International Journal of Innovative Research in Science and Engineering
Volume 7, Issue 12, 2021
www.ijirse.com
ISSN: 2454-9665

COVID 19 DATA ANALYSIS USING MACHINE LEARNING

DARSHANA A, DHARANI M

CSE, BANNARI AMMAN INSTITUTE OF TECHNOLOGY, (INDIA)

CSE, BANNARI AMMAN INSTITUTE OF TECHNOLOGY, (INDIA)

ABSTRACT

The outbreak of COVID-19 has impacted globally. As virus change its variant it makes more difficult to understand the new variant and find the vaccine. As the number of new variant increases many countries are affected. Therefore ,it is necessary to understand the behavior of the spread of COVID-19 as well as the projection of infections and deaths. This paper builds various analysis and visualization along with predictive models that can predict the cases with higher accuracy. Linear regression, support vector Regression, Holts forecasting have been built on various countries along with India. These models are predicting the number of confirmed, recovered, and death cases based on the data available from April 2021 to December 2021. we also compare different time series methodologies to predict the number of confirmed cases of and deaths of COVID-19 in India. These models are found to be more accurate and effective and will be able to predict the number of cases in future with minimal error. The developed machine learning models can effectively predict the positive and confirmed cases in future with minimal error. The developed machine learning models can effectively predict the positive and confirmed cases.

1. INTRODUCTION

COVID 19 has a huge impact and spread around the world in very short period . As early of April 2021, more than 250 million people had been diagnosed and more than 3 million peoples have died. Unlike other viruses , corona virus has been mutated. The first wave broke out around March 2020 . After a series of measures, the spread had come alleviated to some extent. In April 2021 the spread broke out again .In the second wave of epidemic , the number of confirmed cases increased dramatically in Asian and European countries which is a worrying situation . The number of confirmed and diagnosed people in each countries is shown in Fig 1.1. Now different vaccine has been developed and people started vaccinated , but still there are many problems with spreading . We need to minimize the spread through protocols like isolation , maintaining social distance and wearing a mask . predicting future trend , producing medical supplies helping government are extremely important .

1.1 METHODS

In this experiment various python libraries like Numpy, pandas , seaborn , matplotlib and machine learning models such as Linear regression , support vector regression Holt forecasting Algorithms are used . NumPy is a python package it acts with numeric data that is it performs scientific estimations. Pandas module is also a python library it works with tabular data. It is used to evaluate data. Seaborn is an open source python library. It is applied for data visualization and data analysis. Seaborn runs easily with data frames and pandas library .The graph created can also be modified easily. Matplotlib is also python library it is used for data visualization.

1.2 MODELS

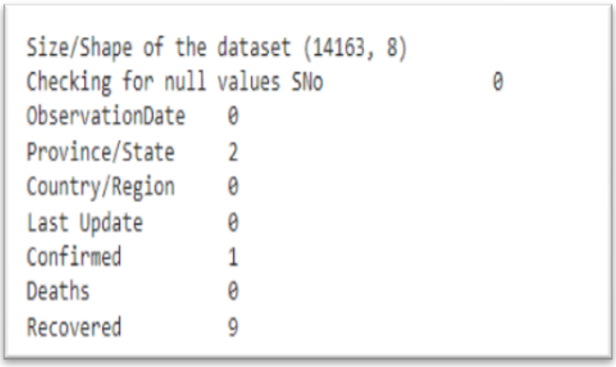
Here linear regression algorithm is used to predict the covid 19 data. It is supervised machine learning .Here it achieves the task to calculate a dependent variable value (Confirmed Cases) based on given   independent variable value (Observation Date) .so this regression method obtains a linear correlation between observation Date(input) and Confirmed cases(output). Support Vector Machine (SVM) used for both Regression and Classification. Here, we have one unrelated variable **confirmed cases** and one related variable **observation Date**. In this problem, we have to train a SVR model with this data to understand the correlation between the confirmed cases ,observation date and be able to predict confirmed cases . Holt method forecasting is a approach to model and predict the performance of a structure of values over period—a time series.

Prerequisites:

- ✓ NumPy
- ✓ Pandas
- ✓ Seaborn
- ✓ Matplotlib
- ✓ Linear Regression
- ✓ Support vector regression
- ✓ Holt forecasting algorithm

1.3 DATASET

1.3.1 Data collection



```
Size/Shape of the dataset (14163, 8)
Checking for null values SNo                0
ObservationDate    0
Province/State     2
Country/Region     0
Last Update        0
Confirmed          1
Deaths             0
Recovered          9
```

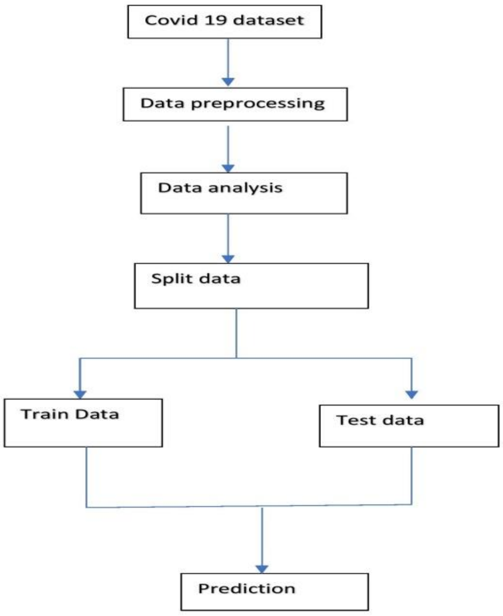
Fig 1

The dataset is the statistical report of COVID-19 cases of 10 countries, which is available in online .These datas are updated daily .The dataset ranges from 01/04/2021 to 05/12/2021.It consists of 14161 rows and 8 columns which contains observation date, state/province, country, Last updated date, confirmed cases, recovered, and deaths. These dataset is first trained and then used for prediction.

1.3.2 Data preparation

After the collection of data , Dataset is checked manually in order to minimize the null values and errors in models. We separate the data into three scenarios like consecutive days , one day gap , and then consecutive days which may be give various results.

2. METHODOLOGY PROPOSED



3. EVALUATION

3.1 Linear Regression:

This statistical model is used to show the relationship between dependent and independent variable. Here confirmed cases is taken as dependent variable which is y and observation date acts as a independent variable x. Hypothesis function for linear regression:

$$y=\mu1+ \mu2.x$$

μ1- Intercept

μ2- co-efficient of x

3.2 Support Vector Regression:

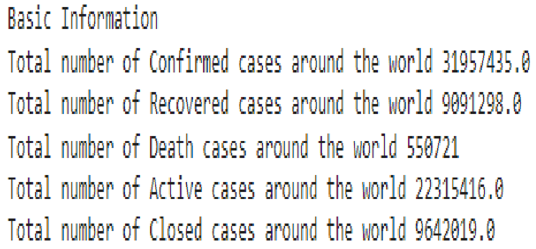
Support Vector Regression is used in support vector machine which minimize the error between real and predicted value.

$$y= wx+b$$

Support vector regression model tries to satisfy the condition $-a < y-wx+b < a$. It used the points with this boundary to predict the value. Here confirmed cases is a independent variable and one dependent variable observation Date. SVR is trained with data to understand the correlation between the confirmed cases and observation date.

4. DATA ANALYSIS AND VISUALIZATION

Analysis and visualization is done with the help of python libraries . Analysing by ploting the graphs make it very clear and comfortable.



```
Basic Information
Total number of Confirmed cases around the world 31957435.0
Total number of Recovered cases around the world 9091298.0
Total number of Death cases around the world 550721
Total number of Active cases around the world 22315416.0
Total number of Closed cases around the world 9642019.0
```

We can clearly see that at the end of November covid 19 reaches 150 million and deaths upto 5 million which is really worrying .

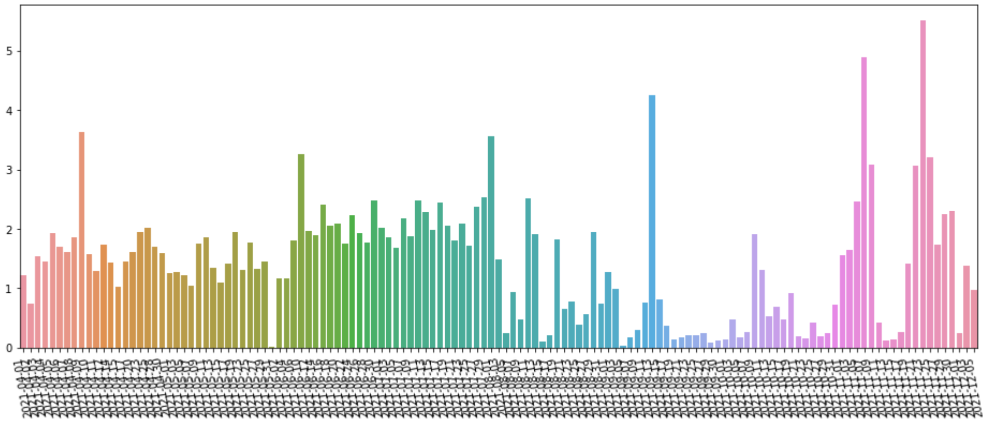


Fig 2 –Distribution of Active cases

Fig 2 represents the distribution of Active cases in April the the spread is very much increased and In India the second outbreak was at the peak level. And at the middle of june it gradually decreases. As the variant changes into new variant omicron the spread was slightly increases in several countries so in the first weak of December there is a slight increase in some increases which shows that safety protocols should be undertaken.

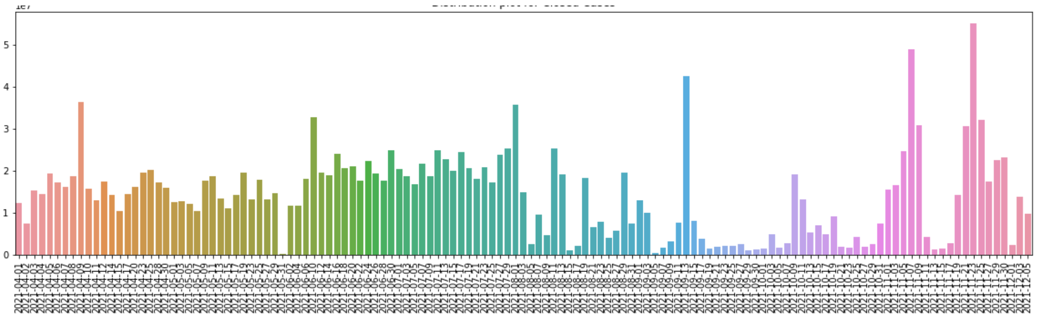


Fig 3 –Distribution of Closed cases

As the active cases increases the medical requirements starts decreasing which lead to high mortality rate . In between may and June the mortality rate increased to peak level . Many countries faces scarcity in medical requirements like oxygen cylinders, poor hospital facilities etc. India face many issues and many countries start helping other in order to fulfill their needs.

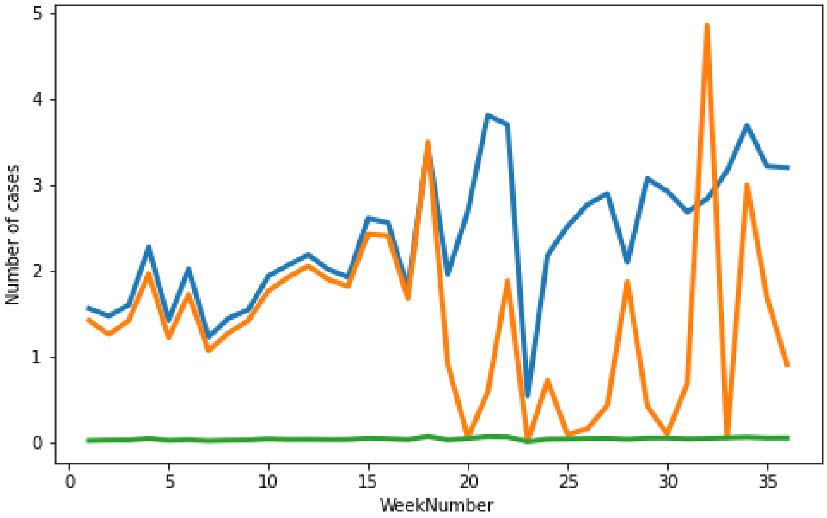


Fig 4-Weak progress of different types of cases.

Weakly progression of cases will increase the level of understanding , Here the graph Fig 4 clearly shows that the Number of confirmed and closed cases are increasing when compared to mortality rate .

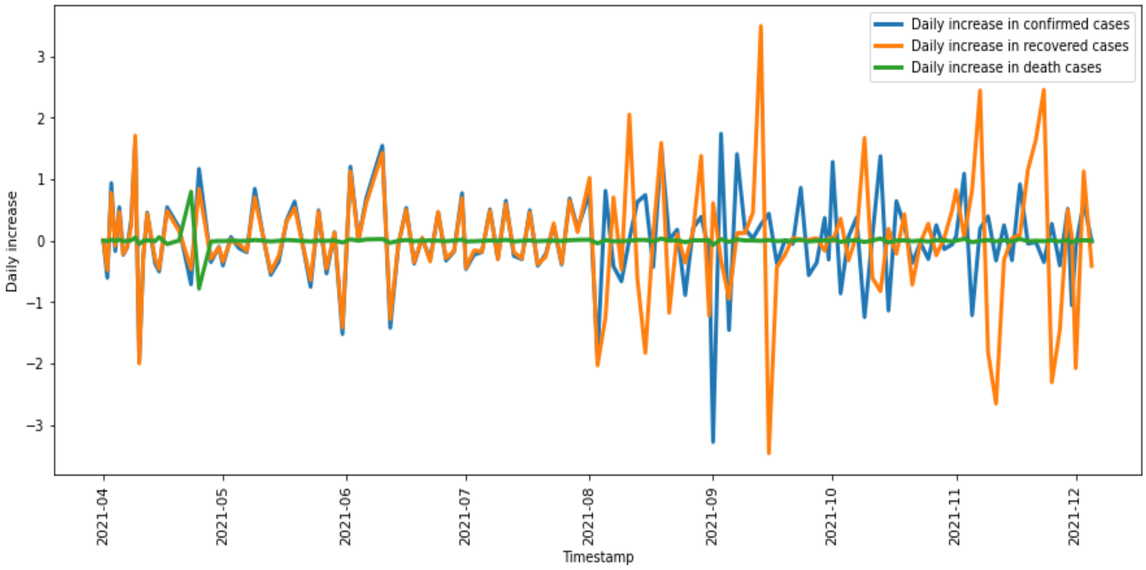
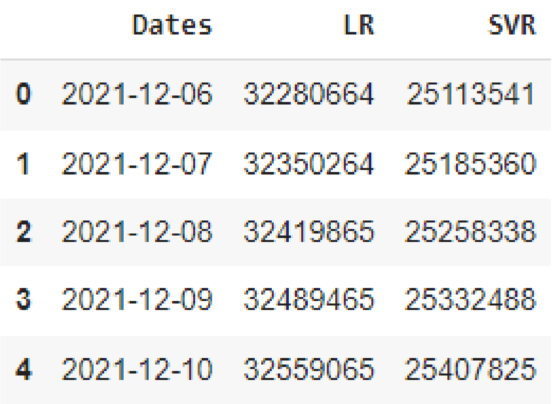


Fig 5 – Daily increase

As the number of recovered cases increases the number of death cases decreases as shown in Fig 5 . The number of confirmed cases start decreasing as the people start vaccinated . Strict lock down protocols and continuous vaccination of people builds end for the second wave of COVID-19.

5. DATA PREDICTION

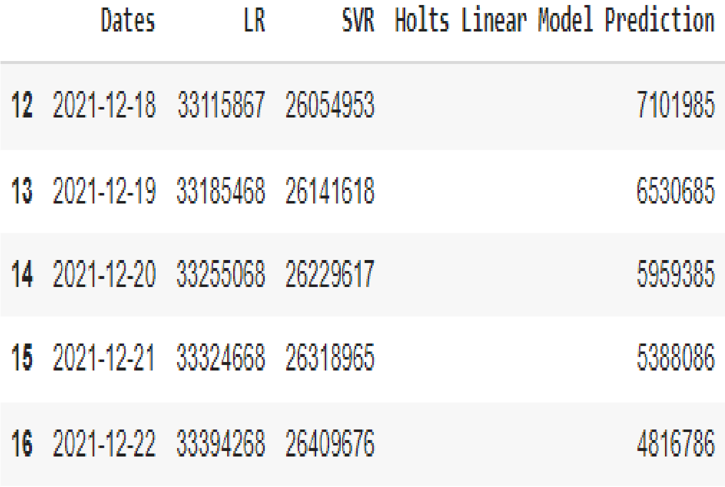
Machine learning models like Linear regression and support vector regression is used for prediction . For more accuracy Holt method forecasting algorithm is used. Holt-winters is one of the most popular forecasting techniques . Holt method is used to predict the behavior of sequence of values over time –a time series. Forecasting always requires a model, and Holt-Winters is a way to model three aspects of the time series: a typical value (average), a slope (trend) over time, and a cyclical repeating pattern (seasonality).



	Dates	LR	SVR
0	2021-12-06	32280664	25113541
1	2021-12-07	32350264	25185360
2	2021-12-08	32419865	25258338
3	2021-12-09	32489465	25332488
4	2021-12-10	32559065	25407825

Fig 6

Fig 6 shows the comparison of predicted values in Linear regression and Support vector regression . Comparing with linear regression SVR gives more accuracy and reduces the error between real and predicted values.



	Dates	LR	SVR	Holts Linear Model Prediction
12	2021-12-18	33115867	26054953	7101985
13	2021-12-19	33185468	26141618	6530685
14	2021-12-20	33255068	26229617	5959385
15	2021-12-21	33324668	26318965	5388086
16	2021-12-22	33394268	26409676	4816786

Fig 7

For more accuracy the predicted values are compared with Holts linear model. These models can be used only for short time predictions and not for long term situations. Predicted values using three models are compared and its clear that Holts model give more accuracy than other two models.

5.1 ANALYSIS OF RESULTS AND DISCUSSION

Virus has the capability to change its variant according to environment and medical factors. So it take several days to understand the nature of new variant . By Analyzing the various dataset can give awareness among the doctors and citizens. These people might think they have been healthy at home because they did not go to hospital for COVID-19 tests. Early release of intervention intensity might increase a risk of the third breakout. Hence the spread is Un accurate and partially predictable.

6. CONCLUSION

Analysis and prediction will help to improve the understanding of this disease and describe the psychological impacts of this pandemic and how these could change as the disease spreads. It gives the clinical review and guidelines for the present pandemic situation .This Data analysis will provide many inputs for future analysis efforts. Prediction of trend data can help the government and citizens to ensure safety by taking precautionary measures in future.

REFERENCES

[1] Hasdeu S, Tortosa F. Riesgo of publication slant on therapeutic interventions for COVID-19. Rev Panama Salud Publica. 2021;45:e157.

[2] Fantin R, Brenes-Camacho G, Barboza-Solís C. Functions by COVID-19: distribution by age and universality of medical coverage in 22 countries. Rev Panama Salud Publica. 2021;45:e42.

[3] Rikitu Terefa D, Shama AT, Feyisa, COVID-19 Vaccine Uptake and Associated Factors Among Health Professionals in Ethiopia,5531-5541.