CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

Accredited with Grade A by NAAC
Accredited with Grade A by KCG

CERTIFICATE

This is to certify that the report entitled "**PREDICTION OF CASES AND HEALTHCARE FECILITY USING MACHINE LEARNING**" is a bonafied work carried out by **VANSH DESAI(19CE018)** under the guidance and supervision of **Prof MAYURI POPPAT** for the subject **Software Group Project – I (CE244)** of 3rd Semester of Bachelor of Technology in **Computer Engineering** at Chandubhai S. Patel Institute of Technology (CSPIT), Faculty of Technology & Engineering (FTE) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the supervision of,

Prof MAYURI POPPAT
U & P U. Patel Dept. of
Computer Engineering, CSPIT, FTE,
CHARUSAT, Changa, Gujarat

Dr. Ritesh Patel
Head - U & P U. Patel Department of Computer Engineering,
CHARUSAT, Changa, Gujarat.

Chandubhai S. Patel Institute of Technology (CSPIT)

Faculty of Technology & Engineering (FTE), CHARUSAT

At: Changa, Ta. Petlad, Dist. Anand, Pin: 388421. Gujarat.

A
Project Report
On
"PREDICTION OF CASES AND HEALTHCARE FECILITY USING MACHINE LEARNING"

Prepared by
19CE018 (VANSH DESAI)

Under the Supervision of

Prof. MAYURI POPAT

Submitted to

Charotar University of Science & Technology (CHARUSAT)
for the Partial Fulfillment of the Requirements for the
Degree of Bachelor of Technology
in Computer Engineering
CE244 – Software Group Project-I
of 3rd Semester of B.Tech CE

Submitted to

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY
Accredited with Grade A by NAAC
Accredited with Grade A by KCG

U & P U. PATEL DEPARTMENT OF COMPUTER ENGINEERING
Faculty of Technology & Engineering (FTE), CHARUSAT
Chandubhai S. Patel Institute of Technology (CSPIT)
At: Changa, Dist: Anand, Pin: 388421.

July-November, 2020

Table of Contents

ABSTRACT

Currently, the world is facing a health and economic crisis due to the spread of the virus SARS-CoV-2 which causes a disease referred to as COVID-19 . By the end of April 2020, the virus has spread to over 3.3 million people in worldwide and has killed over 230,000 . During this pandemic, governments and hospitals have struggled to allocate scarce resources, including tests, treatment in intensive care units (ICUs) and ventilators .

As the virus continues to spread, predicting hospitalizations, mortality, and other patient outcomes becomes important for several reasons: (i) using risk profiles to inform decisions on who should be tested (for the virus and/or antibodies)and at which frequency, (ii) providing more accurate estimates of who is more likely to be hospitalized and the type of care they may need, (iii) informing plans for staffing, resources, and prioritizing the level of care in extremely resource-constrained settings. Equally importantly, as societies adapt to the pandemic, predictive models can (i) assess individual risk so that social distancing measures can transition from “blanket” to more targeted (e.g., deciding who can return to work, who is advised to stay at home, who should be tested, etc.) and (ii) direct policy decisions on who should receive priority for vaccination, which will be critical as initial vaccine production may not suffice to vaccinate everybody.

The main aim of project is to developed model which predict ventilators and intubation condintion we are able to devlop good models

Acknowledgement

With immense pleasure, i would like to present this report on the project report of "**PREDICTION OF CASES AND HEALTHCARE FECILITY USING MACHINE LEARNING**". I express our gratitude to this college for providing us the proper resources and environment for the partial completion of our project.

I have received good support, motivation, guidance & encouragement from many people. First of all I would like to thank **verzeo machine learning internship for knowledge**. Then I would like to thank **Prof. Mayuri Popat** , my project advisor for guiding me in each & every step in the process of my project by giving me guidelines and by sharing his knowledge with me. Thank you for your advice, guidance and assistance.

CHAPTER 1: INTRODUCTION

1.1 PROJECT SUMMARY

To contribute to the current human crisis our attempt in this study is to develop a forecasting system for COVID-19. This System work in two part. First part is the forecasting is done for the two important variables of the disease for the coming 10 days:
1) the number 0f New confirmed cases.
2) the number of recoveries
In second part we predict Outcomes such as admission to an intensive care unit (ICU), invasive ventilation, death and greater disease severity have been studied in the context of COVID-19 disease.

1.2 OBJECTIVE

The findings of these studies suggested that the majority of patients who had moderate to severe respiratory failure required invasive mechanical ventilation and patients with severe illness required admission to the ICU. Furthermore, patients with comorbidity are at higher risks for poor outcomes: respiratory failure, cardiovascular diseases, diabetes, and kidney injury seem to be highly associated with the intubation of patients with COVID-19

1.3 SCOPE

Machine learning in medicine has recently made headlines. Google has developed a machine learning algorithm to help identify cancerous tumors on mammograms. Stanford is using a deep learning algorithm to identify skin cancer. A recent JAMA article reported the results of a deep machine-learning algorithm that was able to diagnose diabetic retinopathy in retinal images. It's clear that machine learning puts another arrow in the quiver of clinical decision making.

1.4 ABOUT DATASET

1.4.1 ABOUT DATASET FIRST PART

The aim of this study is the future forecasting of COVID-19spread focusing on the number of new positive cases, the number of deaths, and the number of recoveries. This dataset is obtained by Kaggle. This dataset has information from the states and union territories of India at daily level.

1.4.2 ABOUT DATASET SECOND PART

We use a dataset that has been open for the general public by the Mexican Government (and updated daily). These data include information about every person who has been tested for SARS-CoV-2 in Mexico. They include demographic information such as: Age, Location, Nationality, the use of an indigenous language; as well as information on pre-existing conditions, including whether the patient has: diabetes, chronic obstructive

pulmonary disease (COPD), asthma, hypertension, obesity, pregnancy, chronic renal failure, other prior diseases, and whether was or is using tobacco. In addition, the data report the dates on which the patient first noticed symptoms, the date when the patient arrived to a care unit, Finally, it contains fields showing the result of the SARS-CoV-2 test, weather the patient was hospitalized, has pneumonia, needed a ventilator, and if she/he was treated in an ICU.

CHAPTER 2: PROJECT MANAGEMENT

2.1 PROJECT PLANNING

If we want to build good ML model the data is appropriate that's why we use Kaggle dataset which has less null value and build like that if we do changes in dataset, we don't do major changes in model. The flow of machine learning project is

2.1.1 Gathering data

1st part data has no null values in it is pre-processed dataset so no pre-processing done in dataset.2nd part has lots of pre-processing 80% data has null value we remove that kind of column which has more null value like pregnancy have 78% value is null so I remove that column. And we have total 1048575 values. This is the information of table.

sex 0
entry_date 0
date_symptoms 0
intubed 828270
pneumonia 15
age 125
diabetes 3216
copd 2840
asthma 2823
hypertension 2977
other_disease 4847
cardiovascular 2918
obesity 2864
renal_chronic 2861
tobacco 3039
result 0
icu 828279

2.1.2 Data pre-processing

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.

We will remove row with dropna() function and make two datasets

1.Person who goes to icu or need intubation

2.Person who positive for covid

2.1.3 Researching the model that will be best for the type of data

Here I observe one problem Class imbalance

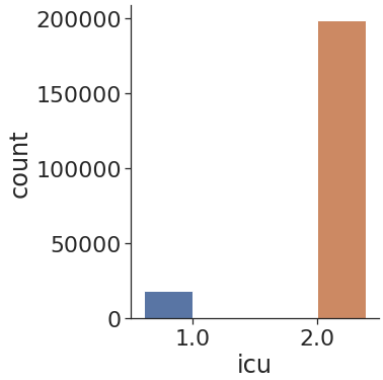


Fig 1: count verses icu

In figure one there are very less people who need icu in dataset. So here 1 is minority class and 2 is majority class so I use us method from imblearn library called SMOTE which help to solve this problem if I don't solve that problem algoritham can underestimate and give result as any one class.

2.1.4 Training and testing the model

In sklearn.preprocessing library feature called MinMaxScaler is used for data fitting.We use that library because when we mix categorical and normal variable we have convert normal variable into categorical variable minmaxscaler help to convert on hot encoding.

From sklearn.model_selection library train_test_split is feature Used to splitting in test and train.

2.1.5 Evaluation

In this we use six algorithams for evoluations

1) K Neighbors Classifier
2) Random Forest Classifier
3) Decision Tree Classifier
4) Gaussian Naive Bayes
5) Support vector machine

CHAPTER 3: SYSTEMS REQUIREMENTS STUDY

3.1 HARDWARE REQUIREMENTS

1)laptop with good performance

3.2 SOFTWARE REQUIREMENTS

1) ANACONDA PAYTHON 3.7
2) GOOGLE COLAB
3) FBPROPHET LIBRARY
4) MATPLOTLIB LIBRARY
5) IMBLERN LIBRARY

CHAPTER 4: SYSTEMS ANALYSIS

4.1 STUDY OF CURRENT SYSTEM

4.1.1 ANALYSIS OF DATA

Our data is like that

	sex	pneumonia	age	diabetes	copd	asthma	hypertension	other_disease	cardiovascular	obesity	renal_chronic	tobacco	result	gap_of_date	icu	intubed
1	1	1.0	69.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1	6.0	2.0	2.0
3	1	1.0	79.0	1.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	2.0	1	135.0	2.0	2.0
4	2	1.0	71.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1	5.0	2.0	2.0
7	1	1.0	61.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1	61.0	2.0	2.0
21	2	1.0	82.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	1	1.0	2.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1048537	2	1.0	80.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	2.0	1	0.0	2.0	2.0
1048540	2	2.0	46.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1	0.0	2.0	2.0
1048541	1	2.0	55.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1	141.0	2.0	2.0
1048550	2	1.0	73.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	1	1.0	2.0	2.0
1048572	1	1.0	67.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	1	5.0	2.0	2.0

217425 rows × 16 columns

Fig:4.1 Fedding Dataset

We have data 'sex','pneumonia', 'age', 'diabetes', 'copd', 'asthma', 'hypertension','other_disease','cardiovascular', 'obesity', 'renal_chronic', 'tobacco', 'result','gap_of_date','icu','intubed' variable in dataset.

4.1.2 CORELATIONS BETWEEN VARIABLE

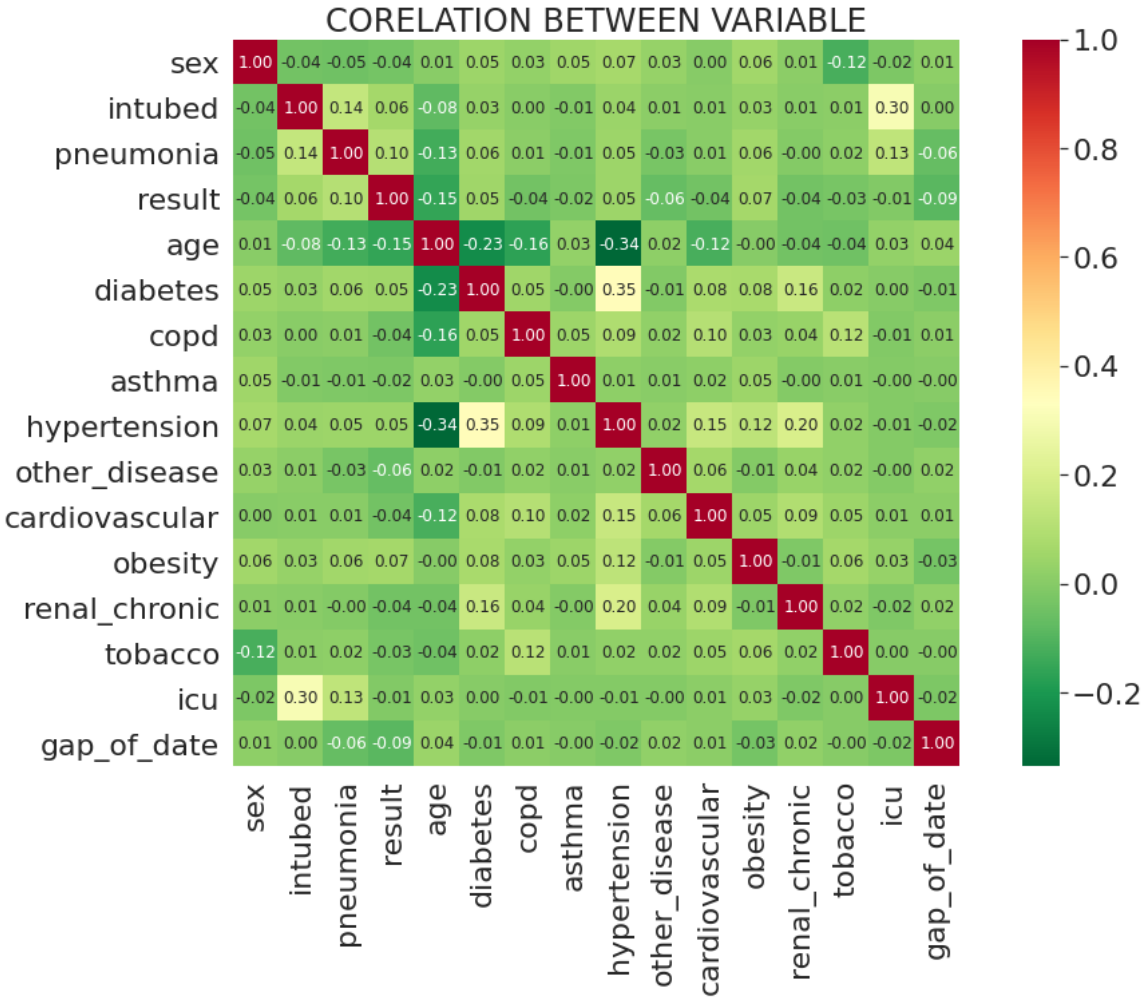


Fig:4.2: CORELATION MATRIX

CORELATION MATRIX is relation between two variables here we can see that diabetes and hypertension is highly corelated. Minus corelation means inversely corelated. Correlation can be an important tool for feature engineering in building machine learning models. Predictors which are uncorrelated with the objective variable are probably good candidates to trim from the model (shoe size is not a useful predictor for salary). In addition, if two predictors are strongly correlated to each other, then we only need to use one of them (in predicting salary, there is no need to use both age in years, and age in months). Taking these steps means that the resulting model will be simpler, and simpler models are easier to interpret.

4.1.3 CHI2 TEST FOR FEATURE SELECTION

Feature selection is an important problem in machine learning, where we will be having several features in line and have to select the best features to build the model. The chi-square test helps you to solve the problem in feature selection by testing the relationship between the features.

1) Steps to perform the Chi-Square Test:
2) Define Hypothesis.
3) Build a Contingency table.
4) Find the expected values.
5) Calculate the Chi-Square statistic.
6) Accept or Reject the Null Hypothesis

	feature_of_intubed	Score
0	sex	32.696564
1	pneumonia	470.264337
2	age	4852.142843
3	diabetes	18.423871
4	copd	0.105983
5	asthma	0.178373
6	hypertension	37.631617
7	other_disease	0.546692
8	cardiovascular	0.294525
9	obesity	10.423235
10	renal_chronic	0.441706
11	tobacco	0.458553
12	result	87.090254

Fik:4.3 chi2 of intubation variable

	feature_score_of_icu	Score
13	gap_of_date	9469.100606
14	icu	8947.557152
2	age	1078.222790
1	pneumonia	638.786040
9	obesity	24.859495
0	sex	16.056936
12	result	5.241762
10	renal_chronic	1.495859
6	hypertension	1.017907
8	cardiovascular	0.450435
3	diabetes	0.364181
4	copd	0.266274
7	other_disease	0.082229
5	asthma	0.027608
11	tobacco	0.008018

Fig:4.4 chi2 of icu variable

Here we can see that different variable important for different features. from that model select for model

4.1.4 RESULTS OF ALGORITHAMS

1) Intubed

1) K Neighbors Classifier

```
          1     2     3     4     5     6     7     8     9    10    11    12    13
Accuracies for different values of n are: [0.75866391 0.6883868  0.81218811 0.79317006 0.83038979 0.82338737
 0.83732321 0.83527653 0.84027826 0.83788663 0.84106014 0.84041624
 0.84241693 0.84242842 0.84264689]
KNearestNeighors performs best at n = 15  with a accuracy of  0.8426468897320916
F1 score : 0.5950889886600645
array([[ 2641,  2040],
       [11645, 70644]])
```

Fig 4.5 result knn for Intubed Variable

2) Random Forest Classifier

```
Random Forest Algorithm Accuracy Score : 84.13%
F1 score : 0.5922496518176814
array([[ 2599,  2115],
       [11687, 70569]])
```

Fig 4.6 result RFC for Intubed Variable

3) Decision Tree Classifier

```
Decision Tree Test Accuracy 83.91%
f1 score: 0.6007277371520456
array([[ 2888,  2593],
       [11398, 70091]])
```

Fig 4.7 result dtc for Intubed Variable

4)Naïve bayes

```
Accuracy of Naive Bayes: 84.06%
f1 score :  0.6362859514146071
array([[ 3960,  3533],
       [10326, 69151]])
```

Fig 4.8 result naïve bayes for Intubed Variable

5)Support vector machine

```
Test Accuracy of SVM Algorithm: 84.61%
F1 score: 0.6310007417876187
array([[ 3591,  2687],
       [10695, 69997]])
```

Fig 4.9 result svm for Intubed Variable

6) result of fbprphet for case prediction

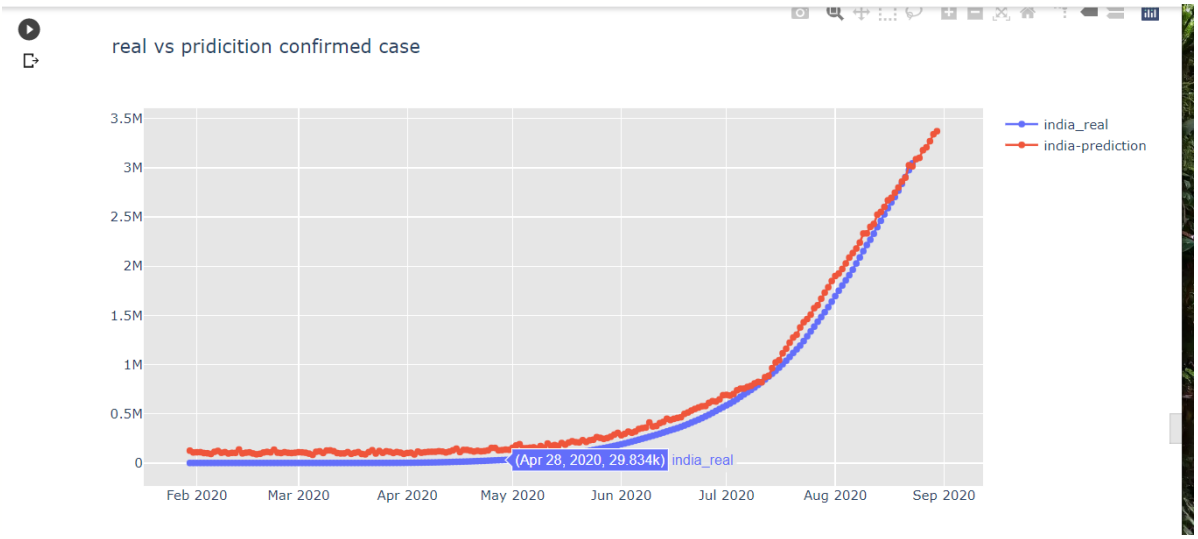
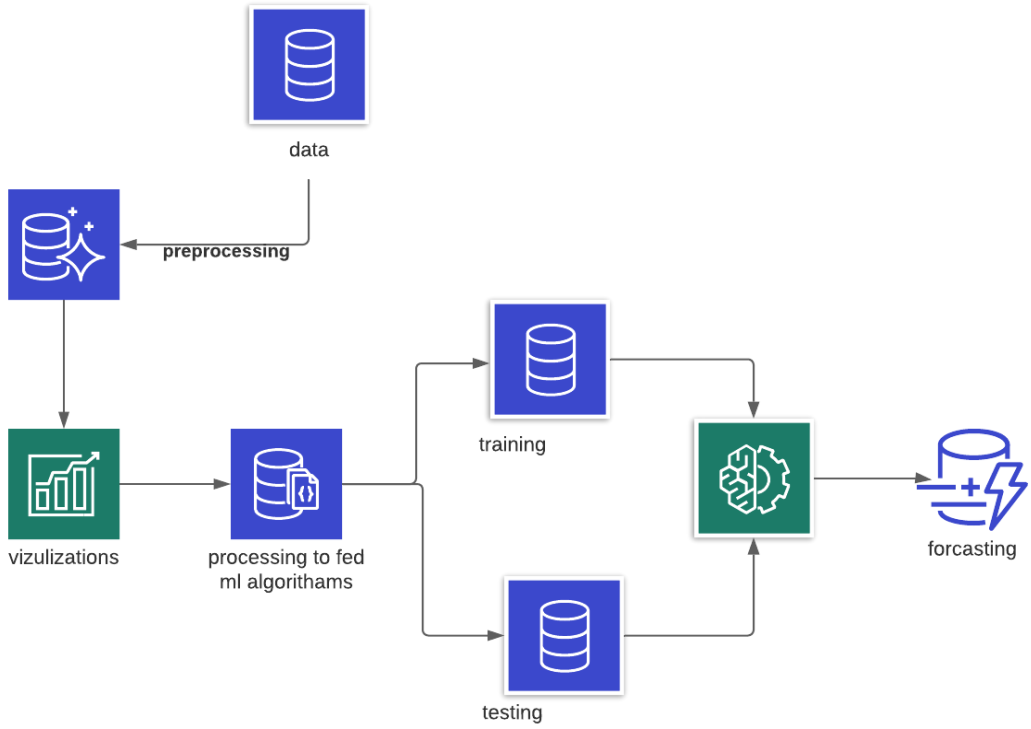


Fig 4.10 fbprophet case pridiction

4.2 FLOWCHART



CHAPTER 5: LIMITATIONS AND FUTURE ENHANCEMENT

5.1 LIMITATIONS

- The Model is not perfectly accurate
- To removing null value very complicated
- Data must be very accurate
- To train data is very difficult

5.2 FUTURE ENHANCEMENTS

Machine learning is future grooming technology. That kind of model help us fight against epidemic. That kind of model increase the surveillance of epidemic. The essence of public health surveillance is the use of data to monitor health problems to facilitate their prevention or control. Data, and interpretations derived from the evaluation of surveillance data, can be useful in setting priorities, planning, and conducting disease control programs, and in assessing the effectiveness of control efforts. For example, identifying geographic areas or populations with higher rates of disease can be helpful in planning control programs and targeting interventions, and monitoring the temporal trend of the rate of disease after implementation of control efforts.

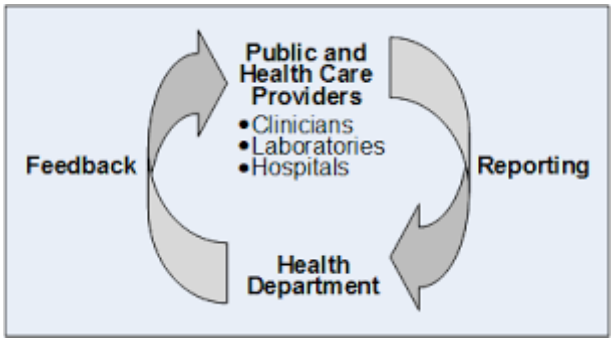


Fig 5.1 surrvielence method of epidemiology

CHAPTER 6: CONCLUSION AND DISCUSSION

6.1 Self-Analysis of Project Viabilities

There are some good statistical method to analysis project viabilities.

6.1.1 F1 SCORE

Here F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

F1 = 2 * (precision * recall) / (precision + recall)

Algoritham	F1 SCORE
1) K Neighbors Classifier	0.5950889886600645
2) Random Forest Classifier	0.5922496518176814
3) Decision Tree Classifier	0.6007277371520456
4) Gaussian Naive Bayes	0.6362859514146071
5) Support vector machine	0.6310007417876187

TABLE 6.1 INTUBED F1 SCORE

6.1.2 CONFUSION MATRIX

n the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes

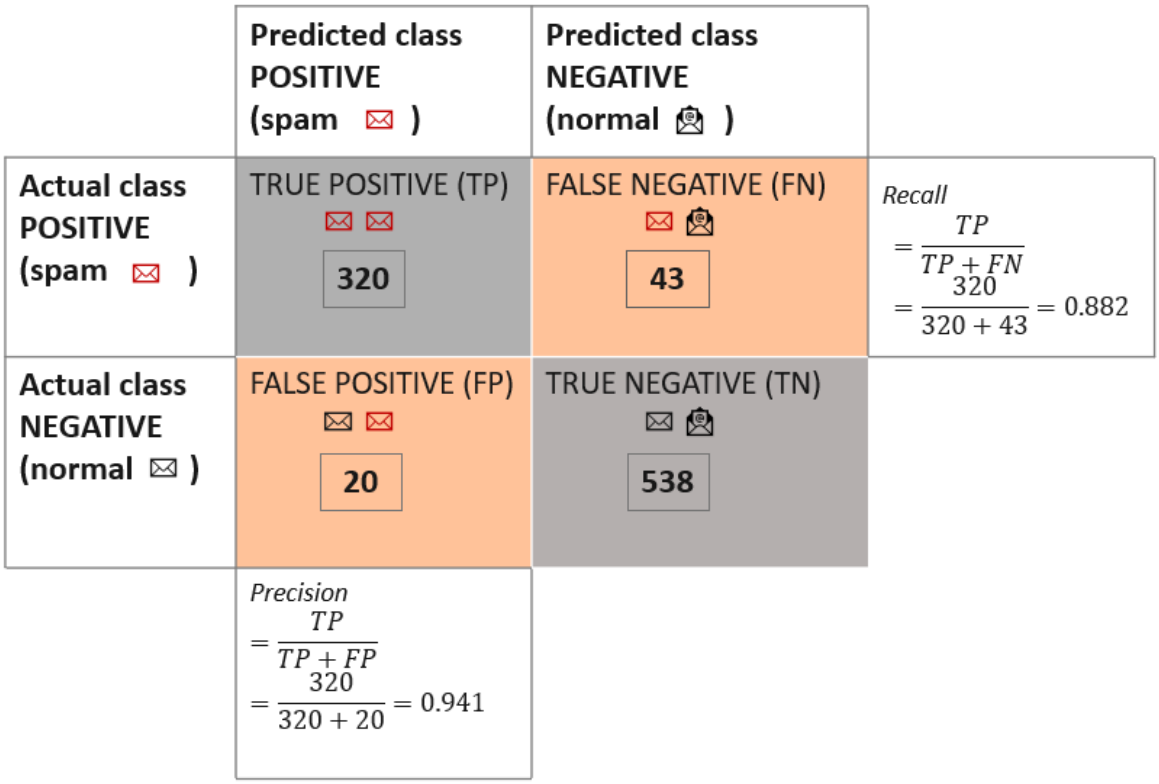


FIGURE 6.1 CONFUSION MATRIX

Algoritham	CONFUSION MATRIX
1) K Neighbors Classifier	[2641, 2040], [11645, 70644]
2) Random Forest Classifier	[2599, 2115], [11687, 70569]
3) Decision Tree Classifier	[2888, 2593], [11398, 70091]
4) Gaussian Naive Bayes	[3960, 3533], [10326, 69151]
5) Support vector machine	[[3591, 2687], [10695, 69997]]

TABLE 6.1 CONFUSION MATRIX OF INTUBED

6.2 SUMMARY OF PROJECT WORK

I develop models to identify the medical risk of a patient with (or suspected for) COVID-19. We hope this work canhelp hospitals and policymakers to distribute more effectively their limited resources including tests, ICU beds andventilators, as well as, to motivate countries and healthcare systems to standardize and share data with the medical informatics community.

CHAPTER 7: SCREENSHOTS

Fig 7.1 visualise case geographically

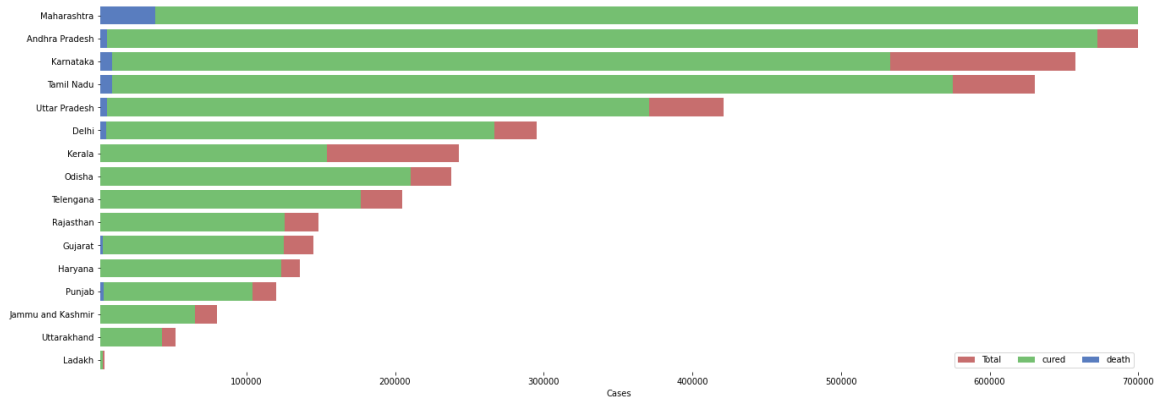


Fig 7.2 case overall india

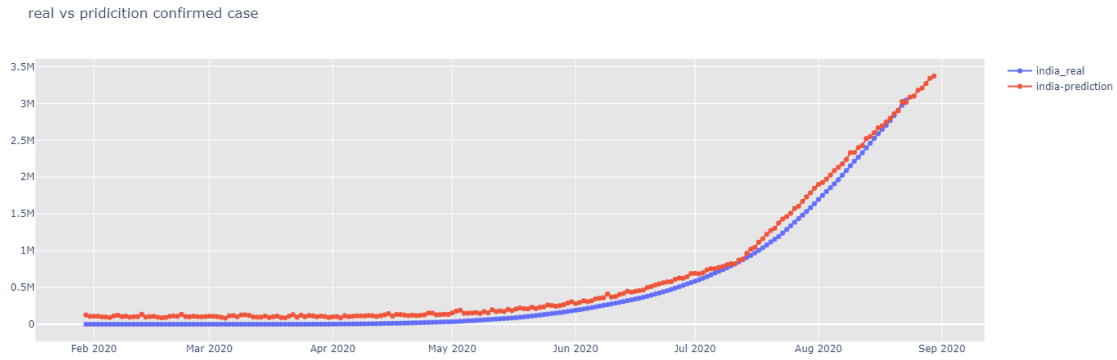


Fig 7.3 real vs predicted confirmed case

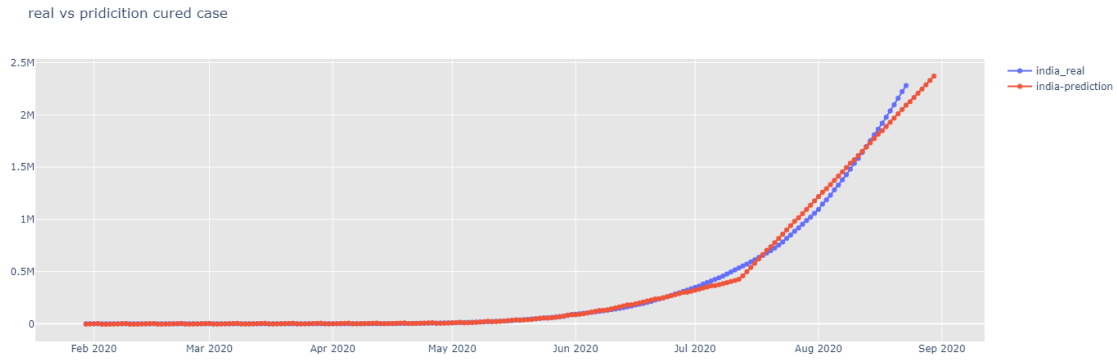


Fig 7.4 real vs predicted cured case

```
      0         1         2         3         4         5         6         7         8
Accuracies for different values of n are: [0.83656433 0.7645165  0.83540301 0.79296309 0.82050132 0.78790387
 0.80894561 0.77852133 0.79625158 0.7696792 ]
KNearestNeighors performs best at n = 1  with a accuracy of  0.8365643325284581
F1 score : 0.5634904338690149
array([[ 1984,  8708],
       [ 5506, 70772]])
```

Fig 7.5 result RFC for ICU Variable

```
Decision Tree Test Accuracy 86.13%
f1 score: 0.5803631430636049
array([[ 2103,  7650],
       [ 5387, 71830]])
```

Fig 7.6 result DTC for ICU Variable

```
confusion_matrix(y_pred, y_test)

Random Forest Algorithm Accuracy Score : 86.13%
F1 score : 0.5905702473405907
array([[ 2094,  6670],
       [ 5396, 72810]])
```

Fig 7.7 result RFF for ICU Variable

```
Accuracy of Naive Bayes: 79.81%
f1 score :  0.6064358658427799
array([[ 4358, 14424],
       [ 3132, 65056]])
```

Fig 7.7 result Naïve bayes for ICU Variable

CHAPTER 8: REFRENCES

[1] WHO announces COVID-19 outbreak a pandemic, (2020).
[2] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect.
Dis. (2020). https://doi.org/10.1016/S1473-3099(20)30120-1.
[3] COVID-19 Global Cases by Johns Hopkins University, 2020. https://www.gisaid.org/epiflu-applications/globalcases-
covid-19/.
[4] At the Top of the Covid-19 Curve, How Do Hospitals Decide Who Gets Treatment? - The New York Times, (n.d.).
https://www.nytimes.com/2020/03/31/us/coronavirus-covid-triage-rationing-ventilators.html (accessed April 29, 2020).
[5] The Hardest Questions Doctors May Face: Who Will Be Saved? Who Won't? - The New York Times, (n.d.).
https://www.nytimes.com/2020/03/21/us/coronavirus-medical-rationing.html (accessed April 29, 2020).
[6] Z. yong Huang, S. Lin, L. li Long, J. yang Cao, F. Luo, W. cheng Qin, D. ming Sun, H. Gregersen, Predicting the
morbidity of chronic obstructive pulmonary disease based on multiple locally weighted linear regression model with
K-means clustering, Int. J. Med. Inf. 139 (2020) 104141.
https://doi.org/10.1016/j.ijmedinf.2020.104141