



Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Aplicación y valoración de algoritmos de Machine Learning
para la predicción de gravedad en accidentes
automovilísticos.

Junio-2023

Tesis profesional

Para obtener el título de Licenciatura en Ingeniería en
Ciencias de la computación.

Presenta:

Juan José Osorio Hernández

Director de Tesis

Dra. María Teresa Torrijos Muñoz

Asesor

M. I. Carlos Armando Ríos Acevedo

Agradecimientos

Es increíble como esta sección aparece en el inicio de este documento de tesis, cuando realmente fue lo último que termine realizando.

Me gustaría iniciar agradeciendo a mi madre, quien en toda mi carrera profesional me apoyo de tantas maneras incontables, lo que soy actualmente es gracias a ella y a los consejos que día a día me da, agradecer también a mis hermanos y hermana, a mi padre, a mi tía y mi tío Gau que fue un pilar muy importante en el desarrollo de mi carrera, muchas gracias, en serio esto es para ustedes y por ustedes.

Júntate con lobos y aullar te enseñas, con este refrán quiero agradecer también a todos mis amigos con los que tuve el gusto de compartir clases, momentos, conocimientos y grandes experiencias que ayudaron a explotar mis habilidades.

A mi asesora, la Dra. María Torrijos Muñoz agradecerle enormemente la paciencia que me tuvo durante todo este tiempo, por todos los consejos, por la confianza que deposito en mí y por su ardua dedicación en la realización de esta tesis.

Tomare este espacio para agradecer a mi novia Jessica que me acompañó en desveladas, que en días me dio una razón, una idea, un consejo que me ayudaba a avanzar, su apoyo incondicional en cada momento, muchas gracias.

Finalmente agradecer a cada profesor con los que tuve el gusto de tomar clases, sus conocimientos, sus consejos todo lo que me compartieron forjo al estudiante que soy ahora, muchas gracias a cada uno de ustedes.

Índice General

Introducción	7
Objetivos	8
Objetivo general	8
Objetivos específicos	8
Estructura de la tesis.....	9
Capítulo 1 Estado del arte	10
Accidentes de Tránsito Terrestre en Zonas Urbanas y Suburbanas	10
Tendencias tecnológicas.....	11
Inteligencia artificial	11
Machine Learning	13
Tipos de Machine Learning.....	14
Aplicaciones.....	15
Redes Neuronales.....	16
Tipos de Redes Neuronales	17
Aplicaciones de las Redes Neuronales.....	17
Deep Learning.....	18
Aplicaciones.....	18
Accuracy	19
Desbalanceo de datos	19
Capítulo 2 Extracción y limpieza de datos	20
Adquisición de datos	20
Limpieza de datos	24
Validación de valores nulos	24
Conversión de valores categóricos.....	25
Balanceo de datos.....	27
Selección de Características.....	29
División de datos.....	30
Capítulo 3 Modelos a implementar	31
Arboles de Decisión	31
KNN	31
Redes Neuronales.....	32
Random Forest	33

Capítulo 4 Resultados	34
Arboles de Decisión	34
KNN	36
Redes Neuronales	38
Función identity	40
Función Relu	41
Función Tanh	43
Random Forest	45
Capítulo 5 Conclusiones.....	47
Conclusiones Modelos	47
Conclusiones generales	48
Referencias	49

Índice de ilustraciones

Ilustración 1. Tecnologías Top para el 2022	11
Ilustración 2. Tipos de enfoque de IA	12
Ilustración 3. Aplicaciones de IA	13
Ilustración 4. Partes principales del ML	14
Ilustración 5. Aplicaciones del ML	15
Ilustración 6. Modelo de una ANN [8]	16
Ilustración 7. Tipos de Redes Neuronales	17
Ilustración 8. Adquisición de datos [11]	20
Ilustración 9. Proceso limpieza de datos	24
Ilustración 10. Visualización valores nulos	24
Ilustración 11. Totales valores nulos	25
Ilustración 12. Columnas categóricas	26
Ilustración 13. Detalle Variables Categóricas Antes de Convertir	26
Ilustración 14. Detalle Variables Categóricas Después de Convertir	26
Ilustración 15. Desbalance de datos	27
Ilustración 16. Datos balanceados	28
Ilustración 17. Variables Relevantes	29
Ilustración 18. División de datos	30
Ilustración 19. Árbol de decisiones	31
Ilustración 20. Método Elbow	32
Ilustración 21. Accuracy Árbol de Decisiones	34
Ilustración 22. Tiempo de ejecución árbol de decisiones	35
Ilustración 23. Matriz de Confusión Profundidad 5	35
Ilustración 24. Matriz de Confusión Profundidad 40	35
Ilustración 25. Matriz de confusión K=5	36
Ilustración 26. Matriz de confusión KNN K=3	37
Ilustración 27. Accuracy Logistic	38
Ilustración 28. Tiempo de ejecución Logistic	39
Ilustración 29. Matriz de confusión Logistic 10 neuronas	39
Ilustración 30. Matriz de confusión Logistic 30 neuronas	39
Ilustración 31. Accuracy Identity	40
Ilustración 32. Tiempo de ejecución Identity	40
Ilustración 33. Matriz de confusión Identity 30 neuronas	41
Ilustración 34. Matriz de confusión Identity 40 neuronas	41
Ilustración 35. Accuracy Relu	41
Ilustración 36. Tiempo de ejecución Relu	42
Ilustración 37. Matriz de confusión Relu 20 neuronas	42
Ilustración 38. Matriz de confusión Relu 30 neuronas	42
Ilustración 39. Accuracy Tanh	43
Ilustración 40. Tiempo de ejecución Tanh	43
Ilustración 41. Matriz de confusión Tanh 40 neuronas	44
Ilustración 42. Matriz de confusión Tanh 10 neuronas	44
Ilustración 43. Accuracy Random Forest	45
Ilustración 44. Tiempo ejecución Random Forest	46
Ilustración 45. Matriz de confusión Random Forest, Arboles=40	46

Ilustración 46. Matriz de confusión Random Forest, Arboles=10	46
Ilustración 47. Precisión Modelos	48

Índice de tablas

Tabla 1. Estructura dataset	24
Tabla 2. Total, Instancias Desbalanceadas	28
Tabla 3. Total, Instancias Balanceadas	28
Tabla 4. Total instancia Train-Test	30
Tabla 5. Tiempo de ejecución KNN.....	37

Introducción

Se define como accidente automovilístico o vial al suceso entre la colisión de un vehículo contra uno o más sectores de la vialidad, como lo pueden ser otro vehículo, algún individuo, animal o escombros en el camino. Los accidentes de este tipo provocan daños materiales, pérdidas humanas lesiones de gravedad para los involucrados en el accidente. Normalmente estos están acompañados por corresponsabilidades del conductor, además de algunas otras fuentes externas, como puede ser el clima, conducir en estado de ebriedad, hacer caso omiso a las señalizaciones de vialidad, conducir a exceso de velocidad, realización de maniobras peligrosas, etc.

La Organización de las Naciones Unidas, desde 1995 estableció el tercer domingo de noviembre como el Día Mundial en Recuerdo de las Víctimas de los Accidentes de Tráfico [9], a fin de sensibilizar a la población mundial sobre los riesgos y consecuencias que ocasionan estos eventos viales.

En México, el programa de información del INEGI Accidentes de tránsito terrestre en zonas urbanas y suburbanas [9] comparte los siguientes datos sobre accidentes automovilísticos.

- En 2020 se registraron 301,678 accidentes de tránsito en las zonas urbanas de México. Durante ese año, uno de cada 100 eventos de tránsito correspondió a accidentes en los que se registraron pérdidas humanas, mientras que en 18 de cada 100 hubo víctimas heridas.
- El total de víctimas muertas y heridas en los accidentes de tránsito ocurridos en zonas urbanas durante 2020 fue de 75 761 personas, de las cuales 3 826 fallecieron en el lugar del accidente (5.1%) y 71 935 presentaron algún tipo de lesión (94.9%).
- En 2020 se reportaron 301 678 accidentes, de los cuales 245 297 registraron solo daños materiales (81.3%); en 52 954 se identificaron víctimas heridas (17.6%), y los 3 427 accidentes restantes corresponden a eventos con al menos una persona fallecida (1.1%) en el lugar del accidente.

A partir de la información que ha sido generada y almacenada se pueden hacer inferencias que ayuden a predecir la gravedad de un accidente con base en el comportamiento de sus variables.

La realización de este trabajo de tesis recolecta y transforma datos, para poder implementar y evaluar una serie de algoritmos de machine learning y así obtener el modelo con el mejor porcentaje de precisión al momento de predecir la severidad de un accidente automovilístico. Para esto se usará información extraída de los datos abiertos del portal del INEGI del año 2021, referentes a Accidentes de Tránsito Terrestre en Zonas Urbanas y Suburbanas [9].

Objetivos

Objetivo general

Valoración de diferentes algoritmos de machine learning para identificar qué modelo genera una mejor precisión de aprendizaje y predecir la gravedad de un accidente automovilístico con una precisión superior al 80%.

Objetivos específicos

1. Aplicar técnicas de pre procesamiento de datos como es la exploración, conversión de datos, imputación.
2. Dividir el conjunto total de datos en 75% para entrenamiento y un 25% para pruebas.
3. Implementar balanceo de datos acorde a los modelos.
4. Implementar algoritmos de selección de características para generar datos de manera consistente y de calidad para la aplicación de algoritmos de machine learning.
5. Entrenar modelos de machine learning supervisados para el dominio del problema.
6. Predicción de la gravedad de un accidente con diferentes algoritmos de machine learning
7. Proporcionar un reporte sobre el modelo con mejor porcentaje de exactitud al predecir la gravedad de un accidente automovilístico.

Estructura de la tesis

Capítulo 1

En el capítulo 1 se presenta el estado del arte para la realización de este trabajo de tesis partiendo de los accidentes automovilísticos en México hasta la definición de algunos términos y algunos temas que se requieren para la comprensión de este trabajo, así como tendencias tecnológicas actuales.

Capítulo 2

En esta sección se detallan los procesos para la limpieza de datos, como la validación de datos nulos, procesos de conversión de datos categóricos, el balanceo de la información, la selección de características y la división de los conjuntos de prueba y entrenamiento. Estos procesos son considerados parte fundamental del desarrollo de este trabajo de tesis y de cualquier otro trabajo de machine learning pues la calidad de datos influye en el desempeño y comportamiento de los algoritmos.

Capítulo 3

En esta sección se revisa el funcionamiento y comportamiento de los modelos a utilizar

Capítulo 4

En este capítulo se muestran los resultados obtenidos al entrenar los modelos y su comportamiento al predecir nuevas instancias. Se hace uso de herramientas gráficas como lo es la matriz de confusión y además se revisan los tiempos de ejecución de cada modelo con base a la variación de algunos de sus parámetros más importantes.

Capítulo 5

En esta última sección de este trabajo de tesis se agregan las conclusiones generales del trabajo realizado en los capítulos anteriores además se incluyen trabajos futuros que pudieran proporcionar mejores resultados en trabajo o problemas similares

Capítulo 1 Estado del arte

Accidentes de Tránsito Terrestre en Zonas Urbanas y Suburbanas

Se define como accidente automovilístico o vial al suceso entre la colisión de un vehículo contra uno o más sectores de la vialidad, como lo pueden ser otro vehículo, algún individuo, animal o escombros en el camino. Los accidentes de este tipo provocan daños materiales, pérdidas humanas lesiones de gravedad para los involucrados en el accidente. Normalmente estos están acompañados por responsabilidades del conductor, además de algunas otras fuentes externas, como puede ser el clima, conducir en estado de ebriedad, no respetar las señalizaciones de vialidad, conducir a exceso de velocidad, imitar maniobras peligrosas, etc.

La Organización de las Naciones Unidas, desde 1995 estableció el tercer domingo de noviembre como el Día Mundial en Recuerdo de las Víctimas de los Accidentes de Tráfico [9], a fin de sensibilizar a la población mundial sobre los riesgos y consecuencias que ocasionan estos eventos viales.

En México, el programa de información del Instituto Nacional de Estadística y Geografía (INEGI), Accidentes de tránsito terrestre en zonas urbanas y suburbanas [9] comparte los siguientes datos sobre accidentes automovilísticos.

- En 2020 se registraron 301,678 accidentes de tránsito en las zonas urbanas de México. Durante ese año, uno de cada 100 eventos de tránsito correspondió a accidentes en los que se registraron pérdidas humanas, mientras que en 18 de cada 100 hubo víctimas heridas.
- El total de víctimas muertas y heridas en los accidentes de tránsito ocurridos en zonas urbanas durante 2020 fue de 75 761 personas, de las cuales 3 826 fallecieron en el lugar del accidente (5.1%) y 71 935 presentaron algún tipo de lesión (94.9%).
- En 2020 se reportaron 301 678 accidentes, de los cuales 245 297 registraron solo daños materiales (81.3%); en 52 954 se identificaron víctimas heridas (17.6%), y los 3 427 accidentes restantes corresponden a eventos con al menos una persona fallecida (1.1%) en el lugar del accidente.

Tendencias Tecnológicas

Año con año la tecnología avanza y el mundo de las Ciencias de la Computación crece de forma exponencial. Ante esto se debe estar preparado para poder adaptarse y saber cómo afrontar estas nuevas tendencias, tomarlas como una ventaja y crecer para un bien común o estar preparados para ser superados por ellas.

Garnet [1] es una empresa estadounidense que realiza investigaciones y análisis de Tecnologías de la Información (IT), cuenta con una extensa base de datos de información del mercado y realiza análisis de benchmarking sobre IT, finanzas, ventas, marketing y operaciones. En un estudio [2] que realiza acerca de cuáles serán las tecnologías top para el 2022, con base en tendencias tecnológicas actuales y en función de las prioridades de los CEO para sus organizaciones se puede encontrar la IA como una de las tecnologías con más valor de negocio y como una de las tecnologías que cuenta con más campo de estudio.



Ilustración 1. Tecnologías Top para el 2022

Inteligencia Artificial

La inteligencia artificial no es un área nueva en las ciencias de la computación como todos creen, tampoco tiene que ver con robots conquistando el mundo y revelándose contra la humanidad. John McCarthy [3] ofrece la siguiente definición de la inteligencia artificial

“Es la ciencia y la ingeniería de la fabricación de máquinas inteligentes, especialmente programas informáticos inteligentes. Está relacionada con la tarea

similar de usar computadoras para entender la inteligencia humana, pero la IA no tiene que limitarse a métodos que son biológicamente observables”

Sin embargo, años antes de esta definición, todo inicio con el trabajo trascendental "Maquinaria computacional e inteligencia" de Alan Turing, a menudo citado como el padre de la informática, el cual se hace la siguiente pregunta “¿Puede pensar una máquina?”, lo que se generó a partir de esta pregunta es la conocida “Prueba de Turing”, generando la interrogante, si un humano podría distinguir entre la respuesta de texto de una computadora y la de un humano.

En un estudio [5] realizado por Stuart Russell y Peter Norvig, se profundiza en cuatro posibles objetivos o definiciones de IA, que se diferencian de los sistemas informáticos en la base de la racionalidad y el pensamiento vs. el actuar. La definición de Alan Turing está bajo la categoría de “Sistemas que actúan como humanos”.

Enfoque humano	Sistemas que piensan como los humanos
	Sistemas que actúan como los humanos
Enfoque ideal	Sistemas que piensan racionalmente
	Sistemas que actúan racionalmente

Ilustración 2. Tipos de enfoque de IA

En su forma más simple, la inteligencia artificial es un campo que combina la ciencia informática y los conjuntos de datos robustos para permitir la resolución de problemas. También abarca los subcampos del machine learning y el Deep learning, que se mencionan frecuentemente junto con la inteligencia artificial. Estas disciplinas están compuestas por algoritmos de IA que buscan crear sistemas expertos que hagan predicciones o clasificaciones basadas en datos de entrada.

La IA al ser un campo un extenso se divide en 2 tipos: IA débil e IA robusta. La IA débil, también llamada IA estrecha (ANI), es un tipo de IA entrenada y enfocada a realizar tareas específicas. La IA débil impulsa la mayor parte de la IA que hoy nos rodea. "Estrecha" podría ser una descripción más precisa de este tipo de IA, ya que es cualquier cosa menos débil; se puede encontrar en algunas aplicaciones muy robustas, como Siri de Apple, Alexa de Amazon, IBM Watson y vehículos autónomos. La IA robusta está compuesta por la inteligencia artificial general (IAG) y la súper inteligencia artificial (SIA). La inteligencia artificial general (IAG), o la IA general, es una forma teórica de IA en la que una máquina tendría una inteligencia igual a la de los humanos; sería autoconsciente y tendría la capacidad de resolver problemas, aprender y planificar para el futuro. La súper inteligencia artificial (SIA), también conocida como súper inteligencia, superaría la inteligencia y la capacidad del

cerebro humano. Si bien la IA robusta todavía es completamente teórica y no tiene ejemplos prácticos de uso actualmente, no significa que los investigadores de IA no estén también explorando su desarrollo.

Hoy en día existen numerosas aplicaciones de la IA, algunos de estos ejemplos más comunes son; reconocimiento de biométricos (voz, fácil, dactilar), servicio al cliente, visión artificial, motores de recomendación, comercio de acciones automatizado, son solo algunos ya que es increíble la velocidad a la que está creciendo y adaptándose la IA en diferentes campos.

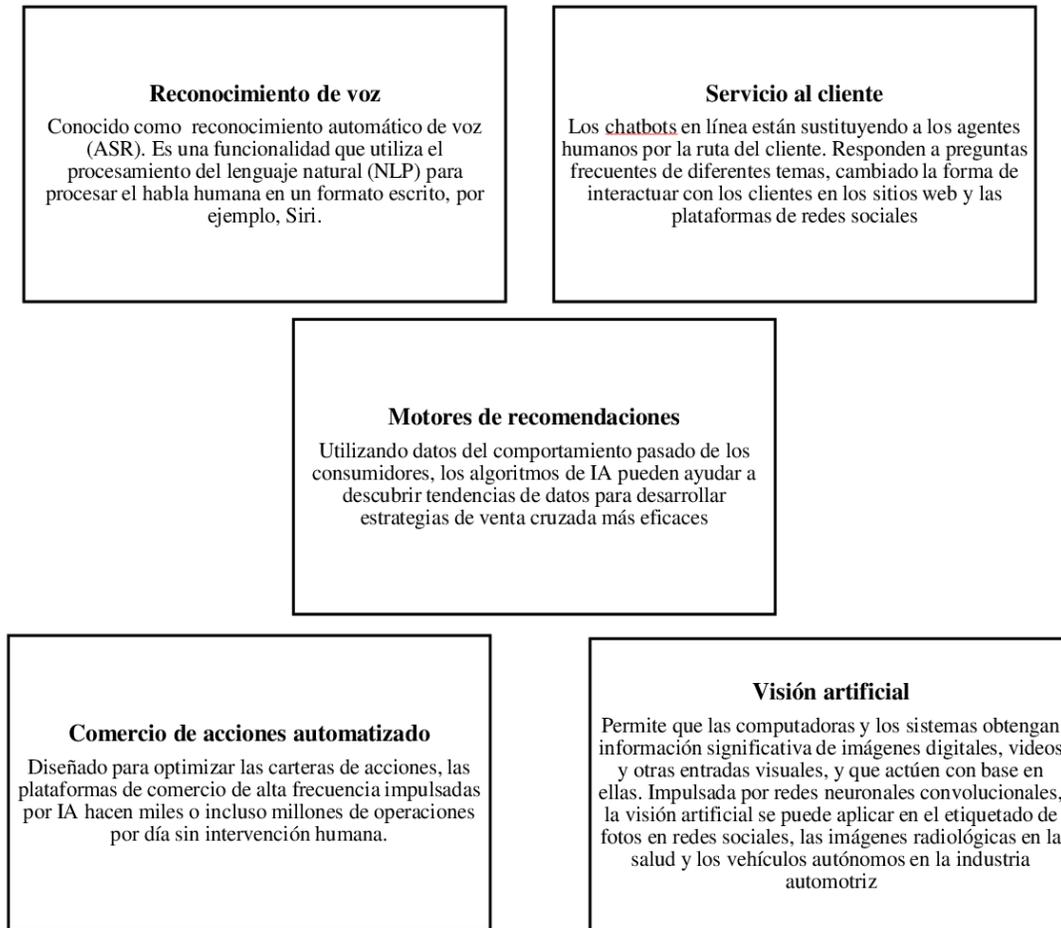


Ilustración 3. Aplicaciones de IA

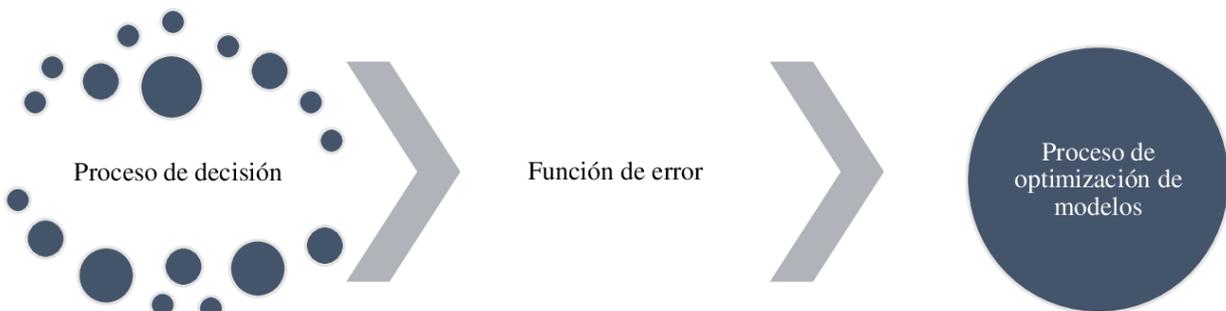
Machine Learning

El machine learning es una rama de la inteligencia artificial (IA) y la ciencia de computación que se centra en el uso de datos y algoritmos para imitar la forma en que los seres humanos aprenden, con una mejora gradual de su precisión. Este término fue acuñado por Arthur L.

Samuel en 1959, en su investigación [6] en donde jugó el juego de damas en una computadora IBM 7094 en 1962, y perdió ante la computadora.

Machine learning es un componente importante del creciente campo de la Ciencia de Datos. Mediante el uso de métodos estadísticos, los algoritmos se capacitan para hacer clasificaciones o predicciones, descubriendo conocimientos clave dentro de los proyectos de minería de datos. Estos conocimientos posteriores impulsan la toma de decisiones dentro de aplicaciones y empresas, lo que es ideal para influir en las métricas. A medida que los big data continúan expandiéndose y creciendo, la demanda del mercado de los científicos de datos aumentará, requiriendo que ayuden en la identificación de las preguntas de negocios más relevantes y posteriormente los datos para responderlas.

UC Berkeley interrumpe el sistema de aprendizaje de un algoritmo de machine learning en tres partes principales. Un proceso de decisión, una función de error, un proceso de optimización de modelos.



En general, los algoritmos de machine learning se utilizan para realizar una predicción o clasificación. Basándose en los datos de entrada, que pueden estar etiquetados o no, el algoritmo generará una estimación sobre un patrón en los datos

Sirve para evaluar la predicción del modelo. Si hay ejemplos conocidos, una función de error puede hacer una comparación para evaluar la precisión del modelo

Si el modelo puede ajustarse mejor a los puntos de datos del conjunto de entrenamiento, las ponderaciones se ajustan para reducir la discrepancia entre el ejemplo conocido y la estimación del modelo. El algoritmo repetirá este proceso de evaluación y optimización, actualizando los pesos de forma autónoma hasta que se haya cumplido un umbral de precisión

Ilustración 4. Partes principales del ML

Tipos de Machine Learning

Como tal podemos encontrar que el machine learning se puede dividir en 4 modelos, las cuales son; supervisado, no supervisado, semisupervisado y por refuerzo. Cada uno de estos tiene una interacción distinta al utilizarse para un conjunto de datos.

- Machine learning supervisado

Se define por su uso de los conjuntos de datos etiquetados para entrenar los algoritmos para clasificar datos o predecir resultados con precisión. A medida que se introducen datos de entrada en el modelo, adapta sus pesos hasta que el modelo se haya ajustado correctamente.

- Machine learning no supervisado

Utiliza algoritmos de machine learning para analizar y agrupar conjuntos de datos sin etiquetar. Estos algoritmos descubren patrones ocultos o agrupaciones de datos sin necesidad de intervención humana.

- Aprendizaje semisupervisado

El aprendizaje semisupervisado ofrece un término medio entre el aprendizaje supervisado y no supervisado. Durante el entrenamiento, utiliza un conjunto de datos etiquetado más pequeño para guiar la clasificación y la extracción de características de un conjunto de datos más grande y sin etiquetar.

- Machine learning por refuerzo

El machine learning por refuerzo es un modelo de machine learning de comportamiento que es similar al aprendizaje supervisado, pero el algoritmo no se entrena utilizando datos de muestra. Este modelo aprende a medida que utiliza el método de prueba y error. Se reforzará una secuencia de resultados satisfactorios para desarrollar la mejor recomendación o política para un problema determinado.

Aplicaciones



Ilustración 5. Aplicaciones del ML

Tipos de redes neuronales

Al igual que el machine learning, las redes neuronales se pueden clasificar en diferentes tipos, las cuales se utilizan para distintos fines, en esta clasificación podemos encontrar; redes neuronales de propagación hacia delante o multicapa (MLP), redes neuronales convolucionales (CNN), redes neuronales recurrentes (RNN).

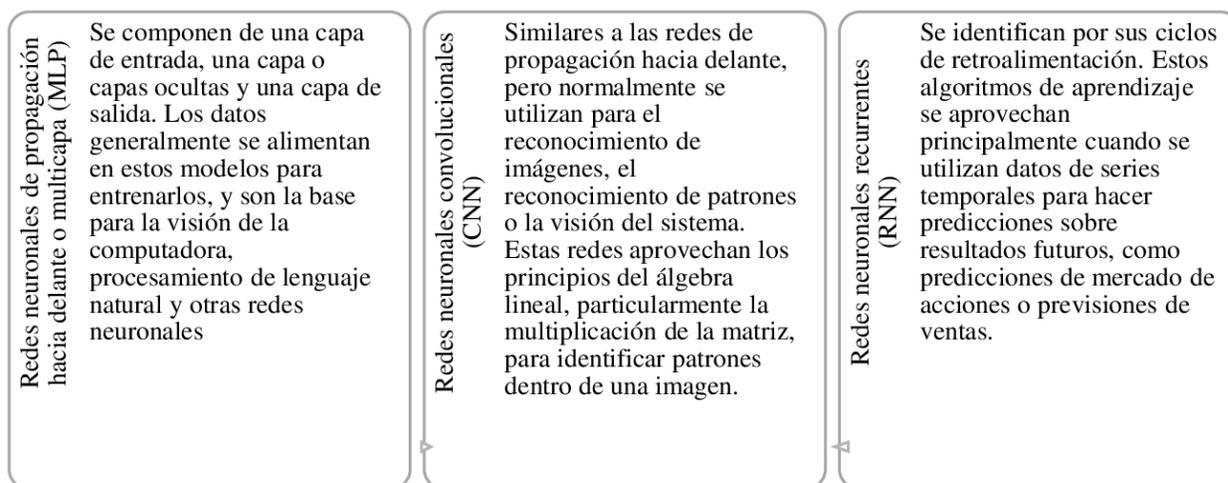


Ilustración 7. Tipos de Redes Neuronales

Aplicaciones de las Redes Neuronales

Las redes neuronales se aplican en diversas áreas desde hace ya muchos años, como es el caso de la Biología, que la usa principalmente para aprender más acerca del cerebro. Otra área en el que es habitualmente utilizada es en el campo del medio ambiente, donde se analiza la tendencia y patrones o la predicción del tiempo. En medicina, está funcionando excelentemente para la predicción de tumores o anomalías, así como para elaborar diagnósticos y tratamientos complejos a seguir, en función de unos síntomas. En el campo de las finanzas, se utiliza habitualmente en todo lo relacionado con la previsión de la evolución de precios, valoración de riesgo de créditos o identificación y falsificaciones.

En el ámbito de la empresa y más concretamente en el del marketing, tiene diversos usos:

- Predicción de ventas.
- Identificar patrones de comportamiento.
- Reconocimiento de caracteres escritos.
- Predicción del comportamiento del consumidor.
- Personalización de estrategias de marketing.
- Crear y entender segmentos de compradores más sofisticados.
- Automatizar actividades de marketing.
- Creación de contenido.

Deep Learning

El Deep learning es una sub área del machine learning que en esencia es una red neuronal con tres o más capas, al ser una red neuronal el Deep learning de igual manera intenta imitar el cerebro humano, aunque lejos de igualar su capacidad, lo que permite es aprender cantidades de datos enormes, agrupar datos y hacer predicciones con una precisión muy alta.

El Deep Learning usa una red neuronal artificial que se compone de un número de niveles jerárquicos. En el nivel inicial de la jerarquía la red aprende algo simple y luego envía esta información al siguiente nivel. El siguiente nivel toma esta información sencilla, la combina, compone una información algo un poco más compleja, y se lo pasa al tercer nivel, y así sucesivamente. Esto es un potenciador para aplicaciones y servicios de inteligencia artificial, esto mejora la automatización, realizando tareas analíticas y físicas sin que haya intervención humana.

Aplicaciones

Las aplicaciones de aprendizaje profundo del mundo real son parte de nuestra vida diaria, pero en la mayoría de los casos, están tan bien integradas en productos y servicios que los usuarios no son conscientes del complejo procesamiento de datos que se lleva a cabo en segundo plano. Algunos de estos ejemplos incluyen los siguientes:

- Utilización de imágenes en lugar de palabras clave para buscar productos de una empresa, o artículos similares.
- Monitorización en tiempo real de reacciones en canales online durante el lanzamiento de productos.
- Orientación de anuncios y predicción de las preferencias de los clientes.
- Identificación y seguimiento de los niveles de confianza de los clientes, sus opiniones y actitud en diferentes canales online y servicios de soporte automatizado al cliente.
- Detección de fraudes, recomendaciones a clientes, gestión de relaciones con los clientes, etc.
- Análisis de imágenes médicas, como radiografías y resonancias magnéticas, aumentando la precisión diagnóstica, en un menor tiempo y con un menor coste que los métodos tradicionales.
- Detección, predicción y prevención de amenazas sofisticadas en tiempo real en el campo de la ciberseguridad.
- Identificación en textos de sentimientos positivos y negativos, temas y palabras clave.

Accuracy

En machine learning las métricas son estrategias que ayudan a medir el desempeño de los modelos, existen una gran cantidad de estas y cada una de ellas tienen sus especificaciones y dependiendo de estas se decide cuáles son las más adecuadas para valorar un algoritmo o modelo, para los algoritmos implementados en este trabajo nos enfocaremos en la métrica del accuracy.

El accuracy o también conocido como precisión, es una métrica que se puede utilizar para evaluar el desempeño de los modelos de clasificación, por lo que esta métrica sirve para medir la exactitud en la que un modelo realizó las predicciones.

Desbalanceo de datos

El desbalanceo [10] de datos es un problema en donde encontramos la desproporcionalidad de las observaciones, lo puede tener repercusiones en la eficiencia del modelo. Para este problema existen técnicas de balanceo de datos las cuales ayudan a eliminar o agregar observaciones para lograr un conjunto de datos balanceado.

Oversampling es una técnica de balanceo que se enfoca en igualar las observaciones, incrementando las observaciones de la clase minoritaria de forma aleatoria.

Capítulo 2 Extracción y limpieza de datos.

Adquisición de datos

Para la adquisición de datos, se utilizó la información que comparte el INEGI por medio del dataset abierto *Accidentes de tránsito terrestre en zonas urbanas y suburbanas* [11]. Este contiene información del año 2021 sobre los accidentes automovilísticos a nivel nacional, entidad federativa y municipio.

The screenshot shows the INEGI website interface. The main content area is titled 'Accidentes de Tránsito Terrestre en Zonas Urbanas y Suburbanas'. It features a search bar, a 'Ver más' button, and a 'Temas relacionados' dropdown. Below this, there are tabs for 'Documentación', 'Tabulados', 'Microdatos', 'Datos abiertos', 'Publicaciones', and 'Herramientas'. The 'Datos abiertos' tab is active, displaying a search bar and a list of data files. The list includes a summary row and four specific files: 'Accidentes de tránsito terrestre en zonas urbanas y suburbanas' (1997 - 2021, 116 MB CSV), 'Base municipal de Accidentes de Tránsito Georreferenciados' (2021, 20.1 MB SHP), 'Base municipal de Accidentes de Tránsito Georreferenciados' (2020, 14.5 MB SHP), and 'Base municipal de Accidentes de Tránsito Georreferenciados' (2019, 17.3 MB SHP).

Ilustración 1. adquisición de datos [11]

El dataset está estructurado por un total 356315 registros, conformado con 42 columnas que a continuación se detallan las características de cada una.

Columna	Descripción
COBERTURA	Área geográfica a la que están referidos los indicadores estadísticos

ID_ENTIDAD	Clave de la entidad federativa según el Catálogo de Entidades, Municipios y Localidades del INEGI
ID_MUNICIPIO	Clave del municipio según el Catálogo de Entidades, Municipios y Localidades del INEGI
ANIO	Los cuatro dígitos correspondientes al año en que ocurrió el accidente
MES	Correspondiente al mes de referencia en que ocurrió el accidente
ID_HORA	La hora (sin los minutos) en que ocurrió el accidente, con rango: 00-23 horas. Clave 99 Hora no especificada
ID_MINUTO	Los minutos en que ocurrió el accidente, con rango: 00-59. Clave 99 Minutos no especificados
ID_DIA	Número correspondiente al día del mes en que ocurrió el accidente, con rango: 01 a 28, 30 ó 31 según corresponda al mes de referencia. Clave 32 Día no especificado.
DIASEMANA	El día de la semana en que ocurrió el accidente
URBANA	Es el área habitada o urbanizada que, partiendo de un núcleo central, presenta continuidad física en todas direcciones hasta ser interrumpida, en forma notoria, por terrenos de uso no urbano como bosques, sembradíos o cuerpos de agua. Se caracteriza por presentar asentamientos humanos concentrados de más de 15,000 habitantes. En estas áreas, se asienta la administración pública, el comercio organizado y la industria. Cuenta con infraestructura, equipamiento y servicios urbanos, tales como drenaje, energía eléctrica, red de agua potable, escuelas, hospitales, áreas verdes y de diversión, etc.
SUBURBANA	Son aquellas zonas donde la población es de 2,500 a 14,999 habitantes, las viviendas se encuentran dispersas y en algunas ocasiones carecen de algunos servicios.
TIPACCID	<p>Corresponda al tipo de accidente de tránsito, de acuerdo con las siguientes descripciones: 1) Colisión con vehículo automotor: Encuentro violento, accidental o imprevisto de dos o más vehículos en una vía de circulación, del cual resultan averías, daños, pérdida parcial o total de vehículos o propiedades, así como lesiones leves y/o fatales a personas. Puede ser lateral, frontal o por alcance. 2) Colisión con peatón: Evento vial donde un vehículo de motor arrolla o golpea a una persona que transita o que se encuentra en alguna vía pública, provocando lesiones leves o fatales. 3) Colisión con animal: Es aquel accidente en el que un vehículo de motor arrolla a cualquier tipo de animal provocando daños materiales, inclusive lesiones leves o fatales a personas ocupantes o no del vehículo. 4) Colisión con objeto fijo: Encuentro violento de un vehículo de motor con cualquier tipo de objeto, que por sus características se encuentre sujeto al piso o asentado en él, tales como postes, guarniciones, señales de tránsito, árboles, contenedores de basura, etc. También se incluye en este tipo de colisión, el percance de un automotor en movimiento contra otro estacionado. 5) Volcadura: Es el tipo de accidente que debido a las circunstancias que lo originan, provocan que el vehículo pierda su posición normal, incluso dé una o varias volteretas. 6) Caída de pasajero Accidente: donde una o más personas que viajan en el vehículo, (excluyendo al conductor), caen fuera del mismo. No se considera este tipo de accidente si la caída fue por consecuencia de otro tipo de accidente. 7) Salida del camino: Evento en donde el vehículo, por causas circunstanciales, abandona de manera violenta e imprevista la vía de circulación por la cual transita. Incluso si por la acción del vehículo cae a una zanja, cuneta, barranca, etc. 8) Incendio: Es el accidente ocasionado por un corto circuito, derrame de combustible o cuestiones desconocidas, que propician la generación de fuego mediante el cual se consume parcial o totalmente el vehículo automotor. Nota: No se clasifique este accidente en este tipo, si el incendio es resultado de una colisión con otro vehículo automotor en circulación, o si el fuego se produce después de una colisión, volcadura o salida del camino. 9) Colisión con ferrocarril: Choque de un vehículo automotor con una locomotora, vagón, góndola o cualquier otro vehículo clasificado como transporte ferroviario. 10) Colisión con motocicleta: Percance vial en donde un vehículo automotor de cualquier tipo tiene un encuentro</p>

	violento, accidental o imprevisto con una motocicleta. Incluso se puede dar el caso de que sea entre dos motocicletas. 11) Colisión con ciclista: Hecho en el cual un vehículo automotor de cualquier tipo arrolla a un ciclista sobre la vía de circulación o en un cruce vial. 12) Otro: Cualquier otro tipo de accidente que no pueda ser clasificado en los 11 incisos descritos anteriormente, tales como derrumbes, deslaves o cualquier otro objeto que caiga sobre los vehículos en circulación y como consecuencia se produzca algún accidente vial.
AUTOMOVIL	Automóvil. Comprende los vehículos de motor destinados principalmente al transporte de personas, que cuentan hasta con 7 asientos (incluyendo el del conductor).
CAMPASAJ	Camioneta de pasajeros. comprende todos los vehículos de motor destinados primordialmente al transporte de personas y que tengan de 8 a 15 asientos (incluyendo el del conductor).
MICROBUS	Microbús. Autobús de menor tamaño usado por lo general en el transporte urbano de pasajeros y que tenga de 16 a 20 asientos (incluyendo el del conductor).
PASCAMION	Camión urbano de pasajeros. Comprende los autobuses urbanos y suburbanos y en general los vehículos que tengan de 21 a 29 asientos, destinados al transporte público y privado de personas, los cuales cuentan con rutas fijas.
OMNIBUS	Ómnibus. Comprende los vehículos automotores con 30 o más asientos, destinados al transporte público y privado de personas, con destinos establecidos, así como horarios de llegada y salida.
TRANVIA	Tren eléctrico o trolebús. Comprende los vehículos de motor destinados al transporte de personas, propulsados por energía eléctrica captada de cables aéreos, que no circulan sobre rieles (este tipo de transporte sólo se encuentra registrado en el Ciudad de México y Guadalajara).
CAMIONETA	Camioneta de carga. Son aquellas que están destinadas exclusivamente al transporte de carga; se identifican de acuerdo al tamaño y a la capacidad de hasta 999 kilogramos.
CAMION	Camión de carga. Comprende los vehículos de propulsión mecánica propia, destinados exclusiva o principalmente al transporte de carga, con capacidad de 1,000 hasta 5,000 kilogramos.
TRACTOR	Tractor con o sin remolque. Comprende los vehículos de propulsión mecánica propia, diseñados exclusiva o principalmente para remolcar otros vehículos (excluye los tractores agrícolas, industriales y de construcción).
FERROCARRI	Ferrocarril. Medio de transporte sobre rieles para el transporte de pasajeros y carga, que recorre distancias relativamente largas a velocidades medias.
MOTOCICLET	Motocicleta. Vehículo automotor de dos, tres o cuatro ruedas, cuyo peso no excede los 400 kilogramos.
BICICLETA	Bicicleta. Vehículo de dos o tres ruedas generalmente iguales, movidas por pedales y una cadena, el cual es propulsado por el esfuerzo humano.
OTROVEHIC	Otro. Considérese cualquier otro tipo de vehículo no descrito en la clasificación anterior; por ejemplo, las ambulancias, grúas, vehículos de tracción animal o humana, carro de bomberos, tractores agrícolas, industriales y de construcción, etc.
CAUSAACCI	La causa presunta o determinante puede considerarse como: El motivo principal que causó el accidente, ya sea por condiciones inseguras o actos irresponsables potencialmente prevenibles, atribuidos a conductores de vehículos, así como a peatones o pasajeros, falla de vehículos, condiciones del camino, circunstancias climatológicas, etc.
CAPAROD	Superficie de rodamiento en donde ocurrió el accidente de tránsito. Pavimentada Conjunto de capas de material rígido (concreto hidráulico) o flexible (carpeta asfáltica) compactado sobre el subsuelo, que permite el tránsito adecuado de vehículos y su carga. No pavimentada. Camino acondicionado con materiales naturales (piedra, bola, tezontle, etc.), para el tránsito de vehículos y/o personas.
SEXO	Genero del conductor presunto responsable de ocasionar el accidente.
ALIENTO	Sobriedad del conductor presunto responsable del accidente

CINTURON	El Conductor presunto responsable del accidente usaba el Cinturón de seguridad
ID_EDAD	La edad del conductor presunto responsable, la cual debe estar anotada con un número arábigo de 12 a 98. La clave 0 se refiere a los registros en donde el conductor se fugó y los registros con clave 99 se refiere Se ignora la edad del conductor
CONDMUERTO	Número de conductores del automóvil, camioneta de pasajeros, microbús, camión urbano de pasajeros, ómnibus, tren eléctrico o trolebús, camioneta de carga, camión de carga, tractor con o sin remolque, ferrocarril o motocicleta involucrado, que a consecuencia del evento muere en el lugar del accidente.
CONDHERIDO	Número de conductores heridos del automóvil, camioneta de pasajeros, microbús, camión urbano de pasajeros, ómnibus, tren eléctrico o trolebús, camioneta de carga, camión de carga, tractor con o sin remolque, ferrocarril o motocicleta involucrados, que resulta heridos a consecuencia del accidente.
PASAMUERTO	Número de personas que son transportadas en algún vehículo de motor, sin considerar al conductor, que a consecuencia del evento muere en el lugar del accidente.
PASAHERIDO	Número de personas que son transportadas en algún vehículo de motor, sin considerar al conductor, que resulta heridos a consecuencia del accidente.
PEATMUERTO	Número de personas que transita por sus propios medios de locomoción por alguna calle, avenida, boulevard, glorieta, etc., que a consecuencia del evento muere en el lugar del accidente.
PEATHERIDO	Número de personas que transita por sus propios medios de locomoción por alguna calle, avenida, boulevard, glorieta, etc. que resulta heridos a consecuencia del accidente.
CICLMUERTO	Número de personas que va operando o circulando en alguna bicicleta, triciclo, etc., que a consecuencia del evento muere en el lugar del accidente.
CICLHERIDO	Número de personas que va operando o circulando en alguna bicicleta, triciclo, etc. que resulta heridos a consecuencia del accidente.
OTROMUERTO	Número de personas que indirectamente estuvieron involucrados en el accidente y que por su naturaleza no pueden ser incluidas en la clasificación señalada anteriormente, tales como personas que se encuentran en el interior de casas, negocios, comercios, así como todos aquellos individuos que por cuestiones laborales realizan actividades de mantenimiento, limpieza y/o construcción de las vías de comunicación, que a consecuencia del evento muere en el lugar del accidente.
OTROHERIDO	Número de personas que indirectamente estuvieron involucrados en el accidente y que por su naturaleza no pueden ser incluidas en la clasificación señalada anteriormente, tales como personas que se encuentran en el interior de casas, negocios, comercios, así como todos aquellos individuos que por cuestiones laborales realizan actividades de mantenimiento, limpieza y/o construcción de las vías de comunicación, que resulta heridos a consecuencia del accidente.
NEMUERTO	Número de víctimas muertas en las cuales la fuente informante no realizó la clasificación, que a consecuencia del evento muere en el lugar del accidente.
NEHERIDO	Corresponde a las víctimas heridas en las cuales la fuente informante no realizó la clasificación, que resulta heridos a consecuencia del accidente.
CLASACC	Los accidentes se clasifican en Fatales: Se refiere a todo accidente de tránsito en el cual una o más personas fallecen en el lugar del evento; No fatales: Se refiere a todo accidente de tránsito en el cual una o más personas resultan con lesiones con o sin consecuencia de muerte y Sólo daños: Se refiere a todo accidente en el que se ocasionaron daños materiales a vehículos automotores, propiedad del estado, inmueble particular y otros.

Tabla 1. Estructura dataset

Limpieza de datos

La limpieza de datos o también conocido como data cleasing, es uno de los procesos que define el éxito de un modelo de machine learning, en este se busca transformar, eliminar datos incorrectos o incompletos, seleccionar aquellos datos que tienen mayor relevancia para el modelo. El proceso de limpieza propuesto es el siguiente.



Ilustración 2. Proceso Limpieza de datos

Validación de valores nulos

En este primer punto se busca validar aquellos valores nulos que se encuentren en el dataset y en caso de encontrar valores nulos, proceder con ellos de manera que se puedan transformar

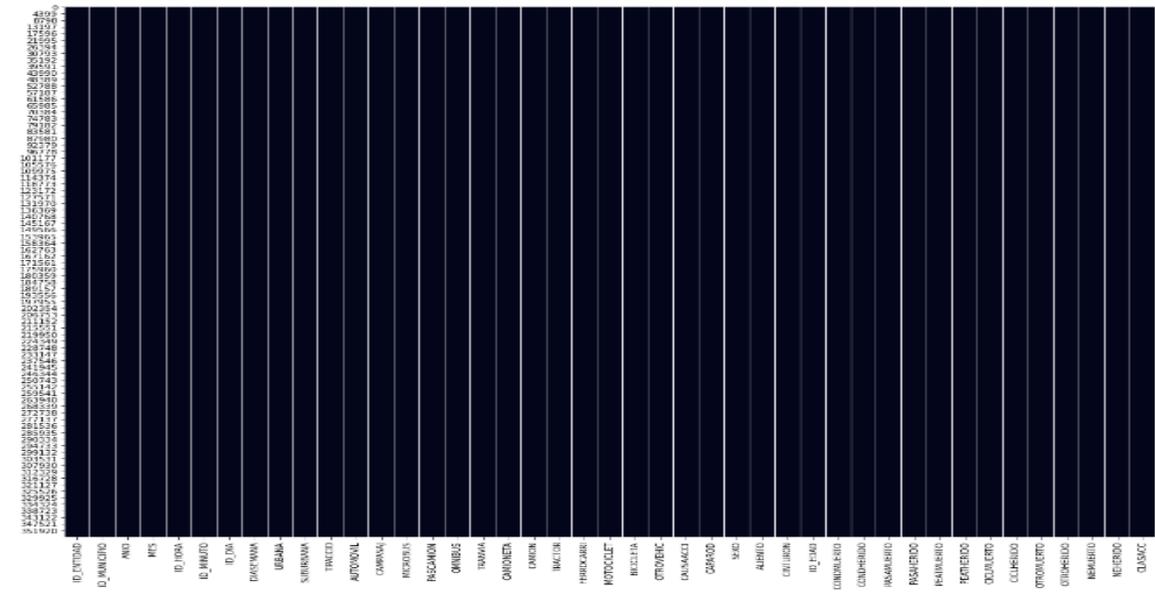


Ilustración 3. Visualización valores nulos

Se puede apreciar que no contamos con valores nulos, pero revisaremos a detalle.

```
In [113]: 1 #Ahora los obtenemos a nivel de suma
          2 df_accidentes.isnull().sum()

Out[113]: ID_ENTIDAD      0
          ID_MUNICIPIO    0
          ANIO            0
          MES             0
          ID_HORA         0
          ID_MINUTO       0
          ID_DIA          0
          DIASEMANA       0
          URBANA          0
          SUBURBANA       0
          TIPACCID        0
          AUTOMOVIL       0
          CAMPASAJ        0
          MICROBUS        0
          PASCAMION       0
          OMNIBUS         0
          TRANVIA         0
          CAMIONETA       0
          CAMION          0
          TRACTOR         0
          FERROCARRI      0
          MOTOCICLET      0
          BICICLETA       0
          OTROVEHIC       0
          CAUSAACCI       0
          CAPAROD         0
          SEXO            0
          ALIENTO        0
          CINTURON       0
          ID_EDAD         0
          CONDMUERTO     0
          CONDHERIDO     0
          PASAMUERTO     0
          PASAHERIDO     0
          PEATMUERTO     0
          PEATHERIDO     0
          CICLMUERTO     0
          CICLHERIDO     0
          OTROMUERTO     0
          OTROHERIDO     0
          NEMUERTO       0
          NEHERIDO       0
          CLASACC        0
          dtype: int64
```

Ilustración 4. Totales Valores Nulos

Se confirma que, para este caso, no se cuenta con valores nulos en el dataset, por lo que se puede continuar con el siguiente paso.

Conversión de valores categóricos

Los modelos trabajan con variables numéricas, por lo que se tiene que transformar aquellas variables que sean categóricas. Para esto primero identificamos las variables tipo object.

```

1 #Identificamos las variables categoricas
2 for i in df_accidentes.select_dtypes(include='object').columns:
3     print(i)

```

DIASEMANA
 URBANA
 SUBURBANA
 TIPACCID
 CAUSAACCI
 CAPAROD
 SEXO
 ALIENTO
 CINTURON
 CLASACC

Ilustración 5. Columnas Categóricas

Se encontraron únicamente un total de 10 variables categóricas, las cuales se pueden apreciar a detalle.

	DIASEMANA	URBANA	SUBURBANA	TIPACCID	CAUSAACCI	CAPAROD	SEXO	ALIENTO	CINTURON	CLASACC
0	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con vehículo automotor	Conductor	Pavimentada	Hombre	Sí	Se ignora	Sólo daños
1	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con vehículo automotor	Conductor	Pavimentada	Hombre	No	Se ignora	Sólo daños
2	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con peatón (atropellamiento)	Peatón o pasajero	Pavimentada	Hombre	No	Se ignora	Fatal
3	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con objeto fijo	Conductor	Pavimentada	Hombre	No	No	No fatal
4	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con peatón (atropellamiento)	Peatón o pasajero	Pavimentada	Se fugó	Se ignora	Se ignora	No fatal
5	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con objeto fijo	Conductor	Pavimentada	Hombre	Sí	Se ignora	Sólo daños
6	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con objeto fijo	Conductor	Pavimentada	Hombre	No	No	Fatal
7	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con vehículo automotor	Conductor	Pavimentada	Se fugó	Se ignora	Se ignora	No fatal
8	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con vehículo automotor	Conductor	Pavimentada	Hombre	No	Se ignora	Sólo daños
9	Viernes	Accidente en intersección	Sin accidente en esta zona	Colisión con peatón (atropellamiento)	Conductor	Pavimentada	Hombre	No	Se ignora	No fatal

Ilustración 6. Detalle Variables Categóricas Antes de Convertir

Al convertir estas variables la información queda de la siguiente manera:

	DIASEMANA	URBANA	SUBURBANA	TIPACCID	CAUSAACCI	CAPAROD	SEXO	ALIENTO	CINTURON	CLASACC
0	5	0	3	8	1	2	1	3	2	3
1	5	0	3	8	1	2	1	1	2	3
2	5	0	3	7	5	2	1	1	2	1
3	5	0	3	6	1	2	1	1	1	2
4	5	0	3	7	5	2	3	2	2	2
5	5	0	3	6	1	2	1	3	2	3
6	5	0	3	6	1	2	1	1	1	1
7	5	0	3	8	1	2	3	2	2	2
8	5	0	3	8	1	2	1	1	2	3
9	5	0	3	7	1	2	1	1	2	2

Ilustración 7. Detalle Variables Categóricas Después de Convertir

Como parte del desarrollo de trabajo de esta tesis se desea predecir la gravedad de un accidente automovilístico, así que la clasificación de la gravedad de un accidente queda de la siguiente forma.

Clasificación de accidente automovilísticos.

- Certificado cero -> 0
- Fatal -> 1
- No fatal -> 2
- Sólo daños -> 3

Esto es un punto importante ya que posteriormente solo se mencionarán los valores asignados a la clase del tipo de accidente.

Balanceo de datos

Para poder obtener los mejores resultados de los modelos se deben balancear los datos esto, con el fin de tener una distribución equilibrada de la información.

Por lo cual, se procede a revisar la distribución de la variable a predecir, en este caso la gravedad del accidente.

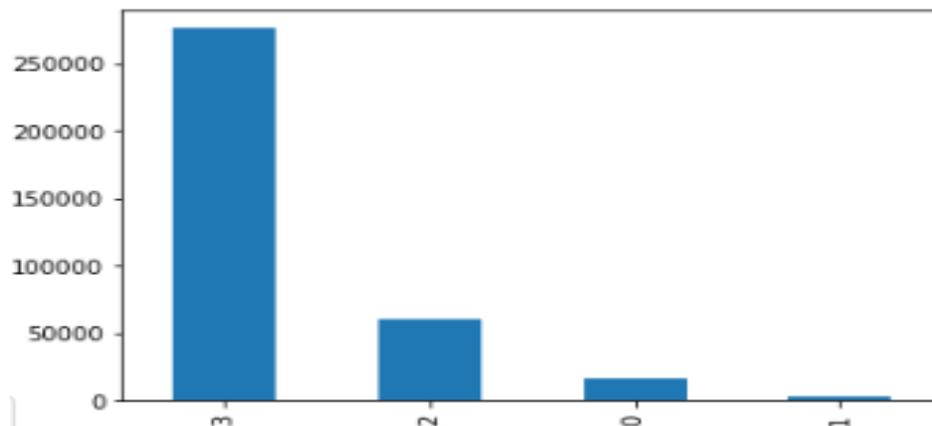


Ilustración 8. Desbalance de datos

El análisis realizado sobre la información nos arroja un desbalanceo de los datos en donde se puede apreciar que la clase 3 cuenta con el mayor número de instancias, mientras que la clase uno es la que menos instancias tiene. El detalle de cada lo clase se puede observar a continuación.

Clase	Instancias
3	275982
2	60584
0	15900
1	3849

Tabla 2. Total, Instancias Desbalanceadas

Con el fin de tener balanceada la información, se aplica la técnica de balanceo Oversampling, esta ayuda a igualar las instancias incrementando las observaciones de la clase minoritaria.

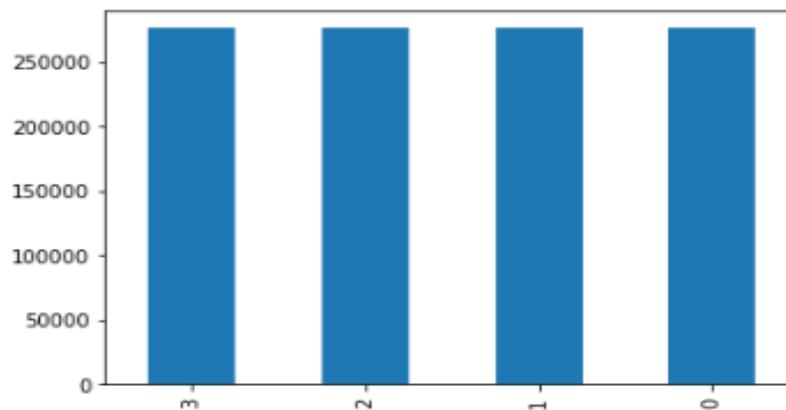


Ilustración 9. Datos balanceados

Se aprecia una distribución uniforme por lo que la información se encuentra ya balanceada, en cada clase se identifican las instancias como se muestra a continuación.

Clase	Instancias
3	275982
2	275982
0	275982
1	275982

Tabla 3. Total, Instancias Balanceadas

Como Oversampling funciona aumentando la clase minoritaria, se puede notar que las clases se quedan con el número de instancias que había en la clase 3, que era la que tenía la mayor cantidad de instancias.

Selección de Características.

La selección de características es el proceso en donde se eligen las variables que tienen más relevancia en la predicción de un modelo de machine learning, por lo cual al aplicar este proceso lo que estamos obteniendo es un subconjunto de la información original [11], además de que buscamos mejorar la precisión del modelo y reducir la complejidad computacional del mismo.

Entre los métodos de selección de características se encuentran los métodos de envoltura en donde su principal propiedad es el uso de un algoritmo de machine learning [12], en donde el rendimiento de este es el criterio principal para elegir las mejores características.

Un ejemplo de estos métodos es, la selección hacia adelante, en donde su característica principal es que es un proceso iterativo, en este se inicia sin ninguna característica y en cada iteración de agregan nuevas características, esto ocurre hasta que el agregar una nueva característica no mejore el rendimiento del modelo [12].

Para el proceso de selección de características se utilizó la selección hacia adelante, también conocida como Forward Selection, y se utilizó el modelo de RandomForestClassifier [13] como algoritmo para calcular las mejores variables, de esto se obtuvieron los siguientes resultados

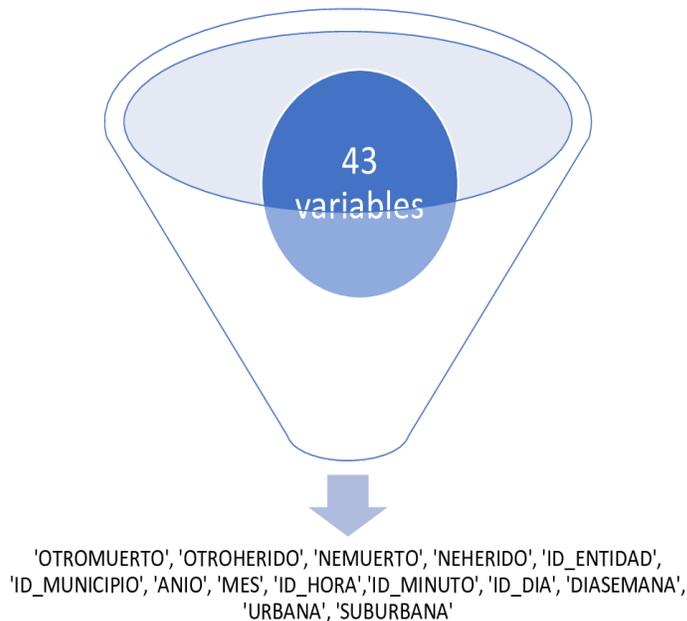


Ilustración 10. Variables Relevantes

Se puede apreciar que al aplicar la selección de características se eliminaron 29 variables las cuales resultaron ser irrelevantes, redundantes o incluso perjudiciales para la precisión de los modelos.

División de datos

Para la implementación de los modelos, se requiere contar con un conjunto de entrenamiento y prueba, tanto para las variables predictoras, como para la variable a predecir. Para esto se crearon los siguientes subconjuntos.

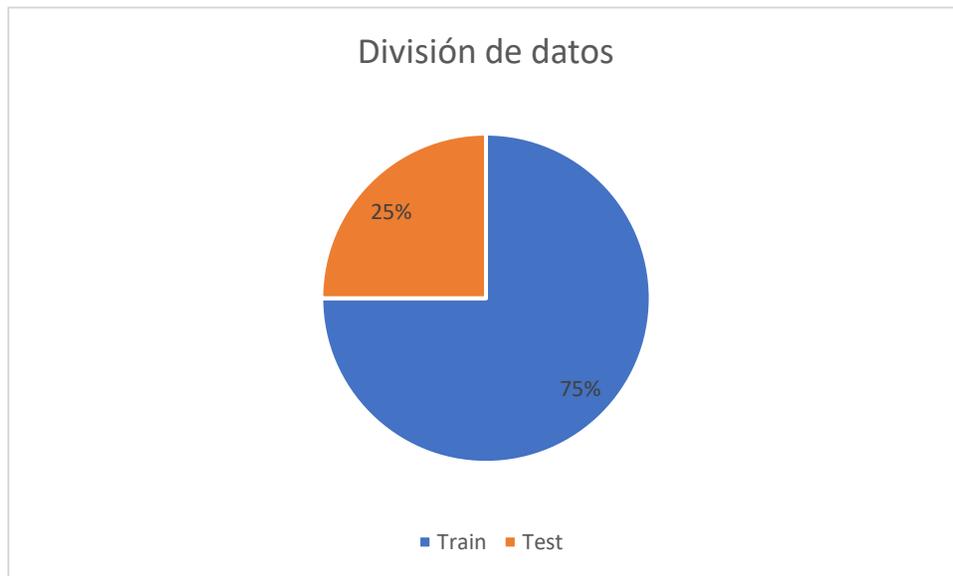


Ilustración 11. División de datos

Dando como resultado los siguientes datasets:

TRAIN	TEST
827946	275982

Tabla 4. Total, instancia Train-Test

Capítulo 3 Modelos a implementar

Arboles de Decisión

Los árboles de decisión son un modelo que utiliza estructuras de árbol para representar decisiones y sus consecuencias. Los árboles de decisión son una forma común de árboles binarios utilizados en el aprendizaje automático.

En un árbol de decisión, cada nodo interno representa una pregunta o decisión, y cada rama representa una posible respuesta a esa pregunta. Las hojas del árbol representan las predicciones o decisiones finales del modelo.

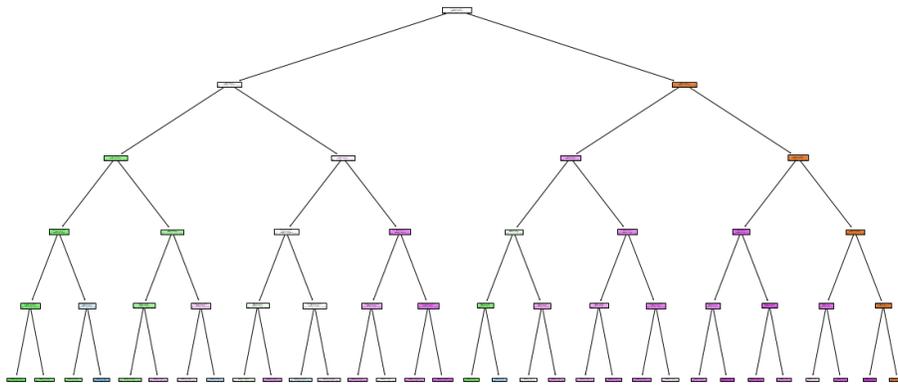


Ilustración 12. Árbol de decisiones

KNN

K-Nearest Neighbors (KNN) es un modelo que se basa en la idea de que los puntos de datos similares están cerca unos de otros. En este modelo toma en cuenta los k vecinos más cercanos al punto de datos de entrada que se está clasificando o prediciendo. La elección [14] del valor de k es un parámetro importante en el algoritmo KNN, y puede afectar significativamente la precisión de las predicciones.

Si k es demasiado pequeño, el modelo puede sobre ajustarse a los datos de entrenamiento y ser sensible al ruido en los datos. Si k es demasiado grande, el modelo puede ser demasiado simplista y no capturar suficientemente bien las características importantes de los datos.

Se utilizó el método del Codo también conocido Elbow para obtener el K adecuado, mismo que será utilizado en pruebas posteriores.

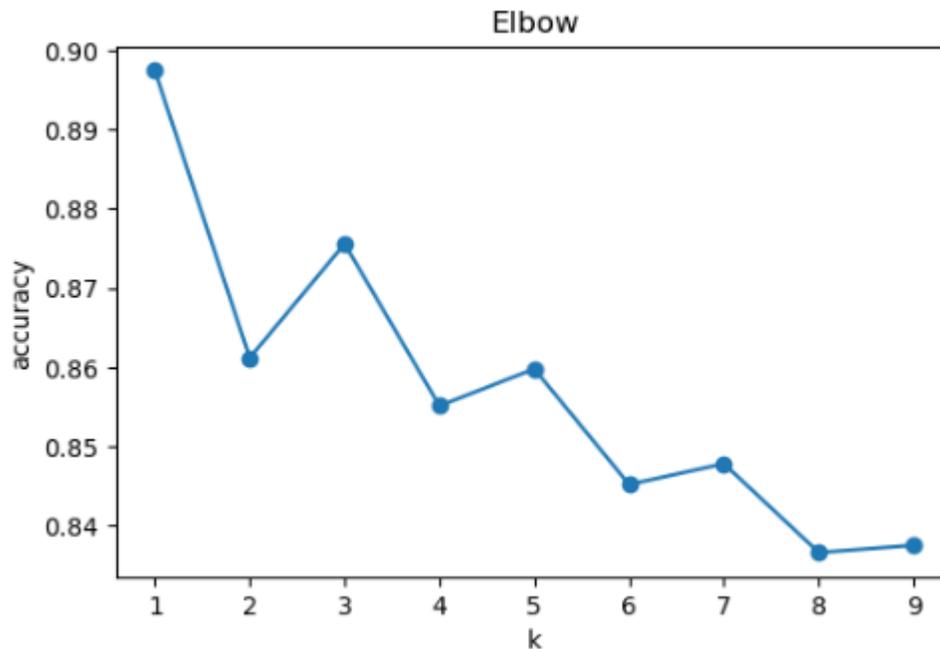


Ilustración 13. Método Elbow

Como se observa en el gráfico, con $K=3$ se obtiene un accuracy mayor a 80%, por lo que este valor será utilizado en las pruebas siguientes.

Redes Neuronales

Las redes neuronales también conocidas como redes neuronales artificiales (ANN) o simuladas (SNN) reflejan cómo se comporta el cerebro humano, imitando la forma en que las neuronas biológicas se señalan entre sí, esto permite que algoritmos reconozcan patrones y resuelvan diferentes tipos de problemas además que éstas son el corazón de los algoritmos del Deep learning.

Las ANN están compuestas por capas de nodos, que contienen una capa de entrada, una o más capas ocultas, y una capa de salida. Cada nodo, o neurona artificial, se conecta a otro y tiene un peso y un umbral asociados.

Un parámetro importante en las redes neuronales es la función de activación, esta sirve para generar una salida con base a la neurona y las variables de entrada que se tienen. Algunas de estas funciones son;

1. Función de activación lineal: esta función lineal que se le conoce también como de identidad, ya que genera una salida prácticamente igual a la entrada.
2. Función de activación sigmoidea: es una función no lineal también es conocida como función de logística y tiene una peculiaridad ya que su grafica tiene una forma de "S". Su principal funcionamiento es que generar los valores de entrada entre 0 y 1, por ejemplo, si se evalúan valores negativos la función será igual a cero, si se evalúa en cero la función dará 0.5 y en valores altos su valor es aproximadamente a 1.
3. Función de activación ReLU: es una función no-lineal que le da valores de cero a todas las entradas negativas y las entradas positivas las deja sin cambios.
4. Función de activación tangente hiperbólica: es una función no lineal que genera valores de entrada entre -1 y 1, por ejemplo, si se evalúan valores negativos la función será igual a -1, si se evalúa en cero la función dará 0 y en valores altos su valor es aproximadamente a 1, se dice también que es un escalamiento de la función logística [15]

Random Forest

Es un modelo que consiste en múltiples árboles de decisión y los utiliza para hacer predicciones.

El funcionamiento principal de Random Forest es la combinación de múltiples árboles de decisión lo que ayuda a reducir el riesgo de un sobreajuste y claro, mejora la precisión del modelo. Es más preciso que un árbol de decisión ya que al ser entrenado con versiones diferentes tiene más diversidad.

Al final por medio de una votación se decide cual es la categoría correcta.

Capítulo 4 Resultados

En esta sección se proporcionan y analizan los resultados obtenidos tras entrenar los modelos de machine learning y aplicar una serie de pruebas variando algunos de los parámetros más importantes que se requieren para el entrenamiento de estos.

Un variable que se tomó en cuenta fue el tiempo de ejecución que tomo cada modelo en cada prueba que se realizó. Es importante resaltar que la ejecución de cada modelo se realizó bajo las mismas condiciones.

Arboles de Decisión

Se realizaron 8 pruebas variando el parámetro de *max_depth* que representa el nivel de profundidad del árbol, para esto se inició con una profundidad de 5, incrementando de 5 en 5 en cada prueba realizada hasta llegar a una profundidad de 40. Los resultados obtenidos de estas pruebas concentrándonos en el accuracy fueron los siguientes.

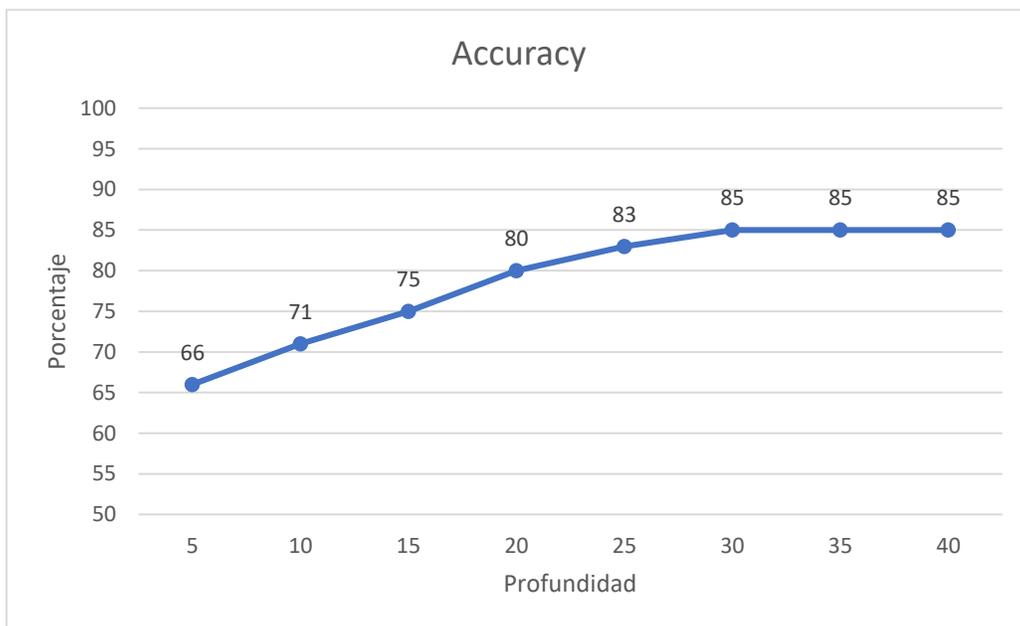


Ilustración 14. Accuracy Árbol de Decisiones

Se puede apreciar cómo es que aumentando el nivel de profundidad el accuracy aumentaba a la par. Llegando a una profundidad de 35, el accuracy converge al 85%, por lo que podríamos deducir que aumentar la profundidad del árbol no mejoraría el accuracy.

Ahora se describe el comportamiento del tiempo de ejecución de cada prueba, tomando en cuenta la profundidad del árbol.

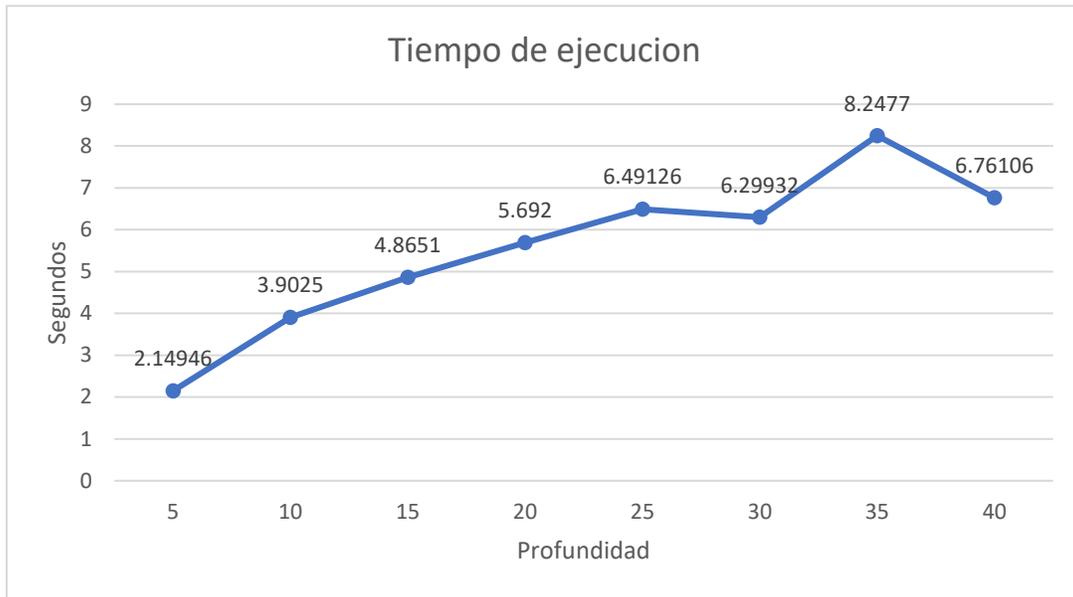


Ilustración 15. Tiempo de ejecución árbol de decisiones

Como se puede observar el tiempo de ejecución para arboles binarios es bajo ya que incluso con una profundidad de 40, no tomo más de 10 segundos lograr un entrenamiento óptimo para alcanzar un porcentaje de accuracy mayor a 80%.

Finalmente, para poder evaluar el comportamiento de este modelo se revisó la matriz de confusión, en este caso solo se muestra la matriz de confusión referentes a una profundidad de 5 y de 40.

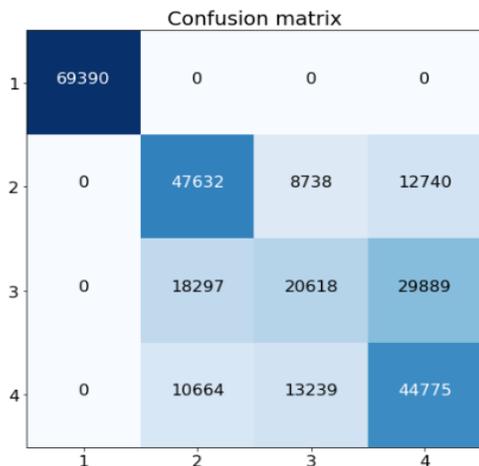


Ilustración 16. Matriz de Confusión Profundidad 5

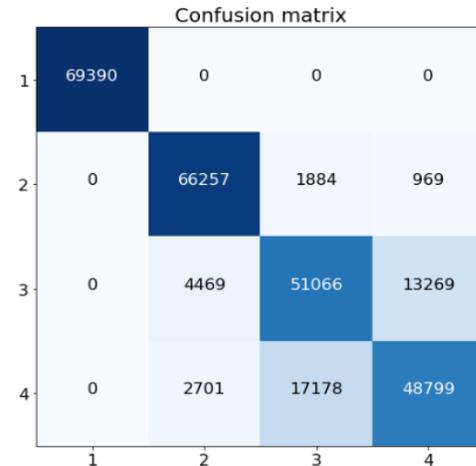


Ilustración 17. Matriz de Confusión Profundidad 40

Se puede apreciar una mejora en la predicción de las clases 2, 3 y 4 esto generado de una reducción de falsos positivos y falsos negativos y por en un aumento los verdaderos positivos.

KNN

Toma el turno del modelo de los K-vecinos más cercanos, por sus siglas en inglés KNN. Para KNN se realizaron 2 pruebas variando el parámetro de **n_neighbors** que representa el número de vecinos cercanos, se utilizaron como valores 3 y 5 vecinos cercanos, aunque previamente se calculó este valor ideal para el número de vecinos por medio del método del codo, se utilizó otro valor para poder apreciar la diferencia que existe entre K. Además de se utilizó el parámetro de distancia tipo euclidiana.

Los resultados obtenidos de estas pruebas puntualizando en el accuracy fueron los siguientes.

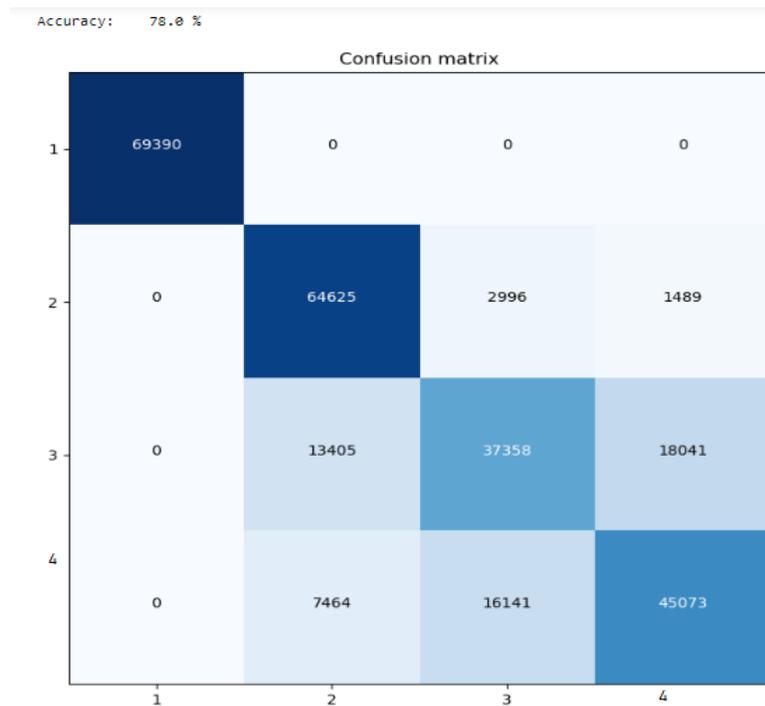


Ilustración 18. Matriz de confusión K=5

Con 5 vecinos se alcanzó un accuracy de 78%, aunque había aun un gran número de falsos positivos y falsos negativos para las clases 2, 3, y 4. Se puede apreciar que la clase 1 es la única obtuvo una totalidad de valores estimados de forma correcta por el modelo

En cambio, probando con 3 vecinos se alcanzó un accuracy de 80%, justo como se predijo con el método del codo.

Accuracy: 80.0 %

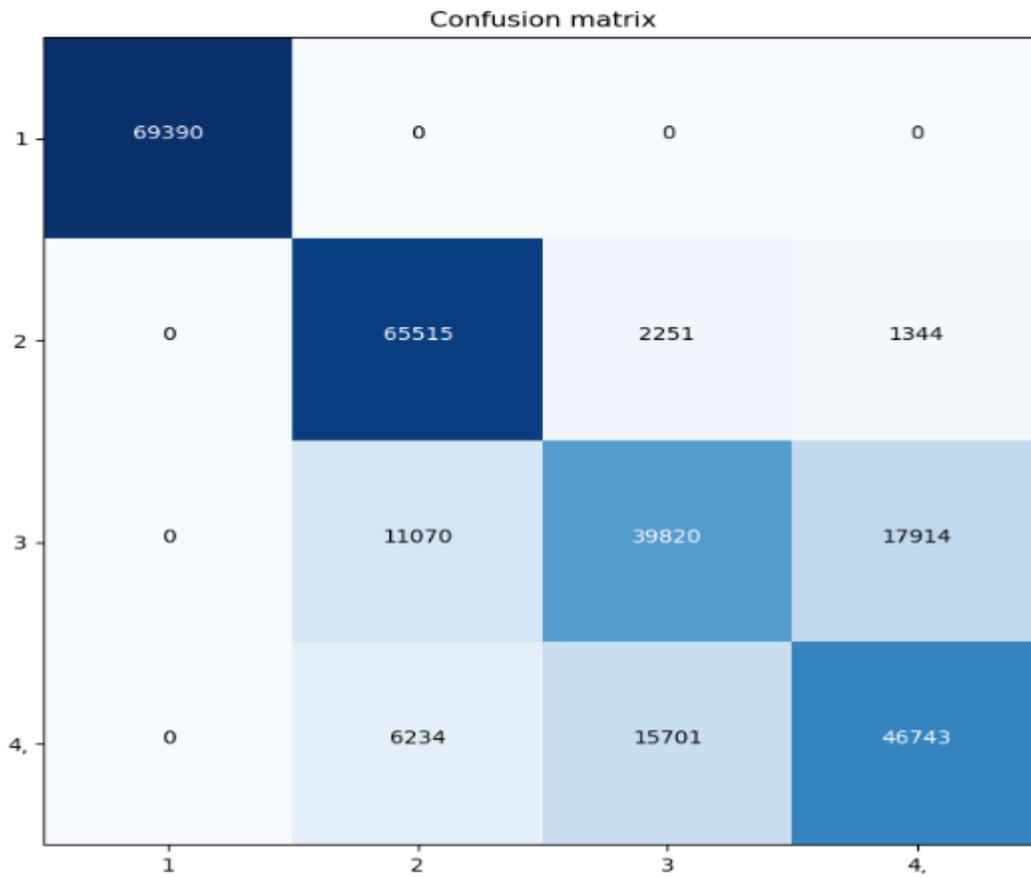


Ilustración 19. Matriz de confusión KNN K=3

Se revisa rápidamente el tiempo de ejecución de las 2 pruebas realizadas

Número de vecinos cercanos	3	5
Tiempo de ejecución (segundos)	1049.081	1068.122

Tabla 5. Tiempo de ejecución KNN

Redes Neuronales

Toca el turno de las redes neuronales, para este caso se realizaron 4 sets de pruebas por función de activación, en cada set se variaba el parámetro de **hidden_layer_sizes** que representa el número de neuronas en cada capa oculta. Otros parámetros que se tomaron en cuenta para las pruebas fueron **solver** que representa el tipo de algoritmo para calcular los pesos en la red, este configurado en **lbfgs**, **max_iter** que representa el número de iteraciones por de entrenamiento, establecido en 500 y **learning_rate_init** que establece que tan rápido pueden cambiar los parámetros a medida que aprende el modelo, en este caso establecido en 0.01.

Función logistic.

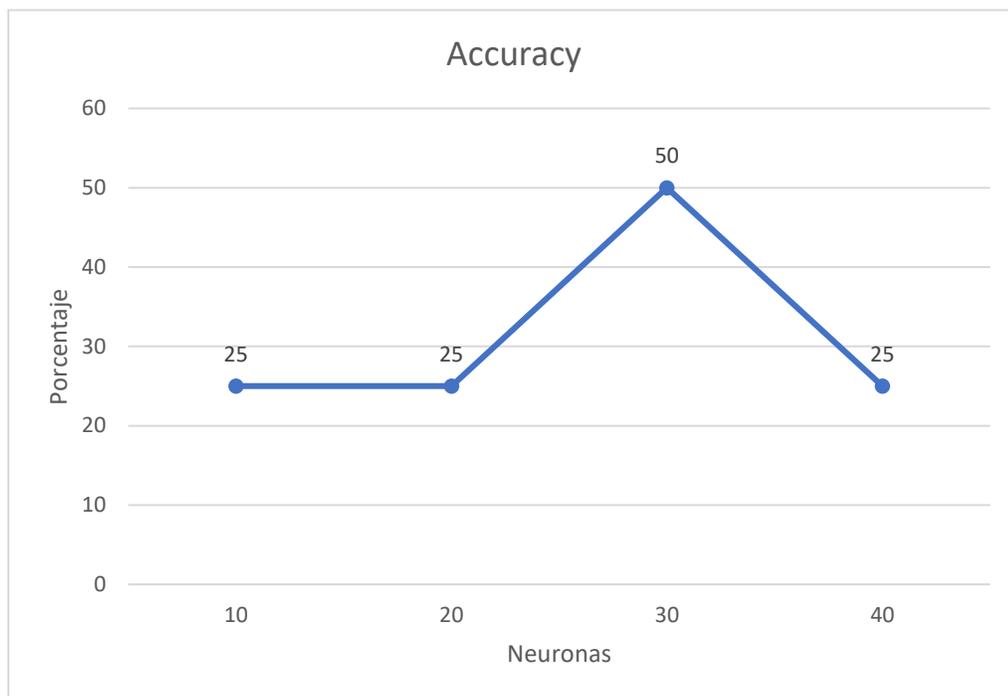


Ilustración 20. Accuracy Logistic

Como se puede observar la red neuronal no alcanza más allá de un 50% accuracy, por lo que podemos concluir que la función de activación no fue la mejor para la información que se está generando a través de la red. Una hipótesis es que esto se deba a la información que se está manejando.

Se revisa rápidamente el tiempo de ejecución

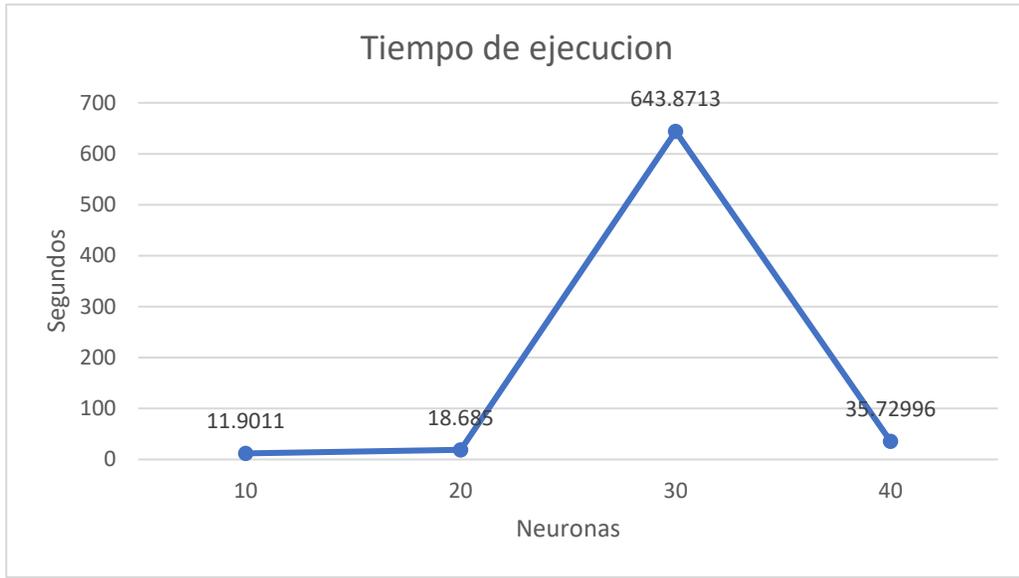


Ilustración 21. Tiempo de ejecución Logistic

Finalmente, para poder evaluar el comportamiento de este modelo se revisa la matriz de confusión, en este caso solo se muestra la matriz de confusión referentes a de 10 y de 30 neuronas, ya que en estas se encontró una diferencia significativa.

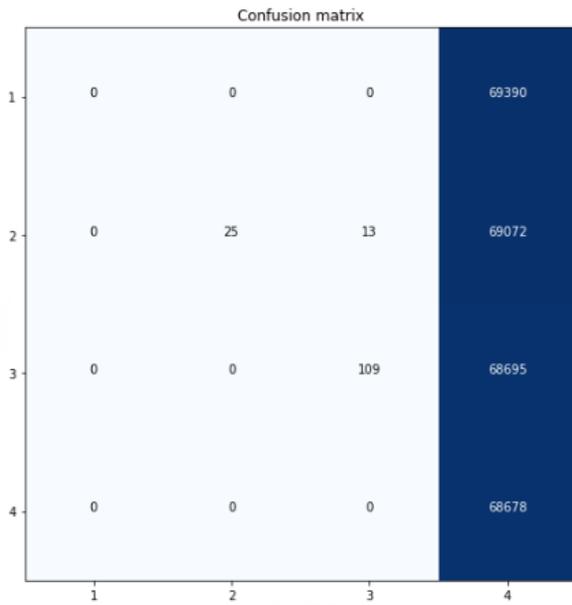


Ilustración 22. Matriz de confusión Logistic 10 neuronas

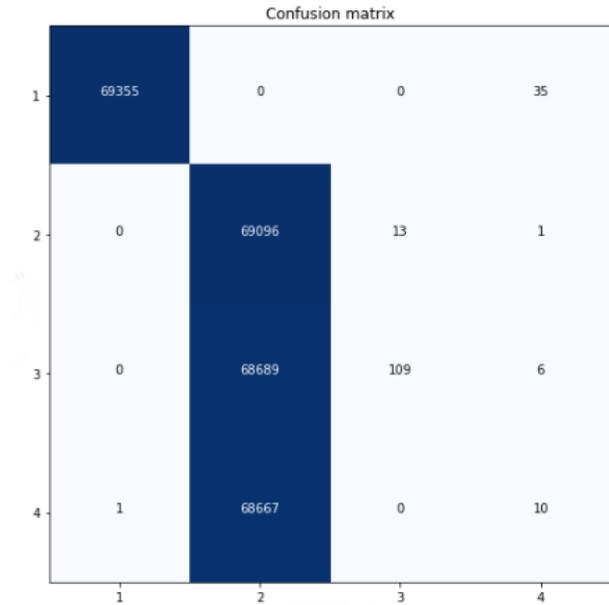


Ilustración 23. Matriz de confusión Logistic 30 neuronas

Se puede apreciar que en ambas pruebas este tiene una deficiencia debido al alto número de falsos positivos y falsos negativos lo que genera un alto número de casos mal clasificados.

Función identity

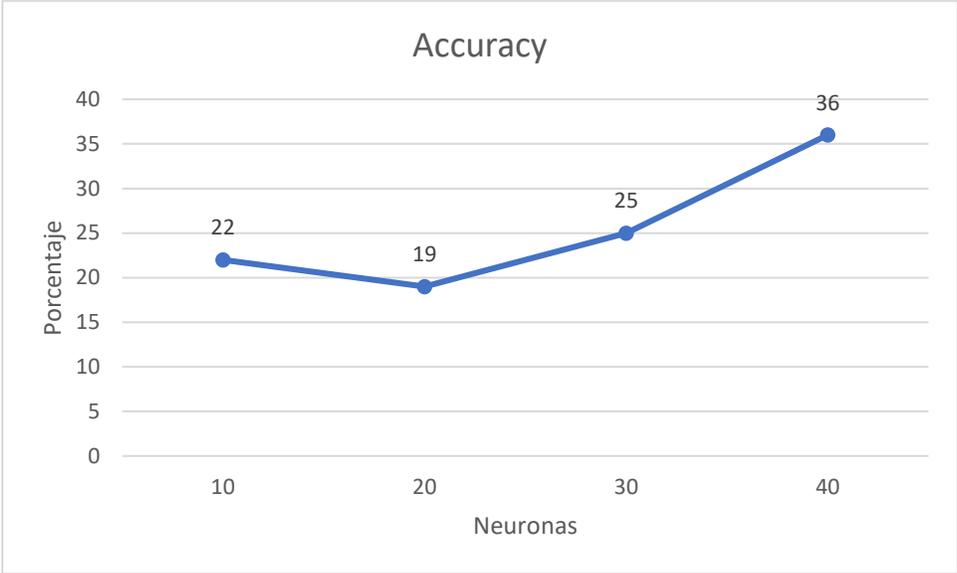


Ilustración 24. Accuracy Identity

La evidencia presentada muestra que la red puede mejorar su precisión aumentando el número de neuronas. En las pruebas realizadas se alcanzó un máximo de 36% de accuracy, si el comportamiento de la red no cambia, se podría concluir que se tiene una alta probabilidad de mejorar la accuracy.

Toca revisar el tiempo de ejecución de la red

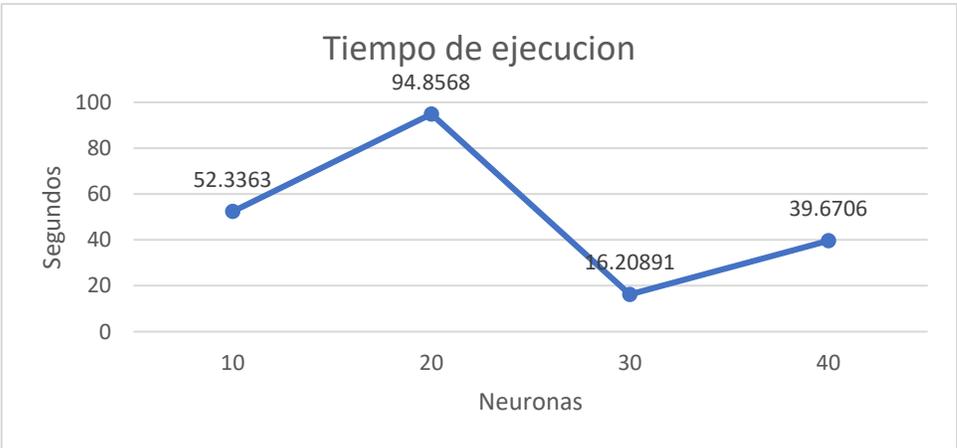


Ilustración 25. Tiempo de ejecución Identity

Se revisa como fue el comportamiento de la red con su matriz de confusión de 40 y de 30 neuronas

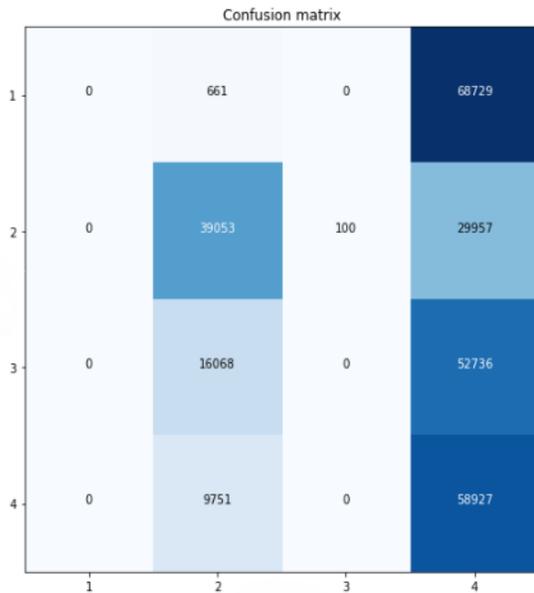


Ilustración 27. Matriz de confusión Identity 40 neuronas

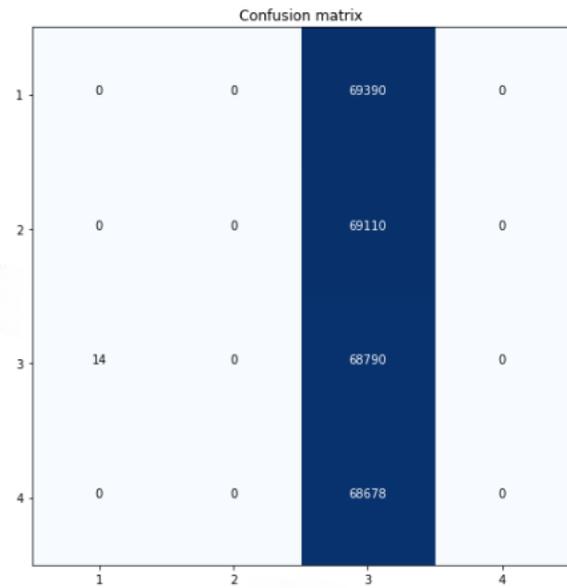


Ilustración 26. Matriz de confusión Identity 30 neuronas

Se aprecia un alto número de clases mal precedidas, lo que ya se podía deducir con el bajo accuracy obtenido en las pruebas.

Función Relu

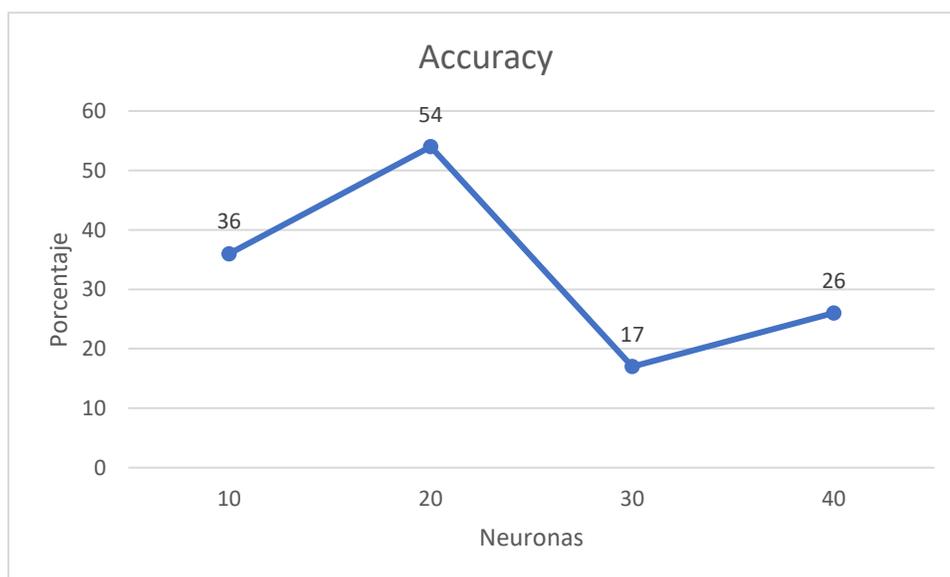


Ilustración 28. Accuracy Relu

Relu es la función con la que se pudo alcanzar un accuracy mayor al 50%, tomando en cuenta que con funciones anteriores no se logró pasar más allá del 36%, aunque se puede apreciar que el aumentar las neuronas no mejoro el accuracy, si no que bajo considerablemente, pudiendo mantener un buen accuracy con 20 neuronas.

En cuestión de tiempo de ejecución, el comportamiento de la red fue el siguiente

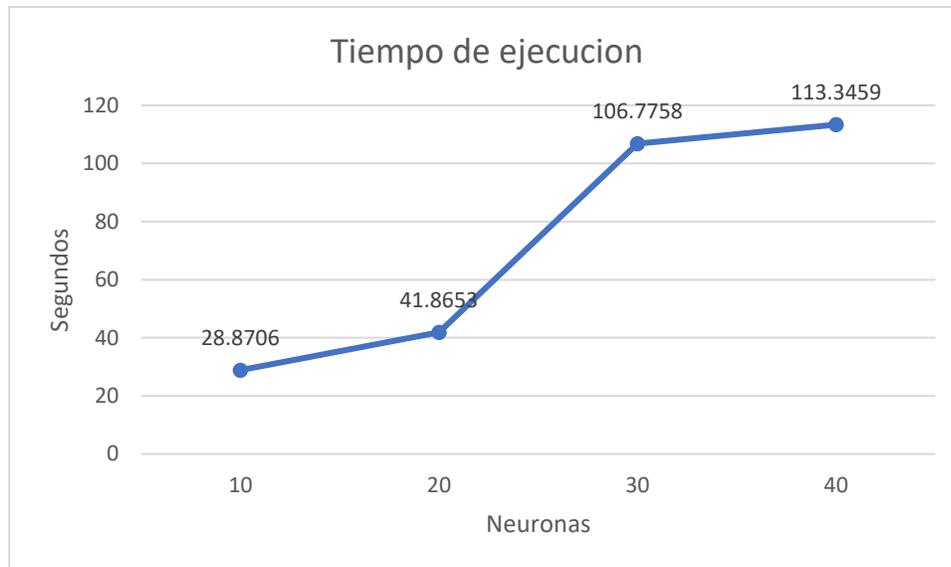


Ilustración 29. Tiempo de ejecución Relu

Finalmente se revisó la matriz de confusión de la red, en este caso con 20 y 30 neuronas.

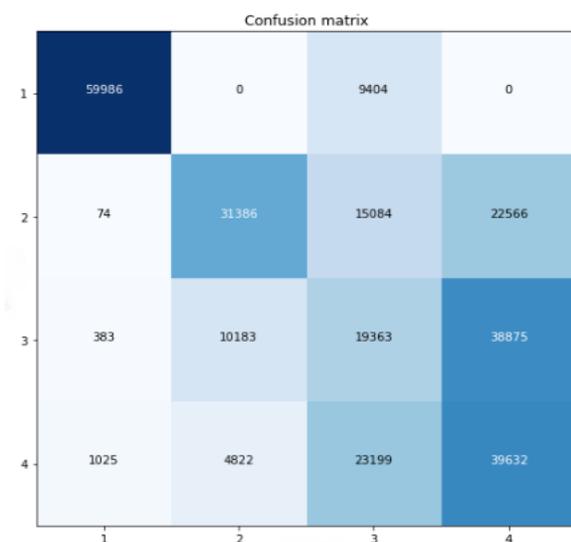


Ilustración 30. Matriz de confusión Relu 20 neuronas

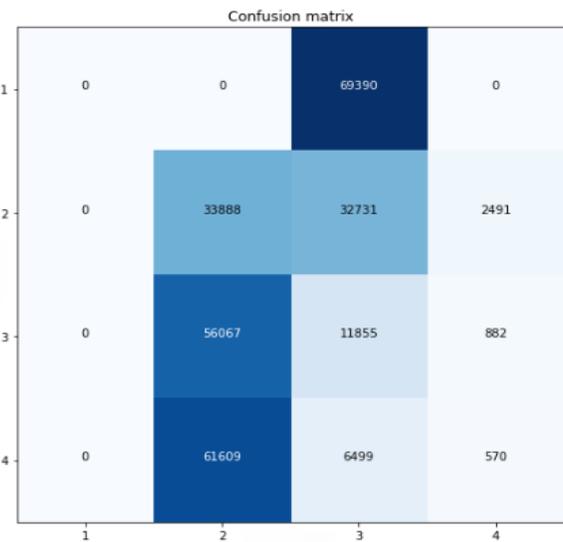


Ilustración 31. Matriz de confusión Relu 30 neuronas

Se puede apreciar que para la clase 1 con 20 neuronas, existe un alto de número de clases precedidas correctamente con respecto a la prueba de 30 neuronas, donde se observa una nula predicción. Además de que con 30 neuronas existe un alto número de falsos positivos y negativos, mientras que usando 20 neuronas se decrementaron estos y se aumentaron los verdaderos positivos.

Función Tanh

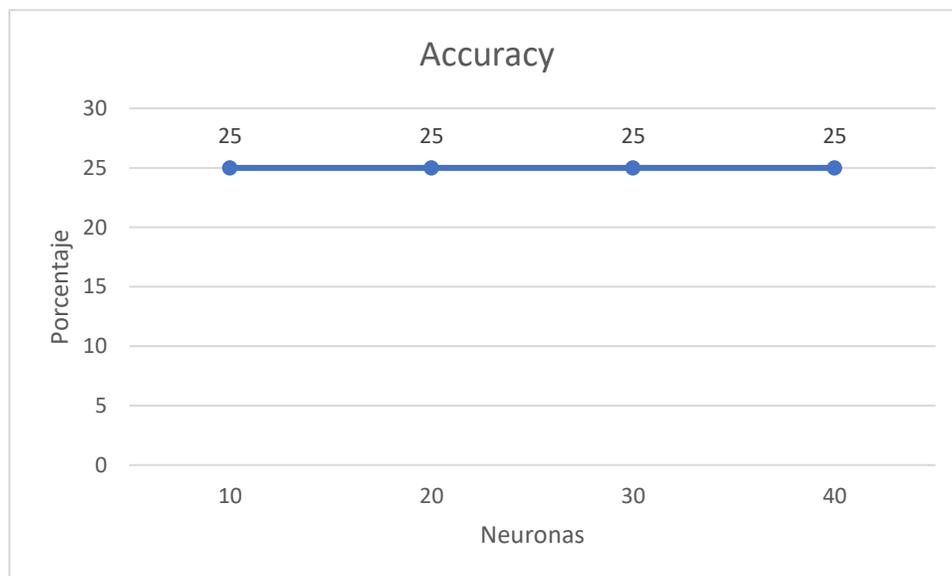


Ilustración 32. Accuracy Tanh

Como se aprecia, el comportamiento de la red con esta función de activación no existe una mejora en el accuracy aumentando el número de neuronas, el accuracy converge al 25%, por lo que se puede deducir que aumentar el número de neuronas de la red no mejoraría el accuracy.

Se revisa rápidamente el tiempo de ejecución

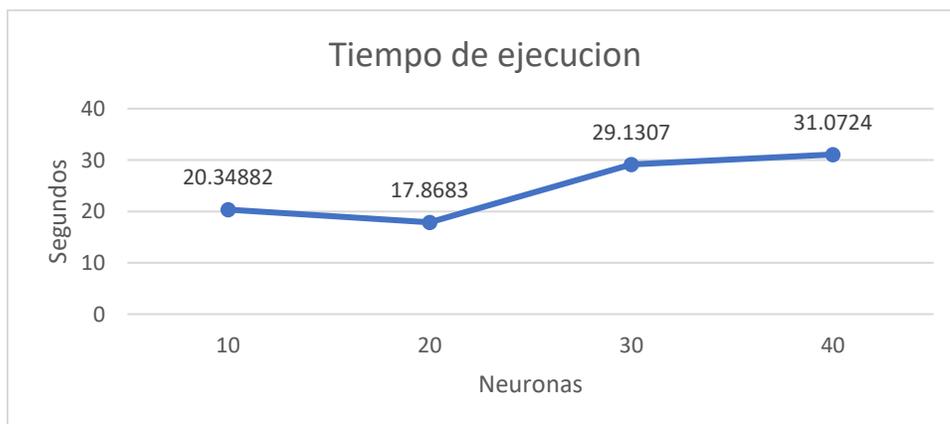


Ilustración 33. Tiempo de ejecución Tanh

Toca el turno de evaluar el comportamiento de la red, así que se revisa la matriz de confusión, en este caso solo se muestra la matriz de confusión referentes a 10 y de 40 neuronas, esto con el fin de revisar si existe algún cambio en su comportamiento.

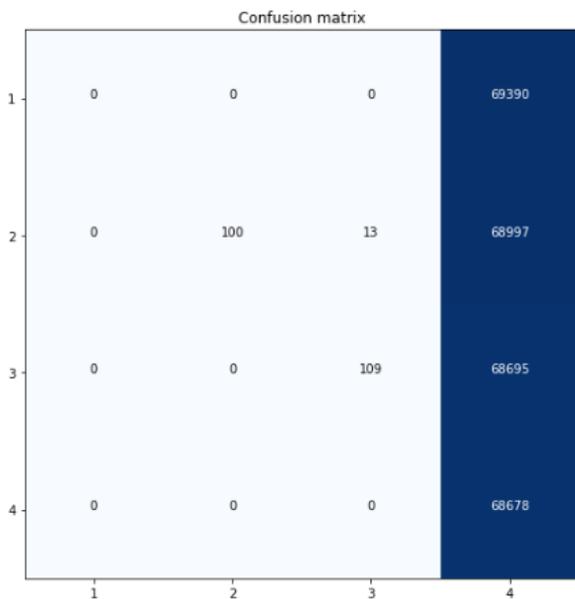


Ilustración 35. Matriz de confusión Tanh 10 neuronas

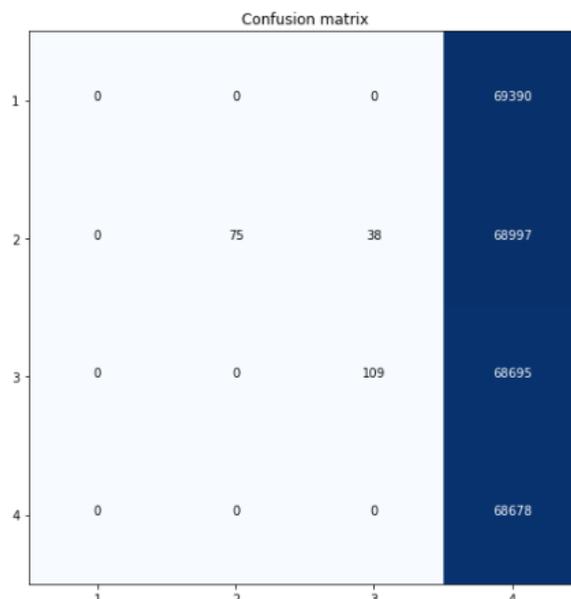


Ilustración 34 Matriz de confusión Tanh 40 neuronas

Por las razones expuestas en el punto anterior se puede ver que la red no se pudo adaptar a los datos, evidenciando un alto número de falsos positivos y negativos.

Random Forest

Este modelo no se tenía considerado dentro de los modelos a probar, pero como parte de las pruebas que se realizaron y con el fin de encontrar el mejor modelo se decidió utilizar. Para las pruebas que se realizaron se probó variando el parámetro de **n_estimators** que corresponde al número de árboles en el bosque, con el parámetro de **criterion** que corresponde a la función de separación, establecida en entropy.

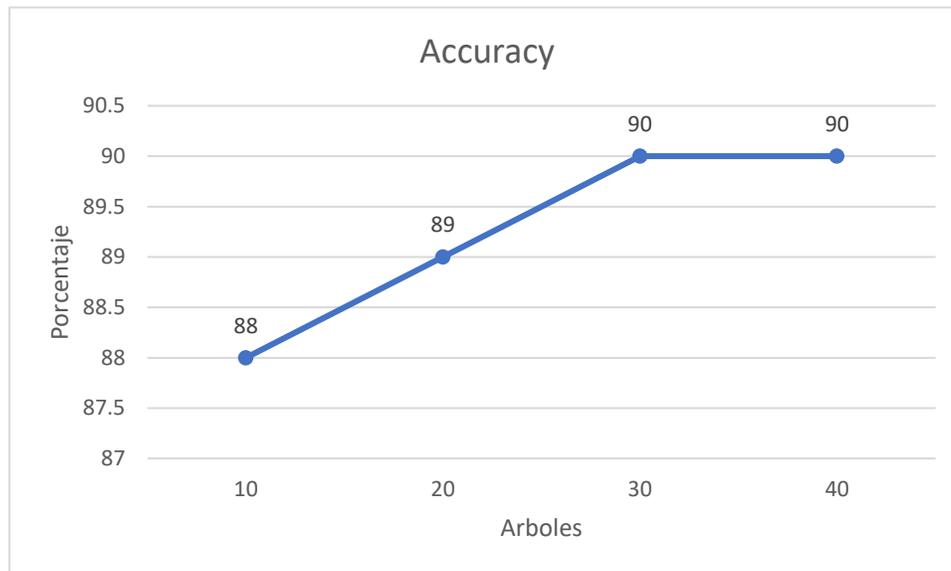


Ilustración 36. Accuracy Random Forest

Se aprecia que este modelo desde la primera prueba con 10 árboles se alcanza un accuracy de más del 80%, al llegar a 30 árboles el accuracy converge al 90%, por lo que se podría deducir que el aumentar el número de árboles no mejoraría más el accuracy del modelo.

Ahora se revisa el tiempo de ejecución del modelo, en donde se puede notar que aumenta conforme se incrementan los árboles. Aunado este tiempo de ejecución es bajo en comparación de los otros modelos probados.

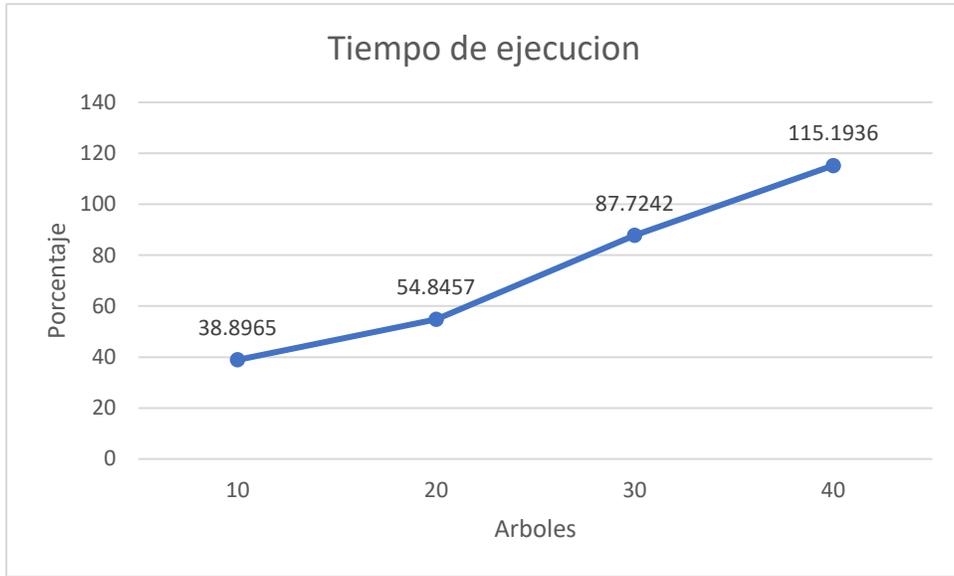


Ilustración 37. Tiempo ejecución Random Forest

Finalmente se revisa la matriz de confusión del modelo, en este caso se revisan las referentes a 10 y 40 árboles.

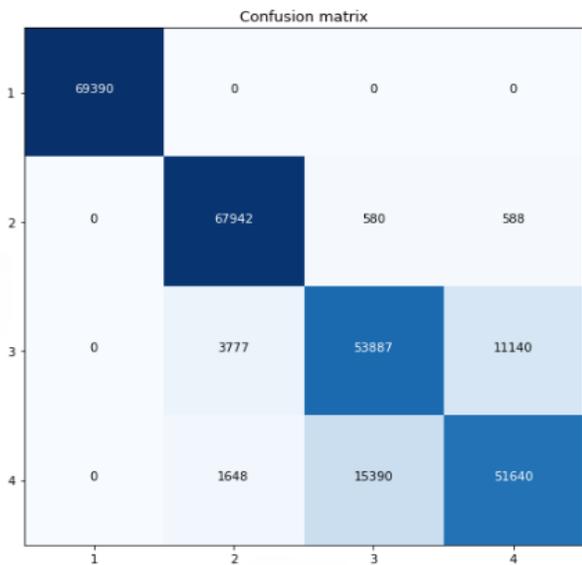


Ilustración 39. Matriz de confusión Random Forest, Arboles=10

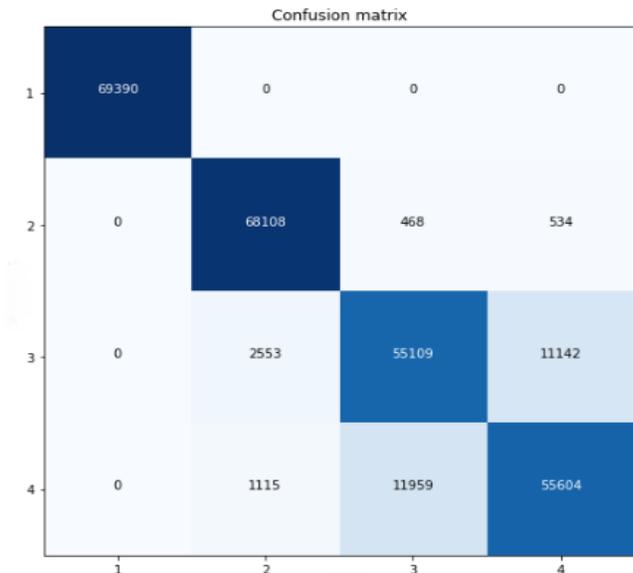


Ilustración 38. Matriz de confusión Random Forest, Arboles=40

Los resultados indican un alto número de verdaderos positivos lo que se ve reflejado en un alto número de clases predecidas correctamente.

Capítulo 5 Conclusiones

El principal objetivo del desarrollo de este trabajo de esta tesis fue obtener el mejor modelo de machine learning con la mejor precisión de aprendizaje al momento de predecir la gravedad de un accidente automovilístico.

Con el fin de mejorar la precisión de los modelos se realizaron procesos de limpieza de datos, como fue la validación de valores nulos y conversión de valores categóricos. Sin embargo, para obtener una mejor precisión de los modelos y evitar un alto coste computacional se aplicaron técnicas de balanceo de datos, así como técnicas de selección de características para obtener únicamente las variables más relevantes para el modelo.

Los modelos implementados en en este trabajo de tesis fueron; árboles de decisión, redes neuronales, KNN y Random Forest. De la implementación de estos modelos se obtuvieron los siguientes resultados.

Conclusiones Modelos

Para árboles de decisión se realizaron 8 pruebas variando en el parámetro de la profundidad del árbol, en donde logramos alcanzar un accuracy máximo de 88% y un tiempo de ejecución máximo de 8.2 segundos.

KNN mostró una ligera reducción en su precisión, ya que con $K=5$ alcanzamos un máximo accuracy 80%, referente al tiempo de ejecución fue uno de los más costosos con un tiempo de ejecución de 1068.1 segundos

Mientras que para las redes neuronales se realizaron cuatro sets pruebas por función de activación, en donde se cambió por ejecución el número total de neuronas. Los resultados obtenidos fueron que la red neuronal que utiliza la función de activación Relu alcanzó un accuracy máximo de 54%, además de un tiempo de ejecución máximo de 113.34 segundos.

Por su parte random forest mostró el mejor comportamiento referente al accuracy, pues se alcanzó un máximo de 90% y en tiempo de ejecución 115.11 segundos.

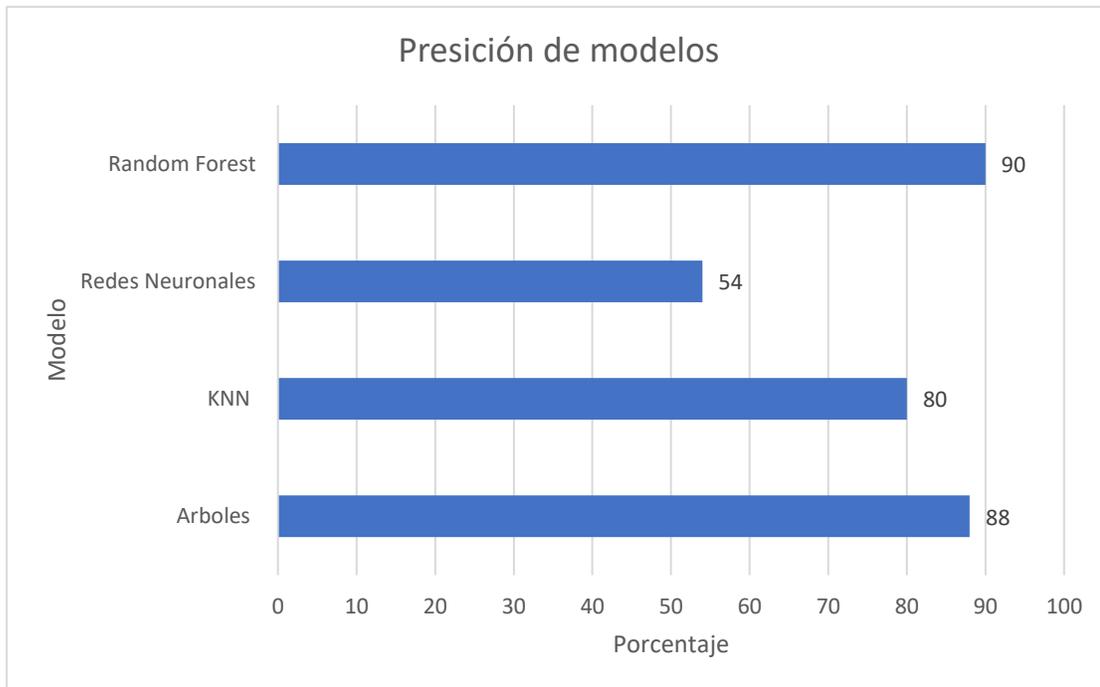


Ilustración 40. Precisión Modelos

En conclusión, Random Forest se consagro como el mejor modelo para poder calcular la gravedad de un accidente automovilístico además de que fue el que generó un menor costo computacional en relación con su precisión.

Por su parte la red neuronal que usa la función de la tangente hiperbólica fue el modelo con la presión más baja, alcanzando un máximo del 25% en su precisión.

Conclusiones generales

El machine learning es un área de las Ciencias Computacionales la cual nos puede ayudar a relacionar la información con los datos que tenemos, aprender y predecir patrones, a tomar decisiones más informadas y precisas, es por eso por lo que en este trabajo de tesis se aplicaron diferentes modelos a la predicción de accidentes automovilísticos alcanzando una precisión de más del 80% lo que podría ayudar a la aplicación de campañas con el fin de hacer conciencia en la población sobre las variables que pueden desencadenar en un accidente automovilístico y la gravedad de este mismo.

Los resultados fueron satisfactorios, pero principalmente en el desarrollo de este trabajo de tesis se obtuvo un amplio conocimiento de los modelos y un enfoque dirigido en la aplicación del machine learning en distintas áreas.

Referencias

- [1] Gartner, "About Gartner," Nov. 30, 2021. Disponible: <https://www.gartner.com/en/about> (Acceso Nov. 30, 2021)
- [2] Gartner, "Gartner Top Strategic Technology Trends for 22," Nov. 30, 2021. Disponible: <https://www.gartner.com/en/information-technology/insights/top-technology-trends>. (Acceso Nov. 30, 2021)
- [3] J. McCarthy, "WHAT IS ARTIFICIAL INTELLIGENCE?," Stanford University, Stanford CA, 2004, Disponible: https://borghese.di.unimi.it/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_what_i_sai.pdf
- [4] IBM, "¿Qué es la inteligencia artificial (IA)?," Jun. 3, 2020. Disponible: <https://www.ibm.com/mx-es/cloud/learn/what-is-artificial-intelligence> (Acceso Nov. 30, 2021)
- [5] S. Russell, y N. Peter. *Artificial Intelligence: A Modern Approach*, 3rd Edition. New Jersey: Prentice Hall, 2010. Disponible: <https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf>
- [6] A. Samuel, "Some studies in machine learning using the game of checkers," Jul. 10, 2019. Disponible: https://hci.iwr.uniheidelberg.de/system/files/private/downloads/636026949/report_frank_gabel.pdf (Acceso Dic. 01, 2021)
- [7] UC Berkeley School of Information, "What Is Machine Learning?," Jun. 26, 2020. Disponible: <https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/>. (Acceso Dic. 01, 2021)
- [8] IBM, "Redes Neuronales," Ago. 17, 2020. Disponible: <https://www.ibm.com/mx-es/cloud/learn/neural-networks>. (Acceso Dic. 01, 2021)
- [9]"GEORREFERENCIACIÓN DE ACCIDENTES DE TRÁNSITO EN ZONAS URBANAS", INEGI, Ciudad de México, COMUNICADO DE PRENSA, 653/21, noviembre de 2022. Accedido el 1 de febrero de 2023. [En línea]. Disponible: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/accidentes/ACCIDENTES_2021.pdf
- [10] J. Olivares López, "Estrategias para el análisis de sentimiento en textos extraídos de Twitter utilizando técnicas de aprendizaje profundo.", Tesis de Licenciatura, Benemérita Universidad Autónoma de Puebla, Puebla, 2022.
- [11] M. Lucía Violini, "Selección de características. Su aplicación a clasificación de texturas", Tesis de Licenciatura, Universidad Nacional de la Plata, Chile, 2014. Accedido el 26 de febrero de 2023. [En línea].

Disponible: http://sedici.unlp.edu.ar/bitstream/handle/10915/63236/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

[12] L. Gonzalez. "Métodos de Selección de Características". Aprende IA. <https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/#:~:text=La%20Selección%20de%20Características%20es,una%20mejor%20comprensión%20del%20proceso> (accedido el 26 de febrero de 2023).

[13] "RandomForestClassifier". <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> (accedido el 26 de febrero de 2023).

[14] S. Raschka y V. Mirijallili, Python Machine Learning. Aprendizaje Automatico y aprendizaje profundo con Python , scikit-learn y Tensor Flow, 2a ed. España: Marcombo, 2019.

[15] "Redes neuronales". <https://bootcampai.medium.com/redes-neuronales-13349dd1a5bb#:~:text=Funci%C3%B3n%20Lineal&text=Por%20lo%20tanto%2C%20esta%20funci%C3%B3n,de%20un%20n%C3%BAmero%20de%20ventas>. <https://bootcampai.medium.com/redes-neuronales-13349dd1a5bb#:~:text=Función%20Lineal&text=Por%20lo%20tanto,%20esta%20función,%20un%20número%20de%20ventas>. (accedido el 26 de febrero de 2023).