ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

AN EMPIRICAL ANALYSIS OF PREDICTION OF COVID 19 USING APPLICATIONS OF DATA SCIENCE

Dr. Rajeshri Pravin Shinkar
Assistant Professor, SIES (Nerul) College of Arts, Science & Commerce

ABSTRACT
There are several COVID-19 outbreak models for predictions which are being officially used all over the world to help the public for informed decisions and take the right safety measures in advanced. This paper describes the analysis of Prediction of Covid 19, with the help of various applications of data science like regression analysis and the time series Forecasting. In Regression there are two different models 1. Linear and 2. Polynomial. These two-regression model evaluation is done by R squared score function and error values function. This paper deals with the data set "Covid 19 dataset for India". Regression model will result in confirmed, recovered and death cases. Forecasting will be for future trend of these cases, for this model author used forecasting model of tableau.

Keywords: *Covid 19, Data Science, Regression analysis, forecasting, tableau*

INTRODUCTION
Covid 19 the word when we used the all public get scared till the date, it is really a disastrous year (2020) for the human being. With the help of Data Science and its various tools and methods are used for centralized the activities of Covid 19 research. During pandemic World Health Organization (WHO) continuously provides the time line for Covid 19 response activities for various information for public safety.

This paper is useful for future aspect of any kind of viruses if affected, for this, the author anlayzed India's Covid 19 data set for the regression analysis with error analysis and the accuracy of the data. Paper also deals with forecasting the pattern of the corona virus cases by using the data science technique time series forecasting with the approach of Tableau. Predictions and Forecasting are useful for various types of corona virus infected cases like confirmed, cured, active and death cases with the available data.

LITERATURE REVIEW
[1]. Garima Jain, Bhawana Mallick (2016)

Review on Weather Forecasting Techniques

(International Journal of Advanced Research in Computer and Communication Engineering)

Vol 5, issue 12, December 16, ISO 3297:2007 certified ISSN(online) 2278-1021 ISSN(print)- 2319-5940 PP 177-182

[2]. Janani. B, Priyanka Sebastian (2014)

Analysis on the weather forecasting and techniques

(Interntaional Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3 Issue 1, January 2014)

[3]. Himani Sharma, Sunil Kumar (2015)

A Survey on Decision Tree Algorithms of Classification in Data Mining

(International Journal of Science and Research IJSR) ISSN (online) 2319-7064 PP 2094-2099

[4]. Agboola A.H Gabriel A. J., Aliyu E.O., Alese B.K.(2013)

Development of a fuzzy logic based rainfall prediction model

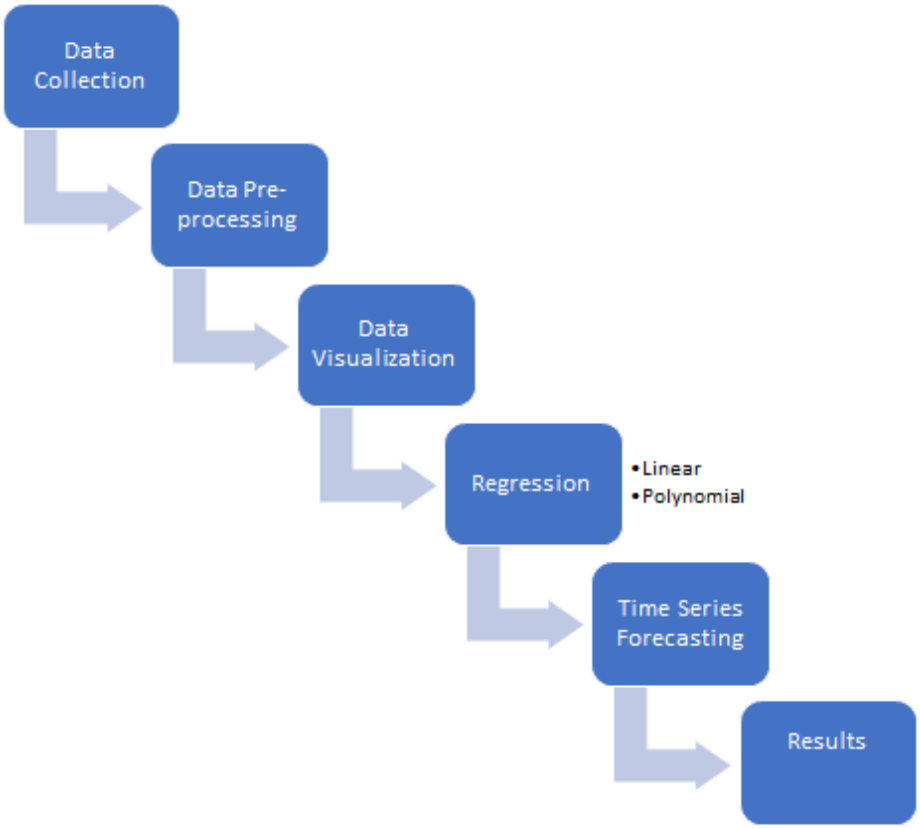
(International Journal of Engineering and Technology Volume 3 No. 4, April,2013, ISSN 2049-3444 PP 427-430

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

MATERIALS & METHODOLOGY

Step down process for the Prediction and Forecasting Analysis of Covid 19



1. **Data Collection:** Gathering of information is the most important step and it is the first step for prediction and forecasting results. There are different forms of data, data may be structured, unstructured and semi structured. The process of gathering and evaluating data on a certain variable in an organised way to provide answers to important questions and evaluate outcomes is known as data collecting. The goal of any data collection should be to collect high-quality data that can be evaluated to come up with solid, believable responses to the issues raised.

The first step in the process is gathering information about the ongoing Covid-19 outbreak in India, which was obtained from Kaggle. The columns of this dataset include the total number of confirmed, cured, and death cases of Covid-19 patients daily from March 12, 2020, to September 30, 2020, across all states.

A second dataset includes columns for total samples, positive results, and negative results from state-by-state testing carried out throughout India.

2. Data Pre-processing

Data pre processing is a second step where data cleaning plays a vital role, raw data may consist of inconsistent data, redundant data, noisy data, missing value, corrupted data. To clean these types data pre processing is very important. Data preparation is a process where these anomalies of data get removed and get cleaned data or final data for processing. These data suitable for the running the application in Machine learning. The study goes through two main reasons for data preparation

1. Data various issues

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

Most likely "a diamond in the rough," the data retrieved during the data retrieval process. Common data problems including incorrect data entry, excessive white spaces, impossible values, missing values, and outliers must be validated. Simple modelling techniques are used to find and pinpoint these data issues; diagnostic charts can be especially helpful. Data cleaning is used in the data pre-processing stage to eliminate redundant or null values. The process is then carried out by using the heatmap and built-in functions like is null to remove any null values from the data ().

3. Data Analysis preparation

The act of cleaning, transforming, and modelling data to unearth useful information for business choices is referred to as data analysis. Data analysis is to extract useful information from data so that decisions can be made based on that information. Once the data has been collected, cleaned, and processed, it is prepared for analysis. Data analysis software and tools can be used to examine this data at this phase in order to better comprehend, decipher, and draw conclusions based on the specifications.

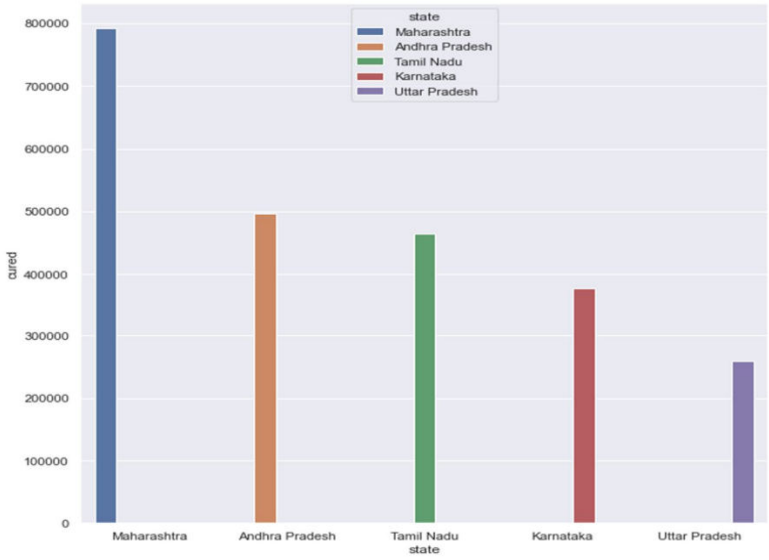
4. Data Visualization

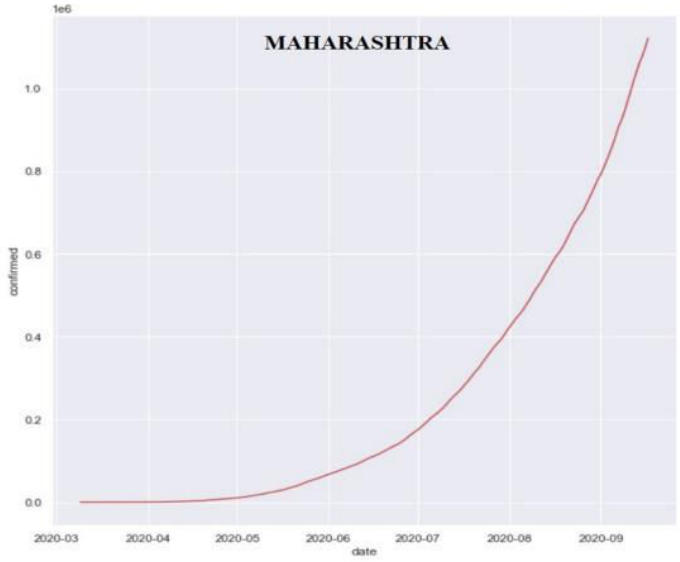
Another of the most crucial tools for defining a qualitative understanding is data visualisation. This can be helpful when attempting to study and extract information from a dataset as well as while looking for patterns, corrupt data, outliers, and other things.

In market research, categorical and numerical data can be represented, which increases the impact of insights while reducing analytical risk. Each one is an excellent component that enables users to assess the state and effects of numerous factors simultaneously. EDA works best when the data is properly understood before attempting to extract as many insights as feasible. Making meaning of the data is essential. EDA 1 and EDA 2 are the two steps of the process. The 'covid 19 India' dataset was put through the intensive pre-processing methods indicated above for EDA 1. Plotting a bar graph with variables 'x' and 'y' as 'states' and 'confirmed cases', respectively, allowed researchers to identify the number of “confirmed cases” in eight Indian states: Maharashtra, Andhra Pradesh, Tamil Nadu, Karnataka, Uttar Pradesh Delhi, West Bengal, and Telangana. This made it easier to determine which India had the most confirmed cases. In July, August, and September of 2020, Maharashtra had the highest number of covid-19 affected cases. Plotting a bar graph in a similar manner was done for the columns representing patients who had been cured and those who had died. Separate research was carried out, taking into account the three states of Maharashtra, West Bengal, and Mizoram, in order to go deeper into the process of identifying the severity of covid-19 in India. The number of new affected cases from March to September is depicted in the line graphs and bar graphs. According to statistics, West Bengal, which had an average of 2,12,383 instances till September, and Mizoram, which had 1506 cases up until September 17th, 2020, were the states with the most newly reported cases. Maharashtra came in second with the fewest. Through statistical analysis, "What happened?" is made clear. There are numerous ways to alter histograms with matplotlib. To calculate and produce a histogram of our variable, we utilise the matplotlib.pyplot.hist() function.‘confirmed’. The hist () function returns a patches object that gives us access to the produced objects' attributes and allows us to customise the plot. Line plots and scatter plots can help us analyse our model. A collection of data or a subset of data is examined. In a scatter plot, also known as a scatter chart or scatter graph, dots are used to represent the values for two different numerical variables. The locations of each dot on the horizontal and vertical axes represent the values for each data point. Values of each location indicates the horizontal and vertical axes. Scatter plots are used to see relation between the two various variables.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022





5. Regression:

Regression is of two types linear and polynomial.

a. LINEAR REGRESSION

Linear regression consists of two variables i.e x and y, by using regression analysis we can relate these two variables with the help of machine learning model. These variables are known as dependent and independent variables. When only one independent variable is provided, linear regression must be used to determine its linear connection to the dependent variable. The number of cases that could occur in the state of Maharashtra is predicted using the dataset known as "state-wise testing details" that was gathered from Kaggle. The majority of the datasets are in CSV format, and we use the pandas package to read these files: tests = pd.read csv('covid 19 India.csv'). Additionally, we have identified the "Confirmed, Cured, and Death cases" column attributes for this dataset, with "Months" acting as the independent variable X and "Confirmed" acting as the dependent variable Y. Based on the day they were officially confirmed, the cases in this dataset were categorised. After importing linear regression from the scikit-learn library, we split the data into training and testing portions using the function train test split (). 20% of the data in this model is utilised for testing, while the remaining 80% is used for training. The model is created to forecast the number of new cases that COVID-19 can infect in the state of Maharashtra based on the data provided. Finally, after execution, our

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

model is finished, and we can forecast the number of cases because we have x test data. To determine the accuracy of our predictions, we must compare the y forecast values to the original values. Predictions indicate that before the overall number of active cases begins to grow, we may observe a peak or plateau around 11,21,221 confirmed cases around the middle of September. Based on the data that has been fed into the model, it can forecast the outcome, which is the number of confirmed instances. Any data must constantly be treated because it is never entirely pure. A simple ML algorithm can be used to make predictions using this model as a starting point. Getting 91% accurate forecasts will be simpler with better data and more sophisticated machine learning systems. The total number of samples taken determines how many new instances there are. As a result, the frequency of new occurrences rises.

b. Polynomial

The second type of regression is known as polynomial regression, where it used for getting the relationship between the dependent variable y and the independent variable x with the 2^{nd}, 3^{rd} till 5^{th} degree polynomial of value x. Generally we are using least-squares method for fitting these types of models in machine learning. By using polynomial regression we can minimize the fluctuation of the estimator values of the coefficients very easily. We can use the anticipated value of y up to nth degree of polynomial function.

RESULTS

TABLE I. ACTUAL DATA

Months of 2020	Types of Cases			
	Confirmed	*Active*	*Cured*	*Death*
April	13788	10902	2451	435
May	94822	55396	36529	2896
June	351945	152324	188964	10656
July	1023435	353318	644520	25597
August	2604826	665008	1889706	50111
September	4970459	935915	3953097	81445
October	7315171	789945	6413690	111536

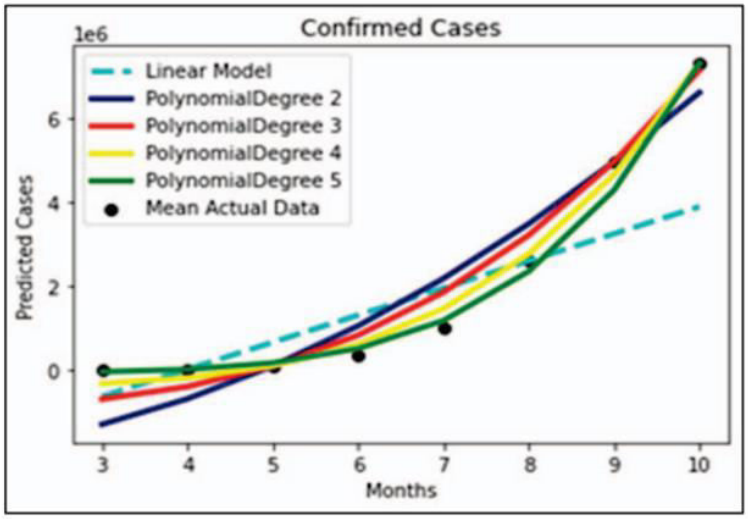


Figure: Regression models applied for Confirmed Cases

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

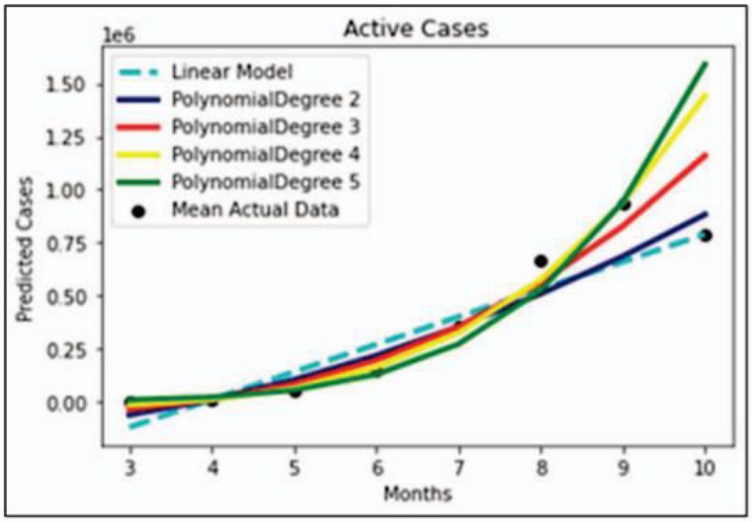


Figure: Regression models for Active Cases

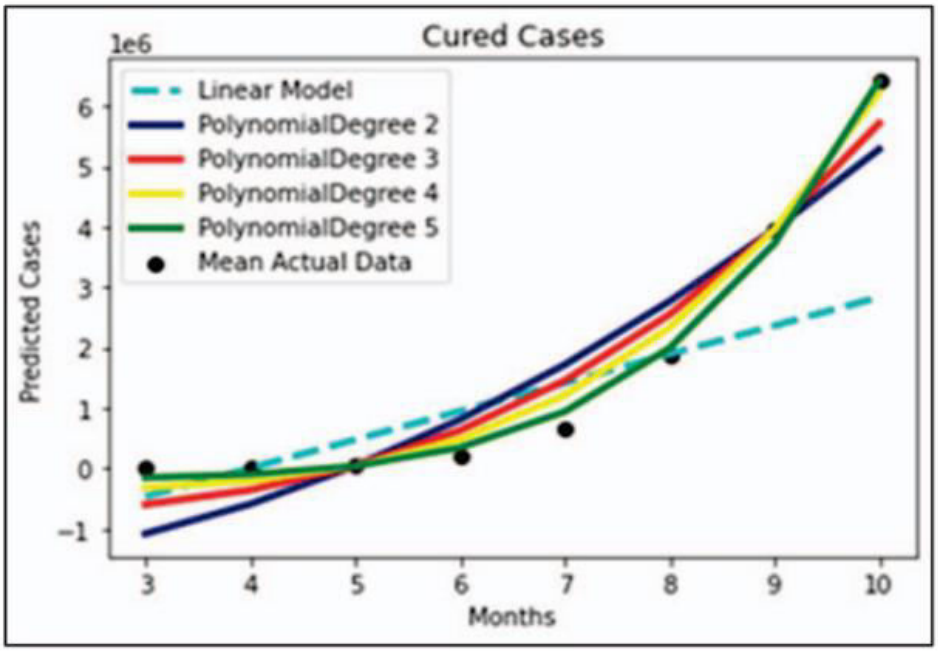


Figure: Regression model for Cured Cases

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

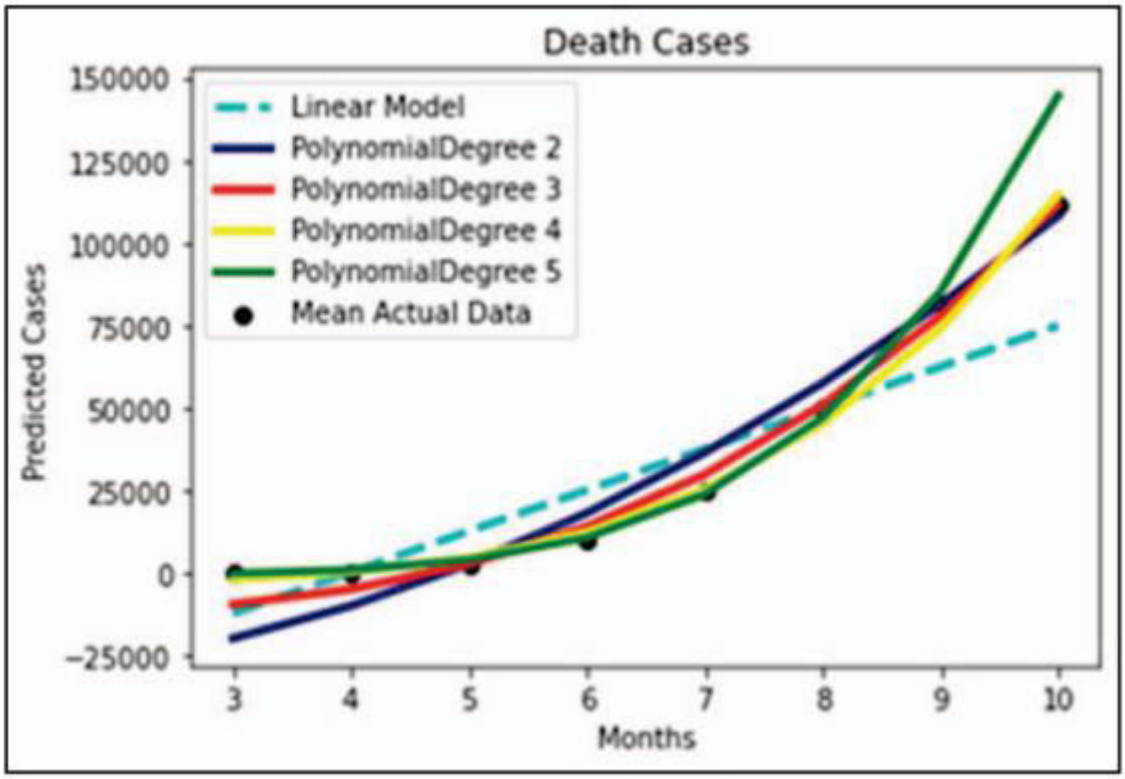


Figure: Regression models for Death Cases

Months of 2020	Types of Cases			
	Confirmed	*Active*	*Cured*	*Death*
April	13846	22001.5	-101771	1149
May	168800	55275	36546.6	4192
June	511824	128933.9	342742.5	10928
July	1177884	271959	937293	24007
August	2355050.5	524735.6	1988076.5	47121
September	4293344	940951	3718271	85182
October	7313589.6	1589497.9	6414256.8	144488

Table: Predicated Data by Polynomial Degree 5

Models		Types of Cases			
		Confirmed	*Active*	*Cured*	*Death*
Polynomial	*Degree 2*	5.77	0	144	1.24
	Degree 3	1.92	0	52.65	0.32
	Degree 4	0.46	0	17.48	0
	Degree 5	0	0	4.34	0
Linear		5.69	0	153	1.39

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

Table: Mean Square Error (MSE)

Mean Square Error shows the formulation of errors generated after the prediction of confirmed, actual and cured and death cases by using regression models, up to the 4th degree polynomial models and linear regression model. If we increase the polynomial the MSE decrease.

Models		Confirmed		Active		Cured		Deaths	
		R^2 score	*MAPE*	*R^2 score*	*MAPE*	*R^2 score*	*MAPE*	*R^2 score*	*MAPE*
Polynomial	*Degree 2*	0.968357651	679.76	0.87213159	13.3	0.957003654	3395.68	0.976890065	315.66
	Degree 3	0.971244641	392	0.932280645	14.84	0.967077464	2052	0.976893557	161
	Degree 4	0.973863222	193.8	0.96323221	12.25	0.968015478	1182.8	0.978630318	9.45
	Degree 5	0.974727757	16.53	0.965823518	20.55	0.968675839	589.44	0.978745042	33.34
Linear		0.791163744	125.66	0.85897801	27.9	0.744672132	231.84	0.857393043	10.39

Table: Accuracy of Covid 19 Analysis for India Based on Mean Average Percentage Error and R2 score

6. Time Series Forecasting

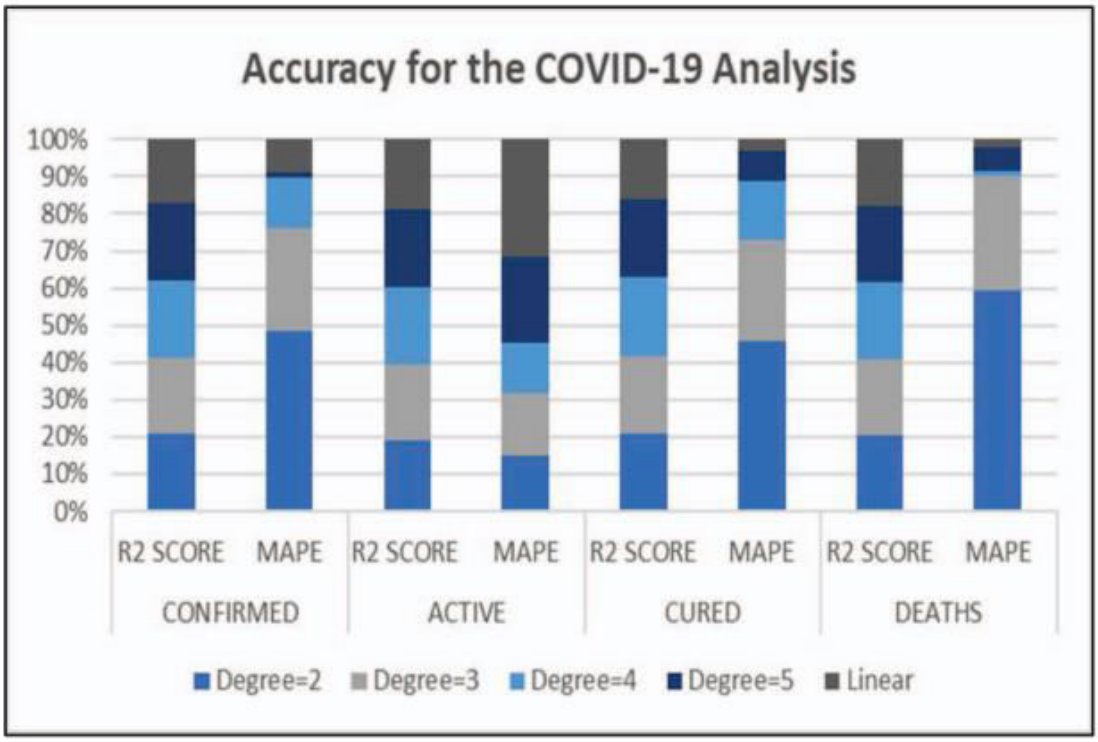


Figure: Performance evaluation of the regression models

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

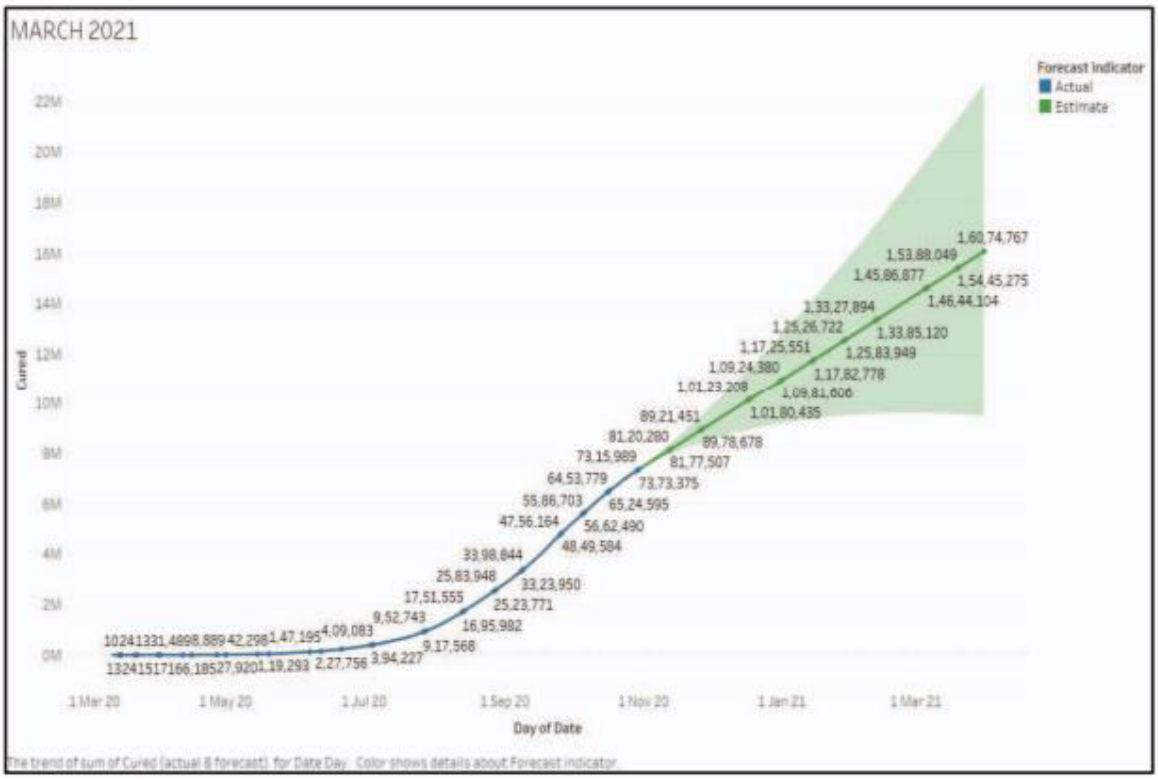


Figure: Forecasting of Cured Cases till March 2021

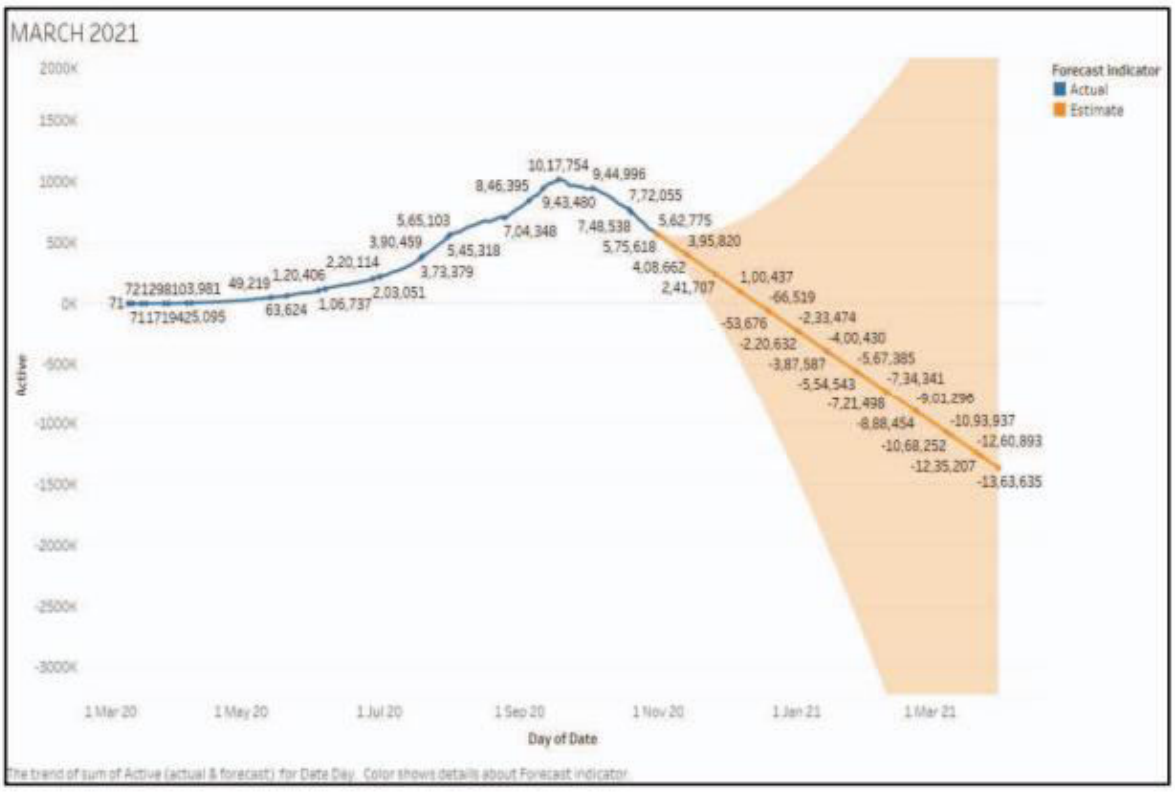


Figure: Forecasting of Active Cases till March 2021

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 10, Dec 2022

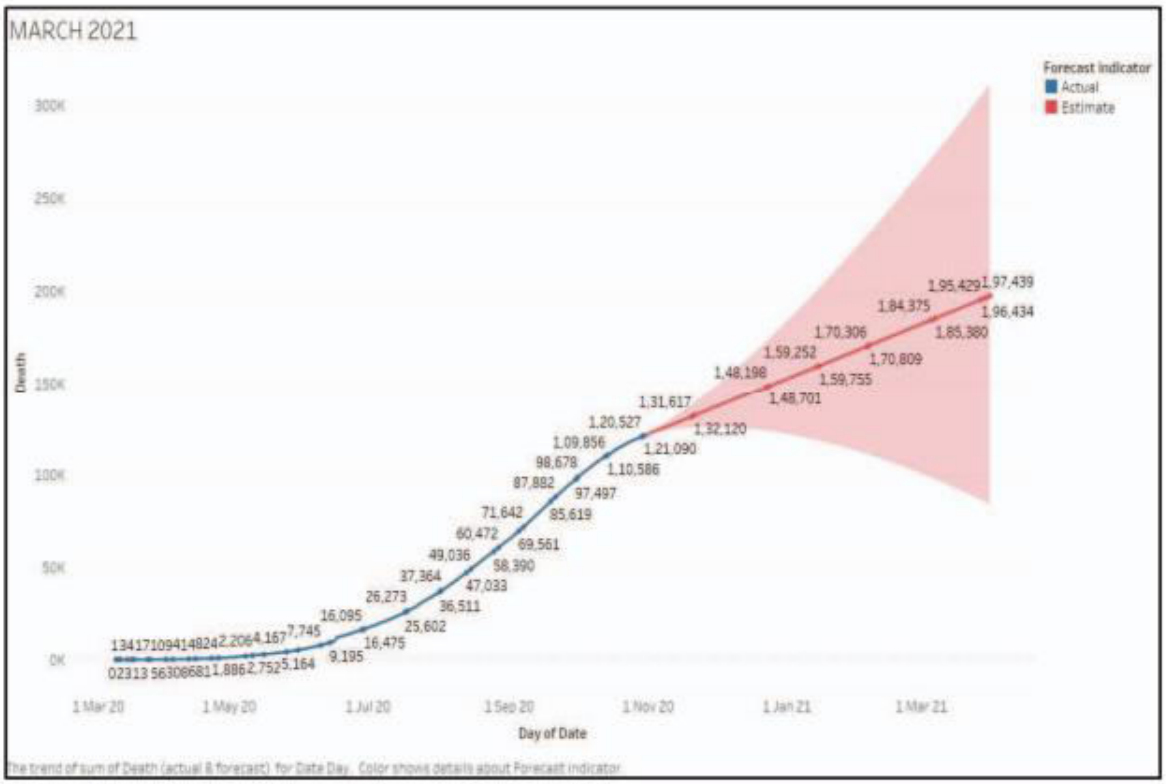


Figure: Forecasting of Death Cases till March 2021

CONCLUSION

To take any decision we required the historical data. By analysing the Covid 19 data set we can ensure about our safety by taking precautionary measure for the future purpose. Analysis is done with the help of application of data science i.e. linear and polynomial regressions where accuracy, R^2 score and MAPE.

We can conclude with the help of Accuracy of Covid 19 analysis for India table that polynomial regression is better than linear regression. By using Tableau forecasting and its results found satisfactory. Whereas the error rate in the future can be reduced as the size of the dataset increases day by day.

FUTURE ASPECT

In future we don't want pandemic situation, with the help of various applications of data science, machine learning and deep learning we will predict or forecast it very easily and we can handle the situation very well.

REFERENCES

[1] Deep Learning applications for COVID-19, Connor Shorten, Taghi M. Khoshgoftaar & Borko Furht.

[2] Dataset Reference: https://www.kaggle.com/

[3] Prediction and analysis of COVID-19 cases using deep learning models: 'A descriptive case study of India'. By Parul Arora et al. Chaos Solitons Fractals. 2020 Oct

[4] Machine learning-based prediction of COVID-19 diagnosis based on symptoms Yazeed Zoabi, Shira Deri-Rozov & Noam Shomron.

[5] Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting Publisher: IEEE.Cite This: Saud Shaikh; Jaini Gala; Aishita Jain; Sunny Advani; Sag

[6] COVID-19 in India: State Wise Analysis and Prediction by Palash Ghosh, Ph.D., Rik Ghosh, MSc, and Bibhas Chakraborty, PhD