DOCUMENTO FINAL DE GRADO

MODELO DE ESTIMACIÓN DE CONTAGIOS Y MUERTES EN UN ESCENARIO DE PANDEMIA – CASO COVID-19 EN COLOMBIA

LAURA SOFIA LOPEZ ZUÑIGA

CORPORACION UNIVERSITARIA COMFACAUCA
FACULTAD DE INGENIERIAS
INGENIERIA MECATRONICA
POPAYAN
2023

DOCUMENTO FINAL DE GRADO

MODELO DE ESTIMACIÓN DE CONTAGIOS Y MUERTES EN UN ESCENARIO DE PANDEMIA – CASO COVID-19 EN COLOMBIA

LAURA SOFIA LOPEZ ZUÑIGA

Trabajo de grado para optar al título de Ingeniería Mecatrónica

Director de tesis:

PhD. JAVIER ANDRES MUÑOZ CHAVES

CORPORACION UNIVERSITARIA COMFACAUCA
FACULTAD DE INGENIERIAS
INGENIERIA MECATRONICA
POPAYAN
2023

Contenido

Listado de figuras

Pag.

Listado de tablas

Pag.

RESUMEN

Para marzo del 2020, la Organización Mundial de la Salud (OMS) declaró la epidemia de coronavirus como una pandemia mundial. Esto debido a que el brote de COVID-19 se había extendido por 118 países de todo el mundo, con un total a esa fecha de 125.260 casos confirmados. Sin embargo, los casos en todo el mundo superaron los 16 millones, de los cuales más de 650.000 resultaron mortales. A nivel nacional desde los primeros meses del año 2020 se reportaron casos de contagios por COVID-19 y desde ese entonces las cifras se incrementaron rápidamente. Con base en esto, el objetivo de este proyecto fue desarrollar un modelo que estime de la cantidad de contagios y muertes por COVID-19 en Colombia como apoyo al sistema de salud, identificando posibles contagios y muertes en el futuro. Las predicciones se basaron en el dataset publicado por el ministerio de Salud y Protección de Colombia. Este conjunto de datos contiene 3’514.639 casos reportados de los cuales 95.437 fueron notificadas como muertes, 3’450.758 como contagios y 3’380.599 como recuperados. Se realizaron 2 experimentos en diferentes espacios de tiempo, la primera predicción se realizó a finales del año 2021 desde octubre hasta diciembre y para el siguiente espacio de tiempo, la segunda predicción tuvo lugar desde finales del mes de mayo hasta finales del mes de junio del mismo año. Se seleccionó la metodología CRISP-DM para modelar los datos. Esta consta de seis fases, que van desde la comprensión del problema que, en este trabajo, es la dinámica de propagación del virus por COVID-19 en Colombia, y su relación con las decisiones en torno a la salud pública; La comprensión de datos, que es donde se recopilaron los datos y se analizó la existencia de valores nulos o valores fuera de rango, los cuales pueden convertirse en ruido para el proceso; la fase de preparación de datos abarca las tareas generales de selección de datos a los que les aplicó la técnica de modelado. En las fases de modelado y evaluación se eligió el modelo de inteligencia artificial más apropiado para el proyecto, y se evaluó su desempeño. Finalmente, los modelos obtenidos se analizaron con los criterios de Error Porcentual Absoluto Medio y Error Cuadrático Medio, donde es posible observar que con base a la métrica MAPE la predicción se puede considerar como muy precisa, una vez que se obtiene valores por debajo de 10, así como se contempla para la predicción en los casos de muertes durante ambos espacios de tiempo, el valor de MAPE no sobrepasa el 4%.

Palabras clave: COVID-19, Deep Learning, Red Neuronal Recurrente, LSTM, Contagios, Muertes.

Capítulo 1 . GENERALIDADES

1.1. PLANTEAMIENTO DEL PROBLEMA

La información sobre contagios y muertes por Covid-19 a nivel mundial sugieren que es necesario contar con un sistema de predicción en vista de que es crucial comprender la dinámica de transmisión de la infección. Con ello se espera poder identificar si las medidas de control tomadas por los entes encargados están teniendo un efecto positivo, de lo contrario, si no se cuenta con un sistema de apoyo, el proceso de la toma de decisiones se puede tornar problemático en vista de que las estimaciones deben basarse en la situación y en los datos que se encuentran en constante cambio. Como consecuencia de esto, en el ámbito de salud, si no se cuenta con información confiable no se podrá calcular la demanda de los servicios médicos ni garantizar una buena distribución de suministros (Correa & Muñoz, 2020). Por otra parte, la economía también se vería afectada en la medida de que los casos se incrementen y que las medidas de contención como los cierres de las ciudades, los confinamientos y las restricciones en la movilidad aumenten, causando un impacto negativo al comercio nacional (PAHO, 2020).

Desde el inicio de la reciente pandemia en el año 2020, se han empleado diversas herramientas matemáticas que contribuyen en el proceso de predecir la dimensión de la misma. Estas herramientas permiten crear modelos epidemiológicos tanto determinísticos (Lenis, 2020), como estocásticos (Fedosov, Fedossova, & Buitrago, 2020). En los modelos determinísticos se consideran a los individuos como un conjunto, en cambio en los estocásticos son considerados individualmente (Manrique, 2020). Además, en un modelo determinístico se pueden controlar los factores que intervienen en el estudio, por esa razón, se pueden predecir con mayor fiabilidad los resultados. Por el contrario, en un modelo estocástico no es posible controlar los factores que intervienen y en consecuencia no produce resultados únicos, por esa razón usar los modelos determinísticos es ideal para el estudio de la pandemia. Como un ejemplo está el modelo SIR que es el modelo matemático más simple que permite describir cómo evoluciona la pandemia, descrito por William Ogilvy Kermack y Anderson Gray McKendrick en 1927 y recibe el nombre de SIR debido a que divide la población en tres categorías: susceptibles (S), infectados (I), y recuperados (R) (Montesinos, 2007).

La deficiencia que tienen estos modelos matemáticos se puede abordar desde diferentes perspectivas, una de ellas es que debido a que son modelos demasiado simplificados, se ignoran factores que tienen una gran relevancia en la propagación de la enfermedad, por ejemplo, se obvian los cambios en los comportamientos sociales y eso hace que el modelo no pueda tener una respuesta precisa en la

estimación. Otro argumento, es que los datos en los que se basa el modelo son suposiciones que la mayoría de veces no son ciertas. En particular, se puede hacer énfasis en que la población a estudiar se considera cerrada, es decir, con un régimen de aislamiento total, pero en la mayoría de ciudades esto no se desarrolló de esta manera, lo que hace que la población en realidad sea más vulnerable. Por último, está el argumento de que los individuos recuperados ya se encuentran inmunizados, pero existe la posibilidad de que presenten una reinfección haciendo que las condiciones cambien completamente el panorama (Moein, 2021).

Dadas las dificultades y limitaciones expuestas anteriormente, una alternativa viable es emplear la inteligencia artificial, también conocida como IA, que puede ser considerada como la capacidad que tiene una máquina de imitar el comportamiento humano basado en algoritmos que los programadores ordenan a las máquinas. Dentro de la inteligencia artificial está el machine learning, que se trata del autoaprendizaje basado en algoritmos, esto significa que el sistema aprende desde su experiencia. Por otro lado, está el Deep learning que es el sistema de aprendizaje profundo, es decir, este aprende de su experiencia, pero a partir de una gran base de datos o de gran cantidad de información proporcionada en la entrada (Neha, Reecha, & Jindal, 2021). Estudios previos han empleado diferentes técnicas de Deep learning para modelar los contagios y muertes por Covid-19 (Wang, 2020) (Luo, Zhang, Fu, & Rao, 2021). Las limitaciones que tienen estos estudios es que se realizaron en diferente ubicación geográfica, por lo tanto, los datos de contagios previos son diferentes, los tiempos de estudio se manejan en periodos distintos e indudablemente las medidas de prevención por cada país fueron manejadas de manera independiente y esto afecta a la propagación del virus.

En este proyecto, se busca desarrollar un sistema de estimación de parámetros, como la cantidad de contagios y muertes por COVID-19 en Colombia, de manera más precisa y en periodos de tiempo más extensos. Esto implica la necesidad de contar con conjuntos de datos de entrenamiento más amplios que los utilizados en un modelo matemático básico. Por esta razón, se requiere el empleo de estrategias avanzadas, lo que hace fundamental disponer de bases de datos sólidas con estadísticas y datos reales de contagios y hospitalizaciones. Además, es esencial tener conocimiento acerca de los aspectos epidemiológicos de la enfermedad para comprender el ritmo de propagación de la pandemia y lograr un pronóstico preciso de los contagios en el futuro (Ruifang, Xinqi, Peipei, Haiyan, & Chunxiao, 2021).

Considerando los fundamentos mencionados, se presenta la siguiente pregunta de investigación: ¿Cómo desarrollar un modelo de estimación de la cantidad de contagios y muertes en un escenario de pandemia como el causado por el COVID-19 en Colombia?

1.2. JUSTIFICACION

Las cifras que serán mencionadas a continuación pusieron en estado de alarma a toda la población mundial, lo que obligó a varios países a tomar medidas drásticas para poder frenar la rápida propagación en los contagios, y de esa forma disminuir las muertes ocasionadas por COVID-19. Debido a que, durante un largo periodo se presentan grandes variaciones en las cifras de nuevos contagios desde que fue notificado el primer caso de contagio en Wuhan el 31 de diciembre del 2019, la pandemia se ha convertido en una fuerte amenaza para la salud pública mundial. Los históricos muestran que las cifras aumentaron exponencialmente en muy poco tiempo. Los casos confirmados en todo el mundo superaron los 16 millones, de los cuales más de 650.000 resultaron mortales, estos solamente hasta el 29 de julio del 2020. La World Health Organization (WHO) presenta la situación en cifras por regiones, en África con 738.344 casos y 12.519, América con 8.840.524 casos y 342.635 muertes, mediterráneo oriental 1.507.734 casos y 38.815 muertes, Europa 3.283.277 casos y 211.616 muertes, Sudeste de Asia 1.892.056 casos y 42.233 muertes, Pacifico occidental 295.613 y 8.262 muertes. Para el 27 de diciembre del 2020 se habían registrado más de 79,2 millones de casos y más de 1,7 millones de muertes desde el inicio de la pandemia (WHO, 2020).

A nivel nacional, el 6 de marzo del 2020 fue reportado el primer caso de contagio de COVID-19 en Colombia (Minsalud, 2020), al 29 de marzo del 2020 las cifras ya habían aumentado a 702 casos, 10 muertes y solo 10 recuperados. Aunque se tomaron las medidas oportunas los casos siguieron aumentando, así que, para el 30 de junio, es decir, tan solo 3 meses desde el primer reporte, ya las cifras alcanzaban los 97.846 casos y 3.334 defunciones. Para estas fechas los casos estaban siendo atendidos en casa y las ocupaciones de camas UCI, pese a que habían aumentado, se seguían manejando cifras que no marcaban gravedad en comparación con las reportadas en el segundo semestre del año 2020; las cuales corresponden a 1.614.822 casos y 42.620 defunciones. Para este mismo periodo de tiempo, se tiene que en relación con los casos activos se puede observar que la proporción de casos en hospitalización general y UCI se incrementa con la edad, es así como del total de pacientes en hospitalización, el 72,6% son personas mayores de 50 años y de las personas en UCI el 81,4% son personas de 50 años y más (OPS, 2020).

En este sentido, se puede establecer la importancia de desarrollar un modelo que permita estimar el número de posibles casos de contagio y las posibles muertes ocasionadas por una pandemia como la del COVID-19, puesto que esto contribuye tanto en las decisiones que deben tomar los gobernantes sobre las medidas necesarias para disminuir la propagación del virus, como para una buena

preparación de insumos y personal a nivel hospitalario. Así mismo, es importante desarrollar un modelo de estimación de casos de contagios y muertes dentro de una pandemia a nivel nacional, debido a que el sistema de salud con el que se cuenta actualmente no está totalmente preparado para enfrentar estas situaciones, dado que, los virus tienen la capacidad de propagarse de diferentes formas, entre las cuales están: por contacto, por vía aérea**,** por contaminación de agua o alimentos, entre otras. Lo que hace que su propagación sea relativamente rápida. Por lo tanto, si se tiene una estimación del número de posibles pacientes contagiados, dentro de los cuales habrá pacientes que van a requerir atención en unidades de cuidados intensivos, y quienes solamente requieren atención hospitalaria, los centros de salud sabrán cómo manejar la situación con respecto a la ocupación, los medicamentos necesarios para el tratamiento y los insumos médicos que entre ellos están, los ventiladores, bombas de infusión, máquinas de hemofiltrado, entre otros. Favoreciendo incluso el caso de que se presente alguna pandemia futura (Huang, 2020; ArunKumar, 2021).

Con base en esto, se plantea el desarrollo de un sistema que permita estimar parámetros como la cantidad de contagios y muertes por COVID-19 en Colombia. Este trabajo toma relevancia, dada la realidad del sistema de salud nacional, que enfrenta limitaciones para cubrir de manera oportuna todos los casos de contagio que surgen a diario, así, resulta indispensable recurrir a herramientas que proporcionen cierto grado de apoyo y ayuda en esta tarea. (Hoz, 2021). En este caso, como ya se ha mencionado previamente, su principal propósito es brindar información pertinente para lograr una estimación cercana a la realidad.

Capítulo 2 . OBJETIVOS

Objetivo general

- Desarrollar un modelo de estimación de contagios y muertes en un escenario de pandemia como la ocasionada por el COVID-19.

Objetivo especifico

- Definir un dataset para el entrenamiento de un modelo estimador de la cantidad de contagios y muertes por COVID-19 en Colombia.

- Proponer un modelo basado en redes neuronales recurrentes que permita estimar el número de contagios y muertes por COVID-19 en Colombia.

- Validar el sistema de estimación sobre la cantidad de contagios y muertes ocasionadas por COVID-19 en Colombia.

Capítulo 3 MARCO TEORICO

A seguir, se proporcionará una contextualización al trabajo mediante la definición de términos clave, lo cual es importante para comprender los conceptos fundamentales y las bases teóricas que sustentan la investigación.

3.1 CORONAVIRUS (COVID 19)

La enfermedad por coronavirus (COVID-19) es una enfermedad infecciosa causada por el virus SARS-CoV-2. Las personas que contraen el virus experimentan una enfermedad respiratoria que va de leve a moderada y la recuperación no requiere tratamiento especial. Sin embargo, algunos casos requieren atención médica. El virus se propaga desde la boca o nariz de una persona infectada en pequeñas partículas líquidas liberadas cuando tose, estornuda, habla, canta o respira. Estas partículas van desde gotículas respiratorias más grandes hasta los aerosoles más pequeños. Cualquier persona puede contraer COVID-19 y enfermar gravemente o morir a cualquier edad (WHO, 2023).

3.2. SINTOMAS

Los signos y síntomas de la enfermedad por coronavirus 2019 (COVID-19) pueden aparecer entre dos y catorce días después de la exposición al virus. El virus se puede transmitir incluso antes de la aparición de los síntomas, ese período entre la exposición y la aparición de estos se llama período de incubación. Entre los síntomas más habituales, se pueden incluir los siguientes: fiebre, tos, cansancio, falta de aire o dificultad para respirar, dolores musculares, escalofríos, dolor de garganta, dolor de cabeza, náuseas, diarrea, además de la pérdida del sentido del gusto o del olfato. Algunas personas pueden no tener síntomas en absoluto, pero aun así pueden contagiar a los demás (Clinic, 2023).

3.3. PREVENCIÓN

Para prevenir la infección y frenar la transmisión de la COVID-19 se pueden tener en cuenta las recomendaciones dadas por la Organización Mundial de la Salud, en las cuales esta, mantenerse al menos a un metro de distancia de los demás, utilizar una mascarilla bien ajustada cuando no sea posible el distanciamiento físico o cuando se encuentre en lugares cerrados, lavarse las manos regularmente con agua y jabón o usar un desinfectante de manos a base de alcohol, cubrirse la boca y la nariz al toser o estornudar. Desde diciembre del 2020 recibir la vacuna contra la COVID-19 entró a ser parte de las recomendaciones para reducir el riesgo de contraer y trasmitir el virus, Las vacunas contra la COVID-19 inducen inmunidad contra el virus SARS-Cov-2, es decir, reducen el riesgo de que de este cause síntomas y tenga consecuencias para la salud (WHO, 2023).

3.4. INTELIGENCIA ARTIFICIAL

La inteligencia Artificial, es una rama multidisciplinar que involucra campos como las ciencias de la computación y de la información, la lógica, la matemática, la estadística, la biología, la psicología, la filosofía, la lingüística y otras áreas. En sinergia con tecnologías avanzadas, busca que los equipos informáticos y distintos dispositivos tecnológicos, realicen tareas que normalmente requerirían inteligencia humana, como, por ejemplo, las capacidades de aprender, razonar, resolver problemas, la percepción visual, el reconocimiento de voz, la toma decisiones y la traducción de idiomas. En conclusión, la IA involucra los sistemas que tienen la capacidad de imitar el comportamiento humano inteligente (García F. R., 2020).

3.4.1. MACHINE LEARNING O APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es un componente importante en la ciencia de datos. Mediante el uso de métodos estadísticos, los algoritmos se entrenan para hacer clasificaciones o predicciones y descubrir información clave en proyectos de minería de datos (IBM, 2023). Por tanto, los programadores han perfeccionado cada vez más la capacidad de las máquinas para estudiar datos. Hoy en día, las aplicaciones modernas de la inteligencia artificial ya han proporcionado coches que se conducen solos y asistentes virtuales, y ha ayudado a detectar fraudes y a gestionar recursos como la electricidad de forma más eficiente (Economist, 2023).

3.4.2. DEEP LEARNING

El aprendizaje profundo es un subcampo del aprendizaje automático que se centra en el desarrollo y entrenamiento de redes neuronales artificiales, en particular redes neuronales profundas donde se pretende imitar la estructura y el funcionamiento de las redes neuronales del cerebro humano en la resolución de problemas complejos. los algoritmos de aprendizaje profundo se utilizan para aprender automáticamente y extraer características de grandes cantidades de datos, lo que les permite hacer predicciones, reconocer patrones y realizar diversas tareas. Adicionalmente cada modelo creado debe cumplir estrictamente dos tareas principales del aprendizaje profundo: extraer información de la entrada y producir una salida con sentido (Mijwil, 2022). Dentro de las características claves del aprendizaje profundo esta, que los algoritmos aprenden automáticamente representaciones jerárquicas de los datos, lo que significa que pueden descubrir y extraer características relevantes a partir de datos de entrada sin procesar. Por otro lado, los modelos de aprendizaje profundo requieren grandes cantidades de datos para el entrenamiento, con el propósito de generalizar bien nuevos datos no vistos.

3.4.3. REDES NEURONALES

Una red neuronal es un modelo computacional inspirado en la estructura y el funcionamiento de las redes neuronales del cerebro humano (Picton, 1994). Se utiliza en el aprendizaje automático y la inteligencia artificial para procesar y analizar datos, hacer predicciones y resolver tareas complejas. Las redes neuronales están formadas por nodos interconectados, también conocidos como neuronas o neuronas artificiales, organizados en capas las cuales se compone de, una capa de entrada, una capa de procesamiento o capas ocultas y una capa de salida (Irene Casas, 2020). La capa de entrada recibe los datos brutos o características que procesará la red neuronal, las capas ocultas realizan cálculos sobre los datos de entrada, transformándolos en una representación más útil para la tarea y finalmente llevando la información a la capa de salida. Además, cuenta con otros componentes como pesos y sesgos que, durante el entrenamiento, estos pesos y sesgos pueden ser ajustados para optimizar el rendimiento de la red. Las neuronas aplican una función de activación a la suma de sus entradas para producir su salida, las funciones de activación más comunes son la sigmoidea, la ReLU (unidad lineal rectificada) y la tanh (tangente hiperbólica) (Aggarwal, 2018). Por último, las redes neuronales necesitan un conjunto de datos para entrenarse pues la red aprende a hacer predicciones precisas ajustando sus pesos y sesgos en función de los patrones brindados por los datos de entrenamiento.

3.4.4. RED NEURONAL RECURRENTE

Las redes neuronales recurrentes (RNN) son un tipo de redes neuronales artificiales diseñada para procesar secuencias de datos y cuentan con memorias capaces de captar información con el propósito de almacenarla. A diferencia de las redes neuronales feedforward, las RNN tienen conexiones que hacen un bucle sobre sí mismas, lo que les permite mantener un estado oculto que representa la información de pasos de tiempo anteriores en una secuencia, es decir, las redes neuronales recurrentes son lo suficientemente potentes como para aprovechar la información de una secuencia relativamente larga, ya que toman información de entradas anteriores para influir en la entrada y salida actual, haciendo que el resultado dependa de entradas anteriores y tomar decisiones basadas tanto en la información actual como en la pasada (Runjie Zhu, 2020). Dependiendo de la tarea específica que se le haya asignado las RNN pueden producir una salida en cada paso temporal o generar una única salida al final de la secuencia. Por ejemplo, en el modelado del lenguaje o en el análisis de un sentimiento al final de analizar una frase (Aggarwal, 2018). A pesar de su eficacia en el manejo de datos secuenciales, las RNN tradicionales tienen limitaciones y esto ha llevado al desarrollo de arquitecturas de RNN más avanzadas, como las redes de memoria a largo plazo (LSTM) y las redes de unidades recurrentes con compuerta (GRU).

3.4.5. REDES DE MEMORIA A LARGO PLAZO Y CORTO PLAZO - LSTM

Las redes LSTM, del inglés Long–Short Term Memory, son un tipo de arquitectura de red neuronal recurrente (RNN) que fueron inventadas para resolver el problema de la dependencia a largo plazo y son usadas en actividades como clasificar, procesar y hacer predicciones basadas en datos de series temporales, es conveniente usar las redes LSTM ya que están diseñadas para resolver el problema del gradiente de fuga y capturar dependencias de largo alcance en datos secuenciales (Luo, Zhang, Fu, & Rao, 2021). Esta red incorpora partes funcionales internas que le permiten controlar y manipular la información a lo largo del tiempo que a su vez la hace capaz de añadir o quitar información del estado de la celda.

Las partes funcionales internas clave de un modelo LSTM son: El estado de la célula, que es el componente primario de una LSTM, sirviendo como una especie de cinta transportadora para que la información fluya a través de la red, el estado de la célula es un vector que se actualiza en cada paso temporal, pero que está controlado cuidadosamente para retener o descartar información. El estado oculto el cual funciona como la memoria de la LSTM, capturando la información relevante de los pasos de tiempo anteriores y la entrada actual. Este estado oculto se utiliza comúnmente para hacer predicciones o tomar decisiones. Al mismo tiempo que tres puertas, puerta del olvido, puerta de entrada y puerta de salida, en donde la puerta del olvido determina qué información del estado anterior de la célula debe descartarse y qué información debe conservarse. La puerta de entrada decide qué nueva información debe añadirse al estado de la célula y finalmente la puerta de salida controla qué partes del estado de la célula se utilizarán para producir el estado oculto para el paso de tiempo actual (Ullán, González, & González, 2022). La arquitectura LSTM está diseñada para una amplia gama de tareas de datos secuenciales, como el procesamiento del lenguaje natural, el análisis de series temporales y el reconocimiento del habla.

La figura a seguir presenta de forma esquemática una red neuronal, la red neuronal recurrente y la red LSTM.

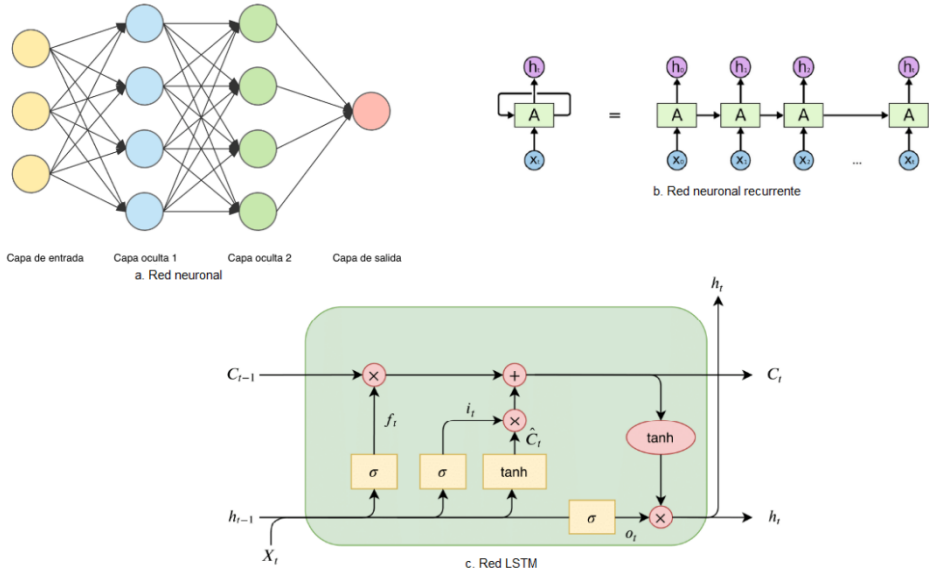


Figura 1. Redes: a.Red Neuronal, b.Red Neuronal Recurrente, c.Red LSTM

3.5. MÉTRICAS DE DESEMPEÑO

En el ámbito de las redes neuronales, las métricas de desempeño son herramientas esenciales para evaluar, ajustar y comunicar la calidad de las predicciones, ayudan a tomar decisiones informadas sobre la selección del modelo, el ajuste de hiperparámetros y su optimización, lo que en última instancia conduce a predicciones más precisas y fiables. La elección de las métricas adecuadas depende de las características específicas de los datos y de los objetivos del análisis, debido a ello, las métricas usadas en este trabajo son aplicadas para problemas de regresión, los cuales son, un tipo de problema de aprendizaje automático supervisado en el que su objetivo es predecir un valor numérico continuo, a diferencia de los problemas de clasificación, que consisten en asignar datos a categorías o etiquetas. A continuación, se presentan las métricas de rendimiento usadas en este estudio: Error Absoluto Medio (MAE), Error Cuadrático Medio (RMSE) y Error Porcentual Absoluto Medio (MAPE). Estas métricas proporcionan una medida cuantitativa sobre qué tan cerca se encuentran las predicciones de la red de los valores reales, permitiendo así una evaluación objetiva y una optimización del rendimiento del modelo.

3.5.1. ERROR ABSOLUTO MEDIO (MAE)

Es el valor promedio de la diferencia absoluta entre los valores de la estimación real y los valores predichos. Donde n es el número de observaciones, y $y_t - y_{\boldsymbol{forecast}(t)}$ es el error entre el valor real y el valor de la predicción.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - y_{\boldsymbol{forecast}(t)}\right| \qquad \textit{Ecuación 1}$$

3.5.2. ERROR CUADRATICO MEDIO (RMSE)

Es una medida frecuentemente utilizada de las diferencias entre los valores predichos por un modelo o un estimado y los valores observados. Es la raíz cuadrada del error cuadrático medio.

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(y_t - y_{forecast(t)})^2}{n}} \quad \textit{Ecuación 2}$$

3.5.3. ERROR PORCENTUAL ABSOLUTO MEDIO (MAPE)

El MAPE cuantifica la precisión como un porcentaje que se puede calcular como un error porcentual acumulado para cada marco temporal. Es decir, representa el error medio en términos porcentuales.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_t - y_{forecast(t)}}{y_t} \right| \quad \textit{Ecuación 3}$$

Capítulo 4 . MARCO ANTECEDENTES

En un enfoque general, se pueden encontrar estudios donde se ha utilizado la inteligencia artificial para crear modelos de predicción de los casos de Covid-19 alrededor del mundo. Por ejemplo, basándose en Redes Neuronales Recurrentes (RNN), ArunKumar, et al proponen RNN de aprendizaje profundo para predecir los casos confirmados acumulados para 10 países que estuvieron muy afectados por el COVID-19, como son, EE.UU., Brasil, India, Rusia, Sudáfrica, México, Perú, Chile, Reino Unido e Irán. Durante el estudio se tienen en cuenta factores como la edad, las instalaciones sanitarias, las medidas tomadas por los dirigentes y demás factores que afectan en la propagación de la pandemia de COVID-19 (ArunKumar, 2021).

Cada vez se han aplicado diferentes métodos de aprendizaje profundo para predecir y estimar el número de casos de COVID-19 para contagios, muertes y recuperados, entre ellos, Shastri propone crear modelos basados en aprendizaje profundo para la predicción de casos confirmados y mortales en India y EE. UU (Shastri, 2020). Para diseñar los modelos propuestos, se utilizan variantes de memorias a corto plazo (LSTM) basadas en redes neuronales recurrentes (RNN). Para obtener los resultados pronosticados se utilizan LSTM apiladas, LSTM bidireccionales y LSTM convolucionales. Con el propósito de seleccionar el mejor modelo se compara el Error Medio Porcentual Absoluto (MAPE) de cada modelo. En el cual el modelo LSTM convolucional obtiene mejores resultados que los otros dos modelos, con una tasa de error que oscila entre el 2,0% y el 3,3% para los conjuntos de datos. Asimismo, en la investigación realizada por Parbat y Chakraborty (Parbat & Chakraborty, 2020) usaron un modelo de regresión de vectores de soporte con un dataset que recopila datos desde el 1 de marzo de 2020 al 30 de abril de 2020 y ese modelo obtuvo una precisión aproximada del 97% en la predicción de muertes, recuperados, número acumulado de casos confirmados y una precisión del 87% en la predicción de nuevos casos diarios.

En un enfoque especifico, se describen estudios correspondientes a las predicciones para los casos positivos de Covid-19 ejecutadas particularmente con redes Long–Short Term Memory. Entre ellos (Gautam, 2020) que implementa el Aprendizaje por Transferencia para la red LSTM, con el fin de aprender el comportamiento de una región y después facilitar el pronóstico para una región completamente diferente. En el entrenamiento se utiliza la API Keras con el backend TensorFlow siendo que, para otras evaluaciones se han utilizado varias librerías de Python. Por otra parte, las redes se entrenaron en la GPU disponible en el entorno de Google Colab. Gautam principalmente utiliza dos modelos basados en casos de Italia y Estados Unidos. Para el entrenamiento, toma un periodo de 4 meses

aproximadamente, donde se garantiza un periodo activo tanto de propagación como de muertes. Con respecto a métricas de desempeño, el modelo italiano tiene un MAE de 0,46 y un RMSE de 0,69, frente al MAE y el RMSE de 0,49 y 0,7, respectivamente, del modelo estadounidense.

En (Luo, Zhang, Fu, & Rao, 2021) se establecen modelos de predicción para los casos diarios de contagios en América, empleando los algoritmos de memoria a largo plazo (LSTM) y refuerzo de gradiente extremo (XGBoost). Con la intención de evaluar el rendimiento del modelo se emplearon cuatro métricas como MAE, MSE, RMSE y MAPE. En esta investigación el número de casos diarios confirmados se recogen del sitio web de la OMS, donde los datos están disponibles en un formato de serie temporal con fecha, mes y año para garantizar el componente temporal. Esos datos son de un rango de 6 meses, desde el 1 de abril de 2020 hasta el 30 de septiembre de 2020. El modelo propuesto se desarrolla con librerías de código abierto como Numpy, Pandas y Keras. La conclusión de las pruebas realizadas en la investigación muestra que el MAPE de los algoritmos LSTM y XGBoost alcanza el 2,32% y el 7,21%, respectivamente.

En (Ruifang, Xinqi, Peipei, Haiyan, & Chunxiao, 2021), se han aplicado los modelos LSTM y LSTM-Markov. El modelo de Markov se empleó para reducir el error de predicción en el número acumulado de casos confirmados del modelo LSTM, por lo tanto, para evitar el sobreajuste se usa el optimizador ADAM con un dropout en 0,02, la capa oculta en 1 y el número de nodos en la capa oculta es 4. Las estadísticas se tomaron de un estudio de la Universidad John Hopkins, donde se incluyen los cuatro países con más casos confirmados en el mundo: Estados Unidos, Gran Bretaña, Brasil y Rusia, con rango de 9 meses incluyendo fechas desde el 1 de marzo de 2020 al 31 de diciembre de 2020. Los experimentos se ejecutan en bibliotecas de código abierto como NumPy, Pandas y TensorFlow. Para verificar la eficacia del método propuesto se comparó el número acumulado de casos infectados predichos por los dos modelos y la RMSE del modelo LSTM-Markov es casi el 40% en relación al modelo LSTM, lo que demuestra que la precisión de la predicción mejora considerablemente con el modelo LSTM-Markov.

En diferentes estudios se obtuvieron importantes avances como, la obtención de modelos con buena capacidad de generalización con el fin de ser aplicados en otros países en la medida en que compartan distribuciones similares. Por ejemplo, a nivel de Latinoamérica en países como Brasil, Colombia, Ecuador se lograron diferentes estudios vinculados con la predicción de casos y muertes por Covid-19 que tenían como finalidad hacer predicciones efectivas que permitan ayudar en la toma de decisiones a las autoridades de la salud. Con este fin, se utilizaron diferentes modelos según los criterios de cada uno de los autores, entre ellos están: Modelo Autorregresivo AR, modelo de Medias Móviles MA y el modelo Media Móvil Integrada Autorregresiva Estacional con Regresores Exógenos (SARIMAX), Memoria a largo plazo para entrenamiento de datos-SAE (LSTM-SAE), Modelo de

predicción BROWN, Modelo LSTM-Markov (García & Daza, 2021; Ruifang, Xinqi, Peipei, Haiyan, & Chunxiao, 2021; Diaz, 2020; Gadelha, 2020). En particular en Colombia se presentó un pronóstico haciendo uso del modelo lineal de Brown, donde se utilizó la base de datos de las personas contagiadas entre el periodo del 6 de marzo del 2020 hasta el 10 de mayo del mismo año, si bien no se manejó el mismo modelo es un avance indispensable para las predicciones en el territorio nacional. Hasta el momento haciendo uso de técnicas de aprendizaje profundo se han desarrollado eficazmente modelos de predicción de contagios y muertes por COVID-19 en países como Estados Unidos, Gran Bretaña, Brasil y Rusia, aun así, para Colombia todavía no existen suficientes estudios que muestren que las técnicas de aprendizaje profundo permitan generar modelos de estimación de contagios y muertes debido a una situación de pandemia como la causada por el COVID-19.

En general, diversos estudios se han desarrollado para sistemas de apoyo con el fin de estimar contagios y muertes en un escenario de pandemia como la ocasionada por el COVID-19. La tabla 1, presenta algunos desarrollos que sirvieron como base para el desarrollo del proyecto.

Tabla 1. Características generales de los antecedentes

Título Documento	Referencias	Características importantes	Elementos relacionados con el proyecto
Forecasting of COVID-19 using deep layer Recurrent Neural Networks with Gated Recurrent Units and Long Short-Term Memory cells.	(ArunKumar, Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells, 2021)	-Estos modelos se desarrollaron en la biblioteca de aprendizaje PyTorchdeep basada en Python y las simulaciones se realizaron Google COLAB - Para casos confirmados, el modelo LSTM funciona mejor para países como EE. UU., Brasil, Sudáfrica, Perú, Chile e Irán, por otro lado, el modelo GRU funciona mejor para India, Rusia, México y el Reino Unido.	-Los datos fueron recuperados de la base de datos COVID-19 de la Universidad John Hopkins -Datos acumulados de aproximadamente 6 meses, para una predicción de 60 días.
Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study	(Shastri, 2020)	- Los datos sin procesar se normalizaron con MinMaxScaler. - Los experimentos se llevan a cabo en Google Colaboratory usando Python 3.0 con bibliotecas de código abierto como Tensorflow, Pandas, Numpy y keras.	-Los datos se tomaron del Ministerio de Salud y Bienestar Familiar, Gobierno de India y Centros para el Control y la Prevención de Enfermedades, Departamento de Salud y Servicios Humanos de EE. UU. -Para los casos confirmados usaron datos de 5 meses y para casos de muertes aproximadamente 4 meses.
The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method.	(Ruifang, Xinqi, Peipei, Haiyan, & Chunxiao, 2021)	-Los experimentos se realizan en bibliotecas de código abierto como NumPy, Pandas y TensorFlow. -El modelo LSTM-Markov puede mejorar la exactitud y precisión de la predicción de tendencias a mediano y largo plazo de COVID-19.	-Las estadísticas utilizadas en este estudio fueron recopiladas por la Universidad John Hopkins -El estudio se realizó en cuatro países: Estados Unidos, Gran Bretaña, Brasil y Rusia -Los datos corresponden a 9 meses
A python based support vector regression model for prediction of COVID19 cases in India	(Parbat & Chakraborty, 2020)	-Se utiliza un modelo de regresión de vectores de soporte para predecir -El modelo trabaja con los datos de las muertes, recuperados y confirmados, tanto casos diarios como acumulado. -Tiene una precisión aproximada del 97% en la predicción de muertes, recuperados y casos confirmados acumulados y una precisión del 87% en la predicción de nuevos casos diarios.	-El modelo ha sido desarrollado en Python 3.6.3 -Los parámetros de rendimiento del modelo se calculan con MSE, RMSE, puntuación de regresión y precisión porcentual

Transfer Learning for COVID-19 cases and deaths forecast using LSTM network	(Gautam, 2020)	-Las pruebas se realizaron con dos modelos, uno basado en datos de Italia y el otro con datos de estados unidos. -Según la investigación, el total de datos es aproximado a 4 meses.	-Se utilizaron varias bibliotecas de Python. -Las redes fueron entrenadas en GPU disponible en el ambiente de Google Colab. -Los parámetros de rendimiento fueron MAE Y RMSE con valores de 0.46 y 0.69 respectivamente.
Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms	(Luo, Zhang, Fu, & Rao, 2021)	-Se utilizó un modelo LSTM con y aumento de gradiante extremo (XGBoost). -Para el análisis de la serie de tiempo se tomaron datos desde el 3 de enero hasta 30 de septiembre del 2020	-El modelo propuesto se desarrolla con LSTM y XGBoost que se realizan con bibliotecas de código abierto como Numpy, Pandas y Keras. -Se emplearon los parámetros más comunes para evaluar el desempeño de la predicción como MAE, MSE, RMSE y MAPE. -Según el resultado del parámetro MAPE el modelo alcanza un valor de 2.32%

Capítulo 5 . METODOLOGIA

En esta sección se presenta el proceso metodológico para este trabajo de investigación. Así, se aplicará la metodología CRISP-DM por sus siglas Cross Industry Standard Process for Data Mining, la cual es una metodología exhaustiva y ampliamente aceptada para llevar a cabo proyectos de minería de datos. CRISP-DM incluye una guía articulada en seis fases: comprensión del problema, comprensión de datos, preparación de datos, modelado, evaluación del modelo e implementación del mismo. La secuencia de las fases es flexible, es decir, los proyectos avanzan y retroceden entre fases si es necesario (Núñez, 2021) . En este estudio la metodología se ajustó a un escenario de pandemia, específicamente enfocado en el caso del COVID-19 en Colombia, es decir, cada una de las fases cumple sus objetivos y tareas específicas dentro del marco de los datos reportados en pandemia para obtener resultados precisos y actualizados en la evolución de la pandemia del COVID-19 en Colombia, ver figura 2.

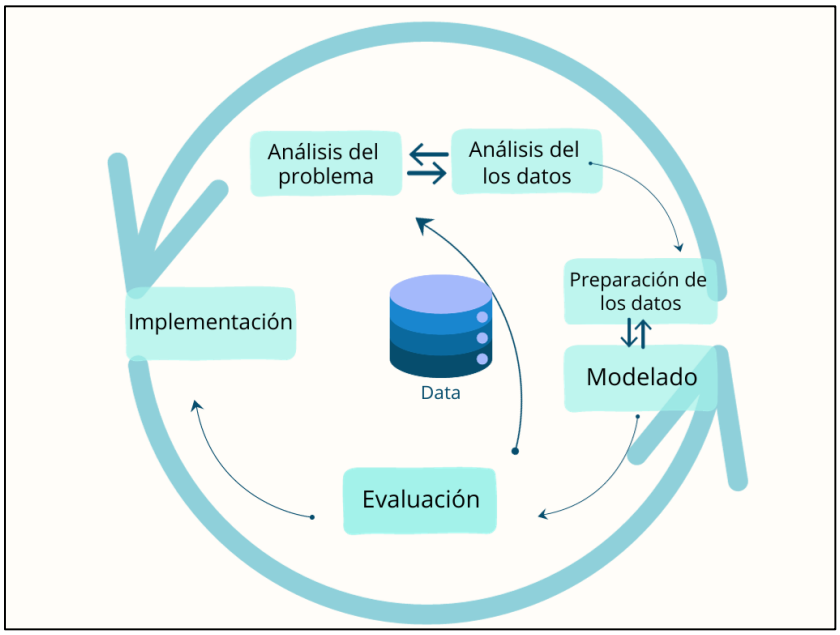


Figura 2. Diagrama de la Metodología utilizada. Fuente propia.

5.1. DATASET

El conjunto de datos utilizado se obtuvo del sitio web del Ministerio de Salud y Protección de Colombia, los datos cuentan con información como la fecha de diagnóstico, la fecha de reporte en la web, departamento, sexo, entre otros. Este conjunto de datos contiene 3’514.639 casos reportados de los cuales 95.437 fueron

notificadas como muertes, 3’450.758 como contagios y 3’380.599 como recuperados. El dataset se dividió 80% para entrenamiento, 20% para test.

5.2. PROCESAMIENTO DE LOS DATOS

La limpieza de datos, también conocida como preprocesamiento de datos, es un paso crucial en el proceso de aprendizaje automático, ya que garantiza que los datos utilizados para el entrenamiento y las pruebas sean precisos, coherentes y fiables, debido a que la calidad de los datos influye directamente en el rendimiento. Las normas y técnicas específicas de limpieza de datos varían en función del tipo de datos y la naturaleza del problema, sin embargo, hay varios pasos comunes que se deben seguir como:

a. Identificar los datos que faltan y hacer su respectivo tratamiento, que dependiendo de la naturaleza de los datos se puede optar por asignar unos nuevos datos o excluir los puntos que faltan.
b. Ubicar valores atípicos, los cuales pueden eliminarse, transformarse o sustituirse por valores más representativos para minimizar su impacto en el modelo.
c. Escalar las características numéricas, considerando estandarizarlas o normalizarlas para asegurar que las características tienen las mismas unidades o escalas. Entre los métodos de escalado habituales se incluyen el escalado Min-Max o la estandarización Z-score.
d. Dividir los datos en conjuntos de entrenamiento y de prueba para evaluar el rendimiento del modelo en datos no vistos.

En esta sección, se realizó la organización y preparación de los datos para obtener un conjunto más homogéneo y reducido que el original, debido a que en el dataset original había información de algunas fechas por fuera del año 2021, además de 21 columnas donde se incluían datos que no iban a ser útiles para este estudio, por ejemplo, nombre del departamento, edad, sexo, pertenencia étnica, entre otros (ver figura 3). Eliminar esos datos fue necesario dado que solo se trabajaría con dos grupos específicos: contagios y muertes. Para los cuales únicamente se necesitaba, la fecha de reporte web y la fecha de contagio o muerte según correspondiera. Para lograr la reducción, se estandarizó el formato de fecha (año/mes/día) en todos los datos. Además, se procedió a sustituir los valores nulos en el dataset y, de ser necesario, se eliminaron aquellos que no aportaban información útil. La figura 4 muestra una sección de los datos filtrados a partir del dataset original.

```
#*
datos1 =datos.reset_index(drop=True)
datos1['fecha reporte web'] = pd.to_datetime(datos1['fecha reporte web'], format = '%Y-%m-%d')
datos1.head(15)
```

fecha reporte web	ID de caso	Código DIVIPOLA departamento	Nombre departamento	Código DIVIPOLA municipio	Nombre municipio	Edad	Unidad de medida de edad	Sexo	Tipo de contagio	...	Código ISO del país	Nombre del país	Recuperado	Fecha de inicio de síntomas	Fecha de muerte	Fecha de diagnóstico	Fecha de recuperación	Tipo de recuperación	Pertenencia étnica	Nombre del grupo étnico
2021-03-04	2,265,685	11	BOGOTA	11,001	BOGOTA	49	1	M	Comunitaria	...	NaN	NaN	Recuperado	2021-02-25 00:00:00	NaN	2021-03-03 00:00:00	2021-03-16 00:00:00	Tiempo	6	NaN
2021-03-04	2,265,686	11	BOGOTA	11,001	BOGOTA	49	1	M	Relacionado	...	NaN	NaN	Recuperado	2021-02-23 00:00:00	NaN	2021-03-03 00:00:00	2021-03-16 00:00:00	Tiempo	6	NaN
2021-03-04	2,265,687	11	BOGOTA	11,001	BOGOTA	51	1	F	Comunitaria	...	NaN	NaN	Recuperado	2021-02-24 00:00:00	NaN	2021-03-03 00:00:00	2021-03-16 00:00:00	Tiempo	6	NaN
2021-03-04	2,265,688	11	BOGOTA	11,001	BOGOTA	51	1	F	Relacionado	...	NaN	NaN	Recuperado	2021-02-27 00:00:00	NaN	2021-03-03 00:00:00	2021-03-13 00:00:00	Tiempo	6	NaN
2021-03-04	2,265,689	11	BOGOTA	11,001	BOGOTA	51	1	F	Comunitaria	...	NaN	NaN	Recuperado	2021-03-01 00:00:00	NaN	2021-03-03 00:00:00	2021-03-16 00:00:00	Tiempo	6	NaN
2021-03-04	2,265,690	11	BOGOTA	11,001	BOGOTA	52	1	F	Relacionado	...	NaN	NaN	Recuperado	2021-02-26 00:00:00	NaN	2021-03-03 00:00:00	2021-03-16 00:00:00	Tiempo	6	NaN
2021-01-15	1,851,419	52	NARIÑO	52,227	CUMBAL	24	1	M	Relacionado	...	NaN	NaN	Recuperado	2021-01-08 00:00:00	NaN	2021-01-14 00:00:00	2021-01-26 00:00:00	Tiempo	1	PASTO
2021-01-15	1,851,420	52	NARIÑO	52,227	CUMBAL	24	1	M	Comunitaria	...	NaN	NaN	Recuperado	2021-01-08 00:00:00	NaN	2021-01-13 00:00:00	2021-01-26 00:00:00	Tiempo	1	Por definir

Figura 3. Dataset original. Fuente propia.

```
#CONTEO MUERTES POR DIA
mxfecha_filtrado=m_filtrado.groupby(['Fecha de muerte']).count()
print(mxfecha_filtrado)
```

Fecha de muerte	fecha reporte web
2021-01-01	107
2021-01-02	110
2021-01-03	122
2021-01-04	153
2021-01-05	158
...	...
2021-12-27	78
2021-12-28	79
2021-12-29	66
2021-12-30	78
2021-12-31	56

[365 rows x 1 columns]

```
#CONTEO CONTAGIOS POR DIA
cxfecha_filtrado=c_filtrado.groupby(['Fecha de diagnóstico']).count()
print(cxfecha_filtrado)
```

Fecha de diagnóstico	fecha reporte web
2021-01-01	5038
2021-01-02	11720
2021-01-03	10809
2021-01-04	16688
2021-01-05	22234
...	...
2021-12-27	5655
2021-12-28	7150
2021-12-29	8155
2021-12-30	3791
2021-12-31	88

[365 rows x 1 columns]

Figura 4. Datos filtrados de muertes y contagios diarios. Fuente propia

5.3. ARQUITECTURA LSTM

Con base en la literatura, en este estudio se utilizó un modelo Long Short Term Memory (LSTM), dado que es considerada una mejora de la Redes Neuronales Recurrentes, aportando un estado de celda (cell state) para mantener el estado a lo largo del tiempo. Adicionalmente la LSTM contiene un procesador que determina si la información que le está llegando es útil o no. La célula, como es comúnmente llamada, cuenta con tres componentes, puerta de entrada (input gate), puerta de olvido (forget gate) y puerta de salida (output gate). Esas puertas son las encargadas de determinar si hay una nueva entrada, de eliminar selectivamente la información en el caso de que no sea importante o dejar información que alterara la salida (Gautam, 2020). El modelo estructural de esta arquitectura se muestra en la Figura 5.

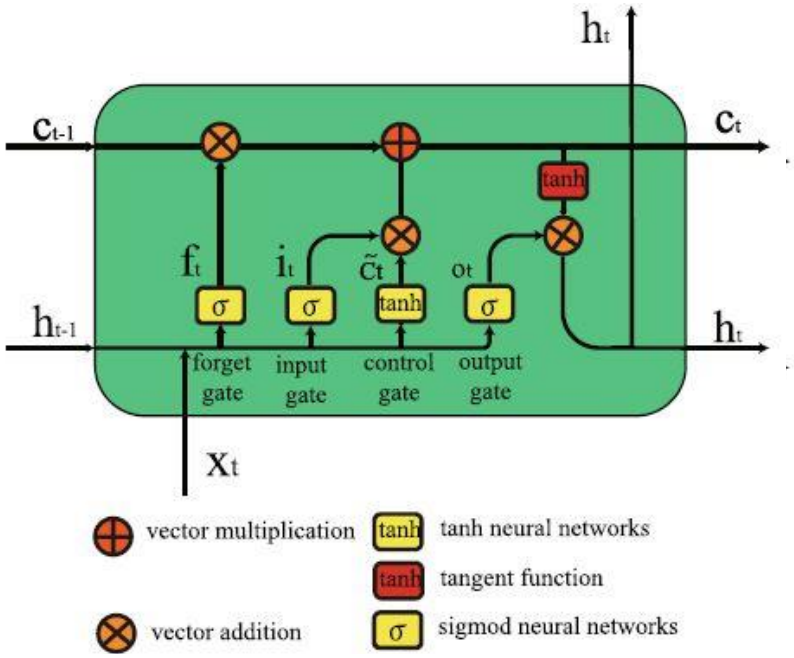


Figura 5. Arquitectura de la red LSTM. Fuente (Luo, 2021)

5.4. ENTRENAMIENTO DEL MODELO

El entrenamiento del modelo se ejecutó en la herramienta de Google, Google Colab, que es un desarrollador de programas Python donde se pueden ejecutar códigos usando sencillamente un navegador web. Colab es una versión alojada en la nube de Jupyter Notebook, adicionalmente brinda acceso gratuito a la infraestructura informática como almacenamiento, memoria, capacidad de procesamiento, unidades de procesamiento de gráficos (GPU) y unidades de procesamiento de tensor (TPU). El modelo se entrenó utilizando los datos reportados por el Ministerio de Salud en el año 2021 para contagios y muertes. De aquí, se empleó el 80% de los datos para el entrenamiento y el 20% para hacer las pruebas.

5.5. METRICAS DE DESEMPEÑO

Las métricas que se usaron fueron las más comunes para evaluar los sistemas de regresión. El error absoluto medio (MAE), que proporciona un error absoluto medio entre el valor previsto y el valor real, el error cuadrático medio, (RMSE) es una medida de diferencia entre los valores predichos y los valores observados y por último el error porcentual absoluto medio (MAPE), que es el que cuantifica la precisión en forma de porcentaje que puede determinarse como un error porcentual absoluto acumulado para cada tiempo. Las ecuaciones 4 a 6, muestran la forma matemática en que estas métricas son determinadas.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |datosTest - datosPredicción|$$

Ecuación 4. Error absoluto medio para la predicción de Muertes y Contagios.

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(datosTest - datosPredicción)^2}{n}}$$

Ecuación 5.Error cuadrático medio para la predicción de Muertes y Contagios.

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n} \left|\frac{datosTest - datosPrediccion}{datosTest}\right|$$

Ecuación 6.Error porcentual absoluto medio para la predicción de Muertes y Contagios.

5.6. UMBRALES DE SELECCIÓN

Para evaluar el rendimiento del modelo utilizado, se analizaron los resultados de predicción según los criterios de Error Porcentual Absoluto Medio (MAPE) y Error Cuadrático Medio (RMSE). Según (Hansun & Young, 2021) la interpretación de los valores MAPE puede ir desde valores menores a 10 para una predicción muy precisa, cifras entre 10-20 corresponden a una buena predicción, entre 20-50 reflejan una predicción razonable y finalmente las cifras mayores a 50 serían predicciones inexactas. Por lo tanto, siempre se busca que los valores sean lo más pequeños posibles. Sin embargo, al usar la métrica RMSE se pueden obtener valores inusuales debido a que esta métrica es más sensible a valores atípicos ya que se está empleando una potencia. No obstante, los valores no solo dependen de que tan exacto pueda ser el modelo y sus resultados, sino también, de la magnitud de los datos, en este caso la cantidad de casos reportados por día. En la tabla 2 se muestran los valores de MAPE asociados a lo que se determina como una buena o mala predicción según corresponda.

Tabla 2. Interpretación rangos de los criterios MAPE y RMSE. Fuente propia

Rangos	Interpretación
<10	Predicción muy precisa
10-20	Buena predicción
20-50	Predicción razonable
>50	Predicción inexacta

Capítulo 6 RESULTADOS

6.1. PREPROCESAMIENTO DE LOS DATOS

En este trabajo, se utilizaron los datos que están reportados en el sitio web del Ministerio de Salud y Protección de Colombia para el periodo de enero a diciembre de 2021 y se pudieron descargar en formato CSV. Los datos se dividieron en dos grupos, contagios y muertes. Cada conjunto tiene la “fecha de reporte web” y la “fecha de contagio” o la “fecha de muerte” según corresponda. A continuación, los datos descargados se pre procesaron para sustituir los valores nulos o eliminar datos que no aporten información útil.

A seguir, se filtraron los datos haciendo uso de la función Rolling de Pandas para calcular la media móvil. Para este estudio se crea la ventana cada 10 días y así se logra reducir algunos picos que se pueden presentar en las gráficas. En las figuras 6 y 7 se observa el cambio que presentan las gráficas de cantidad de muertes y contagios en función de la fecha, usando la función Rolling.

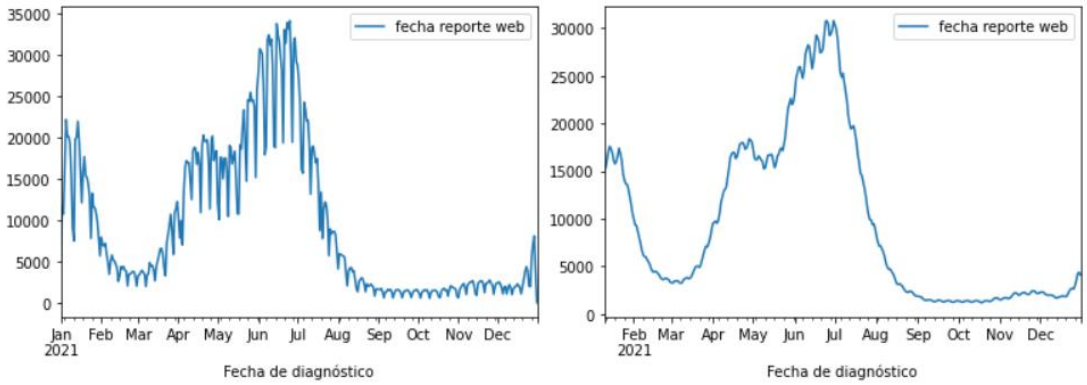


Figura 6. Graficas contagios reportados antes y después de usar la función Rolling. Fuente propia.

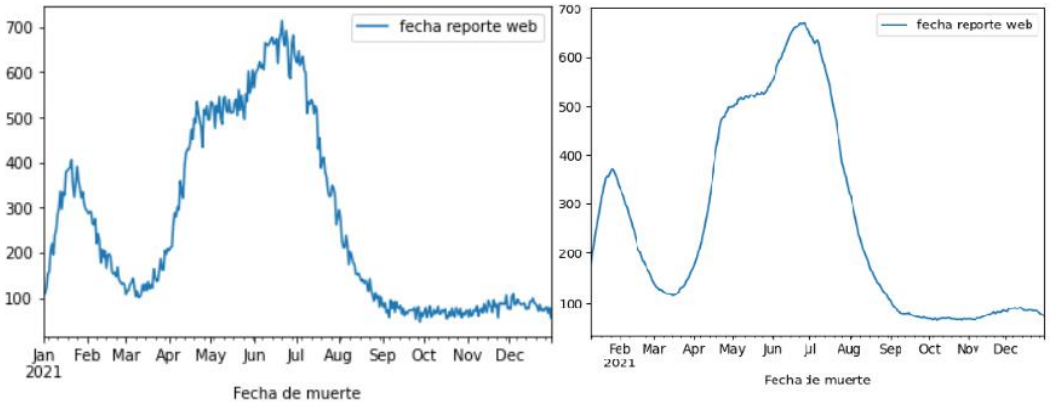


Figura 7. Graficas de muertes reportadas antes y después de usar la función Rolling. Fuente propia.

6.2. ENTRENAMIENTO Y EJECUCION DEL MODELO

La proporción para el entrenamiento y prueba fue de 80% y 20% respectivamente, a continuación, fue necesario escalar los datos en un rango entre -1 y 1 usando la librería de Scikit-Learn. Para el entrenamiento es necesario dividir los datos en vectores, que en este caso se usó un step de 20. Es decir, se toma ese rango determinado para lograr predecir el valor siguiente, y a continuación comparar el nuevo dato predicho con el dato reportado, así sucesivamente se recorre todo el dataset con el fin de obtener un error (ver figura 8).

Como es mencionado previamente, se usó una arquitectura de red LSTM, que empieza con 256 neuronas y se van disminuyendo gradualmente hasta llegar a 32. A su vez, se usa un dropout de 0.2 para evitar overfitting y al final se usa una red densa para obtener la salida. La función de perdida es el error cuadrático medio (MSE) que es la más usada para hacer regresiones con el optimizador Adam. Adicionalmente, se entrena el modelo en 15 o 20 épocas de entrenamiento con un batch size de 32. Finalmente, antes de graficar la predicción es necesario realizar la inversión de los datos que fueron normalizados previamente, con el fin de tener una predicción con cifras reales. La tabla 3 muestra todos los parámetros empleados para el modelo propuesto, para la elección de estos hiperparámetros es esencial tener en cuenta la naturaleza de los datos y el problema específico que se intenta resolver.

Tabla 3. Hiperparámetros para el modelo propuesto. (Fuente propia)

Hiperparámetros	
Nº de unidades	256/128/64/32
Dropout	0.2
Loss	mse
Optimizador	adam
Epochs	20
Batch_size	32
Activación	ReLU
Escalización	(-1,1)
Tamaño de datos para Test	0.2 = 20%

6.3. RESULTADOS GRAFICOS

El pre-procesamiento y el entrenamiento del modelo LSTM se realizó para los dos grupos de datos que se tienen, muertes y contagios, en dos espacios de tiempo diferentes.
El primer espacio de tiempo abarca datos de enero hasta diciembre del año 2021 y la primera predicción se hizo a finales del año 2021 desde octubre hasta diciembre, tanto para muertes como contagios (ver figuras 8a y 8b).
Para el segundo espacio de tiempo los datos van desde enero hasta junio del año 2021, por ende, la segunda predicción tuvo lugar desde finales del mes de mayo hasta finales del mes de junio del mismo año (ver figuras 8c y 8d).
En estas figuras, de color verde se identifican los datos seleccionados para el entrenamiento del modelo, en color azul están los datos que se han reportado y por último, de color rojo los datos predichos por el sistema desarrollado, que se esperan sean los más parecidos posibles a los de color azul (reportados).

Una primera examinación, entre los valores reportados y predichos en cada uno de los experimentos se realizó comparando dos puntos de fechas al azar. En las tablas 4 y 5, se muestran los valores correspondientes obtenidos de la comparación, tanto para muertes como contagios, en los dos periodos de tiempo evaluados.
De esta forma es posible observar que es esencial hacer una buena división del dataset para entrenamiento y validación, con el fin de asegurar que el modelo se ajuste bien a los datos de predicción.

Por lo dicho anteriormente la predicción para el primer espacio de tiempo presenta mejores resultados en cuanto a la curva azul que son los datos reportados y la curva roja que es la predicción, en la figura 8a y 8b se observa que las curvas están prácticamente sobrepuestas, este comportamiento se ve reflejado en la tabla 4 donde los valores esperados y predichos de los grupos de muertes y contagios se diferencian por muy pocas unidades. Con el segundo espacio de tiempo se puede notar que se presentan algunas diferencias entre la curva azul y la curva roja, sin embargo, la curva roja que es la de predicción maneja la misma tendencia que la curva azul, es decir, que los datos irían en aumento a medida que transcurra el tiempo. A pesar de la diferencia que se ve en las curvas, en la tabla 4 donde se comparan los valores reportados y los datos de la predicción se comprueba que aún son datos cercanos, es decir que el modelo tiene buen desempeño en ambos casos así el total de datos se modifique.

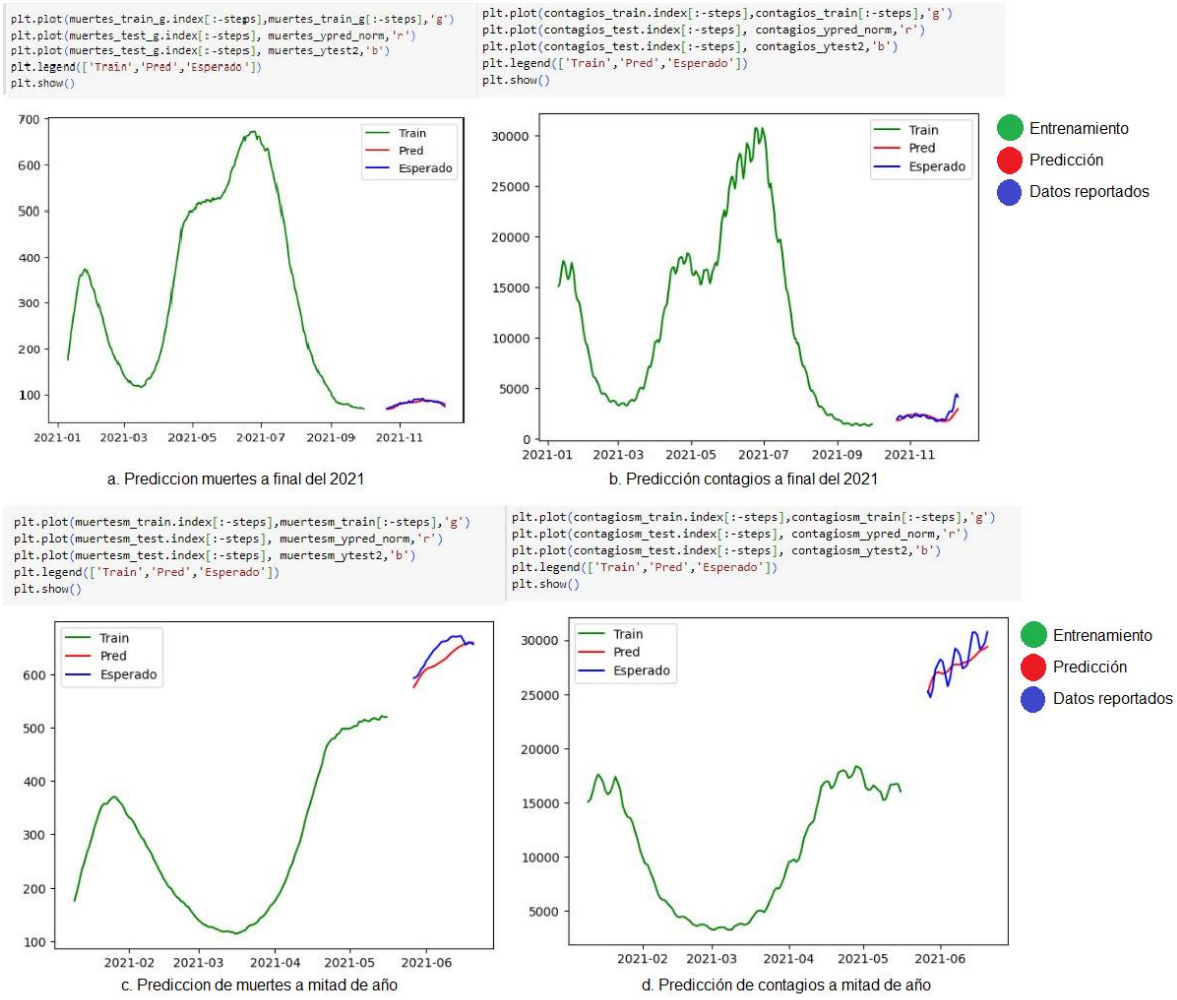


Figura 8. Graficas de predicción de final y mitad de año

Tabla 4. Predicciones de muertes y contagios a fin de año. Fuente propia.

MUERTES 2021 ENERO-DICIEMBRE			
REPORTADOS		PREDICCIÓN	
30/10/2021	31/10/2021	30/10/2021	31/10/2021
79	80	78	79
10/11/2021	11/11/2021	10/11/2021	11/11/2021
83	82	83	84

CONTAGIOS 2021 ENERO-DICIEMBRE			
REPORTADOS		PREDICCIÓN	
28/10/2021	29/10/2021	28/10/2021	29/10/2021
2.225	2.256	2.219	2.273
15/11/2021		15/11/2021	
2.231		2.235	

Tabla 5.Predicciones de muertes y contagios a mitad de año. Fuente propia.

MUERTES 2021 ENERO-JUNIO			
REPORTADOS		PREDICCIÓN	
30/05/2021		30/05/2021	
615		605	
16/06/2021	17/06/2021	16/06/2021	17/06/2021
655	660	655	657

CONTAGIOS 2021 ENERO-JUNIO	
REPORTADOS	PREDICCIÓN
17/06/2021	17/06/2021
29.190	29.000
8/6/2021	8/6/2021
27.769	27.727

El análisis de los resultados obtenidos en este estudio se ve relacionado directamente a las medidas emitidas por el ministerio de salud y el ministerio del interior, tales como establecer directrices diferentes para aquellos lugares con ocupación mayor al 70%, al 80% y al 85% de Unidades de Cuidados Intensivos (UCI). De igual manera, el Gobierno nacional recomendó instaurar opcionalmente medidas de pico y cédula y establecer restricciones nocturnas a la movilidad. Simultáneamente se establecen los criterios y condiciones para la distribución, asignación y entrega de las vacunas en el territorio colombiano en el marco del Plan Nacional de Vacunación Contra el COVID-19, el cual se compone de dos fases y cinco etapas, teniendo como priorización los grupos de riesgo. El manejo que se le dio a estas políticas se vieron reflejados en las cifras a principio de año donde hay una curva descendente que abarca aproximadamente 3 meses incluyendo febrero, marzo y abril que fueron los meses donde empezó la jornada de vacunación a la cual las personas asistieron responsablemente. Mientras que el plan de vacunación avanzaba apropiadamente, a finales del mes de abril había iniciado el paro nacional que se prolongó al menos por dos meses, en el que se expresaron y se fueron sumando actores y demandas diversas a lo largo y ancho del territorio nacional. Se puede caracterizar como una movilización y protesta masiva, fundamentalmente pacífica, la más grande en varios decenios, tanto a nivel nacional, como local y regional, que puso en evidencia la inconformidad de amplios sectores sociales populares y medios de Colombia. Del mismo modo, el impacto en las cifras de contagios y muertes a mitad de año fue notable debido a que hubo un pico que alcanzo alrededor de los 35000 contagios y las 700 muertes entre los meses de junio y julio (Ver figuras 6 y 7), que según expresó el Ministerio de Salud las manifestaciones obstaculizaron, en algunas partes del país, la distribución de las vacunas contra el COVID-19, así como la entrega de oxígeno y medicamentos, y

pidió que se respetara la misión médica afectando la ejecución del plan de vacunación y a la atención de pacientes. Finalmente, se puede observar una disminución de casos a finales del año 2021 (Ver figuras 6 y 7), lo cual está relacionado con el reporte de vacunación alcanzado al 13 de julio de 2021 con un total de 24,187,021 personas inmunizadas, reduciendo significativamente el riesgo de infección o complicaciones en caso de contraer el virus.

6.4. DESEMPEÑO DEL MODELO

En la Tabla 6, se presentan las métricas de desempeño obtenidas a partir de las predicciones de casos de contagio y muertes. En el caso de las muertes, al final del año, los valores de MAPE fueron de 3.55% y los de RMSE de 3.38, mientras que para los contagios fueron de 13.8% (MAPE) y 452.08 (RMSE). Por otro lado, a mitad de año, las predicciones mostraron para muertes un MAPE de 2.97% y un RMSE de 21.07, y para los contagios un MAPE de 3.21 y un RMSE de 1128.03.

Así es posible observar que, en ambos experimentos, los errores porcentuales absolutos medios (MAPE) y los errores cuadráticos medios (RMSE) para el grupo de contagios reflejaron valores elevados. Esto podría atribuirse posiblemente a la magnitud de las cifras, que es mayor, así como a la presencia de cambios más frecuentes a lo largo del tiempo.

Tabla 6. Métricas de desempeño de los experimentos. Fuente propia.

		MAPE	RMSE	INTERPRETACIÓN
Enero-Diciembre	MUERTES	3,55	3,38	PREDICCIÓN MUY PRECISA
	CONTAGIOS	13,8	452,08	BUENA PREDICCIÓN
Enero-Junio	MUERTES	2,972	21,07	PREDICCIÓN MUY PRECISA
	CONTAGIOS	3,216	1128,03	PREDICCIÓN MUY PRECISA

De esta forma es posible observar que en base a la métrica MAPE, como reportado por (Hansun & Young, 2021) la predicción se puede considerar como muy precisa, una vez que se obtiene valores por debajo de 10. Además, como se ha señalado anteriormente, la métrica RMSE puede generar valores poco convencionales, ya que es más susceptible a valores anómalos debido al uso de una potencia en su cálculo.

La Tabla 7 compara los resultados, del modelo propuesto para las predicciones de contagios y muertes por COVID-19 en el año 2021 y otros trabajos en la literatura. (Prasanth, Singh, & Kumar) propusieron un método de previsión para los casos de infección, usando google trends, junto con datos del centro europeo para la

prevención y el control de enfermedades. Desarrollaron un modelo hibrido GWO-LSTM en el que los parámetros de la red en algunos ensayos fueron optimizados con el optimizador Grey Wolf. Obteniendo para el total de casos acumulados un RMSE 0,035 y un MAPE de 14,161%, en el caso del total de muertes acumuladas, los resultados fueron RMSE de 0,074 y MAPE de13,752%. (Luo, Zhang, Fu, & Rao). Establecieron modelos de predicción en casos infectados diariamente aplicando los algoritmos LSTM con un refuerzo XGBoost, para el entrenamiento utilizaron un dataset recuperado del sitio web de la OMS donde se reportan datos desde el 3 de enero de 2020 hasta el 30 de septiembre de 2020 y emplearon 4 parámetros de rendimiento entre ellos RMSE y MAPE. Como resultado obtienen que el modelo LSTM obtuvo mejores resultados en términos de precisión con RMSE de 981 y MAPE de 2,32%. (Naeem, Mashwani, ABIAD, & Shah), propusieron predecir el COVID-19 basándose en 4 métodos diferentes, entre los cuales estaba LSTM, donde los resultados mostraron que ese era el mejor en comparación con los demás modelos, obteniendo un RMSE 687,54 y un MAPE 14,98% para los casos diarios confirmados y para las muertes reportadas los valores fueron RMSE 0,027 y MAPE 0,018%. El rendimiento medido por el error porcentual medio (MAPE) del modelo propuesto es superior a los resultados reportados en la literatura, dado que se entrenó en un conjunto de datos más grande y diverso con una gama más amplia de entradas. Sin embargo, es preciso aclarar que las comparaciones fueron realizadas con estudios que cuentan con características similares, tanto de rango de tiempo como la estructura del modelo, es decir que el modelo no tenga ninguna variación adicional. Con respecto a los resultados arrojados por la métrica de Error cuadrático medio (RMSE) en donde los valores del modelo propuesto suelen ser mayores a los reportados en la literatura, podría estar influido por la magnitud de los datos con la que se trabajó, debido a que los rangos de tiempo son mayores y esto implica tener un mayor flujo de datos.

Tabla 7. Métricas de desempeño RMSE y MAPE

Modelos	Rango de datos	RMSE	MAPE (%)	Referencias
LSTM- Contagios	*Enero 2021 a diciembre 2021*	452,08	13,8	*Este trabajo*
LSTM- Muertes	*Enero 2021 a diciembre 2021*	3,38	3,55	*Este trabajo*
LSTM- Contagios	*Enero 2021 a Junio 2021*	1128,03	3,21	*Este trabajo*
LSTM- Muertes	*Enero 2021 a Junio 2021*	21,07	2,97	*Este trabajo*
LSTM- Contagios	*Febrero 2020 a Mayo 2020*	0,035	14,16	(Prasanth, Singh, & Kumar)
LSTM- Muertes	*Febrero 2020 a Mayo 2020*	0,074	13,75	(Prasanth, Singh, & Kumar)
LSTM-XGBoost- Contagios	*Enero 2020 a Septiembre 2020*	981	2,32	(Luo, Zhang, Fu, & Rao)
LSTM- Contagios	*Abril 2020 a Agosto 2020*	687,54	14,98	(Naeem, Mashwani, ABIAD, & Shah)

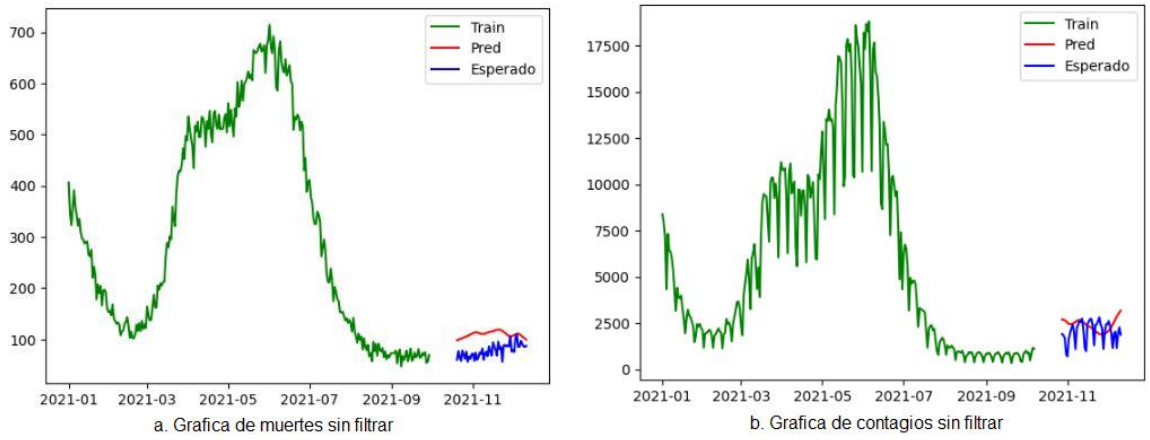
LSTM- Muertes	*Abril 2020 a Agosto 2020*	0,027	0,018	(Naeem, Mashwani, ABIAD, & Shah)

Es importante resaltar que antes de obtener los resultados definitivos de este estudio, que concluyen que el modelo genera predicciones bastante cercanas a los valores informados, se llevaron a cabo otros experimentos con los mismos grupos. Estos experimentos utilizaron métricas similares, pero se basaron en datos sin ningún tipo de filtrado. En otras palabras, se utilizaron datos reportados diariamente sin aplicar ningún proceso de suavizado a las curvas. La única verificación realizada fue que no faltara ningún informe para ningún día del año ni hubiera valores nulos reportados.

Estos experimentos generaron predicciones que, en términos visuales, no parecían estar muy alejadas de los datos reales en cuanto a la tendencia de las curvas (ver figura 9).

```
plt.plot(muertes_train_g.index[:-steps],ytrain_m1_norm,'g')
plt.plot(muertes_test_g.index[:-steps], ypred_m1_norm,'r')
plt.plot(muertes_test_g.index[:-steps], ytest_m1,'b')
plt.legend(['Train','Pred','Esperado'])
plt.show()
```

```
plt.plot(train_conta.index[:-steps],ytrain_c1_norm,'g')
plt.plot(test_conta.index[:-steps], ypred_c1_norm,'r')
plt.plot(test_conta.index[:-steps], ytest_c1,'b')
plt.legend(['Train','Pred','Esperado'])
plt.show()
```



a. Grafica de muertes sin filtrar

b. Grafica de contagios sin filtrar

Figura 9. Graficas sin aplicar ningún filtro

Sin embargo, al aplicar las métricas de desempeño, los valores obtenidos no coincidieron con las expectativas. Específicamente, usando los datos sin ningún proceso de filtrado, los resultados de las métricas de desempeño, arrojaron que el MAPE para el grupo de muertes por Covid-19 fue del 45% y para los contagios fue del 46%, lo que sugiere una predicción razonable (Hansun & Young, 2021).

Así, aunque se emplearon parámetros idénticos en ambos experimentos, como el tamaño de los conjuntos de entrenamiento y prueba, el rango de escalado de datos (-1 a 1), el número de pasos, el optimizador utilizado y la estructura del modelo LSTM, los resultados exhibieron variaciones debido al pre-procesamiento previo al que se sometieron los datos. En este contexto, la aplicación del filtrado de datos en intervalos de 10 días, logra la reducción de ciertos picos o interferencias que

estaban presentes en los datos originales. De esta forma, se resalta la importancia de un adecuado pre-procesamiento, es decir, la limpieza de datos nulos que no contribuyen con información valiosa al análisis.

7. CONCLUSIONES

- A causa de la rápida propagación de contagio que tuvo el COVID-19 durante los años 2020 y 2021 y la poca preparación que se tenía en cuanto al manejo de una pandemia, es importante contemplar el uso de modelos de aprendizaje profundo entrenados con datos reales para la predicción de las tendencias pandémicas. Con la intención de contribuir en la toma de decisiones que deben tomar los gobernantes respecto a las medidas necesarias para disminuir la propagación.

- En esta investigación se presentó un modelo LSTM para la predicción de contagios y muertes por covid-19 en el año 2021, usando datos reportados en el sitio web del Ministerio de Salud y Protección de Colombia. Se desarrollaron 4 experimentos con la misma configuración del dataset y de la red para garantizar una buena predicción. El modelo desarrollado para la predicción de muertes en el espacio de tiempo de enero hasta junio del año 2021 obtuvo un error porcentual absoluto medio (MAPE) del 2,97% el cual es relevante, en comparación con el estudio presentado por (Luo, Zhang, Fu, & Rao, 2021) donde utiliza datos de nuevos contagios, maneja aproximadamente el mismo rango de tiempo y aplica un modelo LSTM, obteniendo un MAPE del 2,32%.

- Emplear unas buenas técnicas de pre-procesamiento fue determinante para la obtención de buenos resultados debido a que el pre-procesamiento ayuda a limpiar el conjunto de datos, eliminando valores atípicos e incoherencias, asegurando de esta forma que el modelo se alimente de datos de alta calidad evitando que la red aprenda patrones falsos. Adicionalmente en las LSTM el escalado y la normalización de las características es fundamental para garantizar que las entradas tengan escalas similares y así mejorar el entrenamiento de la red.

- Considerando los buenos resultados que arrojó el estudio, eventualmente se puede usar este modelo para hacer diferentes predicciones, en la medida en que se tenga en cuenta las características que debe tener el dataset. Es importante contemplar que las muestras que se tomaron son relativamente pequeñas, en el primer experimento se usaron datos de enero a octubre para entrenamiento y para el segundo solamente fueron los primeros cinco meses.

8. RECOMENDACIONES

Los resultados que se obtuvieron haciendo uso de redes neuronales para la predicción de enfermedades, en este caso puntual la del COVID-19, demuestran la importancia de su uso, debido a la capacidad de analizar patrones y grandes conjuntos de datos. Aquí se presentan algunas recomendaciones para trabajos futuros:

- Se propone combinar varios modelos LSTM, experimentar con arquitecturas de red más profundas, añadiendo más capas a la red o haciendo uso de diferentes arquitecturas de redes neuronales como las unidades recurrentes controladas (GRU), para capturar diversos patrones y mejorar la precisión de la predicción.

- Los modelos basados en LSTM proporcionan información valiosa a las autoridades de salud pública para la toma de decisiones, por eso, se sugiere probar este modelo con otros conjuntos de datos, que pueden ser datos reportados de la misma enfermedad, pero en un país diferente, incluso se pueden hacer diferentes validaciones con los reportes de otra enfermedad.

- En cuanto a la comunicación clara que debe tener una herramienta de este tipo para la comunidad, se sugiere crear cuadros de mando (dashboard) o visualizaciones fáciles de usar que comuniquen las predicciones del modelo en tiempo real y de forma clara, para que se pueda generar información fiable.

9. REFERENCIAS

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning.* Springer International Publishing.

ArunKumar. (2021). Forecasting of COVID-19 using deep layer Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells. *ElsevierLtd.*

ArunKumar. (2021). Forecasting the dynamics of cumulative COVID-19 cases for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA).

Clinic, M. (2023, January 20). *Mayo Clinic* . Retrieved from https://www.mayoclinic.org/es-es/diseases-conditions/coronavirus/symptoms-causes/syc-20479963

Correa, J., & Muñoz, M. (2020). SARS-COV.2/COVID-19 en Colombia: tendencias, predicciones y tensiones sobre el sistema sanitario. *Revista de salud publica* .

Diaz, J. (2020). Uso de modelo predictivo para la dinámica de transmisión del Covid-19 en Colombia. . *Repertorio de medicina y cirugía*.

Economist, T. (2023). Machine learning and artificial intelligence in a brave new world. *The economist*. doi:https://www.sas.com/en_us/insights/articles/analytics/machine-learning-and-artificial-intelligence-in-a-brave-new-world.html

Fedosov, V., Fedossova, A., & Buitrago, O. (2020). Modelamiento estocástico de la evolución de la trasmisión de virus altamente contagiosos en lugares concurridos. *Revista UIS Ingenierías* , 13.

Gadelha, I. (2020). Forecasting Covid-19 Dynamics in Brazil: A Data Driven Approach. *Environmental Research and Public Health*.

García, F. R. (2020). Revisión de las tecnologías presentes en la industria 4.0. *Revista UIS Ingenierías, 19*(2), 177-191. doi:https://doi.org/10.18273/revuin.v19n2-2020019

García, M., & Daza, K. (2021). *Predicción, análisis y pronóstico de COVID-19 utilizando un modelo de machine learning basado en el análisis forecasting sobre series temporales.* Guayaquil: Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales.

Gautam, Y. (2020). Transfer Learning for COVID-19 cases and deaths forecast using LSTM. *Elsevier Ltd.*

Hansun, S., & Young, J. (2021). Predicting LQ45 financial sector indices using RNN-LSTM. *Journal of Big Data*.

Hoz, F. d. (2021). ¿Cómo le fue al sistema de salud colombiano durante la pandemia? *Epicrisis*.

Huang, J. (2020). Global prediction system for COVID-19 pandemic. *PubMed Central* .

IBM. (2023). *IBM*. Retrieved from https://www.ibm.com/topics/machine-learning#anchor--412854573

Irene Casas. (2020). Networks, Neural. In A. Kobayashi, *International Encyclopedia of Human Geography* (pp. 381-385).

Lenis, D. O. (2020). Predicciones de un modelo SEIR para casos de covid-19 en Cali, Colombia.

Luo, J., Zhang, Z., Fu, Y., & Rao, F. (2021). Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results in Physics, Volume 27*. doi:104462

Manrique, F. (2020). Modelo SIR de la pandemia. *Revista de salud publica* , 9.

Mijwil, M. M. (2022). Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning. *Iraqi Journal for Computer Science and Mathematics*. doi:https://doi.org/10.52866/ijcsm.2022.01.01.013

Minsalud. (2020, 03 06). *Ministerio de salud* . Retrieved from https://www.minsalud.gov.co/Paginas/Colombia-confirma-su-primer-caso-de-COVID-19.aspx

Moein, S. N. (2021). *Ineficiencia de los modelos SIR en el pronóstico de la epidemia de COVID-19: un estudio de caso de Isfahan.* Scientific Reports.

Montesinos, L. C. (2007). Modelos matemáticos para enfermedades infecciosas. *Salud publica de Mexico*, 9.

Naeem, M., Mashwani, W. K., ABIAD, M., & Shah, H. (2022). Soft Computing Techniques for Forecasting of COVID-19 in Pakistan. *Alexandria Engineering Journal*. doi:https://doi.org/10.1016/j.aej.2022.07.029

Neha, S., Reecha, S., & Jindal, N. (2021). Machine Learning and Deep Learning Applications-A Vision. 5.

Núñez, E. B. (2021). CRISP-DM y K-means neutrosofía en el análisis de factores de riesgo de pérdida de audición en niños. *Asociación Latinoamericana De Ciencias Neutrosóficas.*, 9.

OPS. (2020, 12 29). *Organizacion Panamericana de la salud* . Retrieved from https://www.paho.org/es/colombia

PAHO. (2020). *Pan American Health Organization*. Retrieved from https://iris.paho.org/bitstream/handle/10665.2/52276/PAHOEIHISCOVID-19200007_eng.pdf?sequence=8

Parbat, D., & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19. *Chaos, Solitons & Fractals*.

Picton, P. (1994). What is a Neural Network? In P. Picton, *Introduction to Neural Networks* (pp. 1-12). Red Globe Press London.

Prasanth, S., Singh, U., & Kumar, A. (2020). Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach. *Elsevier*, 12.

Ruifang, M., Xinqi, Z., Peipei, W., Haiyan, L., & Chunxiao, Z. (2021). *The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method.* Scientific Reports.

Runjie Zhu, X. T. (2020). Chapter seven - Deep learning on information retrieval and its applications. In *Deep Learning for Data Analytics* (pp. 125-153). Academic Press.

Shastri, S. (2020). Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. *Elseiver*.

Ullán, P. G., González, R. L., & González, M. A. (2022). Redes LSTM para la predicción de series. In H. S. Blas, & P. G. Ullán, *Avances en Informática y* (p. 131). Master universitario en sistemas inteligentes .

Wang, P. (2020). Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran.

WHO. (2020, 12 27). *World health organization*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

WHO. (2023, 03). *World Health Organization*. Retrieved from https://www.who.int/health-topics/coronavirus#tab=tab_1