Revista de Ingeniería y Tecnologías para el Desarrollo Sustentable (REINGTEC) 8 (2020) 51 - 53

ISSN: 2448 – 7198

ANÁLISIS ESTADÍSTICO DE LA PROPAGACIÓN DE LA PANDEMIA COVID-19 EN MÉXICO APLICANDO MODELOS DE MACHINE LEARNING

STATISTICAL ANALYSIS OF THE SPREAD OF THE COVID-19 PANDEMIC IN MEXICO APPLYING MACHINE LEARNING MODELS

Christian Elías Cruz González.

Instituto Tecnológico Superior de Occidente del Estado de Hidalgo (División de Ingeniería en Tecnologías de la Información y Comunicaciones), Paseo del Agrarismo 2000, Carr. Mixquiahuala – Tula, km 2.5, Mixquiahuala de Juárez, C.P. 42700, Hidalgo, México. cecruz@itsoeh.edu.mx

RESUMEN. *Este trabajo propone un modelo estadístico de Densidad de Kernel para predecir la tendencia del número de contagios de Covid-19 en México. El modelo propuesto hace uso del conjunto de datos (en inglés dataset) proporcionado por el CSSE de la Universidad Johns Hopkins. El modelo de Densidad de Kernel expuesto en este trabajo de investigación, hace uso de la información obtenida durante los primeros 50 días del brote (desde el 27 de febrero hasta el 16 de abril), para estimar los nuevos casos de contagio durante los siguientes 150 días (hasta el 13 de septiembre). Esta información es extrapolada al campo del machine learning, mediante el lenguaje de programación Python, para generar un modelo capaz de predecir el desarrollo de la pandemia en México. El objetivo principal de la investigación es «realizar un análisis estadístico del proceso de propagación de la pandemia Covid-19, utilizando modelos matemáticos, así como modelos computacionales de machine learning, para predecir un patrón de comportamiento en el número de contagios diarios que permita a las autoridades competentes tomar decisiones informadas en favor de las familias mexicanas». La metodología utilizada en esta investigación es la siguiente: 1) analizar matemáticamente el dataset, el cual, posee los datos mundiales segmentados cronológicamente, 2) hallar la ecuación que describe el crecimiento exponencial de contagios diarios en el país, 3) evaluar el modelo propuesto calculando el error cuadrático medio entre las predicciones que arroja el modelo y el número real de casos registrados diariamente.*

Palabras clave: *Covid-19, estadística, machine learning*

ABSTRACT. *This work presents a statistical Kernel Ridge model to predict the trend of the number of Covid-19 infections in Mexico. The proposed model uses a dataset provided by the CSSE of Johns Hopkins University. The Kernel Ridge model described in this research makes use of the information recorded during the first 50 days of the outbreak (from February 27th to April 16th), to estimate new cases of contagion during the next 150 days (until September 13th). This information is extrapolated to the field of machine learning, using the Python programming language, to generate a model capable of predicting the development of the pandemic in Mexico. The main goal of the research is «to carry out a statistical analysis of the propagation process of the Covid-19 pandemic, using mathematical models, as well as computer models of machine learning, to predict a pattern of behavior in the number of daily infections that allows the competent authorities take data-driven decisions in favor of Mexican families». The methodology used in this research is the following: 1) mathematically analyze the dataset, which has the world data segmented chronologically, 2) derive the equation that describes the exponential growth of daily infections in the country, 3) transfer the data to the Kernel Ridge model, finally, 4) evaluate the proposed model by computing the mean square error between the predictions made by the model and the real number of daily recorded cases.*

Key words*: Covid-19, statistic, machine-learning.*

INTRODUCCIÓN

El 27 de febrero del 2020[1] se reportó en México el primer caso de SARS-CoV-2, mejor conocido como COVID-19. Ante esta situación, científicos, médicos, virólogos, y expertos de diversas áreas, empezaron una labor extensa para garantizar el cuidado de la salud de los mexicanos.

La rama tecnológica fue impactada de forma positiva, al ser el área responsable de garantizar el trabajo a distancia, pero también al permitir su uso para facilitar la toma de decisiones de las diversas instancias gubernamentales y de salud[2].

En este sentido, matemáticos, estadísticos, y científicos de datos, se dieron a la tarea de generar modelos predictivos[3 4 5] que ayuden a comprender el comportamiento de la pandemia COVID-19 en México, asegurando así, un panorama más amplio que garantice el cuidado de la salud en el país.

ISSN: 2448 – 7198

El presente trabajo propone el uso del modelo de Kernel Ridge (Densidad de Kernel, en español), como un modelo para la predicción del total de casos confirmados de SARS-CoV-2 en México, tomando como dataset de entrenamiento, todos los datos obtenidos de los datasets del CSSE de la Universidad Johns Hopkins durante los primeros 50 días de la pandemia en el país; con la finalidad de identificar la tendencia en los siguientes 150 días.

Si bien, las pandemias suelen comportarse de manera exponencial[6], la hipótesis aquí planteada es que, al menos en México, el comportamiento de esta pandemia será de menor impacto, debido a que el primer caso presentado en México se encuentra separado por 102 días respecto al primer caso a nivel mundial.

El objetivo principal de la investigación es «realizar un análisis estadístico del proceso de propagación de la pandemia Covid-19, utilizando modelos matemáticos, así como modelos computacionales de machine learning, para predecir un patrón de comportamiento en el número de contagios diarios que permita a las autoridades competentes tomar decisiones informadas en favor de las familias mexicanas».

METODOLOGÍA

La metodología utilizada en esta investigación es la siguiente:

1) Analizar matemáticamente el dataset, el cual posee los datos mundiales segmentados cronológicamente.
2) Hallar la ecuación que describe el crecimiento exponencial de contagios diarios en el país.
3) Evaluar el modelo propuesto calculando el error cuadrático medio entre las predicciones que arroja el modelo y el número real de casos registrados diariamente.

RESULTADOS Y DISCUSIÓN

Para poder garantizar la eficacia del modelo a discutir, en un primer momento (al día 50 desde el primer caso confirmado) se realizó un análisis de la información brindada en el dataset, para descartar variables que, de acuerdo con el estado del arte actual[7 8 9], no influyen de manera significativa (como por ejemplo, algunas características geográficas y demográficas) en la propagación del virus entre los habitantes. Esto permitió reducir el dataset de 35 columnas a 18.

A continuación, se realizó el proceso matemático para generar la Ec. 1, que es el resultado de procesar todos los datos del dataset de entrenamiento en un proceso de regresión exponencial.

$$f(t) = 1.3678e^{0.1993t}$$

Ec 1. Ecuación exponencial.

Seguidamente, mediante le uso del lenguaje de programación Python, y utilizando la herramienta sklearn, se creó el modelo de regresión de la manera que se describe a continuación:

Se cargó el dataset, se redimensionó para obtener la forma necesaria de los datos en el entrenamiento. Después de eso, los datos se preprocesaron en un modelo polinómico de grado 10 (con el propósito de ajustar los datos de manera no lineal, evitando así el underfitting del modelo). Entonces, se entrenó al modelo con los datos preprocesados.

Finalmente, se le provee el dato del tiempo de los 150 días siguientes, con la finalidad de obtener la tendencia deseada.

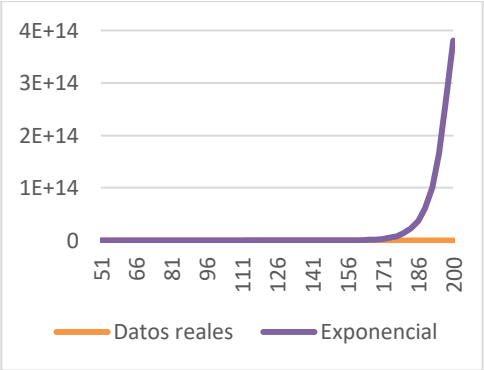

Figura 1. Comparación entre los datos reales y la función exponencial

ISSN: 2448 – 7198

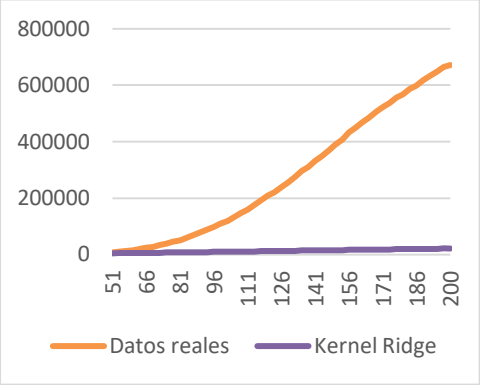

Figura 2. Comparación entre los datos reales y el modelo de Kernel Ridge

Tal como puede apreciarse en las figuras (1)-(2), los datos generados por el modelo distan completamente de los datos reales causados por la pandemia. Igualmente puede observarse que un modelo de regresión exponencial no se adapta correctamente a los datos; en consecuencia, se considera que la hipótesis planteada en el día 50 de contagios, fue errónea.

CONCLUSIONES

Se concluye que la pandemia COVID-19 en México se presentó de una forma totalmente inesperada. De los análisis realizados se descubre que la pandemia COVID-19 en México tiene un comportamiento posiblemente sigmoidal, de acuerdo con las observaciones de los datos reales y su distribución en la figura 2.

Se observa que los datos generados por el modelo de Kernel Ridge, describen la distribución de los datos reales de manera tolerable, durante los 50 días siguientes, sin embargo, no comparten dicho comportamiento durante todo el experimento; por consiguiente, el modelo no explica la propagación del virus.

Se recomienda realizar más investigaciones al respecto, para evitar su propagación ante el panorama de un posible rebrote en la población.

AGRADECIMIENTOS Y/O RECONOCIMIENTOS

Se agradece particularmente el apoyo recibido por el Dr. Francisco Javier Cuadros Romero, por las aportaciones realizadas a la investigación.

REFERENCIAS

1. BBC News. (2020). Coronavirus en México: confirman los primeros casos de covid-19 en el país. BBC. Disponible en: https://www.bbc.com/mundo/noticias-america-latina-51677751. Accesado: 20 abril 2020.
2. El Financiero (2020). Crean Inteligencia Artificial que detecta COVID-19. El Financiero. Disponible en: https://www.elfinanciero.com.mx/ciencia/crean-inteligencia-artificial-que-detecta-covid-19. Accesado: 3 octubre 2020.
3. Pérez, K. (2020). Modelo de profesor Tec da 3 escenarios para picos de COVID en México. Tecnológico de Monterrey. Disponible en: https://tec.mx/es/noticias/guadalajara/investigacion/modelo-de-profesor-tec-da-3-escenarios-para-picos-de-covid-en. Accesado 8 mayo 2020.
4. Alarid Escudero, F. (2020). CIDE y Stanford desarrollan modelo matemático de proyecciones sobre COVID-19. Centro de Investigación y Docencia Económicas. Disponible en: https://www.cide.edu/saladeprensa/cide-y-stanford-desarrollan-modelo-matematico-de-proyecciones-sobre-covid-19/. Accesado: 20 abril 2020.
5. Davidson H. (2020). This article is more than 7 months old First Covid-19 case happened in November, China government records show - report. The Guardian. https://www.theguardian.com/world/2020/mar/13/first-covid-19-case-happened-in-november-china-government-records-show-report. Accesado: 20 abril 2020.
6. Guirao Piñera a. (2020). Entender una epidemia El coronavirus en España, situación y escenarios. Universidad de Murcia. https://digitum.um.es/digitum/bitstream/10201/88621/1/Entender%20una%20epidemia_Guirao2020.pdf
7. Guarneros Olmo, F. (2020). El matemático de la UNAM que pronosticó el aceleramiento del coronavirus en México habló de cuándo llegará el pico máximo y la probable vuelta a la normalidad. Infobae. Disponible en: https://www.infobae.com/america/mexico/2020/04/10/el-matematico-de-la-unam-que-pronostico-el-aceleramiento-del-coronavirus-en-mexico-hablo-de-cuando-llegara-el-pico-maximo-y-la-probable-vuelta-a-la-normalidad/. Accesado: 20 abril 2020.
8. Chow Tong Y. (2020). Mathematical Analysis, Model and Prediction of COVID-19 Data. MedRxiv. Disponible en: https://www.medrxiv.org/content/10.1101/2020.08.04.20168195v1. Accesado: 1 septiembre 2020.
9. Okuonghae D. (2020). Analysis of a mathematical model for COVID-19 population dynamics in Lagos, Nigeria. ScienceDirect. Disponible en: https://www.sciencedirect.com/science/article/pii/S0960077920304306. Accesado: 1 septiembre 2020.