Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

Regression Analysis of COVID-19 using Machine Learning Algorithms

Ekta Gambhir
B. Tech Student - CSE
Dr. Akhilesh Das Gupta Institute of Technology and Management
New Delhi, India
ektagambhir99@gmail.com

Ritika Jain
B. Tech Student - EEE
Dr. Akhilesh Das Gupta Institute of Technology and Management
New Delhi, India
jainritika181999@gmail.com

Alankrit Gupta
B. Tech Student - CSE
Dr. Akhilesh Das Gupta Institute of Technology and Management
New Delhi, India
alankritgupta091099@gmail.com

Uma Tomer
Assistant Professor - CSE
Dr. Akhilesh Das Gupta Institute of Technology and Management
New Delhi, India
uma.tomer@gmail.com

***Abstract*— The outbreak of the Novel Coronavirus or the COVID-19 in various parts of the world has affected the world as a whole and caused millions of deaths. This remains an ominous warning to public health and will be marked as one of the greatest pandemics in world history. This paper aims to provide a better understanding of how various Machine Learning models can be implemented in real-world situations. Apart from the analysis done on the world figures, this paper also analyzes the current trend or pattern of Covid-19 transmission in India. With the help of datasets from the Ministry of Health and Family Welfare of India, this study puts forward various trends and patterns experienced in different parts of the world. The data to be studied has been obtained for 154 days i.e. from January 22, 2020, till June 24, 2020. For future references, the data can be further analyzed, and more results can be obtained.**

Keywords— ***COVID-19, Machine Learning, Data Analysis, Trend Analysis***

I. INTRODUCTION

According to the World Health Organization (WHO), viral and infectious diseases continue to appear and pose a serious threat to public health and well-being. Coronavirus is a broad family of viruses which causes ailments ranging from common cold and flu to severe respiratory issues. According to NCBI, "In the last 20 years, there have been several viral epidemics that have been reported such as the Severe Acute Respiratory Syndrome Coronavirus or better known as SARS-CoV which was declared a pandemic by WHO in 2002 - 2004 and H1N1 influenza in 2009. With most recently, Middle East Respiratory Syndrome Coronavirus better known as MERS- CoV which hit its first outbreak in Saudi Arabia in 2012" [1].

In the chronology of modern times, cases of unrecognized low respiratory infections were first detected during the mid December 2019 in Wuhan, the largest metropolitan city in Hubei province of China. This strange new pneumonia was named "COVID-19" by WHO. WHO declared this surge a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 as it had affected almost 20 countries of the world [2]. There are no specific treatments of this virus so far, but one can reduce the spread of infection by maintaining personal hygiene and social distancing. There have been recoveries around the world, but the pandemic is still not under control.

Since this pandemic has affected the whole world not only in terms of health and hygiene but also in terms of the global economy. Apart from the adverse effects of COVID-19, there have been certain constructive influences around the world. As the world was facing loses, our nature gained something from this pandemic, the harmful particulate matter was eliminated from the environment and most importantly the largest ever ozone hole detected was closed during this pandemic. So it becomes really important to understand the features and characteristics of this disease and predict/estimate the further spread of this disease around the world and how it is going to impact the coming generations and the lives of the people when things become normal.

The timeline of the events of COVID-19 across different nations [2] is shown in Fig 1. and the percentage of confirmed cases per country is shown by Fig 2.

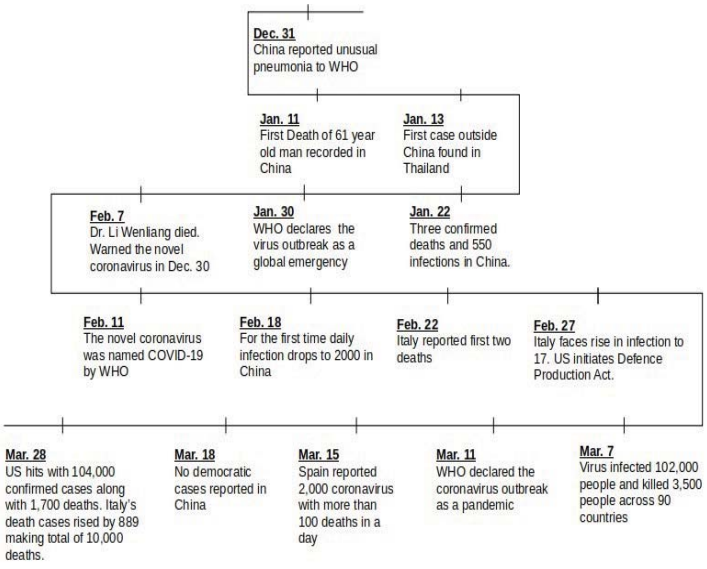

Fig. 1. Timeline of the events of COVID-19 across different nations

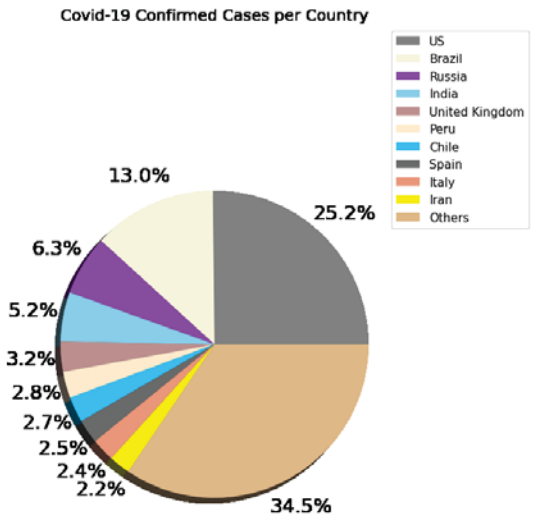

Fig. 2. Percentage of confirmed cases per country.

A. Data Acquisition and Description

The dataset available from the data repository for the "2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), also supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL)" is parameterized dataset having relevant parameters such as Province/State, Country/Region, Latitude, Longitude and dates. Separate datasets have been used for Confirmed, Death, and Recovered cases along with the number of cases on each day. The total dataset used in the study is obtained for 154 days i.e. from January 22, 2020, till June 24, 2020. The data from these datasets were merged to obtain the parameterized dataset of the world from January 22, 2020, till June 24, 2020.

The reason behind choosing regression analysis for the current problem statement is the type of dataset. The dataset is a continuous dataset and regression analysis is best suited when a continuous dependent variable is wanted to predict from several independent variables. The coefficients of the dependent and independent variables in the regression equation (1) determine the relationship between the dependent and independent variables.

$$Y = \theta o + \theta_1 X + \theta_2 X^2 + \ldots + \theta_m X^m + residual\ error \quad (1)$$

Here:

θo is the bias

θ_1 , θ_2 , ... , θ_m are the weights in the equation of polynomial regression and m is the degree of a polynomial

B. Feature Selection

This step involves feature extraction and selection to obtain the best results from our model. Having good and best features allows us to illustrate the underlying structure of the data. Feature Engineering affects the performance of the model significantly. It may involve splitting some feature or aggregating some features to produce new features or collecting data from external sources. Dimensionality reduction helps in evaluating and drawing conclusions easily from the dataset. To draw better conclusions from this dataset, the irrelevant parameters such as the Longitude and Latitude were removed, and the dates were converted to a date-time object.

II. Model Implementation and Analysis

A. Implementation of Support Vector Machine Algorithm in Python

Support Vector Machine Algorithms are supervised machine learning models that are associated with data classification and regression analysis. It constructs a hyperplane or set of hyperplanes in an N-dimensional space for classification or outlier detection. The best params obtained for SVM in Python is shown by Fig 3 and Fig 4. shows the predicted values using the SVM Algorithm. Fig 5. Shows the graphical representation of predictions using the Support Vector Machine Algorithm.

```
svm_search
```

```
RandomizedSearchCV(cv=3, error_score=nan,
                   estimator=SVR(C=1.0, cache_size=200, coef0=0.0, degree=3,
                                 epsilon=0.1, gamma='scale', kernel='rbf',
                                 max_iter=-1, shrinking=True, tol=0.001,
                                 verbose=False),
                   iid='deprecated', n_iter=40, n_jobs=-1,
                   param_distributions={'C': [0.01, 0.1, 1, 10],
                                        'epsilon': [0.01, 0.1, 1],
                                        'gamma': [0.01, 0.1, 1],
                                        'kernel': ['poly', 'sigmoid', 'rbf'],
                                        'shrinking': [True, False]},
                   pre_dispatch='2*n_jobs', random_state=None, refit=True,
                   return_train_score=True, scoring='neg_mean_squared_error',
                   verbose=1)
```

```
svm_search.best_params_
```

```
{'shrinking': True, 'kernel': 'poly', 'gamma': 1, 'epsilon': 1, 'C': 0.1}
```

```
svm_confirmed =svm_search.best_estimator_
svm_pred = svm_confirmed.predict(future_forecast)
```

```
svm_confirmed
```

```
SVR(C=0.1, cache_size=200, coef0=0.0, degree=3, epsilon=1, gamma=1,
    kernel='poly', max_iter=-1, shrinking=True, tol=0.001, verbose=False)
```

Fig. 3. Shows the best params obtained for SVM in Python

```
svm_test_pred
```

```
[ 5740749.  5885419.  6032500.  6182012.  6333975.  6488409.  6645334.
  6804770.  6966736.  7131253.  7298340.  7468018.  7640306.  7815225.
  7992793.  8173032.  8355961.  8541600.  8729969.  8921088.  9114976.
  9311655.  9511143.  9713460.  9918627. 10126663. 10337589. 10551424.
 10768188. 10987902. 11210584. 11436256. 11664936. 11896645. 12131403.
 12369229. 12610145. 12854168. 13101321. 13351621.]
```

Fig. 4. Shows the predicted values using the SVM Algorithm.

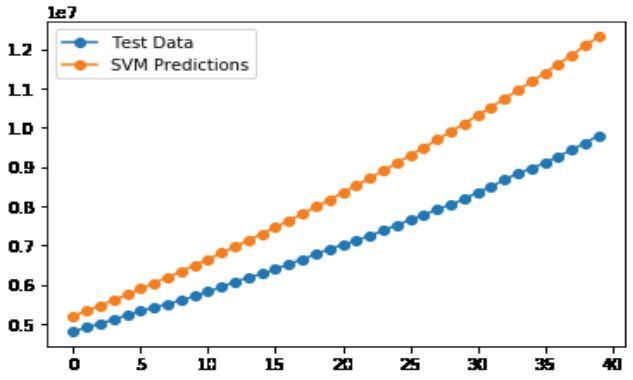

Fig. 5. Predictions using the Support Vector Machine Algorithm.

B. *Implementation of the Polynomial Regression Algorithm in Python*

Polynomial Regression can be expressed as a special case of Linear Regression. In Linear Regression it works on continuous data is known and the two variables (target variable and independent variable) are correlated. What if have known that variables are correlated but the relationship does not look linear, so polynomial regression to fit a polynomial equation can be used to our dataset. Polynomial Regression is a supervised Machine learning Algorithm that is trained based on prior data and then tested on another dataset to validate its accuracy.

The train and test data have been transformed for polynomial regression in Python which is shown by Fig 6. Fig 7. shows the predicted values from August 7, 2020, to August 28, 2020, which are graphically represented as shown in Fig 8. Thus, the polynomial Regression Algorithm shows an accuracy of 93%.

```
# transform our data for polynomial regression
poly = PolynomialFeatures(degree=3)
poly_X_train_confirmed = poly.fit_transform(X_train_confirmed)
poly_X_test_confirmed = poly.fit_transform(X_test_confirmed)
poly_future_forecast = poly.fit_transform(future_forecast)
```

Fig. 6. Shows the transformation of train and test data for polynomial regression in Python

38	08/07/2020	17926785.0
39	08/08/2020	18136758.0
40	08/09/2020	18347945.0
41	08/10/2020	18560347.0
42	08/11/2020	18773964.0
43	08/12/2020	18988795.0
44	08/13/2020	19204840.0
45	08/14/2020	19422100.0
46	08/15/2020	19640575.0
47	08/16/2020	19860264.0
48	08/17/2020	20081167.0
49	08/18/2020	20303286.0
50	08/19/2020	20526618.0
51	08/20/2020	20751165.0
52	08/21/2020	20976927.0
53	08/22/2020	21203903.0
54	08/23/2020	21432094.0
55	08/24/2020	21661500.0
56	08/25/2020	21892120.0
57	08/26/2020	22123954.0
58	08/27/2020	22357003.0
59	08/28/2020	22591266.0

Fig. 7. Shows the predicted values from August 7, 2020 to August 28, 2020

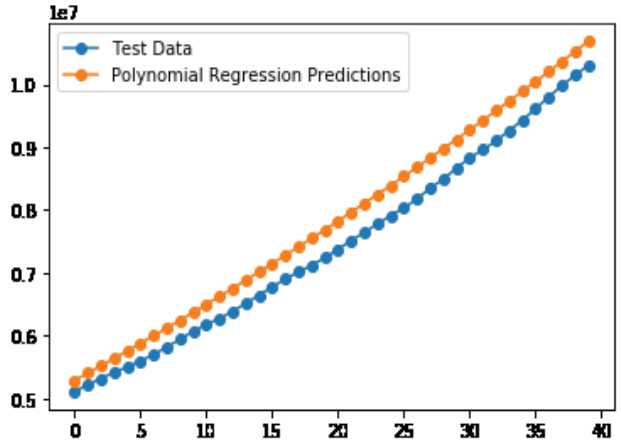

Fig. 8. Shows the graphical representation of predictions using Polynomial Regression

III. DAILY ESCALATION IN WORLD

To see the daily changes in the cases across World, the trend was plotted for Confirmed, Death and Recovered cases.

Fig 9. shows the daily increase in the number of Confirmed cases while Fig 10. shows the daily increase in the number of Confirmed Deaths and Fig 11. shows the daily increase in the number of Confirmed Recoveries across the world.

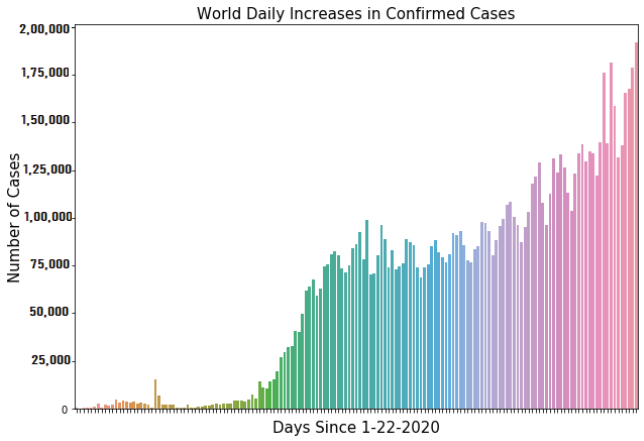

Fig. 9. Shows the daily increase in the number of Confirmed cases

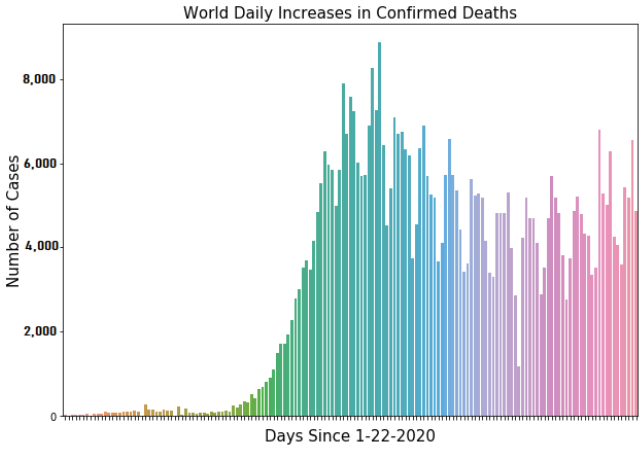

Fig. 10. Shows the daily increase in the number of Confirmed Deaths.

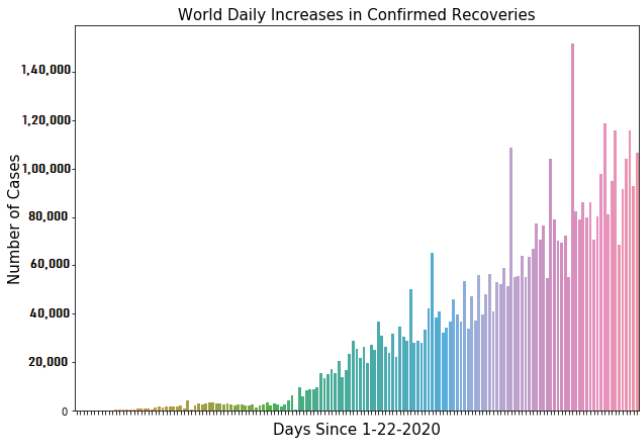

Fig. 11. Shows the daily increase in the number of Confirmed Recoveries.

IV. CASE STUDY: CURRENT TREND IN INDIA

A. Data Acquisition, Description and Feature Selection

This segment describes the collection, processing and assessing the data/statistics of India for evaluating the current trend of COVID-19 infections in India. Real-time data has been collected from the website of the Ministry of Health and
Family Welfare of India, https://www.mohfw.gov.in/. The information has been scraped from the website using the BeautifulSoup module in Python. The link to the scrapped data and analysis of the current trend in India: https://github.com/Ritikajain18/COVID-19-INDIA/blob/master/covid_scrapped_data.ipynb. The first case of COVID-19 in India was detected on January 30, 2020, in Kerala when a student from Wuhan came back to India. The data from January 30, 2020, to June 24, 2020, has been used to evaluate the statistics of India on the ongoing pandemic. The data consists of all the States and Union Territories (State/UTs) of India with the number of Active, Recovered, Deaths and a Total number of Confirmed cases.

To tally and easily work upon the acquired data, the data is sorted as shown in Fig 12. so that the top 10 States/UTs with maximum cases are obtained.

	Name of State/UT	Active cases	Recovered cases	Deaths	Total Confirmed Cases
1	Maharashtra	61807	67706	6283	135796
2	Delhi	23820	36602	2233	62655
3	Tamil Nadu	27181	34112	794	62087
4	Gujarat	6232	19909	1684	27825
5	Uttar Pradesh	6152	11601	569	18322
6	Rajasthan	2966	11910	356	15232
7	West Bengal	5102	8687	569	14358
8	Madhya Pradesh	2342	9215	521	12078
9	Haryana	4940	5916	169	11025
10	Karnataka	3527	5730	142	9399

Fig. 12. Shows the sorted data of top 10 States/UTs with maximum cases

B. Reason for studying State/UT-wise Data

India is the second-most populous country with a total area of 3.287 million km² so analyzing the spread of infection state-wise will help the government to utilize the available resources efficiently [3]. As far as health care facilities are considered, India doesn't have the best facilities in the world and taking into account the population of the country it becomes essential to utilize the resources efficiently and effectively in the best possible way. Despite not having the best health care facilities like most of the developed nations like the USA, Italy, Germany and many other countries, India has been administering this pandemic in the best way possible.

The further investigation involved the daily increase in the number of confirmed, death and recovered cases. The graphs have been plotted for the same. Fig 13. and Fig 14. shows the daily increase in the number of Confirmed and Recovered cases in India, respectively.

Taking into account the different states of India, charts were plotted for these States/UTs in terms of Total Confirmed Cases, Active Cases and Total Closed cases. Fig 15. Shows the plot for the number of Total Confirmed Cases. This graph clearly shows that Maharashtra has the maximum number of Cases amongst the top 10 States followed by Delhi and Tamil Nadu. Fig 16. Shows the plot for the number of Active Cases and Fig 17. Shows the plot for the number of Closed Cases.

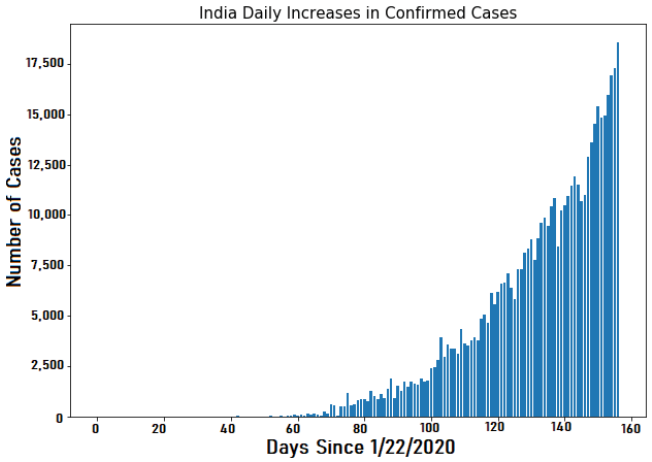

Fig. 13. Daily increase in confirmed cases

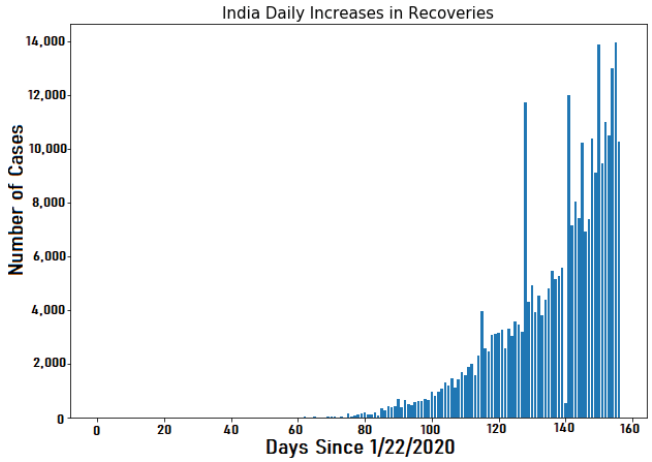

Fig. 14. Daily increase in recovered cases

One of the most astounding aspects of COVID-19 in the world is that despite being the second-most populous country in the world, the mortality rate of India is the lowest with more than 4 lakh cases. Fig 18. shows the plot for mortality rate. The graph distinctly shows that Gujarat has the highest mortality rate followed by Maharashtra.

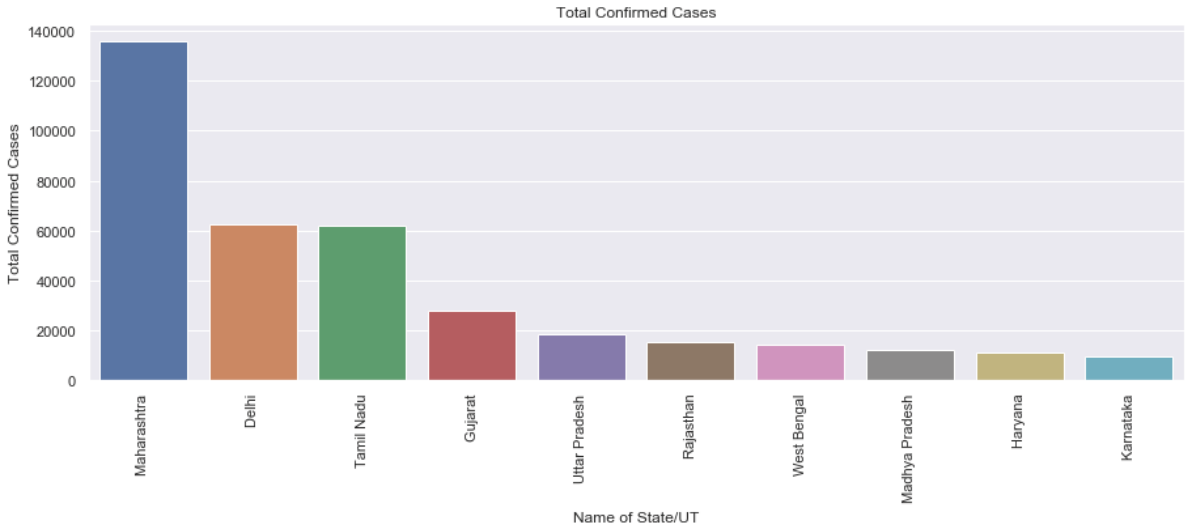

Fig. 15. Plot for the number of Total Confirmed Cases.

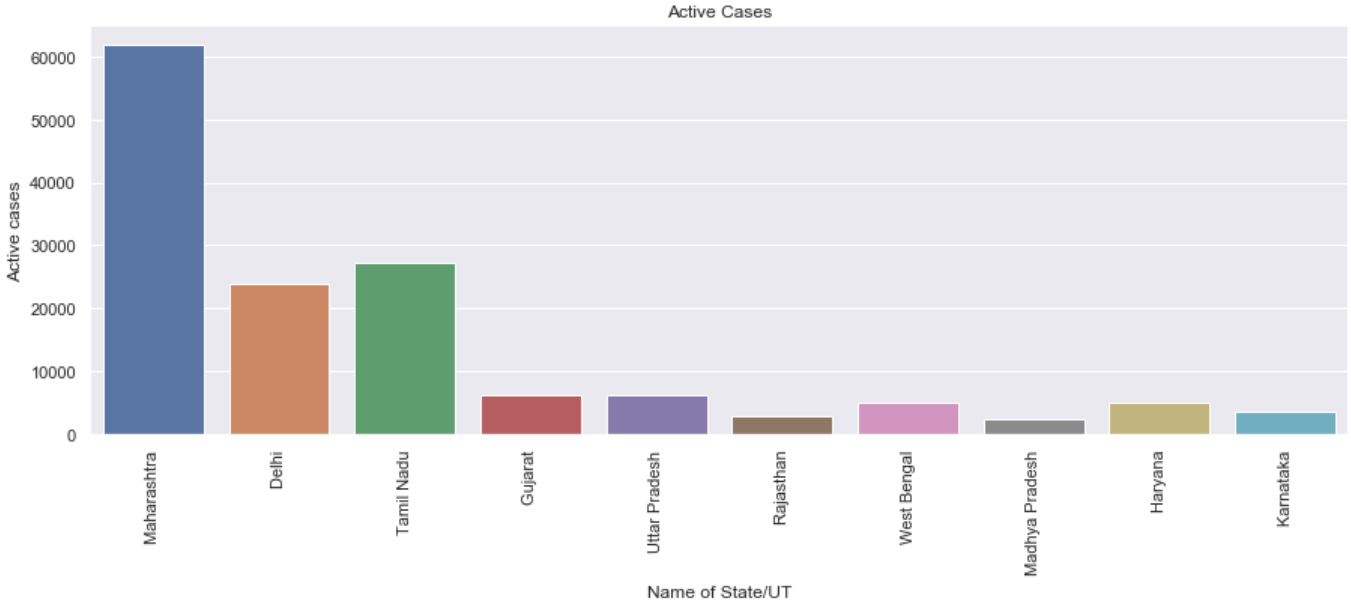

Fig. 16. Plot for the number of Active Cases.

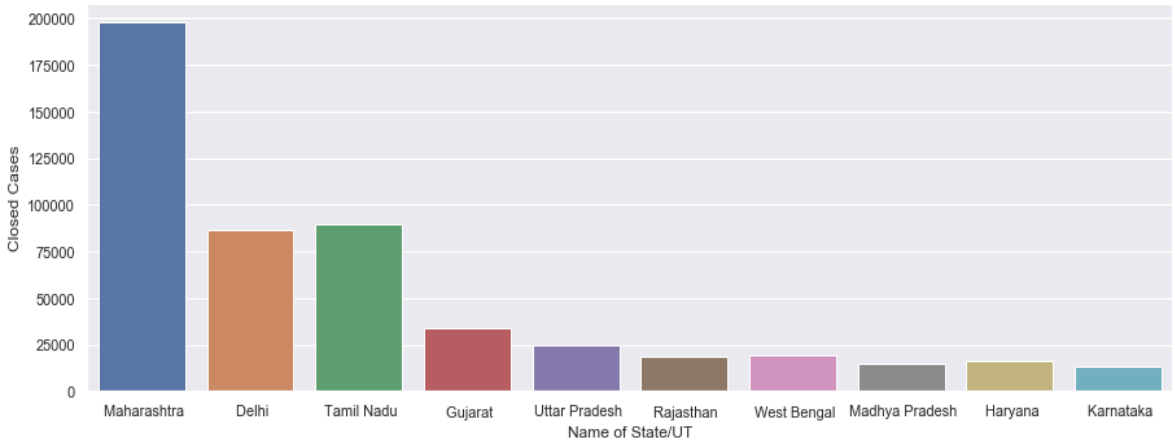

Fig. 17. Plot for the number of Closed Cases

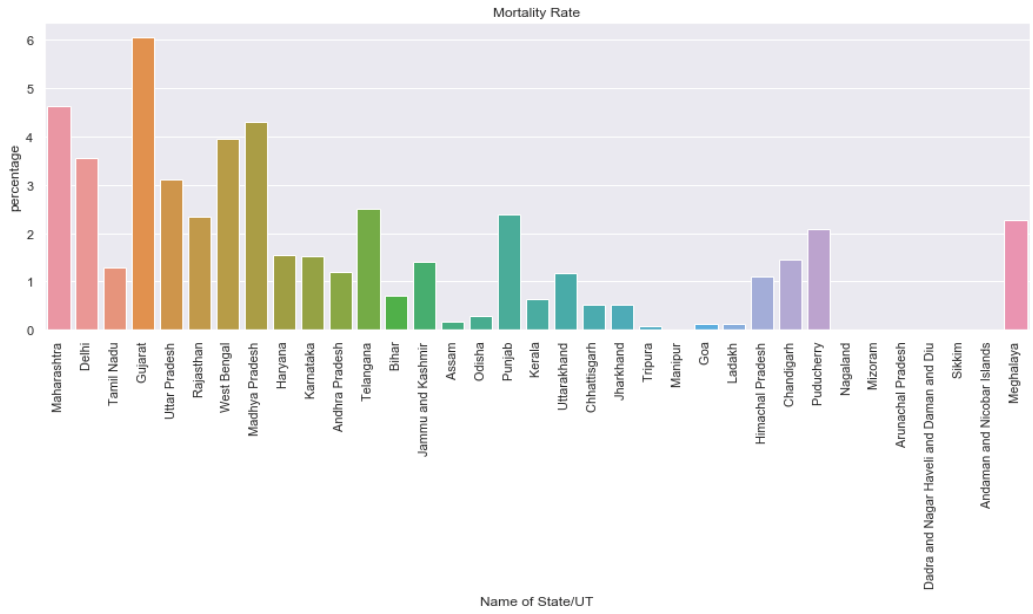

Fig. 18. Shows the plot for mortality rate.

V. Conclusion and Future Scope

This research paper successfully analyzed the current trend of the transmission of Covid-19 in the world. Anticipating the further spread of the Covid-19 or better known as Novel Coronavirus will help in taking necessary actions to control the spread. The paper also presented a comprehensive study of the spread of the virus outbreak situation in India which will further help in taking necessary steps to manage the huge population of India. For the same, two Machine Learning models were used but in future Deep Learning models or hybrid two or models can be used to forecast the further spread of the virus. Real-time data has been used for this analysis and not the conventional data because the data validity of real time data highly depends on the time and its response time requirements come from the external world. Although the real time data can be relaxed in a few cases unlike the conventional data which needs to satisfy every case, real time data is approximately correct with performance metrics calculated as a number of transactions missing their deadlines per unit time. Through our analysis can be concluded that the Polynomial Regression Algorithm as compared to the Support Vector Machine Algorithm, shows an accuracy of approximately 93% by predicting the rise in cases for the next 60 days i.e. for the months of July and August. The case study in this paper puts forward and answers the question that "Why has the data analysis for India been done State/UTs wise?". Also, this analysis raises the question "Why is the mortality rate of India the lowest in the world despite being the second most

populous country in the world?". Further, more attributes can be included in the study in order to add more accuracy during the process. Further investigation may involve the analysis on India dataset by predicting the number of cases in future and how the mortality rate varies with the rise in the number of cases. Hope this article contributes to the world's response to this epidemic and puts forward some references for further research in future.

References

[1] Features, Evaluation and Treatment Coronavirus (COVID-19) . Available: https://www.ncbi.nlm.nih.gov/books/NBK554776/, May 18 **(2020)**.

[2] N. S Punn, S. K. Sonbhadra, S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms", medRxiv, Available: doi**:** https://doi.org/10.1101/2020.04.08.20057679, June 1 **(2020)**.

[3] P. Ghosh, R. Ghosh, B. Chakraborty, "COVID-19 in India: State-wise Analysis and Prediction", medRxiv, Available: doi: https://doi.org/10.1101/2020.04.24.20077792, May 19 **(2020).**

[4] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U.Reuter, T. Rabczuk, P. M. Atkinson, "COVID-19 Outbreak Prediction with Machine Learning", Available at SSRN: https://ssrn.com/abstract=3580188 or http://dx.doi.org/10.2139/ssrn.3580188, April 19 **(2020).**

[5] F. Petropoulos, S. Makridakis, "Forecasting the novel coronavirus COVID-19", Available: https://doi.org/10.1371/journal.pone.0231236, Published: March 31 **(2020).**

[6] Arti. M. K, K. Bhatnagar, "Modeling and Predictions for COVID 19 Spread in India", Available: DOI: 10.13140/RG.2.2.11427.81444, Published: April **(2020).**

[7] H. Shekhar, "Prediction of Spreads of COVID-19 in India from Current Trend", medRxiv, Available: doi: https://doi.org/10.1101/2020.05.01.20087460, May 06, **(2020).**

[8] Yan, L., Zhang, H., Goncalves, J. *et al.* "An interpretable mortality prediction model for COVID-19 patients". Nature Machine Intelligence 2, Available: https://doi.org/10.1038/s42256-020-0180-7, pp. 283–288, Published: 14 May **(2020).**

[9] L. Li, Z. Yang, Z. Dang, C. Meng, J. Huang, H. Meng, D. Wang, G. Chen, J. Zhang, H. Peng, Y. Shao, "Propagation analysis and prediction of the COVID-19", Infectious Disease Modelling vol. 5, Available: https://doi.org/10.1016/j.idm.2020.03.002, pp. 282-292, (**2020**).

[10] S. Zhao, H. Chen, "Modeling the epidemic dynamics and control of COVID-19 outbreak in China". Quantitative Biology, vol. 8, Issue 1, Available: https://doi.org/10.1007/s40484-020-0199-0, pp. 11–19 March (**2020**).

[11] J. Xie, Z. Tong, X. Guan. et al. "Critical care crisis and some recommendations during the COVID-19 epidemic in China". Intensive Care Med, vol.6 Issue 6, Available: https://doi.org/10.1007/s00134-020-05979-7, pp. 837–840, June (**2020**).

[12] W.C.Roda, M. B.Varughese, D. Han, M. Y. Lia, "Why is it difficult to accurately predict the COVID-19 epidemic?", Infectious Disease Modelling, vol. 5, Available: https://doi.org/10.1016/j.idm.2020.03.001, pp. 271-281, **(2020).**

[13] W. Naudé, "Artificial intelligence vs COVID-19: limitations, constraints and pitfalls", Available: doi: 10.1007/s00146-020-00978-0, pp. 1–5, Apr 28 **(2020)**.

[14] R. Gupta, S.K. Pal, G. Pandey, "A Comprehensive Analysis of COVID-19 Outbreak Situation in India", Available: DOI: 10.1101/2020.04.08.20058347, Published: April **(2020).**

[15] Bindhu, V. "Biomedical Image Analysis Using Semantic Segmentation." Journal of Innovative Image Processing (JIIP) 1, no. 02 (2019): 91-101.

[16] Chandy, A. (2019),"A Review On Iot Based Medical Imaging Technology For Healthcare Applications", Journal of Innovative Image Processing (JIIP), 1(01), 51-60.