**CENTRE OF DEVELOPMENT OF ADVANCED COMPUTING, MOHALI**



**ADVANCED COURSE ON DATA SCIENCE AND ANALYTICS (HP-KVN)**

**“PROJECT REPORT ON DETECTION OF COVID-19 VIRUS USING MACHINE LEARNING”**

## SUBMITTED TO: SUBMITTED BY:

MR. SANJAY MADAN Arpit shrivastav

ROHAN ABBHI

CHANCHAL SHARMA

# ACKNOWLEDGEMENT

It is our proud privilege and duty to acknowledge the kind of help and guidance received from several people in preparation of this report. It would not have been possible to prepare this report in this form without their valuable help, cooperation and guidance.

First, and foremost, we wish to record our sincere gratitude to C-DAC Coordinators for their constant support and encouragement in preparation of this report and for making available videos and interface facilities needed to prepare this report.

Their contributions and technical support in preparing this report are greatly acknowledged.

# TABLE OF CONTENTS

* **Introduction to covid 19**
* **Incubation period of covid 19**
* **How it is transmitted**
* **Symptoms of covid 19**
* **Introduction to ml**
* **Use of machine learning in covid 19**
* **Project objective**
* **Dataset description**
* **Libraries**
* **Machine learning workflow**
* **Methodology**
* **Implementation**
* **Model building**
* **Confusion matrix and Classification report**
* **Conclusion**
* **References**

**Abstract**

**COVID-19 outbreaks only affect the lives of people, they result in a negative impact on the economy of the country. On Jan. 30, 2020, it was declared as a health emergency for the entire globe by the World Health Organization (WHO). By Apr. 28, 2023, more than 3 million people were infected by this virus and there was no vaccine to prevent. The WHO released certain guidelines for safety, but they were only precautionary measures. The use of information technology with a focus on fields such as data Science and machine learning can help in the fight against this pandemic. It is important to have early warning methods through which one can forecast how much the disease will affect society, on the basis of which the government can take necessary actions without affecting its economy. In this chapter, we include methods for forecasting future cases based on existing data. Machine learning approaches are used and two solutions, one for predicting the chance of being infected and other for forecasting the number of positive cases, are discussed. A trial was done for different algorithms, and the algorithm that gave results with the best accuracy are covered in the chapter. The chapter discusses autoregressive integrated moving average time series for forecasting confirmed cases for various states of India. Two classifiers, random forest and extra tree classifiers, were selected; both have an accuracy of more than 90%. Of the two, the extra tree classifier has 93.62% accuracy. These results can be used to take corrective measures by different governmental bodies. The availability of techniques for forecasting infectious disease can make it easier to fight COVID-19.**

**Introduction to COVID-19**

**COVID-19 is not just a name now. It has become a deadly widespread virus that has affected tens of thousands of people all over the world. Its origin was Wuhan City, China in Dec. 2019. When people were unaware of the virus, COVID-19 started to spread from one person to another; it has slowly reached almost all countries and has become a pandemic.**

**COVID-19 is the short form for corona virus disease 2019, an illness caused by a novel corona virus (nCoV) now known as severe acute respiratory syndrome corona virus 2 (SARS-CoV-2); formerly called 2019-nCoV. COVID-19 was not the formal name of this virus; it was called SARS-CoV-2 by the International Committee on Taxonomy of Viruses because its symptoms were related to the virus that caused the SARS outbreak in 2003. However, this virus had not previously appeared in humans, and this time, they were severely infected by the virus, so to avoid confusion with other viruses, the World Health Organization (WHO) named it COVID-19 to communicate with the public.**

**During its early stages, COVID-19 was first identified as only an outbreak of respiratory illness cases in Wuhan City, Hubei Province, China. On Dec. 31, 2019, China reported about this respiratory disease to the WHO. It was declared to be COVID-19, a global health emergency, by the WHO on Jan. 30, 2020. According to records of WHO, in 2009, H1N1 was declared to be a global pandemic after which, on Mar. 11, 2020, COVID-19 was declared a global pandemic by the WHO.**

### Incubation period of COVID-19

The incubation period is the time between when someone catches the virus and when symptoms start to appear. As reported by the WHO, this virus has an incubation period of 2–14 days in the human body.

According to the Centers for Disease Control and Prevention (CDC), mild symptoms of the virus start appearing within 5 days and become worse afterward.

However, more recent data on patients showed that the incubation period had increased from 14 to 20 or 28 days as the virus started mutating, and after many negative tests, it suddenly revealed a positive result.

It was reported that patients have tested positive for the virus without having symptoms owing to a strong immunity system. As, the symptoms of this virus do not appear with strong immunity system, so if we come in contact with the person affected by the virus but with strong immunity system then we can definitely get infected.

### How it is transmitted

### The corona virus is transmitted from person to person when they are directly in contacted with each other or when the infected person sneezes or coughs. It is a respiratory disease, so it directly affects the respiratory system. According to CDC, nCoV is reported to be highly contagious, which means it spreads easily from among persons. It can also be spread when a person touches a surface or edible items that have come into contact with an infected person.

### Symptoms of COVID-19

The most common symptoms of COVID-19 are coughing and sneezing, fever, and breathing problems. In addition to these symptoms, diarrhea, hearing problems, a loss of sense of smell, chest pains, and nasal congestion are experienced.

The WHO released precautionary measures to avoid infection from COVID-19 virus. They include covering the face with a mask or cloth, avoid handshaking and instead bowing with namaste, following social distancing, and enforcing a lockdown.

**Introduction to machine learning**

According to Arthur Samuel (1959), ML is the field of study that gives computers the ability to learn without being explicitly programmed. Thus, we can define ML as the field of computer science in which machines can be designed that can program themselves.

The process of learning is simply learning from experience or observations from previous work, such as examples, or instruction, to look for patterns in data and with the help of examples, provided the system can make better decisions. The basic aim of ML is to make computers learn automatically with no human intervention and to adjust perform actions accordingly past data are used to train the model, and then this trained model is used to test new data and then for prediction. The trained ML model's performance is evaluated using some portion of available past data (which is not present during training). This is usually referred as the validation process.

In this process, the ML model is evaluated for its performance measure, such as accuracy.

Accuracy describes the ML model's performance over unseen data in terms of the ratio of the number of correctly predicted features and total available features to be predicted.

An external file that holds a picture, illustration, etc.
Object name is f20-01-9780128245361.jpg

### Some machine learning methods:

### ML algorithms can be divided into supervised or unsupervised learning:

### Supervised ML algorithms is a type of ML technique that can be applied according to what was previously learned to get new data using labeled data and to predict future events or labels. In this type of learning, supervisor (labels) is present to guide or correct. For this first analysis, the known training set and then the output values are predicted using the learning algorithm. The output defined by the learning system can be compared with the actual output; if errors are identified, they can be rectified and the model can be modified accordingly.

### Unsupervised ML algorithms: In this type, there is no supervisor to guide or correct. This type of learning algorithm is used when unlabeled or unclassified information is present to train the system. The system does not define the correct output, but it explores the data in such a way that it can draw inferences (rules) from datasets and can describe hidden structures from unlabeled data.

### Semi supervised ML algorithms are algorithms that are between the category of supervised and unsupervised learning. Thus, this type of learning algorithm uses both unlabeled and labeled data for training purposes, generally a small amount of labeled data and a large amount of unlabeled data. This type of method is used to improve the accuracy of learning.

### Reinforcement ML algorithms is a type of learning method that gives rewards or punishment on the basis of the work performed by the system. If we train the system to perform a certain task and it fails to do that, the system might be punished; if it performs perfectly, it will be rewarded. It typically works on 0 and 1, in which 0 indicates a punishment and 1 indicates a reward.

### It works on the principle in which, if we train a bird or a dog to do some task and it does exactly as we want, we give it a treat or the food it likes, or we might praise it. This is a reward. If it did not perform the task properly, it might be scolded as a punishment by us.

**Use of machine learning in Covid-19**

ML is used in various fields, including medicine to predict disease and forecast its outcome. In medicine, the right diagnosis and the right time are the keys to successful treatment. If the treatment has a high error rate, it may cause several deaths. Therefore, researchers have started using artificial intelligence applications for medical treatment. The task is complicated because the researchers have to choose the right tool: it is a matter of life or death.

For this task, ML achieved a milestone in the field of health care. ML techniques are used to interpret and analyze large datasets and predict their output. These ML tools were used to identify the symptoms of disease and classify samples into treatment groups. ML helps hospitals to maintain administrative processes and treat infectious disease.

ML techniques were previously used to treat cancer, pneumonia, diabetes, Parkinson disease, arthritis, neuromuscular disorders, and many more diseases; they give more than 90% accurate results in prediction and forecasting.

ML techniques were previously used to treat cancer, pneumonia, diabetes, Parkinson disease, arthritis, neuromuscular disorders, and many more diseases; they give more than 90% accurate results in prediction and forecasting. ML is also used to diagnose the disease based on x-ray images. For instance, chest images of patients can be used to detect whether a patient is infected with COVID-19.Moreover, social distancing can be monitored by ML; with the help of this approach, we can keep ourselves safe from COVID-19.

**Project Objective**

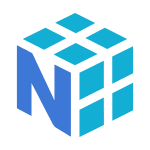
* The Aim of the project is to detect the covid-19 patients on the basis of symptoms.
* To detect whether the patient is Covid positive or not.
* The goal of this study is the evaluation of different patients with a focus on symptoms and pre-existing illness at admission, this is important for initial assessment and adequate emergency care.

**Dataset Description**

* We have taken the Dataset from Kaggle that is in the form of CSV format.
* Dataset contains 5434 rows and 21 columns .
* There is no missing values for the provided input dataset.
* Covid-19 is the variable which notifies whether a particular patient is covid positive or not. And we will be developing our models to detect.

**LIBRARIES**

**NUMPY**



NumPy is the fundamental package for scientific computing in Python. It is a python library that provides a multidimensional array object,

*import numpy as np*

various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

### PANDAS



When working with tabular data, such as data stored in spreadsheets or databases, pandas is the right tool for you. pandas will help you to explore, clean, and process your data. In pandas, a data table is called a DataFrame.

*import pandas as pd*

To load the pandas package and start working with it, import the package. The community agreed alias for pandas is pd, so loading pandas as pd is assumed standard practice for all of the pandas documentation.

### MATPLOTLIB



Matplotlib is a library for making 2D plots of arrays in Python. Although it has its origins in emulating the MATLAB®1 graphics commands, it is independent of MATLAB, and can be used in a Pythonic, object oriented way.

import matplotlib.pyplot as plt

Although Matplotlib is written primarily in pure Python, it makes heavy use of NumPy and other extension code to provide good performance even for large arrays.

### SEABORN



Seaborn is built on top of Python’s core visualisation library Matplotlib. It is meant to serve as a complement, and not a replacement. However, Seaborn comes with some very important features. Let us see a few of them here .

import seaborn as sns

The features help in -

* Built in themes for styling matplotlib graphics
* Visualizing univariate and bivariate data
* Fitting in and visualising linear regression models
* Plotting statistical time series data
* Seaborn works well with Num Py and Pandas data structures
* It comes with built in themes for styling Matplotlib graphics

### SCIKIT-LEARN



### SKLEARN.METRICS

from sklearn.metrics import classification\_report , confusion\_matrix

The [**sklearn.metrics**](https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics) module implements several loss, score, and utility functions to measure classification performance. Some metrics might require probability estimates of the positive class, conf idence values, or binary decisions values. Most implementations allow each sample to provide a weighted contribution to the overall score, through the sample\_weight parameter.

### SKLEARN.ENSEMLE

*from sklearn.ensemble import RandomForestRegressor*

Random forest regressor.

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

### SKLEARN.NEIGHBORS

*from sklearn.neighbors KNeighborsClassifier*

Scikit-learn implements two different nearest neighbors classifiers: [**KNeighborsClassifier**](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier) implements learning based on the k nearest neighbors of each query point, where k is an integer value specified by the user. [**RadiusNeighborsClassifier**](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.RadiusNeighborsClassifier.html#sklearn.neighbors.RadiusNeighborsClassifier) implements learning based on the number of neighbors within a fixed radius r of each training point, where r is a floating-point value specified by the user.

The k-neighbors classification in [**KNeighborsClassifier**](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier) is the most commonly used technique. The optimal choice of the value k is highly data-dependent: in general a larger k suppresses the effects of noise, but makes the classification boundaries less distinct.

**SKLEARN.TREE**

*from sklearn.tree import DecisionTreeClassifier*

**Decision Trees (DTs)** are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-regression). The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

### SKLEARN. LINEAR\_MODEL

*from linear\_model import LogisticRegression*

Logistic Regression (aka logit, MaxEnt) classifier.

In the multiclass case, the training algorithm uses the one-vs-rest (OvR) scheme if the ‘multi\_class’ option is set to ‘ovr’, and uses the cross-entropy loss if the ‘multi\_class’ option is set to ‘multinomial’. (Currently the ‘multinomial’ option is supported only by the ‘lbfgs’, ‘sag’, ‘saga’ and ‘newton-cg’ solvers.)

This class implements regularized logistic regression using the ‘liblinear’ library, ‘newton-cg’, ‘sag’, ‘saga’ and ‘lbfgs’ solvers. **Note that regularization is applied by default**. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted

### SKLEARN.SVM

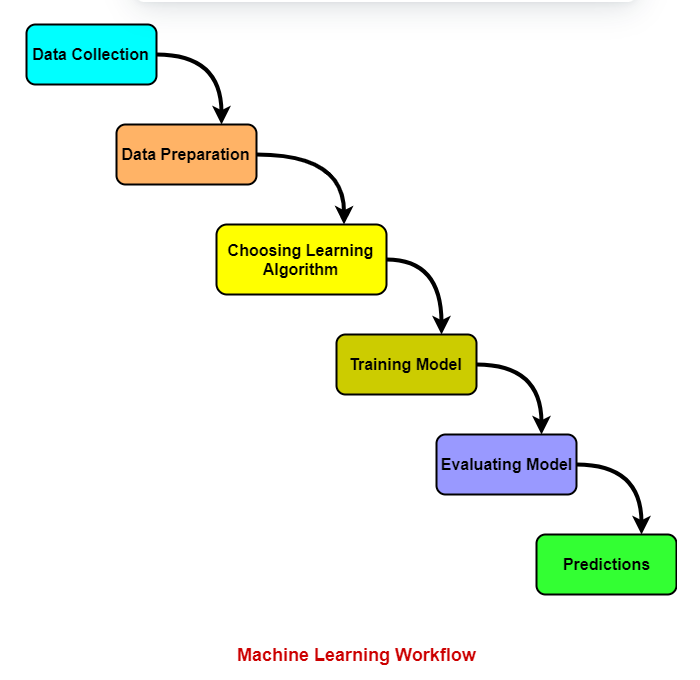
*from sklearn.svm import SVM*

**Support vector machines (SVMs)** are a set of supervised learning methods used for [classification](https://scikit-learn.org/stable/modules/svm.html#svm-classification), [regression](https://scikit-learn.org/stable/modules/svm.html#svm-regression) and [outliers detection](https://scikit-learn.org/stable/modules/svm.html#svm-outlier-detection).

The advantages of support vector machines are:

* Effective in high dimensional spaces.
* Still effective in cases where number of dimensions is greater than the number of samples.
* Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
* Versatile: different [Kernel functions](https://scikit-learn.org/stable/modules/svm.html#svm-kernels) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

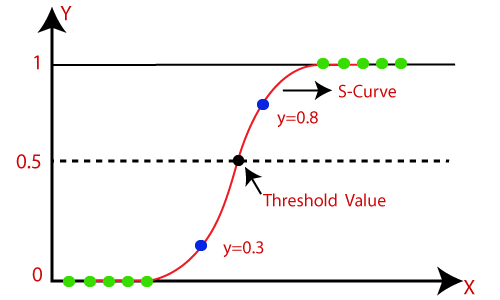
## MACHINE LEARNING WORKFLOW



# METHODOLOGY

## LOGISTIC REGRESSION

* 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.
* 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.



## Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

Logistic Regression in Machine Learning

* In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

Logistic Regression in Machine Learning

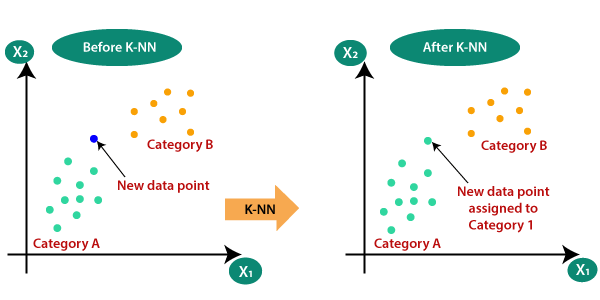
* But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

Logistic Regression in Machine Learning

The above equation is the final equation for Logistic Regression.

## KNN ALGORITHM

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



## 

## DECISION TREE

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.**
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.



**IMPORTANT DEFINITIONS :**

**INFORMATION GAIN:**

* Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
* It calculates how much information a feature provides us about a class.
* According to the value of information gain, we split the node and build the decision tree.
* A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

**Information Gain= Entropy(S)- [(Weighted Avg) \*Entropy (each feature)**

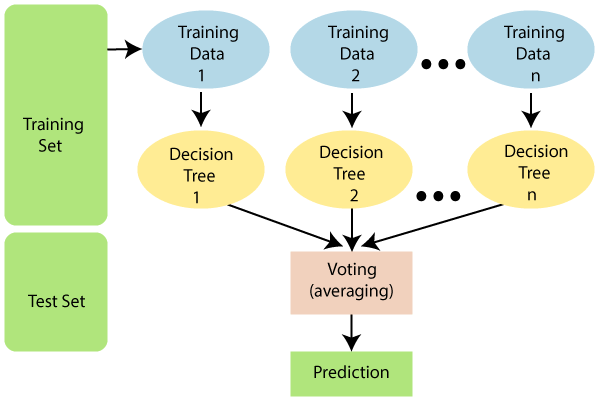
**GINI INDEX:**

* Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm.
* An attribute with the low Gini index should be preferred as compared to the high Gini index.
* It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
* Gini index can be calculated using the below formula:

**Gini Index= 1- ∑jPj2**

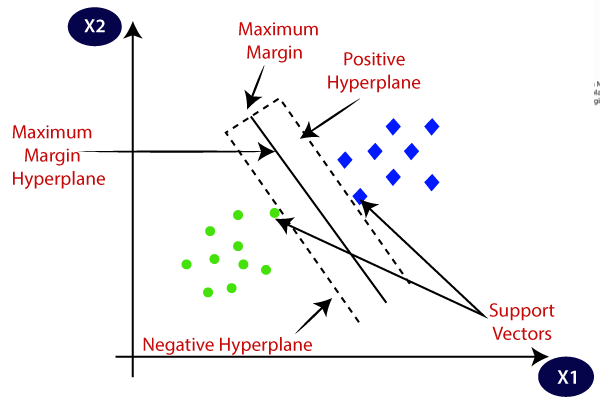
## RANDOM FOREST

* Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*
* As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
* The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



## SVM (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.



## TYPES OF SVM

* **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

**HYPERPLANE AND SUPPORT VECTORS IN THE SVM ALGORITHM:**

* **Hyperplane:** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

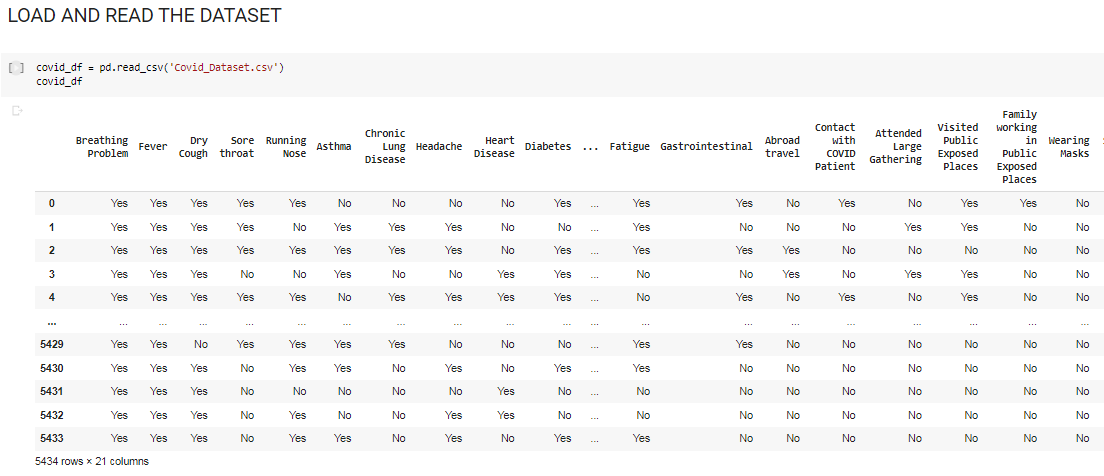
We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

* **Support Vectors:**

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

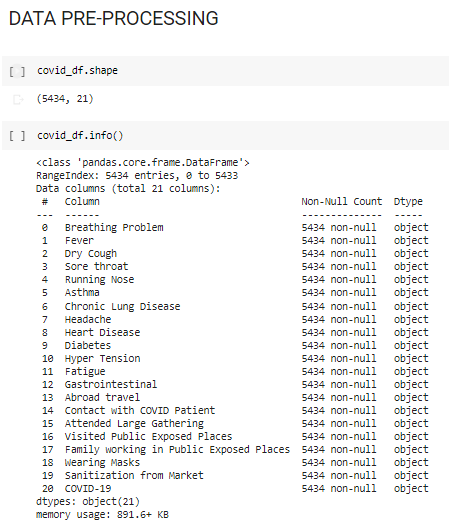
# IMPLEMENTATION

## READING DATASET



## 

## DATA PREPROCESSING



## 

## 

## 

## 

## 

## 

## 

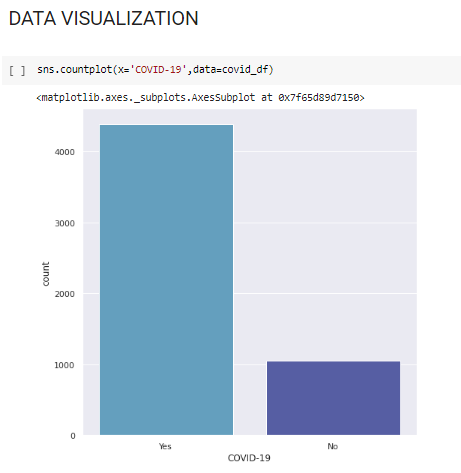
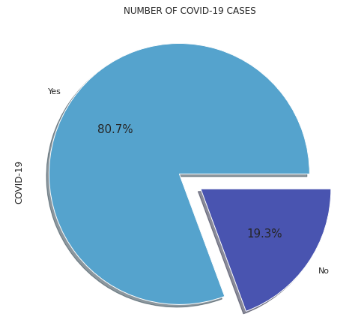
## EXPLORATORY DATA ANALYSIS

Checking null values



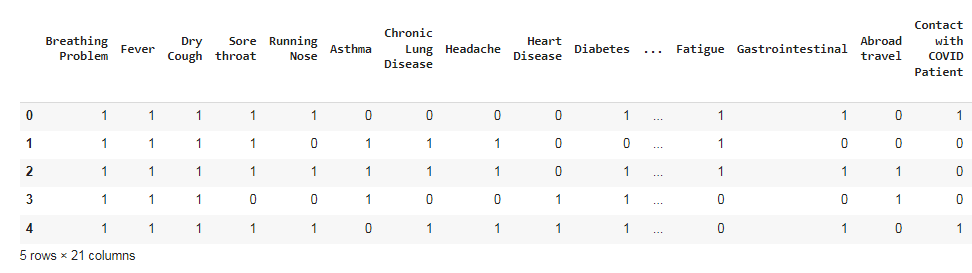
Return 0 As there are No Null values

**DATA VISUALIZATION**

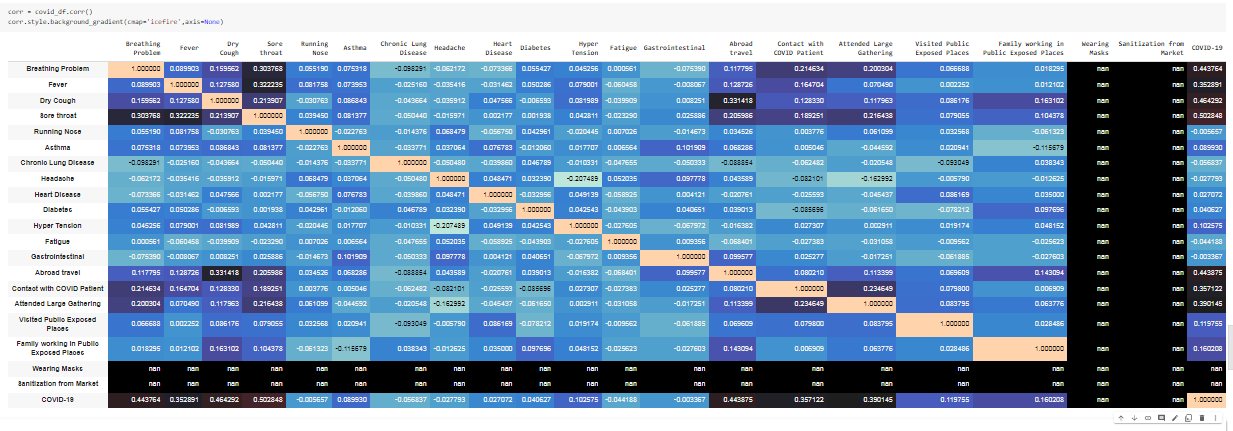
 ****

Here we are visualizing number of covid positive and negative patients

**FEATURE TRANSFORMATION**

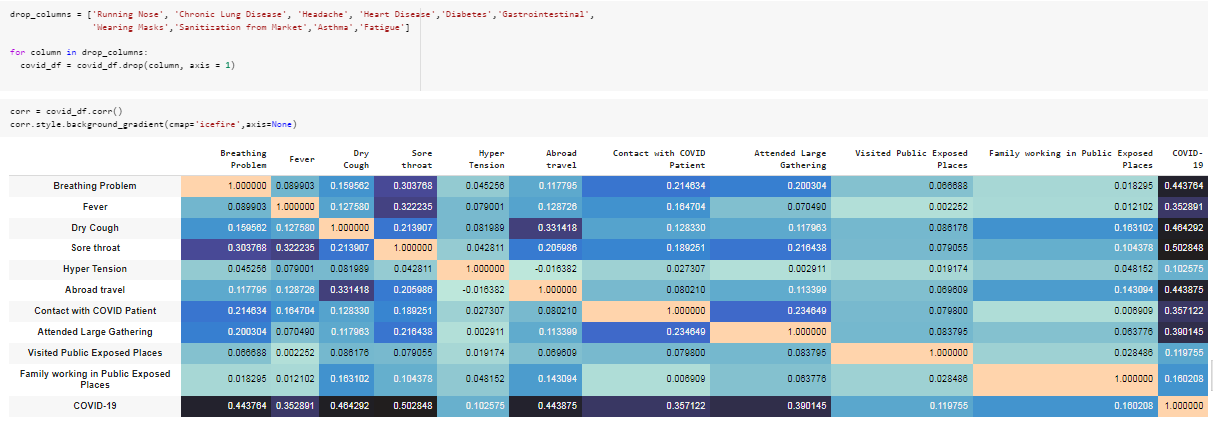


Here we convert the categorical data(yes/no) into numeric binary form(1/0).



This shows the correlation matrix of the dataset. This matrix explains that attribute class is independent of both the amount and time of the transaction was made. It is even clear from the matrix, the class of the transaction is depending on PCA applied attributes.

**FEATURE SELECTION**



Here we drop those columns those have weak correlation with our target variable(COVID-19).

## SPLITTING THE DATA FOR TRAINING AND TESTING

## 

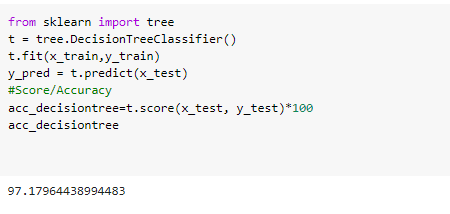
## MODEL BUILDING

## 

* LOGISTIC MODEL

## 

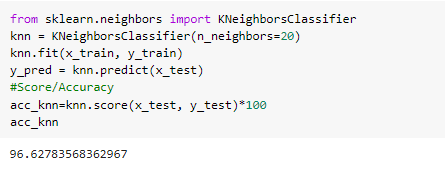
* DECISION TREE MODEL



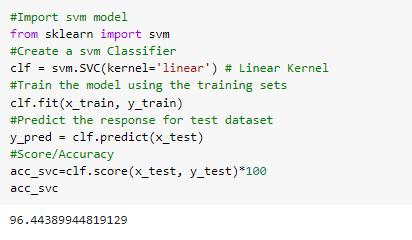
* RANDOM FOREST MODEL

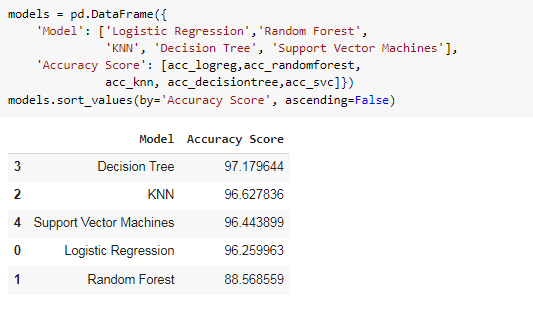
# 

* KNN MODEL



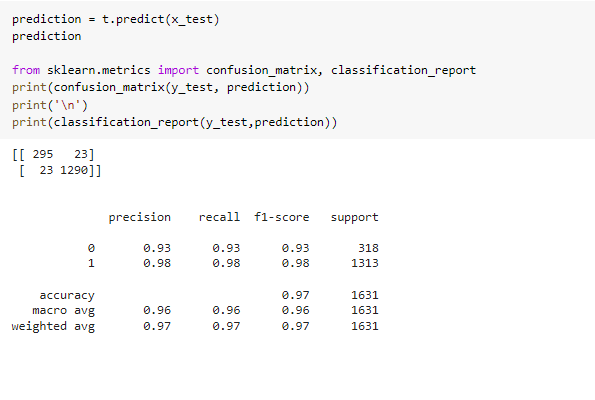
* SVM MODEL



* **COMPARISON TABLE**

We observed that the decision tree model gave the best accuracy among all models.So we choose decision tree for making prediction.

**CONFUSION MATRIX AND CLASSIFICATION REPORT**

**CONCLUSION**

The conclusion of our project is we have detected the most common severe symptoms a covid patient would have and on the basis of those severe symptoms we detected whether a patient is covid positive or not.

# 

# 

# 

# 

# 

# REFERENCE

1. <https://scikit-learn.org/>
2. <https://seaborn.pydata.org/>
3. <https://matplotlib.org/>
4. <https://pandas.pydata.org/>
5. https://numpy.org/
6. https://www.kaggle.com/
7. https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/