

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

Artificial Neural Network Model for the Prediction and Detection of Cases and Outbreaks of Arboviral Diseases (Zika, Dengue, and Chikungunya) in Colombia

Juan David Garcia-Lopez ^{1a}, Andrés Felipe Fandiño-Plata ^{1b}

¹ Grupo de Investigación Ópalo, Facultad de Fisicomecánicas, Universidad Industrial de Santander, Colombia.
Correos electrónicos: juandagar98@hotmail.com ¹, anfepapla@hotmail.com ²

Recibido: 18 enero, 2024. Aceptado: 18 enero, 2024. Versión final: 29 enero, 2024.

Resumen

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.

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

Abstract

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

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

1. 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., 2020), 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.

2. Revisión de la literatura

La construcción de la ecuación de búsqueda para encontrar documentación relevante sobre la predicción de casos de Dengue, Zika y Chikunguña implicó la propuesta de tres grupos de términos clave. Se utilizó la herramienta SCOPUS de la Biblioteca virtual de la Universidad Industrial de Santander para llevar a cabo esta búsqueda y análisis de información.

El objetivo era extraer referencias bibliográficas relacionadas con modelos de predicción, utilizando términos clave como Neural Network, Dengue, Zika, Chikunguña, Prediction, LSTM, Forecasting, entre otros.

Dentro de la revisión de literatura se halló que la mayoría de las investigaciones sólo se enfocaban en investigaciones relacionadas al Dengue, dejando de lado proyecciones de nuestro interés relacionadas con el Zika y el Chikunguña, sin embargo, y según (Mussumeci & Codeço Coelho, 2020) “...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 diferentes investigaciones fueron desarrolladas desde diversas magnitudes geográficas, desde análisis de diferentes ciudades, como (Xu et al., 2020), hasta regiones específicas, como el caso de (Mussumeci & Codeço Coelho, 2020), o incluso una ciudad en particular, (Taylor’s University (Subang Jaya et al., n.d.)). Se identificaron dos grandes grupos de datos de entrada: ambientales y epidemiológicos.

Mientras algunos autores como (Xu et al., 2020) clasifican las variables en estos dos grupos, otros añaden una categorización más detallada que incluye datos socioeconómicos, como lo es el caso de (Taylor’s University (Subang Jaya et al., n.d.)), quien dedica un apartado de la investigación a describir la naturaleza de cada uno de estos dos grupos, clasificándolos como “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). 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, lo cual nos sirvió como guía para determinar las bases de datos para nuestra investigación. En el caso de registro diario de contagios en

la ciudad entre los años 2002-2012, 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 y velocidad diaria del viento.

Los modelos de predicción usados por (Taylor’s University (Subang Jaya et al., n.d.)) fueron algoritmos de DL, ARIMA (autoregressive integrated moving average) 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, además del uso de regresión lineal simple (LR) y árbol de decisión (DT). Al comparar el resultado de los tres modelos de predicción propuestos, se planteó 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,

Por su parte, Xu et al. (2020) comparó el rendimiento de diferentes métodos de predicción para casos de contagio de Dengue en 20 ciudades chinas durante 24 meses, de 2017 a 2018. Los métodos evaluados incluyeron 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. La data utilizada por (Xu et al., 2020) se dividió en dos grupos principales: datos ambientales y datos epidemiológicos. Los datos epidemiológicos incluyeron características demográficas básicas y el momento de los eventos relacionados con la enfermedad, como fecha de inicio, diagnóstico y, en algunos casos, fecha de muerte. Para los datos meteorológicos, se consideraron 15 variables, de las cuales se eliminaron 5 mediante filtrado de alta correlación y baja varianza, dejando un total de 10 variables meteorológicas. 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., 2020), “...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 %."

Finalmente, se discute un enfoque alternativo, el autor Amin Samina, quien surge como co- autor en los documentos (Amin, Irfan Uddin, et al., 2020) y (Amin, Uddin, et al., 2020), utiliza datos de redes sociales, como Twitter, para predecir brotes de enfermedades. Este método emplea técnicas de Procesamiento de Lenguaje Natural y Aprendizaje Profundo para analizar sentimientos y opiniones de los usuarios, 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-of-words (CBOW). Se destaca la eficiencia y disponibilidad de esta información en tiempo real, lo que podría mejorar la detección temprana de brotes epidémicos.

En resumen, estos estudios demuestran una variedad de enfoques para predecir la incidencia de enfermedades transmitidas por mosquitos, desde modelos tradicionales hasta técnicas innovadoras que aprovechan datos no estructurados de redes sociales.

3. Recolección preprocesamiento y análisis de datos

Este trabajo de investigación utiliza microdatos proporcionados por SISPRO, el sistema integrado de información de la protección social del gobierno de Colombia. La base de datos incluye información demográfica relevante para el estudio de enfermedades arbovirales como Zika, Dengue y Chikungunya. Se emplearon datos meteorológicos, considerando factores como temperatura, precipitación, índice de vegetación, humedad y velocidad del viento, utilizando la data del satélite ArcGIS World Geocoder.

La ventana de tiempo para la recolección de datos abarcó desde 2007 hasta 2021. Se seleccionaron 20 localidades (municipios y ciudades) con base en el umbral mínimo de contagios totales durante los 15 años, siendo Cali, Medellín, Ibagué, Cúcuta, Bucaramanga, Villavicencio, Barranquilla, Neiva, Floridablanca, Cartagena, Sincelejo, Valledupar, Yopal, Palmira, Armenia, Pereira, Montería, Girón, Santa Marta y Soledad las elegidas. Aunque representan el 1.79% del total de localidades en Colombia, acumulan el 47.38% de los contagios registrados (588,220).

3.1. SISPRO

Se accede a la base de datos de SISPRO tras completar capacitaciones proporcionadas por la entidad. Se extrae información relevante sobre casos de enfermedades arbovirales, como Dengue, Zika y Chikunguña, incluyendo datos como número de casos, área de ocurrencia, año, tipo de evento, edad, género, departamento y municipio.

La información inicial, presentada en tablas dinámicas, se organiza y filtra para facilitar el análisis descriptivo. Debido a la complejidad de la data, se segmenta por año de ocurrencia, generando 15 documentos Excel. Se eliminan codificaciones alfanuméricas y registros no informativos, como aquellos definidos como "-No Definido" en el lugar de ocurrencia.

Se detectan inconsistencias en la variable "Edad", incluyendo registros con edades superiores a 300 años. Se intenta corregir mediante una fórmula condicional, dividiendo edades mayores a 120 años por 10. Sin embargo, esto genera sesgos en la data. Se opta por una solución alternativa: la eliminación de ese tipo de registros, quienes superaban los 70.000 contagios (70.133 para ser exactos), siendo un 5,35% de los registros de datos de contagios totales, eliminando el sesgo y manteniendo la integridad de otros análisis, como la distribución de contagios por género, el comportamiento temporal de los contagios y la proporción de contagios por ubicación y área de ocurrencia.

3.2. ArcGIS World Geocoder

El estudio utiliza el satélite de la NASA ArcGIS World Geocoder para recopilar datos meteorológicos esenciales, tomando variables clave como temperatura, presión, velocidad del viento, precipitación y humedad. Se toma la humedad relativa a 2 metros, medida en porcentaje para facilitar el análisis, y selecciona el valor promedio de precipitaciones diarias mes a mes. La temperatura se aborda considerando diversas categorías, como la temperatura medida a 2 metros sobre la tierra en grados Celsius. Se incluyen datos de presión superficial en kPa y velocidad del viento a 10 metros sobre el suelo, con las variables de velocidad máxima y mínima medida en metros por segundo. Este enfoque metodológico robusto proporciona una base integral para el análisis meteorológico en el contexto del estudio.

4. Análisis Descriptivo

Se inicia un análisis descriptivo de los casos de Dengue, Zika y Chikunguña en Colombia durante el periodo 2007-2021 en diversas localizaciones. Se exploran las tendencias, correlaciones y comportamientos de las variables en el conjunto de datos categorizado, sin plantear una hipótesis específica.

El análisis inicia con la evaluación temporal de los contagios, evidenciando picos notables en los años 2010, 2013, 2016 y 2019, siendo el año 2016 el de mayor incidencia.

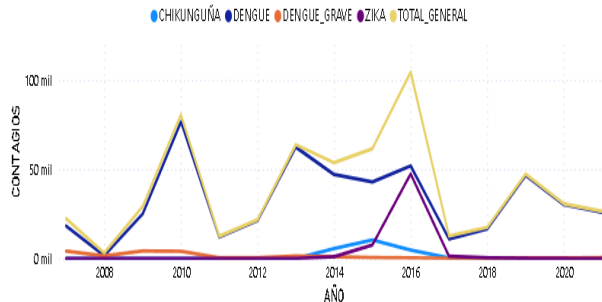


Figura 1. Evolución de los contagios desde el 2007 al 2021. Información recopilada de la base de datos SISPRO y analizada en Power BI.

Se examina la distribución de contagios según el género, mostrando proporciones equitativas entre géneros.

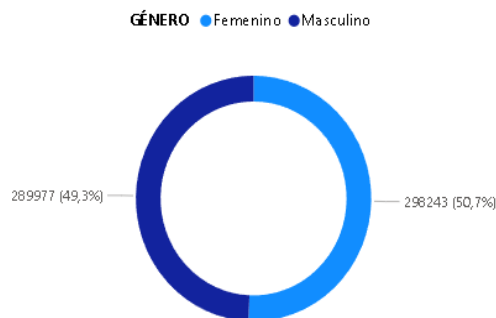


Figura 2. Distribución de los contagios según el género femenino y masculino. Información recopilada de la base de datos SISPRO y analizada en Power BI.

Posteriormente, se aborda la segmentación por área de ocurrencia. "Cabecera" concentra el 94.58% de los contagios, seguido por "Centro Poblado" (3.92%) y "Área Rural Dispersa" (1.5%).

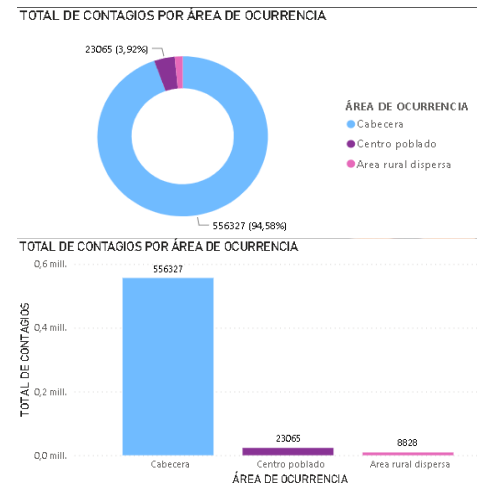


Figura 3. Distribución de los contagios según el área de ocurrencia. Información recopilada de la base de datos SISPRO y analizada en Power BI.

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 19). 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.

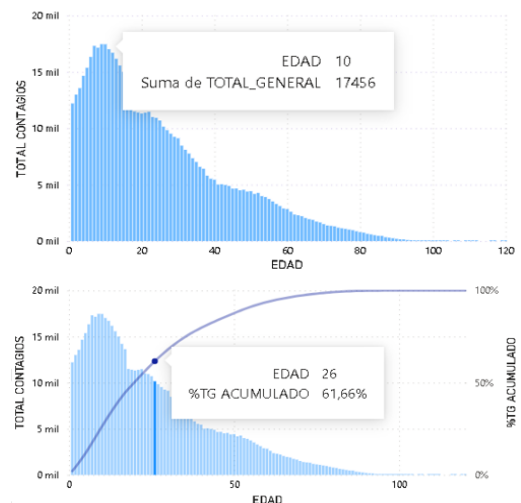


Figura 4. Edad y total porcentual acumulado. Información recopilada de la base de datos SISPRO y analizada en Power BI.

5. Reducción de variables

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.” (aprendeIA, 2023)

En la Figura 5, podemos apreciar la matriz de correlación con cada una de las variables numéricas de la data del proyecto, donde se puede analizar 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.

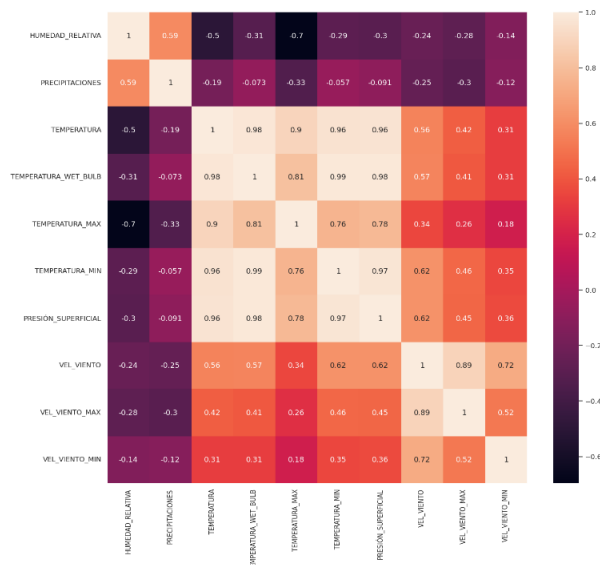


Figura 5. Matriz de correlación de variables numéricas. Información recopilada de la base de datos final y analizada en Google Colab.

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.

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” (Ruiz-López et. al, 2016).

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 artículo 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).

6. Creación de los modelos

En esta sección, se detalla el procesamiento de datos para el análisis de series temporales de casos de dengue en 20 ciudades colombianas entre 2007 y 2021. Se describen las etapas esenciales, desde la carga y normalización de datos hasta la implementación de modelos GRU, LSTM y RNN en Google Colab. La selección del modelo se basa en métricas estándar como MSE, RMSE, MAE y R2,

respaldada por el tiempo de ejecución.

La carga inicial y ordenamiento de datos se realiza mediante Pandas, seguido de una normalización aplicando el Logaritmo Natural más uno ($\ln(x + 1)$), ilustrado a continuación:

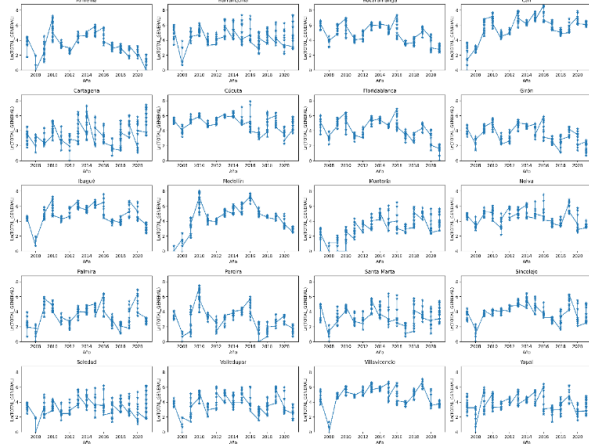


Figura 6. Logaritmo natural de los casos de dengue para las ciudades a través de los años.

Se incorpora una técnica adicional de normalización utilizando 'sklearn' y 'MinMaxScaler', re-escalando los valores de casos y variables adicionales a un rango de 0 a 1.

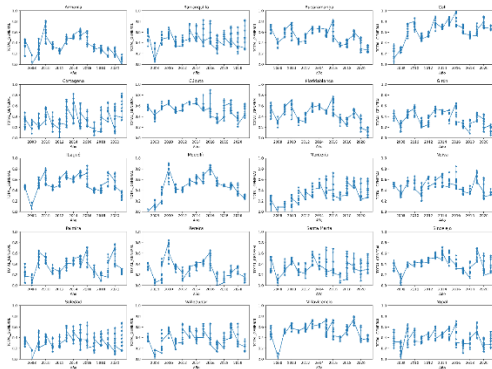


Figura 7. Representación de re-escalamiento para casos de dengue en las ciudades.

Se ha identificado ausencia de datos para algunos meses en ciudades específicas durante los años 2007 y 2008. Por lo cual, se excluyen estos dos años del análisis para asegurar un conjunto de datos completo y normalizado antes de la implementación en la red neuronal.

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 1.

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

Fuente: elaboración propia.

6.1. 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_GRU; 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 3.

El consumo de recursos computacionales en Google Colab para las tres redes neuronales recurrentes (GRU, LSTM Y RNN) se resume en la Tabla 2, los datos se obtuvieron del entrenamiento base y de la malla de experimentos (Google Colab, Arboviral_GRU; Arboviral_LSTM; Arboviral_RNN).

Tabla 2.

Costo computacional

Entrenamiento	Tiempo de ejecución (s)
Guía RNN	461.494
Guía LSTM	706.771
Guía 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 1. Fuente: elaboración propia.

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.

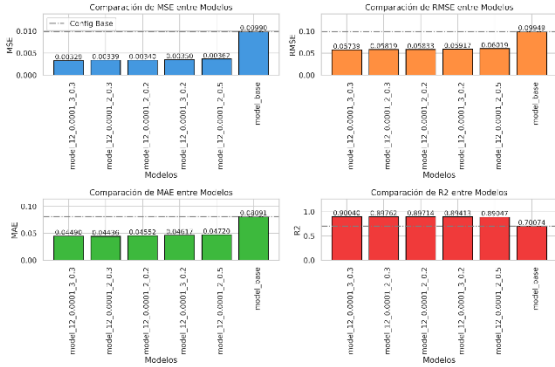


Figura 8. Métricas para las mejores 5 configuraciones, más la base (Modelos LSTM)

En la Figura 8 se presentan los resultados obtenidos para la malla de experimentos con la red LSTM, la Figura 9 los resultados de los experimentos con la red GRU y la Figura 10 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 5 muestra un resumen de los mejores modelos para cada caso.

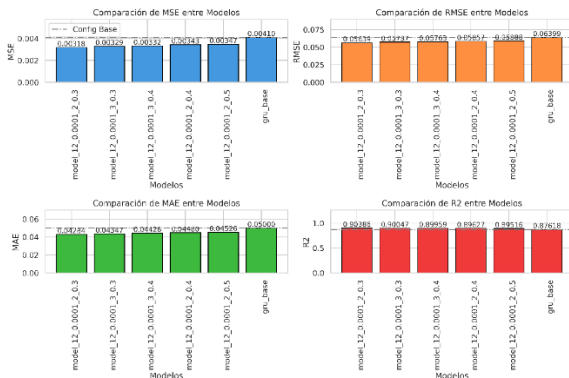


Figura 9. Métricas para las mejores 5 configuraciones más la base (Modelos GRU)

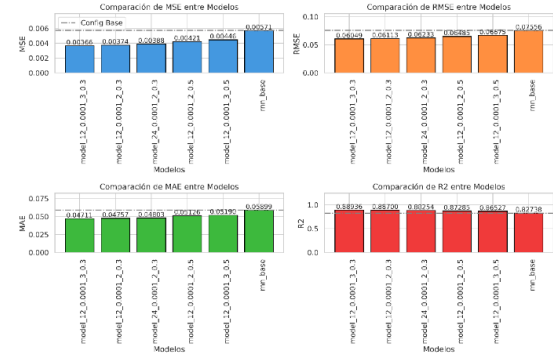


Figura 10. Métricas para las mejores 5 configuraciones más la base (Modelos RNN)

Tabla 3.
Resumen de rendimiento para los mejores modelos de cada arquitectura

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

Fuente: elaboración propia.

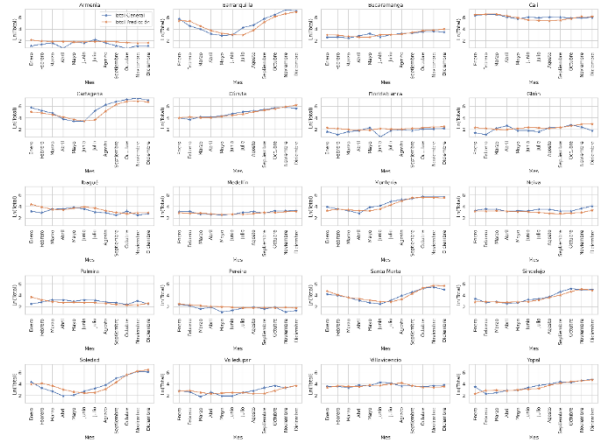


Figura 11. Métricas para las mejores 5 configuraciones más la base (Modelos RNN)

Para obtener una vista completa de los resultados de inferencia en el conjunto de validación, se pueden consultar las Figuras 11, 12 y 13, 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.

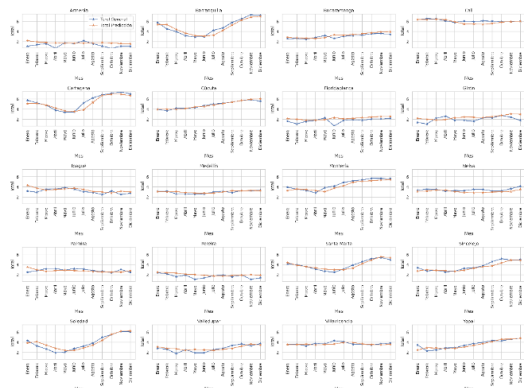


Figura 12. Comparación entre casos originales y casos predichos por red LSTM ($L_n + 1$)

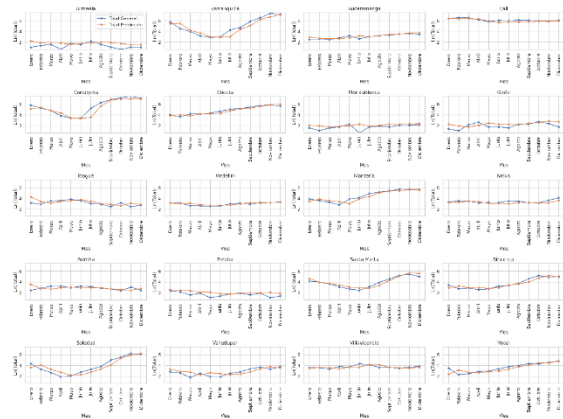


Figura 13. Comparación entre casos originales y casos predichos por GRU ($L_n + 1$)

7. Conclusiones

A partir de los resultados obtenidos en las métricas MAE, MSE, RMSE y R^2 para las diferentes arquitecturas (RNN, LSTM y GRU) de los mejores modelos (Ver Tabla 3), se observa que sus RMSE son similares, con valores de 5.639%, 5.739% y 6.049% respectivamente (Ver Figuras 8, 9, 10 y Tabla 3), 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 3, `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 2).

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 8, 9 y 10). Estos valores indican que la malla de experimentos detallada en la Tabla 1 fue fundamental para mejorar las arquitecturas de referencia.

Referencias

- Boletín Epidemiológico (n.d.). 2022 Boletín epidemiológico semana 52. https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2022_Bolet%C3%ADn_epidemiologico_semana_52.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., 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>
- AWS. (s.f.). Neural 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/>
- 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 health and safety at work.
- 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>
- Google Colab. (2023). Dengue GRU. https://drive.google.com/file/d/19pm9pDphM4KbnVTA_d14mS85R-BHFIYpB/view?usp=sharing
- Google Colab. (2023). Dengue LSTM. <https://drive.google.com/file/d/1KwYNoyGqT3BltybazAvGXl75SyxZF75n/view?usp=sharing>
- Google Colab. (2023). Dengue RNN. <https://drive.google.com/file/d/11jkt6qzrfMZiScwgTINNC1TTvzcpGLrx/view?usp=sharing>
- 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.

- Iberdrola. (s.f.). Machine learning aprendizaje automatico. Recuperado el 13 de enero de 2023, <https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico>
- IMB. (s.f.). Deep learning. Recuperado el 13 de enero de 2023, de IMB: <https://www.ibm.com/cloud/deep-learning>
- 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.
- 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_afd91f8ebbb4c78b29f7da3ed840d67.html
- 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>
- 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., Gutiérrez R, C., & Gonella-Díaz, 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\(DENV\)](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/indicador-es-dengue/boletin-anual-arbovirosis-2022.html>
- PAHO. (s.f.). Zika Recuperado el 16 de enero de 2023, <https://www.paho.org/es/temas/zika#:~:text=La%20fiebre%20del%20Zika%20es,no%20purulenta%20que%20ocurre%20entre>
- 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
- 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>
- 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>
- 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.
- 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.
- 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>
- 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#.