

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 18/04/2021 | Aceptado: 30/05/2021 | Publicado: 30/09/2021

Identificadores persistentes:
ARK: [ark:/42411/s6/a43](https://nbn-resolving.org/urn:nbn:org:ark:42411/s6/a43)
PURL: [42411/s6/a43](https://nbn-resolving.org/urn:nbn:org:ark:42411/s6/a43)

Predicción de mortalidad a causa del Covid 19 en Perú utilizando redes neuronales artificiales

Prediction of mortality due to Covid 19 in Peru using artificial neural networks

Cesar Mayta Avalos ¹[\[0000-0002-5722-1854\]*, Jesús Cristian Valdivia Mamani ²\[\\[0000-0001-9566-6988\\]\]\(https://orcid.org/0000-0001-9566-6988\), Fernando Rosales Castilla ³\[\\[0000-0003-0668-2885\\]\]\(https://orcid.org/0000-0003-0668-2885\), Milca Gines Colana ⁴\[\\[0000-0002-3596-2803\\]\]\(https://orcid.org/0000-0002-3596-2803\)](https://orcid.org/0000-0002-5722-1854)

¹ Universidad Jorge Basadre Grohman de Tacna. cesar.mayta@unjbg.edu.pe

² Universidad Jorge Basadre Grohman de Tacna. jesus.valdivia@unjbg.edu.pe

³ Universidad Jorge Basadre Grohman de Tacna. fernando.rosales@unjbg.edu.pe

⁴ Universidad Jorge Basadre Grohman de Tacna. milca.gines@unjbg.edu.pe

* Autor para correspondencia: cesar.mayta@unjbg.edu.pe

Resumen

Con el desarrollo de la pandemia en Perú, la cantidad de fallecidos ha ido en aumento y lamentablemente no se han tomado las medidas adecuadas, esto por no tener una herramienta que nos permita saber la cantidad de fallecidos posibles en un tiempo determinado. El objetivo del presente artículo es proponer una herramienta capaz de predecir la cantidad de fallecidos por COVID-19 en función del tiempo. La metodología utilizada fue redes neuronales artificiales utilizando series temporales con información obtenida del Ministerio de Salud del estado peruano a través de su portal de datos abiertos. Los resultados alcanzados tuvieron un error cuadrático medio de 0.0037 y pérdida de 0.0480. Los resultados obtenidos a lo largo del artículo confirman la validez de esta herramienta y la efectividad en la predicción de la cantidad de fallecidos a causa del COVID 19.

Palabras clave: Inteligencia Artificial, Series Temporales, COVID 19, Predicción, Pronóstico

Abstract

With the development of the pandemic in Peru, the number of deaths has been increasing and unfortunately the appropriate measures have not been taken, this because we do not have a tool that allows us to know the number of

possible deaths in a given time. The objective of this article is to propose a tool capable of predicting the number of deaths from COVID-19 as a function of time. The methodology used was artificial neural networks using time series with information obtained from the Ministry of Health of the Peruvian state through its open data portal. The results achieved had a mean square error of 0.0037 and a loss of 0.0480. The results obtained throughout the article confirm the validity of this tool and its effectiveness in predicting the number of deaths from COVID 19.

Keywords: Artificial Intelligence, Time Series, COVID 19, Prediction, Forecast

Introducción

Hoy en día, la prioridad de todas las naciones, es lograr el máximo nivel de salud, tanto en el aspecto físico, mental y social. A nivel mundial se toman decisiones a favor del pueblo, con el fin de lograr este objetivo. Durante muchos años hemos sido afectados por diversos tipos de enfermedades, de los cuales hemos salido airosos a pesar de las pérdidas. Esto conlleva a preguntarnos: ¿Estaremos preparados para una pandemia de grandes magnitudes? ¿Nuestro sistema de salud llegaría a colapsar?

En diciembre de 2019, se detectó por primera vez un brote causado por el nuevo coronavirus humano del síndrome respiratorio agudo grave de tipo 2 (SARS-CoV-2), en la ciudad de Wuhan, provincia de Hubei, en China [1]. Desde entonces se ha propagado por todo el mundo produciendo una severa crisis económica, social y de salud, nunca antes vista. Debido a su rápida extensión, la Organización Mundial de la Salud declaró la pandemia por la enfermedad de coronavirus del 2019 (COVID-19) el 11 de marzo del 2020. El COVID-19 hizo su aparición en Perú un poco más tarde en comparación con otros países. Al primer caso confirmado, se tomaron medidas como la instauración de un estado de emergencia sanitaria, inmovilización total obligatoria, medidas de higiene, cierre de fronteras y de aeropuertos. A pesar de todas las medidas tomadas por el estado peruano, se tuvo la tasa de mortalidad más alta a nivel mundial durante la primera ola de contagios.

Con el desarrollo de la pandemia en Perú, la cantidad de fallecidos ha ido en aumento, según el registro de fallecimientos del Ministerio de Salud (MINSa) [2], las últimas cifras han sido alarmantes alcanzando un total de fallecidos de 193230 hasta el 03 de julio del 2021. Como estado, nuestro sistema de salud falló, nuestras entidades hospitalarias no se dieron abasto, incluso instituciones funerarias colapsaron, nuestro sistema educativo no está preparado para adversidades, nuestro sistema económico es ineficaz. ¿Cuántas personas más tendrán que morir para tener un buen sistema, para así prever las muertes de más peruanos?

Lamentablemente en nuestro país se tiene el pensamiento arcaico de “ver para creer”. Es por eso que teniendo cifras adecuadas, que nos permitan saber la cantidad de fallecidos posibles en un tiempo determinado, se podrían plantear nuevas estrategias. Para así evitar la aglomeración de pacientes y la desesperación de un pueblo que no solo se queja por la enfermedad, sino también por la falta de dinero, la crisis que produjo esta pandemia. Actualmente no se tiene un modelo predictivo que permita conocer el número de fallecidos diarios por COVID 19 en el Perú, al no tener este modelo no se podría estimar si la tasa de mortalidad incrementará en el tiempo.

La inteligencia artificial es una nueva forma de resolver problemas dentro de los cuales se incluyen los sistemas expertos, el manejo y control de robots y los procesadores, que intenta integrar el conocimiento en tales sistemas, en otras palabras, un sistema inteligente capaz de escribir su propio programa. Un sistema experto definido como una estructura de programación capaz de almacenar y utilizar un conocimiento sobre un área determinada que se traduce en su capacidad de aprendizaje.

El objetivo de la investigación es proponer una herramienta capaz de predecir la cantidad de fallecidos por COVID-19 en función del tiempo. Para el cumplimiento de dicho objetivo, el presente artículo consta de cuatro secciones, aparte de la introducción y se estructura de la siguiente manera: La sección Materiales y métodos o Metodología computacional, describe el desarrollo de las actividades realizadas a lo largo del trabajo de investigación, partiendo de conceptos previos y definición de las herramientas utilizadas, la sección de Análisis de los resultados, muestra la descripción del *dataset* utilizado, su respectivo tratamiento, entrenamiento de la herramienta y predicción; Finalmente en la sección de conclusiones, se presenta en base al objetivo planteado.

Materiales y métodos o Metodología computacional

Conceptos

Para dar solución al problema identificado es necesario conocer el concepto de Inteligencia Artificial, Series Temporales, Estacionalidad y las Principales librerías utilizadas en Python.

Inteligencia Artificial.

Una de las definiciones que se puede considerar más ajustada a la realidad es la reflejada en la *Encyclopedia Of Artificial Intelligence*:

“La Inteligencia Artificial es un campo de la ciencia y la ingeniería que se ocupa de la comprensión, desde el punto de vista informático, de lo que se denomina comúnmente comportamiento inteligente. También se ocupa de la creación de artefactos que exhiben este comportamiento”.

Otros autores prefieren otras definiciones como:

“La Inteligencia Artificial es el estudio de las ideas que permiten ser inteligentes a los ordenadores” (H.Winston).

“Parte de la informática que estudia procesos simbólicos, razonamiento no algorítmico y representaciones simbólicas de conocimiento” [3].

Series Temporales:

Como lo describe [4], una serie temporal es un conjunto de observaciones de una variable tomadas al largo de intervalos regulares de tiempo, como el número de automóviles vendidos por un fabricante cada mes durante los últimos diez años. Las series temporales aparecen prácticamente en todos los campos de actividad. El interés de su análisis estadístico radica en el estudio del comportamiento de la serie, lo que permite explicar sus variaciones y, sobre todo, en la posibilidad de predecir valores futuros.

Estacionalidad:

La estacionalidad se define por una fluctuación cíclica o periódica de la serie temporal que se repite de forma regular.

Si no fuera por la estacionalidad, el análisis de las series temporales se convertiría en una tarea muy simple.[5]

Un ejemplo claro de variación estacional es la que se produce a lo largo del año, particularmente en climas templados y fríos, en cuanto la temperatura, las precipitaciones inundaciones y sequías periódicas, la duración del día y la noche y la ecología, y que dan lugar a las estaciones del año.

Herramientas

En esta investigación propone como aporte una herramienta para determinar la predicción del crecimiento de los fallecimientos debido al Covid 19 en el mes de Julio, usando para ello Series temporales mediante herramientas de machine learning. Se utilizó como lenguaje de programación Python, la cual utiliza programación multiparadigma ya que soporta parcialmente la orientación a objetos, programación imperativa y en menor medida la programación funcional.

Principales librerías de Python para Machine Learning:

- **Pandas:** Librería más utilizada para el tratamiento de datos en Python, una de sus grandes virtudes que tiene esta librería es la carga de datos como los archivos de texto plano como CSV “Comma Separated Values”[6].
- **Numpy:** Librería que por excelencia tiene su virtud en el procesamiento de arrays. Debido a que contiene una gran colección de funciones que permite realizar cálculos matemáticos complejos sobre arrays multidimensionales. [7]
- **Matplotlib:** Esta librería es importante en las tareas de visualización y entre sus cualidades destacan que es open source y trabaja a bajo nivel.
- **Seaborn:** Es una librería gráfica basada en matplotlib, especializada en la visualización de datos estadísticos. Se caracteriza por ofrecer un interfaz de alto nivel para crear gráficos estadísticos visualmente atractivos e informativos.
- **Skelearn:** Es una librería que cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad.
- **Keras:** Es una interfaz de alto nivel para manipular las redes neuronales, con keras podemos ver fácilmente si nuestras ideas darán buenos resultados inmediatos.[8]

Google Colaboraty

Como se indica. [9], utilizaremos la herramienta de Google Colaboraty también llamada “Colab” la cual nos permitirá ejecutar y programar en Python desde nuestros navegadores mediante una Jupyter Notebook.

Análisis y resultados

Diseño y Población

El presente estudio fue descriptivo y estuvo basado en un análisis de series de tiempo correspondiente al período entre el 3 de marzo del 2020 al 3 de julio del 2021 en Perú. Se utilizó un dataset de los fallecidos por COVID 19, extraídos del portal del Gobierno del Estado Peruano en la sección de datos abiertos proporcionados por el Ministerio de Salud.

Tratamiento del Dataset

El dataset consta de 193,230 registros pertenecientes al periodo del 3 de marzo del 2020 al 3 de julio del 2021, a continuación, se muestra el diccionario de datos:

Tabla 1. Diccionario de Datos.

Variable	Descripción
FECHA_CORTE	Fecha de corte de información
UUID	ID de la persona fallecida por covid-19
FECHA_FALLECIMIENTO	Fecha de fallecimiento que ocurre por covid-19
EDAD_DECLARADA	Edad de la persona fallecida por covid-19
SEXO	Sexo de la persona fallecida por covid-19
CLASIFICACION_DEF	Criterios utilizados para la confirmación de la defunción por covid-19
UBIGEO	Código de Ubicación Geográfica que denotan "DDppdd" (Departamento, provincia, distrito), fuente INEI
DEPARTAMENTO	Departamento donde reside la persona fallecida por covid-19
PROVINCIA	Provincia donde reside la persona fallecida por covid-19
DISTRITO	Distrito donde reside la persona fallecida por covid-19

Fuente: Minsa

Se ejecutaron las acciones de retirar aquellas filas con datos anómalos y anormales con la finalidad de no afectar considerablemente a los resultados, la estrategia desarrollada para dicho tratamiento consistió en eliminarlas del dataset.

Seguidamente y disponiendo de un dataset depurado se procedió a eliminar las columnas FECHA_CORTE, UUID, CLASIFICACION_DEF, UBIGEO, DEPARTAMENTO, PROVINCIA y DISTRITO, la justificación de lo señalado se basa en el aspecto que la serie de tiempo será aplicada sobre la columna FECHA_FALLECIMIENTO.

Se realiza una conversión de la columna FECHA_FALLECIMIENTO que es de tipo INT a formato DATETIME, el cual nos permitirá realizar el agrupamiento de cantidad de fallecidos por fecha.

A continuación, se procede a crear un índice con la correspondiente columna de FECHA_FALLECIMIENTO, es importante señalar que una de las características más poderosas y recomendadas de las series temporales es la indexación basada en el tiempo, el uso de fechas y horas para organizar y acceder de forma intuitiva a los datos.

A partir del índice creado iniciamos las acciones de obtener la cantidad de filas agrupadas por fecha con las que se va a procesar siendo un total de: 479 filas.

La reducción de resolución es volver a muestrear con *resample* un conjunto de datos de series de tiempo a un marco de tiempo más amplio, en el caso aplicado se está pasando de días a meses originando un número reducido de filas.

Estacionalidad

En el siguiente modelo se ajusta la media de los fallecidos por COVID 19 del año 2020 con el año 2021, observando que las líneas de color azul y naranja no coinciden, no son exactas, pero sí muestran una misma tendencia. Esto muestra un comportamiento de un periodo de repetición por lo cual esta serie temporal es considerada Estacionaria.



Figura 1. Comportamiento de la Serie temporal de los fallecidos por COVID 19 del año 2020 y 2021.

PREPROCESAMIENTO

Debido a que la cantidad de variables de entrada es una sola columna, se necesita hacer un preprocesamiento de nuestro *dataset* para tener más variables de entrada. Para este fin estamos considerando el hacer una conversión de series a aprendizaje supervisado considerando 7 días.

El resultado de esta conversión es el siguiente:

	var1(t-7)	var1(t-6)	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
7	-1.000000	-1.000000	-1.000000	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166
8	-1.000000	-1.000000	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899
9	-1.000000	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899	-0.990533
10	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899	-0.990533	-0.990533
11	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899	-0.990533	-0.990533	-0.988166

Figura 2. Resultado del preprocesamiento.

Entrenamiento

Se aplicó el modelo de series temporales utilizando un 60% para entrenamiento y 40% para pruebas. Este procesamiento se hizo sobre un total 479 registros de agrupamiento de fallecimientos por fecha.

Con esto hemos procesado para entrenamiento 287 registros y para la validación 192 días teniendo como resultado del entrenamiento los siguientes valores

0s 2ms/step - loss: 0.0553 - mse: 0.0049 - val_loss: 0.0480 - val_mse: 0.0037

El siguiente gráfico nos muestra en los puntos verdes el valor esperado del entrenamiento y los puntos rojos el valor resultante de la red neuronal que es valor de validación de prueba y como se puede apreciar sin muy cercanos.

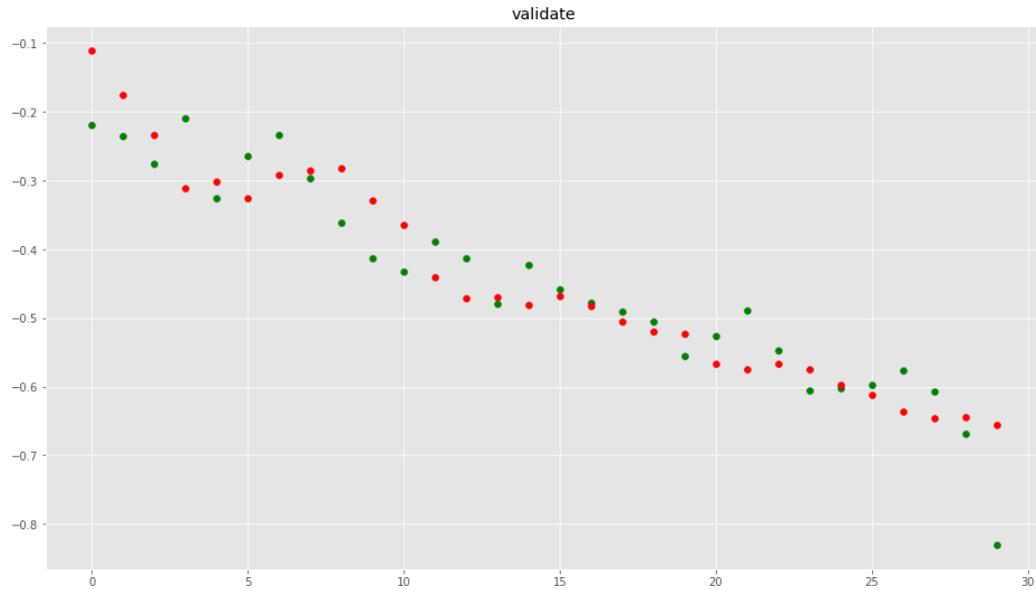


Figura 3. Tendencia de valores esperados vs valores resultantes de la red neuronal.

Respecto a la pérdida (*loss*) se puede apreciar el siguiente gráfico que muestra en azul la pérdida del valor esperado y en rojo la pérdida del valor hallado en la prueba de nuestro modelo lo que indica que también son similares



Figura 4. Tendencia de valor de Loss (pérdida) de datos esperados y resultantes de la red neuronal.

En este gráfico tenemos la variación del valor de error cuadrático medio que es la suma de las diferencias entre el valor esperado vs el valor de validación elevado al cuadrado el cual nos da al final un valor de precisión de 0.037

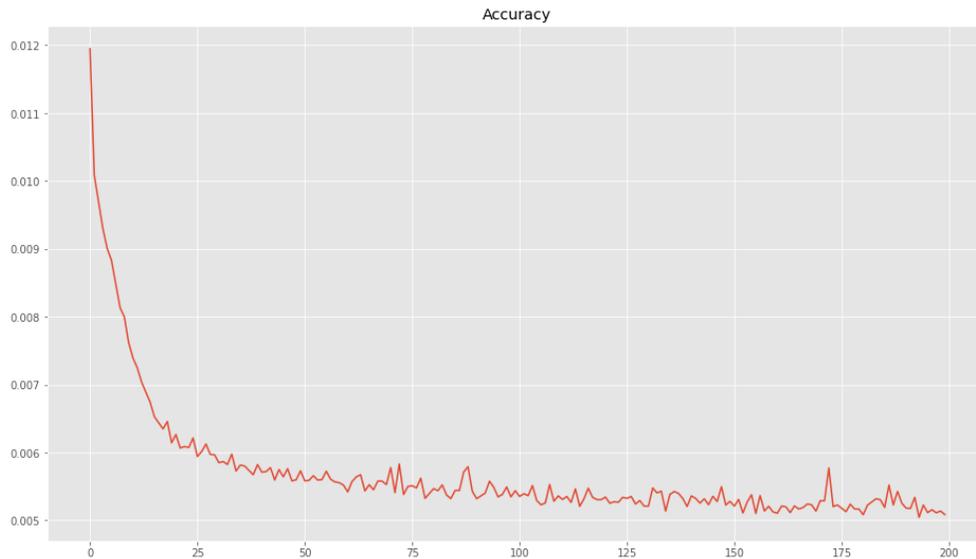


Figura 5. Comportamiento del error cuadrático medio.

Finalmente, para poder visualizar en la Figura 6 la diferencia entre los valores que nuestro modelo predijo en el entrenamiento vs el valor real de fallecidos tenemos el siguiente cuadro que nos muestra que mantiene la misma tendencia.

Predicción

Como nuestro modelo ya se encuentra entrenado y validado, procedemos a realizar la predicción; ya que, nuestro *dataset* solo nos muestra la cantidad de fallecidos hasta el 03 de Julio del 2021, se realizó la predicción de la cantidad de fallecidos para el periodo restante hasta fines del mes de Julio del 2021. Para la predicción se utilizará los últimos 30 días de nuestro *dataset* para predecir el mes de Julio.

En la Figura 7 se muestra la distribución de elementos por día, mostrando una tendencia a la baja en la cantidad de fallecidos, Se puede evidenciar que para fines del mes de Julio se llegará a los 110 fallecidos, la mitad de la cantidad de fallecidos a inicios del mismo mes que es 220; asimismo, observamos dos picos durante el mes, al tercer día del mes

en el que hay un aumento de 208 a 211 fallecidos y el segundo pico lo vemos al sexto día con un aumento de 180 a 182 fallecidos.

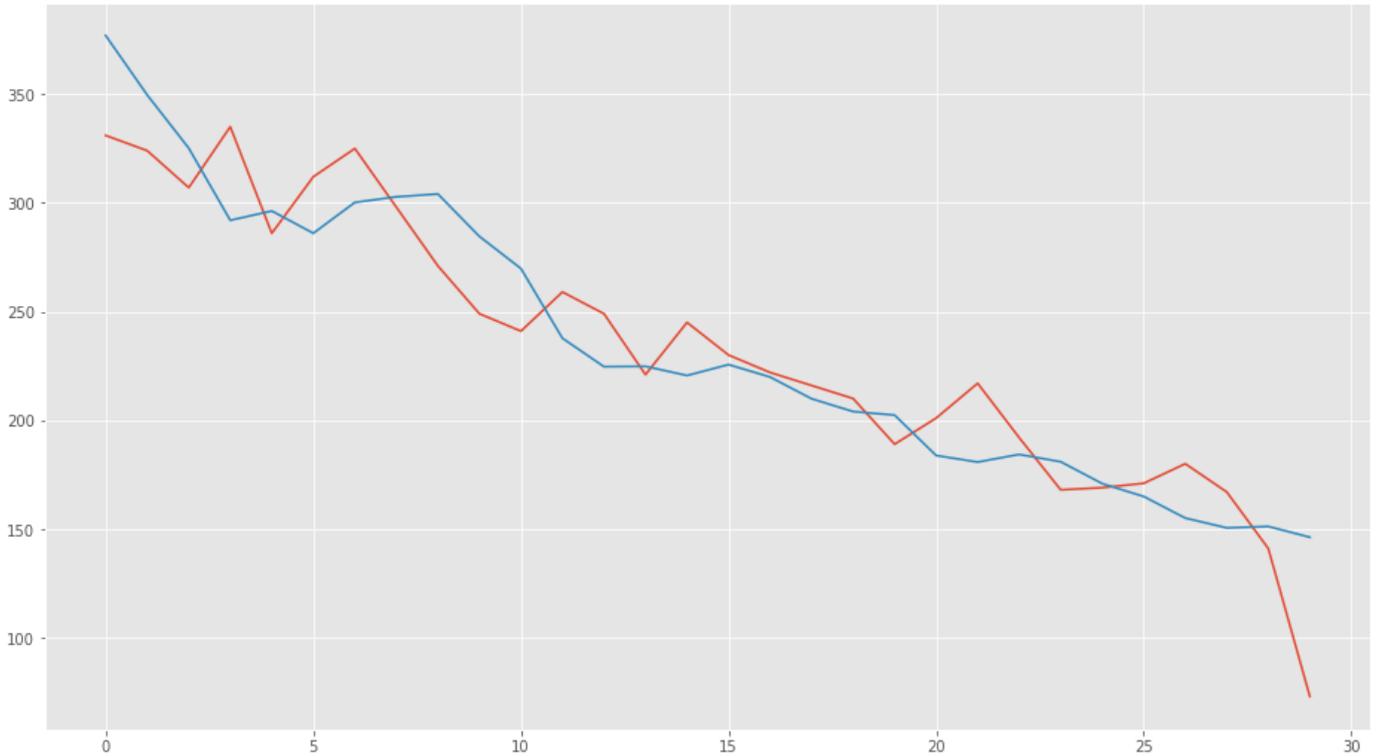


Figura 6. Tendencia de cantidad de fallecidos de la predicción del entrenamiento vs el valor real.

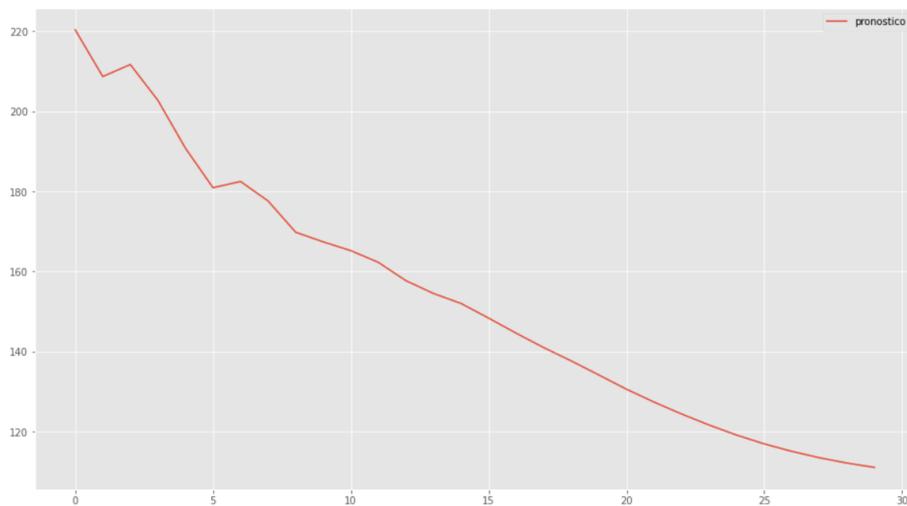


Figura 7. Pronóstico para el mes de Julio 2021.

Conclusiones

Con los datos disponibles hasta la fecha se presenta una herramienta para la predicción de la cantidad de fallecidos a causa del COVID 19 basada en redes neuronales artificiales, que complementa a la información proporcionada por el ministerio de salud a través del portal de datos abiertos.

Los resultados obtenidos a lo largo del artículo confirman la validez de esta herramienta y la efectividad en la predicción de la cantidad de fallecidos a causa del COVID 19, lo cual se evidencia en la sección de análisis y resultados.

Los modelos identificados a lo largo del artículo presentan un horizonte de estimación que depende de la estacionalidad, en aras de obtener una buena predicción. Para el desarrollo de la presente herramienta se utilizó con la cantidad de fallecidos por Covid 19 durante la pandemia, por lo que dicha estacionalidad es temporal debido a que al término de la pandemia volvería a sus valores normales.

En otras palabras, se requiere construir un modelo que simule de forma más precisa utilizando un *dataset* más amplio que no considere el tiempo de pandemia para que se pueda aplicar para futuros años post pandemia.

Referencias

- [1] Organización Mundial de la Salud, “coronavirus COVID 19” July, 2021. [Online]. Available: https://www.who.int/es/emergencias/diseases/novel-coronavirus-2019?gclid=CjwKCAjw55-HBhAHEiwARMCszrbbBSFmekHH9cphVjelvC85L8pGGpKMcOMiNDkbJPAMYeUrpSEXaRoCT7MQAvD_BwE. [Accessed Jul. 09, 2021].
- [2] Ministerio de salud, “datos abiertos,” July, 2021. [Online]. Available: <https://www.datosabiertos.gob.pe/dataset/fallecidos-por-covid-19-ministerio-de-salud-minsa/resource/4b7636f3-5f0c-4404-8526>. [Accessed Jul. 09, 2021].
- [3] R. Pino, A. Gómez, N.de Abajo, "Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva," Universidad de Oviedo, pp. 01, 2001.
- [4] C. Guisande, A. Vaamonde, A. Barreiro, "Tratamiento de datos con R, Statistica y SPSS," Ediciones Diaz de santos, pp. 585, 2013.
- [5] J. Arnau, "Diseños de Series Temporales: Técnicas de Análisis," Edicions Universitat Barcelona, pp. 92, 2001.
- [6] W. McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython," O'Reilly Media, pp. 04, 2012.
- [7] F. Nelli, "Python Data Analytics: With Pandas, NumPy, and Matplotlib," Apress, pp. 47, 2018.

[8] J. Torres, "DEEP LEARNING Introducción práctica con Keras," CC BY-NC-SA, pp. 97, 2018.

[9] B. Auffarth, "Artificial Intelligence with Python Cookbook: Proven recipes for applying AI algorithms and deep learning techniques using TensorFlow 2.x and PyTorch 1.6," Packt Publishing Ltd, pp. 10, 2020.

Roles de Autoría

Cesar Mayta Avalos: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Validación, Redacción - borrador original. **Jesús Cristian Valdivia Mamani:** Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Fernando Rosales Castilla:** Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Milca Gines Colana:** Conceptualización, Análisis formal, Investigación, Metodología, Redacción - borrador original.