International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

TIME SERIES ANALYSIS AND PREDICTION OF COVID-19 USING FORECASTING MODELS

1 K YASASWINI, Assistant Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India, yasaswini@sreyas.ac.in

2 R Sai Chandu, 3 D Rakesh Kumar, 4 M Hemanth, Department of CSE, Sreyas Institute of Engineering and Technology, Telangana, India.

ABSTRACT: This research is focused on the data analytics for the available data for COVID-19 pandemic disease. In this research work, Python and its libraries are applied for the explanatory data analytics of this secondary dataset. Considering the variation of the scenario with time, it has been observed to analyze the data with the time series analysis in order to forecast the future effect of Coronavirus globally or individually. This analysis has been conducted using seven forecast methods. But the method with the least errors are reported here, viz.SARIMAX. It is so much required to forecast the future possibilities in such random and unique occurrence of the pandemic around the world. This analysis helps many researchers and scientists to understand the statistical forecasting which will be a great support for future preparedness. This analysis will be helpful for the organizational and social entities to tackle this pandemic across the country . We also prepare a dashboard on the cases, vaccinations across the country.

1. INTRODUCTION

Coronavirus disease (COVID-19) is a new disease caused by the SARS-COV-2 virus. The virus first originated in Wuhan, Hubei province in December 2019. While at first it is just a series of pneumonia with unknown cause in Wuhan, it quickly became an international crisis in less than a month. Almost six million people have been infected with over three hundred thousand deaths worldwide. In effect, countries have been locked down, public places have been closed, and various other activity-limiting policies have been implemented to slow down the spread of the disease. The COVID-19 virus spreads primarily through droplets that come out from a person’s mouth or nose as they sneeze or cough. It may not sound deadly if people play safe and not coughing or sneezing carelessly, but the fact that it has spread through the globe denies the fact that COVID-19 cannot be treated as deadly. Currently, COVID-19 is a green research topic as the whole world is suffering and struggling in this time due to this Pandemic disease. As there is no medication and strategy to get rid of this global crisis, it is important to predict the future possibilities for future preparedness. This work is useful for the prediction of future cases to support several social entities and organizations, viz. hospitals, pharmaceuticals, NGOs, Government bodies, etc for their readiness to combat. Machine learning is an application of AI that provides the ability to automatically learn and improve from experience without being explicitly programmed. The whole process of machine learning is explained in the fig: 1. In case of a normal process the input is given as data and the output would the answers or results in a generalized manner. But in machine learning the user makes the model to analyze so it gives data along with answers as input and the output will be the optimized solutions.

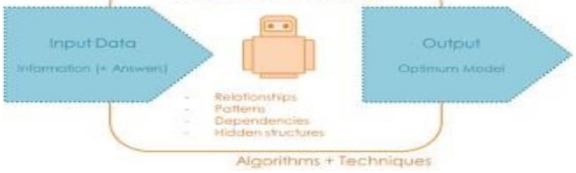


Fig.1: Example figure

The dataset is large, and the analysis can be done using different techniques and perspectives. To analyze this dataset, it calls the statistical data analytics techniques for future predictions. This work tests different forecasting techniques using Python libraries [3]–[4][5] to found the one with the least errors in forecasting. This research is focused on the data analytics for the available data for COVID-19 pandemic disease. In this research work, Python and its libraries are applied for the exploratory data analysis of this secondary dataset. Considering the variation of the scenario with time, it has been observed to analyze the data with the time series analysis (TSA) in order to forecast the future effect of Coronavirus globally or individually. This analysis has been conducted using seven forecast methods. But only the methods with the least errors are reported here, viz.SARIMAX.

2. LITERATURE REVIEW

ARIMA Model in Predicting Banking Stock Market Data:

Banking time series forecasting gains a main rule in finance and economics which has encouraged the researchers to introduce a fit models in forecasting accuracy. In this paper, the researchers present the advantages of the autoregressive integrated moving average (ARIMA) model forecasting accuracy. Banking data from Amman stock market (ASE) in Jordan was selected as a tool to show the ability of ARIMA in forecasting banking data. Therefore, Daily data from 1993 until 2017 is used for this study. As a result this article shows that the ARIMA model has significant results for short-term prediction. Therefore, these results will be helpful for the investments.

Predicting tourism demand by A.R.I.M.A. models:

The paper provides a short-run estimation of international tourism demand focusing on the case of F.Y.R. Macedonia. For this purpose, the Box–Jenkins methodology is applied and several alternative specifications are tested in the modelling of original time series and international tourist arrivals recorded in the period 1956–2013. Upon the outcomes of standard indicators for accuracy testing, the research identifies the model of A.R.I.M.A.(1,1,1) as most suitable for forecasting. According to the research findings, a 13.9% increase in international tourist arrivals is expected by 2018. The forecasted values of the chosen model can assist in mitigating any potential negative impacts, as well as in the preparation of a tourism development plan for the country.

Real-time prediction of influenza outbreaks in Belgium Epidemics:

Seasonal influenza is a worldwide public health concern. Forecasting its dynamics can improve the management of public health regulations, resources and infrastructure, and eventually reduce mortality and the costs induced by influenza-related absenteeism. In Belgium, a network of Sentinel General Practitioners (SGPs) is in place for the early detection of the seasonal influenza epidemic. This surveillance network reports the weekly incidence of influenza-like illness (ILI) cases, which makes it possible to detect the epidemic onset, as well as other characteristics of the epidemic season. In this paper, we present an approach for predicting the weekly ILI incidence in real-time by resorting to a dynamically calibrated compartmental model, which furthermore takes into account the dynamics of other influenza seasons. In order to validate the proposed approach, we used data collected by the Belgian SGPs for the influenza seasons 2010–2016. In spite of the great variability among different epidemic seasons, providing weekly predictions makes it possible to capture variations in the ILI incidence. The confidence region becomes more representative of the epidemic behavior as ILI data from more seasons become available. Since the SIR model is then calibrated dynamically every week, the predicted ILI curve gets rapidly tuned to the dynamics of the ongoing season. The results show that the proposed method can be used to characterize the overall behavior of an epidemic.

Monitoring the SARS epidemic in China: a time series analysis:

In this article, we studied three types of time series analysis methods in modeling and forecasting the severe acute respiratory syndrome (SARS) epidemic in mainland China. The first model was a Box-Jenkins model, autoregressive model with order 1 (AR(1)). The second model was a random walk (ARIMA(0,1,0)) model on the log transformed daily reported SARS cases and the third one was a combination of growth curve fitting and autoregressive moving average model, ARMA(1,1). We applied all these three methods to monitor the dynamic of SARS in China based on the daily probable new cases reported by the Ministry of Health of China.

Prediction and analysis of Coronavirus Disease 2019:

The outbreak of Corona Virus Disease 2019 (COVID-19) in Wuhan has significantly impacted the economy and society globally. Countries are in a strict state of prevention and control of this pandemic. In this study, the development trend analysis of the cumulative confirmed cases, cumulative deaths, and cumulative cured cases was conducted based on data from Wuhan, Hubei Province, China from January 23, 2020 to April 6, 2020 using an Elman neural network, long short-term memory (LSTM), and support vector machine (SVM). A SVM with fuzzy granulation was used to predict the growth range of confirmed new cases, new deaths, and new cured cases. The experimental results showed that the Elman neural network and SVM used in this study can predict the development trend of cumulative confirmed cases, deaths, and cured cases, whereas LSTM is more suitable for the prediction of the cumulative confirmed cases. The SVM with fuzzy granulation can successfully predict the growth range of confirmed new cases and new cured cases, although the average predicted values are slightly large. Currently, the United States is the epicenter of the COVID-19 pandemic. We also used data modeling from the United States to further verify the validity of the proposed models.

3. METHODOLOGY

The mathematical modelling of coronavirus disease-19 (COVID-19) pandemic has been attempted by a wide range of researchers from the very beginning of cases in India. Initial analysis of available models revealed large variations in scope, assumptions, predictions, course, effect of interventions, effect on health-care services, and so on. Thus, a rapid review was conducted for narrative synthesis and to assess correlation between predicted and actual values of cases in India. This review has clearly shown the importance of assumptions and strong correlation between short-term projections but uncertainties for long-term predictions. The results for long-term predictions could not be synthesized as very few studies have provided the same. The shortterm predictions may be revised as more and more data become available. The assumptions too will expand and firm up as the pandemic evolves because at the start of pandemic, data are sparse and making correct assumptions is difficult. Models with more realistic assumptions may be developed subsequently. There is a case for state-specific models in our country owing to the large variation in assumptions for each state.

This research is focused on the data analytics for the available data for COVID-19 pandemic disease. In this research work, Python and its libraries are applied for the exploratory data analysis of this secondary dataset. Considering the variation of the scenario with time, it has been observed to analyze the data with the time series analysis (TSA) in order to forecast the future effect of Coronavirus globally or individually. This analysis has been conducted using seven forecast methods. But only the methods with the least errors are reported here, viz. SARIMAX This model also provides a dashboard built using streamlit library in python which helps to draw meaningful insights from data. This work is useful for the prediction of future cases to support several social entities and organizations, viz. hospitals, pharmaceuticals, NGOs, Government bodies, etc for their readiness to combat.

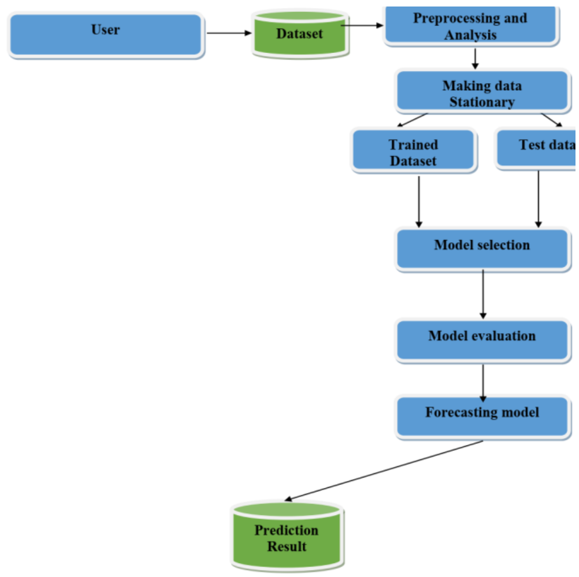


Fig.2: System architecture

MODULES:

DATASET:

From the Covid-19 India API the data set is extracted. Basically it consists of day wise confirmed cases,recoverd cases and deaths in India from the starting day of pandemic.It alos consists of statewise daily cases and deaths..

PREPROCESSING AND ANALYSIS:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task. We will analyse the data by plotting the graphs and check whether there is any trend or seasonality and draw some meaningful sights from the graph.;

MAKING DATA STATIONARY:

Non-stationarity is when the statistical properties of a series, e.g the mean, variance, and covariance (or the process generating the series) changes over time. Non-stationary series are typically difficult to model and forecast and are therefore required to be made stationary to obtain meaningful results as many statistical tools and processes require stationarity. A proven method of stationarizing a non-stationary series is through the use of differencing.

TRAINING DATA:

Simply put, training data is used to train an algorithm. Generally, training data is a certain percentage of an overall dataset along with testing set. As a rule, the better the training data, the better the algorithm or classifier performs. Assuming that your test set meets the preceding two conditions, your goal is to create a model that generalizes well to new data. Our test set serves as a proxy for new data. For example, consider the following figure. Notice that the model learned for the training data is very simple. This model doesn't do a perfect job—a few predictions are wrong.

TESTING DATA:

Some test data is used to confirm the expected result, i.e., When test data is entered the expected result should come and some test data is used to verify the software behavior to invalid input data. Test data is generated by testers or by automation tools which support testing. Most of the times in regression testing the test data is re-used, it is always a good practice to verify the test data before re-using it in any kind of test.

MODEL SELECTION:

Model selection is the process of selecting one final time series forecasting model from among a collection of candidate models for a training dataset. Model selection is a process that can be applied both across different types of models (e.g. Naïve forecasting,ARIMA,SARIMA, etc.)

MODEL EVALUATION:

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data. Methods for evaluating a model's performance are divided into 2 categories: namely, holdout and Cross-validation. Both methods use a test set (i.e data not seen by the model) to evaluate model performance.

FORECASTING MODEL:

Forecasting models are used to forecast future data as a function of past data. They are appropriate to use when past numerical data is available and when it is reasonable to assume that some of the patterns in the data are expected to continue into the future.

4. IMPLEMENTATION

CLASSICAL METHODS

Naïve model:

In most cases, naïve models are applied as a random walk (with the last observed value used as a unit for the next period forecast) and a seasonal random walk (with a value from the same period of the last observed time span used as a unit of the forecast).

Exponential Smoothing Model:

The foundation of machine learning time series classification. Forecasts are made on the basis of captured weighted averages and according to weights decreasing as the observer tracks back in time. A number of extensions of the simple exponential smoothing (SES) have been introduced to include the trend/damped trend and seasonality.

ARIMA/SARIMA:

ARIMA stands for the combination of Autoregressive (AR) and Moving Average (MA) approaches within building a composite model of the time series. ARIMA models includeparameters to account for season and trend (for instance, dummy variables for weekdays and they're distinguishing). In addition, they allow for the inclusion of autoregressive and moving average terms to handle the autocorrelation embedded in the data.

SARIMA stands for Seasonal Autoregressive Integrated Moving Average: it widens the application of the ARIMA by including a linear combination of seasonal past values and/or forecast errors.

Linear Regression method:

Another time series forecasting example. Linear regression is the simple statistical technique commonly used for predictive modeling. Breaking it down to basics, it comes to providing an equation of independent variables, on which our target variable is built upon.

MACHINE LEARNING METHODS

Multi-Layer Perceptron (MLP):

As an applied machine learning approach, the MLP model implies the triple structure of the initial layer of the network which takes in an input, a hidden layer of nodes, and an output layer used to make a prediction.

Recurrent Neural Network (RNN):

RNNs are basically neural networks with memory that can be used for predicting timedependent targets. Recurrent neural networks can memorize the previously captured state of the input to make a decision for the future time-step. Recently, lots of variations have been introduced to adapt Recurrent Networks to a variety of domains.

Long Short-Term Memory (LSTM):

LSTM cells (special RNN cells) were developed to find the solution to the issue with gradients by presenting several gates to help the model make a decision on what information to mark as significant and what information to ignore.

GRU is another type of gated recurrent network. Besides methods mentioned above, Convolutional Neural Network models, or CNNs for short, as well as decision tree-based models like Random Forest and Gradient Boosting variations (LightGBM, CatBoost, etc.) can be applied to time series forecasting.

5. EXPERIMENTAL RESULTS

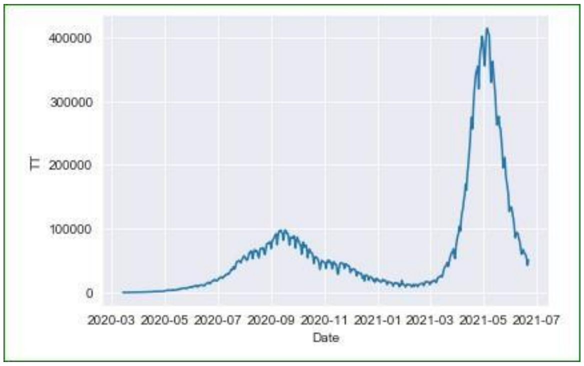


Fig.3: daily analysis of covid 19 cases

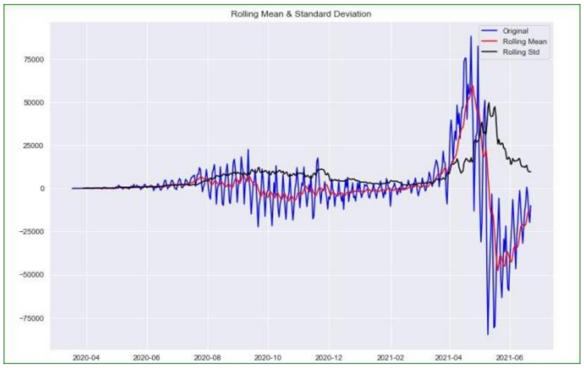


Fig.4: rolling mean and standard deviation after making data stationary

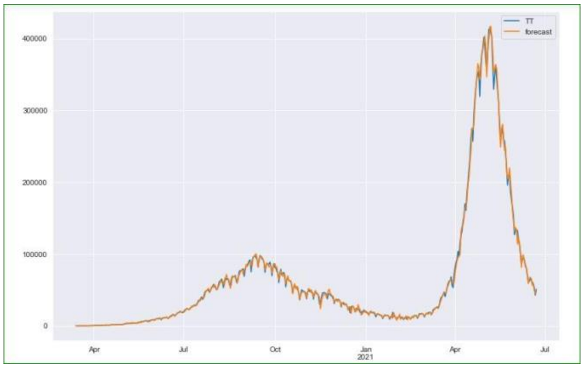


Fig.5: prediction of our model

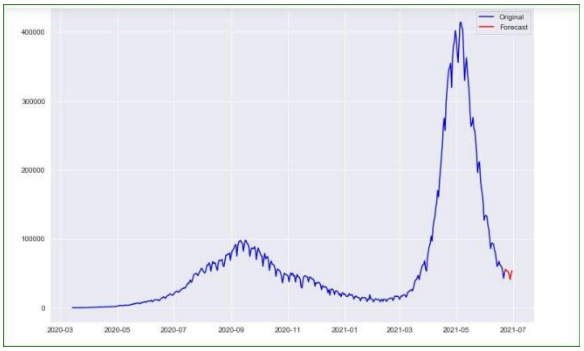
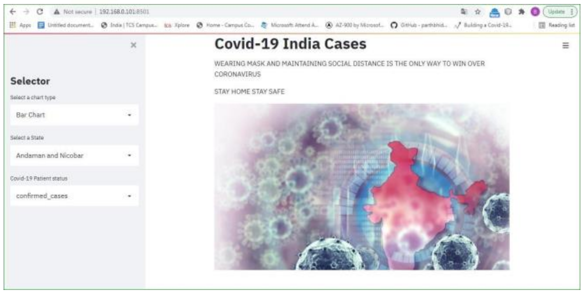


Fig.6: forecasting of upcoming covid-19 cases



Covid-19 India Cases

WEARING MASK AND MAINTAINING SOCIAL DISTANCE IS THE ONLY WAY TO WIN OVER CORONAVIRUS

STAY HOME STAY SAFE

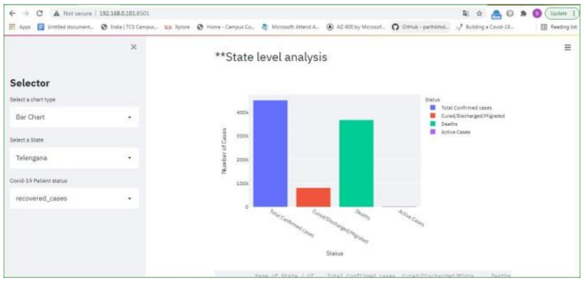
Selector

Select a chart type: Bar Chart

Select a State: Andaman and Nicobar

Covid-19 Patient status: confirmed_cases

Fig.7: screenshot of covid 19 dashboard



**State level analysis

Selector

Select a chart type: Bar Chart

Select a State: Telengana

Covid-19 Patient status: recovered_cases

Name of State / UT Total Confirmed cases Cured/Discharged/Migra... Deaths

Fig.8: creenshot of dashboard showing covid-19 statistics in telangana

6. CONCLUSION

The first case of the coronavirus found in the Wuhan city of China in the seafood market, slowly this virus spread more than 200 countries. India is also affected by this virus the number of the confirmed cases found in India is the 3,00,27,883 in which 2,89,87,479 cases are recovered and in which 3,90,691 people died. The symptoms of the coronavirus are the same for all ages person, but some people fill different symptoms. The common symptoms are Fever, tiredness, cough, while some people may experience diarrhea, sore throat, runny nose, nasal congestion, loss in taste and smell, etc. This findings of this research works are helpful for future prediction of the confirm cases of the COVID-19 which is the helpful for the government entities and other healthcare organizations to try to mitigate the problem of the outbreak. Moreover, the codes written in Python libraries are helpful to understand the forecasting methods and the differences. In this work, worldwide data is used and tested for four most famous model to forecast the future scenario. The dataset may be increasing with time, the codes developed in this work are just need a re-run in the future to see the future value of the confirm cases anytime. In the present work, four model are used for the prediction and finally those method in which error will be minimum is chosen. Based on the MSE values, the SARIMAX model is found the best model as the error in this model is very low as compare to the other models. At present, the SARIMAX model is found the best one. The other methods are less suitable for the prediction of the epidemic because data point is very less. In future, as the data points will be very high, any other model may also come up with the great insights.

7. FUTURE SCOPE

This paper discusses the various time series forecasting models such as ARIMA and SARIMAX which were applied to the data set. It utilizes the time series data of previous dates and then tries to predict the possible cases of COVID-19 in the future. In the present work, four models are used for the prediction and finally those method in which error will be minimum is chosen. Based on the MSE values, the SARIMAX model is found the best model as the error in this model is very low as compare to the other models. At present, the SARIMAX model is found the best one. The other methods are less suitable for the prediction of the epidemic because data point is very less. In future, as the data points will be very high, any other model may also come up with the great insights. There can be other machine learning algorithms MLP, RNN and LSTM which can be used for perediction. In the future we can also add predict the spread by using the rate of vaccination process presently going on in our country. We can also develop the dashboard by adding vaccination details along with hospital details and beds etc and we can make that available on the internet so that people get awareness of the pandemic.

REFERENCES

1. J. Brownlee, "How to identify and remove seasonality from time series data with python", Machine Learning Mastery.
2. Almasarweh M, Wadi SAL ARIMA Model in Predicting Banking Stock Market Data Modern Applied Science, 12 (2018), p. 309
3. Petrevska B. Predicting tourism demand by A.R.I.M.A. models Economic ResearchEkonomska Istraživanja, 30 (2017), pp. 939-950 Jan
4. Petrevska B Forecasting international tourism demand: The evidence of Macedonia UTMS Journal of Economics, 3 (2012), pp. 45-55
5. Miranda GHB, Baetens JM, Bossuyt N, Bruno OM, Baets BD Real-time prediction of influenza outbreaks in Belgium Epidemics, 28 (2019), p. 100341
6. J. Duk Seo, "Trend Seasonality Moving Average Auto Regressive Model: My Journey to time Series Data with Interactive code", Towards data science.
7. J. Brownlee, "How to identify and remove seasonality from time series data with python", Machine Learning Mastery.
8. D. Lai, "Monitoring the SARS epidemic in China: a time series analysis", Journal of Data Science, vol. 3, pp. 279-293, 2005.
9. J. Lin, K. Li, Y. Jiang, X. Guo and T. Zhao, "Prediction and analysis of Coronavirus Disease 2019", China university of Geoscience (Beijing), pp. 315-326, 2019.
10. 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-theArt", SN Computer Science, vol. 1, no. 4, pp. 1-15, 2020