Predicción del fin de la vacunación del COVID-19 en México y latinoamérica

14 de junio del 2021

Diego Fernando Montaño Pérez
Ingeniería en Tecnologías Computacionales
Instituto Tecnológico y de Estudios Superiores de Monterrey
Monterrey Nuevo León México
A01282875@itesm.mx

Jose Alberto Gonzalez Arteaga
Ingeniería en Tecnologías Computacionales
Instituto Tecnológico y de Estudios Superiores de Monterrey
Monterrey Nuevo León México
A01038061@itesm.mx

Jamir Leal Cota
Ingeniería en Tecnologías de Información
Instituto Tecnológico y de Estudios Superiores de Monterrey
Monterrey Nuevo León México
A00826275@itesm.mx

ABSTRACT

El evento de la pandemia SARS-CoV2 2019 sigue desencadenando casos dignos de investigación por el gran impacto que tuvo a todos los habitantes del planeta. El manejo de la vacunación así como su rendimiento en la comunidad de latinoamérica es un caso de estudio por la diversidad de gobiernos y población. La investigación se sustenta en múltiples modelos de inteligencia artificial, principalmente interpolación lineal y ARIMA para la fecha estimada de inmunidad poblacional en una región (70% de la población de la región). Los resultados obtenidos muestran países como Brasil, Chile y Argentina que tienen un muy buen rendimiento con relación a su población. Así mismo, muestra países con una necesidad de agilizar la vacunación tanto como Colombia y Bolivia. Las estimaciones reflejan el rendimiento de cada uno de los países y nos dan una idea de sus procesos de vacunación.

Palabras clave

Predicción, Vacunas, COVID-19, México, Latinoamerica, Aprendizaje supervisado, ARIMA, Forecasting.

1 Introducción

1.1 Situación global y local de las vacunas contra el COVID-19

En 2019 en la ciudad de Wuhan, China, empezó el brote de una nueva enfermedad que provoca neumonía por causas desconocidas, tiempo después al virus que provocaba esto se le denominó por el nombre de SARS-CoV2 la cual se empezó a propagar por todo el mundo rápidamente. Fue hasta el último trimestre del 2020 cuando se empezó a autorizar el uso de las vacunas en desarrollo, actualmente a inicios del 2021 hay 10 vacunas autorizadas para el uso público entre las cuales están las desarrolladas por Pfizer, Moderna, AstraZeneca, Centro Gamaleya, etc.. El proceso de vacunación es lento y por lo tanto no hay exactitud de cuando el mundo entero podrá estar vacunado contra esta enfermedad que mantuvo a gran parte del mundo encerrado durante el 2020. En México para el momento que se escribe el artículo se tiene registrado que tan sólo un 11.3% de la población está completamente vacunada y al menos 35.2 millones de mexicanos tienen una sola dosis (Our World in Data,2021).

1.2 Definición del problema

Analizando la manera en la que la pandemia ha impactado a México tenemos como consecuencias principales una crisis económica y de salud que esta ha dejado, especialmente el segundo y tercer trimestre del 2020. La evolución de la misma tiene una relación directa con la vacunación, la cual ha estado incrementando de manera paulatina a lo largo del 2021.

México al igual que muchos otros países han estado en un constante cambio en su situación de control, ya que se han presentado diferentes olas de contagios intermitentes. El análisis para la predicción de los resultados de un procedimiento como en este caso es el incremento de la población vacunada puede ser de gran utilidad para grupos de interés, como el gobierno, las empresas, grupos vulnerables, personal del sector salud y población general. De esta manera podrían tomar mayores precauciones con

respecto a los daños económicos que podría dejar el aislamiento debido a los incrementos en los contagios.

De acuerdo con la organización mundial de turismo (UNWTO por sus siglas en inglés), México ocupó el octavo lugar a nivel mundial como país más visitado por turistas extranjeros, dejando una derrama económica de $19,571 millones de dólares y proyectando un crecimiento del 11%. Enlazando los datos con la situación actual el impacto sobre la disminución del turismo en el país se considera directa sobre la economía a nivel nacional, creando repercusiones que podrían tomar años en recuperarse.

1.3 Objetivo

El objetivo del proyecto es conseguir la estimación de una fecha para la cual se tenga vacunada al 70% de la población mexicana. De esta manera se considera la salida de la pandemia, un posible fin del aislamiento y disminución de la necesidad de los protocolos correspondientes relacionados con la prevención de los contagios del COVID-19 (uso de mascarillas, respeto de aforos, entre otras).

1.4 Motivación

Se consideró de gran utilidad aplicar técnicas y modelos de aprendizaje automático, para realizar las estimaciones. Las cuales se realizaron a partir de un dataset obtenido de la plataforma Kaggle, el cual se centra sus variables principalmente en el tema del seguimiento de la vacunación tanto por país como por fabricantes de la misma vacuna. Dicho dataset está respaldado por la organización *Our World in Data*.

Haciendo uso de la información y teniendo la hipótesis generadas se considera que se puede hacer una limpieza de datos para hacer análisis de series temporales, generando predicciones sobre el tiempo hasta conseguir la fecha en la cual se llegará a tener al porcentaje deseado de población que cuente con la vacuna. Dicho análisis se podría aplicar no solo a México, sino también a cada país que cuenta con su respectiva información sobre la vacunación.

1.5 Resultados esperados

Hacer solo uso del lenguaje de programación Python para la creación de los modelos y usar las librerías de *Pandas* y *Scikit-learn*.
La técnicas y modelos a usarse sean de *Time Series, Forecasting(Recursive multi-step forecasting & forecasting autorregresivo) y el algoritmo de Expectation Maximum.*

Se espera que el modelo de predicción a crear tenga una precisión igual o mayor a un 90%, la cual nos ayude a determinar una fecha esperada para el fin de la vacunación, se espera que esta sea antes de diciembre del 2024.

2 Conceptos Previos

Predecir las series de tiempo consiste en tratar de crear un modelo que nos permita encontrar su valor en un tiempo $t + l$ donde t es el tiempo actual y l es la duración hasta donde se desea obtener el valor. Este tipo de técnicas son de gran utilidad como base para la planificación, ya que permiten obtener con una alta precisión el comportamiento de la serie que se está analizando en un tiempo futuro $> t$.

En la sucesión de los datos, las observaciones que se han obtenido son discretas a intervalos equidistantes de tiempo. El objetivo de predecir series temporales es crear una función o modelo $\hat{z}_t(l)$ que permita minimizar la media cuadrada de las desviaciones entre los valores reales (actuales) y predecidos por la función, expresado de la siguiente manera: $z_{t+l} - \hat{z}_t(l)$. (Box y Jenkins, 1970).

Los modelos autorregresivos integrados de media móvil (ARIMA por sus siglas en inglés) resultan de gran utilidad para realizar predicciones relacionadas con el tiempo, al igual que el suavizamiento exponencial. Sin embargo, los modelos ARIMA se aproximan mediante la búsqueda de una auto correlación entre los valores de los datos mientras que los basados en suavizamiento exponencial utilizan mayormente las tendencias y la estacionalidad como eje central para las predicciones.

Para hacer uso de ARIMA es importante tener en cuenta el concepto de estacionario, ya que este indica cuando este tipo de modelos son de utilidad. Una serie temporal se considera estacionaria cuando sus propiedades son independientes del tiempo en el que se está analizando. Partiendo de esta regla, los modelos no estacionarios son los que cuentan con una tendencia o son estacionales.

Se considera aplicar este tipo de modelo específicamente cuando las series de tiempo son no estacionarias, ya que el término integrador se refiere al número de diferencias (una diferencia corresponde a la longitud del intervalo temporal) necesarias para convertir la serie en estacionaria. Cuando no es necesario hacer uso de este porque nuestra serie temporal ya es estacionaria se emplea el modelo ARMA.

ARIMA es un modelo que basa su complejidad en otros tres más simples, ya que hace uso de la autoregresión de sus propios valores para realizar predicciones. Mientras que también emplea el la media móvil, uno de los modelos más simples en el análisis de series temporales. Como se mencionó anteriormente la integración convierte la serie temporal no estacionaria en una estacionaria, ya que para modelar idealmente las series temporales deben cumplir con esta propiedad.

Existe una variación de este modelo donde se agrega el componente estacional, separando dos principales parámetros que se deberán seleccionar, los cuales se presentan de la siguiente manera:

$$SARIMA(p,\ d,\ q)(P,\ D,\ Q,\ s)$$

Describiendo esto tenemos que p corresponde a P, al igual que qy Q. El parámetro d se encargará de convertir el modelo en estacionario, mientras que D tiene como intención eliminar la estacionalidad de la serie, s corresponde a la duración de una temporada (también conocido como periodo).

3 Metodología

Para el desarrollo del proyecto se utilizó la herramienta de Google Colab para ejecutar el código y las librerías de Python, se decidió programar en Python ya que es el lenguaje que mejor sabía usar el equipo.
Las librerías utilizadas fueron la siguientes:

1. pandas
2. Numpy
3. Seaborn
4. Matplotlib
5. Skforecast
6. Sklearn
7. Statsmodels
8. Pmdarima

Para trabajar con las bases de datos se tuvo que descargar desde la página de Kaggle, una vez obtenidas se inició el proceso de exploración analitica de la informacion.
En la primera fase se exploraron todos los datos que se tenían en ambas bases de datos, la base de datos *country_vaccinations.cvs* contiene columnas repartidas por fechas del país, número total de habitantes vacunados, vacunaciones diarias, etc.
La de *country_vaccinations_by_manufacturer.cvs* contiene datos sobre los fabricantes de las vacunas, por lo que fue desechada pues no contenía datos relevantes para la investigación.

La primera base de datos se filtro para solo tener datos de México, se buscó que la información dada por el proveedor de la fuente de datos fuera correcta en cuanto a las fechas y se encontró que se tienen datos desde el 24 de diciembre del 2020 hasta el 31 de mayo del 2021, para tener un panorama más profundo de los datos que se dan se decidió realizar un mapa de calor (**Figura I**) para graficar la correlación entre los datos y saber con mayor precisión si estos nos servirán.

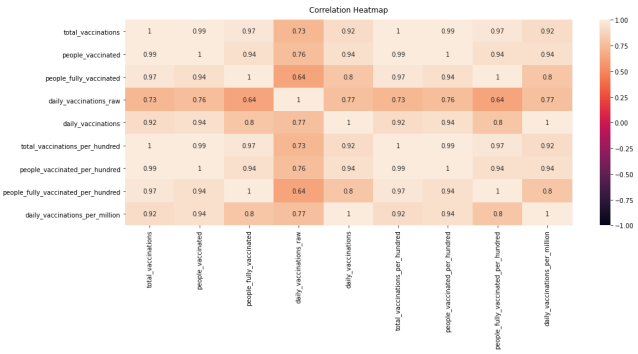

Figura I: **Mapa de calor de correlación de datos de México en *country_vaccinations.cvs***

Teniendo en cuenta que los datos servirán para el proyecto se empezó con el preprocesamiento de datos, por lo que primero se buscaron los datos nulos o faltantes que pudieran haber y se graficaron todas las columnas para poder observar a detalle la distribución (**Figura II**) que tenían estos datos.

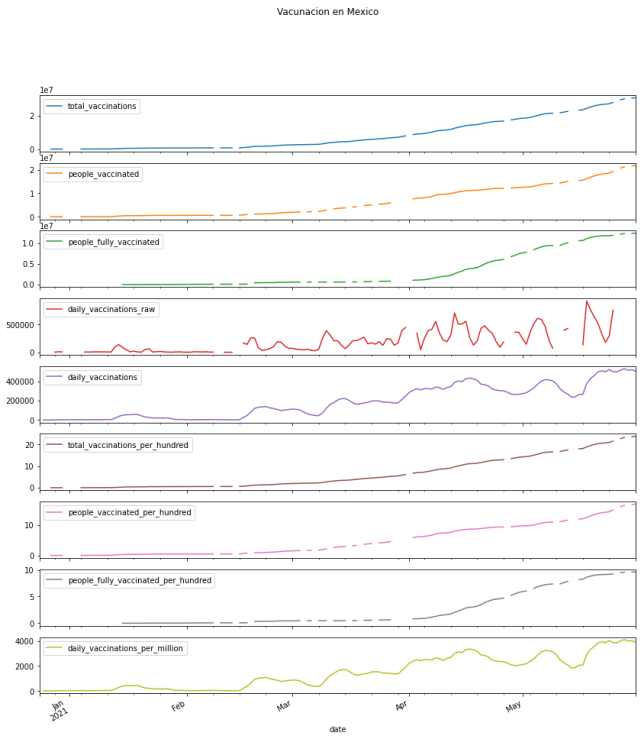

Figura II: **Distribución de datos de México en *country_vaccinations.cvs***

Empezando el preprocesamiento de datos lo primero que se realizó fue la imputación de datos faltantes usando interpolación con diferentes métodos, entre ellos fueron el *rolling average y mean*, *lineal, tiempo, cuadrático, cúbico, slineal, akima, polinomial y spline.* Se aplicaron estas técnicas en todas las columnas pero la que terminó siendo más importante para poder hacer nuestra predicción a futuro es la de gente vacunada diariamente.

4 Resultados

Después de aplicar el preprocesamiento de los datos se rellenaron los valores de tiempo faltantes con los obtenidos del modelo con los mejores resultados. La siguiente tabla muestra la comparación de los diferentes modelos que fueron aplicados (orden por precisión descendente):

Método aplicado	**Precisión (R-squared)**
RollingMean	0.469632
RollingMedian	0.464575
InterpolateLinear	0.436166
InterpolateTime	0.436166
InterpolateSLinear	0.436166
InterpolateAkima	0.414266
FillBackfill	0.396812
FillFfil	0.374594
FillMean	0.287655
FillMedian	0.263478
InterpolateQuadratic	0.263478
InterpolateCubic	0.192907
InterpolateSpline5	-1.542447
InterpolatePoly5	-1.542407
InterpolatePoly7	-0.685922
InterpolateSpline4	-0.475822

RollingMean fue el método que tuvo mayor precisión, el cual es calcular la media de un periodo determinado, se seleccionó una duración para el periodo de 24 días. La misma técnica pero haciendo uso de la mediana obtuvo casi el mismo desempeño, al igual que las que emplean interpolación lineal.

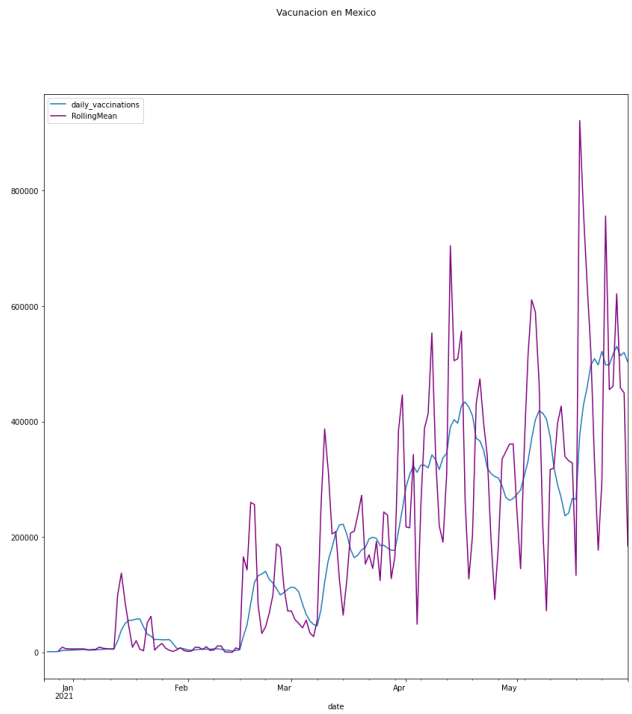

Figura III: **Media usando como intervalo de tiempo 24 días.**

Como se puede observar en la figura X, una media con una ventana de tiempo reducida funciona como una tendencia que se comporta de manera similar a la variable original.

Sin embargo, se eligió como método principal la interpolación lineal para llenar los valores faltantes puesto que hacer uso de variables estadísticas crea un efecto de "suavizado" en las predicciones, como se aprecia en la figura X, en otras palabras los incrementos o decrementos drásticos son minimizados. Se muestran las gráficas después de aplicar la interpolación lineal.

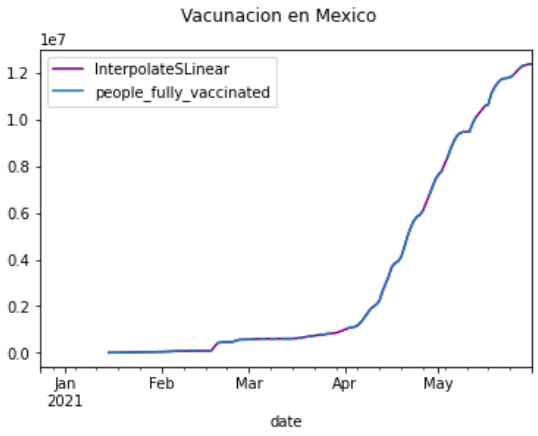

Figura IVI: **Población completamente vacunada con los datos interpolados.**

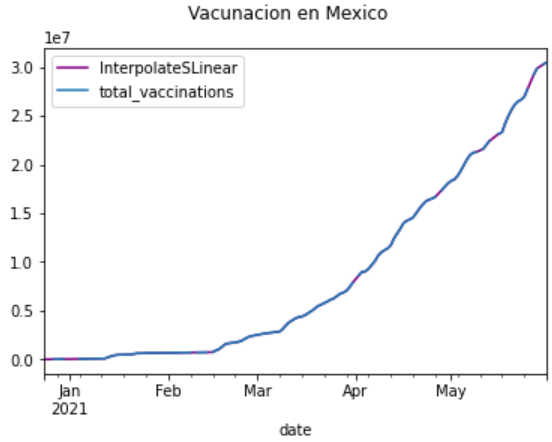

Figura VI: **Vacunas totales con los datos interpolados.**

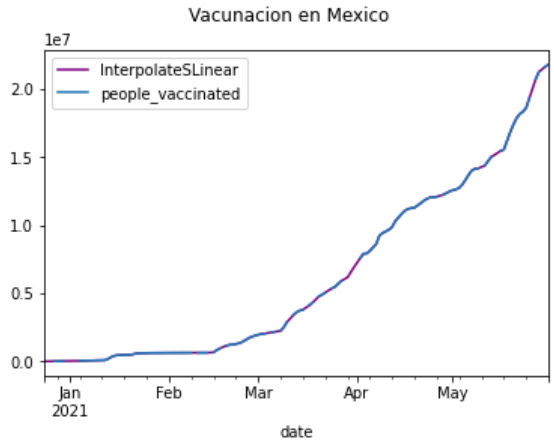

Figura VII: **Población vacunada con los datos interpolados.**

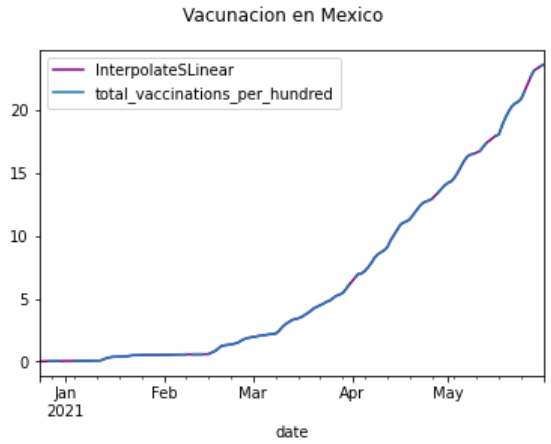

Figura VIII: **Población vacunada por centena con datos interpolados.**

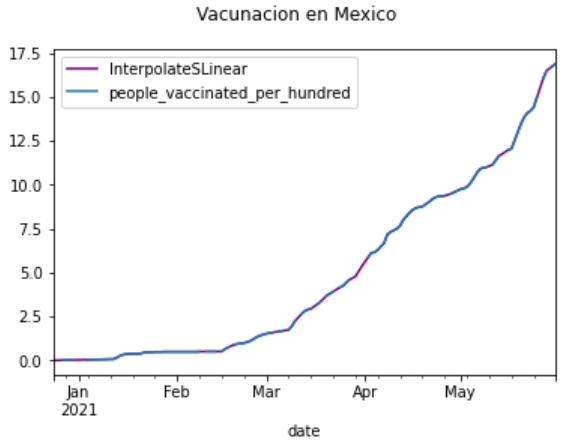

Figura IX: **Personas vacunadas por centena con los datos interpolados.**

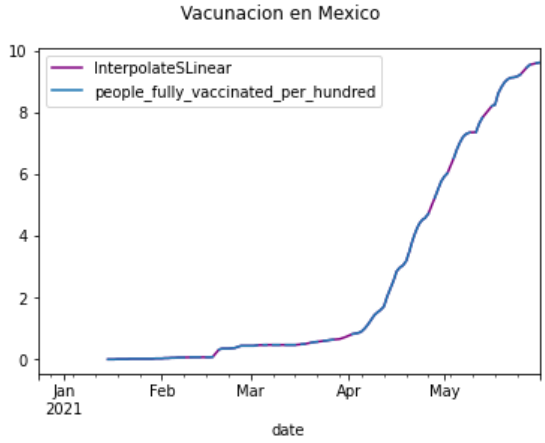

Figura X: **Personas totalmente vacunadas con los datos interpolados.**

Como se puede observar en las gráficas mostradas los valores que en un principio parecían incompletos se han llenado con los valores obtenidos por la interpolación lineal, dejando el dataset completo y listo para ser usado por los modelos de predicción.

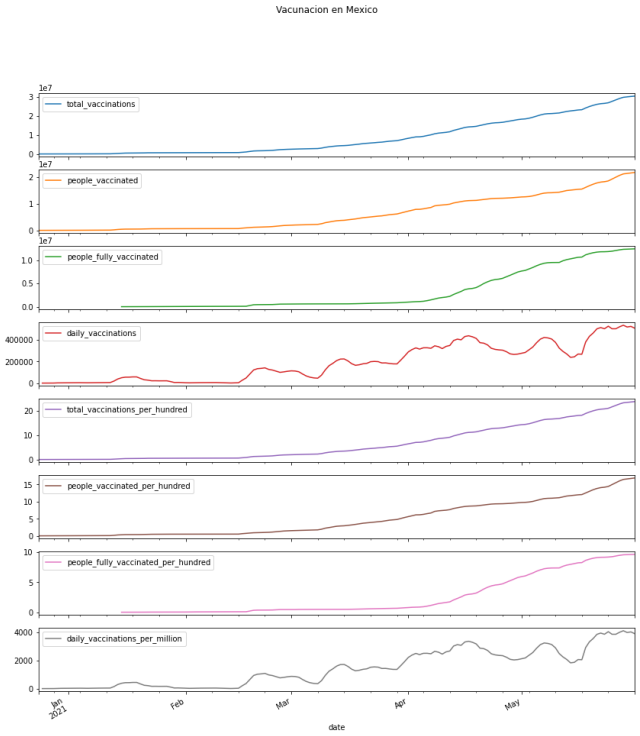

Figura XI: **Se muestran las variables a lo largo del tiempo.**

Una vez que se tenían los datos preparados para realizar el entrenamiento de los modelos. Se realizaron las pruebas con diferentes modelos simples de regresión lineal, el modelo con mejores resultados fue el regresor Lasso. Se presenta la gráfica comparada con los valores reales.

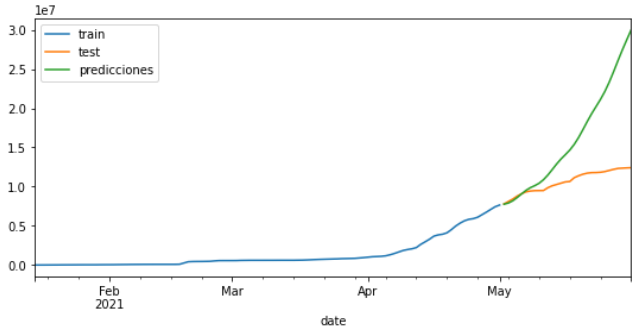

Figura XII: **Se muestra la regresión lineal lasso como predicción (verde) contra los valores reales (naranja).**

Como se puede observar la predicción haciendo uso de un modelo demasiado simple que no está centrado en la evolución temporal, sino más bien tratándolo como cualquier otra serie no temporal, muestra un error muy alto.

Para obtener los mejores hiper parámetros a utilizar en ARIMA, se aplicó una búsqueda aleatoria para encontrarlos. Los resultados obtenidos fueron los siguientes:

$$SARIMAX(p = 4,\ d = 2,\ q = 2)$$

Dichos parámetros son los que fueron ingresados en el modelo estacional autorregresivo integrado de media móvil con factores exógenos.

Medida analizada	Regresión Lasso	SARIMAX
Error medio cuadrado (MSE)	58855086696481.59	2328261250492.38
Error medio absoluto (MAE)	5480745.09	1185238.67

En la tabla superior se muestra como SARIMAX tiene un desempeño sustancialmente más alto en cuanto a la evaluación de la predicción comparada con los valores reales, contra la regresión lasso.

Los valores de predicción haciendo uso del modelo para la serie temporal se muestran en la siguiente gráfica:

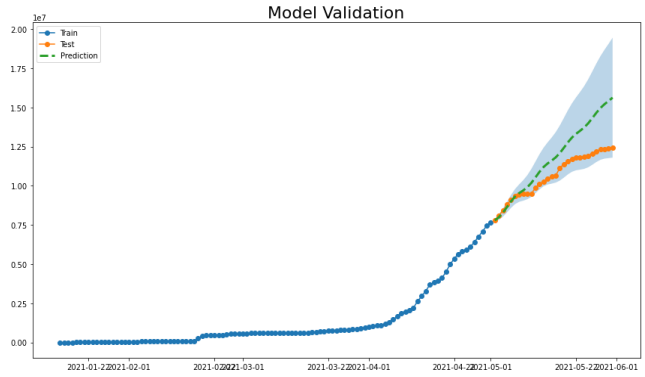

Figura XIII: **Se muestran los valores predecidos por el modelo (verde) contra los reales (naranja).**

Se logró un desempeño más alto del modelo, que incluso hace un emparejamiento con el borde inferior del mismo.

Tratando de predecir la hipótesis planteada, la cual indicaba la población totalmente vacunada con un porcentaje mayor al 70%, la fecha fue el **18 de agosto del 2023**.

Todo el proceso realizado para hacer la predicción de la vacunación en México se realizó con otros países de latinoamérica, obteniendo las siguientes fechas para el mismo porcentaje de la población vacunada (70%) haciendo uso de SARIMAX:

País	Fecha para alcanzar el 70% de la población vacunada	Población completamente vacunada (millones de habitantes)
Argentina	2022-08-21	31.5
Aruba	2021-06-13	0.0745
Bolivia	2026-09-09	8.6
Brazil	2022-05-20	148
Chile	2021-07-13	13.26
Colombia	2026-10-03	35.1
Costa Rica	2022-01-10	3.5
Peru	2023-12-29	22.7
Ecuador	2022-06-03	12.17
México	2023-08-18	90.29

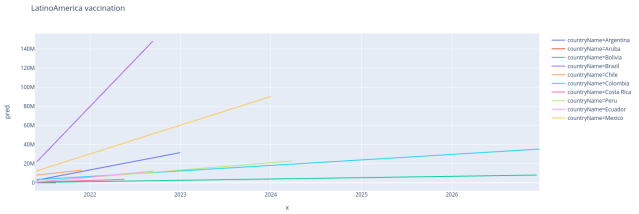

Figura XIV: **Predicción de países latinoamericanos separados por color.**

Se puede catalogar como casos de éxito Argentina, Brasil y Chile, los cuales de acuerdo a su población tienen un rendimiento muy bueno con la aplicación de vacunas. Por otro lado, hay países como Bolivia y Colombia que tienen un rendimiento muy pobre y tienen una necesidad de agilizar su aplicación. Los demás países se encuentran en un rendimiento promedio en relación a sus habitantes.

5 Conclusiones y Reflexiones

De acuerdo con la hipótesis planteada podemos rechazarla, ya que en el caso de México obtuvimos una fecha estimada más cercana a la actualidad, así como uso de diferentes modelos de los planeados inicialmente. El poder adaptativo del proyecto es fundamental para poder realizar un

producto de calidad, en nuestro caso tuvimos que probar con múltiples modelos tanto para el manejo de datos faltantes, así como la implementación del forecasting en time series. Así mismo, es importante evaluar el rendimiento del modelo y poder discriminar para utilizar el modelo que se apegue más a lo necesario en la problemática a solucionar.

Las estimaciones esperadas son muy interesantes y sin duda son datos valiosos para nuestra comunidad en latinoamérica, ya que nos puede dar una idea no solo de una fecha estimada donde más del 70% de la población está vacunada, sino de cómo cada país ha ido ha ido manejando la pandemia de acuerdo con sus herramientas. Si bien es cierto que cada uno de los países tienen diferencias de gobierno, tamaño de población y posición geográfica, es importante realizar este tipo de comparaciones para poder prestar atención a los casos de éxito y así poder identificar patrones y tratar de replicarlos.

Sin duda es un proyecto muy valioso el cual desencadena futuras hipótesis y premisas que pueden generar aún más valor a los datos obtenidos por y para la comunidad de Latinoamérica.

Referencias

[1] Box, G. E. P., & Jenkins, G. M. 1970. Time series analysis: Forecasting and control. San Francisco: Holden-Day. DOI: http://www.ru.ac.bd/stat/wp-content/uploads/sites/25/2019/03/504_05_Box_Time-Series-Analysis-Forecasting-and-Control-2015.pdf

[2] Centers for Disease Control and Prevention. (n.d.). Diferentes vacunas contra el COVID-19. Centers for Disease Control and Prevention. DOI: https://espanol.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html

[3] Hyndman, R.J., & Athanasopoulos, G. 2018 Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. DOI: https://www.OTexts.com/fpp2

[4] Markos, O. 2020. Time series forecasting- SARIMA vs Auto ARIMA models. Medium. DOI: https://medium.com/analytics-vidhya/time-series-forecasting-sarima-vs-auto-arima-models-f95e76d71d8f

[5] Peixeiro, M. 2021. The Complete Guide to Time Series Analysis and Forecasting. Medium. DOI: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

[6] Pulagam, S. 2020. Time Series forecasting using Auto ARIMA in python. Towards Data Science - Medium. DOI: https://towardsdatascience.com/time-series-forecasting-using-auto-arima-in-python-bb83e49210cd

[7] Redacción Nacional CONECTA. 2021. Vacunas contra COVID: características y diferencias. Tecnológico de Monterrey. DOI: https://tec.mx/es/noticias/nacional/salud/vacunas-contra-covid-caracteristicas-y-diferencias

[8] World Tourism Organization. 2017, UNWTO Tourism Highlights, 2017 Edition, UNWTO, Madrid, DOI: https://www.e-unwto.org/doi/pdf/10.18111/9789284419029