A Project/Dissertation ETE Report

On

CORONA VIRUS OUTBREAK PREDICTION USING M.L.

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

Submitted to:

Ms. Vaishali Gupta

Submitted by:

Aman Kumar

Vivek Kumar

19SCSE1010688 / 19SCSE1010398

SCHOOL OF COMPUTING SCIENCE & ENGINEERING

Galgotias University, Greater Noida

December – 2021

ACKNOWLEDGEMENT

In performing our assignment, we had to take the help and guideline of some respected person, who deserve our greatest gratitude. The completion of this assignment gives us much Pleasure. We would like to show our gratitude Ms. Vaishali Gupta, Project Instructor, Galgotias University who introduced us to the Methodology of work, and whose passion for the "underlying structures" had lasting effect and for giving us a good guideline for assignment throughout numerous consultations. We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us in writing this project.

Many people, especially our classmates and team members itself, have made valuable comment suggestions on this proposal which gave us an inspiration to improve our assignment. We thank all the people for their help directly and indirectly to complete our assignment.

Abstract

On this research, predicting the corona virus is divided into many phases, state wide analysis which covers active , total, cured and death cases, also the increase in cases on daily basis which covers India confirmed, death and recovered cases. This also covers thread of novel coronavirus cases worldwide and predicts the outcome of outbreak in the forthcoming days. Google Colab and an open source platform called Prophet are used to implement this model. We use Prophet for forecasting the time series data for predicting the outbreak of virus. Google Colab provides us Jupyter note book environment which is suitable for machine learning concepts and it is a free online cloud. Support Vector Machine, Regression and Data Visualization are the Machine learning concepts used for better outcome of the model. This model is all the source analysis of spread of virus with total, new, active cases with prediction of future out break and weekly study epidemic.

Keywords: Visualization, Outbreak, Prediction, Regression

List of Figures

TABLE OF CONTENTS

INTRODUCTION

In this era of automation, computing and knowledge science have necessary role within the health care business. These technologies square measure thus socially connected that medical professionals will simply manage their roles and patient care. All health care organizations exerting to develop an automatic system that may be accustomed settle for the challenges faced in health care. Scientists square measure functioning on machine learning (ML) to develop good solutions to diagnose and treat sickness. cubic centimeter is capable of police work sickness and virus infections additional accurately so patients' sickness may be diagnosed at an early stage, the harmful stages of diseases may be avoided, and there may be fewer patients. within the same manner, cubic centimeter may be accustomed change the task of predicting COVID-19 infection and facilitate forecast future infection tallies of COVID-19. In this paper, Coronavirus irruption prediction is completed by victimization the datasets obtained from Tableau code. The prediction of irruption of coronavirus is important so as to require safety measures during this pandemic and keep the increasing active cases. Not solely the person's health is degraded, increasing cases makes it hard to manage and treat every individual person because the government, scientists and officers don't seem to be still keen on studying the virus and manufacturing the vaccine. This model effectively forecasts the irruption in future and by analyzing the cases we will either take safety measures or go beneath internment because the initial cases were terribly venturesome. several scientists and universities are researching the traits of virus with their experiments.

1.2 Formulation of Problem

Corona Virus disease (COVID-19) is an infectious disease caused by a newly discovered virus, which emerged in Wuhan, China in December of 2019.

Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes, so you might have heard caution to practice respiratory etiquette (for example, by coughing into a flexed elbow).

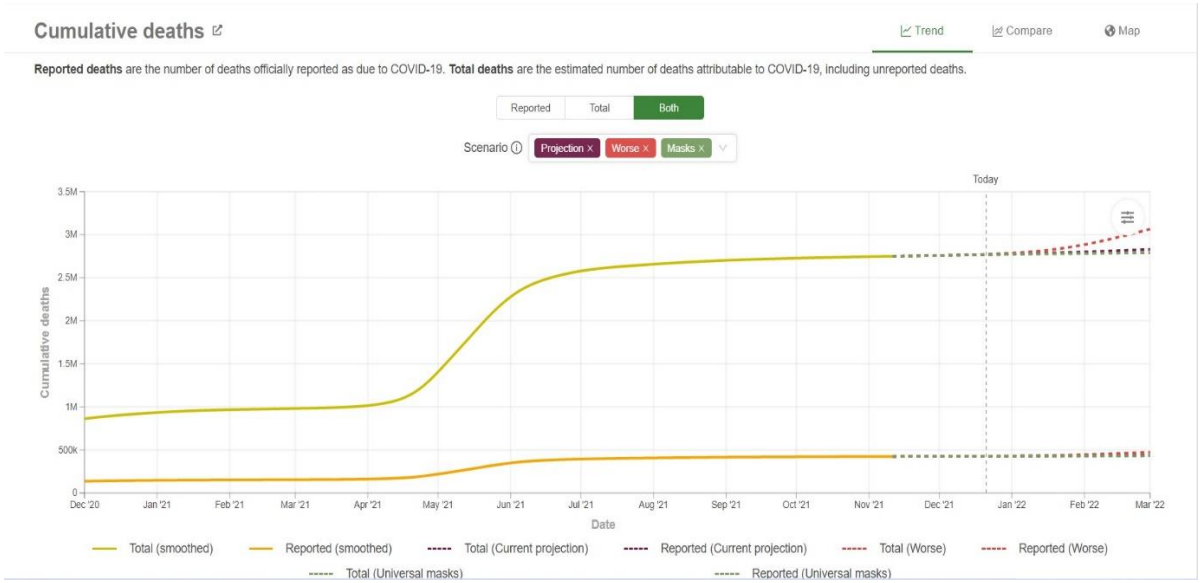
1.2.1 Tools and Technology Used

The machine learning algorithms play a large role in prediction or forecasting the long run outcomes or analyzing the things terribly accurately and expeditiously. during this model, we have a tendency to use the machine learning algorithms like Support Vector Machine, Regression and knowledge visual image. The bar graph plays a large half in representation of the model. This model is comparative analysis of cases in several countries and analysis of active cases and total cases in Indian states. The cured proportion and death proportion within the Indian states is analyzed. the overall cases and new cases area unit shown for India, Italy, Asian country and city. The worldwide novel coronavirus cases area unit analysed for recovered, death and confirmed and that we predict the long run outcome of occurrence. throughout this analysis and prediction we found that India is doing higher up to now in dominant the occurrence compared to alternative countries.

Forecasting

For forecasting through ML, time series analysis may be used, which is an important part of ML. It is a univariate type of regression in which the target feature (dependent feature) is forecast using only one input feature (independent feature), which is time [41e43]. It is used to forecast future event values, and it has an important role for forecasting the existence of respiratory diseases such as COVID-19. Positive cases are increasing daily, so it is necessary to forecast whether the ratio by which the number is increasing is continuing based on prior observations. It is helpful for the government, because based on the forecast, it can plan for resources to control the spread of disease and act for the future so that the growth rate of the infection decreases without affecting more people [30,32,35]. Forecasts depend completely on past trends, so forecast values cannot be guaranteed. However, this forecasted approximation of events may help authorities to assess forthcoming resource planning to compete with any pandemic situation such as COVID-19. We used the most widely used forecasting method, called the ARIMA model for time series forecasting. ARIMA is used for time series data to predict future trends [41e50]. ARIMA is a form of univariate regression analysis that predicts future values based on differences between values rather than actual values. It combines three terms: (1) autoregression (AR), showing changing a variable that revert values on the basis of its prior values; (2) integrated (I), or replacing data values on the basis of the difference between data values and previous values; and (3) moving average (MA), a successive average taken on successive time frames of constant size of a time series previously available.

Fig.1 shows the forecast for INDIA based on data up to Nov, 2021 is the future forecast.


Cumulative deaths
Trend
Compare
Map
Reported deaths are the number of deaths officially reported as due to COVID-19. Total deaths are the estimated number of deaths attributable to COVID-19, including unreported deaths.
Reported
Total
Both
Scenario
Projection x
Worse x
Masks x
Today
3.5M
3M
2.5M
2M
1.5M
1M
500k
0
Cumulative deaths
Dec '20
Jan '21
Feb '21
Mar '21
Apr '21
May '21
Jun '21
Jul '21
Aug '21
Sep '21
Oct '21
Nov '21
Dec '21
Jan '22
Feb '22
Mar '22
Date
Total (smoothed)
Reported (smoothed)
Total (Current projection)
Reported (Current projection)
Total (Worse)
Reported (Worse)
Total (Universal masks)
Reported (Universal masks)

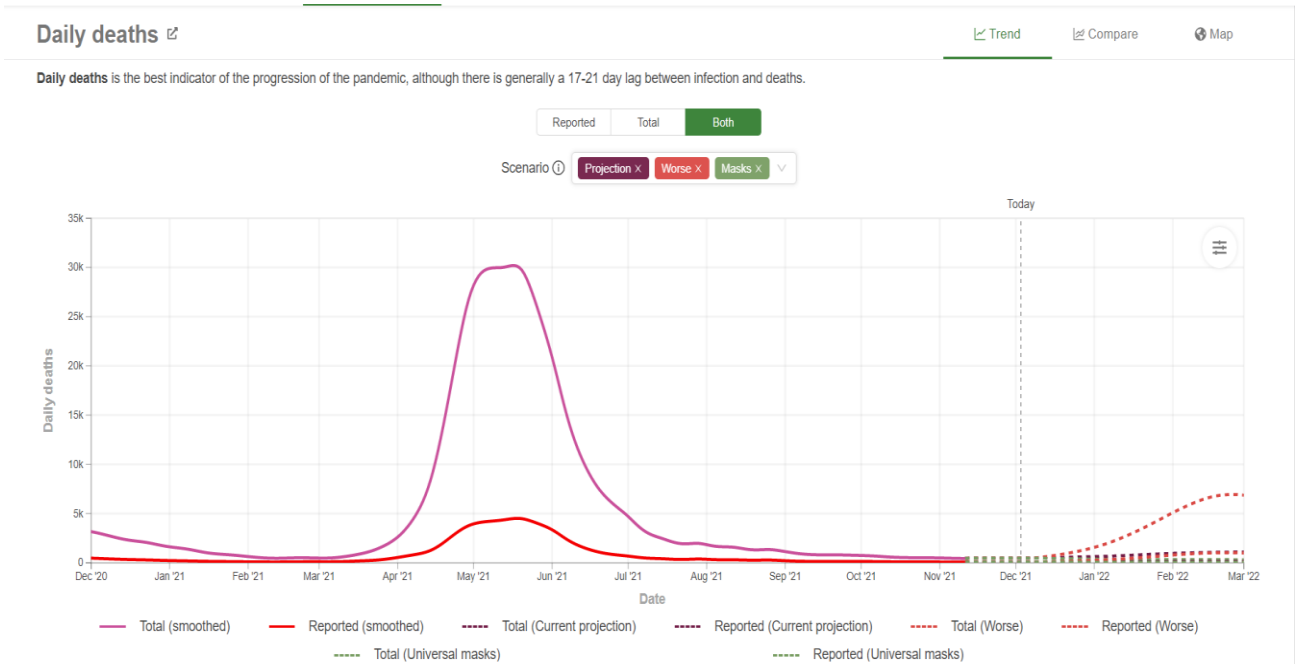

Daily deaths
Trend
Compare
Map
Daily deaths is the best indicator of the progression of the pandemic, although there is generally a 17-21 day lag between infection and deaths.
Reported
Total
Both
Scenario
Projection x
Worse x
Masks x
Today
35k
30k
25k
20k
15k
10k
5k
0
Daily deaths
Dec '20
Jan '21
Feb '21
Mar '21
Apr '21
May '21
Jun '21
Jul '21
Aug '21
Sep '21
Oct '21
Nov '21
Dec '21
Jan '22
Feb '22
Mar '22
Date
Total (smoothed)
Reported (smoothed)
Total (Current projection)
Reported (Current projection)
Total (Worse)
Reported (Worse)
Total (Universal masks)
Reported (Universal masks)

Fig.1

Literature survey

The model takes the available data which includes latitude, longitude, days, cases, deaths, positive outcomes from the datasets and validates it by training the model, then the hyper parameters are applied with data visualization, regression and support vector machine which by using prophet predicts the outbreak of cases that are confirmed, death and recovered. To find the correlation between each and every features heat map is used. Heat map uses the correlation matrix. We visualize the results in the form of graphs and histogram. The system architecture is shown here which takes us through the steps through which we predict the coronavirus outbreak. A strong model that predicts how the virus could spread across different countries and regions is required. The model that predicts the spread of the virus in the next days is our proposed system.

Use Case Diagram of Corona virus Outbreak prediction using M.L.

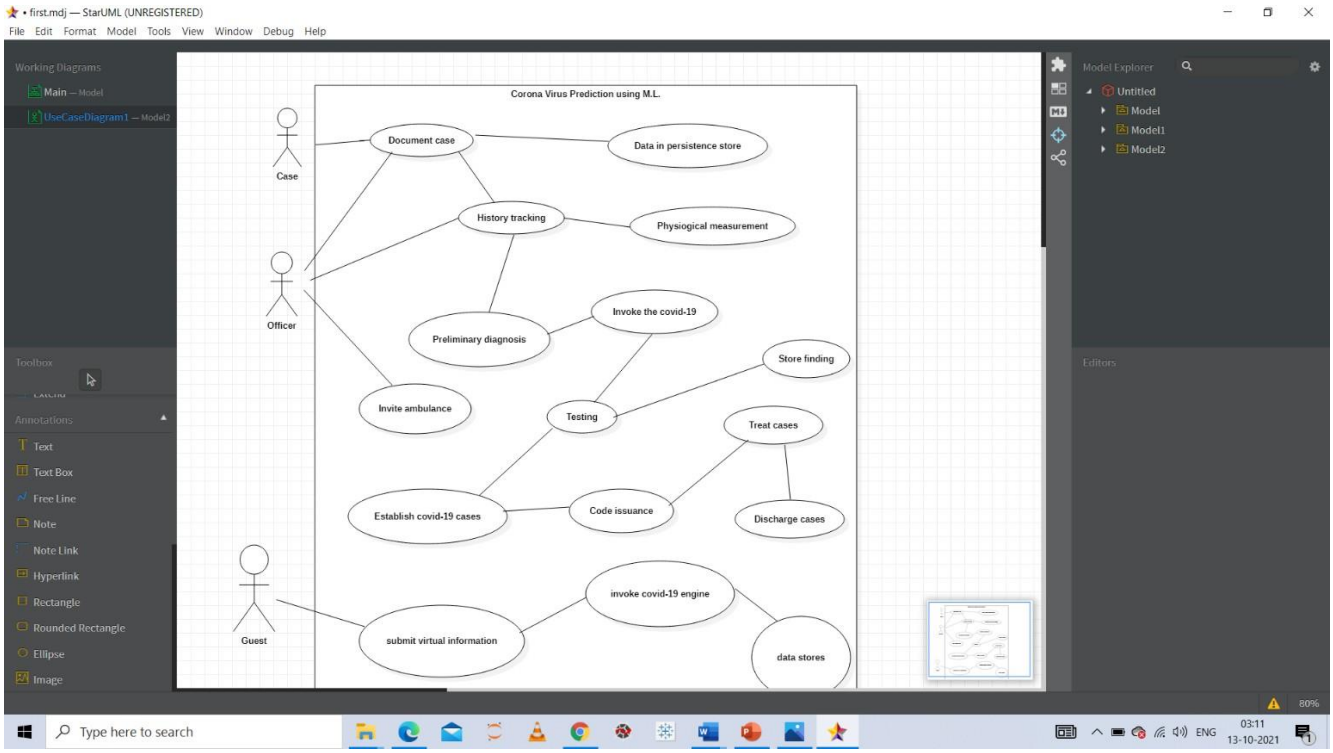


Fig.2

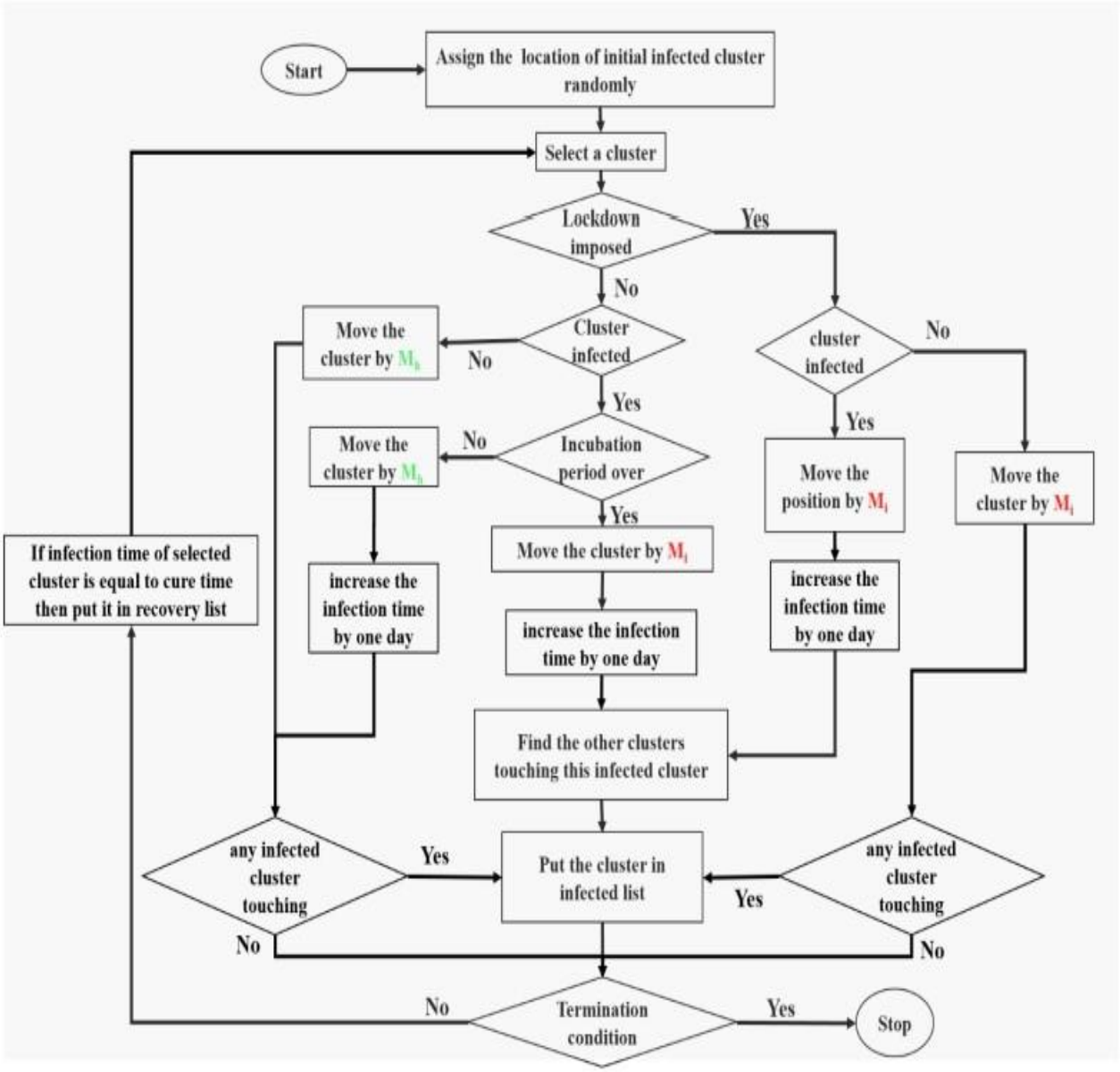


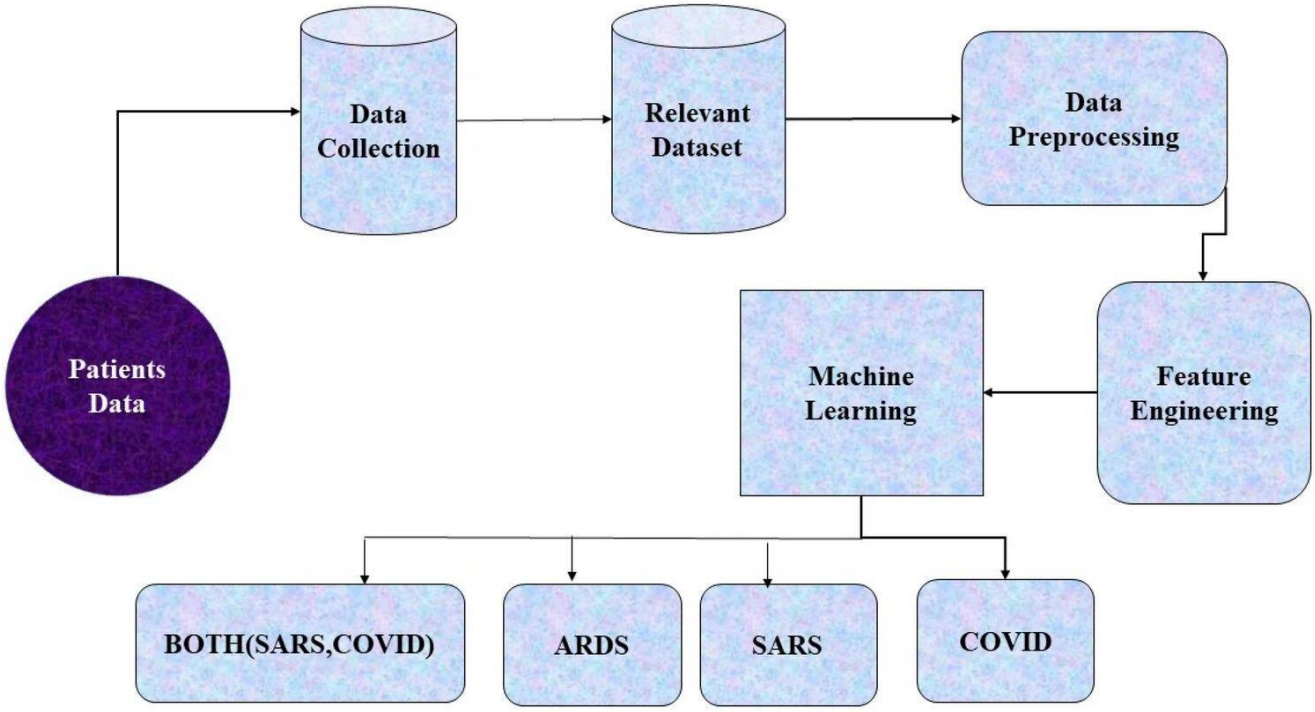
Fig.3

Flow Chart Diagram for Corona virus outbreak prediction using M.L.

Fig.4

UML diagram for corona virus outbreak prediction using M.L.

Data flow diagram



Hardware and Software Requirement specifications:

Software Requirements:

Name of component	Specification
Operating System	Windows 98, Windows XP, Windows7, Linux, Windows 10
Language	Python
Database	MySQL Server
Browser	Any of Mozilla, Opera, Chrome, etc
Web Server	Tomcat 7
Software Development Kit	Anaconda Navigator
Scripting Language Enable	Python
Database JDBC Driver	MySQL, Oracle

Hardware Requirements:

Name of component	Specification
Processor	Pentium III 630MHz
RAM	128 MB
Hard disk	20 GB
Monitor	15” color monitor
Keyboard	122 keys

Results

In this project we finally made the model that predicts the coronavirus outbreak and analyses the cases of virus in various area and time in various calculative forms.

The trend of coronavirus cases in India from figure shows us that the Trend of cases in India.

	State/UTs	Total Cases	Active	Discharged	Deaths	Active Ratio	Discharge Ratio	Death Ratio
0	Andaman and Nicobar	7620	9	7482	129	0.120000	98.190000	1.690000
1	Andhra Pradesh	2049314	11655	2023496	14163	0.570000	98.740000	0.690000
2	Arunachal Pradesh	54572	440	53856	276	0.810000	98.690000	0.510000
3	Assam	601787	4487	591432	5868	0.750000	98.280000	0.980000
4	Bihar	725952	55	716237	9660	0.010000	98.660000	1.330000
5	Chandigarh	65224	42	64363	819	0.060000	98.680000	1.260000
6	Chhattisgarh	1005325	285	991475	13565	0.030000	98.620000	1.350000
7	Dadra and Nagar Haveli and Daman and Diu	10670	0	10666	4	0.000000	99.960000	0.040000
8	Delhi	1438821	392	1413342	25087	0.030000	98.230000	1.740000
9	Goa	176318	830	172176	3312	0.470000	97.650000	1.880000
10	Gujarat	825916	156	815678	10082	0.020000	98.760000	1.220000
11	Haryana	770863	278	760711	9874	0.040000	98.680000	1.280000
12	Himachal Pradesh	218898	1793	213430	3675	0.820000	97.500000	1.680000
13	Jammu and Kashmir	329320	1388	323510	4422	0.420000	98.240000	1.340000
14	Jharkhand	348215	78	343002	5135	0.020000	98.500000	1.470000
15	Karnataka	2975067	12594	2924693	37780	0.420000	98.310000	1.270000
16	Kerala	4664944	144075	4495904	24965	3.090000	96.380000	0.540000
17	Ladakh	20795	73	20515	207	0.350000	98.650000	1.000000
18	Lakshadweep	10361	4	10306	51	0.040000	99.470000	0.490000
19	Madhya Pradesh	792519	115	781882	10522	0.010000	98.660000	1.330000
20	Maharashtra	6547793	40252	6368530	139011	0.610000	97.260000	2.120000
21	Manipur	120426	2207	116365	1854	1.830000	96.630000	1.540000
22	Meghalaya	81161	1693	78068	1400	2.090000	96.190000	1.720000
23	Mizoram	93660	16841	76512	307	17.980000	81.690000	0.330000
24	Nagaland	31219	365	30189	665	1.170000	96.700000	2.130000
25	Odisha	1025874	5099	1012583	8192	0.500000	98.700000	0.800000
26	Puducherry	126308	828	123640	1840	0.660000	97.890000	1.460000
27	Punjab	601600	290	584797	16513	0.050000	97.210000	2.740000
28	Rajasthan	954322	76	945292	8954	0.010000	99.050000	0.940000
29	Sikkim	31412	600	30425	387	1.910000	96.860000	1.230000
30	Tamil Nadu	2662177	17192	2609435	35550	0.650000	98.020000	1.340000
31	Telengana	665749	4620	657213	3916	0.690000	98.720000	0.590000
32	Tripura	84127	260	83054	813	0.310000	98.720000	0.970000
33	Uttar Pradesh	1709800	159	1686749	22892	0.010000	98.650000	1.340000
34	Uttarakhand	343530	209	335927	7394	0.060000	97.790000	2.150000
35	West Bengal	1568321	7580	1541963	18778	0.480000	98.320000	1.200000

Fig.6

In figure 7 The correlation matrix of dataset features are given

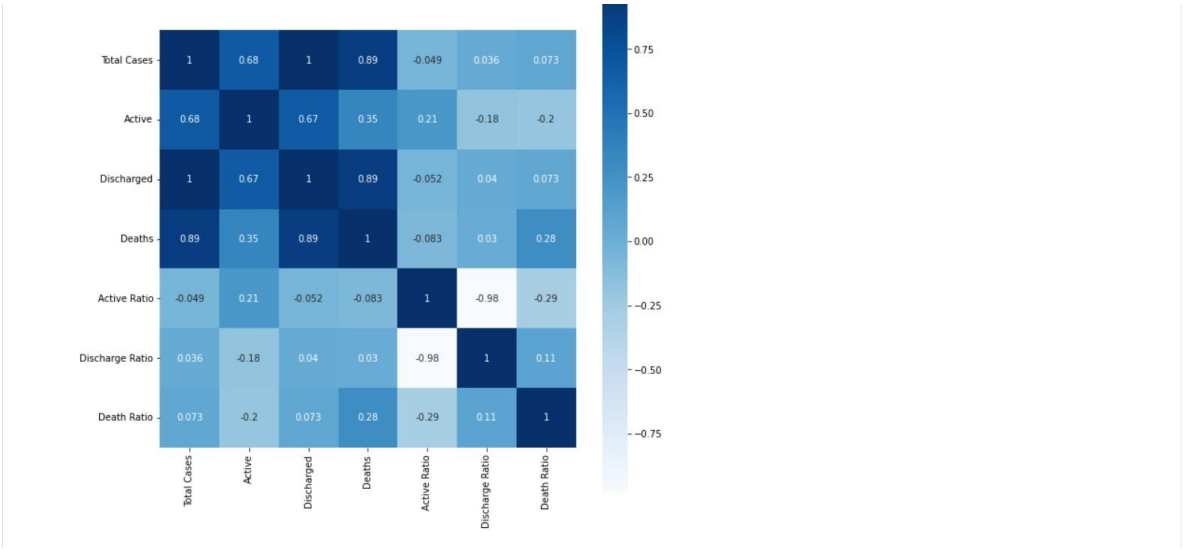


Fig.7

We can observe from fig 8 that according to reports Maharashtra and Kerala has recorded with highest number of active coronavirus cases.

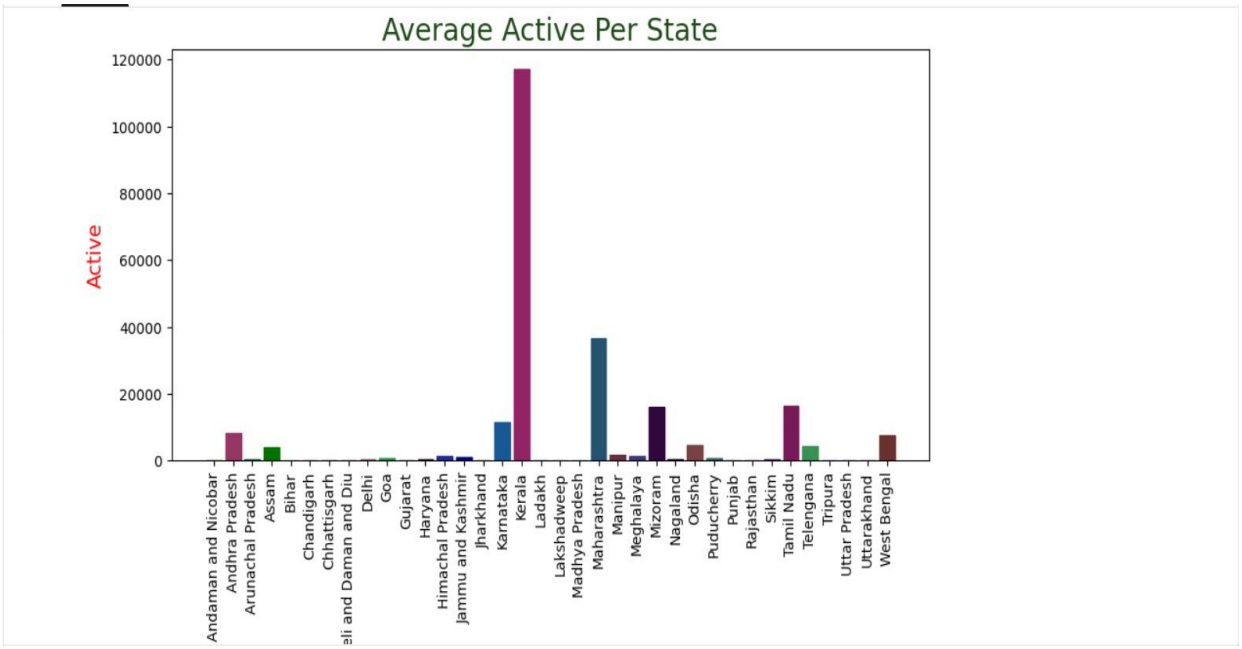
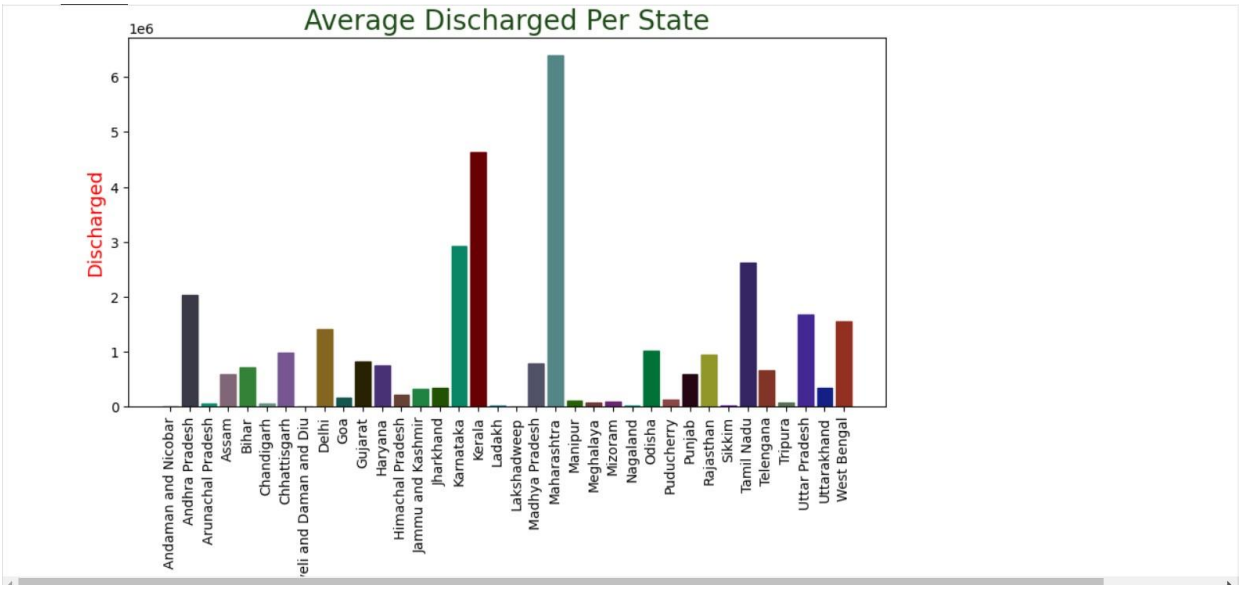


Fig.8

From fig 9 and fig 10, we can see that death cases are very few in comparison to discharge cases. From the reported cases, the rate of cured cases is high.



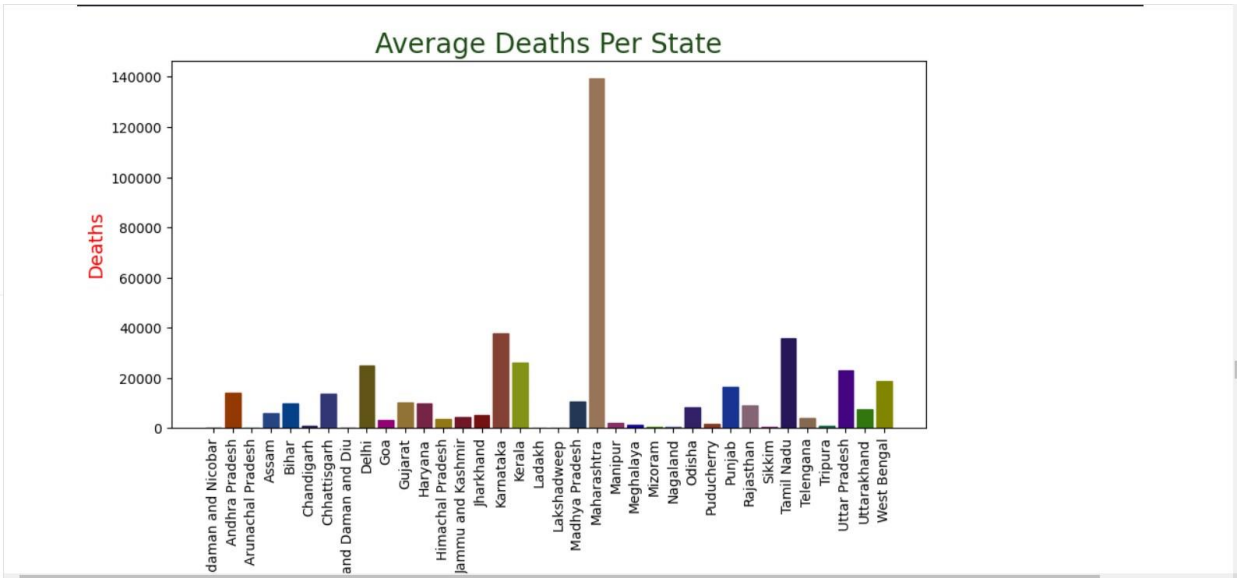
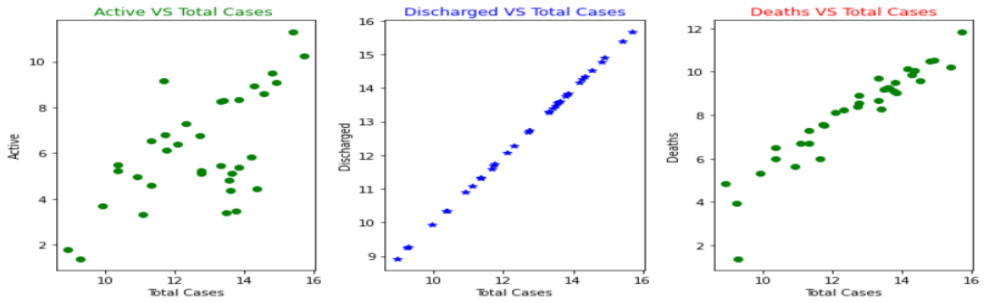


Fig.9&10



Conclusion

COVID-19 pandemic now appears to be serious infected spread disease like any other wide-spread diseases. Because of the rapid rise in the number of cases during pandemic, causes struggle in healthcare sector to identify suitable and appropriate treatment. Machine Learning methodologies are commonly used as alternative methods for classification and prediction. In this paper, Supervised Machine Learning classification models for COVID-19 virus was developed using KNN, SVM, Decision Tree and Random Forest models. All the above mentioned models are modeled using 80% data for training and 20% data for testing. The machine learning model developed with K-Nearest Neighbor performs better among all models employed for the COVID-19 data set analysis, in terms of accuracy of 98.34%, recall with 97% and F1 score with 0.97. In contrast SVM emerged to be the best model among all the models in term of precision with 97% and has less performance in other metrics compared to KNN. This study helps us to understand and stop the serious wide spread situation by utilizing the most common features that directly affects the spread of disease. This also serves the purpose of humans in predicting the disease that impacts the disease spread abnormalities and high recovery ratio.

The future work is to provide better significant results using several other machine learning models for finding estimates that helps the clinicians, medical and governmental organizations to look forward for real time preparations ahead at this sort of pandemic disease in future. Also Deep Learning models can be employed to predict the COVID cases

Reference

1. “A Tutorial on Support Vector Regression”, Alex J. Smola, Bernhard SScholkopf–Statistics and Computing archive Volume 14 Issue 3, August 2004,p.199-222.

2. Bishop, pattern recognition and machine learning, chapter 7Sparse Kernel Machines.

3. Chang and Lin, LIB SVM: A library for Support Vector Machines.

4.Crammerand Singer On the Algorithmic Implementation of Multi class Kernel-based Vector Machines , JMLR 2001.

5. Fan, Rong-En ,etal.,”LIB LINEAR: A library for large linear classification.”, Journal of machine learning research 9.Aug(2008):1871-1874.

6. Yogesh Kumaret.al “Machine Learning aspects and its application towards different research areas”.

7. Schölkopfet.al New Support Vector Algorithms.

8. Wu, Lin and Weng,” Probability estimates for multi-class classification by pair wise coupling”, JMLR5:975-1005, 2004.

9. S. M. Kissler et al., “Projecting the transmission dynamics of SARSCoV-2 through the post pandemic period,” Science, vol. 368, no. 6493, pp. 860– 868, 2020.