ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

"Análisis de accidentes de tránsito en el cantón Guayaquil usando Machine Learning"

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

David Argüello – Enrique Alcívar

GUAYAQUIL - ECUADOR

Año: 2022 - 2023

DEDICATORIA

A mis abuelos que son mi compañía y apoyo, a mi madre por impulsarme a no rendirme, y a mi padre que está en cielo siendo mi guía.

David Argüello F.

A mi padre y madre que fueron el apoyo en este viaje, a mis hermanos quienes siempre estuvieron en los momentos buenos y malos y a Dios por haberme guiado a lo largo de mi vida.

Enrique Alcívar S.

AGRADECIMIENTOS

A mi familia por todo el apoyo que me han dado, a Enrique por embarcarse en esta aventura sin rumbo, a Melanie por ser mi pilar y luz en el camino hasta esta meta de ser ingeniero, y a mí por no rendirme en todo este trayecto.

David Argüello F.

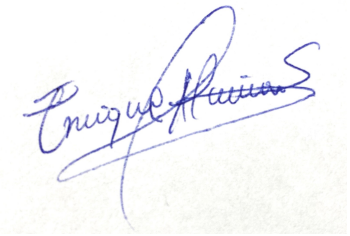
A mis padres por haberme dado la fortaleza para seguir adelante en aquellos momentos de debilidad, a David por la paciencia y la atención y a Dios por ser el pilar de todo.

Enrique Alcívar S.

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; Nosotros, David Argüello y Enrique Alcívar, damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

David Arguello Fiallos



Enrique Alcívar Salinas

EVALUADORES

Firmado electrónicamente por:
SANDRA LORENA
GARCIA BUSTOS

Ph.D. Sandra García Bustos
PROFESOR DE LA MATERIA

Firmado electrónicamente por:
MARIELA ALEXANDRA
GONZALEZ NARVAEZ

M.Sc. Mariela González
PROFESOR TUTOR

RESUMEN

A medida que la población incrementa, las opciones de financiamiento incrementan y la ciudad se expande, los accidentes crecen con estos factores. Debido a esto, hemos decidido estudiar este fenómeno de los siniestros de tránsito desde otra perspectiva que no sea la tradicional estadística descriptiva sino usando un enfoque de Machine Learning, para poder entender a través de los datos que realmente determina el nivel de siniestralidad de un accidente de tránsito en el cantón Guayaquil, usando los datos accidentes de tránsitos durante el 2021 y 2022 que maneja la Agencia Nacional de Tránsito. Dentro de este estudio, se utilizaron distintos métodos de aprendizaje supervisado y contrastándolos entre sí para determinar cuál tiene una precisión más alta en base al nivel de siniestralidad real de un accidente. Se encontró que estos métodos son muchos más precisos que la estadística clásica al momento de describir un siniestro, nos ayudan a identificar de manera más certera las variables que definen el nivel de siniestralidad de un accidente. Estos resultados pueden ser incluidos dentro de algún aplicativo móvil o web que sea de consumo para los miembros de rescate y control, de esta manera podríamos tener equipos que estén preparados del escenario que van a encontrar en el lugar del accidente. Se pueden tomar medidas preventivas en las zonas más frecuentes de accidentes, como tener una estación de ambulancias aledañas y disminuir el número de fallecidos.

Palabras claves: Nivel de siniestralidad, métodos de aprendizaje automático, estadística descriptiva, tasa de mortalidad, accidentes de tránsito.

ABSTRACT

As population grows, financial opportunities increases and the city expands, the traffic accidents grow with these factors, due to this we have decided to study this phenomenon of traffic accidents from another perspective different from traditional descriptive statistics instead using Machine Learning approach to be able to understand through data what really determines the severity of a traffic accident in Guayaquil using the data from National Traffic Agency (ANT) yearly report of traffic accidents. Within this paper we are using different supervised methods and comparing them with each other to see which one has best accuracy on forecasting the severity of an accident. We found that these methods are most accurate than conventional statistics when we try to describe an accident, them help us to identify in a most accurate way which variables define the severity of an accident. These results could be included inside of a web/mobile app to be use by rescue teams like police or ambulances. Using them we could aware those rescue teams what they could expect to see at the time they arrived to the place of accident.

Key words: Severity of an accident, machine learning methods, descriptive statistics, reduce death rate in traffic accidents.

ÍNDICE GENERAL

ABREVIATURAS

ANT	Agencia Nacional de Tránsito.
ATM	Agencia de Tránsito y Movilidad de Guayaquil.
CTE	Comisión de Tránsito del Ecuador.
DNCTSV	Dirección Nacional de Control de Tránsito y Seguridad Vial (Policía Nacional).
OISEVI	Observatorio Iberoamericano de Seguridad Vial.
OMS	Organización Mundial de la Salud.
LOTTTSV	Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial.
RLOTTTSV	Reglamento a Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial.
INEC	Instituto Nacional de Estadísticas y Censos.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

Los accidentes de tránsito se llevan cada año la vida de aproximadamente 1.3 millones de personas en el mundo (WHO, 2022), a su vez entre 20 y 50 millones sufren lesiones no fatales, muchas de ellas provocan una discapacidad.

Las lesiones por accidentes de tránsito causan pérdidas económicas significativas a las personas, sus familias y naciones enteras. Estas pérdidas están relacionadas con el costo del tratamiento de los heridos, muertos o discapacitados y la pérdida de productividad, así como el tiempo que los familiares lesionados tienen que dejar el trabajo o la escuela para cuidarlos. Para los países, los accidentes de tránsito representan un 3% de su PIB (WHO, 2022).

Según la OMS (Organización Mundial de la Salud) la tasa de mortalidad por accidentes de tránsito en Ecuador es de 21.3 decesos por habitantes (WHO, 2022).

Por su parte el INEC (Instituto Nacional de Estadísticas y Censos) indica que los accidentes de tránsito son una de las 10 causas principales de muertes en Ecuador, siendo como principal causa de siniestros la impericia e imprudencia del conductor, con un total de 9,281 siniestros, que representan el 43.5%; seguido de no respetar las señales de tránsito con 4,476 siniestros, con el 21% del total de siniestros de tránsito nacionales (INEC, 2021).

En Ecuador se han registrado 14,388 accidentes durante el año, cifra que se divide en 1,470 accidentes letales y 12,412 accidentes con heridos según la Agencia Nacional de Tránsito (ANT, 2022). Esto representa que seis personas pierden la vida cada día en las carreteras del país (ANT, 2022).

Mientras que, en la ciudad de Guayaquil, en el mismo período se han registrado 2793 accidentes, de los cuales resultaron 1848 personas lesionadas y 137 fallecidas. Comparando estos datos con los del 2021, se han registrado 118 accidentes menos,

pero un incremento de aproximadamente el 10% en relación al número de fallecidos (ANT, 2022).

Frente a esta situación, la Agencia de Tránsito Municipal (ATM) propone generar talleres, capacitaciones continuas y campañas que tengan como objetivo inculcar en los ciudadanos una responsabilidad vial (Universo, 2022).

Para tomar decisiones basadas en datos, y utilizar los recursos de una manera óptima, es importante realizar un análisis de contraste aplicando métodos estadísticos supervisados y no supervisado para poder predecir el escenario de un accidente, en términos de la situación de los ocupantes del vehículo, en base a factores directos como tipo de vehículos involucrados, estado del conductor o factores exógenos como lugar, hora y fecha.

1.1 Descripción del problema

Según la Comisión Económica para América Latina, los accidentes de tránsito son la segunda causa de muerte para personas entre 5 y 29 años y pueden convertirse en la tercera causa de discapacidad para personas entre 30 y 44 años en el mundo. Esto representa en términos monetarios, altos costos a la economía de los países rondando los $518MM (miles de millones), de los cuales $65MM corresponden a países de medianos o bajos ingresos, monto que supera al que reciben como ayuda de fondos internacionales. A nivel mundial, la cantidad de fallecidos por accidentes de tránsito es alrededor de 1.2M de personas anualmente. Para países en vías de desarrollo donde el número de vehículos está creciendo rápidamente, la seguridad vial se convierte en un factor clave que deben controlar mediante medidas de prevención y generando conciencia dentro de los habitantes. (Planzer R, 2005)

Dentro de Ecuador, un gran número de los siniestros que se registran en el país se deben a la falta de una cultura de prevención. Según el director del Observatorio de Seguridad Vial (OSEVI), Yasmany García, este problema se lo soluciona en conjunto, tanto de parte de las autoridades como por parte de los ciudadanos, ya que la seguridad vial es una responsabilidad de todos incluyendo hasta de los peatones.

Con la llegada de nuevas marcas de vehículos para el mercado ecuatoriano, cada vez se ven más autos en las calles, incrementando el tráfico y se puede asumir que de la

misma manera incrementan las probabilidades de accidentes dentro del país. Para mayo del presente año, las ventas de vehículos nuevos registran un crecimiento interanual del 23.8% contra el mismo mes del 2021 (Automegazine, 2022). Dentro del top de marcas más vendidas se encuentran Chery, JAC, JETOUR, GREATWALL, SHINERAY, DFSK, SOUEAST, las cuales una década atrás no se las conocía.

En la actualidad los accidentes de tránsito pueden estar relacionados con muchos más factores exógenos que no se toman en consideración dentro del panorama convencional, como, por ejemplo: el incremento en la cantidad de autos dentro de la ciudad, los conductores que tienen licencias compradas o la implementación de nuevos paneles de entretenimiento dentro de los autos.

1.2 Justificación del problema

Los accidentes de tránsito constituyen unas de las 10 causas principales de muerte en Ecuador (INEC, 2021).

En la ciudad de Guayaquil en lo que respecta al periodo enero-agosto del 2022 el 83% de los siniestros son de conductores de género masculino con edades entre 21 y 32 años siendo una de las 4 causas principales de siniestro el conducir un vehículo superando los límites máximos de velocidad, realizar cambios bruscos o indebidos de carril, no respetar las señales reglamentarias de tránsito (Pare, ceda el paso, luz roja del semáforo, etc.) y conducir bajo la influencia de alcohol, sustancias estupefacientes o psicotrópicas y/o medicamentos; además de tener mayor incidencia durante los fines de semana con periodos de 6 am, 2 – 7 pm. (ANT, 2022).

Adicionalmente la ATM detalla que existen cinco vías con mayor siniestralidad. La principal es la vía Perimetral donde se reportaron 159 accidentes con 146 lesionados y 19 muertos; seguida de la vía Daule con 91 accidentes, la autopista Narcisa de Jesús con 90 siniestros de tránsito, Francisco de Orellana con 83 accidentes y la Casuarina con 52 casos. (Universo, 2022).

Sería factible la sectorización de estas zonas con más índice de siniestralidad puesto que con la identificación de estas se determinan las intersecciones, vías y tramos más conflictivos que registran los mayores números de accidentes y fatalidades. De esta

manera contribuir al mejoramiento de la seguridad vial y por ende proyectarnos a mejorar la calidad de vida de la comunidad en general (Pulgarín, 2022).

La conducta de manejo es otro factor importante al momento de evitar siniestros de tránsito, Carro & Ampudia (2019) aportan evidencia inicial sobre el papel que juega el sexo y los contextos de conducción en las conductas de riesgo durante el manejo de un automóvil, sobre todo de estos últimos, en donde se apreciaron diferencias importantes en el uso del cinturón de seguridad, la conducción con ambas manos al volante y la velocidad, en Guayaquil hasta agosto de los 137 perecidos, 121 fallecieron por el no uso de cinturón(ANT, 2022).

- **Estado del arte**

En el 2016 se realizó un estudio exploratorio, descriptivo y transversal con representación espacial sobre la accidentabilidad, lesividad y letalidad por accidentes de tránsito registrados en las provincias y cantones del Ecuador, a partir de fuentes oficiales de información secundaria, donde se identificaron aquellos cantones con mayores tasas de letalidad, con esto se obtuvo información útil para identificar y aportar soporte en establecer programas y tomar correctivos puntuales en la seguridad vial (Gómez A, 2016).

En otro estudio se utilizó un análisis espacial exploratorio de los siniestros por accidentes de tránsito en las Provincias de la Región Amazónica del Ecuador, de forma similar al estudio previamente descrito, para analizar la distribución espacial de los siniestros de tránsito permitiendo la detección y análisis de puntos de riesgo geográficos a partir de variables sociodemográficas, densidad poblacional, desarrollo económico e infraestructura y estado de las carreteras, permitiendo a los organismos oficiales establecer acciones y políticas de seguridad vial en áreas geográficas determinadas (Galarza L., et. al, 2017).

En el ámbito internacional, se propuso analizar, desde una perspectiva multidimensional, los datos acerca de los fallecidos por siniestros viales en la provincia de Buenos Aires. Como primera instancia se buscó construir una tipología de siniestros viales, y, en segundo lugar, se analizó la distribución de los conglomerados en distintas zonas de la provincia aplicando métodos de clasificación jerárquica, mostrando como resultados una base de segmentación para orientar esfuerzos focalizados entre distintos grupos de conductores (Montes S. & Ledesma R., 2020).

1.3 Objetivos

1.3.1 Objetivo General

Analizar los accidentes de tránsito ocurridos en el cantón Guayaquil usando técnicas de Machine Learning, para poder identificar factores principales relacionados con el estado post accidente de los tripulantes.

1.3.2 Objetivos Específicos

- Clasificar por nivel de siniestralidad los accidentes para identificar qué combinación de factores se pueden considerar letales si ocurriese un accidente.
- Predecir el número de accidentes de tránsito a través de métodos de aprendizaje estadístico.
- Identificar las zonas más propensas a tener un accidente fatal según sus características.

1.4 Marco Teórico

1.4.1 Seguridad Vial

Son medidas asociadas con el control de la velocidad, el diseño de las construcciones, la seguridad de los vehículos, el acatamiento de las leyes, la atención de emergencia tras un accidente de tránsito (Salve VIDAS, 2017).

1.4.2 Siniestros de tránsito

Todo suceso o acción no intencionada que, como consecuencia de una o varias causas y con independencia de su gravedad, se produzca en vías o lugares destinados al uso público o privado y cause la muerte, lesiones de diversa índole o naturales a las personas y daños a la propiedad de vehículos, carreteras o infraestructuras, que involucren a los usuarios de la vía, del vehículo, de la vía y/o del medio ambiente (RLOTTTSV, 2016).

1.4.3 Tipos de choques entre vehículos

Dentro de los choques o colisiones que son el impacto de dos vehículos, lejos de la idea que todos son lo mismo, hay distintos tipos de choques según las circunstancias en las que se dan. Según RLOTTSV, cuando alguien choca al vehículo que está detrás de él, se conoce como choque por alcance. Cuando dos vehículos que vienen en direcciones opuestas se impactan, hay dos escenarios: cuando se impactan y su eje longitudinal coincide, cuando se impactan frente a frente, se conoce como choque frontal longitudinal, pero cuando el choque no es

exactamente frente a frente, se conoce como choque frontal excéntrico. Así mismo cuando un vehículo impacta al otro por un lado hay dos posibles escenarios, cuando uno de los vehículos impacta perpendicularmente al otro se conoce como choque lateral perpendicular, pero cuando no forman precisamente 90 grados entre el frente un vehículo contra el lateral del otro se conoce como choque lateral angular.

No necesariamente deben impactarse con el frente del vehículo, existen situaciones en las cuales se pueden impactar paralelamente, como son los roces. Si los vehículos van en la misma dirección se conocen como roces negativos y el acto de dos vehículos afectando la integridad del otro paralelamente se conoce como rozamiento.

1.4.4 Siniestros no necesariamente contra otro vehículo

Uno de los puntos que menciona RLOTTTSV es que no todos los siniestros que afecten la integridad de un vehículo deben ser contra otro, existen situaciones externas que pueden afectarlo. Cuando un vehículo se impacta contra un objeto fijo se conoce como estrellamiento, puede ocurrir yendo en línea recta o por pérdida de carril, que es la salida del vehículo de la calzada normal de circulación.

1.4.5 Consecuencias de siniestros en los vehículos.

Posterior a los choques o siniestros pueden ocasionar volcamientos, que pueden ser laterales o longitudinales que pueden ser por partes o ciclo completo del volcamiento, si un carro queda asentado sobre un lateral del mismo, se conoce como volcamiento lateral ¼. Y de la misma manera, si se da toda la vuelta y queda asentado sobre el mismo eje de las llantas será un volcamiento lateral en ciclo completo.

1.4.6 Siniestros desde el punto de vista de personas

Dentro de los siniestros no solo involucran a vehículos, también son siniestros los accidentes que afecten la integridad de las personas. Cuando un vehículo pasa con su rueda o ruedas por encima de una persona (o animal) se considera

arrollamiento, pero si solo lo impacta se conoce como atropello. Pero si el pasajero pierde el equilibrio y esto lleva a que tenga un descenso violento hacia la calzada se conoce como caída de pasajero.

1.4.7 Personajes dentro del esquema de tránsito

Dentro del tránsito, las personas ocupan un rol según su perfil. Dentro del vehículo quien va al volante se conoce como conductor y los demás son pasajeros. Quienes transitan por las calles sin ir en un vehículo son los peatones según RLOTTTSV, pero OISEVI (Observatorio Iberoamericano de Seguridad Vial) también incluye a esta categoría a personas con discapacidades que van en un vehículo como silla de ruedas o personas que van en patines/patinetas también son consideradas como peatones.

1.4.8 Consecuencias para involucrados en un siniestro

Todos los involucrados que tengan alguna consecuencia en su integridad física post accidente son considerados como víctimas. Dentro de las mismas según la severidad de las consecuencias tienen una categoría. Si una de las víctimas muere a consecuencia de un siniestro vial en el momento o durante los 30 posteriores días según OISEVI, se considera como fallecido, siempre y cuando las autoridades descarten suicidio. Así mismo, las víctimas que como consecuencia del siniestro no fallezcan, pero sufran lesiones se consideran como lesionados.

1.4.9 Tipos de vehículos involucrados

Hace referencia a la clasificación vehicular dada por la norma NTE INEN (Servicio Ecuatoriano de Normalización) 2656:2016 que identifica a los vehículos automotores median sus características generales de diseño y uso, estos vehículos se encuentran considerados en el RLOTTTSV (Reglamento a Ley Orgánica de Transporte Terrestre, Tránsito y Seguridad Vial) Dentro de esta clasificación se encuentran:

- Automóvil: Vehículo liviano destinado al transporte de un grupo reducido de personas.
- Bicicleta: Vehículo de tracción humana de dos o más ruedas en línea.

- Bus: Vehículo automotor diseñado para el transporte de pasajeros compuesto por un chasis y una carrocería acondicionada para transportar al conductor y a más de 36 asientos.
- Camión: Vehículo a motor construido especialmente para el transporte de carga con capacidad superior a 3.5 Ton.
- Camioneta: Vehículo a motor construido para el transporte de carga, con capacidad máxima de 3.5 Ton.
- Emergencias: Pertenece a la Policía Nacional o Cuerpo de Bomberos y las ambulancias institucionales que porten los distintivos especiales determinados para el efecto.
- Especial: Vehículos que pertenecen a las categorías M, N u O destinados al transporte de pasajeros o mercancías. Ej: Casas rodantes, estaciones médicas móvil, entre otros.
- Furgoneta: Vehículo ligero diseñado para el transporte de pasajeros y mercancías. Con una capacidad máxima de 18 asientos.
- Motocicleta: Vehículo de dos a cuatro ruedas cuya masa en vacío no exceda de 400 kg. Según OISEVI Se incluyen los vehículos con una cilindrada inferior a 50cc que no estén incluidos en la definición de ciclomotor.
- Vehículo Deportivo: Vehículo fabricado con carrocería cerrada o abierte con techo fijo o desmontable.

1.4.10 Machine Learning (ML)

Según (IBM, 2021), Machine Learning es una forma de inteligencia artificial que permite a un sistema ‘aprender’ de los datos en lugar de aprender mediante programación explícita. Para lograr esto, los modelos de ML necesitan un conjunto de datos para entrenarse y por lo general se validan con una misma porción de ese conjunto de datos que no fueron parte del entrenamiento.

Machine Learning cubre las siguientes técnicas de aprendizaje:

- Aprendizaje Supervisado: Por lo general, comienza con un conjunto de datos establecido y una variable resultante que nos permite una cierta comprensión de cómo se clasifican los datos. El aprendizaje supervisado

tiene la meta de encontrar patrones dentro de los datos que se puedan aplicar a un proceso de analítica.

- Aprendizaje No supervisado: Se utiliza cuando el problema requiere de una cantidad masiva de datos sin etiquetar. Entre ellos están los métodos de clustering y asociación.
- Aprendizaje de Refuerzo: Es un modelo de aprendizaje conductual. Este algoritmo recibe una retroalimentación del análisis de datos, conduciendo al usuario hacia el mejor resultado. El sistema como tal aprende a través de la prueba y error ya que no tiene un modelo de entrenamiento, usa secuencia de decisiones exitosas que conducen al fortalecimiento del proceso.
- Deep Learning: Es un método más específico de ML que incorpora redes neuronales en capas sucesivas para aprender de los datos de manera iterativa. Esta técnica resulta cuando se trata de encontrar patrones dentro de datos no estructurados. Las redes neuronales están diseñadas para emular cómo funciona el cerebro humano. Las redes neuronales y el deep learning se utilizan a menudo en el mundo de reconocimiento de imágenes, voces y aplicaciones de visión de computadoras para detectar cosas específicas (IBM, 2021).

CAPÍTULO 2

2. METODOLOGÍA

En este capítulo se presentan los procedimientos utilizados para la construcción de los modelos de pronóstico de severidad de los accidentes según la información de la Agencia Nacional de Transito (ANT). Para el pronóstico que buscamos hemos utilizado los modelos de SVM, Random Forest, Regresión Multinomial.

2.1 Supporting Vector Machine (SVM)

Esta técnica trata de encontrar alguna curva que permita separar los distintos grupos de datos que se encuentren en el set de datos. Este algoritmo tiene bondades de ajuste usando clasificadores lineales y no lineales (Géron A., 2017).

2.1.1 SVM – Linear Classifier

Esta técnica trata de ajustar una línea recta que pueda separar los distintos grupos que hayan dentro de los datos. Para entender un poco más este concepto veamos la *Imagen 1*, en la parte izquierda podemos ver los grupos Iris-Versicolor e Iris-Setosa, los cuales pueden ser claramente separados con una línea recta, y vemos que hay varias rectas que podrían separarlos correctamente (como la línea roja o la línea rosada) pero de la misma manera hay rectas como la punteada de verde que ofrece una predicción tan errada que incluso está separando dentro del mismo grupo de Iris-Versicolor (Géron A., 2017).

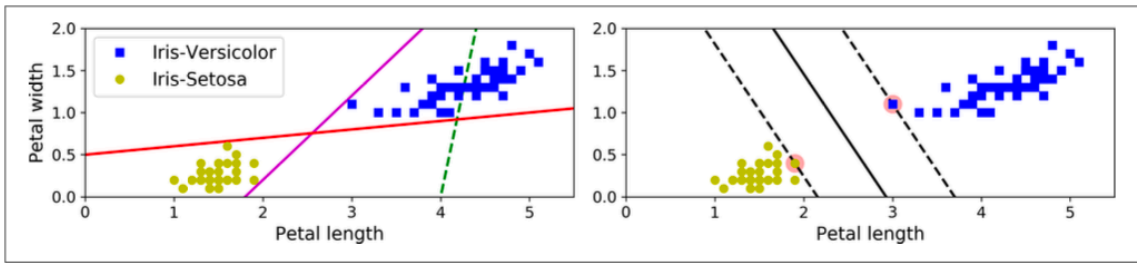


Imagen 1. Gráfico de Petal Length vs Petal Width del dataset Iris.

2.1.2 SVM – Non Linear Classifier

Pese a que los clasificadores lineales son muy eficientes, hay algunos sets de datos que no están ni cerca a ser linealmente separables por lo que se puede abordar estas situaciones agregando variables para usar un clasificador polinomial. Por ejemplo, en la *Imagen 2* en la parte izquierda, si tratáramos de usar una sola variable explicativa y tratáramos de usar un clasificador lineal, no existe alguna línea recta que pueda separar los dos grupos correctamente, pero si agregamos una variable explicativa adicional vemos en la parte derecha que podemos dibujar una línea horizontal que separe estos grupos correctamente (Géron A., 2017).

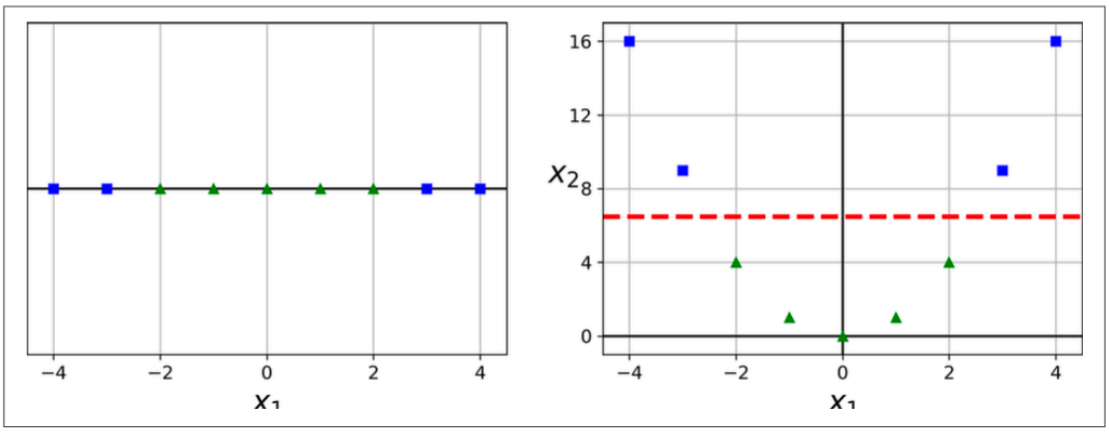


Imagen 2. Scatterplot de grupos con una y dos variables (lineal vs polinomial).

Bajo esta misma idea de usar más de una variable explicativa, también se nos pueden presentar datos donde una línea recta no pueda separar los grupos correctamente, y es en estos casos donde usamos clasificadores no lineales que cumplan con separar los grupos correctamente como podemos ver en la *Imagen 3*

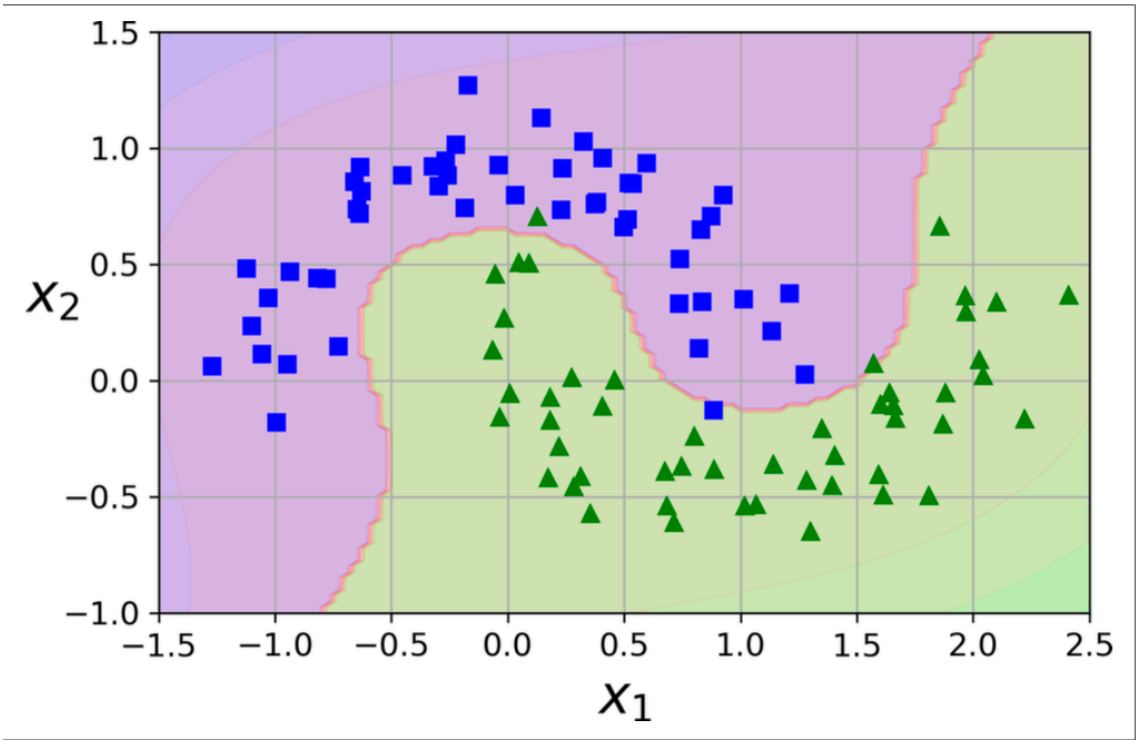


Imagen 3. Scatterplot de grupos separados por una curva.

2.2 Random Forest

Esta ténica usa como base Decisions Trees pero entrena distintos árboles en distintas sub-muestras del set de datos y toma el pronóstico más votado entre los distintos árboles. (Géron A., 2017).

2.2.1 Decision Tree

Son algoritmos de probabilidades o árboles como tal, donde se va avanzando dentro del árbol según el orden del algoritmo comenzando con la pregunta relacionada a la variable que más variabilidad represente y al final del árbol se termina con una rama que cumple con las características de la observación y lleva a una respuesta específica.

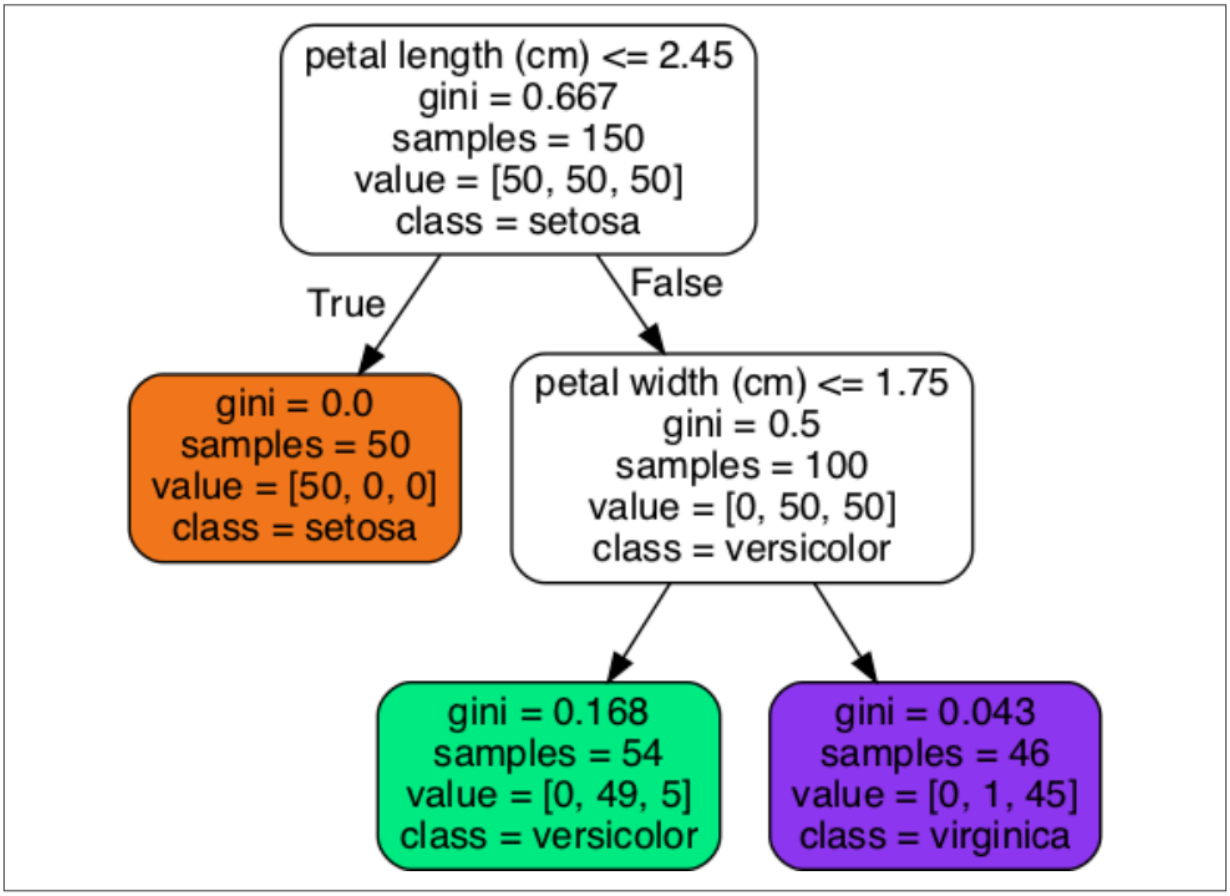


Imagen 4. Decision Tree de clasificación de especie de plantas

2.3 Regresión Logística

Este modelo estadístico estima, como tal, la probabilidad de que un objeto con ciertas características (las variables independientes) pertenezca a un determinado grupo. Se toma que si la probabilidad es >= 0.5 entonces pertenece al grupo en caso contrario no pertenece al grupo (Géron A., 2017).

2.3.1 Regresión Logística Multinomial

Este modelo estadístico es parecido a la regresión logística convencional, pero con la diferencia que es un modelo más general que no está restringido a dar una respuesta binaria (IBM,2021).

2.4 Clustering Jerárquico

Según (IBM,2021) es un procedimiento que trata de identificar grupos relativamente homogéneos de variables, basándose en las características seleccionadas mediante un algoritmo que comienza con cada variable en un cluster distinto y combina clusters hasta quedarse solo con uno final. Es posible analizar las variables directamente como vienen en los datos o aplicar dentro de una variedad de transformaciones de estandarización.

2.5 Distancia de Gower

Es un coeficiente el cual mide la similitud entre dos unidades de muestreo. La matriz de similitudes entre todos los pares de unidades de muestreo se lo encuentra como positivo y semidefinido (a excepción de los casos donde haya valores faltantes) (Gower J.C., 1971).

2.6 Mapa de concentración

Es una herramienta visual, la cual nos permite ver y cuantificar dónde están sucediendo las cosas. Al momento de crear mapas sobre comportamientos de la población, se puede observar si existen tendencias o patrones de donde ocurre más frecuentemente el fenómeno a estudiar.

2.7 Selección de variables

Para la selección de variables usamos un modelo Tree-Based Estimators para poder identificar que variables contienen la mayor cantidad de variabilidad con respecto a nuestro resultado.

2.7.1 Random Forest Feature Importance

Como lo revisamos previamente, Random Forest son una colección de árboles de decisión, que van generando distintos conjuntos de datos aleatorios y se producen votos de los distintos árboles para decidir la clase final de la variable de respuesta. De la misma manera, pueden ir seleccionando al azar variables que generen subconjuntos aleatorios y una clase final que sea la variable de respuesta, estos

mismos van identificando que variables son las más importantes al momento de generar una clase final que concuerde con la variable de respuesta.

Usando este modelo hemos seleccionado las variables:

- Mes: Mes del año que ocurrió el accidente (enero, febrero, marzo,…, diciembre).
- Día: Día de la semana que ocurrió el siniestro (lunes, martes, miércoles, jueves, viernes).
- Periodo: Hora del día (0, 1,2, 3…, 21, 22,23).
- Feriado: Un indicador si ocurrió durante un feriado o no (sí, no).
- Código Causa: Categorías de accidentes.
- Clase Final: Categoría de tipo de accidente.
- Zona: Zona dentro del cantón Guayaquil.
- Tipo de Vehículo: Clasificación de vehículo afectado en base a RLOTTTSV
- Suma de Vehículos: Cantidad de vehículos involucrados.
- Edad: Edad del conductor (años).
- Sexo: Identificador del género del conductor (Masculino, Femenino).
- Casco: Identificador si tenía o no el casco puesto (sí, no).
- Participante: Involucrado dentro del accidente de quien se reporta el estado final. (Conductor, pasajero, peatón)
- Cinturón: Indicador si tenía el cinturón puesto (sí, no).

El software que se utilizó fue R version 4.1.2 por medio del IDE RStudio 2021.09.1
Los paquetes/librerías usadas fueron:
Readxl, ggplot2, scales, caret, randomForest, e1071, dplyr, nnet, MASS, ggmap, RColorBrewer, osmdata, pacman, ClustOfVar, cluster.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

En este capítulo se muestran los resultados de la evaluación de los modelos de aprendizaje estadístico para predecir el nivel de siniestralidad de los siniestros de tránsito.

Para el análisis se recogieron los siniestros y accidentes registrados por la ANT desde enero del 2021 hasta agosto del 2022, de los cuales se obtuvieron 7,142 registros para la ciudad de Guayaquil, de los cuales 4,126 tenían información faltante y 15 de ellos no correspondían con el límite cantonal por lo que se procedió con su eliminación obteniendo 3,001 registros.

Como variable dependiente se tomó CONDICIÓN, la cual consta de 3 niveles de severidad (ileso, lesionado, fallecido), mientras que las variables regresoras que se utilizaron fueron Día, Mes, Período, Feriado, Clase final, Código de causa, Número de personas afectadas, Zona, Tipo de Vehículo, Suma de Vehículo, Edad, Sexo, Casco, Cinturón, Participante.

Modelo de Random Forest

En la Tabla 1. se muestra la matriz confusión del modelo de clasificación con la función Random Forest, en el cual por default se escogieron 500 árboles con 4 variables probadas en cada división obteniendo una tasa de error OOB (Out of bag) del 3.19%, mientras que la disminución total de las impurezas de los nodos a partir de la división en la variable mediante el índice de Gini muestra que el número de lesionados por siniestros, el uso del cinturón de seguridad, fueron aquellas variables que mayor importancia presentaron en el modelo, obteniendo un performance del 94.87%.

Tabla 1. Matriz de Confusión metodología Random Forest

Fuente: Elaboración propia

	FALLECIDO	ILESO	LESIONADO
FALLECIDO	27	4	0
ILESO	18	150	0
LESIONADO	12	12	677

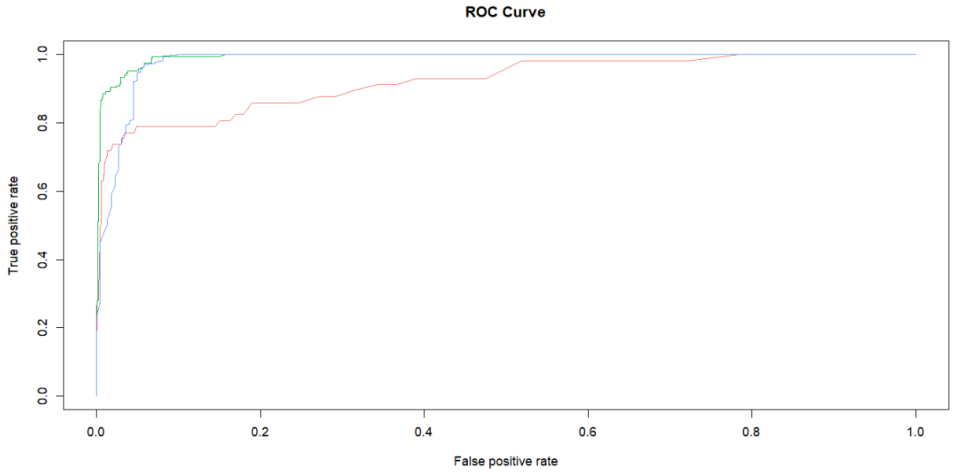


Imagen 5. Curva ROC modelo Random Forest con ntrees=500

3.1 Modelo Support Vector Machine

En la tabla Tabla 2. se muestra la matriz confusión del modelo de clasificación con la función tune a partir de un kernel lineal que realiza 10-cross-validation para identificar el valor óptimo de penalización. Entre sus argumentos están: el modelo svm y un vector ranges con los valores de los hiperparámetros que se quieren evaluar, en este caso se establecieron costos de 0.001, 0.01, 0.1, 1, 5, 10, 15 y 20, siendo el costo 1 según la imagen 5 con menor error perfomance 0.0323, con 356 vectores para las 3 clases, obteniendo un accuracy del 95.22%

Tabla 2. Matriz de Confusión metodología SVM

Fuente: Elaboración propia

	FALLECIDO	ILESO	LESIONADO
FALLECIDO	26	2	0
ILESO	19	154	0
LESIONADO	12	10	677

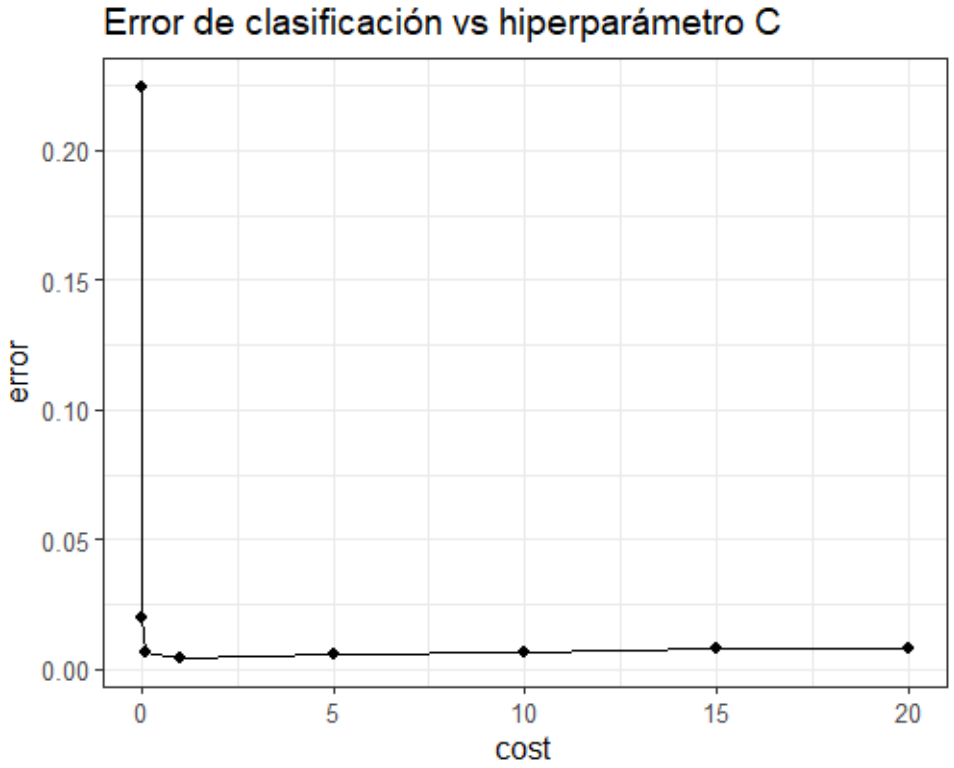


Imagen 6. Error de clasificación vs hiperparámetro C SVM

3.2 Modelo de Regresión Multinomial

Para el método de regresión multinomial, se utilizó la función multinom() del paquete nnet , en el cual ajusta un modelo multinomial log-lineal a través de redes neuronales luego de 100 iteraciones el modelo obtuvo un AIC de 776.23 y un accuracy del 94.56%.

Tabla 3. Matriz de Confusión metodología Regresión Multinomial

Fuente: Elaboración propia

	FALLECIDO	ILESO	LESIONADO
FALLECIDO	28	3	0
ILESO	17	147	1
LESIONADO	12	16	676

3.3 Clustering Jerárquico

Para tratar de establecer patrones entre los siniestros de tránsito se utilizó la función hclust() del paquete hclust el cual genera un análisis de clustering jerárquico utilizando un conjunto de disimilitud para los n objetos que se agrupan. Inicialmente, cada objeto se asigna a su propio cluster y luego el algoritmo procede de manera iterativa, en cada etapa uniendo los dos grupos más similares, continuando hasta que forme un solo grupo. En cada etapa, las distancias entre los conglomerados se vuelven a calcular mediante la fórmula de actualización de disimilitud de Lance-Williams, según el método de conglomerado particular que se utilice (Murtagh, F., 1985).

La matriz de disimilaridades se calculó a través de la métrica de Gower usada en datos mixtos y para la determinación del número de clusters se realizó por medio de 50 muestras bootstrap mostrando como resultado 6 clusters (Imagen 8).

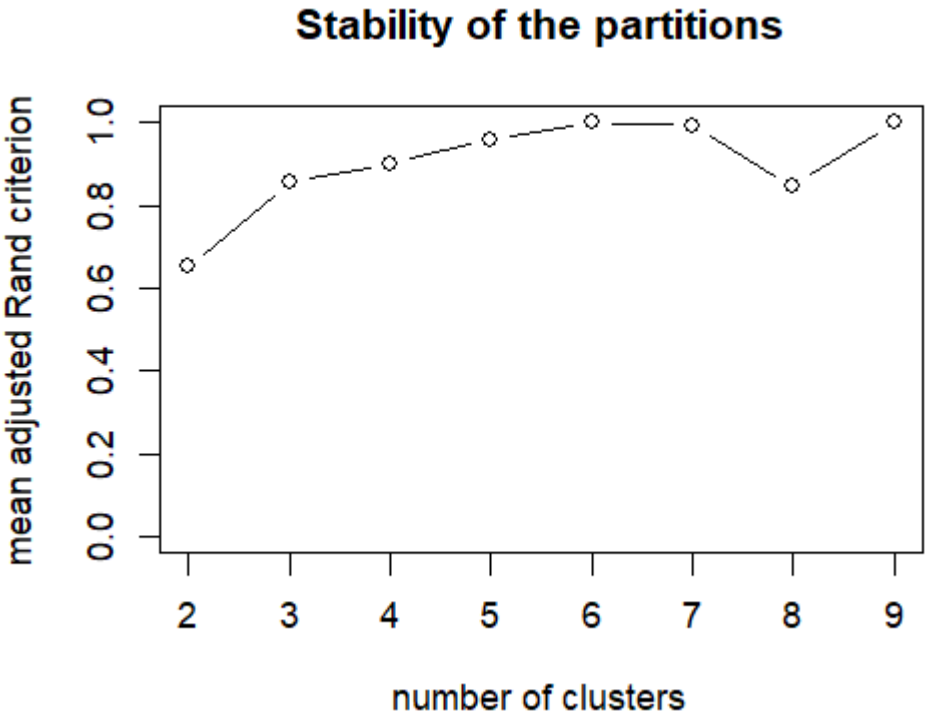


Imagen 7. Estabilidad de las particiones de los clusters bajo el criterio de media ajustada de Rand

	PERIODO	CLASEF	CODCAUSA	TIPOV	EDAD	SEXO	CINTURÓN	CONDICIÓN
CLUSTER 1	13 a 17	CHOQUE LAT	23,19	MOTO	30	H	N/A	LESIONADO
CLUSTER 2	7 a 10, 14 a 19	ATROPELLOS	9	AUTO	36	M	NO	LESIONADO-FALLECIDO
CLUSTER 3	0, 21 a 23	PERDIDA DE PISTA	6,9	AUTO	38	H	SI	ILESO
CLUSTER 4	7, 12 a 15	CHOQUE LAT	19,23	AUTO	39	H	SI	LESIONADO-ILESO
CLUSTER 5	18 A 22, 6 A 8	CHOQUE POSTERIOR	6,11,12	MOTO	33	H	N/A	LESIONADO
CLUSTER 6	14, 16 A 20	ATROPELLOS	9	MOTO	32	H	N/A	LESIONADO

Imagen 8. Representación de las características asignadas a cada cluster

3.4 Mapa de Calor

Usando un mapa de calor a partir de nuestros datos (Imagen 9), podemos observar la zona con más accidentes letales que se encuentra localizada en la zona norte de

la ciudad, agrupando sectores desde Mapasingue hasta Pascuales. Podemos asumir que estos siniestros ocurren debido al tipo de vías, estas son callejones aún de tierra y sin señalización, además hay abundancia de vehículos modificados como tricimotos o triciclos sin dejar a un lado los scooters (o bicimotos) que aún no están regularizados y los pueden usar niños o grupo de personas sin ninguna medida de seguridad. Sin embargo, dentro de todo el cantón se han registrado siniestros fatales, lo que no implica que únicamente en la zona norte ocurren este tipo de accidentes.

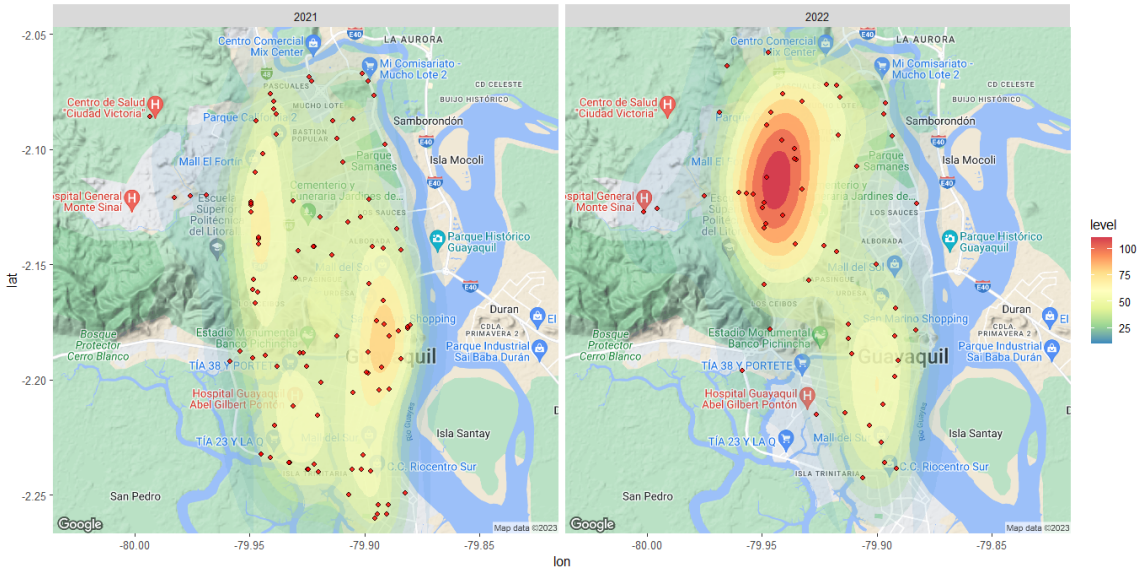


Imagen 9. Mapa de calor de accidentes de tránsito en el cantón Guayaquil

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- Los modelos realizados cumplen con el objetivo de predecir de manera correcta la severidad, usando SVM se obtuvo el mayor accuracy debido a la naturaleza de los datos; sin embargo, para el análisis se utilizó la metodología por Random forest por ser la más adecuada ya que al tener gran cantidad de predictores el modelo maneja implícitamente la colinealidad en las variables.
- Las variables que mayor importancia en el modelo de Random forest fueron: el uso de cinturón, la clase final, el casco y el participante.
- El análisis de clúster clasificó en 6 grupos los siniestros de tránsito, siendo los más destacados el clúster 2 representado por hombres de 36 años fallecidos en automóviles por atropellos.

4.2 Recomendaciones

- Tener una base de datos representativa, incluyendo otros factores que son importantes al momento de analizar siniestros de tránsito, por ejemplo, estado de la calzada, clima, velocidad, entre otros.
- Podríamos tomar nuestro modelo y validarlo con los datos de accidentes de todo el país para ver si se comporta de la misma manera los resultados de certeza sobre la siniestralidad del accidente.
- Limitaciones de software de uso libre para generar los gráficos, ya que tuvimos que suscribirnos a una API de Google que tiene un límite de uso gratuito y comienzan a generar una factura mensual de consumo.

- Podríamos replicar el modelo en otro lenguaje de programación para poder implementarlo como una API, dentro de una aplicación que se pueda ejecutar en versión web/móvil y que esté disponible para ambulancias, hospitales y miembros de las entidades de tránsito para que de esta manera puedan estar preparados al tipo de accidente que van a atender.
- Para mejorar el desempeño del modelo de Random Forest se puede convertir la columna ‘causa_probable’ en dummy para llevarla a 29 columnas binarias, de esta manera el modelo podría mejorar su accuracy ya que está diseñado para aguantar set de datos extensos en columnas.

BIBLIOGRAFÍA

ANT. (2022). *Agencia Nacional de Tránsito*. https://www.ant.gob.ec/visor-de-siniestralidad-estadisticas/

AutoMagazine. (2022). https://automagazine.ec/las-ventas-de-vehiculos-en-ecuador-a-mayo-de-2022/

Carro, E., y Ampudia, A. (2018). Conductas de riesgo al conducir un automóvil en zonas urbanas del sur de Tamaulipas y la Ciudad de México. 18. *https://doi.org/https://doi.org/10.29059/cienciauat.v13i2.988*

El Universo. (2022). https://www.eluniverso.com/guayaquil/comunidad/siete-vias-de-guayaquil-concentran-el-51-de-muertes-por-accidentes-de-transito-para-la-perimetral-y-narcisa-de-jesus-se-analiza-bajar-limites-de-velocidad-nota/

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. Jupyter. https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/

Gómez, A., Galarza, L., Merino, P., y Algora, A. (2017). Estudio geoespacial de los accidentes de tránsito en la Región Amazónica Ecuatoriana. *CienciAmérica*, 6. http://201.159.222.118/openjournal/index.php/uti/article/view/80/66

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27(4), 857–871. https://doi.org/10.1109/ultsym.1987.199076

IBM. (2022). *IBM*. https://www.ibm.com/mx-es/analytics/machine-learning

INEC. (2021). *Instituto de Estadísticas y Censos.* https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Estadistica%20de%20Transporte/2021/2021_NOTA%20T%C3%89CNICA_SINIESTROS.pdf

IT4AGRI. (2021). https://it4agri.com/feature-selection-using-tree-based-method-and-recursive-feature-eliminationrfe/

Montes, S., y Ledesma, R. (2017). MUERTES POR SINIESTROS DE TRÁNSITO EN LA PROVINCIA DE BUENOS AIRES EN 2017: UN ANÁLISIS MEDIANTE

MÉTODOS DE CLASIFICACIÓN JERÁRQUICA. 10. http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1853-810X2021000100051

OMS. (2017). Organización Mundial de la Salud. https://apps.who.int/iris/bitstream/handle/10665/255308/9789243511702-spa.pdf;sequence=1

OMS (2022). https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries

Planzer, R. (2005). *Repositorio de CEPAL.* https://repositorio.cepal.org/bitstream/handle/11362/6296/S05804_es.pdf?sequence=1&isAllowed=y

Primicias. (2022): https://www.primicias.ec/noticias/sociedad/muertes-accidentes-transito-ecuador-movilidad/

APÉNDICES

Código en R usado para elaboración del heat map

```
mean.longitude <- mean(fallecidos$LONGITUD)
mean.latitude <- mean(fallecidos$LATITUD)

pacman::p_load(ggmap, osmdata)
drone.map <- ggmap(get_googlemap(center = c(mean.longitude, mean.latitude), zoom = 1
## Convert into ggmap object
#drone.map <- ggmap(drone.map, extent="device", legend="none")

drone.map <- drone.map + stat_density2d(data=fallecidos,
                                        aes(x=LONGITUD, y=LATITUD, fill=..level.., a
## Define the spectral colors to fill the density contours
drone.map <- drone.map + scale_fill_gradientn(colours=rev(brewer.pal(7, "Spectral"))
drone.map <- drone.map + geom_point(data=fallecidos,
                                    aes(x=LONGITUD, y=LATITUD), fill="red", shape=21
## Remove any legends
drone.map <- drone.map + guides(size=FALSE, alpha = FALSE)
drone.map <- drone.map + facet_wrap(~AÑO)
print(drone.map)
```