Validation of the Prediction of Effectiveness of Statistical Time Series Models Using an Artificial Neural Network Model.

Validación de la Predicción de Efectividad de los Modelos Estadísticos de Serie de Tiempo Mediante un Modelo de Red Neuronal Artificial.

Lorenzo Cevallos-Torrez, MSc[1], Angel Ochoa-Flores, MSc[1], Luis Chóez-Acosta, MSc[1] ,
Darwin Patiño-Pérez, Ph.D[1], Miguel Botto-Tobar,MSc[1], Dalva Icaza-Rivera, MAE[2],Liliana Sarmiento-Barreiro, MSIG[3],
[1]Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Guayaquil,Ecuador, lorenzo.cevallost@ug.edu.ec, angel.ochoaf@ug.edu.ec,luis.choeza@ug.edu.ec,darwin.patinop@ug.edu.ec,miguel.bottot@ug.edu.ec,
[2]Universidad Estatal de Milagro, Facultad de Administración, Milagro, Ecuador, dicazar@unemi.edu.ec,
[3]Universidad de Guayaquil, Facultad de Ciencias Médicas, Guayaquil,Ecuador,liliana.sarmientob@ug.edu.ec

Abstract. – The objective of this project was to predict the number of cases of infections and deaths from covid-19 through the application of artificial intelligence techniques in order to validate the effectiveness of a statistical model and counteract congestion in the health area within the territory. Ecuadorian, the rapid spread that caused serious consequences in the health systems and the virus triggered a global health crisis, the drastic impact on people's lives caused the application of Artificial Neural Networks-RNA techniques to obtain rapid diagnoses and effective. Historical data from the Ecuadorian state about the infections and deaths recorded per day were taken, the data was processed using the time series statistical method technique and later in the RNA models for the generation of the prediction and validation of the statistical method, the results obtained from each of the neural networks provided a feasible forecast that was close to the real values. The main conclusions show that the techniques applied in this project are efficient when predicting the number of cases of infection and death from covid-19 based on historical data and that the use of neural networks is very useful for solving various predictive problems.

Keywords: covid-19, time series, artificial intelligence, neural networks, prediction.

Resumen. – El objetivo de este proyecto fue predecir el número de casos de contagios y muertes por covid-19 mediante la aplicación de técnicas de inteligencia artificial con la finalidad de validar la efectividad de un modelo estadístico y contrarrestar el congestionamiento en el área de salud dentro del territorio ecuatoriano, la rápida propagación que ocasiono graves consecuencias en el sistemas de salud y del virus desencadeno una crisis sanitaria a nivel mundial, el drástico impacto en la vida de las personas ocasiono que se apliquen técnicas de Redes Neuronales Artificiales-RNA para poder obtener diagnósticos rápidos y efectivos. Se tomaron datos históricos del estado ecuatoriano acerca de los contagios y muertes registrados por día, se procesaron los datos mediante la técnica de método estadísticos de series de tiempo y posteriormente en los modelos de RNA para la generación de la predicción y validación del método estadístico, los resultados obtenidos de cada una de las redes neuronales proporcionaron un pronóstico factible y cercano a los valores reales. Las principales conclusiones muestran que las técnicas aplicadas en este proyecto son eficientes al momento de predecir el número de casos en contagios y muerte por covid-19 en base a datos históricos y que el uso de las redes neuronales resulta muy útil para resolver diversos problemas predictivos.

Palabras Claves: covid-19, series de tiempo, inteligencia artificial , redes neuronales, predicción.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LEIRD2022.1.1.192
ISBN: 978-628-95207-3-6 ISSN: 2414-6390

I. INTRODUCCION

A finales del 2019 se descubrió en China un nuevo tipo de coronavirus que sería la principal causa de la pandemia del 2020[1]. Según se indica que los coronavirus son una extensa familia de virus que causan infecciones respiratorias que pueden ir desde el resfriado común hasta enfermedades más graves como el síndrome respiratorio de Oriente Medio (MERS)[2] y el síndrome respiratorio agudo severo (SARS)[3].

El COVID-19 es causado por el coronavirus 2 del síndrome respiratorio agudo severo (SARS-CoV2), su forma es redonda u ovalada y a menudo polimórfica, tiene un diámetro de 60 a 140 nm, la proteína espiga que se encuentra en la superficie del virus [4] y forma una estructura en forma de barra, es la estructura principal utilizada para la tipificación, la proteína de la nucleocápside encapsula el genoma viral y puede usarse como antígeno de diagnóstico. Los contagios de covid-19 se puede transmitir por diferentes medios de propagación, como pueden ser por aerosoles, por contacto físico con una persona infectada o tocar alguna superficie contaminada con el virus[5].

La importancia de los modelos actuales para poder predecir la efectividad de un modelo estadístico de series de tiempo en casos de contagios y muerte por covid-19[6] es poder manejar o gestionar grandes volúmenes de datos sin la intervención manual del ser humano o el conocimiento acerca del proceso. Los modelos de redes neuronales[7] a partir de su predicción, son de gran ayuda para la toma de decisiones por parte de los especialistas; los apoyados en este tipo de tecnología, ahora tienen más tiempo para

dedicarse a idear otros tratamientos potenciales[8] que permitan contrarrestar estragos de la pandemia a la cual enfrenta el Ecuador.

A) Trabajos Relacionados

Los autores de [9] desarrollaron una comparativa de las redes neuronales artificiales (RNA) tipo perceptrón multicapa (MLP) [10] y funciones de base radial (RBF)[11] aplicadas a la predicción de series de tiempo utilizando *resilient backpropagation* como algoritmo de aprendizaje para la red MLP y una combinación entre el algoritmo de los k-emanes[12] y el método de la matriz pseudo-inversa para la RBF y para la implementación de dichas redes utilizaron un sistema basado en la arquitectura cliente-servidor previendo la integración con aplicaciones web 2.0. Para evaluar el desempeño de las redes neuronales artificiales en cuanto a tiempo y calidad de respuesta utilizaron conjuntos de datos de diferentes características y cantidad de datos. Dicha comparación afirma que es preferible el uso de las RBF, debido a que obtiene mejores tiempos de ejecución y en cuanto a la calidad de respuesta ambas obtienen resultados similares. Sin embargo, existen varios modelos además de las RBF Y MLP, tales como Feedforward[13], Fuzzy[14] y la Red Neural Recurrente[15] por lo cual a través de la presente investigación se determinará el mejor ellos como método para validar series de tiempo.

En [16] se propone el uso de Redes Neuronales Recurrentes (RNN) como nuevo enfoque para tratar el problema del análisis y predicción de series temporales. Como casos de estudio y para demostrar el grado de éxito del uso de RNN aplican este enfoque al estudio del consumo eléctrico en la población de Sóller (Mallorca) y en el estudio del consumo eléctrico en la isla de Tenerife. El objetivo fue mostrar cómo, con esta metodología se puede predecir el consumo eléctrico de las poblaciones con un grado de precisión que ronda el 93%. Para ello, proponemos el uso de un tipo de red neuronal recurrente, conocida como “Long Short Term Memory Network”[17] (LSTM), por lo tanto siendo las redes neuronales recurrentes un nuevo enfoque para el análisis y predicción de series de tiempo mediante la presente investigación se evalúa la RNN[18] en un diferente caso de estudio es decir en casos de contagios y muertes por covid-19 con el objetivo de demostrar su efectividad ante la predicción de series de tiempo y otros modelos de Redes neuronales propuestas.

Los autores [19] proponen usar integralmente la estrategia de regularización de descomposición de pesos y validación cruzada con el fin de controlar integralmente el problema del sobreajuste en las redes neuronales tipo perceptrón multicapa para el pronóstico de series de tiempo con el fin de evaluar la capacidad de la propuesta, se pronostica una serie de tiempo tradicional y los resultados evidencian que la combinación de ambas técnicas permiten encontrar modelos con mejor capacidad de generalización que aproximaciones tradicionales, sin embargo mediante esta estrategia para generalizar y obtener mejor precisión se ve afectado el procedimiento automático para lo cual en el presente trabajo de investigación aplicaremos el modelo de perceptrón multicapa para el pronóstico en tiempos de pandemia como son los casos de contagios por Covid-19.

II. MATERIALES Y METODOS

A. Dataset

Los datos utilizados para esta investigación consisten en los reportes diarios subidos en página del ministerio de salud y reportes de ecuacovid subidos a repositorios de github, con la cual se conformó una base de datos con un total de 497 registros en un archivo de Excel y posteriormente convertido en un archivo csv (comma-separated values) o valores separados por comas, comprendidos en los periodos de 29 de febrero del 2020 hasta 23 de julio del 2021.

Aplicando la metodología KDD "Knowledge Discovery in Databases-KDD" (Descubrimiento de conocimiento en base de datos)[20] el cual consiste en los pasos de: importación y muestreo de datos, calidad de datos, transformación, modelización, evaluación e implementación; utilizando los pasos mencionados se procede a realizar el análisis de los modelos generados de serie de tiempo obtenido mediante el lenguaje de programación R en su versión 4.1.0 y redes neuronales Perceptrón Multicapa, Recurrente, Feedforward y Fuzzy obtenidos con el lenguaje de programación Python en su versión 3.9.

Se han realizado varios estudios para el pronóstico de casos y muertes por covid-19, en[21] se menciona la metodología utilizada para la elaboración de modelos predictivos de contagios y defunciones para la epidemia de COVID-19 en España basados en curvas de Gompertz. Pronosticando el final de la epidemia entre los meses de junio y julio de 2020. Por otro lado, según estudios realizados por [22] se propone como procedimiento para predecir el número de contagios y defunciones por covid-19, una curva de regresión que ajuste o explique cualitativa y cuantitativamente los datos conocidos. Mientras que en [23] se exponen unos modelos estadísticos como promedio móvil simple de 2 términos, métodos de suavizado exponencial lineal simple, suavización exponencial de Brown, modelo de media móvil integrado regresivo automático (ARIMA).

B. Modelos

1)Serie de Tiempo

En [24] se indica que una serie de tiempo es un conjunto de observaciones de una variable medida en puntos sucesivos en el tiempo o a lo largo de periodos sucesivos. Sin embargo en [25] se indica que este tipo de análisis permite estudiar la relación potencialmente causal entre diferentes variables que cambian en el tiempo y que se relacionan entre sí. Es la técnica más importante para hacer inferencias acerca del futuro, predicción, con base en lo que ha ocurrido en el pasado y se aplica en diferentes disciplinas del conocimiento. Por ejemplo, las series de tiempos son

bastante aplicados en el campo de la economía, utilizados también para medir indicadores, como índices de desempleo, índice de precios, pronósticos.

Las series constan de componentes como tendencia, ciclo, estacionalidad y movimientos irregulares o aleatorios [26]. En el caso de la tendencia, esta indica la dirección hacia la cual se dirige la serie de tiempo, característica que lo convierte en el componente más importante. Puede ser creciente, decreciente, constante, lineal, curvilínea, entre otras; se llama también tendencia a largo plazo y se representa con Tt [27]. Mientras que el ciclo indica las variaciones que ocurren en una serie de tiempo en períodos más prolongados. Cuando la métrica es en años son variaciones mayores de un año, comúnmente de 2 a 10 años. La serie sube y baja suavemente a manera de onda siguiendo la tendencia. Dicho ciclo puede ser causado por diversos cambios y se representa con Ct [27]. Sin embargo, la estacionalidad, indica las variaciones que ocurren acorto plazo en una serie de tiempo con respecto a la línea de tendencia general. Ocurre en períodos fijos como días, semanas, meses, trimestres o años y se representa con Et [27]. Finalmente, los movimientos irregulares (aleatorios), son oscilaciones de una serie temporal a corto plazo y que se atribuyen a factores imprevisibles o aleatorios. Corresponde al efecto de diversos factores a menudo desconocidos y se representa con At [27].

2)Tipos de Aprendizaje

El aprendizaje supervisado según [28] indica que se le dice "supervisado" ya que para su entrenamiento necesita de un conjunto de datos previamente etiquetado y clasificado. Sobre este grupo de datos conocido como "conjunto de datos de entrenamiento" el algoritmo realizará predicciones y las comparará con las etiquetas, con el error obtenido y a través de sucesivas iteraciones irá ajustando el modelo logrando así un aprendizaje progresivo. Mientras que el aprendizaje no supervisado, es una técnica de aprendizaje automático en la que los usuarios no necesitan supervisar el modelo[29]. En cambio, permite que el modelo funcione por sí solo para descubrir patrones e información que antes no se había detectado. Se trata principalmente de los datos sin etiquetar. Los algoritmos de aprendizaje no supervisado permiten a los usuarios realizar tareas de procesamiento más complejas en comparación con el aprendizaje supervisado[30].

3)Redes Neuronales Artificiales

En [31] se define a la red neuronal como un modelo computacional de procesamiento paralelo, el cual está compuesto por un conjunto de elementos neuronales simples. Este tipo de redes se relacionan con el cerebro, por la manera en la que se adquiere conocimiento mediante un proceso de aprendizaje y la cohesión presente entre las neuronas (peso sináptico), la cual se emplea para almacenar conocimiento. La red esta formada por una serie de capas, en donde las neuronas se encuentran ubicadas en varias capas, de manera que las neuronas de una capa están conectadas con las neuronas de la capa siguiente, a las que pueden enviar información. Cada neurona de la red (Fig.1) es una unidad de procesamiento de información que recibe información a través de las conexiones con las neuronas de la capa anterior [32], la red está conformada por la capa de entrada, que es quien recibe información del exterior. En las redes biológicas, esta sería una tarea de las dendritas [31] ; por otra parte, están las capas ocultas, la cuáles están encargadas de realizar el trabajo de la red. En las redes biológicas, está sería el soma. Y finalmente la capa de salida, que proporciona el resultado del trabajo de la red al exterior y envía información hacia otras neuronas. En las redes biológicas, esta sería una actividad realizada por el axón.

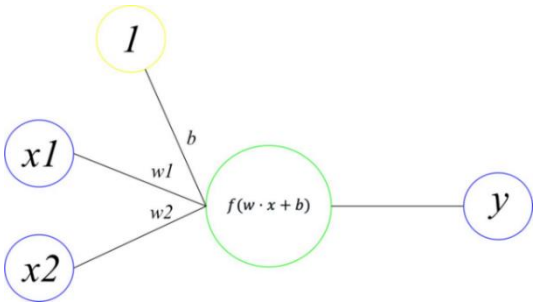

Fig. 1 Neurona Artificial

Tipos de Redes Neuronales

El Perceptrón Multicapa.- El Perceptrón multicapa es una red neuronal unidireccional constituida por tres capas o más: una capa de entrada, otra de salida y el resto de las capas intermedias denominadas capas ocultas. El tipo de aprendizaje de esta red es supervisado y se hace mediante el algoritmo denominado retropropagación de errores (backpropagation)[33].

Fuzzy o Difusas.- Es una combinación de las redes neuronales con el concepto de la Lógica Difusa. En [34] se menciona que el conocimiento obtenido por una red en el proceso de entrenamiento resulta en una gran cantidad de parámetros imposibles de interpretar con simples palabras. Por otra parte, una regla difusa consiste en sentencias lógicas del tipo if-then que se ajustan perfectamente al lenguaje natural. Hay que destacar que los sistemas difusos no pueden aprender de manera autónoma. Estas dos perspectivas son combinadas en los sistemas neurodifusos, con el fin de lograr simultáneamente el aprendizaje automático y la interpretación de los resultados por parte del usuario. Las reglas obtenidas de esta forma pueden revelar patrones en los datos de un modelo.

Feedforward.- En [35] definen que una red neuronal feed-forward consiste en una red neuronal compuesta de varias capas de neuronas colocadas en serie. La capa de neuronas que reciben el aprendizaje, los datos o señales del entorno se denomina capa sensorial. La capa de neuronas encargadas de proporcionar la respuesta de la red se denomina capa de salida. Una capa de neuronas intermedias entre la capa sensorial y la capa de salida que no tiene conexión directa con el entorno, es decir no recibe estímulos ni genera una respuesta final de la red, se denomina la capa oculta. Las neuronas de la capa oculta, gracias a que se encuentran en el intermedio de la red neuronal, proporciona grados de libertad a la red en general las cuales le permite aprender y representar más fehacientemente determinadas características del entorno que trata de modelar.

Recurrentes.- Las redes neuronales recurrentes[36] tienen caminos de retroalimentación entre todos los elementos que las conforman. Una sola neurona está entonces conectada a las neuronas posteriores en la siguiente capa, las neuronas pasadas de la capa anterior y a ella misma a través de vectores de pesos variables que sufren alteraciones en cada epoch con el fin de alcanzar los parámetros o metas de operación. Realizan el intercambio de información entre neuronas de una manera mucho más compleja y por sus características[37], dependiendo del tipo de algoritmo de entrenamiento que se elija, pueden propagar la información hacia delante en el tiempo, lo cual equivale a predecir eventos.

La Función de Activación

Es la que provee o devuelve la salida de una red, la cual será validada si supera un valor de umbral determinado para decidir si debe o no enviar los resultados a la siguiente neurona, en [31] se sugiere que el estado de activación de la neurona para un determinado instante de tiempo t puede ser expresado de la siguiente manera:

$$\sum(w.x) + b \quad (1)$$

Donde:
w = Peso asignado a cada nodo.
x = Valor de entrada de cada nodo
b = Es el número que anima a algunas neuronas activarse más fácilmente que otras.

Aunque hay muchas funciones de activación, hay tres más usado:

1)Función de Umbral**.-** Esta función devuelve el valor 0 siempre que la suma de los pesos sea menor que el valor marcado, de lo contrario, la función devolverá el valor 1.

2)Función Sigmoidea**.-** Se usa comúnmente en redes Multicapa o red con señal continua.

$$y = \frac{1}{1+e^{x}} \quad (2)$$

Donde:
y = Resultado de la función sigmoidea
e = Valor de Euler cuya aproximación es 2.7182
x = Valor de entrada de cada nodo

3)Función Tangente Hiperbólica.- Misma función que la anterior también se utiliza en redes con señal continua y su principal característica es que se pueden devolver valores negativos de hecho, los valores ingresados se convierten en una escala de [-1, 1]. Es muy útil en redes neuronales recurrentes.

$$f(x) = \frac{2}{1+e^{2x}} - 1 \quad (3)$$

Donde:
f(x) = Resultado de la función tangente hiperbólica
e = Valor de Euler cuya aproximación es 2.7182
2x = Valor de entrada de cada nodo multiplicado por 2.

Métricas de Validación

Para la validación de las redes neuronales se lo realizo mediante los criterios de raíz de error cuadrático medio y error porcentual medio para hacer una estimación de la varianza de error que tiene los resultados de cada red neuronal.

RMSE.- La raíz del error cuadrático expresada en (4)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - F_t)^2} \quad (4)$$

Donde:
n: Es la cantidad de datos
A_t: Es el valor predicho para la t-ésima observación en el conjunto de datos.
F_t: Es el valor observado para la t-ésima observación en el conjunto de datos.

MAE.-El error absoluto medio está expresado en (5)

$$MAE = \frac{\sum_{I=1}^{N}|x_i - y_i|}{N} \quad (5)$$

Donde:
x: Es el conjunto de valores reales
y: Es el conjunto de valores pronosticados
N: es la cantidad total de datos

MAPE.-Es la media de error porcentual absoluto expresada en (6), brinda información y permite la identificación del tamaño de los errores de pronóstico, comparándolos con los valores reales de la serie. La aplicación de éste se hace importante cuando el tamaño de la variable del pronóstico es relevante para evaluar la eficiencia del pronóstico.

$$MAPE = \frac{l}{n}\sum_{t=1}^{n}\left|Z_t - \frac{\hat{z}_t}{Z_t}\right| \quad (6)$$

Donde:
n: Es la cantidad de datos
Z_t: Es el valor estimado
$\hat{z}$_t: El valor de las ventas estimado

III. METODOLOGIA

Se implementó el proceso metodológico denominado "Knowledge Discovery in Databases - KDD" con el objetivo de obtener el mejor modelo de red neuronal para validar el modelo estadístico de series de tiempo, a través de las 6 fases de la cual se compone, los cuales se explican de forma detallada en el siguiente apartado:

Fase 1. Importación y muestreo de datos: Los datos se obtuvieron de sitios web del estado como www.salud.gob.ec y reportes de ecuacovid subidos a repositorios de github, correspondientes entre el 29 de febrero del 2020 y el 23 de julio del 2021, información del número de casos confirmados y fallecidos por Covid-19 en el Ecuador, dejando en evidencia al grupo de personas más vulnerables ante este virus denominado Covid-19, siendo estos datos de suma importancia para el pronóstico del modelo estadístico de serie de tiempo y además para el análisis y entrenamiento de cada una de las redes neuronales como lo son Perceptrón

multicapa, Fuzzy neural networks, Feedforward y Recurrent networks. La muestra es del mismo tamaño de la población la cual se conforma por un total de 497 registros.

Fase 2. Calidad de datos: La integridad o calidad de los datos determina la confiabilidad y un óptimo rendimiento frente a los modelos que los implementen, del mismo modo, una vez creado el dataset, se procedió a realizar el procedimiento respectivo en esta fase el cual implica la identificación y preprocesamiento de los datos con el objetivo de garantizar resultados pertinentes y confiables, los cuales se describen a continuación:

1)La revisión de los registros de la base de datos obtenida se llevó a cabo de forma manual, mediante el uso del aplicativo Microsoft Office Excel para hojas de cálculos, que ofrece herramientas que son de ayuda permitiendo corroborar que existían inconsistencias en algunos de ellos tales como; campos vacíos como casos confirmados por rango de edad, campos que nos ayudan a identificar el grupo de personas por edad más afectados y verificar que las columnas de los campos Total Casos confirmados, Total Fallecidos y Vacunados tenga consistencias con los valores numéricos demás columnas.

2)Identificadas las inconsistencias de la base de datos, debido a que la base de datos no posee un gran volumen de tamaño, se procedió a realizar la depuración manejando las correcciones respectivas de forma manual, dichos campos vacíos que pertenecen a datos cuantitativos, para lo cual se optó por completar los campos con información adecuada, se observó que los datos iban en forma creciente por lo que se rellena las celdas vacías, con el valor del último registro, puesto que fueron días que no hubo incremento, por lo que los valores se mantienen igual con respecto día anterior y finalmente para la consistencia en los registros, se verifica que tantos la suma de las columnas de las provincias como las columnas de las edades sea igual a los campos de Total Casos confirmados y Total Fallecidos, y se realiza el mismo procedimiento con las columnas Primera Dosis y Segunda Dosis para la columna de Vacunados tenga consistencia.

Fase 3. Transformación: Esta fase una vez totalmente preparada la base de datos se procede a seleccionar los datos más significativos para ser primeramente implementada al modelo estadístico de series de tiempo y posterior a cada una de las redes neuronales propuestas para esta investigación y realizar el entrenamiento correspondiente. En el proceso de la transformación se busca identificar las variables que serán utilizadas en la siguiente fase de modelización por medio de la minería de datos.

La base de datos original inicialmente consistía en los siguientes atributos: "Fecha del registro, N° de casos confirmados de 0-11 meses, De 1-4 años, De 5-9 años, De 10-14 años, De 15-19 años, De 20-49 años, De 50-64 años, N° de casos confirmados mayores a 65 años, N° de casos confirmados en Azuay, Bolívar, Cañar, Carchi, Chimborazo, Cotopaxi, El Oro, Esmeraldas Galápagos, Guayas, Imbabura, Loja, Los Ríos, Manabí, Morona Santiago, Napo, Orellana, Pastaza, Pichincha, Santa Elena, Sto. Domingo, Sucumbíos, Zamora Chinchipe, Femenino, Masculino, Primera Dosis, Segunda Dosis, Vacunados, Total Casos Confirmados, Total Fallecidos". Para la selección de las variables significativas se realizó lo siguiente:

1)Se realizó una comprobación de correlación entre las variables del dataset, y verificar cuales son las variables que más se correlacionan entre si tomando en consideración las variables más relevantes para un criterio de atención médica que desea analizar, mediante el proceso de coeficiente de correlación de Pearson.

2) Una vez realizada la comprobación de correlación se procedió a armar un nuevo dataset conformados con las variables elegidas por la correlación con mayor relevancia para el proceso de la modelización dejando en el nuevo dataset las siguientes variables: De 0-11 meses confirmados, De 1-4 años confirmados, De 5-9 años confirmados, De 10-14 años confirmados, De 15-19 años confirmados, De 20-49 años confirmados, De 50-64 años confirmados, Más de 65 años confirmados, femenino, masculino, vacunados, casos confirmados, muertes confirmadas.

Fase 4. Modelización: Una vez armado el nuevo dataset se procede a crear los modelos que se utilizaran, empezando con el modelo de series de tiempo, para lo cual se utilizó el lenguaje de programación R en su versión 4.1.0 con el IDE de desarrollo R Studio. Para la selección del modelo debido a que se desea analizar una serie de tiempo que contiene varias variables, se investigó sobre los diferentes modelos de series de tiempo, en la cual se eligió el modelo VAR, o modelo de Vectores Autorregresivos el cual es el modelo que más adecuado para el análisis de series de tiempos multivariantes, siendo estos una extensión de los modelos AR y se procede a realizar un análisis de las tendencias de las series de tiempo de cada variable del dataset.

Luego del análisis de las tendencias de las series de tiempo, se procede a realizar la prueba de Phillips Perron para comprobar si es estacionario. Flores [38] nos indica que: "La prueba P-P de Phillips & Perron es una prueba de raíz unitaria. Es decir, se utiliza en el análisis de series de tiempo para probar la hipótesis nula de que una serie de tiempo es integrada de orden 1. Se basa en la prueba de Dickey & Fuller, de que la hipótesis nula es $p = 0$, donde delta es la primera diferencia del operador". En [39] se sugiere que "El valor p se define como la probabilidad de observar un valor estadístico de prueba mayor o igual al encontrado. Tradicionalmente, el valor de corte para rechazar la hipótesis nula es 0.05, lo que significa que cuando no hay diferencia, se espera un valor tan extremo para el estadístico de prueba en menos del 5% de las veces". Según [40] afirma que: "un proceso o modelo es estacionario en sentido estricto si las funciones de distribución conjuntas como la media, las varianzas, las covarianzas y las funciones de distribución completas son constantes o invariantes con respecto al desplazamiento en el tiempo". Posteriormente, se verifico que todas las variables no eran estacionarios por lo que se procedió a realizar diferenciaciones para lograr que sea estacionario, se utiliza la función ndiffs para verificar cuantas diferencias son necesarias.

Ahora determinaremos el orden de retraso para lo cual se usa para los criterios de información (AIC, HQ, SC y FPE) el Criterio de información de Akaike (AIC), Criterio de información de Hannan y Quinn (HQ), el criterio información bayesiana de Schwarz (SC) y el error de predicción final de Akaike (FPE), se escoge el valor mínimo de regazo. Estos criterios de información nos indica que la información relativa que se pierde con las diferentes especificaciones. Finalmente, se procedió con la realización del pronóstico para los próximos 30 días de nuestro dataset y se lo guarda en un archivo csv.

Redes Neuronales Artificiales Implementadas

Luego de haber realizado las series de tiempos y obtener los pronósticos se procede a realizar la comprobación con las redes neuronales artificiales en Tabla I se reflejan los modelos Perceptrón Multicapa, Red Neuronal Recurrente, Fuzzy y Feedforward, para lo cual se utilizó el lenguaje de programación python en su versión 3.9, en el entorno de desarrollo Google Colab.

TABLA I
MODELOS DE REDES NEURONALES

Algoritmo Perceptrón multicapa

```
model = Sequential()
model.add(Dense(PASOS, input_shape=(1,PASOS),activation='relu'))
model.add(Dense(PASOS-5, activation='relu'))
model.add(Flatten())
model.add(Dense(1, activation='relu'))
model.compile(loss='mean_squared_error',optimizer='Adam',metrics=["mse"])
model.summary()
```

Algoritmo Red Neuronal Recurrente

```
model = Sequential()
model.add(LSTM(PASOS, return_sequences=True, input_shape=(1,PASOS)))
model.add(LSTM(PASOS-5, return_sequences=False))
model.add(Flatten())
model.add(Dense(1))
model.compile(loss='mean_squared_error',optimizer='adam',metrics= ["mse"])
model.summary()
```

Algoritmo Feedforward

```
model = Sequential()
model.add(Dense(PASOS, input_shape=(1,PASOS),activation='tanh'))
model.add(Flatten())
model.add(Dense(1, activation='tanh'))
model.compile(loss='mean_absolute_error',optimizer='adam',metrics=["mse"])
```

Algoritmo Red Neuronal Fuzzy

```
from FuzzyLayer import FuzzyLayer
from DefuzzyLayer import DefuzzyLayer
model = Sequential()
model.add(FuzzyLayer(PASOS*3, input_shape=(1,PASOS)))
model.add(Flatten())
model.add(DefuzzyLayer(1))

model.compile(loss='logcosh',optimizer='rmsprop',metrics=['mean_squared_error'])
model.summary()
```

Se realizó el respectivo entrenamiento de cada red neuronal con el respectivo dataset, y generó su propio pronóstico de las variables dependientes que se quieren estudiar las cuales son casos confirmados y las muertes confirmados, luego se hace una comparativa de los resultados obtenidos de la series de tiempo con los resultados de la red neuronal para determinar cuál de las redes neuronales fue la más precisa con respecto al pronóstico del modelo de series de tiempo, mediante métricas de margen de error.

Para el entrenamiento de las redes neuronales es recomendable realizar un normalizado al dataset, en que los valores son transformados en rangos de 0 y 1, para el modelo de Red Recurrente y Perceptrón Multicapa, mientras que en los modelos de Feedforward y Fuzzy, debido a que se utiliza la función de activación de tangente hiperbólica el cual usa un rango de -1 a 1, es necesario normalizar los datos de entradas, para disminuir la posibilidad de tener resultados incorrectos. Luego el dataset se lo divide en dos subconjuntos, teniendo una para entrenamiento con un 93.88% de los datos y otro para validación con un 6.12% de los datos.

Fase 5. Evaluación

Análisis de los resultados

En la Tabla II se reflejan los porcentajes de precisión reflejados por cada uno de los modelos de redes neuronales.

TABLA II
METRICAS DE EVALUACION

Algoritmo	MAPE		MAE		RMSE	
	Casos	Muertes	Casos	Muertes	Casos	Muertes
Feedforward	0.74	10.78	3555.49	2480.27	3588.82	2809.8
Perceptrón	0.79	13.6	3816.98	3032.55	4131.7	3439.45
Recurrente	0.76	19.77	3555.49	4529.02	3588.82	4556.88
Fuzzy	0.78	14	3759.86	3213.42	3807.04	3266.76

CASO DE ESTUDIO

Los resultados reflejan que las redes neuronales se adaptan muy bien para el pronóstico proporcionando un alto nivel de precisión, siendo estos factibles para validar el método estadístico de series de tiempo, aunque como se puede observar cuentan con una diferencia entre los modelos analizados. Destacando el modelo con los menores márgenes de error es el modelo Feedforward teniendo un error porcentual absoluto medio de 0.74 en la variable de casos confirmados y 10.78 en la variable de muertes confirmadas, en error absoluto medio se tiene 3555.49 en la variable de casos confirmados y 2480.27 muertes confirmadas, en raíz del error cuadrático medio se tiene 3588.82 en la variable de casos confirmados y 2809.8 en la variable de muertes confirmadas.

A continuación, se describe el proceso que se realiza para el análisis de casos de contagios y muertes por covid-19 de la población de Ecuador. El objetivo de este estudio es hacer una predicción o estimación futura aproximada de la cantidad total de casos en contagios y muertes por días, a partir de la información de la serie histórica de casos registrados desde el inicio de la pandemia en un periodo determinado de tiempo. Este análisis tiene como fin proporcionar una herramienta adicional en la toma de decisión, como por ejemplo anticiparse a los hechos y tomar

medidas correspondientes en el área. Para este fin, se emplea series de tiempo donde se registra el número de casos de contagios y muertes por covid-19 en el país que comprende el periodo de tiempo entre 29/02/2020, hasta el 23/07/2021. Estos datos son proporcionados por los sitios web del estado ecuatoriano.

IV. Resultado, Discusion y Conclusion

Resultados

El modelo de VAR (vectores auto regresivos) fueron utilizado para la series de tiempo fueron aplicados mediante la herramienta tecnológica R Studio entorno de desarrollo integrado (IDE) para el lenguaje de programación R, mientras que los modelos de las redes neuronales de Perceptrón Multicapa, Recurrente, Feedforward, Fuzzy, fueron codificados mediante IDE de desarrollo cloud Google Colab para el lenguaje de programación Python, con el objetivo de predecir la alza de casos de contagios de COVID-19, para la toma oportuna de decisiones por parte de las autoridades y personal médico, y lograr bajar las casos de contagios a tiempo.

Se realizó una exhaustiva investigación mediante las revisiones sistemáticas y documentales los cuales fueron nuestra fuente de información necesaria para definir las variables de la presente investigación y escoger los modelos de redes neuronales las cuales nos ayudan a validar el método estadístico de series de tiempo.

Se llevo a cabo el levantamiento de información correspondiente a casos de contagios y fallecimiento por covid-19 mediante sitios web del estado, conformando de tal manera el dataset con las características necesarias y posteriormente aplicado en el análisis y diseño de la investigación siguiendo las 6 fases de la metodología KDD.

Mediante el modelo estadístico de series de tiempo se realizó el pronóstico para las variables relacionadas con contagio y muerte por covid-19 definidas en el dataset y de tal forma se obtuvo el modelo predictivo que representa los datos observados en un periodo de tiempo.

Se procedió a evaluar las variables mediante cada uno de los modelos de redes neuronales propuestos en esta investigación definiendo una función en Python con el propósito de validar el modelo estadístico de series de tiempo y obtener el mejor modelo de red neuronal. Para lo cual se obtuvo en los resultados obtenidos en la Tabla 2 de la evaluación todos los modelos se obtuvo buenos resultados con pocas diferencias, al analizar los resultados se puede concluir que el mejor modelo para validar la red neuronal es el modelo Feedforward obteniendo en validación de casos confirmados resultados 0.74 en error porcentual absoluto medio, 3555.49 en error absoluto medio y 3588.82 en raíz del error cuadrático medio, mientras que para muertes confirmadas se tiene 10.78 en error porcentual absoluto medio, 2480.27 en error absoluto medio y 2809.8 en raíz del error cuadrático medio.

Conclusiones

1) En la recopilación de información bibliográfica de los diferentes modelos de redes neuronales, para validar aquellos métodos estadísticos basados en pronóstico y que se ajusten al estudio de los casos de contagios y muerte por covid-19, se realizó en base a la información encontradas de diversas fuentes confiables tales como: Google Scholar, Scielo, Sciencedirect, Redalyc, entre otros. En base al tema de investigación, se determinó las modelos de redes neuronales a utilizar más comunes para pronósticos de series de tiempos.

2)Se obtuvo información de los casos de contagios y muertes a nivel en sitios web del estado como www.salud.gob.ec y reportes de ecuacovid subidos a repositorios de github, para conformar la base de datos utilizada en este proyecto. Para la depuración de la base de datos se utilizó la metodología Knowledge Discovery in Databases (KDD), el cual se realizó en 6 fases, permitiendo obtener una dataset con datos más limpio con las variables significativas para su aplicación en los algoritmos.

3)Se realizó una predicción de las variables más significativas para los casos y muertes confirmados con covid-19, en el que se obtuvo un modelo predictivo, mediante el modelo VAR (modelo de Vectores Autorregresivos) el cual es utilizado para series de tiempos multivariables, en el lenguaje de programación R con el entorno de desarrollo R studio, guardando los resultados del pronóstico en un generado por R en formato csv.

4)Se validó el modelo estadístico mediante una realización de pronósticos en los modelos de redes neuronales propuestos siendo estos Perceptrón Multicapa, Recurrente, Feedforward, Fuzzy; utilizando el lenguaje de programación Python en el entorno de desarrollo en la nube de Google Colab y posteriormente comparando los resultados obtenidos de cada red neuronal, para comprobar cual se ajusta mejor a las características predictivas del modelo estadístico.

5)De acuerdo al análisis realizado en la evaluación del modelo estadístico de series de tiempo tipo VAR frente a los diferentes modelos de redes neuronales se determinó en base a los resultados que el mejor modelo es el Feedforward obteniendo en validación de casos confirmados resultados 0.74 en error porcentual absoluto medio, 3555.49 en error absoluto medio y 3588.82 en raíz del error cuadrático medio, mientras que para muertes confirmadas se tiene 10.78 en error porcentual absoluto medio, 2480.27 en error absoluto medio y 2809.8 en raíz del error cuadrático medio.

6)Los modelos de RN el Perceptrón Multicapa, Red Recurrente, Feedforward y Fuzzy son muy potentes al momento de realizar predicciones, pero evidentemente al compararlos de acuerdo con la eficiencia, en este caso el modelo Feedforward obtiene mejor precisión ante el estudio de los casos de contagios y muerte por covid-19.

REFERENCIAS

[1] X. Huang, F. Wei, L. Hu, L. Wen, and K. Chen, "Epidemiology and clinical characteristics of COVID-19," *Archives of Iranian Medicine*, vol. 23, no. 4. 2020.

[2] N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, and E. Petersen, "COVID-19, SARS and MERS: are they closely related?," *Clinical Microbiology and Infection*, vol. 26, no. 6. 2020.

[3] J. Shang *et al.*, "Cell entry mechanisms of SARS-CoV-2," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 21, 2020.

[4] A. M. R. Pérez, T. J. J. Gómez, and G. R. A. Dieguez, "Características clínico-epidemiológicas de la COVID-19," *Rev. Habanera Ciencias Médicas 19(2)e_3254*, pp. 1–15, 2020.

[5] J. R. Vielma Guevara, J. del C. Villarreal Andrade, and L. V. Gutiérrez Peña, "Pandemia por el SARS-CoV-2: aspectos biológicos, epidemiológicos y clínicos," *Observador del Conocimiento. Revista Especializada de Gestión Social del Conocimiento*, vol. 5, no. 3. 2020.

[6] NCT05014373, "Philippine Trial to Determine Efficacy and Safety of Favipiravir for COVID-19," *https://clinicaltrials.gov/show/NCT05014373*, 2021.

[7] J. A.-T. Barrera, "Redes Neuronales Artificiales," *Univ. Guadalajara*, p. 276, 2016.

[8] U. of W. Hospitals, "Tratamientos y profilaxis (prevención) para COVID-19 - UW Health COVID-19 Information." [Online]. Available: https://coronavirus.uwhealth.org/es/sintomas-y-cuidados/tratamientos-y-profilaxis-prevencion-para-covid-19/. [Accessed: 18-Feb-2022].

[9] "'Clinical Characteristics and Outcomes of Hospitalized Patients with COVID-19 Treated in Hubei (epicenter) and outside Hubei (non-epicenter): A Nationwide Analysis of China' European Respiratory Journal 2020." [Online]. Available: https://erj.ersjournals.com/content/early/2020/04/01/13993003.00562-2020.figures-only.

[10] F. Haghighat, "Predicting the trend of indicators related to Covid-19 using the combined MLP-MC model," *Chaos, Solitons and Fractals*, vol. 152, 2021.

[11] Q. Jiang, L. Zhu, C. Shu, and V. Sekar, "An efficient multilayer RBF neural network and its application to regression problems," *Neural Comput. Appl.*, vol. 34, no. 6, 2022.

[12] D. M. Polo, L. P. Caballero, and E. M. Gómez, "Comparación de Redes Neuronales aplicadas a la predicción de Series de Tiempo Comparison of Neural Network applied to prediction of Time Series," vol. 13, pp. 88–95, 2015.

[13] H. Z. Alemu, W. Wu, and J. Zhao, "Feedforward neural networks with a hidden layer regularization method," *Symmetry (Basel).*, vol. 10, no. 10, 2018.

[14] N. S. Hamde, A. Kumar, and S. Maithani, "Fuzzy machine learning approach for transitioned building footprints extraction using dual-sensor temporal data," *SN Appl. Sci.*, vol. 3, no. 4, 2021.

[15] I. Bonet Cruz, S. Salazar Martínez, A. Rodríguez Abed, R. Grau Ábalo, and M. M. García Lorenzo, "Redes neuronales recurrentes para el analisis de secuencias," *Rev. Cuba. Ciencias Informáticas*, vol. 1, no. 4, 2007.

[16] L. Tian *et al.*, "RNA-dependent RNA polymerase (RdRp) inhibitors: The current landscape and repurposing for the COVID-19 pandemic," *European Journal of Medicinal Chemistry*, vol. 213. 2021.

[17] N. Becerra Yoma Prof. and L. Mendoza Inzunza Dra., "Inteligencia artificial aplicada a la medicina respiratoria," *Rev. Chil. enfermedades Respir.*, vol. 37, no. 4, 2021.

[18] O. Araque, R. Barbado, J. Fernando Sánchez-Rada, and C. A. Iglesias, "Applying Recurrent Neural Networks to Sentiment Analysis of Spanish Tweets Aplicación de Redes Neuronales Recurrentes al Análisis de Sentimientos sobre Tweets en Español," *Work. Semant. Anal. SEPLN*, 2017.

[19] A. F. Velásquez J. D, Fonnegra R, "Pronóstico de series de tiempo con redes neuronales regularizadas y validación cruzada," pp. 267–279, 2013.

[20] Yuli Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika," *J. Edik Inform.*, vol. 2, no. 2, 2019.

[21] B. B. Anderson Camilo and I. de Telecomunicaciones, "Diseño e implementación de un modelo analítico predictivo para el apoyo en la toma de decisiones enfocado en las empresas de telecomunicaciones," 2020.

[22] G. E. Chanchí Golondrino, W. Y. Campo Muñoz, and L. M. Sierra Martinez, "Aplicación de la regresión polinomial para la caracterización de la curva del COVID-19, mediante técnicas de machine learning," *Investig. e Innovación en Ing.*, vol. 8, no. 2, 2020.

[23] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset," *Data Br.*, vol. 29, 2020.

[24] A. Alonso Rodríguez, "Análisis de las series temporales a la luz de Deep Learning," *Anu. jurídico y económico Escur.*, no. 52, 2019.

[25] and G. R. J. C. González-Avella, J. M. Tudurí, "Análisis de Series Temporales Usando Redes Neuronales Recurrentes," pp. 1–6, 2017.

[26] B. Montesdeoca, S. Tutor, C. Guerra, and A. Fecha, "Estudios de predicción en series temporales de datos meteorológicos utilizando redes neuronales recurrentes (Bachelor's thesis).," *Univ. Las Palmas Gran Canar. España*, 2016.

[27] "PRONÓSTICO DE SERIES DE TIEMPO CON REDES NEURONALES REGULARIZADAS Y VALIDACIÓN CRUZADA." [Online]. Available: https://redib.org/Record/oai_articulo591245-pronóstico-de-series-de-tiempo-con-redes-neuronales-regularizadas-y-validación-cruzada. [Accessed: 14-Oct-2022].

[28] D. L. PRADO PIN and M. P. VILLAVICENCIO

GALÁN, “DESARROLLO DE UN MODELO PREDICTIVO BASADO EN APRENDIZAJE DE MAQUINA SUPERVISADO, PARA EL ANÁLISIS DE DATOS DE TRISOMÍA 21 EN PROCESOS DE SALUD PRENATAL,” 2020.

[29] V. Roman, “Aprendizaje No Supervisado en Machine Learning: Agrupación | by Victor Roman | Ciencia y Datos | Medium,” *Medium*, 2019.

[30] A. R. Valdez Alvarado, “Machine Learning para Todos.,” *Reseaechgate*, no. January 2019, 2019.

[31] E. S. M, A. Serna, and E. Acevedo, “Principios y características de las redes neuronales artificiales,” *Desarro. e Innovación en Ing. Segunda Edición*, 2017.

[32] “Visualización e interpretación de redes neuronales convolucionales mediante dropout espacial.” [Online]. Available: https://riull.ull.es/xmlui/handle/915/21778. [Accessed: 25-Mar-2022].

[33] P. A. Blanco, “Algoritmo de Retropropagación,” *Univ. Pennsylvania Pinkel*, 2014.

[34] G. Morales, “Introducción a la lógica difusa,” *Cent. Investig. y Estud. Av. México*, 2002.

[35] M. Gabella, “Topology of Learning in Feedforward Neural Networks,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 8, 2021.

[36] O. Araque, R. Barbado, J. Fernando Sánchez-Rada, and C. A. Iglesias, “Applying Recurrent Neural Networks to Sentiment Analysis of Spanish Tweets,” *TASS 2017 Work. Semant. Anal. SEPLN*, 2017.

[37] B. Li and X. Zhuang, “Multiscale computation on feedforward neural network and recurrent neural network,” *Front. Struct. Civ. Eng.*, vol. 14, no. 6, 2020.

[38] T. Del Barrio Castro, P. M. M. Rodrigues, and A. M. R. Taylor, “On the Behaviour of Phillips-Perron Tests in the Presence of Persistent Cycles,” *Oxf. Bull. Econ. Stat.*, vol. 77, no. 4, 2015.

[39] J. K. Afriyie, S. Twumasi-Ankrah, K. B. Gyamfi, D. Arthur, and W. A. Pels, “Evaluating the performance of unit root tests in single time series processes,” *Math. Stat.*, vol. 8, no. 6, 2020.

[40] O. P. Cedeño-Fuentes, L. Arboleda-Castro, I. Jacho-Sánchez, and P. Novoa-Hernández, “Optimización de modelos de Stackelberg no estacionarios mediante un algoritmo evolutivo auto-adaptativo,” *TecnoLógicas*, vol. 20, no. 39, 2017.