iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39272

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Covid-19 Analysis and Prediction using Data Science and Machine Learning

Akshata Kulkarni[1], Atharva Kulkarni[2], Vaishnavi Kumbhar[3]
[1, 2, 3]*Marathwada Mitra Mandals College of Engineering*

Abstract: Officials around the world are using several COVID-19 outbreak prediction models to make educated decisions and enact necessary control measures. In this study, we developed a Machine Learning model which predicts and forecasts the COVID-19 outbreak in India, with the goal of determining the best regression model for an in-depth examination of the novel coronavirus. Based on data available from January 31 to October 31, 2020, collected from Kaggle, this model predicts the number of confirmed cases in Maharashtra. We're using a Machine Learning model to foresee the future trend of these situations. The project has the potential to demonstrate the importance of information dissemination in improving response time and planning ahead of time to help reduce risk.

I. INTRODUCTION

The year 2020 has proven to be a disaster for humanity. COVID-19 was discovered in India during the last week of January 2020, when a group of Indian students went to Kerala from Wuhan, China. Since then, the global response to the COVID-19 pandemic has been heavily reliant on Data Science in general.

We attempt to centralize the various COVID-19 research activities using Data Science, where we define Data Science to include the various methods and tools—including those from Machine Learning (ML), statistics, modeling, simulation, and data visualization using python —that can be used to store, process, highlight and extract insights from the data. We highlight common issues and traps seen across the surveyed works, publicly available datasets. All of this is done in an interactive computing environment- Jupyter Notebook, in which users execute code to see what happens which leads to new results. In light of this, our research addresses two critical challenges.

To begin, the first step is to provide strong insights using Exploratory Data Analysis (EDA) which is concerned with employing descriptive statistics, visual techniques, and simple modeling as the data may be in a graphical format which is the easiest way to understand large scale data.

Because Machine Learning and Data Science are often used to process massive amounts of data, that is why a powerful low-level language is recommended. Python– a general-purpose language used by data scientists and developers, because its straightforward syntax makes it simple to communicate. The second step is developing a predictive model based on Machine Learning techniques that were created to accurately forecast the number of COVID 19 positive cases in the country.

The majority of previous research and media attention centered on the total number of infections in the country. However, considering India's size and diversity, it's vital to look at the disease's spread in each state independently, as the situations vary greatly. The goal of this study is to evaluate data on the number of sick people in each Indian state (limited to only those states with sufficient data for prediction) and forecast the number of illnesses in that state over the next 8 days. The majority of Indian states are relatively substantial in terms of both area and population. When analyzing coronavirus infection data, assuming that the entire country is on the same page may not provide us with the most accurate picture. This is because the first infection, new infection rate, progression over time, and preventive actions taken by state governments and the general public differ from state to state. Each state must be addressed independently. It will allow the government to make the best use of the limited resources available. We anticipate that state-by-state projections will help state governments allocate their limited healthcare resources more effectively.

II. METHODOLOGY

Understanding the what, why, and how of this project is the first step. It’s not always possible to plan every piece of the Data Science process ahead of time. Data can be stored in a variety of formats, from basic text files to database tables. The goal now is to collect all of the information that is required.

©IJRASET: All Rights are Reserved

A. *Data Collection*

Data collection is the process of obtaining and analyzing data on a certain variable in a structured manner, allowing one to answer pertinent questions and assess consequences. All data gathering should aim to acquire high-quality evidence that can be analyzed to produce convincing and reliable answers to the questions that have been addressed. The process begins with acquiring data for the ongoing Covid-19 outbreak in India which was collected from Kaggle; the columns of this dataset comprise the total number of confirmed, cured, and death cases of Covid-19 patients throughout all states daily from March 12, 2020, to September 30, 2020. Another dataset comprises state-by-state testing conducted across India, with columns such as total samples, positive and negative results.

B. *Data Pre-processing*

Real-world data sometimes contains noise, missing values, and is in an unsuitable format that cannot be used directly in Machine Learning models. Data preparation is the process of preparing it for usage. It is a necessary task for cleaning data and making it suitable for a Machine Learning model, which improves the model's accuracy and efficiency. Before any data analysis process can begin, these are the primary issues that must be thoroughly understood and resolved.

The study goes through two primary reasons for data preparation in-depth:

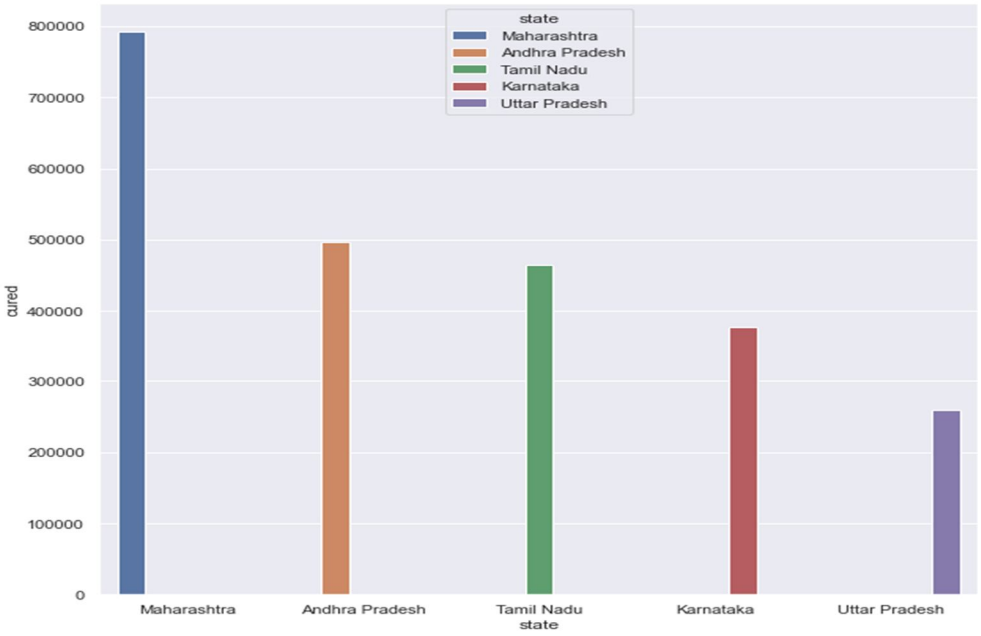
(I)Data issues (II) Data Analysis Preparation

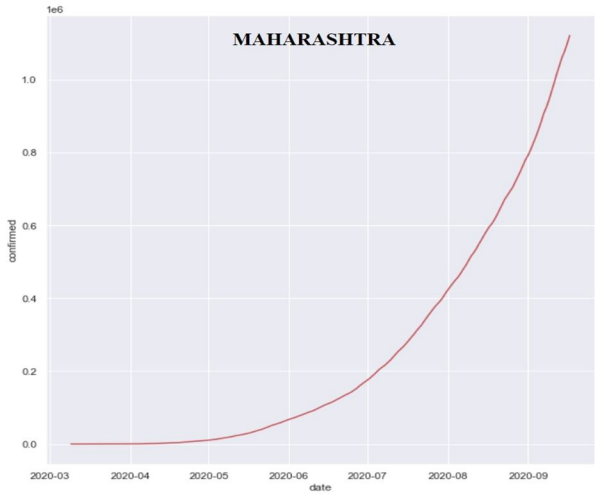
1) *Data Issues:* The data obtained during the data retrieval step is most likely "a diamond in the rough". Validation is required for common data issues such as data entry errors, superfluous white spaces, impossible values, missing values, and outliers. To uncover and identify these data flaws, simple modeling techniques are applied; diagnostic charts can be particularly useful. In the section of data pre-processing, redundant or null values were removed by data cleaning. The process is then continued by removing any null or NaN values from the data using the heatmap and built-in functions such as isnull ().
2) *Data Analysis Preparation:* Data analysis is described as the act of cleansing, transforming, and modeling data to uncover usable information for business decisions. The goal of data analysis is to extract usable information from data and make decisions based on that knowledge. The data is ready for analysis once it has been gathered, cleansed, and processed. During this phase, this data can be analyzed using data analysis tools and software to help understand, analyze, and draw conclusions based on the requirements.

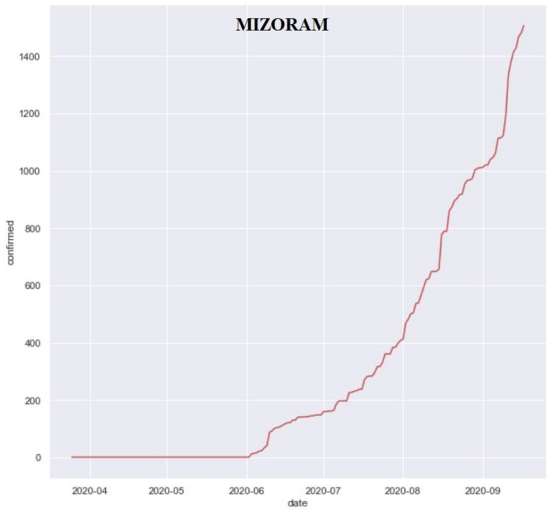
III. DATA VISUALISATION

Data visualization is one of the most important instruments for determining a qualitative understanding. This can be useful when trying to examine a dataset and extract information about it, as well as spotting patterns, corrupt data, outliers, and other things. Both numerical and categorical data can be visualized in market research, which serves to boost the impact of insights while also lowering the risk of analysis. Each represents a high-quality component that enables clients to evaluate the condition and impact of a large number of variables at the same time. EDA is best to first understand the data and then try to glean as many insights as possible from it. It is all about making sense of the data. The procedure is divided into two stages: EDA 1 and EDA 2. For EDA 1, the dataset 'covid_19_india' was subjected to the above-mentioned extensive pre-processing techniques. The number of confirmed cases in eight Indian states, namely Maharashtra, Andhra Pradesh, Tamil Nadu, Karnataka, Uttar Pradesh Delhi, West Bengal, and Telangana, were determined by plotting a Bar graph with variables ‘x’ and ‘y’ as ‘states’ and ‘confirmed cases’ respectively. This facilitated the process of identifying the greatest number of confirmed cases in India. It was recorded that Maharashtra had the most number of covid-19 positive cases in July, August, September 2020. A similar process of plotting a bar graph was carried out for the columns of death cases and cured patients. To delve deeper into the process of understanding the severity of covid-19 in India, a separate analysis was conducted, taking three states into account: Maharashtra, West Bengal, and Mizoram. The Line Graphs and Bar Graphs show the number of new positive cases from March to September. It was recorded that Maharashtra had the highest number of new cases, followed by West Bengal, which had an average of 2,12,383 cases until September, and Mizoram, which had 1506 cases until September 17th, 2020. Statistical Analysis demonstrates "What happened?" through the use of historical data in the form of dashboards, histograms, scatter plots, line plots, etc. For EDA 2, matplotlib provides a range of different methods to customize histograms. The matplotlib.pyplot.hist() function is used to compute and generate a histogram of our variable ‘confirmed’. The hist () function returns a patches object that gives us access to the attributes of the created objects, enabling us to alter the plot according to our preference. Similarly, scatter plots and line plots assist us in analyzing our model. It examines a set of data or a subset of data. Dots are used to indicate values for two different numeric variables in a scatter plot (also known as a scatter chart or scatter graph). The values for each data point are indicated by the position of each dot on the horizontal and vertical axes. Scatter plots are used to see how variables relate to one another.

©IJRASET: All Rights are Reserved

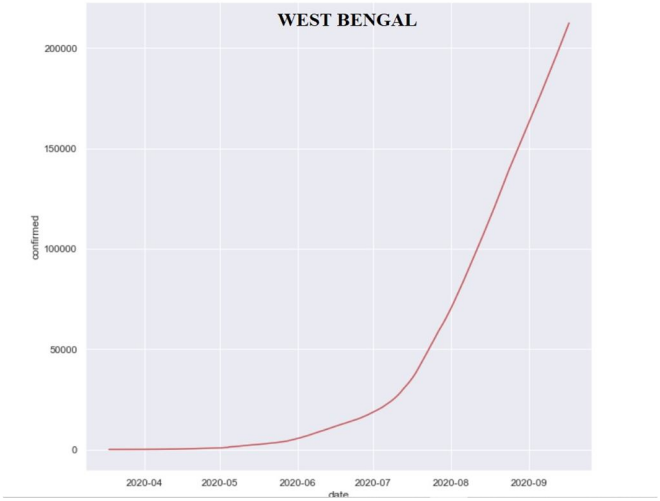






©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com



IV. PREDICTION USING LINEAR REGRESSION MODEL DEVELOPMENT

Linear Regression is a supervised Machine Learning model that finds the best fit linear line between the independent and dependent variables, that is the linear relationship between the dependent and independent variables. In this model, we have used Multiple Linear Regression where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable. The dataset 'state-wise testing details' obtained from Kaggle is used to forecast the number of cases that could occur in the state of Maharashtra. Most of the datasets are in the CSV file, for reading this file we use pandas library; tests = pd.read_csv('covid_19_india.csv'). Furthermore, we have defined the column attributes of this dataset as "Confirmed, Cured, and Death cases", with "Months" as the independent variable X and "Confirmed" as the dependent variable Y. The cases in this dataset were classified based on the date they were confirmed. We used the scikit-learn train_test_split () function to divide the data into training and testing parts after importing linear regression from the sci-kit-learn library. In this model, 20% of the data is used for testing and the other 80% is used for training. Based on the data provided, the model is developed to train the dataset to forecast the number of new instances that can be infected by COVID-19 in the state of Maharashtra. Finally, after execution, our model is complete; now that we have x test data, we can predict the number of cases. We must compare the y prediction values to the original values to calculate the accuracy of our model, which was implemented by a concept known as the r2 score.

According to predictions, we may see a peak or plateau of 11,21,221 confirmed cases around the middle of September, before the total number of active cases starts to rise. This model can predict the outcome that is the number of confirmed cases based on the data that has been fed to it. Any data is not completely in its pure form and always needs to be processed. This model is a basic example of how predictions can be made using a simple ML algorithm. With more accurate data and advanced Machine Learning algorithms, it will be easier to obtain 91% accurate predictions. The number of new instances is determined by the total number of samples collected. As a result, the rate of new instances increases exponentially.

V. CONCLUSION

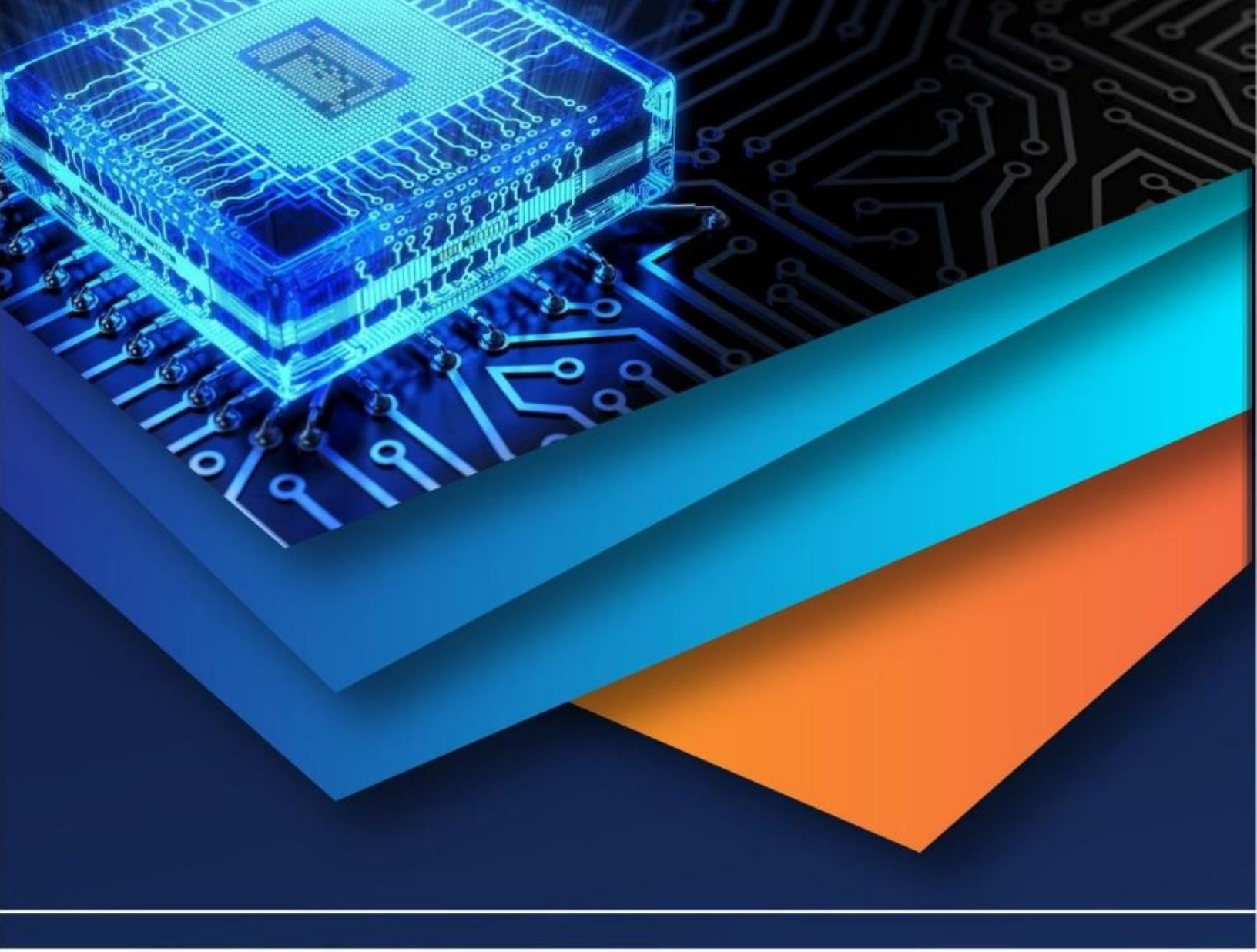
With data analytics and data mining, information and communication technology aids in the decision-making process based on historical data. The amount of data available is enormous, making gathering information and extracting an intriguing pattern from it a difficult undertaking. The current data on confirmed, recovered, and death in India over a long period, aids in predicting and forecasting the near future. The model's accuracy could be improved by including more features such as many hospitals, the infected person's immune system, the patient's age, gender, and the efforts taken to combat the virus's spread, among others, to make it entirely informative. Data is available on a massive scale that must be reduced and made understandable for each individual. The information we obtain is displayed in graphical form, which is the simplest way to comprehend enormous amounts of information. Given the importance of data openness inside the government, it is also our responsibility to avoid spreading false information and to remain calm in this situation. The research has demonstrated the relevance of information dissemination in terms of increased response times and reducing risk by planning ahead of time.

©IJRASET: All Rights are Reserved

REFERENCES

[1] Prediction and analysis of COVID-19 positive cases using deep learning models: 'A descriptive case study of India'. By Parul Arora et al. Chaos Solitons Fractals. 2020 Oct
[2] Machine learning-based prediction of COVID-19 diagnosis based on symptoms Yazeed Zoabi, Shira Deri-Rozov & Noam Shomron.
[3] Deep Learning applications for COVID-19, Connor Shorten, Taghi M. Khoshgoftaar & Borko Furht.
[4] Analysis and prediction of COVID-19 trajectory: A machine learning approach Ritanjali Majhi,Rahul Thang.
[5] Dataset Reference: https://www.kaggle.com/
[6] 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
[7] COVID-19 in India: State Wise Analysis and Prediction by Palash Ghosh, Ph.D., Rik Ghosh, MSc, and Bibhas Chakraborty, PhD

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)