Modelos Deep Learning para el proceso de vigilancia de enfermedades arbovirales: aplicación a lo datos de Dengue, Zika y Chicungunya en Santander

Sonia Isabel Polo Triana

Trabajo de investigación para optar por el título de Magíster en Ingeniería Industrial

Director:

Henry Lamos Díaz, PhD Matemática - Física

Universidad Industrial de Santander

Facultad de ingenierías físico-mecánicas

Escuela de estudios industriales y empresariales

Maestría en ingeniería industrial

Bucaramanga

2023

Agradecimientos

Culminar esta etapa y este proyecto no podría haber sido posible sin el apoyo de muchas personas.

En primer lugar y ante todo mi más profundo sentimiento de agradecimiento a mi tutor el profesor Henry Lamos, especialmente por su paciencia durante todo este proceso, por su orientación, por alentarme a continuar y recordarme que siempre se puede ser mejor, pero sobre todo por confiar en que podía lograrlo.

Agradezco a mi familia, especialmente a mis padres y hermano por su amor y apoyo en esta etapa.

A David (Q.E.P.D) quien me motivó y gracias a él inicié este camino. Siempre tendrás un lugar especial en mi mente y mi corazón.

A mis compañeros de la cohorte 11 “La manada” Andrés, Juan Gabriel, Silvia, Daniela, Samuel, Javier, Gerardo, con quienes inicié este proceso y con quienes compartí momentos muy amenos e hicieron este recorrido mucho más agradable.

A la Universidad Industrial de Santander, a la Escuela de Estudios Industriales y Empresariales y a todos los docentes por su empeño, dedicación y esfuerzo para formar profesionales de alto valor dispuestos dar lo mejor de sí para contribuir a la sociedad.

A todos aquellos que de alguna u otra forma me acompañaron en este proceso.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Lista de Anexos

Resumen

Para el desarrollo del presente trabajo, se crearon modelos predictivos basados en Deep Learning (LSTM, RNN, GRU, CNN) para la predicción de casos de Dengue, Zika y Chikungunya en Santander, a partir de variables climáticas (Temperaturas, precipitaciones y humedad relativa). Para esto se siguió la metodología KDD (Knowledge Discovery in Databases). Como primera instancia, se recolectó y organizó el conjunto de datos y se realizó el análisis estadístico descriptivo. Posteriormente, se realizó el preprocesamiento de los datos, se hicieron transformaciones de estos y se determinó la importancia de las características para escoger las variables a utilizar. Luego, se realizó la construcción de los modelos para lo cual se realizó la selección de los hiperparámetros. Los modelos Deep Learning se compararon con un modelo de red neuronal tradicional (MLP), el cual presentó los mejores resultados en cuatro de los seis municipios de estudio.

Palabras clave: Deep Learning, Predicción, Dengue

Abstract

For the development of this work, predictive models based on Deep Learning (LSTM, RNN, GRU, CNN) were created for the prediction of cases of Dengue, Zika and Chikungunya in Santander, based on climatic variables (temperatures, rainfall and relative humidity). For this purpose, the KDD (Knowledge Discovery in Databases) methodology was used. As a first step, the data set was collected and organized and the descriptive statistical analysis was performed. Subsequently, the data was preprocessed, data transformations were performed and the importance of the characteristics was determined in order to choose the variables to be used. Then, the construction of the models was carried out, for which the hyperparameters were selected. The Deep Learning models were compared with a traditional neural network model (MLP), which presented the best results in four of the six study municipalities.

Keywords: Deep Learning, Prediction, Dengue fever

Introducción

La epidemiología es una disciplina que se encarga del estudio de los factores, predicciones y control de los factores relacionados con la salud y la enfermedad, así como también del estudio de la dinámica y de la distribución (propagación) de las enfermedades, en este caso virales de una población. En este sentido, la vigilancia de enfermedades infecciosas es un proceso integral en el que la información sobre brotes y vectores de enfermedades infecciosas se recopila, analiza e interpreta de forma continua y sistemática. Además requiere que, los resultados se compartan rápidamente a las personas que los necesitan para prevenir y controlar estas enfermedades (Chae et al., 2018). La base para el control de enfermedades infecciosas ha sido principalmente la implementación de sistemas de los sistemas de vigilancia que rastrean enfermedades, patógenos y resultados clínicos. Sin embargo, estos sistemas son de tipo tradicional y tienen el inconveniente de retrasos severos y falta de resolución espacial.

En salud pública, los sistemas de vigilancia se basan principalmente en datos recopilados y codificados manualmente, lentos para acumular y difíciles de difundir para su análisis. Además, los informes de estos sistemas tienden a ser nacionales o regionales con poca información sobre enfermedades a nivel local (Bansal et al., 2016). En este sentido, el desarrollo acelerado de nuevos sistemas que dependen de grandes flujos de datos, el aumento de la disponibilidad de información del sector salud, como lo son datos móviles, registros de salud electrónicos y redes sociales, ha fomentado la generación de estrategias innovadoras con el objetivo de difundir rápidamente información de vigilancia para una rápida intervención en salud pública (Wyber et al., 2015)

Teniendo en cuenta que Colombia es un país con varias regiones en las cuales las condiciones son las idóneas para la propagación de enfermedades de este tipo, se presenta el caso del departamento de Santander en donde hasta la semana epidemiológica 52 de 2019 se habían

notificado 9757 casos de Dengue, 41 casos de Chicungunya y 36 casos de Zika(Instituto Nacional de Salud, 2019). Considerando que son muchos los factores que influyen la incidencia de las enfermedades infecciosas, entre los más representativos las variables meteorológicas, los factores socioeconómicos y la densidad de población de vectores, es complejo develar las relaciones existentes entre los casos que se puedan presentar y estas variables. Razón por la cual el modelo clásico de serie temporal no puede ajustarla fácilmente. Por consiguiente, han venido surgiendo estudios que han utilizado técnicas del campo del aprendizaje profundo para predecir enfermedades infecciosas.

El Deep Learning o aprendizaje profundo es un campo de investigación vital dentro de los métodos de aprendizaje automático, dado su rendimiento demostrado en diferentes dominios y los rápidos progresos de las mejoras metodológicas, muchos aspectos del aprendizaje profundo representan una gran oportunidad para la vigilancia epidemiológica, se ha probado que produce resultados satisfactorios cuando se utiliza para realizar tareas que son complejas para los métodos de análisis convencionales (Coccia, 2020).

Por lo anteriormente expuesto, el objeto de esta propuesta de investigación fue aplicar modelos Deep Learning para el proceso de vigilancia de enfermedades arbovirales, específicamente Dengue, Zika y Chikungunya, en el departamento de Santander.

De este modo, para el cumplimiento de los objetivos propuestos, se siguió la metodología denominada descubrimiento de conocimiento en base de datos (Knowledge Discovery in Databases KDD, por sus siglas en inglés), considerando las fases principales por la que está compuesta: abstracción y comprensión del tema de estudio; recolección e integración de los datos; preprocesamiento de los datos, construcción de modelos Deep Learning y; difusión y uso. Esta última fase comprendió el diseño de una herramienta para apoyar el proceso de visualización del

comportamiento de los modelos y de los análisis descriptivos que permitirán obtener información del comportamiento de las enfermedades arbovirales.

El documento que se presenta a continuación se encuentra estructurado de la siguiente manera: en la sección 1 se presenta la revisión de literatura que abarca los conceptos principales y los diferentes estudios realizados alrededor del tema de investigación; la sección 2 comprende el planteamiento del problema y la justificación de la temática abordada; mientras que los objetivos e Hipótesis a probar, son descritos en la sección 3 y 4 respectivamente. En la sección 5 se realiza la descripción de la metodología que se llevó a cabo para cumplir con los objetivos establecidos, en la sección 6 donde se muestran los resultados, finalmente en la sección 7 y 8 se presentan las conclusiones y respectivas recomendaciones.

1. Revisión de Literatura

En este apartado se presenta la revisión de la literatura que permita identificar los avances que ha realizado la comunidad científica en vigilancia de enfermedades arbovirales y la predicción mediante el uso de métodos de Deep Learning. En primer lugar, se realiza una introducción a los términos y conceptos principales que hacen parte de la presente investigación, posteriormente se revisa la literatura reciente sobre la aplicación de tecnologías Deep Learning para avanzar en el dominio de la vigilancia de enfermedades arbovirales y/o infecciosas.

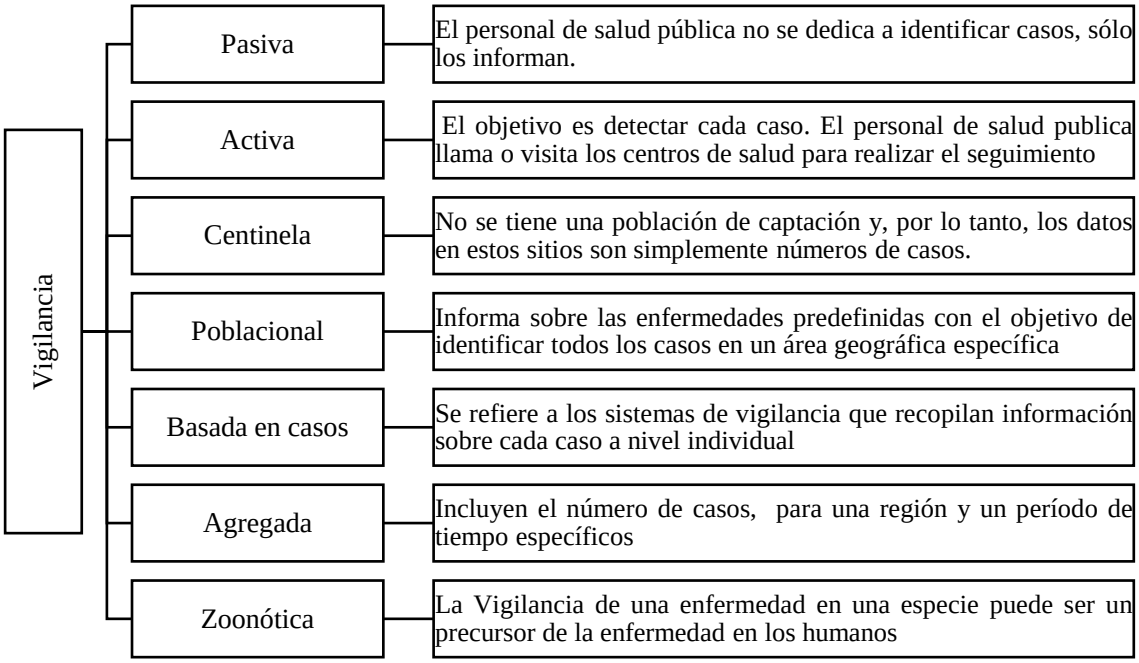
1.1 Vigilancia de enfermedades arbovirales y/o infecciosas

La vigilancia de enfermedades infecciosas es una herramienta epidemiológica importante para controlar la salud de una población, sus objetivos son: (1) describir la carga actual y la epidemiología de la enfermedad, (2) monitorear las tendencias e (3) identificar brotes y nuevos patógenos. En primer lugar, describir la carga y la epidemiología (incluida la estacionalidad, la distribución por edad, los grupos de edad, entre otros.) de la enfermedad es fundamental para demostrar la necesidad y abogar por intervenciones, como la vacunación y la administración masiva de medicamentos(Murray & Cohen, 2017).

En segundo lugar, la vigilancia de enfermedades infecciosas se utiliza para controlar las tendencias de las enfermedades, las cuales no solo significan el número de casos, sino también la etiología de los casos. La vigilancia también monitorea el control, la eliminación y la erradicación de enfermedades. Finalmente, un aspecto clave de la vigilancia de enfermedades infecciosas es el ciclo de detección, respuesta y prevención de brotes. La vigilancia continua de una enfermedad puede facilitar la detección temprana de un brote,

permitiendo una respuesta más rápida y, por lo tanto, mitigando el brote(Murray & Cohen, 2017).

Figura 1. Métodos de Vigilancia de enfermedades Infecciosas



Fuente: (Murray & Cohen, 2017)

Un sistema de vigilancia ideal es representativo de la población, flexible, económico y resistente, con informes oportunos y validación de sus productos (Simonsen et al., 2016). Desde el siglo XIX se han venido desarrollando sistemas de vigilancia de enfoque sistémico, donde en un principio, los médicos informaban constantemente las incidencias semanales de enfermedades y muertes con una estratificación cada vez más amplia por causa, edad y sexo, y alrededor de 1900 la codificación de la causa de muerte evolucionó en lo que ahora se conoce como Clasificación Internacional de Enfermedades (World Health Organization, 2016). Durante el siglo XX, los avances en potencia informática y tecnología de la información permitieron el desarrollo de informes electrónicos de enfermedades comunes

por parte de los médicos, facilitando así su comunicación las autoridades de salud pública y el público de manera oportuna, un ejemplo de esto fue el sistema francés Sentinelles (Simonsen et al., 2016).

La vigilancia de enfermedades infecciosas puede tener diferentes enfoques basados en la epidemiología y la presentación clínica de la enfermedad y los objetivos de la vigilancia (Figura 1). En relación a la forma en como es difundida la información de la vigilancia de enfermedades, se destacan los boletines e informes (semanales o mensuales, anuales o semestrales), la literatura y conferencias científicas, métodos que, aunque mejoran la riqueza del conocimiento disponible y avanzan en la investigación de una determinada enfermedad, no son lo suficientemente oportunos como para movilizar una respuesta a un brote.

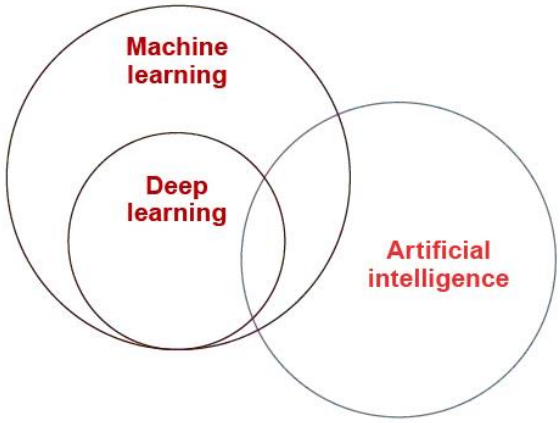
En este sentido, el desarrollo acelerado de nuevos sistemas que dependen de grandes flujos de datos, sumado al aumento de la disponibilidad de datos sobre salud, como lo son datos móviles, registros de salud electrónicos y redes sociales, se han venido generando estrategias innovadoras con el objetivo de difundir más rápidamente la información de vigilancia para una rápida intervención de salud pública (Wyber et al., 2015).

Por ejemplo, el Programa de Monitoreo de Enfermedades Emergentes (ProMED) consolida y verifica informes de los medios de comunicación, observadores y noticias, y los difunde por correo electrónico y su sitio web, o la plataforma en línea HealthMap la cual permite a los usuarios ver la distribución geográfica de múltiples enfermedades. Según Hall, Correa, Yoon, & Braden (2012) estas nuevas fuentes de datos e información proporcionan oportunidades para mejorar el conocimiento de la salud dado que otorga más detalles de los eventos de interés con respecto a tiempo, lugar y persona.

1.2 Deep Learning: Marco General

En el dominio de la Inteligencia Artificial, el aprendizaje automático desarrolla algoritmos y modelos estadísticos que los sistemas informáticos utilizan para realizar de manera efectiva una tarea específica basándose en patrones e inferencia. Un campo de investigación vital dentro de los métodos de aprendizaje automático es la Tecnología de Deep Learning o Aprendizaje Profundo el cual desarrolla redes neuronales profundas, redes neuronales recurrentes y redes neuronales convolucionales para múltiples aplicaciones, tales como visión artificial, reconocimiento de voz, procesamiento de lenguaje natural, reconocimiento de audio, filtrado de redes sociales, bioinformática, análisis de imágenes médicas, inspección de materiales, etc. (Coccia, 2020).

Figura 2. El Deep Learning como un subconjunto del aprendizaje automático y la inteligencia artificial.



Fuente: (Trask, 2019)

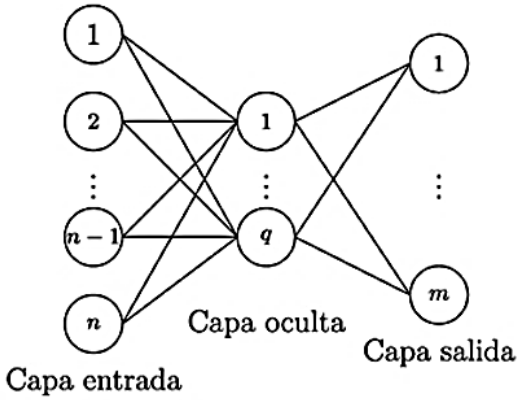
Aunque el aprendizaje profundo es un subcampo bastante antiguo del aprendizaje automático, solo se destacó a principios de la década de 2010. En los pocos años transcurridos desde entonces, ha logrado una revolución en el campo, con resultados notables

en problemas de percepción como la vista y la audición, problemas que involucran habilidades que parecen naturales e intuitivas para los humanos pero que han sido esquivas durante mucho tiempo para las máquinas. En particular, el aprendizaje profundo ha logrado los siguientes avances, todo en áreas históricamente difíciles de aprendizaje automático (Chollet, 2017)

- Clasificación de imágenes a nivel casi humano
- Reconocimiento de voz a nivel casi humano
- Transcripción de escritura a mano de nivel casi humano
- Traducción automática mejorada
- Conversión de texto a voz mejorada
- Asistentes digitales como Google Now y Amazon Alexa
- Conducción autónoma a nivel casi humano
- Orientación de anuncios mejorada, como la utilizada por Google, Baidu y Bing
- Resultados de búsqueda mejorados en la web
- Capacidad para responder preguntas en lenguaje natural.

La red neuronal es el método base del Deep Learning o aprendizaje profundo. Una red neuronal (Figura 3) es un conjunto de funciones lineales intercaladas con funciones de activación no lineales. La unidad básica de una red neuronal es la neurona. Una neurona realiza una combinación lineal de sus entradas y una función de activación. Al conjunto de neuronas que reciben el mismo grupo de entradas se le denomina capa. Las neuronas de una misma capa también generan salidas que servirán como entrada de la siguiente capa.(Bermejo & Vizcarra, 2020).

Figura 3. Representación de una red neuronal



Fuente: (Vivas et al., 2014)

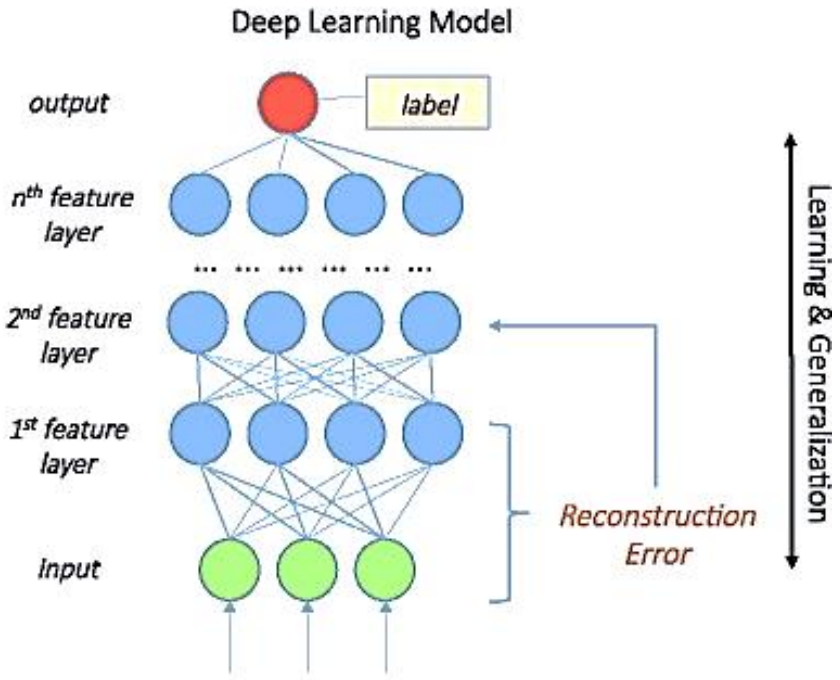
La capa de entrada recibe la representación numérica de los datos que se procesarán. La capa intermedia o capa oculta realiza una combinación lineal de las salidas de la anterior capa y luego ejecuta una función de activación. Cabe señalar que una red puede tener varias capas ocultas en su arquitectura. Finalmente, La capa de salida de la misma forma recibe como entradas el resultado de la anterior capa. Sin embargo, utiliza una función de activación diferente. Generalmente las funciones de activación de la capa de salida realizan una normalización las salidas en una distribución de probabilidades. El proceso que consiste en la entrada a la red, el cálculo secuencial de todas las capas ocultas de la red hasta obtener un resultado de la capa de salida se conoce como propagación hacia adelante (Vivas et al., 2014).

De acuerdo con lo expuesto por Miotto et al. (2017) el aprendizaje profundo se distingue del aprendizaje automático tradicional en la forma en que se aprenden las representaciones de los datos. En otras palabras, el aprendizaje profundo permite que los modelos computacionales que se componen de múltiples capas de procesamiento basadas en redes neuronales aprendan representaciones de datos con múltiples niveles de

abstracción (LeCun et al., 2015). Las principales diferencias entre el aprendizaje profundo y las redes neuronales artificiales tradicionales (ANN, Artificial Neural Networks, por sus siglas en inglés) son el número de capas ocultas, sus conexiones y la capacidad de aprender abstracciones significativas de las entradas.

De acuerdo con Bengio (2009) las redes neuronales tradicionales generalmente están limitadas a tres capas y están entrenadas para obtener representaciones supervisadas que están optimizadas solo para la tarea específica y generalmente no son generalizables. De manera diferente, cada capa de un sistema de aprendizaje profundo (Figura 4) produce una representación de los patrones observados en función de los datos que recibe como entradas de la capa a continuación, al optimizar un criterio local no supervisado.

Figura 4. Arquitectura de redes neuronales profundas



Fuente: (Miotto et al., 2017)

Una particularidad del aprendizaje profundo es que estas capas de características no están diseñadas por ingenieros humanos, sino que se aprenden de los datos mediante un procedimiento de aprendizaje de propósito general. Las redes neuronales profundas procesan las entradas de manera no lineal en capas para pre-entrenar (inicializar) los nodos en capas ocultas posteriores, y de esta forma aprender 'estructuras profundas' y representaciones que son generalizables. Estas representaciones luego se introducen en una capa supervisada para ajustar toda la red utilizando el algoritmo de retro propagación hacia representaciones que están optimizadas para la tarea específica de extremo a extremo (Miotto et al., 2017).

Dentro de las arquitecturas Deep Learning o aprendizaje profundo se encuentran: i) Redes neuronales convolucionales (CNN, Convolutional Neural Networks); ii) Redes Neuronales Recurrentes (RNN, Recurrent Neural Neworks); iii) Máquina de Boltzmann restringida (RBM, Restricted Boltzmann machine); iv) Autocodificadores (AE, Autoencoders) Para poder comprender mejor estos conceptos, a continuación, se explica en detalle cada uno.

- **Redes neuronales convolucionales (CNN, Convolutional Neural Networks)**

Al aumentar el número de capas en una red neuronal para hacerla más profunda, aumenta la complejidad de la red y esto permite modelar funciones que son más complicadas. Sin embargo, el número de pesos y sesgos aumentará exponencialmente. De hecho, aprender problemas tan difíciles puede volverse imposible para las redes neuronales normales, por lo cual las redes neuronales convolucionales conducen a la solución. Las CNN se basan en conexiones locales y pesos vinculados a través de las unidades seguidas de agrupación de características (submuestreo) para obtener descriptores invariantes de traducción (Prasoon et

al., 2013). La arquitectura básica de CNN consta de una capa convolucional y de agrupación, opcionalmente seguida de una capa totalmente conectada para la predicción supervisada. En la práctica, las CNN están compuestas por más de 10 capas convolucionales y de agrupación para modelar mejor el espacio de entrada. Las aplicaciones más exitosas de CNN se obtuvieron en visión artificial (Krizhevsky et al., 2012). Las CNN generalmente requieren un gran conjunto de datos de documentos etiquetados para recibir la capacitación adecuada.

- **Redes Neuronales Recurrentes (RNN, Recurrent Neural Neworks)**

Las RNN son redes neuronales en las que los datos pueden fluir en cualquier dirección. Se componen por una red que realiza la misma tarea para cada elemento de una secuencia, y cada valor de salida depende de los cálculos anteriores. El concepto básico que subyace a las RNN es utilizar información secuencial (Prasoon et al., 2013). En una red neuronal normal, se supone que todas las entradas y salidas son independientes entre sí. Las RNN se denominan recurrentes ya que repiten la misma tarea para cada elemento de una secuencia, y la salida se basa en los cálculos anteriores. Por lo tanto, se puede decir que las RNN tienen una "memoria" que captura información sobre lo que se ha calculado previamente. En teoría, las RNN pueden usar información en secuencias muy largas, pero en realidad, pueden mirar hacia atrás solo unos pocos pasos debido a la desaparición y la explosión de los problemas de gradiente (Miotto et al., 2017).

Las redes recurrentes LSTM, (LSTM, Long Short-Term Memory) también abordaron este problema modelando el estado oculto con celdas que deciden qué guardar (y qué borrar) en la memoria dado el estado anterior, la memoria actual y el valor de entrada. Estas variantes son eficientes para capturar dependencias a largo plazo y dieron excelentes resultados en aplicaciones de procesamiento de lenguaje natural (Sutskever et al., 2014). Las redes de

memoria a largo y corto plazo se consideran una de las arquitecturas de aprendizaje profundo más avanzadas para las tareas de aprendizaje secuencial, como el reconocimiento de voz o la predicción de series de tiempo (Graves et al., 2013). Las LSTM se han utilizado para pronosticar las tendencias de la gripe y las epidemias de enfermedad de manos, pies y boca con éxito (L. Liu et al., 2018; M. Wang et al., 2019).

- **Máquina de Boltzmann restringida (RBM, Restricted Boltzmann machine)**

Un RBM es un modelo estocástico generativo que aprende una distribución de probabilidad sobre el espacio de entrada (Yoo et al., 2014). Los RBM son una variante de las máquinas de Boltzmann, con la restricción de que sus neuronas deben formar un gráfico bipartito. Los pares de nodos de cada uno de los dos grupos (es decir, unidades visibles y ocultas) pueden tener una conexión simétrica entre ellos, pero no hay conexiones entre nodos dentro de un grupo. Esta restricción permite algoritmos de entrenamiento más eficientes que la clase general de máquinas Boltzmann, lo que permite conexiones entre unidades ocultas. Los sistemas de aprendizaje profundo obtenidos al apilar RBM se denominan Redes de creencias profundas(Hinton & Osindero, 2006).

- **Autocodificadores (AE, Autoencoders)**

Un AE es un modelo de aprendizaje no supervisado donde el valor objetivo es igual a la entrada (Cheng et al., 2016). Los EA están compuestos por un decodificador, que transforma la entrada en una representación latente, y por un decodificador, que reconstruye la entrada a partir de esta representación. Los EA están entrenados para minimizar el error de reconstrucción. Al restringir la dimensión de la representación latente para que sea diferente de la entrada (en consecuencia, de la salida), es posible descubrir patrones

relevantes en los datos. Los EA se utilizan principalmente para el aprendizaje de representación (Bengio et al., 2013).

1.2.1 Conceptos relacionados a redes neuronales

- **Función de Activación**

Una función de activación en una red neuronal es una función matemática que se aplica a la salida de cada neurona en una red neuronal. La función de activación define cómo se calcula la salida de una neurona a partir de la entrada y pesos asociados. La función de activación permite que la salida de una neurona sea no lineal, lo que es crucial para la capacidad de una red neuronal para aprender relaciones complejas y no lineales entre las entradas y las salidas. La función activación calcula el estado de actividad de una neurona; transformando la entrada global (menos el umbral, Θi) en un valor (estado) de activación, cuyo rango normalmente va de (0 a 1) o de (–1 a 1). Esto es así, porque una neurona puede estar totalmente inactiva (0 o –1) o activa (1)(Damián, 2001).

Hay muchas funciones de activación diferentes, cada una con sus propias características y usos. Algunas de las funciones de activación más comunes son la función sigmoide, la función ReLU (rectified linear unit) y la función tangente hiperbólica. La elección de la función de activación adecuada depende del tipo de problema que se esté tratando y de las características de los datos de entrada. En general, se recomienda experimentar con diferentes funciones de activación para encontrar la mejor solución para un problema dado.

- **Backpropagation**

Backpropagation es un algoritmo utilizado para entrenar redes neuronales artificiales. Es un método de retro propagación que se utiliza para ajustar los pesos de las conexiones entre las neuronas en una red neuronal con el objetivo de minimizar una función de pérdida.

El proceso de entrenamiento se realiza en dos fases: la primera es la propagación hacia adelante, en la que se realiza una predicción basada en los pesos actuales de la red y los datos de entrada. La segunda fase es la retro propagación, en la que se ajustan los pesos de la red utilizando el gradiente de la función de pérdida y un algoritmo de optimización, como gradiente descendente estocástico. Este proceso se repite hasta que el error sea aceptablemente pequeño, lo que indica que la red está bien entrenada. Backpropagation es un algoritmo esencial en el aprendizaje profundo y es ampliamente utilizado en diversas aplicaciones, incluyendo reconocimiento de imágenes, procesamiento de lenguaje natural y análisis de datos (Hossain et al., 2013).

- **Funciones de pérdida**

El problema de aprendizaje de una red neuronal se presenta como un problema de búsqueda u optimización y se utiliza un algoritmo para navegar por el espacio de posibles conjuntos de pesos que el modelo puede usar para hacer predicciones buenas o lo suficientemente buenas. Normalmente, un modelo de red neuronal se entrena utilizando el algoritmo de optimización de descenso de gradiente estocástico y los pesos se actualizan utilizando el algoritmo de retro propagación de error (Brownlee, 2018). En el contexto de un algoritmo de optimización, la función utilizada para evaluar una solución candidata (es decir, un conjunto de pesos) se denomina función objetiva, la cual se trata de maximizar o

minimizar, esto significa que se busca una solución candidata que tenga la puntuación más alta o más baja, respectivamente. Por lo general, con las redes neuronales, se busca minimizar el error. Como tal, la función objetivo a menudo se conoce como una función de costo o una función de pérdida y el valor calculado por la función de pérdida se conoce simplemente como pérdida. La función de costo reduce todos los diversos aspectos buenos y malos de un sistema posiblemente complejo a un solo número, un valor escalar, lo que permite clasificar y comparar las soluciones candidatas (Chollet, 2017). Existen varias funciones matemáticas que pueden usarse, la elección de una depende del problema que se esté resolviendo. Algunas de estas funciones son:

- **Pérdida de entropía cruzada (o pérdida logarítmica):** La pérdida de entropía cruzada a menudo se conoce simplemente como entropía cruzada, pérdida logarítmica, pérdida logística. Cada probabilidad predicha se compara con el valor de salida de la clase actual (0 o 1) y se calcula una puntuación que penaliza la probabilidad en función de la distancia del valor esperado. La penalización es logarítmica, ofreciendo una puntuación pequeña para pequeñas diferencias (0,1 o 0,2) y una puntuación enorme para una gran diferencia (0,9 o 1,0). La pérdida de entropía cruzada se minimiza, donde los valores más pequeños representan un mejor modelo que los valores más grandes. Un modelo que predice probabilidades perfectas tiene una entropía cruzada o pérdida logarítmica de 0.0. Esta función se usa para problemas de clasificación.
- **La pérdida media de error al cuadrado, o MSE.** se calcula como el promedio de las diferencias al cuadrado entre los valores predichos y los actuales. El resultado es siempre positivo independientemente del signo de los valores

predichos y actuales y un valor perfecto es 0.0. El valor de pérdida se minimiza, aunque se puede utilizar en un proceso de optimización de maximización haciendo que la puntuación sea negativa. Esta función se usa para problemas de regresión.

1.3 Aprendizaje Automático, Aprendizaje Profundo y las enfermedades infecciosas

Como se mencionó anteriormente, a nivel mundial está generando datos a un ritmo asombroso y aún en aumento. Si bien estos 'grandes datos' han desbloqueado nuevas oportunidades para comprender la salud pública, tienen un potencial aún mayor para la investigación y la práctica(Mooney & Pejaver, 2018), en este sentido un creciente número de investigaciones ha venido realizando estudios de vigilancia de enfermedades infecciosas, con el objetivo de complementar los sistemas de información y/o vigilancias existentes. Por un lado, se han venido realizando estudios sobre la detección de enfermedades infecciosas utilizando fuentes de datos como las consultas de búsqueda en Internet (Lampos et al., 2015; Rohart et al., 2016; Teng et al., 2017) y el Big data de las redes sociales. (Tenkanen et al., 2017) informan que los Big data de las redes sociales son relativamente fáciles de recopilar y pueden usarse libremente.

Samaras, García-Barriocanal, & Sicilia (2020) prueban crear un sistema de vigilancia electrónica obteniendo y analizando datos en línea. El objetivo consistió en reunir evidencia sobre qué tipo de fuente de datos conduce a mejores resultados. Se utilizaron los datos sobre influenza en Grecia recopilados de Google y Twitter y se compararon con los datos de influenza de la autoridad oficial de Europa. Los datos se analizaron utilizando dos modelos: el modelo ARIMA calculó estimaciones basadas en sumas semanales y un modelo

aproximado personalizado que usa sumas diarias. También Zhao et al.(2020) en su estudio proponen un nuevo marco de red neuronal semi-supervisada que integra las fortalezas de la epidemiología computacional y las técnicas de minería de redes sociales para el modelado epidemiológico de la influenza. Este marco aprende los estados de salud de los usuarios de las redes sociales y las acciones de intervención en tiempo real, regularizadas por el modelo de enfermedad subyacente y la red de contactos.

En particular, un estudio de Shin et al. (2016) evaluaron la posibilidad de utilizar un sistema de vigilancia digital basado en búsquedas web y datos de redes sociales para monitorear un brote de MERS, encontrando que la enfermedad y los datos de Twitter están altamente correlacionados concluyendo que, existe la posibilidad de utilizar sistemas de vigilancia digital para monitorear enfermedades infecciosas en el futuro. De otra parte, Teng et al. (2017)desarrollaron un modelo de pronóstico dinámico para el virus Zika basado en datos de búsqueda en línea en tiempo real de Google Trends (GT). Fue diseñado para proporcionar vigilancia y detección de la enfermedad y números predictivos de casos de infección, lo que les daría tiempo suficiente para implementar intervenciones.

Adicionalmente, se encuentran estudios de predicción de enfermedades basados en factores ambientales (Huang et al., 2013), dado que se ha demostrado que los datos meteorológicos comprenden un factor que tiene una gran influencia en la aparición de enfermedades infecciosas (T. Liu et al., 2015), por ejemplo el estudio realizado por Huang et al. (2013) informó que las tendencias en la fiebre del dengue muestran una fuerte correlación con la temperatura y la humedad.

A su vez, distintos autores han querido ir más allá y no sólo considerar una fuente o tipo de información, con el objetivo de mejorar los resultados de las predicciones, se han integrado diferentes variables en un mismo estudio, tal es el caso de Shi et al. (2016) quien

utilizó datos meteorológicos, datos de vigilancia de vectores y estadísticas nacionales basadas en población, con ello estableció un conjunto de modelos estadísticos utilizando el método de operador de contracción y selección menos absoluto (LASSO) para pronosticar la incidencia semanal de notificaciones de dengue en un horizonte temporal de 3 meses en Singapur. Así mismo, Li et al. (2019) proporcionaron una predicción muy acertada del dengue mediante la integración de un modelo aditivo generalizado (GAM) de la dinámica de mosquitos con un modelo compartimental de transmisión viral susceptible-infectado-recuperado (SIR) en China continental. Por su parte, Soliman, Lyubchich, & Gel. (2020) buscaron mejorar los pronósticos del Zika al introducir los conceptos de análisis de datos topológicos y, específicamente, homología persistente de variables atmosféricas, en el modelo de propagación del virus. Para ello, introdujeron un nuevo concepto de números acumulativos de Betti que luego se integran como descriptores topológicos en tres modelos predictivos de aprendizaje automático: bosque aleatorio, regresión potenciada generalizada y red neuronal profunda. Mientras tanto, Jiang, Hao, Ding, Fu, & Li (2018) adoptaron tres modelos de aprendizaje automático para mapear la probabilidad de un brote epidémico de Zika a nivel global, emparejando capas covariables multidisciplinarias de alta dimensión con datos completos de ubicación sobre la infección por virus Zika registrada en humanos.

Sin embargo, considerando que son muchos los factores que influyen la incidencia de las enfermedades infecciosas, entre los más representativos se encuentran las variables meteorológicas, los factores socioeconómicos y la densidad de población de vectores, es complejo develar las relaciones existentes entre los casos que se puedan presentar y estas variables. Razón por la cual el modelo clásico de serie temporal no puede ajustarla fácilmente. A raíz de esto, han venido surgiendo estudios que han utilizado técnicas del campo del aprendizaje profundo para predecir enfermedades infecciosas.

El Deep Learning o aprendizaje profundo es un método de análisis y se está utilizando activamente en una variedad de campos. Este método produce resultados satisfactorios cuando se utiliza para realizar tareas que son complejas para los métodos de análisis convencionales (Coccia, 2020). Por ejemplo, Zhang & Nawata (2017) realizó un estudio comparativo sobre la predicción de los brotes de influenza, para lo cual utilizó seis modelos diferentes: media móvil integrada autorregresiva (ARIMA), regresión de vectores de soporte (SVR), bosque aleatorio (RF), aumento de gradiente (GB), red neuronal artificial (ANN) y Memoria a corto y largo plazo (LSTM) con ajuste de hiperparámetro. Como resultado encontraron que los modelos de aprendizaje automático (SVR, RF, GB) lograron los MAPE (Mean Absolute Percentage Error) más bajos y los MAPE de los modelos de aprendizaje profundo (ANN, LSTM) fueron menores que los de los modelos de aprendizaje automático. Asimismo, Augusta, Deardon, & Taylor (2019) hacen uso de métodos supervisados de estadística y aprendizaje automático con un enfoque de aprendizaje profundo para la simulación de epidemias a través de dos poblaciones de granjas porcinas en Iowa. Entre tanto, Scarafoni, Telfer, Ricke, Thornton, & Comolli (2019) aplicaron las redes neuronales convolucionales (CNN), para identificar el tropismo del huésped para los virus de influenza A y humanos basados en secuencias de proteínas, mientras que (Porrello et al., 2019) las usaron para el modelado de la presencia de culicoides.

Estudios como el de Chae et al. (2018), reconocieron que los informes de enfermedades infecciosas de algunas organizaciones médicas incluían retrasos que pueden ocurrir en el sistema de informes, por lo cual construyeron un modelo de pronóstico de enfermedades infecciosas mediante la optimización de los parámetros de los algoritmos de aprendizaje profundo al considerar grandes bases de datos, incluidos los datos de las redes sociales.

Por su parte, Xu et al. (2020).propusieron un modelo basado en LSTM (Long short-term memory), que permitió predecir eficientemente los casos mensuales de dengue utilizando datos meteorológicos y casos de dengue en 20 ciudades de China continental. Mientras que C. Wang, Qi, & Zhu (2020) con el objetivo de mejorar la precisión y el poder predictivo de la aparición de enfermedades cardio pulmonares, establecieron cuatro modelos diferentes de aprendizaje profundo (DL) para capturar dependencias inherentes a largo plazo en secuencias y posibles relaciones complejas entre las diferentes categorías agrupadas de enfermedades respiratorias y del sistema circulatorio en Nanjing desde 2013 hasta 2018. Adicionalmente, L. Wang, Chen, & Marathe (2020) desarrollaron el pronóstico de epidemias basado en el aprendizaje profundo guiado por la teoría con información sintética TDEFSI[1] para el ILI (Influenza-Like Illness), este marco de predicción integra las fortalezas de las redes neuronales profundas y simulaciones de alta resolución de procesos epidémicos a través de redes, produce pronósticos espaciotemporales precisos de alta resolución utilizando datos de series temporales de baja resolución.

Por su parte, A. Arista-Jalife et al. (2020) presentan una propuesta basada en redes neuronales profundas (DNN, Deep Neural Networks) que es capaz de identificar el mosquito Aedes aegypti y Aedes albopictus en la etapa larval, que es fácilmente desechable, restringido a cuerpos de agua e incapaz de transmitir enfermedades según los Centros para el Control y la Prevención de Enfermedades. (Gourisaria et al., 2020) utilizan las redes neuronales profundas en la imagen microscópica de las células sanguíneas infectadas con malaria para predecir si un organismo sufre de malaria o no.

[1] Theory-guided Deep Learning-based Epidemic Forecasting with Synthetic Information

De manera general, cada una de las aplicaciones tanto del aprendizaje automático como del aprendizaje profundo, comparten un objetivo común, y es evitar, prevenir y controlar los brotes de enfermedades que puedan generarse, y que esto puede ser logrado a través de la estandarización efectiva del cuidado de la salud, los planes de mejoramiento y los sistemas de vigilancia (Tsui et al., 2011). Se resalta que la propagación de enfermedades infecciosas tiene un impacto significativo a nivel individual y social. De ahí que los sistemas de datos son catalogados como importantes para aproximarse a la prevención, detección, respuesta y gestión de brotes de enfermedades infecciosas de plantas, animales y humanos (O'Shea, 2017). Es importante tener presente que, es fundamental que los algoritmos de vigilancia posean dos "propiedades conflictivas": ser robustos ante varios patrones en el ajuste de los procesos de referencia y ser sensibles en la detección de cambios en el proceso (como brotes). En este sentido, como consecuencia del avance de las tecnologías de información en salud, sistemas de información médica y sistemas de recolección de datos sindrómicos y salud pública, hay grandes retos y oportunidades para robustecer los algoritmos de vigilancia en términos de monitorear y tener la trazabilidad de los cambios de patrones (Tsui et al., 2011).

La revisión de literatura deja entrever como se está transformando la forma en que se realizan la prevención y el control de las epidemias, a través de la aplicación de Inteligencia Artificial y las tecnologías de Big data. Los enfoques y razonamientos Deep Learning asociados al desafío de la detección, pronóstico y contención de amenazas de epidemias son actualmente un campo en auge, como lo demuestran los documentos incluidos en este apartado. Existen muchos aspectos del Deep Learning que desde la epidemiología de enfermedades arbovirales podrían ser de gran utilidad como lo es su rendimiento superior,

un esquema de aprendizaje de extremo a extremo con aprendizaje integrado, la capacidad de manejar datos complejos y de múltiples modalidades (Coccia, 2020).

2. Planteamiento y Justificación del Problema de Investigación

Las enfermedades arbovirales son un grupo de enfermedades ocasionadas por virus que se trasmiten entre hospederos vertebrados por medio de artrópodos(Arredondo-García et al., 2016), es decir, son trasmitidos por una gran variedad de vectores que comprenden todo insecto u otro animal que normalmente sea portador de un agente infeccioso y que constituya un riesgo para la salud pública (Organización Mundial de la Salud, 2016). Los virus del Dengue, Chikungunya y Zika causan infecciones peligrosas en regiones tropicales y subtropicales de todo el mundo. La Organización Mundial de la Salud estima que una de cada tres personas en toda la población humana está en peligro de contraer una de estas enfermedades por una sola picadura de mosquito(Antonio Arista-Jalife et al., 2020).

Considerando la magnitud de las enfermedades arbovirales, es importante tener presente que el dengue ha sido una amenaza global desde la Segunda Guerra Mundial. Según un análisis reciente de la distribución global y la carga del virus del dengue, se reportan aproximadamente 390 millones de casos de dengue anualmente en todo el mundo, especialmente en Asia y América del Sur. (Xu et al., 2020). Con relación al virus del Zika, antes de 2015, se habían producido brotes en África, el sudeste de Asia y las islas del Pacífico. En mayo de 2015, se confirmó la presencia de la enfermedad en Brasil. Posteriormente, se extendido por las Américas, con epidemias en muchos países. La Organización Mundial de la Salud declaró al Zika, y su posible vínculo con defectos de nacimiento, una emergencia de salud pública internacional en febrero de 2016(Teng et al., 2017). En Colombia cerca del 85% del territorio se encuentra por debajo de los 1600 m.s.n.m (metros sobre el nivel del mar) y se caracteriza por condiciones climáticas, geográficas y epidemiológicas aptas para la transmisión de este tipo de enfermedades (INS, 2019) y Santander por ejemplo, hasta la semana epidemiológica 52 de 2019 se habían notificado 9757 casos de Dengue, 41 casos de

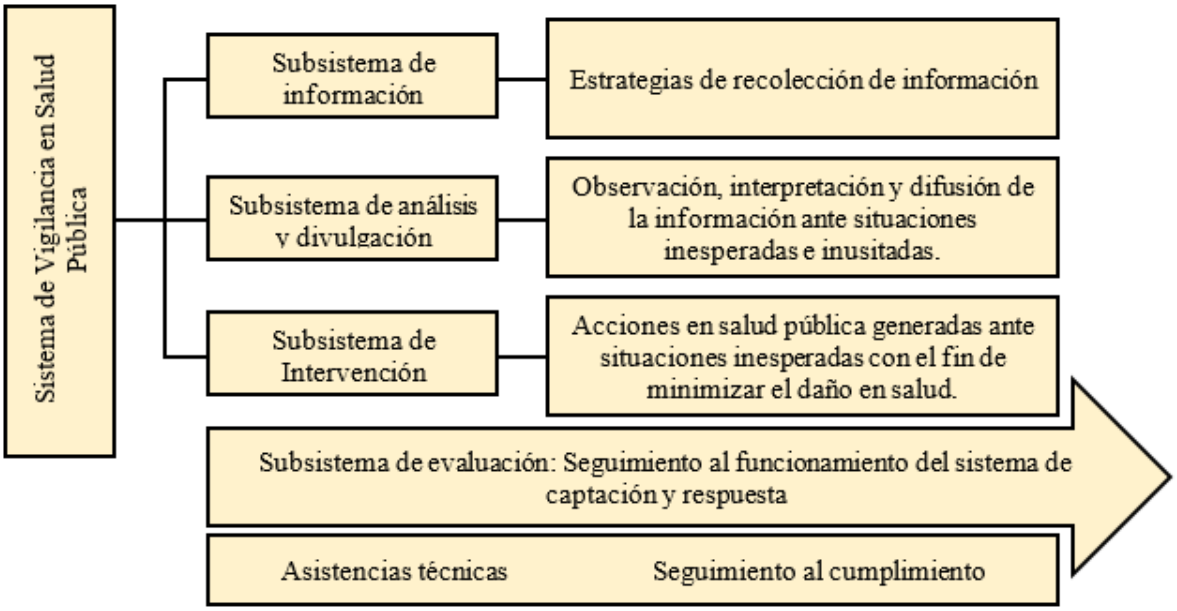
Chicungunya y 36 casos de Zika (Instituto Nacional de Salud, 2019). Actualmente, estas infecciones virales no se pueden prevenir con vacunas y no existe un tratamiento directo que pueda disminuir eficazmente la infección viral, por lo cual la mejor y quizás la única línea de defensa contra estas enfermedades es la vigilancia, la prevención, el control y la supresión efectiva de sus vectores(Antonio Arista-Jalife et al., 2020).

La base para el control de enfermedades infecciosas ha sido principalmente la implementación de sistemas de los sistemas de vigilancia, los cuales rastrean enfermedades, patógenos y resultados clínicos. Sin embargo, los sistemas de vigilancia tradicionales son conocidos por los retrasos severos y la falta de resolución espacial. Los informes de enfermedades infecciosas de algunas organizaciones médicas generalmente se encuentran incompletos y pueden producirse demoras en el sistema de notificación. Como una referencia a lo expuesto, en el sistema tradicional de vigilancia de la influenza, transcurren alrededor de dos semanas entre el momento en que se hace un informe y cuando se difunde lo que dificulta una respuesta rápida a las enfermedades infecciosas(Teng et al., 2017). Los sistemas de vigilancia se basan principalmente en datos recopilados y codificados manualmente, lentos para acumular y difíciles de difundir para su análisis. Además, los informes de estos sistemas tienden a ser nacionales o regionales con poca información sobre enfermedades a nivel local (Bansal et al., 2016).

En Colombia, el proceso de Vigilancia en Salud Pública -VSP se encuentra integrado por cuatro subsistemas los cuales articulan acciones de recolección de datos, análisis, interpretación y divulgación de información, de acuerdo con las características propias de los eventos a vigilar, las condiciones epidemiológicas, las posibilidades reales del desarrollo de intervenciones de control, prevención y atención de cada sector y entidad territorial(Instituto Nacional de Salud, 2018). En la Figura 5 se describen los cuatro subsistemas que integran la

VSP de Colombia con los respectivos lineamientos requeridos en cada uno de ellos. En la figura 6 se presenta el flujo de información para el proceso de notificación.

Figura 5. Componentes del sistema de vigilancia en salud pública de Colombia



Fuente: INS 2018

Figura 6. Flujo de la notificación de Eventos de Interés en Salud Pública mediante correo electrónico y portal web, Colombia, 2018-2020*

UPGD: Unidad Primaria Generadora de datos; UNM: Unidad Notificadora Municipal; UND: Unidad Notificadora Departamental-Distrital; INS: Instituto Nacional de Salud

Fuente: INS 2018

Simplemente para el proceso de notificación todos los integrantes de VSP para realizar un reporte deben cumplir con una serie de protocolos en términos de estructura de datos, responsabilidad, clasificación, periodicidad y destino. Los que hace que el proceso de notificación sea extenso y se produzcan demoras.

En este sentido, la vigilancia es el resultado de la recolección, validación, análisis e interpretación de datos de salud y enfermedades (Tsui et al., 2011) en pro de minimizar los contagios. Sin embargo, considerando todas las etapas que se deben cumplir durante el proceso de notificación, es extenso el tiempo requerido para completar dicho proceso, haciendo que para los sistemas de vigilancia sea difícil actuar de inmediato y generalmente dificulta la realización de acciones adecuadas y oportunas, debido a informes faltantes y retrasados (Chae et al., 2018). Además, en la mayoría de ocasiones no se conocen las tendencias de las enfermedades infecciosas, lo que significa que la predicción no es fácil.

Por otra parte, no siempre es posible confiar en la medicina para desarrollar rápidamente vacunas u otros tratamientos, la mejor prevención es detectar las pandemias tempranas posibles y detener la transmisión. Al bloquear la transmisión, eventualmente también se podría reducir la mutación de los virus y así mantener el virus en una etapa en la que las vacunas podrían ayudar a combatir(Agrebi & Larbi, 2020).

Teniendo en cuenta que el valor de la vigilancia implica la entrega efectiva y eficiente de información útil (H. I. Hall et al., 2012), y que es la característica esencial de la práctica epidemiológica y de la salud pública (Porta, 2008), los sistemas de vigilancia se deben caracterizar por ser flexibles al incremento en las necesidades de información, así como utilizar las tecnologías apropiadas que permitan difundir datos oportunamente. Así, la brecha en el conocimiento de la salud puede ocurrir cuando los sistemas de vigilancia carecen de oportunidad, completitud, fácil adaptación, o eficiencia, y para afrontar esto, la vigilancia en salud pública puede beneficiarse de los avances en ciencias de información, tecnologías y el aumento de bases de datos y fuentes de datos (H. I. Hall et al., 2012).

Por tanto, es importante enfocar esfuerzos en la vigilancia de los vectores y las enfermedades que estos transmiten, con el fin de implementar estrategias preventivas que

permitan mitigar su impacto en la población. Cabe resaltar que muchas de estas enfermedades son prevenibles a través de medidas de protección fundamentales (WHO, 2017); sin embargo, su vigilancia y prevención también implica la vigilancia de sus vectores, lo que dificulta controlarlas y evitar su expansión (Angulo et al., 2013). En este sentido, emerge un área interdisciplinar denominada la epidemiología computacional del Big Data, que utiliza modelos computacionales y big data para entender y controlar la difusión espacio-temporal de una enfermedad a través de la población. Por lo tanto, se está optando por desarrollar tecnologías en bases de datos, así como aplicar técnicas de analítica como minería de datos y aprendizaje automático, que permiten manejar cantidades masivas de información, construir modelos y apoyar la toma de decisiones.

De acuerdo a la discusión presentada, se considera necesario desarrollar un modelo de predicción de enfermedades infecciosas basado en datos en tiempo real. Además, la predicción del alcance de las enfermedades infecciosas, lleva a una disminución de los costos sociales, lo cual redunda en beneficio de la salud para la sociedad en general. De modo que, esta investigación busca apoyar las etapas de análisis, interpretación de datos y difusión de la información del proceso de vigilancia de enfermedades arbovirales en el departamento de Santander a través del uso de técnicas Deep Learning que faciliten el análisis de patrones y tendencias con el fin de ser soporte para la toma de decisiones de las autoridades de salud pública.

3. Objetivos

3.1 Objetivo General

Aplicar modelos Deep Learning para el proceso de vigilancia de enfermedades arbovirales, específicamente Dengue, Zika y Chikungunya, en el departamento de Santander.

3.2 Objetivos Específicos

- Identificar modelos Deep Learning en la vigilancia de enfermedades arbovirales a través de una revisión de literatura.
- Formular modelos Deep Learning para la predicción de la propagación de las zonas geográficas de incidencia, la probabilidad de ocurrencia de infección y la estimación de la población humana proyectada que potencialmente podría estar expuesta a enfermedades arbovirales en el departamento de Santander.
- Validar el desempeño de los modelos Deep Learning mediante la utilización de datos históricos del departamento de Santander.
- Diseñar un prototipo de arquitectura visual apoyado en la plataforma shinyapps.io los modelos Deep Learning de mejor desempeño en la predicción de las enfermedades arbovirales que apoye el proceso de toma de decisiones de vigilancia en salud pública.

4. Hipótesis

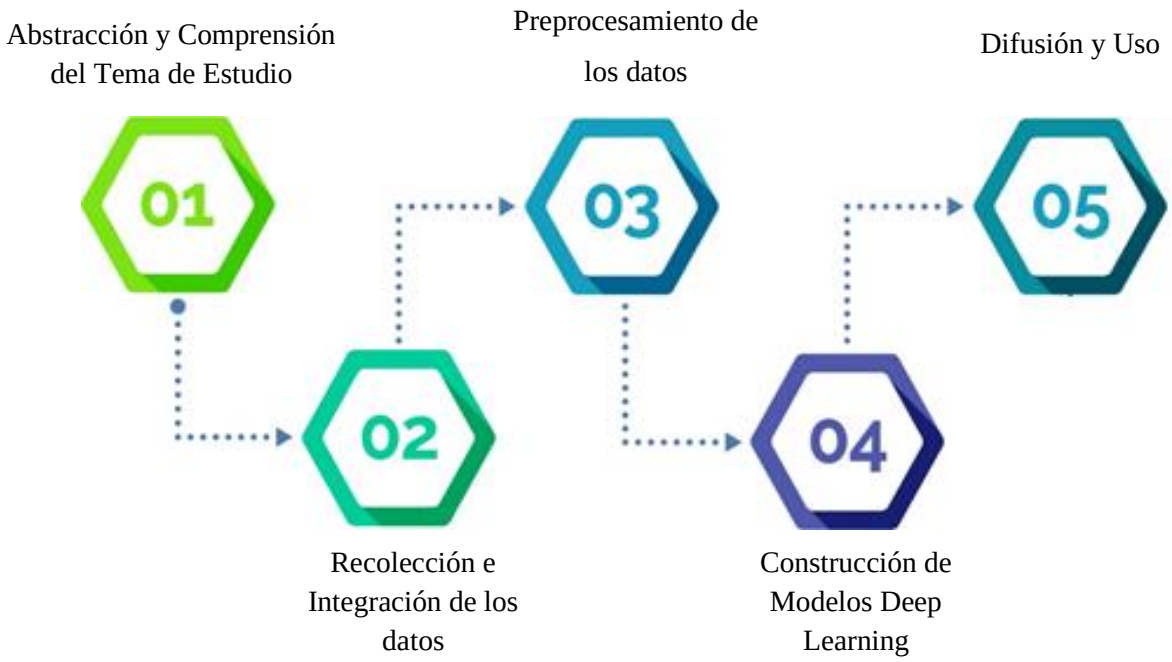
Los modelos Deep Learning mejoran el análisis e interpretación de datos y difusión de la información en el proceso de vigilancia en salud pública en el departamento de Santander respecto a las redes neuronales artificiales.

5. Metodología

5.1 Proceso KDD (Knowledge Discovery in Databases)

La presente investigación siguió un proceso llamado Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases – KDD, por sus siglas en inglés), el cual consiste en encontrar un "modelo" válido, útil y entendible que describa patrones novedosos de acuerdo con la información. El proceso de KDD es de naturaleza iterativa e interactiva. Se compone de varios pasos o etapas diferentes y se detallan a continuación:

Figura 7. Procedimiento Metodológico



5.1.1 Fase I. Abstracción y Comprensión del tema de Estudio

Esta fase consistió en entender la problemática a afrontar y el contexto para proponer soluciones viables y reales. Reconocer las propiedades y limitaciones del escenario en estudio, para con base en ello, definir las metas a alcanzar y el proceso para lograrlo. Como primera fase de la investigación, se realizó una revisión de literatura orientada a la aplicación de técnicas y/o herramientas Deep Learning para el pronóstico y vigilancia de enfermedades arbovirales. Para ello, se seleccionó la siguiente ecuación de búsqueda:

((TITLE-ABS-KEY (“Deep Learning”) AND ("infectious disease" OR "arboviral disease*" OR epidemic) AND (surveillance OR propagation OR "predict*" OR "forescast*")).* A partir de la ecuación de búsqueda elegida y con apoyo de la base de datos Scopus, Science Direct y PubMed, se realizó la revisión (primer capítulo del documento), la cual permitió identificar los modelos y variables más utilizadas para el pronóstico y vigilancia de enfermedades arbovirales, además de las métricas empleadas para la evaluación de desempeño de los modelos.

5.1.2 Fase II. Recolección e Integración de los datos

Esta fase abarcó la gestión de las bases de datos y la respectiva recolección e integración de los datos. De manera general, se utilizaron variables climáticas: Precipitaciones, Temperaturas, Humedad relativa, y los datos correspondientes a los casos de dengue.

Para los datos de casos de dengue, Zika y Chikungunya, se gestionó la adquisición de bases de datos de los registros de SIVIGILA (Sistema Nacional de Vigilancia en Salud Pública); para esto, fue necesario asistir a una capacitación sobre “Estadísticas de afiliados a

salud BDUA" para luego solicitar el acceso a la base de datos, una vez realizada la solicitud, fue proporcionado un usuario y contraseña para acceder a los cubos de datos del Sistema Integrado de Información de la Protección Social – SISPRO. La información de los casos es reportada por semana epidemiológica. Con relación a los casos de Dengue, se contó con información desde 2007 al 2020, para los casos de Zika desde el 2015 al 2020, y para los casos de Chikungunya desde el 2014 al 2020.

Con relación a las variables climáticas, la información se recopiló del portal abierto de información meteorológica del Instituto de Hidrología, Meteorología y Estudios Ambientales -IDEAM (http://dhime.ideam.gov.co/atencionciudadano/). Estas variables fueron descargadas originalmente en escala diaria, por lo cual, fue necesario agrupar los datos para obtener valores promedio, máximos y mínimos para cada semana epidemiológica. Se registró la temperatura promedio diaria máxima, la temperatura promedio diaria mínima, la temperatura máxima y la temperatura mínima de la semana y la humedad relativa promedio máxima y mínima. Para el caso de las precipitaciones, se obtuvo tanto el promedio de los días de la semana epidemiológica como la sumatoria o acumulado total de lluvias para la semana.

Considerando que la resolución espacial de los datos correspondientes a las variables climáticas se presenta a nivel municipal, los modelos predictivos, con uso de estas variables se generaron al mismo nivel, específicamente para los municipios de Bucaramanga, Barrancabermeja, Floridablanca, Piedecuesta, Girón y Lebrija, primordialmente por dos razones; en primer lugar, fueron los municipios para los cuales se encontró mayor cantidad de información, debido a la gran cantidad de datos faltantes, pues para una gran cantidad de municipios no se encontraron datos registrados y en otros casos se encontraron datos solo para ciertos periodos de tiempo. Por lo cual fue necesario construir bases de datos para cada

municipio con diferentes ventanas temporales y diferentes variables. En segundo lugar, los municipios seleccionados representan el 78%, 87% y 64% de participación departamental para los casos de Dengue, Zika y Chicungunya respectivamente. La siguiente tabla resume el conjunto completo de todas las variables con la notación empleada para cada una.

Tabla 1 *Conjunto de variables recolectadas*

Variable	Descripción	Notación
Temperatura Seca Máxima Diaria Promedio	Promedio de las temperaturas secas máximas de los días correspondientes a cada semana epidemiológica.	Temp_Seca_Max_Dia_Prom
Temperatura Seca Máxima (Máxima Semana)	Dato máximo de temperatura seca registrado entre los días correspondientes a cada semana epidemiológica.	Temp_Seca_Max_Sem
Temperatura Seca Mínima Diaria Promedio	Promedio de las temperaturas secas mínimas de los días correspondientes a cada semana epidemiológica.	Temp_Seca_Min_Dia_Prom
Temperatura Seca Mínima (Mínima Semana)	Dato mínimo de temperatura seca registrado entre los días correspondientes a cada semana epidemiológica.	Temp_Seca_Min_Sem
Temperatura Máxima Diaria Promedio	Promedio de las temperaturas máximas de los días correspondientes a cada semana epidemiológica.	Temp_Max_Dia_Prom
Temperatura Máxima (Máxima Semana)	Dato máximo de temperatura registrado entre los días correspondientes a cada semana epidemiológica.	Temp_Max_Sem
Temperatura Mínima Diaria Promedio	Promedio de las temperaturas mínimas de los días correspondientes a cada semana epidemiológica.	Temp_Min_Dia_Prom
Temperatura Mínima (Mínima Semana)	Dato mínimo de temperatura registrado entre los días correspondientes a cada semana epidemiológica.	Temp_Min_Sem

Variable	Descripción	Notación
Promedio Diario Precipitaciones	Promedio de las precipitaciones registradas en los días correspondientes a cada semana epidemiológica.	Prom_Dia_Preci
Suma Total Precipitaciones Semana	Sumatoria de las precipitaciones registradas en los días correspondientes a cada semana epidemiológica.	Total_Preci_Sem
Humedad Relativa Máxima	Promedio de las humedades máximas registradas en los días correspondientes a cada semana epidemiológica.	Hum_Rel_Max
Humedad Relativa Mínima	Promedio de las humedades mínimas registradas en los días correspondientes a cada semana epidemiológica.	Hum_Rel_Min
Casos de Dengue	Número total de personas reportadas con Dengue en la semana epidemiológica.	Casos_Dengue

A continuación, se presentan las variables y ventanas de tiempo disponibles para cada municipio de estudio.

Tabla 2 *Conjunto de variables disponibles por municipio*

Municipio	Variables	Ventana Temporal
Barrancabermeja	Temp_Seca_Max_Dia_Prom Temp_Seca_Max_Sem Temp_Seca_Min_Dia_Prom Temp_Seca_Min_Sem Temp_Max_Dia_Prom Temp_Max_Sem Temp_Min_Dia_Prom Temp_Min_Sem Prom_Dia_Preci Total_Preci_Sem Hum_Rel_Max Hum_Rel_Min Casos_Dengue	2008 - 2019
Girón		2014 - 2020
Lebrija	Temp_Seca_Max_Dia_Prom Temp_Seca_Max_Sem Temp_Seca_Min_Dia_Prom Temp_Seca_Min_Sem Temp_Max_Dia_Prom Temp_Max_Sem	2008 - 2015

Municipio	Variables	Ventana Temporal
	Temp_Min_Dia_Prom Temp_Min_Sem Prom_Dia_Preci Total_Preci_Sem Casos_Dengue	
Bucaramanga	Prom_Dia_Preci Total_Preci_Sem	2008 - 2019
Floridablanca		
Piedecuesta		

5.1.3 Fase III. Preprocesamiento de los datos

Considerando que la calidad del modelo de predicción generado y el conocimiento descubierto no depende estrictamente de la técnica de análisis utilizada, sino también, de la calidad de los datos analizados. La gestión de la calidad de los datos tanto de las variables climáticas como de los casos de dengue fue un desafío, considerando que las bases de datos presentaban problemas como datos incorrectos o datos faltantes.

En esta investigación, los modelos se trabajaron con el set de datos departamental, en el cual sólo se consideraron los datos históricos de los casos de Dengue, Zika y Chikungunya del departamento de Santander, en este caso, no se consideraron las variables climáticas, debido a que la información de estas se encuentra a nivel municipal; por su parte, los modelos con el set de datos municipal, incluyeron las variables climáticas y los datos históricos de los casos de Dengue, Zika y Chicungunya por municipio. El set de datos municipal, se construyó para los municipios de Bucaramanga, Floridablanca, Girón, Piedecuesta, Barrancabermeja y Lebrija.

Con relación a la limpieza de los datos, se siguió la metodología propuesta por Corrales, Corrales, y Ledezma, (2018) para la limpieza de datos en modelos de regresión, en donde, una vez seleccionados los periodos de tiempo con mayor información, se continuó

con la identificación de valores perdidos y su tratamiento, que para este caso consistió en la imputación a no más del 3% de las observaciones. Para las variables climáticas se tomó el valor medio entre la observación previa y siguiente (imputación por vecino más cercano). Para los casos de Dengue, Zika y Chicungunya no se encontraron valores perdidos.

Para la normalización de los datos, se utilizó la función MinMaxScaler de la librería sklearn, en Python, para cambiar la escala de los datos del rango original y hacer que todos los valores quedaran dentro del rango de 0 y 1; esto, con el fin de que todos los atributos quedaran en un rango fijo.

Además, considerando que, diversos estudios han demostrado una no simultaneidad temporal por la asociación estadística entre los casos de dengue y las variables climáticas (Cardona Acosta, 2015), se consideraron rezagos para cada una de las variables de 1 a 6 semanas. Para los municipios de Bucaramanga, Floridablanca y Piedecuesta, solo se contaba con la variable precipitaciones, se realizó un modelo para cada uno de los rezagos.

Sin embargo, para los municipios de Barrancabermeja, Girón y Lebrija fue necesario identificar cual rezago y combinación de variables era el más adecuado para los modelos de predicción. Para esto, se realizó un modelo Random Forest mediante el cual se evaluó la importancia de las características, cuya técnica consiste en asignar una puntuación a las características de entrada en función de su utilidad para predecir una variable de destino. De esta forma, eliminar de forma definitiva variables que no eran de interés o no podían aportar información relevante para los modelos.

En este proceso, se consideró la importancia de la característica por permutación, la cual mide la importancia de una característica calculando el aumento en el error de predicción del modelo después de permutar la característica. Una característica es “importante” si cambiar sus valores aumenta el error del modelo, porque en este caso el modelo se basó en

la característica para la predicción. Una característica es “no importante” si cambiar sus valores deja el error del modelo sin cambios, porque en este caso el modelo ignoró la característica para la predicción (Breiman, 2001 tomado de Molnar, 2021). Así mismo, se consideró la importancia basada en los valores SHAP (SHapley Additive exPlanations) cuyo enfoque se basa en la teoría de juegos, se encarga de encontrar la contribución marginal promedio de un parámetro de entrada (variable) a través de todas las combinaciones posibles, el objetivo es lograr interpretar la predicción de un modelo mediante la contribución de cada parámetro de entrada (Molnar, 2021). Finalmente, también se tuvo en cuenta la correlación de Pearson entre las variables predictoras y la variable respuesta (Visualizar los resultados de importancia de las características en los Anexos B, C, y D). De cada una de las técnicas utilizadas para determinar la importancia de las características se tomaron las 2 características principales, las cuales se presentan en la siguiente tabla.

Tabla 3 Variables más importantes por municipio

Técnica	**Variables más importantes por municipio**		
	Barrancabermeja	Girón	Lebrija
Importancia por Permutación	• Humedad relativa mínima con rezago 6. • Humedad relativa mínima	• Humedad relativa máxima con rezago 1. • Humedad relativa máxima con rezago 2.	• Temperatura seca mínima semana con rezago 1. • Temperatura seca mínima semana con rezago 6.
Importancia por Valores SHAP	• Temperatura seca mínima diaria promedio con rezago 2. • Humedad relativa mínima con rezago 6.	• Humedad relativa máxima con rezago 5. • Humedad relativa máxima con rezago 2.	• Temperatura mínima semana con rezago 3. • Temperatura mínima semana con rezago 4.
Correlación de Pearson	• Humedad relativa mínima con rezago 6.	• Humedad relativa máxima con rezago 4.	• Temperatura seca mínima diaria

	• Humedad relativa mínima con rezago 4.	• Humedad relativa máxima con rezago 5.	promedio con rezago 3. • Temperatura seca mínima diaria promedio con rezago 4.

Una vez definidas las características de mayor importancia para cada municipio, se plantearon diferentes escenarios, con diferentes combinaciones de variables, con el fin de identificar la combinación ideal de atributos predictores. A continuación, se presentan los escenarios planteados.

Tabla 4 Escenarios planteados para construcción de los modelos

Escenarios	**Municipios**		
	Barrancabermeja	Girón	Lebrija
Escenario 1	• Humedad relativa mínima • Humedad relativa mínima con rezago 4. • Humedad relativa mínima con rezago 6. • Temperatura seca mínima diaria promedio con rezago 2.	• Humedad relativa máxima con rezago 1. • Humedad relativa máxima con rezago 2. • Humedad relativa máxima con rezago 4. • Humedad relativa máxima con rezago 5.	• Temperatura seca mínima semana con rezago 1. • Temperatura seca mínima semana con rezago 6. • Temperatura mínima semana con rezago 3. • Temperatura mínima semana con rezago 4. • Temperatura seca mínima diaria promedio con rezago 3. • Temperatura seca mínima diaria promedio con rezago 4.
Escenario 2	• Humedad relativa mínima	• Humedad relativa máxima con rezago 5.	• Temperatura seca mínima semana con rezago 1.

	• Humedad relativa mínima con rezago 6.	• Humedad relativa máxima con rezago 2.	• Temperatura seca mínima semana con rezago 6.
Escenario 3	• Humedad relativa mínima con rezago 6. • Temperatura seca mínima diaria promedio con rezago 2.	• Humedad relativa máxima con rezago 1. • Humedad relativa máxima con rezago 2.	• Temperatura mínima semana con rezago 3. • Temperatura mínima semana con rezago 4.
Escenario 4	• Humedad relativa mínima con rezago 4. • Humedad relativa mínima con rezago 6.	• Humedad relativa máxima con rezago 4. • Humedad relativa máxima con rezago 5.	• Temperatura seca mínima diaria promedio con rezago 3. • Temperatura seca mínima diaria promedio con rezago 4.

Para cada municipio los modelos fueron entrenados y probados de acuerdo con los escenarios propuestos y las combinaciones de las características predictoras.

5.1.4 Fase IV. Construcción de los Modelos Deep Learning

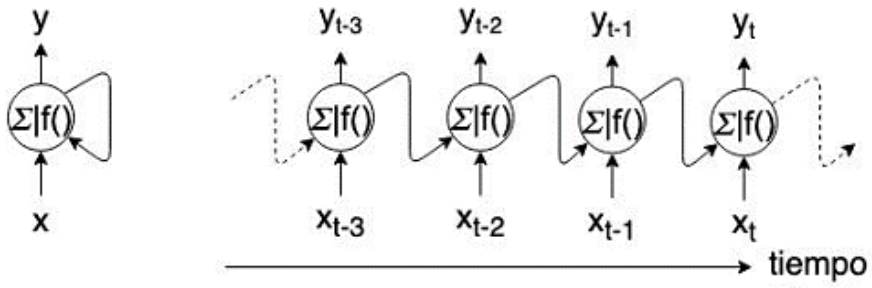
Con base en la revisión de literatura, se seleccionaron los siguientes modelos Deep Learning para el análisis de la información de enfermedades arbovirales en Santander, considerando que han sido frecuentemente utilizados para la predicción de series temporales y han reportado resultados satisfactorios.

- **Redes Neuronales Recurrentes (RNN)**

Una red neuronal recurrente es una red neuronal que contiene ciclos internos que realimentan la red, generando así, memoria. Dicha memoria le permite a la red aprender y generalizar a lo largo de secuencias de entradas en lugar de patrones individuales. El estado

de la red neuronal recurrente es reestablecido cada vez que procesan dos secuencias diferentes e independientes, por lo que se considera a una secuencia como un dato singular, una sola entrada en la red. Dicho dato ya no es procesado en un solo paso, sino que la red internamente cicla sobre la secuencia de elementos (Fierro, 2020). En cada paso temporal t, la neurona recurrente recibe las entradas x_t y también su propia salida producida por el paso temporal previo y_{t-1}. Cada neurona recurrente posee dos grupos de pesos sinápticos, uno para las entradas x_t y otro para las salidas del paso temporal previo y_{t-1}.

Figura 8. Neurona recurrente (izq.) - Neurona recurrente desplegada en el tiempo (der.)



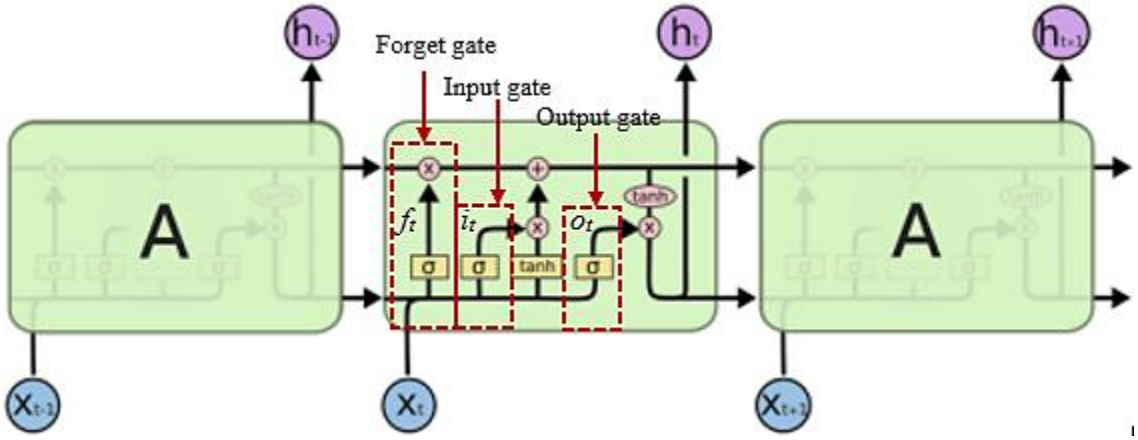
Fuente: Tomado de (Fierro, 2020)

Las redes neuronales recurrentes presentan un inconveniente, ya que, en teoría deberían retener información en un tiempo t sobre entradas procesadas varios pasos temporales atrás, en la práctica, dichas dependencias a largo plazo resultan imposibles de aprender. A este problema se lo conoce como `Desvanecimiento del Gradiente'. La red eventualmente pierde su capacidad de ser entrenada a medida que se le agregan capas (Fierro, 2020).

- **Redes Neuronales Recurrentes de Memoria a Largo y Corto Plazo (LSTM)**

Fueron propuestas por Hochreiter y Schmidhuber (1997) para dar solución al problema de "desvanecimiento del Gradiente", el cual compromete la eficiencia de la red de neuronas recurrente. El principio detrás del modelo LSTM es la celda de memoria o memory cell, en donde cada celda tiene un grupo de operaciones muy específicas que permiten controlar el flujo de información. Estas operaciones, llamadas puertas permiten decidir si cierta información es recordada u olvidada.

Figura 9. Arquitectura de una Red LSTM



Fuente: Tomado de (Chumino et al., 2021)

Dichas puertas son las siguientes(Guerrero, 2020):

- **Puerta de olvido o forget gate (f):** Se usa para ocultar el estado ct-1 de modo que solo se considere parte del estado en el paso t. Es decir, esta puerta decide qué información permanece y cual se olvida.
- **Puerta de entrada input gate (i):** Se usa para ocultar qué información nueva se almacenará en el oculto (hidden state) de la célula.

- **Puerta de salida o output gate (o):** Esta puerta se encarga de decidir cuál será el estado oculto de la célula en el paso de tiempo siguiente.

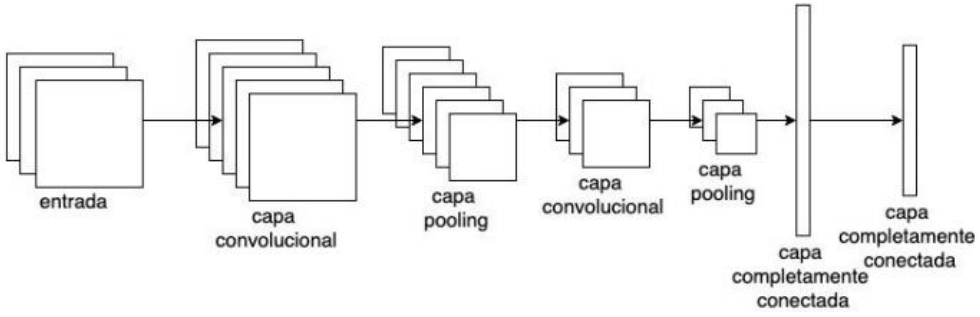
- **Redes Neuronales Convolucionales (CNN)**

Son un tipo de red neuronal diseñado originalmente para procesar datos de imágenes, pero las mismas propiedades que hacen propicias a las redes neuronales convolucionales para problemas de visión de computadoras las hacen muy relevantes para procesar señales. El tiempo puede ser tratado como una dimensión espacial, como la altura o el ancho de una imagen 2D. Estas son las redes convolucionales 1D. Las capas de convolución 1D obtienen nuevas secuencias convolucionadas a través de filtros que interpretan ciertas características de las secuencias originales que permiten reconocer patrones locales en la misma.

En una red neuronal convolucional hay tres tipos de capas, las capas convolucionales, de pooling y densamente conectadas. En La capa convolucional las neuronas no están conectadas a cada uno de los valores de entrada, solo a los valores dentro de sus campos receptivos. En la siguiente capa convolucional, cada neurona está conectada solamente a neuronas ubicadas dentro de un sector reducido de la capa anterior. La diferencia fundamental entre una capa densamente conectada y una convolucional es que las capas densas aprenden patrones globales de sus entradas, mientras que las convolucionales aprenden filltros (o Kernels) que modifican la señal original, generando descriptores o mapas de características. Esta arquitectura le permite a la red concentrarse en características más simples en las primeras capas y agruparlas luego en características más complejas en las capas siguientes.

La capa de pooling tiene como objetivo reducir la muestra anteriormente procesada, lo que disminuye la carga computacional, utilización de memoria y número de parámetros, extrayendo subsecuencias de una entrada y devolviendo el valor máximo o promedio. En una arquitectura típica de una red neuronal convolucional se apilan u grupo de capas convolucionales, luego una capa de pooling. Dicha estructura mencionada normalmente se repite. La entrada se reduce más y más a su paso por la red, a la vez que aumenta la cantidad de mapas de características. Al final de la pila una red neuronal densamente conectada es agregada.

Figura 10. Arquitectura red convolucional



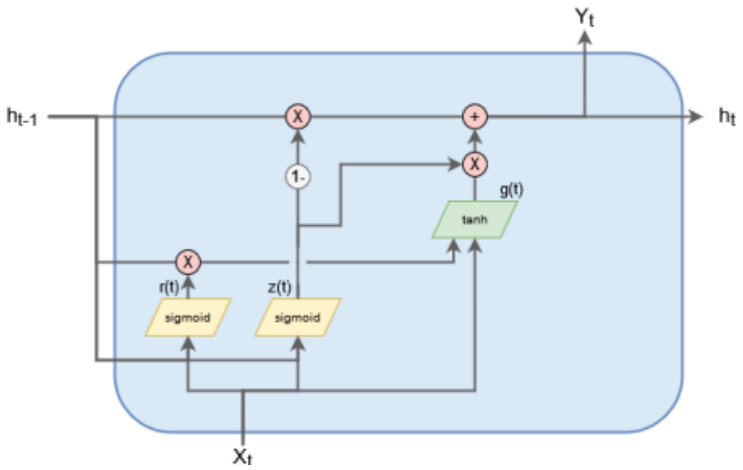
Fuente: Tomado de (Fierro, 2020)

- **Redes de Unidades recurrentes con compuertas (GRU)**

Las GRUs son neuronas recurrentes con una estructura un poco diferente que las LSTM, son más recientes y eliminan algunas de las operaciones de las LSTM; sin embargo, ha demostrado que pueden trabajar con el mismo rendimiento en general. En este caso la celda solo trabaja con un único vector de estado $h_{(t)}$ y con dos puertas que modificarán el flujo de dicho vector. La primera de ellas, update gate, cuya salida es $z_{(t)}$ sirve como controladora de las puertas input gate y forget gate de modo que su rango de valores de salida

[0 - 1] controla si se elimina información del estado o se añade. La segunda, reset gate, cuya salida es $r_{(t)}$, controla qué parte del estado anterior se le mostrará a la capa principal con función tanh $g_{(t)}$, que será la que dé lugar a la salida de la celda(Cea Morán, 2020).

Figura 11. Estructura de una celda GRU



Fuente: Tomado de (Cea Morán, 2020)

Se realizaron modelos a nivel departamental con la variable únicamente de casos, para Dengue, Zika y Chikungunya. A nivel municipal, para los municipios con solo la variable precipitaciones (Bucaramanga, Floridablanca y Piedecuesta) se creó un modelo para cada rezago de 1 a 6 y para cada tipo de red elegida (RNN, LSTM, CNN y GRU), es decir, 24 modelos por municipio. Para los municipios con todas las variables climáticas, creó un modelo para cada tipo de red elegida y para cada escenario definido (4 escenarios), es decir, 16 modelos por municipio; sin embargo, con estos municipios también se quiso probar el nivel predictivo con la variable precipitaciones, por tanto, se crearon 24 modelos adicionales, para un total de 64 modelos generados.

Para la creación de los modelos, fue necesario estructurar los datos como un problema de aprendizaje supervisado. En donde, a cada conjunto de variables de entrada X le corresponde un conjunto de variables de salida, y un algoritmo es utilizado para aprender una función de mapeo desde la entrada a la salida.

$$Y = f(X)$$

En los modelos de redes neuronales, los datos de las muestras utilizadas como entrada y las de salida, en principio, deben quedar estructurados tridimensionalmente. Las tres dimensiones de dicha entrada son:

Muestras. Una secuencia es una muestra. Un lote se compone de una o más muestras. La cantidad de muestras de entrada y de salida es la misma. Esto significa que a cada muestra de entrada le corresponde una muestra de salida.

Pasos temporales. Un paso temporal es un punto de observación en una muestra. Una muestra se compone de múltiples pasos temporales. La cantidad de pasos temporales de las muestras de salida es la cantidad de predicciones, o pasos temporales a futuro que el modelo aprende a predecir a partir de una muestra de entrada. La cantidad de pasos temporales de las muestras de entrada es la cantidad de observaciones de cada variable que el modelo utiliza para aprender a predecir la salida.

Características. Una característica es una variable de una observación. Un paso temporal se compone de una o más características. Las características presentes en las muestras de salida son las que el modelo aprende a predecir a partir de una muestra de entrada.

La estructura de datos que representa un lote de muestras de entrada se resume comúnmente de la siguiente manera:

Figura 12. Estructura de las muestras

[muestras, pasos temporales de entrada, características de entrada]

La cantidad de pasos temporales de las muestras de entrada en el presente trabajo se escogió mediante un proceso de prueba y error, mediante la verificación de los resultados obtenidos utilizando varios pasos temporales diferentes. Como resultado de esta exploración, se seleccionaron 7 pasos de tiempo, para ser incorporados en los modelos.

Horizonte de predicción

Existen tres maneras alternativas para producir predicciones de períodos múltiples con modelos de aprendizaje automático (Bontempi et al., 2013). La estrategia recursiva involucra añadir la última predicción del último paso de tiempo (timestep) como entrada de la siguiente predicción; de esta manera se establece un modelo de una única salida y un sistema recursivo de predicción hasta el límite definido. La estrategia directa, donde se entrena un modelo para cada paso de tiempo a predecir. Una combinación de las anteriores dos, donde se utilizan varios modelos y un sistema recursivo entre los propios modelos.

En la presente investigación, se utilizó el enfoque recursivo de predicción considerando que una de sus principales ventajas es su simplicidad y reducida carga computacional. Como desventaja se tiene que, a medida que el horizonte predictivo se incrementa, la precisión de las nuevas predicciones tiende a deteriorarse.

Optimización de hiperparámetros

Todo el afinamiento de los hiperparámetros realizado en este proyecto, se hizo con el algoritmo RandomizedSerachCV de la librería de Sklearn en Python. Ya que los modelos utilizados requieren mucha capacidad computacional, debido a la gran cantidad de parámetros para probar. Mediante la función RandomizedSerachCV es posible realizar una búsqueda aleatoria que intenta identificar la mejor estructura para el modelo probando, no todas las combinaciones posibles de los valores, sino solo un cierto número de ellos. De esta forma, aun cuando el número total de posibles combinaciones sea elevado, es posible limitar el tiempo de entrenamiento.

A continuación, se presenta el conjunto de hiperparámetros para el afinamiento junto con los parámetros seleccionados:

Tabla 5. Hiperparámetros modelos LSTM – RNN - GRU

Hiperparámetros	Valores Evaluados	Valor Seleccionado
Número de neuronas en capa	[50, 70, 100]	100
Función de Activación	['sigmoid','relu', 'tanh']	tanh
Inicializador de los pesos (kernel initializer)	['uniform', 'normal', 'zero']	uniform
Número de capas de la red	[3, 5, 7]	3
Tasa de abandono (Dropout rate)	[0.1, 0.3, 0.5]	0.5

Los hiperparámetros de los modelos RNN, LSTM y GRU, dieron de manera similar para todos los municipios y para el set de datos departamental. Sin embargo, si se presentaron algunas variaciones en los hiperparámetros de los modelos de red convolucional.

Tabla 6. Hiperparámetros modelos CNN

Hiperparámetros	**Valores Evaluados**	**Valores seleccionados**	
		B/manga B/meja P/cuesta Lebrija	Girón F/blanca Santander
Función de Activación	['relu','tanh', 'sigmoid', 'linear']	linear	tanh
Tasa de abandono (Dropout rate)	[0.1, 0.3, 0.5]	0.5	0.1
Número de capas convolucionales (1D) y Maxpooling	[2,3,4]	3	2
Numero de filtros en capa convolucional	[60,40,80,120,220]	80	220
Numero de neuronas en la capa densa	[128,64,32,256]	256	128
Optimizador	['adam', 'sgd', 'rmsprop']	rmsprop	rmsprop
padding	['valid', 'same', 'causal']	same	same

Arquitecturas de los modelos

A continuación, se presentan las implementaciones con Keras de las distintas arquitecturas de los modelos, tomando como base los resultados de la evaluación de hiperparámetros.

Técnica	Arquitectura	Configuración de parámetros
LSTM **RNN** **GRU**	• 1 capa de entrada • 3 capas intermedias (LSTM, RNN, GRU) con 100 neuronas. • 1 capa densa de salida con una neurona.	*units=100,* *activation='tanh',* *return_sequences = True* *kernel_initializer="uniform"* *dropout= 0.5* *optimizer = "adam"* *epochs=100* *kernel_regularizer=None,* *recurrent_regularizer=None,* *bias_regularizer=None,* *activity_regularizer=None,* *kernel_constraint=None,* *recurrent_constraint=None,* *bias_constraint=None,*
CNN	• 1 capa de entrada. • 3 capas Conv1D y MaxPooling1D con 80 filtros. • 1 capa densa con 256 neuronas. • 1 capa densa de salida, con 1 neurona.	*Filters=80,* *kernel_size= 2* *strides=1,* *padding=' same',* *activation= 'linear',* *use_bias=True, , kernel_regularizer=None,* *bias_regularizer=None,* *activity_regularizer=None,* *kernel_constraint=None,* *bias_constraint=None,* *kernel_regularizer=Ninguno,*

Es importante resaltar que, debió a la rápida convergencia evidenciada en los modelos durante las pruebas preliminares realizadas, los modelos se trabajaron con 100 epochs. Dicha convergencia puede estar asociada a diferentes factores, dentro de los cuales se destacan la capacidad, puesto que los modelos de Deep Learning tienen una gran capacidad de modelado,

lo que les permite aprender patrones complejos en los datos de entrenamiento con menos iteraciones. Así mismo, para la cantidad de datos con la que un modelo Deep Learning puede trabajar, los dataset utilizados son relativamente pequeños, lo cual pudo facilitar el aprendizaje del modelo sin tantas iteraciones haciendo que la convergencia sea más rápida.

Para la ejecución de los modelos, los datos se separaron entre el 80% para el entrenamiento y el 20% para las pruebas.

Métricas de evaluación

Como describen Kuhn y Johnson, al crear sistemas de predicción, el objetivo es generar salidas cercanas al valor real, con menor tasa de error, ya que la probabilidad de predecir valores iguales a los observados tiende a 0.

Las técnicas de error absoluto medio (RMSE) y error absoluto medio (MAE) son las formas evaluativas de aprendizaje automático y modelos de aprendizaje profundo (Carvajal et al., 2018).

- **MAE**

  El error absoluto medio, corresponde al promedio de las diferencias absolutas entre los valores reales y las predicciones. MAE es una puntuación lineal que significa que todas las diferencias individuales son ponderadas igualitariamente en el promedio. Se calcula utilizando la siguiente ecuación:

$$MAE = \frac{1}{N}\sum|y_i - y_i'|$$

- **RMSE**

El error cuadrático medio o la desviación cuadrática media es una de las medidas más utilizadas para evaluar la calidad de las predicciones. Muestra qué tan lejos caen las predicciones de los valores verdaderos medidos usando la distancia euclidiana. Formalmente se define de la siguiente manera:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y'_i - y_i)^2}{n}}$$

donde N es el número de puntos de datos, y(i) es la i-ésima medida y y´(i) es su predicción correspondiente.

- **Coeficiente de determinación (R^2)**

Es la proporción de la varianza total de la variable explicada por la regresión. Sirve para reflejar la bondad del ajuste de un modelo a la variable que se pretende explicar. Se define de la siguiente manera:

$$R^2 = \frac{\sum_{t=1}^{T}\left(\hat{Y}_t - \bar{Y}\right)^2}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

5.1.5 Fase V. Difusión y uso.

En esta última fase, se diseñó un prototipo de arquitectura visual como medio de presentación de los resultados y el análisis de la información. Todo el código fuente de la herramienta de visualización y los modelos puede ser encontrado en el siguiente enlace Drive: https://drive.google.com/drive/folders/1ou5T1AD-33s5FGQ008a7y3xVOwTwA7u-?usp=sharing

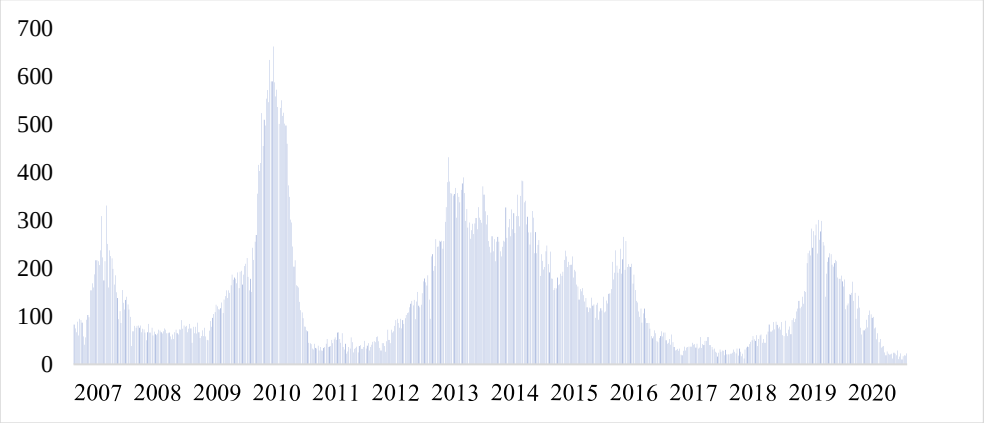
6. Resultados y discusión

6.1. Análisis descriptivo de los datos

En primer lugar, se realizó un análisis descriptivo de los datos con el fin de conocer cada una de las variables, así como su comportamiento.

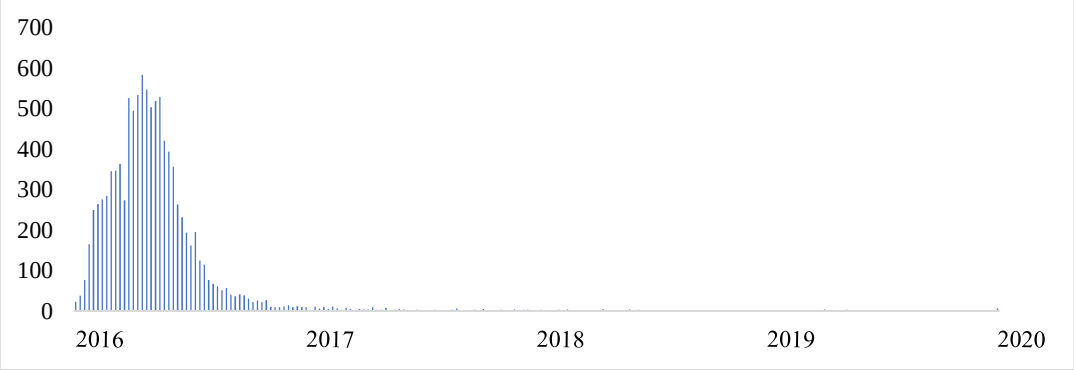
6.1.1 Análisis descriptivo de casos de Dengue, Zika y Chikungunya

Figura 13 *Comportamiento de casos de Dengue Santander*



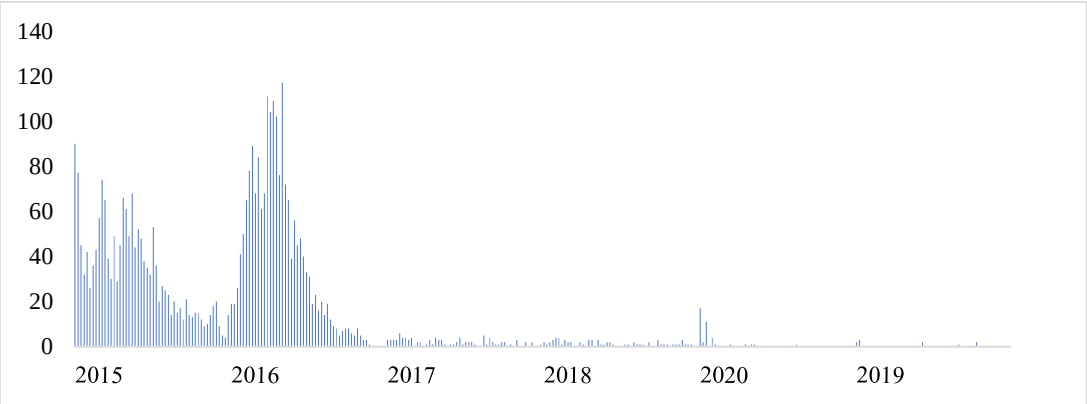
La base de datos del departamento de Santander de casos de dengue se compone de 728 datos, correspondientes a 52 semanas epidemiológicas por año, desde el año 2007 al 2020. En total, se reportan 105.965 casos de dengue en el periodo 2007-2020. En la figura es posible evidenciar como se registran ciclos de transmisión epidémicos cada tres años, con brotes epidémicos en 2007, 2010, 2013, 2016 y 2019. Siendo el más relevante el registrado en el 2010. El valor máximo de número de casos reportado en el periodo evaluado, se presentó en la semana 19 del año 2010, donde se reportaron 663 casos. Mientras que el valor mínimo reportado ha sido 10 casos en la semana 48 del año 2020. El promedio de casos por semana epidemiológica es de 145 casos.

Figura 14 *Comportamiento de casos de Zika Santander*



Con relación a los casos de Zika, en total en el departamento se han reportado 10.398 casos. El primer reporte en Santander se generó en la semana 45 del año 2015, presentándose en total 32 casos este año en el departamento. En el año 2016 se reportaron 10.086 casos, siendo este el único año donde se ha presentado brote en el departamento, representando un porcentaje del 97% de los casos reportados desde el 2015. En el periodo 2017-2020 se tiene un reporte de 280 casos.

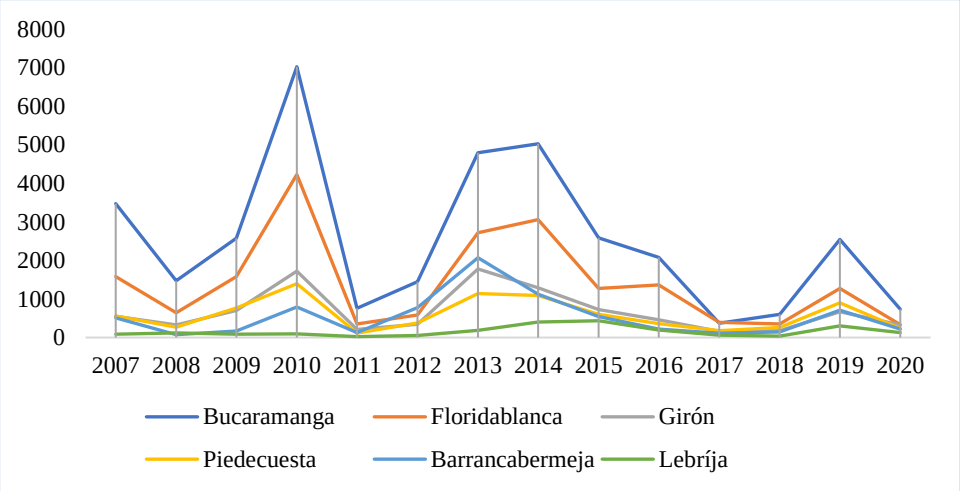
Figura 15 *Comportamiento de casos de Chikungunya Santander*



Respecto a los casos de Chikungunya, en total en el departamento se han reportado 3.989 casos. El primer reporte en Santander se generó en la semana 36 del año 2014 ,

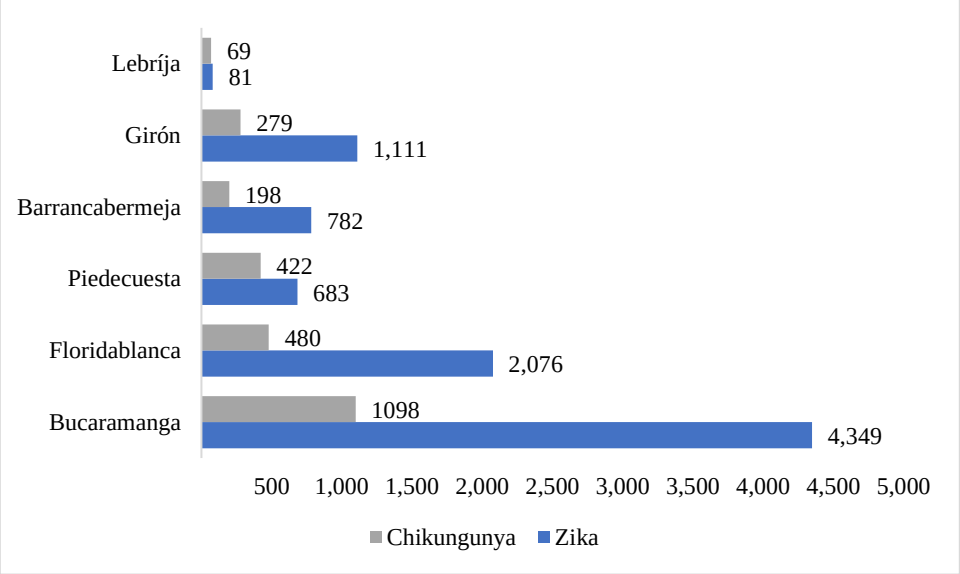
presentándose en total 134 casos este año en el departamento. En los años 2015 y 2016 se reportaron 1.727 y 1.915 casos respectivamente, representando un porcentaje del 91% de los casos totales reportados. En el periodo 2017-2020 se tiene un reporte de 213 casos.

Figura 16. *Comportamiento de casos de Dengue en municipios*



A nivel municipal, también es posible apreciar los ciclos de transmisión epidémicos cada tres años, con brotes epidémicos en 2007, 2010, 2013, 2016 y 2019. Bucaramanga es el municipio que representa la mayor proporción de casos seguido de Floridablanca, pues son las áreas con mayor concentración de población, lo cual facilita la existencia y proliferación de criaderos potenciales del vector, mayor disponibilidad de poblaciones vulnerables a la infección y al contacto hombre-vector-virus.

Figura 17 *Casos totales de Zika y Chikungunya periodo 2014-2020 en Municipios*



Los seis municipios estudiados, representan en total el 87% y 64% de los casos de Zika y Chicungunya en el departamento, respectivamente. Sólo Bucaramanga abarca el 42% de los casos de Zika y el 28% de los casos de Chikungunya reportados en el periodo 2014-2020.

6.1.2 Análisis descriptivo variables climáticas

Las variables de temperatura, humedad relativa y precipitaciones se obtuvieron para los municipios de Barrancabermeja, Lebrija y Girón. Para los municipios de Piedecuesta, Floridablanca y Bucaramanga, solo se contó con disponibilidad de la variable precipitaciones. A continuación, se presenta el análisis descriptivo de las mimas.

- **Barrancabermeja, Girón y Lebrija**

La base de datos de los municipios de Barrancabermeja, Girón y Lebrija, se componen de 624 registros que corresponden a 52 registros por año, desde el 2008 a 2019. Cada uno de los 52 registros corresponde a una semana epidemiológica registrada.

Para cada semana epidemiológica y de cada año mencionado, se registran las siguientes características:

- Casos Dengue (Casos)
- Promedio Diario Precipitaciones (Lt/m2)
- Suma Total Precipitaciones Semana (Lt/m2)
- Temperatura Máxima Diaria Promedio (°C)
- Temperatura Máxima (Máxima Semana) (°C)
- Temperatura Mínima Diaria Promedio (°C)
- Temperatura Mínima (Mínima Semana) (°C)
- Temperatura Seca Máxima Diaria Promedio (°C)
- Temperatura Seca Máxima (Máxima Semana) (°C)
- Temperatura Seca Mínima Diaria Promedio (°C)
- Temperatura Seca Mínima (Mínima Semana) (°C)
- Humedad Relativa Máxima (%)
- Humedad Relativa Mínima (%

Figura 18 Comportamiento variables climáticas Barrancabermeja

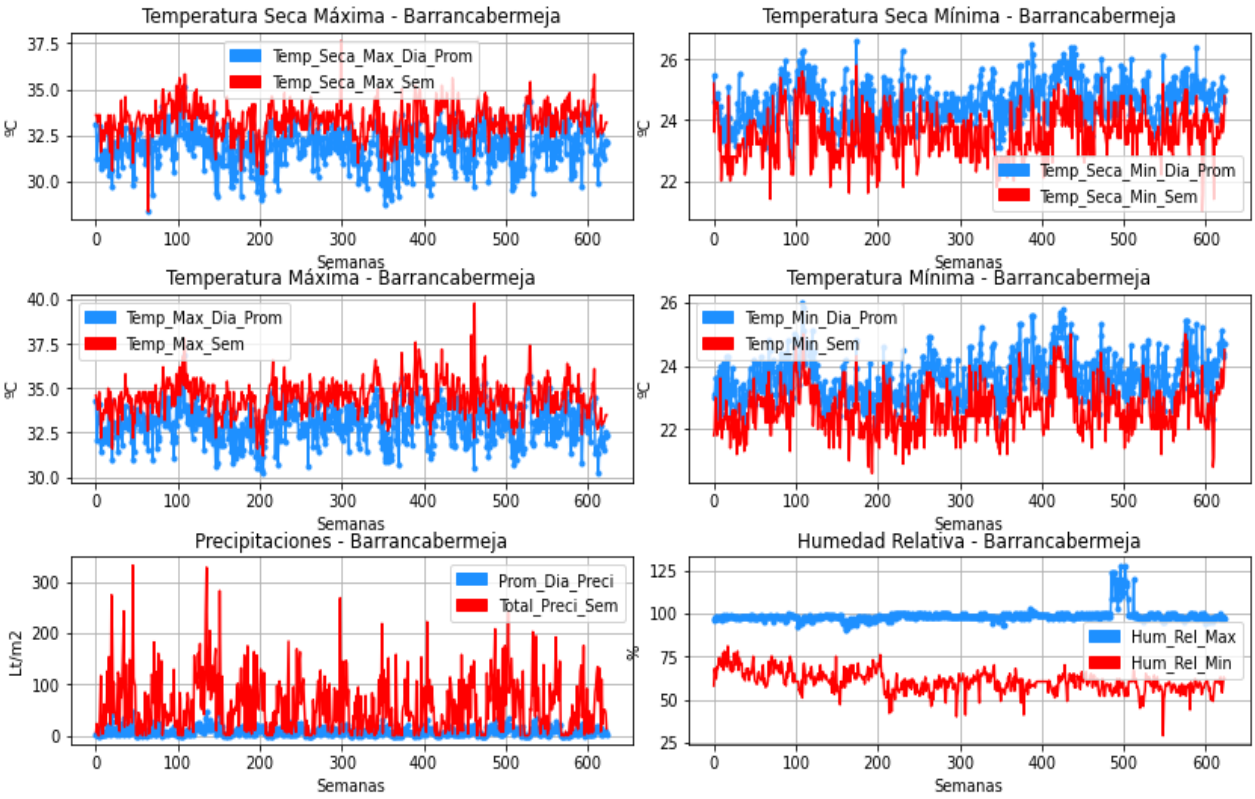


Tabla 7 Estadísticos básicos variables climáticas Barrancabermeja

	Media	Desviación Estándar	Valor Mínimo	Valor Máximo
Temp_Seca_Max_Dia_Prom	31,8	1,07	28,4	35,1
Temp_Seca_Max_Sem	33,2	0,91	28,4	37,6
Temp_Seca_Min_Dia_Prom	24,6	0,67	22,8	26,6
Temp_Seca_Min_Sem	23,5	0,78	21,0	25,8
Prom_Dia_Preci	8,4	7,9	0,0	47,3
Total_Preci_Sem	58,5	55,3	0,0	331,7
Temp_Max_Dia_Prom	32,9	1,04	30,2	36,9
Temp_Max_Sem	34,4	0,98	31,2	39,8
Temp_Min_Dia_Prom	23,6	0,72	22,1	25,9
Temp_Min_Sem	22,6	0,75	20,6	25,0
Hum_Rel_Max	98,5	3,95	90,0	128
Hum_Rel_Min	60,3	6,16	29,0	81,0

El municipio de Barrancabermeja cuenta con temperaturas máximas que varían entre los 28,4 °C y los 36,9 °C y las temperaturas mínimas entre los 22,6°C y los 25,8 °C. presenta precipitaciones promedio semanales de 58,5 mm. Los años 2008 y 2010 fueron los años con mayor cantidad de precipitaciones con promedios de 73,5 mm y 91,8 mm respectivamente. Con relación a la humedad relativa, se aprecian valores superiores al 100% lo cual podría deberse a errores de registro. En promedio la humedad relativa máxima alcanza valores superiores al 95%.

Figura 19 Comportamiento variables climáticas Girón

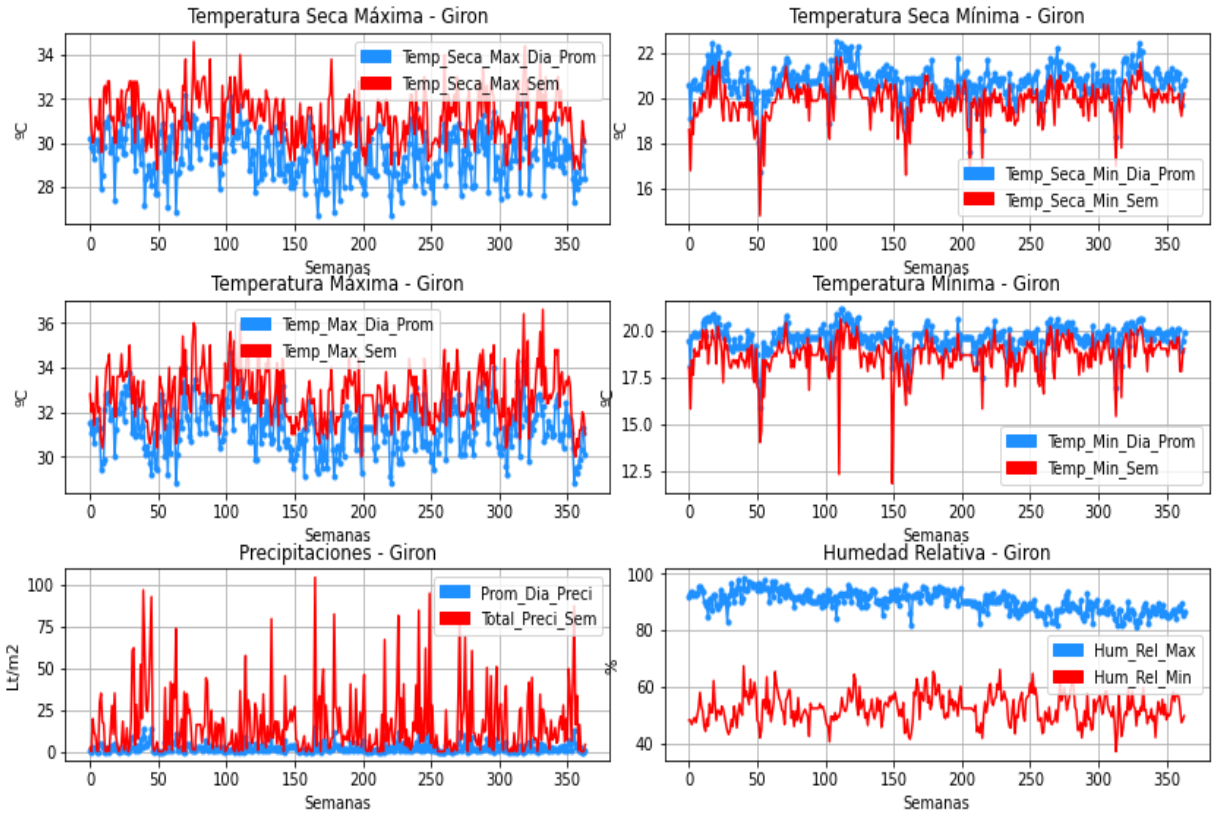


Tabla 8 Estadísticos básicos variables climáticas Girón

	Media	Desviación Estándar	Valor Mínimo	Valor Máximo
Temp_Seca_Max_Dia_Prom	29,4	1,08	26,7	32,2
Temp_Seca_Max_Sem	31,1	1,11	28,8	34,6
Temp_Seca_Min_Dia_Prom	20,7	0,72	16,7	22,5
Temp_Seca_Min_Sem	19,8	0,83	14,8	21,8
Prom_Dia_Preci	2,29	2,76	0,0	14,9
Total_Preci_Sem	16,01	19,3	0,0	104,3
Temp_Max_Dia_Prom	31,3	1,07	28,8	34,7
Temp_Max_Sem	32,7	1,21	30,0	36,6
Temp_Min_Dia_Prom	19,5	0,68	15,9	21,1
Temp_Min_Sem	18,6	1,0	11,8	20,6
Hum_Rel_Max	90,1	3,5	81,4	98,7
Hum_Rel_Min	52,2	5,06	37,0	67,3

El municipio de Girón, cuenta con rangos de temperatura máximas similares a los de Barrancabermeja, los valores varían entre los 28,8 °C y los 36,6 °C, sin embargo, las temperaturas mínimas son más bajas, encontrándose entre los 11,8 °C y los 21,8 °C y la temperatura promedio no supera los 30°C. Presenta precipitaciones promedio semanales de 16,01 mm. De manera general, el mayor número de observaciones correspondió a semanas con precipitación menor a 10 mm, el año 2014 correspondió al año con mayores precipitaciones con un promedio de 18.75mm. Respecto a la humedad relativa el valor máximo reportado es de 98,7%.

Figura 20 Comportamiento variables climáticas Lebrija

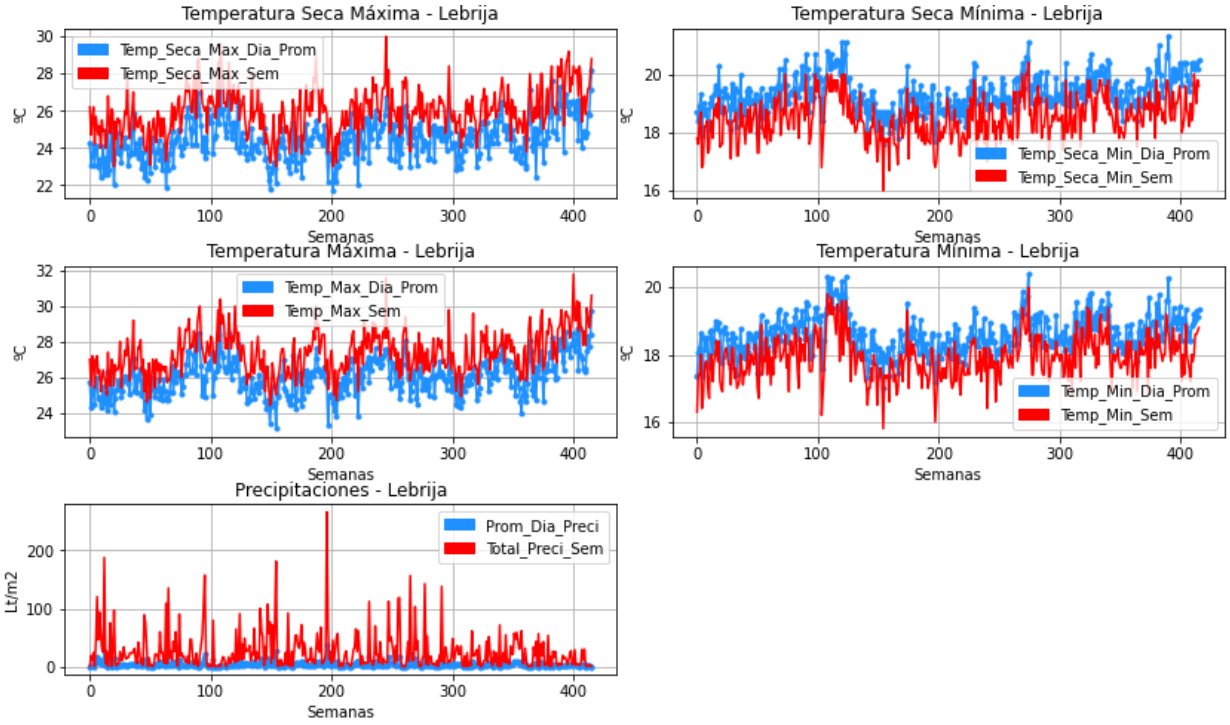


Tabla 9 Estadísticos básicos variables climáticas Lebrija

	Media	Desviación Estándar	Valor Mínimo	Valor Máximo
Temp_Seca_Max_Dia_Prom	24,4	1,10	21,7	28,2
Temp_Seca_Max_Sem	25,9	1,21	23,0	30,0
Temp_Seca_Min_Dia_Prom	19,3	0,65	17,6	21,3
Temp_Seca_Min_Sem	18,5	0,72	16,0	20,4
Prom_Dia_Preci	3,57	4,61	0,0	38,03
Total_Preci_Sem	25,03	32,33	0,0	266,2
Temp_Max_Dia_Prom	26,0	1,06	23,1	29,7
Temp_Max_Sem	27,3	1,21	24,4	31,8
Temp_Min_Dia_Prom	18,6	0,58	17,0	20,39
Temp_Min_Sem	17,9	0,65	15,8	20,0

El municipio de Lebrija, cuenta con rangos de temperatura más bajas que los municipios de Barrancabermeja y Girón, el promedio de sus temperaturas máximas no supera

los 30°C y las temperaturas mínimas de mantienen por debajo de los 20°C. Con relación a las precipitaciones, el mayor número de observaciones correspondió a semanas con precipitación entre los 0 y 15 mm, el año 2010 correspondió al año con mayores precipitaciones con un promedio de 30,2 mm.

- **Bucaramanga, Floridablanca y Piedecuesta**

Figura 21 Comportamiento variables climáticas Bucaramanga, Florida, Piedecuesta

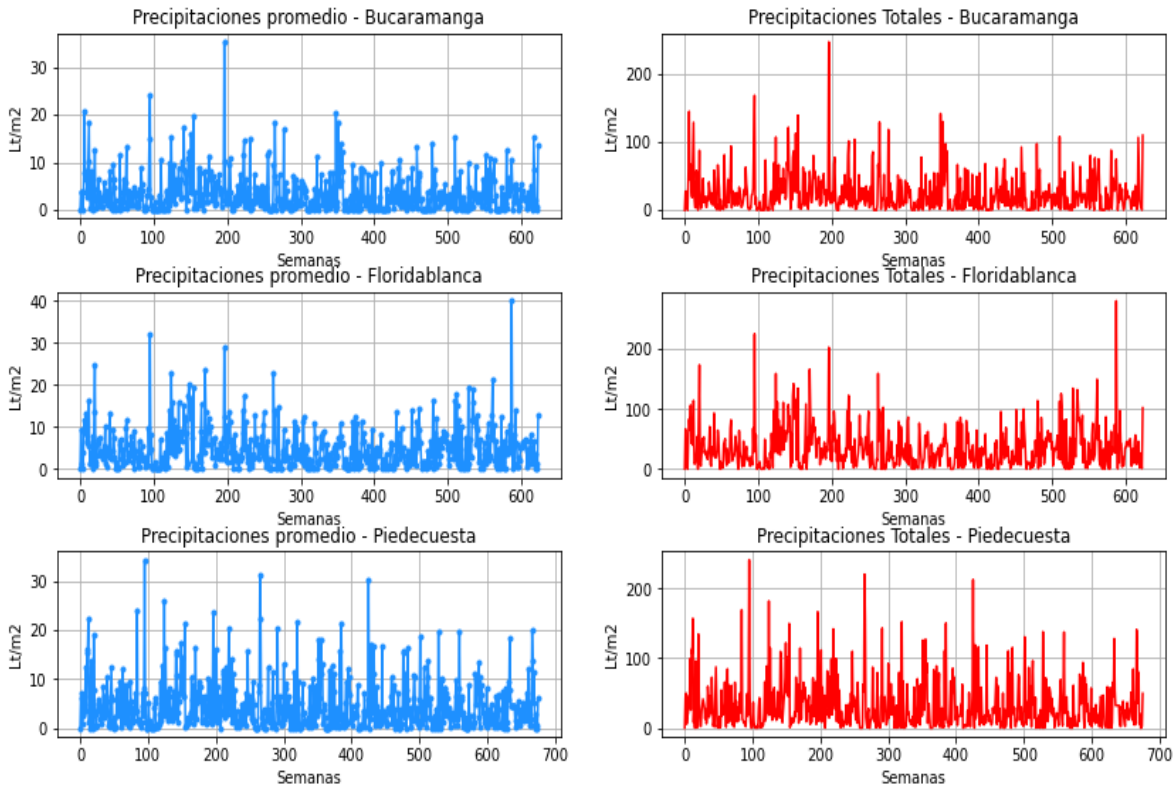


Tabla 10 Estadísticos básicos variables climáticas Bucaramanga, Florida, Piedecuesta

Municipio		Media	Desviación Estándar	Valor Mínimo	Valor Máximo
Bucaramanga	Prom_Dia_Preci	3,52	4,04	0,0	35,2
	Total_Preci_Sem	24,7	28,3	0,0	246,5
Floridablanca	Prom_Dia_Preci	4,92	4,88	0,0	40,01
	Total_Preci_Sem	34,5	34,2	0,0	280,1
Piedecuesta	Prom_Dia_Preci	4,59	5,04	0,0	34,4
	Total_Preci_Sem	32,2	35,4	0,0	241,1

Los municipios de Bucaramanga, Floridablanca y Piedecuesta, presentan un comportamiento similar en las precipitaciones. A lo largo de todo el periodo evaluado no se evidencian grandes cambios en el comportamiento de las precipitaciones.

6.3 Resultados de los modelos

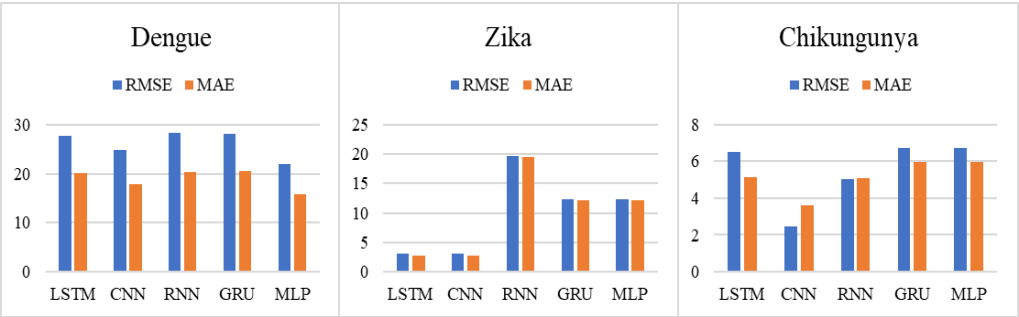
6.3.2 Modelo departamental para Dengue, Zika y Chikungunya

Con el objetivo de probar la hipótesis planteada en la sección 4, adicionalmente a los modelos Deep Learning, se consideró un modelo de red neuronal tradicional MLP, al cual se le conoce también como perceptrón multicapa. A continuación, se presenta la comparación del desempeño de los algoritmos.

Tabla 11. Resultados modelo departamental para Dengue, Zika y Chikungunya

Modelos	Métricas	Dengue	Zika	Chikungunya
LSTM	RMSE	27,63	3,07	6,50
	MAE	20,06	2,81	5,13
	R2(%)	87,04%	-3,58%[2]	8,74%
CNN	RMSE	24,89	3,16	2,47
	MAE	17,82	2,69	3,63
	R2(%)	89,48%	-3,87%	81,10%
RNN	RMSE	28,30	19,58	5,03
	MAE	20,25	19,53	5,08
	R2(%)	86,41%	-185,64%	28,37%
GRU	RMSE	28,05	12,23	6,73
	MAE	20,50	12,14	5,96
	R2(%)	86,65%	-71,77%	2,32%
MLP	RMSE	22,04	12,23	6,73
	MAE	15,87	12,14	5,96
	R2(%)	91,76%	-71,77%	2,32%

Figura 22. Comportamiento RMSE-MAE modelo departamental



Al revisar los resultados, con relación a los datos del Dengue se evidencia que los valores del RMSE y MAE para todos los modelos estuvieron por debajo de 30, considerándose una precisión alta, teniendo en cuenta que el promedio de casos de Dengue por semana epidemiológica es de 145 casos. Así mismo, los valores de R2 para todos los modelos fue superior al 85%, lo que demuestra la aptitud de los modelos. El modelo MLP

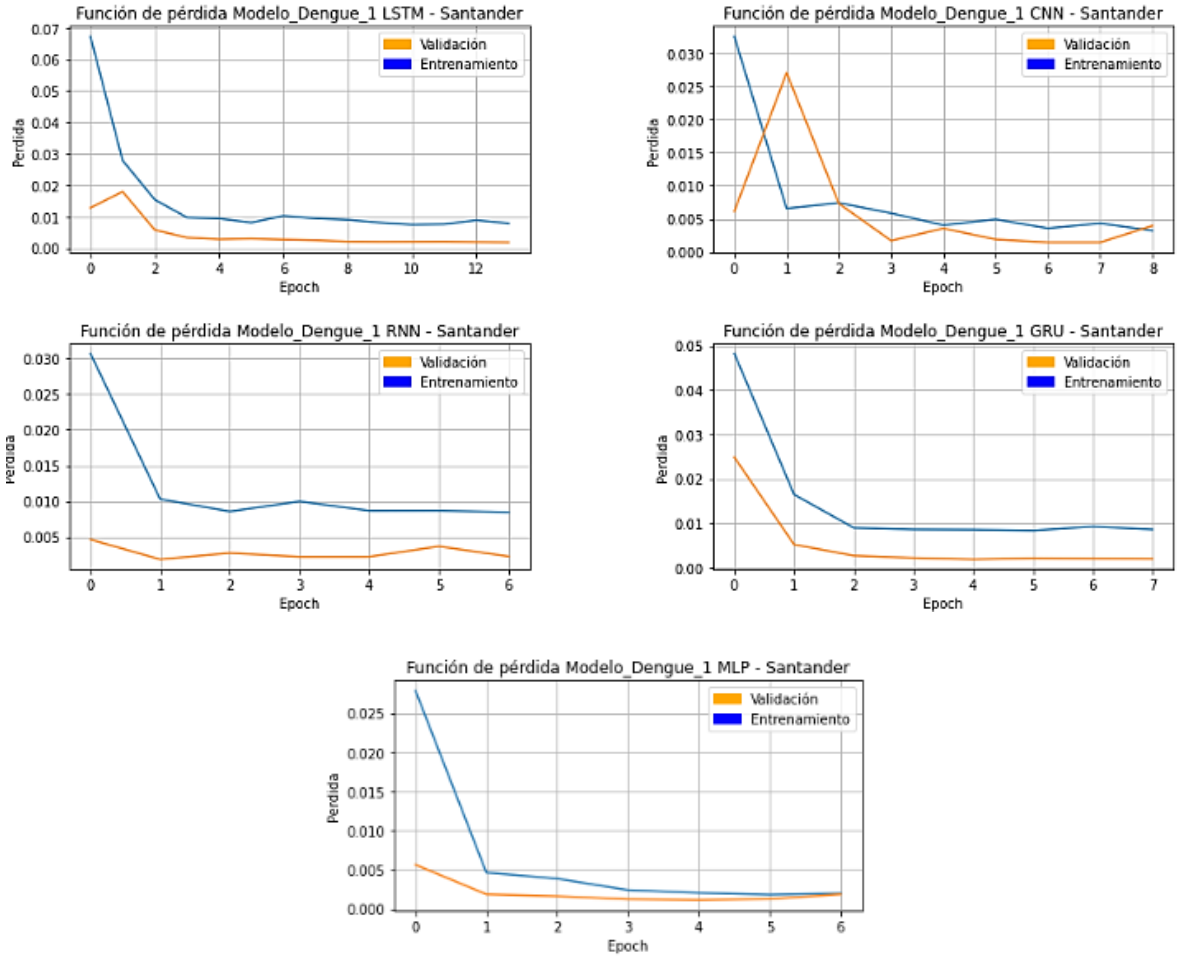
[2] Los valores negativos para la métrica R2 Score significan que el modelo es tan malo que su predicción es peor que simplemente hacer una predicción basada en la media de los datos de respuesta. En este caso, el modelo está haciendo predicciones peores que el caso de un modelo constante y su R2 score es negativo.

fue superior a todos los demás, con los valores de error más pequeños y un ajuste por encima del 90%, de los modelos Deep Learning, CNN presentó los mejores resultados. La superioridad del modelo MLP sobre los modelos Deep Learning, particularmente con los datos de dengue, podría deberse a que sobre los datos que se están trabajando tienen una dificultad relativamente baja, por tanto, la arquitectura del modelo MLP se habría adaptado mejor.

Respecto a los modelos para Zika y Chicungunya, los resultados fueron desalentadores, pues los modelos no tuvieron la capacidad de ajustarse a los datos, una de las principales razones, se debe a que la cantidad de semanas con observaciones de cero notificaciones de casos, al contar con tan pocos datos, el modelo no puede aprender del conjunto de datos de entrenamiento, por ejemplo en el caso del conjunto de datos del Zika, se tiene gran cantidad de observaciones para el año 2016, el promedio de casos por semana epidemiológica es de 193, mientras que para los años 2017, 2018 y 2019 el promedio de casos por semana disminuye abruptamente a 3.6, 1.07 y 0.57 respectivamente, de modo que, al tomar los datos de 2016, para predecir años posteriores los errores sean muy grandes, pues el modelo se entrena con datos que están entre los 9 y 584 casos, para predecir datos que están entre 0 y 11 casos. Ahora, al intentar predecir un año con sus mismos datos, por ejemplo, tomar 42 semanas (80%) para predecir, las próximas 10 semanas (20%), el modelo no cuenta con suficientes datos para aprender e igualmente, el conjunto de datos de validación se queda con muy pocos ejemplos en comparación con el conjunto de datos de entrenamiento, no proporcionando suficiente información para evaluar la capacidad del modelo para generalizar. Esto mismo ocurre con el conjunto de datos de Chikungunya, donde solo se cuenta con datos representativos para los años 2015-2016.

El desempeño de los modelos de regresión también es evaluado mediante las curvas de aprendizaje, estas curvas permiten establecer si existe sobre o sub-aprendizaje. Para los modelos de predicción de dengue, se evidencia en general un buen ajuste, considerando que las curvas de validación y entrenamiento disminuyen hasta un punto de estabilidad con una brecha mínima entre los dos valores de pérdida finales. En las gráficas se visualiza que una pérdida de validación inferior a la pérdida de entrenamiento. En este caso, indica que el conjunto de datos de validación puede ser más fácil de predecir para el modelo que el conjunto de datos de entrenamiento.

Figura 23. Funciones de pérdida modelo Dengue



Para el caso de los modelos para Zika y Chicungunya, como se mencionó anteriormente los modelos no tuvieron la capacidad de ajustarse debido a la complejidad de los conjuntos de datos, se evidencian valores ruidosos y fluctuaciones aleatorias en las curvas de perdida, lo que indica que los modelos no lograron aprender el conjunto de datos de entrenamiento en absoluto, pues no proporciona información suficiente para aprender el problema, en relación con el conjunto de datos de validación utilizado para evaluarlo, a su vez, el conjunto de datos de validación es relativamente poco representativo.

Figura 24. Funciones de pérdida Modelos Zika

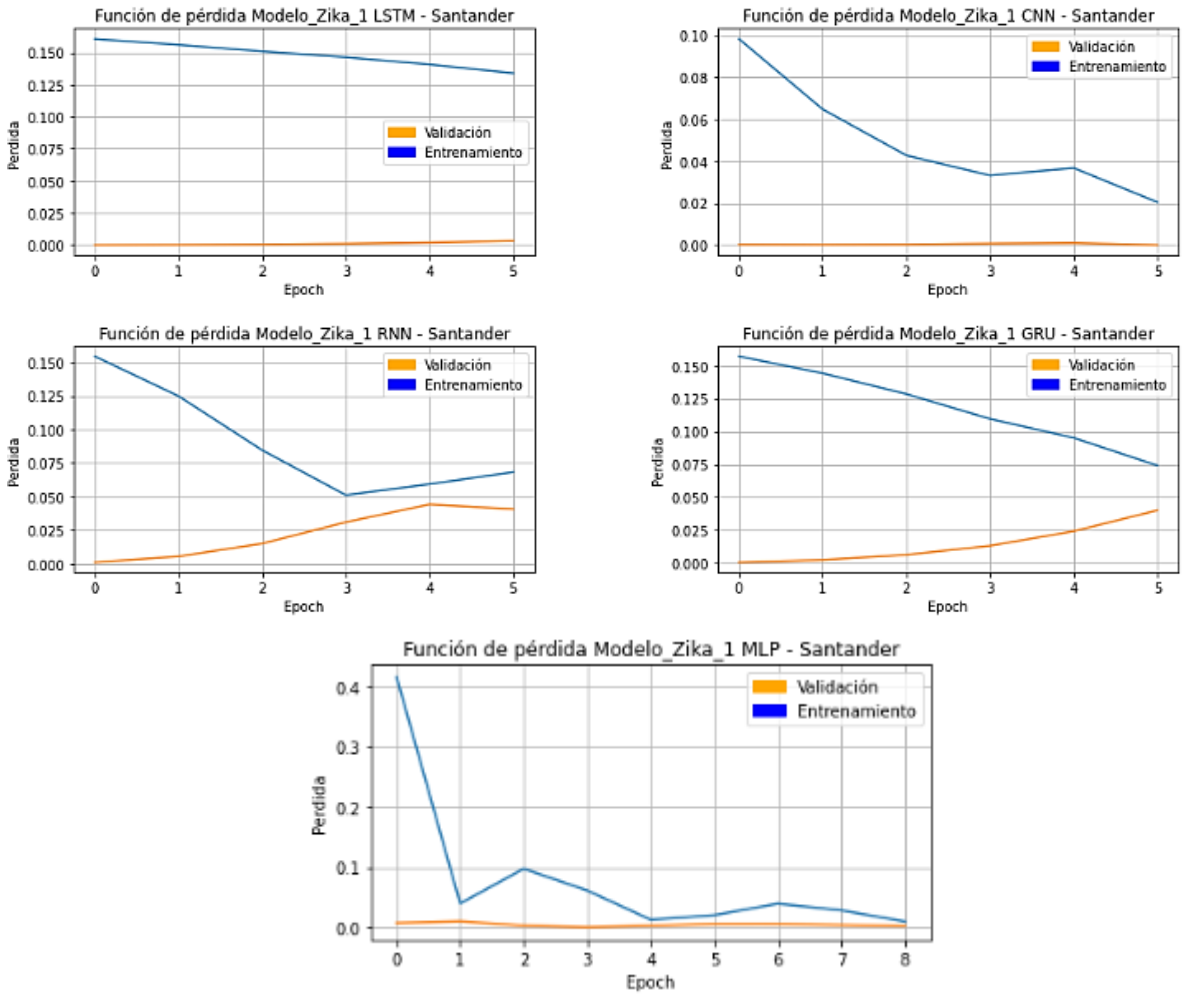
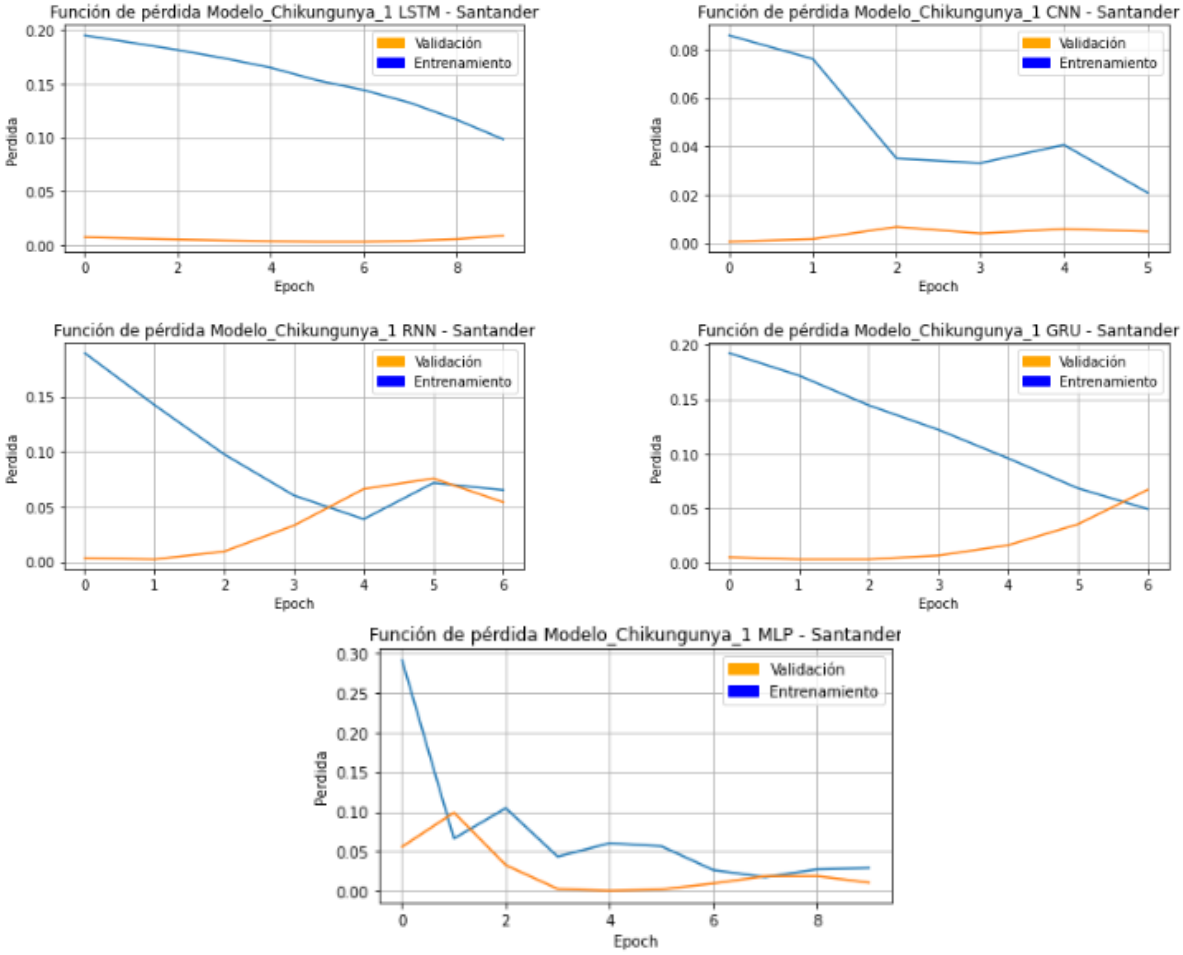
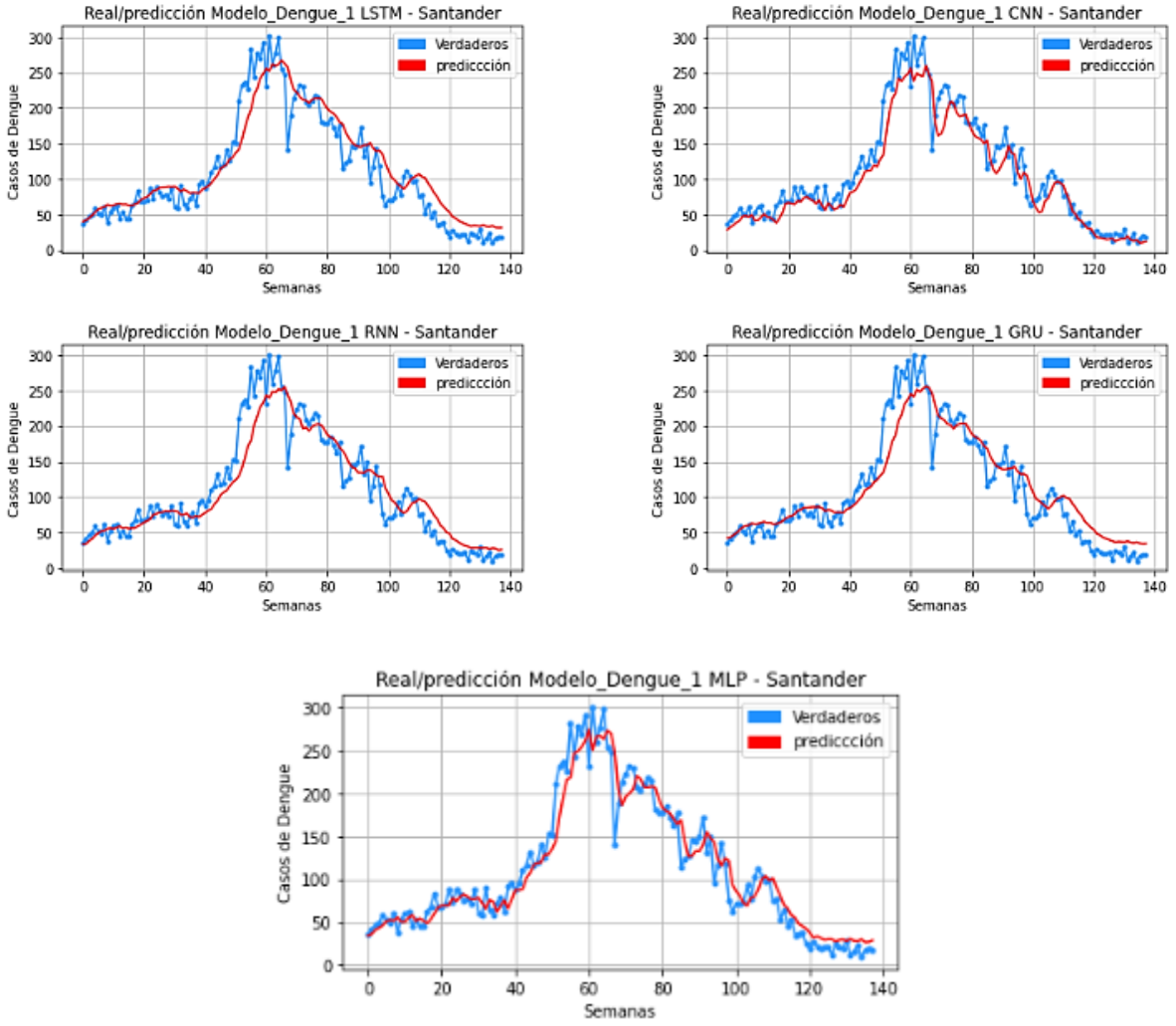


Figura 25. Funciones de pérdida Modelos Chikungunya



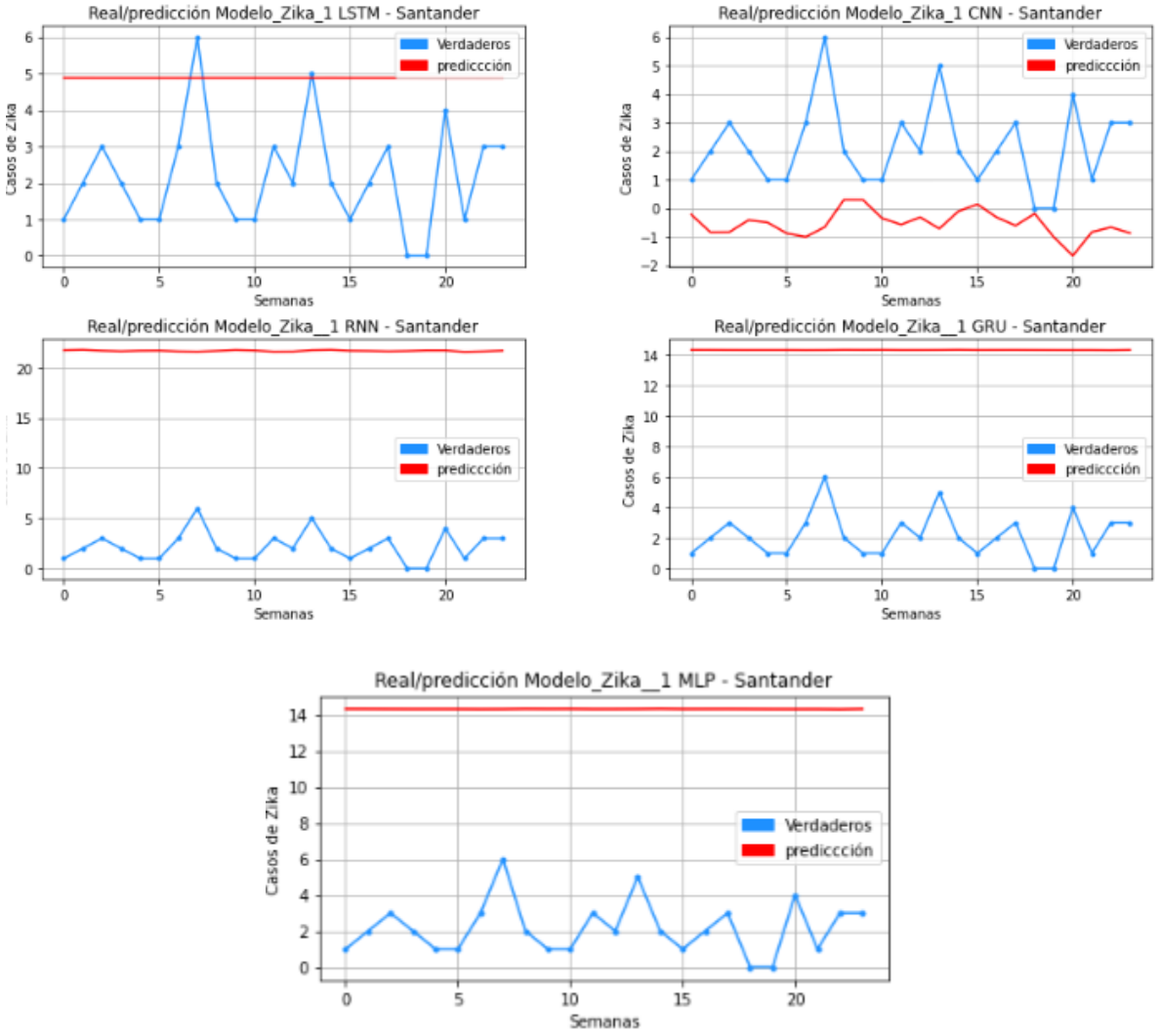
A continuación, se muestran las gráficas de predicción para cada uno de los modelos evaluados, en donde se comparan los datos de validación verdaderos versus las predicciones realizadas por los modelos, para el Dengue se visualiza como los modelos MLP y CNN lograron un mejor ajuste respecto a los demás modelos, se evidencia como MLP se ajusta mejor a los picos altos, mientras que CNN se ajusta mejor a los picos bajos, probablemente a esto se deba los valores de error mas bajos del MLP, al ajustarse mejor a los valores altos, los errores son mejores.

Figura 26. Predicciones Modelo Dengue



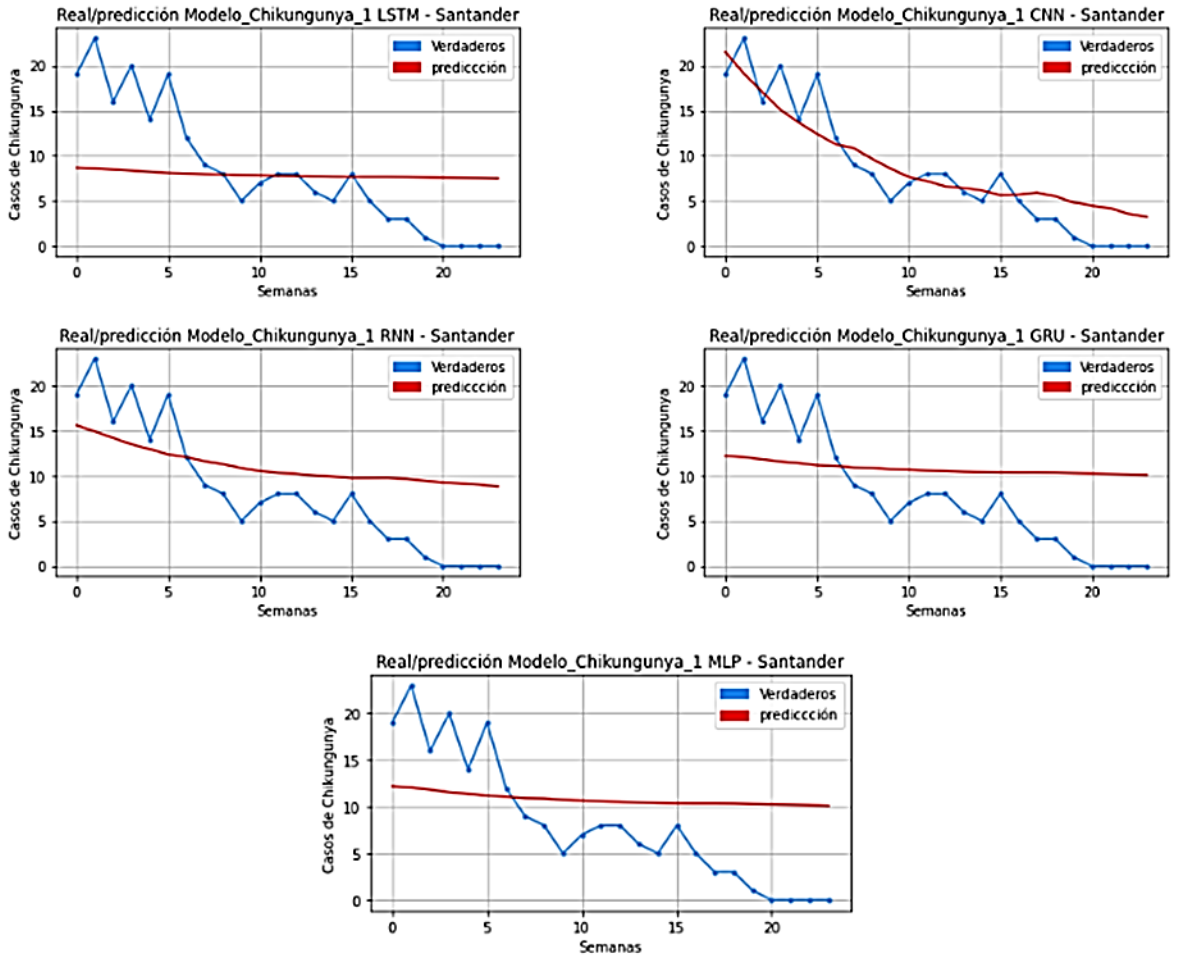
Por su parte las predicciones de los modelos para Zika, se evidencia como, al ser los valores de entrenamiento tan altos (valores del año 2016 y parte de 2017), e intentar predecir valores tan bajos (valores año 2017-2018) los modelos se quedan sin capacidad de predicción, haciendo que los valores predichos estén muy por encima de los valores reales, exceptuando el modelo CNN el cual a diferencia de todos los demás predijo valores muy por debajo.

Figura 27. Predicciones Modelos Zika



Con relación a los datos de Chikungunya, se evidencia como el comportamiento de los valores predichos es casi una línea recta, indicando que los modelos predijeron y se mantuvieron en valores medios, sin embargo, se visualiza como el modelo CNN fue el único que mantuvo la tendencia de los datos reales y presentó el mejor ajuste, dando cuenta de la gran capacidad de este modelo para trabajar con datos más complejos.

Figura 28. Predicciones Modelos Chikungunya



6.3.3 Modelos a nivel Municipal

Respecto a los modelos a nivel municipal, se debe resaltar que no fue posible realizar los modelos para los datos de Zika y Chikungunya, dado que a nivel departamental donde la cantidad de datos es mayor, no se lograron predicciones adecuadas, debido a la gran cantidad de semanas con observaciones de cero notificaciones de casos, a nivel municipal esta cantidad es mucho mayor. Por tanto, sólo se trabajó con los datos de Dengue.

Como se mencionó en la metodología, para los municipios de Barrancabermeja, Girón y Lebrija, se consideraron diferentes escenarios, con diferente combinación de variables, partiendo de los resultados obtenidos del análisis de importancia de las variables.

- ***Resultados modelos Barrancabermeja***

Tabla 12. Resultados modelos Barrancabermeja

Modelo		**Escenarios**			
		1	**2**	**3**	**4**
LSTM	RMSE	3,91	4,58	3,73	4,19
	MAE	3,06	3,79	2,60	3,44
	R2(%)	64,86%	51,83%	68,04%	59,67%
CNN	RMSE	3,61	3,68	3,78	3,30
	MAE	2,83	3,06	2,98	2,70
	R2(%)	70,01%	68,91%	67,19%	74,98%
RNN	RMSE	4,15	3,81	4,07	4,14
	MAE	3,16	2,94	3,24	3,35
	R2(%)	60,43%	66,69%	61,92%	60,68%
GRU	RMSE	4,23	4,06	4,02	3,83
	MAE	3,54	3,35	3,25	3,05
	R2(%)	58,88%	62,16%	62,78%	66,23%
MLP	RMSE	4,01	3,31	3,45	3,09
	MAE	2,90	2,62	2,53	2,23
	R2(%)	63,09%	74,73%	72,58%	77,98%

Figura 29. Desempeño RMSE Modelos Barrancabermeja

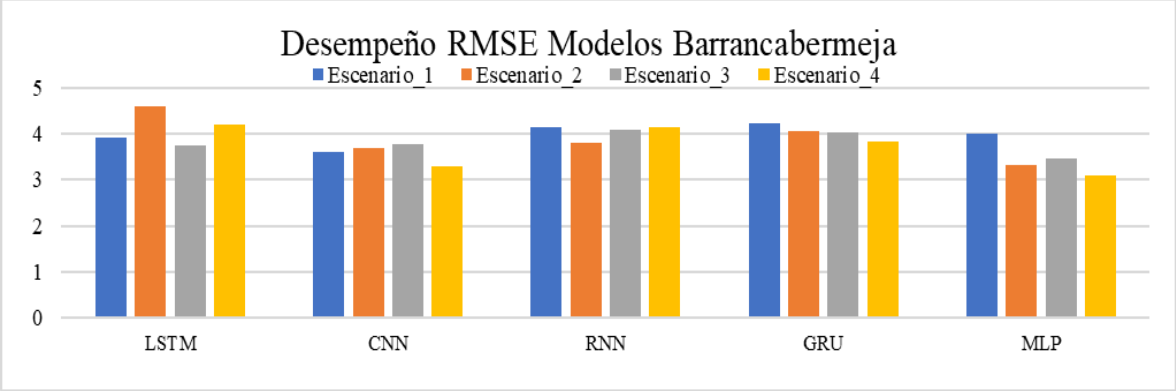
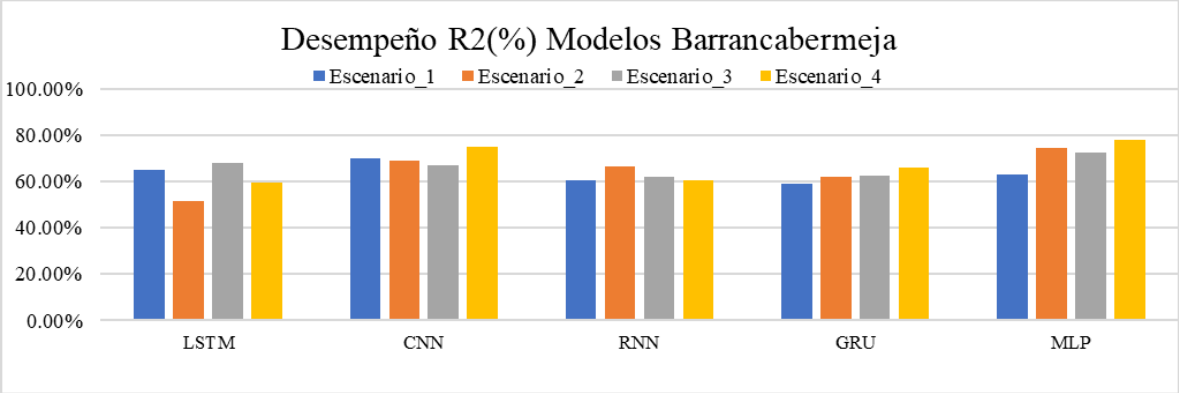


Figura 30. Desempeño R2(%) Modelos Barrancabermeja



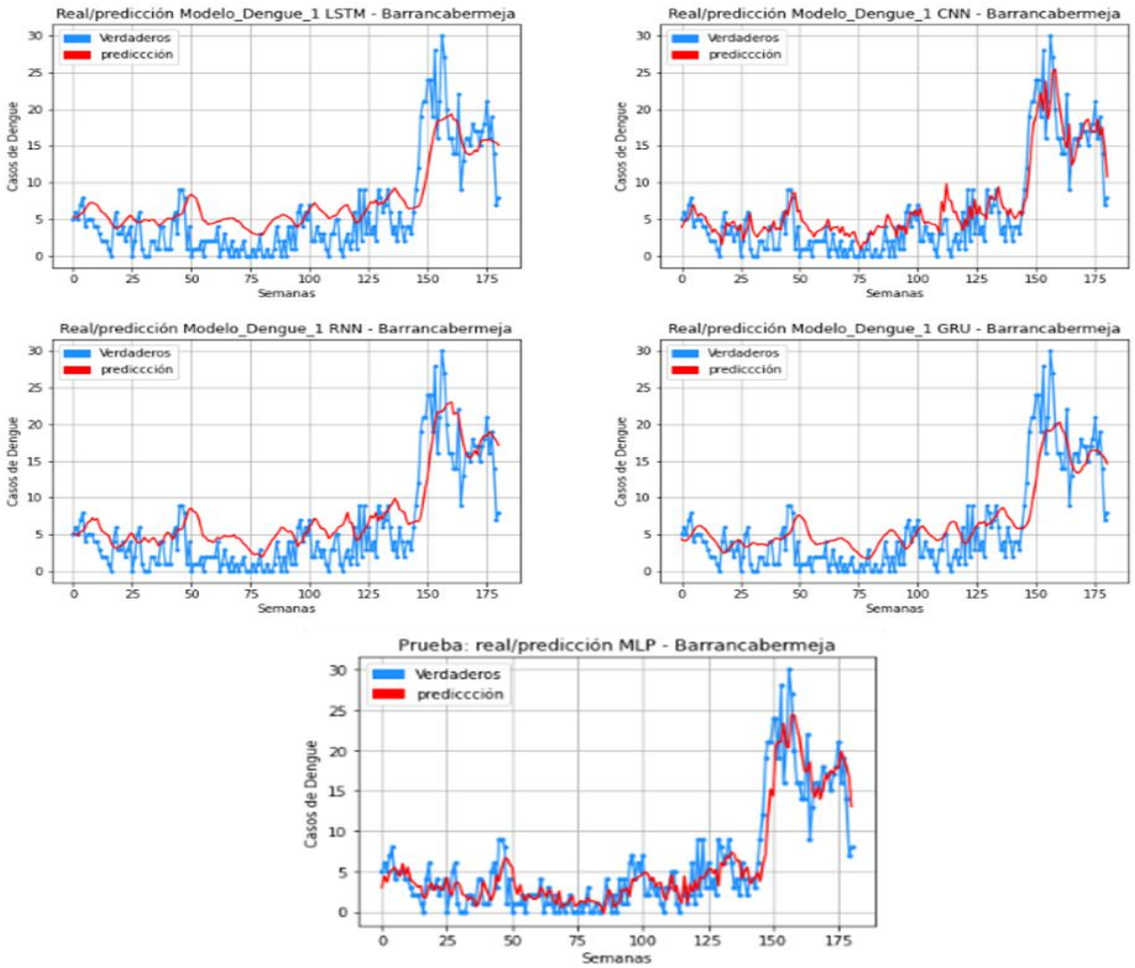
El mejor comportamiento de manera general para Barrancabermeja se presenta en el escenario 4 donde el promedio de error de los modelos fue de 3.71, mientras que el de los escenarios 1, 2 y 3 fue de 3.98, 3.89 y 3.81 respectivamente. Esto demuestra que, para el municipio de Barrancabermeja, la variable predictora más importante es la humedad relativa mínima, pues el escenario 4 lo compone la humedad relativa mínima con rezago 4 y con rezago 6.

Con base en los modelos, los mejores desempeños los tuvieron MLP y CNN con promedios de error de 3.59 y 3.46 respectivamente. De manera general, el promedio de ajuste

de los modelos estuvo por encima del 65%, sin embargo, fue en el escenario 4 donde los modelos MLP y CNN alcanzaron su mayor ajuste con 77.98% y 74.98% respectivamente.

A continuación, se presentan las gráficas de predicción para el escenario 4. Se observa como los modelos MLP y CNN presentan un mejor ajuste a los datos respecto a los demás modelos. Se evidencia que el modelo CNN predice valores superiores a los valores reales, haciendo que sea mejor prediciendo los picos de casos, mientras que el modelo MLP predice valores intermedios, sin embargo, no es tan adecuado prediciendo los picos ya sea para valores altos o bajos.

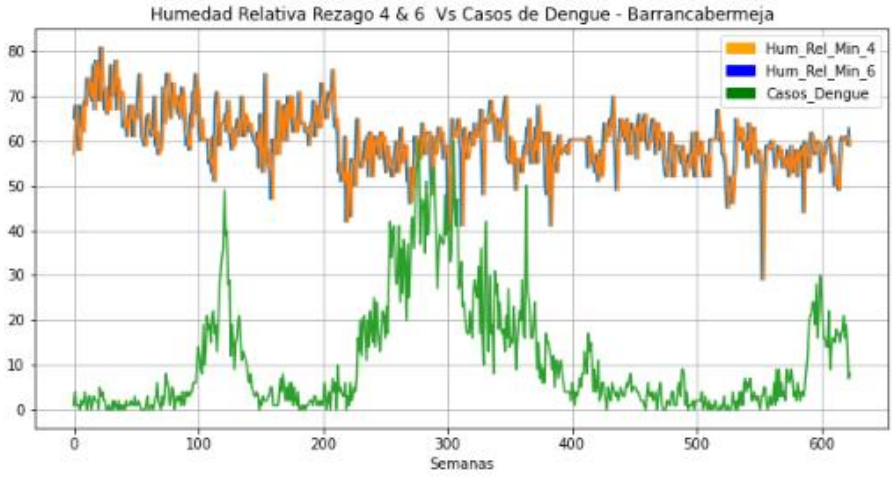
Figura 31. Graficas de Predicción Modelos Barrancabermeja – Escenario 4



El hecho de que la humedad relativa mínima para el municipio de Barrancabermeja, cuyo valor máximo ha alcanzado 81% y en promedio es del 60%, sea la variable que mejores resultados arroja, podría deberse a que, dado que el municipio maneja temperaturas altas, en promedio por encima de los 30°C, se generan condiciones ideales de humedad, como lo explica Naish et al. (2014) la presión del vapor de saturación aumenta exponencialmente a medida que aumenta la temperatura, la humedad relativa varía en función tanto de la temperatura como del contenido real de vapor de agua del aire.

Al realizar la gráfica de la humedad relativa mínima con rezago 4 y 6 versus los casos de Dengue, se evidencia como en los picos donde mayor cantidad de casos hubo, la humedad se encontraba entre el 50% y el 60%. Estos resultados no están alejados de lo encontrado por otros autores, de manera similar, en el estudio de Noureldin & Shaffer (2019) sus resultados mostraron que la humedad relativa ≥56% se asoció significativamente con un mayor número de casos de dengue, así mismo Ouattara et al. (2022) encontraron un mayor número de casos de dengue cuando la humedad relativa se estuvo entre el 45 % y el 70 %.

Figura 32. Comportamiento Humedad relativa rezagos 4 & 6 Vs Casos Dengue - Barrancabermeja



- ***Resultados Modelos Girón***

Tabla 13. Resultados modelos Girón

Modelo		Escenarios			
		1	**2**	**3**	**4**
LSTM	RMSE	4,47	4,49	4,50	4,02
	MAE	3,30	3,45	3,36	3,07
	R2(%)	55,25%	54,86%	54,54%	63,74%
CNN	RMSE	4,65	3,61	3,95	3,69
	MAE	3,46	2,83	2,90	2,81
	R2(%)	51,58%	70,68%	64,91%	69,45%
RNN	RMSE	4,08	4,00	3,95	4,35
	MAE	3,16	2,97	2,96	3,53
	R2(%)	62,67%	64,14%	65,04%	57,56%
GRU	RMSE	4,04	4,06	4,17	4,20
	MAE	2,98	3,07	3,22	3,20
	R2(%)	63,28%	63,01%	60,90%	60,48%
MLP	RMSE	4,39	3,86	4,45	4,30
	MAE	3,48	2,94	3,57	3,34
	R2(%)	56,70%	66,51%	55,56%	58,45%

Figura 33. Desempeño RMSE Modelos Girón

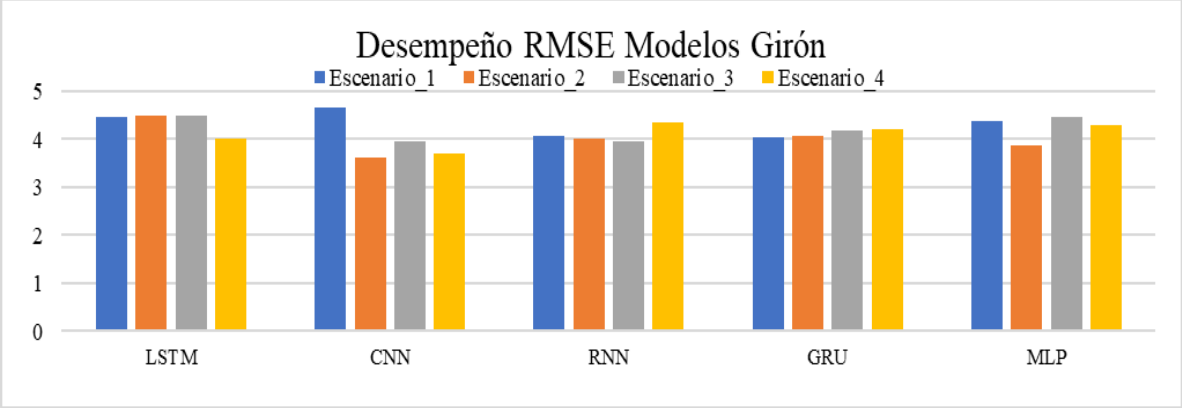
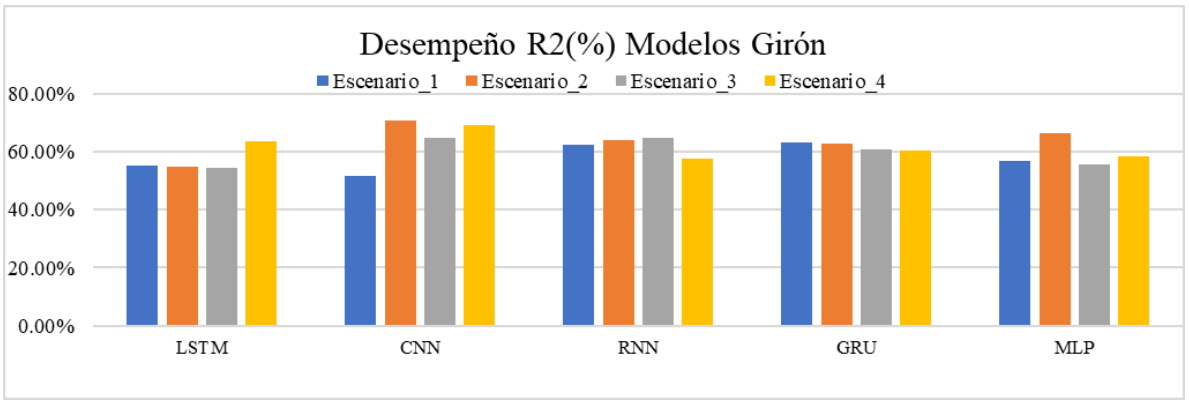


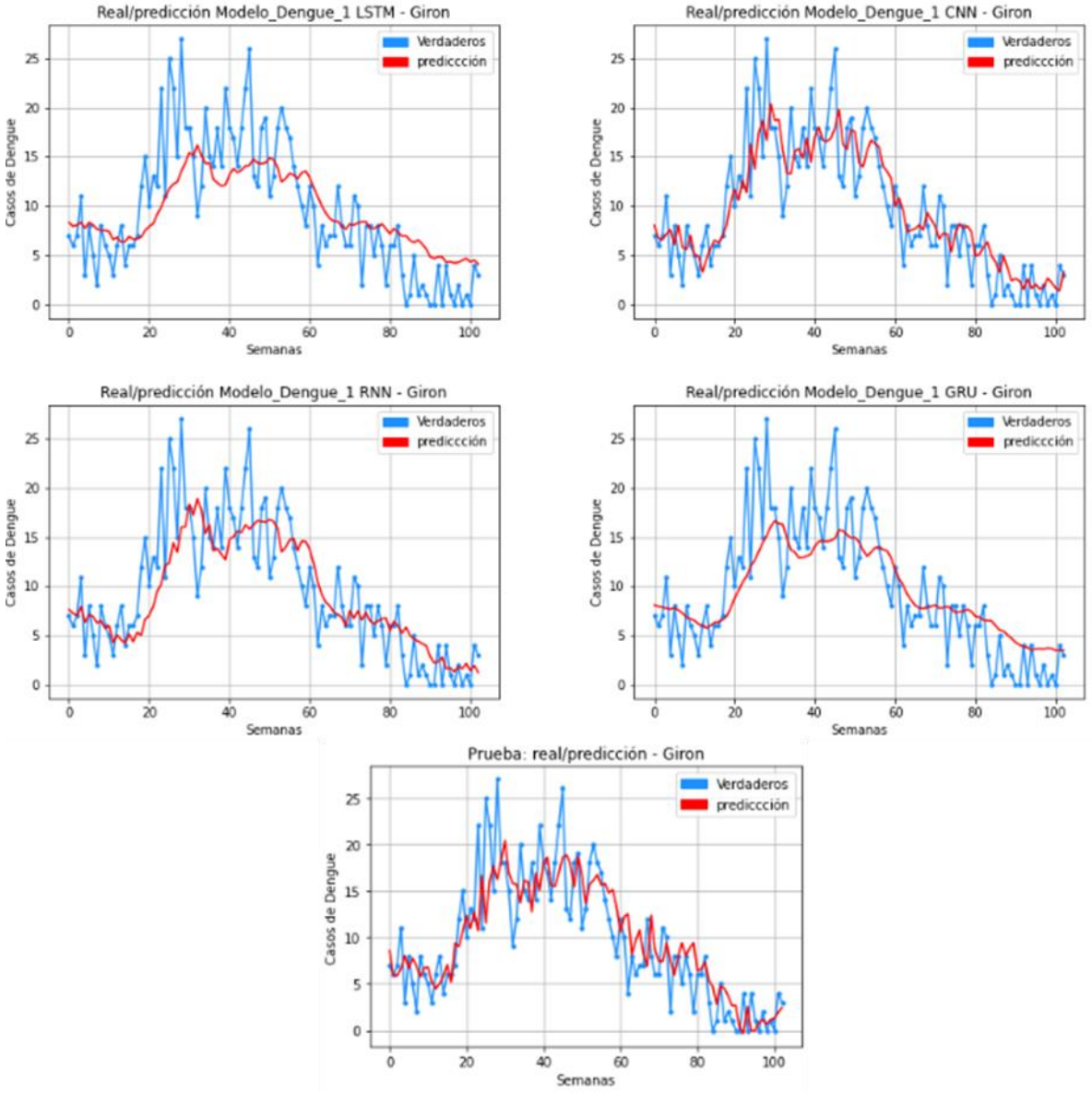
Figura 34. Desempeño R2(%) Modelos Girón



Para el municipio de Girón, el mejor comportamiento lo presentó el escenario 2 donde el promedio de error de los modelos fue de 4.0, mientras que el de los escenarios 1, 3 y 4 fue de 4.33, 4.20 y 4.11 respectivamente. Al igual que para Barrancabermeja, la humedad relativa es la variable que genera las mejores predicciones, sin embargo, para girón es la humedad relativa máxima, el escenario 2 lo compone la humedad relativa máxima con rezago 5 y con rezago 2.

Con relación a los modelos, el promedio de ajuste de los modelos estuvo por encima del 60%, los mejores desempeños los tuvieron CNN y RNN con promedios de error de 3.9 y 4.0 respectivamente. Sin embargo, específicamente para el escenario 2, en modelo MLP presentó un valor de error menor (3.86) y por consiguiente un mejor ajuste que el RNN. A continuación, se presentan las gráficas de predicción para el escenario 2.

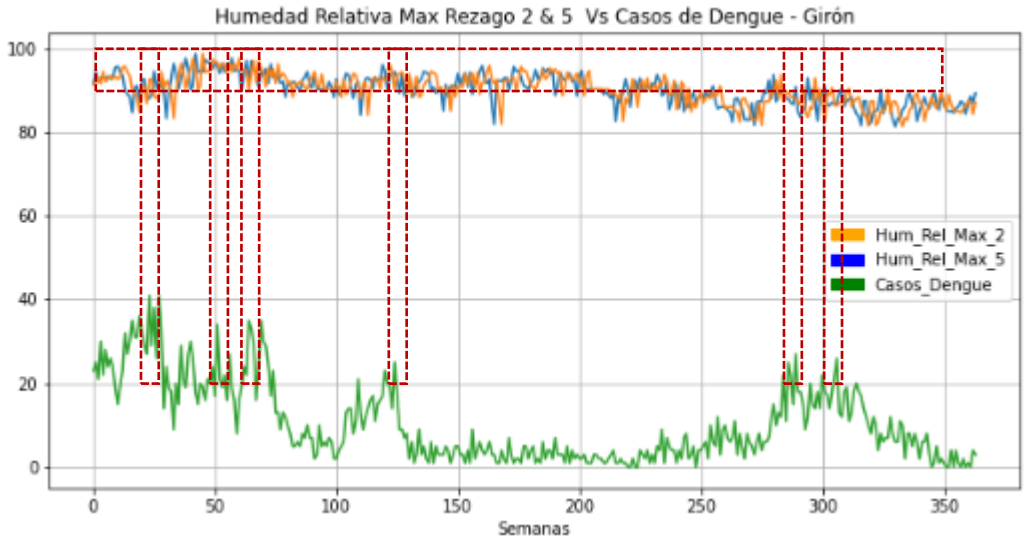
Figura 35. Graficas de Predicción Modelos Girón – Escenario 2



Al realizar la gráfica de la humedad relativa máxima con rezago 2 & 5 versus los casos de Dengue, es posible identificar que para las semanas donde mayor cantidad casos se presentaron la humedad relativa se mantuvo por encima del 90%, exceptuando los picos donde la cantidad de casos fue inferior a 30. Siendo esto contrario a los resultados reportados por el municipio de Barrancabermeja, donde un mayor número de casos de Dengue se asocia a humedades relativas por debajo del 60%. Sin embargo, se han reportado estudios donde

indican que cuando la humedad relativa es cercana al 84% la supervivencia diaria del mosquito varía entre el 91% y 95%, mientras que en valores más bajos de humedad (35%) la supervivencia baja a un 60% y 69% (Canyon et al., 1999; Citado en Vásques Brenes, 2020). Por su parte, Costa et al. (2010) mencionan que humedades relativas elevadas (80%) se asocian con mayores de oviposición, debido a que cuando los niveles de humedad son bajos, se reduce la producción de huevos debido a que se requiere mayores necesidades para el mantenimiento del cuerpo.

Figura 36. Comportamiento Humedad relativa rezagos 2 & 5 Vs Casos Dengue - Girón



- ***Resultados Modelos Lebrija***

Tabla 14. Resultados modelos Lebrija

Modelo		Escenarios			
		1	**2**	**3**	**4**
LSTM	RMSE	5,21	5,23	4,71	5,04
	MAE	3,90	3,94	3,55	3,79
	R2(%)	-42,74%	-43,71%	-16,58%	-33,41%
CNN	RMSE	4,68	4,16	3,92	3,86
	MAE	3,51	3,24	3,03	3,03
	R2(%)	-15,13%	8,98%	19,42%	21,56%
RNN	RMSE	5,29	4,51	5,00	5,34
	MAE	3,99	3,35	3,75	4,01
	R2(%)	-46,88%	-6,92%	-31,37%	-49,63%
GRU	RMSE	4,89	4,75	5,00	4,60
	MAE	3,65	3,36	3,77	3,45
	R2(%)	-25,35%	-18,28%	-31,50%	-11,19%
MLP	RMSE	4,92	5,33	4,95	4,69
	MAE	3,62	3,95	3,66	3,38
	R2(%)	-25,48%	-47,14%	-27,11%	-14.14%

Figura 37. Desempeño RMSE Modelos Lebrija

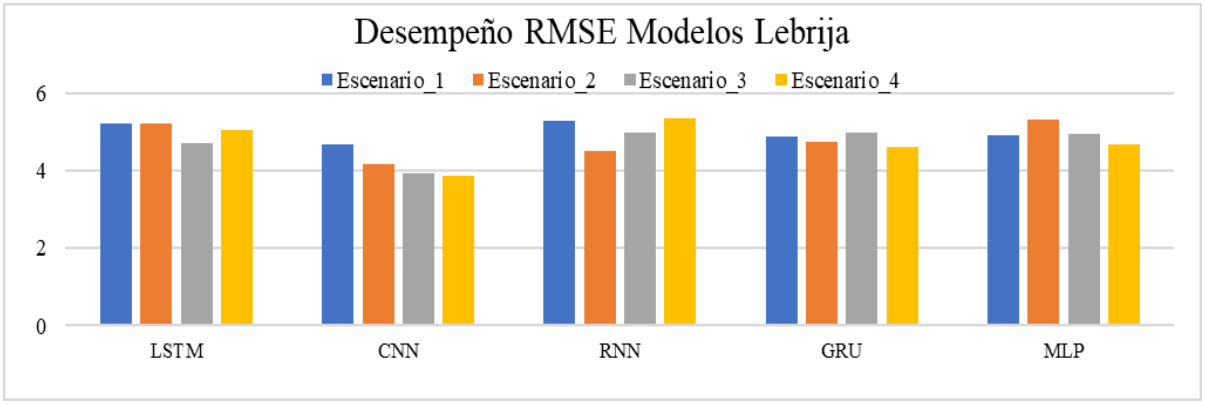
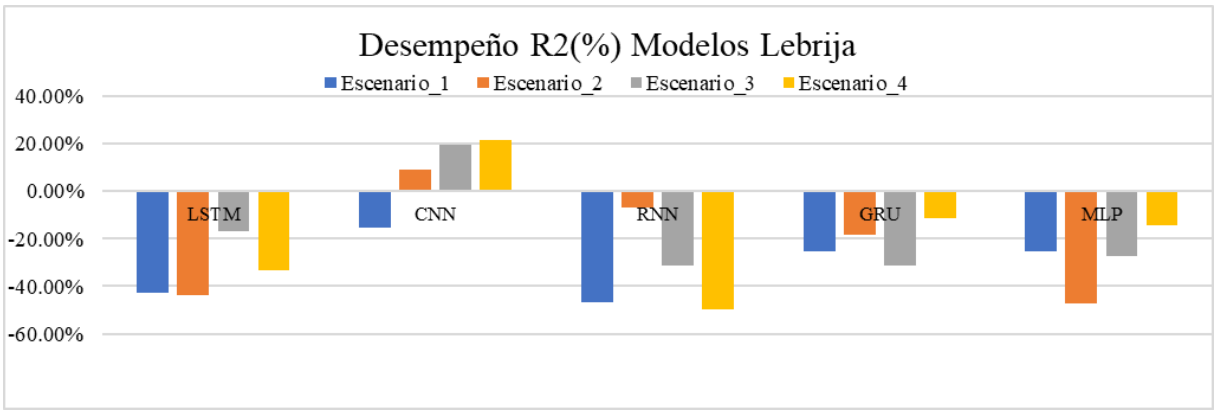


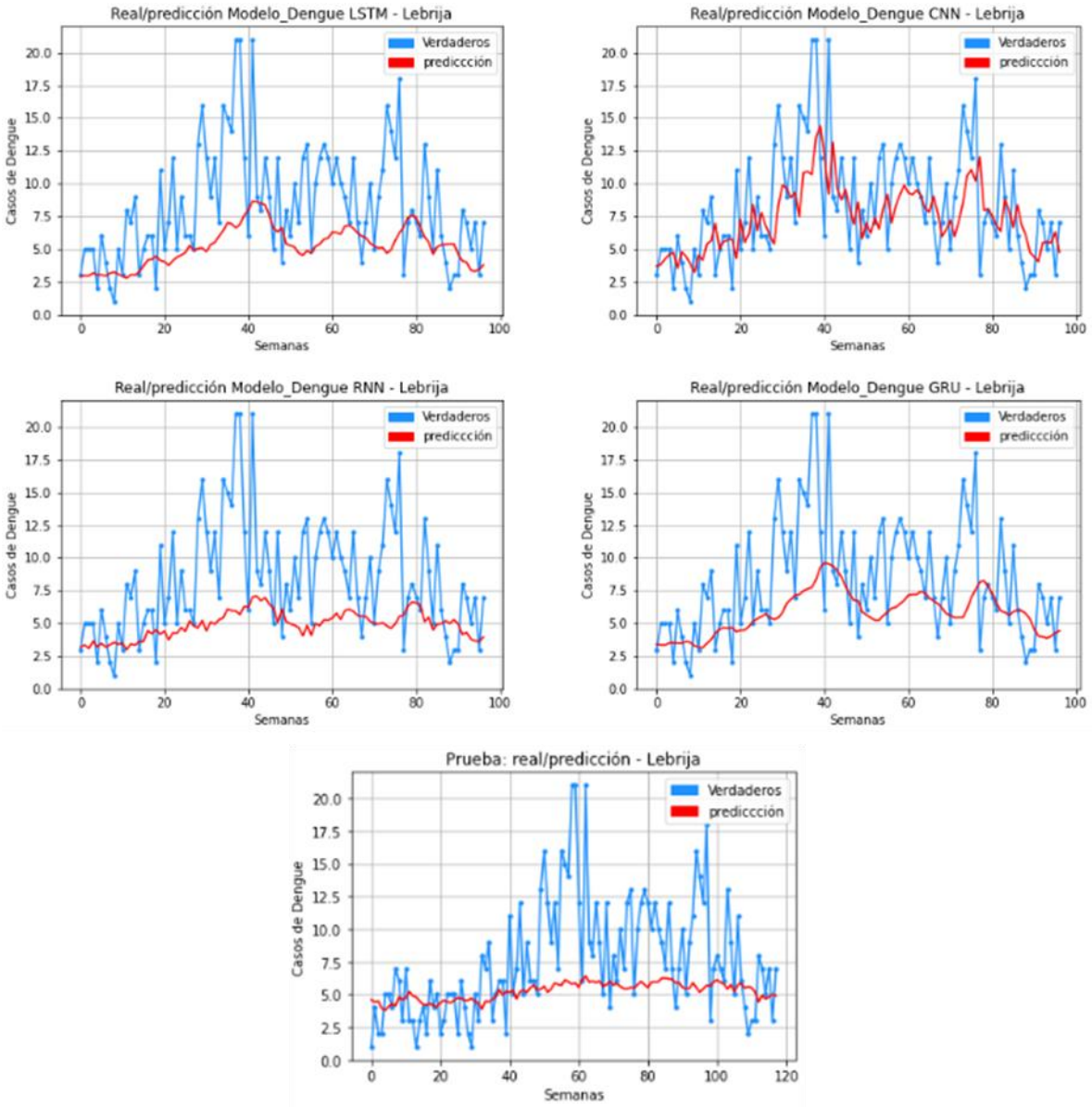
Figura 38. Desempeño R2(%) Modelos Lebrija



A diferencia de los municipios de Barrancabermeja y Girón, la cantidad de casos de Lebrija es mucho más bajo, así como el periodo evaluado. El total de casos semanal hasta el año 2013 es inferior a 8 casos, los años 2014 y 2015 el promedio es de 8 casos por semana y se presentan algunas semanas con valores superior a 20 casos. Así mismo, entre 2009 y 2012 se reportan solo 256 casos, presentándose varias semanas con reporte de cero casos. Al tratarse de valores tan bajos, dificulta mucho más la predicción. Esto se refleja en los valores de error, que se encuentran por encima del promedio semanal de los años con menos casos, y muy cercanos al promedio de los años con mayores casos. Así mismo, los valores del R2 reflejan que los modelos no contaron con la capacidad de ajustarse a los datos, exceptuando el modelo CNN que fue el único modelo que para los escenarios 2, 3 y 4 no presentó valores negativos. Presentando el mejor ajuste para el escenario 4, logrando explicar el 21,56% de variación de los casos.

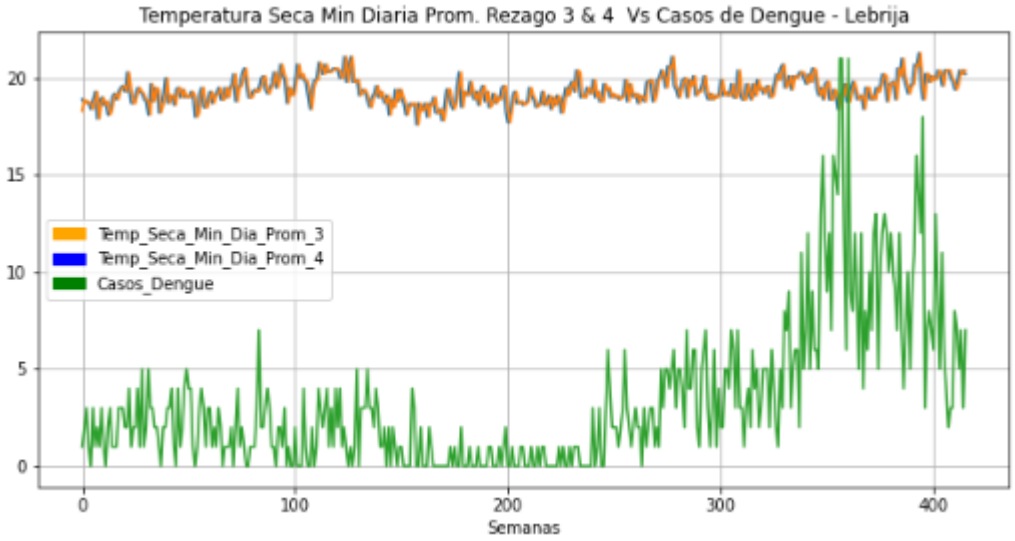
A continuación, se presenta las gráficas de las predicciones para el escenario 4.

Figura 39. Graficas de Predicción Modelos Lebrija – Escenario 4



El escenario 4 de Lebrija se compone de la Temperatura seca mínima diaria promedio con rezago 3 y esta misma variable con rezago 4. En Lebrija, la temperatura seca mínima varía entre los 17,3 °C y los 21,3°C. En la siguiente figura se presenta el comportamiento de la temperatura mínima y los casos de dengue.

Figura 40. Temperatura Seca Min Diaria Prom. Rezago 3 & 4 Vs Casos de Dengue - Lebrija



En este caso en particular, es posible visualizar más claramente como los casos de dengue se mantienen por debajo de los 8 casos por semana y presentan un aumento en las últimas semanas del periodo evaluado (2014-2015), por su parte la temperatura no presenta cambios abruptos, ni tampoco a lo largo de la ventana temporal y se mantiene relativamente constante, el promedio de temperatura en Lebrija es 19,3 °C lo que podría explicar los bajos niveles de casos en este municipio, dado que la temperatura se encuentra por debajo de los niveles óptimos para el mosquito. De acuerdo con Yang et al.(2009) el rango ideal de temperatura para la supervivencia de los mosquitos adultos, se encuentra entre los 15°C y 30°C y Mordecai et al. (2019) en su estudio con relación a la influencia general de la temperatura en la transmisión de enfermedades mencionan que la transmisión alcanzó su punto máximo a temperaturas intermedias entre 22.7 y 29.1 °C, la temperatura media óptima fue de 25.2 °C. De manera similar cuanto más baja es la temperatura, las hembras *Ae. aegipty* son capaces de sostener el vuelo, pero solo por cortos períodos de tiempo experimentando

una reducción en la movilidad y por lo tanto una mayor dificultad para alimentarse (Brady et al., 2014; citado en Vásques Brenes, 2020).

- ***Resultados Modelos Bucaramanga***

Tabla 15. Resultados modelos Bucaramanga

Modelo		**Rezago en semanas**					
		1	**2**	**3**	**4**	**5**	**6**
LSTM	RMSE	13,25	13,94	13,34	14,98	19,73	12,53
	MAE	10,96	11,63	10,77	13,20	18,38	9,97
	R2(%)	61,76%	57,68%	61,27%	51,13%	15,28%	65,82%
CNN	RMSE	8,35	9,91	9,15	9,13	9,80	8,00
	MAE	5,71	6,96	7,24	6,60	7,57	5,37
	R2(%)	84,81%	78,62%	81,75%	81,85%	79,07%	86,04%
RNN	RMSE	14,48	12,81	12,94	14,15	10,66	15,01
	MAE	12,65	10,56	10,85	11,51	7,73	11,88
	R2(%)	54,34%	63,89%	63,56%	56,38%	75,25%	50,97%
GRU	RMSE	11,51	11,81	11,83	11,55	12,45	11,55
	MAE	8,87	8,25	9,35	9,03	10,15	8,45
	R2(%)	71,12%	69,61%	69,53%	70,95%	66,25%	70,93%
MLP	RMSE	8,14	7,99	8,57	8,11	7,94	8,12
	MAE	5,33	5,67	5,68	5,50	5,71	5,71
	R2(%)	85,55%	86,07%	83,99%	85,68%	86,24%	85,62%

Figura 41. Desempeño RMSE Modelos Bucaramanga

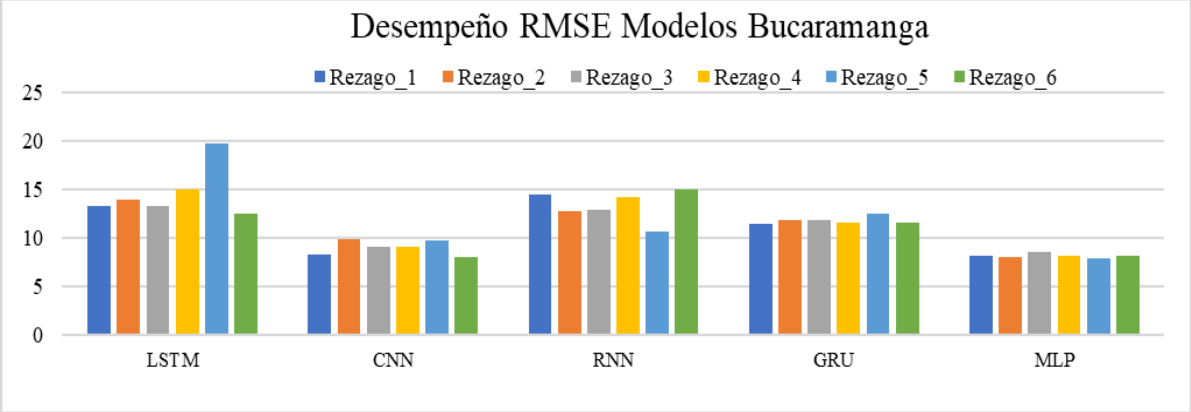
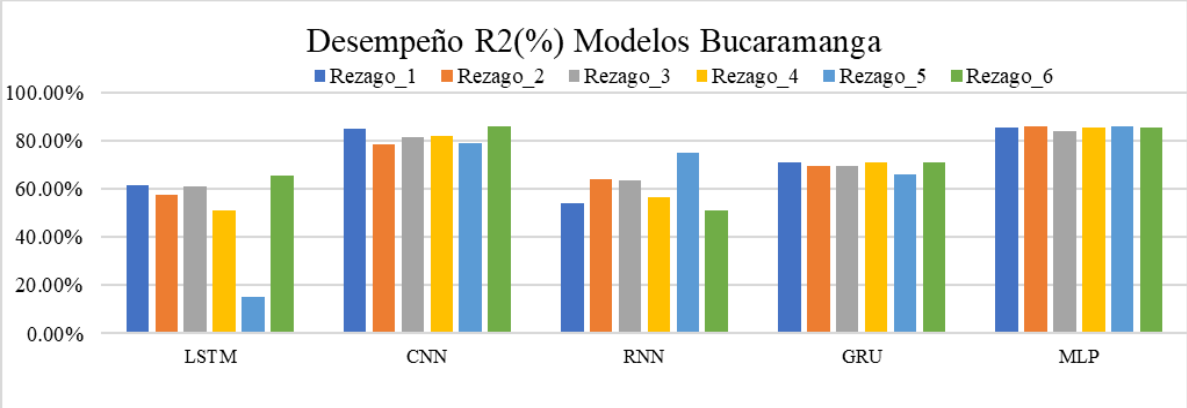


Figura 42. Desempeño R2(%) Modelos Bucaramanga

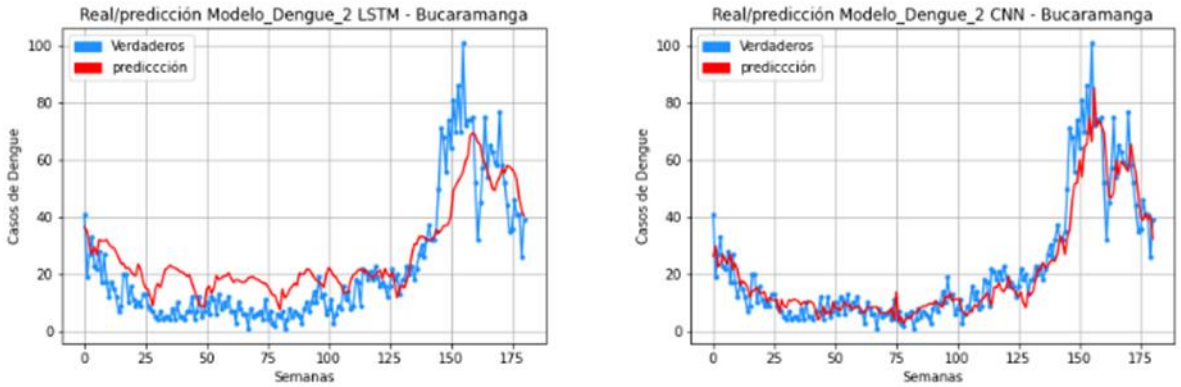


De manera general, los errores cuadrados medios para el municipio de Bucaramanga estuvieron por debajo de 15, exceptuando el modelo LSTM trabajado con el Rezago 5 que obtuvo un valor de 18.38. Para este caso, los mejores desempeños los presentaron los modelos MLP y CNN, cuyos valores de error estuvieron por debajo de 10. De igual forma esto se ve reflejado en el ajuste de los modelos, ambos presentaron un ajuste por encima del 78%, siendo superiores los ajustes de MLP, exceptuando el rezago 6, donde el ajuste del MLP fue de 85.62% y el de CNN de 86.04%. Con relación al comportamiento de los modelos

con base en los rezagos, no se evidencian diferencias importantes entre uno y otro, sim embargo, es notorio como el rezago 6, presentó los valores de error más bajos y los mejores ajustes, en todos los modelos, exceptuando el modelo RNN. Al no encontrar diferencias significativas entre los valores de los rezagos, podría pensarse en que la influencia de las precipitaciones sobre la cantidad de casos de Dengue en el municipio de Bucaramanga, no es tan relevante.

A continuación, se presentan las gráficas de predicción para el rezago 6. Las graficas evidencian como los modelos MLP y CNN, son los que mejor ajuste presentan, sin embargo, es notorio como el modelo CNN presenta mejor capacidad para capturar los picos, así como la disminución durante el tiempo de prueba evaluado. Por su parte, se evidencia una sub-predicción del modelo MLP en su mayoría los valores predichos se encuentran por debajo de los valores reales.

Figura 43. Graficas de Predicción Modelos Bucaramanga – Rezago 6



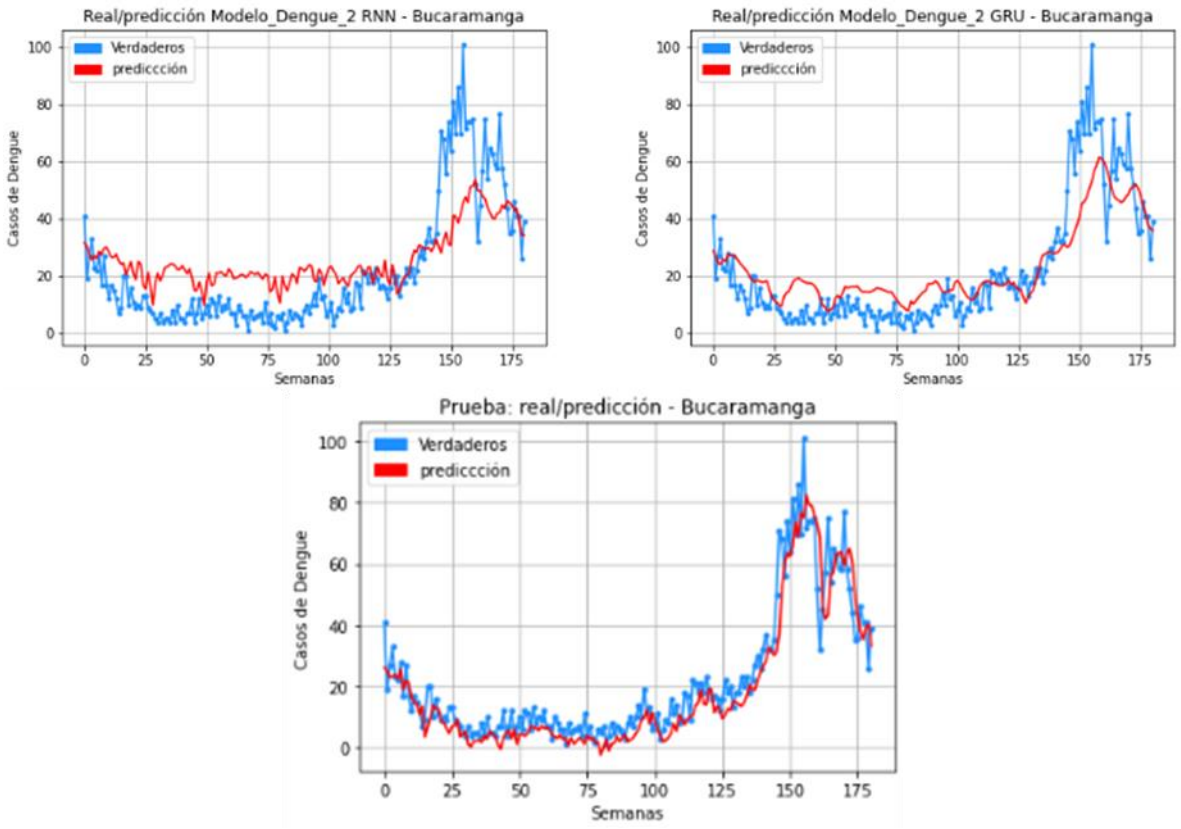
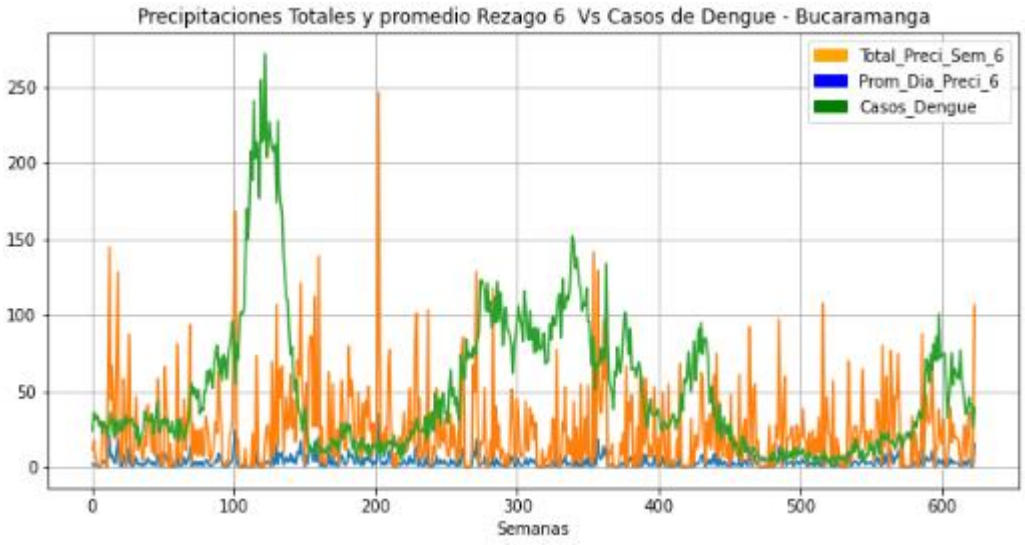


Figura 44. Comportamiento precipitaciones Vs Casos Dengue Rezago 6 - Bucaramanga



- ***Resultados Modelos Floridablanca***

Tabla 16. Resultados modelos Floridablanca

Modelo		Rezago en semanas					
		1	**2**	**3**	**4**	**5**	**6**
LSTM	RMSE	7,60	8,52	9,23	7,57	7,06	6,92
	MAE	5,64	6,17	7,91	6,35	5,53	5,19
	R2(%)	41,73%	26,91%	14,22%	42,26%	49,71%	51,74%
CNN	RMSE	5,49	5,49	5,25	5,17	5,54	5,97
	MAE	4,12	4,40	4,06	4,15	4,17	4,45
	R2(%)	69,64%	69,60%	72,15%	73,03%	69,09%	64,02%
RNN	RMSE	7,50	7,30	6,54	7,56	6,84	7,56
	MAE	5,88	5,74	5,13	6,29	5,31	6,15
	R2(%)	43,37%	46,34%	56,90%	42,39%	52,81%	42,36%
GRU	RMSE	6,95	6,77	7,46	7,04	7,91	6,65
	MAE	5,71	5,24	6,07	5,64	6,59	4,87
	R2(%)	51,30%	53,74%	43,97%	50,02%	36,99%	55,35%
MLP	RMSE	5,64	6,30	5,97	5,44	5,27	5,74
	MAE	4,19	4,79	4,47	4,06	3,99	4,48
	R2(%)	67,92%	59,92%	64,11%	70,18%	71,97%	66,80%

Figura 45. Desempeño RMSE Modelos Floridablanca

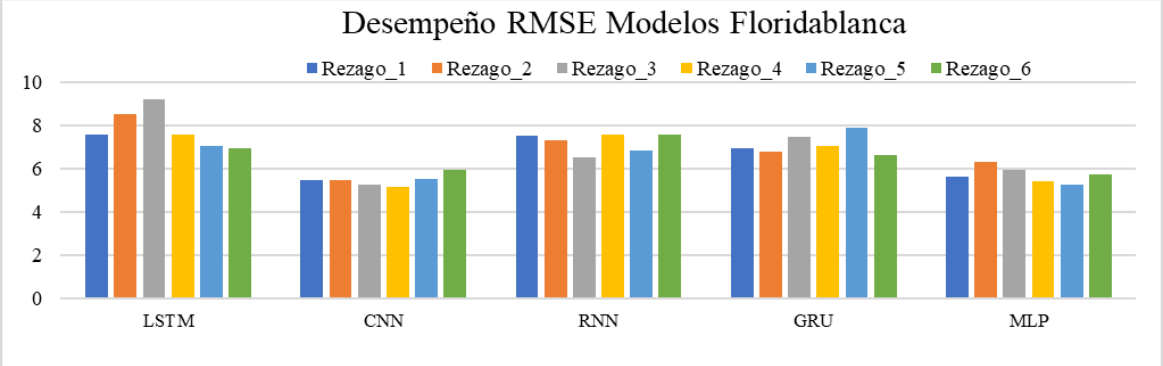
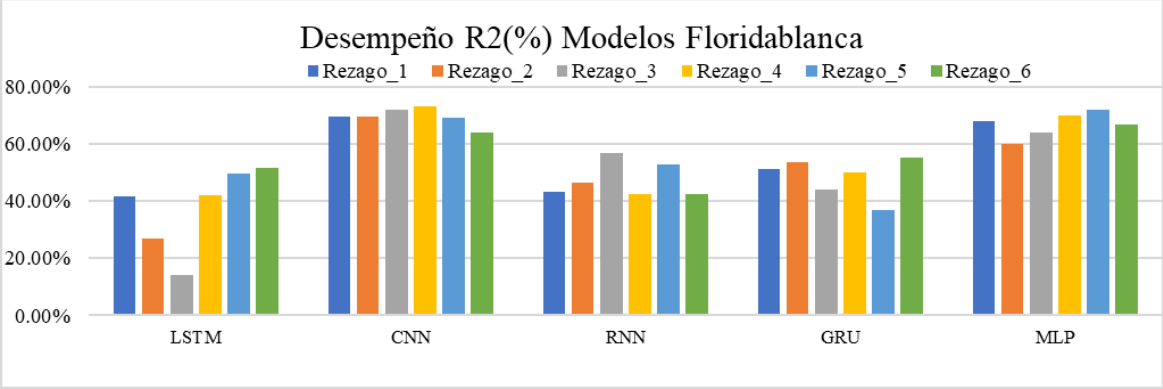


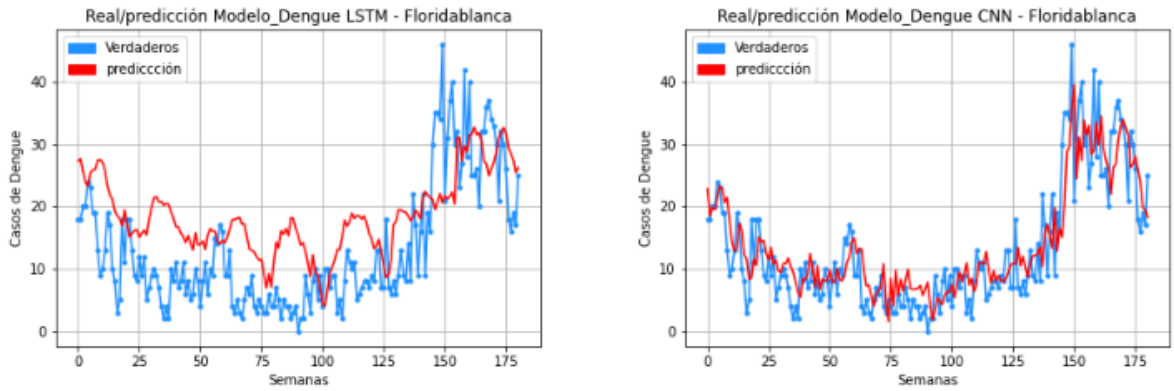
Figura 46. Desempeño R2(%) Modelos Floridablanca



En Floridablanca, al igual que en Bucaramanga y análisis anteriores, el modelo MLP, seguido de CNN presentaron los mejores desempeños, mientras que para los modelos LSTM, RNN y GRU los errores estuvieron por encima de 7, MLP y CNN tuvieron errores de 5.4 y 5.7 respectivamente. Así mismo, este buen desempeño se evidencia en los ajustes, pues son los dos únicos modelos que superan el 60% en el valor de R2. De manera similar a Bucaramanga, no se evidencian diferencias importantes entre uno y otro, sin embargo, en este caso el rezago 5, presentó los valores de error más bajos y los mejores ajustes.

A continuación, se presentan las gráficas de predicción para el rezago 5.

Figura 47. Graficas de Predicción Modelos Floridablanca – Rezago 5



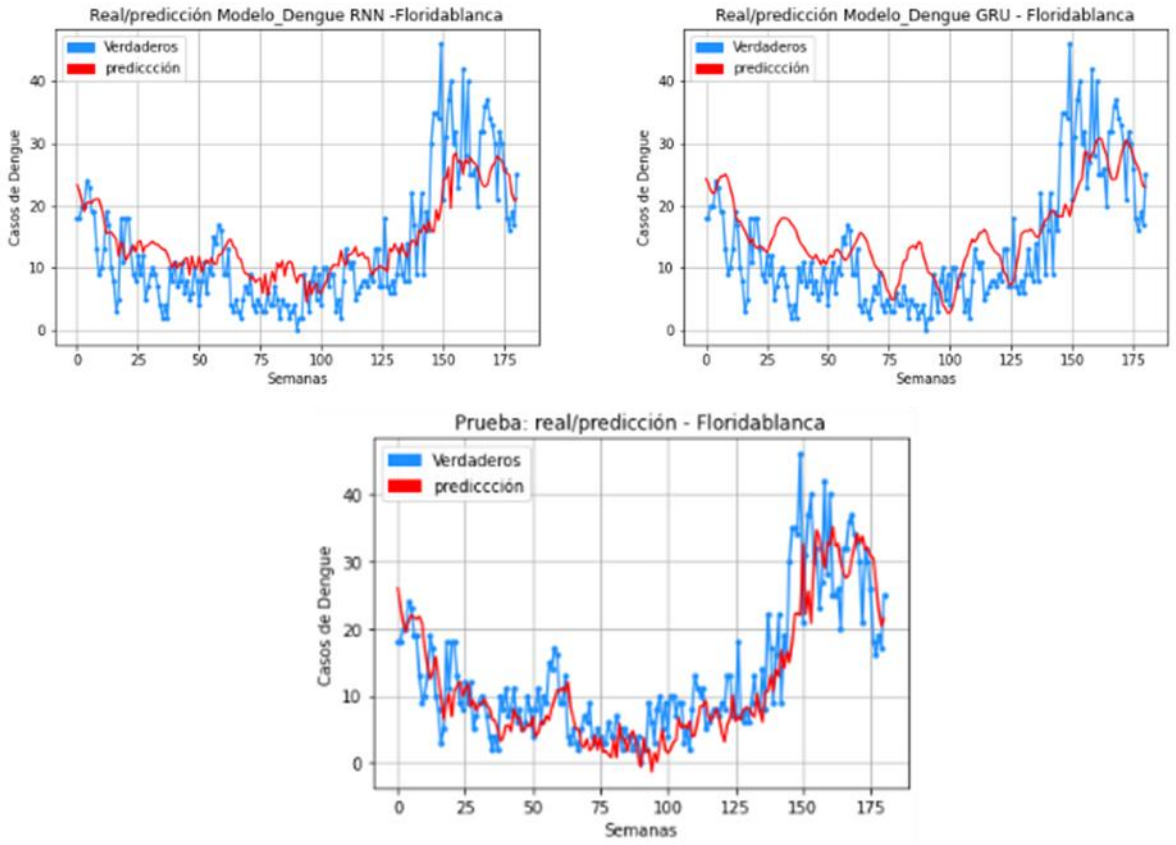
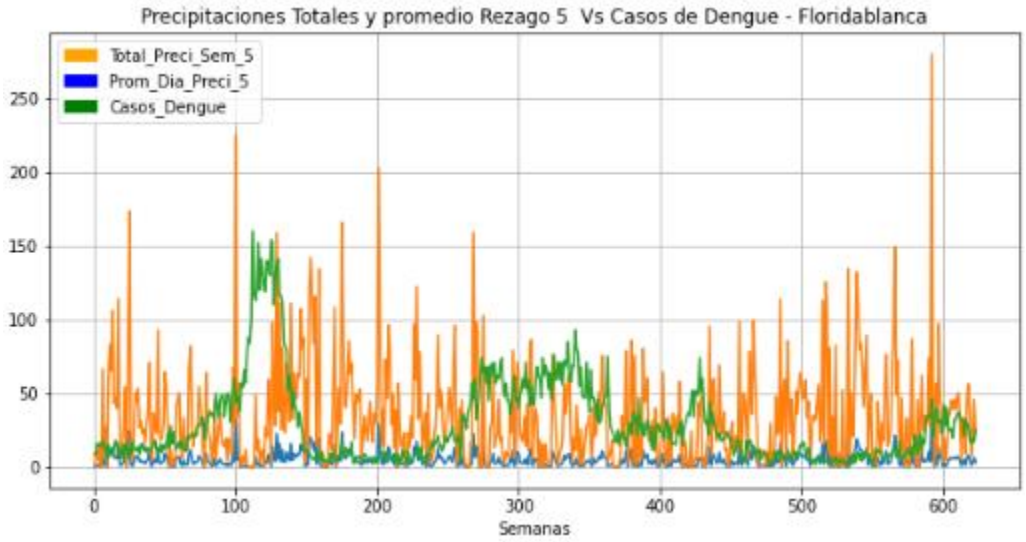


Figura 48. Comportamiento precipitaciones Rezago 5 Vs Casos Dengue - Floridablanca



- ***Resultados Modelos Piedecuesta***

Tabla 17. Resultados modelos Piedecuesta

Modelo		**Rezago en semanas**					
		1	2	3	4	5	6
LSTM	RMSE	5,27	4,67	5,89	5,92	5,36	5,32
	MAE	3,80	3,19	4,15	4,15	3,57	3,73
	R2(%)	59,96%	68,54%	49,99%	49,62%	58,59%	59,17%
CNN	RMSE	4,09	4,26	4,04	4,31	3,99	4,05
	MAE	2,91	2,66	2,63	2,84	2,62	2,61
	R2(%)	75,84%	73,83%	76,43%	73,30%	77,05%	76,32%
RNN	RMSE	4,81	5,00	4,61	5,26	5,10	4,78
	MAE	3,27	3,80	3,23	3,29	3,50	3,26
	R2(%)	66,69%	63,95%	69,41%	60,17%	62,53%	67,02%
GRU	RMSE	4,75	5,11	4,76	4,83	4,83	4,95
	MAE	3,34	3,69	3,22	3,55	3,39	3,53
	R2(%)	67,48%	62,43%	67,30%	66,38%	66,37%	64,71%
MLP	RMSE	4,18	4,00	3,93	5,45	4,03	4,03
	MAE	2,78	2,52	2,47	4,14	2,58	2,56
	R2(%)	74,81%	76,95%	77,78%	57,26%	76,59%	76,63%

Figura 49. Desempeño RMSE Modelos Piedecuesta

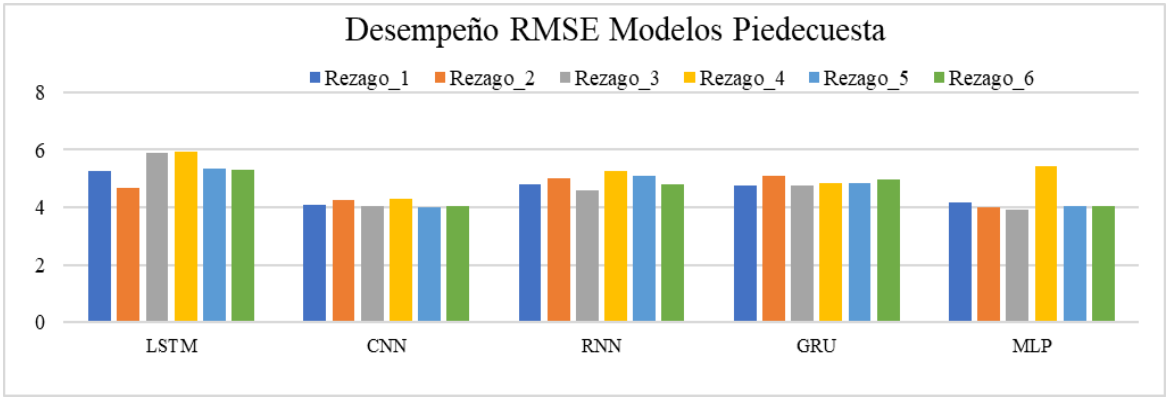
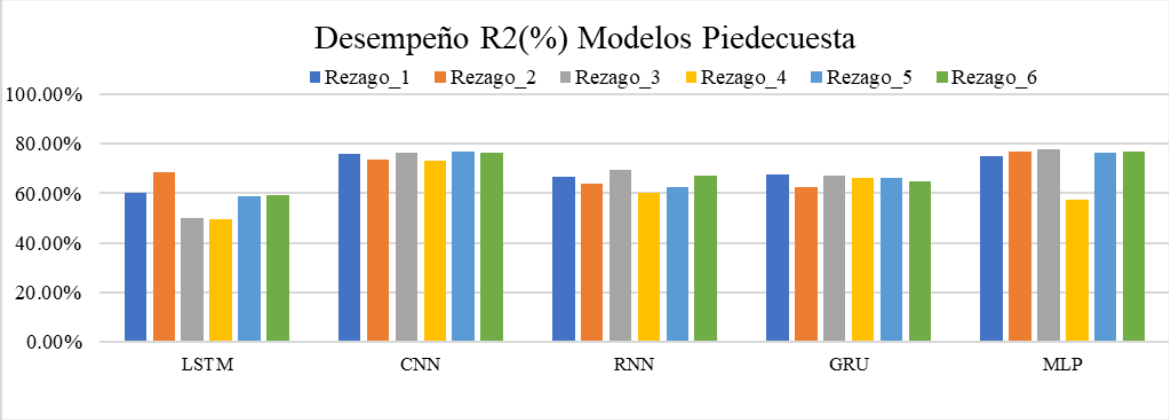


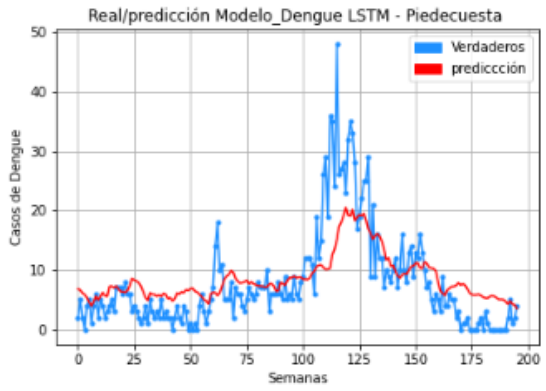
Figura 50. Desempeño R2(%) Modelos Piedecuesta

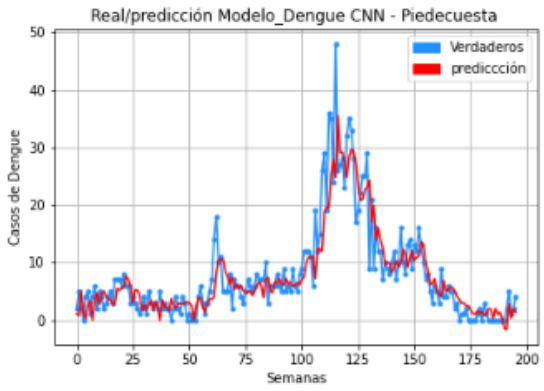


Nuevamente en Piedecuesta, los modelos CNN y MLP presentan los mejores ajustes, sin embargo, en este caso en modelo CNN presentó el mejor desempeño, a diferencia de los demás modelos, sus ajustes estuvieron por encima del 70% y sus RMSE no superaron el valor de 4. Respecto a los rezagos, en promedio el rezago 3 fue quien obtuvo los valores de error mas bajos con todos los modelos, exceptuando con la red LSTM.

A continuación, se presentan las gráficas de predicción para el rezago 3. Se videncia como el modelo CNN presenta una mejor adaptación cuando hay descenso en los casos, sin embargo, al igual que el MLP, no se ajusta tan bien en los picos.

Figura 51. Graficas de Predicción Modelos Piedecuesta – Rezago 3





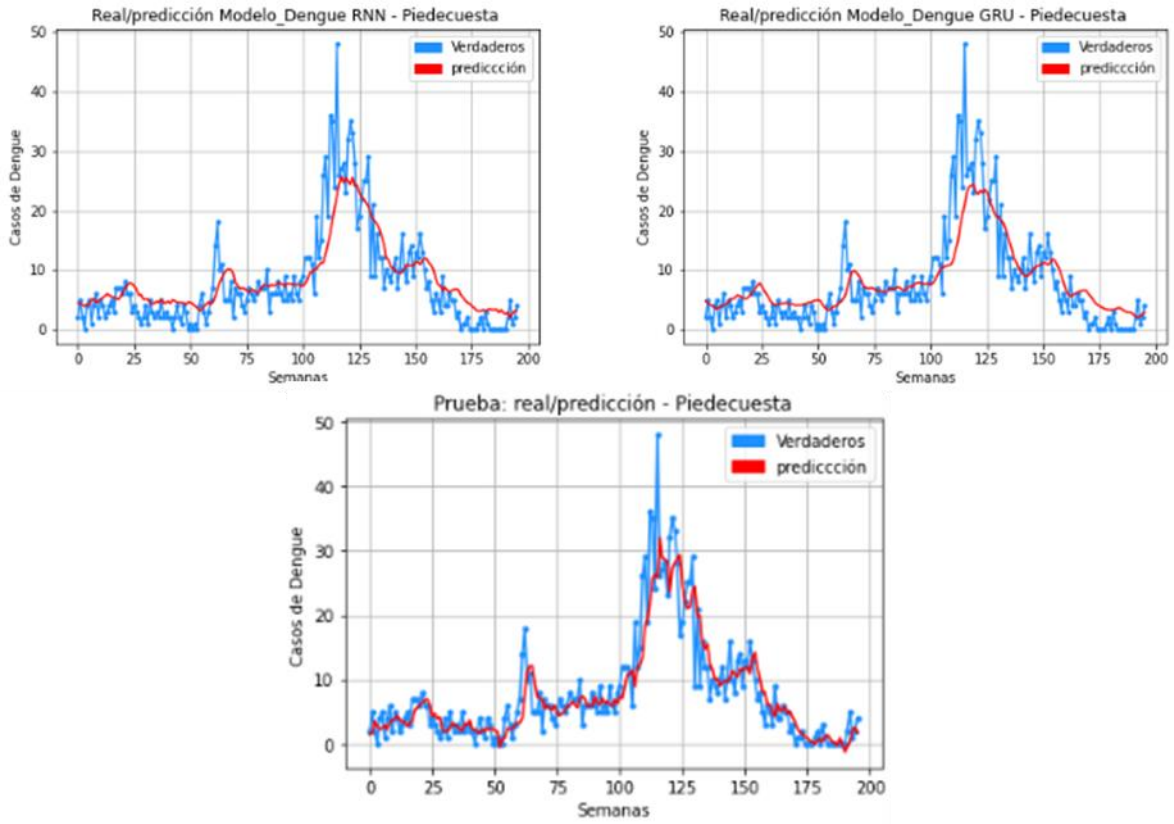
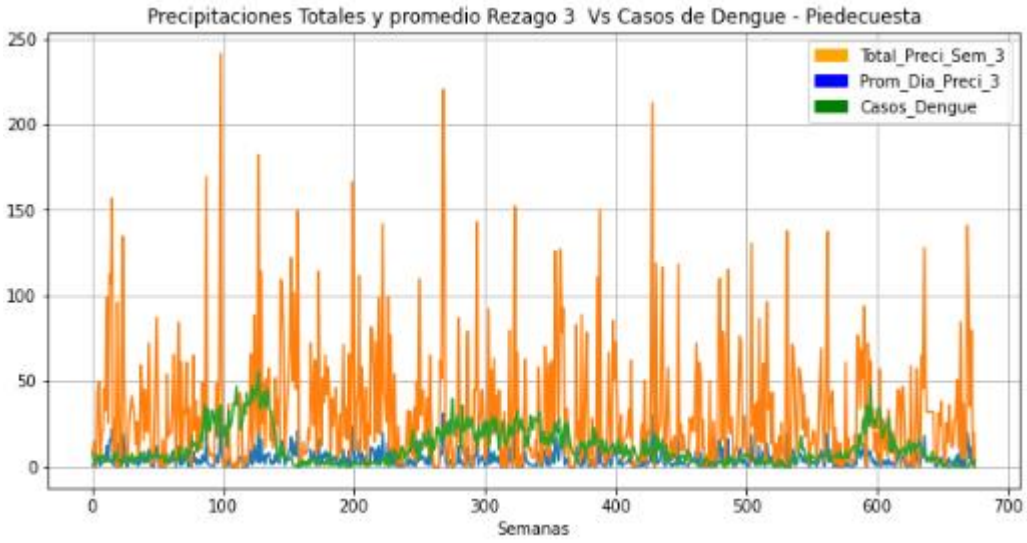


Figura 52. Comportamiento precipitaciones Rezago 3 Vs Casos Dengue - Piedecuesta



A modo de análisis general, considerando las gráficas de comportamiento de las precipitaciones versus los casos de Dengue, para los municipios de Bucaramanga, Floridablanca y Piedecuesta, en ninguno de los tres casos se evidencian fluctuaciones o cambios que den indicios sobre como las precipitaciones influyen sobre la cantidad de casos de Dengue, así mismo, de acuerdo con el análisis de correlaciones (Ver anexo E) las precipitaciones no se asociaron estadísticamente con el Dengue. Sin embargo, lo que esto puede significar es que, a pesar de que se requiere agua para la reproducción de mosquitos, pues de acuerdo con Morin et al.(2013) y Wijayanti et al. (2016) las precipitaciones proporcionan hábitats para la fase acuática del ciclo de vida y tienen una influencia importante en la distribución del vector. Se tiene que, en estos tres municipios no se presentan grandes variaciones en el comportamiento de las lluvias y durante gran parte del año casi que permanece constante el nivel de lluvias, por lo cual las variaciones de una semana a otra no afectan significativamente a la incidencia del Dengue.

A pesar de lo anterior, los resultados de los modelos con la variable precipitaciones no son desalentadores, pues en promedio para cada uno de los municipios los valores del R2 se encuentran por encima del 60%.

Por otra parte, con el objetivo de complementar el análisis, se corrieron modelos sólo con la variable precipitaciones para los municipios de Barrancabermeja, Girón y Lebrija, con el fin de conocer el comportamiento y capacidad de predicción de los modelos, a pesar de que la variable precipitaciones no fue relevante, de acuerdo con los resultados de la importancia de las características para estos municipios. A continuación, se presentan los resultados.

Tabla 18. Resultados modelos sólo Precipitaciones Barrancabermeja

Modelo		Rezago en semanas					
		1	2	3	4	5	6
LSTM	RMSE	3,96	4,14	3,67	3,98	4,03	3,86
	MAE	2,96	3,10	2,78	3,15	3,01	3,03
	R2(%)	63,96%	60,54%	69,08%	63,65%	62,57%	65,74%
CNN	RMSE	3,04	3,40	3,17	2,97	3,04	2,99
	MAE	2,25	2,38	2,36	2,14	2,30	2,28
	R2(%)	78,75%	73,37%	76,89%	79,70%	78,72%	79,39%
RNN	RMSE	3.81	4,10	3,71	3,64	3,67	3,96
	MAE	2.90	3,19	2,81	2,72	2,74	3,23
	R2(%)	66,68%	61,31%	68,34%	69,48%	68,98%	63,85%
GRU	RMSE	3.79	3,73	3,82	3,66	3,82	3,57
	MAE	2.05	2,99	2,76	2,74	3,10	2,74
	R2(%)	66,94%	68,04%	66,44%	69,24%	66,46%	70,73%

Figura 53. Desempeño Precipitaciones RMSE Modelos Barrancabermeja

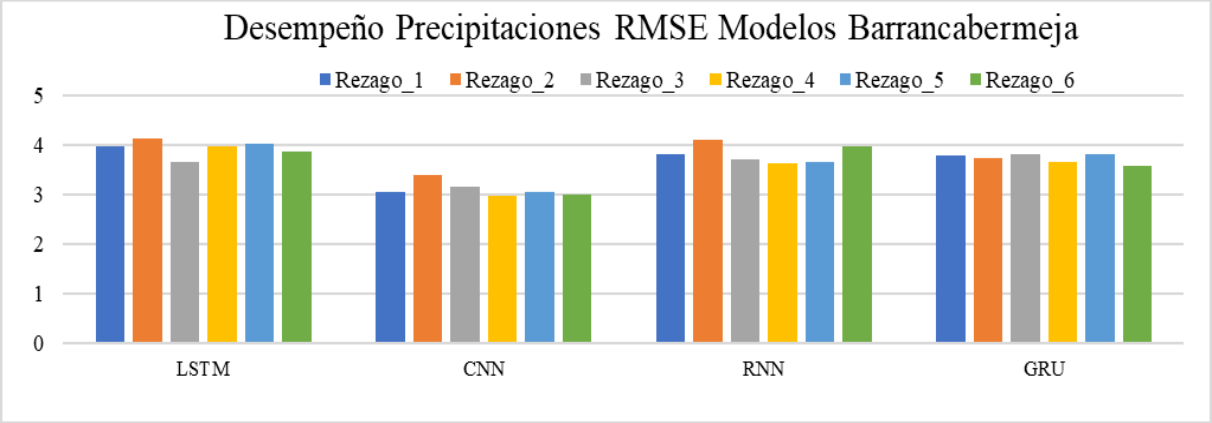


Figura 54. Desempeño Precipitaciones R2(%) Modelos Barrancabermeja

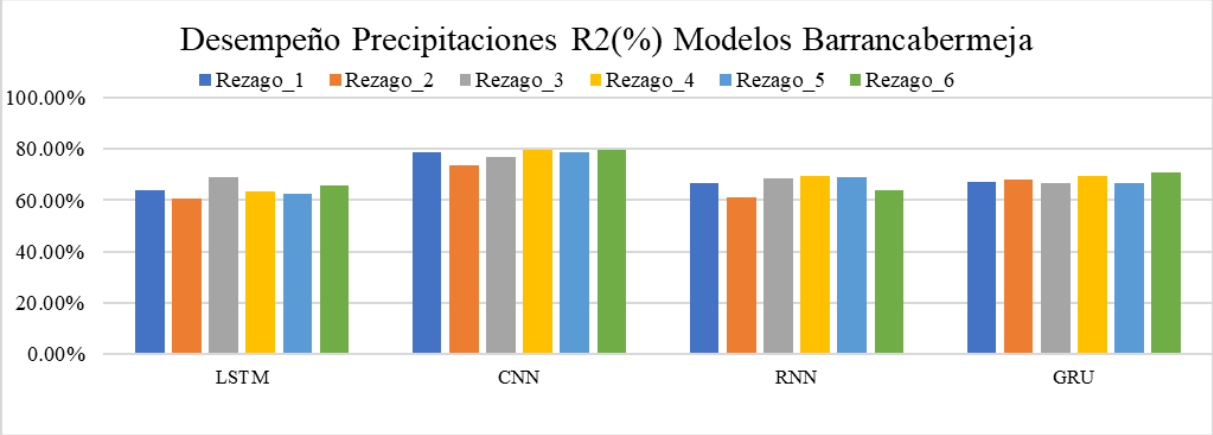


Tabla 19. Resultados modelos sólo Precipitaciones Lebrija

Modelo		**Rezago en semanas**					
		1	2	3	4	5	6
LSTM	RMSE	5,09	5,01	5,16	5,04	5,16	5,00
	MAE	3,82	3,76	3,92	3,81	3,90	3,77
	R2(%)	-36,08%	-31,74%	-39,82%	-33,23%	-39,97%	-31,55%
CNN	RMSE	3,76	3,76	3,87	3,79	3,76	3,74
	MAE	3,07	3,06	3,17	3,09	3,07	3,03
	R2(%)	25,82%	25,88%	21,24%	24,66%	25,52%	26,36%
RNN	RMSE	4,81	4,61	4,54	4,46	5,15	4,87
	MAE	3,55	3,44	3,42	3,36	3,91	3,66
	R2(%)	-21,36%	-11,73%	-8,36%	-4,36%	-39,32%	-24,60%
GRU	RMSE	5,11	4,73	5,00	4,39	4,13	4,81
	MAE	3,86	3,53	3,78	3,35	3,56	3,61
	R2(%)	-37,28%	-17,51%	-31,29%	-1,30%	-17,32%	-21,48%

Figura 55. Desempeño Precipitaciones RMSE Modelos Lebrija

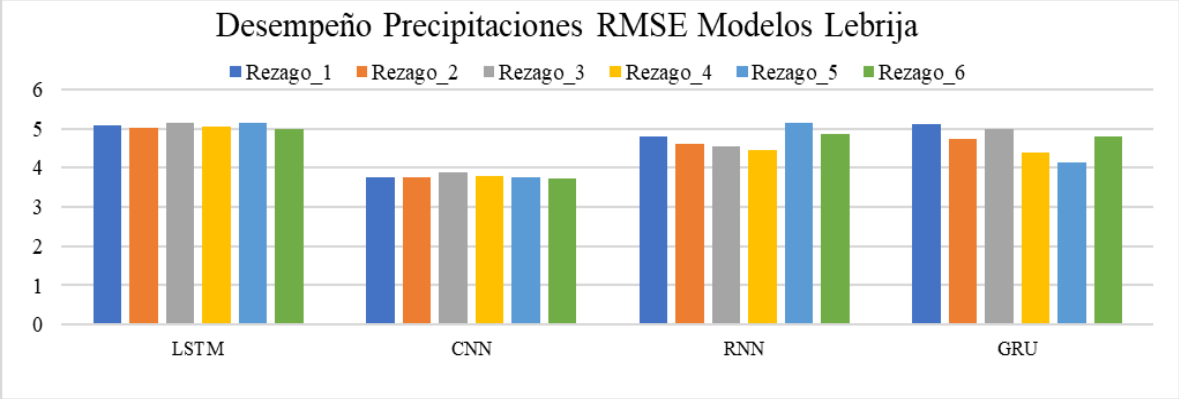


Figura 56. Desempeño Precipitaciones R2(%) Modelos Lebrija

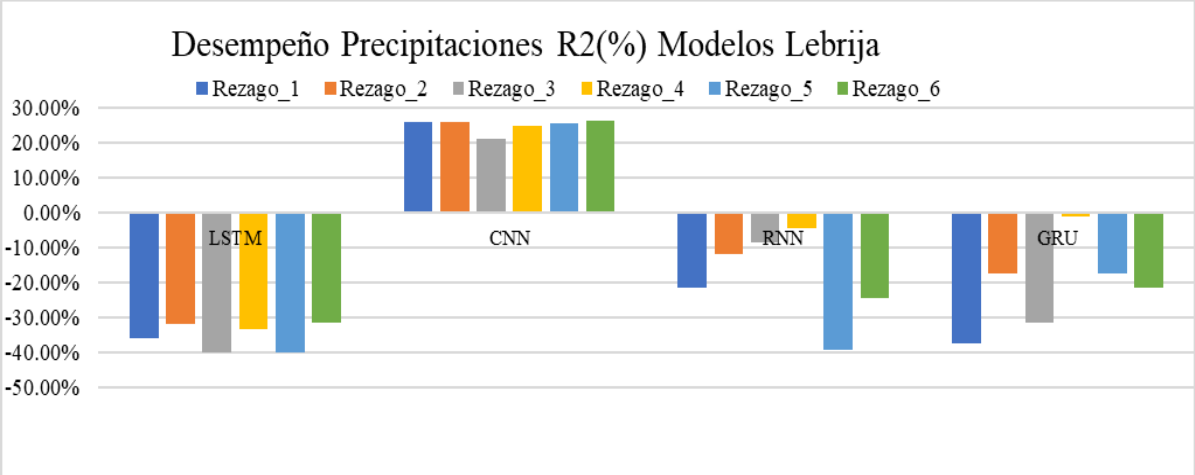


Tabla 20. Resultados modelos sólo Precipitaciones Girón

Modelo		**Rezago en semanas**					
		1	**2**	**3**	**4**	**5**	**6**
LSTM	RMSE	4,22	4,23	4,13	4,03	4,11	4,33
	MAE	3,16	3,16	3,07	3,07	3,14	3,38
	R2(%)	60,01%	59,80%	61,67%	63,49%	62,02%	57,84%
CNN	RMSE	3,86	3,61	3,78	3,73	3,77	3,77
	MAE	3,08	2,84	2,84	2,85	2,96	2,96
	R2(%)	66,64%	70,69%	67,95%	68,79%	68,04%	68,17%
RNN	RMSE	4,03	4,13	4,17	3,91	4,45	4,16

Modelo		Rezago en semanas					
		1	2	3	4	5	6
	MAE	3,03	3,18	3,14	2,94	3,26	3,22
	R2(%)	63,61%	61,74%	60,91%	65,62%	55,66%	61,08%
GRU	RMSE	4,10	4,05	4,01	4,07	4,13	4,27
	MAE	3,19	3,12	3,08	3,22	3,17	3,49
	R2(%)	62,22%	63,15%	63,87%	62,81%	61,72%	59,10%

Figura 57. Desempeño Precipitaciones RMSE Modelos Girón

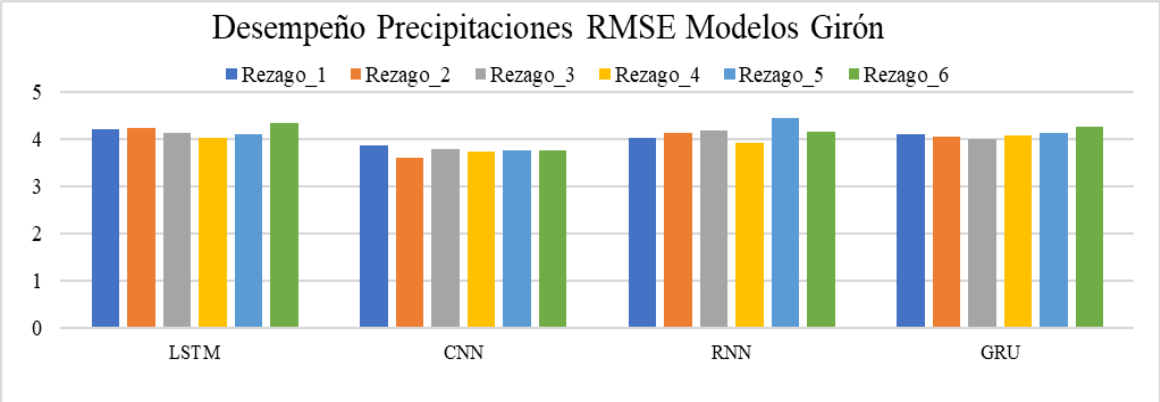
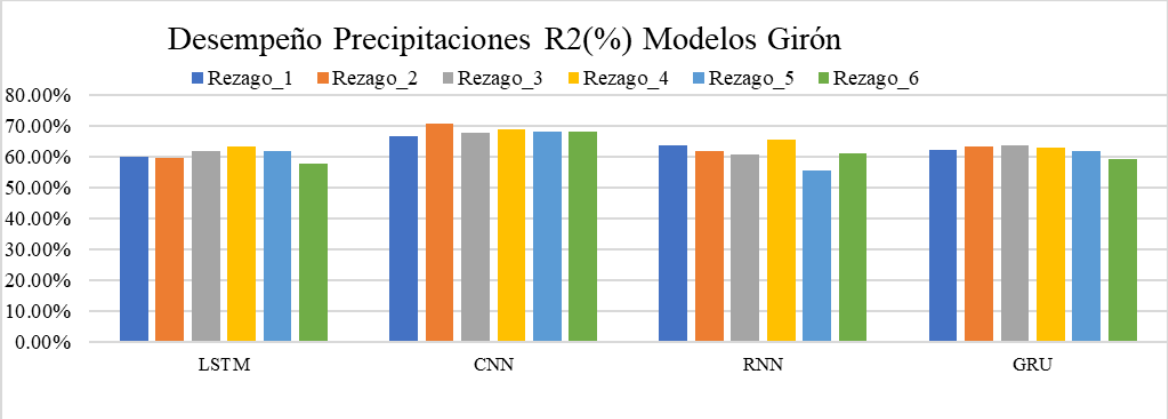


Figura 58. Desempeño Precipitaciones R2(%) Modelos Girón



Los resultados para los municipios de Barrancabermeja, Lebrija y Girón haciendo uso sólo de la variable precipitaciones, sorprendentemente dieron mejores resultados que los modelos generados con las variables de mayor importancia, en Barrancabermeja, por ejemplo, el promedio de error del modelo CNN con las variables de mayor importancia (escenarios) fue de 3.5, mientras que con el uso de sólo precipitaciones el promedio de error fue de 3.1 y los valores de R2 estuvieron por encima del 73%. En Lebrija, el modelo CNN que fue el que mejor se ajustó, con la variable precipitaciones, no presentó ningún valor negativo, el promedio de ajuste fue del 24.9%, sin embargo, en los modelos de los escenarios el mayor valor R2 logrado fue del 21.56%. y lo mismo sucedió con el municipio de Girón.

Un factor común que se evidencio con estos modelos de solo precipitaciones para estos municipios es que en los tres casos el valor del RMSE más bajo se presentó con el rezago 4, el RMSE para este rezago en Girón fue de 3.94, el Lebrija de 4.42 y en Barrancabermeja de 3.56.

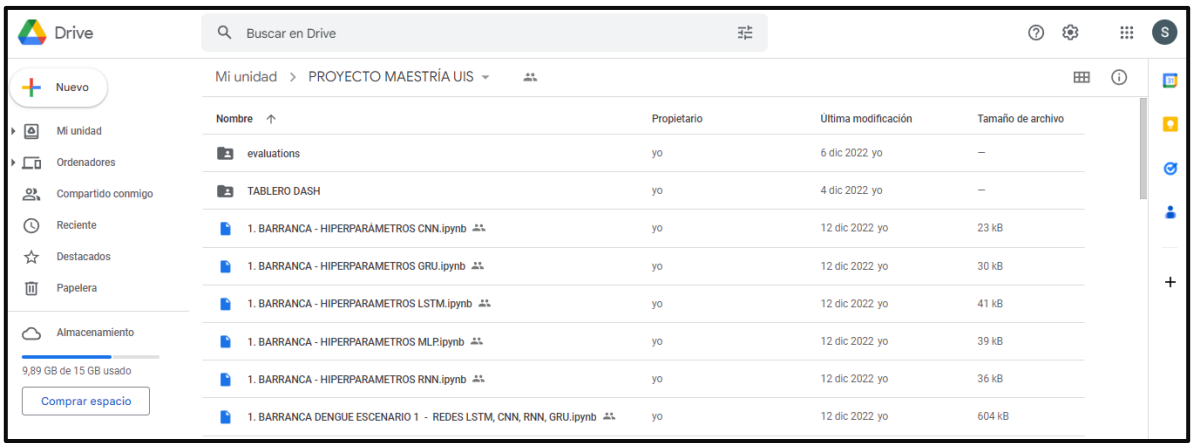
De manera general, se aprecia que, para las precipitaciones y demás variables, todos los municipios y modelos presentan los mejores desempeños para rezagos superiores a 3 semanas, esto tiene sentido pues como lo explican Padilla et al. (2012) el huevo completa su desarrollo embrionario en 48 horas, si el ambiente es húmedo y cálido, pero ese tiempo puede prolongarse hasta cinco días con temperaturas más bajas. Se adhiere a las paredes de los recipientes, como albercas, tanques bajos, floreros, llantas y otros, que contienen agua limpia. La fase de huevo dura uno a dos días; le sigue la fase de larva, cuyo desarrollo consta de cuatro estadios en el agua, donde se alimentan de material orgánico sumergido o acumulado en las paredes de los recipientes. Las larvas crecen y completan su desarrollo entre cinco y siete días, dependiendo de la temperatura ambiente donde se encuentren. En condiciones de temperatura de 24 a 28 °C, el estadio de pupa dura entre uno y dos días. Este es el último

estadio dentro del agua, en el cual no se alimentan y se producen modificaciones anatómicas y fisiológicas, hasta que emergen los adultos. Todo este proceso tiene una duración aproximada de 15 días (2 semanas).

6.4 Herramienta de Visualización

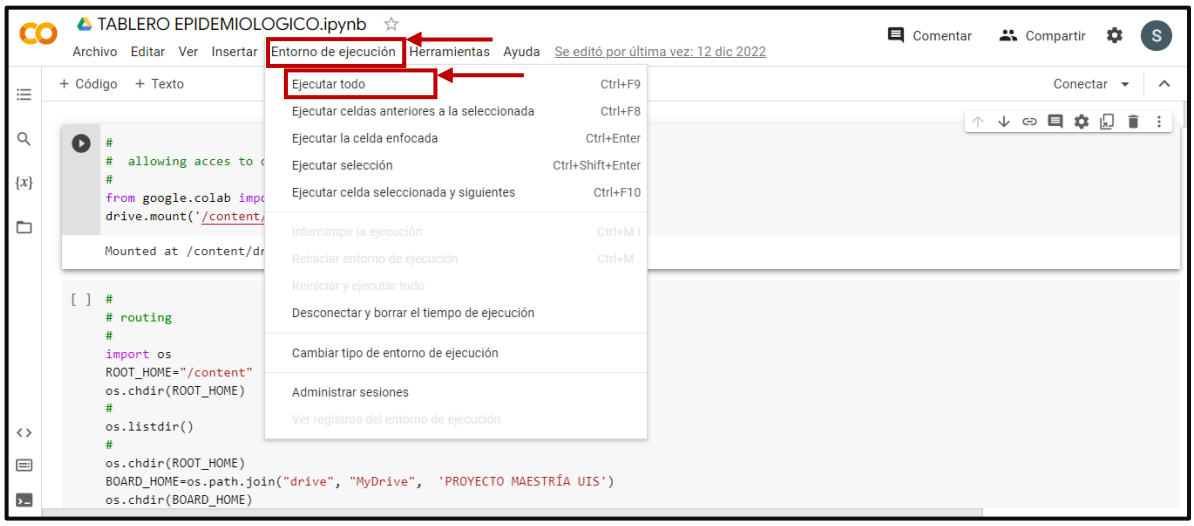
En esta última fase, se diseñó un prototipo de arquitectura visual como medio de presentación de los resultados y el análisis de la información. La herramienta consta de dos partes, una que contiene el análisis descriptivo, en donde es posible la visualización del comportamiento de los casos por municipio, por sexo, por área de ocurrencia y grupo etario. La segunda parte corresponde al análisis predictivo, donde es posible la selección del municipio para el cual se requiere hacer la predicción, las variables a utilizar, la proporción de datos de entrenamiento y prueba y el modelo de predicción. Todo el código fuente de la herramienta de visualización y los modelos puede ser encontrado en el siguiente enlace Drive: https://drive.google.com/drive/folders/1ou5T1AD-33s5FGQ008a7y3xVOwTwA7u-?usp=sharing

Para la visualización del tablero, es necesario correr el código, el cual al final indicará el link de acceso al mismo. En primer lugar, una vez abierto el enlace de la carpeta drive, será posible ver los códigos trabajados y una carpeta llamada “TABLERO DASH”. Como se muestra a continuación:



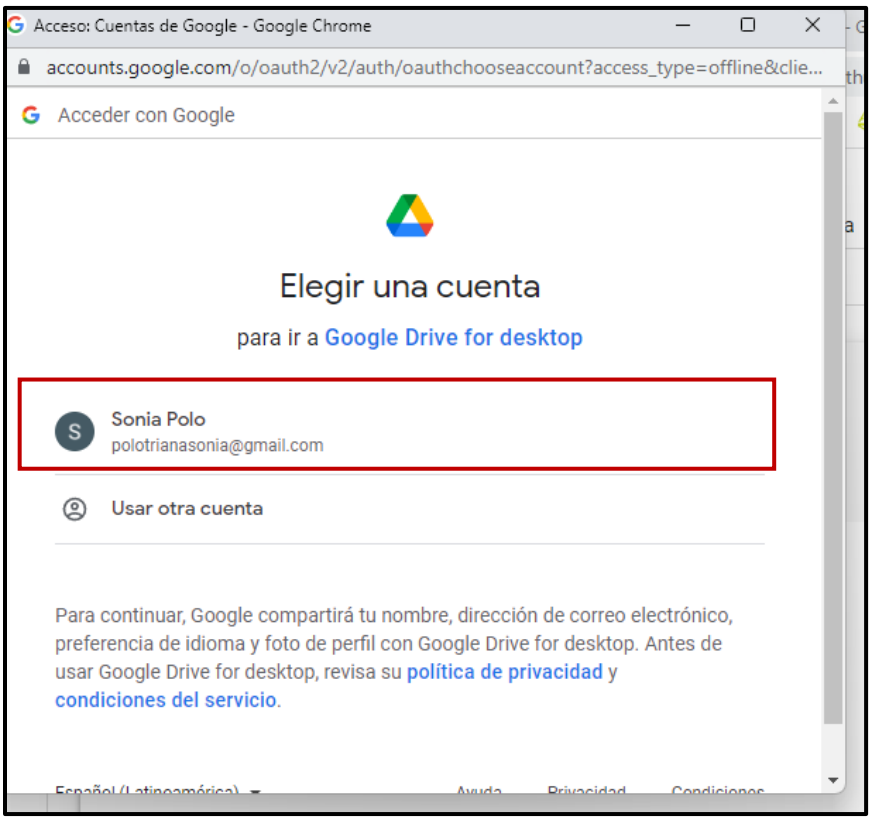
Una vez allí, se deberá ingresar a la carpeta “TABLERO DASH”, en donde se encontrará un cuaderno de Google Colab, que contiene el código llamado “TABLERO EPDEMIOLÓGICO", el cual deberá abrir.

Una vez abierto el código, deberá dar clic en la pestaña llamada “Entorno de ejecución” y posteriormente en la opción ejecutar todo. Como se muestra a continuación:

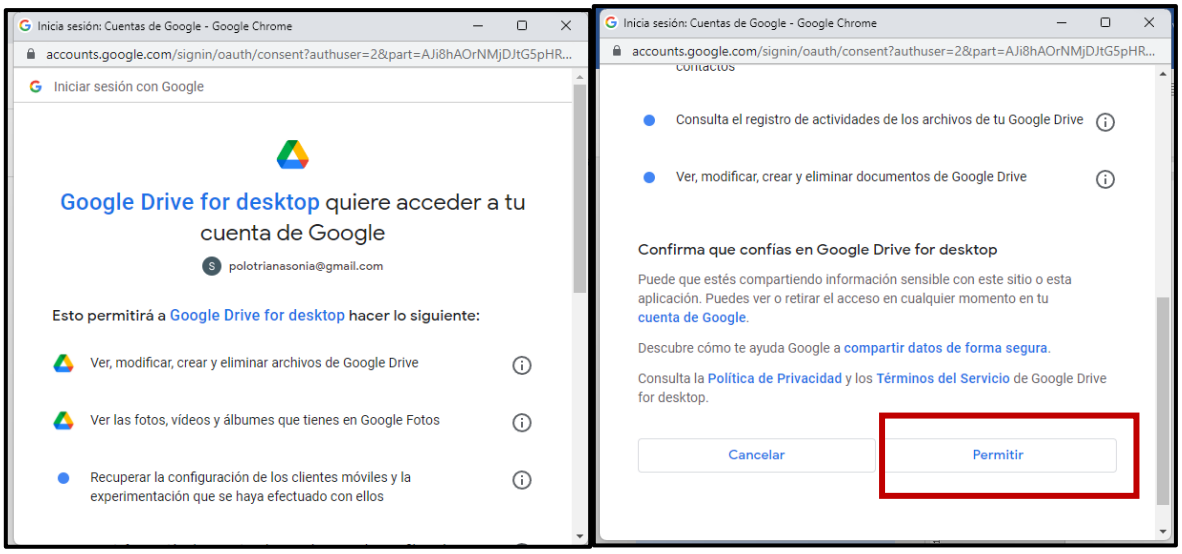


Una vez hecho esto, el código iniciará a correr, como primera instancia le pedirá que le permita el acceso a los archivos del Google drive, esto con el fin de acceder a los códigos de los modelos, a lo cual deberá escoger la opción “Conectar con Google Drive”.

Luego, le pedirá que elija la cuenta de acceso y dará clic sobre la cuenta que allí le aparezca.

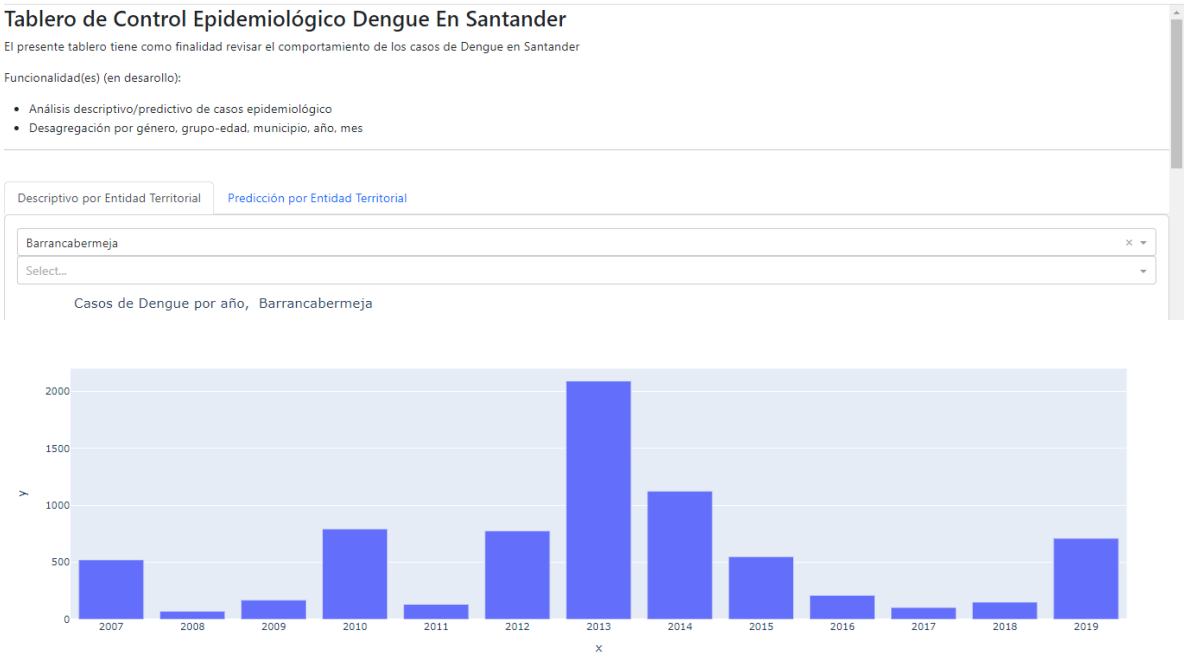


Una vez hecho esto le aparecerá otra ventana solicitando acceso, en la cual deberá deslizar la barra derecha hasta el final y dar clic en “Permitir”.



Posterior a esto, el código ejecutará y deberá deslizar la barra derecha para dirigirse al final del código en donde encontrará el link de acceso:

Una vez en el tablero, se encontrará con una ventana como la siguiente:



La herramienta se compone de dos pestañas, la primera llamada “Descriptivo por entidad territorial” donde será posible seleccionar el municipio y visualizar el comportamiento de los casos de dengue, hacia la parte inferior de la página se encuentran las gráficas de casos por mes, por sexo, por área y grupo etario.

La segunda pestaña llamada “Predicción por entidad territorial” se creó con el fin de que el usuario pueda interactuar con los modelos y generar predicciones, por lo tanto, tiene las opciones de seleccionar, el municipio, las variables con las cuales quiere que se ejecute el modelo, el porcentaje de datos a usar para entrenamiento y finalmente el modelo con el cual quiere realizar la predicción. Una vez seleccionadas las opciones deseadas, será posible visualizar la gráfica de predicción en la parte inferior y comparar los resultados del error RMSE.

1. Seleccionar municipio.

2. Seleccionar las variables predictoras deseadas para realizar la predicción.

3. Indicar el porcentaje de datos con los cuales desea realizar el entrenamiento, valores entre 0.5 y 0.9 (50% y 90% del dataset) y finalmente seleccionar el modelo con el cual desea realizar la predicción.

7. Conclusiones

En el presente estudio, se utilizaron diferentes modelos de Deep Learning y un modelo de red neuronal tradicional con el fin de evaluar la capacidad de predicción de los modelos y las asociaciones entre el Dengue y la variabilidad climática. Los resultados demostraron que la incorporación de periodos de rezago permite un mejor desempeño en la predicción de los patrones temporales. Con relación a los modelos, el algoritmo de red neuronal tradicional (MLP) presentó un mejor desempeño que los algoritmos de Deep Learning en los municipios de Bucaramanga, Barrancabermeja, Floridablanca y Girón, sin embargo, para los municipios de Lebrija y Piedecuesta, el mejor desempeño lo presentó el modelo CNN, cabe aclara que en estos municipios la proporción de casos respecto a los demás es mucho menor, demostrando así que los modelos Deep Learning tienen mayor capacidad para adaptarse a bases de datos más complejas, en este caso, esto hace referencia a la baja cantidad de casos y mayor cantidad de semanas con reportes de cero casos. Esto es debido a sus capas convolucionales, las cuales constituyen un elemento esencial para este tipo de redes neuronales, específicamente las convoluciones 1D, que se caracterizan por aplicarse en una única dimensión, en general de naturaleza temporal como lo fue este estudio. Así mismo, diferentes estudios indican que este tipo de convolución presenta un rendimiento superior para el uso de datos limitados y altas variaciones adquiridas de distintos ámbitos.

En este tipo de trabajo, las etapas previas que abarcan las actividades de selección, preprocesamiento y transformación de los datos, son de vital importancia, pues de ello depende que el conjunto de datos sea consistente y confiable.

Para cada municipio de estudio se encontraron algunas variables más relevantes que otras en la predicción de casos de dengue, esto lleva a que se deba realizar una depuración

personalizada de los datasets en cada municipio. Esto a su vez demuestra que de la variabilidad climática en la dinámica de transmisión de enfermedades como el Dengue siguen siendo discutidos por la comunidad científica. La relevancia de este trabajo radica en que se ratifica y complementa la comprensión de la dinámica de dengue en la región y puede ayudar en el desarrollo de sistemas de alerta temprana basados en factores climáticos.

Por otra parte, a pesar de la poca asociación estadística entre las precipitaciones y el Dengue, los modelos trabajados con esta variable presentaron desempeños superiores que los demás, esto demuestra la capacidad de los modelos para capturar características intrínsecas de los datos y generalizar brindando así predicciones confiables.

Con relación al zika y al Chikungunya, no se han vuelto a presentar brotes desde el 2016, por lo cual no fue posible el análisis a nivel municipal y a nivel departamental no se obtuvieron resultados confiables. Sin embargo, esta disminución tan abrupta sobre todo en el caso del Chikungunya, se debe a que después de contagiarse la persona desarrolla los anticuerpos que se encargaran de proteger a lo largo de la vida. De acuerdo con la evidencia disponible hasta el momento, habría inmunidad de por vida.

Predecir enfermedades no es una tarea fácil. Como se evidenció en los resultados, la capacidad de predicción presentó variaciones dependiendo del municipio de estudio, lo cual indica que adicional a los factores climáticos, existe la influencia de otros factores intrínsecos en la dinámica epidemiológica local, dichos factores no fueron incluidos en el presente estudio, sin embargo, serán considerados en el trabajo futuro, variables como las condiciones socioeconómicas, movilidad poblacional, las campañas de salud pública durante periodos determinados. Esto con el objetivo de disponer de un mayor horizonte de información epidemiológica y climática, con el cual se podría estimar de manera más real el comportamiento de los casos de dengue a futuro.

Adicionalmente, para futuras investigaciones el reto consiste en el desarrollo de una herramienta de pronóstico que sea funcional para todos los municipios del departamento y a nivel nacional, mediante la inclusión de las variables mencionadas anteriormente, así como información meteorológica satelital.

Con relación al prototipo de arquitectura visual, permite la generación de pronósticos haciendo uso de las diferentes variables, sin embargo, se espera crear sinergias que permitan el mejoramiento continuo de la herramienta, de manera que en el futuro esta sea capaz de reestimar las probabilidades de riesgo de aparición de brotes para cada municipio de manera automática. En esta dirección, será posible centrar los esfuerzos en optimizarla, con previsiones cada vez más ajustadas al tiempo que se logre garantizar su continuidad. En este sentido, será una herramienta fácilmente exportable y adaptable a otros departamentos o incluso a nivel nacional.

8. Recomendaciones

Considerando la constante evolución de la tecnología de la información, también se sugiere para futuros trabajos la investigación de nuevas técnicas Deep Learning junto con nuevas combinaciones y configuraciones de las técnicas utilizadas por este trabajo. Esto con el objetivo de una correlación más precisa con los casos de dengue en las diferentes escalas espaciales de análisis. Dicha información contribuiría en gran medida en los programas de prevención y control de la enfermedad

Se recomienda seguir fortaleciendo las acciones de prevención y control con un abordaje integral desde los determinantes sociales de la salud, que incluyan estrategias y acciones intersectoriales que permitan modificar los determinantes sociales y ambientales de estas enfermedades. Por tal motivo, es relevante la búsqueda de mecanismos de coordinación interinstitucional sobre todo en materia de manejo de información que permitan el diseño e implementación de un sistema informático para la integración de diversas bases y capas de datos.

9. Resultado Asistencia a Congreso Internacional

A continuación, se presentan la asistencia a un congreso durante mis estudios de maestría, donde se presentó una ponencia relacionada con los resultados parciales de este trabajo de investigación.

- **Congreso Internacional**

IEOM Society International

The 3rd South American International Conference on
Industrial Engineering and Operations Management
Asuncion, Paraguay, July 18-21, 2022, Host: Asuncion National University

Certificate of Presentation

This is to certify that

Ruth Aralí Martínez Vega, PhD in Public Health Sciences with area concentration in Infectious Diseases, Professor School of Medicine University of Santander UDES
Henry Lamos Diaz, PhD Physics – Mathematics, Professor of Industrial Engineering, Industrial University of Santander UIS
Sonia Isabel Polo Triana, Master's Student in Industrial Engineering UIS

Delivered an Oral Presentation entitled "ID 271 Prediction of Dengue cases in Five Municipalities of Santander, Colombia using machine learning models." at the 3rd South American IEOM Paraguay Conference.

Dr. Jorge Kurita, Conference Chair, Research Faculty, Dept. of Industrial Engineering, Asuncion National University, Paraguay
Dr. Ahad Ali, Conference Co-Chair, Assoc. Professor and Director of IE, Lawrence Tech University, USA, Executive Director, IEOM Society
Prof. Vitor M. Caldana, Conference Program Chair, IFSP – Instituto Federal de São Paulo, Sorocaba, Sao Paulo, SP, Brazil
Prof. Don Reimer, Program Co-Chair, Director of Membership and Chapters – IEOM Society and Adjunct Prof. LTU, MI, USA

Sponsors and Partners

IEOM Society International, 21411 Civic Center Dr., Suite # 205, Southfield, Michigan 48076, USA, www.ieomsociety.org

10. Referencias

Agrebi, S., & Larbi, A. (2020). Use of artificial intelligence in infectious diseases. In *Artificial Intelligence in Precision Health*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-817133-2.00018-5

Angulo, V. M., Esteban, L., Urbano, P., Hincapié, E., & Núñez, L. A. (2013). Escenarios de transmisión de las principales enfermedades transmitidas por vectores en Colombia, 1990-2016. *Biomédica*, *33*(4), 24. https://doi.org/http://dx.doi.org/10.7705/biomedica.v33i4.836

Arista-Jalife, A., Nakano, M., Garcia-Nonoal, Z., Robles-Camarillo, D., Perez-Meana, H., & Arista-Viveros, H. A. (2020). Aedes mosquito detection in its larval stage using deep neural networks. *Knowledge-Based Systems*, *189*. https://doi.org/10.1016/j.knosys.2019.07.012

Arista-Jalife, Antonio, Nakano, M., Garcia-Nonoal, Z., Robles-Camarillo, D., Perez-Meana, H., & Arista-Viveros, H. A. (2020). Aedes mosquito detection in its larval stage using deep neural networks. *Knowledge-Based Systems*, *189*, 104841. https://doi.org/10.1016/j.knosys.2019.07.012

Arredondo-García, J., Méndez-Herrera, A., & Medina-Cortina, H. (2016). Arbovirus en Latinoamérica Correspondencia. *Acta Pediátrica de México*, 111–131.

Augusta, C., Deardon, R., & Taylor, G. (2019). Deep learning for supervised classification of spatial epidemics. *Spatial and Spatio-Temporal Epidemiology*, *29*, 187–198. https://doi.org/10.1016/j.sste.2018.08.002

Bansal, S., Chowell, G., Simonsen, L., Vespignani, A., & Viboud, C. (2016). Big data for infectious disease surveillance and modeling. *Journal of Infectious Diseases*, *214*(Suppl 4), S375–S379. https://doi.org/10.1093/infdis/jiw400

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 1–127.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*, 1798–1828.

Bermejo, D., & Vizcarra, G. (2020). *Modelo Basado en Aprendizaje Profundo para el Análisis de Sentimientos de Tuits en Español* [Universidad Nacional del Altiplano]. http://repositorio.unap.edu.pe/bitstream/handle/UNAP/12719/Bermejo_Escobar_Danitza_Yvette_Vizcarra_Aguilar_Gerson_Waldyr.pdf?sequence=1&isAllowed=y

Bontempi, G., Ben Taieb, S., & Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. *Lecture Notes in Business Information Processing*, *138 LNBIP*(June 2014), 62–77. https://doi.org/10.1007/978-3-642-36318-4_3

Brady, O. J., Golding, N., Pigott, D. M., Kraemer, M. U. G., Messina, J. P., Jr, R. C. R., Scott, T. W., Smith, D. L., Gething, P. W., & Hay, S. I. (2014). Global temperature constraints on Aedes aegypti and Ae . albopictus persistence and competence for dengue virus transmission. *Parasites and Vectors*, 1–17. https://parasitesandvectors.biomedcentral.com/articles/10.1186/1756-3305-7-338

Breiman, L. (2001). Random Forests. *Machine Learning*, 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2018). Better Deep Learning : Train Faster, Reduce Overfitting, and Make Better Predictions. *Machine Learning Mastery With Python*, *1*(2), 539.

Canyon, D. V., Hii, J. L. K., & Müller, R. (1999). Adaptation of Aedes aegypti (Diptera: Culicidae) oviposition behavior in response to humidity and diet. *Journal of Insect Physiology*, *45*(10), 959–964. https://doi.org/10.1016/S0022-1910(99)00085-2

Cardona Acosta, L. A. (2015). *Evaluación de factores ambientales y climáticos como*

elementos de riesgo asociados con la transmisión del dengue y la leishmaniasis a diferentes escalas temporales y espaciales en Colombia. [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/55890/32295880.2016.pdf?sequence=1&isAllowed=y

Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC Infectious Diseases*, *18*(1), 1–15. https://doi.org/10.1186/s12879-018-3066-0

Cea Morán, J. J. (2020). *Redes neuronales recurrentes para la generación automática de música* [Universidad Politécnica de Madrid]. https://oa.upm.es/63687/1/TFM_JUAN_JULIAN_CEA_MORAN.pdf

Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International Journal of Environmental Research and Public Health*, *15*(8). https://doi.org/10.3390/ijerph15081596

Cheng, J.-Z., Ni, D., & Chou, Y. (2016). Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT. *Scans. Sci Rep*, 244–254.

Chollet, F. (2017). *Deep Learning with Python*. https://www.manning.com/books/deep-learning-with-python

Chumino, V., Rodriguez, R., & Stalker, A. (2021). *Redes neuronales recurrentes aplicadas a sistemas de localización indoor en redes WLAN* [Universidad de la República]. https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/29502

Coccia, M. (2020). Deep learning technology for improving cancer care in society: New

directions in cancer imaging driven by artificial intelligence. *Technology in Society*, *60*(July 2019), 101198. https://doi.org/10.1016/j.techsoc.2019.101198

Costa, E. A. P. de A., Santos, E. M. de M., Correia, J. C., & de Albuquerque, C. M. R. (2010). Impact of small variations in temperature and humidity on the reproductive activity and survival of Aedes aegypti (Diptera, Culicidae). *Revista Brasileira de Entomologia*, *54*(3), 488–493. https://doi.org/10.1590/S0085-56262010000300021

Damián, M. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. *Historia*, 55. ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf

Fierro, A. A. (2020). Predicción de Series Temporales con Redes Neuronales [Universidad Nacional de La Plata]. In *Facultad de Informática Universidad Nacional de La Plata Argentina*. http://sedici.unlp.edu.ar/bitstream/handle/10915/114857/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Gourisaria, M. K., Das, S., Sharma, R., Rautaray, S. S., & Pandey, M. (2020). A deep learning model for malaria disease detection and analysis using deep convolutional neural networks. *International Journal on Emerging Technologies*, *11*(2), 699–704.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks; Proceedings of the 2013. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.

Guerrero, J. (2020). *Redes recurrentes* [Universidad de Sevillla]. https://idus.us.es/bitstream/handle/11441/115230/TFG DGMyE Pérez Guerrero%2C Jesús.pdf?sequence=1&isAllowed=y

Hall, H., Correa, A., Yoon, P., & Braden, C. (2012). Lexicon, Definitions, and Conceptual

Framework for Public Health Surveillance. *Centers for Disease Control and Prevention (CDC)*. doi: https://doi.org/su6103a5 [pii]

Hall, H. I., Correa, A., Yoon, P. W., & Braden, C. R. (2012). Lexicon, Definitions, and Conceptual Framework for Public Health Surveillance. In *Mmwr 2012* (Vol. 61, pp. 10–14). Centers for Disease Control and Prevention (CDC). https://doi.org/su6103a5 [pii]

Hinton, G., & Osindero, S. T. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput*, 1527–1554.

Hossain, A., Rahman, M., Kumar Prodhan, U., & Khan, F. (2013). Implementation of back-propagation neural Network for isolated bangla speech recognition. *International Journal of Information Sciences and Techniques*, *3*(4), 1–9. https://doi.org/10.5121/ijist.2013.3401

Huang, X., Williams, G., Clements, A. C. A., & Hu, W. (2013). Imported dengue cases, weather variation and autochthonous dengue incidence in Cairns, Australia. *PLoS ONE*, *8*(12), 1–7. https://doi.org/10.1371/journal.pone.0081887

INS. (2019). *Vigilancia Enfermedades Transmisibles*. Direcciones-Vigilancia.

Instituto Nacional de Salud, I. (2018). *Lineamientos Nacionales 2018 Vigilancia en Salud Pública.*

Instituto Nacional de Salud, I. (2019). *Boletín Epidemiológico Semanal, Semana epidemiológica 52 de 22 al 28 de diciembre de 2019*. https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2019_Boletin_epidemiologico_semana_52.pdf

Jiang, D., Hao, M., Ding, F., Fu, J., & Li, M. (2018). Mapping the transmission risk of Zika virus using machine learning models. *Acta Tropica*, *185*, 391–399. https://doi.org/10.1016/j.actatropica.2018.06.021

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep

convolutional neural networks. *Adv Neural Inf Process Syst*, 1097–1105.

Lampos, V., Miller, A. C., Crossan, S., & Stefansen, C. (2015). Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, *5*, 1–10. https://doi.org/10.1038/srep12760

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 436–444.

Li, R., Xu, L., Bjørnstad, O. N., Liu, K., Song, T., Chen, A., Xu, B., Liu, Q., & Stenseth, N. C. (2019). Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(9), 3624–3629. https://doi.org/10.1073/pnas.1806094116

Liu, L., Han, M., Zhou, Y., & Wang, Y. (2018). LSTM Recurrent Neural Networks for Influenza Trends Prediction. *Bioinform. Res. Appl*, 259–264.

Liu, T., Zhang, Y., Lin, H., Lv, X., Xiao, J., Zeng, W., Gu, Y., Rutherford, S., Tong, S., & Ma, W. (2015). A large temperature fluctuation may trigger an epidemic erythromelalgia outbreak in China. *Scientific Reports*, *5*, 1–8. https://doi.org/10.1038/srep09525

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Molnar, C. (2021). *Interpretable Machine Learning. A guide for Making Black Box Models Explainable*. https://fedefliguer.github.io/AAI/shap.html

Mooney, S. J., & Pejaver, V. (2018). Big Data in Public Health: Terminology, Machine Learning, and Privacy. In *Annual Review of Public Health* (Vol. 39). https://doi.org/10.1146/annurev-publhealth-040617-014208

Mordecai, E. A., Caldwell, J. M., Grossman, M. K., Lippi, C. A., Johnson, L. R., Neira, M.,

Rohr, J. R., Ryan, S. J., Savage, V., Shocket, M. S., Sippy, R., Stewart Ibarra, A. M., Thomas, M. B., & Villena, O. (2019). Thermal biology of mosquito-borne disease. *Ecology Letters*, *22*(10), 1690–1708. https://doi.org/10.1111/ele.13335

Morin, C. W., Comrie, A. C., & Ernst, K. (2013). Climate and dengue transmission: Evidence and implications. *Environmental Health Perspectives*, *121*(11–12), 1264–1272. https://doi.org/10.1289/ehp.1306556

Murray, J., & Cohen, A. L. (2017). Infectious Disease Surveillance. *International Encyclopedia of Public Health*, *January*, 222–229. https://doi.org/10.1016/B978-0-12-803678-5.00517-8 PMCID: PMC7149515

Naish, S., Dale, P., Mackenzie, J. S., McBride, J., Mengersen, K., & Tong, S. (2014). Climate change and dengue: A critical and systematic review of quantitative modelling approaches. *BMC Infectious Diseases*, *14*(1), 1–14. https://doi.org/10.1186/1471-2334-14-167

Noureldin, E., & Shaffer, L. (2019). Role of climatic factors in the incidence of dengue in port sudan city, sudan. *Eastern Mediterranean Health Journal*, *25*(12), 852–860. https://doi.org/10.26719/emhj.19.019

O'Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International Journal of Medical Informatics*, 15–22. doi:https://doi.org/10.1016/j.ijmedinf.2017.01.019

Organización Mundial de la Salud, O. (2016). *Reglamento Sanitario Internacional (2005) 3 Ed.* https://apps.who.int/iris/bitstream/handle/10665/246186/9789243580494-spa.pdf;jsessionid=633A8E92C4A748860815280A00350353?sequence=1

Ouattara, C. A., Traore, T. I., Traore, S., Sangare, I., Meda, C. Z., & Savadogo, L. G. B. (2022). Climate factors and dengue fever in Burkina Faso from 2017 to 2019. *Journal*

of Public Health in Africa, *13*(1), 2–5. https://doi.org/10.4081/jphia.2022.2145

Padilla, J. C., Rojas, D. P., & Sáenz-Gómez, R. (2012). *Dengue en Colombia: Epidemiología de la reemergencia a la hiperendemia*.

Porrello, A., Vincenzi, S., Buzzega, P., Calderara, S., Conte, A., Ippoliti, C., Candeloro, L., Di Lorenzo, A., & Capobianco Dondona, A. (2019). Spotting insects from satellites: Modeling the presence of culicoides imicola through deep CNNs. *Proceedings - 15th International Conference on Signal Image Technology and Internet Based Systems, SISITS 2019*, 159–166. https://doi.org/10.1109/SITIS.2019.00036

Porta, M. (2008). *Dictionary of Epidemiology*. Oxford University Press, USA.

Prasoon, A., Petersen, K., & Igel, C. (2013). Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv*, 246–253.

Rohart, F., Milinovich, G. J., Avril, S. M. R., Lê Cao, K. A., Tong, S., & Hu, W. (2016). Disease surveillance based on Internet-based linear models: An Australian case study of previously unmodeled infection diseases. *Scientific Reports*, *6*(April), 1–11. https://doi.org/10.1038/srep38522

Samaras, L., García-Barriocanal, E., & Sicilia, M.-A. (2020). Comparing Social media and Google to detect and predict severe epidemics. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-61686-9

Scarafoni, D., Telfer, B. A., Ricke, D. O., Thornton, J. R., & Comolli, J. (2019). Predicting Influenza A Tropism with End-to-End Learning of Deep Networks. *Health Security*, *17*(6), 468–476. https://doi.org/10.1089/hs.2019.0055

Schmidhuber, S., & Hochreiter, J. (1997). LONG SHORT-TERM MEMORY. *Neural Computation*.

Shi, Y., Liu, X., Kok, S. Y., Rajarethinam, J., Liang, S., Yap, G., Chong, C. S., Lee, K. S., Tan, S. S. Y., Chin, C. K. Y., Lo, A., Kong, W., Ng, L. C., & Cook, A. R. (2016). Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in Singapore. *Environmental Health Perspectives*, *124*(9), 1369–1375. https://doi.org/10.1289/ehp.1509981

Shin, S. Y., Seo, D. W., An, J., Kwak, H., Kim, S. H., Gwack, J., & Jo, M. W. (2016). High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Scientific Reports*, *6*(September 2015), 1–7. https://doi.org/10.1038/srep32920

Simonsen, L., Gog, J. R., Olson, D., & Viboud, C. (2016). Infectious disease surveillance in the big data era: Towards faster and locally relevant systems. *Journal of Infectious Diseases*, *214*(Suppl 4), S380–S385. https://doi.org/10.1093/infdis/jiw376

Soliman, M., Lyubchich, V., & Gel, Y. R. (2020). Ensemble forecasting of the Zika space-time spread with topological data analysis. *Environmetrics*. https://doi.org/10.1002/env.2629

Sutskever, I., Vinyals, O., & Le, Q. (2014). equence to sequence learning with neural networks. *Adv Neural Inf Process Syst*, 3104–3012.

Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., An, X., Feng, D., & Tong, Y. (2017). Dynamic forecasting of zika epidemics using google trends. *PLoS ONE*, *12*(1), 1–10. https://doi.org/10.1371/journal.pone.0165085

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-18007-4

Trask, A. W. (2019). *Grokking Deep Learning* (C. Taylor (ed.)).

Tsui, K. L., Wong, S. Y., Jiang, W., & Lin, C. J. (2011). Recent research and developments in temporal and spatiotemporal surveillance for public health. *IEEE Transactions on Reliability*, *60*(1), 49–58. https://doi.org/10.1109/TR.2010.2104192

Vásques Brenes, P. A. (2020). *Uso del aprendizaje automatizado y de variables climáticas como herramienta para la predicción del riesgo de Dengue en Costa Rica* [Universidad de Costa Rica]. https://www.kerwa.ucr.ac.cr/bitstream/handle/10669/82278/TFG. Paola Vasquez Brenes.pdf?sequence=1&isAllowed=y

Vivas, H., Martínez, H., & Pérez, R. (2014). Structured secant method for the multilayer perceptron training. *Revista de Ciencias*, 131–150.

Wang, C., Qi, Y., & Zhu, G. (2020). Deep learning for predicting the occurrence of cardiopulmonary diseases in Nanjing, China. *Chemosphere*, *257*. https://doi.org/10.1016/j.chemosphere.2020.127176

Wang, L., Chen, J., & Marathe, M. (2020). TDEFSI: Theory-guided Deep Learning-based Epidemic Forecasting with Synthetic Information. *ACM Transactions on Spatial Algorithms and Systems*, *6*(3). https://doi.org/10.1145/3380971

Wang, M., Wang, H., Wang, J., Liu, H., Lu, R., Duan, T., Gong, X., Feng, S., Liu, Y., Cui, Z., Li, C., & Ma, J. (2019). A novel model for malaria prediction based on ensemble algorithms. *PLoS ONE*, *14*(12). https://doi.org/10.1371/journal.pone.0226910

WHO. (2017). *Vector-borne diseases*. News.

Wijayanti, S. P. M., Sunaryo, S., Suprihatin, S., McFarlane, M., Rainey, S. M., Dietrich, I., Schnettler, E., Biek, R., & Kohl, A. (2016). Dengue in Java, Indonesia: Relevance of Mosquito Indices as Risk Predictors. *PLoS Neglected Tropical Diseases*, *10*(3), 1–15. https://doi.org/10.1371/journal.pntd.0004500

Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., & Anthony, L. (2015). *Big data in global health : improving health in low- and middle-income countries*. *November 2014*, 203–208.

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). https://doi.org/10.3390/ijerph17020453

Yang, H. M., Macoris, M. L. G., Galvani, K. C., Andrighetti, M. T. M., & Wanderley, D. M. V. (2009). Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue. *Epidemiology and Infection*, *137*(8), 1188–1202. https://doi.org/10.1017/S0950268809002040

Yoo, Y., Brosch, T., & Traboulsee, A. (2014). Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation. *International Workshop on Machine Learning in Medical Imaging, Boston*, 117–124.

Zhang, J., & Nawata, K. (2017). A comparative study on predicting influenza outbreaks. *BioScience Trends*, *11*(5), 533–541. https://doi.org/10.5582/bst.2017.01257

Zhao, L., Chen, J., Chen, F., Jin, F., Wang, W., Lu, C.-T., & Ramakrishnan, N. (2020). Online flu epidemiological deep modeling on disease contact network. *GeoInformatica*, *24*(2), 443–475. https://doi.org/10.1007/s10707-019-00376-9

Anexos

Anexo A: Revisión de Literatura

Con el objetivo de conocer y explorar la aplicación de tecnologías Deep Learning en el sector salud, específicamente en la vigilancia de enfermedades arbovirales y/o infecciosas, se realiza una revisión de literatura de tal manera que se puedan identificar oportunidades de investigación y obtener un conocimiento sobre estas técnicas. A continuación, se presenta el protocolo de la revisión:

Ecuación de búsqueda

La ecuación de búsqueda utilizada fue la siguiente:

((TITLE-ABS-KEY (“Deep Learning”) AND ("infectious disease*" OR "arboviral disease*" OR epidemic) AND (surveillance OR propagation OR "predict*" OR "forescast*"))

La base de datos que se usó para la búsqueda de bibliografía fue www.SCOPUS.com, la cual permitió exportar los resultados y analizarlos con el software de análisis bibliométrico, que en este caso fue R-studio, con la librería bibliometrix. Seguidamente se aplican los criterios de inclusión y exclusión:

- **Criterios de inclusión:**

a) Selección de tipos de documentos como artículos científicos y artículos de revisión.

b) Documentos en idioma inglés y español.

d) Documentos relacionados con temáticas como ingeniería, medicina, ciencias computacionales, específicamente aquellos relacionados con la aplicación de técnicas Deep Learning en enfermedades infecciosas.

- **Criterios de exclusión:**

a) Documentos relacionados con enfermedades diferentes a las infecciosas y/o arbovirales como el Cáncer.

b) Trabajos relacionados con la vigilancia de enfermedades a partir del procesamiento de imágenes.

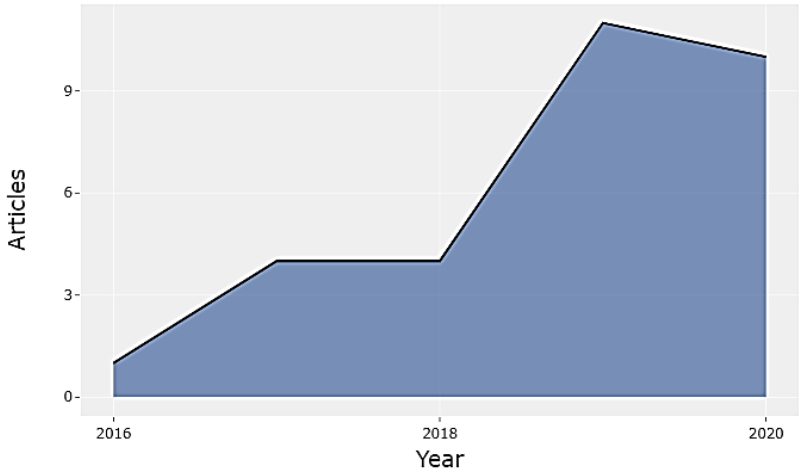
c) Documentos tipo cartas, resúmenes, artículos o revisiones de conferencia.

Ejecución de la revisión

La búsqueda arrojó un total de 71 documentos, a los cuales se les hace una revisión de su título y resumen descartando aquellos que no cumplen con los criterios establecidos. De este ejercicio resultan 30 documentos que son estudiados en su totalidad. Cabe resaltar que se realizó de manera paralela un ejercicio de bola de nieve donde se identifican trabajos de interés para la investigación, pero al no encontrarse en Scopus, aunque se encuentran referenciados, no fueron incluidos en el presente análisis bibliométrico.

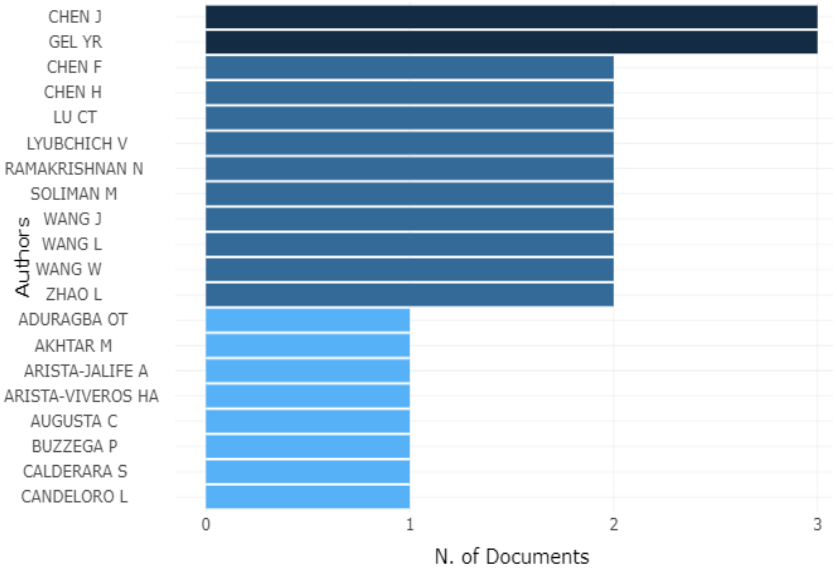
A continuación, se presentan los resultados más relevantes

Figura 59. Producción anual del tema de investigación



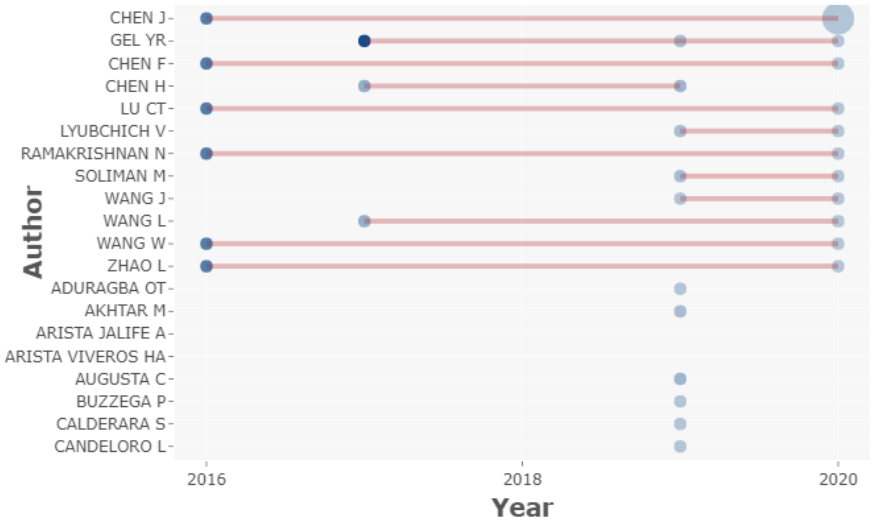
La figura 9 muestra que las aplicaciones de las técnicas Deep Learning en la vigilancia de enfermedades arbovirales es un tema de investigación reciente y que ha venido aumentado el interés de los investigadores por esta temática. En la 135igura 10 se detallan los autores más relevantes (eje y) y el número de publicaciones (eje x). El primer lugar lo lideran 2 autores con 3 publicaciones cada uno.

Figura 60. Autores principales relacionados al tema de investigación



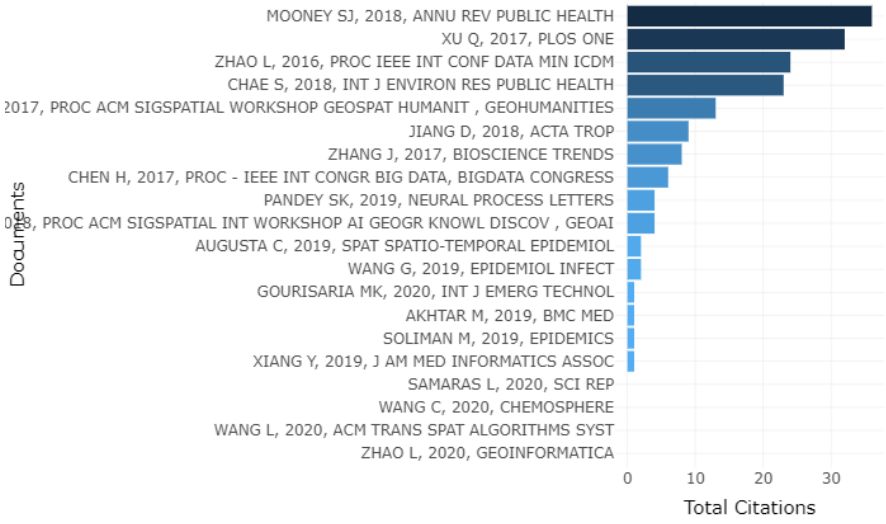
En La figura 11 se detalla el nivel de actividad por año para cada autor. Se evidencia que seis autores han tenido actividad constante desde el 2016 tres de ellos desde el 2017, mientras que la mayoría solo han tenido publicaciones recientes.

Figura 61. Producción a través del tiempo de los autores principales



Con relación a los artículos más citados, se destaca el trabajo de Mooney & Pejaver (2018) en el cual se realiza una revisión de los usos del Big Data en la Salud Pública.

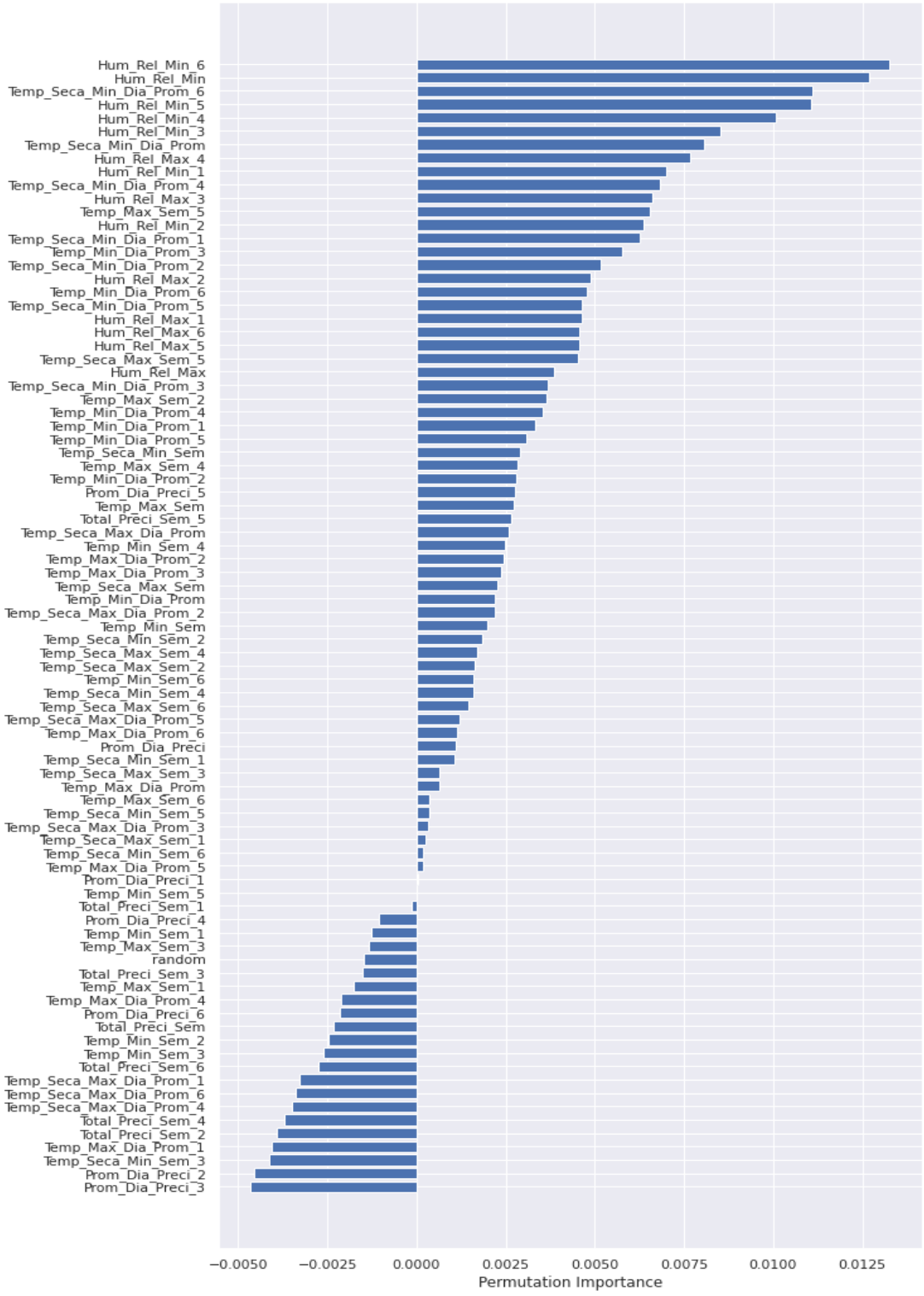
Figura 62. Artículos más citados



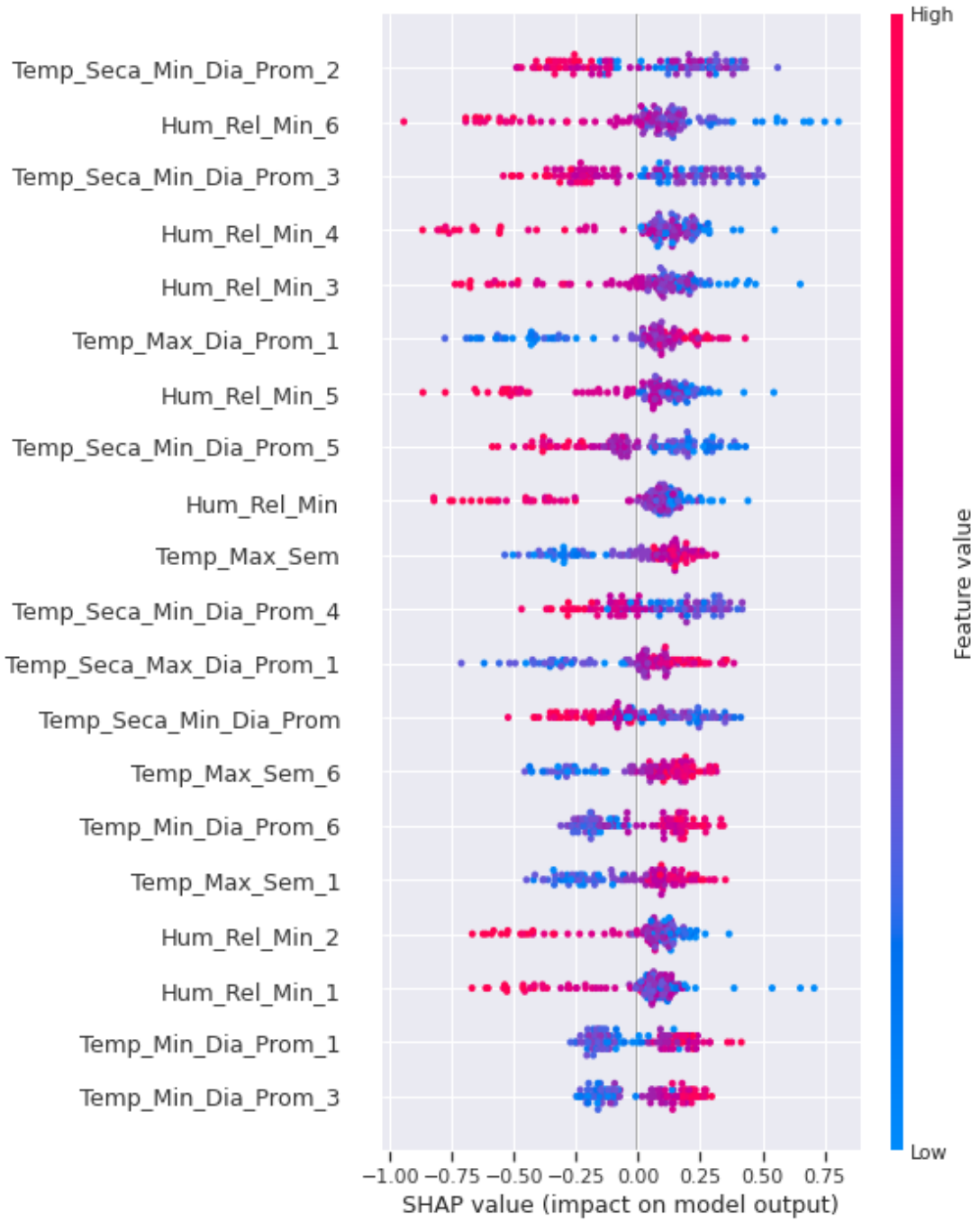
Finalmente, la nube de palabras, donde se organiza las palabras más usadas (título, abstract y keywords) y se les asigna un tamaño acorde con su frecuencia de aparición, por lo tanto, las palabras más grandes son las más usadas en todos los documentos de la búsqueda.

Anexo B: Resultados Importancia de la Características Barrancabermeja

- **Importancia de las características por permutación – Barrancabermeja**



- **Importancia de las características por valores SHAP – Barrancabermeja**

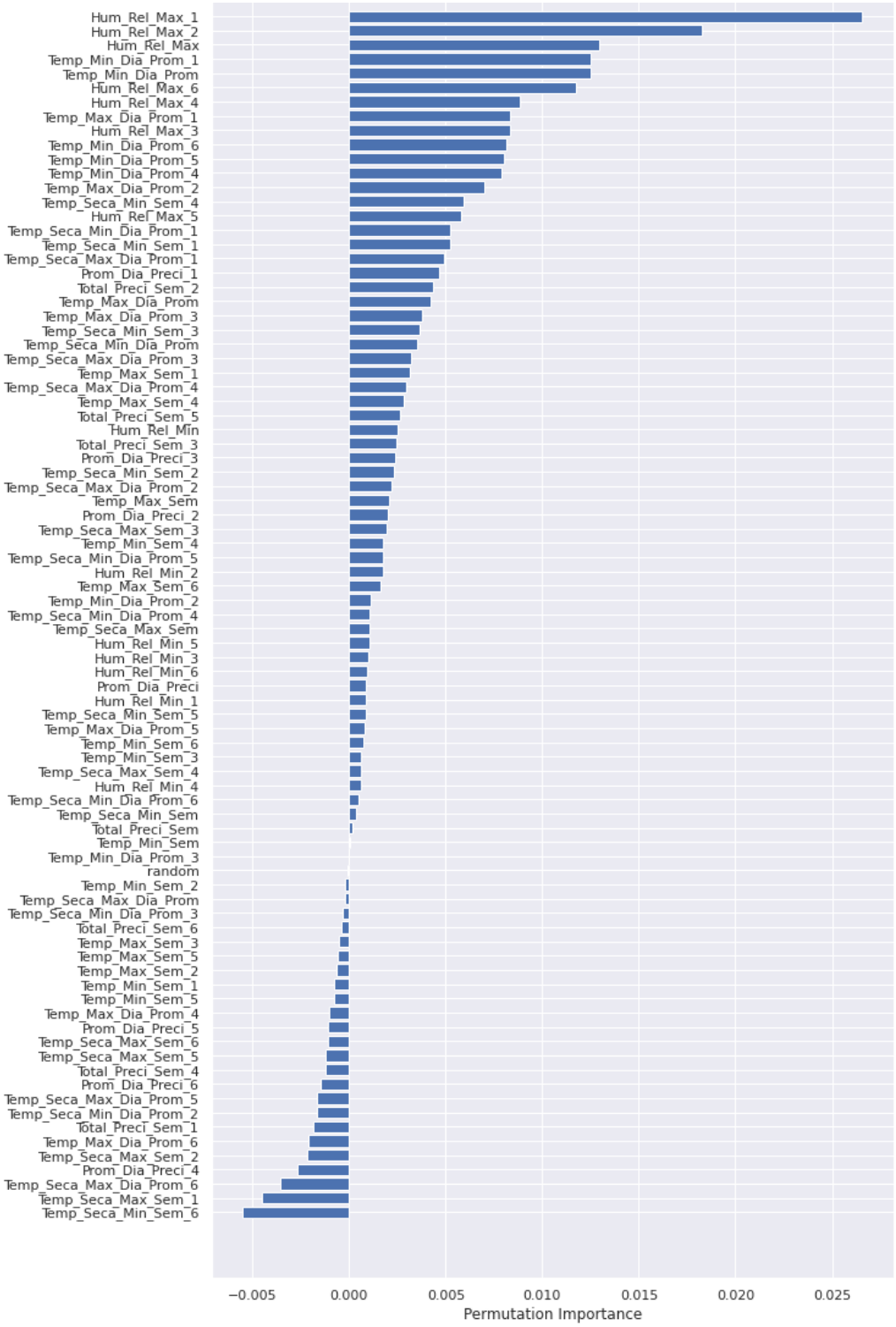


- **Análisis de Correlaciones - Barrancabermeja**

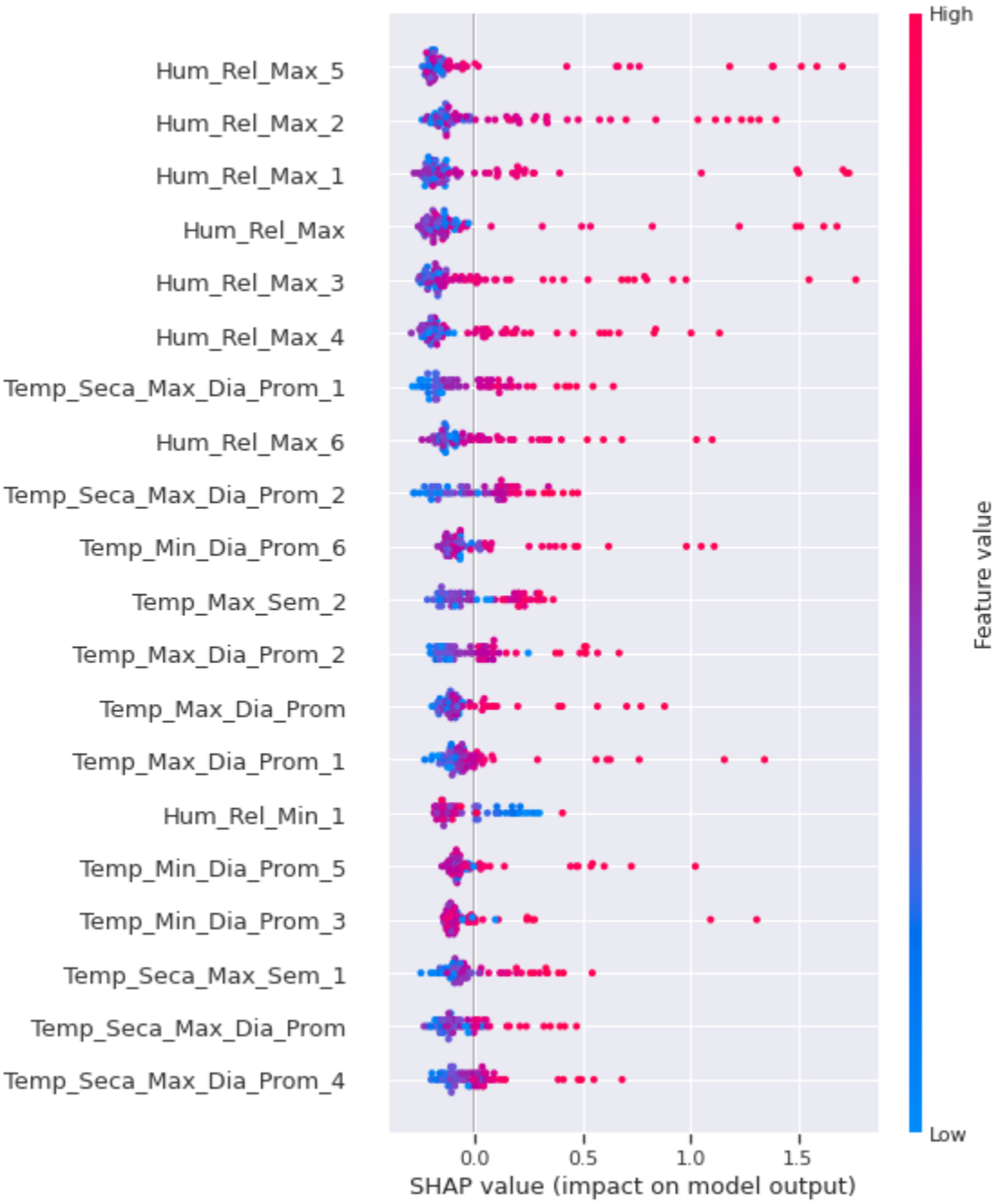
Temp_Max_Sem_1	0.076039
Temp_Min_Dia_Prom_1	0.075566
Temp_Seca_Max_Sem_6	0.074100
Temp_Max_Sem	0.073516
Temp_Max_Sem_2	0.072343
Temp_Seca_Max_Dia_Prom	0.069847
Temp_Seca_Max_Dia_Prom_1	0.068800
Temp_Max_Sem_6	0.068341
Temp_Seca_Max_Sem_1	0.066713
Temp_Seca_Max_Sem	0.064881
Temp_Max_Dia_Prom	0.063053
Temp_Min_Dia_Prom	0.062969
Temp_Seca_Max_Sem_5	0.062067
Temp_Max_Sem_5	0.061091
Temp_Seca_Max_Dia_Prom_2	0.059918
Temp_Seca_Max_Sem_2	0.059401
Temp_Min_Dia_Prom_6	0.058985
Temp_Min_Dia_Prom_3	0.058883
Temp_Max_Dia_Prom_1	0.057385
Temp_Min_Dia_Prom_2	0.054772
Temp_Max_Sem_3	0.053675
Temp_Max_Sem_4	0.050430
Temp_Min_Dia_Prom_5	0.050152
Temp_Seca_Max_Sem_4	0.049641
Temp_Seca_Max_Dia_Prom_3	0.048272
Temp_Max_Dia_Prom_2	0.044450
Temp_Min_Dia_Prom_4	0.043531
Temp_Seca_Max_Sem_3	0.043412
Temp_Max_Dia_Prom_3	0.041180
Temp_Seca_Max_Dia_Prom_5	0.040419
Total_Preci_Sem_5	-0.031139
Prom_Dia_Preci_1	-0.031378
Temp_Seca_Min_Sem_3	-0.031425
Total_Preci_Sem_6	-0.031489
Temp_Seca_Min_Sem	-0.031710
Temp_Seca_Min_Sem_5	-0.032802
Temp_Seca_Min_Dia_Prom_5	-0.033006
Prom_Dia_Preci_5	-0.035166
Prom_Dia_Preci_6	-0.035365
Temp_Seca_Min_Dia_Prom_6	-0.037521
Temp_Seca_Min_Sem_6	-0.037791
Hum_Rel_Max_1	-0.060949
Hum_Rel_Max_3	-0.064428
Hum_Rel_Max_2	-0.070262
Temp_Seca_Max_Dia_Prom_6	0.034157
Temp_Max_Dia_Prom_5	0.033716
Temp_Seca_Max_Dia_Prom_4	0.033689
Semana_Epi	0.026691
Temp_Max_Dia_Prom_6	0.025918
Temp_Max_Dia_Prom_4	0.021353
Año	0.013319
Temp_Min_Sem_1	-0.006433
Temp_Min_Sem_2	-0.006745
Temp_Min_Sem	-0.007829
Temp_Seca_Min_Dia_Prom_1	-0.013482
Temp_Seca_Min_Sem_1	-0.014482
Temp_Min_Sem_4	-0.015246
Temp_Min_Sem_3	-0.016123
Temp_Min_Sem_6	-0.016636
Total_Preci_Sem_4	-0.018665
Temp_Seca_Min_Dia_Prom_2	-0.019382
Temp_Seca_Min_Dia_Prom	-0.019939
Total_Preci_Sem_2	-0.021930
Prom_Dia_Preci_4	-0.023021
Total_Preci_Sem_3	-0.024778
Temp_Seca_Min_Dia_Prom_3	-0.025049
Temp_Min_Sem_5	-0.026358
Prom_Dia_Preci_2	-0.026414
Total_Preci_Sem	-0.026658
Total_Preci_Sem_1	-0.027574
Temp_Seca_Min_Dia_Prom_4	-0.029009
Prom_Dia_Preci_3	-0.029346
Temp_Seca_Min_Sem_2	-0.029541
Temp_Seca_Min_Sem_4	-0.030770
Prom_Dia_Preci	-0.030869
Hum_Rel_Max	-0.071485
Hum_Rel_Max_4	-0.077675
Hum_Rel_Max_5	-0.078021
Hum_Rel_Max_6	-0.078467
Hum_Rel_Min_2	-0.156961
Hum_Rel_Min	-0.165334
Hum_Rel_Min_1	-0.166278
Hum_Rel_Min_3	-0.172941
Hum_Rel_Min_5	-0.173411
Hum_Rel_Min_4	-0.175615
Hum_Rel_Min_6	-0.190853

Anexo C: Resultados Importancia de la Características Girón

- **Importancia de las características por permutación – Girón**



- **Importancia de las características por valores SHAP – Girón**

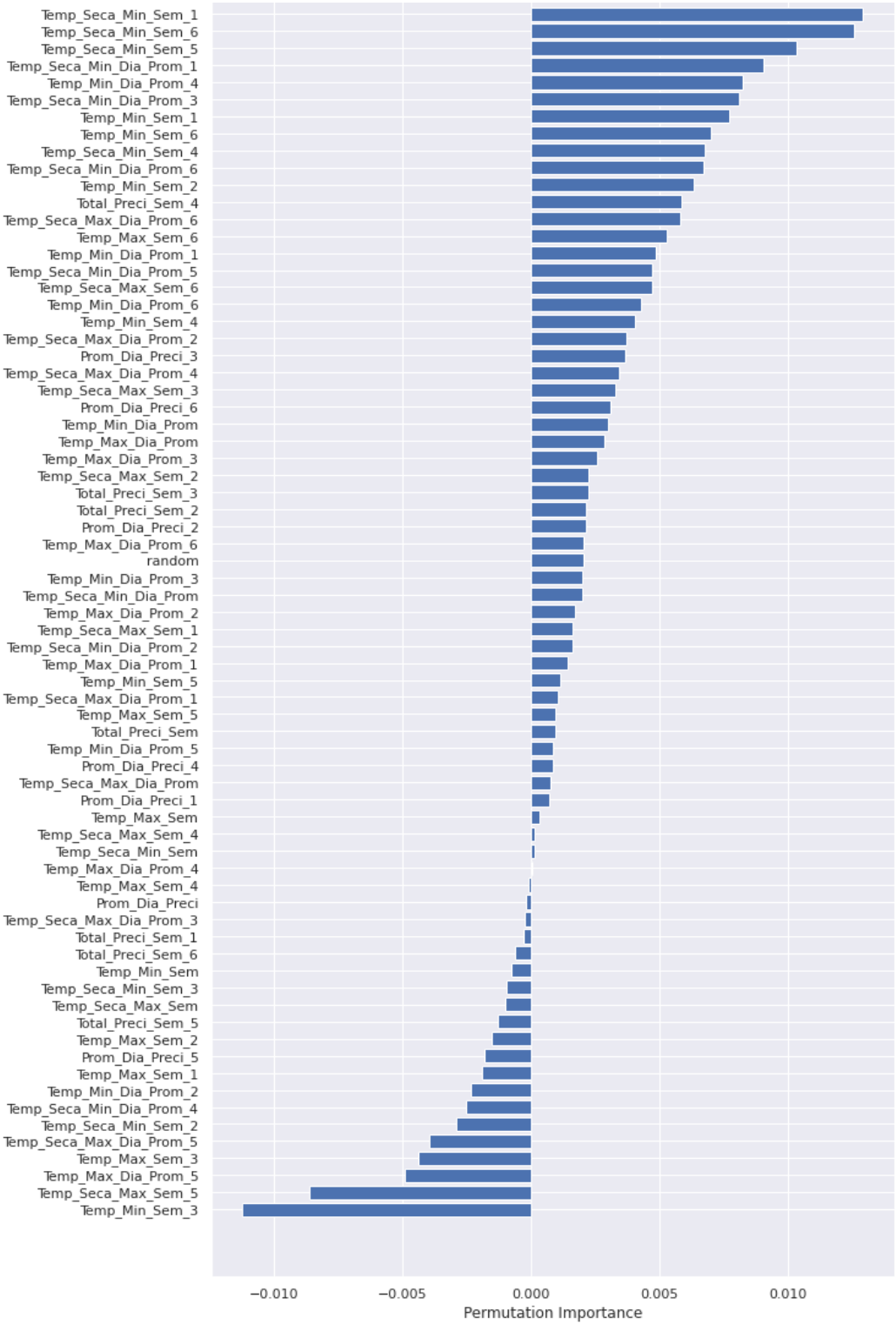


- **Análisis de Correlaciones – Girón**

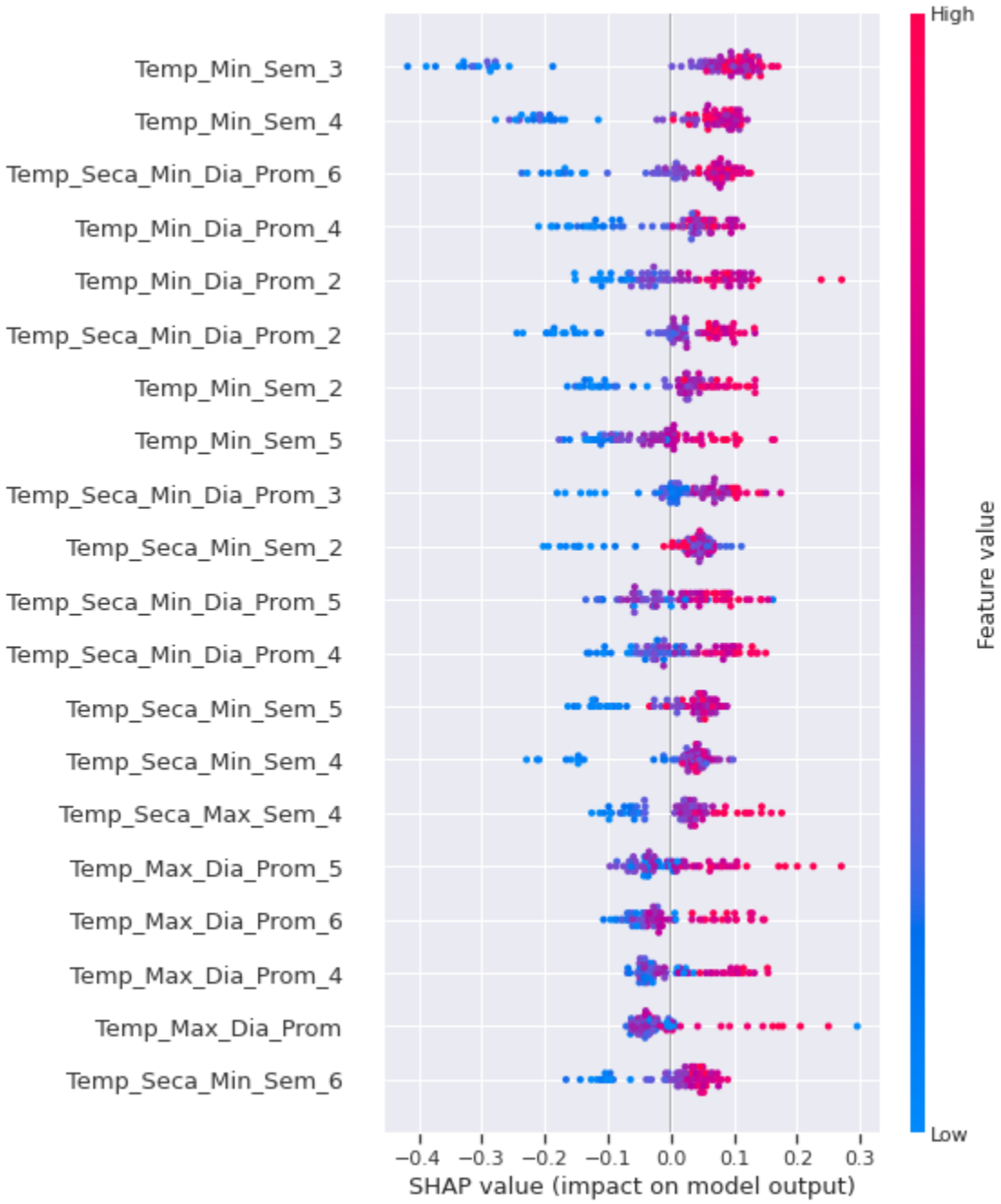
Temp_Seca_Min_Dia_Prom	0.095978
Hum_Rel_Max_4	0.247883
Temp_Seca_Max_Dia_Prom_6	0.093343
Hum_Rel_Max_5	0.245520
Temp_Min_Dia_Prom_2	0.090022
Hum_Rel_Max_3	0.238640
Temp_Min_Dia_Prom_1	0.087686
Hum_Rel_Max_6	0.237988
Temp_Seca_Min_Dia_Prom_2	0.084329
Hum_Rel_Max	0.225757
Temp_Min_Dia_Prom	0.083386
Hum_Rel_Max_2	0.215679
Temp_Seca_Min_Dia_Prom_6	0.078853
Hum_Rel_Max_1	0.213170
Temp_Max_Dia_Prom_5	0.078239
Temp_Seca_Max_Dia_Prom_1	0.210439
Temp_Max_Sem_4	0.077332
Temp_Seca_Max_Dia_Prom_2	0.204388
Temp_Min_Dia_Prom_5	0.066900
Temp_Seca_Max_Dia_Prom	0.203482
Temp_Min_Dia_Prom_3	0.066507
Temp_Max_Dia_Prom_1	0.200619
Temp_Min_Dia_Prom_4	0.066213
Temp_Max_Dia_Prom	0.195877
Temp_Seca_Max_Sem_6	0.066144
Temp_Seca_Max_Dia_Prom_4	0.187669
Temp_Min_Sem_6	0.060859
Temp_Seca_Max_Dia_Prom_3	0.185045
Temp_Seca_Min_Dia_Prom_3	0.057838
Temp_Max_Dia_Prom_2	0.184313
Temp_Seca_Min_Dia_Prom_4	0.056168
Temp_Seca_Max_Sem	0.165881
Temp_Seca_Min_Dia_Prom_5	0.045316
Temp_Seca_Max_Sem_1	0.163803
Temp_Max_Dia_Prom_6	0.042814
Temp_Seca_Max_Sem_2	0.158611
Total_Preci_Sem_6	0.031496
Temp_Max_Dia_Prom_4	0.151395
Prom_Dia_Preci_6	0.031402
Temp_Max_Dia_Prom_3	0.149295
Temp_Min_Sem_1	0.024717
Temp_Max_Sem	0.138495
Temp_Max_Sem_5	0.021150
Temp_Seca_Max_Sem_4	0.137469
Temp_Min_Sem_2	0.016380
Temp_Seca_Max_Sem_3	0.134639
Temp_Min_Sem_4	0.007336
Temp_Max_Sem_2	0.129298
Temp_Min_Sem	0.005891
Temp_Max_Sem_1	0.126930
Hum_Rel_Min_6	0.005825
Temp_Seca_Max_Dia_Prom_5	0.124853
Temp_Min_Sem_5	0.001613
Temp_Seca_Max_Sem_5	0.107254
Prom_Dia_Preci	-0.002715
Temp_Min_Dia_Prom_6	0.107135
Total_Preci_Sem	-0.002983
Temp_Seca_Min_Dia_Prom_1	0.101875
Temp_Min_Sem_3	-0.005382
Temp_Max_Sem_3	0.101065
Temp_Seca_Min_Sem_1	-0.008197
Prom_Dia_Preci_1	-0.010539
Total_Preci_Sem_2	-0.039679
Total_Preci_Sem_1	-0.011083
Prom_Dia_Preci_2	-0.039740
Temp_Seca_Min_Sem_6	-0.012844
Temp_Seca_Min_Sem_4	-0.041459
Temp_Max_Sem_6	-0.017159
Temp_Seca_Min_Sem_2	-0.044102
Temp_Seca_Min_Sem	-0.020204
Temp_Seca_Min_Sem_3	-0.050510
Total_Preci_Sem_3	-0.026441
Hum_Rel_Min_3	-0.067018
Prom_Dia_Preci_3	-0.026449
Hum_Rel_Min_4	-0.068338
Hum_Rel_Min_5	-0.030369
Hum_Rel_Min	-0.071102
Prom_Dia_Preci_5	-0.031112
Hum_Rel_Min_2	-0.083677
Total_Preci_Sem_5	-0.031205
Hum_Rel_Min_1	-0.092519
Temp_Seca_Min_Sem_5	-0.032926
Total_Preci_Sem_4	-0.034247
Prom_Dia_Preci_4	-0.034479

Anexo D: Resultados Importancia de la Características Lebrija

- **Importancia de las características por permutación – Lebrija**



- **Importancia de las características por valores SHAP – Lebrija**



- **Análisis de Correlaciones - Lebrija**

Temp_Seca_Min_Dia_Prom_3	0.250694
Temp_Seca_Min_Dia_Prom_4	0.247290
Temp_Seca_Min_Dia_Prom_2	0.245513
Temp_Seca_Min_Dia_Prom_5	0.241142
Temp_Seca_Min_Dia_Prom_6	0.240956
Temp_Seca_Max_Dia_Prom_6	0.228850
Temp_Seca_Min_Dia_Prom_1	0.227605
Temp_Min_Dia_Prom_5	0.225745
Temp_Seca_Min_Dia_Prom	0.220425
Temp_Max_Dia_Prom_5	0.219199
Temp_Seca_Min_Sem_5	0.216753
Temp_Max_Dia_Prom_6	0.216464
Temp_Min_Dia_Prom_3	0.212153
Temp_Seca_Max_Dia_Prom_5	0.211651
Temp_Max_Dia_Prom_4	0.210650
Temp_Seca_Max_Dia_Prom_4	0.208281
Temp_Min_Dia_Prom_6	0.207551
Temp_Seca_Min_Sem_4	0.206039
Temp_Min_Dia_Prom_4	0.205993
Temp_Seca_Min_Sem_6	0.205442
Temp_Max_Dia_Prom_3	0.203545
Temp_Max_Sem_4	0.201158
Temp_Min_Dia_Prom_2	0.200824
Temp_Max_Sem_6	0.200175
Temp_Seca_Max_Sem_4	0.199759
Temp_Min_Sem_4	0.198496
Temp_Max_Sem_3	0.197184
Temp_Seca_Max_Dia_Prom_3	0.196358
Temp_Min_Sem_5	0.194610
Temp_Seca_Min_Sem_3	0.194498
Temp_Min_Sem_3	0.194456
Temp_Seca_Max_Sem_6	0.192200
Temp_Seca_Min_Sem_1	0.191131
Temp_Max_Sem	0.188411
Temp_Min_Dia_Prom_1	0.185426
Temp_Seca_Max_Dia_Prom_2	0.184765
Temp_Seca_Max_Dia_Prom	0.184506
Temp_Min_Sem_6	0.184391
Temp_Seca_Max_Sem_5	0.183293
Temp_Max_Dia_Prom	0.182269
Temp_Seca_Min_Sem_2	0.180403
Temp_Min_Sem_2	0.179762
Temp_Min_Sem_1	0.179150
Temp_Max_Sem_5	0.179138
Temp_Seca_Max_Sem_3	0.178749
Temp_Max_Sem_2	0.178363
Temp_Max_Dia_Prom_2	0.178038
Temp_Seca_Max_Dia_Prom_1	0.177857
Temp_Seca_Max_Sem_2	0.171145
Temp_Max_Dia_Prom_1	0.169631
Temp_Seca_Max_Sem	0.164550
Temp_Seca_Min_Sem	0.163214
Temp_Max_Sem_1	0.159457
Temp_Min_Dia_Prom	0.152179
Temp_Seca_Max_Sem_1	0.149272
Semana_Epi	0.123654
Temp_Min_Sem	0.114318
Prom_Dia_Preci_2	-0.063264
Total_Preci_Sem_2	-0.063335
Prom_Dia_Preci_3	-0.071185
Total_Preci_Sem_3	-0.071651
Total_Preci_Sem_1	-0.075281
Total_Preci_Sem_5	-0.075440
Prom_Dia_Preci_1	-0.075721
Prom_Dia_Preci_5	-0.076023
Total_Preci_Sem	-0.082236
Prom_Dia_Preci	-0.082265
Total_Preci_Sem_4	-0.095164
Prom_Dia_Preci_4	-0.095392
Total_Preci_Sem_6	-0.097799
Prom_Dia_Preci_6	-0.097959

Anexo E: Resultados Análisis de Correlaciones Bucaramanga, Floridablanca y Piedecuesta

Bucaramanga

Total_Preci_Sem_1	-0.003866
Prom_Dia_Preci_1	-0.003911
Prom_Dia_Preci	-0.007573
Total_Preci_Sem	-0.007839
Prom_Dia_Preci_2	-0.011427
Total_Preci_Sem_2	-0.011648
Total_Preci_Sem_4	-0.021885
Prom_Dia_Preci_4	-0.022010
Prom_Dia_Preci_3	-0.022916
Total_Preci_Sem_3	-0.023043
Total_Preci_Sem_5	-0.030913
Prom_Dia_Preci_5	-0.030994
Total_Preci_Sem_6	-0.039253
Prom_Dia_Preci_6	-0.039469

Floridablanca

Prom_Dia_Preci_2	-0.061136
Total_Preci_Sem_2	-0.061706
Prom_Dia_Preci	-0.063144
Total_Preci_Sem	-0.063880
Prom_Dia_Preci_4	-0.064432
Total_Preci_Sem_4	-0.064917
Prom_Dia_Preci_1	-0.070908
Total_Preci_Sem_1	-0.071684
Prom_Dia_Preci_3	-0.072096
Total_Preci_Sem_3	-0.072783
Semana_Epi	-0.075163
Prom_Dia_Preci_5	-0.078919
Total_Preci_Sem_5	-0.079531
Prom_Dia_Preci_6	-0.092193
Total_Preci_Sem_6	-0.092626

Piedecuesta

Prom_Dia_Preci_2	0.013305
Prom_Dia_Preci_4	0.013305
Total_Preci_Sem_2	0.012389
Prom_Dia_Preci_5	-0.005645
Total_Preci_Sem_5	-0.006459
Semana_Epi	-0.015582
Prom_Dia_Preci_1	-0.016064
Total_Preci_Sem_1	-0.016489
Prom_Dia_Preci_3	-0.017774
Total_Preci_Sem_3	-0.018507
Prom_Dia_Preci_6	-0.020350
Total_Preci_Sem_6	-0.020949
Prom_Dia_Preci	-0.023251
Total_Preci_Sem	-0.023918
Total_Preci_Sem_4	-0.035886