Revista Amazonía Digital

Vol. 1 Núm. 2: e195 (2022)
https://doi.org/10.55873/rad.v1i1.195
e-ISSN: 2810-8701
Universidad Nacional Amazónica de Madre de Dios

CC BY

Artículo de revisión / Review article

Análisis de datos y pronóstico de casos de la Covid-19 en el departamento de Madre de Dios de Perú utilizando técnicas LSTM

Data analysis and forecast of Covid-19 cases in the Madre de Dios Department of Peru using LSTM techniques

Jose Carlos Navarro-Vega [1*]; Nelly Jacqueline Ulloa-Gallardo [1]; Diego Raphael Paz-Bustamante [1]; Diego Gustavo Zegarra-Conde [1]; Wilder Nina-Choquehuayta [1]

[1]Universidad Nacional Amazónica de Madre de Dios, Madre de Dios, Perú

Recibido: 10/05/2022
Aceptado: 19/06/2022
Publicado: 25/07/2022

*Autor de correspondencia: jcnavarro@unamad.edu.pe

Resumen: Actualmente la Covid-19 está causando grandes pérdidas a nivel mundial, es por ello que diferentes trabajos permiten predecir o pronosticar el comportamiento de cantidad de infectados utilizando técnicas de forecasting, dentro del campo de Inteligencia Artificial se están permitiendo tomar medidas de control en los diferentes países. En este trabajo se propuso un modelo de aprendizaje profundo para pronosticar los casos diarios en las regiones de Madre de Dios. Los datos utilizados pertenecen al conjunto de datos abiertos Covid-19 del Ministerio de Salud de Perú (MINSA). El conjunto de datos incluye los períodos de inicio de marzo de 2020 a fines de diciembre de 2021. Se utilizó un LSTM utilizando variables de Fecha, Departamento, Provincia, Distrito, Casos, IP. ID y con tamaño de ventana de 5 días, se obtuvo una precisión de 94,67% con los datos de entrenamiento y un 92,31%.

Palabras clave: Covid-19; LSTM; infectados; pandemia; pronóstico

Abstract: Currently, Covid-19 is causing great losses worldwide, which is why different works that allow predicting or forecasting the behavior of the number of infected using forecasting techniques within the Artificial Intelligence field are allowing control measures to be taken in the different countries. In this work, a deep learning model was proposed to forecast daily cases in the regions of Madre de Dios. The data used belongs to the covid-19 open data set, from the Peruvian Ministry of Health (MINSA). The data set includes the periods from the beginning of March 2020 to the end of December 2021. An LSTM was used using variables of Date, Department, Province, District, Cases, IP. ID and with a window size of 5 days, an accuracy of 94.67% was obtained with the training data and 92.31%.

Keywords: Covid-19; LSTM; infected; pandemic; forecast

Cómo citar / Citation: Navarro-Vega, J.C., Ulloa-Gallardo, N.J., Paz-Bustamante, D. R., Zegarra-Conde & D.G., Nina-Choquehuayta, W. (2022). Análisis de datos y pronóstico de casos de la Covid-19 en el departamento de Madre de Dios de Perú utilizando técnicas LSTM. *Revista Amazonía Digital*, 1(2), e195. https://doi.org/10.55873/rad.v1i1.195

1. Introducción

El 11 de marzo de 2020, la Organización Mundial de la Salud (OMS) declaró al virus *SARS-CoV-2* (nombre científico de la enfermedad de coronavirus 2019 - COVID-19) como pandemia mundial. El alarmante aumento de pacientes infectados y fallecidos por el mortal virus, se ha visto afectado en casi todos los países. Para diseñar las mejores estrategias y la toma de decisiones de manera oportuna es fundamental recurrir a los modelos de predicción que, analizan hechos pasados para mejorar las predicciones sobre hechos que ocurrirán en el futuro, y proporcionan información para estar prevenidos frente a posibles amenazas (Shinde et al., 2020). A través del COVID-19 se han podido percibir los numerosos problemas que atraviesan las personas en el mundo. Los impactos negativos del virus en diversos aspectos en la vida del ser humano se ven afectados fundamentalmente en la salud y economía. El pronóstico preciso de la cantidad de casos confirmados puede mejorar la toma de decisiones a los gobiernos sobre las intervenciones de manera oportuna (Lalmuanawma et al., 2020). Los científicos y especialistas de la salud están en la búsqueda de una nueva tecnología a fin de ayudar a reducir la pandemia de COVID-19. La aplicación de la Inteligencia Artificial (AI) y Machine Learning (ML) ha permitido a los investigadores darle un nuevo enfoque en la lucha contra un nuevo brote de Coronavirus. Su objetivo fue analizar la importancia y el rol que tomarán la IA y ML como herramientas tecnológicas y métodos en la detección, el pronóstico y el seguimiento de casos de la COVID-19 (Lalmuanawma et al., 2020).

Existen estudios como de Arora et al. (2020) que, utilizó deep learning para pronosticar el número de casos positivos de la COVID-19 en la India, clasifica los estados por diferentes zonas según la propagación de casos positivos y facilita la identificación de puntos calientes de la COVID-19. También, Ayyoubzadeh et al. (2020) pronostica la incidencia de la COVID-19 en Irán mediante deep learning, permitiendo predecir los nuevos casos diarios empleando el modelo de regresión lineal y memoria a corto y largo plazo (acrónimo en inglés LSTM). Fanelli y Piazza (2020) señalan que, analizaron la evolución en el tiempo de la COVID-19 en Italia, Francia y China. Se basaron en el modelo de campo medio simple para recopilar información sobre la propagación de la epidemia, de manera específica sobre el tiempo y el pico más alto de infectados confirmados, pronosticaron que 2,500 unidades de ventilación deberían considerar en su planificación, las autoridades de Italia.

Para 10 países altamente y densamente poblados desarrollaron un sistema de predicción de los brotes de la COVID-19, utilizaron 9 algoritmos diferentes de aprendizaje automático y desarrollaron un conjunto de modelos a fin de predecir el incremento de nuevos casos, obtuvieron una precisión promedio de 87,9% ± 3,9% en los 10 países. Los modelos de predicción propuestos pueden ayudar a las partes interesadas a planificar ante un súbito brote y garantizar una mejor gestión de los recursos (Khakharia et al., 2021).

Según Wang et al. (2020) señalan que, hicieron uso de los datos publicados de la COVID-19 por la Universidad Johns Hopkins, para desarrollar un modelo de pronóstico mediante el método de Deep learning (aprendizaje profundo), en base al conjunto de entrenamiento de casos confirmados construyeron un modelo mejorado basado en la memoria a corto plazo (LSTM), con un pronóstico de precisión solo en los próximos 30 días, la integración de un mecanismo de actualización continua con LSTM que permite obtener proyecciones a largo plazo. Pronostican en Perú que la pandemia continuará hasta noviembre del 2020, en Irán la caída del número de casos positivos por día sea menor a 1 000 en noviembre, y en Rusia habrá aumentos más de 2 000 en diciembre. Finalmente, señalan que las decisiones tomadas por el gobierno, en el control estricto puede reducir de manera significativa la transmisión del COVID-19. Por otro lado, Sujath et al. (2020) presenta un modelo que podría ser de utilidad en las predicciones referente a la propagación de la COVID-19. Utilizaron los modelos de regresión lineal (RL), perceptrón multicapa (MLP) y el vector Auto-Regresión (VAR) que fueron aplicados con los datos de la página de Kaggle. Concluyeron que, el método MLP presentó mejores resultados de predicción

en comparación con LR y VAR utilizando las aplicaciones de WEKA y Orange, que permitió predecir los patrones potenciales de los efectos de la COVID-19 en la India.

El Perú, se encamina a combatir con esta pandemia que estamos librando, la COVID-19, este virus mortal para las personas, el cual no mide condición social, raza, ni situación económica. En la actualidad, donde tenemos un mundo globalizado, y con el avance de la alta tecnología y de conocimiento especializado, se requiere elaborar un proyecto tecnológico para poder combatir la COVID-19. Como se conoce la COVID-19 atacó diferentes departamentos de nuestro Perú, el sector salud, SALUD DIRESA Madre de Dios, cada vez es más importante el procesamiento de datos de personas que se contagiaron con COVID-19 en sectores públicos y privados mediante herramientas digitales. Y cada vez son más variadas las fuentes y su naturaleza. Se tratan datos semiestructurados de registros médicos que fueron digitalizados, de historias clínicas.

En este contexto, la presente investigación tiene como finalidad realizar una análisis de los datos publicados por el Ministerio de Salud (MINSA) para obtener insights y luego proponer un modelo de pronóstico para los casos positivos de COVID-19 utilizando mediante el uso de técnicas de deep learning en el departamento de Madre de Dios, la técnica a usar será LSTM, posteriormente para un mejor análisis se realizará un mapa sectorizado con el pronóstico de casos positivos de COVID-19, clasificado por categorías en función al número de casos positivos (leve, moderada y severa). Según los resultados obtenidos las autoridades y funcionarios de la Dirección Regional de Madre de Dios tomarán las mejores decisiones para aplicar estrategias de prevención y contener el aumento del número de pacientes con el nuevo coronavirus.

2. Materiales y métodos

En la Tabla 1 se observa diferentes artículos científicos revisados y analizados, cada artículo tiene el tipo de metodología utilizada para pronosticar datos recopilados de la plataforma de datos abiertos proporcionados por el Ministerio de Salud de cada país, y el porcentaje de efectividad obtenido con cada modelo presentado. En el primer artículo, Khakharia et al. (2021) proponen utilizar nueve técnicas de Machine Learning: Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regressor (SVR), Linear Regressor polynomial (LRP), Bayesian Ridge Regression (BRR), Li- ear Regression (LR), Random Forest Regressor (RFR), Holt-Winter Exponential Smoothing (HW), and Extreme Gradient Boost Regressor (XGB), donde ARMA con el país de Ethiopia se tuvo mayor precisión llegando a 99,93%.

En el segundo artículo, Sujath et al. (2020) proponen tres técnicas: Linear regression (LR), multilayer perceptron (MLP), and VAR, donde se puede visualizar que los tres métodos utilizados obtuvieron el mismo porcentaje de efectividad, 95%. En el tercer artículo, Kafieh et al. (2021) proponen tres técnicas: multilayer perceptron, random forest y different versions of long short-term memory (LSTM) y cuatro métricas Including Mean Average Percentage Error (MAPE), Root Mean Square Error (RMSE), Normalized RMSE (NRMSE) y R2. El mejor desempeño se encontró para una versión modificada de LSTM, denominada M-LSTM (modelo ganador), para pronosticar la trayectoria futura de la pandemia en los países mencionados. Se puede concluir que el método MLP está dando mejores resultados de predicción que los métodos LR y VAR.

En el cuarto artículo, Fanelli y Piazza (2020) analizan la dinámica temporal del brote de la enfermedad por coronavirus 2019 en China, Italia y Francia en la ventana de tiempo del 22/01 al 15/03/2020, en donde sus experimentos estiman que 2500 unidades de ventilación deberían representar una cifra justa para el requisito máximo que deben considerar las autoridades sanitarias de Italia para su planificación estratégica. En el quinto artículo, Mendieta M. (2020) se propone dos técnicas: Modelos Lineales (MCL) y no lineales (MCNL). Se pudo ver que el método MCNL, obtuvo un 96,3 % de efectividad en el tema de predicción. En el sexto artículo, Garrido et al. (2022) proponen un modelo: Susceptible, Expuesto, Infectado y Recuperado (SEIR). Se puede

obtener que el modelo pudo ser capaz de obtener un procedimiento claro de la evolución del proceso hospitalario el cual permite valorar de forma cualitativa la evolución de la pandemia en España.

En el séptimo artículo, Franco et al. (2020) plantean tres modelos: SIR (Susceptible-Infectious-Recovered), SEIR-Extendido y log-lineal. Se puede afirmar que, las aproximaciones que se presentan a partir de los datos y modelos propuestos no necesariamente predicen por completo el comportamiento de la pandemia, debido a que estos pueden ser afectados por diversos factores y variables (sociales, económicas y culturales) que no pueden ser introducidos dentro del modelo, además la variación de los datos es constante, lo que puede cambiar de forma drástica las tendencias y proyecciones. En el octavo artículo, Cocconi & Roark (2020) analizan dos modelos: LG (Regresión Logística generalizada) y Modelo de Gompertz. Se pudo obtener al momento de validar los modelos propuestos, que en ambos casos una precisión aproximada de un 99%.

En el noveno artículo, Cruz-Mendoza et al. (2020) proponen dos modelos: Recurrent Neural Network (RNN) y LSTM utilizando MATLAB y GOOGLE COLAB. Se pudo obtener resultados que muestran que el mejor enfoque de capacitación se obtiene con Colab, y el mejor enfoque de prueba se obtiene con MATLAB. En el décimo artículo, Aguilar I. et al. (2021) plantea cinco modelos: Red neuronal de convolución temporal (TCN), MAE, MAD, MSLE y RMSLE. Se pudo obtener que el modelo presentado puede ser utilizado por cualquier región como una herramienta para evaluar las tendencias dinámicas en los casos diarios de *SARS-CoV-2*. Además, el modelo puede ser aplicado como parte de la toma de decisiones de políticas.

Tabla 1. Revisión de técnicas de predicción de casos de COVID-19

Artículo	Técnicas	Dataset/países	Evaluación	Resultado
Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning (Khakharia et al., 2021).	Arima, Arma, Brr, Hw, Lrp, Lr, Rfr, Svr, Xgr	Bangladesh, India, China, Pakistan, Germany, Nigeria, Ethiopia.	Accuracy	87,9% ± 3,9% 99,93%
A machine learning forecasting model for COVID-19 pandemic in India (Sujath et al., 2020).	Lr, Mlp Y Var	India	Confidence Interval (CI)	95 %
COVID-19 in Iran: Forecasting Pandemic Using Deep Learning (Kafieh et al., 2021).	Mape, Rmse, Nrmse, R2 Mlp, Lr Y Var	Irán, Germany, Italy, Japan, Korea, Switzerland	Confidence Interval (CI)	95 %
Analysis and forecast of COVID-19 spreading in China, Italy and France. (Fanelli & Piazza, 2020).	(Sird)	China, Italy y France	Infectivity	80%−90%
Estudio sobre modelos predictivos para la COVID-19 en Cuba Study on predictive models for COVID-19 in Cuba (Mendieta et al., 2020).	Mcl Y Mcnl	CUBA	Efectividad	96,3%
Mathematical model optimized for prediction and health care planning for COVID-19 (Garrido et al., 2022).	Seir	España	Intervalo de confianza.	95 %
Modelos de predicción del impacto y evolución del COVID-19 en República Dominicana (Franco et al., 2020).	Sir, Seir - Extendido Y Log-Lineal.	República Dominicana.	Factor de crecimiento.	94,6 %
Predicción de contagios, recuperaciones y casos fatales de COVID-19 en Argentina a través del	Lg y Gompertz	Argentina	Intervalo de proyección	99 %

uso de modelos de regresión no lineal como base para la planificación de recursos hospitalarios (Cocconi & Roark, 2020).				
LSTM performance analysis for predictive models based on Covid-19 dataset (Cruz-Mendoza et al., 2020).	Lstm y Rnn	Perú	Regarding iterations	80 %
Forecasting SARS-CoV-2 in the peruvian regions: a deep learning approach using temporal convolutional neural networks (Aguilar I. et al., 2021)	Tcn, Mae, Mad, Msle y Rmsle	Perú	Prediction intervals	97,33 %

3. Resultados

La metodología planteada para el análisis y pronóstico de casos contagiados de COVID-19 en el departamento de Madre de Dios se muestra en la Figura 1, donde se utilizamos datos abiertos de la página de MINSA. En la etapa de análisis procedimos a calcular la cantidad de fallecidos y casos positivos en la primera, segunda y tercera ola, para establecer insights respecto a la edad, condición y sexo referentes al departamento de Madre de Dios. Utilizando los datos del MINSA hasta la fecha de 10/07/2022, en el Perú se reporta un total de casos positivos de 3, 675,152 y un total de 213,685 fallecidos. En el departamento de Madre de Dios un total de 18, 058 de casos de COVID-19, donde se tiene 861 casos de fallecidos, y se ha realizado pruebas positivas de (3, 880); (9, 950) y (4, 228) de PCR, PRUEBA RÁPIDA Y ANTÍGENO respectivamente.

Después en la etapa de pronóstico establecimos las variables que tuvieron mayor correlación para establecer un modelo utilizando la técnica de LSTM como se muestra en la Figura 2, y las variables utilizadas que se utilizaron fueron Fecha, Departamento, Provincia, Distrito, Casos, IP. ID y W5, como se describen en la Tabla 2. Obtuvimos una precisión del 94,67% con los datos de entrenamiento y un 92,31% con los datos de pruebas, considerando una ventana de 5 días.

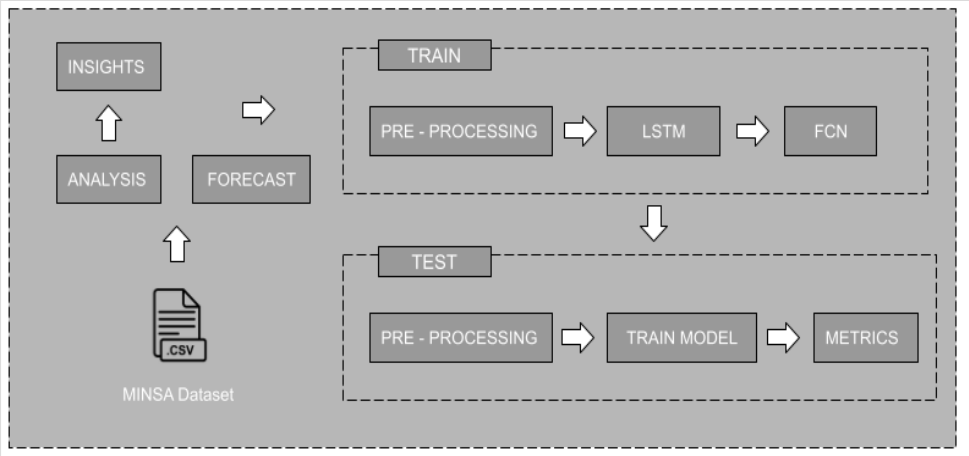

Figura 1. Metodología de la propuesta

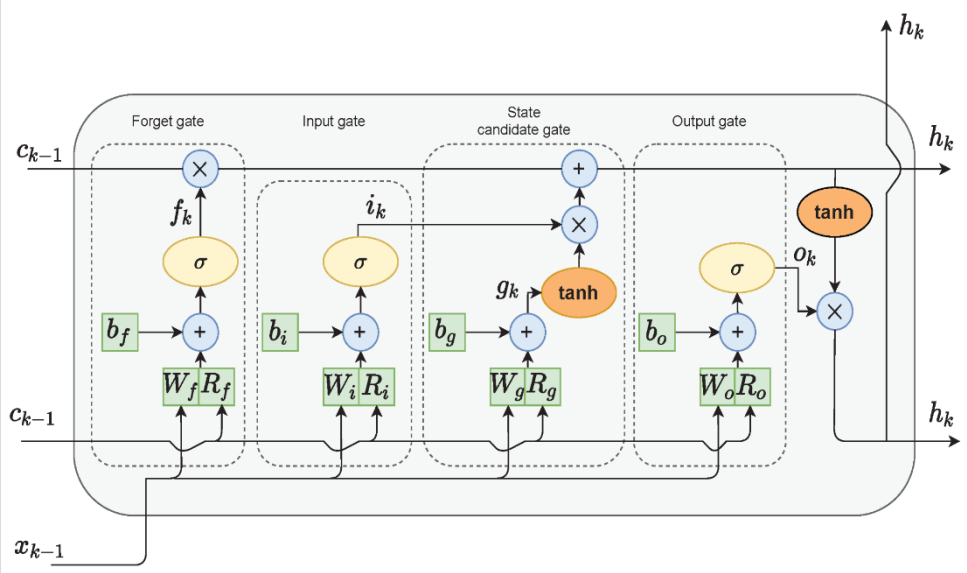

Figura 2. Arquitectura de LSTM

Tabla 2. Descripción de las variables para el modelado y pronóstico

Variable	Descripción
Fecha	Información del día, mes año
Departamento	Nombre del departamento
Provincia	Nombre de la provincia
Distrito	Nombre del distrito
Casos	Total de casos por día
IP	Número de provincias infectadas por departamento
ID	Número de distritos infectadas por departamento
W5	Tamaño de ventana de 5 días de casos previos

4. Discusión

En el estado de arte se establecen diferentes técnicas utilizando datos de diferentes países como se muestra en la Tabla 1, para esta sección de la discusión tomamos datos de solo Perú, donde en el artículo de Cruz-Mendoza et al. (2020) propusieron que RNN + LSTM obtiene una precisión de 80% y en el artículo de Aguilar I. et al. (2021) propusieron una red neuronal de convolución temporal (TCN) utilizando diferentes métricas como MAE, MAD, MSLE y RMSLE, donde se obtuvo 96,11% and 97,33% en la etapa de entrenamiento y prueba, a diferencia del trabajo presentado que se obtuvo 94,67% y 92,31% en la etapa de entrenamiento y prueba, donde tiene un comportamiento intermedio.

5. Conclusiones

La revisión de estudios de investigación recientemente publicados entre los años 2020 y 2021 sobre la aplicación de modelos y técnicas en el pronóstico de los casos de COVID-19, permiten conocer su precisión y efectividad. El análisis de los datos publicados por el Ministerio de Salud en la plataforma de gobierno abierto de Perú, permitió el insight sobre la pandemia de COVID-19 y la propuesta de la técnica LSTM para pronosticar los casos confirmados y fallecidos por la COVID-19 en Madre de Dios. El uso de métricas de clasificación para verificar la eficiencia de la técnica propuesta con la tasa de error. Para los casos confirmados, la precisión fue de 92,31%. El pronóstico puede ayudar a las autoridades de turno a realizar una mejor planificación estratégica ante un aumento inesperado de casos.

Financiamiento

La presente investigación estuvo financiada por el Vicerrectorado de Investigación de la Universidad Nacional Amazónica de Madre de Dios con Resolución del Vicerrectorado de Investigación N° 145-2020-UNAMAD-VRI.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Contribución de autoría

P-B, D. R. y Z-C, D.G.: Recopilación y curación de datos.

U-G, N. J.: Conceptualización del tema desarrollado.

N-C, W.: Diseño de la metodología y aplicación técnicas computacionales, redacción del borrador inicial del artículo científico.

N-V, J.C.: Administración y supervisión del proyecto, revisión y edición del artículo.

Referencias bibliográficas

Aguilar I., L., Ibáñez-Reluz, M., Z. Aguilar, J. C., Zavaleta-Aguilar, E. W., & Aguilar, L. A. (2021). Forecasting SARS-CoV-2 in the peruvian regions: a deep learning approach using temporal convolutional neural networks. *Selecciones Matemáticas*, *8*(1), 12–26. https://doi.org/10.17268/sel.mat.2021.01.02

Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons and Fractals Fractals*, *139*, 110017. https://doi.org/10.1016/j.chaos.2020.110017

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & R Niakan Kalhori, S. (2020). Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health and Surveillance*, *6*(2), e18828. https://doi.org/10.2196/18828

Cocconi, M., & Roark, G. (2020). *Predicción de contagios, recuperaciones y casos fatales de COVID-19 en Argentina a través del uso de modelos de regresión no lineal como base para la planificación de recursos hospitalarios*. XIII COINI 2020 UTN FRBA - Congreso Argentino Internacional de Ingeniería Industrial.

Cruz-Mendoza, I., Quevedo-Pulido, J., & Adanaque-Infante, L. (2020). LSTM perfomance analysis for predictive models based on Covid-19 dataset. *2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 1–4. https://doi.org/10.1109/INTERCON50315.2020.9220248

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons and Fractals*, *134*, 109761. https://doi.org/10.1016/j.chaos.2020.109761

Franco, E. F., Calderón, V. V., & Ramos, R. T. (2020). Modelos de predicción del impacto y evolución del COVID-19 en República Dominicana. *Ciencia, Ambiente y Clima*, *3*(1), 5–21. https://doi.org/10.22206/cac.2020.v3i1.pp5-21

Garrido, J. M., Martínez-Rodríguez, D., Rodríguez-Serrano, F., Pérez-Villares, J. M., Ferreiro-Marzal, A., Jiménez-Quintana, M. M., & Villanueva, R. J. (2022). Mathematical model

optimized for prediction and health care planning for COVID-19. *Medicina Intensiva (English Edition)*, *46*(5), 248–258. https://doi.org/10.1016/j.medine.2022.02.020

Kafieh, R., Arian, R., Saeedizadeh, N., Amini, Z., Serej, N. D., Minaee, S., Yadav, S. K., Vaezi, A., Rezaei, N., & Haghjooy Javanmard, S. (2021). COVID-19 in Iran: Forecasting Pandemic Using Deep Learning. *Computational and Mathematical Methods in Medicine*, *2021*, 1–16. https://doi.org/10.1155/2021/6927985

Khakharia, A., Shah, V., Jain, S., Shah, J., Tiwari, A., Daphal, P., Warang, M., & Mehendale, N. (2021). Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning. *Annals of Data Science*, *8*(1), 1–19. https://doi.org/10.1007/s40745-020-00314-9

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, *139*, 110059. https://doi.org/10.1016/j.chaos.2020.110059

Mendieta, J. F. M., Cortes Cortes, M. E., Cortes Iglesias, M., Perez Fernandez, A. del C., & Manzano Cabrera, M. (2020). Study on predictive models for COVID-19 in Cuba. *Medisur-Revista De Ciencias Medicas De Cienfuegos*, *18*(3), 431–442. https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt/grc-741565

Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanien, A. E. (2020). Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. *SN Computer Science*, *1*(4), 197. https://doi.org/10.1007/s42979-020-00209-9

Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, *34*(7), 959–972. https://doi.org/10.1007/s00477-020-01827-8

Wang, P., Zheng, X., Ai, G., Liu, D., & Zhu, B. (2020). Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. *Chaos, Solitons and Fractals*, *140*, 110214. https://doi.org/10.1016/j.chaos.2020.110214