

# **DATA ANALYSIS ON COVID-19**

Harshit Arora (04317704422)

## **Abstract**

The Covid-19 pandemic has shaken the world completely. No one knew what was coming and everyone was running helter-skelter. The governments were paralyzed and the infrastructure required to deal with this problem was absent completely. Our Project seeks to uncover the mystery using the application of data analysis to solve it. Patterns reveal what the common issues are and common symptoms and everything that is common comes out in a visual representation. It's these representations which make complex things easy and digestible to people from non tech backgrounds. A huge dataset of people suffering from Corona virus to give us better ways of fighting the pandemic.

In-fact data science is the new method of diagnostic and can lead to even better cure for diseases. It's this frontier we seek to find from our project.

## **Introduction**

Covid-19 cases are increasing day by day in and all over the world, millions of people are dying and the economy is experiencing free-fall. It has been spreading like water in the open ocean and it seems like there is no stopping it, But thankfully since the inception of this epidemic many countries have properly managed the databases of each and every patient and their health history.

We today have advanced computational infrastructure and data science algorithms through which we can analyse these data sets and gain insight-full information so that we can help the society. People have proposed many interesting models and trend prediction methods.

The project will help us in recognizing the insights that will be gained by using data science algorithms on the data, these insights will help us in identifying and

giving an idea of how the number of covid cases are impacted as possibility of being diagnosed positive on the basis of the symptoms .

## Literature Survey

According to the research paper [1], the authors, R. Wang, G. Hu, C. Jiang, H. Lu and Y. Zhang, have compared the prediction of patterns by using 3 methods and comparing their graphs with each other. These models are the conventional logical regression model, the Particle Swarm Optimization SIR model and the Lowest Square approach SIR model. The chart ultimately shows some patients with a novel form of X-axis coronary pneumonia, and Y-axis date. By seeing the three patterns we come to know that the data is plotted in the form of a curve.

“The public figures of daily updated confirmed instances of Covid-19 from University John Hopkins were analysed in this study article [2] proposed by V.Z.Marmarelis.[2]. RM as described by Riccati Equation, is the main modelling element for the method. The public figures of daily updated confirmed instances of Covid-19 from University John Hopkins were analysed in this study article [2] proposed by V.Z.Marmarelis et al. [2]. RM, as described by Riccati Equation, is the main modelling element for the method. Further by applying the equation we find 5 different parameters and their dependence on the no. of cases increasing day by day”.

Everyone analysed knowledge on coronary disease and sustainable therapy utilising research articles from Gerry Wolfe\*, Ashraf elnashar\*, Will Schreiber\* Izzat Alsmadi\*. " Guided by COVID-19 Literary Clustering of the Datasets from Kaggle based on COVID-19. [3] The data were further divided into four: (1) Mobility social distances, (2) Health and COVID; (3) Economic impact; and (4) Vulnerable population, and were utilised in a second dataset from MTI. The document has been analysed and text has been processed in order to produce tokens for clustering and the use of the K-Median method to label data to assist extract and analyse categorised data.

According to Tuli,[4] the epidemic may be tracked extremely efficiently via Shrestha et al Machine Learning (ML) and Cloud Computing, anticipate an

outbreak of the illness, and create appropriate policies to regulate its expansion. Then given the array, face extraction and collection is done. They have proposed a Machine Learning model that can be run continuously on Cloud Data Centers (CDCs) for accurate spread prediction and proactive development of strategic response by the government and citizens. The dataset used by them in this case study, World in Data by Hannah Ritchie. They have also used a cloud framework and azure instances for real time analysis of data. The research paper [5] Francisco Nauber, Bernardo Gois et al. have emphasised the rising popularity of epidemic behaviour prediction research due to their capacity to anticipate the natural course of viruses.

This study presents several predictor approaches with machine training, logistic regression, filters, and epidemiological models in order to explain COVID-19's behaviour. The research paper [6], the authors Yazeed Zoabi, Shira Deri-Rozov and Noam Shomron have acknowledged that accurate SARS-CoV-2 screening allows for fast and efficient COVID-19 diagnosis and reduces the strain on health care systems. Prediction models using many characteristics have been created to assess the likelihood of infection. The model projected 0.90 auROC in the forwardlooking test set (area under the receiver operating characteristic curve). The research paper [7], authors Enis Karaarslan and Doğan Aydın mentioned that The incident at COVID-19 showed that the world was unwilling to disseminate the virus so rapidly. One crucial factor in mitigating the detrimental impacts of an epidemic or pandemic is the effective use of information technology. They suggested a management epidemic system (EMS), which relies on the unfettered and timely flow of information between states and organisations. They have been using an MPISA paradigm, which allows different platforms to be integrated and gives the solution for issues of scalability and interoperability. [8] This paper Describes the use of a new epidemiological compartment-based model for the estimation of the propagation of the coronavirus COVID-19, that is, SEIAR (Susceptible Exposed Asymptomatic Infectious Recovered). This is accomplished through the heuristic approach of differential evolution. In this way the day(s) when that number reaches its maximum, the associated value and the future evolution of its spread may be evaluated in approximate order for different situations. The [9] authors Ayyoubzadeh S et al have Used computerised data mining technologies for improved insights on the outbreak of COVID-19 in each country and globally for

the management of the health catastrophe. Google Trends website collected data. For estimating the number of positive COVID-19 instances, linear regression and long-term memory (LSTM) models were utilised. [10] The study document [7] by Amir-Sardar Kwekha Rashid, Heam N Abduljabbar and Bilal Alhayani shows that in COVID-19 research, hypotheses may be proved to be deterministic, transforming into clear findings and predictions. The outcomes of supervised learning algorithms are better than those of 92.9% of uncontrolled learning algorithms. The assistance for the development of standard diagnostic procedures like IgM, IgG, X-ray chest, CT-scans and RT-PCR can be seen as an artificial intelligence and deep learning. The CNN Algorithms selected to perform this study are MobileNet, DenseNet, Xception, ResNet, InceptionV3, InceptionResNetV2, VGGNet, NASNet.

## Methodology

We clean the data by using excel cleaning methods. The process can be explain in following points :

1. First, Take the dataset, remove redundant data and organise the data according to our needs.
2. Second, Load the dataset on the Jupyter Notebook and apply data visualization techniques to understand the data better.

### Description of the Process

We are building our own COVID Analysis System using Jupyter Notebook. We can describe the process in following steps :

Step 1: Cleaning the dataset The very first step in our project is to get a reliable and authentic dataset for the analysis. Our search for dataset ended on [11] which is govt website which has provided dataset for free use and is absolutely authentic. Then next thing we did was to clean the dataset and remove unwanted columns from dataset for faster computation.

Step 2: Data Visualization Here, we use the dataset and check the consistency of the dataset by checking the values out of the dataset randomly. Then we do data visualization for better understanding of data by the use of various plots, graph and heatmaps. All this graphs and plots gets us an insight into huge datasets easily.

# Results and analysis

jupyter Final Python Project Last Checkpoint Last Sunday at 1:22 AM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

## Corona Virus Pandemic - In India

### An Exploratory Data Visualization and Analysis

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus.

Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness. Anyone can get sick with COVID-19 and become seriously ill or die at any age.

In this notebook, We will take a look at the current situation in India. We will take a look at the regions which are most hampered by the outbreak and how numbers have steadily climbed in the country.

### PROGRAMMING LANGUAGE AND MODULES INCLUDED IN THIS PROJECT :

```
In [37]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
%matplotlib notebook
```

### Data Sets Included in this Project :

```
In [39]: covid19_df = pd.read_csv(r"C:\Users\VARSHIT\Downloads\COVID19-EDA-INDIA-master\COVID19-EDA-INDIA-master\covid_19_india.csv")
Individuals_df = pd.read_csv(r"C:\Users\VARSHIT\Downloads\COVID19-EDA-INDIA-master\COVID19-EDA-INDIA-master\IndividualDetails.csv")
excel_file = pd.ExcelFile(r"C:\Users\VARSHIT\Downloads\COVID19-EDA-INDIA-master\COVID19-EDA-INDIA-master\ISPA.xlsx")
covid_df = pd.ExcelFile(r"C:\Users\VARSHIT\Downloads\COVID19-EDA-INDIA-master\COVID19-EDA-INDIA-master\Covid cases in India.xlsx")
dod_india = pd.read_excel(r"C:\Users\VARSHIT\Downloads\COVID19-EDA-INDIA-master\COVID19-EDA-INDIA-master\per_day_cases.xlsx", parse_vaccine_df = pd.read_csv(r"C:\Users\VARSHIT\Downloads\COVID19-EDA-INDIA-master\COVID19-EDA-INDIA-master\covid_vaccine_statewise.csv")
Indian_states_df = excel_file.parse('Sheet1')
```

In [ ]:

Type here to search

GBP/INR +0.84%

22:12 15-06-2023

jupyter Final Python Project Last Checkpoint Last Sunday at 1:22 AM (autosaved) Logout

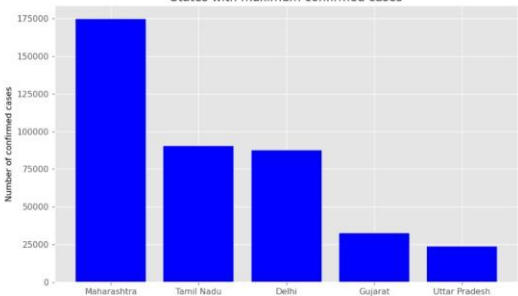
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

## STATEWISE FIGURES

```
In [21]: covid19_df_latest = covid19_df_latest.sort_values(by=['confirmed'], ascending = False)
plt.figure(figsize=(10,6), dpi = 90)
plt.bar(covid19_df_latest['State/UnionTerritory'][:5], covid19_df_latest['confirmed'][:5], align='center', color='blue')
plt.ylabel('Number of confirmed cases')
plt.title('States with maximum confirmed cases')
plt.show()

<IPython.core.display.Javascript object>
```

### States with maximum confirmed cases

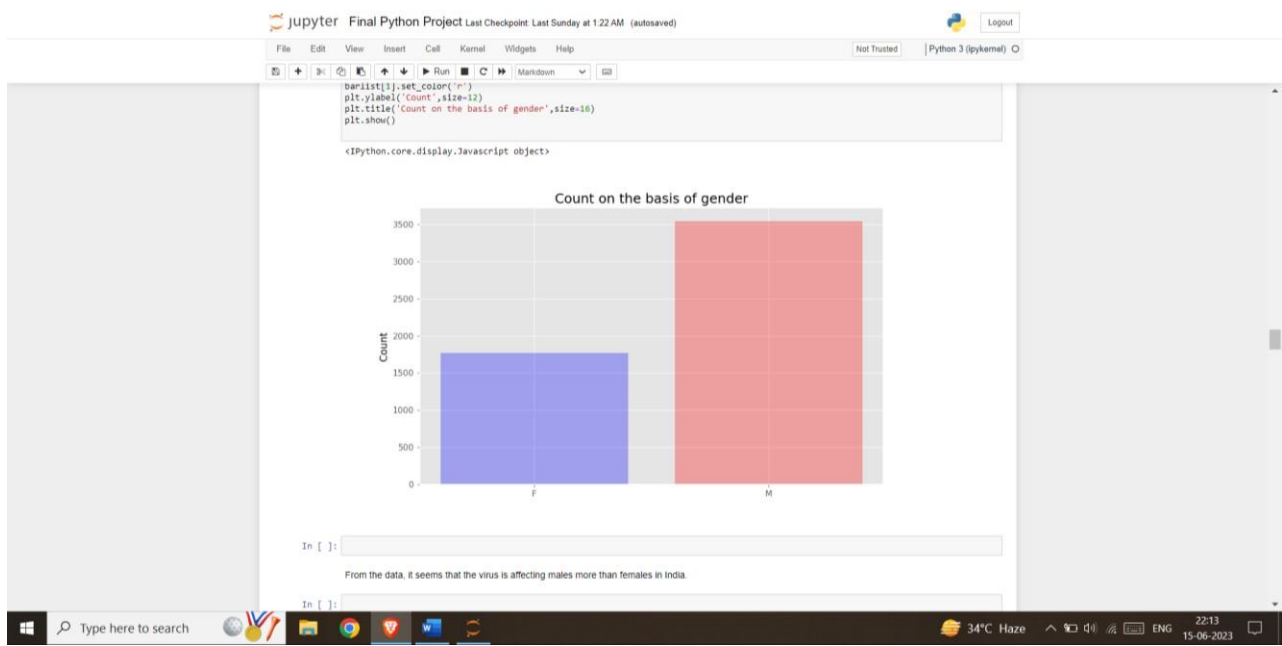
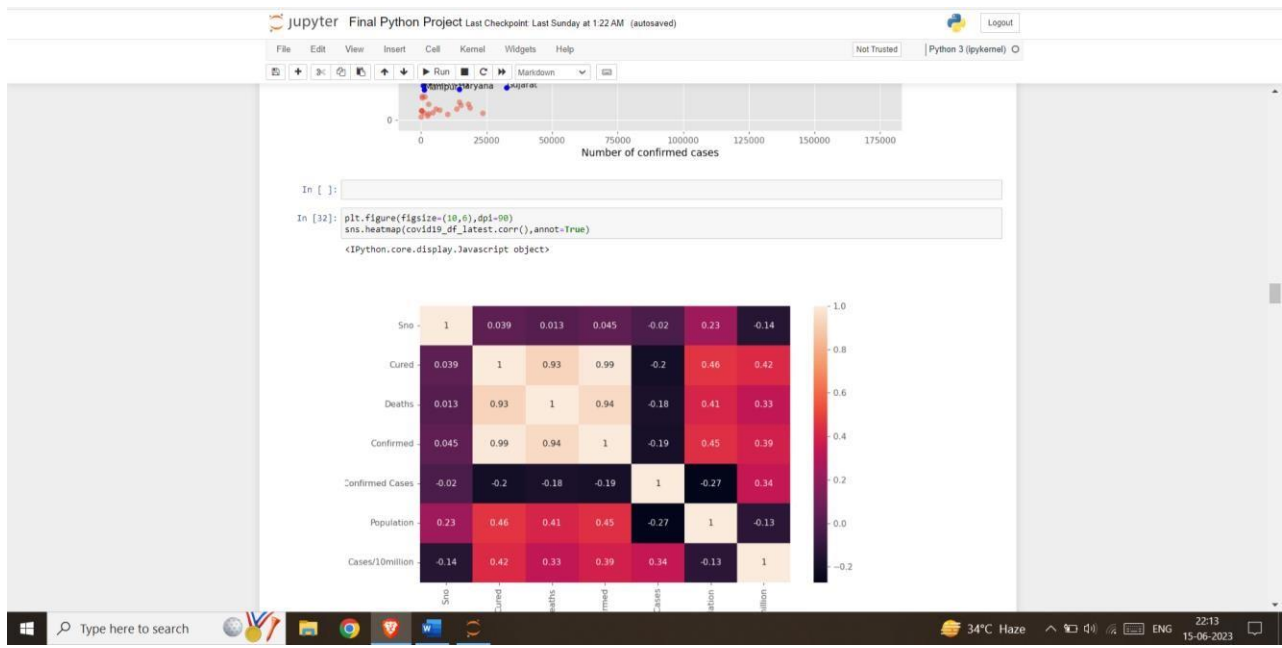


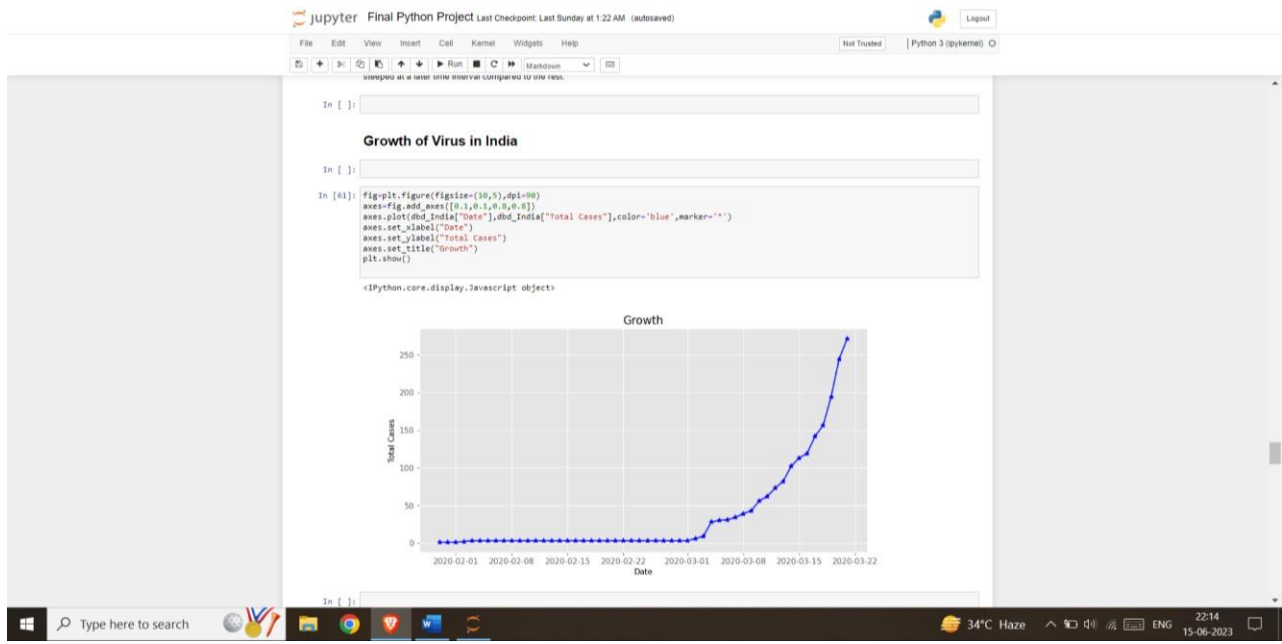
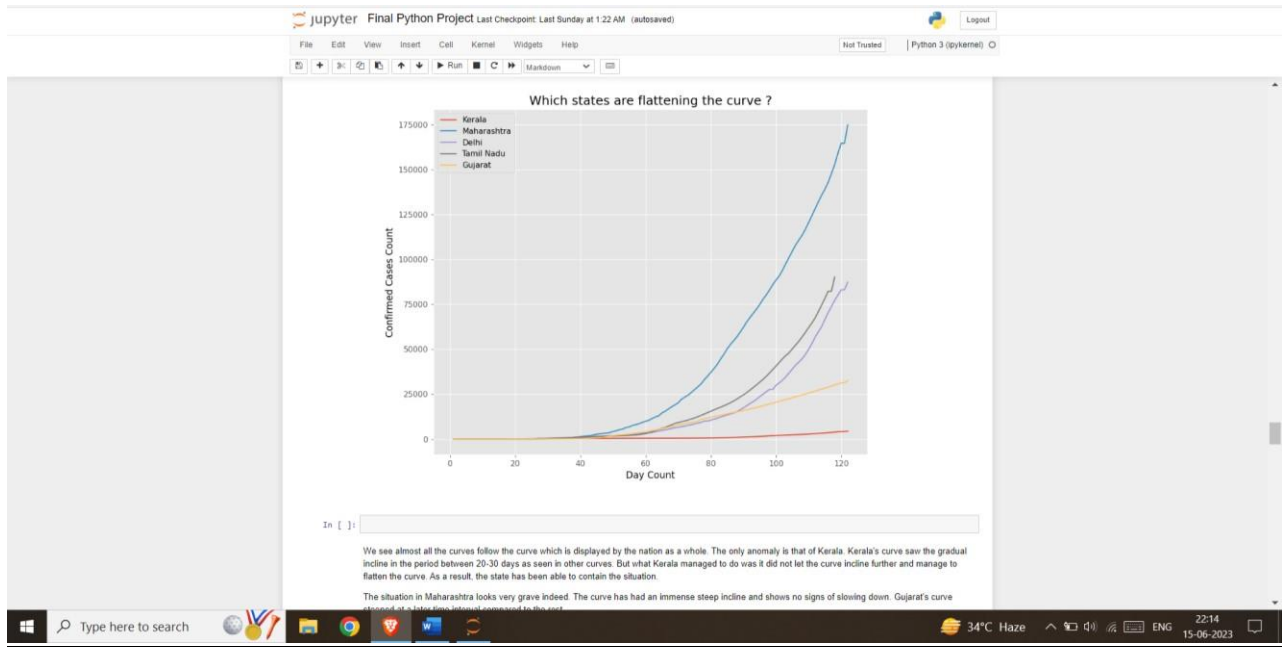
| State/UnionTerritory | confirmed |
|----------------------|-----------|
| Maharashtra          | 175000    |
| Tamil Nadu           | 90000     |
| Delhi                | 85000     |
| Gujarat              | 30000     |
| Uttar Pradesh        | 25000     |

Type here to search

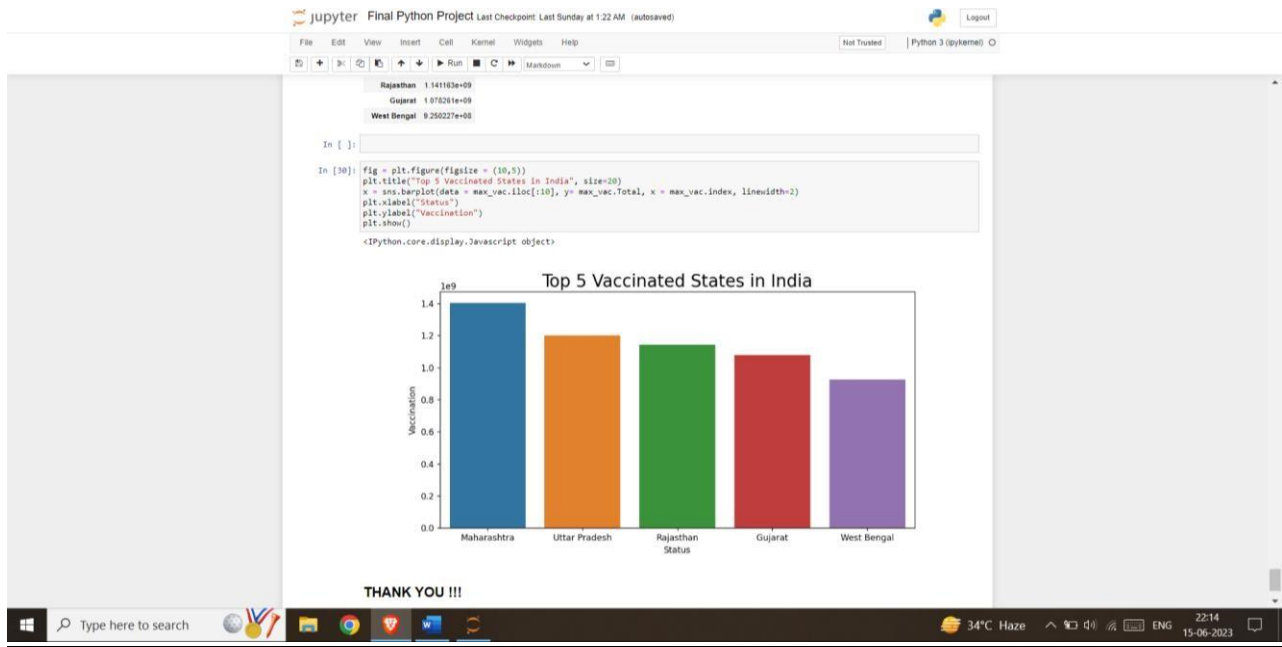
GBP/INR +0.84%

22:12 15-06-2023









## **Conclusion and future work**

The Covid - 19 Pandemic is a huge struggle for all of us. The project we are making will seek to find the answers to the most pertinent questions as to what is it that makes the covid 19 such a tragedy and what all people are the ones who are most affected by it. It will seek to find the appropriate response which can be mounted by the authorities concerned and we can reach to a place of proper discussion about the problem and solve it in the best possible manner out there. It will also lead to a solution to any medical condition we might encounter later on in our lives where we can apply data sciences for medical diagnostics. This project saves on the already limited resources that India have and prevents the spread as people can use it to get an idea that they should go and get tested .It also helps unhealthy and infected people to isolate themselves. Using this system we can effectively and efficiently mitigate the burden on our healthcare system which is completely stressed out.

We are using Machine Learning to give predictions on the basis of data taken from government website. using the accuracy graph we finally use the algorithm with best accuracy in this case (Decision Tree Classifier) to predict the person is either ve or +ve on the basis of symptoms.

**Computing Accuracy :** In this step we compute accuracy of all the algorithms by checking the four algorithms mentioned here: Logistic Regression, KNN, Random Forest Classifier, Decision tree Algorithm , we selected these algorithms on the basis of their qualities of regression & classification.

**Predicting Covid +ve or -ve :** All we need to do is plot a graph of accuracy of all the algorithms and use the algorithm with best accuracy to predict whether a person has corona or not. We take input of 5 symptoms in binary values and using our predictor we predict the person is positive or negative on the basis of these 5 symptoms.

## References

- [1]R. Wang, G. Hu, C. Jiang, H. Lu and Y. Zhang, "Data Analytics for the COVID-19 Epidemic," 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 2020,pp.1261-1266,doi:10.1109/COMPSA C48688.2020.00-83. [2]V. Z. Marmarelis, "Predictive Modeling of Covid-19 Data in the US: Adaptive Phase-Space Approach," in IEEE Open Journal of Engineering in Medicine and Biology, vol. 1, pp. 207-213, 2020, doi: 10.1109/OJEMB.2020.3008313.
- [3]G. Wolfe, A. Elnashar, W. Schreiber and I. Alsmadi, "COVID-19 Candidate Treatments, a Data Analytics Approach," 2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA), 2020, pp. 139146, doi: 10.1109/MCNA50957.2020.9264290.
- [4]Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic. Internet of Things,11,100222.<https://doi.org/10.1016/j.iot.2020.100222>
- [5]Nauber Francisco Santiago Valdir, Santiago Saulo Melo, Raphael Costa, Marcelo Oliveira, Oliveira Francisco the Chagas, Bernardo Gois, Alex Lima, Kennedy Santos. Douglas Marques Henrique, João Alexandre Lôbo Marques