Exploratory Data Analysis and Forecasting of Covid-19 Vaccination using ARIMA and LSTM

Journal of Development Economics and Management Research Studies (JDMS)

A Peer-Reviewed Open Access International Journal

ISSN: 2582 5119 (Online)

Crossref Prefix No: 10.53422

09(11), 77-88, January-March, 2022

@Center for Development Economic Studies (CDES)

Reprints and permissions

https://www.cdes.org.in/

https://www.cdes.org.in/about-journal/

Exploratory Data Analysis and Forecasting of Covid-19 Vaccination using ARIMA and LSTM

Dr. Neha Sharma[1], and Ms. Snehal Patel [2]

Abstract

The Covid-19 outburst seemed in Wuhan in December 2019 and spread rapidly all over the world. The Covid-19 ailment does now have clinically proven vaccines and medication for treatment [1]. WHO recommends that initial vaccination should arrange groups at the highest risk of introduction to infection in each country, including health workers, older persons and those with other health issues? In this study, we going to apply ARIMA and LSTM. ARIMA (Autoregressive Integrated Moving Average), is a class of classical that captures a suite of unlike standard temporal structures in period series data. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network proficient in knowledge order dependence in arrangement prediction problems. This is a behaviour required in multifaceted problem domains like machine translation, speech recognition, and more. LSTMs are a composite area of deep learning.

Keywords: ARIMA, Exploratory Analysis, LSTM, Vaccine Forecasting.

1. Introduction

Coronavirus (COVID-19) is an infectious disease caused by a newly learnt virus. Most people contaminated with COVID- 19 will experience modest respiratory illness and recuperate without the need for the superior treatment. Grown-up people, as well as those with inferior risk of heart disease, diabetes, chronic obstructive pulmonary disease, and cancer, are more probable to be seriously ill. The best way to avoid and delay transmission is to better comprehend the COVID-19 virus, the disease it reasons and how it spreads. Defend yourself and others from infection by washing your hands or using alcohol-based medicine often and deprived of touching your face [2]. The COVID-19 virus is spread mostly through saliva droplets or runny nose when an infected person coughs or sneezes, so it is significant that you

[1] Director, Shanti Business School, Ahmedabad, Gujarat, India.

[2] Pursuing MSC in Big Data, FOM University, Germany.

re-practice the rehearsal of breathing (for example, by coughing on a flexible Elbow). Impartial access to harmless and effective vaccines is important in ending the COVID-19 epidemic, so it is very hopeful to see so many vaccines show progress. Vaccination does not mean that a person can lose their breath and put themselves and others in danger, especially as research lasts those vaccines can guard not only from disease but also from infection and broadcast. But it is not the vaccine that will halt the epidemic, it is the vaccination [2]. One must guarantee the obtainability of vaccines equitably and openness, and ensure that each nation has access to them and can transport them to shield its people, starting with the most exposed. With the support of the analytical and visual course through analysis and model, data will change numbers into patterns and patterns into understanding.

This project is to examine and picture the newest global scenario in the contest against the "covid-19 vaccine epidemic" in which we will foresee the timing of each country's vaccination process could be accomplished. In the present context, numerous countries are facing a shortage of strategies. Studies show that more than 950 million times the COVID-19 vaccine has been dispersed worldwide hence, it is essential and beneficial to do additional research and investigate the country's smart analytics and predict other vaccine-level forecasts to regulate when the country will be entirely vaccinated. In this analysis, we working to be using different methods and methods to enhance the exactness rate and to observe the variations and outcomes of the diverse models applied. We will be undertaking time-series forecasting for vaccination progress through the assistance of machine learning algorithms such as ARIMA modelling and LSTM which is a decent model from the neural network, this will help us expect further vaccination progress rate. A basic exploratory analysis will also be part of this work to observe and analyze the overall statistics of vaccines being cast-off and rolled out by different regions of the world and the topmost vaccines used by a precise country to aid fight the contest of covid- 19 pandemic affections [3]. By the termination, we will be able to gain so excessive identifications about the vaccine and vaccinations trendy across the globe. In the next units, the changes in the process and results of various models will be further deliberated.

2. Methodology

2.1Description of Data

The dataset applied emphases on the COVID-19 vaccines and vaccinations. The historical data for covid-19 vaccine and vaccination has been mined from Kaggle (an Online community of data scientists and machine learning practitioners). The data set comprises data of all the diverse vaccines rolled out to countries, total vaccinated, daily Vaccinated, manufacture of the vaccine and more arenas are obtainable in the dataset. This dataset involves data of all the possible countries consuming access to the vaccine or the vaccination procedure.

2.2Exploratory Data Analysis

Exploratory Data Analysis (EDA) in Python is the primary phase in your data analysis course established by “John Tukey” in the 1970s. In statistics, exploratory data analysis is a tactic to analyse facts groups to abridge their chief features, frequently with visual means. By the name itself, we can get to distinguish that it is a stage in which we need to discover the data set. Exploratory Data Analysis is a vital step beforehand you jump to machine learning or demonstrating your data [5]. By undertaking this you can get to know whether the selected features are decent enough to model, are all the landscapes required, are there any connections grounded on which we can either go back to the Data Pre-processing step or transfer on to modelling. Later Exploratory Data Analysis is accomplished and visions are drawn, its structures can be used for supervised and unsupervised machine learning. EDA comprise some basic steps such as data description, handling missing data and outliers and understanding the elementary relation amid data through some visual presentation. About basic libraries used for the data, exploratory analysis were pandas, NumPy, matplotlib, seaborn, plotly and some additional consequently [9].

A. Data collection

Data collected from the prevalent data science and machine learning tool stage known as Kaggle is from where we have composed the data from and applied the data for all-purpose events and further investigate and knowledge mining.

B. Feature engineering and selection

When analysing and forecasting a model, at times one may not need all the features or input variables for the analysis or modelling hence we can study the features and decide which of those features are needed for the model. Using a dataset without doing feature engineering and selection may cause some uncertain results or results that are not very impactful.

C. Data mining and exploratory analysis

After removing variables that are not relevant for model making. cleaning of data is performed and thereafter we can perform some exploratory data analysis to overcome certain relations and relationships of the variables.

D. Select appropriate machine learning technique

For our modelling, we have chosen ARIMA and LSTM model as it is best suited for time series and real-time data. Both models are very impactful as they consist of crucial and relevant data needed for future insights.

E. Representation of result in build model and visual form

We have made good use of certain python libraries to visually represent our results and outputs to easily understand and identify patterns and insights. Those libraries are MATPLOTLIB, SEABORN, PYPLOT and PYWAFFLE to enhance the proportion of vaccines rolled out in different parts of the globe [9]. A simple summary model is generated for the forecasting of the vaccination rate and a simple differencing plot of making the dataset stationary.

2.3ARIMA

The algorithm applied in the project is the ARIMA (Auto-Regressive Integrated Moving Average) model which is a step-by-step model for analysing and predicting time series data. It explicitly provides a list of common structures in time-series data, and as a result, provides a simple yet powerful way to make accurate time forecasts. The ARIMA model is characterized by 3 terms: p, d, q where, p the AR command order is the MA word order and d is the number of variations required to make the time series stop. In our practice, we have used forecasts for a series of unconventional times as they make predictions only based on previous data collected. While working on the model the first step was to stabilize the data so that there would be no difference in prediction, this is done by removing the lag values (previous values), this step can be repeated until we get static data. Performing data suspension is important otherwise the output will always be 0. This way we can calculate the difference between the lag values. Once the data is stationary and there are no variations found then the model can be created as we have already discovered values of p,q and d. ARIMA() is used and imported from the stats model in python. The model summary discloses a lot of information. The table in the centre is the coefficients table [8]. Wherever the standards under 'coef' are the masses of the particular terms. The Arima model consists of some important information like a dependent variable, no of observations, model, date, coefficient and the best model found in a total fit time. The

ARIMA technique analyses and estimates similarly spaced univariate time series data, handover function data, and intervention data by means of the autoregressive integrated moving-average (ARIMA) or autoregressive moving-average (ARMA) model. An ARIMA model foresees a value in a response time series as a lined combination of its past values, past errors (also called shocks or innovations), and current and past values of other time series. The ARIMA procedure delivers an all-inclusive set of gears for univariate time series model identification, parameter estimation, and foretelling, and it offers great elasticity in the kinds of ARIMA or ARIMAX Getting Started: ARIMA Procedure 187 models that can be analysed. The ARIMA procedure provisions seasonal, subset, and factored ARIMA models; interference or interrupted time series models; several regression analyses with ARMA faults; and well-spoken transfer function models of any difficulty [10]. In an ARIMA model, the upcoming value of a variable is a linear grouping of past values and past errors, expressed as follows:

$$W_t = \mu + \frac{\theta(B)}{\phi(B)} a_t$$

Equation (1). ARIMA Equation

Where:

t index times

W_t is the response series Y_t or a difference of the
response series μ is the mean term
B is the backshift operator; that is $BX_t = X_{t-1}$

$\phi(B)$ is the autoregressive operator, represented as a polynomial in the backshift operator: $\phi(B) = 1 - \phi 1 \, B - \ldots - \phi P \, B^P$
$\theta(B)$ is the autoregressive operator, represented as a polynomial in the backshift operator:

$\theta(B) = 1 - \theta \, 1 \, B - \ldots - \theta \, P \, B^P$

a_t is the independent disturbance, also called the random error

2.3 **LSTM:**

LSTM is observed as one of the most substantial options for forecasting happenings, LSTM networks are well-suited to classifying, processing and creating predictions founded on time series data, meanwhile, there can be lags 6 of unknown duration between significant proceedings in a time series. Although LSTMS are a kind of RNN and function likewise to traditional RNN, its gating instrument is what sets it separately. This feature discourses the "short-term memory" problem of RNN. The pattern moves the past veiled state into the specific stage of the structure in this sort of architecture as shown in figure 1.

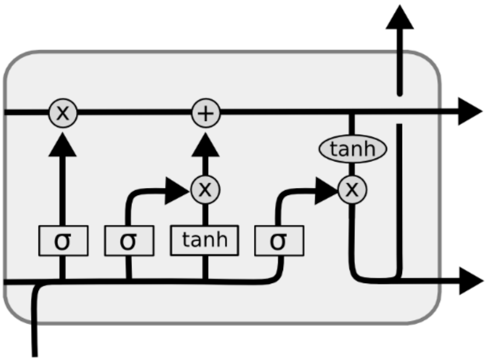


Figure 1 Long Short Memory Network Cell State

An LSTM (Long-term storage of short-term memory) has a similar control flow as a recurrent neural network. It courses data passing on material as it broadcasts onward. The variations are the actions within the LSTM's cells. These operations are used to let the LSTM keep or overlook data. The stages of the approach are drawn as follows:

- ✓ Adapts abstracts from a list of strings into lists of integers (arrangements). Create features and labels from sequences
- ✓ Build LSTM model with embedding LSTM and Dense layers Load in pre-trained embedding.
- ✓ Train model to foresee following work in order.

Each LSTM recurrent unit too upholds a vector called the Internal Cell State which theoretically designates the information that was designated to be reserved by the earlier LSTM recurrent unit [6].

3. Results

3.1 Visualization from Exploratory Analysis

With the help of performing exploratory analysis and the algorithm, we have found some good insights and basic understanding of how the vaccination and the vaccination manufacturers can expect. Some insights discovered during this project research were [9]:

1. Top vaccines in the fight to covid-19.
2. Proportions of the different vaccines.
3. Daily & total vaccinations according to different regions.
4. Vaccinations used by a particular country.
5. Most used vaccination.
6. Total vaccinations of a country grouped by the vaccine.
7. How many people will be vaccinated shortly?
8. When will we be able to achieve a 100% vaccination rate per 100 people (estimations as per previous data)?

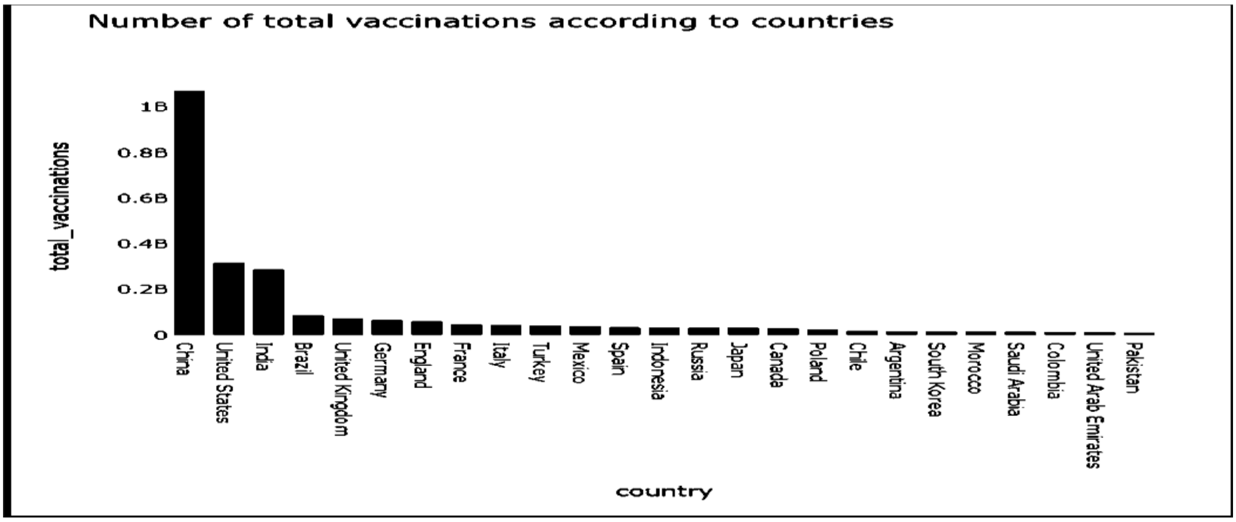


Figure 2. Total vaccinations as per Country

8

this graph represents the number of total vaccinations according to the different and top countries in the vaccination progress.

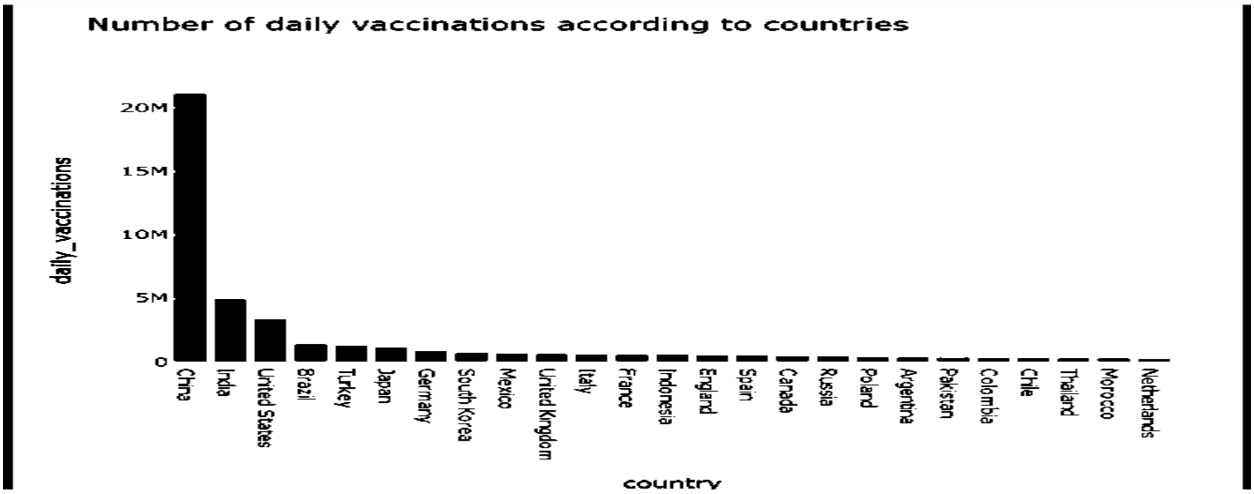


Figure 3. Daily vaccinations as per Country

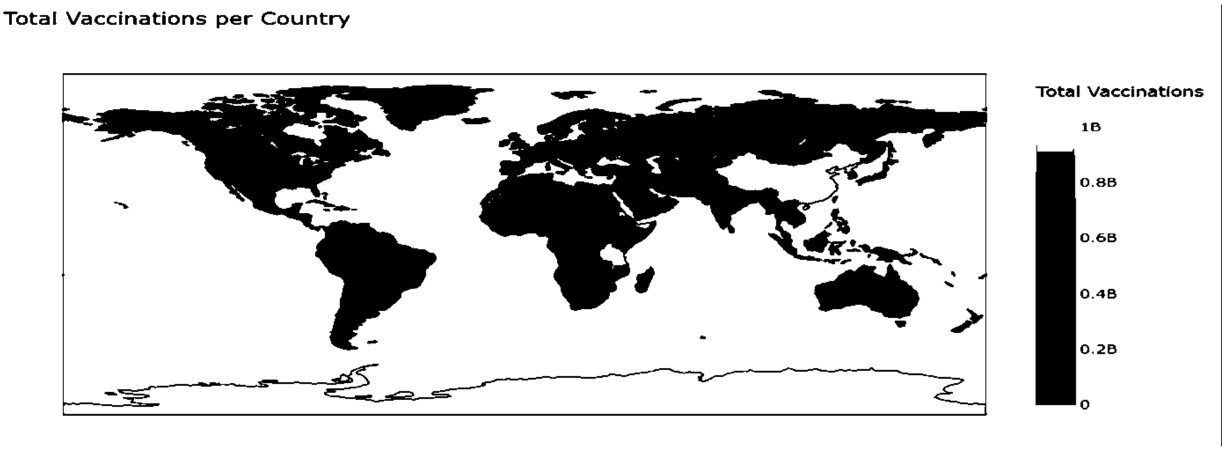


Figure 4. Total vaccinations as per Country displayed on Maps

3.2Visualization for ARIMA

ARIMA models are denoted with the notation ARIMA (p, d, q). These three parameters reason for seasonality, tendency, and noise in data:

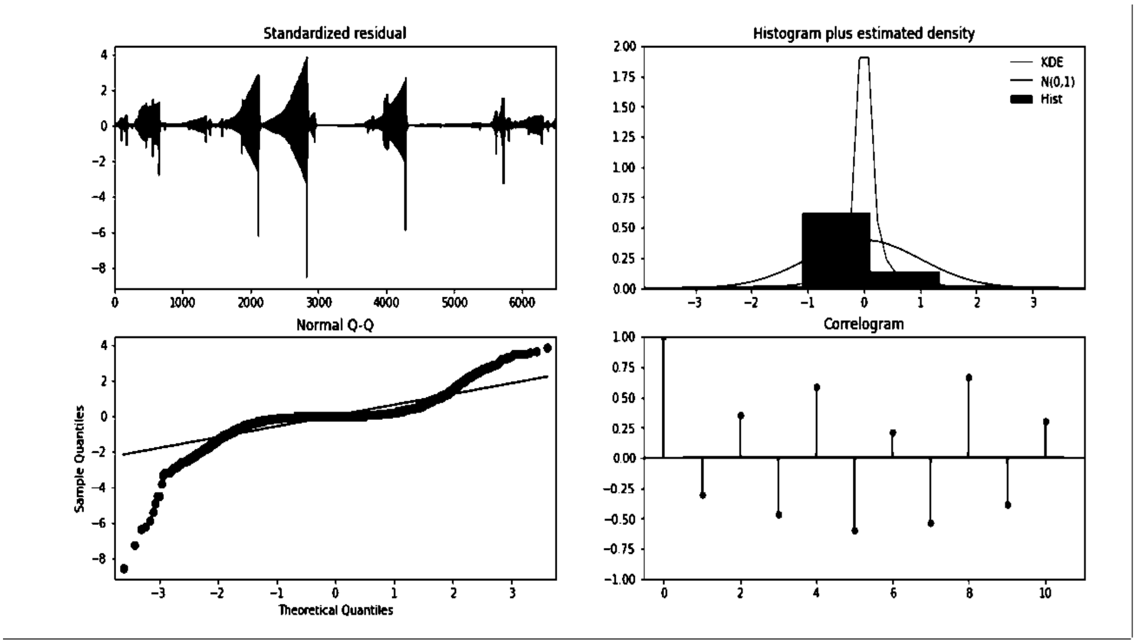


Figure 5. Residual and Correlation

To help us understand the accuracy of forecasts, we have tried to check the past pattern of location, vaccine and total vaccination in different regions.

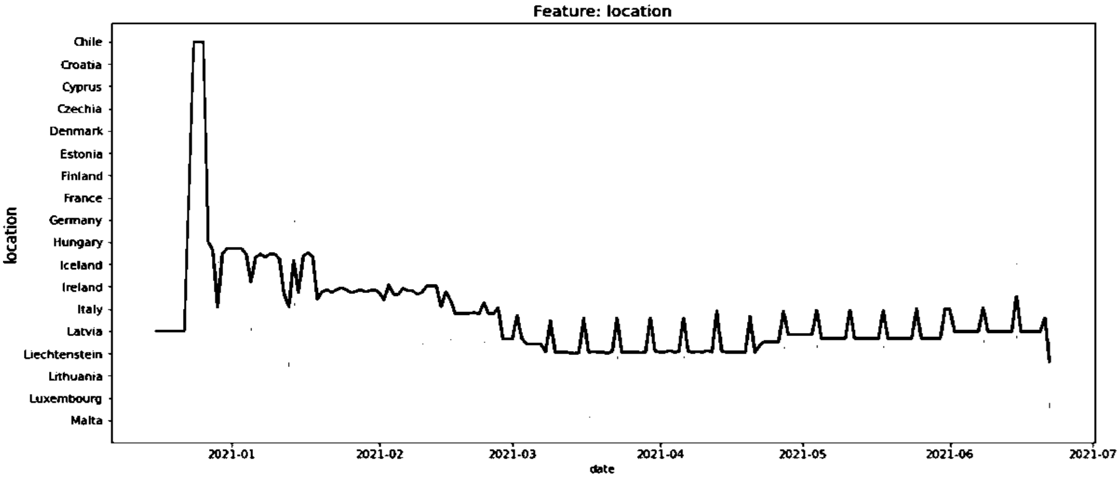


Figure 6. Prediction of vaccination rate per location**.**

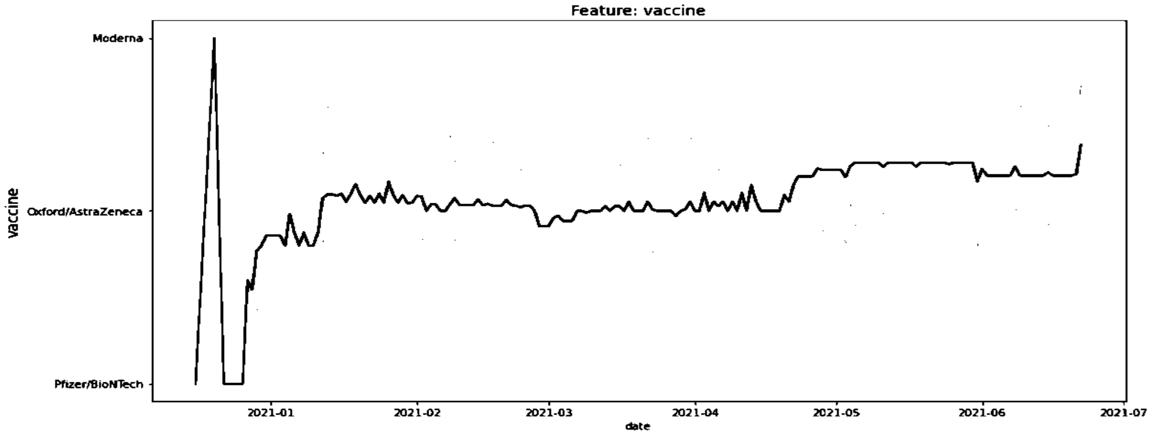


Figure 7. prediction of vaccine rate used

To build the ARIMA model, we need to validate and train the dataset. During the training of the data, validating it infuses new data into the model that would have not been evaluated early. Validation provides the first test against unseen data in neural networks. This type of data helps to build up the model it evaluates the data continuously to learn the nature and behaviour of the data and then adjust accordingly as per the intended purpose. The figure below shows the validation and training of data at intervals.

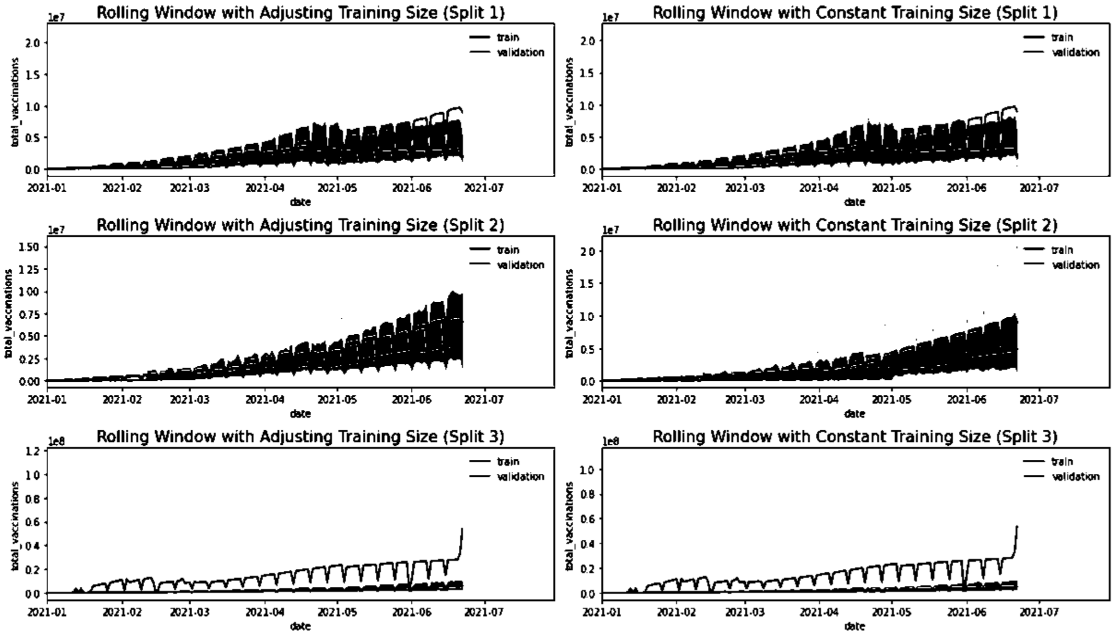


Figure 8. training and validating data

3.3Visualization for LSTM

The standard metrics that are used to compare the time series data will be Root Mean Squared Error and Mean Absolute Percentage Error. The lower the value of RMSE the better the fit of the model[6].

Root Mean Squared Error:

The RMSE is a quadratic marking rule which actions the average greatness of the error. The calculation for the RMSE is given in equally of the references. Articulating the formula in words, the alteration between forecast and corresponding observed values are respectively squared and then averaged completed the sample. Lastly, the square root of the average is occupied. Meanwhile, the errors are squared beforehand they are averaged, the RMSE stretches a relatively high weight to great faults. These incomes the RMSE is most valuable when great errors are mainly unwanted.

$$RMSE = (\sqrt{\frac{\Sigma\ (predicted - actual)^2}{total\ predictions}})$$

Equation (2) Root Mean Squared Error Equation

Mean Absolute Error:

When using the MAE, the error balances linearly. Therefore, an error of 10, is 10 times inferior to an error of 1. In both cases, the error is distinct in the same component of dimension as the target variable. So, the query you need to ask yourself, are advanced errors really that are significant? This depends on the field of your problem.

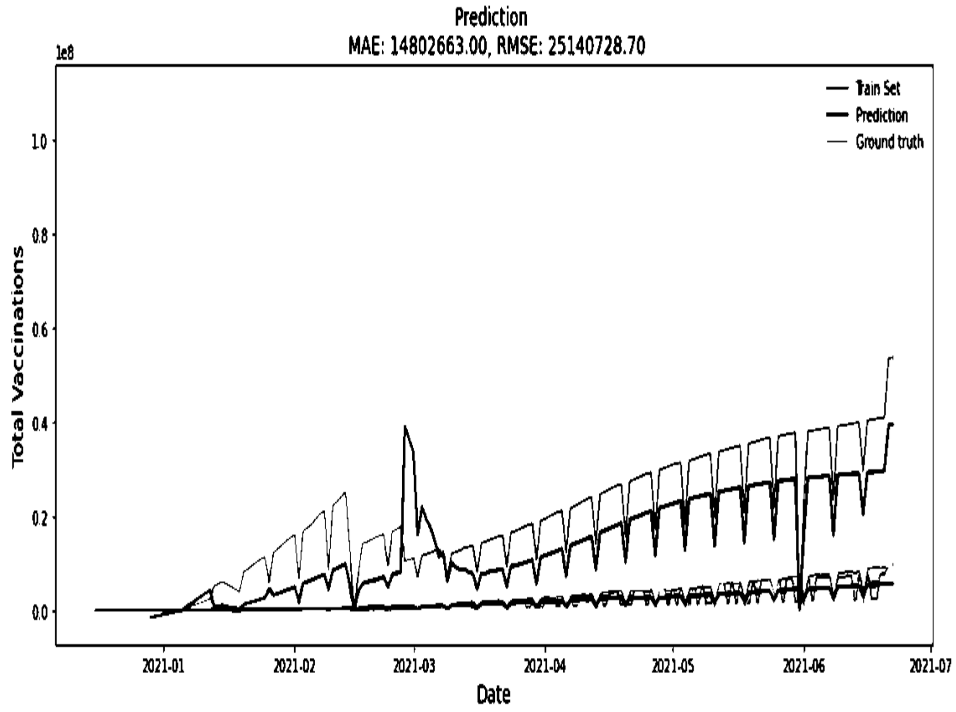
$$MAE = \frac{1}{2}\sum_{J=1}^{N} |YJ - \hat{Y}J|$$

Equation (3) Mean Absolute Error Equation

12

The MAE and the RMSE can be applied together to classify the difference in the errors in a set of estimates. The RMSE will endlessly persevere greater or equivalent to the MAE; the higher variance amid them, the better the change within unrelated errors within the model. If the RMSE=MAE, then all the faults are of the alike degree.

Figure 9. total vaccination prediction



Comparison of both models using different techniques to get to predicted results.

Covariance Type:	opg					
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8105	0.031	26.565	0.000	0.751	0.870
ar.L2	-0.4467	0.039	-11.418	0.000	-0.523	-0.370
ar.L3	0.6150	0.022	27.968	0.000	0.572	0.658
ma.L1	-1.3881	0.011	-126.389	0.000	-1.410	-1.367
ma.L2	0.7287	0.010	75.674	0.000	0.710	0.748
sigma2	2.349e+13	5.29e-15	4.44e+27	0.000	2.35e+13	

Table 1. ARIMA results

```
Model: "sequential"
Layer (type)          Output Shape        Param #
lstm (LSTM)            (None, 1, 128)     92672
lstm_1 (LSTM)           (None, 64)        49408
dense (Dense)          (None, 25)        1625
dense_1 (Dense)         (None, 1)          26
Total params: 143,731
Trainable params: 143,731
Non-trainable params: 0
MAE:1480266.00 ,RMSE:25140728.70
```

Table 1. LSTM results

4. Conclusion

The covid-19 outburst has exaggerated more than 210 countries as of now. The rapid increase in the number of cases caused healthcare industries to collapse. Countries have faced major problems such as medicine, personnel and hospital capacity. Each country was taken on measures to reduce the spread of the virus. In addition, countries are making predictions for the future situation of the covid-19 outbreak by using machine learning methods. Thus, the burden of the health sector can be reduced by foreseeing future situations and making strategies and plans. In this study, we have just done experimentation on the dataset for self-learning purposes. This paper can be used for further research or findings however the analysis and findings don't 100% guarantee that the results generated are accurate this paper is not created for any false impact and does not impact social, medical and welfare society. This entire study is based on experimental analysis using exploratory data analysis with python and good machine learning tools to find good practice. More research on this will be addressed as this was one of the most significant aspects of the previous few years and still, a very concerning matter for future forecasting purposes as this matter is impacting and still an ongoing situation throughout each part of the world today.

References

1. Simran Preet Kaur, Vandana Gupta (2020). "COVID-19 Vaccine: A comprehensive status report", Section of Microbiology, Ram Lal Anand College, University of Delhi, Benito Juarez Road, New Delhi 110021, India, Virus Research, 288, (2020).
2. WHO (World Health Organisation) (2021). " Evaluation of COVID-19 vaccine Effectiveness", c World Health Organization, 2021. Some rights reserved. This effort is obtainable under the CC BY-NC-SA 3.0 IGO licence, WHO reference number: WHO/2019-nCoV/ vaccine effectiveness/measurement/2021.1
3. Pinar Cihan (2021). "Fuzzy Rule-Based System for Predicting Daily Case in COVID-19 Outbreak",Dept. of Computer Engineering, Tekirda_ Nam_K Kemal University, Tekirda, Turkey,pkaya@nku.edu.tr.

4. Gaurav Pant, Alka, Deviram Garlapati, Ashish Gaur, Kaizar Hossain, Shoor Vir Singh, Ashish Kumar Gupta (2020). "Air quality assessment among populous sites of major metropolitan cities in India during COVID-19 pandemic confinement", Environmental Science and Pollution Research (2020) 27:44629–44636, Springer-Verlag GmbH Germany, part of Springer Nature 2020.
5. Alissar Naaser ULCO-lasl, Denis Hamad, Chaibann nasr (2020). "Visualization methods for exploratory data analysis",0-7803-9521-2/06/$20.00 §2006 IEEE.
6. Poorna Shankar and Pawar, (2011). Classification of Global Carbon Emissions using Artificial Neural Networks. International Journal of Computer Applications, [online] Available at:
7. <https://www.ijcaonline.org/archives/volume29/number3/3544-4859> [Accessed 20 August 2021].
8. Poornashankar (2013). Performance Analysis of Different Feed-Forward Networks in Non-Linear Classification. International Journal of Soft Computing and Engineering (IJSCE), [online] Available at:<https://www.ijsce.org/wp-content/uploads/papers/v3i2/B1532053213.pdf> [Accessed 20 August 2021].
9. Support.sas.com. 2021. PROC ARIMA. [online] Available at: <https://support.sas.com/documentation/onlinedoc/ets/132/arima.pdf> [Accessed 20 August 2021].
10. Cheng, D., Schretlen, P., Kronenfeld, N., Bozowsky, N., & Wright, W. (2013). Tile based visual analytics for Twitter big data exploratory analysis. 2013 IEEE International Conference on Big Data. doi:10.1109/bigdata.2013.66917.
11. Ariyo, A., Adewumi, A. and Ayo, C., (2014). Stock Price Prediction Using the ARIMA Model. 2014 UKSim- AMSS 16th International Conference on Computer Modelling and Simulation.

***