International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN : 2456-172X | Vol. 6, No. 2, June 2021 – August 2021 | Pages 1-17

Machine Learning Approach for Prediction of Health Parameters of Covid-19 Patients.

M V Jerabandi, Ashwini Kumbar, Preeti Bhomakkanavar, Sampreeta Patil
mark.jerabandi@gmail.com
Dept of CSE, Rural Engineering College
Hulkoti, Karnataka, India

ABSTRACT

Coronavirus disease, caused by severe acute respiratory syndrome has become an unprecedented public health crisis. The covid-19 pandemic has impacted the lives and health of persons worldwide, with potential for further effects in the future. The world health organization (WHO) has reported covid-19 as an epidemic that puts a heavy burden on healthcare sectors in almost every country. Investigation studies determined that artificial intelligence and machine learning techniques can play a key role in reducing the effect of the virus spread.ML is capable of detecting disease and virus infections more accurately so that patient's disease can be diagnosed at an early stage. here we include methods for forecasting future cases based on the existing data. machine learning approaches are used for predicting the chance of being infected and forecasting the number of positive cases. a trial was done for different algorithms and the algorithms that gave results with the best accuracy are covered. The availability of techniques for forecasting infectious disease can make it easier to fight covid-19.

1.0 Introduction

Machine Learning and Artificial intelligence are considered as an integral part of future technologies. Artificial Intelligence is an area Focused on developing Intelligent machines that work and react like humans. Machine learning is subfield of Artificial intelligence. The goal of machine learning generally is to understand the structure of data fit that data into models that can be understood and computational approaches. In Traditionally computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for specific range because of this machine learning facilities computers in building models from sample data in order to automate decision based on data inputs.

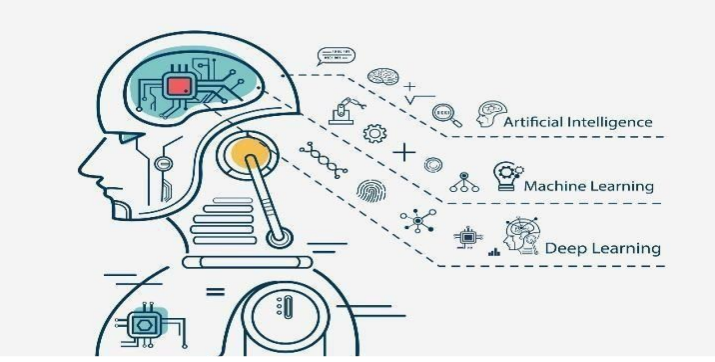

Fig.1 How does machine learning work?

Machine learning is a form of artificial intelligence that teaches computers to think in a similar way to how humans do learning and improving upon fast experiences.it works by exploring data, identifying patterns and involves minimal human interventions.almost any task that can be completed with a data defined pattern or set of rules can be automated with machine learning.this allows companies to transform processes that were previously only possible for humans to perform think responding to customer service calls,book keeping and reviewing resumes.

It is broadly classified into four types:

• **Supervised Machine Learning**.

Supervised Machine Learning Supervised Learning is a Machine Learning model that is built to give out predictions. This algorithm is performed by taking a labelled set of data as input and also known responses as output to learn the regression/classification model. It develops predictive models from classification algorithms and regression techniques.

• **Unsupervised Machine Learning.**

This is a data driven approach that works better when provided with sufficient data. For example, the movies in Netflix.com are suggested based on the principal of clustering of movies where several similar movies are grouped based on customer's recently watched movie list. It mostly discovers the unknown patterns in the data but most of the time these approximations are weak when compared with the supervised learning.

• **Semi-Supervised Machine Learning**

The name "semi-supervised learning" comes from the fact that the data used is between supervised and unsupervised learning. Semi-supervised algorithm has the tendency to learn both from labelled and unlabeled data. Semi-supervised machine learning gives high accuracy with a minimum annotation work. Semi-supervised machine learning uses mostly unlabeled data together combined with labelled data to give better classifiers. As less annotation work is enough to give good accuracy, humans have less work to do here.

• **Reinforcement Machine Learning.**

Reinforcement learning learns its behavior from a trial-and-error method in a dynamic environment. Here, the problem is solved by taking an appropriate action in a certain situation to maximize the output and to obtain the acquired results. In Reinforcement Learning, there is presentation of the input or output data. Instead, when the desired action is chosen, the agent is immediately told the reward and the next state are not considering the long terms actions. For the agent to act optimally it should have the knowledge about states, rewards, transitions and actions actively. Formally, the model consists of a discrete set of environment states, S, a set of scalar reinforcement signals typically {0;1} or the real numbers.

The pandemic covid-19 is a global challenge which has infected and killed people worldwide. some people do not show any symptom while some have a fever, cough, sore throat, general weakness and fatigue in most severe cases, severe phenomena, acute respiratory distress syndrome sepsis and septique shock all leading to death it has adversely affected the economy and social integrity of countries. There is rising concern about the mental health challenges of the general population along with health workers and family of infected people. this study aims to determine effect of covid-19 on mental health of people in India. It also focuses on the stigma and discriminating factors in our society and ways to cope with such conditions. The COVID-19 pandemic has led to a dramatic loss of human life worldwide and presents an unprecedented challenge to public health, food systems and the world of work. The economic and social disruption caused by the pandemic is devastating: tens of millions of people are at risk of falling

into extreme poverty, while the number of undernourished people, currently estimated at nearly 690 million, could increase by up to 132 million by the end of the year. Machine learning (ML) is a subfield of AI that focuses on algorithms that enable computers to define a model for complex relationships or patterns from empirical data without being explicitly programmed. Deep learning (DL), a subcategory of ML, achieves great power and flexibility compared to conventional ML models by drawing inspiration from biological neural networks to solve a wide variety of complex tasks, including the classification of medical imaging and natural language processing (NLP) AI techniques have been employed in the health care domain on different scales ranging from the prediction of disease spread trajectory to the development of diagnostic and prognostic models. These technologies and a wide range of data types, including data from social media, radiological images, omics, drug databases, and public health agencies, have been used for disease prediction. Several studies have focused on reviewing publications that discuss AI applications to support the COVID-19 response.

1.2 Literature Survey

In this section, a brief discussion of various research work carried out in the field of machine learning approach for prediction of health of COVID-19 is presented here.

1. **Artificial Intelligence and machine learning to fight covid-19:**

This article was published in 3 april,2020 by Ahmad Alimadadi,sachin aryal,Ishan manandhar.Coronavirus disease 2019,caused by severe acute respiratory syndrome coronavirus 2(SARS-cov-2), has become an unprecedented Public health crisis.The large-scale data of COVID-19 patients can be integrated and analysed by advanced machine learning algorithms to better understand the pattern of viral spread,further improve diagnostic speed and accuracy,develop novel effective therapeutic approaches and potentially identify the most suscepctible people based on personalized genetic and physiological characteristics. Machine learning techniques used here are taxonomic classification of covid-19 genomes,CRISPR-based covid-19 detection assay,survival prediction of severe covid-19 patients,and discovering potential drug candidates against covid-19.The US federal agencies are already promoting the formations of consortia and funding opportunities in addition to these initiatives integrating covid-19 related clinical data with existing banks,such as UK biobank,with pre-existing data of those patients,such as their genotype and physiological characteristics, could maximize our efforts towards a faster and feasible approach for meanginful data-mining by bioinformaticoians and computational scientists. A centralized collection of worldwide covid-19 patients' data will be beneficial for future artificial intelligence and machine learning research to develop a predictive, diagnostic, and therapeutic strategies against covid-19.

2. **COVID-19 Patient Health Prediction using Boosted Random Forest Algorithm.**

This article was published by 3 July 2020 by Celestine Iwendi,ali Kashif Bashir, Atharva peshkar.This paper purposes a fine-tuned Random forest model boosted by the Adaboost algorithm.The model uses the covid-19 patient's geographical,travel,health and demographical data to predict the severity of the case and the possible outcome,recovery or death.The model has an accuracy of 94% and a f1 score of 0.86 on the dataset used.The data analysis reveals a positive correlation between patients gender and death and also indicates that the majoritpatients are aged between 20 and 70 years. The author proposed a Generative Adversarial Network based fine-tuned model for detecting pneumonia from chest x-ray scans which is one of the symptoms of Covid-19 infection. we have used the pre-processed dataset to train multiple ML classification models.the models included in this study include decision tree

classifier, support vector classifier, Gaussian naïve bayes classifier, and boosted random forest classifier.

3. **Machine learning approaches in covid-19 severity risk prediction.**

The purpose of this study is to develop and test machines learning based models for covid 19 severity prediction. It consists of 337 test samples of covid-19 positive patients. Machine learning models were trained to predict the severity of sickness using patient data. A new feature engineering method based on topological data analysis called uniform manifold approximation and projection (UMAP) shown that it achieves better results and it has 100% accuracy. Machine learning classifier such as xGBoost, Adaboost, Random Forest and extra tree. This proposed approach aims to assist hospitals and medical facilities in determining who should be seen first and who has higher priority for admission to the hospitals.

4. **Severity detection for the corona virus disease patients using machine learning model based on blood and urine tests.**

This article was published in 31 July 2020 by Haochen Yao, Nan Zhang, this study investigated the detection of severely ill patients with covid-19 from those with mild symptoms using clinical information and the blood urine test data the clinical information of consisted of age, sex, body temperature, heart rate, respiratory rate and blood pressure. Support vector machine is a supervised machine learning that may accomplish both classification and regression tasks. SVM tries to find a hyperplane to separate data by the highest margin. This algorithm has been widely used to build the prediction models using the data of blood test and urine test. Random forest algorithm, K nearest algorithm, boosting based algorithm are implemented using python programming language and Scikit-learn package. The experimental data demonstrated strong correlations with the covid-19 severeness and the final covid-19 severeness detection model achieved the accuracy 0.8148 on the independent test dataset using only 28 clinical biomarkers.

5. **Cardiac involvement in covid-19 patients**

This article was published in 19 april,2020 by Ghazal aghagoli BS, Benjamin Gallo marin AB. This review seeks to gather and distil the existing body of literature that describes the cardiac implications of covid-19.Covid-19 patients with pre-existing cardiovascular disease are counted in greater frequency in intensive care unit settings,and ultimately suffer greater rates of mortality.The study found that patients with cardiac injury had higher mortality than those without 51.2% vs 4.5%,and the cox regression model shows that patients with cardiac injury are at a higher risk of death. Cardiac risk factors have been identified that predict the susceptibility to covid-19 infection and illness severity. according to the centers for disease control and prevention, elderly patients with comorbidities are at a higher risk to become infected by covid-19.

6. **Immediate and long-term consequences of covid-19 infections for the development of neurological disease.**

This article was published in 4 june,2020 by Michael T. Heneka, Robert Brown. Increasing evidence suggests that infection with sars-cov-2 causes neurological deficits in a substantial proportion of affected patients.While these symptoms arise acutely during the course of infection and problems of brain.covid-19 effected cases experience that high levels of proinflammatory cytokines and acute respiratory dysfunction.and requires ventilation.During the acute phase of covid-19 infection,36% of cases develop neurological symptoms of which 25% can be directly affected by central nervous system. The patients surviving with covid-19

are at a higher risk for subsequent development of neurological disease and in particular Alzheimer's disease.

7. COVID-19 and the effects on pulmonary function following infection.

This article was published in September 2021 by Kristyn L. Lewis, Neal M. Patel. This study aims to compare pre-infection and post-infection pulmonary function tests (PFT)in covid19 infected patients to better delineate between pre-existing abnormalities and effects of the virus. This was a retrospective multi-center cohort study.patients were identified based on having covid-19 and a pre and post infection PFT within one year of infection.this study says that there is no difference between the pre and post PFT data,specifically with the forced vital capacity.here there could be a relationship with ceratin underlying lung diseases and decreased lung function following infection.this information should aid clinicians in their interpretation of pulmonary function tests obatained by covid-19 infection.

8. 6-month consequences of covid-19 patients discharged from hospital.

This article was published in 16 jan,2021 by chowlin huang MD,prof Bin cao MD.The aim of this study is that to describe the long-term consequences of patients with covid-19 who have been discharged from hospital and to investigate the associated risk factors in particular disease severity.This study is the largest cohort study(n=1733) with the longest follow-up duration for the consequences of adult patients discharged from hospital recovering from covid-19.here they found that 76% of patients reported at least one symptom at 6 months after symptom onset and the proportion was higher in women. The common symptoms were fatigue, muscle weakness and sleep difficulties,23% of patients reported anxiety, depression at follow-up patients with pulmonary diffusion abnormally during follow-up is higher in patients. multivariable adjusted linear regression models was used here.

9. Covid-19 candidate treatments, a data analytic approach.

This paper focuses in evaluvating a repository of research papers to extract knowledge related to covid-19 and possible treatments.they have used 2 datasets here.covid-19 pulmonary risk literature clustering from Kaggle and another dataset is Maryland transportation institute dataset.the training methods used here are NLP via the spacy library,k-means,perceptron,support vector machine,decision tree classifier,random forest and libraries are pandas,numpy,matplotlib and several other libraries.decision tree classifier consistently provides the greater accuracy score .they built the model using tensorflow and keras using the sequential class and it has 98.65 accuracy. the end they analysed that the risk factors are age, weight, smoking, diabetes, chronic, respiratory diseases, asthma and immunity.

10. Follow-up studies in covid-19 recovered patients.

This article was published in august 2020 by Vellingiri Balachandra.This study suggests that it is highly important to provide counselling,mortal support to the recovered patients and society to restore to normally.The study describes that SARS-Cov-2 mainly affects the people who are previously having medical issues related to lungs,kidney,heart and the GI tract.the patients recovered from covid-19 are still carrying the virus in their body which could make it even more difficult to control the spread of the disease.most of the covid-19 recovered patients are experiencing stress for several weeks and this will usually disappear within a short period and physiological symptoms like depression,fear and anxiety may persist for a longer time. Here they recommend some precautions like covid-19 recovered patients should be treated with the atmost care, rapid follow-up should be done and which includes like nucleic acid tests, home monitoring programs for recovered patients can help them to improve their diet and

physical activity, proper counselling and education about the ill-effects of smoking and alcohol consumptions must be given to the recovered patients.

1.3 Motivation

It is important to understand different reasons of mental health issues, ways in which they are affecting our society and resilience of people and the ways they try to cope up with such situations. therefore, the current study will aim to show impact of covid-19 on mental health of people in India. Human beings are social species which require most satisfying environment with social relationship and physical well beings the pandemic has affected lifestyles, education, currier, development and economy in few months as there as being sudden increase in number of patients. Isolation, contact, restrictions, and economic shutdown impose a complete change to the psychosocial environment. these measures have the potential to threaten the mental health of children, adolescent and elders significantly.so that our main aim of this project is Predicting of various health issues on the basis of analysis.

- Following up the studies in covid-19 recovery patients by considering all the parameters. Covid-19 x-ray, CT scans of infections to the organs of the body before and after covid19

1.4 Problem statement:

Design and development for prediction of various health parameters of covid-19 patients using machine learning approach

1.5 Objectives:

- Identifying the most suitable machine learning techniques for prediction to perform on clinical reports of patients
- To test the performance of machine learning algorithms such as k nearest neighbor, Random Forest and extra tree classifier, logistic regression model and support vector machine
- Analysing the datasets and to predict the various health parameters of covid-19 patients.

2.0 Proposed Methodology

The proposed methodology for the prediction of Covid-19 is depicted in below. At the initial stage user will gather the dataset required to predict the quality of the water. Once the data has been loaded, system will pre-process the data and extraction of features will be done. After the features required for the prediction have been extracted, system will compare the features with model and prediction will be given as final result in the end. Graphical visualization are being made to compare the performance parameters of all Machine Learning algorithms considered in our work.

About Dataset original dataset was taken from https://www.kaggle.com/sudalairajkumar/novelcoronavirus-2019-datasetvirus-2019-dataset

data.csv -> original dataset
train.csv -> training dataset (contains data for 'death' and 'recovered' columns)
test.xlsx -> test dataset (does not contains data for 'death' and 'recovered' columns)
final.csv -> result obtained by running the model on test.xlsx

The dataset has been compiled from various sources including the World Health Organization and John Hopkins University. However, this dataset has been pre-processed further by us to meet the needs of this study. Fever, cough, cold, fatigue, body pain, and malaise were the most common symptoms that were noticed in patients whose data is available in this dataset.

2.1 Data exploration

A quatitative study was conducted to build up a trusted model to forecast covid-19 diagnosis, from the signs and symptoms that patients had. Here we use some of the dataset and the dataset include the patient's information like age, body temperature, heart rate, city scan, computer diagnosis, clinically obtained values was regarded as a feature in this study.our starategy includes the following stages like data collection, data pre-processing, classification and performance evaluation. The classification stage can be accomplished either by building statistical model. A block diagram of the purposed work is illustrated below.

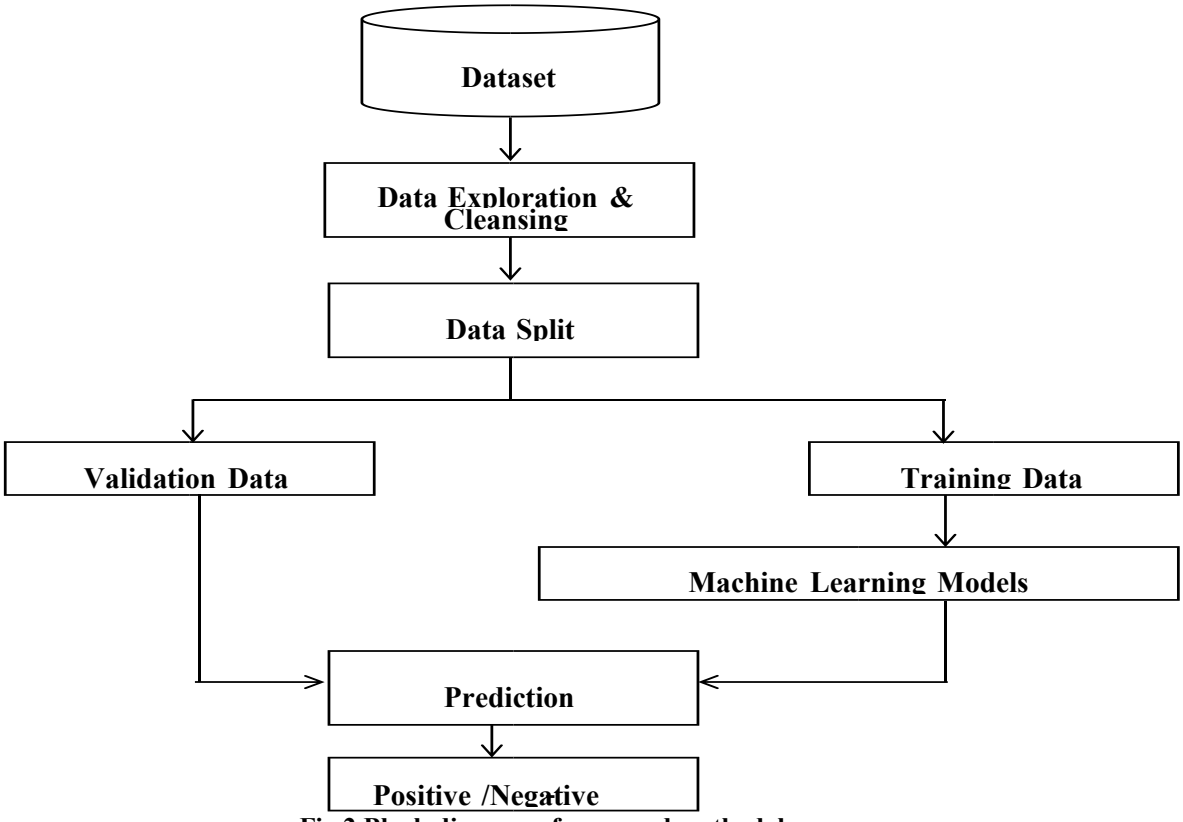

Fig.2 Block diagram of proposed methodology

2.1.1. Data Cleaning

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data cleaning. Professional data scientists usually invest a very large portion of their time in this step because of the belief that "Better data beats fancier algorithms". If we have a well-cleaned dataset, there are chances that we can get achieve good results with simple algorithms also, which can prove very beneficial at times especially in terms of computation when the dataset size is large. Obviously, different types of data will require different types of cleaning. However, this systematic approach can always serve as a good starting point.
Steps involved in Data Cleaning:

Fig. 3 Cycle of Data Cleaning Process

Removal of unwanted observations: This includes deleting duplicate/ redundant or irrelevant values from your dataset. Duplicate observations most frequently arise during data collection and irrelevant observations are those that don't actually fit the specific problem that you're trying to solve. Redundant observations alter the efficiency by a great extent as the data repeats and may add towards the correct side or towards the incorrect side, thereby producing unfaithful results. Irrelevant observations are any type of data that is of no use to us and can be removed directly.

Fixing Structural errors: The errors that arise during measurement, transfer of data, or other similar situations are called structural errors. Structural errors include typos in the name of features, the same attribute with a different name, mislabeled classes, i.e. separate classes that should really be the same, or inconsistent capitalization. For example, the model will treat America and America as different classes or values, though they represent the same value or red, yellow, and red-yellow as different classes or attributes, though one class can be included in the other two classes. So, these are some structural errors that make our model inefficient and give poor quality results.

Managing Unwanted outliers: Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. Generally, we should not remove outliers until we have a legitimate reason to remove them. Sometimes, removing them improves performance, sometimes not. So, one must have a good reason to remove the outlier, such as suspicious measurements that are unlikely to be part of real data.

Handling missing data: Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important. The two most common ways to deal with missing data are: Dropping observations with missing values. The fact that the value was missing may be informative in itself. Plus, in the real world, you often need to make predictions on new data even if some of the features are missing!

Imputing the missing values from past observations. Again, "missingness" is almost always informative in itself, and you should tell your algorithm if a value was missing. Even if you build a model to impute your values, you're not adding any real information. You're just reinforcing the patterns already provided by other features. Missing data is like missing a puzzle piece. If you drop it, that's like pretending the puzzle slot isn't there. If you impute it,

that's like trying to squeeze in a piece from somewhere else in the puzzle. So, missing data is always an informative and an indication of something important. And we must be aware of our algorithm of missing data by flagging it. By using this technique of flagging and filling, you are essentially allowing the algorithm to estimate the optimal constant for missingness, instead of just filling it in with the mean.

Data collection: Data collection was an essential and protracted process. The accuracy of the data collection is essential to maintain cohesion.as clinical information of patients was not publicly available; it was an inflexible and tedious process to collect the data.an intense search was conducted on various datasets together open-source clinical information of patients diagnosed with covid.

Data pre-processing: Data pre-processing is an important process in development of machine learning model. the data collected is often loosely controlled with controlled with out-of-range value, missing value etc. imputation of missing values in our data missing values have been handled by using simple imputer from sklearn python package.the missing values replaced by using main starategy. encoding categorical data.

Classification: Classification is a process related to categorization, the process in which negative and positive outcomes are recognized and understood. This is achieved by invoking various machine learning models. machine learning is widely used in different applications due to its powerful prediction and high accuracy while statistical analysis shows cases emphasis in models that can be interpreted easily with uncertainty and precision by using some models and algorithms such as listed below.

Data Split: The cleansed data is split into two categories in the ratio of 80:20 for training and validation respectively.

2.1.2. Machine learning models

Three machine learning models namely logistic regression, decision tree and XGB are trained to see the performance on collected data and the input data is assessed to predict whether the person is suffering from Parkinson or not.

Logistic Regression Algorithm: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. o Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems .In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

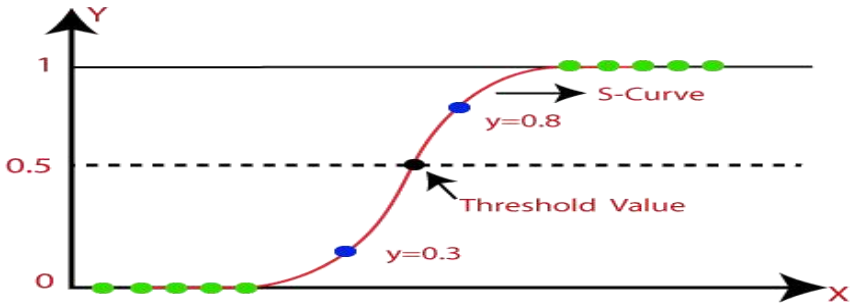

Fig. 4 Logistic Function

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function. o In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Decision Tree Algorithm: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. the decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. o A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

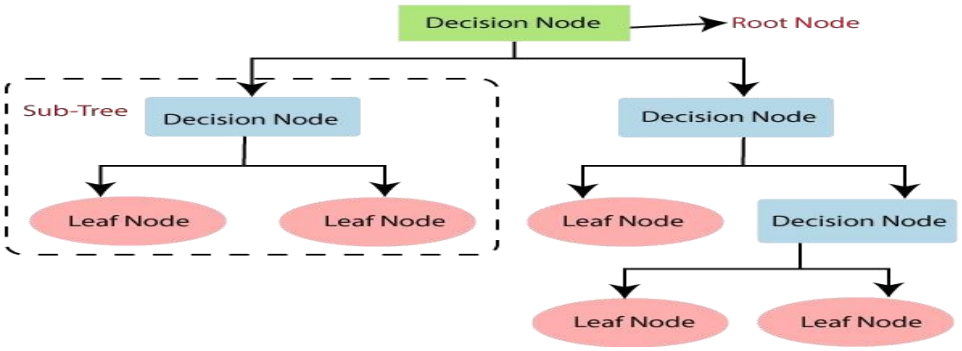

Fig.5 Structure of a Decision tree

Why we use Decision Trees? There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand. The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Gaussian Naive Bayes: It is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. We have explored the idea behind Gaussian Naive Bayes along with an example. Before going into it, we shall go through a brief overview of Naive Bayes. Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier.

Bayes Theorem: Bayes Theorem can be used to calculate conditional probability. Being a powerful tool in the study of probability, it is also applied in Machine Learning.

The Formula For Bayes' Theorem Is

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

where:

$P(A) =$ The probability of A occurring
$P(B) =$ The probability of B occurring
$P(A|B) =$ The probability of A given B
$P(B|A) =$ The probability of B given A
$P\left(A \bigcap B\right)) =$ The probability of both A and B occurring

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be-

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Sometimes assume variance is independent of Y (i.e., σi), or independent of Xi (i.e., σk) or both (i.e., σ). Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution. An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution.

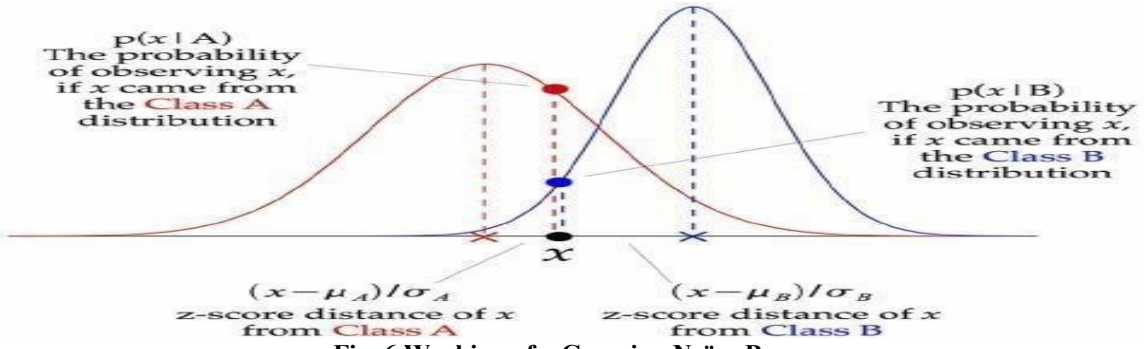

Fig. 6 Working of a Gaussian Naïve Bayes

The above illustration indicates how a Gaussian Naive Bayes (GNB) classifier works. At every data point, the z-score distance between that point and each class-mean is calculated, namely the distance from the class mean divided by the standard deviation of that class. Thus, we see that the Gaussian Naive Bayes has a slightly different approach and can be used efficiently

Boosted Random Forest: Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Features of Random Forest

- Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.

Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.

- Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

- Stability- Stability arises because the result is based on majority voting/ averaging.

Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

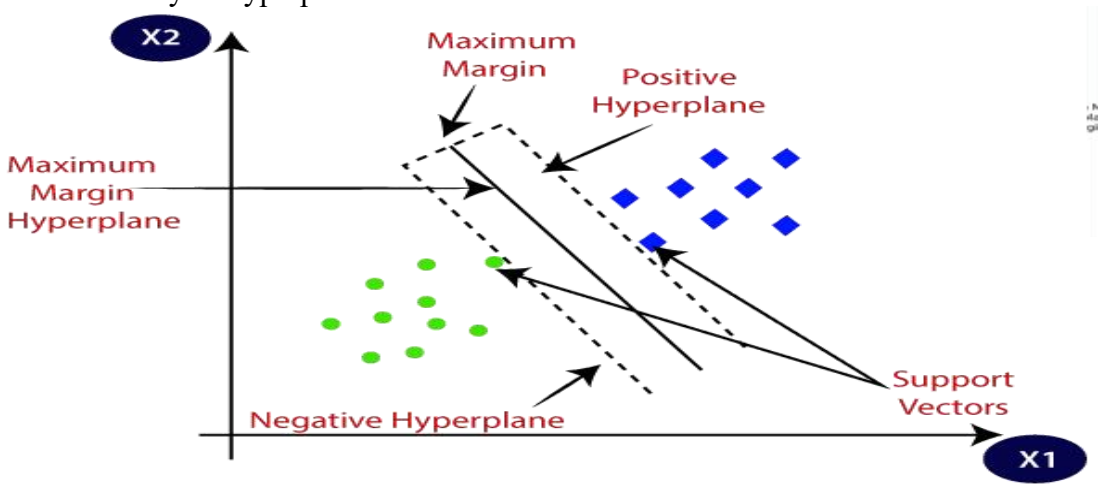

Fig. 7 Working of SVM

2.2 Tools and technologies used

2.2.1. Libraries

NumPy: NumPy library is an important foundational tool for studying Machine Learning. Many of its functions are very useful for performing any mathematical or scientific calculation. As it is known that mathematics is the foundation of machine learning, most of the mathematical tasks can be performed using NumPy.

Pandas: Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays

Matplotlib: It is an open-source plotting library in Python introduced in the year 2003. It is a very comprehensive library and designed in such a way that most of the functions for plotting in MATLAB can be used in Python. It consists of several plots like the Line Plot, Bar Plot, Scatter Plot, Histogram etc. through which we can visualize various types of data.

Sklearn: Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

Keras: Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result as fast as possible is key to doing good research.

Google Collaboratory: Google Collab is a online notebook-like coding environment that is well-suited for machine learning and data analysis. It comes equipped with many Machine Learning libraries and offers GPU usage. It is mainly used by data scientists and ML engineers.

Features

- Running the cell: To “Run the Cell” (a cell is a place where you enter either your code or text to be executed), you can either press “Ctrl +Enter” or “Shift +Enter”. I personally use “Shift +Enter” because when you press “Shift +Enter” it runs that particular cell and automatically creates a new cell, which is a handy feature to get things done quickly.
- Uploading files: This can be a really painful task if you follow the steps of some random online tutorials. But I have an easy solution which is a one-step process, you do not need to write extra codes, or perform some extra tasks, whereas you just have to upload manually to the Collab. Here is the way to do it.
- Importing the libraries without installing them: Google Colab gives you a unique feature that not even the best IDE provides such as “Importing the libraries without installations”, you don’t have to manually install any libraries prior to importing them, all you have to do is just tell for example import pandas as pd, per se and then your job is done, gone are those days where you have to type pip install pandas in your command prompt, which is disgusting. This is indeed one of the best features of Google Colab
- Faster GPUs. Access to faster GPUs and TPUs means you spend less time waiting while your code is running.

3.0 Results and Discussion

Evaluation Metrics: The purpose of the following study is to accurately predict the outcome of a particular patient depending on multiple factors, including but not limited to travel history, demographics etc. Since this is a very crucial prediction, accuracy is very important. Thus, for the purpose of evaluating the model we considered three evaluation metrics for this study.The

following terms are used in the equations: TP, True Positive; TN, True Negative; FP, False Positive; and FN, False Negative.

Accuracy: Given a dataset consisting of (TP + TN) data points, the accuracy is equal to the ratio of total correct predictions (TP + TN + FP + FN) by the classifier to the total data points. Accuracy is an important measure which is used to assess the performance of the classification model. Accuracy is calculated as shown in Equation (1) as follows:

Accuracy=TP + TNTP + TN + FP + FN 0.0<Accuracy<1.0

(1)Accuracy=TP + TNTP + TN +FP + FN 0.0<Accuracy<1.0

Precision: Precision is equal to the ratio of the True Positive (TP) samples to the sum of True Positive (TP) and False Positive (FP) samples. Precision is also a key metric to identify the number of correctly classified patients in an imbalanced class dataset. Precision is calculated as given in Equation (2) as follows:

Precision=TPTP + FP (2)Precision=TPTP + FP

Recall: Recall is equal to the ratio of the True Positive (TP) samples to the sum of True Positive (TP) and False Negative (FN) samples. Recall is a significant metric to identify the number of correctly classified patients in an imbalanced class dataset out of all the patients that could have been correctly predicted. Recall is calculated as given in Equation (3) as follows:

Recall=TPTP + FN (3)Recall=TPTP + FN(3)

F1 Score: F1 Score is equal to the harmonic mean of Recall and Precision value. The F1 Score strikes the perfect balance between Precision and Recall thereby providing a correct evaluation of the model's performance in classifying COVID-19 patients. This is the most significant measure that we will be using to evaluate the model. F1 Score can be calculated as shown in Equation (4) as follows:

F1 Score=2 × Precision × Recall Precision + Recall (4)F1 Score=2 × Precision × Recall Precision + Recall (4)

Results: We have used the pre-processed dataset to train multiple ML classification models. The models included in this study include: Decision Tree Classifier, Support Vector Classifier, Gaussian Naïve Bayes Classifier, and Boosted Random Forest Classifier.

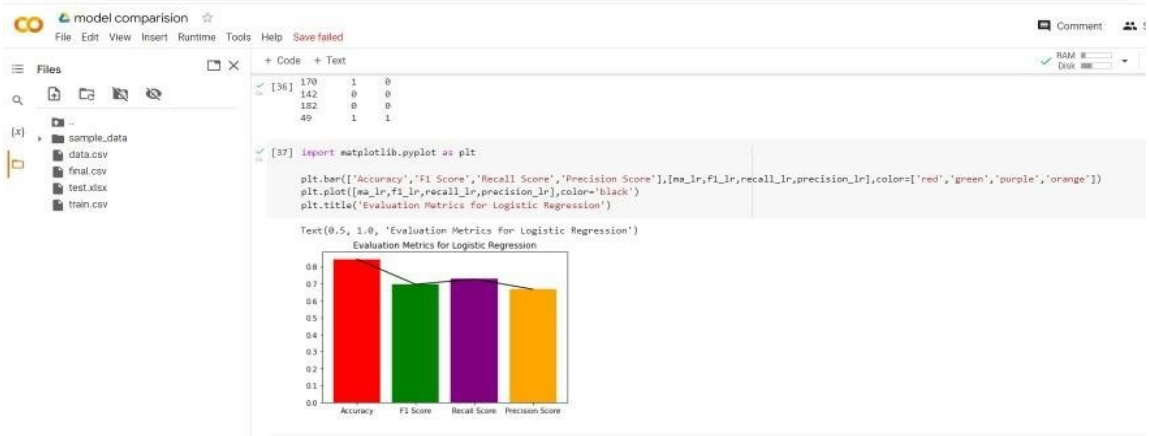

Fig.8 Evaluation metrics for Logistic Regression

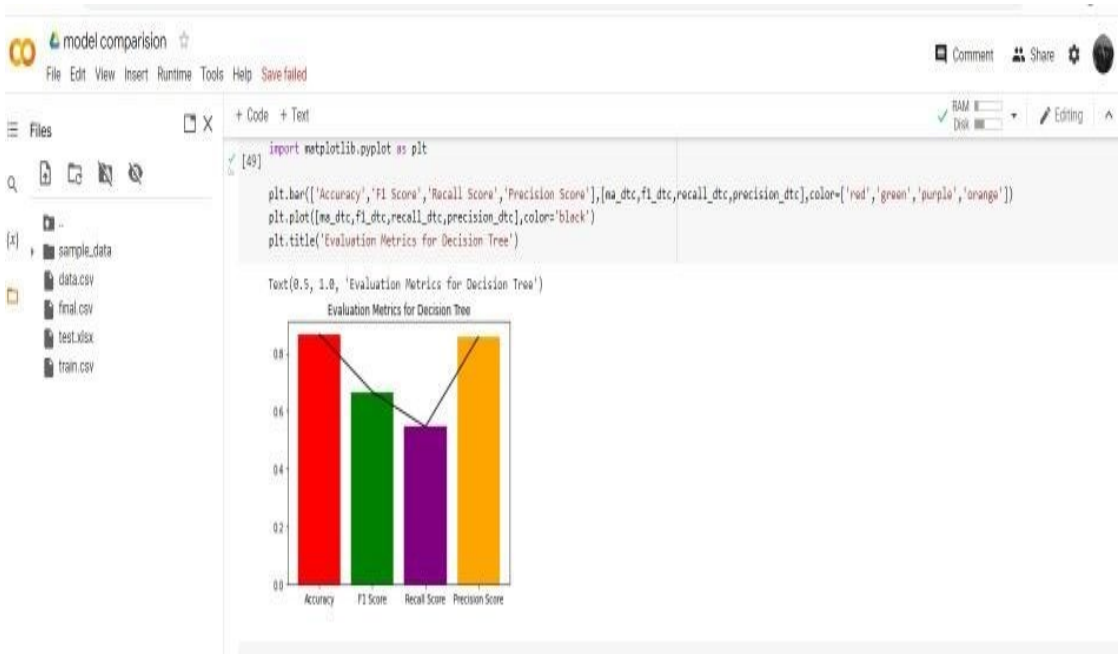

Fig. 9 Evaluation metrics of Decision tree

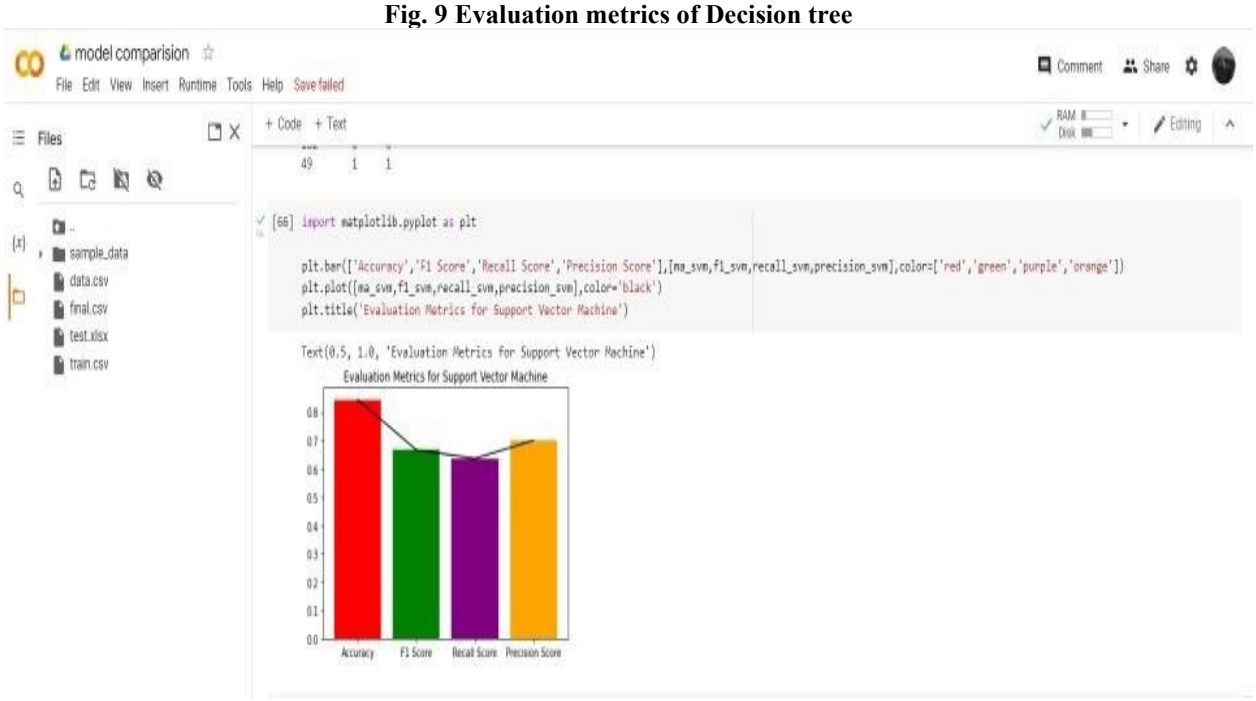

Fig. 10 Evaluation metrics of Support vector machine

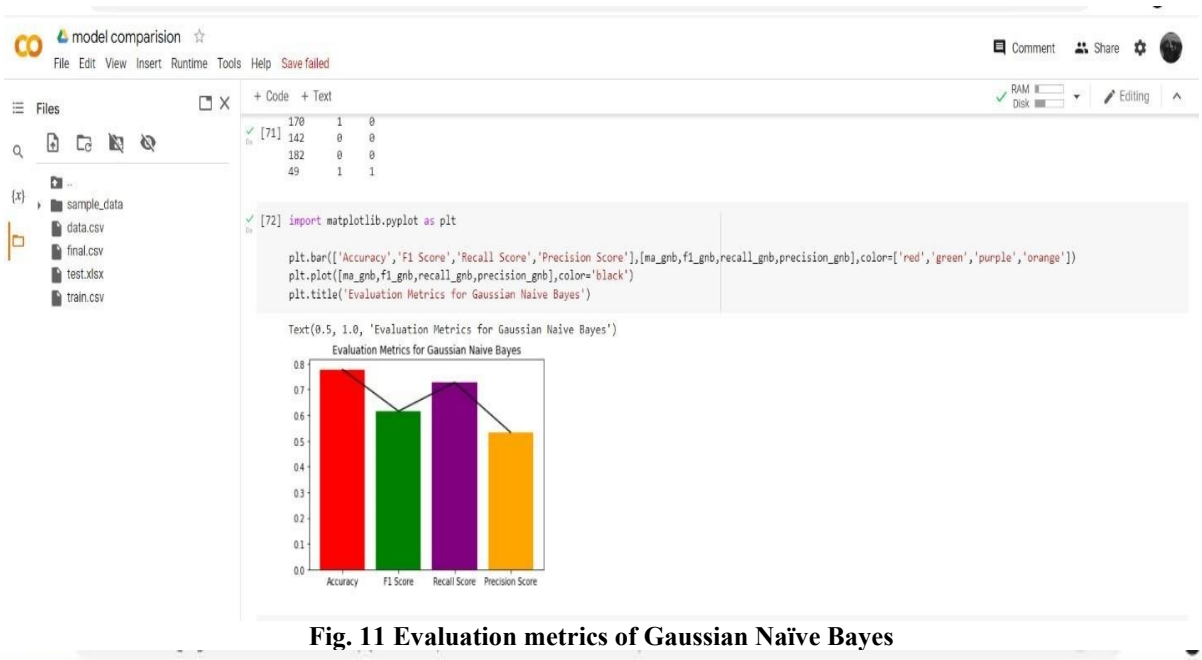

Fig. 11 Evaluation metrics of Gaussian Naïve Bayes

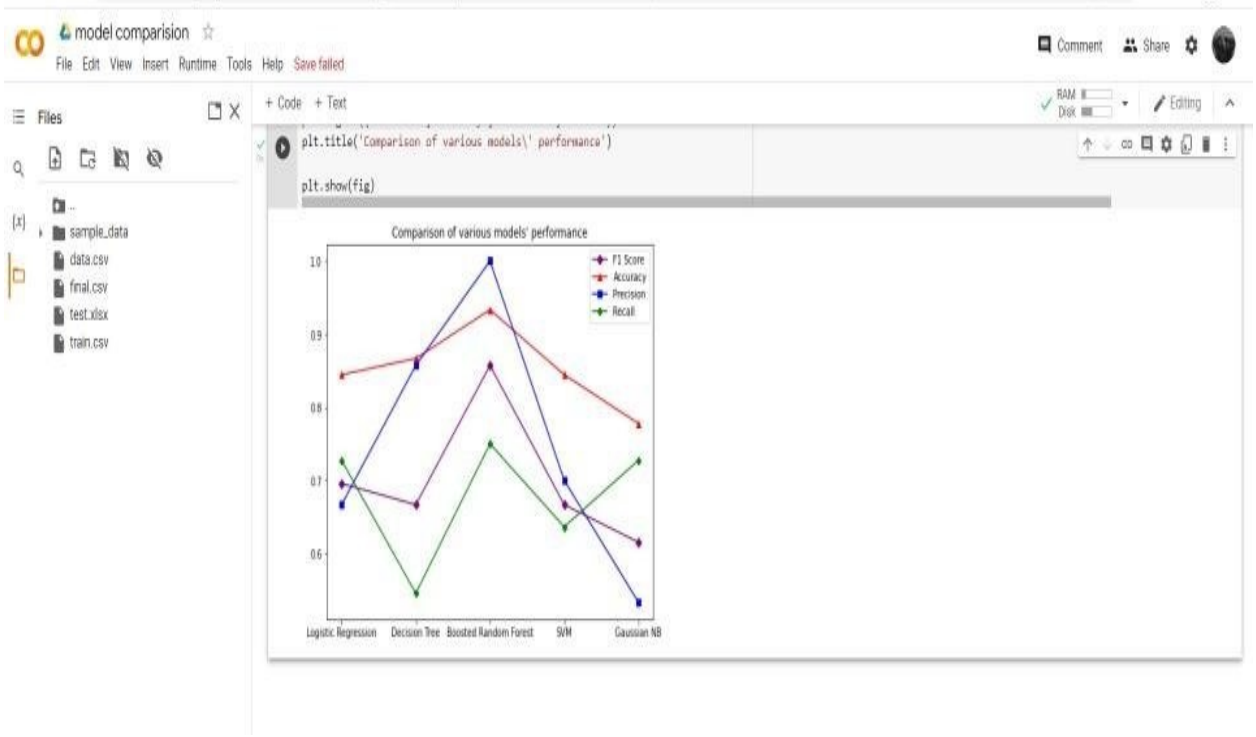

Fig. 12 Comparison of various models' performance

Since Boosted Random Forest algorithm is the best performing model, we will fine tune the model for better performance on the dataset.

4.0 Conclusion and Future Work

During the severe acute respiratory syndrome-new coronavirus-2 pandemic, clinicians turned to more quick diagnosis approaches due to a lack of laboratory diagnostic instruments and a long wait period. Although techniques based on proteomic analysis can efficiently diagnose COVID19 at an early stage, it is equally crucial to recognize serious COVID-19 patients before

they display severe symptoms. In this study, a set of methods for pre-processing data, manipulating categorical variables, and a feature selection procedure based on various statistical, mathematical and data analysis algorithms was performed to identify the most efficient feature engineering algorithm, for a prognostic prediction of severity. We utilize many Machine Learning algorithms to construct a predictive model to classify the data once pre-processed and reduced.

In terms of accuracy, sensitivity, specificity, and roc curve, the proposed system has proven successful and high performances.

Future work will focus on creating a pipeline that combines CXR scanning computer vision models with these types of demographic and healthcare data processing models. These models will then be integrated into applications that will support the growth of mobile healthcare. This can provide a step toward a semi-autonomous diagnostic system that can provide rapid screening and detection for COVID-19 affected regions and prepare us for future outbreaks.

References:

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. (2020) 395:497–506. Doi:10.1016/S0140-6736(20)30183-5
2. Prediction of COVID-19 Severity Using Chest Computed Tomography and Laboratory Measurements: Evaluation Using a Machine Learning Approach JMIR Med Inform 2020
3. WJ Wiersinga, A Rhodes, AC Cheng, SJ Peacock, HC Prescott Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review
4. M. Bansal Cardiovascular disease and COVID-19 Diabetes Metab. Syndr., 14 (3) (2020)
5. Mao L, Jin H, Wang M, Hu Y, Chen S, He Q, et al. Neurologic manifestations of hospitalized patients with coronavirus disease 2019 in Wuhan, China. JAMA Neurol.
6. Hazardous Postoperative Outcomes of Unexpected COVID-19 Infected Patients: A call for Global consideration of sampling all asymptomatic patients before surgical treatment by chen nahson, Arie Bittermanm Riad Haddad, David Hazzan, Ofer Lavie.
7. Follow-up studies in covid-19 pateints recovered pateints by Iyermahalaxmi, Mahadevi Subramanya.
8. Covid-19 patient health prediction using boosted Random Forest Algorithm by Celestine iwendi, Ali Kashif Bashir, R. Sujatha.
9. Comparing machine learning algorithms for predicting covid-19 mortality by Khadijeh Moulaei, Mostafa Shanbehzadeh.
10. 6-month consequences of covid-19 in patients discharged from hospital: a cohort study by Chaloin Huang MD, Lixue Huang MD.