ANALYSIS OF COVID-19 BASED UPON SYMPTOMS

A project report submitted in partial fulfillment of the requirements for the award of the degree of

MASTER OF COMPUTER APPLICATIONS
(CA 656)

BY

Meghaj Kumar Mallick(MCA/25017/18)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

BIRLA INSTITUTE OF TECHNOLOGY, MESRA
JAIPUR CAMPUS, JAIPUR
SP-2021

DECLARATION CERTIFICATE

This is to certify that the work presented in the project entitled "Analysis of COVID-19 based upon the Symptoms**"** in partial fulfillment of the requirement for the award of Degree of Master of Computer Applications of Birla Institute of Technology, Mesra, Ranchi, Extension Center Jaipur is an authentic work carried out under our supervision and guidance.

To the best of my knowledge, the content of this project does not form a basis for the award of any previous degree to anyone else.

Date: 04/05/2021

Meghaj Kumar Mallick (MCA/25017/18)

Name of Supervisor and Signature:

Dr. Madhavi Sinha
Associate Professor & In-charge,
Department of Computer Science & Engineering
Birla Institute of Technology, Mesra, Jaipur Campus

Name of Supervisor and Signature:

Dr. Piyush Gupta
Associate Professor,
Department of Computer Science & Engineering
Birla Institute of Technology, Mesra, Jaipur Campus

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to Dr. Piyush Gupta (Project Coordinator) Sir & to Dr. Madhavi Sinha (H.O.D, Computer Science.BIT Jaipur Campus) Mam, for their giudence and support to my project "Analysis of COVID-19 Based upon the symptoms" at Birla Institute of Technology, Mesra Jaipur Campus .They help me in completing this project.

I would also like to extend my gratitude to my parents and my friends who helped me in completing my project within the given time

Meghaj Kumar Mallick (MCA/25017/18)

INDEX PAGE

1. ABSTRACT

The novel corona virus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. This project is based upon the analysis & prediction of the COVID-19 (Corona Virus) disease by their symptoms.

As this COVID-19 is spread from person to person, Artificial intelligence based electronic devices can play a pivotal role in preventing the spread of this virus. As the role of healthcare epidemiologists has expanded, the pervasiveness of electronic health data has expanded too .

The increasing availability of electronic health data presents a major opportunity in healthcare for both discoveries and practical applications to improve healthcare. This data can be used for training machine learning algorithms to improve its decision-making in terms of predicting diseases.

This project describes the application of machine learning. We will use various machine leaning algorithm such as logistic regression, random forest, & naive bayes algorithm

In this process, we propose a machine-learning model that predicts a positive COVID-19 infection by asking basic questions that are based upon the symptoms of the disease i.e. fever, difficulty in breathing, dry cough etc. We are also going to use the ensemble technique of bagging approach to improve the final result. Finally we can predict the result on a GUI.

2. INTRODUCTION

The outbreak of the novel corona virus in early December 2019 in the Hubei province of the People's Republic of China has spread worldwide. This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected.

Corona viruses are a large family of viruses that are known to cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome(MERS) and Severe Acute Respiratory Syndrome(SARS) . These two diseases are spread by the corona viruses named as MERS-CoV and SARS-CoV. SARS was first seen in 2002 in China and MERS was first seen in 2012 in SaudiArabia. The latest virus seen in Wuhan, China is called SARS-COV-2 and it causes corona virus.

In this project we will use machine learning approach to identify the symptoms provided by the users. This entire process is done by collecting the data from the user. These data will help to indentify whether any person is suffering from COVID-19 or not, which is based upon some predefined standard symptoms. These symptoms are based on the guidelines given by the World Health Organization (WHO) & the Ministry of Health and Welfare, India.

In this project we will get to know about the dataset contains seven major variables that will bring an impact on whether any person is suffering from corona or not
.

- **Country**: List of the countries a person has visited.
- **Age**: Classification of the age group for each person, based on WHO age standard group.
- **Symptoms**: According to WHO there are five major symptoms such as Fever, Tiredness, Difficulty in Breathing, Dry Cough & Sore throat.
- **Other Symptoms:** Other symptoms include Pain, Nasal Conjection etc.
- **Severity**: The level of severity, Mild, Moderate & Severe.
- **Contact**: Whether person came to contact with a COVID-19 patient.

We need two kinds of data in **csv** (comma separated values) format such as **raw data & cleaned data**. In raw data it contains all possible labels of variables, which is used to generate cleaned data. The cleaned data contains all possible from raw data, which can be used for analysis. The cleaned data might contain some dummy variables.

3. MOTIVATION

We all have been affected by the current COVID-19 pandemic. However, the impact of the pandemic and its consequences are felt differently depending on our status as individuals and as members of society.

Research is continuing to find a cure for this disease while there is no exact reason for this outbreak. As the number of cases to test for Corona virus is increasing rapidly day by day, it is impossible to test due to the time and cost factors.

Thus, we need to create an application that could analysis this disease & it will help to save life of many persons. The project will focus upon the analysis & prediction of COVID-19 which based upon the symptoms.

The main goal of this thesis is to develop a machine learning model that could predict whether a patient is suffering from COVID-19. To develop such a model, a literature study alongside an experiment is set to identify a suitable algorithm. To assess the features that impacts the prediction model.

4. AIM

- The aim is to provide a machine learning model that can easily analysis the symptoms of COVID-19.
- The prediction is performed using the clinical information of the patients.
- The goal is to identify whether a patient can potentially be diagnosed with COVID-19.

5. OBJECTIVE

- The objective of this project is to easily identify the COVID-19 disease by its symptoms.
- It will help to reduce the risk of getting affected. By using the machine leaning model we will try to indentify the symptoms of this disease.
- Prediction of COVID-19 by using Machine Learning could help increase the speed of disease identification resulting in reduced mortality rate

6. SOFTWARE REQUIRMENT SEPCIFICATION

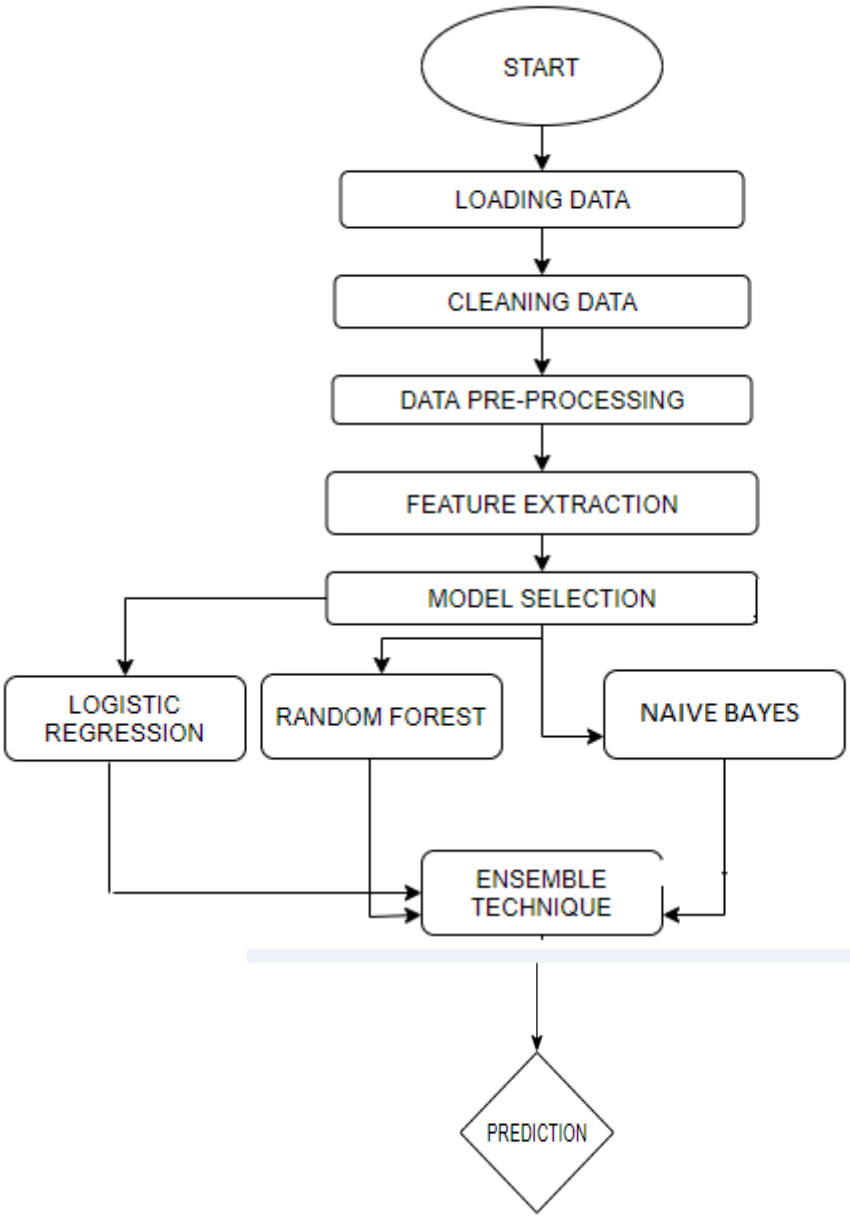
6 (A) SOFTWARE & HARDWARE REQUIREMENT

- Package Requirement : Tinker, NumPy, Scikit Learn, Matplot Library
- Platform: Jupyter-Notebook, Anaconda, Python-IDE.
- Operating System : Microsoft Windows 7 or Above
- Processor : Intel Core i3 or above
- RAM : 4 GB or above
- Hard Disk : 250 GB or above

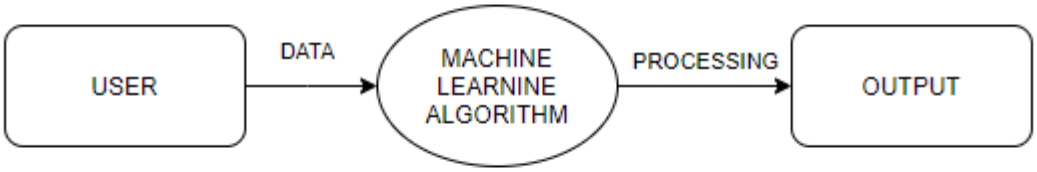
6 (B) SCOPE & FUTURE WORKS

- This research focuses on development of a machine learning model for predicting COVID-19 in patients.
- We also work to identify the features from the clinical information of patients that would influence the predictive result of COVID-19.
- This study does not focus on outer factors such as weather or any environmental factors that might influence results.
- This project will be helpful in predicting the second wave of corona virus by using the clinical data of patient along with CT –Scan Result. It will also require the use of sensors.

6 (C) SOFTWARE DESIGN

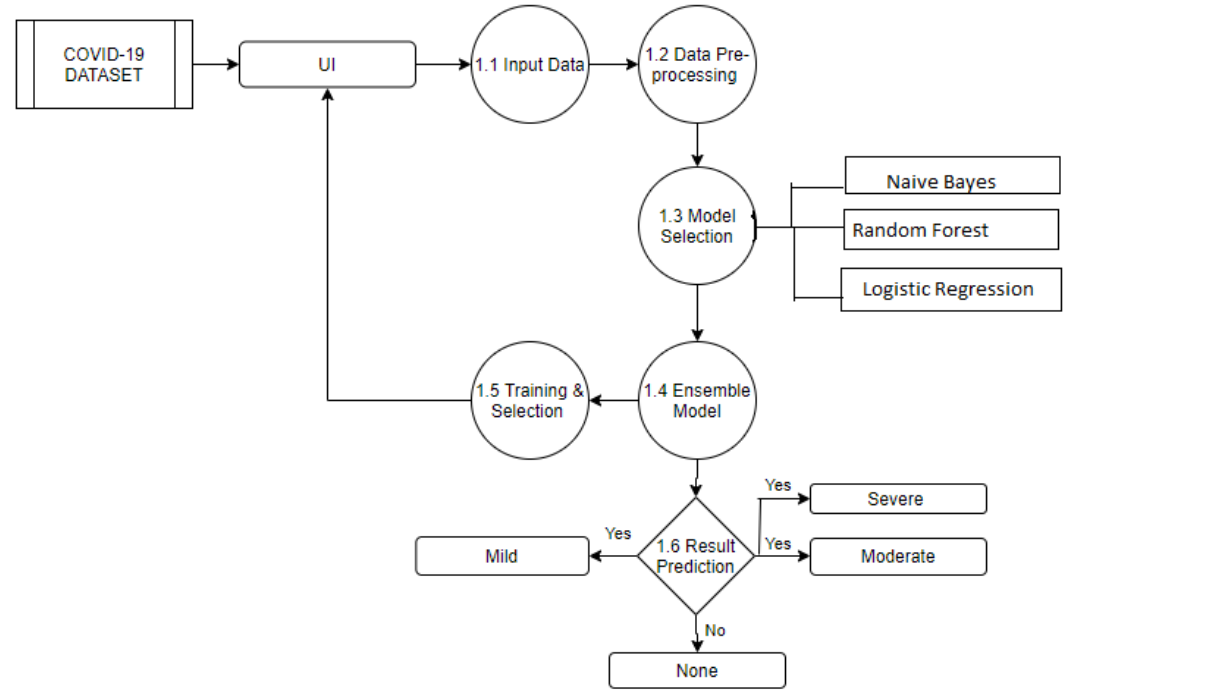


6 (D) CONTEXT DIAGRAM (DFD LEVEL 0)



- The context level diagram (Data Flow Diagram Level 0) consist of two external entities the user interface & the output block.
- The machine learning algorithm process block represents the use of these algorithms such as logistic regression, random forest, naive bayes..
- The output is obtained after processing.

6 (E) CONTEXT DIAGRAM (DFD LEVEL 1)

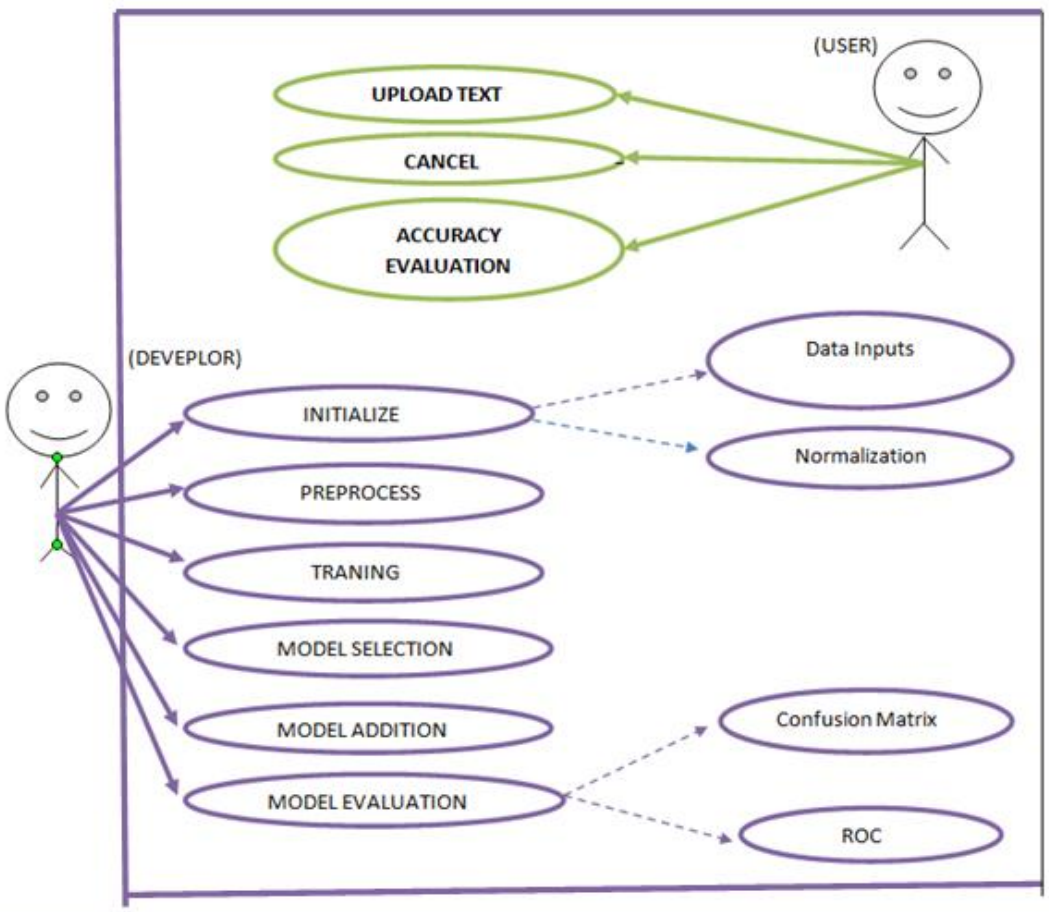


- The context diagram (DFD Level 1) contains 2 external entity UI block & Result Prediction block; five processes block Input data, Data Pre-processing, Model

selection, Ensemble Model & Training and Selection block.

- The COVID-19 Dataset consists of text data that is downloaded from various sources.
- The UI interface inputs the data from the users.
- From 1.1 process block the data is being provided by the user.
- The next 1.2 block preprocess the data & for selecting the model, 1.3 block works the machine learning algorithm.
- The next block 1.4 uses ensemble technique of bagging approach for final prediction of the data.
- The next block of 1.5 is used to train our machine learning model.
- Finally we get our results as per the given data.

6 (F) USE CASE DIAGRAM



6 (F) NON FUNCTIONAL REQUIRMENTS

As the name suggest these are requirement that are not directly interacted with the specific function in this project.

- Performance: The symptoms are taken as input from the users and they are treated as the important feature for analysis.
- Availability: The results are only accurate if the features of the input provided by the users are true and correct.
- Flexibility: It provides the users very comfortable way to get symptoms from user interface and analysis them.
- Learn ability: The software is easy to use & reduced the learning work.

7. SOFTWARE DESIGN SPECIFICATIONS

7 (A) THE DATASET

- Data collection was an essential and protracted process. Regardless the field of research, accuracy of the data collection is essential to maintain cohesion.
- These data will help to identify whether any person is having COVID-19 or not based upon some pre-defined standard symptoms by World Health Organization (WHO) and Ministry of Health & Family Welfare, India.
- The data-set is a combined multi-dimensional data. It contains fields with textual data and some with precise values. The data set for COVID-19 symptoms have been downloaded from world health organization.
- The attributes that were considered in the data-set for the machine learning model are presented in Table 1.

SR. No.	Feature Name	Feature Description
1	Person Gender	The gender of a person
2	Person Age	Classification of age according to WHO age group standard
3	Country Visited	List of country a person has visited.
4	Symptoms	According to WHO there are tiredness, difficulty in breathing, dry cough, sore throat, pain, nasal congestion, runny nose, diarrhoea etc.
5	Severity	Levels of severity are mild, moderate & severe.
6	Contact	Has the person came in contacted with COVID-19 patient

7 (B) LOADING & CLEANING THE DATA

The main reason behind data processing is that **data almost never comes in a form that is ready for us** and our personal experience, a large amount of time spent on a data science project is on manipulating data

.

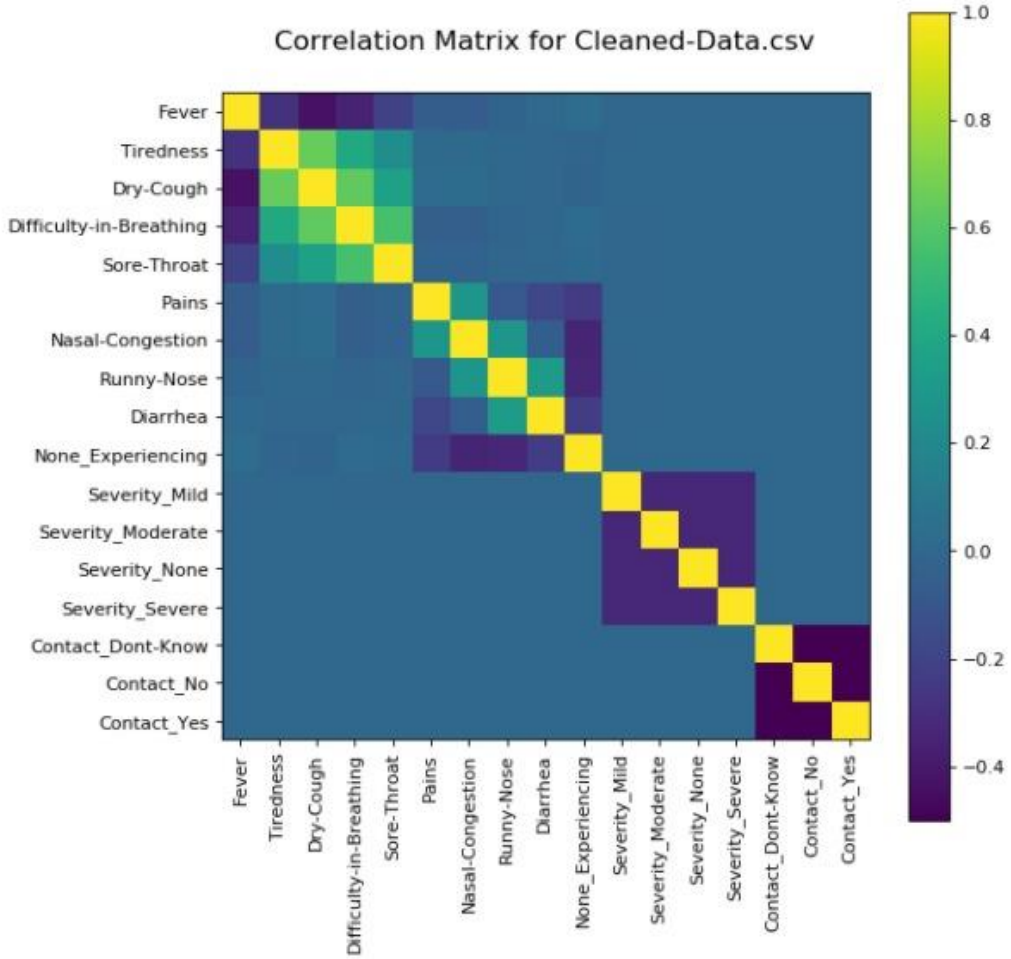
- The dataset is being download from various sites,

.

- While loading our dataset we have raw, separate values (.csv files). Where we have 3,16,000 records on patients varying from different regions of the world.

- We have approximately twenty-five plus features which are acting as symptoms & we have to classify the data into four major categories based on our feature.

- We have binary data for each and every dependent & independent variables



7 (C) DATA PRE-PROCESSING

- Data pre-processing occurs to be a curial step while implementing in every data science project.
- In our case due to binary data sight in every feature, we don't require much pre-processing of our data.
- We have first tried to use feature engineering to handle our categorical data input.
- There are some null values after that we have gone for a basic dimensionally deduction by admitting the non reliable feature of the data.
- At last we have also gone through capturing feature i.e. very important for some analysis about the most relevant feature acting in our results.

```
# Show the the tope 5 obersrvation of the dataset
display(df.head(5))

#Show data desccription
display(df.describe())

#show data shape
display(df.shape)

# show data type
display(df.dtypes)
```

	Fever	Tiredness	Dry-Cough	Difficulty-in-Breathing	Sore-Throat	None_Sympton	Pains	Nasal-Congestion	Runny-Nose	Diarrhea	...	C
0	1	1	1	1	1	0	1	1	1	1	...	1
1	1	1	1	1	1	0	1	1	1	1	...	1
2	1	1	1	1	1	0	1	1	1	1	...	1
3	1	1	1	1	1	0	1	1	1	1	...	1
4	1	1	1	1	1	0	1	1	1	1	...	1

7 (D) MODEL SELECTIONS

- Due to the crucial medical condition we cannot rely completely on a single model.

- Hence we have to go for the ensemble technique which can help in adapting the result from different models and finally improve our result.
- We have to use several machine learning algorithms such as logistic regression, random forest & naive bayes to create a model.
- This model will help to analyse & predict the final result using the bagging approach of the ensemble technique.
- Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.
- Model selection is a process that can be applied both across different types of models and across models of the same type configured with different model hyper parameters (e.g. different kernels in an SVM)
- For this task we need to compare the relative performance between models.
- Therefore the loss functions and the metric that represent it, becomes fundamental for selecting the right and non-over-fitted model.

7 (E) RANDOM FOREST ALGORITHM

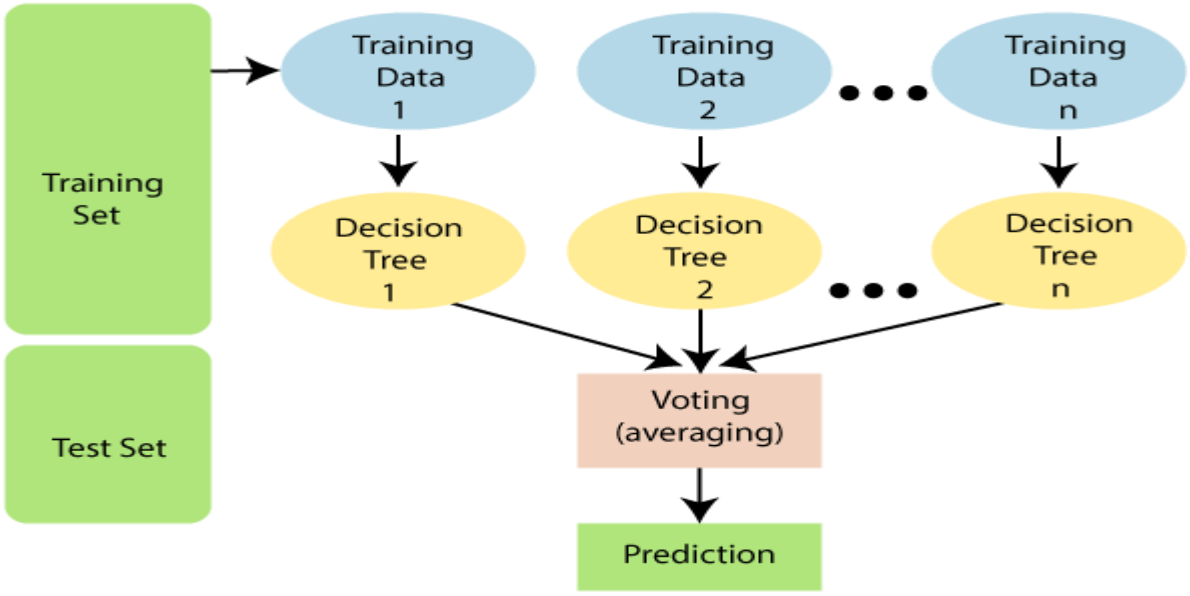
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest.

Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

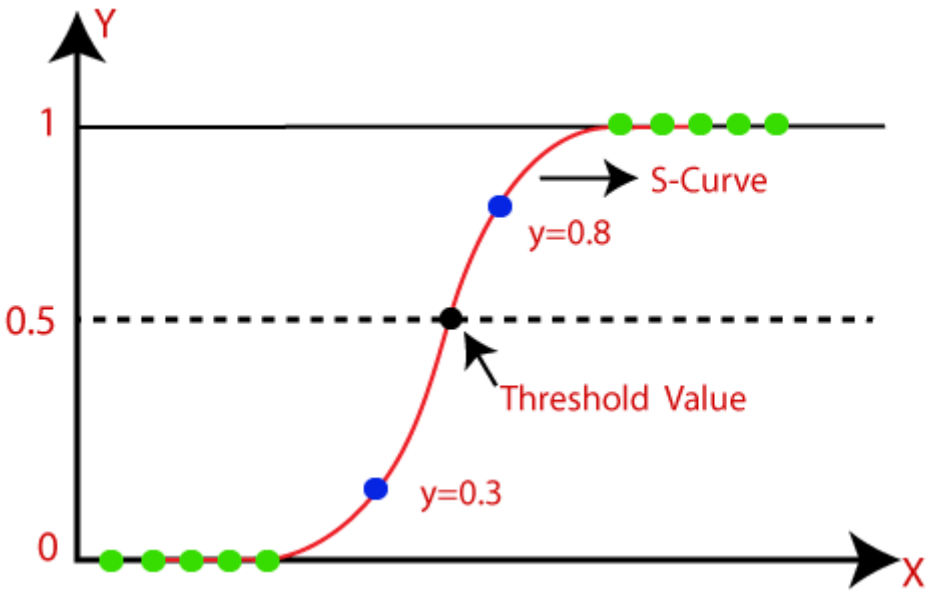


7 (F) LOGISTIC REGRESSION ALGORITHM

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features(or inputs), X.

Contrary to popular belief, logistic regression IS a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc



7 (G) NAIVE BAISED ALGORITHM

Naive Bayes algorithm is a classification technique based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other. In simple words, the assumption is that the presence of a feature in a class is independent to the presence of any other feature in the same class.

In Bayesian classification, the main interest is to find the posterior probabilities i.e. the probability of a label given some observed features,

$P(L \mid features)$. With the help of Bayes theorem, we can express this in quantitative form as follows −

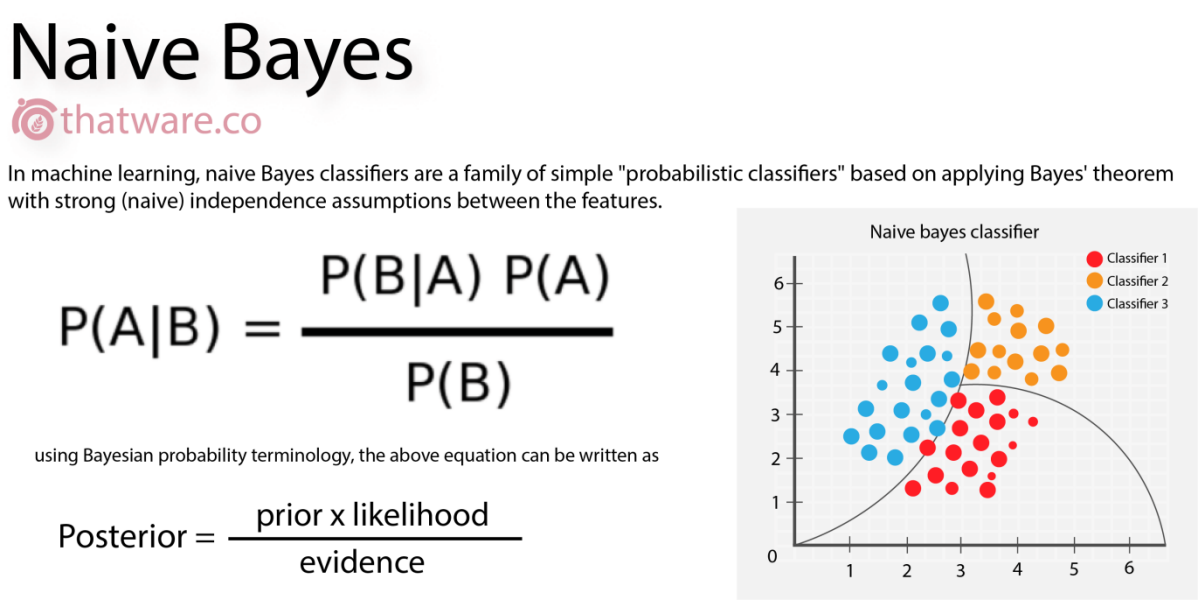
P(L|features)=P(L)P(features|L)P(features)P(L|features)=P(L)P(features|L)$P(features)$

Here, $P(L \mid features)$ is the posterior probability of class.

$P(L)$ is the prior probability of class.

$P(features \mid L)$ is the likelihood which is the probability of predictor given class.

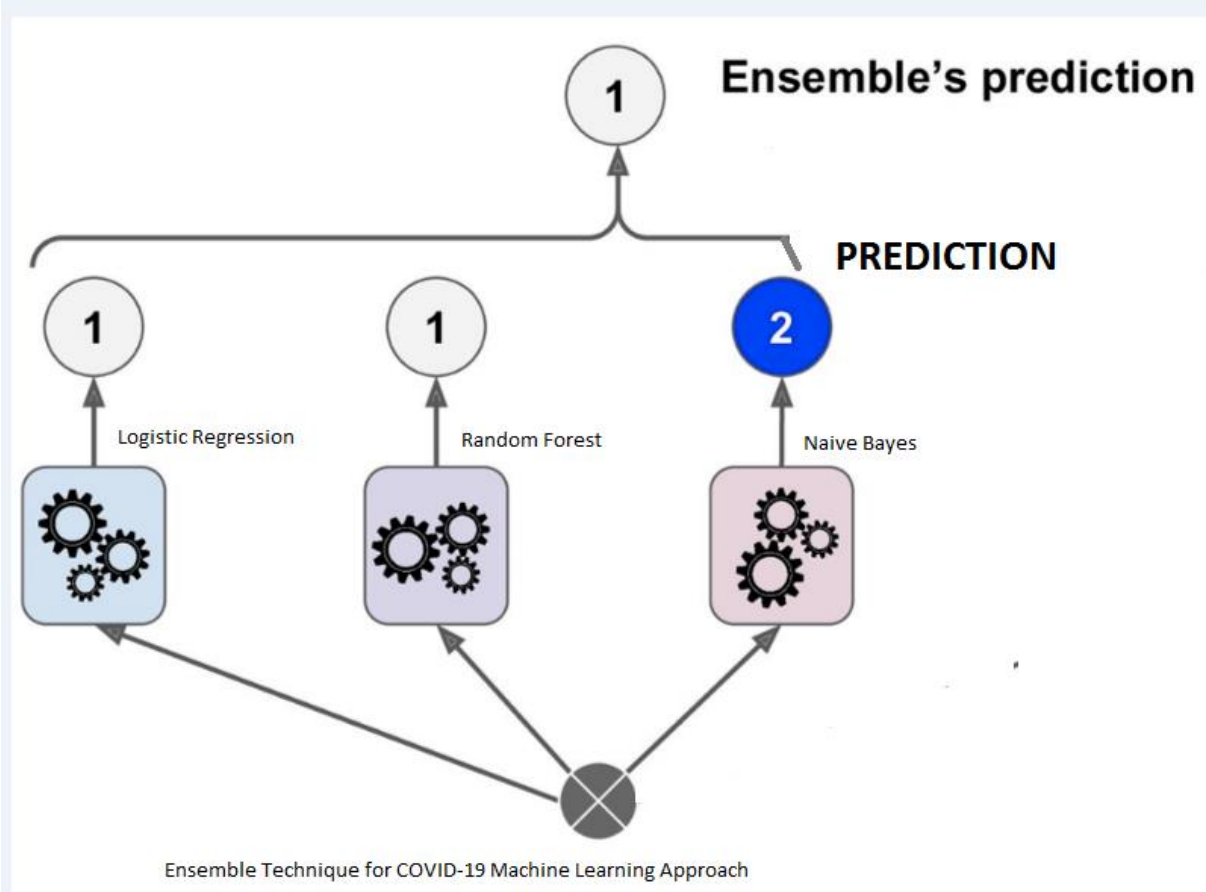
$P(features)$ is the prior probability of predictor.



7 (H) ENSEMBLE TECHNIQUE

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. The ensemble method usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods

Ensembles offer two specific benefits on a predictive modelling project, and it is important to know what these benefits are and how to measure them to ensure that using an ensemble is the right decision on your project.



Ensemble Technique for COVID-19 Machine Learning Approach

- In this project we have to use multiple machine learning algorithms to predict the result.
- The algorithms are Random Forest, Logistic Regression & Naive Bayes which will be implemented to test the accuracy of the data provided by the users.
- Finally the result is obtained by ensemble technique which help us in improving our result.

8. PREDICTION & IMPLEMENTATION

Finally after evaluating the final result came out by the bagging approach is very accurate we will get our predictable result .Our data is in the binary format so the values will be stored in the array.

Step I: Prepare to use Google Colab

```
#from google.colab import drive
#drive.mount('/content/drive')
```

```
# unzip dataset files from google drive to content folder in colab

import os

if os.path.exists("/content/covid19-data/")==False:
  print("unzip files!")
  !unzip -q "/content/covid19-data.zip"
```

Step II: Import the data from library

```
import pandas as pd
import numpy as np
import datetime
import requests
import warnings

import matplotlib.pyplot as plt
import matplotlib
import matplotlib.dates as mdates
import seaborn as sns
import squarify
import plotly.offline as py
import plotly_express as px


from IPython.display import Image
warnings.filterwarnings('ignore')
%matplotlib inline
```

Step III: Read the datasets

Country	Age	Gender	Symptoms	Experienci...	Severity	Contact
China	0-9	Male	Fever,Tiredness,Dry-Cough,Difficulty-in-Breathing,Sore-Throat	Pains,Nasal-Congestion,Runny-Nose,Diarrhea	Mild	Yes
Italy	10-19	Female	Fever,Tiredness,Dry-Cough,Difficulty-in-Breathing	Pains,Nasal-Congestion,Runny-Nose	Moderate	No
Iran	20-24	Transgender	Fever,Tiredness,Dry-Cough	Pains,Nasal-Congestion	Severe	Dont-Know
Republic of Korean	25-59		Fever,Tiredness	Pains	None	
France	60+		Fever	Nasal-Congestion,Runny-Nose,Diarrhea		
Spain			Tiredness,Dry-Cough,Difficulty-in-Breathing,Sore-Throat	Nasal-Congestion,Runny-Nose		
Germany			Tiredness,Dry-Cough,Difficulty-in-Breathing	Nasal-Congestion		

Step IV: Pre-process the data

```
# Show the the tope 5 obersrvation of the dataset
display(df.head(5))

#Show data desccription
display(df.describe())

#show data shape
display(df.shape)

# show data type
display(df.dtypes)
```

	Fever	Tiredness	Dry-Cough	Difficulty-in-Breathing	Sore-Throat	None_Sympton	Pains	Nasal-Congestion	Runny-Nose	Diarrhea	...	C
0	1	1	1	1	1	0	1	1	1	1	...	1
1	1	1	1	1	1	0	1	1	1	1	...	1
2	1	1	1	1	1	0	1	1	1	1	...	1
3	1	1	1	1	1	0	1	1	1	1	...	1
4	1	1	1	1	1	0	1	1	1	1	...	1

Step V: Analysis of Data

```
for i in df.columns:
    print('Attribute name:',i)
    print('--------------------')
    print(df[i].value_counts())
    print('--------------------')
```

```
Attribute name: Fever
--------------------
0    217800
1     99000
Name: Fever, dtype: int64
--------------------
Attribute name: Tiredness
--------------------
0    158400
1    158400
Name: Tiredness, dtype: int64
--------------------
Attribute name: Dry-Cough
--------------------
1    178200
0    138600
Name: Dry-Cough, dtype: int64
--------------------
```

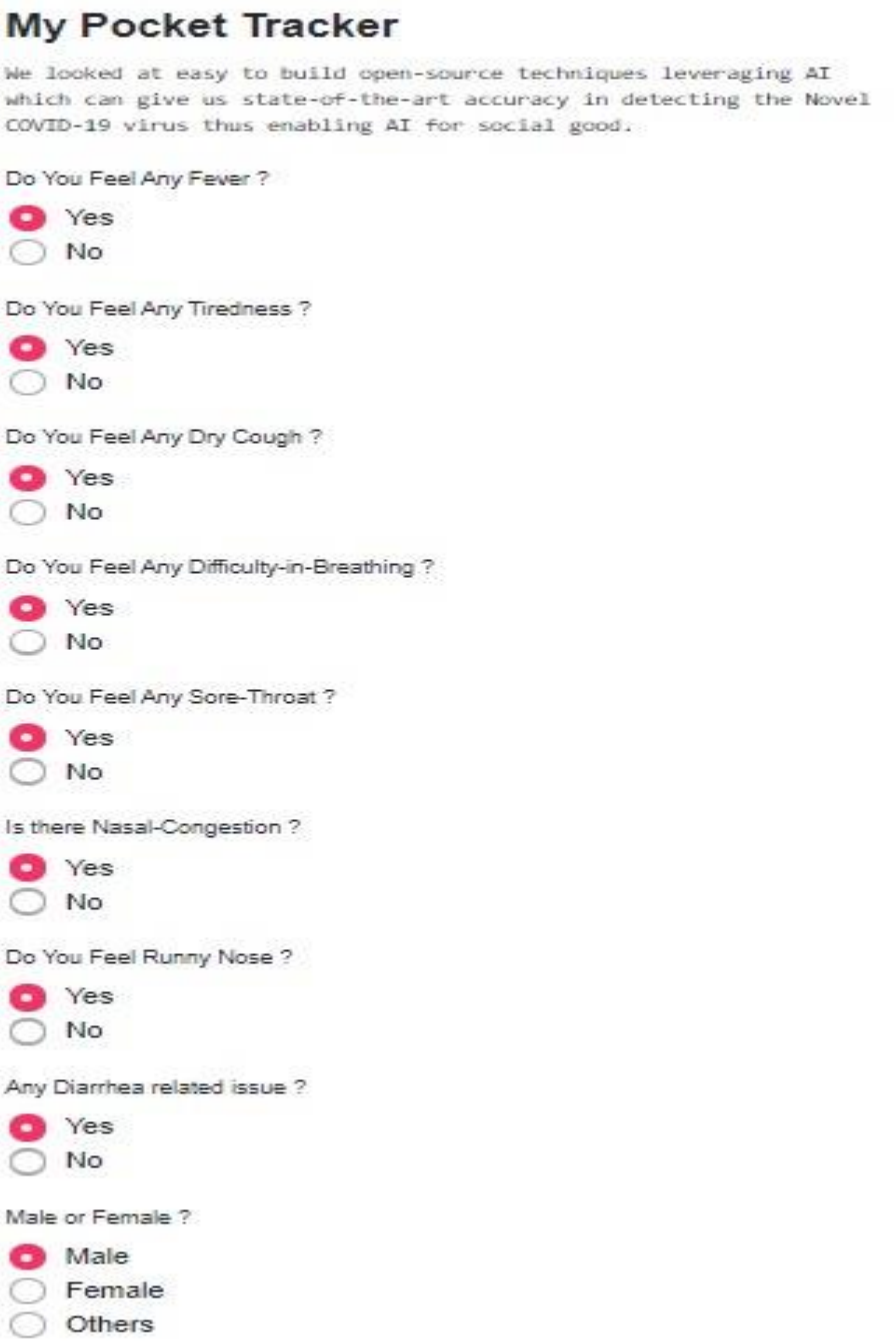
Step VI: Model Implementation

```
#RandomForestClassifier

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df=pd.read_csv("../input/covid19-symptoms-checker/Cleaned-Data.csv")
x=df.iloc[:,0:13].values
y=df.iloc[:,13].values
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=.25,random_state=0)
from sklearn.preprocessing import StandardScaler
sc_x= StandardScaler()
x_train=sc_x.fit_transform(x_train)
x_test=sc_x.fit_transform(x_test)

from sklearn.ensemble import RandomForestClassifier
classifier =  RandomForestClassifier(n_estimators=10)
classifier.fit(x_train,y_train)
y_pred=classifier.predict(x_test)
y_pred
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test,y_pred)
cm
from sklearn.metrics import accuracy_score
accuracy_score(y_test,y_pred)
```

Step VII: Graphical User Interface (Front-End Design)



- This is the front end interface of our project in which the user can select the symptoms .
- This is easy to use and collect the data from the users.

9. LIMITATIONS

- This entire project of COVID-19 is based upon the first wave of corona virus i.e. which can be determined by their symptoms.
- This project can only determine the result when the input values are correct.
- The machine learning model can be used to analysis the symptoms of this disease only when the correct option is being selected by the user.
- To predict the actual result we have to use the clinical dataset which contains the patient medical reports such as CT scanning & RT-PCR test data.

10. CONCLUSION

- In this research, a systematic literature review has been conducted to identify the suitable algorithm for prediction of COVID-19 in patients.
- The selected algorithms were trained with the patient clinical information about the basic symptoms that indicated the infection in a person.

.

- A prediction system that could find the possibility of outbreak of novel diseases that could harm mankind through socio-economic and cultural factor consideration can be developed.
- It is recommended to work on calibrated and ensemble methods that could resolve problems faster with better outcomes than the existing algorithms

11. **REFERNCES**

[1] Punn, N. S, Sonbhadra, S. K. & Agarwal, S. COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms. medRxiv, https://doi.org/10.1101/2020.04.08.20057679 (2020).

[2] Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. https://doi.org/10.1016/S1473-3099 (20)30120-1

[3] Mei, X. et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat. Med. 26, 1224–1228 (2020)

[4] Feng, C. et al. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. medRxiv, https://doi.org/10.1101/2020.03.19.20039099 (2020).

[5] Hastie, T., Tibshirani, R. & Friedman, J. In The Elements of Statistical Learning: Data Mining, Inference, and Prediction (eds. Hastie, T.,Tibshirani, R. & Friedman, J.)337–387 (Springer,2009).

[6] Google covid-19 search trends symptoms dataset. https://goo.gle/covid19symptomdataset