International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF COVID-19 USING REGRESSION ANALYSIS

Bathula Preetham Kumar Reddy[1]

[1]*U.G Student, Department of ECE, Sreenidhi Institute of Science and Technology, Hyderabad*

---***---

Abstract – *During the epidemic outbreak, panic situations all over the world will arise. It causes a great life loss and affects the economies of all the countries. It can last from a few days to years. There is a lot of research going on to eradicate the outbreak. It is very evident that the spread of this can reach peaks in upcoming days. With the help of prediction, the governments and people can know the impact in coming dates and they can take care of themselves. Using mathematical models and machine learning algorithms such as Linear Regression, SVM, LSTM, and Naïve Bayes can help us to predict future circumstances. The linear regression model is advantageous to predict the number of affected cases of COVID-19. We create an end to end model for predicting the cases of COVID-19 based on the dataset from reputed organizations. This project is aimed to make a user-friendly application for people to know the impact in the coming days and take care of themselves.*

Key Words: COVID-19, Machine learning, Linear Regression, Polynomial Regression, Feature Engineering, API modeling, Model deployment.

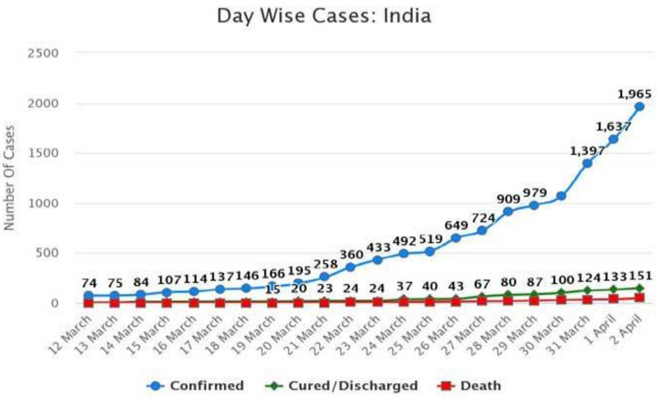
1. INTRODUCTION

Artificial Intelligence is very much coming to existence and being used in research areas. It can also be used to analyze hard time situations. COVID-19 outbreak is very much unexpected. In the initial phase of the virus outbreak, there was not much detailed information about the virus available. As the days passed, the facts of the virus and the nature of the virus are known. The initial symptoms are fever, cough, and shortness of breath. Apart from this, it may include pneumonia and acute respiratory distress syndromes. Now, in India, a lot of positive cases are found which do not have any symptoms. As the population of India is very high, there is a chance of rapid multiplication of cases. Even though the government is implementing lockdowns and imposing restrictions, the spread rate is increasing exponentially. Even doctors and frontline workers are getting affected while serving people.

In this paper, we analyze the outbreak of COVID-19 from the initial phase and build a predictive model that uses Regression techniques. Finally, we deploy the model to make it accessible to the general audience for predicting the affected cases in the upcoming days.

2. EXISTING METHODS

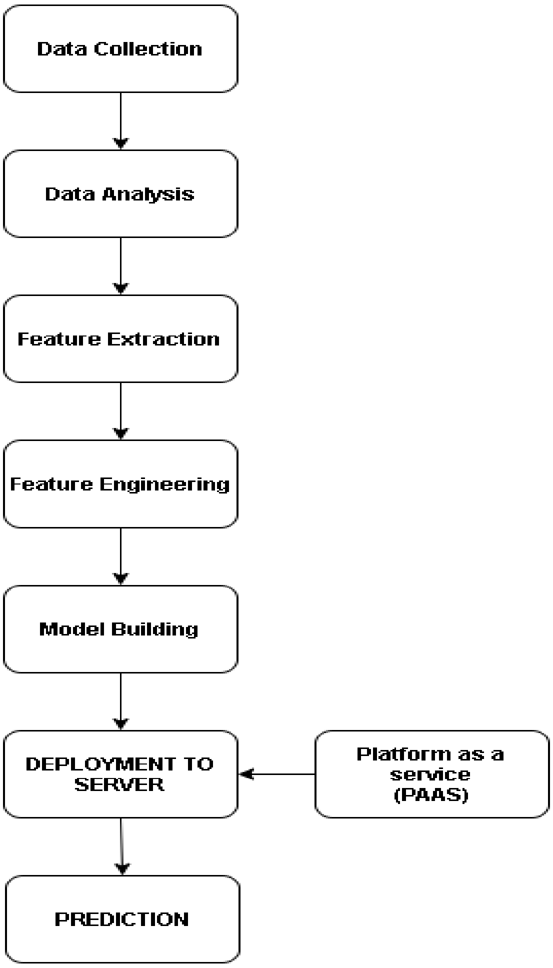
At present, people are depending on news channels, surveys, and social media platforms to know the number of affected cases all over the world. Some of the mathematical models and statistical tools are used to estimate the cases but they cannot be used by general people. These methods mostly show the variation in the past analysis.



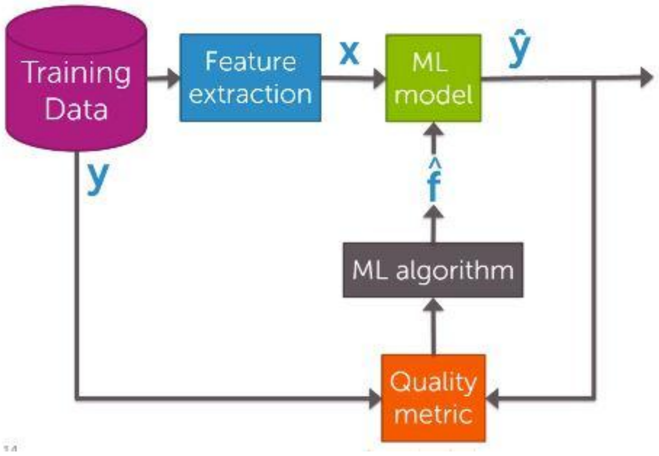
To estimate the affected cases based on the previous trend, we design a model and make it accessible to all by developing an API through our new system.

3. PROPOSED SYSTEM

Here, we build a prediction system based on the previous data. Firstly we collect the datasets either in form of .json or .csv format. Now we extract the features required for the prediction model. After extraction of features, make the featured data to the usable form. Apply various regression techniques and measure the accuracy then decide the best model and fit the data.

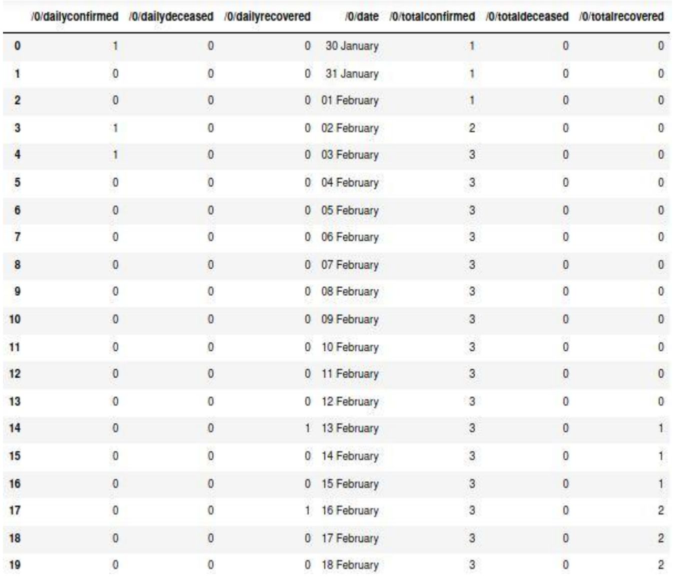


After fitting the model, deploy the model to cloud service by dumping it with the help of libraries. After deployment, it is ready to use.



3.1 DATA SOURCE

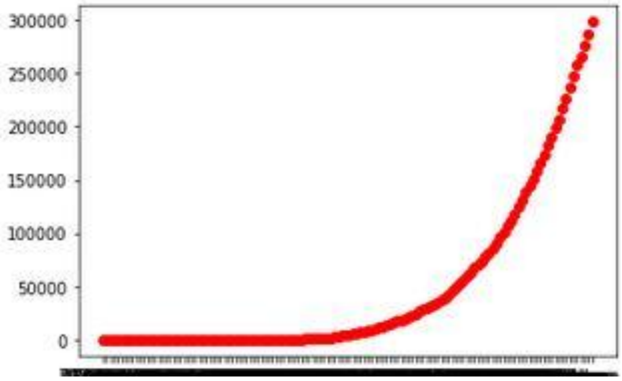
The dataset which is used in this project is sourced from a volunteer-driven, crowd-sourced database for COVID-19 stats & patient tracing in India. The dataset in the database is in JSON format. It consists of the number of samples collected each day and the total number of positive cases. Dataset is also available in Kaggle.



	/0/dailyconfirmed	/0/dailydeceased	/0/dailyrecovered	/0/date	/0/totalconfirmed	/0/totaldeceased	/0/totalrecovered
0	1	0	0	30 January	1	0	0
1	0	0	0	31 January	1	0	0
2	0	0	0	01 February	1	0	0
3	1	0	0	02 February	2	0	0
4	1	0	0	03 February	3	0	0
5	0	0	0	04 February	3	0	0
6	0	0	0	05 February	3	0	0
7	0	0	0	06 February	3	0	0
8	0	0	0	07 February	3	0	0
9	0	0	0	08 February	3	0	0
10	0	0	0	09 February	3	0	0
11	0	0	0	10 February	3	0	0
12	0	0	0	11 February	3	0	0
13	0	0	0	12 February	3	0	0
14	0	0	1	13 February	3	0	1
15	0	0	0	14 February	3	0	1
16	0	0	0	15 February	3	0	1
17	0	0	1	16 February	3	0	2
18	0	0	0	17 February	3	0	2
19	0	0	0	18 February	3	0	2

3.2 DATA ANALYSIS

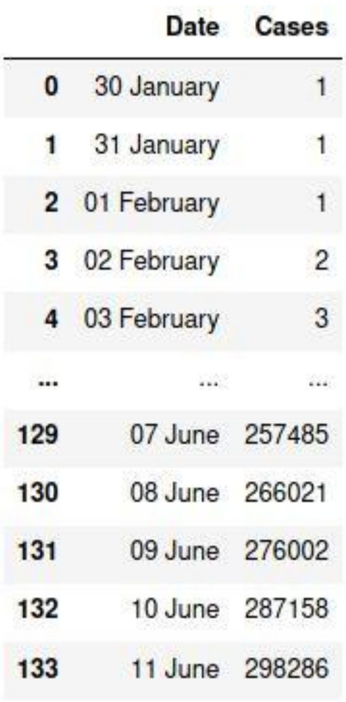
Using data visualization libraries of python such as seaborn, analyze the data and decide the most influential factors based on the statistical parameters such as Correlation and variance. We can observe that more dense data is present in an initial state, it may lead to biasing. So we consider removing that dense data.



3.3 FEATURE EXTRACTION

For building a predictive model using any technique, we need to know the dependent and independent features. Independent features will influence the result. The result is known as a dependent feature. Select the features based on data analysis.

As we need to predict the number of cases based on the date, we need to select the date and Confirmed cases as independent and dependent parameters respectively.

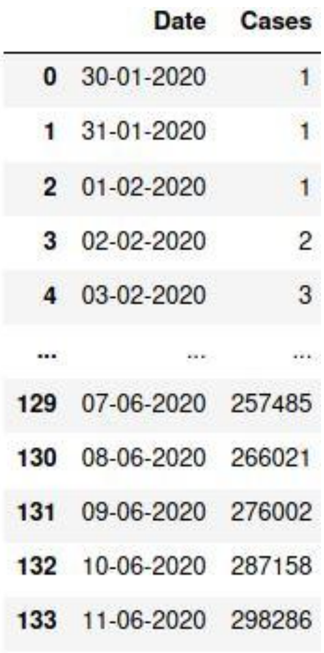


	Date	Cases
0	30 January	1
1	31 January	1
2	01 February	1
3	02 February	2
4	03 February	3
...	...	...
129	07 June	257485
130	08 June	266021
131	09 June	276002
132	10 June	287158
133	11 June	298286

3.4 FEATURE ENGINEERING

The selected parameters should be brought into a proper readable format. The Date feature is in string format, so it should be converted to either YYYY-MM-DD or DD-MM-YYYY.

We should convert the date to Gregorian ordinal number for mathematical alignment.



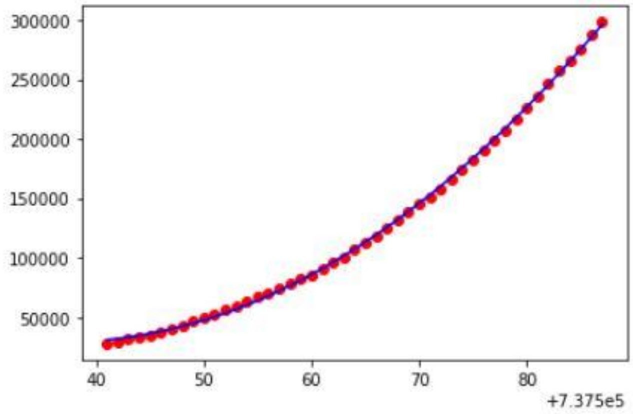
	Date	Cases
0	30-01-2020	1
1	31-01-2020	1
2	01-02-2020	1
3	02-02-2020	2
4	03-02-2020	3
...	...	...
129	07-06-2020	257485
130	08-06-2020	266021
131	09-06-2020	276002
132	10-06-2020	287158
133	11-06-2020	298286

3.5 LINEAR REGRESSION (MODEL BUILDING)

Linear Regression is used to predict the amount or count of the cases by using the data (features). The normal linear regression cannot be used for high variance data, so we use polynomial regression of degree 4.

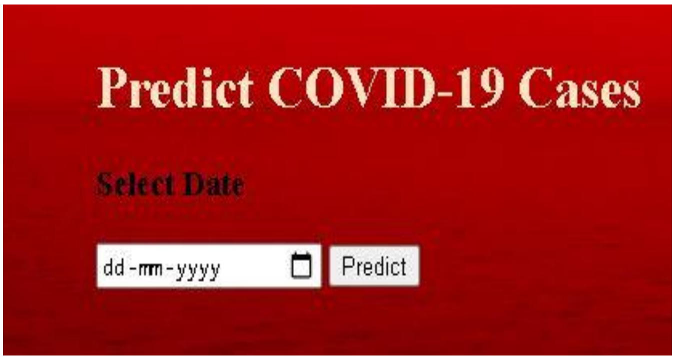
$$Y=\theta_0+\theta_1X+\theta_2X^2+\ldots+\theta_mX^m+\text{residual error}$$

The validation set performance on train dataset looks as follows:



3.6 MODEL DEPLOYMENT

The regression model is dumped using python libraries and with the help of Web framework, we deploy our model API to Platform as a service (PAAS).



4. CONCLUSION

In this paper, we have implemented an end to end application for predicting the number of affected COVID-19 cases using linear regression. Despite who the person is, with the help of this application anyone can predict the number of cases.

For future enhancements, based on the switching of lockdown and imposing restrictions, we can study the fluctuation in the number of death rates and affected cases.

REFERENCES

[1] S. K. Bandyopadhyay and S. Dutta, "Machine learning approach for confirmation of COVID-19 cases: Positive, negative, death and release, "medRxiv, 2020

[2] Yan, X., & Su, X. (2009). “*Linear regression analysis: theory and computing*.” World Scientific.

[3] Andreas Buja, Dianne Cook, and Deborah F. Swayne(1996). “Journal of Computational and Graphical Statistics.”

[4] Chand, Nimai & Das Adhikari, Nimai & Alka, Arpana & Kurva, Vamshi Kumar&Nayak, Hitesh&Kushwaha, Jitendra & Nayak, Ashish & Nayak, Sankalp & Shaj, Vaisakh & Rishav, Kumar. (2018). “Epidemic Outbreak Prediction Using Artificial Intelligence.” International Journal of Computer Science and Information Technology.10. 10.5121/ijcsit.2018.10405.