IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007029 |

Time Series Analysis and Prediction of Covid-19 Using Sarima Model

G.Soujanya [1], T.Bala Sai Kumar [2], L.Uday Kiran [3] ,A.Adharsh [4],S.Jaya Sayee Teja [5]

Associate Professor, Department of Computer Science and Engineering, JB Institute of Engineering and Technology, Moinabad, Telangana, India[1]

B.Tech. Student, Department of Computer Science and Engineering, JB Institute of Engineering and Technology, Moinabad, Telangana, India[2,3,4,5]

ABSTRACT: This exploration is centered around the information investigation for the accessible information for COVID-19 pandemic sickness. In this examination work, Python and its libraries are applied for the exploratory information investigation of this optional dataset. Thinking about the variety of the situation with time, it has been seen to investigate the information with the time series examination to gauge the future impact of Coronavirus universally or separately. This examination has been led utilizing seven conjecture techniques. Yet, the technique with the least errors are accounted for here, viz.SARIMA. It is so much needed to gauge the future prospects in such arbitrary and special event of the pandemic all throughout the globe. This examination helps numerous analysts and researchers to comprehend the measurable estimating which will be an extraordinary help for future readiness. This examination will be useful for the authoritative and social elements to handle this pandemic across the country.We additionally set up a dashboard on the cases and vaccinations all over the country

KEYWORDS: COVID-19,Timeseries forecasting , SARIMA , Adfuller test ,Streamlit.

I.INTRODUCTION

Covid infection (COVID-19) is another sickness brought about by the SARS-COV-2 infection. The infection initially started in Wuhan, Wuhan territory in December 2019. While from the outset it's anything but a progression of pneumonia with obscure reason in Wuhan, it's anything but a worldwide emergency in under a month. As a result, nations have been secured, public spots have been shut, and different other movement restricting strategies have been executed to hinder the spread of the illness. The COVID-19 infection spreads essentially through drops that come out from an individual's mouth or nose as they sniffle or hack. It may not sound destructive if individuals play protected and not hacking or wheezing recklessly, however the way that it has spread through the globe denies the way that COVID-19 can't be treated as lethal. Presently, COVID-19 is a green exploration theme as the entire world is enduring and battling in this time because of this Pandemic infection. As there is no drug and technique to dispose of this worldwide emergency, foresee the future opportunities for future readiness. This work is valuable for the forecast of future cases to help a few social substances and associations, viz. emergency clinics, drugs, NGOs, Government bodies, and so forth for their preparation to battle.

Time series forecasting is considered one of the most applied data science techniques that are used in different industries such as finance, supply chain management, production, and inventory planning. Stock prices forecasting, weather forecasting, business planning, resource allocation are only a few of the many possible applications for time series forecasting.The predictive models based on machine learning found wide implementation in time series projects required by various businesses for facilitating predictive distribution of time and resources.

Time Series pertains to the sequence of observations collected in constant time intervals be it daily, monthly, quarterly or yearly. Time Series Analysis involves developing models used to describe the observed time series and understand the "why" behind its dataset. This involves creating assumptions and interpretations about a given data. Time Series Forecasting makes use of the best fitting model essential to predicting the future observation based on complex processing current and previous data.

The dataset is large, the investigation should be possible utilizing various strategies and viewpoints. To break down this dataset, it calls the measurable information examination strategies for future expectations. This work tests distinctive guaging strategies utilizing Python libraries to establish the one with minimal mistakes in estimating. This

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007029 |

exploration is centered around the information examination for the accessible information for COVID-19 pandemic sickness. In this examination work, Python and its libraries are applied for the exploratory information investigation of this optional dataset. Thinking about the variety of the situation with time, it has been seen to dissect the information with the time series examination (TSA) to conjecture the future impact of Coronavirus universally or separately. This examination has been led utilizing four estimate strategies. In any case, just the one technique with the least errors are accounted for here, viz. SARIMA..

II.RELATED WORK

This work analyses the optional dataset taken from the COVID-19 India API . The work proposes an exploratory information investigation that depends on a period series examination of the COVID-19 dataset. Python and its libraries are used for this exploratory information investigation and time series examination as these libraries are productive and powerful in such sort of information investigation for future forecasts. The accompanying python libraries are discovered valuable in this work:

The Wes McKinney composed the pandas library codes which is an open-source and free for use. It's anything but an in vogue python library that is utilized for information investigation, control, and cleaning. There are a great deal of capacities and strategies accessible in the pandas, which are exceptionally valuable for information investigation.

NumPy library is utilized for Scientific registering; its complete name is the Numeric Python. NumPy is written in the Python and the C language, it's anything but an easy method to store the broad information as it takes next to no memory to store.

Sklearn library is utilized in AI for information mining and information investigation. The fundamental highlights of the sklearn are identified with dimensionality decrease, relapse, prerecession, grouping, and bunching. Math module is utilized for every one of the numerical activities like likelihood, geometry, and measurement activity.

Matplotlib library is utilized to make the diagrams, John D Hunter composed the codes and presented this library. It is very simple to use for the individuals who definitely know the MATLAB and different charts plotting instruments.

Seaborn library is for the most part used to improve the information perception. It is more instinctive than the Matplotlib, it is more enlightening and alluring illustrations instrument .

In our investigation, the overall dataset is utilized for better forecasts and near the real world. For future preparing in this undertaking, we utilized TSA since we have only one variable,i.e., time. TSA assists with tracking down the future worth that is reliant upon the time very much like an expectation of the stock cost or business gauge. Besides, when any information is gathered in equivalent stretch actually like month to month, feebly, and yearly then we can utilize TSA for determining future worth. More often than not series have prepared irregularity and abnormality related with them. Some of them do have a cyclic example. The pattern is a development to somewhat sequential qualities over an extensive stretch of the time. At the point when the time series examination shows the vertical example or descending, it is known as the upswing or downtrend individually. Irregularity implies rehashing designs in the fixed time stretch. For instance, winters or Summers happens each year at a set time; which acquires the increment the convenient interest of the blowers or airconditioners. Anomaly is additionally called commotion, so these are whimsical in nature, or unsystematic remaining occurs for a brief span. It is non-rehashing and not unsurprising. For instance, in the state of catastrophic event, the interest for the prescriptions and medical services items increments. The last part of the time series examination is the Cyclic. The cyclicity communicates the repeatitive here and there developments which neither have any fixed example nor any fixed timings.

At the point when the qualities are the steady and worth as the capacity (sinx, cosx), then, at that point we can't utilize the Time Series Analysis. Any sort of factual model that will apply to time series, there ought to be stationarity in TSA information. It could be a likelihood that our information conduct follow a similar example later on. For checking the stationarity, there are two well known strategies, i.e., the Rolling Statistics and ADCF test.

The above point is invaluable to know for this exploration work. There are four strategies for the Time Series Analysis

A. Holt's Linear Trend Method

This strategy is utilized when the dataset has a pattern that implies the dataset pursues the expanding or diminishing direction.

B. Holt's Winter Model

This strategy is utilized when the dataset has the irregularity.

C. Arima Model

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007029 |

ARIMA is perhaps the best technique for the TSA. This is the mix of the two models Autoregressive (AR) and Moving Average (MA). 'AR' and 'MA' are two separate models, which creates another model by restricting both through coordinating by 'I'. AR implies the connection between's the past period and the current. ARIMA models have three-boundary p, q, and d. The boundary, p, presents the autoregressive slacks, q presents the moving average and d presents the request for separation. On the off chance that we take the combination by only one request, the worth of d would be one, assuming separate it in the request for two, we have the worth of the d would be 2, etc. On the off chance that it is needed to anticipate the worth of the p, the PACF chart is required. While for the expectation of the worth of the q, the ACF chart is to be plotted.

D.SARIMA Model

Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an expansion of ARIMA that unequivocally upholds univariate time series information with an occasional segment. It adds three new hyperparameters to indicate the autoregression (AR), differencing (I) and moving average (MA) for the seasonal segment of the series, just as an extra boundary for the time of the irregularity.

III.METHODOLOGY

The suggested system is a COVID-19 India API data-based Timeseries Forecasting Model. We have gathered data required for our forecasting from the public API provided in the Github.We carefully examine our dataset because the characteristics of the data can strongly affect the model results.The dataset consists of many columns, we remove some of them which are not useful for our forecasting.We should also be sure to check for and deal with any missing values. There are many other data preparation steps to consider depending on our analytical approach and business objectives. Using the pandas package, we took some preparation steps with dataset so that it's slightly cleaner than most real-life datasets. we checked for missing data and included only two columns: 'Date' and 'Confirmed'.Finally, we indexed our data with time so that rows will be indicated by a date rather than just a standard integer,such that the values in the first column will be in YYYY-MM-DD date format.Firstly we will make our data stationary using Adfuller test.We also plot the ACF and PACF for finding the correlation between the data and we find p and q values.Finally we fit the dat to our SARIMA model and we will predict the future cases. The system modules used for our proposed VQA system are:

- **Dataset Preparation And Time Series Analysis:** We have gathered data required for our forecasting from the public API provided in the Github.We carefully examine our dataset because the characteristics of the data can strongly affect the model results. Using the pandas package, we took some preparation steps with dataset so that it's slightly cleaner than most real-life datasets. we checked for missing data and included only two columns: 'Date' and 'Confirmed'.Finally, we indexed our data with time so that rows will be indicated by a date rather than just a standard integer,such that the values in the first column will be in YYYY-MM-DD date formatThe next step is simply to plot the dataset. In this project, we use the matplotlib and pyplot packages.We can observe whether the data has any trend or seasonality We also decompose the data by using seasonal decompose function from statsmodel package.Then we perform Augmented Dickey-Fuller test to determine whether the processed data is stationary or not with different levels of confidenceand next we will make our data stationary by detrending and differencing techniques.We will finally divide the data set into training and testing data sets for fitting into the model.

- **Training the Model:** This phase involves training and testing the model.We use SARIMA for training.We will find RMSE values for predicting the errors and find the perfect model for our data.

- **Model Prediction:** We will use the model to predict future values based on previously observed values.We will finally predict the future Covid-19 cases and we will find the upper and lower boundaries of our prediction and we will visualize our predictions

- **Dashboard using Streamlit:** We will prepare a dashboard using the data we have and we will also make our predictions visible in that dashboard.We use streamlit library in python for preparing the dashboard. Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps. This dashboard contains data visualization of covid of all the states of India. Data is visualized with the help of a bar chart, pie chart, and line graph.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007029 |

IV. EXPERIMENTAL RESULTS

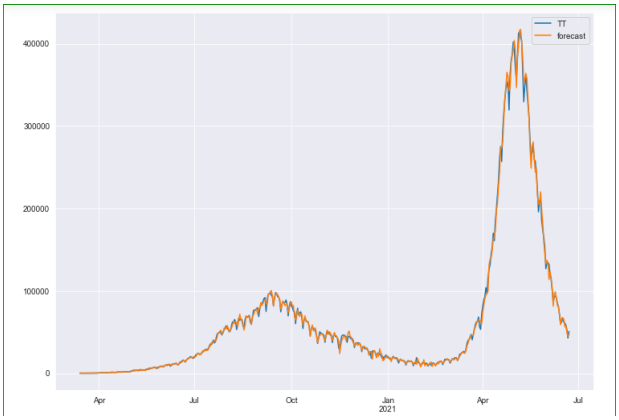

Fig.1.Image of our forecasting model showing its predictions for the actual data

	date	values
0	2021-06-23	56005.509063
1	2021-06-24	53967.064311
2	2021-06-25	52917.423785
3	2021-06-26	52135.282284
4	2021-06-27	47820.067708
5	2021-06-28	40951.270879
6	2021-06-29	48304.383655
7	2021-06-30	53922.449676

Fig 2:Smaple of our predictions for the next seven days

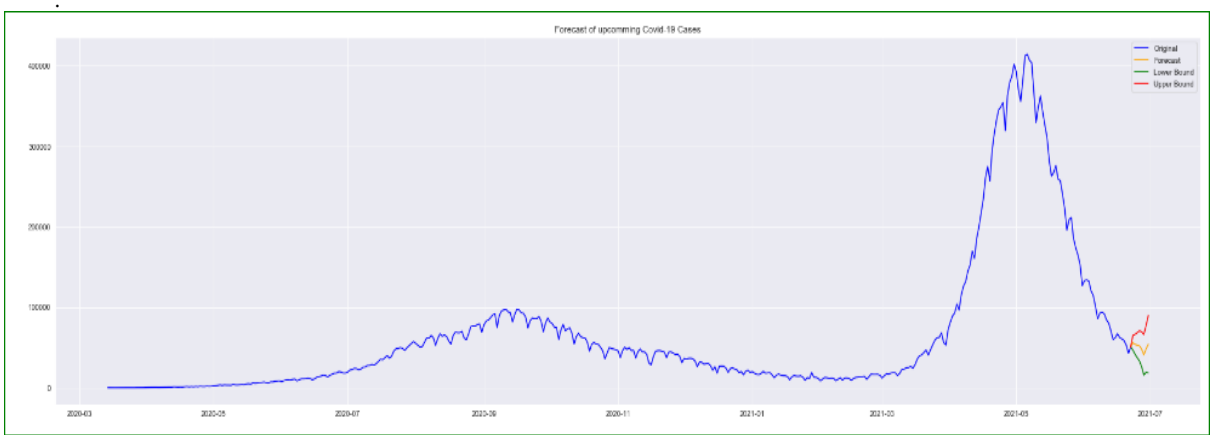

Fig.3.Graph showing confidence intervals of our predictions

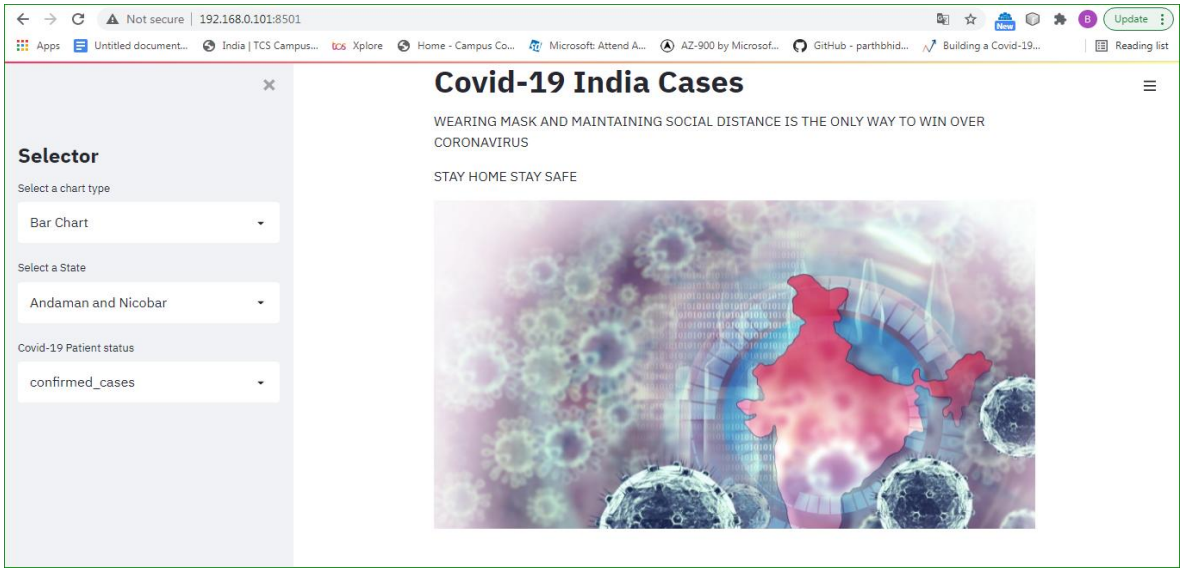

Fig.4.Screenshot of our Dashboard

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

| DOI:10.15680/IJIRSET.2021.1007029 |

V.CONCLUSION

This research works are useful for future expectation of the affirm instances of the covid-19 which is the useful for the public authority elements and other medical care associations to attempt to moderate the issue of the flare-up. also, the codes written in python libraries are useful to comprehend the guaging techniques and the distinctions. in this work, overall information is utilized and tried for four most popular model to forcast the future scenarion. the dataset might be expanding with time, the codes created in this work are simply need a re-run in the future to see the future worth of the affirm cases whenever. in the current work, four model are utilized for the forecast lastly those strategy wherein blunder will be least is picked. in light of the mse esteems, the sarimax model is tracked down the best model as the mistake in this model is exceptionally low as contrast with the other models.at present, the sarimax model is tracked down the best one. different strategies are less appropriate for the expectation of the pestilence on the grounds that datapoint is extremely less. in future, as the datapoints will be exceptionally high, some other model may likewise think of the extraordinary bits of knowledge.

REFERENCES

[1] G. Singh, "7 methods to perform Time Series forecasting (with Python codes)", Analytics Vidhya.
[2] A. Jain, "A comprehensive beginner's guide to create a time series Forecast", Analytics Vidhya
Almasarweh
[3] M, Wadi SAL ARIMA Model in Predicting Banking Stock Market DataModern Applied Science, 12 (2018), p. 309
[4] J. Duk Seo, "Trend Seasonality Moving Average Auto Regressive Model: My Journey to time Series Data with Interactive code", *Towards data science*.
[5] G. R. Shinde, A.B. Kalamkar, P.N. Mahalle, N. Dey, J. Chaki and A.E. Hassanien, "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art", SN Computer Science, vol. 1, no. 4, pp. 1-15, 2020.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details