Modelos de aprendizaje computacional para la predicción de siniestralidad vial en Bogotá D.C.

Sebastián Rodrigo Castellanos Cardozo

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Bogotá D.C., Colombia
2023

Modelos de aprendizaje computacional para la predicción de siniestralidad vial en Bogotá D.C.

Sebastián Rodrigo Castellanos Cardozo

Trabajo de grado presentado como requisito para optar al título de:
Ingeniero de Sistemas y Computación

Director:
Juan Camilo Ramírez, Ph.D.
Asesora:
Rosalba Cruz, Esp

Línea de Investigación:
Inteligencia Computacional
Grupo de Investigación:
LACSER - Laboratory for Advanced Computational Science and Engineering Research

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Bogotá D.C., Colombia
2023

(Dedicatoria)

A mis padres, hermanos y mi niño de hace 10 años que añoraba triunfar con la informática.

Agradecimientos

A mi director de proyecto de grado Juan Camilo Ramírez, que me brindó toda la ayuda necesaria para la realización de este trabajo. También, quiero agradecer al rapero, poeta y sociólogo Ignacio José Fornés Olmo a.k.a Nach, por acompañarme en los momentos de mayor incertidumbre y bloqueo mental. Sin más que añadir, de la manera más visceral y explícita en la escritura (como lo puede llegar a ser la poesía y que muchas veces es necesaria para nuestra existencia como seres racionales, pensantes y sentimentales), quiero terminar con dos versos de su libro 'Hambriento':

≪Quiero que mi cama sea una aeronave
y elija espacios y tiempos imposibles.
Que Dalí pinte las paredes de mi mente
mientras Morfeo me sacude divertido.

Quiero soñarlo todo y al revés,
y que me sueñe el mundo esta noche.
Por eso mientras duerma seré dos,
porque estaré aquí y también muy lejos.≫

Nach

Resumen

Actualmente los siniestros viales son un problema que afectan negativamente el ámbito socioeconómico de cualquier país o ciudad, por lo cual todos los gobiernos buscan soluciones factibles que les permitan disminuir el riesgo de mortalidad y accidentalidad. Las soluciones que han implementado muchos gobiernos se basan en la resolución del artículo 74/299 "Mejoramiento de la seguridad vial", proclamando el Decenio de Acción de Seguridad Vial 2021-2023, realizado en la Asamblea General de las Naciones Unidas. Gobiernos como el de Bogotá D.C. obtuvieron resultados favorables, que ayudaron a disminuir el riesgo de accidentalidad vial casi en un 50 %. Aunque estas medidas resultan ser eficientes, la gran cantidad de factores que potencialmente contribuyen a la ocurrencia de cada accidente vial puede ser grande, por lo cual no es un problema trivial anticipar o predecir cada uno de estos siniestros, aunque esto sería deseable para poder diseñar medidas preventivas efectivas. Para abordar este problema se pueden utilizar modelos predictivos empleando técnicas de aprendizaje automático que permitan estimar el riesgo de ocurrencia de estos siniestros a partir de sus factores contribuyentes potenciales. Actualmente, no existe un modelo de aprendizaje automático que permita predecir al ciento por ciento el riesgo de accidentalidad en Bogotá D.C. a partir de datos georreferenciados (por ejemplo, zona geográfica, hora, características del vehículo, *etc.*), por ello, este proyecto de grado que forma parte del proyecto de investigación *'Implementation of an intelligent transportation system (ITS) for the prediction of traffic accident risk in Bogotá D.C., Colombia'*, actualmente en ejecución en la Facultad de Ingeniería de Sistemas de la Universidad Antonio Nariño. En el proyecto de grado se desarrollaron modelos de aprendizaje automático que contribuyen al propósito del proyecto de investigación de predecir estos fenómenos a partir de sus factores geoespeaciales (como lo pueden ser el clima, estado de la carretera, diseño de la vía, *etc.*).

Palabras clave: Aprendizaje automático, modelos de predicción, siniestralidad vial, redes neuronales, bosques aleatorios, máquinas de vectores de soporte.

Abstract

Currently road accidents are a problem that negatively worsens the socioeconomic environment of any country or city, which is why all governments seek feasible solutions that allow them to reduce the risk of mortality and accidents. The solutions that many governments have implemented are based on the resolution of article 74/299 "Improvement of road safety", proclaiming the Decade of Action for Road Safety 2021-2023, held at the United Nations General Assembly. Although the measures of governments such as that of Bogotá D.C. turn out to be efficient, the large number of factors that potentially contribute to the occurrence of each road accident can be large, so it is not a trivial problem to anticipate or predict each of these accidents, although this would be desirable in order to design effective preventive measures. To address this problem, predictive models can be used using machine learning techniques that allow estimating the risk of occurrence of these casualties based on their potential contributing factors. Currently, there is no machine learning model that allows one hundred percent to predict the risk of accidents in Bogotá D.C. from georeferenced data (for example: geographical area, time, vehicle characteristics, *etc.*), therefore, this degree project that is part of the research project 'Implementation of an intelligent transportation system (ITS) for the prediction of traffic accident risk in Bogotá D.C., Colombia', currently running at the Faculty of Systems Engineering of the Antonio Nariño University, it intends to develop automatic learning models that can be used as a tool for road authorities in decision-making, in order to avoid the risk of deaths or injuries due to these casualties.

Keywords: Machine learning, prediction models, road accident rate, neural networks, random forests, support vector machines.

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

Los siniestros viales representan un grave problema de salud pública y son una de las principales causas de muertes y lesiones en el mundo, dejando cerca de un 1,3 millones de personas fallecidas y 50 millones más con lesiones y heridas graves en el mundo anualmente (Naciones Unidas, 2023). Actualmente, los gobiernos de todo el mundo buscan asegurar que todos los campos de la sociedad puedan estar involucrados en las acciones relacionadas con la seguridad vial e impulsar nuevas políticas y medidas para reducir las muertes y lesiones, para alcanzar una meta en común: disminuir a la mitad el número de víctimas mortales y lesionadas para el año 2030 (Naciones Unidas, 2023). Con esto, se llevó a cabo por medio de la resolución 74/299 de la Asamblea General de las Naciones Unidas, la realización de un Segundo Decenio de Acción para la Seguridad vial 2021-2030, con la finalidad de lograr reducir las defunciones y traumatismos que generan los siniestros viales en un 50 % durante este período (Organización Mundial de la Salud, 2021). Los países y ciudades orientados por este plan están obteniendo reducciones en las muertes, como, por ejemplo, Bogotá D.C. consiguió reducir a la mitad las muertes en diez años, todo esto gracias a la toma de acciones integradas como: mejoras técnicas de gran alcance y cambios en las reformas regulatorias (Naciones Unidas, 2023). La Agencia Nacional de Seguridad Vial, a través del Observatorio Nacional de Seguridad Vial reportó durante los primeros seis meses del año 2022 308 muertes ocasionadas por siniestros viales en Bogotá D.C., y un total de 612 fallecimientos en todo el año. Lo que muestra un aumento de 27 personas en comparación con el mismo período del año anterior (Malaver, 2023). Aunque los cambios en las reformas regulatorias tuvieron un cambio favorable en la disminución de mortalidad de siniestros viales, todavía hay factores que no se contemplan como cambios geotemporales, los diferentes tipos de actores viales, *etc.*

Los factores geotemporales (como pueden llegar a ser el clima o el estado de las vías) que potencialmente contribuyen a la ocurrencia de cada siniestro vial pueden ser muy numerosos y tener asociado un gran volumen de datos, lo cual hace necesario el uso de modelos predictivos utilizando técnicas de aprendizaje automático para su análisis con el propósito de estimar el riesgo de ocurrencia de estos fenómenos (Díaz, 2021).

Este proyecto de grado tiene como propósito entrenar varios modelos para predecir el riesgo de accidentalidad en las vías de Bogotá D.C. haciendo uso del aprendizaje automático supervisado. Los datos que fueron usados como parte del proyecto de grado que forma parte del proyecto de investigación *'Implementation of an intelligent transportation system (ITS)*

for the prediction of traffic accident risk in Bogotá D.C., Colombia' de la Universidad Antonio Nariño, tienen como punto de partida los datos de acceso público suministrados por el gobierno distrital. Posteriormente estos datos fueron utilizados para entrenar y evaluar varios modelos de aprendizaje automático con el propósito de determinar la factibilidad del uso de este tipo de herramientas para apoyar el propósito del proyecto de investigación de predecir estos siniestros a partir de sus factores geoespaciales.

El resto del documento está organizado de la siguiente manera: El Capítulo 2 contiene el planteamiento del problema, donde se abarca la formulación del problema, la pregunta de investigación, la justificación, el objetivo general y los objetivos específicos, y los alcances y limitaciones del mismo. El Capítulo 3 contiene el marco de referencia, donde se abarca el marco teórico, el estado del arte y el marco legal. El Capítulo 4 contiene la metodología, donde se explica detallamente el paso a paso del proceso que se realizó para el preprocesamiento de los datos, el entrenamiento de los modelos y la elección del mejor modelo y por último, el Capítulo 5 contiene las conclusiones.

2 Planteamiento del problema

2.1. Formulación del problema

Los siniestros viales provocan graves pérdidas, tanto humanas como económicas, configurándose en un problema de alta complejidad para grandes zonas urbanas alrededor del mundo, incluyendo a Bogotá D.C., Colombia, por lo cual su reducción significativa es una de las metas contempladas en el objetivo once 'Ciudades y comunidades sostenibles', que hace parte de los objetivos de desarrollo sostenible, que pertenecen a la agenda promulgada por las Naciones Unidas (2023). La predicción precisa de los accidentes de tráfico tiene un gran potencial para ofrecer protección a la seguridad pública y pérdidas económicas. Aunque hay un conjunto amplio de estrategias gubernamentales adoptadas tradicionalmente para la prevención de estos incidentes, la comprensión exhaustiva de este fenómeno, incluyendo todos sus factores contribuyentes, como el comportamiento de los actores viales, las condiciones climáticas, el estado del vehículo o de la vía, *etc.*, es objeto de investigación para lograr la predicción posterior del mismo, y, así, también medidas preventivas más eficaces que las que se tienen actualmente (Khayesi y Meleckidzedeck, 2021). Sin embargo, esta predicción no es trivial debido a la compleja causalidad de los accidentes de tráfico con múltiples factores geotemporales, incluidas las correlaciones espaciales, las interacciones dinámicas temporales, las variaciones por tipo de usuarios y las influencias externas en los datos heterogéneos relevantes para el tráfico (Khayesi y Meleckidzedeck, 2021).

En la literatura académica no hay modelos de aprendizaje automático para la predicción de siniestralidad en Bogotá D.C. utilizando datos actualizados hasta el año 2023. El proyecto de grado pretende ayudar a respaldar otras investigaciones previas en lo que respecta a la predicción de siniestros viales principalmente en Bogotá D.C., entrenando modelos de aprendizaje automático con bases de datos públicas provenientes de las autoridades gubernamentales que permitan suministrar la información necesaria para la toma de decisiones. Aunque todos los modelos entrenados permiten suministrar información relevante se debe escoger el modelo con el mejor desempeño que permita prevenir los posibles incidentes en la vía. Además, de permitir contribuir al propósito del proyecto de investigación, el cual se centra en realizar la correcta predicción de estos siniestros a partir de sus diferentes factores geoespaciales.

2.2. Pregunta de investigación

¿Qué técnica de aprendizaje automático produce el modelo de predicción del riesgo de siniestralidad vial en Bogotá D.C. con el mejor desempeño, en términos del puntaje F1 o la curva ROC, a partir de la información geotemporal asociada a cada accidente utilizando datos actualizados hasta 2023?

2.3. Justificación

Los accidentes de tráfico son la principal causa de muerte entre las personas de 20 a 40 años en la ciudad de Bogotá D.C. según el reporte de Fallecidos y Lesionados 2022 de la Agencia Nacional de Seguridad Vial ofrecida por el Ministerio de Transporte (Agencia Nacional de Seguridad Vial, 2022). Tan solo en el mes de diciembre de 2022, 591 personas fallecieron a causa de siniestros viales. En los cuales se encuentran diferentes tipos de usuario, donde destacan los motociclistas con una cifra de 239 fallecidos que corresponden al 39,67 % de la Participación Actor Vial del año 2022, seguidos por los peatones con 218 fallecidos que corresponden al 36,14 %, ciclistas con 94 fallecidos que corresponden al 15,55 %, conductores de transporte de carga con 7 fallecidos que corresponden al 1,16 % y el restante 7,48 % que corresponde a pasajeros de transporte público y otros (Agencia Nacional de Seguridad Vial, 2022). Esta problemática no solo provoca graves pérdidas humanas, sino también genera un gran impacto en la economía de Bogotá D.C., ya que la repercusión que generan los siniestros viales sobre la economía asciende a los 23,9 billones de pesos anuales, lo que equivaldría al 3,6 por ciento del Producto Interno Bruto de Colombia (Portafolio.co, 2020). La seguridad vial es una de las principales metas que hacen parte del objetivo once 'Ciudades y comunidades sostenibles', promulgados por las Naciones Unidas (Naciones Unidas, 2023). Aunque sea una meta contemplada a largo plazo, hoy en día no existen modelos de aprendizaje automático que permitan predecir el riesgo de accidentalidad vial en Bogotá D.C. utilizando datos actualizados hasta 2023. Con este proyecto de grado se pretende mediante la implementación de estos en un software con datos obtenidos de la exploración, análisis e investigación preliminar de las bases de datos históricas de siniestralidad vial pueda ser usado por entidades e instituciones como el Observatorio Nacional de Seguridad Vial como herramienta para tener el registro y análisis histórico de siniestros viales en Bogotá D.C para poder tomar posibles elecciones que ayuden a prevenir, controlar y disminuir el riesgo de muerte o lesión, permitiendo, así, tener un impacto positivo socioeconómico para Bogotá D.C. Además, de permitir sentar las bases para usar en futuros modelos, que, usando los modelos empleados en este proyecto de grado puedan ser mejorados para permitir ser usados no solo en Bogotá D.C., sino en todas las ciudades que estén afectadas por problemas similares.

2.4. Objetivos

2.4.1. Objetivo general

Diseñar y evaluar modelos de aprendizaje automático para la estimación del riesgo de accidentalidad vial en Bogotá D.C. utilizando datos históricos de accidentes actualizados hasta 2023 obtenidos de fuentes distritales de acceso público para identificar la técnica de mayor confiabilidad utilizando métricas de desempeño empleadas en la literatura científica, incluyendo la curva ROC y el puntaje F1.

2.4.2. Objetivos específicos

- Organizar un historial de accidentes automovilísticos en Bogotá D.C. a partir de los datos públicos actualizados hasta 2023 del gobierno distrital, utilizando técnicas de preprocesamiento de datos para usarse en el posterior entrenamiento de los modelos predictivos de riesgo de siniestralidad vial.
- Diseñar modelos de aprendizaje automático para determinar el de mejor desempeño, utilizando modelos computacionales predictivos sobre el conjunto de datos previamente construido.
- Ejecutar un plan de evaluación de los modelos predictivos implementados, utilizando métricas de desempeño reportadas en la literatura científica para estimar la confiabilidad de estos.

2.5. Alcances y limitaciones del proyecto

2.5.1. Alcances

- La realización del diseño e implementación de los diferentes clasificadores (incluyendo redes neuronales y bosques aleatorios) se hicieron usando bibliotecas conocidas de libre distribución en el lenguaje Python, como SciPy o Scikit-Learn.
- La evaluación de los datos se dará por distintos modelos entre los cuales estuvieron: SVM (Support Vector Machines), redes neuronales, bosques aleatorios, entre otros.
- Analizar y evaluar los modelos definidos bajo el uso de métricas como: recall, puntuación F1, área bajo la curva ROC, entre otras.

2.5.2. Limitaciones

- No se tiene contemplado la recopilación de datos, dado que se utilizarán datos de acceso público.

- No se realizará soporte y/o mantenimiento sobre el código una vez se realice la entrega del proyecto de grado.

3 Marco de referencia

3.1. Marco teórico

3.1.1. Aprendizaje automático

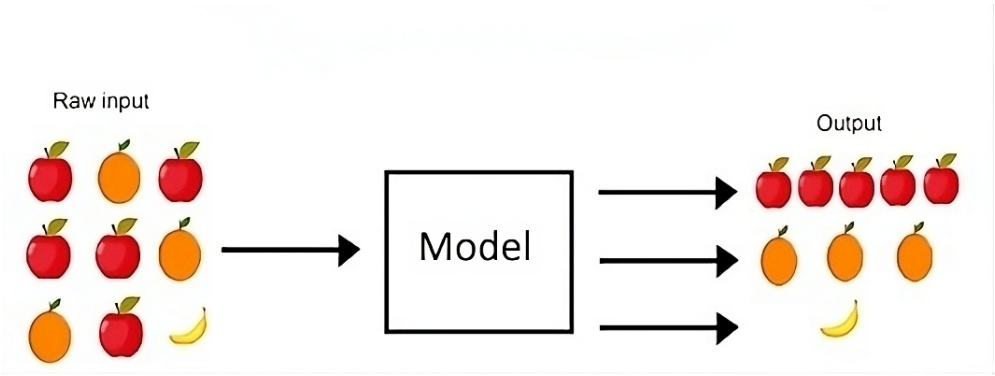
El aprendizaje automático es una rama de la Inteligencia Artificial, donde se desarrollan modelos entrenados mediante ciclos de retroalimentación que le permite aprender de la experiencia adquirida (Norman, 2021). Haciendo uso de un determinado tipo de aprendizaje, la máquina puede ser capaz de resolver el problema, y asimismo poderlo reconocer, para tener la capacidad de suministrar una solución con todo lo aprendido (Norman, 2021). Todos estos cambios por los que pasa el modelo al ser entrenado no solo se refieren al perfeccionamiento de sus capacidades y habilidades para realizar tareas específicas, sino que también pueden involucrar transformaciones en la representación de hechos del propio sistema.

Con esto, se permite que se puedan generar algoritmos que tengan la capacidad de aprender sin necesidad de una programación previa explícita. El desarrollador ya no tendrá la necesidad de tener que desarrollar la lógica para cada uno de los escenarios posibles que pueda tener el entorno en el cual se está trabajando, sino que lo único que tendrá que hacer es entrenar al algoritmo con un base de datos que contenga una gran cantidad de información para que tenga la posibilidad de aprender por su cuenta (Sandoval, 2018). Lo cual resulta fundamental para los modelos a desarrollar en este proyecto de grado, ya que, un humano promedio no tendría ni el tiempo ni la capacidad necesaria para la gran afluencia en el procesamiento de datos, operaciones, *etc.*

En este contexto, surgen dos tipos de aprendizaje automático, los cuales son: el supervisado y el no supervisado. El aprendizaje supervisado es aquel en el que se entrena un modelo predictivo, otorgándole características (preguntas) y etiquetas (respuestas) para que por su propia cuenta realice una predicción a futuro, con las características previas usadas. Por ejemplo, la clasificación es un tipo de aprendizaje supervisado (el cual se usó en el desarrollo de los modelos de este proyecto de grado) en el cual se recibe una entrada con información y a partir de esta información es capaz de clasificar ya sea por una o más características. Para entender de una manera más explícita lo previamente mencionado se tiene el siguiente ejemplo: un conjunto de datos de frutas {(manzana, “fruta roja”), (naranja “fruta naranja”), (banana, “fruta amarilla”)}, donde el modelo clasificador previamente entrenado es capaz

de clasificar las frutas en grupos, como lo muestra la Figura **3-1**.

Figura 3-1: Ejemplo de modelo clasificador usando aprendizaje supervisado, clasificando el conjunto de datos por tipo de frutas.

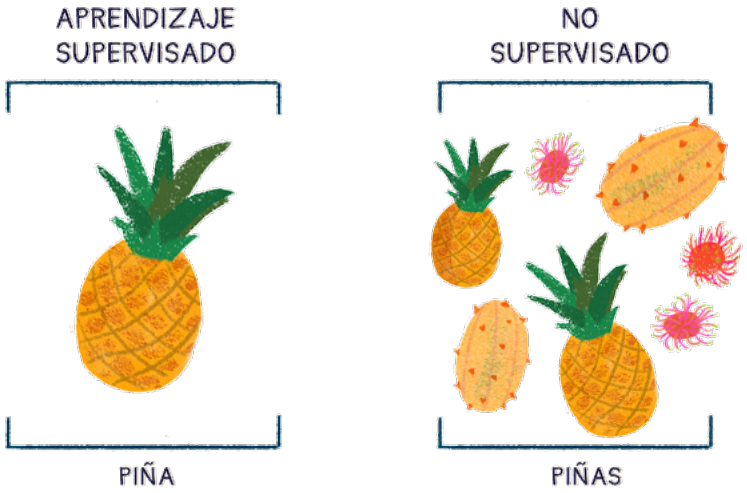


Fuente: Obtenido de pyCodeMates (2023).

Teniendo este conjunto de datos, el algoritmo puede clasificarlos mediante aprendizaje automático, con el objetivo de entrenar y desarrollar un modelo que le permita predecir la etiqueta correspondiente a la nueva fruta que será añadida al conjunto de datos previamente obtenido (Bobadilla, 2020).

Por otro lado, el aprendizaje no supervisado es aquel al que solo se le concede al algoritmo las características, mas no las etiquetas. Con este enfoque, se pretende una agrupación de datos obtenidos de las características previamente suministradas. Así, el algoritmo puede clasificar los datos de acuerdo con las características que puedan compartir con los grupos, asumiendo que posiblemente pueda pertenecer al mismo grupo, como por ejemplo, agrupar productos en un comercio electrónico, clasificar tipos de clientes en un servicio online, *etc.* (Sandoval, 2018). En la Figura **3-2** se puede observar la diferencia entre un modelo que usó aprendizaje supervisado y otro que usó aprendizaje no supervisado.

Figura 3-2: Diferencia entre un modelo que usó aprendizaje supervisado y no supervisado, clasificando la fruta de acuerdo al método de clasificación de cada modelo.



Fuente: Obtenido de Google (2023).

En la Figura **3-2**, se está entrenando un modelo de aprendizaje automático que sea capaz de reconocer diferentes frutas. Para hacer esto, el modelo podría ser entrenado mostrándole fotos de piñas que estén etiquetadas como frutas. Al mostrarle estas imágenes, el modelo al usar aprendizaje supervisado puede aprender a asociar las características específicas de las piñas, como su forma y textura rugosa con pinchos (Google, 2023). Posteriormente, cuando se encuentra con una fruta similar en un frutero, el modelo de aprendizaje supervisado debería ser capaz de identificarla correctamente basándose en el conocimiento previo que adquirió durante el entrenamiento con las imágenes de las piñas. En cambio, el modelo al usar aprendizaje no supervisado agrupa objetos similares visualmente, como pueden ser las bolas rugosas con pinchos, pero que no puede lograr identificarlos como piñas sin alguna identificación o etiqueta explícita. A pesar de ser capaz de reconocer similitudes visuales, el modelo requiere de información adicional para ser capaz de asignar nombres precisos a las categorías (Google, 2023).

Para la creación de los modelos de aprendizaje automático que fueron usados en este proyecto de grado resulta relevante el aprendizaje supervisado, ya que mediante el entrenamiento de los modelos se pueden clasificar los factores asociados a los siniestros viales.

3.1.2. Variable independiente y dependiente

La variable independiente es aquella variable que el investigador puede poner a prueba para comprobar o demostrar una hipótesis. Esta puede ser una cualidad, característica o propiedad que puede afectar a las variables dependientes, haciendo que estas puedan llegar a ser modificadas (Mimenza, 2019).

La variable dependiente es aquella cualidad o característica que es modificada por la variable independiente (Mimenza, 2019).

Para entender de una manera más explícita las variables independientes y dependientes se usarán en un hipotético caso de siniestro vial posibles variables relacionadas a este proyecto de grado como ejemplo. En este caso, se define como variable dependiente la gravedad del siniestro vial y las variables independientes serán: la velocidad del vehículo (velocidad a la que se desplazaba en el momento del aciddente), el clima (la condición o condiciones meteorológicas en el momento del accidente como lo puede ser la niebla, lluvia, *etc.*) y el tipo de vehículo (carro, moto, cicla, bús, *etc.*).

3.1.3. Normalización

Es una técnica que se aplica en la fase de preparación de los datos, y que se usa para facilitarle el aprendizaje a los modelos. Aquí, los datos usados para el entrenamiento se colocan en una escala similar o común, sin cambiar las diferencias entre los intervalos originales. Por ejemplo, suponiendo que tenemos un conjunto de datos de entrada donde este cuenta con una columna con valores entre 0 y 1, y otra columna con valores entre 100.000 y 1.000.000, a simple vista, se puede observar la notable diferencia entre la escala de estos números. Esto puede traer consigo problemas al tratar de realizar la combinación de los valores. Para evitar esto, se normalizan los datos con una escala en común, donde se realiza la creación de nuevos valores que permitan mantener las relaciones originales de los datos, y a su vez, que también permitan conservar valores dentro de la escala, que se aplica en todas las columnas numéricas que posteriormente usará el modelo (Microsoft, 2022). En este caso, la clase *StandardScaler* de la biblioteca Scikit-Learn permite realizar esta normalización.

3.1.4. Positivo y negativo en modelos de aprendizaje automático

En problemas de clasificación binaria (aquella tarea que divide un conjunto de datos en dos grupos o clases), es decir, aquellos donde solamente hay dos clases o categorías, es común referirse a una de estas como la "positiva" y la otra como la "negativa", sin que estos términos impliquen necesariamente una preferencia por una o por la otra. Por lo tanto, la clase positiva y negativa se refiere a la forma en que el modelo puede etiquetar los conjuntos de datos. Por ejemplo, se puede definir como clase positiva "El correo es spam", "Es un perro" o "Está enfermo", y como clase negativa "El correo no es spam", "No es un perro" o "No está enfermo" (Google, 2022).

3.1.5. Técnica de reducción de dimensionalidad

Es aquel proceso necesario en el aprendizaje automático que ayuda a reducir el número de las variables aleatorias o características de un conjunto de datos que no estén proporcinando

información nueva para el modelo. Este proceso resulta relevante, ya que ayuda a reducir la redundancia de los modelos de aprendizaje automático al eliminar características irrelevantes, y a su vez también ayuda a mejorar el rendimiento del mismo (Samina y cols., 2014).

3.1.6. Métricas básicas de desempeño en modelos de aprendizaje automático

En las ecuaciones posteriores a este ítem del documento se usan nombres genéricos de la nomenclatura inglesa, los cuales son:

- ***True Negative*** [TN]: Estos valores son identificados o clasificados por el algoritmo como negativos (en este caso se le asignan el valor de 0) y que realmente son negativos.
- ***True Positive*** [TP]: Estos valores son identificados o clasificados por el algoritmo como positivos y que realmente sí son positivos.
- ***False Positive*** [FP]: Estos valores con identificados o clasificados por el algoritmo como positivos, pero que realmente no lo son.
- ***False Negative*** [FN]: Es lo contrario a los falsos positivos. Es decir, son aquellos valores que el algoritmo identifica o clasifica como negativos, pero realmente son positivos.

En la Figura **3-3** se puede observar en un ambiente cotidiano el positivo y negativo.

Figura 3-3: Ejemplo cotidiano de los diferentes tipos de resultados en un modelo de aprendizaje automático.

Fuente: Obtenido de Vaid (2019).

3.1.7. Matriz de confusión

Es un herramienta indispensable en modelos de aprendizaje automático supervisado. En ella se representa en forma de tabla todas las predicciones que el algoritmo logró reconocer en relación con resultados reales. Gracias a esto, permite analizar de manera eficiente el desempeño del algoritmo al realizar la tarea de clasificación, identificando en el proceso los aciertos y errores que obtuvo para su posterior corrección al ser entrenado (Heras, 2020).

3.1.8. Precisión

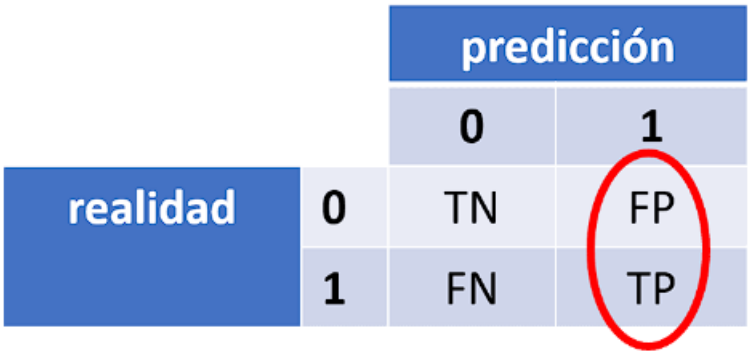
Es una métrica que permite medir la calidad del modelo de aprendizaje automático empleado, enfocándose exclusivamente en las tareas de clasificación (Heras, 2020). Además, de que esta métrica tiene la particularidad de ser útil cuando se requiere minimizar los falsos positivos. En la ecuación 3-1 se puede observar cómo se calcula la precisión:

$$precision = \frac{TP}{TP - FP} \tag{3-1}$$

Los valores obtenidos se representan mediante una matriz de confusión, la cual sirve como ayuda para evaluar y analizar los resultados de un modelo o algoritmo de aprendizaje automático supervisado, para así determinar los aciertos o errores obtenidos previamente del aprendizaje con los datos suministrados. En las columnas de la tabla se podrá visualizar el número de predicciones, mientras que en las filas el número real de las instancias (Arce, 2019).

En la Figura **3-4** se puede observar la demostración de precisión en la matriz de confusión.

Figura 3-4: Matriz de confusión con resultados obtenidos a partir del uso de la métrica de precisión.



Fuente: Obtenido de Heras (2020).

3.1.9. Recall

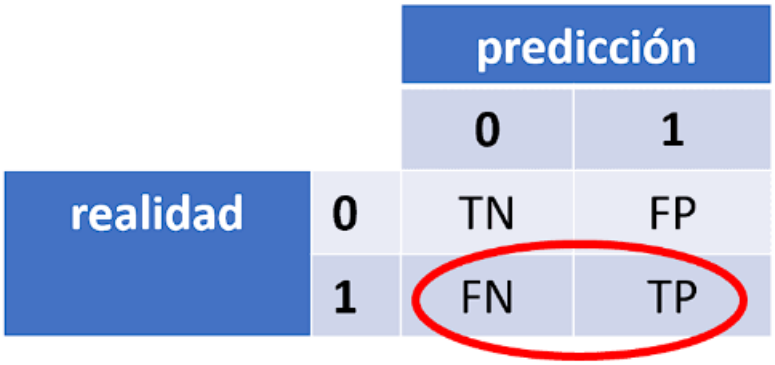
Es una métrica que informa sobre la relación de ejemplos positivos que están identificados correctamente por el modelo predictivo de aprendizaje automático (Gupta y cols., 2021). Esta métrica tiene la particularidad de ser especialmente útil cuando se requiere minimizar los falsos negativos.

En la ecuación 3-2 se puede observar cómo se calcula el *recall*:

$$recall = \frac{TP}{TP + FN} \tag{3-2}$$

En la Figura **3-5** se puede observar la demostración de recall en la matriz de confusión.

Figura 3-5: Matriz de confusión con resultados obtenidos a partir del uso de la métrica *recall*.



Fuente: Obtenido de (Heras, 2020).

3.1.10. Puntuación F1

La puntuación F1 es una métrica de aprendizaje automático, donde se utiliza un algoritmo que permita asignar con precisión los datos para posteriormente etiquetarlos y definirlos (Lipton y cols., 2014). Esta métrica se usa para combinar medidas de precisión y *recall* en un mismo valor, lo cual es útil cuando se desea equilibrar ambas medidas (Heras, 2020).

En la ecuación 3-3 podemos observar cómo se calcula la puntuación F1:

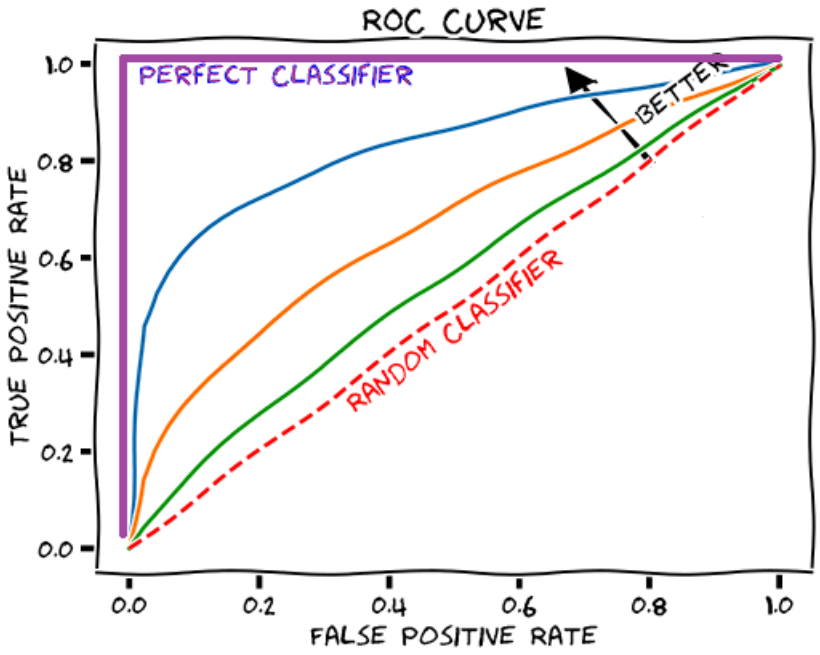
$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3-3}$$

3.1.11. Curva ROC

La curva ROC es una representación gráfica que se usa para evaluar el rendimiento de los diferentes modelos. En ella, se tienen tasas de verdaderos positivos, las cuales especifican la tasa a la que el modelo logra predecir como positivo y que realmente es positivo. En cambio, las tasas de falsos positivos especifican la tasa a la que el modelo predice como positivo cuando en realidad es negativo (MathWorks, 2023). Con esto, se pretende obtener una tasa de falsos positivos en 0, mientras que los verdaderos positivos tienden aproximarse a 1, que es donde se debería situar el modelo perfecto (J. M. Pérez y Martin, 2023). De esta manera, se reconoce si el modelo está realizando el correcto procesamiento de los datos. La curva ROC además permite evaluar el desempeño y calidad de un clasificador binario y también de un clasificador multiclase.

En la Figura **3-6** se puede observar un ejemplo de la curva ROC.

Figura 3-6: Ejemplo curva ROC.



Fuente: Obtenido de Sanahuja (2021).

En la Figura **3-6** se pueden observar cinco curvas ROC de diferente color (Rojo, verde, naranja, azul y morado). El AUROC para una curva dada es simplemente el área debajo de ella. El peor AUROC es 0.5 y el mejor AUROC es 1.0.:

- Un AUROC de 0.5 (área bajo la línea punteada roja) corresponde a un lanzamiento de moneda, es decir, un modelo prácticamente inútil.

- Un AUROC menor a 0.7 es un rendimiento subóptimo.
- Un AUROC de 0.70 a 0.80 es un buen rendimiento.
- Un AUROC mayor a 0.8 es un rendimiento excelente.
- Un AUROC de 1.0 (área bajo la línea morada) corresponde a el clasificador ideal o perfecto.

3.1.12. Python

Es un lenguaje de programación de alto nivel que fue desarrollado por Guido van Rossum. Es conocido, también por ser un lenguaje multiparadigma, ya que permite la programación orientada a objetos (POO), programación imperativa y programación funcional (I. C. Pérez y cols., 2014).

Gracias a su amplio soporte por parte de la comunidad, Python puede usarse para desarrollar interfaces gráficas, librerías matemáticas, desarrollo de software a nivel de aplicaciones de escritorio y web, creación de videojuegos, Inteligencia Artificial, *etc.* En el caso particular de este proyecto de grado, se usó Python para el desarrollo de modelos de aprendizaje automático, el cual, al tener una gran soporte y librerías especializadas en el procesamiento de datos como Panda o Scikit-Learn facilita el desarrollo y evaluación de modelos de aprendizaje capaces de predecir el riesgo de accidentalidad en Bogotá D.C., Colombia.

3.1.13. Scikit-Learn

Scikit-Learn es una biblioteca de Python, la cual abarca una extensa gama de algoritmos de aprendizaje automático, incluyendo: clasificación, regresión, reducción de dimensionalidad y agrupamiento. Además, de que brinda los módulos necesarios para el preprocesamiento de datos, la posibilidad de extraer características, optimizar hiperparámetros y la evaluación de modelos. Todos estos algoritmos de aprendizaje automático permiten ser usados para problemas supervisados y no supervisados de escala mediana (Pedregosa y cols., 2011).

Scikit-Learn se basa en las populares bibliotecas de Python NumPy y SciPy. NumPy por su parte permite operaciones eficientes entre arreglos y matrices multidimensionales. En cambio, SciPy proporciona los módulos necesarios para la computación científica (Hackeling, 2014).

3.2. Estado del arte

En el campo, muchos profesionales han investigado el problema de la predicción precisa de los accidentes viales. Para ello, se deben tener en cuenta múltiples factores para una predicción precisa, tales como el comportamiento de los conductores, condiciones climáticas,

estructuras viales, el flujo de tránsito, *etc.* Aunque algunos profesionales tienen en cuenta estos factores clave, no tienen en cuenta los factores indirectos (Yu y cols., 2021).

En el año 2022 se propuso un modelo de aprendizaje automático predictivo con la capacidad de poder analizar y determinar la gravedad de los accidentes, el número de víctimas y la cantidad de vehículos involucrados (Ardakani y cols., 2022). Los autores usaron el conjunto de datos de accidentes de tráfico entre 2005 y 2014 de Reino Unido para entrenar su modelo. Allí, establecieron y evaluaron cuatro técnicas de aprendizaje, las cuales son árboles de decisión, clasificación de bosques aleatorios, regresión logística multinomial y clasificaciones naïve Bayes para así escoger el modelo predictivo más adecuado. Al estudiar la precisión de los modelos, los autores determinaron que el clasificador naïve Bayes es el que tiene peor rendimiento. En la Figura **3-7** se puede observar los resultados obtenidos de cada modelo.

Figura 3-7: Resultados de la precisión de los modelos predictivos.

Modelo predictivo	Precisión de la gravedad del accidente	Precisión del número de vehículos	Número de bajas Precisión
Clasificador de árboles de decisión	85.4774%	64.6837%	83.8944%
Clasificador de bosque aleatorio	85.5798%	64.5471	83.9474%
Regresión logística multinomial	85.5142%	64.0328%	83.9804%
Clasificador Naive Bayes	<20%	<20%	<20%

Fuente: Obtenido de Ardakani y cols. (2022).

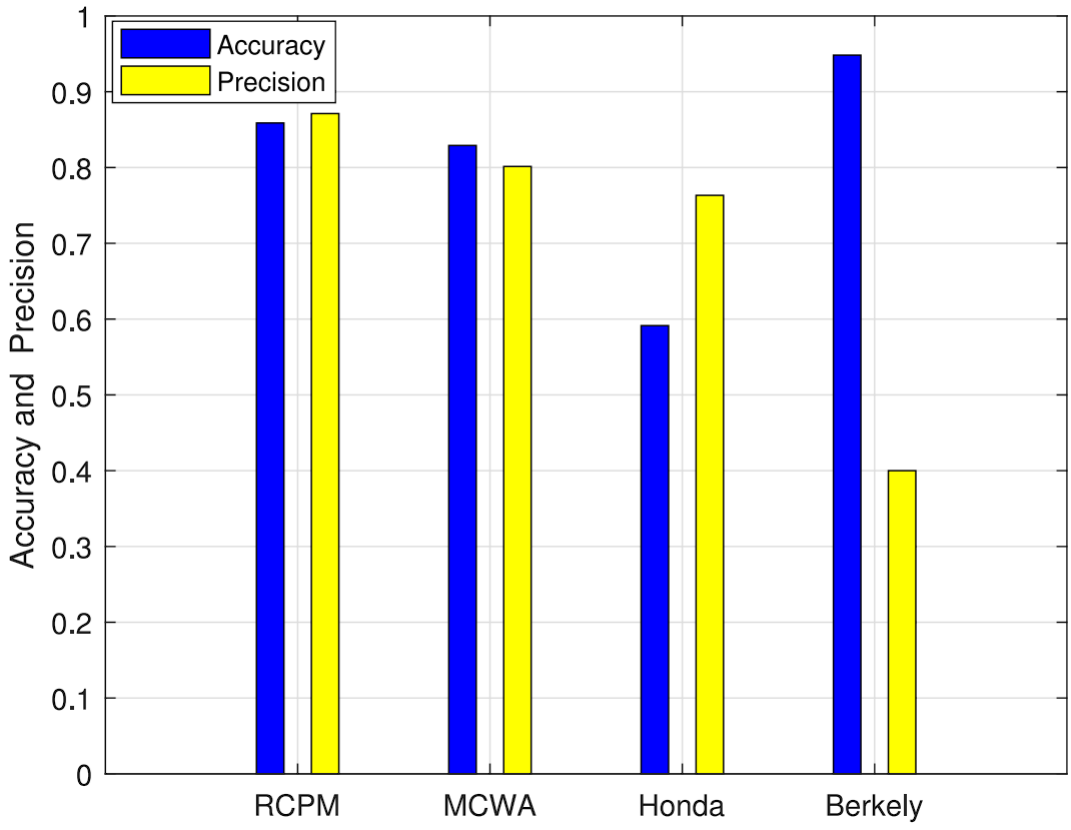
De los cuatro modelos, tres de ellos resultaron relativamente buenos con cerca de un 60 % a 80 % de precisión (Ardakani y cols., 2022). Al tener tres modelos con buenos resultados, para los autores resulta relevante proponer una estrategia mejorada con la que se pretende reducir la proporción desequilibrada presentada en las etiquetas de los datos con los que se entrenaron los modelos predictivos.

En otro caso, se realizó un mecanismo de predicción de colisiones en tiempo real utilizando aprendizaje automático para un sistema de transporte inteligente. Allí, los autores buscaban predecir una de las causas más relevantes de accidentes de tráfico: una posible colisión en la parte trasera del vehículo. Así, de esta manera, pretenden avisar a los conductores de un posible choque en tiempo real. Los autores, declaran que la razón principal de este tipo de siniestros viales, es por la tardía reacción que los conductores tienen para reaccionar

ante cualquier peligro, haciendo que su respuesta de frenado sea muy lenta (Xin y cols., 2020).

Para la correcta predicción de colisiones en tiempo real, propusieron un mecanismo de predicción de colisiones traseras con aprendizaje automático (RCPM) basándose en una red neuronal convolucional (CNN), en la que los datos de trayectoria real se utilizan para realizar los ajustes pertinentes a la red neuronal. Hicieron la recopilación del conjunto de datos de tráfico real obtenido de la Administración de Carreteras de EE. UU., la simulación de próxima generación (NGSIM), para realizar el debido entrenamiento del modelo. Los autores realizaron la comparación de *recall* y precisión de su modelo con los modelos de Honda, Berkely y MCWA. La Figura **3-8** muestra la comparación de los diferentes modelos.

Figura 3-8: Comparación de exactitud y precisión de los modelos de predicción de colisiones.

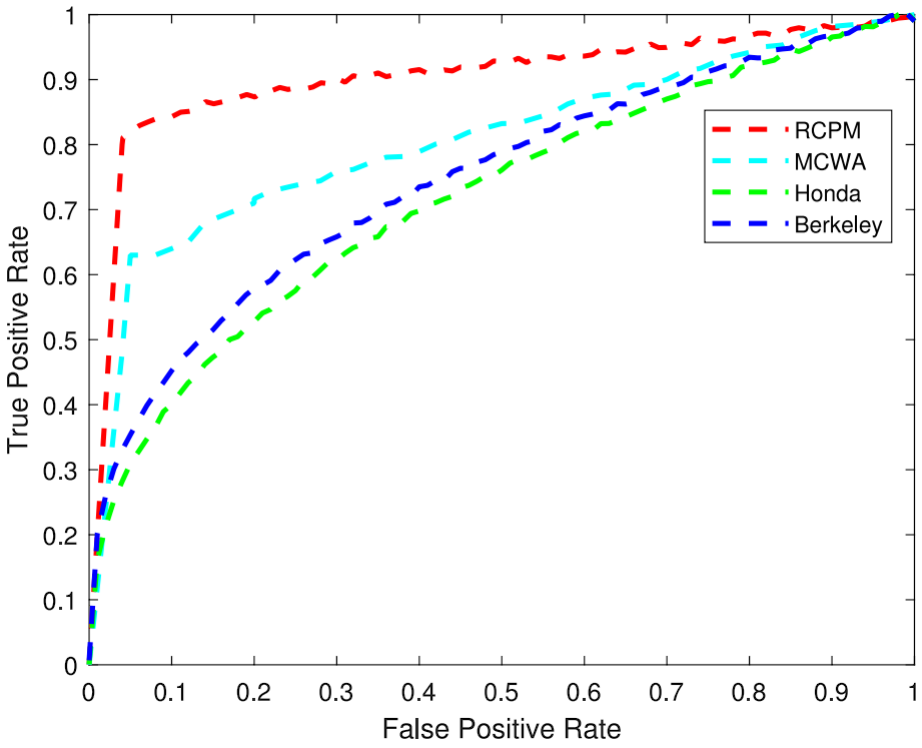


Fuente: Obtenido de Xin y cols. (2020).

La exactitud y precisión del modelo de MCWA es mucho mejor que la de Honda y Berkely, ya que no es sensible a los diferentes PRT (Percepción-Tiempo de Reacción) de los humanos. Sin embargo, el modelo de MCWA no tiene en consideración el problema del desequilibrio de las clases, y el espacio de búsqueda de la red neuronal de percepción de tres capas con 5 neuronas de entrada es muchísimo menor que el modelo RCPM propuesto por los autores (Xin y cols., 2020). También, se realizó la prueba de los modelos usando la métrica de la

curva ROC, obteniendo los siguientes resultados:

Figura 3-9: Comparación de las curvas ROC de cada modelo de predicción de colisiones.



Fuente: Obtenido de Xin y cols. (2020).

La curva con más desviación diagonal de 45 grados (la más cercana a la esquina superior izquierda) es la que logra el mejor rendimiento. La curva ROC del modelo de los autores es la que presenta el mayor acercamiento a la esquina superior derecha, por lo que el RCPM es el más optimo. El presente proyecto de grado tiene como particularidad realizar el respectivo estudio y entrenamiento para posteriormente diseñar y evaluar modelos de aprendizaje automático para la estimación de los siniestros viales que se están ocasionando en Bogotá D.C., para así realizar una predicción más precisa que ayude a las entidades gubernamentales a prevenir este tipo de accidentes.

A pesar de los antecedentes vistos, el proyecto de grado "Modelos de aprendizaje computacional para la predicción de siniestralidad vial en Bogotá D.C.resulta relevante en el contexto presentado. Aunque se han realizado modelos de aprendizaje automático para diferentes ubicaciones en el mundo con el fin de prevenir accidentes viales, no se ha realizado un modelo entrenado especificamente para la ciudad de Bogotá D. C. con datos actualizados hasta 2023.

3.3. Marco legal

3.3.1. Ley estatutaria 1581 de 2012 (Habeas Data)

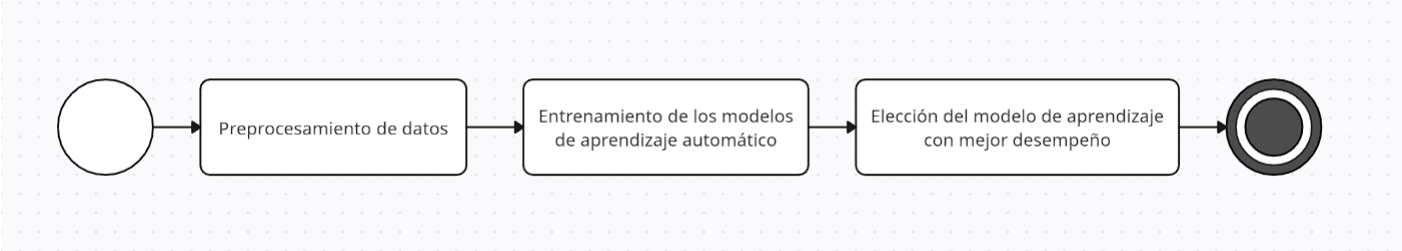
Esta ley tiene como objetivo el derecho que tienen todas las personas a conocer, actualizar y rectificar la información que se haya reunido sobre ellas en bases de datos o archivos que puedan ser susceptibles a tratamiento de datos por entidades públicas o privadas. Además, de poder restringir la libre distribución de sus datos, con el fin de evitar el uso con fines diferentes a los originalmente brindados (Ley 1581, 2012).

4 Metodología

La metodología que será usada está diseñada específicamente para este proyecto de grado, en relación con las tareas a realizar en cada paso.

En la Figura **4-1** se puede observar el flujo de los pasos de la metodología.

Figura 4-1: Modelo BPMN (Business Process Modeling and Notation) de los pasos realizados en la metodología del proyecto de grado.



Fuente: Elaboración propia.

4.1. Preprocesamiento de datos

Teniendo en cuenta que los datos que fueron usados para el entrenamiento de los modelos de aprendizaje automático provienen del portal Datos Abiertos Secretaría Distrital de Movilidad [1], el cual es un repositorio de datos públicos que es administrado por la Secretaria de Movilidad de Bogotá, D.C., se procederá a realizar el debido procesamiento de datos para entrenarlos.
Antes de continuar, resulta relevante explicar las columnas del conjunto de datos usado:

- X e Y: Son las coordenadas del punto exacto donde ocurrió el accidente.
- Formulario: Aquel registro que se identifica con un código alfanumérico donde las autoridades pertinentes guardan en un reporte la información relevante del siniestro vial.
- Código accidente: Código numérico que se usa para identificar el siniestro vial.

[1] https://datos.movilidadbogota.gov.co/datasets/movilidadbogota::accidente-1/explore

- Fecha de ocurrencia del accidente: Fecha exacta de cuando se originó el siniestro vial.
- Hora de ocurrencia del accidente: Hora exacta en franja a. m. o p. m. de cuando ocurrió el siniestro.
- Año de ocurrencia del accidente: Año en el cual se originó el siniestro vial.
- Mes de ocurrencia del accidente: Mes en el cual se originó el siniestro vial.
- Día de ocurrencia del accidente: Día en el cual se originó el siniestro vial.
- Dirección: Dirección apróximada de donde se originó el siniestro vial.
- Gravedad: Tipo de gravedad del siniestro vial, en el cual se encuentran categorizados de la siguiente manera: Con heridos, Con daños y Solo muertos.
- Clase de accidente: Tipo de accidente del siniestro vial, en el cual se encuentran categorizados de la siguiente manera: Atropello, Autolesión, Caída de ocupante, Choque, Incendio, Otro y Volcamiento.
- Localidad: Localidad perteneciente a Bogotá, D.C. donde se originó el siniestro vial.
- Latitud y Longitud del siniestro vial.
- CIV (Código de Identificación Vial): Código que está compuesto por números y letras que ayudan a identificar cuadras, calles, y edificaciones dentro de Bogotá, D.C.

Las tareas realizadas para el correcto preprocesamiento de datos fueron las siguientes:

- Definir la variable dependiente y las variables independientes.
- Normalizar los datos de entrada y salida.
- Realizar la partición de los datos que serán usados para entrenamiento y para pruebas.

De esta manera, se entregó como cumplimiento lo siguiente:

- Los datos procesados para el posterior entrenamiento.

Para el caso de la primera tarea, resulta relevante indicar cuál fue la variable independiente usada del conjunto de datos obtenido al realizar el preprocesamiento de datos. Para este caso, se tomó como variable independiente la columna denominada: *Type*.

Para el caso de la segunda tarea, al realizar la normalización se obtuvo el siguiente conjunto de datos (el cual fue el que se usó para entrenar y evaluar los modelos de aprendizaje automático en las fases posteriores a esta):

Figura 4-2: Conjunto de datos preprocesado

ID	Time	Year	Month	Weekday	Severity	Type	Latitude	Longitude
10533629	13:30:00	2020	November	Tuesday	Vehicle damage	Crash	4.682	-74.042
4416604	08:40:00	2015	June	Thursday	Vehicle damage	Crash	4.68959145	-74.05637597
4472710	11:58:00	2017	February	Saturday	Vehicle damage	Crash	4.69564938	-74.05152074
4448666	16:00:00	2016	June	Thursday	Vehicle damage	Crash	4.68487898	-74.05717989
4468964	20:00:00	2016	December	Friday	Vehicle damage	Crash	4.69360635	-74.05086959
4510438	23:10:00	2018	March	Monday	Injured	Crash	4.67529655	-74.06540965
4487069	07:50:00	2017	July	Monday	Vehicle damage	Crash	4.69597481	-74.06887244
4401969	16:00:00	2015	January	Monday	Vehicle damage	Crash	4.68818243	-74.04295422
10563307	10:22:00	2022	February	Tuesday	Injured	Crash	4.688	-74.05
4433352	16:00:00	2015	December	Wednesday	Vehicle damage	Crash	4.68608391	-74.05204021
4500309	09:18:00	2017	November	Tuesday	Vehicle damage	Crash	4.67938676	-74.05882981
4479567	11:15:00	2017	April	Sunday	Vehicle damage	Crash	4.69109341	-74.04933765
10453203	13:30:00	2018	July	Tuesday	Vehicle damage	Crash	4.67782839	-74.05395828
10453981	12:00:00	2018	July	Tuesday	Vehicle damage	Crash	4.69459177	-74.06452395
10452690	12:10:00	2018	June	Thursday	Vehicle damage	Crash	4.69466985	-74.04371931
10451506	12:40:00	2018	June	Monday	Injured	Crash	4.67721947	-74.05171663
10451983	18:20:00	2018	June	Friday	Vehicle damage	Crash	4.68331101	-74.04692095
10454606	11:00:00	2018	March	Sunday	Vehicle damage	Crash	4.67782839	-74.05395828
10453095	10:30:00	2018	July	Tuesday	Vehicle damage	Crash	4.68766238	-74.05091984
10454282	07:50:00	2018	July	Friday	Vehicle damage	Crash	4.68759462	-74.06023788
10453921	15:45:00	2018	July	Wednesday	Vehicle damage	Crash	4.69387615	-74.05591409

Fuente: Elaboración propia.

4.2. Entrenamiento de los modelos de aprendizaje automático

Teniendo los datos procesados que serán usados, se procederá a realizar el debido proceso de entrenamiento y asimismo su evaluación en las pruebas pertinentes. Las tareas realizadas para el correcto entrenamiento de los modelos de aprendizaje automático fueron las siguientes:

- Definir cada modelo de aprendizaje.
- Ingresar los datos procesados en las actividades de entrenamiento y prueba.
- Realizar el debido proceso de entrenamiento y prueba para cada uno de los modelos definidos.

Para el caso de la primera tarea de la fase de entrenamiento de los modelos de aprendizaje automático, se usaron los siguientes modelos: SVM (*Support Vector Machines*), MLP (*Multilayer Perceptron*) y RFC (*Random Forest Classifier*), los cuales fueron entrenados con los datos preprocesados que se obtuvieron en la primera fase de la metodología (cabe aclarar que estos datos preprocesados fueron almacenados en archivos con extensión *csv*).

Para el caso de la tercera tarea de la fase de entrenamiento de los modelos de aprendizaje automático, se usaron los siguientes parámetros haciendo uso de *GridSearchCV* y *PCA*

(Principal Components Analysis), que vienen incluidos en la biblioteca de Scikit-Learn, las cuales permiten optimizar y ajustar los mejores parámetros para cada modelo usado.

SVM (*Support Vector Machines*)

- C: [1, 10, 100].
- Kernel: ['linear'].
- CV: 10.

MLP (*Multilayer Perceptron*)

- Hidden_layer_sizes: 1.
- Activation: 'logistic'.
- Solver: 'adam'.
- Random_state: 1.
- CV: 10.

RFC (*Random Forest Classifier*)

- n_estimators: [200, 700].
- n_features: 6.
- n_informative: 3.
- n_redundant: 0.
- n_repeated: 0.
- n_classes: 2.
- Random_state: 0.
- Shuffle: False.
- n_jobs: -1.
- Max_features: ['auto', 'sqrt', 'log2'].

- n_estimators: 50.
- Oob_score: True.
- CV: 10.

De esta manera, se entregó como cumplimiento lo siguiente:

- Los modelos predictivos entrenados con su correspondiente evaluación de desempeño.

4.3. Elección del mejor modelo de aprendizaje

Con los modelos definidos y ya entrenados, se estudió con base en los resultados del punto anterior cuál de ellos fue el más óptimo. Las tareas para realizar la correcta la elección del mejor modelo de aprendizaje fueron:

- Generar un archivo CSV (valores separados por comas) con las mediciones de las mediciones de las métricas.
- Elegir el modelo con el mejor desempeño predictivo.

Para el caso de la primera tarea de la fase de elección del mejor modelo de aprendizaje automático, se obtuvieron los siguientes resultados que se pueden observar en la Tabla **4-1**:

Tabla 4-1: Resultados de los modelos de aprendizaje automáticos

Modelo	Accuracy	Precision	Recall	F1 Score
SVM (Support Vector Machines)	0.753	0.692	0.789	0.951
MLP (Multilayer Perceptron)	0.845	0.827	0.818	0.893
RFC (Random Forest Classifier)	0.996	0.901	0.967	0.976

Fuente: Elaboración propia.

De esta manera, se entregó como cumplimiento lo siguiente:

- Modelo de aprendizaje automático supervisado que servirá como herramienta para reducir siniestros viales.

Antes de finalizar este capítulo, resulta relevante mencionar un ejemplo práctico hipotético sobre la manera en que el modelo funcionará. El modelo de aprendizaje automático escogido (RPC), queda listo para operar de la siguiente manera: El modelo recibirá como entrada un conjunto de datos con las siguientes características:

- Mes en el que ocurrió el siniestro vial.

- Día en el que ocurrió el siniestro vial.
- Hora del siniestro vial.
- Latitud y Longitud del siniestro vial.
- Zona de la ciudad en que ocurrió el siniestro vial.
- Tipo de accidente.

Para este ejemplo práctico hipotético se usarán los siguientes valores:

- Mes en el que ocurrió el siniestro vial: Noviembre.
- Día en el que ocurrió el siniestro vial: Martes.
- Hora del siniestro vial: 13:15:00.
- Latitud y Longitud del siniestro vial: 4.68959145, -74.05637597.
- Zona de la ciudad en que ocurrió el siniestro vial: Norte.
- Tipo de accidente: Choque.

Con los valores previamente mencionados el modelo será capaz de predecir el posible siniestro vial en base a la variable independiente denominada Tipo de accidente. Y dará como resultado un valor entre 0 y 1 (teniendo en cuenta que entré más cercano sea el valor a 1 será más confiable la predicción del siniestro y entre más cercano sea el valor a 0 más desconfiable será la predicción del siniestro).

5 Conclusiones

Al emplear métodos sencillos y eficientes, junto con un preprocesamiento de datos cuidadoso permite que los datos nulos, duplicados e irrelevantes del conjunto de datos sean excluídos, haciendo que los modelos fueran más óptimos a la hora de realizar el entrenamiento y su posterior elección, eligiendo al mejor modelo en base con las métricas usadas para evaluarlos. Además, resulta relevante mencionar la actualización del conjunto de datos usado, ya que al ser obtenidos de una fuente de bases de datos como la del gobierno distrital (que por lo general, se actualizan mensualmente), haciendo que los resultados obtenidos de los modelos siempre estén actualizados.

Como se ha remarcado en varias secciones de este documento, la fase de preprocesamiento de datos es una de las partes fundamentales al realizar el debido entrenamiento de los modelos de aprendizaje automático. Por ello, esta es una de las fases que más requieren trabajo y la cual genera más problemas (como lo pueden ser las columnas o variables redundantes y que no resultan relevantes para los modelos), y de los cuales muchos de ellos están ligados a el conjunto de datos crudo que se obtiene directamente de la plataforma de Datos de la Secretaría de Movilidad. Para realizar una solución eficiente, se optó por hacer uso de técnicas como *One Hot Encoding* (la cual permite que los atributos categóricos o nominales sean convertidos a números para que el modelo pueda procesarlos) y *Principal Component Analysis* o Análisis de componentes principales (que ayuda a reducir el número de columnas obtenidas al aplicar la técnica *One Hot Encoding*, permitiendo simplificar y visualizar los datos sin que estos pierdan su información importante). También, resulta relevante el hecho de la fase de entrenamiento, ya que al manejar muchos datos hace que indudablemente los modelos se demoren más. Para solucionar esto se optó por usar Google Colab, la cual es una plataforma enfocada al desarrollo de programas en Python, y que además cuenta con los recursos necesarios en lo que respecta a procesador y ram para realizar este entrenamiento. Esta al ser una plataforma web, permite que se hagan entrenen varios modelos sin afectar significativamente algún equipo de manera local.

Los tiempos empleados para el entrenamiento y desarrollo de cada modelo resultan bastantes extensos. Esto dependiendo de los parámetros usados para cada uno, el número de la muestra usada y el equipo de cómputo usado para este fin. Para aquellos que deseen ejecutar los modelos, tienen que disponer de un tiempo considerable (con considerable, me refiero a una estimación entre 6 a 20 horas usando Google Colab).

El modelo de aprendizaje automático que resultó relevante para realizar la predicción de siniestros viales según las métricas de desempeño usadas y con el conjunto de datos obtenido del preprocesamiento de datos es el RFC (*Random Forest Classifier*). En todas las métricas (teniendo en cuenta que se evaluan en un rango de 0 a 1), este fue el que obtuvo las mejores métricas al realizar la evaluación de todos los modelos. Además, que también el tiempo empleado para el entrenamiento y desarrollo fue mucho menor que el resto.

Este proyecto de investigación resulta relevante, ya que al hacer parte del grupo de investigación LACSER (*Laboratory for Advanced Computational Science and Engineering Research*) de la Univsersidad Antonio Nariño, los resultados obtenidos ayudan al macroproyecto del cual hace parte (Movilidad inteligente y sostenible). Ya que uno de los principales problemas que enfrenta Bogotá, D.C. y en general a nivel global son los muertos y lesionados provenientes de siniestro viales, haciendo que se configure como un problema de alta complejidad. Como se mencionó en previos capítulos, este proyecto de grado pretende permitir sentar las bases para usar en futuros modelos, que, usando los modelos empleados en este proyecto de grado puedan ser mejorados para permitir ser usados no solo en Bogotá, D.C., sino en todas las ciudades que estén afectadas por problemas similares.

Referencias

Agencia Nacional de Seguridad Vial. (2022). *Estadísticas fallecidos y lesionados 2021 2022.* Descargado de https://ansv.gov.co/es/observatorio/estad%C3%ADsticas/fallecidos-y-lesionados-2021-2022

Arce, J. I. B. (2019, 7). *Health big data.* Descargado de https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

Ardakani, S. P., Liang, X. N., Mengistu, K. T., So, R. S., Wei, X., él, B., y Cheshmehzangi, A. (2022). Road car accident prediction using a machine-learning-enabled data analysis. *MDPI*, 15-5939.

Bobadilla, J. (2020). *Machine learning y deep learning: Usando python, sickit y keras.* Ra-ma Editorial. Descargado de https://books.google.com.co/books?hl=es&lr=&id=iAAyEAAAQBAJ&oi=fnd&pg=PA11&dq=python+y+machine+learning&ots=QhCay_nK3s&sig=xCghH3NOfc8p_pC-VK2GF_ocHTQ&redir_esc=y#v=onepage&q=python%20y%20machine%20learning&f=false

Díaz, J. (2021, 10). *Con una precisión sin precedentes: Una tecnología logra predecir accidentes de tráfico.* Descargado de https://www.elconfidencial.com/tecnologia/novaceno/2021-10-15/accidente-trafico-inteligencia-artificial_3306727/

Google. (2022, 9). *Clasificación: Verdadero o falso y positivo o negativo.* Descargado de https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative?hl=es-419#:~:text=A%20true%20positive%20is%20an,incorrectly%20predicts%20the%20positive%20class.

Google. (2023). *Los distintos métodos usados para enseñar a la ia.* Descargado de https://atozofai.withgoogle.com/intl/es/learning/

Gupta, A., Anand, A., y Hasija, Y. (2021, 4). *Recall-based machine learning approach for early detection of cervical cancer.* Descargado de https://ieeexplore.ieee.org/abstract/document/9418099

Hackeling, G. (2014). *Mastering machine learning with scikit-learn* (2.ª ed.; Packtx, Ed.). Descargado de https://books.google.com.co/books?hl=es&lr=&id=9-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=Gavin+Hackeling+2014&ots=FNiAQtbWOk&sig=

rVgO2hK53hCHS7xCUOe_Yul3080&redir_esc=y#v=onepage&q=Gavin%20Hackeling%202014&f=false

Heras, J. M. (2020, 10). *iartificial.* Descargado de https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

Khayesi, J. A., y Meleckidzedeck. (2021, 5). *El papel del sistema de las naciones unidas en la mejora de la seguridad vial.* Descargado de https://www.un.org/es/cr%C3%B3nica-onu/el-papel-del-sistema-de-las-naciones-unidas-en-la-mejora-de-la-seguridad-vial-para

Ley 1581. (2012, 10). *Ley 1581 de 2012.* Descargado de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Lipton, Z. C., Elkan, C., y Naryanaswamy, B. (2014, 2). *Thresholding classifiers to maximize f1 score.* Descargado de https://arxiv.org/abs/1402.1892

Malaver, C. (2023, 8). *En 2023 se han registrado 308 fallecidos por siniestros viales en bogotá.* Descargado de https://www.eltiempo.com/bogota/en-2023-se-han-registrado-308-fallecidos-en-bogota-en-siniestros-viales-795217

MathWorks. (2023). *Evaluación del rendimiento de los modelos de clasificación de machine learning.* Descargado de https://la.mathworks.com/discovery/roc-curve.html

Microsoft. (2022, 9). *Componente normalizar datos.* Descargado de https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/normalize-data

Mimenza, O. C. (2019, 7). *Variable dependiente e independiente: qué son, con ejemplos.* Descargado de https://psicologiaymente.com/miscelanea/variable-dependiente-independiente

Naciones Unidas. (2023). *Naciones unidas de colombia.* Descargado de https://colombia.un.org/es/sdgs/11

Norman, A. T. (2021). *Aprendizaje automático en acción* (Tektime, Ed.).

Organización Mundial de la Salud. (2021). *Plan mundial: Decenio de acción para la seguridad vial 2021-2030.* Descargado de https://cdn.who.int/media/docs/default-source/documents/health-topics/road-traffic-injuries/21323-spanish-global-plan-for-road-safety-for-web.pdf?sfvrsn=65cf34c8_35&download=true

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011, 10). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2825-2830. Descargado de https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/

Portafolio.co. (2020, 11). *Lo que le cuestan los accidentes viales a la economía de colombia.* Descargado de https://www.eltiempo.com/economia/sectores/lo-que-le-cuestan-los-accidentes-viales-a-la-economia-colombiana-549422

pyCodeMates. (2023, 5). *The top 5 must known classification algorithms in machine learning.* Descargado de https://www.pycodemates.com/2022/10/top-5-must-known-classification-algorithms-machine-learning.html

Pérez, I. C., Ricardo, Y. D., y García, R. A. B. (2014). *El lenguaje de programación python/the programming language python.* Descargado de https://www.redalyc.org/pdf/1815/181531232001.pdf

Pérez, J. M., y Martin, P. P. (2023, 1). La curva roc/roc curve. *ScienceDirect*, *49*. Descargado de https://www.sciencedirect.com/science/article/abs/pii/S1138359322001952

Samina, K., Tehmin, K., y Shamila, N. (2014). *A survey of feature selection and feature extraction techniques in machine learning.* Descargado de https://ieeexplore.ieee.org/document/6918213/authors#authors

Sanahuja, P. M. (2021, 1). *Entendiendo la curva roc y el auc: Dos medidas del rendimiento de un clasificador binario.* Descargado de https://polmartisanahuja.com/entendiendo-la-curva-roc-y-el-auc-dos-medidas-del-rendimiento-de-un-clasificador-binario-que-van-de-la-mano/#:~:text=El%20AUC%20significa%20%C3%A1rea%20bajo,una%20tarea%20de%20clasificaci%C3%B3n%20dada.

Sandoval, L. J. (2018, 10). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Artículos de Revista (ITCA-FEPADE)*, 36-40. Descargado de http://redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Vaid, N. K. (2019, 5). *Statistical performance measures.* Descargado de https://neeraj-kumar-vaid.medium.com/statistical-performance-measures-12bad66694b7

Xin, W., Liu, J., Qiu, E., Mu, C., Chen, C., y Zhou, P. (2020). A real-time collision prediction mechanism with deep learning for intelligent transportation system. *IEEE*, 9497-9508.

Yu, L., Du, B., Hu, X., Sun, L., Han, L., y Lv, W. (2021, 1). Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing*, *423*, 135-147. Descargado de https://www.sciencedirect.com/science/article/abs/pii/S092523122031451X doi: 10.1016/J.NEUCOM.2020.09.043