccionadas se puede apreciar en la ilustración 2.

Tabla 1.Variables seleccionadas para el análisis (2022).

Datos De Movilidad Accidentes	Datos De Movilidad Lesionados
Hora_Ocurrencia_Acc, Ano_Ocurrencia_Acc, Mes_Ocurrencia_Acc, Dia_Ocurrencia_Acc, Gravedad, Clase_Acc, Localidad, Latitud, Longitud, Fecha.	Hora_Ocurrencia_Acc, Ano_Ocurrencia_Acc, Mes_Ocurrencia_Acc, Dia_Ocurrencia_Acc, Clase_Acc, Localidad, Condición (rol en el siniestro), Genero, Edad, Latitud, Longitud, Fecha.

Ilustración 2. Categorización de las variables.



Con el fin de contextualizar visualmente cada siniestro, se utilizó la herramienta de Power Maps para generar un mapa de calor/densidad con diferentes capas que representan los puntos de concentración de accidentalidad, lesiones y mortandad en el conglomerado del periodo de tiempo analizado (2015-2022). Ver Ilustraciones 3, 4 y 5.

Ilustración 3.Mapa de calor accidentes

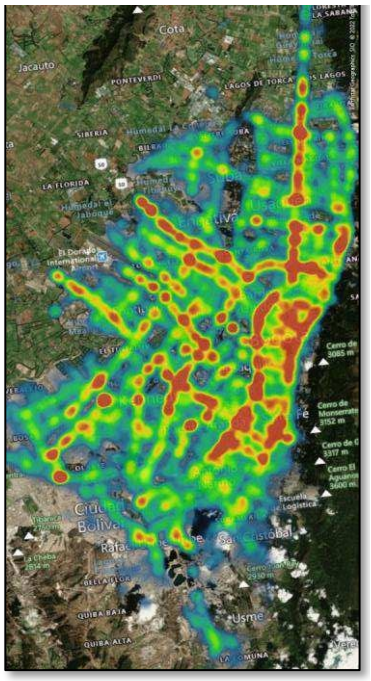


Ilustración 4. Mapa de calor lesiones

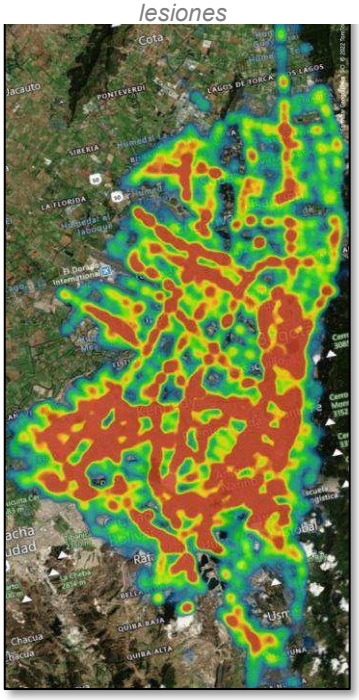
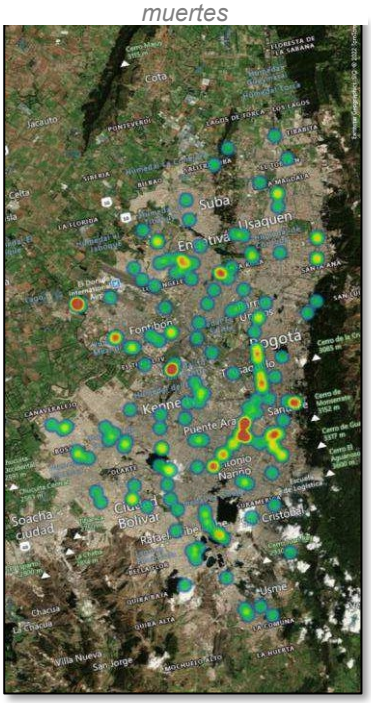


Ilustración 5. Mapa de calor muertes



En estos mapas de calor señalados, se perciben los puntos de alta concentración en accidentalidad, representados con el color rojo; estos ocurren en las avenidas principales (Autopista Norte, Avenida Caracas, Avenida Boyacá), con mayor presencia a simple vista en las localidades de Kennedy, Teusaquillo, Engativá, y Santa Fe. Por su parte, los lesionados están uniformemente distribuidos, ya que a simple vista no evidencian un patrón singular por localidad; los accidentes con mortandad se ven aglomerados entre localidades como La Candelaria, Kennedy, Los Mártires y Engativá.

Ahora bien, dado que al inicio del estudio no se tenía conocimiento ni ningún acercamiento sobre el funcionamiento del software Python, se utilizó R Studio para empezar a hacer un análisis a profundidad de la información. Allí se filtraron y manipularon las variables a estudiar para cada base de datos de manera independiente (considerando que si se acoplaban, el resultado del análisis se basaría en consideración a vacíos o datos nulos), se detectaron las diferentes incongruencias que tenía cada base de datos para eliminar y reemplazar algunos datos inexistentes o completarlos dependiendo del caso y tipo de variable; así entonces, se convirtieron algunos datos a idioma inglés para aquellos valores de tipo texto, se transformaron los factores, se extrajeron datos específicos de la fecha y se definieron los tipos de variables con los respectivos formatos, bien sean numéricos, tipo texto, de factores, horas y fechas. Es importante denotar que para la variable de edad se dividieron las edades con la condición etérea según las etapas que indica (Ministerio de salud, n.d.). El diagnóstico estuvo compuesto por un análisis descriptivo estático en el que se incluyen tablas de contingencia, diagramas como circular, barras, combinados y árboles, entre otros, con el fin de determinar y representar la distribución de frecuencias de accidentalidad según cada variable seleccionada previamente. En adición al análisis exploratorio, se ejecutó un análisis inferencial que evalúa la correlación entre las posibles combinaciones de las variables mediante las hipótesis:

H0: Las variables comparadas son independientes

H1: Las variables comparadas no son independientes

Estas hipótesis se evaluaron con el valor de probabilidad (P-val) mediante la prueba Chi Cuadrado (chisq.test en R) trayendo a consideración que las variables son de tipo categórico. Es importante recalcar que todas las posibles combinaciones de las variables fueron probadas en este estadístico para evaluar la independencia o relación entre las mismas y de esta forma seleccionar las variables de mayor importancia, necesarias para hacer el modelo de ML.

Finalmente, para revisar la selección correcta de las variables o eliminación concreta, se comprobaron los resultados de asociación entre variables y se ahondó en las categorías de las variables con el MCA. La idea principal de hacer uso del MCA no solo es proporcionar una representación gráfica ante la relación de categorías y variables, simplificar la información e identificar patrones, sino que también busca entender la varianza de las categorías ante diferentes dimensiones. Para ejecutar este análisis múltiple fue necesario implementar un muestreo aleatorio estratificado por proporciones dependientes al año debido al tamaño de los datos. Esto se realizó mediante la librería sampler con la que se obtuvo una muestra total de 383 observaciones para corroborar que representan adecuadamente las distribuciones de los datos poblacionales. Así entonces, los valores propios demuestran que resulta difícil explicar la variabilidad de las variables categóricas en pocas dimensiones, puesto que para poder explicar al menos el 70% de la varianza de los datos se necesitaría de 33 dimensiones aproximadamente, por lo que para efectos de practicidad se evaluaron las primeras 2 dimensiones para corroborar la similitud con el análisis descriptivo previo referente a la descripción de los datos.

7.1.1. ANÁLISIS DE RESULTADOS DESCRIPTIVOS

Para la base de accidentes se observa un patrón evidente en el que el tipo de accidente de choque es el factor de mayor representación con el 85,85% del total de accidentes, donde la gravedad más destacada es de tipo solo daños, la localidad en la que prevalecen dichos accidentes es Kennedy en el tramo horario de 12 a 4 de la tarde los viernes. El mes de octubre (9%) y el año 2018 (14%) fueron las categorías de tiempo con mayor participación, sin embargo, no resultan factores diferenciales o significativos, al igual que el día del mes.
Los datos obtenidos de la base de lesionados muestran resultados consistentes en términos del día de la semana, el tipo de accidente y la ubicación. Sin embargo, resulta interesante destacar que la información temporal revela detalles significativos. Por ejemplo, en cuanto al año, la mayoría de los casos se

concentran en el año 2022. Asimismo, al analizar el horario, se observa que el período de 6 a 8 de la mañana es cuando se registra el mayor número de lesionados. Además, refleja que el factor humano entra a ser un factor diferenciador, más específicamente el sexo masculino (65%), donde en su mayoría son motociclistas adultos (49%) los accidentados.

Para complementar la información previa, se combinaron ambas bases de datos en Pyhton (con variable única llamada “Formulario”). En este análisis se observaron cosas de mayor detalle como por ejemplo que existe una clusterización de la información por localidad según el gráfico de coordenadas (ver Ilustración 4). Se vuelve a confirmar lo expuesto en previas investigaciones sobre 3 horarios de mayor accidentalidad que son de 6 a 8 am, de 12 del mediodía a 2 de la tarde y de 4 a 6 pm; además, por medio de tokenización de la dirección se descubrió que hay más accidentes en vías de tipo calle con el 53% que de tipo carrera (44%). En este estudio conjunto también se observa que la clase de accidente de choque, el año 2018, la severidad de solo daños y la localidad de Kennedy son categorías que no cambian respecto al análisis anterior y siguen siendo variables imperativas para el estudio.

Ilustración 4. Clusterización de accidentes y lesiones según la latitud, longitud y tipo de dirección

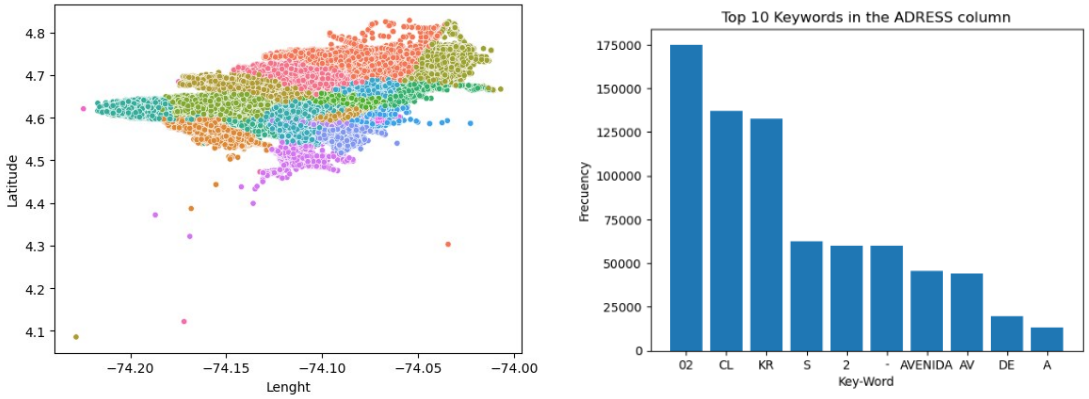


Ilustración 5. Accidentes y lesiones semanales por hora

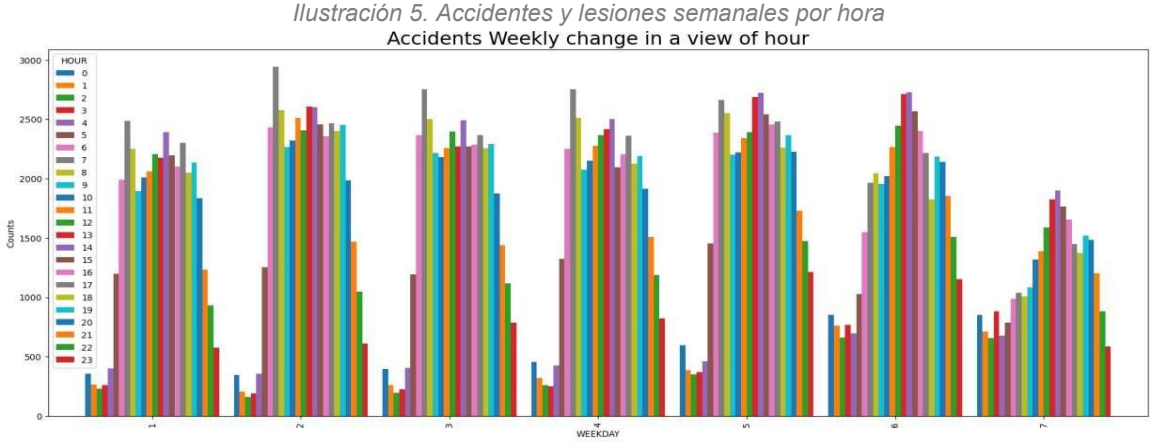


Ilustración 6. Top 10 de localidades con más accidentes y lesiones según la severidad.

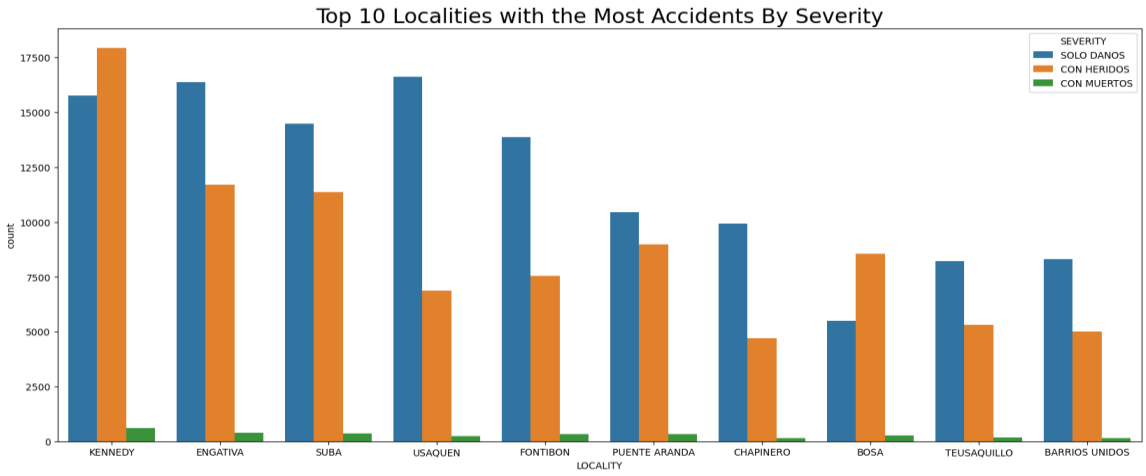
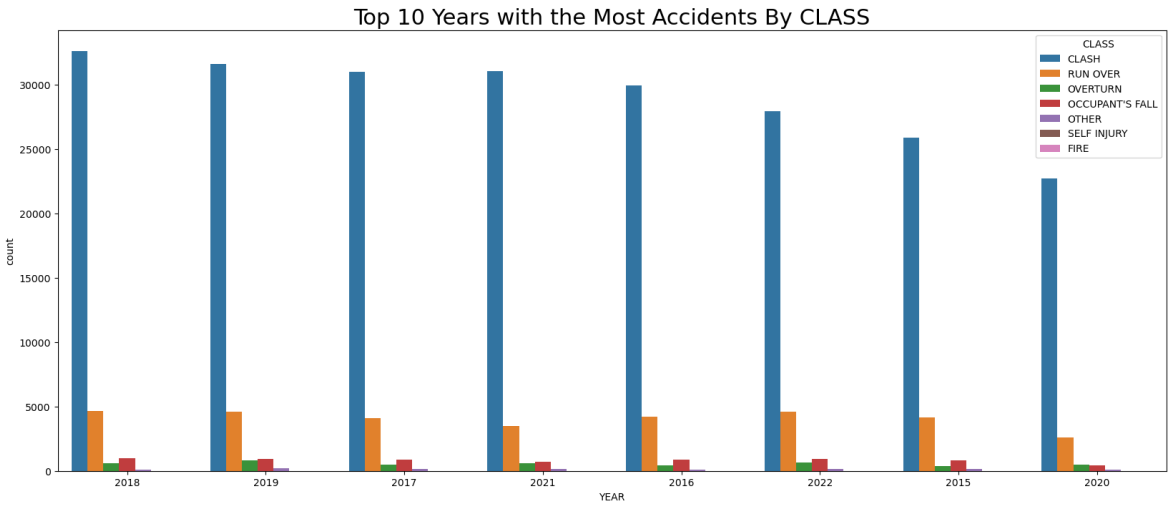


Ilustración 7. Top 10 de años con más accidentes y lesiones según la clase



Ahora bien, con respecto a las pruebas inferenciales en ambas bases se encontró que sí existe correlación/independencia ante todas las posibles combinaciones, a excepción de la clase y la localidad con un $p - val > \alpha$ (donde $\alpha = 5\%$) en la base de accidentes que indica que no hay independencia entre estas dos variables mencionadas. Cabe denotar que la correlación con los días parece no ser tan significativa, ya que su distribución de variabilidad es constante.

En la línea paralela del MCA, los resultados indican que la dimensión con mayor capacidad de varianza explicada es la dimensión #1 con el 3.3.% (ver Ilustración 10). De igual manera, las dimensiones 1 y 2 no son suficientes para retener la variación total contenida de los datos, ya que solamente representan el 6.2%. Ante esta situación, los resultados que se muestran posteriormente son un ejemplo generalizado con las primeras 2 dimensiones para corroborar la similitud con el análisis descriptivo previo referente a la descripción de los datos, debido a que la cantidad de dimensiones para explicar un porcentaje significativo tiene gran escala.

Ilustración 8. Scree plot por dimensión

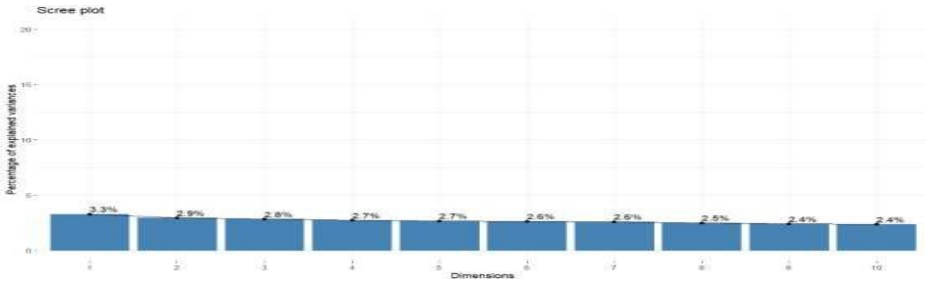


Ilustración 9. Índice de correlación por dimensión por variable

Categorical variables (eta2)

	Dim.1	Dim.2	Dim.3
ANO_OCURRENCIA_ACC	0.090	0.149	0.220
HORA_OCURRENCIA_ACC	0.174	0.395	0.388
MES_OCURRENCIA_ACC	0.163	0.315	0.241
DIA_OCURRENCIA_ACC	0.026	0.265	0.278
SEVERITY	0.519	0.001	0.027
CLASE_ACC	0.518	0.130	0.123
LOCALIDAD	0.414	0.444	0.365

En la Ilustración 11 referente a la correlación entre variables y dimensiones, se identifica que la variable con mayor representación en la dimensión 1 es la de Severidad, mientras que para la dimensión 2 es la localidad en la que ocurrió el accidente. También se observa que la correlación entre variables tiene 2 agrupaciones, por un lado, la de tiempo (Año, Mes, Día y Hora en el que ocurre el siniestro) y por otro lado el espacio (localidad, clase y severidad del accidente).

Para la correlación entre las categorías/factores de cada variable, se observa correlación en la gran mayoría de ellos (ver Ilustración 10). Por medio de las curvas de densidad (Ilustración 12) se detallan las zonas de las categorías que están altamente concentradas, donde se corrobora mayor correlación en los factores denotados en la previa sección tales como la localidad de Kennedy, el choque como tipo de accidente, los viernes, entre otros. También se identifica que no existe correlación entre variables como las localidades de Antonio Nariño, Candelaria, Ciudad Bolivar, hora de 4 a 6 de la mañana, tipo de accidente conductor. Las Ilustraciones 12 y 13 permiten identificar la calidad de representación (Cos^2) de cada categoría, en este caso se observa una fuerte calidad que supera el 49% en accidentes de tipo choque, con heridos y solo daños; esto indica que estas categorías son las que explican mayoritariamente la correlación en las dimensiones 1 y 2.

Ilustración 10. correlación entre variables y dimensiones

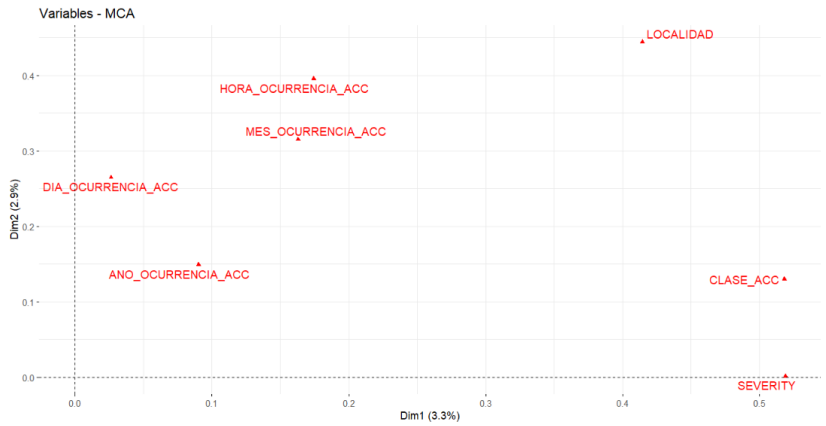


Ilustración 11. Visualización de la calidad de representación de los factores.

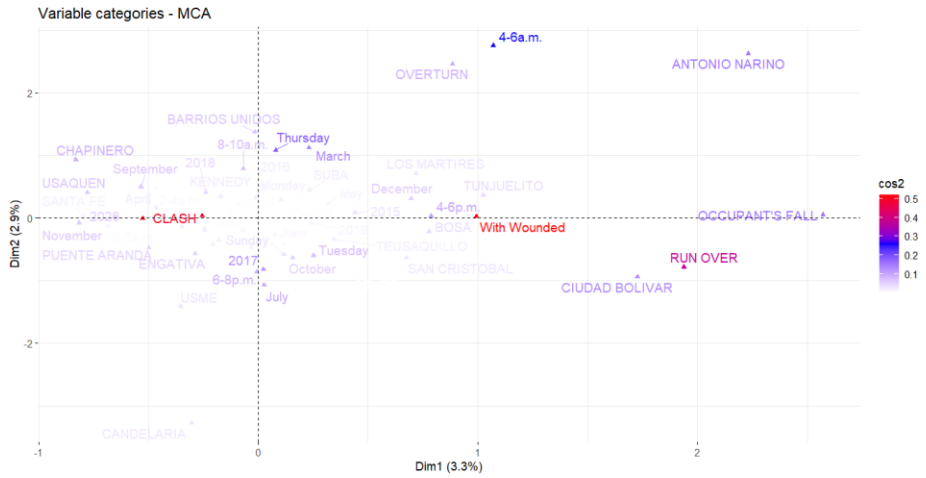


Ilustración 12. MCA plot con curvas de representación

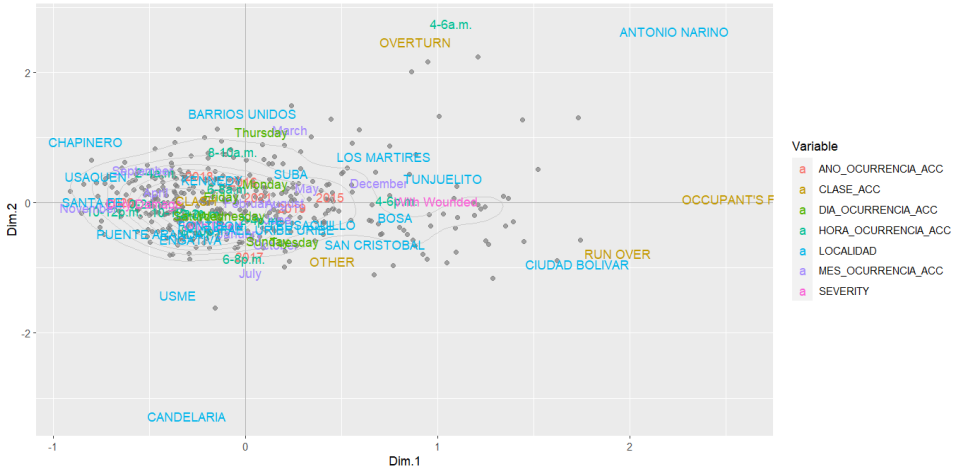
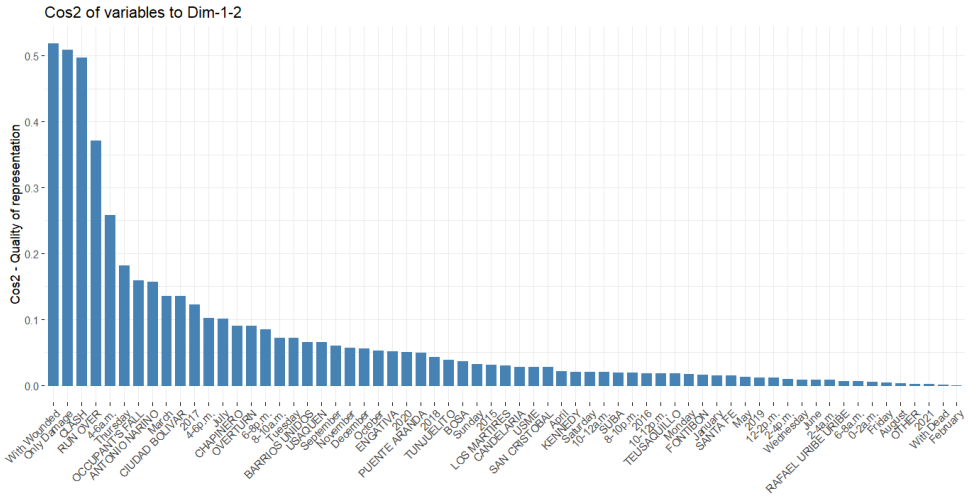


Ilustración 13. Calidad de representación de los factores



Al integrar ambas fuentes de datos con el fin de llevar a cabo el proceso de ruteo, se concluyó que las variables de condición, género y edad no son apropiadas para su utilización. Estas variables en primer lugar no son primordiales y, en segundo lugar, presentan un 54% de información faltante (Nan), lo cual se traduce en ambigüedad porque no representa utilidad ni confianza en el diseño de rutas esperado. Con base en este análisis, se determinó que las variables de selección que se usarán en la metodología predictiva son: la clase de accidente, la localidad, la hora, el mes, el día de la semana, la dirección del accidente y como variable de salida la severidad.

7.2. METODOLOGÍA PREDICTIVA

Con el preprocesamiento descrito previamente y las variables seleccionadas se consideró primeramente un modelo de asignación de los centros de distribución a una muestra de las tiendas de reparto con el fin de reducir el tamaño del modelo dada su complejidad. Este modelo considera como ventana de tiempo una hora pico (de 6 a 8 am), pero está planteado en un escenario ideal que no contempla restricciones de movilidad de camiones o aspectos similares.

La asignación se hace debido a que actualmente existen 3 centros de distribución aledaños a la capital colombiana, por lo que se entra en un supuesto de que los 3 centros no solo reparten los mismos tipos de productos, sino que también tienen la misma capacidad para suplir cualquier tienda de la ciudad de Bogotá. Bajo este supuesto, se seleccionó una muestra de 5 tiendas de la ciudad distribuidas proporcionalmente en diferentes UPZ y localidades de la metrópoli; se propusieron 3 rutas diferentes para llegar única y exclusivamente desde el CEDI hacia la tienda, haciendo uso manual de Google Maps con sus respectivas distancias y tiempos (tanto en hora muerta como en hora de mayor accidentalidad según lo estudiado en la sección previa). Así entonces, se procedió a hacer un problema de asignación óptima con su respectivo modelamiento matemático propuesto de la siguiente manera:

Índices

P: plantas/CEDI {1,2,3}, donde 1=Sibaté, 2=Tocancipá, 3=Funza

T: Tiendas {1,2,3,4,5} donde 1=Kennedy, 2=Engativá, 3=Suba, 4=Chapinero, 5=San Cristobal

Parámetros

D_{TP}: Distancia de la planta P a la tienda T

Variable de decisión (Binaria)

X_{PT}: si la planta P atiende a la tienda T

Función objetivo

Minimizar la suma ponderada de la distancia total recorrida, de esta manera, el modelo resuelve el problema de asignación y cantidad de plantas/CEDI's a abrir.

$$Min\ Z = \sum_{P=1}^{3} \sum_{T=1}^{5} D_{TP} * X_{PT}$$

Restricciones

1. Una tienda solo puede ser atendida por una planta

$$Demanda \quad \sum_{P=1}^{3} X_{PT} = 1 \qquad \forall T\{1, \dots, 5\}$$

Los resultados obtenidos mediante este modelo de asignación permitieron minimizar la cantidad de nodos a predecir. Dicha minimización posibilitó la generación de nuevas matrices que reflejan la interconexión entre tiendas y las posibles rutas considerando la dirección de la ruta (doble vía). Para llevar a cabo este proceso, se acudió nuevamente a la herramienta de Google Maps con el fin de extraer la ruta de menor distancia en el mismo tramo de recorrido con el horario

seleccionado (6 a 8 am), allí también se buscaron y registraron en la hoja de Excel los datos esenciales de la ruta (tiempos, distancias, nodos).

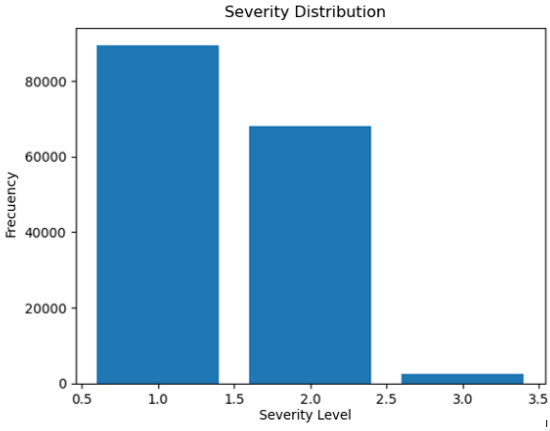
Así entonces, al tener menos nodos con el modelo de asignación, fue posible estandarizar los valores de la dirección y el horario apilado en 2 horas para simplificar la tarea de predicción de la clase y precisar mejor el resultado, más específicamente al “dumizar” las categorías de las variables, puesto que esto puede facilitar su procesamiento. Dicha estandarización fue desafiante, ya que, en el proceso de depuración de datos, se identificó una notable diversidad en la forma de redacción de las direcciones, por lo que, para abordar la diversidad mencionada, se diseñó un diccionario integral que engloba las diversas formas de expresar una dirección; es indispensable que si una persona desea replicar este modelo tenga total comprensión de los datos y conozca bien las direcciones a las que se haga referencia.

El diccionario integral mencionado expresa la variedad de formas en las que se puede expresar una sección de la dirección de búsqueda, como, por ejemplo: 'esperanza|galan|ac 24|cl 24 |cl 24-|cl 24 -' (donde el separador “|” significa o), junto con su key-value genérico que se le asigne, en este caso: "Av La Esperanza". Con este diccionario se buscó crear un nuevo dataframe que contuviese exclusivamente los patrones de direcciones/nodos seleccionados. Este nuevo dataframe se diseñó para sustituir la dirección original por el key-value genérico correspondiente extraído del diccionario; además contiene las variables 'Class', 'Locality','Hour', 'Month', 'Weekday','Severity','Match' y hace la respectiva dumización de las mismas a excepción de la variable de salida (Severity). Para ejecutar la predicción de clasificación de la severidad fue necesario sustituir los valores correspondientes a "Solo Daños: 1", "Con Heridos: 2" y "Con Muertos: 3”. Con esto, se evaluó la distribución de las clases en el nuevo conjunto de datos para determinar su equidad y precisión, es decir, verificar si los valores de entrenamiento estaban balanceados. Este aspecto es crítico, ya que, de haber un desequilibrio, este podría influir en el sesgo, el rendimiento o la relevancia del

modelo de predicción de la clase.

Los resultados, como se muestran en la Ilustración 14, indicaron un desequilibrio en los datos, lo cual es comprensible y realista, dado que los accidentes con víctimas mortales son menos frecuentes. Para abordar este desequilibrio de clases a predecir, se optó por una estrategia que mantiene la proporción de clases sin necesidad de recurrir directamente a oversampling/undersampling considerando que el tamaño de entrenamiento de los datos; esta estrategia se basa en ajustar un parámetro llamado class_weight con 'balanced'. El uso de 'balanced' permite que el modelo ajuste automáticamente los pesos de las clases de forma inversamente proporcional a sus frecuencias; es decir, las categorías menos frecuentes tienen mayor peso en el modelo. Este parámetro está disponible para los learners[4] seleccionados (Decisión Tree, Random Forest, Logistic Regression), a excepción de K-Nearest Neighbors (KNN) que es el único modelo de clasificación que requiere oversampling, el cual se hace mediante la librería "imblearn".

Ilustración 14. Distribución del nivel de severidad



Dado que se cuenta con 4 posibles modelos seleccionados según la revisión de la literatura, se procede a evaluar los mismos para identificar cuál de ellos se adapta mejor al escenario planteado en el presente documento. Los modelos se entrenaron con el 80% de los datos, separando al conjunto de prueba de forma estratificada, es decir, con la misma proporción de clases con las que se entrena. Se emplearon métricas de evaluación ligadas con la naturaleza específica del ejercicio, tales como Accuracy, Precision, Recall y F1-Score. Con este análisis, se identificó el modelo más eficaz, y se sometió a una evaluación adicional que

[4] Modelos de predicción

realiza múltiples divisiones aleatorias del conjunto de datos en conjuntos train y test: el Monte Carlo Cross Validation con 5 iteraciones y métrica de F1-Score (una de las métricas más usadas para evaluar datos desbalanceados) para tener mayor certeza y confianza de los datos obtenidos en la predicción de la clase. Adicionalmente, para comprender mejor el comportamiento de predicción del modelo de clasificación seleccionado, se visualizó la matriz de confusión que determina los valores predichos como verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

En este orden de ideas, se procedió a generar las instancias de predicción de manera individual, ya que cada predicción puede incluir o no especificaciones adicionales que afecten el valor de salida. En este caso particular, las especificaciones fueron: el tramo horario (de 6 a 8 am) y una dirección genérica específica (ej: Av Boyacá); no obstante, en caso de que existan más especificaciones, el modelo está diseñado para recibirlas a través de un bucle. De esta forma entonces, el modelo entrenado toma como entrada la instancia creada y arroja el resultado de la clasificación predicha, junto con la probabilidad asociada a dicha predicción; esta es la probabilidad del evento que se almacena en la hoja Excel junto con el nivel de severidad predicho y se utilizan para incrementar los tiempos en el diseño de ruteo que se ejecutará posteriormente.

7.2.1. ANÁLISIS DE RESULTADOS PREDICTIVOS

El resultado del modelo de asignación por parámetros de distancia indica que no se debe abrir o asignar ninguna tienda al CEDI 2 que refiere al centro de distribución ubicado en Tocancipá, además, la ruta que minimiza la función objetivo de la distancia total recorrida es la Ruta 1, como se muestra a continuación:

Tabla 2. Resultado del modelo de asignación.

Ruta	Función Objetivo	Asignación CEDI's
R1	102,515	1,3,3,3,1
R2	112,710	1,3,3,1,1
R3	110,485	1,3,3,1,1

La interpretación de la asignación previa se resume en que Sibaté se encargará exclusivamente de atender las tiendas de San Cristóbal y Kennedy, mientras que Funza asumirá la responsabilidad de las demás tiendas (Chapinero, Engativá y Suba); todo esto en un recorrido que abarca la distancia total de 102,5 km.Con los resultados de la asignación previa, se derivó la reducción de nodos y se creó la nueva base de datos según el diccionario creado; con esta nueva base de tamaño 160179 filas x 7 columnas se trabajará la predicción de clasificación del modelo. Así pues, se obtuvieron las variables Dummy para las variables de tipo "Object", es decir las variables de tipo categóricas, lo que resultó en un total de 129 columnas, excluyendo la variable de salida.

Los resultados de la distribución de la variable de salida (Severidad), revelaron que los datos están desbalanceados como se expuso en la Ilustración 16; el 60% de los datos representaron al primer nivel de severidad referente a "Solo Daños", mientras que el 43% a "Con heridos" y el porcentaje mínimo restante se asoció al nivel "Con Muertos". Teniendo esto en consideración, se evaluaron los "learners" o modelos de ML con los scores establecidos, tanto para el escenario que incorpora el parámetro de desbalanceo como para el escenario que los excluye y trabaja bajo el supuesto de datos balanceados; esto se hizo para ver si las métricas difieren entre sí. Los resultados del modelo se muestran en la tabla número 3.

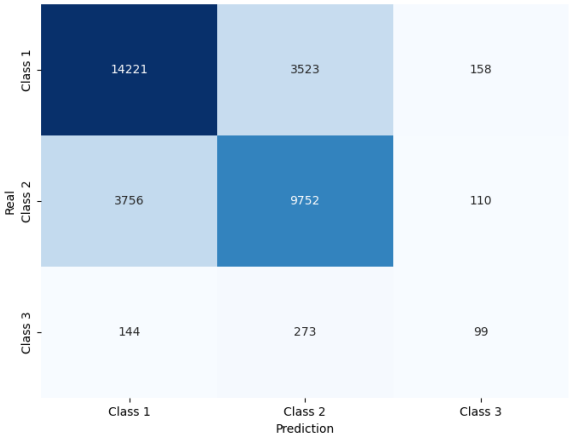
Tabla 3. Comparativo del rendimiento de los learners

Escenario	Métrica	DecisionTree	RandomForest	Logistic	KNN
Sin	accuracy	0.726	0.756	0.704	0.669
Parámetros	precision	0.725	0.752	0.708	0.663
de	recall	0.725	0.755	0.704	0.669
Desbalance	f1	0.725	0.751	0.679	0.660

Con	accuracy	0.715	0.751	0.568	0.629
Parámetros	precision	0.728	0.749	0.658	0.659
de	recall	0.715	0.751	0.568	0.629
Desbalance	f1	0.719	0.750	0.576	0.642

En la Tabla 3, se evidencia que, la diferencia entre los scorings que evalúan el rendimiento de cada uno de los modelos varía ligeramente al contemplar o no los parámetros de desbalance. No obstante, para obtener resultados realistas, los resultados a usar en el ejercicio de predicción son exclusivamente aquellos obtenidos con los parámetros de desbalance, más específicamente el mejor modelo de clasificación seleccionado, el cual es el Random Forest (RF) con la métrica del F1-Score. La elección de la métrica F1 se fundamenta en el contexto de datos desbalanceados y en su rendimiento obtenido del 75%. Para tener mayor confianza en esta métrica, se incorporó la validación cruzada Monte Carlo (MCCV), corroborando el rendimiento del modelo RF con un promedio de validación del 74.9%. Posteriormente, se realizó una matriz de confusión que permite visualizar el comportamiento del modelo, es decir qué clases se están prediciendo correctamente; la matriz resultante se muestra en la Ilustración 15:

Ilustración 15. Matriz de confusión del modelo seleccionado Random Forest



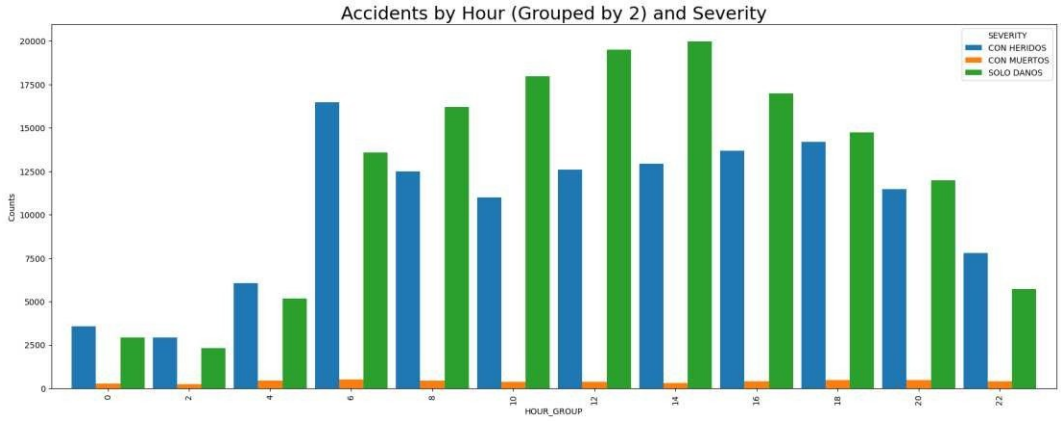
En la matriz anterior se visualizan las filas que corresponden al nivel de severidad

real, mientras que la columna corresponde a la predicción que hace el modelo de clasificación con respecto al nivel de severidad; por ejemplo, la interpretación para el nivel de severidad Con heridos (Clase 2) es:

- Falsos Negativos para "Nivel 1" no predicho como "Nivel 2": 3756
- Verdaderos Positivos para "Nivel 2": 9752
- Falsos Positivos para "Nivel 3" predicho como "Clase 2": 110

Finalmente, al analizar la predicción de clase para las nuevas instancias, se observó un dato curioso: dadas las especificaciones de las instancias a predecir, se observó que el nivel de severidad que se clasificó mayoritariamente fue el 2 que corresponde a accidentes con heridos. Para verificar si este patrón de predicción era consistente con la realidad o si representaba un error del modelo, se realizó una prueba modificando el horario de “6 a 8 am” por “12 a 2 pm”; ante este escenario se observó que la clase de mayor predicción era la 1 con un aumento del 10% en promedio sobre la probabilidad predicha. Con lo anterior se validó el correcto funcionamiento del modelo, debido a que esta diferencia en la clase predicha según las especificaciones efectivamente tiene relación con el horario en el que mayor conglomeración vehicular hay durante el día. Como se observa en la Ilustración 16, de 10 am a 4 pm hay un aumento notable de accidentes de tráfico explicado por levantamiento de restricciones de horarios pico, lo que conlleva a la disminución de congestión de vehículos, aumento de la velocidad del recorrido y carencia en la precaución de los conductores a la hora de manejar, lo que genera un aumento en la frecuencia de accidentes que permiten al modelo tener mayor precisión en la clase a predecir. Ante este panorama, se confirma que los resultados del escenario inicial fueron correctos y realistas según los datos de entrenamiento, a pesar de que la probabilidad de predicción sea ligeramente menor. Es importante denotar con este análisis que el modelo es sensible a las especificaciones particulares generadas en la instancia que se intenta predecir.

Ilustración 16. Accidentes por tramo horario según la clase de severidad



7.3. METODOLOGÍA PRESCRIPTIVA

En vista de lo anteriormente expuesto, se procede a realizar el modelo de ruteo mediante la metodología del Traveling Salesman Problem con la heurística de Greedy,. Para ello, como se mencionó anteriormente, hay 2 probabilidades en juego: la primera es una probabilidad acumulada o por frecuencia y la segunda se establece con la probabilidad de la severidad predicha.

Para calcular la probabilidad faltante referente a la probabilidad por frecuencia, fue necesario implementar un diccionario que contara la cantidad total de accidentes que han existido en cada tramo basado en el key-value genérico de la dirección. Esta información se exportó al Excel global y se asoció cada nodo con su respectiva frecuencia para cada tramo del recorrido según las tiendas a visitar con la asignación previa. Así entonces, con la suma total de accidentes se logró calcular y registrar la probabilidad por frecuencia.

Ahora, considerando que la probabilidad predicha por el nivel de severidad es un evento calculado de manera independiente, se crean factores esperados de la siguiente manera:

- Para severidad 1, el factor es 0.5.
- Para severidad 2, el factor es 1.

- Para severidad 3, el factor es 1.5.

Estos valores relacionan el nivel de severidad predicho con la importancia que se le da a la probabilidad de predicción, además, permiten reflejar tiempos realistas que representan un impacto proporcional según la importancia de la severidad del accidente en vía (Solo daños, Con heridos o Con muertos). Por lo que el propósito del factor es darle mayor o menor peso al valor de probabilidad predicha según el nivel de severidad, y por ende controlar el impacto que puede generar en el aumento de tiempo de un recorrido modelado en un escenario realista.

Así entonces, la fórmula utilizada para el aumento del tiempo original en minutos (**T'**) en función de las probabilidades de accidente y la severidad pronosticada se planteó de la siguiente manera:

$$T' = T * (1 + \mathrm{PA} + \mathrm{F} * \mathrm{PS})$$

donde:

- **T** Es el tiempo Original
- **PA** es la probabilidad por frecuencia de accidente
- **F** es el factor asociado al nivel de severidad
- **PS** es la probabilidad de la severidad predicha

La justificación por la que los tiempos asociados a cada probabilidad no se multiplican, sino que se suman, radica en que los eventos son independientes y el método de cálculo es distinto para cada uno; por ejemplo, la probabilidad de accidente tiene una escala normalizada en un rango de 0 a 1 global, es decir, con respecto a la suma de todas las posibles instancias en el tramo de ruta, mientras que la probabilidad de severidad posee una escala de 0 a 1 para cada instancia individual.

En ese sentido, se crearon las nuevas matrices de tiempo, tanto sin accidentes, como con penalización por probabilidad de accidentes de tráfico y severidad de

este, en el tramo de tiempo seleccionado sobre el cual se obtuvieron todos los datos (6 a 8 am). Seguido a esto, se codificó el modelo TSP dirigido (con especificación de direcciones y valores cambiantes según la dirección), integrando como parámetros las matrices de tiempo calculadas; este modelo, no solo muestra el tiempo total del recorrido, sino que también proporciona una visualización del grafo basado únicamente en los nodos, mas no en espacios geográficos.

7.3.1. ANÁLISIS DE RESULTADOS PRESCRIPTIVOS

El modelo TSP fue probado en 4 escenarios distintos, el primero no contempla ninguna ventana de tiempo, únicamente contempla distancias, mientras que los demás si están planteados en escenarios de tiempos bien sea en hora muerta, en hora concurrida sin contemplar probabilidad alguna de accidentes y en hora concurrida considerando accidentalidad vial. Los resultados del modelo ejecutado arrojaron los siguientes resultados:

Tabla 4. Resultados del modelo TSP.

Escenarios	**Sibaté**	**Funza**
Con Distancias	Ruta: Sibaté -> Kennedy -> San Cristóbal Distancia: 30.02	Ruta óptima: Funza -> Engativá -> Suba -> Chapinero Distancia: 33.72
Con Tiempos en hora muerta (IDEAL)	Ruta: Sibaté -> Kennedy -> San Cristóbal Tiempo: 54.1	Ruta óptima: Funza -> Suba -> Engativá -> Chapinero Tiempo: 96.24
Con Tiempos en hora pico	Ruta: Sibaté -> Kennedy -> San Cristóbal Tiempo: 111.1	Ruta óptima: Funza -> Suba -> Engativá -> Chapinero Tiempo: 204.12
Con Tiempos y Probabilidad de accidentes	Ruta: Sibaté -> Kennedy -> San Cristóbal Tiempo: 169.18	Ruta óptima: Funza -> Engativá -> Suba -> Chapinero Tiempo: 286.19

En el resumen expuesto a priori (Tabla 4), se detalla que existe una diferencia

significativa en el valor de la función objetivo de tiempo según los escenarios planteados, puesto que, al considerar la presencia de accidentes en la vía, el tiempo tiende a incrementarse. No obstante, resulta interesante que las variaciones en el orden de los nodos a recorrer son mínimas, lo cual puede explicarse debido a la cantidad de nodos a visitar, ya que existe evidencia significativa para asegurar que, a mayor cantidad de nodos, mayor probabilidad de variación en el orden de la ruta. A modo de ejemplo, las tiendas asignadas al CEDI de Sibaté no varían su orden ante ningún escenario, mientras que las de Funza si tienen dicha variabilidad.

Para finalizar, la representación visual de nodos del diseño de ruteo que tenía por objeto el presente estudio para la muestra seleccionada y, que el modelo TSP calculó como óptimo abarcando e integrando las probabilidades de accidentes viales en los tiempos de recorrido es el siguiente:

Ilustración 17. TSP del CEDI Sibaté

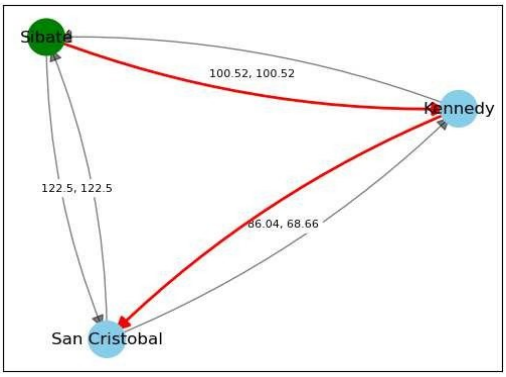
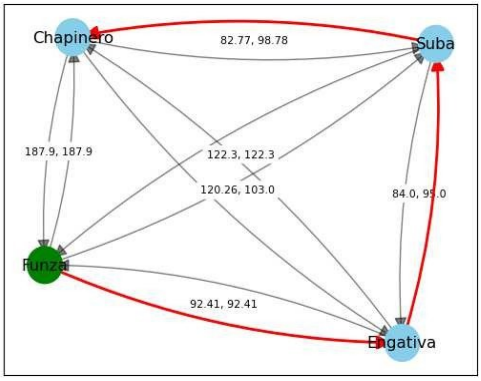


Ilustración 18. TSP del CEDI Funza



Para efectos prácticos y de claridad de visualización ante los resultados descritos, se tomó un mapa prediseñado de Moquete (2024) y se ubicaron los respectivos puntos del resultado, junto con la ruta propuesta como se detalla a continuación:

Ilustración 19. Mapa de localidades con el resultado del modelo

8. CONCLUSIONES Y TRABAJO FUTURO

De forma sucinta, el análisis descriptivo identificó las variables con mayor incidencia que generan un sesgo o diferencia en la frecuencia de accidentalidad (bien sea mayor o menor), las cuales fueron la localidad de Kennedy, los accidentes de tipo choque con clasificación de solo daños. Por su parte, en la etapa predictiva se realizó una tarea adicional en la que se implementó un modelo de asignación de CEDI's a la muestra de tiendas para reducir la cantidad de variables, descubriendo que solo era necesario abrir 2 centros de distribución. En lo que respecta al modelo seleccionado para la predicción de clase fue Random Forest ya que demostró un buen rendimiento que tuvo un promedio de validaciones de Monte Carlo del 74,9% con la métrica del F1- Score, teniendo en cuenta que los datos de la variable de salida (Severidad) estaban desbalanceados. Las predicciones de clase y su probabilidad asociada se utilizaron para calcular el tiempo de incremento del recorrido considerando un factor de peso según la severidad predicha; la fórmula utilizada que incrementa el

tiempo original tiene en cuenta una probabilidad por frecuencia en conjunto con la probabilidad predicha.

En esta misma línea, el modelo de ruteo TSP encontró la ruta óptima asociando los accidentes viales y ofreció una visualización previa de grafos con nodos que indican la ruta seleccionada y el orden de la dirección de los nodos a visitar.

Para finiquitar, es importante destacar la relevancia de este estudio, ya que además de estar actualizado, contribuye a la comprensión de la problemática y a la toma de decisiones que busquen evadir y reducir la accidentalidad vial en la capital colombiana mediante herramientas de optimización y uso aplicado de Machine Learning. La metodología empleada puede ser replicada en otras ciudades y contextos geográficos similares bajo las condiciones planteadas. De la misma manera, se precisa como investigación futura la generalización del modelo sin necesidad de muestreo de tiendas, en un escenario dinámico, con recolección de datos automatizada, variación de la temporalidad y especificaciones para el modelo de predicción según las oportunidades de mejora. Así mismo, si se desea evaluar series temporales sería interesante utilizar técnicas de modelado y análisis de un call center aplicados al estudio de accidentes viales.

8.1. REFERENCIAS

Aarón, M. A., Gómez, C. A., Fontalvo, J., & Gómez, A. J. (2019). Vehicular mobility analysis using simulation in the department of la guajira: The case of riohacha and maicao. *Informacion Tecnologica*, *30*(1), 321–332. https://doi.org/10.4067/S0718-07642019000100321

Abadi, A., Rajabioun, T., & Ioannou, P. A. (2015). Traffic Flow Prediction for Road Transportation Networks with Limited Traffic Data. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 653–662. https://doi.org/10.1109/TITS.2014.2337238

Ahmadi, A., Jahangiri, A., Berardi, V., & Machiani, S. G. (2020). Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Journal of Transportation Safety and Security*,

12(4), 522–546. https://doi.org/10.1080/19439962.2018.1505793

Alisha Sikri, N. P. Singh, Surjeet Dalal. (2023). Chi-Square Method of Feature Selection: Impact of Pre-Processing of Data. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 241–248. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2680

AlKheder, S., AlRukaibi, F., & Aiash, A. (2020). Risk analysis of traffic accidents' severities: An application of three data mining models. *ISA Transactions*, *106*, 213–220. https://doi.org/10.1016/J.ISATRA.2020.06.018

Angarita-Zapata, J. S., Maestre-Gongora, G., & Calderín, J. F. (2021). A bibliometric analysis and benchmark of machine learning and automl in crash severity prediction: The case study of three colombian cities. *Sensors*, *21*(24). https://doi.org/10.3390/s21248401

Arévalo-Támara, A., Orozco-Fontalvo, M., & Cantillo, V. (2020). Factors influencing crash frequency on colombian rural roads. *Promet - Traffic - Traffico*, *32*(4), 449–460. https://doi.org/10.7307/PTT.V32I4.3385

Basso, F., Pezoa, R., Varas, M., & Villalobos, M. (2021). A deep learning approach for real-time crash prediction using vehicle-by-vehicle data. *Accident Analysis and Prevention*, *162*. https://doi.org/10.1016/J.AAP.2021.106409

Becerra, B. X. (2022). *Bogotá es la cuarta ciudad del mundo con peor tráfico vehicular según Traffic Index*. https://www.larepublica.co/globoeconomia/bogota-cuarta-ciudad-en-el-mundo-con-el-peor-trafico-vehicular-segun-nuevo-ranking-3325102

Bustos, C., Rhoads, D., Solé-Ribalta, A., Masip, D., Arenas, A., Lapedriza, A., & Borge-Holthoefer, J. (2021). Explainable, automated urban interventions to improve pedestrian and vehicle safety. *Transportation Research Part C: Emerging Technologies*, *125*. https://doi.org/10.1016/j.trc.2021.103018

Bernal, I. (2022). Al año se pierden en promedio 74 horas por el tráfico, sostiene estudio de WeWork. *La República*. https://www.larepublica.co/empresas/al-ano-se-pierden-en-promedio-74-horas-por-el-trafico-sostiene-estudio-de-wework-3368109

Carvajal, G. A., Sarmiento, O. L., Medaglia, A. L., Cabrales, S., Rodríguez, D. A., Quistberg, D. A., & López, S. (2020). Bicycle safety in Bogotá: A seven-year analysis of bicyclists' collisions and fatalities. *Accident Analysis and Prevention*, *144*. https://doi.org/10.1016/j.aap.2020.105596

Cela, J., & Montoya-Torres, J. (2022). Predicción de la accidentabilidad en ciudades basada en datos de tráfico. *Universitat Oberta de Catalunya*.

Chaparro, M., Hernández-Vásquez, A., & Parras, A. (2018). Geospatial and environmental analysis of road traffic accidents in the city of Resistencia, Argentina. *Salud Colectiva*, *14*(1), 139–151. https://doi.org/10.18294/sc.2018.1207

Coursera. (2022). *¿Qué es el modelado estadístico?* | *Coursera*. Coursera. https://www.coursera.org/articles/statistical-modeling

Das, S., Avelar, R., Dixon, K., & Sun, X. (2018). Investigation on the wrong way driving crash patterns using multiple correspondence analysis. Accident; analysis and prevention, 111, 43–55. https://doi.org/10.1016/j.aap.2017.11.016

DNP. (2020). *DNP advierte que se avecina colapso de movilidad en las principales capitales.* | *DNP*. Departamento Nacional de Planeación GOV.CO. https://www.dnp.gov.co/Paginas/DNP%20advierte%20que%20se%20avecina%20colapso%20de%20movilidad%20en%20las%20principales%20capitales.aspx

Ghaemi, Y., El-Ocla, H., Yadav, N. R., Madana, M. R., Raju, D. K., Dhanabal, V., & Sheshadri, V. (2021). Intelligent transport system using time delay-based multipath routing protocol for vehicular ad hoc networks. *Sensors*, *21*(22). https://doi.org/10.3390/S21227706

Giripunje, L. M., Vidyarthi, A., & Shandilya, S. K. (2022). Adaptive Congestion Prediction in Vehicular Ad-hoc Networks (VANET) Using Type-2 Fuzzy Model to Establish Reliable Routes. *Wireless Personal Communications*, *125*(4), 3527–3548. https://doi.org/10.1007/S11277-022-09723-W/TABLES/7

Guerra, A., Gadhiya, V., & Srisurin, P. (2022). CRASH PREDICTION ON ROAD SEGMENTS USING MACHINE LEARNING METHODS. *ASEAN Engineering Journal*, *12*(3), 27–37. https://doi.org/10.11113/AEJ.V12.17601

Ismail, S., Dzulkifli, D. (2021). Optimization of Route Selection and Carbon Emission Release for Waste Collection Systems. Springer link. https://doi.org/10.1007/978-3-030-67307-9_12

Jiang, S., Zhang, Y., Liu, R., Jafari, M., & Kharbeche, M. (2022). Data-Driven Optimization for Dynamic Shortest Path Problem Considering Traffic Safety. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2022.3165757

Klinjun, N., Kelly, M., Praditsathaporn, C., & Petsirasan, R. (2021a). Identification of factors affecting road traffic injuries incidence and severity in southern Thailand based on accident investigation reports. *Sustainability (Switzerland)*, *13*(22). https://doi.org/10.3390/su132212467

Klinjun, N., Kelly, M., Praditsathaporn, C., & Petsirasan, R. (2021b). Identification of factors affecting road traffic injuries incidence and severity in southern Thailand based on accident investigation reports. *Sustainability (Switzerland)*, *13*(22). https://doi.org/10.3390/su132212467

Leonavičienė, T., Pukalskas, S., Pumputis, V., Kulešienė, E., & Žuraulis, V. (2020a). Investigation of factors that have affected the outcomes of road traffic accidents on Lithuanian roads. *Baltic Journal of Road and Bridge Engineering*, *15*(5), 1–20. https://doi.org/10.7250/bjrbe.2020-15.504

Leonavičienė, T., Pukalskas, S., Pumputis, V., Kulešienė, E., & Žuraulis, V. (2020b). Investigation of factors that have affected the outcomes of road traffic accidents on Lithuanian roads. *Baltic Journal of Road and Bridge Engineering*, *15*(5), 1–20. https://doi.org/10.7250/bjrbe.2020-15.504

Li, D., Wu, J., & Peng, D. (2021). Online Traffic Accident Spatial-Temporal Post-Impact Prediction Model on Highways Based on Spiking Neural Networks. *Journal of Advanced Transportation*, *2021*. https://doi.org/10.1155/2021/9290921

Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis and Prevention*, *135*. https://doi.org/10.1016/J.AAP.2019.105371

Liao, X., Zhou, T., Wang, X., Dai, R., Chen, X., & Zhu, X. (2022). Driver Route Planning Method Based on Accident Risk Cost Prediction. *Journal of Advanced Transportation*, *2022*. https://doi.org/10.1155/2022/5023052

Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. *Accident Analysis & Prevention*, *160*, 106322. https://doi.org/10.1016/J.AAP.2021.106322

Medina-Salgado, B., Sánchez-DelaCruz, E., Pozos-Parra, P., & Sierra, J. E. (2022). Urban traffic flow prediction techniques: A review. *Sustainable Computing: Informatics and Systems*, *35*. https://doi.org/10.1016/J.SUSCOM.2022.100739

Mercado, L. (2023). Zonas de Bogotá y horarios en los que se registran más accidentes de tránsito, según IA. *El Tiempo.* https://www.eltiempo.com/bogota/bogota-inteligencia-artificial-revela-zonas-y-horarios-con-mas-accidentes-de-transito-808977

Ministerio de salud. (n.d.). Ciclo de Vida. *Minsalud.* https://minsalud.gov.co/proteccionsocial/Paginas/cicloVida.aspx

Mohammed J. Zaki, Wagner Meira, Jr. (2020). Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition. *Cambridge University*. https://dataminingbook.info/book_html/

Moquete, B.M. (2024). Mapa de Bogota. Mapainteractivo.net. https://www.mapainteractivo.net/fotos/mapa-de-bogota.html.

NetworkX. (2023). Introduction NetworkX Library. *NetworkX documentation.* https://networkx.org/documentation/stable/reference/introduction.html

OMS. (2022). *Traumatismos causados por el tránsito*. Organización Mundial de La Salud. https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., Berrio Garcia, S., Barrero, L. H., & Sana, S. S. (2021). Extraction of decision rules using genetic algorithms and simulated annealing for prediction of severity of traffic accidents by motorcyclists. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10051–10072. https://doi.org/10.1007/S12652-020-02759-5

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., & Sana, S. S. (2021a). Prediction of motorcyclist traffic crashes in Cartagena (Colombia): Development of a safety performance function. *RAIRO - Operations Research*, *55*(3), 1257–1278. https://doi.org/10.1051/ro/2021055

Ospina-Mateus, H., Quintana Jiménez, L. A., Lopez-Valdes, F. J., & Sana, S. S. (2021b). Prediction of motorcyclist traffic crashes in Cartagena (Colombia): Development of a safety performance function. *RAIRO - Operations Research*, *55*(3), 1257–1278. https://doi.org/10.1051/RO/2021055

Park, H., Haghani, A., Samuel, S., & Knodler, M. A. (2018). Real-time prediction and avoidance of secondary crashes under unexpected traffic congestion. *Accident Analysis and Prevention*, *112*, 39–49. https://doi.org/10.1016/J.AAP.2017.11.025

Perafan-Villota, J. C., Mondragon, O. H., & Mayor-Toro, W. M. (2022). Corrections: Fast and precise: Parallel processing of vehicle traffic videos using big data analytics (IEEE Transactions on Intelligent Transportation Systems (2021) DOI: 10.1109/TITS.2021.3109625). *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 13911. https://doi.org/10.1109/TITS.2021.3116636

Rashidi, M. H., Keshavarz, S., Pazari, P., Safahieh, N., & Samimi, A. (2022). Modeling the accuracy of traffic crash prediction models. *IATSS Research*. https://doi.org/10.1016/J.IATSSR.2022.03.004

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–269. https://doi.org/10.1016/J.JSR.2021.12.007

Sattar, K., Chikh Oughali, F., Assi, K., Ratrout, N., Jamal, A., & Masiur Rahman, S. (2022). Transparent deep machine learning framework for predicting traffic crash severity. *Neural Computing and Applications*. https://doi.org/10.1007/S00521-022-07769-2

SDM. (2022). *Siniestros Viales Bogota Accidente*. Datos Abiertos Secretaría Distrital de Movilidad. https://datos.movilidadbogota.gov.co/datasets/movilidadbogota::siniestros-viales-bogota-accidente/explore?location=4.456827%2C-74.117650%2C9.98

Semana. (2022). *¡A 2 kilómetros por hora! Los trancones de Bogotá cada día son peores*. Revista Semana. https://www.semana.com/nacion/articulo/a-2-kilometros-por-hora-los-trancones-de-bogota-cada-dia-son-peores/202200/

Shiran, G., Imaninasab, R., & Khayamim, R. (2021). Crash severity analysis of highways based on multinomial logistic regression model, decision tree techniques and artificial neural network: A modeling comparison. *Sustainability (Switzerland)*, *13*(10). https://doi.org/10.3390/SU13105670

Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Machine Learning: Fundamental Concepts and Algorithms, 2nd Edition, Cambridge University Press, March 2020. ISBN: 978-1108473989.

Siamidoudaran, M., & Iscioglu, E. (2019). Injury severity prediction of traffic collision by applying a series of neural networks: The city of London case study. *Promet - Traffic – Traffico*. https://doi.org/10.7307/PTT.V31I6.3032

Terán, J., Navarro, L., Quintero M., C. G., & Pardo, M. (2020). Intelligent driving assistant based on road accident risk map analysis and vehicle telemetry. *Sensors (Switzerland)*. https://doi.org/10.3390/s20061763

Universidad de Granada (2020). Algoritmo Greedy. *DECSAI*. https://elvex.ugr.es/decsai/algorithms/slides/4%20Gredy.pdf

Vásquez, J. M. (2021). *Conoce el modelo de movilidad sostenible que promueve el POT | Bogota.gov.co*. https://bogota.gov.co/mi-ciudad/planeacion/conoce-el-modelo-de-movilidad-sostenible-que-promueve-el-pot

Xie, S., Ji, X., Yang, W., Fang, R., & Hao, J. (2020a). Exploring Risk Factors with Crash Severity on China Two-Lane Rural Roads Using a Random-Parameter Ordered Probit Model. *Journal of Advanced Transportation*, *2020*. https://doi.org/10.1155/2020/8870497

Xie, S., Ji, X., Yang, W., Fang, R., & Hao, J. (2020b). Exploring Risk Factors with Crash Severity on China Two-Lane Rural Roads Using a Random-Parameter Ordered Probit Model. *Journal of Advanced Transportation*, *2020*. https://doi.org/10.1155/2020/8870497

Yang, Y., Wang, K., Yuan, Z., & Liu, D. (2022). Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction. *Journal of Advanced Transportation*, *2022*. https://doi.org/10.1155/2022/4257865

Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access*, *6*, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979

Zhang, K., Hassan, M., Yahaya, M., & Yang, S. (2018). Analysis of Work-Zone Crashes Using the Ordered Probit Model with Factor Analysis in Egypt. *Journal of Advanced Transportation*, *2018*. https://doi.org/10.1155/2018/8570207