Modelo de redes neuronales artificiales para la predicción y detección de casos y brotes de enfermedades arbovirales (Zika, Dengue y Chikunguña) en Colombia

Juan David García López y Andrés Felipe Fandiño Plata

Trabajo de Grado para Optar el título de Ingeniero Industrial

Director

Henry Lamos Díaz

Ph.D en Física - Matemática

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2023

Agradecimientos

A mi familia quisiera expresar mi más profundo agradecimiento a cada uno de ustedes por haberme acompañado en este increíble viaje académico que culmina con la realización de mi tesis de grado. En especial, quiero dirigirme a mi amada madre, cuyo apoyo inquebrantable ha sido mi mayor fortaleza. Tu constante aliento y amor han sido la luz que me ha guiado a lo largo de esta travesía, y no puedo agradecerte lo suficiente por tu sacrificio y dedicación.

A mis amigos, quienes han sido pilares fundamentales en mi vida, les estoy enormemente agradecido. Su presencia, ánimo y palabras de aliento han sido como un bálsamo en los momentos desafiantes, recordándome que no estoy solo en este camino. A mis compañeros de carrera, hemos compartido risas, desafíos y triunfos, creando recuerdos que atesoraré siempre. Juntos hemos superado obstáculos y celebramos éxitos, construyendo una red de apoyo invaluable.

Mi gratitud se extiende también a los destacados docentes que han iluminado mi camino académico. Sus conocimientos, guía y paciencia han sido fundamentales en mi formación, y les agradezco sinceramente por compartir su experiencia y motivarme a alcanzar mis metas. Finalmente, a la Universidad Industrial de Santander, agradezco profundamente por brindarme la oportunidad de adquirir una educación de calidad. Esta institución ha sido el hogar donde he cultivado no solo conocimientos, sino también valores que llevaré conmigo a lo largo de mi vida. Gracias a todos por ser parte esencial de este capítulo significativo en mi vida.

Dedicatoria

A mi amada madre, fuente inagotable de inspiración y coraje, quiero expresar mi más profundo agradecimiento. Tu inquebrantable esfuerzo y valentía han sido la brújula que ha guiado mis propios pasos. Admiro tu dedicación y sacrificio, que han sido faro en los momentos más oscuros y combustible en mis momentos de duda. Esta tesis es un tributo a ti, un reflejo de la fortaleza que has instilado en mí a lo largo de los años. Gracias por ser mi mayor apoyo y mi constante motivación.

A los profesores y tutores que han compartido su conocimiento y experiencia durante mi formación académica, les debo un reconocimiento profundo. Sus enseñanzas han sido fundamentales para mi crecimiento intelectual y profesional. Cada desafío y cada logro son también suyos, ya que han sido guías esenciales en mi camino. Agradezco su paciencia, dedicación y pasión por la enseñanza, elementos cruciales que han dejado una marca imborrable en mi desarrollo académico.

A todas las personas que dedican sus vidas a la ardua labor de la investigación, simplemente gracias. Sus esfuerzos incansables son la fuerza impulsora detrás del avance de la humanidad. Me sobran palabras de elogio para describir la importancia de su trabajo y el impacto positivo que generan en el mundo. Esta tesis es también un homenaje a su dedicación y contribución invaluable al conocimiento y al progreso de la sociedad. Gracias por inspirarme a seguir explorando, cuestionando y aprendiendo.

Contenido

Pág.

Lista de Tablas

Pág.

Lista de Figuras

Pág.

Lista de Apéndices

Los apéndices están adjuntos y pueden ser visualizados en la base de Datos de la Biblioteca UIS

Apéndice A. Solicitud de usuario y contraseña bodega de datos SISPRO.

Apéndice B. Usuario y contraseña de acceso a base de datos SISPRO

Apéndice C. DATA_TOTAL organizada y consolidada

Apéndice D. Diagrama de red consulta 1

Apéndice E. Diagrama de red consulta 2

Apéndice F. Análisis bibliométrico en Power Bi

Apéndice G. Fases de la metodología

Apéndice H. Análisis preliminar de variables en Python

Apéndice I. Análisis 20 ciudades elegidas Power Bi

Apéndice J. Análisis 20 ciudades elegidas sin edad mayor a 120 Power Bi

Apéndice K. Proyecciones del conjunto final de datos Power Bi

Apéndice L. Arboviral_LSTM

Apéndice M. Arboviral_GRU

Apéndice N. Arboviral_RNN

Apéndice O. Artículo científico

Resumen

Título: Modelo de redes neuronales artificiales para la predicción y detección de casos y brotes de enfermedades arbovirales (Zika, Dengue y Chikunguña) en Colombia[*]

Autores: Andrés Felipe Fandiño Plata, Juan David Garcia Lopez[**]

Palabras claves: redes neuronales, enfermedades arbovirales, aprendizaje profundo, predicción.

Descripción:

Las enfermedades arbovirales transmitidas por artrópodos como los insectos se han repetido a lo largo de la historia, afectando a grandes segmentos de la población mundial. Las lesiones resultantes van desde la pérdida de la vida, hasta la discapacidad prolongada, además de los elevados costos en los sistemas de salud pública para su atención y tratamiento.

La presente investigación ejecuta un modelo de red neural de aprendizaje por medio del software Python, el cual agrupa las variables de tipo demográfico de 20 municipios de Colombia y las variables climatológicas seleccionadas de estos, en ambos casos desde el año 2007 al año 2021, a partir de tres gestores de datos ArcGIS World Geocoder, base de datos SISPRO y el DANE. Este modelo de aprendizaje automático de redes neuronales permite analizar y comparar el comportamiento de los brotes de estas enfermedades arbovirales, y así poder plantear una alternativa para su tratamiento y prevención en los sistemas de salud en Colombia.

[*] Trabajo de Grado

[**] Facultad de Ingenierías Fisicomecánicas Escuela de Estudios Industriales y Empresariales Director: Henry Lamos Díaz Ph.D en Física - Matemáticas

Abstract

Title: Artificial neural network model for the prediction and detection of cases and outbreaks of arboviral diseases (Zika, Dengue and Chikungunya) in Colombia[*]

Authors: Andrés Felipe Fandiño Plata, Juan David Garcia Lopez[**]

Keywords: neural networks, arboviral diseases, deep learning, prediction.

Description:

Arboviral diseases transmitted by arthropods such as insects have been repeated throughout history, affecting large segments of the world's population. The resulting injuries range from loss of life to prolonged disability, in addition to the high costs in public health systems for their care and treatment.

This research executes a neural network model of learning through Python software, which groups the demographic variables of 20 municipalities in Colombia and the selected climatological variables of these, in both cases from the year 2007 to the year 2021, based on three data managers ArcGIS World Geocoder, SISPRO database, and DANE. This automatic learning model of neural networks allows analyzing and comparing the behavior of outbreaks of these arboviral diseases, and thus being able to propose an alternative for their treatment and prevention in health systems in Colombia.

[*] Bachelor thesis

[**] Facultad de Ingenierías Fisicomecánicas Escuela de Estudios Industriales y Empresariales Director: Henry Lamos Díaz Ph.D en Física - Matemáticas

Introducción

Las enfermedades arbovirales, son un grupo de enfermedades infecciosas transmitidas por artrópodos. “Las infecciones arbovirales (abreviación del inglés 'arthropod-borne', o sea, 'transmitida por artrópodos') son causadas por uno de los tantos virus transmitidos por artrópodos, tales como mosquitos y garrapatas.” (Department of Health New York State, 2005). Dentro de dicho grupo, podemos destacar la incidencia del dengue, zika y chikunguña (Nuestro objeto de estudio), los cuales y según (María del Carmen Álvarez Escobar et al., n.d.) “son enfermedades del grupo de las arbovirosis, transmitidas por los mosquitos Aedes aegypti y Aedes albopictus”. Estos mosquitos han tenido una exitosa expansión por todo el mundo, “A nivel mundial, Aedes aegypti y Aedes albopictus son dos de las especies más importantes de mosquitos, en lo que se refiere a la transmisión de enfermedades. Ambas se consideran especies invasoras, ya que han colonizado exitosamente muchos sitios fuera de sus ámbitos nativos” (Rey & Lounibos, 2015), sobre todo en áreas tropicales y subtropicales (María del Carmen Álvarez Escobar et al., n.d.).

El número de contagios en el continente americano no es un factor que precise ser ignorado “En la Región de las Américas, entre la semana epidemiológica (SE) 1 y la SE 40 del año 2022, se notificaron un total de 2,780,867 casos de enfermedad por arbovirus. De estos, 2,499,047 (89.9 %) fueron casos de dengue, 250,369 (9.0 %) casos de Chikunguña, y 31,451 (1.1 %) fueron casos de zika”. (Organización Panamericana de la Salud, 2022). A nivel local, vemos que en Colombia la situación no es muy diferente, por ejemplo, según (Ojeda R et al., 2014) “En 2016 Se reportaron en el país 103.822 casos de Dengue (49.9%♀), 19.556 de Chikunguña

(63.3%♀) y 106.559 de Zika (66.4%♀).” , en el 2019 según (Rico-Mendoza et al., 2019), quien en su investigación titulada “Co-circulation of dengue, chikungunya, and Zika viruses in Colombia from 2008 to 2018” expone “En 2016 se reportaron 101.016 casos de dengue al SIVIGILA, de los cuales 59.114 no tenían signos de alarma, 41.003 presentaban señales de alarma y 899 eran dengue grave”, respecto al Chikunguña presentó que “Entre 2014 y 2016 se notificaron 19.435 casos de CHIKV en Colombia” y respecto al Zika muestra los siguientes datos de casos de contagios “Del 9 de agosto de 2015 al 2 de abril de 2016, un total de 65.726 se reportaron casos de ZIKV en Colombia” O según(*2022_Boletín_epidemiologico_semana_52*, n.d.), donde expone “En la semana epidemiológica 52 de 2022 se notificaron 2 058 casos probables de dengue: 1 007 casos de esta semana y 1 051 casos de semanas anteriores”, mientras que por parte del virus del Zika reporta un total nacional de 138 casos reportados y de Chikunguña un total de 94 casos, así podemos seguir enunciando múltiples estudios que evidencian la seria problemática que, en materia de salud, representan estos tres virus para coyuntura nacional.

Por las razones expuestas es imperativo destinar esfuerzos a la investigación de nuevos métodos que puedan ayudar a la búsqueda e identificación de futuros brotes de este tipo de enfermedades, ya que, en materia de costos, y para el caso del Dengue, según (Rodríguez et al., 2016) “El costo financiero total de la enfermedad en Colombia desde una perspectiva social fue de US$ 167,8 millones en 2010, US$ 129,9 millones en 2011 y US$ 131,7 millones en 2012.”. Respecto a los costos de ausentismo e incapacidad laboral (Carolina Sánchez et al., n.d.) expone “El cálculo real de los costos del absentismo es muy difícil de conseguir, teniendo en cuenta la complejidad de este fenómeno.”, no obstante, y según (Rodríguez et al., 2016), se puede generar una estimación de los costos asociados, dichos costos denominados como costos

indirectos comprendieron, costos por pérdida de productividad y absentismo tanto del paciente como del cuidador en caso de episodios no mortales por parte del paciente y el cuidador, estos costos están estimados en US y fueron proyectados con una tasa de cambio del dólar promedio para el año 2012 de 1.798,23 COP por U.S, generando así un costo estimado promedio total para los pacientes de US$ 1.762.657 (US$ 1.364.210– US$ 2.230.539) y para los cuidadores US$ 1.284.073 (US$1.063.016– US$1.542.116) en el año 2010 (con un intervalo de confianza del 95%). Por ello, y comprendiendo el impacto negativo que generan estas enfermedades en la economía nacional, se plantea la viabilidad del uso de Deep Learning (DL) y Machine Learning (ML) para la predicción de futuros brotes, tal como, por ejemplo, lo propone (Xu et al., 2020a), en su trabajo titulado "Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method" o (Taylor's University (Subang Jaya et al., n.d.) en la "2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA) proceedings : 26-28 October 2018 Taylor's University Lakeside Campus, Subang Jaya, Malaysia.", titulada "How to Efficiently Predict Dengue Incidence in Kuala Lumpur" donde hacen uso de estos métodos con el fin de generar estimaciones de brotes y contagios de estas enfermedades, así, en el presente trabajo se pretende usar los resultados de dichas investigaciones con el fin de apoyar a los sistemas de salud de nuestra nación.

Así, en el presente trabajo se ejecutaron y evaluaron los desempeños, bajo diferentes métricas, de 3 diferentes modelos de redes neuronales (LSTM, RNN, GRU) acotando el objeto de estudio, dada la disponibilidad de datos, a las primeras 20 ciudades colombianas con mayor registro de número de contagios de las enfermedades de Dengue, Zika y Chikunguña.

Tabla 1. *Cumplimiento de objetivos.*

OBJETIVOS ESPECÍFICOS	CUMPLIMIENTO
Revisar la literatura científica sobre las investigaciones realizadas frente a la predicción de enfermedades arbovirales (zika, dengue y chikunguña) por medio del uso de redes neuronales.	Capítulo 6
Comparar diferentes arquitecturas de redes neuronales (rnn, gru y lstm) para la predicción de los casos y brotes.	Numeral 8.1.2.
Seleccionar la red neuronal, de acuerdo con ciertas métricas como la eficiencia (diferentes métricas) obtenida para la predicción de los casos de enfermedades arbovirales (zika, dengue y chikunguña) en colombia.	Numeral 8.1.2.
Desarrollar una herramienta computacional de visualización de los modelos construidos.	Apéndice L, Apéndice M, Apéndice N.
Elaborar un artículo basado en la investigación realizada con el fin de ser publicable.	Apéndice O

Nota. Fuente propia

1. Generalidades del proyecto

1.1 Objetivos

1.1.1 Objetivo general

Construir un modelo de red neuronal artificial para la predicción de la cantidad de casos y detección de brotes de enfermedades arbovirales en Colombia.

1.1.2 Objetivos específicos

Revisar la literatura científica sobre las investigaciones realizadas frente a la predicción de enfermedades arbovirales (Zika, Dengue y Chikunguña) por medio del uso de redes neuronales.

Comparar diferentes arquitecturas de redes neuronales (RNN, GRU y LSTM) para la predicción de los casos y brotes.

Seleccionar la red neuronal, de acuerdo con ciertas métricas como la eficiencia (diferentes métricas) obtenida para la predicción de los casos de enfermedades arbovirales (Zika, Dengue y Chikunguña) en Colombia.

Desarrollar una herramienta computacional de visualización de los modelos construidos.

Elaborar un artículo basado en la investigación realizada con el fin de ser publicable.

2. Planteamiento y/o justificación del problema

Las enfermedades arbovirales, en este caso, dengue, zika y chikunguña, transmitidas por los mosquitos Aedes aegypti y Aedes albopictus (María del Carmen Álvarez Escobar et al., n.d.), son enfermedades con gran transmisión a nivel mundial, de hecho, y según (Taylor's University (Subang Jaya et al., n.d.) "Las enfermedades transmitidas por mosquitos se están propagando rápidamente en todas las regiones del mundo con una estimación de 2.500 millones de personas en todo el mundo están en riesgo" y su velocidad de propagación es muy rápida, razón por la cual, es de suma importancia generar pronósticos que ayuden a los sistemas de salud a estar preparados con medidas de contención ante este tipo de eventualidades, "La fiebre del dengue (DF) es una de las enfermedades de más rápida propagación en el mundo, y los pronósticos precisos del dengue de manera oportuna podrían ayudar al gobierno local a implementar medidas de control efectivas" (Xu et al., 2020).

Para algunos países se presentan serias dificultades para atender los requerimientos médicos de su población, usualmente los recursos destinados no son suficientes para financiar sistemas de salud adecuados, de hecho, los dineros suelen aplicarse en su mayoría a poblaciones urbanas, dejando relegadas a las poblaciones rurales, "India y varias naciones donde la población es alta, la atención médica es uno de los principales desafíos a tratar. Los recursos médicos que pone a disposición el gobierno no pueden hacer frente a la alta población. Las zonas rurales son las más afectadas debido a la falta de una infraestructura médica adecuada por parte del sector de la salud pública" (SCAD College of Engineering and Technology & Institute of Electrical and Electronics Engineers, n.d.).

Según (OPS, 2019) y (Centros para el Control y la Prevención de Enfermedades, 2019), el dengue, zika y chichunguña, no tienen una vacuna o tratamientos antivirales específicos que

los combatan, además, (Palomares-Marín J et al., 2018) nos expone que aunque existan vacunas que suelen ser administradas para combatir estas enfermedades, generan una protección no significativa, razón por la cual, el tratamiento de estas enfermedades se resume al cuidado del paciente durante el periodo de padecimiento de los síntomas (Giovanny Rincón-Silva & David Rincón Silva, n.d.), sobrecargando con ello los sistemas de salud de una nación, y por tanto, la capacidad de respuesta de los mismos, por ello, y como se mencionó antes, es imperativo generar planes de acción para que los sistemas de salud estén preparados ante estas eventualidades, partiendo de la mejora en los sistemas usados para identificar las personas infectadas y con ello, la identificación de los posibles brotes de este tipo de enfermedades, razón por la cual, se propone el uso de técnicas de Deep Learning y Machine Learning para este tipo de estudios, “ La identificación de las personas infectadas con dengue se determina mediante pruebas clínicas, pero la técnica propuesta se utiliza para la vigilancia automática y la identificación de las regiones donde la propagación está ocurriendo a un ritmo alarmante y guiar a los profesionales de la salud a tomar las medidas necesarias para controlar la propagación.” (Amin, Uddin, et al., 2020).

En el continente americano, y por tanto en Colombia, se presentan ciertos factores epidemiológicos que son favorables para el desarrollo de este tipo de epidemias, “En el caso de los países americanos se considera que estos son más vulnerables a brotes continuos de la epidemia debido a que aún no se ha generado un sistema ‘inmune grupal’. Es decir, las personas son más vulnerables a infecciones puesto que aún no se han desarrollado estrategias inmunes resistentes al virus.” (Giovanny Rincón-Silva & David Rincón Silva, n.d.).

En Colombia la incidencia de estos virus ha sido notable y ha generado problemas de salud pública en diversas regiones del país. “El dengue se ha notificado constantemente en

Colombia durante las últimas dos décadas causando un promedio de 84.926 casos cada año (1980-2019). El zika se notificó por primera vez en Colombia en 2015 y fue seguido por un brote significativo de 91.711 casos en 2016. El Chikunguña se detectó por primera vez en Colombia en 2013, causando 275.907 casos en ese solo año. Colombia ahora es hiperendémica para el dengue, así como endémica tanto para Zika como para chikungunya." (Morgan et al., 2021). Dejando con ello un elevado costo para el sistema de salud colombiano, "Esto deriva en múltiples traumatismos: desde pérdida de vidas humanas hasta prolongadas incapacidades laborales, pasando por onerosos costos para los sistemas públicos de salud." (Abultaif Amira, 2021)

Por ello, y ante el estado del sistema de salud público colombiano, donde según (Botero et al., n.d.) el sistema de salud colombiano cuenta con una cobertura cercana al 97,8%, pero "hay problemas de cobertura territorial y de calidad de servicios en regiones apartadas; hay problemas institucionales en la arquitectura del sistema, que si bien se han abordado recientemente, todavía siguen gravitando sobre su legitimidad; hay problemas de eficiencia en el sistema, por bajos niveles de productividad de las entidades que lo conforman, y especialmente por el deficiente desempeño de algunos hospitales públicos, aquejados de clientelismo y burocracia; pero ante todo, hay problemas de sostenibilidad financiera, que en cierto modo resultan de los problema ya mencionados, pero reflejan también una inconsistencia fundamental del sistema, le pedimos muchos más servicios de los que estamos dispuestos a pagar."(Alonso Botero et al., 2021). Ante ello, es imperativo, como se ha mencionado reiteradamente, generar planes de acción predictivos que aprovechen el florecimiento de las tecnologías de Deep Learning y Machine Learning para apoyar los sistemas de respuesta por parte del sistema de salud colombiano.

3. Marco teórico

En este fragmento, se determina el marco conceptual con base en las temáticas de utilidad que aportan a la comprensión, interpretación y caracterización de la presente investigación.

3.1. Machine Learning

El Machine Learning o aprendizaje automático es un campo científico y, más particularmente, una subcategoría de inteligencia artificial. Consiste en dejar que los algoritmos descubran «patterns», es decir, patrones recurrentes, en conjuntos de datos. Esos datos pueden ser números, palabras, imágenes, estadísticas, etc.

Todo lo que se pueda almacenar digitalmente puede servir como dato para el Machine Learning. Al detectar patrones en esos datos, los algoritmos aprenden y mejoran su rendimiento en la ejecución de una tarea específica.

En resumen, los algoritmos de Machine Learning aprenden de forma autónoma a realizar una tarea o hacer predicciones a partir de datos y mejorar su rendimiento con el tiempo. Una vez entrenado, el algoritmo podrá encontrar los patrones en nuevos datos.(*Machine Learning: Definición, Funcionamiento, Usos,* n.d.)

3.1.1. Etapas

Hay cuatro etapas principales en el desarrollo de un modelo de Machine Learning. Por lo general, es un Data Scientist quien gestiona y supervisa el proceso.

El primer paso es seleccionar y preparar un conjunto de datos de entrenamiento. Esos datos se utilizarán para alimentar el modelo de Machine Learning para aprender a resolver el problema para el que se ha diseñado. Los datos se pueden etiquetar para indicarle al modelo las

características que debe identificar. También pueden estar sin etiquetar, entonces será el modelo el que deberá detectar y extraer características recurrentes por sí mismo. En ambos casos, los datos deben prepararse, organizarse y limpiarse cuidadosamente. De lo contrario, el entrenamiento del modelo de Machine Learning puede estar sesgado. Los resultados de sus predicciones futuras se verán afectados directamente.(*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

El segundo paso es seleccionar un algoritmo para ejecutar sobre el conjunto de datos de entrenamiento. El tipo de algoritmo que se emplea depende del tipo y del volumen de datos de entrenamiento y del tipo de problema que haya que resolver.(*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

El tercer paso es entrenar el algoritmo. Es un proceso de repetición. Las variables se ejecutan a través del algoritmo y los resultados se comparan con los que debería haber producido. Los «pesos» y el sesgo se pueden ajustar para aumentar la precisión del resultado. Después se vuelve a ejecutar las variables hasta que el algoritmo produzca el resultado correcto en la mayoría de los casos. El algoritmo entrenado es el modelo de Machine Learning.

El cuarto y último paso es el uso y la mejora del modelo. Utilizamos el modelo sobre nuevos datos, cuyo origen depende del problema que haya que resolver. Por ejemplo, en los correos electrónicos se usará un modelo de Machine Learning diseñado para detectar spam.(*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

3.1.2. Principales algoritmos de machine learning

3.1.2.1. Algoritmos de regresión, lineal o logística. Permiten comprender las relaciones entre los datos. La regresión lineal se utiliza para predecir el valor de una variable dependiente en función del valor de una variable independiente. Sería, por ejemplo, para predecir las ventas

anuales de un comercial en función de su nivel de estudios o de experiencia. La regresión logística a su vez se utiliza cuando las variables dependientes son binarias. Otro tipo de algoritmo de regresión llamado máquina de vectores de soporte es pertinente cuando las variables dependientes son más difíciles de clasificar.(*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

3.1.2.2. Árbol de decisiones. Ese algoritmo permite establecer recomendaciones basadas en un conjunto de reglas de decisión partiendo de datos clasificados. Por ejemplo, es posible recomendar por qué equipo de fútbol apostar basándose en datos como la edad de los jugadores o el porcentaje de victorias del equipo.

3.1.2.3. Clustering. Para los datos no etiquetados, a menudo se utilizan los algoritmos de «clustering». Ese método consiste en identificar los grupos con registros similares y etiquetar esos registros según el grupo al que pertenecen. Anteriormente, se desconocen los grupos y sus características. Entre los algoritmos de clustering, encontramos K-medias, TwoStep o incluso Kohonen.(*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

3.1.2.4. Data Mining. Los algoritmos de asociación permiten descubrir patrones y relaciones en los datos, e identificar las relaciones “si/entonces”, llamadas “reglas de asociación». Esas reglas son similares a las que se utilizan en el campo del Data Mining o minería de datos. (*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

3.1.2.5. Redes neuronales. Por último, las redes neuronales son algoritmos en forma de red con varias capas. La primera permite la captación de datos, una o más capas escondidas permiten sacar conclusiones de los datos captados y la última capa asigna una probabilidad a

cada conclusión. Una red de neuronas “profunda” está compuesta por múltiples capas ocultas que permiten afinar los resultados de la anterior. Es la que se utiliza en el campo del Deep Learning.(*Machine Learning: Definición, Funcionamiento, Usos*, n.d.)

3.1.3. Deep learning

Deep learning es un subconjunto de machine learning (que a su vez es parte de la inteligencia artificial) donde las redes neuronales, algoritmos inspirados en cómo funciona el cerebro humano, aprenden de grandes cantidades de datos. Los algoritmos de deep learning realizan una tarea repetitiva que ayuda a mejorar de manera gradual el resultado a través de ‘’deep layers’’ lo que permite el aprendizaje progresivo. Este proceso forma parte de una familia más amplia de métodos de machine learning basados en redes neuronales. Deep learning ha tenido un gran impacto en todas las industrias. Por ejemplo, en las ciencias de la vida, el aprendizaje profundo se puede utilizar para el análisis avanzado de imágenes, la investigación, el descubrimiento de medicinas, la predicción de problemas de salud, así como síntomas de enfermedades, y la aceleración de conocimientos a partir de la secuenciación genómica. En el transporte, puede ayudar a los vehículos autónomos a adaptarse a las condiciones cambiantes; y a su vez ser utilizado para proteger infraestructuras críticas.(*IBM Cloud*, n.d.)

3.2. Redes neuronales artificiales

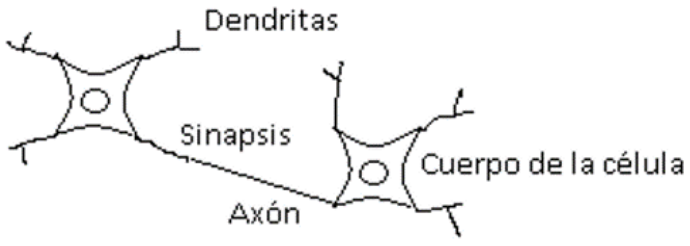
3.2.1. Redes neuronales biológicas

En las redes neuronales biológicas, la información se almacena en los puntos de contacto entre diferentes neuronas en el cerebro, es lo que normalmente se conoce como sinapsis. El premio nobel de medicina de 1906, Santiago Ramón y Cajal fue la primera persona en poder demostrar y validar que el sistema nervioso humano se compone de células individuales a las que se les empezó a llamar neuronas, las cuales se conectan entre sí creando una gran red de comunicación, pero en donde no se ha determinado hasta hoy la forma en que se procesa su información. La sinapsis se ha tratado de llevar a la forma artificial debido a la eficacia de los procesos llevados a cabo por el cerebro. Por lo que se ha desarrollado la teoría de las Redes Neuronales Artificiales (RNA), el objetivo es poder emular las redes neuronales biológicas y hacer posible que estas aprendan tácticas y soluciones basadas en ejemplos de comportamiento típico de patrones. Estos sistemas artificiales no necesitan de una programación previa, teniendo la capacidad de generalizar y aprender de la experiencia. En 1943, Warren McCulloch y Walter Pitts presentaron un modelo de neuronas artificiales y construyeron lo que fue considerado como el primer modelo de una red neuronal implementada en las ciencias de la computación. En 1949 Donald Hebb, empieza a trabajar más a fondo el concepto de aprendizaje de una red neuronal y empieza a definir el trabajo de neuronas interconectadas, aumentando su fuerza sináptica y activaciones de cambio de estas. Entre 1957 y 1959 Frank Rosenblatt desarrolla el Perceptrón, considerada como la red neuronal más antigua pero que en la actualidad sigue utilizándose como identificador de patrones. La historia de las redes neuronales artificiales y su comienzo donde se forjaron los principios y las bases para trabajar con neuronas artificiales se pueden situar desde el año 1936 al año 1986.((*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

3.2.2 Elementos básicos de las redes neuronales

Las redes neuronales artificiales, o como se les conoce generalmente: ANN (Artificial Neural Networks) o RNA (redes neuronales artificiales) tienen sus bases y funcionan de una forma muy similar a las redes neuronales biológicas del cerebro de las personas. Su funcionamiento se da gracias a elementos que se comportan como una neurona biológica en sus funciones principales. Para comprender cuáles con los elementos básicos que compone una red neuronal, primero es necesario conocer el funcionamiento de una neurona. Las neuronas tienen tres componentes principales, las cuales son denominadas dendritas, el cuerpo de la célula o 176 soma y el axón. El punto de conexión entre el axón de una célula y una dendrita de otra célula se llama sinapsis. En términos computacionales, las dendritas, son las receptoras de la red, que cargan de señales eléctricas el cuerpo de la célula. El cuerpo de la célula realiza la suma de esas señales de entrada. El axón es una fibra larga que lleva la señal desde el cuerpo de la célula hasta otras neuronas, como se puede visualizar en la Figura 1.((*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

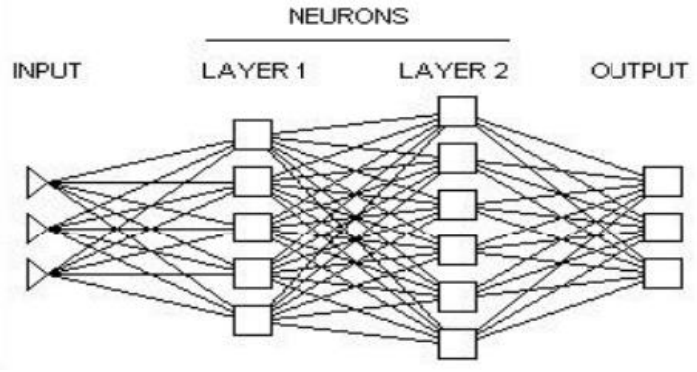
Figura 1. *Componentes principales de una neurona.*



Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 177), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

Una neurona como tal es diminuta en sí, pero cuando muchas se encuentran interconectadas, pueden formar toda una red de comunicaciones que pueden resolver problemas muy complejos. Por ejemplo, el cerebro de una persona contiene billones de neuronas. A esta comunicación entre neuronas se le denomina entonces una red neuronal. Se puede decir, por tanto, que una red neuronal está conformada por neuronas que se encuentran interconectadas y organizadas en tres capas. Los datos ingresan por medio de la "capa de entrada" (input), que pasan a través de la "capa oculta" (layer1, layer2) y salen por la "capa de salida" (output).(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 2. *Capas de una red neuronal.*



Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 177), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

3.2.3. Capas de la neurona artificial

Las neuronas se encuentran compuestas entonces por varias capas, de manera que las neuronas de una capa están conectadas con las neuronas de la capa siguiente, a las que pueden enviar información. Cada neurona de la red es una unidad de procesamiento de información que recibe información a través de las conexiones con las neuronas de la capa anterior.(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

▪ Capa de Entrada: Es quien recibe información del exterior. En las redes biológicas, esta sería tarea de las dendritas.

▪ Capas ocultas: La cuáles están encargadas de realizar el trabajo de la red. En las redes biológicas, está sería el soma.

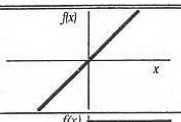
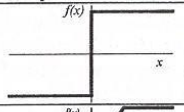
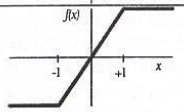
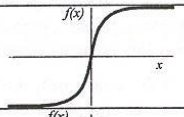
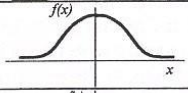
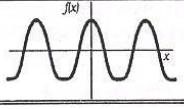
▪ Capa de Salida: Proporciona el resultado del trabajo de la red al exterior y envía información hacia otras neuronas. En las redes biológicas, esta sería una actividad realizada por el axón.(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

3.2.4 Función de activación

A una neurona artificial se le asigna un peso sináptico a las entradas que provienen desde otras neuronas. Este procedimiento es similar al que se realiza en una neurona de un ser humano, a lo que normalmente en la medicina se le conoce como sinapsis. El peso sináptico entonces es un valor numérico y que puede ir cambiando durante la fase de entrenamiento. Este peso hace que la red neural tengo una utilidad y es allí donde se almacena la información.

En un modelo neuronal, se debe disponer de una regla de propagación para combinar las salidas de cada neurona con las ponderaciones establecidas por el patrón de conexión, con eso se especifica la valoración de las entradas que recibe cada neurona. Normalmente puede realizarse una suma de las entradas, teniendo en cuenta el peso sináptico asociado a cada entrada. Aunque otras operaciones también son posibles. Con el valor obtenido con la regla de propagación, esta se filtra con de una función conocida como función de activación. A través de esta función se da la salida de la neurona. Las funciones de activación se escogen dependiendo del objetivo de entrenamiento de la red neuronal. En la Figura 3 se muestran las funciones de activación más utilizadas.(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 3. *Funciones de activación más utilizadas.*

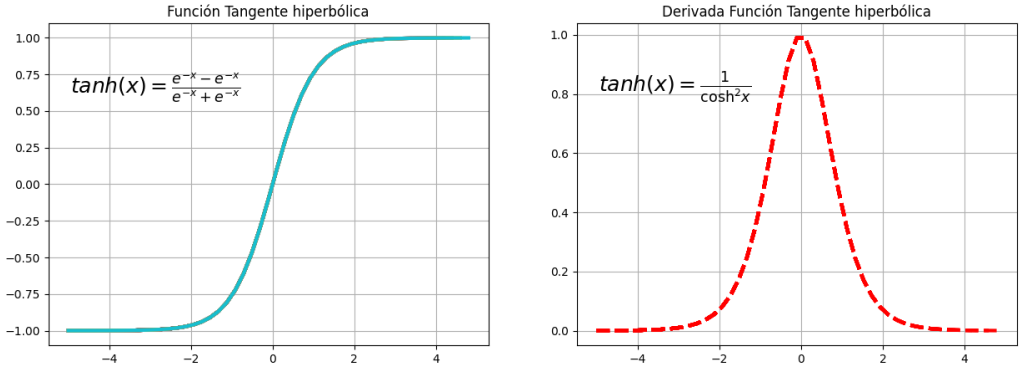
	Función	Rango	Gráfica
Identidad	$y = x$	$[-\infty, +\infty]$	
Escalón	$y = sign(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0, +1\}$	
Lineal a tramos	$y = \begin{cases} -1, & si\ x < -l \\ x, & si\ +l \le x \le -l \\ +1, & si\ x > +l \end{cases}$	$[-1, +1]$	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$ $y = tgh(x)$	$[0, +1]$ $[-1, +1]$	
Gaussiana	$y = Ae^{-Bx^2}$	$[0,+1]$	
Sinusoidal	$y = A\,\mathrm{sen}(\omega x + \varphi)$	$[-1,+1]$	

Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 178), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

3.2.4.1. Función Sigmoide. Esta función también conocida como función logística, está en un rango de valores de salida está entre cero y uno por lo que la salida es interpretada como una probabilidad. Si se evalúa la función con valores de entrada muy negativos, es decir x<0 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. Por lo que esta función se usa en la última capa y se usa para clasificar datos en dos categorías. Actualmente la sigmoide no es una función muy utilizada debido a que no está centrada y esto afecta en el aprendizaje y entrenamiento de la neurona por lo que influye con el problema de desaparición de gradiente.(*Redes Neuronales. Programa de Visión… | by Bootcamp AI | Medium*, n.d.)

3.2.4.2. Función Tangente hiperbólica o Gaussiana. Es una función similar a la Sigmoide, pero produce salidas en escala de [-1, +1]. Además, es una función continua. En otras palabras, la función produce resultados para cada valor de x.

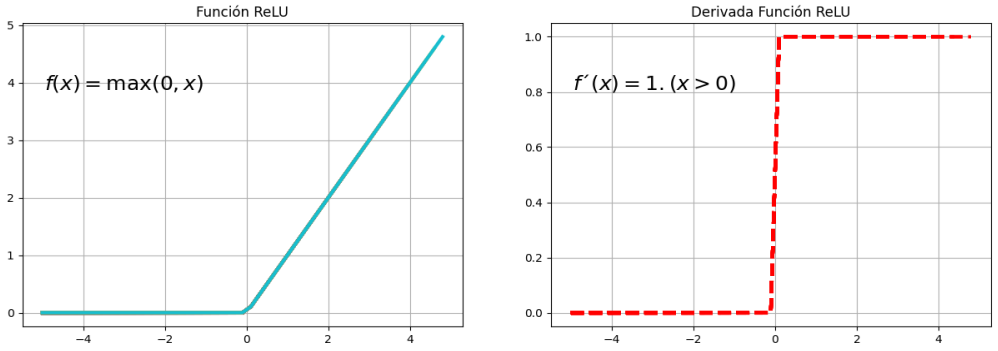
Figura 4. *Función tangente hiperbólica y derivada.*



Nota. Adaptado de La importancia de las funciones de activación en una red neuronal, por Jorge Calvo Martin, 2022, Linkedin (https://www.linkedin.com/pulse/la-importancia-de-las-funciones-activaci%C3%B3n-en-una-red-calvo-martin/?originalSubdomain=es).

3.2.4.3. Función RELU (Rectified Lineal Unit). ReLU es la función de activación más utilizada en el mundo en este momento. Desde entonces, se utiliza en casi todas las redes neuronales convolucionales o el aprendizaje profundo. Como puedes ver, ReLU está medio rectificado (desde abajo). f(z) es cero cuando z es menor que cero y f(z) es igual a z cuando z es superior o igual a cero. Es una función usada en las capas ocultas de nuestra red neuronal, NO en las de salida.

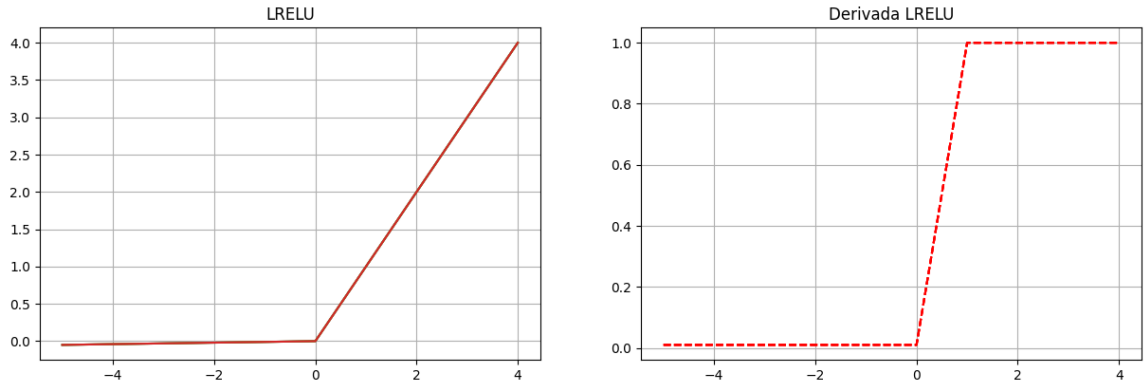
Figura 5. *Función ReLu y derivada.*



Nota. Adaptado de La importancia de las funciones de activación en una red neuronal, por Jorge Calvo Martin, 2022, Linkedin (https://www.linkedin.com/pulse/la-importancia-de-las-funciones-activaci%C3%B3n-en-una-red-calvo-martin/?originalSubdomain=es).

3.2.4.4. Leaky-ReLU. Leaky-ReLU es una mejora del valor predeterminado principal de ReLU, en el sentido de que puede manejar los valores negativos bastante bien, pero aún presenta no linealidad. $LRelu(x)=max(0.01x,x)$

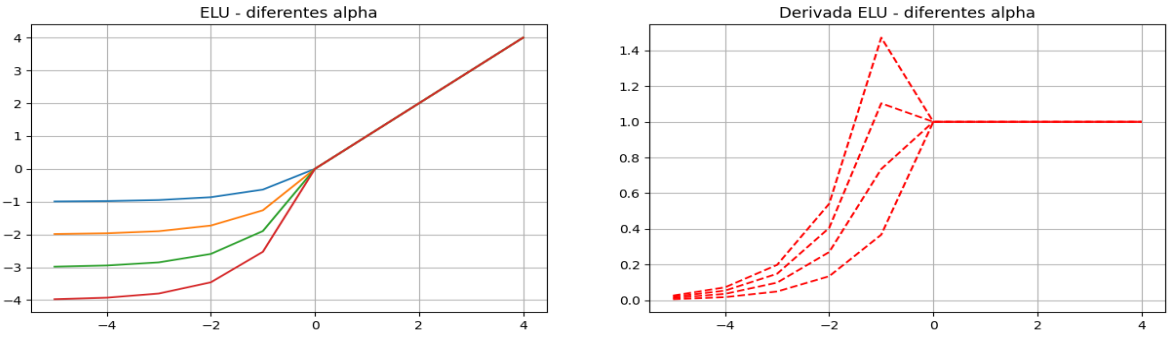
Figura 6. *Función LRELU y derivada.*



Nota. Adaptado de La importancia de las funciones de activación en una red neuronal, por Jorge Calvo Martin, 2022, Linkedin (https://www.linkedin.com/pulse/la-importancia-de-las-funciones-activaci%C3%B3n-en-una-red-calvo-martin/?originalSubdomain=es).

3.2.4.5. Función ELU (Exponential Linear Unit). La Unidad Lineal Exponencial (ELU) es una función de activación para redes neuronales. A diferencia de las ReLU, las ELU tienen valores negativos, lo que les permite acercar las activaciones de unidades medias a cero, como la normalización por lotes, pero con una menor complejidad computacional. Los cambios medios hacia cero aceleran el aprendizaje al acercar el gradiente normal al gradiente natural de la unidad debido a un efecto de cambio de sesgo reducido. Si bien las LReLU y las PReLU también tienen valores negativos, no garantizan un estado de desactivación resistente al ruido. Las ELU se saturan a un valor negativo con entradas más pequeñas y, por lo tanto, disminuyen la variación y la información propagadas hacia adelante.

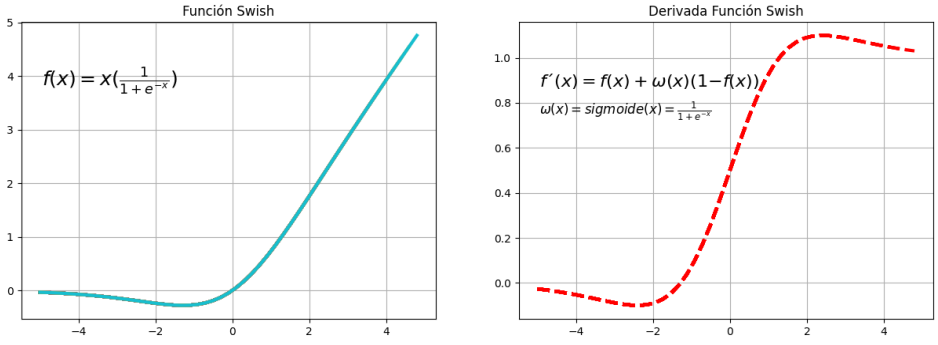
Figura 7. *Función ELU y derivada.*



Nota. Adaptado de La importancia de las funciones de activación en una red neuronal, por Jorge Calvo Martin, 2022, Linkedin (https://www.linkedin.com/pulse/la-importancia-de-las-funciones-activaci%C3%B3n-en-una-red-calvo-martin/?originalSubdomain=es).

3.2.4.6. Función Swish. Swish es una función suave y no monótona que iguala o supera constantemente a ReLU en redes profundas aplicadas a una variedad de dominios desafiantes, como la clasificación de imágenes y la traducción automática. Es ilimitado arriba y acotado abajo y es el atributo no monótono el que realmente crea la diferencia.

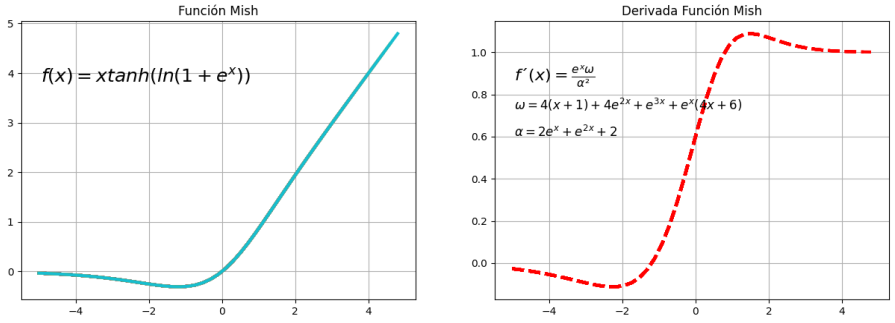
Figura 8. *Función Swish y derivada.*



Nota. Adaptado de La importancia de las funciones de activación en una red neuronal, por Jorge Calvo Martin, 2022, Linkedin (https://www.linkedin.com/pulse/la-importancia-de-las-funciones-activaci%C3%B3n-en-una-red-calvo-martin/?originalSubdomain=es).

3.2.4.7. Función MISH. Mish, una nueva función de activación no monotónica autorregulada inspirada en la propiedad de activación automática de Swish. Mish tiende a igualar o mejorar el rendimiento de las arquitecturas de redes neuronales en comparación con Swish, ReLU y Leaky ReLU a través de diferentes tareas en Computer Vision.

Figura 9. *Función Mish y derivada.*



Nota. Adaptado de La importancia de las funciones de activación en una red neuronal, por Jorge Calvo Martin, 2022, Linkedin (https://www.linkedin.com/pulse/la-importancia-de-las-funciones-activaci%C3%B3n-en-una-red-calvo-martin/?originalSubdomain=es).

3.2.5. Dropout

Dropout es un método que desactiva un numero de neuronas de una red neuronal de forma aleatoria. En cada iteración de la red neuronal dropout desactivara diferentes neuronas, las neuronas desactivadas no se toman en cuenta para el forwardpropagation ni para el backwardpropagation lo que obliga a las neuronas cercanas a no depender tanto de las neuronas desactivadas. Este método ayuda a reducir el overfitting ya que las neuronas cercanas suelen aprender patrones que se relacionan y estas relaciones pueden llegar a formar un patrón muy específico con los datos de entrenamiento, con dropout esta dependencia entre neuronas es

menor en toda la red neuronal, de esta manera las neuronas necesitan trabajar mejor de forma solitaria y no depender tanto de las relaciones con las neuronas vecinas.(Jordi et al., n.d.)

Dropout tiene un parámetro que indica la probabilidad de que las neuronas se queden activadas, este parámetro toma valores de 0 a 1, 0.5 suele usarse por defecto indicando que la mitad de las neuronas se quedaran activadas, si los valores son cercanos a 0 dropout desactivara menos neuronas, si es cercano a 1 desactivara muchas más neuronas. Dropout solo se usa durante la fase de entrenamiento, en la fase de pruebas ninguna neurona se desactiva, pero las escalamos en la probabilidad del dropout para compensar a las neuronas desactivadas durante la fase de entrenamiento.(*Dropout y Batch Normalization*, n.d.)

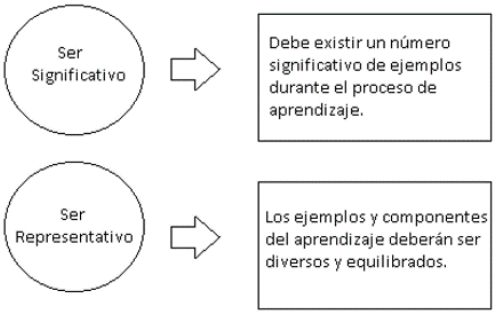
Se puede establecer un dropout diferente por cada capa, dependiendo de lo que necesitemos en cada una, en las capas de entrada suele usarse un dropout muy alto (0.7) para mantener a la mayoría de las neuronas activadas y en capas ocultas un dropout de (0.5). (Dropout Y Batch Normalization, n.d.)

3.2.6. Aprendizaje de las neuronas

Las neuronas artificiales se pueden clasificar de acuerdo con los valores que pueden tomar. Se pueden identificar dos tipos: 1) Neuronas binarias y 2) Neuronas reales. En el caso de las neuronas binarias, únicamente pueden tomar valores que se encuentren dentro del intervalo {-1, 1} o {0, 1}. En el caso de las neuronas reales, estas pueden tomar valores que se encuentren dentro de los intervalos [0, 1] o [-1, 1]. Generalmente, los pesos no se encuentran restringidos a intervalos específicos, aunque para aplicaciones específicas puede ser esto necesario.(Helm et al., 2020)

El proceso de aprendizaje de las redes neuronales artificiales es de carácter secuencial. De esta forma, el aprendizaje se da en todo momento adquiriendo conocimiento a través de las experiencias ocurridas. El conjunto de aprendizaje de una red neuronal artificial contiene dos características denominadas “ser significativo” y “ser representativo” en donde, para que haya aprendizaje, debe existir un número significativos de ejemplos durante el proceso de aprendizaje y estos deben ser diversos y equilibrados como se ve en la Figura 11.((*PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 10. *Características del conjunto de aprendizaje de una red neuronal artificial.*



Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 179), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

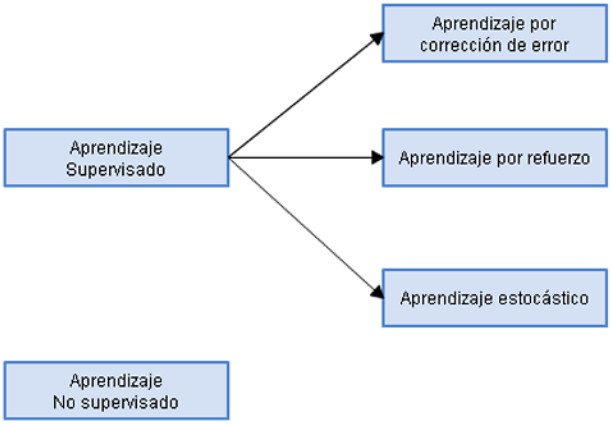
En el proceso de aprendizaje de una red neuronal, debe haber suficientes ejemplos para que la red sea capaz de adaptar sus pesos de forma eficaz, a esto se le llama "Ser significativo". Los ejemplos y componentes del aprendizaje de la red neuronal deben ser diversos y

equilibrados. Por ejemplo, si el conjunto de aprendizaje contiene un número mayor de ejemplos de un tipo que de otro, esta red estará más especializadas en un solo tipo de datos, a estos se le llama “Ser representativo”. Una red neuronal debe utilizar un tipo específico en la etapa de aprendizaje, realizando un entrenamiento para optimizar la función que analiza la salida de la red y poder determinar la eficiencia del aprendizaje.((*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

3.2.7. Tipos de aprendizaje

Las redes neuronales se basan en un algoritmo para aprender durante su etapa de aprendizaje y dependiendo del tipo que se esté utilizando. Se da por entendido que una red aprendió cuando los pesos de las conexiones han cambiado según la regla de aprendizaje utilizada en el entrenamiento y estos permanecen estables. Se puede tener tipos de aprendizaje supervisados y no supervisados y clasificados, como se muestra en la Figura 12. Se aclara que el tipo de aprendizaje supervisado es el más utilizado dentro de las redes neuronales artificiales.((*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 11. *Tipos de aprendizaje de una red neuronal artificial.*



Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 179), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

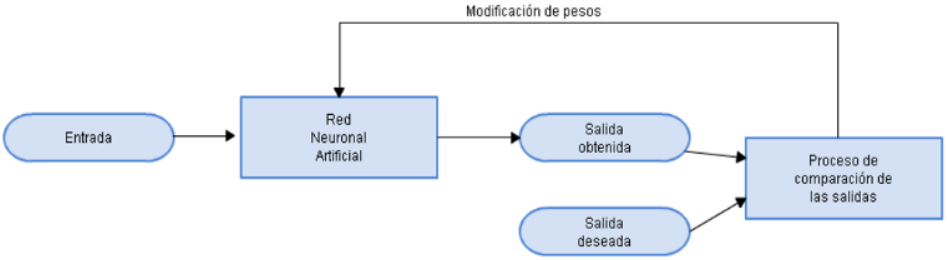
3.2.8. Aprendizaje supervisado

En este tipo de aprendizaje se realiza un entrenamiento de la red neuronal que estará supervisado y controlado por el diseñador de esta, para determinar que la respuesta de la red sea una específica dependiendo de la entrada. En caso de que la respuesta entregada sea diferente a la indicada, se modifican los pesos de las conexiones para aproximar la respuesta a la salida debida. Como se mencionó con anterioridad, a esto se la llama "Ser significativo".(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.,* n.d.)

3.2.9. Aprendizaje por corrección de error.

Durante el entrenamiento, se presenta a la red neuronal artificial, las entradas y salidas deseadas. La finalidad de este aprendizaje por corrección de error es que haya una diferencia mínima entre la salida obtenida y la deseada. Para ello, se hace una comparación de ambas salidas y se ajustan los pesos de las conexiones de la red teniendo en cuenta las diferencias con los valores deseados y los obtenidos (Figura 12).(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.,* n.d.)

Figura 12. *Aprendizaje por corrección de error.*

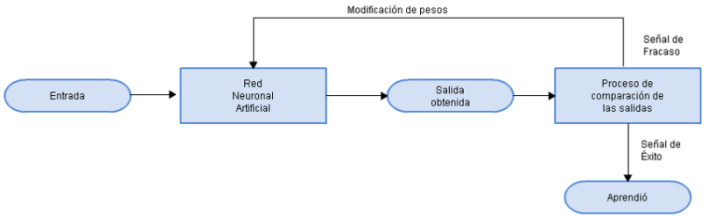


Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 180), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

3.2.10. Aprendizaje por Refuerzo

Para el entrenamiento de una red neuronal a través de este aprendizaje supervisado, no se cuenta con un ejemplo completo de una salida esperada, el diseñador de la red indica, mediante una señal de refuerzo, si la salida que se obtuvo de la red se acerca a la deseada. Las señales se refuerzo son: 1) Éxito o 2) Fracaso. Con esto se ajustan los pesos basándose en un mecanismo de probabilidades para acercarse a la salida deseada. Para este aprendizaje, la señal de refuerzo solo informa si la salida de la red se acerca a la deseada o no. En algunos algoritmos, con la señal de "Fracaso" se siguen modificando los pesos. Con la señal de "Éxito" el sistema ha cumplido con una salida deseada. Este aprendizaje por refuerzo suele ser más lento que el aprendizaje por corrección de error ya que puede recibir señales de Fracaso consecutivas. Entre más veces se reciba una señal de refuerzo, más tiempo tardará la red en aprender (Figura 13). (*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 13. *Aprendizaje por refuerzo.*

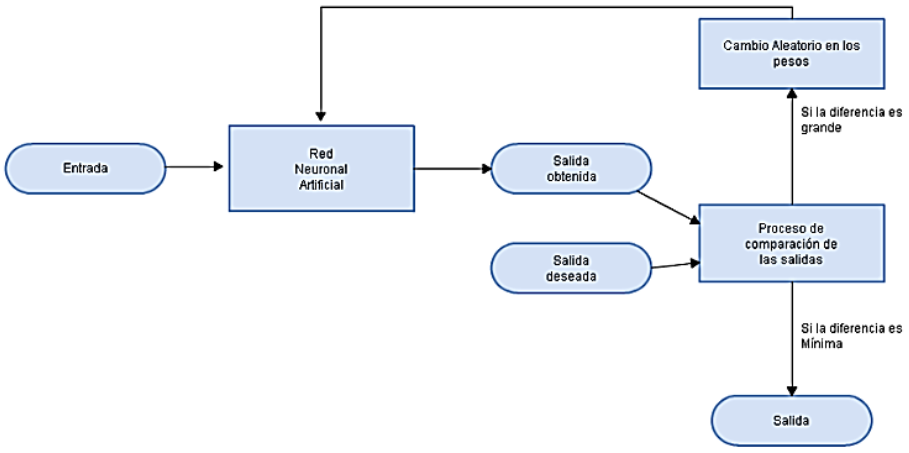


Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 181), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

3.2.11. Aprendizaje estocástico.

Durante el entrenamiento bajo el aprendizaje estocástico, se realizan cambios de forma aleatoria en los pesos de la red y se analiza la salida obtenida en comparación a la salida deseada (Figura 14). Si la diferencia de ambas salidas es mínima, esto significa que la red ha aprendido. Si la diferencia entre las salidas obtenida y esperada es mayor, se aceptarían cambios en el peso en función de una distribución de probabilidades determinadas. Este aprendizaje tiene similitud a los estados energéticos de los sólidos físicos, donde se maneja un estado mínimo de energía. Si después del cambio la energía decrece, se acepta el cambio. Si la energía no decrece, se acepta el cambio en función de una distribución de probabilidades determinada.((*PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 14. *Aprendizaje estocástico.*

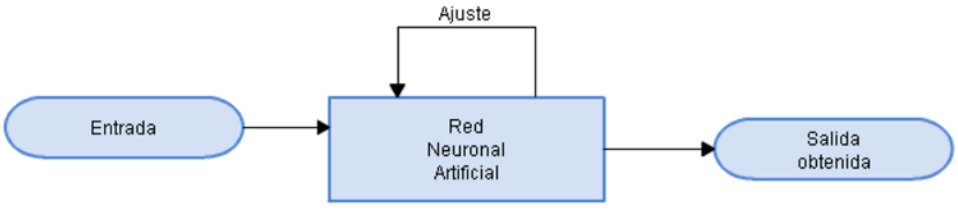


Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 181), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

3.2.12. Aprendizaje no-supervisado.

Como se observa en la Figura 16, este tipo de aprendizaje no requiere de una supervisión y no hay un proceso de comparación de salidas externos, como se puede visualizar en las figuras anteriores. Este tipo de aprendizaje muestra un proceso Autoorganización hasta cierto grado. La red neuronal descubre con los datos de entrada las características, regularidades, correlaciones y categorías, y lo hace de una forma autónoma.(*(PDF) Desarrollos de La Ingeniería Ambiental En La Evaluación de La Calidad de Los Recursos Naturales y La Salud Ambiental.*, n.d.)

Figura 15. *Aprendizaje no-supervisado.*



Nota. Adaptado de Desarrollo e Innovación en Ingeniería (2nd ed, p. 181), Acevedo M, E., Serna A, A., & Serna M, E. 2017, Editorial Instituto Antioqueño de Investigación.

Este tipo de aprendizaje fue desarrollado por Kohonen en el año de 1984 con el apoyo de otros investigadores. En este aprendizaje no se requieren de unas salidas deseadas y debidas. Por lo tanto, no se realizan comparaciones entre las salidas reales y las salidas. El algoritmo de entrenamiento modifica los pesos de la red de tal manera que se produzcan vectores de salida consistentes. Existen algunos algoritmos de aprendizaje no supervisados, pero la gran mayoría de trabajos se basan en el modelo propuesto en 1949 por Hebb que se caracteriza por incrementar el valor del peso de la conexión si las dos neuronas unidas son activadas. Hebb mencionó que, si dos neuronas que se encuentran interconectadas entre sí se activan al mismo tiempo, esto quiere decir que la fuerza sináptica ha incrementado. La forma de corrección utilizada se basa en incrementar la magnitud de los pesos si ambas neuronas están inactivas al mismo tiempo. (Acevedo M et al., 2017)

3.3. Infecciones arbovirales

Las infecciones arbovirales son causadas por uno de los tantos virus transmitidos por artrópodos, tales como mosquitos y garrapatas. Estas infecciones aparecen con más frecuencia durante los meses de clima cálido, cuando los mosquitos y las garrapatas están activos.(*Infecciones Arbovirales (Encefalitis Transmitida Por Artrópodos, Encefalitis Equina Oriental, Encefalitis de St. Louis, Encefalitis de California, Encefalitis Powassan, Encefalitis Del Nilo Occidental)*, n.d.)

3.3.1. Dengue.

El dengue es una infección vírica transmitida a los humanos por la picadura de mosquitos infectados. Los principales vectores de la enfermedad son los mosquitos Aedes aegypti y, en menor medida, Ae. Albopictus, el virus que causa esta enfermedad es el virus del dengue.(*Dengue y Dengue Grave*, n.d.)

3.3.2. Chikunguña.

La fiebre chikunguña es una enfermedad causada por un virus que trasmiten los mosquitos. El nombre significa "doblarse por el dolor" o "caminar encorvado", porque la infección causa dolor intenso de articulaciones y músculos. Otros síntomas pueden incluir fiebre elevada repentina, dolor de cabeza, fatiga, sarpullido, náuseas y ojos rojos, (Pritish K. Tosh, 2022) se propaga por la picadura de los mosquitos Aedes aegypti o Aedes albopictus infectados, que son los mismos vectores del dengue.(*¿Qué Es La Fiebre Chikungunya? ¿Debería Preocuparme? - Mayo Clinic*, n.d.)

3.3.3. Zika.

La fiebre del Zika es una enfermedad viral transmitida por mosquitos del género Aedes causada por el virus Zika (ZIKV), y que consiste en fiebre leve, sarpullido (principalmente maculo-papular), dolor de cabeza, dolor en las articulaciones, dolor muscular, malestar general y conjuntivitis no purulenta que ocurre entre 2 a 7 días después de la picadura del mosquito vector. Una de cada cuatro personas infectadas puede desarrollar síntomas, pero en quienes sí son afectados la enfermedad es usualmente leve, con síntomas que pueden durar entre 2 y 7 días. La

apariencia clínica es muchas veces similar a la del dengue, que también se transmite por mosquitos.(*Servicios de Salud - OPS/OMS | Organización Panamericana de La Salud*, n.d.)

3.4. Regresión lineal múltiple

Un modelo de regresión lineal múltiple es un modelo estadístico versátil para evaluar las relaciones entre un destino continuo y los predictores.(*SERIE DE TALLERES DE MODELOS DE REGRESIÓN LINEAL, MÚLTIPLE Y LOGÍSTICA – Escuela Global*, n.d.)

Los predictores pueden ser campos continuos, categóricos o derivados, de modo que las relaciones no lineales también estén soportadas. El modelo es lineal porque consiste en términos de aditivos en los que cada término es un predictor que se multiplica por un coeficiente estimado. El término de constante (intercepción) también se añade normalmente al modelo.

La regresión lineal se utiliza para generar conocimientos para los gráficos que contienen al menos dos campos continuos con uno identificado como el destino y el otro como un predictor. Además, se puede especificar un predictor categórico y dos campos continuos auxiliares en un gráfico y se pueden utilizar para generar un modelo de regresión adecuado. (IBM, s.f.)

3.5. Train-Test

El procedimiento de división de prueba de tren se usa para estimar el rendimiento de los algoritmos de aprendizaje automático cuando se usan para hacer predicciones sobre datos que no se usan para entrenar el modelo.

Es un procedimiento rápido y fácil de realizar, cuyos resultados le permiten comparar el rendimiento de los algoritmos de aprendizaje automático para su problema de modelado

predictivo. Aunque es fácil de usar e interpretar, hay momentos en los que no se debe usar el procedimiento, como cuando tiene un conjunto de datos pequeño y situaciones en las que se requiere una configuración adicional, como cuando se usa para la clasificación y el conjunto de datos no está equilibrado. (Brownlee, 2020)

3.6. Métricas de evaluación

3.6.1. Error Cuadrático medio (MSE):

El Error Cuadrado Medio (MSE) es el promedio de la diferencia al cuadrado entre el valor real y los valores predichos por el modelo de regresión. La razón de la cuadratura del error es eliminar cualquier signo negativo. Al cuadrar el error, MSE penaliza el error más que MAE. (medium.com, 2022)

Figura 16. *Ecuación MSE.*

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

Nota. Adaptado de Formula, por Not Nice Square Error, 2020 (https://emilia-orellana44.medium.com/not-nice-square-error-2d18c248391c)

3.6.2. Raíz del Error cuadrático medio (RMSE)

La raíz del error cuadrático medio (RMSE) mide la diferencia media entre modelos estadísticos, valores predichos y los valores reales. Matemáticamente, es la desviación estándar de los residuos. Los residuos representan la distancia entre la línea de predicción y los puntos de

datos. RMSE cuantifica qué tan dispersos están estos residuos, revelando qué tan estrechamente se agrupan los datos observados alrededor de los valores predichos. A medida que los puntos de datos se acercan a la línea de regresión, el modelo tiene menos errores, bajando el RMSE. Un modelo con menos error produce predicciones más precisas. (Frost, 2023).

La raíz del error cuadrático medio (RMSE) mide la diferencia media entre modelos estadísticos, valores predichos y los valores reales. Matemáticamente, es la desviación estándar de los residuos. Los residuos representan la distancia entre la línea de predicción y los puntos de datos. RMSE cuantifica qué tan dispersos están estos residuos, revelando qué tan estrechamente se agrupan los datos observados alrededor de los valores predichos. A medida que los puntos de datos se acercan a la línea de regresión, el modelo tiene menos errores, bajando el RMSE. Un modelo con menos error produce predicciones más precisas. (Frost, 2023).

Figura 17. *Ecuación RMSE.*

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

Nota. Adaptado de FÓRMULA RMSE, por Root Mean Square Error (RMSE), 2023 (https://statisticsbyjim.com/regression/root-mean-square-error-rmse/)

3.6.3. Error absoluto medio (MAE)

El error absoluto medio representa el promedio de la diferencia absoluta entre los valores reales y predichos en el conjunto de datos. Mide el promedio de los residuos en el conjunto de

datos. (*MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric Is Better? | by Akshita Chugh | Analytics Vidhya | Medium*, n.d.)

Figura 18. *Ecuación MAE.*

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Nota. Adaptado de FÓRMULA MAE, por MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? ,2020 (https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e)

3.6.4. R – Cuadrado:

El R-cuadrado es una medida estadística de qué tan cerca están los datos de la línea de regresión ajustada. También se conoce como coeficiente de determinación, o coeficiente de determinación múltiple si se trata de regresión múltiple. La definición de R-cuadrado es bastante sencilla: es el porcentaje de la variación en la variable de respuesta que es explicado por un modelo lineal.(*Análisis de Regresión: ¿Cómo Puedo Interpretar El R-Cuadrado y Evaluar La Bondad de Ajuste?*, n.d.) Es decir:

R-cuadrado = Variación explicada / variación total

El R-cuadrado siempre está entre 0 y 100%:

- 0% indica que el modelo no explica ninguna porción de la variabilidad de los datos de respuesta en torno a su media.
- 100% indica que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media.

En general, cuanto mayor es el R-cuadrado, mejor se ajusta el modelo a los datos.(*Análisis de Regresión: ¿Cómo Puedo Interpretar El R-Cuadrado y Evaluar La Bondad de Ajuste?,* n.d.)

3.6.5. Valor-P

En estadística una hipótesis es una afirmación que se realiza sobre algún o algunos parámetros de una distribución de probabilidad de una población de estudio. Una afirmación se rechaza cuando los resultados obtenidos en una muestra de la población no sean compatibles con ella. Cuando una hipótesis se rechaza es necesario señalar el nivel de significancia. Se define como nivel de significancia.

En este punto sabemos que si queremos demostrar una hipótesis solamente hay que plantear su opuesta y demostrar que los datos disponibles no la soportan. Para lo que calcularemos el p-value. Lo que requiere asumir que los datos se comportan de una manera dada, por ejemplo, que son aleatorios. Así si tenemos una moneda y la hipótesis de que esta no está trucada la posibilidad de obtener una cara es de 0,5, dos seguidas 0,25, tres seguidas 0,12. Siendo esta el p-value de la hipótesis en cada uno de los casos, la probabilidad de que se cierta la hipótesis.(*Significado de P-Value En Machine Learning - Analytics Lane*, n.d.)

3.7. Optimización de modelos.

El modelado de optimización es un enfoque matemático que se utiliza para encontrar la mejor solución a un problema desde un conjunto de opciones posibles, teniendo en cuenta restricciones y objetivos específicos. El aprendizaje automático utiliza una variedad de métodos, métricas y funciones para encontrar la combinación ideal de parámetros para mejorar el aprendizaje del modelo con el error más bajo posible. (¿Qué Es El Modelado De Optimización? | IBM, n.d.).

3.7.1. Análisis de correlaciones

El análisis de correlación consiste en un procedimiento estadístico para determinar si dos variables están relacionadas o no. El resultado del análisis es un coeficiente de correlación que puede tomar valores entre -1 y +1. El signo indica el tipo de correlación entre las dos variables. Un signo positivo indica que existe una relación positiva entre las dos variables; es decir, cuando la magnitud de una incrementa, la otra también. Un signo negativo indica que existe una relación negativa entre las dos variables. Mientras los valores de una incrementan, los de la segunda variable disminuyen. Si dos variables son independientes, el coeficiente de correlación es de magnitud cero. La fuerza de la relación lineal incrementa a medida que el coeficiente de correlación se aproxima a -1 o a +1.(*Análisis de Correlación – Conogasi,* n.d.)

Se requiere une entrada/muestra que sea una matriz de n X 2, con n filas (observaciones) y 2 columnas (variables). Se pueden usar recursos/material software estadístico: R, SAS, SPSS, Stata. Muy importante que la matriz no tenga valores ausentes (mismo número observaciones en variable X y en variable Y). (*Análisis de Correlación – Conogasi,* n.d.)

3.7.1.1. Procedimiento. La fórmula general para calcular el coeficiente de correlación entre dos variables es:

Figura 19. *Fórmula para calcular el coeficiente de correlación.*

$$r = Cov_{xy} / S_{xx} S_{yy}$$

Nota. Alquicira, J. (2017, 25 de Mayo) Análisis de correlación. Conogasi, Conocimiento para la vida. Fecha de consulta: Enero 30, 2024. Sitio web: https://conogasi.org/articulos/analisis-de-correlacion-2/

El coeficiente de correlación es el resultado de dividir la covarianza entre las variables X y Y entre la raíz cuadrada del producto de la varianza de X y la de Y. (*Análisis de Correlación – Conogasi*, n.d.)

- Calcular la covarianza entre la variable X y la variable Y (entre las dos columnas de la matriz) de acuerdo con la siguiente fórmula:

Figura 20. *Fórmula para calcular la covarianza entre dos variables.*

$$Cov_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N}$$

Nota. Alquicira, J. (2017, 25 de Mayo) Análisis de correlación. Conogasi, Conocimiento para la vida. Fecha de consulta: Enero 30, 2024. Sitio web: https://conogasi.org/articulos/analisis-de-correlacion-2/

Se calcula la media de todos los valores de X y de Y Se realiza la sumatoria del producto de las diferencias entre cada observación de cada variable y su media correspondiente. La

sumatoria calculada anteriormente se divide entre el número total de observaciones menos 1.(*Análisis de Correlación – Conogasi*, n.d.)

- Calcular la varianza de la variable X y la varianza de la variable Y , y obtener la raíz cuadrada de cada una:

Figura 21. *Producto de desviaciones estándar.*

$$\sqrt{\left(\frac{\sum_{i=1}^{n}(x_i-\tilde{x})^2}{N}\right)}\sqrt{\left(\frac{\sum_{i=1}^{n}(y_i-\tilde{y})^2}{N}\right)}$$

Nota. Alquicira, J. (2017, 25 de Mayo) Análisis de correlación. Conogasi, Conocimiento para la vida. Fecha de consulta: Enero 30, 2024. Sitio web: https://conogasi.org/articulos/analisis-de-correlacion-2/

Para cada variable se calcula la desviación estándar y se multiplican

- Se divide la covarianza entre el producto de las desviaciones estándar Ejemplo en R: Crear dos variables dependientes y utilizar la función cor(), O utilizar la fórmula antes propuesta.(*Análisis de Correlación – Conogasi*, n.d.)

3.7.2. Normalización de datos

La normalización es una técnica de pre-procesamiento de datos que se utiliza para ajustar las características o atributos de los datos dentro de un rango específico y mejorar su interpretación. Por lo tanto, la normalización es un proceso esencial en la ciencia de datos y el aprendizaje automático que se utiliza para mejorar la eficiencia y la precisión de los algoritmos.

(*“¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

En términos simples, la normalización es una forma de escalar y transformar los datos para que estén en un rango común, independientemente de la escala original de los datos. Esta técnica se utiliza para estandarizar los datos y reducir el impacto de las diferencias en la escala y la magnitud de los atributos de los datos. Por ejemplo, si una característica tiene valores en el rango de 1 a 1000 y otra característica tiene valores en el rango de 1 a 5, la característica con valores más grandes tendrá una influencia dominante en el modelo de aprendizaje automático.

En el aprendizaje automático, la normalización es importante porque muchos algoritmos de aprendizaje automático, como la regresión logística, los árboles de decisión y las redes neuronales, requieren que los datos estén normalizados para funcionar correctamente.

Además, la normalización también es importante porque ayuda a mejorar la interpretación de los resultados del análisis de datos. La normalización permite a los investigadores comparar las características de diferentes conjuntos de datos y hacer inferencias sobre su distribución. Esto es especialmente útil en el análisis estadístico y la visualización de datos.

Existen varias técnicas de normalización comunes en la ciencia de datos y el aprendizaje automático, como la normalización de z-score, la normalización min-max y la normalización por desviación estándar. Cada una de estas técnicas tiene sus propias ventajas y desventajas, y la elección de la técnica adecuada depende del conjunto de datos específico y los requisitos del

modelo. Una técnica importante que hay que comprender en el pre-procesamiento de datos. Cuando echamos un vistazo al conjunto de datos de automóviles usados, observamos en los datos que la característica de longitud oscila entre 150-250, mientras que la característica de anchura y altura oscila entre 50-100. Puede que queramos normalizar estas variables para que el rango de los valores sea consistente. (*"¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

3.7.2.1. Normalización de z-score. también conocida como estandarización, es una técnica de normalización utilizada en estadísticas y en el campo del aprendizaje automático. Consiste en transformar los valores de una característica en un conjunto de datos de tal manera que tengan una media de cero y una desviación estándar de uno. Esto se logra restando la media de los datos y dividiendo el resultado por la desviación estándar. (*"¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

Figura 22. *Ecuación de Normalización método Z-Score.*

$$N_i = \frac{\left(X_i - \mu\right)}{\sigma}$$

Nota. Blanco, J. I. (2023, April 28). "¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? Medium. https://jorgeiblanco.medium.com/por-qu%C3%A9-la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0.

La normalización de z-score es útil porque permite comparar características con diferentes unidades y rangos de valores en un mismo conjunto de datos. Además, los datos normalizados son más robustos a los valores atípicos (outliers) que otros métodos de normalización. Esta técnica se utiliza a menudo en la fase de preprocesamiento de datos en el aprendizaje automático, antes de aplicar un modelo de machine learning a los datos. (*"¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

3.7.2.2. La normalización Min-Max. Es una técnica de normalización utilizada en estadísticas y en el campo del aprendizaje automático. Consiste en transformar los valores de una característica en un conjunto de datos de tal manera que estén en un rango de valores específico, típicamente entre 0 y 1. (*"¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

Figura 23. *Ecuación de Normalización método Min — Max.*

$$N_i = \frac{\left(X_i - X_{min}\right)}{X_{max} - X_{min}}$$

Nota. Blanco, J. I. (2023, April 28). "¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? Medium. https://jorgeiblanco.medium.com/por-qu%C3%A9-la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0

La normalización Min-Max es útil porque permite comparar características con diferentes unidades y rangos de valores en un mismo conjunto de datos, y también es una técnica de normalización muy simple de implementar. Sin embargo, la normalización Min-Max puede ser sensible a los valores atípicos (outliers) en los datos, lo que puede afectar la distribución de los valores normalizados. En general, se recomienda utilizar la normalización Min-Max junto con otras técnicas de preprocesamiento de datos, como la detección y eliminación de valores atípicos, para obtener los mejores resultados en el análisis de datos y el aprendizaje automático. (*"¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

3.7.2.3. Escalado simple. es una técnica de preprocesamiento de datos utilizada en estadísticas y en el campo del aprendizaje automático. Consiste en transformar los valores de una característica en un conjunto de datos de tal manera que estén en un nuevo rango de valores específico. Es decir, sólo divide cada valor por el valor máximo de esa característica. Esto hace que los nuevos valores oscilen entre cero y uno. (*"¿Por Qué La Normalización Es Clave e Importante En Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium*, n.d.)

Figura 24. *Ecuación de normalización método escalado simple.*

$$N_i = \frac{\left(X_i\right)}{X_{max}}$$

Nota. Blanco, J. I. (2023, April 28). "¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? Medium. https://jorgeiblanco.medium.com/por-qu%C3%A9-la-

normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0

3.7.3. Reducción de la dimensionalidad

Las técnicas de reducción de la dimensionalidad se refieren al proceso de disminuir el número de dimensiones o características de un conjunto de datos, dentro de sus técnicas se incluyen la extracción y combinación de características, destacando métodos como el Backward Elimination y el Análisis de Componentes Principales – PCA. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.3.1. Multicolinealidad. La multicolinealidad implica una fuerte dependencia lineal entre las variables independientes, lo que resulta en una estimación única de los parámetros y falsas relaciones entre la variable dependiente y los regresores, lo que resulta en inferencias estadísticas poco precisas. Para hacer frente a este problema se suelen usar técnicas de selección de características como el Factor de Inflación de la Varianza y el Backward Elimination. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.3.2. Factor de Inflación de la Varianza (VIF). El Factor de Inflación de la Varianza (VIF) cuantifica la influencia de la varianza de una variable independiente por su interacción con las demás y permite detectar problemas graves de multicolinealidad. En general, los valores VIF superiores a 5 indican que las variables independientes involucradas están altamente correlacionadas. (MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS, n.d.).El VIF se representa:

Figura 25. *Formula del factor de inflación de la varianza (VIF).*

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Nota. Fuente propia.

Donde ${R_i}^2$ representa el coeficiente de determinación no ajustado para la regresión de la i-ésima variable independiente sobre las demás.

3.7.3.3. Backward Elimination. En el análisis de regresión, la eliminación hacia atrás es un método para elegir un subconjunto de características explicativas importantes para el modelo. El método utiliza un modelo de regresión y determina la significancia estadística de las variables independientes a un nivel de significancia del 5 %. Luego, elimina el predictor menos significativo (el valor p más alto) hasta que todos los predictores del modelo son significativos. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.3.4. Forward Selection. El uso de la selección hacia adelante es una forma de evaluar el desempeño de un modelo a medida que se agregan variables. Consiste en un tipo de regresión que comienza con un modelo vacío y luego agrega la variable única que produce la mejora óptima individual para el modelo en cada paso. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.4 Ajuste de hiperparámetros (GridSearchCV)

Los valores de las configuraciones ajustables de un modelo para controlar el proceso de entrenamiento se conocen como hiperparámetros. La búsqueda en rejilla, también conocida como "GridSearch ", es un método de búsqueda exhaustivo que examina todas las combinaciones de valores de hiperparámetros especificados en el conjunto de datos de entrenamiento y validación. Los hiperparámetros más efectivos generarán las métricas de desempeño más efectivas, como el coeficiente de determinación. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.5. Función de error

La función de error o función de pérdida es un método utilizado en Aprendizaje Automático para evaluar la efectividad de un algoritmo para modelar los datos, una alta desviación de los valores reales arrojaría un número grande en la función de pérdida (Parmar, 2018). Las funciones de pérdida más usadas son el error cuadrático medio y el error absoluto medio. La función de pérdida aprende gradualmente a reducir el error en la predicción utilizando algún optimizador. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.6. Optimizadores

Los optimizadores son métodos utilizados en Aprendizaje Automático para modificar los atributos del modelo, como los pesos y la tasa de aprendizaje, con el objetivo de minimizar la función de pérdida general y mejorar la precisión del modelo; algunos optimizadores son el Gradiente Descendente (Stochastic, Minibatch), Adam, Momentum, AdaGrad y RMSProp. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

3.7.7. Tamaño de lote y épocas

En redes neuronales, el tamaño de lote o batch size hace referencia al número de submuestras de datos de entrenamiento para la entrada, un tamaño de lote pequeño acelera el proceso de aprendizaje y uno grande aumenta la precisión del modelo. Por su parte, las épocas son el número de veces que el conjunto de datos completo pasa, hacia adelante y hacia atrás, a través de un modelo de red neuronal. (*MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS*, n.d.)

4. Factores asociados

A continuación, analizamos algunos de los factores geográficos, sociales, culturales y políticos asociados con el aumento del número de personas infectadas con enfermedades arbovirales. Son factores relevantes las variables geográficas, sociales, culturales, económicas y/o políticas que en conjunto o individualmente influyen en el aumento del número de infecciones por enfermedades arbovirales.

Es difícil determinar la prevalencia e importancia de estos factores porque no tenemos datos disponibles, pero según la Organización Panamericana de la Salud, los factores macro asociados a los arbovirus son: latitud, altitud (<2200 m sobre el nivel del mar), temperatura ambiente (15-40 grados centígrados), humedad (media a alta), habitación (urbanización no

planificada y alta densidad de población), condiciones de la vivienda (ventanas sin soportes, agua de lluvia en canalones, basura atascada), botellas en las paredes, suministro de agua (sin protección). almacenamiento de agua en la casa por más de 7 días, falta de suministro de agua por red separada, suministro y uso periódico de tanques o tanques sin tapa), recolección de desechos sólidos, contenedores de almacenamiento inadecuados, recolección inadecuada o inexistente, contenedores pequeños desechados, neumáticos desechados y otros artículos dejados al aire libre; estatus socioeconómico (ingresos bajos o insuficientes) y estatus cultural (favorable a la sociedad en relación con la transmisión del dengue, la transmisión de mosquitos, las complicaciones del dengue y las conductas de riesgo de mortalidad debido a factores económicos y ambientales).

Importancia e incidencia de los factores: Los factores geográficos son de suma importancia, ya que las condiciones climáticas y la topografía pueden influir en la propagación de los mosquitos vectores y la replicación de los virus.

Las áreas con baja altitud, alta temperatura y humedad brindan condiciones favorables para la reproducción de mosquitos y aumentan la incidencia de enfermedades arbovirales.

En el ámbito social, el estatus socioeconómico juega un papel importante. Las personas de bajos ingresos a menudo enfrentan malas condiciones de vida, carecen de medidas preventivas adecuadas y viven en áreas con atención médica limitada, lo que puede contribuir a la propagación de la enfermedad.

Los factores culturales y políticos están estrechamente relacionados con las prácticas de prevención y tratamiento de enfermedades. Una planificación urbana adecuada y un acceso eficiente a los servicios de salud pueden frenar la incidencia creando un entorno más saludable y promoviendo la detección temprana y la gestión de casos.

A nivel micro, los factores individuales del huésped, las características del patógeno y las características del vector influyen directamente en la transmisión y la gravedad de la enfermedad. Factores como el género, la edad, la inmunidad, la ocupación y el comportamiento individual y comunitario pueden influir en la susceptibilidad y el riesgo de infección.

En conclusión, es importante desarrollar estrategias integrales de prevención y control de enfermedades arbovirales que tengan en cuenta aspectos tanto macro como micro y las complejas interacciones entre factores geográficos, sociales, culturales y ambientales.

(Pública, 2020)

4.1 Ámbito Social

La creciente urbanización de la población es uno de los principales factores que contribuyen al aumento de los casos de enfermedades por arbovirus en las Américas. Alrededor del 80% de la población colombiana vive en grandes áreas urbanas, lo que significa que una proporción significativa de la población está infectada con diferentes serotipos circulantes de enfermedades arbovirales. Esta tendencia se ha intensificado en las últimas décadas debido a los flujos migratorios desde las zonas rurales, la migración provocada por los conflictos sociales armados, la falta de oportunidades de empleo en las zonas rurales y la preocupación por la seguridad.

En la mayoría de los municipios donde las enfermedades arbovirales son endémicas, la urbanización se ha dado de manera descontrolada y acelerada, particularmente en las cabeceras municipales. Esto ha creado problemas para que los gobiernos locales brinden servicios de

saneamiento adecuados, como agua limpia y eliminación adecuada de desechos sólidos. Debido a la naturaleza desordenada de la urbanización, satisfacer la demanda, cobertura, frecuencia y calidad de estos servicios se ha convertido en un desafío para las administraciones municipales. La limitada disponibilidad de agua potable en varios municipios en riesgo ha obligado a la población a almacenar agua para consumo humano inmediato y usos domésticos. Este almacenamiento se realiza utilizando varios tipos de contenedores, lo que contribuye al crecimiento y propagación de posibles criaderos de estas enfermedades.

4.2 Ámbito Cultural

En las zonas endémicas de transmisión del país, la práctica de almacenar agua está profundamente arraigada en las creencias culturales de la mayoría de la población. Existen determinados factores que favorecen la continuación de esta práctica, como la falta de concienciación sobre el riesgo individual y colectivo que supone, y el importante desfase entre el conocimiento sobre las medidas de prevención y control necesarias y la puesta en marcha real de acciones periódicas para prevenirlas. Además, el controlar los posibles criaderos de mosquitos transmisores de estas enfermedades.

Existen barreras significativas y profundamente arraigadas en la cultura popular, como el paternalismo hacia la población, relacionado con campañas e intervenciones de control institucional. La creencia es que el gobierno tiene la responsabilidad exclusiva del control del dengue y que la fumigación es el único método efectivo. Tal comportamiento ha creado una

dependencia y una demanda irracional por el uso de insecticidas como medida principal para el control de vectores, lo que induce una falsa sensación de seguridad. Además, la población desconoce sus responsabilidades en la prevención y control del dengue, lo que contribuye al mantenimiento y aumento de la magnitud del problema.

4.3 Ámbito Político

El insuficiente desarrollo de las instituciones de salud a nivel local y su limitada capacidad para responder a la detección y manejo de los brotes de dengue, junto con la ausencia de medidas sostenidas de promoción, prevención y control, son factores importantes que contribuyen a la transmisión endémica o epidémica de las enfermedades arbovirales. El liderazgo del sector salud en la búsqueda de compromisos de otros sectores y la participación comunitaria en la gestión integral e integrada es débil debido a modelos de intervención reduccionistas y acciones sociales que no se alinean con la realidad cultural. Además, existe una falta de apoyo político real y comprometido a nivel local para el desarrollo y mantenimiento de políticas, planes y proyectos de prevención y control de estas enfermedades.

El sector salud ha asumido la responsabilidad de ejecutar medidas de promoción, prevención, vigilancia y control del dengue, debido al abordaje fragmentado y desconectado con otros sectores y la comunidad. Como resultado, el sector depende principalmente del control químico para la prevención y control de la enfermedad, utilizando insecticidas y larvicidas para eliminar o controlar las formas inmaduras y adultas del vector. Estas medidas se complementan con campañas de recolección de basura y chatarra, así como con medidas de información, educación y comunicación, las cuales, a pesar de aumentar el conocimiento sobre el dengue en la población, no han sido efectivas para controlar los criaderos de mosquitos en los hogares y

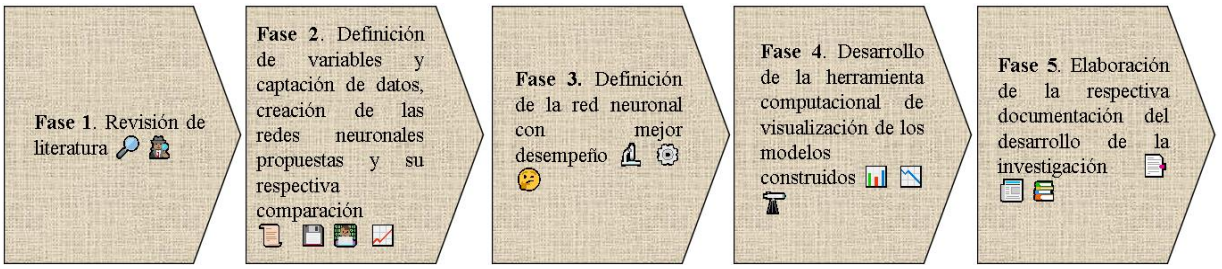
comunidades. Además, estas actividades no son parte de una estrategia integral de cambio de comportamiento. El enfoque del sector salud es reduccionista, y sus acciones sociales no se alinean con la realidad cultural, lo que también contribuye al problema. Adicionalmente, existe una falta de compromiso político decidido a nivel territorial para apoyar el desarrollo y mantenimiento de políticas, planes y proyectos para la prevención y control del dengue.

Si bien la cobertura y accesibilidad del Sistema General de Seguridad Social en Salud (SGSSS) es buena para la población de las zonas urbanas, existen dificultades para acceder a una atención médica oportuna. Además, existe una baja sensibilidad del personal médico para identificar casos y signos de alarma; existe falta de cumplimiento de las guías clínicas oficiales para la atención integral, lo que conduce a prácticas y decisiones médicas inadecuadas durante el manejo clínico, que influyen en el pronóstico de los casos y aumentan la frecuencia de formas graves y muertes por dengue. (SÁNCHEZ CABRERA, 2015)

5. Metodología para la investigación

Con el fin de dar cumplimiento a los objetivos planteados para el desarrollo del proyecto, se propone la metodología KDD usualmente utilizada en proyectos de minería de datos y Machine Learning, tal como se muestra en la Figura 26.

Figura 26. *Fases metodológicas.*



Nota Generada en Microsoft PowerPoint. Emojis obtenidos de: https://getemoji.com/

5.1 Fase 1: Revisión de literatura

- Estudiar el uso que se le ha dado a la literatura disponible sobre los modelos predictivos y redes neuronales artificiales para la predicción de enfermedades arbovirales.
- Establecer la base de datos a usar durante la investigación.
- Definir las palabras y términos claves de búsqueda adecuados en investigaciones de métodos computacionales de Machine Learning y Deep Learning en la predicción de casos de enfermedades arbovirales (Dengue, Zika, Chikunguña)
- Crear la ecuación de búsqueda adecuada acorde a las necesidades de información de la investigación.
- Llevar a cabo un estudio bibliométrico para determinar el estado actual del tema de investigación.

5.2 Fase 2: Definición de variables y captación de datos, creación de las redes neuronales propuestas y su respectiva comparación.

- Definición de variables que han sido de interés para otros autores que han realizado investigaciones previas en esta área de estudio.
- Recopilación de datos predefinidos de las fuentes predeterminadas y posterior creación del dataset.

- Preprocesamiento de los datos obtenidos, pasando por los procesos de: Limpieza de datos y gestión de datos faltantes, recopilación de los datos útiles para nuestra investigación por medio del uso de reducción de variables de dimensionalidad y la Transformación de los datos necesaria por medio de métodos de discretización, normalización o escalado.
- Creación de modelos de redes neuronales artificiales usando el lenguaje de programación Python.
- Determinar las medidas de rendimiento a utilizar para evaluar la calidad del ajuste de los modelos.
- Evaluar los modelos con las medidas de rendimiento correspondientes y determinar las variables que tienen mayor impacto en la predicción de brotes de enfermedades Arbovirales en Colombia (Dengue, Zika y Chikunguña).

5.3 Fase 3: Definición de la red neuronal con mejor desempeño.

- Comparar y elegir el modelo de red neuronal que se ajusta mejor a los valores de brotes de Dengue, Zika y Chikunguña observados.
- Verificar la precisión de las variables identificadas en los modelos con virólogos expertos en la propagación de estas enfermedades.

5.4 Fase 4: Desarrollo de la herramienta computacional de visualización de los modelos construidos

- Escoger entre Power BI, Tableau y Python la herramienta computacional de visualización a utilizar.
- Cargar en la herramienta computacional de visualización escogida la data de los modelos de redes neurales obtenidos.
- Identificar las proyecciones gráficas que mejor se adapten a los objetivos de visualización.

5.5 Fase 5: Elaboración de la respectiva documentación del desarrollo de la investigación.

- Creación de un documento final (Libro) que resuma los trabajos realizados durante la implementación del proyecto.
- Escribir un artículo científico que sea apto para su publicación.
- Entrega de artículo al director y codirector.
- Entrega del libro a director de trabajo de grado.

6. Revisión de literatura

6.1 Ecuación de búsqueda

En la construcción de la ecuación de búsqueda se proponen tres grupos con términos claves buscando encontrar documentación que se encuentre bastante relacionado con el tema de investigación propuesto. Los términos asociados se presentan en la siguiente tabla.

Tabla 2. *Palabras claves y términos asociados.*

Palabra clave	Inglés	Español
Inteligencia artificial	artificial neural network / Recurrent Neural Network / Deep Learning	Red neuronal artificial / Red neuronal recurrente / Aprendizaje profundo
Enfermedades	Dengue / Zika / Chikungunya / Arbovirus	Dengue / Zika / Chikunguña / Arbovirus
Predicción	Prediction / Detection / Forecasting	Predicción / Detección / Pronóstico
Modelos	RNN / GRU / LSTM	RNN / GRU / LSTM

Tabla 3. *Ecuación de búsqueda.*

Fecha	Ecuación	Resultado
Noviembre 3, 2022	TITLE-ABS-KEY (("artificial neural network" OR "Recurrent Neural Network" OR "DEEP LEARNING")) AND TITLE (("Dengue" OR "Zika" OR "Chicunguña" OR "Arboviral")) AND TITLE-ABS-KEY (("Prediction" OR "Detection" OR "Forecasting")) AND TITLE-ABS-KEY (("Diseases" OR "Dengue" OR "Zika" OR "Chicunguña" OR "Arboviral")) AND TITLE-ABS-KEY (("RNN" OR "GRU" OR "LSTM")).	22

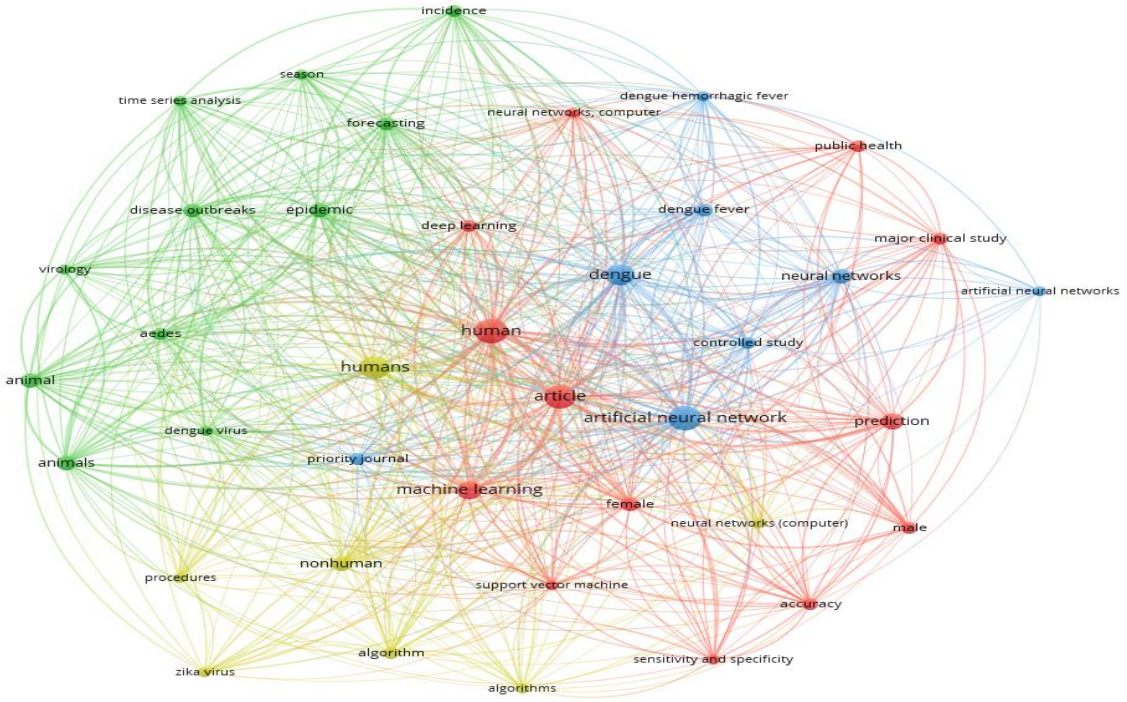
Se utilizó la herramienta SCOPUS un recurso disponible en la Biblioteca virtual de la Universidad Industrial de Santander para realizar la búsqueda y posterior análisis de la información para el desarrollo del presente proyecto. Todo esto, con el fin de encontrar referencias bibliográficas que permitieran extraer la mayor cantidad de información relacionada con modelos para la predicción de futuros casos de Dengue, Zika y Chikunguña de acuerdo con las variables a analizar.

De una revisión de literatura previa se identificaron las siguientes palabras claves: Neural Network, Dengue, Prediction, LSTM y Forecasting, las cuales se utilizaron como base para formular la siguiente ecuación de búsqueda, validada por el director y el codirector del proyecto.

Los parámetros de búsqueda utilizados nos arrojan un resultado total de sólo 22 documentos, una cantidad que puede considerarse pequeña si la comparamos con otras ecuaciones de búsqueda de trabajos de investigación similares realizados con anterioridad, no obstante, las razones para esta detallada delimitación recaen en la calidad de la consulta, pues se buscó la mayor correlación posible entre el tema de investigación y los trabajos previamente realizados en dicho campo de estudio.

Por ejemplo, en una primera aproximación, denominada “Consulta 1” se generó la ecuación de búsqueda (TITLE-ABS-KEY (("artificial neural network" OR "Recurrent Neural Arboviral"))) , de la cual se obtuvieron 268 resultados, ubicados en un espacio de tiempo de 17 años, dicha consulta presentaba una evidente carencia de profundidad en el tema de estudio, pues en un análisis de palabras claves realizado (Ver Figura 27), se pudo determinar la ausencia de términos claves como “forecasting”, “Prediction”, “Deep Learning”, “Machine Learning”, “Artificial neural networks” términos relevantes en este campo de estudio, razón por la cual, dichos términos fueron incluidos en él refinamiento de la ecuación de búsqueda.

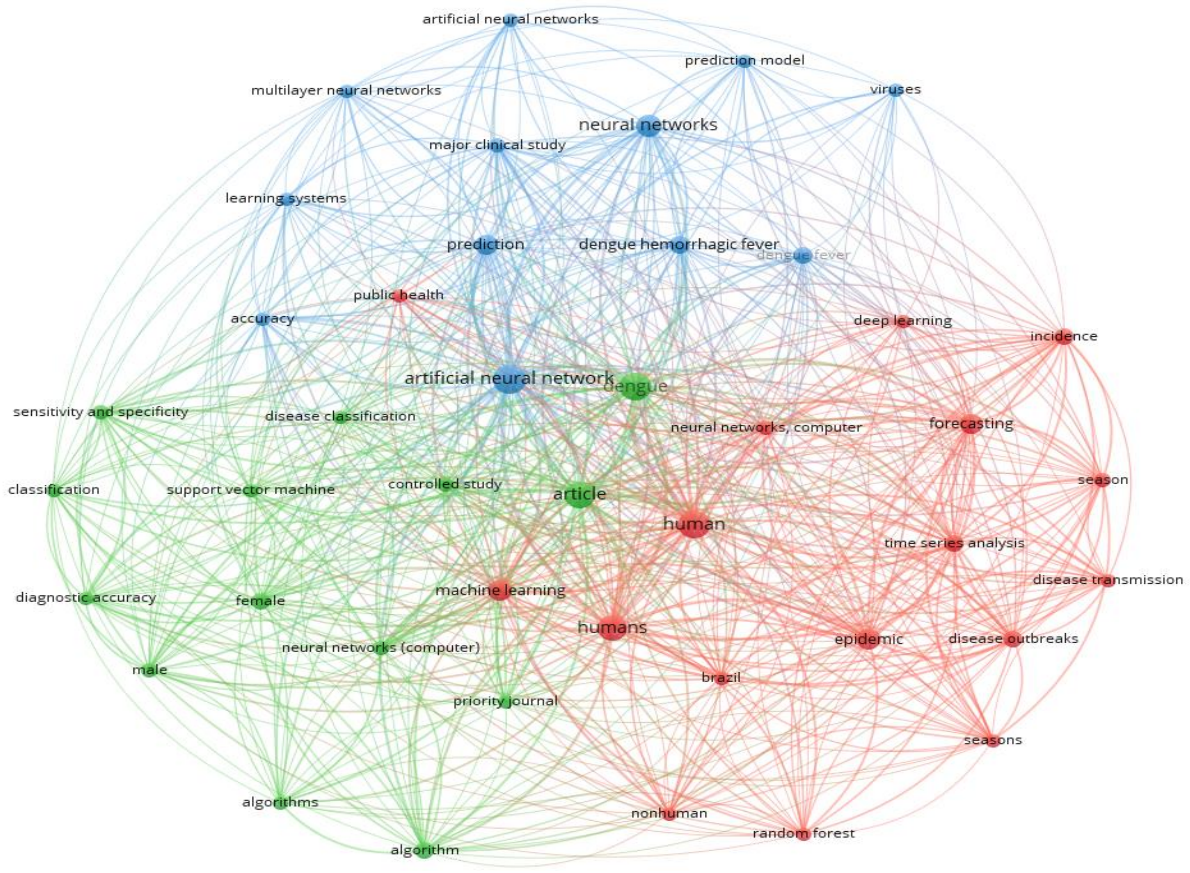
Figura 27. *Diagrama de red con key words Consulta 1.*



Nota: Información recopilada de Scopus

Para una segunda estancia (Consulta 2) y con la inclusión de los términos clave anteriormente mencionados, se realiza una reducción del número de documentos en la búsqueda, hasta llegar a un total de 100 de ellos, los cuales fueron publicados en una ventana de tiempo de 17 años, del 2005 hasta el año 2022. De nuevo se realiza un análisis de palabras clave con los 30 primeros documentos con mayor número de citaciones, generando así un diagrama de red (ver Figura 28). En esta nueva consulta ya se puede apreciar la inclusión de términos claves como “forecasting”, “Prediction”, “Deep Learning”, “Machine Learning”, “Artificial neural networks” entre otros términos de gran relevancia, empero, sigue siendo notable la ausencia de términos relevantes como el tipo de red neuronal a utilizar (RNN, LSTM Y GRU), por ello se procede a generar otro refinamiento de la ecuación de búsqueda en cuestión.

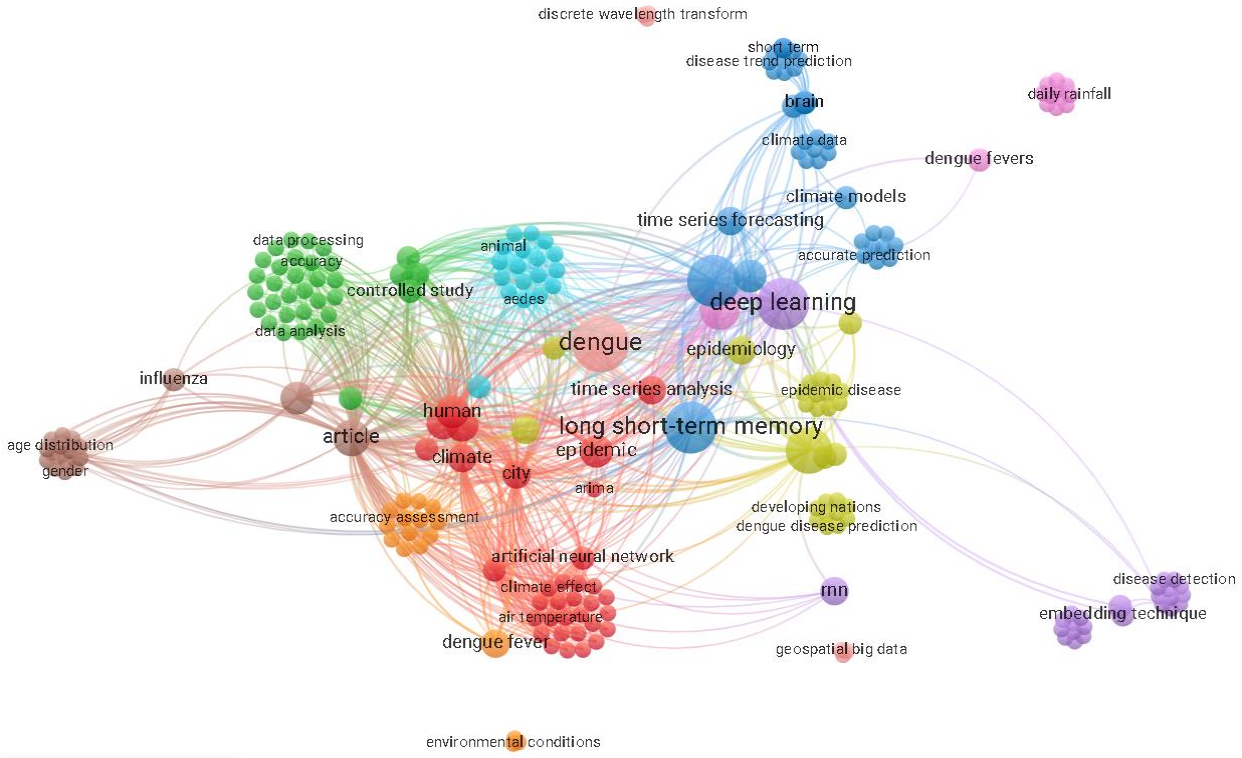
Figura 28. *Diagrama de red con key words Consulta 2.*



Nota: Información recopilada de Scopus. Fuente propia

En este punto de la investigación se llega a la ecuación de búsqueda final, agregando como términos de búsqueda las denominaciones de las redes neuronales a estudiar, en esta ecuación de búsqueda pudimos obtener los ya mencionados anteriormente 22 documentos. De nuevo, se hace un análisis con un diagrama de red, para observar la relación entre las palabras claves de la búsqueda. (Ver Figura 29). Como se puede observar en la Figura 29, las palabras claves están agrupadas en 9 clústeres, dentro de los cuales podemos identificar la presencia de los términos claves sugeridos en aproximaciones literarias realizadas con anterioridad.

Figura 29. *Diagrama de red con key words Consulta 3.*



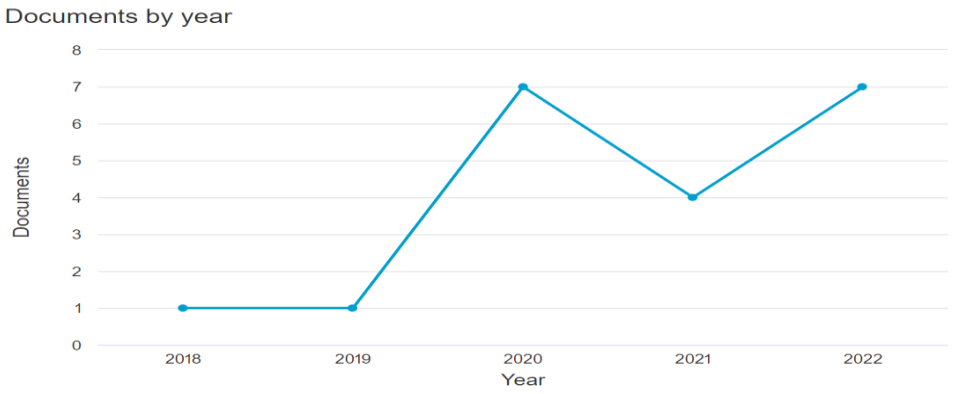
Nota: Información recopilada de Scopus

Dado el pequeño número de documentos obtenidos de la ecuación de búsqueda, se procede a realizar un análisis de todos y cada uno de ellos, independientemente de su campo de acción o de la cantidad de citaciones por artículo, denotando ciertas consideraciones.

6.1.1 Producción científica anual

La producción científica entre los años 2018-2022 referente a la temática en estudio se puede reflejar en el siguiente gráfico:

Figura 30. Producción científica anual.



Nota: Información recopilada de Scopus.

La producción científica anual de documentos referentes a modelos predictivos de enfermedades arbovirales estuvo en aumento entre los años 2019 a 2020, de hecho, el aumento fue de un 600%, con un pico en el año 2020 con un 35% del total de los documentos de investigación analizados (Ver figura 31), siendo junto con el año 2022, los años con mayor producción científica en esta área.

Es interesante analizar el "posible" impacto que generó la pandemia del Covid-19 en el desarrollo de documentos científicos en esta área de estudio, pues los estudios en este campo disminuyeron para el 2021, pasando a representar sólo el 20% del total general de la producción científica (Ver figura 31), lo que se traduce en una reducción del 42,85% de la producción científica de un año a otro, no obstante, esta producción se recuperó en igual proporción para el siguiente año, año 2022, lo cual puede llegar a sugerir la existencia de cierta influencia en la reducción de los trabajos en esta área a causa de la pandemia de COVID 19.

Figura 31. *Producción científica año a año segmentada en artículos o papers de conferencias.*

Year	Article	Conference Paper	Total
2018		5,00%	5,00%
2019		5,00%	5,00%
2020	25,00%	10,00%	35,00%
2021	10,00%	10,00%	20,00%
2022	30,00%	5,00%	35,00%
Total	65,00%	35,00%	100,00%

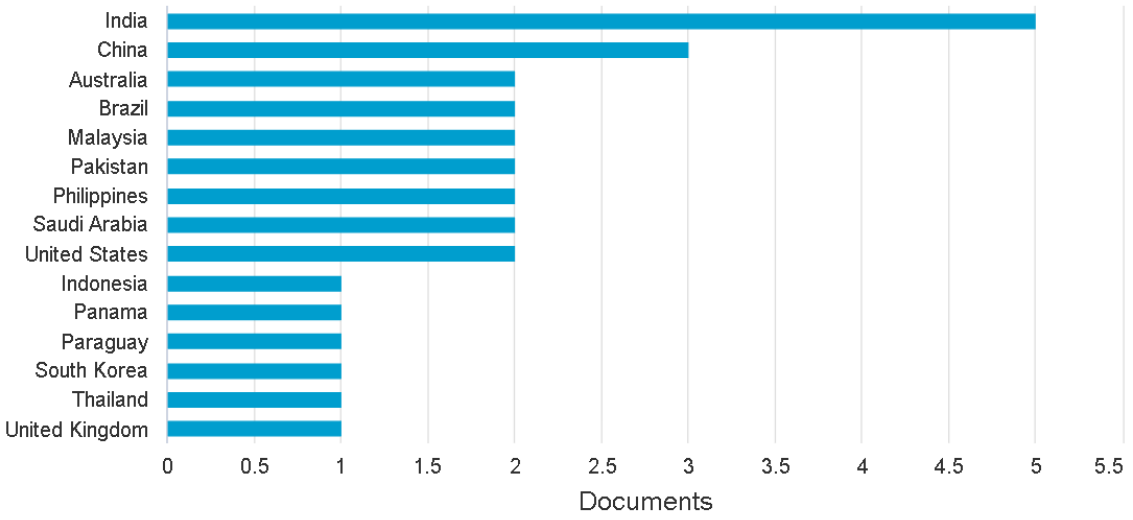
Nota: Información recopilada de Scopus y analizada en Power BI.

Respecto a la naturaleza de los documentos, podemos notar que estos se clasifican en artículos científicos y papers de conferencias, representando el 65% y 35% de los documentos, respectivamente (Ver figura 31), donde podemos apreciar que, los artículos producidos fueron generados a partir del año 2020

6.1.2 Producción científica por país

Con el fin de analizar los aportes realizados por cada uno de los países que participaron en la elaboración de estos documentos, se procede a generar dos visualizaciones, un diagrama de barras, donde está contenido cada uno de los países participantes y la cantidad de documentos elaborados o co- elaborados por cada uno de ellos (Ver figura 32) y un mapamundi (Ver Figura 33), donde se resalta dicha producción científica por país, presentando una segmentación por colores, acorde a la cantidad de material elaborado o co- elaborado por cada país.

Figura 32. *Producción científica por país.*



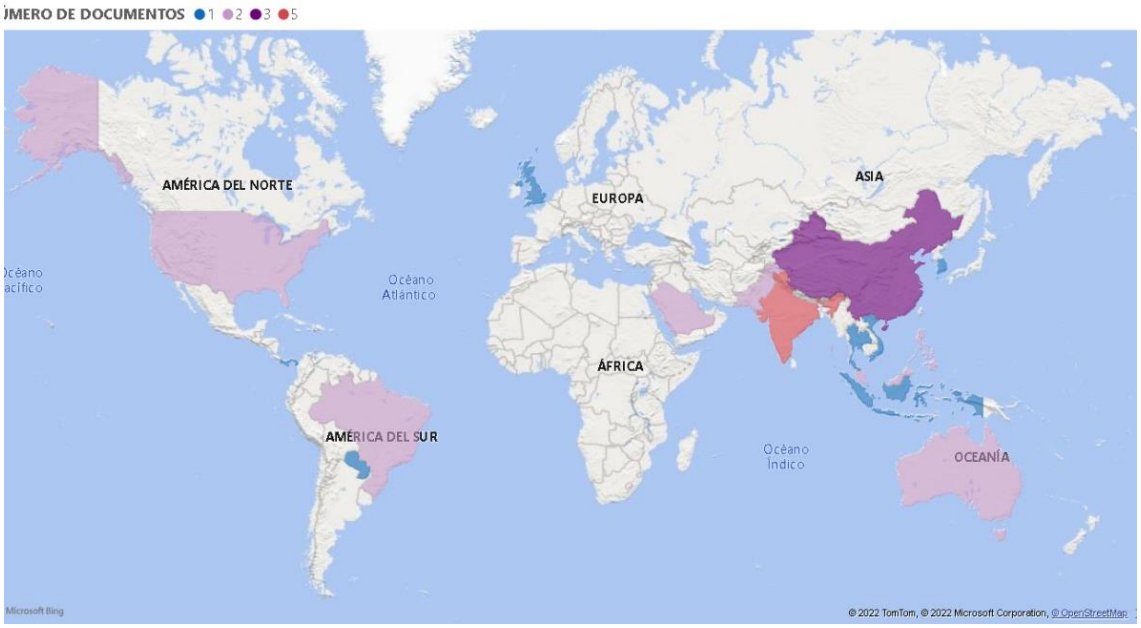
Nota: Información recopilada de Scopus

En este apartado, podemos notar, el gran aporte que genera la India en materia de investigación en este campo científico, ya que 5 documentos, un 25% del total general producido, fueron elaborados en la India o se contó con la colaboración de pares e instituciones de este país (Existen documentos con participación de investigadores e/o instituciones conformadas por equipos de diferentes países), seguido de China, con 3 documentos (15% del total general producido), Australia, Brasil, Malasia, Pakistán, Filipinas, Arabia Saudita , cada uno de ellos con una participación en 2 documentos (10% del total general producido), cerrando así el top 9 de países con mayor contribución en esta área.

Al analizar la Figura 33, es notable el peso que tiene, en este campo de investigación, el continente asiático, puntualmente el sudeste asiático, donde se puede apreciar se aglomera el grueso de países que realizaron aportes en este campo de investigación. De hecho, es allí donde se ubican los países con mayor número de investigaciones en esta área, y donde según la OMS existe una considerable incidencia de contagios de dengue "…En otro estudio sobre la

prevalencia del dengue se estima que 3900 millones de personas están en riesgo de infección por los virus del dengue. Pese a que existe riesgo de infección en 129 países, el 70% de la carga real se concentra en Asia.” (OMS, 2022).

Figura 33. *Producción científica por país.*



Nota: Información recopilada de Scopus y analizada en Power BI

Al observar detalladamente la producción científica latinoamericana, podemos notar que es escasa, ya que trabajos de esta naturaleza sólo han sido realizados en Brasil, Panamá y Paraguay, una razón más para emular este tipo de estudios en Colombia, además de subrayar que la poca literatura hallada concerniente a estas investigaciones, enfocadas específicamente a los métodos de predicción, nos sugiere que es un campo de investigación con un gran margen de acción y mucho por investigar y desarrollar.

6.2 Revisión de literatura preliminar

En esta fase, se analizaron la totalidad de los 20 documentos hallados en la búsqueda de la herramienta Scopus, no obstante, la mayoría de los documentos hallados (17/20), sólo abordan como objeto de estudio las proyecciones realizadas sobre contagios de Dengue y, dado que, nuestra investigación se centra en la predicción de casos y brotes de Zika, Dengue y Chikunguña, se podría suponer que, la literatura hallada no es suficiente, sin embargo, y según (*4 I. Introducción,* n.d.)"…el mosquito Aedes (Principalmente el aegypti y albopictus especies), también transmite otras enfermedades virales graves como Zika, Chikungunya y fiebre amarilla…", teniendo así, como punto de partida para las proyecciones de Zika y Chikunguña las mismas variables y factores (Ambientales y demográficos) que inciden en los estudios realizados sobre el Dengue, razón por la cual y para el desarrollo de nuestra investigación, podemos basarnos en las directrices consignadas en los documentos hallados por nuestra búsqueda.

Las investigaciones halladas han sido abordadas de diversas maneras, por ejemplo, respecto al área geográfica de acción, existen investigaciones con segmentaciones territoriales variadas, desde análisis de diferentes ciudades de un país, como en el caso del documento "Forecast of Dengue Cases in 20 Chinese Cities Based" (Xu et al., 2020a), de regiones específicas de un país, como en "Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression",(Mussumeci & Codeço Coelho, 2020) donde se segmentó el alcance geográfico de la investigación al Estado de Rio de Janeiro, Brasil, hasta casos en los cuales se realizó el estudio sobre una ciudad en específico, por ejemplo en "How to Efficiently Predict Dengue Incidence in Kuala Lumpur"(Taylor's University (Subang Jaya et al., n.d.).

Respecto a los parámetros de los datos de entrada necesarios para la investigación, se halló que, en general, diversos autores suelen clasificarlos en dos grandes grupos, datos

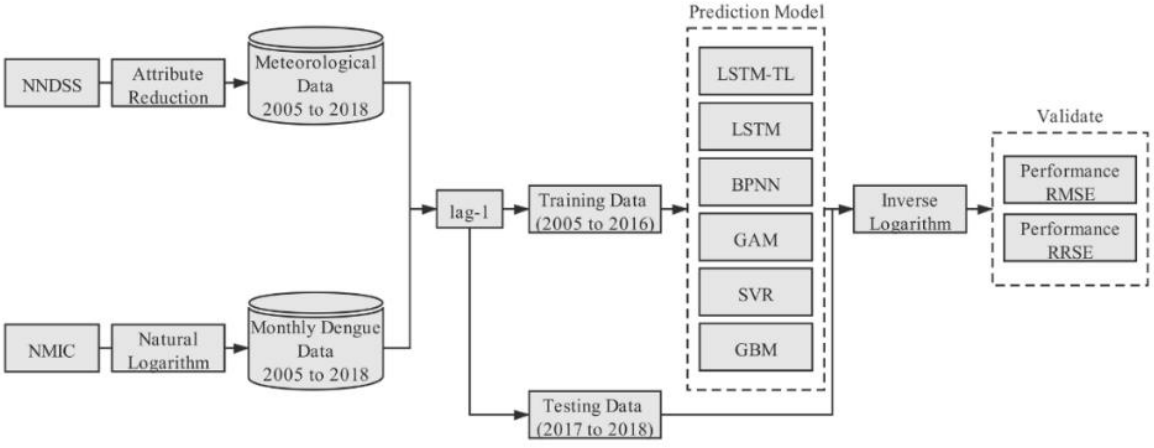
ambientales y datos epidemiológicos, como es el caso de, (Xu et al., 2020a), quien consideró los casos de registros de contagios mensuales y los registros de factores climáticos, mientras que, (SCAD College of Engineering and Technology & Institute of Electrical and Electronics Engineers, n.d.) caracteriza las variables de entrada en los mismos dos grandes grupos, pero, dentro de la data usada para la predicción incluye la contaminación del ambiente como variable intrínseca de los datos ambientales, por su parte, y con una categorización más detallada (Taylor's University (Subang Jaya et al., n.d.), dedica un apartado de la investigación a describir la naturaleza de cada uno de estos dos grupos, clasificándolos cómo "datos medioambientales, y datos epidemiológicos y socioeconómicos", detallando cada uno de los aspectos tenidos en cuenta dentro de cada grupo, por ejemplo, para las variables ambientales, especifica el uso de múltiples datos ambientales (Precipitaciones, temperatura ambiental, altitud, bioma, temperatura de la superficie terrestre, vegetación detectada remotamente, anomalías en la temperatura de la superficie del mar, índice de Oscilación del Sur, humedad, índice de población de mosquitos e índice de población de larvas) y para las variables epidemiológicas y socioeconómicas plantea un amplio uso de variables demográficas y socioeconómicas (Incidencia del dengue, datos del censo (población), límites administrativos, índice de pobreza, acceso a la electricidad, acceso al agua potable, índice de saneamiento, índice de salud infantil, índice de educación infantil, índice de calidad de vida infantil, movimiento de población humana y control de vectores).

El primer artículo analizado exhaustivamente fue (Xu et al., 2020a), se realizó una comparación entre los rendimientos de diferentes métodos de predicción GAM (Modelo Aditivo Generalizado), GBM (Máquina de Aumento de Gradiente), BPNN (Red Neuronal de Retro propagación), SVR (Regresión de Vectores de Soporte) y redes neuronales LSTM, frente al comportamiento real de contagios durante 24 meses de 2017 a 2018 en 20 ciudades chinas. Para

ello, se usó una data que fue divida en dos grandes grupos, datos ambientales y datos epidemiológicos, respecto a estos últimos, se incluyeron características demográficas básicas (género, edad, nacionalidad, y dirección de residencia), además de incluir un apartado con el momento de los eventos relacionados con la enfermedad (fecha de inicio de la enfermedad, diagnóstico y en algunos casos, fecha de la muerte). Para los datos meteorológicos, fueron tenidas en cuenta 15 variables, 5 de las cuales fueron eliminadas de la data usando filtrado de alta correlación y filtrado de baja varianza, dejando como resultado en la data 10 variables meteorológicas (presión máxima, presión media, presión media del agua, temperatura mínima del aire, temperatura máxima del aire, promedio de la temperatura máxima diaria, promedio de precipitación diaria, número de días con lluvia y promedio de humedad relativa).

Se construye una red neuronal tipo LSTM como modelo de predicción, modelo sobre el cual, se contrastó su comportamiento frente a otros modelos de predicción, llevando a cabo un proceso tal cual como se presenta en el marco general del estudio (Ver Figura 34).

Figura 34. *Marco general de estudio usado por* (Xu et al., 2020a)



Flujo de trabajo resumido para la construcción del modelo de pronóstico basado en LSTM para casos de dengue y su comparación con otros modelos candidatos. NNDSS: Sistema Nacional de Vigilancia de Enfermedades de Declaración Obligatoria; NMIC: Centro Nacional de Información Meteorológica; BPNN: Red neuronal de retropropagación; GAM: Modelo Aditivo Generalizado; SVR: Regresión de vectores de soporte; GBM: Máquina potenciadora de gradientes.

Nota: Adaptado por Figure 2 [Imagen], por Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method, 2020, Environmental Research and Public Health (IJERPH | Free Full-Text | Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method ((Xu et al., 2020b)).

Los resultados obtenidos evidencian una tendencia al aumento de los casos en las ciudades objeto de estudio, además, el modelo LSTM, por entrenamiento TL (Transfer Learning), tiene un RMSE más bajo en la mayoría de las ciudades objeto de estudio, de hecho, se determinó que las reducciones previstas en el RMSE (Error Medio Cuadrado), serían considerables (34,6%, 47,4%, 30,3%, 26,9% y 32,5%) para un grupo de ciudades pertenecientes a la provincia de Guangzhou, China. En efecto y citando textualmente a los autores (Xu et al., 2020a), "…De acuerdo con la precisión predictiva para los dos períodos de predicción, el modelo LSTM por entrenamiento TL tiene un RMSE más bajo en la mayoría de las ciudades que el BPNN modelo, modelo GAM, modelo SVR y modelos GBM. Nuestro método LSTM redujo las predicciones de RMSE promedio en un 12,99 % a 24,91 % en comparación con los casos de dengue estimados de otros modelos publicados anteriormente, y las predicciones de RMSE

promedio en el período del brote disminuyeron en un 15,09 % a 26,82 %. En particular, el método LSTM redujo las predicciones de RMSE en un 44,48 % a un 75,56 % en Guangzhou, que tiene la mayor incidencia de dengue en China, y las predicciones de RMSE en el período del brote se redujeron en un 44,75 % a un 75,7 %."

Por otra parte, en el artículo de la conferencia "How to Efficiently Predict Dengue Incidence in Kuala Lumpur" (Taylor's University (Subang Jaya et al., n.d.), se resaltó la importancia de incorporar el uso de Machine Learneing (ML) y Deep Learning (DL) en los métodos de evaluación y predicción de los sistemas de salud en Kuala Lumpur, con el fin de mitigar los daños causados por este tipo de enfermedades "El dengue ha sido una de las principales fuentes para la hospitalización y la muerte (en particular para los niños pequeños) y sigue siendo un importante problema de salud pública para los endémicos países de todo el mundo"(Taylor's University (Subang Jaya et al., n.d.) , para así generar medidas de respuesta previas ante eventuales brotes. En este artículo se evaluó el comportamiento de tres diferentes enfoques de ML y DL, GA_RNN (Genetic Algorithm Based Recurrent Neural Networks), LR (Linear Regression) y DT (Decision Tree).

Para la determinación de la data usada, los autores(Taylor's University (Subang Jaya et al., n.d.), se basaron en investigaciones previas, clasificaron la data en datos ambientales y datos epidemiológicos & socioeconómicos. Respecto a la recopilación y tratamiento de datos, partieron de generar una compilación únicamente de fuentes confiables, como el ayuntamiento de la ciudad de Kuala Lumpur, el Departamento Meteorológico de Malasia y datos para EVI (Índice de vegetación mejorado) de la NASA. En el caso de registro diario de contagios en la ciudad

entre los años 2002-2012, se encontró que, como suele suceder, la data era demasiado ruidosa (un dato curioso hallado durante nuestra investigación, ya que a pesar de ser un problema bastante recurrente en trabajos relacionados con grandes volúmenes de datos, no fue mencionado de forma significativa en los documentos objeto de estudio) razón por la cual, tuvieron que recurrir a usar los promedios de estos datos durante un período de tiempo como entrada. En cuanto a los datos ambientales, se presentó otro típico problema de esta área, los datos faltantes o perdidos, por ello los puntos de datos faltantes se interfirieron mediante interpolación de spline cúbica, aparte de tener que generar promedios para determinar los datos de entrada de variables ambientales como precipitaciones diarias, temperatura diaria, humedad diaria y velocidad diaria del viento, ya que estos fueron obtenidos de diferentes estaciones ambientales a lo largo y ancho de la ciudad.

Los modelos de predicción usados por (Taylor's University (Subang Jaya et al., n.d.) fueron algoritmos de DL , por un lado tenemos redes neuronales recurrentes, las cuales, está comprobado tienen un mejor desempeño comparado con métodos de análisis estadístico, como ARIMA (autoregressive integrated moving average), por ello, y dado el buen desempeño de las LSTM, se propuso generar un proceso de optimización de las mismas por medio del uso de GA(Genetic Algorithm), obteniendo así una Red neuronal recurrente mejorada por algoritmo genético, denominada por sus siglas en inglés como GA_RNN. Respecto al segundo método de predicción usado, se empleó regresión lineal simple (LR), usando 3 parámetros o variables predictoras, índice de vegetación mejorada (EVI), precipitaciones diarias y temperatura. Se denotó que la LR es una buena opción ya que puede tratar con variables observables irregulares y produce un modelo con varianzas bajas. Como tercer y último método de predicción, se usó un

árbol de decisión (DT), ya que, dada la viabilidad del uso de un enfoque basado en reglas para predecir los rangos altos y bajos de incidencia del dengue, se aproximó un modelo de DT con el fin de generar una aproximación de los casos de fiebre de dengue (DF), en su funcionamiento el DT toma los datos históricos de casos de DF y crea pautas para pronosticar en qué rango estarán los contagios diarios a futuro.

Respecto a los resultados experimentales, se tomó como conjunto de datos de entrenamiento, la data de 2002 a 2010 y los datos de 2011 a 2012 fueron tomados como conjunto de datos de prueba. Por último, los resultados obtenidos, demostraron que, por ejemplo, las predicciones generadas por el modelo GA_RNN generan un mejor desempeño que el modelo de LR cuando se compara con el número real de incidentes de dengue del conjunto de datos de entrenamiento, citando textualmente tenemos "…Durante el entrenamiento, los resultados predichos por GA_RNN coincidieron estrechamente con los valores históricos de entrenamiento, mientras que LR produjo valores que probablemente fueron más bajos que los valores históricos."(Taylor's University (Subang Jaya et al., n.d.), ahora, al comparar los tres modelos de predicción propuestos, se propuso como medida de eficiencia los Errores Absolutos Medios (MAES) y los Errores Cuadrados Medios (RMSEs), dando como resultado que, el modelo con el mejor desempeño fue GA_RNN, pues presentó un MAE = 10.95 y un RMSE = 13.06, frente a un DT con el peor desempeño (MAE = 25.32 y RMSE = 34.86), con valores casi 3 veces más altos que el modelo GA_RNN, generando como resultado, que las predicciones de GA_RNN son más útiles para tomar medidas preventivas para reducir el impacto de la propagación del virus del dengue en Kuala Lumpur.

Por otra parte, y usando un enfoque diferente, aparece el autor Amin Samina, quien surge como co- autor en los documentos (Amin, Irfan Uddin, et al., 2020) y (Amin, Uddin, et al., 2020), donde usaron una data no estructurada, en este caso se apoyaron en datos extraídos de las redes sociales, puntualmente, de la plataforma Twitter, de allí extrajeron información en forma de sentimientos, pensamientos u opiniones, simplemente basándose en publicaciones disponibles en esta red social. Para ello, se apoyaron el uso de redes neuronales LSTM y técnicas de incrustación de palabras Word2Vec con Skip-gram (SG) y Word2Vec con Continuous-bag-ofwords (CBOW). Las razones por las cuáles se recomienda emplear este tipo de datos se fundamentan en el constante crecimiento de interacciones y expresión de sentimientos casi que en tiempo real, de parte de diferentes usuarios alrededor del mundo, lo cual suele ser denominado como "huellas" que son dejadas por los internautas, dichas huellas, han venido siendo utilizada para apoyar diferentes predicciones en distintos campos de acción, por ejemplo, en el sector económico "A principios de año, el Banco de Inglaterra anunciaba la creación de un equipo especial de analistas para estudiar los comportamientos de los ciudadanos en las redes sociales con el objetivo de predecir la economía" (BBVA, 2017), ahora está siendo utilizada con el fin de pronosticar brotes de enfermedades arbovirales, de hecho, las redes sociales podrían ser usadas de manera eficiente para tratar personas infectadas por diferentes enfermedades y mitigar los impactos de dicha enfermedad en la salud pública de una región, mejorando así la detección temprana de diferentes brotes, tal como lo propone (Paul et al., n.d.). De hecho, al comparar este enfoque con el enfoque tradicional (se manejan datos epidemiológicos y datos ambientales, denominados estructurados, por lo general, es una data institucional) se puede decir que este enfoque es más eficiente en términos de disponibilidad de la información. "Los enfoques tradicionales para detectar brotes epidémicos son cuando las personas fueron diagnosticadas con

una enfermedad que informan al centro de salud local, que luego puede notificar a los proveedores de atención médica relevantes para responder y brindar servicios para rastrear esa epidemia. Este proceso a menudo toma semanas antes de que se registre la información y, en ciertas situaciones, se pierden vidas valiosas antes de que se tomen las medidas adecuadas." (Amin, Irfan Uddin, et al., 2020).

Respecto al uso de LSTM con Word2Vec, se estaría combinando el uso de minería de datos y técnicas de inteligencia artificial, empleando NLP (Natural Language Processing) y DL (Deep Learning). Respecto a la captación de los datos, el autor (Amin, Irfan Uddin, et al., 2020) hace referencia al uso de la API (Application Programming Interface) de Twitter, la cual es de uso libre. Tweets que contenían palabras como "dengue", "fiebre de dengue", "gripa" o "influenza" fueron almacenadas en un repositorio de datos, junto con datos como la información del usuario del tweet, fecha del tweet, ubicación, entre otros, utilizando métodos de DL y ML, tal como lo recomienda (Hernandez-Suarez et al., 2019). Luego de ello (Amin, Uddin, et al., 2020) realiza los característicos pasos de preprocesamiento de la data mediante técnicas de preprocesamiento NLP (Stop Words Removal, Special Characters, Stemming y Tokenization), con el fin de normalizar los datos a utilizar, para luego entrar en la fase de extracción de características utilizando análisis semántico con técnicas de incrustación de palabras, para así convertir data tipo texto en datos de tipo numérico aptos para la computación de redes neuronales tipo ANN o RNN (Lipton et al., 2015). El siguiente paso del desarrollo fue la fase de entrenamiento, validación y división de pruebas, aquí(Amin, Uddin, et al., 2020) usó el 80% de la data obtenida con el fin de entrenar al modelo (dicho porcentaje puede diferir según los criterios de rendimiento del modelo propuesto), con el fin de utilizar el enfoque de validación

cruzada K-fold y el 20% restante se usa para testear el modelo (Fushiki, 2011), todo ello, con el fin de evitar el desajuste y sobreajuste del modelo (todo este proceso puede apreciarse de mejor manera al observar la Figura 35).

Por último, “LSTM es una variante modificada de RNN, también conocida como red neuronal recurrente elegante, propuesta por (Hochreiter & Schmidhuber, 1997), que tiene un estado de memoria que tiene la capacidad de recordar información y aprender dependencias a largo plazo durante largos períodos” (Amin, Uddin, et al., 2020).

Figura 35. *Metodología propuesta usada por* (Amin, Uddin, et al., 2020)

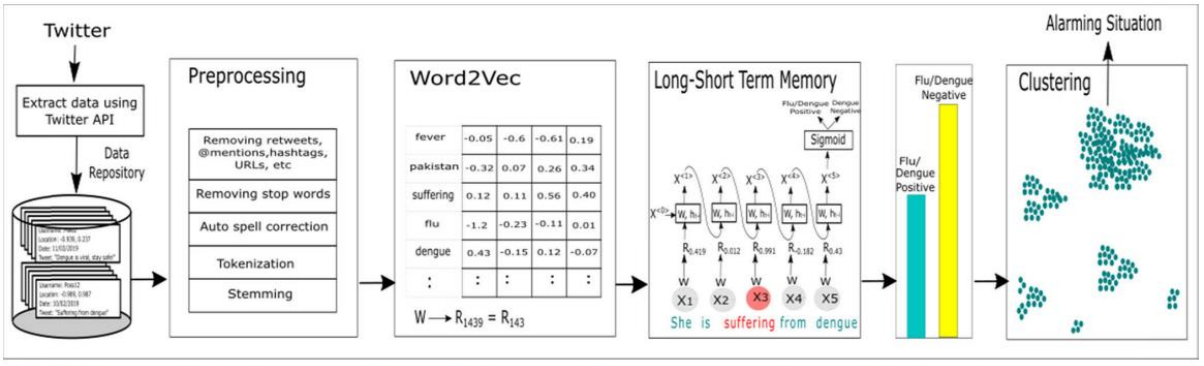


FIGURE 1. **A proposed methodology for identification of disease infected people in tweets and clustering based on regions.**

Nota. Adaptado por Figure 1 [Imagen], por Detecting Dengue/Flu Infections Based on Tweets Using LSTM and Word Embedding, 2020, IEEE Access (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9223762)).

6.3 Identificación de variables

Esta sección de la revisión bibliográfica se examina en detalle las variables que han sido estudiadas dentro de los temas de interés por los distintos autores. A través de un análisis preliminar de la literatura, se identifican las variables directamente relacionadas con el estudio de la predicción de enfermedades arbovirales por medio de redes neuronales y aprendizaje profundo. Megha Chovatiya, Anushka Dhameliya y colaboradores (2019) , quienes decidieron extraer los datos de varios sitios web gubernamentales y centros de salud. Crearon un modelo que aprende de estos datos para predecir la probabilidad de una erupción en condiciones climáticas similares en el futuro. Los posibles resultados de los brotes se muestran utilizando los mapas de calor de Google para resaltar las áreas donde es probable que ocurra el dengue. Los datos utilizados para el sistema incluyen datos meteorológicos (temperatura, presión, humedad) recogidos del sitio web TimeandDate.com, datos de precipitaciones del departamento meteorológico, índice de calidad del aire del Control Central de Contaminación Junta, número de casos de dengue registrados desde la Secretaría de Salud Departamento, Jodhpur. Los datos utilizados son específicos de la ciudad Jodhpur. A continuación, estos datos se han añadido junto con el mes y año como atributo común.

Presentado lo anterior, Jiucheng Xu, Keqiang Xu y colaboradores (2020) en su interés de pronosticar los casos de dengue en 20 ciudades chinas basado en el método de aprendizaje profundo, proponen desarrollar un modelo de pronóstico oportuno y preciso del dengue basado en redes neuronales recurrentes de memoria a corto plazo (LSTM) considerando solo los casos mensuales de dengue y los factores climáticos. Las variables de los datos meteorológicos de estas ciudades se obtuvieron del Centro Nacional de Información Meteorológica (NMIC). Un total de 15 variables meteorológicas velocidad extrema del viento, velocidad máxima del viento,

velocidad media del viento, presión mínima, presión máxima, presión media, presión media del agua, temperatura mínima del aire, temperatura máxima del aire, temperatura media del aire, promedio de la temperatura máxima diaria temperatura, promedio de temperatura diaria más baja, promedio de precipitación diaria, número de días con lluvia y promedio de humedad relativa, se mantuvo sin valores perdidos en los datos sin procesar. El modelo LSTM podría sobre ajustarse si todas las variables se utilizan para el entrenamiento del modelo de red neuronal. El sobreajuste mejora el rendimiento del modelo en el conjunto de entrenamiento; sin embargo, funciona mal en el conjunto de prueba, lo que indica que la capacidad de generalización del modelo es débil. Por lo tanto, la selección de atributos se utilizó para evitar el sobreajuste y eliminar atributos redundantes.

Vicente Navarro Valencia, Yamilka Díaz y colaboradores (2021), tomaron los datos de incidencia de dengue en el Instituto de Salud Memorial Gorgas (ICGES) y el Ministerio de Salud (MINSA). El MINSA estableció un programa de recopilación de datos en 1988. La definición de caso de dengue sigue las pautas de la Organización Mundial de la Salud (OMS). De 1993 a 2011, la clasificación de 1997 fue Dengue (DF), dengue con fiebre hemorrágica (FHD) y dengue shock (DSS) se aplicaron en 3 de 18 casos. A partir de 2012 se aplican las definiciones de dengue con o sin síntomas y dengue grave de la OMS 2009. Con sede en la ciudad de Panamá, el CGES es el lugar principal para realizar pruebas de laboratorio de DENV de pacientes hospitalizados y ambulatorios de los hospitales regionales de Panamá que tenían sospecha clínica de dengue. El conjunto de datos completo contiene variables climáticas desde 1995 hasta el presente, pero organizadas por período de recolección como incidencia de dengue. Se eligió esta estación porque se encontraba a una altura similar a la elevación promedio del área de estudio,

aproximadamente 9 metros sobre el nivel del mar. De esta estación se seleccionaron las series de tiempo: temperatura, precipitación y humedad relativa. Los puntos de datos en blanco se imputaron utilizando el modelo de aprendizaje profundo multivariante LSTM y la estación cercana de Barro Colorado.

Ángelus Ronald Doni y Graciasappan Sasipraba (2020), presentaron un modelo para predecir casos de dengue utilizando redes neuronales artificiales (ARN), los datos proporcionados son población, precipitación, temperatura y humedad. Los datos históricos del dengue también se comparten en función de la entrada y los datos procesados utilizando una capa oculta y el resultado esperado es un pronóstico de casos y muertes por dengue. Los modelos anteriores se implementaron utilizando algoritmos LSTM, SVM, XGBoost, BPNN, GAM y Random Forest. Se compara el rendimiento del algoritmo. Los datos recopilados de diversas fuentes se procesan previamente. Entre las funciones disponibles, se tienen en cuenta las características de influencia aplicando el algoritmo XGBoost. Los datos se dividen en tres categorías: entrenamiento, prueba y validación. Se construye y entrena un modelo en el conjunto de datos de entrenamiento, y la precisión del modelo propuesto se prueba usando el conjunto de datos de prueba. Luego, el modelo se evalúa en el conjunto de datos de validación. Luego aplica las medidas necesarias para evaluar los resultados y generar un motor de recomendación.

A partir de la revisión realizada se definen las variables de incidencia a nivel Colombia para este trabajo de investigación, las cuales tienen como objetivo medir el comportamiento de los casos de Dengue, Zika y Chikunguña con el fin de predecir futuros brotes para la prevención

y control de salud. Las variables como temperatura ambiente, presión, humedad, número de casos registrados, área de ocurrencia, año, semana epidemiológica, edad, municipio y departamento donde reside el infectado, género (sexo), índice de pobreza. Las anteriores variables se encuentran descritas en el apéndice y cabe resaltar que pueden estar sujetas a modificaciones, puesto que al realizar el proceso de la implementación de algoritmos se puede considerar incluir o eliminar algunas de ellas.

7. Recolección, preprocesamiento y análisis de datos

El preprocesamiento de datos es una etapa esencial del proceso de descubrimiento de información o KDD (Knowledge Discovery in Databases, en inglés). Esta etapa se encarga de la limpieza de datos, su integración, transformación y reducción para la siguiente fase de minería de datos. (García et al., n.d.)

Para llevar a cabo una adecuada preparación de los datos, es necesario efectuar actividades como completar los valores faltantes en las celdas, analizar y eliminar los datos que puedan introducir ruido en la información recopilada, revisar y corregir las inconsistencias que puedan dificultar el manejo adecuado de los datos, y ajustar las observaciones que puedan generar redundancias.

Con el propósito de tener los datos a punto para la elaboración de los modelos de predicción de Machine Learning se procede a realizar la etapa de preprocesamiento de datos; limpieza y transformación. Adicionalmente, en este capítulo se lleva a cabo un primer

acercamiento a los patrones que presenta la predicción de las enfermades arbovirales en Colombia.

A continuación, se describen las fases implementadas:

7.1 Recolección de datos

En este trabajo de investigación se analizan los microdatos brindados por SISPRO que es el sistema integrado de información de la protección social del gobierno de Colombia, una base de información con respecto a datos y sistemas de información del sector sobre oferta y demanda de servicios de salud, calidad de los servicios, aseguramiento, financiamiento y promoción social, lo cual incluye una serie de datos de carácter demográfico, necesarios para el estudio en cuestión, tal como lo propone Jiucheng Xu y Keqiang Xu en su investigación Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method (Xu et al., 2020a)

Respecto a los datos meteorológicos usados(Taylor's University (Subang Jaya et al., n.d.) expuso "Además de los dos factores de uso frecuente (temperatura media diaria y precipitación diaria), también tuvimos en cuenta el índice de vegetación mejorado (EVI), la humedad y la velocidad del viento como factores de entrada a nuestro modelos de predicción", por ello, se hizo uso de la data del satélite ArcGIS World Geocoder de donde se obtuvo la información sobre el comportamiento de las variables meteorológicas correspondientes al periodo establecido para la investigación. Con el fin de hallar las poblaciones estimadas, año a año, de cada una de las localidades objeto de estudio, se usó al DANE como fuente de información, ya que, por ejemplo, para (Taylor's University (Subang Jaya et al., n.d.) fue un factor relevante usado durante en el

trabajo de investigación “How to Efficiently Predict Dengue Incidence in Kuala Lumpur”. Por último, y en base a la información disponible, se eligió la ventana de tiempo de recolección de datos del año 2007 al 2021.

Otra de las decisiones tomadas con el fin de facilitar el proceso de recolección de los datos, fue la segmentación de los municipios / ciudades que serían abordados en el estudio, ya que el proceso de la captación de la data del total de los municipios de Colombia resulto muy complejo “…aproximadamente 1120 municipios, incluyendo los 10 Distritos que se cuentan como Municipios.” (CEPAL, n.d.), pues esto implicaría recolectar, por ejemplo, datos meteorológicos de cada variable (12 variables meteorológicas elegidas inicialmente), mes a mes, durante 15 años (2007 – 2021) para los ya mencionados 1120 municipios (1120 Municipios*12 Variables* 15 Años*12 Meses de cada año = 2’419’200 Consultas), además de identificar en qué meses de cada año, no se presentan contagios en un determinado municipio, para así no tenerlo en cuenta a la hora de la consolidación de la data de ese mes. Dado lo anterior, se tomó como criterio de selección, el número de contagios totales (de Dengue, Zika y Chikunguña) durante los 15 años de análisis de datos, determinando como umbral mínimo de selección a aquellas ubicaciones (Municipios y ciudades) con un número mayor a los 10.000 contagios por … (Ver Figura 36)

Figura 36. *24 ciudades de Colombia con mayor número de contagios de Dengue, zika y chikunguña del 2007 al 2021.*

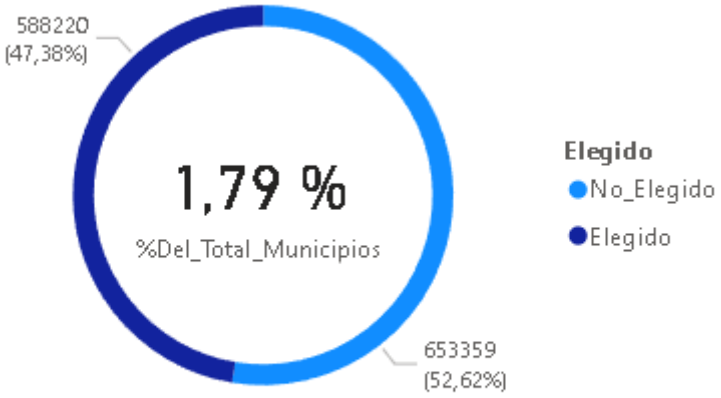
MUNICIPIO	Suma de TOTAL_GENERAL
Cali	126282
Medellín	50385
Ibagué	43793
Cúcuta	43403
Bucaramanga	42725
Villavicencio	39930
Barranquilla	33368
Neiva	29879
Floridablanca	23415
Cartagena	19956
Sincelejo	17021
Valledupar	15931
Yopal	15024
Palmira	15000
Armenia	14284
Pereira	13191
Montería	12396
Girón	11242
Santa Marta	10957
Soledad	10038
Piedecuesta	9725
Acacías	9405
Barrancabermeja	8666
Espinal	8623
Total	**1241579**

Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

Dado lo expuesto previamente, fueron elegidas 20 localizaciones (municipios y ciudades), las cuales ordenadas de forma decreciente en referencia al número de contagios fueron: Cali, Medellín, Ibagué, Cúcuta, Bucaramanga, Villavicencio, Barranquilla, Neiva, Floridablanca, Cartagena, Sincelejo, Valledupar, Yopal, Palmira, Armenia, Pereira, Montería, Girón, Santa Marta y Soledad. Estas 20 localizaciones (Municipios y ciudades), representan el 1,79% del total de las localizaciones de Colombia, pero a nivel de contagios registran 588.220 contagios, lo cual se traduce en un 47,38% de los contagios registrados (Ver Figura 37)

Figura 37. *Comparativa porcentual entre los municipios y sus contagios objeto de estudio con*

respecto al total de municipios y al total de contagios.



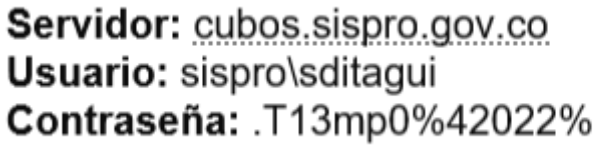
Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI

Una vez segmentados las localizaciones que serían objeto de estudio, se procedió con el preprocesamiento de datos, el cual fue abordado bajo las siguientes actividades:

7.1.1 Extracción de datos de la base de datos de SISPRO

Para tener acceso a la base de datos de SISPRO, se debe seguir un conducto regular, el cual incluye una serie de capacitaciones brindadas por dicha entidad, capacitaciones en las cuales, se recalca la importancia del manejo de los datos, aspectos legales a tener en cuenta y una breve explicación del manejo de estos, junto con la posterior asignación de una clave y usuario de ingreso a la base de datos.

Figura 38. *Datos para acceder a la base de datos SISPRO en Excel.*



Nota: Imagen tomada de las credenciales de acceso enviadas por SISPRO

Una vez se tuvieron las credenciales de acceso, se procedió a realizar la respectiva extracción de la data útil para nuestro estudio (Número de casos, área de ocurrencia, año de contagio, el tipo de evento, dengue, zika, chikunguña, edad y género del contagiado, departamento y municipio de registro del contagio).

7.1.2 Extracción y selección de los datos meteorológicos

Respecto a la elección de este conjunto de datos, nos decantamos por el uso de los datos brindados por ArcGIS World Geocoder, ya que en comparación con los datos brindados por el IDEAM esta fuente (ArcGIS World Geocoder) nos brinda una mejor calidad de la data, sin el considerable número de datos faltantes y valores atípicos hallados en las bases de datos del IDEAM, que aparentemente en algunos casos, no concordaban con la realidad. En cambio, los datos de ArcGIS World Geocoder se presentaban de manera ordenada, en bloque (agrupando diferentes características meteorológicas en un solo documento) y con arreglos estadísticos ya realizados, como los promedios de registros diarios segmentados mes a mes. Todos ellos dados en formato .csv, tratados a brevedad con el uso de Microsoft Excel para Microsoft 365 MSO (versión 2303 compilación 16.0.16227.20202) de 64 bits.

7.1.3 Extracción de los datos de proyecciones de población de las localidades seleccionadas

Debido a la relevancia de esta variable en investigaciones realizadas previamente, se procede a hacer uso de las bases de datos del Departamento Administrativo Nacional de Estadísticas (DANE), puntualmente de los censos y respectivas proyecciones de población realizadas los años 2005 (y sus respectivas proyecciones hasta el año 2017) y año 2018 (y sus respectivas proyecciones hasta el año 2021).

7.2 Limpieza y eliminación de datos

Eliminar las variables innecesarias es una parte esencial del proceso de limpieza de datos, es fundamental para asegurar la integridad de la información. Una buena base de datos facilita la toma de decisiones acertadas y la obtención de conclusiones pertinentes. En la etapa inicial, se procederá a eliminar manualmente de la base de datos aquellas variables que no sean relevantes para el estudio en cuestión, conservando únicamente aquellas que se utilizarán.

7.2.1 Limpieza y eliminación de los datos de SISPRO.

Dado que la información venía presentada en forma de tabla dinámica, se tuvo que hacer un proceso de organización y filtrado de la data, lo cual se realizó haciendo uso de los programas de Microsoft Excel para Microsoft 365 MSO (versión 2303 compilación 16.0.16227.20202) de 64 bits y Microsoft Power Bi Desktop (Versión: 2.116.622.0 64-bit), con el fin de obtener una

data ordenada, apta para su posterior análisis descriptivo y la ejecución de los objetivos propuestos.

Considerando la dificultad de manejar la data en su totalidad (Por su gran tamaño, y dada nuestra capacidad de cómputo con un procesador AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx 2.30 GHz y una memoria RAM instalada de 8 GB, los tiempos de espera entre operaciones realizadas eran demasiado largos), se tuvo que segmentar la data total por año de ocurrencia de los eventos, generando así, un total de 15 documentos tipo Excel a ser tratados.

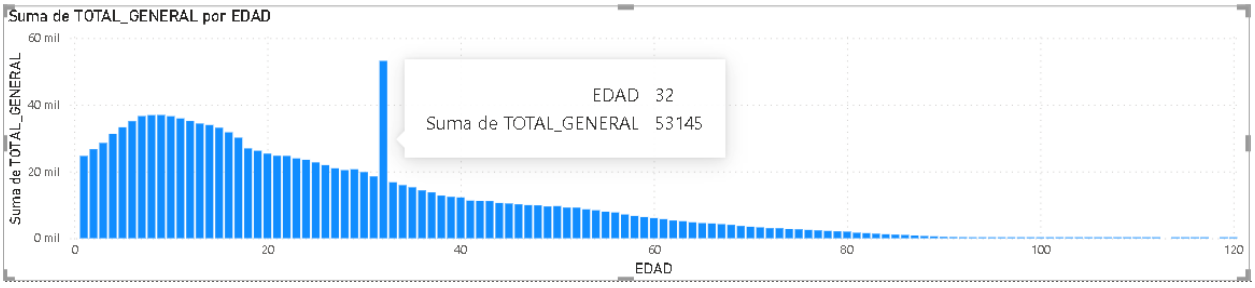
Respecto a la nomenclatura de la data, nos encontramos con un sistema de codificación empleado para identificar tanto a poblaciones como a otra serie de variables presentadas en la data, por ejemplo, respecto a los departamentos, teníamos codificaciones de dos dígitos para la identificación de cada uno de los departamentos (05-Antioquia, 08-Atlántico, …), por su parte, en los municipios se presentaba una nomenclatura similar, pero con una codificación de 5 dígitos (54874 - Villa Del Rosario, 66170 – Dosquebradas, …) y por último, se presentaba otro tipo de codificación, de 1 dígito, para identificar el área de residencia del contagiado, la cual va de 1 a 3 (1 – CABECERA, 2 - CENTRO POBLADO, 3 - AREA RURAL DISPERSA), todas estas codificaciones alfanuméricas (incluyendo el “-”), fueron eliminadas del set de datos.

Dentro de los registros de identificación del lugar de ocurrencia, se identificaron y eliminaron registros que no aportaban ningún tipo de información, ya que estaban definidos como “-No Definido”, además se dejó de tener en cuenta la variable “Departamento” debido a que una vez generada la segmentación de las localizaciones, la variable “Departamento” dejó de

generar un aporte significativo a la data, asimismo contaba con múltiples inconsistencias en sus registros (Nombres de departamentos mal escritos con y sin tildes).

Por último, se identificó una serie de registros con inconsistencias en la edad de los contagiados, ya que registraban, por ejemplo, edades superiores a los 300 años, presentando cierta acumulación en edades entre los 320 a 329 años. Esta inconsistencia se supuso como un error en la digitación, razón por la cual, en una primera estancia, se optó por generar una formula condicional en Power Bi Desktop, donde para edades mayores a 120 años se generaba una división de dicha edad por 10, llevando a estos registros a una serie de valores normales. No obstante, dicha acción nos condujo a generar un comportamiento anormal en la data, ya que, como se ve en la Figura 39, los registros de contagios en función de la edad, presentan un determinado comportamiento, el cual, al llegar al valor correspondiente a la edad de 32 años(Uno de los principales resultados de dividir edades de 320 años en 10), el valor de contagios se dispara abruptamente, lo que sugiere que, las personas de 32 años son más susceptibles a contagiarse de Dengue, Zika o Chikunguña, y esto, contrasta con la realidad.

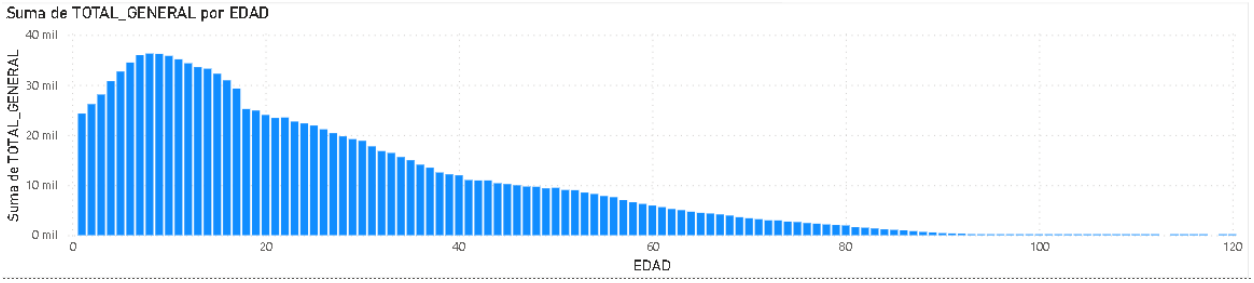
Figura 39. *Datos totales general con respecto a la edad.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

Por consiguiente, se trazó una solución alternativa, una que en lo posible no generara sesgos en los datos, dicha solución se basó en la eliminación de ese tipo de registros, quienes superaban los 70.000 registros de contagios (70.133 para ser exactos), siendo un 5,35% de los registros de datos de contagios totales. Con dicha acción, se logró eliminar el sesgo anteriormente presentado, además de no generar alteraciones significativas en el comportamiento de otra serie de factores analizados, como la proporción de distribución de los contagiados en función de su género, el comportamiento de los contagios y los respectivos picos de contagios a lo largo del tiempo de estudio, la proporción de contagios en función de las ciudades y municipios, la proporción de la distribución de los contagios en función de el área de ocurrencia, entre otros (Ver Figura 40).

Figura 40. *Datos totales general con respecto a la edad corrigiendo el error.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

7.2.2 Extracción y selección de los datos meteorológicos brindados por ArcGIS World Geocoder

El satélite de la NASA ArcGIS World Geocoder fue la base para la obtención de la información de los datos meteorológicos, las variables eran temperatura, presión, velocidad del viento, precipitación y humedad. Entre las cuales encontramos que, para la humedad, se presentaban las variables “specific Humidity at 2 Meters (g/kg)” y “Relative Humidity at 2 Meters (%)” donde hallamos, una que era la humedad específica medida en gramos/kilogramo, y la otra, la variable elegida, que fue la humedad relativa a 2 metros medida de forma porcentual(Relative Humidity at 2 Meters (%)) , ya que en investigaciones anteriores consultadas previamente, se obtuvo que todos tenían en cuenta la humedad, pero medida de forma porcentual ya que facilitaba su análisis. Para definir la variable de precipitaciones, también se contaba con sólo dos categorías, una era el valor promedio de precipitaciones diarias en el mes y la otra era la sumatoria de las precipitaciones de todo el mes, de estas, se escogió el valor promedio de precipitaciones diarias mes a mes para cada uno de los años involucrados en el estudio. Para el caso de la temperatura fue un poco más complicado ya que se tenían siete categorías distintas de temperatura, de las cuales se tuvo en cuenta la temperatura medida a 2 m sobre la tierra en grados Celsius, la temperatura Wet Bulb a 2 metros sobre la tierra en grados Celsius, el máximo de temperatura a 2 metros de la tierra en grados Celsius y el mínimo de la temperatura de la tierra a 2metros de la tierra en grados Celsius. La presión superficial sólo contaba con una categoría que fue la que se tomó en cuenta, medida en kPa, y por último, para el caso de la velocidad del viento, se escogió el dato con mayor proximidad al suelo, que era a los 10 m, por lo tanto, dentro de las categorías que estaban disponibles, que eran 10, se seleccionaron la velocidad del viento a 10 m medida en metros sobre segundo, la máxima velocidad del viento medida a 10 m también

en metros sobre segundos y la mínima velocidad del viento a 10 m sobre el suelo media en metros/segundo.

7.2.3 Extracción y selección de los datos del DANE

De la base de datos del departamento DANE se pretendía obtener una proyección poblacional de cada uno de los municipios seleccionados dentro de la franja de tiempo de 2007 a 2021, basándose en los censos realizados en el año 2005 (con su respectiva proyección poblacional hasta el año 2017) y el censo realizado en el año 2018 (de donde se tomaría la proyección poblacional hasta el año 2021). Esta base datos proporcionados por el departamento DANE también contenía información con respecto al género de las personas que harían parte de esa proyección poblacional, de esa de esa base de datos se obtuvo el total general proyectado para cada uno de los 20 municipios seleccionados como objeto de estudio, y esta variable entra a ser parte de los datos totales para analizar en el modelo de aprendizaje profundo de Red neuronal.

7.3 Análisis descriptivo

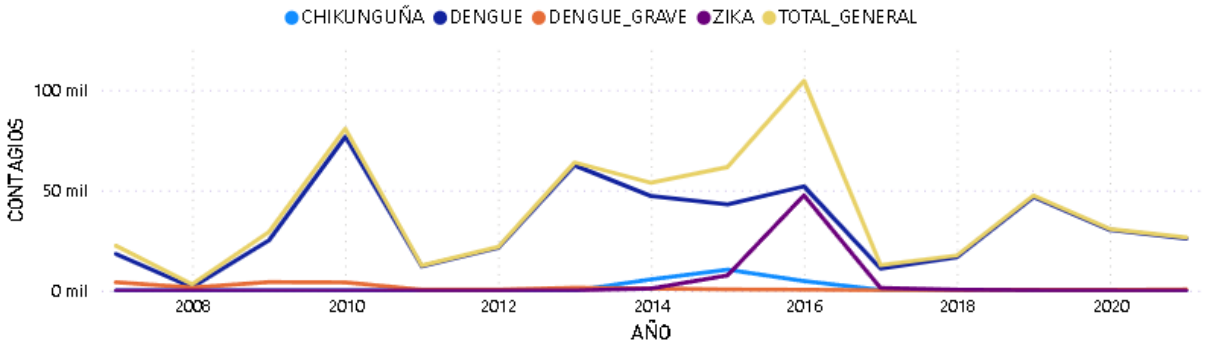
El análisis descriptivo, como su nombre lo indica, consiste en describir las tendencias claves en los datos existentes y observar las situaciones que conduzcan a nuevos hechos. Este método se basa en una o varias preguntas de investigación y no tiene una hipótesis. Además, incluye la recopilación de datos relacionados, posteriormente, los organiza, tabula y describe el resultado. (*¿Qué Es El Análisis Descriptivo?*, n.d.)

Teniendo en cuenta lo anterior, es fundamental llevar a cabo un estudio de carácter descriptivo usando los datos que han sido previamente procesados y categorizados. El objetivo de este análisis es generar una visión diagnóstica que permita observar la tendencia de los casos de Dengue, Zika y Chikunguña en Colombia entre los años 2007 a 2021 en las localizaciones (Municipios y ciudades) elegidas.

En este apartado se analiza la variabilidad, correlación, dependencia de variables y distintos comportamientos que presentan los datos, con el fin de hacer un acercamiento exploratorio al cómo se ha desarrollado la dinámica de contagios en las localizaciones objeto de estudio y en los años preestablecidos.

En una primera instancia, se buscó analizar el comportamiento a lo largo del tiempo (año 2007 a año 2021) de los contagios en las ciudades y municipios elegidos para el estudio, lo cual podemos observar en la Figura 41, en dicha figura se puede observar una serie de picos de contagios, los cuales se produjeron en los años 2010, 2013, 2016 y 2019, siendo el mayor de estos picos el del año 2016.

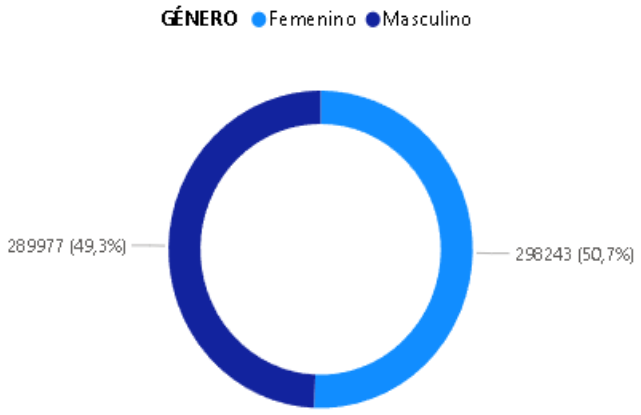
Figura 41. *Picos de contagios.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

Si analizamos la Figura 42, podemos ver la distribución de los contagios acorde al género de los contagiados. En ella, se aprecia un reporte con 289.977 contagios en el género masculino y 298.243 en el género femenino (para las ciudades segmentadas previamente con reportes superiores a los 10.000 contagios), representando el 49,3% y 50,7% respectivamente, en consecuencia, se puede apreciar que no existe una diferencia sustantiva respecto al género del contagiado.

Figura 42. *Distribución de los contagios acorde al género de los contagiados.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

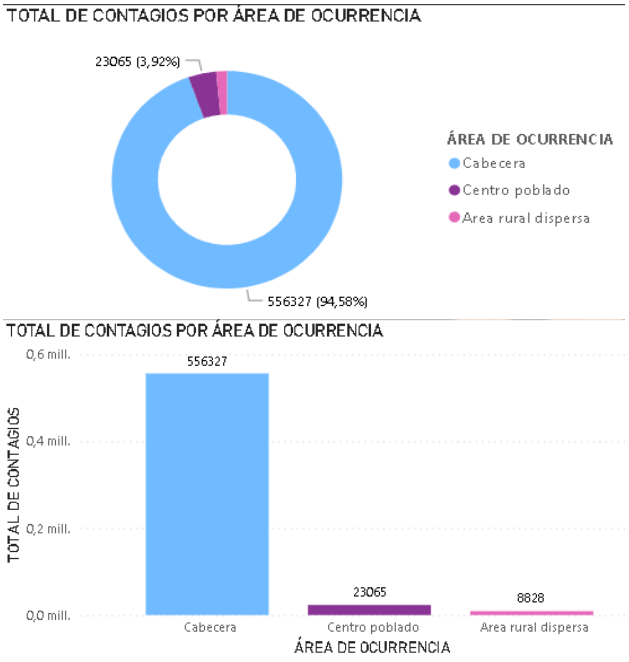
En la Figura 43 se muestra la segmentación de los contagios acorde al área de ocurrencia de los contagios:

- Cabecera: Sector territorial del municipio donde se desarrolla la actividad urbana y se encuentra la sede de los poderes municipales. (*Definición de Cabecera Municipal - Diccionario Panhispánico Del Español Jurídico - RAE*, n.d.)

- Centro Poblado: es un concepto creado por el DANE para fines estadísticos de localización geográfica de núcleos de población. Se define como una concentración de mínimo veinte (20) viviendas contiguas, vecinas o adosadas entre sí, ubicada en el área rural de un municipio o de un Corregimiento Departamental. (Comunitaria et al., n.d.)
- Área rural dispersa: Es la unidad habitacional localizada en el suelo rural de manera aislada que se encuentra asociada a las formas de vida del campo y no hace parte de centros poblados rurales ni de parcelaciones destinadas a vivienda campestre.(Fernando & Usuga, n.d.)

Tenemos un gráfico de barras y uno de anillos, segmentados acorde al área de ocurrencia del contagio, en ella (Figura 43), se puede notar una clara diferencia entre las tres opciones disponibles, como era de esperar el área de ocurrencia “Cabecera” fue el lugar en el que más se registraron contagios, un total de 556.327, representando el 94,58% de registros de contagios totales, seguido del área “Centro poblado” con 23.063 registros de contagios, representando 3,92% de la data, y por último, el área denominada “Área rural dispersa” con 8.828 contagios registrados, un 1,5% de la data total.

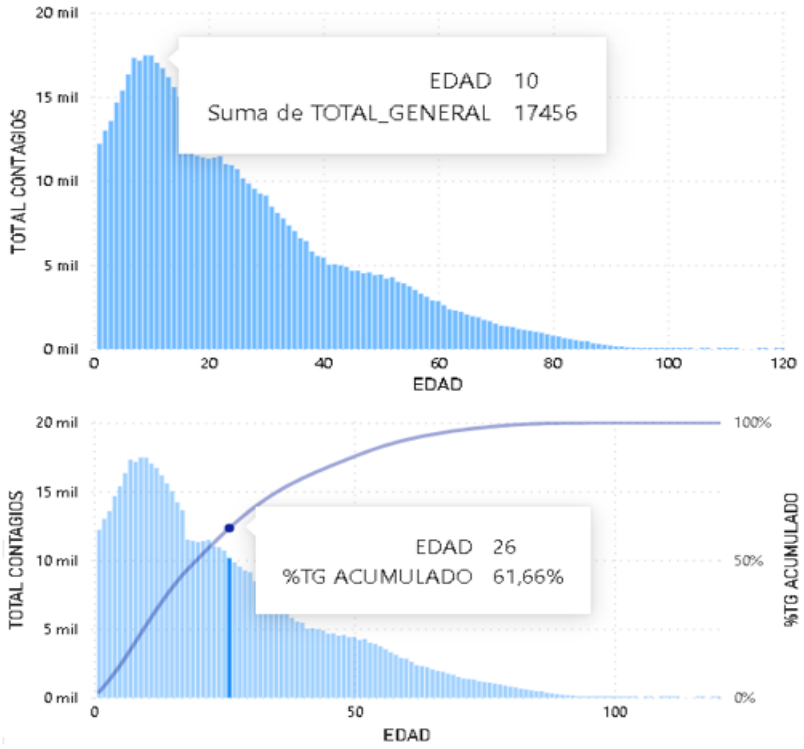
Figura 43. *Total de contagios por área de ocurrencia.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

Al analizar la variable “Edad”, observamos cierta distribución de los registros de contagios, las edades con mayor número de contagios registrados se hallan en los jóvenes, adolescentes y niños, considerando “… in útero y nacimiento, primera infancia (0-5 años), infancia (6 - 11 años), adolescencia (12-18 años), juventud (14 - 26 años), adultez (27 - 59 años) y vejez (60 años y más)” (MINSALUD, s.f.). puntualmente por debajo de los 26 años, de hecho, el pico de los contagios se halla en personas con 10 años, registrando 17.456 contagios (Ver Figura 44). A partir de dicha edad, se empieza a registrar un descenso en el número de contagios a medida que aumentamos la edad, de hecho, las personas menores de 26 años representan el 61,66% del total de los contagios, generando así la noción de que se registran más contagios en personas jóvenes.

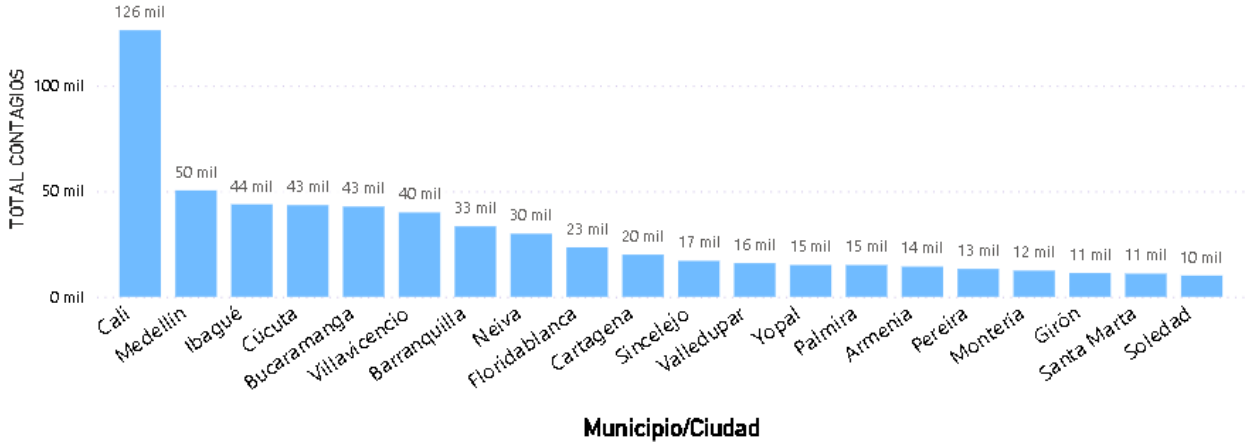
Figura 44. *Edad y total porcentual acumulado.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

Al graficar los contagios en función del municipio/ciudad de ocurrencia, se pueden notar ciertas diferencias entre diferentes localidades, por ejemplo, en la ciudad de Cali se registraron alrededor de 126 mil contagios (Ver Figura 45), doblando así, en número de contagios, a la segunda ciudad con mayor número de contagios, Medellín, con 50 mil, lo que nos sugiere que existen ciertos factores de incidencia que hacen que la ciudad de Cali presente un mayor número de contagios, ya que por ejemplo, en cuestión de población estimada, Medellín supera a Cali por varios miles de habitantes, pero son más los asentamiento de agua y las precipitaciones en la capital del valle del cauca, lo que nos lleva a creer que por ello sea un entorno fructífero para la proliferación del principal vector transmisor de Dengue, Zika y Chikunguña, el mosquito Aedes aegypti.

Figura 45. *Municipio o ciudad.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Power BI.

7.4. Reducción de dimensiones

Con el fin de evitar un sobre ajuste y mejorar el desempeño de los modelos a desarrollar, se procede a realizar una reducción de dimensionalidad “En los problemas de Machine Learning y de la ciencia de datos, el objetivo principal sigue siendo encontrar las características más relevantes que juegan un papel dominante en la determinación e influencia de los resultados de la producción” (aprendeIA, 2023), además, “Cada característica que se incluye en el análisis, puede llegar a incrementar el costo y el tiempo de proceso de los sistemas, por lo que hay una fuerte motivación para diseñar e implementar sistemas con pequeños conjuntos de características.” (Henández, Delgado Trejos, Rivera Piedrahita, & Castellanos, 2006)

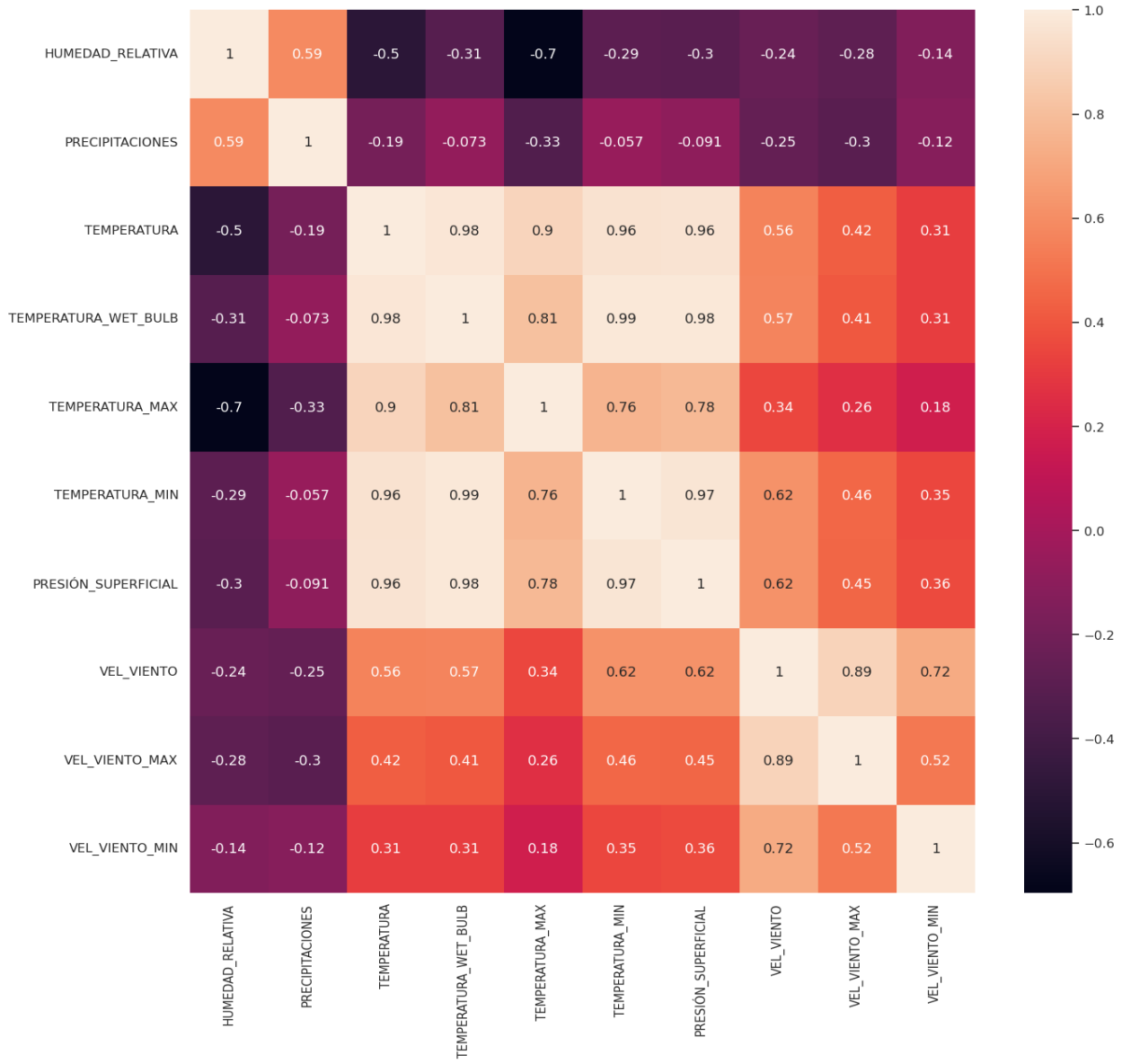
7.4.1. Filtro de alta correlación

En este apartado se realizará un filtro de correlación entre cada una de las variables numéricas del conjunto de datos, con el fin de determinar si existen altas correlaciones entre las variables y así realizar una reducción de estas “Es probable que las columnas de datos con tendencias muy similares también contengan información muy similar, y solo una de ellas bastará para la clasificación.” (*El Caracol Africano (Achatina Fúlica) (Página 2)*, n.d.)

En la Figura 46, podemos apreciar la matriz de correlación con cada una de las variables numéricas de la data del proyecto, donde se puede apreciar que sí existen altas correlaciones entre algunas variables, por ejemplo, las variable ‘TEMPERATURA’ presenta una alta correlación con las variables ‘PRESION_SUPERFICIAL’, 0.96 , ‘TEMPERATURA_MIN’, 0.96, ‘TEMPERATURA_MAX’, 0.9, y ‘TEMPERATURA_WET_BULB’, 0.98, por su parte, la variable ‘TEMPERATURA_WET_BULB’ tiene altas correlaciones con el mismo grupo de variables.

Razón por la cual, y para efectos prácticos, se procedió a eliminar las variables “TEMPERATURA” y “TEMPERATURA_WET_BULB” mencionadas anteriormente del set de datos a tratar.

Figura 46. *Variables.*



Nota: Información recopilada de la base de datos SISPRO y analizada en Google Colab.

7.4.2. Reducción de otras variables

Con el fin de optimizar el rendimiento del modelo, y evitar 'El sobre ajuste', "El sobreajuste ocurre cuando el modelo se corresponde demasiado con un conjunto particular de datos y no se generaliza bien. Un modelo sobredimensionado funcionaría demasiado bien en el

conjunto de datos de formación para que falle en datos futuros y haga que la predicción sea poco fiable" (aprendeIA, 2023), se decide prescindir de las variables 'GENERO', 'EDAD', 'ALTURA_MEDIA_MSNM' y 'AREA_DE_OCURRENCIA', lo cual coincide con la búsqueda de un modelo de comparación con variables similares, dicho modelo de comparación será el modelo tomado como referencia de Jiucheng Xu et al. (2020), además, y en el caso de la variable 'GENERO' se elimina debido a que esta no presenta una diferencia sustantiva entre los géneros Femenino y Masculino, tal como se trató anteriormente (289.977 contagios en el género masculino y 298.243 en el género femenino, representando el 49,3% y 50,7% respectivamente), respecto al 'AREA_DE_OCURRENCIA' se notó que la gran mayoría de los casos se presentaban en el área denominada como "Cabecera" con un 94,58% de los casos y la 'ALTURA_MEDIA_MSNM' de todas las localizaciones escogidas se hallaba en el rango de altura media sobre el nivel del mar medida en metros (msnm) comprendido entre los 2 msnm (ciudad de Cartagena) y 1551 msnm (ciudad de Armenia), lo cual coincide con las condiciones normales del ecosistema habitado por el principal vector de transmisión el mosquito Aedes aegypti a alturas menores a los 2200 msnm "se han adelantado múltiples investigaciones que han permitido identificar muy puntualmente los principales requisitos ambientales que determinan la presencia del mosquito; se sabe que para su reproducción necesitan altitudes menores a los 2.200 msnm, aunque se ha reportado presencia del mosquito a más de 2.300 msnm"(*Dengue (Página 2)*, n.d.)

Una vez finalizado el análisis descriptivo y realizada la reducción de dimensionalidad, y previo a la creación de los modelos, se toma como punto de partida, o modelo base, el estudio de Jiucheng Xu et al. (2020), que resalta la capacidad de realizar pronósticos de casos de Dengue en ciudades chinas mediante el uso del método de Aprendizaje Profundo. En este contexto, se

respalda el análisis de correlación en la sección anterior con el análisis del articulo realizado por Jiucheng Xu et al. (2020), y así mantener únicamente las siguientes variables: ‘MUNICIPIO’, ‘AÑO’, ‘MES’, ‘HUMEDAD_RELATIVA’, ‘PRECIPITACIONES’, ‘TEMPERATURA_MAX’, ‘TEMPERATURA_MIN’, ‘PRESIÓN_SUPERFICIAL’, ‘VEL_VIENTO’, ‘VEL_VIENTO_MIN’, ‘VEL_VIENTO_MAX’ y ‘TOTAL_GENERAL’. La elección de no utilizar todos los encabezados disponibles se realiza con el propósito de optimizar el rendimiento de la red neuronal, esta reducción de la dimensionalidad en el conjunto de datos persigue una aproximación más eficiente en el proceso de entrenamiento, lo que permite a la red neuronal capturar patrones significativos con mayor precisión y robustez (Jiucheng Xu et al., 2020).

8. Creación de los modelos

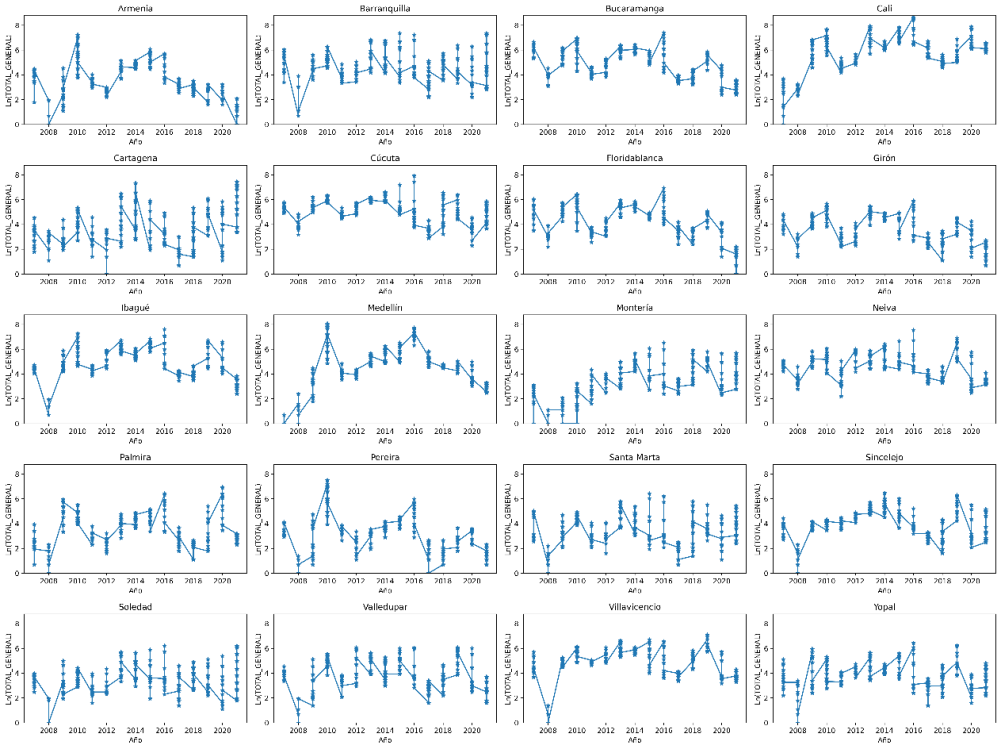
8.1. Definición, Entrenamiento y Validación de los Modelos Predictivos

En esta sección, se detalla el procesamiento de datos, que abarca la verificación del set de datos desde la organización de la data que se va a ingresar a la red neuronal hasta la definición de experimentos, la validación de modelos GRU, LSTM y RNN, y la elección de los modelos definitivos. La selección del modelo con la mejor capacidad de generalización se fundamenta en una evaluación comparativa que se apoya en métricas estándar para modelos de series de tiempo como el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el coeficiente de determinación (R^2) y tiempo de ejecución.

La implementación del código de programación se llevó a cabo en la plataforma Google Colab, esta consta de tres scripts, cada script se fundamenta en las redes de series de tiempo, estas fueron, redes GRU, LSTM y RNN (Google Colab, Arboviral_GR; Arboviral_LSTM; Arboviral_RNN), la cual facilita a los usuarios escribir código en Python de manera organizada y comprensible, además de permitir la compilación y visualización de datos de forma eficiente, Colab permite el acceso a una CPU Intel Xeon a 2,20 GHz con 13 GB de RAM.

8.1.1. Preprocesamiento y normalización de Datos. El procesamiento de los datos se divide en varias etapas esenciales. En primer lugar, se procede a la carga, ordenamiento y normalización de los datos. La carga de los datos se efectúa utilizando la biblioteca Pandas, permitiendo obtener una visión general del registro de casos de dengue en las 20 ciudades a lo largo del período comprendido entre 2007 y 2021. Como parte de la normalización, se aplica el Logaritmo Natural más uno ($\ln(x + 1)$) a los valores de casos, lo que además de facilitar la visualización, contribuye a la homogeneización de los datos, este proceso se ilustra en la Figura 47.

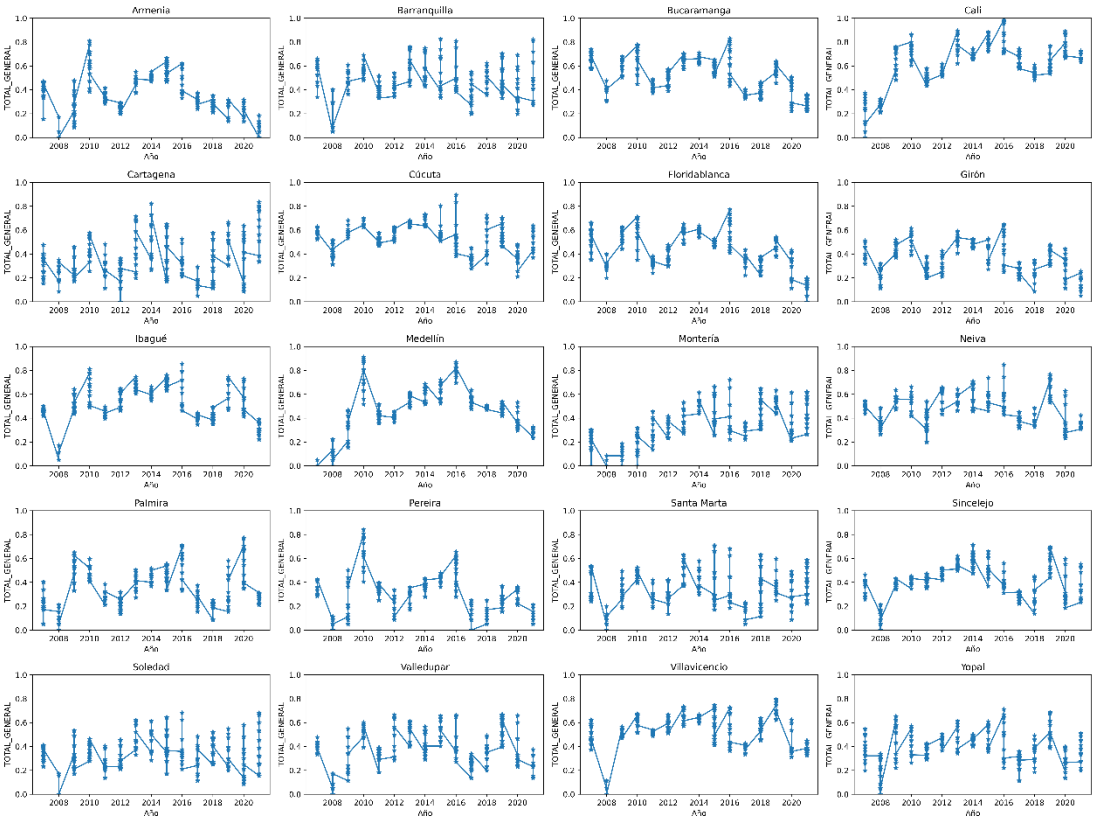
Figura 47. *Logaritmo natural de los casos de dengue para las ciudades a través de los años.*



Nota: Información recopilada de la base de datos y analizada en Google Colab

Además del proceso de normalización anteriormente mencionado, se aplica una técnica adicional de normalización que implica el re-escalamiento de los datos. En el ejemplo presentado, se lleva a cabo el escalamiento de los valores de casos en un rango de 0 a 1 para cada ciudad, y para el resto variables o encabezado de los datos. Para lograr esto, se utilizó la librería ‘sklearn’ y función ‘MinMaxScaler’. Este procedimiento se realiza con el propósito de preservar la integridad y homogeneidad de los datos antes de introducirlos en la red neuronal. Una vez completada la normalización adicional, se obtiene la misma representación gráfica de los datos, como se puede observar en la Figura 48.

Figura 48. *Representación de re-escalamiento para casos de dengue en las ciudades.*

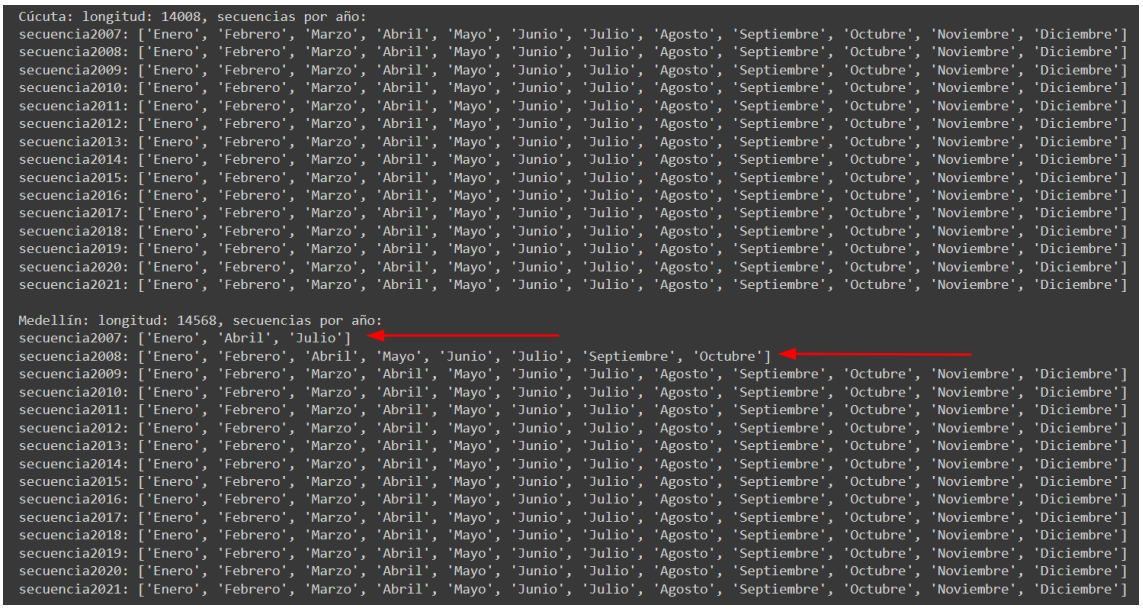


Nota. Información recopilada de la base de datos y analizada en Google Colab.

Luego, se realizó una visualización de los campos con el propósito de verificar la integridad de todos los registros. Durante la inspección se verificó que el conjunto de datos contara con un registro para cada mes en todas las ciudades, lo que equivaldría a un total de 3600 registros (12 meses x 20 ciudades x 14 años). Sin embargo, se constató que la longitud de los datos era de 3533 registros. Al profundizar en la búsqueda de datos faltantes, se identificó que las ciudades de Palmira, Montería, Santa Marta, Valledupar, Sincelejo, Armenia, Barranquilla, Medellín e Ibagué presentaban ausencia de registros para algunos meses en los años 2007 y 2008 (Ver Figura 49). Ante esta situación, se tomó la decisión de excluir estos dos años específicos del

análisis, ya que era necesario contar con un conjunto de datos completo y normalizado antes de incorporarlo a la red neuronal.

Figura 49. *Ejemplo de registros faltantes para Medellín.*



Nota: Información recopilada de la base de datos y analizada en Google Colab

Finalmente, se prepara los datos para la entrada de la red neuronal, se inicia con el orden para construir la serie de tiempo que se presenta a la red neuronal, inicia con año, precedida de municipio y finalmente, mes. El conjunto de entrada de la red está formado por dos vectores, sec_X y sec_y; sec_X representa la serie de tiempo de entrada con forma (2400, 12, 9) y sec_y representa la serie de tiempo de salida, es decir, lo que se espera dada una secuencia sec_X, su forma es (2400, 1), para finalmente separar en conjunto de entrenamiento y validación, el de entrenamiento corresponde con secuencia desde 2009 al 2019; mientras que, en validación son los años 2020 y 2021. Cabe aclarar que, para predecir un mes, es necesario pasar 12 meses anteriores, seguidas de las 8 variables meteorológicas y la variable de casos de dengue.

Tabla 4. *Hiperparámetros guía y experimento con variación en malla para entrenamiento.*

Parámetro	Entrenamiento guía	GRU	LSTM	RNN
Tamaño del batch	24	24 - 12	24 - 12	24 -12
Tasa de aprendizaje	1e-5	1e-4 - 1e-5	1e-4 - 1e-5	1e-4 - 1e-5
Look back	12	12	12	12
Unidades RNN	64	64	64	64
Capas ocultas	1	2 - 3	2 - 3	2 - 3
Dropout	0.4	0.4 - 0.5 - 0.3	0.4 - 0.5 - 0.3	0.4 - 0.5 - 0.3

8.1.2 Definición de los modelos predictivos.

En la definición de los modelos se tomaron tres topologías de red a comparar, una red con capas GRU, LSTM y RNN (Google Colab, Arboviral_GR; Arboviral_LSTM; Arboviral_RNN). Las redes GRU, LSTM y RNN son tipos de redes neuronales recurrentes que son particularmente útiles para tareas de predicción de secuencias (Liu et al., 2018), y como caso particular en este trabajo, la predicción de casos de dengue para Colombia. Para la elección de parámetros que mejor se adapten al problema, inicialmente se tomó la arquitectura de red diseñada por Jiucheng Xu et al., (2020), a partir de esta se realizaron los experimentos en malla tomando la variación hiperparámetros de la red que se muestran en la Tabla 4.

El consumo de recursos computacionales en Google Colab para las tres redes neuronales recurrentes (GRU, LSTM Y RNN) se resume en la Tabla 5, los datos se obtuvieron del entrenamiento base y de la malla de experimentos (Google Colab, Arboviral_GR; Arboviral_LSTM; Arboviral_RNN).

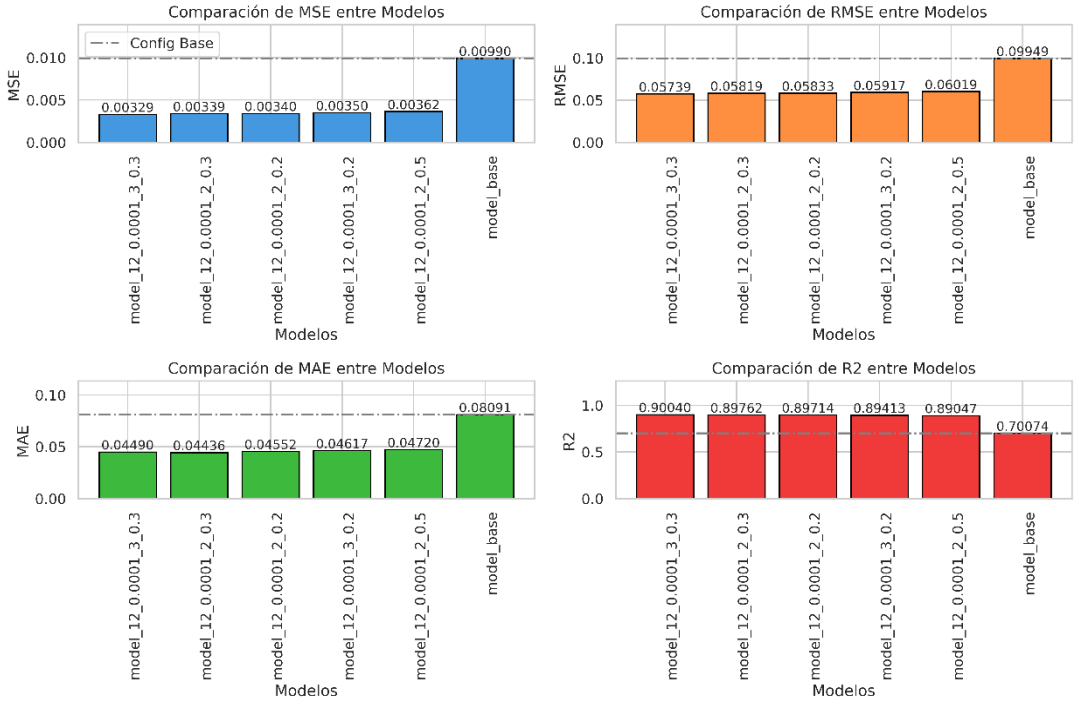
Tabla 5. *Costo computacional.*

Entrenamiento	Tiempo de ejecución (s)
Guia RNN	461.494
Guia LSTM	706.771
Guia GRU	816.476
Malla RNN	4027.52
Malla LSTM	13925.91
Malla GRU	22126.3

Nota. Guía corresponde al entrenamiento unitario con la configuración base del artículo Jiucheng Xu et al., (2020). Malla corresponde a los experimentos realizados en la Tabla 4.

Debido a la gran cantidad de experimentos realizados (72 + 3, 24 por tipo de red más las configuraciones bases), se optó por generar gráficos que representen las métricas de los mejores modelos utilizando cuatro métricas mencionadas previamente en la sección 8.1.

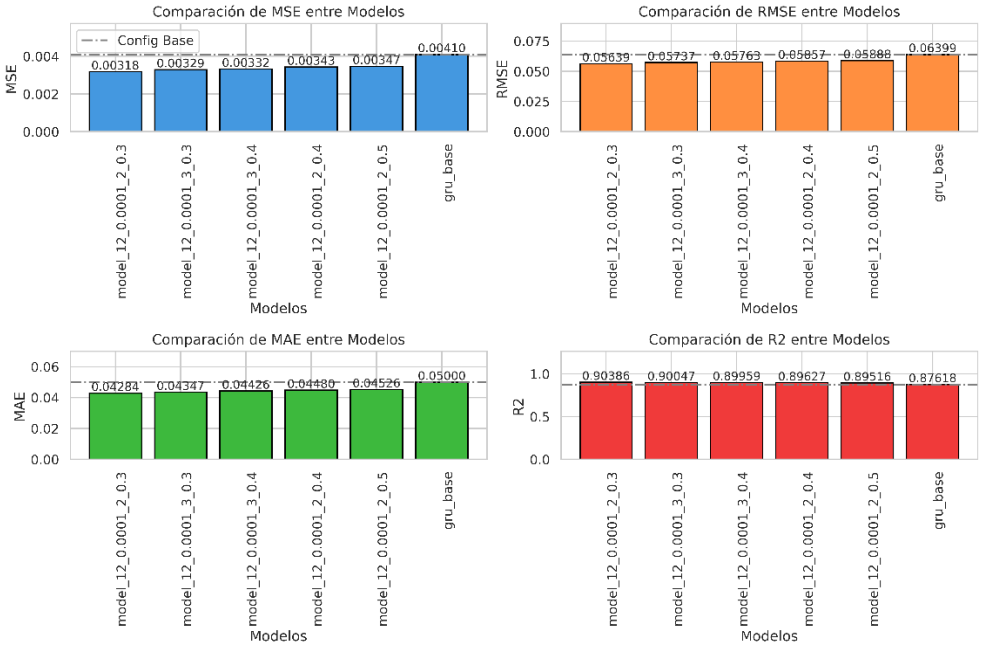
Figura 50. *Métricas para las mejores 5 configuraciones, más la base (Modelos LSTM).*



Nota. Se presentan los resultados obtenidos para la malla de experimentos con la red LSTM. Se muestran las métricas de error medio cuadrático (MSE), error cuadrático medio raíz (RMSE), error medio absoluto (MAE) y coeficiente de determinación (R^2) para las mejores 5 configuraciones, más la configuración base.

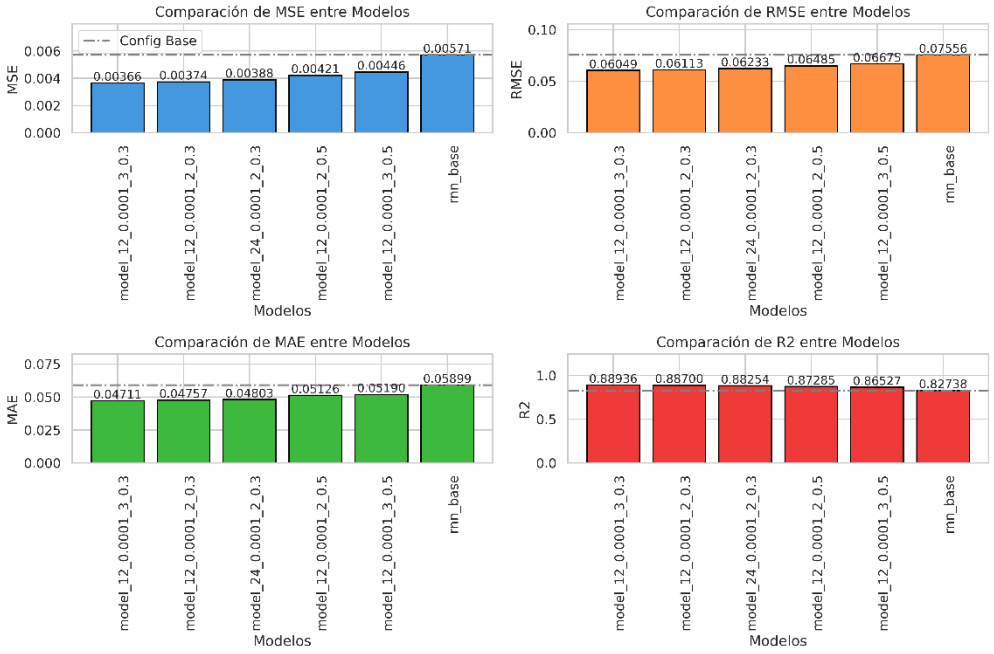
En la Figura 50 se presentan los resultados obtenidos para la malla de experimentos con la red LSTM, la Figura 51 los resultados de los experimentos con la red GRU y la Figura 52 para la RNN. En el eje X de cada figura se representan los nombres de los modelos, siguiendo una jerarquía de parámetros que incluye el tamaño de lote (batchsize), la tasa de aprendizaje (lr), el número de capas (numlayers) y el valor de dropout, es decir, "model_batchsize_lr_numlayers_dropout"; para el eje Y, se muestran los respectivos valores de la métrica para cada caso, lo que facilita la evaluación y comparación de las métricas MAE, MSE, RMSE y R^2 entre las diferentes configuraciones de modelos. La Tabla 6 muestra un resumen de los mejores modelos para cada caso.

Figura 51. *Métricas para las mejores 5 configuraciones más la base (Modelos GRU).*



Nota. Se presentan los resultados obtenidos para la malla de experimentos con la red GRU. Se muestran las métricas de error medio cuadrático (MSE), error cuadrático medio raíz (RMSE), error medio absoluto (MAE) y coeficiente de determinación (R^2) para las mejores 5 configuraciones, más la configuración base.

Figura 52. *Métricas para las mejores 5 configuraciones más la base (Modelos RNN).*



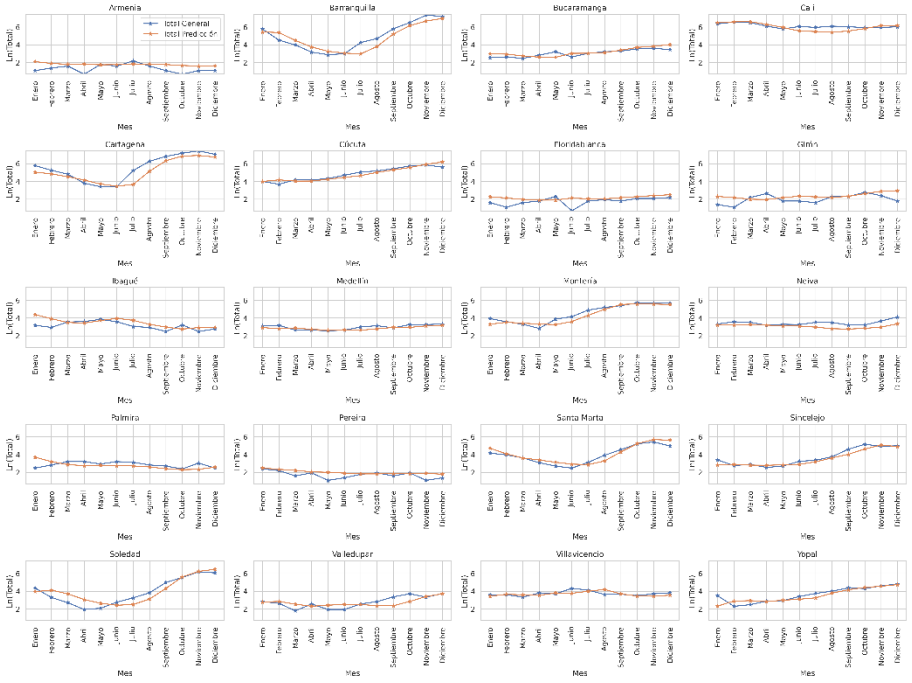
Nota. Se presentan los resultados obtenidos para la malla de experimentos con la red RNN. Se muestran las métricas de error medio cuadrático (MSE), error cuadrático medio raíz (RMSE), error medio absoluto (MAE) y coeficiente de determinación (R^2) para las mejores 5 configuraciones, más la configuración base.

Tabla 6. Resumen de rendimiento para los mejores modelos de cada arquitectura.

Arquitectura: configuración	**MSE**	**RMSE**	**MAE**	**R^2**
LSTM*: model_12_0.0001_2_0.3*	0.00329	0.05739	0.0449	0.90040
GRU*: model_12_0.0001_2_0.3*	0.00318	0.05639	0.04284	0.90386
RNN*: model_12_0.0001_3_0.3*	0.00366	0.06049	0.04711	0.88936

Nota: Nota: La nomenclatura de la arquitectura *model_bsz_lr_numlayer_dropout* se utiliza para especificar los siguientes parámetros: bsz: tamaño del batch, lr: tasa de aprendizaje, numlayer: número de capas ocultas y dropout: cantidad de desconexión de las capas ocultas.

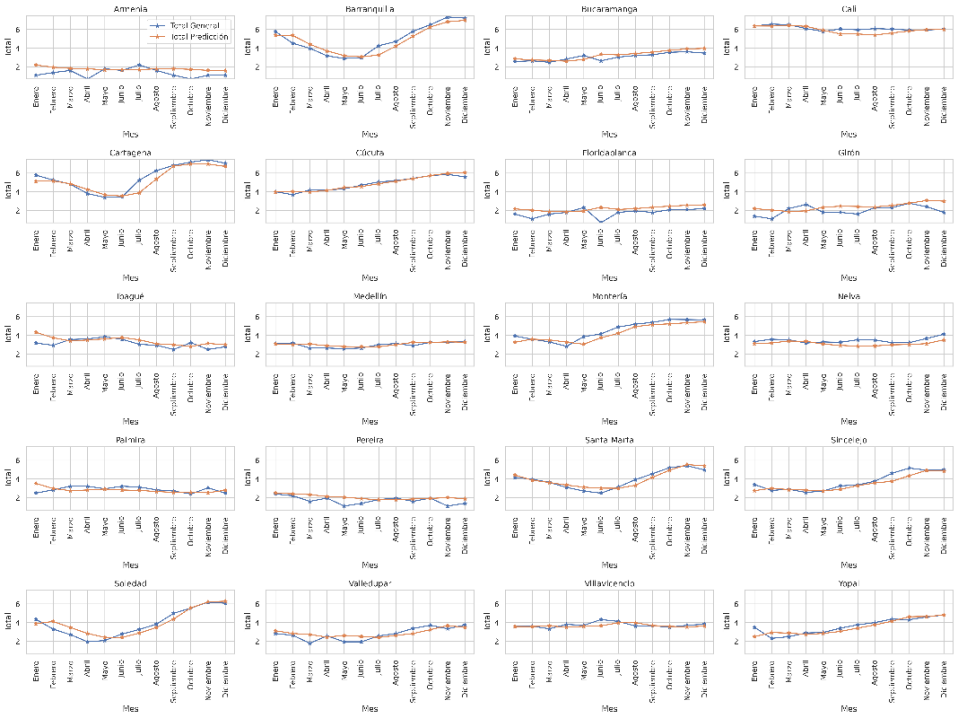
Figura 53. *Comparación entre casos originales y casos predichos por red RNN (Ln +1).*



Nota. se presenta la comparación visual de los resultados reales y predichos de la red RNN para las 20 ciudades con el set de validación, es decir, para los años 2020 y 2021.

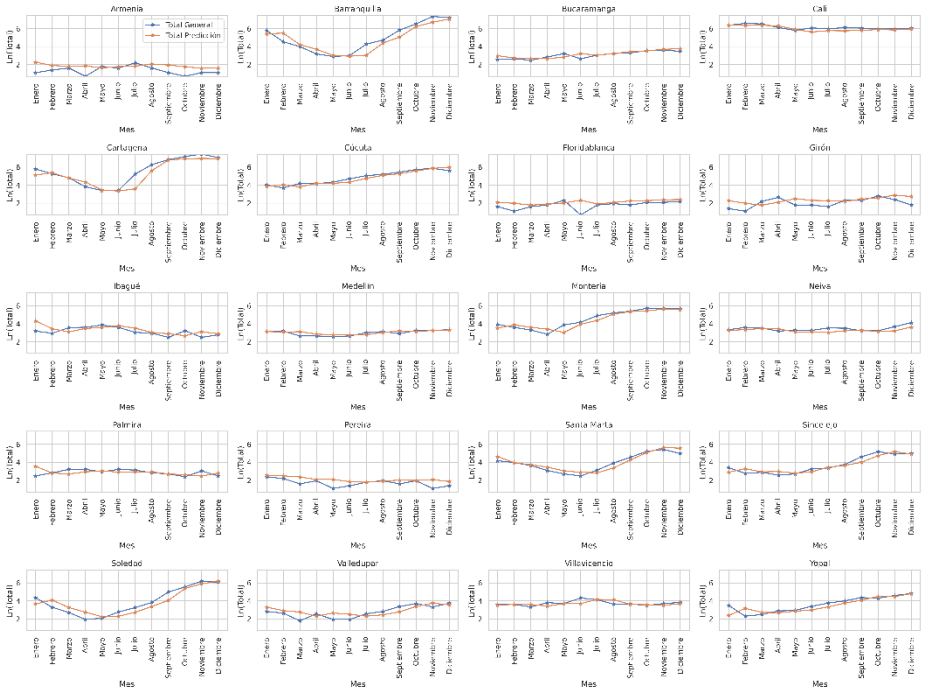
Para obtener una vista completa de los resultados de inferencia en el conjunto de validación, se pueden consultar las Figuras 53, 54 y 55, que corresponden a los modelos RNN *(model_12_0.0001_2_0.3)*, LSTM (*model_12_0.0001_2_0.3*) y GRU (*model_12_0.0001_3_0.3*), respectivamente, cabe aclarar que el eje Y representa los valores del Logaritmo natural + 1 de los casos de dengue con la intención de ver en escala normalizada los valores originales y de predicción.

Figura 54. *Comparación entre casos originales y casos predichos por red LSTM (Ln +1).*



Nota. se presenta la comparación visual de los resultados reales y predichos de la red RNN para las 20 ciudades con el set de validación, es decir, para los años 2020 y 2021.

Figura 55. *Comparación entre casos originales y casos predichos por GRU (Ln +1).*



Nota. Se presenta la comparación visual de los resultados reales y predichos de la red RNN para las 20 ciudades con el set de validación, es decir, para los años 2020 y 2021.

9.Conclusiones

A partir de los resultados obtenidos en las métricas MAE, MSE, RMSE y R^2 para las diferentes arquitecturas (GRU, LSTM y RNN) de los mejores modelos (Ver Tabla 6), se observa que sus RMSE son similares, con valores de 5.639%, 5.739% y 6.049% respectivamente (Ver Figuras 50, 51, 52 y Tabla 6), esto sugiere que las tres arquitecturas tienen un rendimiento comparable en la predicción de casos de enfermedades arbovirales (Dengue, Zika y Chikunguña). No obstante, el modelo GRU destaca por su precisión del 5.639% en RMSE (Ver Tabla 6. model_12_0.0001_3_0.3), aunque requiere más tiempo de ejecución, aproximadamente

816 segundos, en comparación con los 706 segundos de la LSTM y los 461 segundos de la RNN (Ver Tabla 4).

Dado que los modelos exhiben resultados similares, se abre un espacio significativo para la exploración y ajuste de los hiperparámetros de las redes neuronales. Esto señala la posibilidad de mejorar aún más el rendimiento de las redes neuronales mediante la optimización de sus configuraciones.

La selección adecuada de hiperparámetros puede ser esencial para maximizar la precisión en la predicción de brotes de enfermedades arbovirales. Es importante destacar que el enfoque inicial propuesto por el artículo de Xu et al. (2020) proporcionó resultados prometedores. Para las arquitecturas RNN, LSTM y GRU, se obtuvo un RMSE de 9.949%, 6.399% y 7.556% respectivamente (Ver Figuras 50, 51 y 52). Estos valores indican que la malla de experimentos detallada en la Tabla 3 fue fundamental para mejorar las arquitecturas de referencia.

La segmentación geográfica sobre las localizaciones (ciudades y municipios) objeto de estudio, bajo el criterio de trabajar con las localizaciones que reportaran más de 10000 casos de contagios totales en el espacio de tiempo comprendido entre los años 2007 a 2021, facilitó en gran medida el desarrollo de la investigación, pues las 20 localizaciones elegidas representan tan sólo el 1,79 % del total de municipios en Colombia, pero a su vez representan el 47,38 % del total de contagios reportado (Ver Figura 37), lo cual significó trabajar con una población representativa del conjunto de datos y a su vez reducir el costo computacional de los modelos trabajados.

La presente investigación constituye un avance significativo en la ingeniería industrial al implementar un modelo de red neuronal en Python para predecir brotes de enfermedades arbovirales en Colombia. Integrando variables demográficas y climatológicas de 20 municipios colombianos durante los años 2007 al 2021, los resultados obtenidos no solo ofrecen una base sólida para futuros trabajos, sino que también señalan la posibilidad de extender la investigación al campo de la generación de datos para predecir los años posteriores al 2021. Se sugiere que estos trabajos futuros consideren datos de al menos 12 meses previos al mes objetivo, permitiendo así generar predicciones para el mes siguiente a ese período. Este enfoque estratégico no solo fortalecería la base de conocimientos en la disciplina de la ingeniería industrial, sino que también representaría una contribución significativa al desarrollo continuo de soluciones innovadoras en el ámbito de la salud pública. De esta manera, la aplicación de la ingeniería industrial emerge como un recurso valioso para la gestión y análisis de datos esenciales, con un potencial real para mejorar la calidad de vida de la sociedad.

10. Recomendaciones

Se recomienda que el gobierno, las entidades públicas de salud y las entidades territoriales adopten una toma de decisiones basada en datos. Este tipo de investigaciones fortalece la toma de decisiones para reducir la incidencia de enfermedades (En este caso Zika, Dengue y Chikunguña) transmitidas por vectores en la población. Se sugiere replicar este estudio en otros entornos, variando e incorporando nuevos parámetros de entrada, como datos meteorológicos y demográficos, entre los que se podría incluir la radiación solar, el índice de vegetación, el índice de pobreza o el estrato socioeconómico del individuo afectado.

Se aconseja agregar métricas adicionales que brinden una evaluación más completa además de los ya utilizados, que eran el coeficiente de determinación (R2), el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE). Podría incluirse la Puntuación de Información Bayesiana (BIC) para comparar modelos considerando tanto el ajuste como la complejidad, el Error Porcentual Absoluto Medio (MAPE) para evaluar la precisión porcentual promedio del modelo y el Índice de Concordancia de Lin (CCL) para evaluar la concordancia entre las predicciones del modelo y los datos reales.

Referencias Bibliográficas

Acevedo M, E., Serna A, A., & Serna M, E. (2017). Capitulo 10. Principios y características de las redes neuronales artificiales. In Desarrollo e Innovación en Ingeniería (2nd ed.). Editorial Instituto Antioqueño de Investigación. https://www.researchgate.net/profile/Jhon-Fredy-Narvaez/publication/320170890_Desarrollos_de_la_Ingenieria_ambiental_en_la_evaluacion_de_la_calidad_de_los_recursos_naturales_y_la_salud_ambiental/links/59d26bfca6fdcc181ad611ce/Desarrollos-de-la-Ingenieria-ambiental-en-la-evaluacion-de-la-calidad-de-los-recursos-naturales-y-la-salud-ambiental.pdf#page=174

Ai, B. (2021, December 12). Redes neuronales - Bootcamp AI - Medium. Medium. https://bootcampai.medium.com/redes-neuronales-13349dd1a5bb

Alonso Botero, J., Arellano Morales, M., Medina Gaspar, D., & Echeverri Durán, C. (2021). EVALUACIÓN DEL SISTEMA DE SALUD COLOMBIANO: UNA REVISIÓN EN EL MARCO DE LA LEY ESTATUTARIA EN SALUD DE 2015. https://www.anif.com.co/wp-content/uploads/2021/08/anif-doc-final-def.pdf

Amin, S., Irfan Uddin, M., Ali Zeb, M., Alarood, A. A., Mahmoud, M., & Alkinani, M. H. (2020). Detecting dengue/flu infections based on tweets using LSTM and word embedding. IEEE Access, 8, 189054–189068. https://doi.org/10.1109/ACCESS.2020.3031174

Amin, S., Irfan Uddin, M., Ali Zeb, M., Alarood, A. A., Mahmoud, M., & Alkinani, M. H. (2020). Detecting dengue/flu infections based on tweets using LSTM and word embedding. IEEE Access, 8, 189054–189068. https://doi.org/10.1109/ACCESS.2020.3031174

Amin, S., Uddin, M. I., Hassan, S., Khan, A., Nasser, N., Alharbi, A., & Alyami, H. (2020). Recurrent Neural Networks with TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease. IEEE Access, 8, 131522–131533. https://doi.org/10.1109/ACCESS.2020.3009058

Análisis de correlación – Conogasi. (n.d.). Retrieved February 4, 2024, from https://conogasi.org/articulos/analisis-de-correlacion-2/

Análisis de correlación. (2017, May 25). Conogasi. https://conogasi.org/articulos/analisis-de-correlacion-2/

Análisis de Regresión: ¿Cómo Puedo Interpretar el R-cuadrado y Evaluar la Bondad de Ajuste? (n.d.). Retrieved February 4, 2024, from https://blog.minitab.com/es/analisis-de-regresion-como-puedo-interpretar-el-r-cuadrado-y-evaluar-la-bondad-de-ajuste

AWS. (s.f.). Neutral Network. Recuperado el 13 de Enero de 2023, de aws: https://aws.amazon.com/es/what-is/neural-network/

BBVA. (2017). El poder predictivo de las redes sociales. https://www.bbva.com/es/poder-predictivo-redes-sociales/

Blanco, J. I. (28 de 4 de 2023). ¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? Obtenido de https://jorgeiblanco.medium.com/por-qu%C3%A9-la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0

Boletin Epidemiológico (n.d.). 2022 Boletín epidemiológico semana 52. https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2022_Bolet%C3%ADn_epidemiologico_semana_52.pdf

Boletín_epidemiologico_semana_52. (n.d.).

Carolina Sánchez, D., Laboral, A., & Visión Desde Gestión De La Seguridad Y La Salud En El Trabajo, U. la. (n.d.). Absenteeism: a view from the management of healthand safety at work.

Carolina Sánchez, D., Laboral, A., & Visión Desde Gestión De La Seguridad Y La Salud En El Trabajo, U. la. (n.d.). ABSENTEEISM: A VIEW FROM THE MANAGEMENT OF HEALTHAND SAFETY AT WORK.

CEPAL. (n.d.). Colombia - Sistema político electoral. Retrieved April 17, 2023, from https://oig.cepal.org/es/paises/9/system#:~:text=Los%20municipios%20de%20Colombia%20corresponden,que%20se%20cuentan%20como%20Municipios.

Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019). On Empirical Comparisons of Optimizers for Deep Learning. http://arxiv.org/abs/1910.05446

Chugh, A. (8 de 12 de 2020). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? Obtenido de https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Comunitaria, P., El, E. N., De, P., Del, P., Luis, J., & Chávez, L. (n.d.). UNIVERSIDAD DE HUELVA TESIS DOCTORAL Presentada por Cuthbert, D. (1981). Origins of the variance

inflation factor as recalled de Myttenaere, A., Golden, B., le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. Neurocomputing, 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

Definición de cabecera municipal - Diccionario panhispánico del español jurídico - RAE. (n.d.). Retrieved February 4, 2024, from https://dpej.rae.es/lema/cabecera-municipal

Dengue (página 2). (n.d.). Retrieved February 4, 2024, from https://www.monografias.com/trabajos82/dengue-enfermedad/dengue-enfermedad2

Dengue y dengue grave. (n.d.). Retrieved February 4, 2024, from https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue

Desarrollos de la Ingeniería ambiental en la evaluación de la calidad de los recursos naturales y la salud ambiental. (n.d.). Retrieved February 4, 2024, from https://www.researchgate.net/publication/320170890_Desarrollos_de_la_Ingenieria_ambiental_en_la_evaluacion_de_la_calidad_de_los_recursos_naturales_y_la_salud_ambiental

Dropout y Batch Normalization. (n.d.). https://vincentblog.xyz/posts/dropout-y-batch-normalization

Dropout y Batch Normalization. (n.d.). Retrieved February 4, 2024, from https://vincentblog.xyz/posts/dropout-y-batch-normalization

Editor, M. B. (18 de 04 de 2019). MINITAB. Obtenido de https://blog.minitab.com/es/analisis-de-regresion-como-puedo-interpretar-el-r-cuadrado-y-evaluar-la-bondad-de-ajuste

El Bouchefry, K., & de Souza, R. S. (2020b). Learning in Big Data: Introduction to Machine Learning. In Knowledge Discovery in Big Data from Astronomy and Earth Observation (pp. 225–249). Elsevier. https://doi.org/10.1016/B978-0-12-819154-5.00023-0

El caracol africano (Achatina Fúlica) (página 2). (n.d.). Retrieved February 4, 2024, from https://www.monografias.com/trabajos105/caracol-africano-achatina-fulica/caracol-africano-achatina-fulica2

Eom, H., Son, Y., & Choi, S. (2020). Feature-Selective Ensemble Learning-Based Long-Term Regional PV Generation Forecasting. IEEE Access, 8, 54620–54630. https://doi.org/10.1109/ACCESS.2020.2981819

Fernando, J., & Usuga, C. (n.d.). Estado del arte de los sistemas sépticos para el tratamiento del agua residual en zonas rurales. Retrieved February 4, 2024, from www.udea.edu.co

Frost, J. (2023). Estadísticas Por Jim. Obtenido de Error Cuadrado Medio Raíz (RMSE): https://statisticsbyjim.com/regression/root-mean-square-error-rmse/

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

García, S., Ra-Mírez-Gallego, S., Luengo, J., & Herrera, F. (n.d.). Big Data monografía monografía Big Data: Preprocesamiento y calidad de datos. Retrieved February 4, 2024, from www.highlycited.com

Google Colab. (2023). Dengue GRU. https://drive.google.com/file/d/19pm9pDphM4KbnVTAd14mS85R-BHFIYpB/view?usp=sharing

Google Colab. (2023). Dengue LSTM. https://drive.google.com/file/d/1KwYNoyGqT3BltybazAvGXI75SyxZF75n/view?usp=sharing

Google Colab. (2023). Dengue RNN. https://drive.google.com/file/d/11jkt6qzrfMZiScwgTlNNC1TTvzcpgLrx/view?usp=sharing

Guerrero, S. C., & Melo, O. O. (2017). Una metodología para el tratamiento de la multicolinealidad a través del escalamiento multidimensional. Ciencia En Desarrollo, 8, 9–24.

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. Current Reviews in Musculoskeletal Medicine, 13(1), 69–76. https://doi.org/10.1007/S12178-020-09600-8

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., & Villalba, L. J. G. (2019). Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. Sensors (Switzerland), 19(7). https://doi.org/10.3390/s19071746

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., & Villalba, L. J. G. (2019). Using twitter data to monitor natural

disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. Sensors (Switzerland), 19(7). https://doi.org/10.3390/s19071746

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Iberdrola. (s.f.). Machine learning aprendizaje automatico. Recuperado el 13 de Enero de 2023, https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

IBM Cloud. (n.d.). Retrieved February 4, 2024, from https://www.ibm.com/mx-es/cloud

IMB. (s.f.). Deep learning. Recuperado el 13 de Enero de 2023, de IMB: https://www.ibm.com/co-es/cloud/deep-learning

Infecciones Arbovirales (encefalitis transmitida por artrópodos, encefalitis equina oriental, encefalitis de St. Louis, encefalitis de California, encefalitis Powassan, encefalitis del Nilo Occidental). (n.d.). Retrieved February 4, 2024, from https://www.health.ny.gov/es/diseases/communicable/arboviral/fact_sheet.htm

Jordi, A. :, Mata, M., & Torres Cebrián, A. (n.d.). PROJECTE FI DE CARRERA TÍTOL: Diseño e implementación de un sistema de control por voz.

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. http://arxiv.org/abs/1506.00019

Liu, L., Han, M., Zhou, Y., & Wang, Y. (2018). Lstm recurrent neural networks for influenza trends prediction. In Bioinformatics Research and Applications: 14th International

Symposium, ISBRA 2018, Beijing, China, June 8-11, 2018, Proceedings 14 (pp. 259-264). Springer International Publishing.

Machine Learning: definición, funcionamiento, usos. (n.d.). Retrieved February 4, 2024, from https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? | by Akshita Chugh | Analytics Vidhya | Medium. (n.d.). Retrieved February 4, 2024, from https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

Marcano-Cedeno, A., Quintanilla-Dominguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, 2845–2850. https://doi.org/10.1109/IECON.2010.5675075

Meca, I., & Argullo, O. (4 de Mayo de 2018). rstudio. Recuperado el 16 de Enero de 2023, de http://rstudio-pubs-static.s3.amazonaws.com/386432_afd91f8ebbbb4c78b29f7da3ed840d67.html

medium.com. (23 de MAYO de 2022). Obtenido de https://medium.com/@ooemma83/interpretation-of-evaluation-metrics-for-regression-analysis-mae-mse-rmse-mape-r-squared-and-5693b61a9833

Microsoft. (2022, September 26). Ajuste de hiperparámetros de un modelo (v2). Azure. https://learn.microsoft.com/es-es/azure/machine-learning/how-to-tune-hyperparameters

Ministerio de salud y protección social. (s.f.). Fiebre chikunguña Recuperado el 13 de Enero de 2023, https://www.minsalud.gov.co/salud/publica/PET/Paginas/chikunguna.aspx

MODELO PREDICTIVO PARA EL RENDIMIENTO DE CULTIVOS. (n.d.). Retrieved February 4, 2024, from https://noesis.uis.edu.co/server/api/core/bitstreams/236893dc-7de0-41ee-bed8-53932c109b9d/content

Mussumeci, E., & Codeço Coelho, F. (2020). Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression. Spatial and Spatio-Temporal Epidemiology, 35. https://doi.org/10.1016/j.sste.2020.100372

New York State. (s.f.). Infecciones Arbovirales (encefalitis transmitida por artrópodos, encefalitis equina oriental, encefalitis de St. Louis, encefalitis de California, encefalitis Powassan, encefalitis del Nilo Occidental) Recuperado el 13 de Enero de 2023, https://www.health.ny.gov/es/diseases/communicable/arboviral/fact_sheet.htm#:~:text=%C2%BFQu%C3%A9%20son%20las%20infecciones%20arbovirales,tales%20como%20mosquitos%20y%20garrapatas.

Ojeda R, A., Londoño O, R., Gutierrez R, C., & Gonella-Diaza, A. (2014). Follicular dynamics, corpus luteum growth and regression in multiparous buffalo cows and buffalo heifers. Revista MVZ Córdoba, 19(2), 4130–4140. https://doi.org/10.21897/rmvz.106

Ojeda R, A., Londoño O, R., Gutierrez R, C., & Gonella-Diaza, A. (2014). Follicular dynamics, corpus luteum growth and regression in multiparous buffalo cows and buffalo heifers. Revista MVZ Córdoba, 19(2), 4130–4140. https://doi.org/10.21897/rmvz.106

OMS. (10 de Enero de 2022). Dengue y dengue grave. https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue

Organización mundial de la salud. (10 de Enero de 2022). Dengue y dengue grave Recuperado el 13 de Enero de 2023 https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=El%20dengue%20es%20una%20infecci%C3%B3n,virus%20del%20dengue%20(DENV).

Organización Panamericana de la Salud. (2022, December 21). Actualización epidemiológica semanal para dengue, chikunguña y zika en 2022. https://www3.paho.org/data/index.php/es/temas/indicadores-dengue/boletin-anual-arbovirosis-2022.html

PAHO. (s.f.). Zika Recuperado el 16 de Enero de 2023, de https://www.paho.org/es/temas/zika#:~:text=La%20fiebre%20del%20Zika%20es,no%20purulenta%20que%20ocurre%20entre

Parmar, R. (2018, September 2). Common Loss functions in machine learning. Towards Data Science. https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23

Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., & Gonzalez, G. (n.d.). Social media mining for public health monitoring and surveillance. www.worldscientific.com

Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? | by Jorge I. Blanco | Medium. (n.d.). Retrieved February 4, 2024, from https://jorgeiblanco.medium.com/por-qu%C3%A9-la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0

Pritish K. Tosh, M. (22 de Septiembre de 2022). ¿Qué es la fiebre chikungunya? ¿Debería preocuparme? Recuperado el 13 de Enero de 2022, https://www.mayoclinic.org/es-es/diseases-conditions/infectious-diseases/expert-answers/what-is-chikungunya-fever/faq-20109686

¿Qué es el análisis descriptivo? (n.d.). Retrieved February 4, 2024, from https://www.questionpro.com/blog/es/analisis-descriptivo/

¿Qué es la fiebre chikungunya? ¿Debería preocuparme? - Mayo Clinic. (n.d.). Retrieved February 4, 2024, from https://www.mayoclinic.org/es/diseases-conditions/infectious-diseases/expert-answers/what-is-chikungunya-fever/faq-20109686

Redes neuronales. Programa de Visión… | by Bootcamp AI | Medium. (n.d.). Retrieved February 4, 2024, from https://bootcampai.medium.com/redes-neuronales-13349dd1a5bb

Rico-Mendoza, A., Porras-Ramírez, A., Chang, A., Encinales, L., & Lynch, R. (2019). Co-circulation of dengue, chikungunya, and Zika viruses in Colombia from 2008 to 2018. Revista Panamericana de Salud Pública, 43, 1. https://doi.org/10.26633/RPSP.2019.49

Rodríguez, R. C., Carrasquilla, G., Porras, A., Galera-Gelvez, K., Yescas, J. G. L., & Rueda-Gallardo, J. A. (2016). The burden of dengue and the financial cost to Colombia, 2010-2012. In American Journal of Tropical Medicine and Hygiene (Vol. 94, Issue 5, pp. 1065–1072). American Society of Tropical Medicine and Hygiene. https://doi.org/10.4269/ajtmh.15-0280

SÁNCHEZ CABRERA, J. A. (2015). Análisis de las actividades de prevención y control del dengue según conocimientos, actitudes y prácticas en los barrios la Florida y las Palmas II del municipio de Neiva durante el segundo semestre del 2014 [Universidad Santo Tomás].

https://repository.usta.edu.co/bitstream/handle/11634/9501/S%C3%A1nchezJaime2015.pdf?sequence=1&isAllowed=y

SCAD College of Engineering and Technology, & Institute of Electrical and Electronics Engineers. (n.d.). Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI 2019) : 23-25, April 2019.

Seifert, A., & Rasp, S. (2020). Potential and Limitations of Machine Learning for Modeling Warm-Rain Cloud Microphysical Processes. Journal of Advances in Modeling Earth Systems, 12(12). https://doi.org/10.1029/2020MS002301

SERIE DE TALLERES DE MODELOS DE REGRESIÓN LINEAL, MÚLTIPLE Y LOGÍSTICA – Escuela Global. (n.d.). Retrieved February 4, 2024, from https://www.especializacionesglobal.net/courses/curso-completo-de-regresion-lineal-multiple/

Servicios de Salud - OPS/OMS | Organización Panamericana de la Salud. (n.d.). Retrieved February 4, 2024, from https://www.paho.org/es/temas/servicios-salud

Significado de p-value en Machine Learning - Analytics Lane. (n.d.). Retrieved February 4, 2024, from https://www.analyticslane.com/2021/09/24/significado-de-p-value-en-machine-learning/

Taylor's University (Subang Jaya, S., IEEE Consumer Electronics Society. Malaysia Chapter, & Institute of Electrical and Electronics Engineers. (n.d.). 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA) : proceedings : 26-28 October 2018 Taylor's University Lakeside Campus, Subang Jaya, Malaysia.

Team, D. (2023, October 30). Machine Learning: definición, funcionamiento, usos. Formación En Ciencia De Datos | DataScientest.com. https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

Veeramsetty, V., Singal, G., & Badal, T. (2020). Coinnet: platform independent application to recognize Indian currency notes using deep learning techniques. Multimedia Tools and Applications, 79(31–32), 22569–22594. https://doi.org/10.1007/s11042-020-09031-0

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., & Liu, Q. (2020). Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. International Journal of Environmental Research and Public Health, 17(2), 453. https://doi.org/10.3390/ijerph17020453

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., & Liu, Q. (2020a). Forecast of dengue cases in 20 chinese cities based on the deep learning method. International Journal of Environmental Research and Public Health, 17(2). https://doi.org/10.3390/ijerph17020453

Yamila Catela, E., Cimoli, M., & Porcile, G. (2012). Lem Lem Working Paper Series Productivity and structural heterogeneity in the Brazilian manufacturing sector: trends and determinants Productivity and structural heterogeneity in the Brazilian manufacturing sector: trends and determinants#.