ISSN 2007-9737

Enfoque de re-entrenamiento automático de un modelo de Machine Learning para predecir el índice de infección por COVID 19

Erick Palomino Gutierrez, David Calderón Vilca*

Universidad Nacional Mayor de San Marcos,
Depto. Ingeniería de Software,
Perú

{erick.palomino1, hcalderonv}@unmsm.edu.pe

Resumen. Con la llegada de la infección por COVID 19 en el mundo, se vio afectada la estabilidad económica y sanitaria de distintos países del mundo; y a pesar de las múltiples propuestas de predicción de infección por COVID 19, aún existe el problema de la precisión poca duradera de estos trabajos. Este artículo presenta un enfoque de re-entrenamiento automático de un modelo de predicción de índice de infección por COVID 19 en cualquier país del mundo con el objetivo de tener una herramienta de planificación protocolar y estratégica para contrarrestar esta infección y que tenga un precisión que perdure a lo largo del tiempo. Para el desarrollo del modelo se utilizó una red neuronal recurrente LSTM y una fuente de datos constantemente actualizada, fundamental para la aplicación del enfoque. Los modelos producidos por este enfoque mantuvieron durante un mes de re-entrenamiento semanal un coeficiente de determinación en promedio de 0.989, 0.986 y 0.996 en 3 diferentes países: Perú, Brasil y Chile respectivamente. Comparado con otros trabajos, los modelos producidos tienen la ventaja de ser entrenados semanalmente con datos actualizados, manteniendo una precisión duradera y se comprobó, además, la efectividad de un proceso automatizado de re-entrenamiento de modelos de machine learning.

Palabras clave. Enfoque**,** re-entrenamiento automático, machine learning, COVID 19, predicción de infección**.**

Automatic Retraining Approach of a Machine Learning Model for Prediction of the COVID-19 Infection Rate

Abstract. With the arrival of the COVID-19 infection worldwide, the economic and health stability of various countries was affected. Despite numerous proposals for predicting COVID-19 infection, the problem of the short-lived accuracy of these models persists. This article presents an approach for the automatic retraining of a COVID-19 infection rate prediction model for any country in the world. The aim is to provide a tool for protocol and strategic planning to counter this infection, with accuracy that endures over time. The model was developed using a recurrent neural network (LSTM) and a constantly updated data source, essential for the application of the approach. The models produced by this approach maintained an average coefficient of determination of 0.989, 0.986, and 0.996 in three different countries—Peru, Brazil, and Chile, respectively—during a month of weekly retraining. Compared to other studies, the models produced have the advantage of being trained weekly with updated data, maintaining long-term accuracy. Furthermore, the effectiveness of an automated machine learning model retraining process was verified.

Keywords. Approach, automated retraining, machine learning, COVID-19, infection prediction.

1. Introducción

A inicio del 2020 la preocupación de una posible pandemia se iba haciendo cada vez más real, más no convincente para muchas autoridades que escépticas a lo que sucedía y confiadas de sus sistemas sanitarios daban por falso o irrelevante la expansión del virus SARS-COV-2.

A finales del año los estragos económicos que nos dejaba podía medirse en decenas de millones de personas en riesgo de pobreza extrema y alrededor de la mitad de la población mundial con riesgo de perder su modo de subsistencia y afectando directamente a los trabajadores

informales debido a las medidas de protección estatales así como la saturación del sistema de salud tanto estatal como privado [1].

En efectos sanitarios, países con sistemas sanitarios fortalecidos como Estados Unidos contaban ya a inicios del 2021 con medio millón de fallecidos por COVID 19 [2], mientras países con menos población y mayores problemas sanitarios alcanzaban los cientos de miles de decesos [3].

A raíz de los efectos experimentados por el desconocimiento del comportamiento de la infección comenzaron a surgir múltiples investigaciones para conocer y contrarrestar los efectos de la pandemia en la población mundial. Gracias a esto se pudo desarrollar múltiples vacunas que actualmente sirven como barrera y disminuyen enormemente la cantidad de fallecidos [3].

Sin embargo, el virus aún existe y aún sigue propagándose e incluso volviendo a re-infectar a antiguos pacientes como también a personas vacunadas.

Esto permite que cada cierto tiempo se manifieste una “ola” de contagios que es capaz de saturar las instituciones médicas de aquellos países que no cuentan con la capacidad suficiente para atender tanto a los infectados por el virus como a cualquier otra persona que busque cualquier tipo de atención médica [4].

Con el fin de conocer con anticipación el comportamiento de la infección o de poseer información de una aproximación futura del índice de infección o mortalidad del COVID 19 se han realizado distintos trabajos proponiendo modelos tanto computacionales [5, 6] como estadísticos [7], [8] para poder predecir el índice de infección o mortalidad del virus.

De esta forma conociendo el índice de infección de los próximos días podemos prevenir los efectos de una ola de contagios a través de medidas protocolares o planeamiento estratégico para contrarrestar la saturación del sistema sanitario.

A pesar de mostrar excelentes resultados, algunos autores afirman que estos resultados, más específicamente, la precisión obtenida por estos modelos no es duradera [9].

Y esto se debe a dos principales razones, una es el uso de una fuente de datos de entrenamiento antigua o limitada a fechas donde aún no se observaba las variantes del comportamiento de la infección y la otra razón es que la mayoría de los modelos no consideran los eventos o medidas influyentes en el índice de infección.

Sin embargo, existen trabajos que consisten en recolectar y actualizar los datos del índice de infección o muerte en la mayor parte de países del mundo [3], adicionalmente ofrecen un tratamiento a los datos más relevantes para facilitar el procesamiento o entrenamiento de modelos [9].

Esto facilita enormemente la disposición de los datos necesarios para entrenar un modelo de predicción de infección de COVID 19; complementándolo con la técnica más eficiente y de resultados duraderos se obtendría una herramienta robusta para poder prevenir los efectos de una ola de contagios o preparar la infraestructura médica necesaria para enfrentar los casos más graves de contagio, como proponen [10, 11].

Dicho esto, el objetivo de este trabajo es proponer un enfoque de re-entrenamiento haciendo uso del algoritmo o la técnica más óptima, determinada a partir de un conjunto de trabajos anteriores, para poder predecir el índice de infección por COVID 19 en distintos países del mundo.

El contenido siguiente del artículo se encuentra estructurado de la siguiente manera.

La sección 2 describe la información de las investigaciones similares realizadas en todo el mundo comparando técnicas clasificadas en técnicas estadísticas y computacionales y analizando principalmente sus resultados.

La sección 3 explica la determinación de la mejor técnica así como el flujo general de entrenamiento y la condición de re-entrenamiento.

La sección 4 describe los resultados obtenidos de la investigación y su respectiva interpretación.

Finalmente en la sección 5 se discuten los resultados obtenidos comparándolos con los resultados de trabajos similares, además se indican las oportunidades de mejora de este tipo de investigación.

ISSN 2007-9737

2. Trabajos relacionados

2.1. Conjuntos de datos para entrenar modelos de predicción de infección por COVID 19 en el mundo

Existen actualmente dos fuentes centralizadas de datos de infección y muertes por COVID 19 en el mundo, siendo el más conocido el de la universidad de Jhon Hopkins [3] debido a su popular dashboard de consulta de infección en distintos países; sin embargo, lo que resalta de la propuesta de [12] del grupo Our World In Data de la universidad de Oxford es que se enfocan más que solo mostrar los datos reales, puesto que realizan un tratamiento a algunos campos del dataset para que estos puedan ser utilizados para el entrenamiento de distintos modelos u operaciones de análisis estadístico.

En la literatura revisada, se pudo observar que la tendencia de recolección de datos es principalmente orientada a instituciones sanitarias nacionales como [5-22] que proponen distintos modelos de predicción de infección por COVID 19 en distintos paises de mundo; siendo algunas otras propuestas de modelos predictivos de infección por COVID 19 como las de [23-25] las que utilizan un dataset centralizado y constantemente actualizado como los datasets propuestos por [3] o [12].

Las propuestas que utilizan datasets de las entidades médicas estatales directamente presentan una principal limitación de los datos ya que estos deben ser extraídos y organizados antes de poder ser tratados para adaptarlos a alguna técnica de machine learning; a diferencia de las propuestas que utilizan el dataset constantemente actualizado las cuales simplemente pueden realizar el mismo procedimiento de desarrollo o entrenamiento de modelo y obtener un modelo actualizado.

2.2. Modelos de predicción de índice de infección de COVID 19 utilizando técnicas estadísticas

La principal técnica utilizada para predecir el índice de infección por COVID 19 en distintos países del mundo entre estos autores [13-17], [19,23] es la de regresión polinomial teniendo en su mayoría grados polinomiales entre 5 a 11 con resultados bastante precisos, de forma que podemos clasificar la regresión polinomial como la técnica más popular entre las técnicas estadísticas para predecir el índice de infección por COVID 19. Sin embargo, [14, 26] nos explican que estos modelos son funcionales por cortos periodos de tiempo, puesto que con el cambio del comportamiento de la infección a través de tiempo, la predicción de estos modelos va siendo cada vez más imprecisa. Esta idea es reforzada por [27] que nos explica que su modelo matemático alcanza un grado alto de precisión debido al poco cambio de comportamiento de la infección en Japón, específicamente la periodicidad constante entre olas de contagios. Es por esto que [14] nos indica que para mantener la precisión de un modelo es preferible utilizar modelos epidemiológicos.

De esta forma encontramos a [7,9,14] quienes basándose en el modelo epidemiológico SEIR (Susceptible, Exposed, Infective and Recovered individuals) y aprovechando la facilidad de interpretar matemáticamente el comportamiento de la infección de COVID 19 en cualquier parte del mundo construyen sus propios modelos para predecir el comportamiento de la infección del virus obteniendo resultados bastante precisos.

Las métodos utilizados para validar los modelos son el coeficiente de determinación o R^2 con los siguientes resultados: [13] con resultados entre 0.769 a 0.998, [14] con los resultados de sus modelos no lineales entre 0.452 a 0.906, de su modelo de SEIR 0.882 y de sus modelos de regresión entre 0.064 a 0.91, [15] con 0.999 tanto para índice de infectados, recuperados y fallecidos, [17] con 0.983, [18] con 0.91, [19] con resultados entre 0.944 a 0.998, [23] con 0.999 y [24] con 0.996. También se utiliza el error cuadrático medio o RMSE [16] con resultados entre 300 a 600 y finalmente se utiliza el error porcentual [7] con resultados entre 0.29 a 0.15.

2.3. Modelos de predicción de índice de infección de COVID 19 utilizando técnicas de machine learning

En los trabajos de [5,6,8] y [10,20-22,25] podemos encontrar que proponen distintas arquitecturas de red neuronal para poder predecir el índice de infección por COVID 19 en distintos

países del mundo ; siendo el tipo de red neuronal más utilizada el de Multi Layer Perceptron(MLP), como es el caso de los trabajos de [5,6,25] y [22] pero estos modelos, a su vez, son los menos robustos o con menor consideración de los factores exógenos que influyen directamente en el índice de infección de COVID 19.

Por otro lado, resaltan de entre estas propuestas las investigaciones de [20,21] y [10] quienes proponen, a diferencia de los demás, arquitecturas de red neuronal recurrente, más específicamente, el modelo Long-Short Term Memory (LSTM) con el objetivo de predecir el índice de infección por COVID 19 en Estados Unidos y China por parte [10] adicionalmente propone una variación del modelo LSTM, siendo este el modelo BILSTM-GASVR que además de predecir el índice de infección por COVID 19 también tiene como objetivo predecir la cantidad de equipo UCI necesitado para combatir los efectos de la infección. De entre estos últimos trabajos se identificó una propuesta bastante robusta por parte de [21] quienes explican 4 premisas principales:

- La fuerza de infección influye en el índice de infección.
- La fuerza de infección se mide en base a medidas de tendencia central y medidas de dispersión.
- Un infectado solo puede contagiar durante 2 semanas.
- El centro laboral es el principal medio de contagio.

Premisas que son utilizadas para diseñar una arquitectura que considera intrínsecamente los factores humanos, generalmente ignorados [28], que influyen a la infección. Obteniendo, de esta manera, modelos de buena precisión y resultados duraderos.

Los métodos que se utilizaron para medir la precisión de estos modelos son el coeficiente de determinación o R^2 obteniendo resultados como [5] con resultados de 0.999, [6] con resultados entre 0.96 a 0.99, [8] con resultados de 0.999, [25] con resultados entre 0.66 a 0.999, [20] con 0.97, [22] con resultados entre 0.867 a 0.918. También existen propuestas que utilizan la métrica de error cuadrático medio o RMSE como la propuesta de [21] con un resultado de 981 o [10] con un resultado de 131.294.

3. Metodología

Para el desarrollo de este trabajo se realizó una comparación entre distintos métodos para seleccionar el mejor y que sea utilizado en el enfoque de re-entrenamiento, por lo que primero se explica el proceso de selección de este método y luego se continúa con la explicación de la metodología del enfoque de re-entrenamiento automático.

3.1. Determinación del mejor modelo de predicción de índice de COVID 19

En la revisión de la literatura se encontraron múltiples propuestas para predecir el índice de infección en distintos lugares del mundo. Entre los más resaltantes tenemos: la propuesta de [25] con un modelo de regresión Multi Layer Perceptron (MLP) para predecir la cantidad de infectados por COVID 19 en varios países. Para lograrlo utilizaron una entrada de 14 “timesteps” siendo el índice de infección de los 14 días anteriores de la fecha que se busca predecir y adicionalmente en una evaluación de algoritmos de optimización siendo el algoritmo “NAdam” el que ofrecía mejores resultados. [22] propone igualmente un modelo MLP para predecir la infección y el índice de fallecidos por COVID 19 en Bangladesh, además utiliza el método Support Vector Regression (SVR) demostrando por medio de sus resultados que la arquitectura MLP resulta ser más eficiente entre las dos. [8] refuerza la idea de que MLP es el modelo que ofrece mejores resultados para la predicción del índice de infección por COVID 19, contextualizando el problema en México y obteniendo una precisión que tiende al 100%; sin embargo, esta precisión es opacada por el error cuadrático medio que resulta alto por el desarrollo de modelo con datos del índice de infección acumulado.

Por otro lado, [20], compararon 3 modelos de predicción (Crecimiento Logístico, Regresión Polinomial y Redes Neuronales Long Short-Term Memory) de infección por COVID 19 y determinaron que el modelo LSTM es el más

ISSN 2007-9737

Tabla 1. Premisas de diseño de entrada [21]

Número	Premisa
1	La infección depende de la fuerza del contagio en el momento
2	Las series temporales del índice de infección otorgan información de la fuerza del contagio en el momento
3	Se puede asumir que un infectado solo puede contagiar durante 14 días porque es el tiempo promedio hasta su aislamiento
4	El área de trabajo es un medio crítico de transmisión del virus por lo que es más probable que una infección ocurra durante un día de semana

eficiente y preciso de entre los 3. [21] propone de mismo modo una red neuronal LSTM para predecir el índice de infección por COVID 19 en Estados Unidos y además fundamenta de mejor manera la decisión de [25] de utilizar como entrada la infección de los 14 días anteriores a la fecha que se busca predecir agregando 3 valores más: la media de estos 14 días, la desviación estándar de estos 14 y un valor que indique si es día de semana o no. Este último punto es explicado por medio de 4 premisas principales (ver Tabla 1).

De esta forma se obtuvieron un modelo bastante preciso, medido en base al error cuadrático medio y obteniendo un valor de 981.

En este trabajo se optó por utilizar la arquitectura LSTM para la predicción del índice de infección por COVID 19 de [21] tomando en cuenta el análisis de los resultados de las distintas propuestas mencionadas, así como también la consideración de ciertos puntos como el fundamento de sus propuestas, velocidad de desarrollo y uso de recursos eficiente; siendo la propuesta de [21] la que más destacó entre las otras propuestas en las características anteriormente mencionadas.

A esta arquitectura se le eliminó de la entrada el valor que indica si es día de semana o no, pues esto no es relevante en el contexto latinoamericano, ya que la mayoría de la población laboralmente activa es informal y no se cumplen con horarios de trabajo semanales [22]. Adicionalmente se seleccionó como algoritmo de optimización el algoritmo “NAdam” en base a las comparaciones realizadas por [25].

3.2. Metodología de producción de modelos de infección por COVID 19 con re-entrenamiento automático

Para el desarrollo de esta investigación se propone el procedimiento mostrado en la Figura 1.

Donde se inicia con la extracción de los datos del dataset de índice de infección por COVID 19 en distintos países, centralizado y constantemente actualizado propuesto por [12].

El preprocesamiento necesario de estos datos para darle la forma necesaria para el entrenamiento del modelo de predicción de infección por COVID 19 en cualquier país de interés que contenga el dataset, posteriormente se particiona en conjuntos de entrenamiento y validación para finalmente pasar al entrenamiento, validación y producción del modelo de predicción del índice de infección por COVID 19 en cualquier país de interés que tiene como entrada la información de infección en 14 días anteriores a la fecha que se busca predecir y tiene como salida el índice de infección de la fecha que se busca predecir.

Este proceso está programado para ser repetido cada semana y de esta forma se producen modelos cada vez más actualizados con la información de la infección. A continuación se explicará con más detalle cada uno de estos pasos.

3.2.1. Dataset centralizado y constantemente actualizado

El conjunto de datos utilizado es el propuesto por [12] que posee la ventaja actualizar diariamente los datos utilizados en esta investigación. En este dataset encontramos distintos grupos de métricas relacionadas a la infección de COVID 19 en el mundo (véase Tabla 2).

Siendo estos grupos de métricas conjuntos de atributos relacionados a las vacunaciones, tests realizados y sus resultados, hospitalizaciones, casos confirmados, muertes confirmadas, índices de reproducción del virus, métricas de respuestas políticas y otras variables de interés.

Estos datos están organizados en países del mundo y ordenados por las fechas del inicio de la

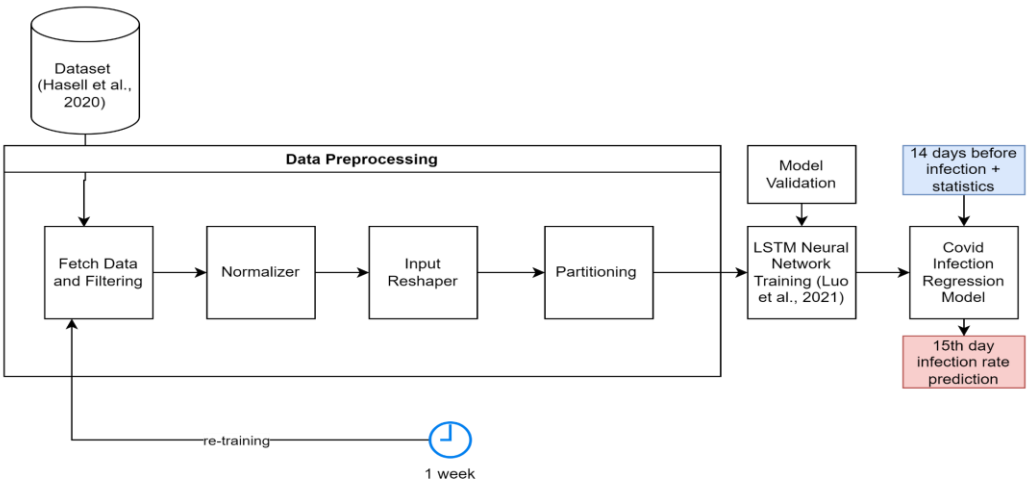

Fig. 1. Metodología de producción de modelos y re-entrenamiento

Tabla 2. Grupos de atributos [12]

Metrics	Source	Updated	Countries
Vaccinations	Official data collated by the Our World in Data team	Daily	218
Tests & positivity	Official data collated by the Our World in Data team	Weekly	193
Hospital & ICU	Official data collated by the Our World in Data team	Daily	47
Confirmed cases	JHU CSSE COVID-19 Data	Daily	217
Confirmed deaths	JHU CSSE COVID-19 Data	Daily	217
Reproduction rate	Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C	Daily	192
Policy responses	Oxford COVID-19 Government Response Tracker	Daily	187
Other variables of interest	International organizations (UN, World Bank, OECD, IHME…)	Fixed	241

infección hasta la actualidad, siendo actualizado cada día. En el caso de este trabajo de investigación se hizo uso del grupo de atributos “Confirmed Cases” (Véase Tabla 2).

Como atributos el grupo “Confirmed Cases” contiene (Véase Tabla 3) los datos del índice de infección acumulada (total_cases), nuevos infectados por día (new_cases), nuevos infectados por día suavizado (new_cases_smoothed) y una variación de estos tres campos por millón (total_cases_per_million, new_cases_per_million y new_cases_smoothed_per_million) de todos los países de los cuales recolecta datos de los reportes oficiales estatales a través de web scraping.

Como se puede observar en la Tabla 3 este dataset contiene campos tratados o suavizados para ser utilizado en análisis estadístico o machine learning, estos campos son modificaciones de los índices de infección y mortalidad que permiten suavizar las curvas del comportamiento de cada conjunto de valores. Debido a que el suavizado de datos nos permite encontrar más fácilmente patrones en el comportamiento de la infección se

ISSN 2007-9737

Tabla 3. Variables del grupo “Confirmad Cases” [12]

Variable	Description
total_cases	Total confirmed cases of COVID-19. Counts can include probable cases, where reported.
new_cases	New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported.
total_cases_per_million	Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.
new_cases_per_million	New confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.
new_cases_smoothed_per_million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable cases, where reported.

utilizó para el entrenamiento del modelo los campos “new_cases_smoothed” y “date” para conocer la fecha en que se recolectaron dichos datos.

3.2.2. Preprocesamiento de datos

Para obtener los datos [12] proporciona una url para descargar directamente el csv actualizado del dataset que contiene los datos de infección ordenados por países. De esta forma este dataset es filtrado según el país del que se busque desarrollar el modelo de predicción, además las columnas que no sean “date” y “new_cases_smoothed” son eliminadas.

Debido a que no existen registros de infección en todos los días de recolección de datos, existen nulos en el campo “new_cases_smoothed”, por ello se realiza un tratamiento de datos en base a los valores más cercanos; es decir, se reemplazarán los nulos del campo “new_cases_smoothed” con el valor de la siguiente fecha más cercana que contenga algún valor (Forward Fill) y si este valor nulo está ubicado en las últimas fechas y no existe una fecha siguiente con algún valor, se utilizará el valor de la fecha anterior más cercana que contenga un valor (Backward Fill).

Con el objetivo de realizar el entrenamiento de la red neuronal artificial, es necesario la normalización de los datos, por lo que se realizó una normalización Min-Max en el campo “new_cases_smoothed”.

Una vez normalizados, los datos que inicialmente están ordenados según las fechas en las que fueron recolectados, son agrupados en una colección de entrada y un valor de salida. De tal forma que la correspondencia entre el valor de salida y la colección de entrada sea la siguiente:

Si tenemos el valor “new_cases_smoothed” de un día “n” como valor de salida, entonces su entrada será la colección de los valores “new_cases_smoothed” de los 14 días anteriores a “n”, la media aritmética y desviación estándar de estos 14 valores (Véase Figura 2).

Esta transformación se realiza con el objetivo de que los datos tengan el formato de entrada de la arquitectura LSTM propuesta en este trabajo, basada en la propuesta de [21]. Es decir deben pasar de tener el formato extraido del dataset [12] para tener la forma indicada en la Figura 3.

3.2.3. Entrenamiento de red neuronal LSTM para predicción del índice de infección por COVID 19

Para la predicción del índice de infección por COVID 19 en distintos países del mundo se diseñó una arquitectura basada en la arquitectura propuesta por [21] (Véase Figura 4.)

Donde tenemos dos capas de LSTM que reciben como entrada el índice de infección de los 14 días anteriores a la fecha que se busca predecir, la media de infección de estos 14 días y su desviación estándar para finalmente predecir la infección del día buscado.

Es decir existe una capa de entrada con 16 nodos que recoge esa colección de entrada y los valores de infección de los 14 días anteriores, junto a las estadísticas pasan a la capa LSTM

date	new_cases_smoothed
…	…
n-2	I(n-2)
n-1	I(n-1)
n	I(n)

Input Reshape

date	Input[1]	Input[2]		Input[14]	Input[15]	Input[16]
15	I(1)	I(2)	…	I(14)	mean(I(1)...I(14))	std(I(1)...I(14))
…	…	…	…	…	…	…
n	I(n-14)	I(n-13)	…	I(n-1)	mean(I(n-14)...I(n-1))	std(I(n-14)...I(n-1))

Fig. 2. Transformación de los datos

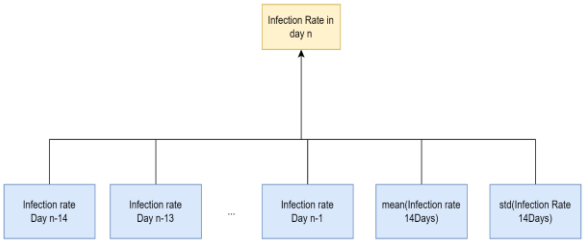

Fig. 3. Formato de entrada modificada [21]

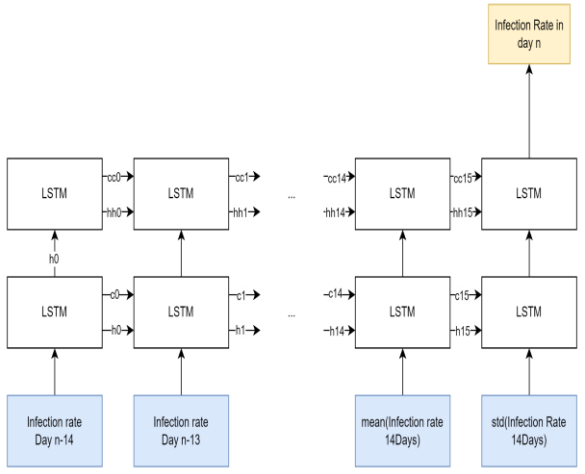

Fig. 4. Arquitectura de red neuronal basada en la propuesta de [21]

donde son las entradas principales de cada nodo LSTM; es decir, el primer nodo recibe la infección de la fecha más antigua de la colección, el segundo nodo el valor de infección del siguiente día y así hasta el nodo 14 donde se recibe el índice de infección del día anterior al que se busca predecir; en los dos últimos nodos LSTM (nodos 15 y 16) se recibe el valor de la media aritmética y la desviación estándar de la infección de los 14 días anteriores respectivamente.

Como se puede observar en la Figura 4, desde el nodo 2 al 16 de la primera capa reciben dos valores de entrada adicional llamados “c” y “h” que representan los vectores de estado (c) y las salidas de los nodos anteriores (h) de la misma capa, siendo este el mecanismo de la arquitectura que permite “recordar” a la red neuronal cual era el comportamiento de la infección en los días anteriores, mejorando de esta forma la capacidad de predecir la infección en el día buscado.

La segunda capa de nodos LSTM recibe como entrada directa los valores “h” de la capa anterior, y al igual que la primera capa LSTM desde el nodo 2 al 16 se recibe como entradas adicionales las salidas “c” y “h” de los nodos anteriores de la misma capa. Finalmente solo el último nodo de esta última capa LSTM envía su salida a una capa densa de un único nodo que nos da como resultado la predicción del índice de infección por COVID 19 en el día posterior a los 14 días utilizados como entrada.

En la Tabla 4 podemos observar el resumen de la arquitectura utilizada para la red neuronal artificial junto a su parametría. Finalmente se utilizó el algoritmo “NAdam” como algoritmo de optimización, acorde a la propuesta de [25], y como función de pérdida el error cuadrático medio (MSE), considerando también 100 “epochs” de entrenamiento para disminuir el tiempo de este proceso.

3.2.4. Predicción de infección por COVID 19

El modelo de predicción de infección por COVID 19 en un país de interés, resultante del flujo de entrenamiento, tiene un esquema de entrada que consiste en una colección de 16 valores numéricos relacionados a la infección en los días anteriores y un valor numérico como salida que viene a ser la predicción de índice de infección (Véase Figura 3).

ISSN 2007-9737

Tabla 4. Resumen Arquitectura

Capa	Hiper Parámetros
Input Layer	Nodos: 16
LSTM 1	Nodos: 16 Función de activación: tanh Función de activación recurrente: sigmoide Output: secuencia de datos (16)
LSTM 2	Nodos: 16 Función de activación: tanh Función de activación recurrente: sigmoide
Dense Layer	Nodos: 1 Función de activación: lineal

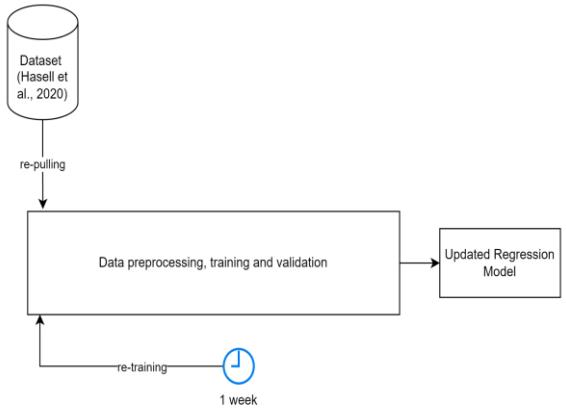

Fig. 5. Re-entrenamiento automático

De forma específica, si queremos predecir por ejemplo el índice de infección por COVID 19 del día 15/12/2022 en el Perú, es necesario conocer el índice de infección de los 14 días anteriores en el Perú, es decir de los días desde el 01/12/2022 al 14/12/2022, y ordenarlos como los primeros 14 elementos de la colección que es la entrada del modelo.

Posteriormente, se calcula la media aritmética y la desviación estándar del índice de infección de los 14 días que son parte de la colección y estos resultados son agregados a la colección en el mismo orden. Hasta ese momento ya tenemos una colección con 16 elementos que incluyen los índices de infección por COVID 19 en el Perú desde el 01/12/2022 hasta el 14/12/2022, la media aritmética y la desviación estándar de ese grupo

Tabla 5. Resultados en 1 mes Perú

Semana	R^2	RMSE	MAPE
1	0.987	468.143	0.189
2	0.992	404.548	0.139
3	0.986	606.971	0.314
4	0.991	466.169	0.293

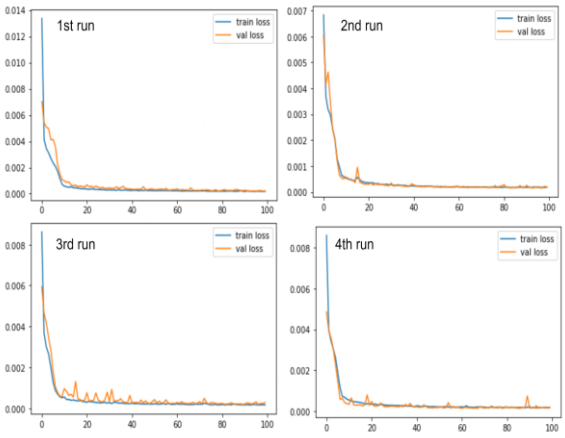

Fig. 6. Función de pérdida Perú

de índices de infección en el mismo orden. Al usar como entrada esta colección al modelo entrenado, este nos dará como resultado un valor numérico que viene a ser la predicción de infección por COVID 19 en el Perú en el día 15/12/2022.

3.2.5. Validación del modelo

Para validar el modelo de predicción de infección por COVID 19 resultante del flujo de entrenamiento se utilizaron las métricas de coeficiente de determinación o R^2, que nos sirve para conocer el ajuste de las predicciones del modelo con los datos reales de infección; la raíz cuadrada del error cuadrático medio o RMSE, que nos permite conocer una aproximación al promedio de la diferencia entre las predicciones y los datos reales; y finalmente el error porcentual absoluto medio o MAPE, que nos permite conocer la magnitud de error en términos porcentuales.

3.2.6. Re-entrenamiento automático

Con el fin de conservar la precisión del modelo resultante del proceso, este será re-entrenado cada semana con el objetivo de obtener

mínimamente la mitad de una entrada totalmente desconocida por el modelo (7 días nuevos). Dejando el flujo del entrenamiento del modelo guiado por un evento de tiempo que se activará cada semana (Véase Figura 5).

Este flujo nos retorna como resultado un modelo de predicción de infección por COVID 19 del país de interés con el que se haya entrenado y que tiene un proceso automático para ser reemplazado por otro modelo de predicción entrenado con nuevos datos cada semana. Siendo estos modelos resultantes capaces de predecir solamente el índice de infección de un día teniendo como información previa la infección de los 14 días anteriores.

4. Resultados

Para medir la capacidad de mantener la precisión y la constante actualización de los modelos, se desarrollaron tres flujos a partir del mismo modelo de la metodología. Es decir se realizó el procedimiento de producción de un modelo de predicción del índice de infección por COVID 19 en el Perú, Brasil y Chile. Procedimientos en los que se aplicó el enfoque de re-entrenamiento y los resultados de estos fueron recopilados a lo largo de 4 semanas.

Sin embargo, antes de revisar los resultados recolectados, se presentará el particionamiento del conjunto de datos para el entrenamiento y validación de los modelos.

4.1. Particionamiento del conjunto de datos

Luego de realizar el preprocesamiento de los datos, se tiene un conjunto de datos formateados de acuerdo a la entrada del modelo de predicción de infección por COVID 19. Este conjunto de datos se particionó en 2 grupos, un 30% del total del conjunto se agruparon como datos de validación del modelo y el otro 70% del conjunto de datos se utilizaron para el entrenamiento de modelo.

4.2. Modelo de predicción de infección por COVID 19 en el Perú

Este modelo tuvo como primer entrenamiento el día 05/10/22, siendo re-entrenado cada semana hasta la fecha del 26/10/22 teniendo como resultados las métricas R^2, RMSE y MAPE de cada entrenamiento. En promedio se obtuvo un coeficiente de determinación o R^2 de 0.989, una raíz del error cuadrático medio o RMSE de 486.458 y un error porcentual absoluto medio o MAPE de 0.234 al predecir el índice de infección por COVID 19 en el Perú. La figura 6 muestra con más detalle el progreso de la métrica MSE durante el entrenamiento tanto para los valores de entrenamiento como validación. Y la tabla 5 las métricas de cada entrenamiento.

4.3. Modelo de predicción de infección por COVID 19 en Brasil

Este modelo tuvo como primer entrenamiento el día 05/10/22, siendo re-entrenado cada semana hasta la fecha del 26/10/22 teniendo como resultados las métricas R^2, RMSE y MAPE de cada entrenamiento. En promedio se obtuvo un coeficiente de determinación o R^2 de 0.986, una raíz del error cuadrático medio o RMSE de 2146.362 y un error porcentual absoluto medio o MAPE de 0.286 al predecir el índice de infección por COVID 19 en Brasil. La figura 7 muestra con más detalle el progreso de la métrica MSE durante el entrenamiento tanto para los valores de entrenamiento como validación. Y la tabla 6 las métricas de cada entrenamiento.

4.4. Modelo de predicción de infección por COVID 19 en Chile

Este modelo tuvo como primer entrenamiento el día 05/10/22, siendo re-entrenado cada semana hasta la fecha del 26/10/22 teniendo como resultados las métricas R^2, RMSE y MAPE de cada entrenamiento. En promedio se obtuvo un coeficiente de determinación o R^2 de 0.996, una raíz del error cuadrático medio o RMSE de 191.173 y un error porcentual absoluto medio o MAPE de 0.217 al predecir el índice de infección por COVID 19 en Chile. La figura 8 muestra con más detalle el progreso de la métrica MSE durante el entrenamiento tanto para los valores de entrenamiento como validación. Y la tabla 7 las métricas de cada entrenamiento.

ISSN 2007-9737

Tabla 6. Resultados en 1 mes Brasil

Semana	R^2	RMSE	MAPE
1	0.980	2420.165	0.391
2	0.988	1982.678	0.329
3	0.986	2285.669	0.268
4	0.989	1896.935	0.154

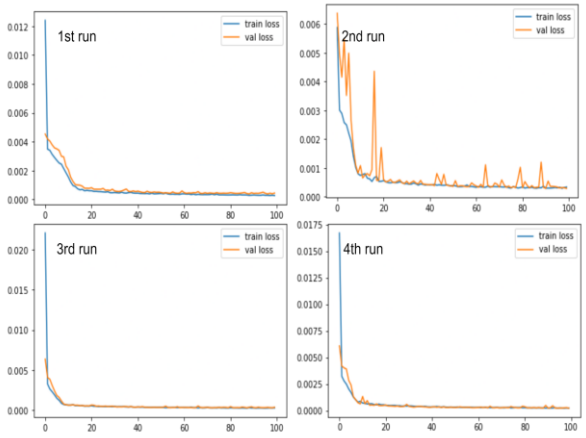

Fig. 7. Función de pérdida Brasil

Analizando los resultados obtenidos, podemos observar que la precisión de los 3 modelos se mantiene a través del tiempo, además podemos notar que la que obtiene mejores resultados en base al R^2 y RMSE es el modelo de Chile; por otro lado, el MAPE nos indica que este modelo posee valores precisos en las métricas de R^2 y RMSE debido al bajo índice de infección del virus en dicho país, esto es debido a que posee valores altos a pesar de la precisión de las otras métricas; de todos modos el modelo de Chile termina siendo el más preciso de entre los tres países. Finalmente podemos notar que el valor de la función de pérdida de los datos de validación durante el entrenamiento en Chile fluctúa, en 3 de sus 4 entrenamientos, en un rango bastante amplio a diferencia de Perú y Brasil que, según la gráfica, este valor se mantiene en un descenso constante.

4.5. Discusión

Este trabajo sirve como evidencia de la capacidad del enfoque de re-entrenamiento propuesto para mantener la precisión de un modelo de predicción, en este caso, de infección por COVID 19 en distintos países del mundo. Sin embargo, para el funcionamiento de un enfoque como este, es necesario de una fuente de datos de fácil acceso y constantemente actualizada como la propuesta de [12]. Aún así, la constante actualización del "conocimiento" de los modelos producidos por este enfoque y la gran precisión obtenida por la arquitectura de la red neuronal hacen que estos modelos puedan ser usados como herramientas de planificación protocolar estatal o particular por un largo periodo de tiempo y no en periodos estáticos como los modelos propuestos por [5-9,13-25].

Los resultados muestran que hay una mejora de la precisión en base al coeficiente de determinación al obtener en el Perú un 0.989, en Chile un 0.996 y en Brasil un 0.986 en comparación con el trabajo de [20] quien, con una arquitectura similar aplicada en Estados Unidos y Wuhan, obtiene un 0.97, y con respecto al RMSE para Perú, Chile y Brasil se obtuvieron en promedio 486.458, 191.173 y 2146.362, valores bastante altos en comparación de [21] que, siendo la base de la arquitectura propuesta en este trabajo, obtuvo valores más bajos como 900 en un país con mayor cantidad de habitantes como es Estados Unidos. Esto último se debe a que la fecha en la que se realizó dicha investigación aún no habían sucedido las mayores olas de infectados en dicho país y por lo tanto el entrenamiento se realizó con valores de infección mucho menores a los considerados en esta investigación.

Pasando a comparaciones más directas en base a los países utilizados, tenemos el trabajo de [29] que utilizando Support Vector Machine y MLP obtiene resultados en base al MAPE de 0.18 en Perú y 0.16 para Chile, mientras este trabajo obtuvo 0.234 y 0.217 respectivamente para cada país, siendo estos valores bastante similares a pesar de la variabilidad de los valores de infección considerados por los modelos por la diferencia de fechas en las que se realizó el entrenamiento. [30] considerando fechas con altos índices de infección pasadas las primeras olas, obtiene como mejor resultado para predecir la infección por COVID 19 en Chile un RMSE de 635 siendo este valor menos preciso que el 191.173 en promedio obtenido por los modelos propuestos en esta investigación.

Tabla 7. Resultados en 1 mes Chile

Semana	R2	RMSE	MAPE
1	0.996	201.660	0.471
2	0.995	189.570	0.065
3	0.997	185.448	0.239
4	0.996	188.015	0.094

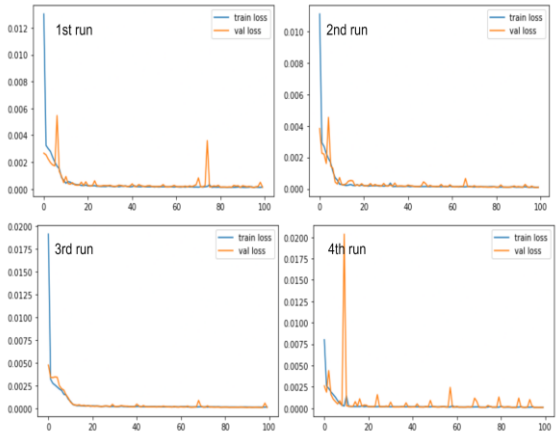

Fig. 8. Función de pérdida Chile

En el trabajo [31], quienes desarrollan, utilizando en enfoque de aprendizaje incremental y re-entrenamiento automático, una propuesta de predicción de curva epidemiológica (índice de infección, índice de curados e índice de fallecidos) de COVID 19 en cualquier país del mundo, obtuvieron como resultados en Croacia durante un re-entrenamiento diario a lo largo de 30 días un coeficiente de determinación de 0.999 en el índice de infección acumulado, 0.999 en el índice de curados acumulados, 0.999 en el índice de muertos; derivando finalmente en una predicción de curva epidemiológica con un coeficiente de determinación de 0.996, similar al resultado del modelo de predicción de COVID en Chile propuesto en este trabajo, a pesar de la mayor densidad poblacional chilena, la cual afecta negativamente a la precisión de los modelos.

Este trabajo usa la arquitectura de red neuronal propuesta por [21] aplicando un enfoque de re-entrenamiento que permite producir modelos que mantengan la precisión durante un largo periodo de tiempo y conozcan los distintos comportamientos de la infección conforme la infección se desarrolla en el mundo. Además, se modifica la arquitectura para tener una entrada más acorde al contexto latinoamericano, sin dejar de considerar las premisas que el mismo autor presenta en su trabajo.

5. Conclusiones

En el presente estudio se propone un enfoque de re-entrenamiento automatizado de modelos de predicción, específicamente, de índice de infección por COVID 19 en cualquier país del mundo. El enfoque, como se pudo observar en los resultados, mantuvo la precisión por encima del 95% de los modelos resultantes del flujo automatizado del re-entrenamiento durante un periodo de tiempo de 1 mes obteniendo de esta forma una herramienta de planificación estratégica o protocolar con precisión duradera y además un modelo o guía de re-entrenamiento automático a cualquier modelo de machine learning que se desee desarrollar; siempre y cuando se tenga una fuente de datos constantemente actualizada. Este último punto puede ser de interés para aquellos trabajos enfocados en aprendizaje basado en incrementos o también llamado “Online Learning” y hasta ajustes de modelos pre-entrenados a través de “Transfer Learning”, buscando siempre que los modelos conozcan nuevos datos que se van generando con el tiempo y sean de importancia o aprovechables por los modelos de predicción.

References

1. **World Health Organization. (2020).** Impact of COVID-19 on People’s Livelihoods, Their Health and Our Food Systems. World Health Organization.
2. **Jones, B. (2022).** The Changing Political Geography of COVID-19 Over the Last Two Years. Pew Research Center - U.S. Politics & Policy.
3. **Dong, E., Du, H., Gardner, L. (2022).** An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. Lancet Infectious Diseases, Vol. 20, No. 5, pp. 533–534. doi: 10.1016/S1473-3099(20)30120-1.

ISSN 2007-9737

4. **El Comercio. (2022).** Servicios saturados en la tercera ola: ¿cuál es la situación en las aseguradoras privadas? El Comercio Perú.
5. **Dhamodharavadhani, S., Rathipriya, R., Chatterjee, J.M. (2022).** COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models. Frontiers in Public Health, Vol. 8. doi: 10.3389/fpubh.2020. 00441.
6. **Shawaqfah, M., Almomani, F. (2022).** Forecast of the Outbreak of COVID-19 Using Artificial Neural Network: Case Study Qatar, Spain, and Italy. Results in Physics, Vol. 27, pp. 104484. doi: 10.1016/j.rinp.2021.104484.
7. **Yu, X., Lu, L., Shen, J., Li, J., Xiao, W., Chen, Y. (2022).** RLIM: A Recursive and Latent Infection Model for the Prediction of US COVID-19 Infections and Turning Points. Nonlinear Dynamics, Vol. 106, No. 2, pp. 1397–1410. doi: 10.1007/s11071-021-06520-1.
8. **Torrealba-Rodriguez, O., Conde-Gutiérrez, R.A., Hernández-Javier, A.L. (2022).** Modeling and Prediction of COVID-19 in Mexico Applying Mathematical and Computational Models. Chaos, Solitons & Fractals, Vol. 138, pp. 109946. doi: 10.1016/j.chaos.2020.109946.
9. **Santosh, K.C. (2022).** COVID-19 Prediction Models and Unexploited Data. Journal of Medical Systems, Vol. 44, No. 9. doi: 10.1007/s10916-020-01645-z.
10. **Zhang, W., Li, X. (2024).** A Data-Driven Combined Prediction Method for the Demand for Intensive Care Unit Healthcare Resources in Public Health Emergencies. BMC Health Services Research, Vol. 24, No. 1, pp. 477. doi: 10.1186/s12913-024-10955-8.
11. **Hassan, S.A.Z. (2024).** An AI Healthcare Ecosystem Framework for Covid-19 Detection and Forecasting Using CronaSona. Medical & Biological Engineering & Computing, Vol. 62, No. 7, pp. 1959–1979. doi: 10.1007/s11517-024-03058-3.
12. **Hasell, J. (2022).** A Cross-Country Database of COVID-19 Testing. Scientific Data, Vol. 7, No. 1. doi: 10.1038/s41597-020-00688-8.
13. **Kashif, M. (2022).** Estimation of Death and Recovery Rates Using Covid-19 Data of Pakistan. Eurasian Journal of Medicine and Oncology. doi: 10.14744/ejmo.2021.89947.
14. **Omara, T., Harby, K.A. (2022).** Using Mathematical and Statistical Model to Forecast the Path of Infection by COVID-19 in the Kingdom of Saudi Arabia. IIUM Medical Journal Malaysia, Vol. 20, No. 2. doi: 10.31436/imjm.v20i2.1683.
15. **Chanchí Golondrino, G.E., Campo Muñoz, W.Y., Sierra Martinez, L.M. (2022).** Aplicación de la regresión polinomial para la caracterización de la curva del COVID-19, mediante técnicas de machine learning. Investigación e Innovación en Ingeniería, Vol. 8, No. 2, pp. 87–105. doi: 10.17081/invinno. 8.2.4103.
16. **Pérez Abreu, R., Estrada, S., de-la-Torre-Gutiérrez, H. (2022).** A Two-Step Polynomial and Nonlinear Growth Approach for Modeling COVID-19 Cases in Mexico. Mathematics, Vol. 9, No. 18, pp. 2180. doi: 10.3390/math9182180.
17. **Nikhil, Saini, A., Panday, S., Gupta, N. (2022).** Polynomial Based Linear Regression Model to Predict COVID-19 Cases. In 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). doi: 10.1109/RTEICT52294.2021. 9574032.
18. **Square Platform LLC, Saqib, M. (2022).** Forecasting COVID-19 Outbreak Progression Using Hybrid Polynomial-Bayesian Ridge Regression Model. doi: 10.21203/rs.3.rs-75292/v1.
19. **Cortés Martínez, A.E., Becerra Huertas, C.E. (2022).** Caracterización de la tendencia del COVID-19 en Colombia con regresiones polinomiales. Gerencia y Políticas de Salud, Vol. 20, pp. 1–12. doi: 10.11144/javeriana.rgps20.ctcc.
20. **Sun, T. (2022).** Analysis of COVID-19 Based on Several Machine Learning Techniques. Journal of Physics: Conference Series, Vol. 1827, No. 1, pp. 012083. doi: 10.1088/1742-6596/1827/1/012083.

21. **Luo, J., Zhang, Z., Fu, Y., Rao, F. (2022).** Time Series Prediction of COVID-19 Transmission in America Using LSTM and XGBoost Algorithms. Results in Physics, Vol. 27, pp. 104462. doi: 10.1016/j.rinp.2021.104462.
22. **Nayan, A.-A., Kijsirikul, B., Iwahori, Y. (2022).** Coronavirus Disease Situation Analysis and Prediction Using Machine Learning: A Study on Bangladeshi Population. International Journal of Electrical and Computer Engineering, Vol. 12, No. 4, pp. 4217–4227. doi: 10.11591/ijece.v12i4.pp4217-4227.
23. **Singh, M., Dalmia, S. (2022).** Prediction of Number of Fatalities Due to Covid-19 Using Machine Learning. In 2020 IEEE 17th India Council International Conference (INDICON). doi: 10.1109/INDICON49873.2020.9342390.
24. **Díaz Pinzón, J.E. (2022).** Predicción del COVID-19 a nivel mundial para el año 2021. Revista Repertorio de Medicina y Cirugía, pp. 131–137. doi: 10.31260/repertmedcir.01217372.1143.
25. **Wieczorek, M., Siłka, J., Woźniak, M. (2022).** Neural Network Powered COVID-19 Spread Forecasting Model. Chaos, Solitons & Fractals, Vol. 140, pp. 110203. doi: 10.1016/j.chaos.2020.110203.
26. **Tang, C., Todo, Y., Kodera, S., Sun, R., Shimada, A., Hirata, A. (2024).** A Novel Multivariate Time Series Forecasting Dendritic Neuron Model for COVID-19 Pandemic Transmission Tendency. Neural Networks, Vol. 179, pp. 106527. doi: 10.1016/j.neunet.2024.106527.
27. **Manabe, H. (2024).** Simple Mathematical Model for Predicting COVID-19 Outbreaks in Japan Based on Epidemic Waves with a Cyclical Trend. BMC Infectious Diseases, Vol. 24, No. 1, pp. 465. doi: 10.1186/s12879-024-09354-5.
28. **Qu, Z., Zhang, B., Wang, H. (2023).** A Multivariate Deep Learning Model with Coupled Human Intervention Factors for COVID-19 Forecasting. Systems, Vol. 11, No. 4, pp. 201. doi: 10.3390/systems11040201.
29. **Jojoa, M., Garcia-Zapirain, B., MDPI, A.G. (2022).** Forecasting COVID 19 Confirmed Cases Using Machine Learning: The Case of America. doi: 10.20944/preprints202009.0228.v1.
30. **Barría-Sandoval, C., Ferreira, G., Benz-Parra, K., López-Flores, P. (2022).** Prediction of Confirmed Cases of and Deaths Caused by COVID-19 in Chile Through Time Series Techniques: A Comparative Study. PLOS ONE, Vol. 16, No. 4, pp. e0245414. doi: 10.1371/journal.pone.0245414.
31. **Šegota, S. et al. (2022).** Automated Pipeline for Continual Data Gathering and Retraining of the Machine Learning-Based COVID-19 Spread Models. EAI Endorsed Transactions on Bioengineering and Bioinformatics, Vol. 1, No. 3, pp. 169582. doi: 10.4108/eai.4-5-2021.169582.

Article received on 02/12/2023; accepted on 14/12/2025.
**Corresponding authors is David Calderón Vilca.*