

**Mini-Project Competition-2022**

**PROJECT SYNOPSIS OF**

**APPLICATION USING**

**MACHINE LEARNING**

Using the concept of python modules and building an ML model to analyse and predict human illness

Semester: IV

Academic year: 2021-22

**Team members:**

Amrutha A 4VV20CI007

Vishwanath K R 4VV20CI061



**Table of contents**

|  |  |  |
| --- | --- | --- |
| **Sl. no.** | **Topic** | **Page number** |
| 1 | Introduction | 3 |
| 2 | Purpose | 4 |
| 3 | Objective | 4 |
| 4 | Approach/ Methodology | 5 |
| 5 | Flow of Model | 6 |
| 6 | Performance measure | 7-12 |
| 7 | Proposed Result | 13 |



**Introduction**

In this evolving era where everything is getting automated, Artificial Intelligence, Machine learning, Data science has made a huge leap in taking a step towards modernization. AI describes computational programs that mimic human intelligence. Machine learning is a subset of Artificial Intelligence which extracts raw data automatically. Application of machine learning is observed in many fields of area, where medical field is one amongst and our field of interest.

Machine learning with its capability of learning based on the sample data, it improves the efficiency, reliability and reduces the cost in computational processes. It can accurately and rapidly generate models through data analysis. As health data is enormous and manually handling is tedious, machine learning tools gives it hands in to us.

Machine learning does data analysis in many forms such as audio-processing (Stereo/Mono), text-processing, image processing (3D/2D processing). Our model does data analysis based on “text processing”. In this project using python modules, dataset the machine learning model can analyse and predict the human disease/illness. The prediction is based on four models i.e., Gaussian Naïve Bayes Classifier, Random Forest Classifier, Support Vector Classifier, Bernoulli and K-Fold cross-validation technique is used to evaluate the model. The mode of all the four model gives the final combined prediction of the illness.



**Purpose**

The main purpose of this project is to promote automated analysis and prediction of human illness rather human involvement to detect the illness based on the symptoms encountered by the individual. This is done by examining the testing dataset with the training dataset. The training dataset contains different set of symptoms for same disease responsible. The testing dataset contains small amount of data of all diseases with the symptoms mentioned.

**Objective**

We consult doctor for the caused illness referring to our symptoms. Doctor based on our explanation to the symptoms, predicts the disease and suggests us to undergo predicted disease related tests. With the help of machine learning if we can build a model which can predict the disease by examining the previous sample tests, then it would be helpful. So to promote such a change this project is a step. Here one can predict his/her illness by prescribing the symptoms.

We use Naïve Bayes classifier, Support Vector Classifier and Random Forest Classifier in order to predict the disease, each model predicts the disease independently. The final model combines all the three models’ predictions in order to provide the final prediction which improves the accuracy and optimises the final machine learning model.



**Methodology**

This project aims to implement an efficient machine learning model which can predict human illness based on the symptoms which the individual provides.

* Gathering the data: Data preparation is the primary step for any machine learning application. We will be using a dataset from Kaggle for this application. This dataset consists of two CSV files one for training and one for testing. There is a total of 133 columns in the dataset out of which 132 columns are symptoms and last column is the prognosis/disease.
* Cleaning the data: Cleaning is the most crucial step in this ML project. The quality of the data determines the quality and accuracy of our machine learning model. So, before feeding the data into our model it is necessary to clean the data first. In our dataset all the columns are numerical and the target column i.e., prognosis is a string type and is encoded to numerical form using a label encoder.
* Model building: After gathering and cleaning the data, the data is ready and can be used to train the ML model. We will be using this cleaned data to train the support vector classifier, Naive Bayes classifier and random forest classifier. We will be using a confusion matrix to determine the quality of the models.
* Inference: After training the three models we will be predicting the disease for the input symptoms by combining the predictions of all the models. This makes our end prediction much more accurate.

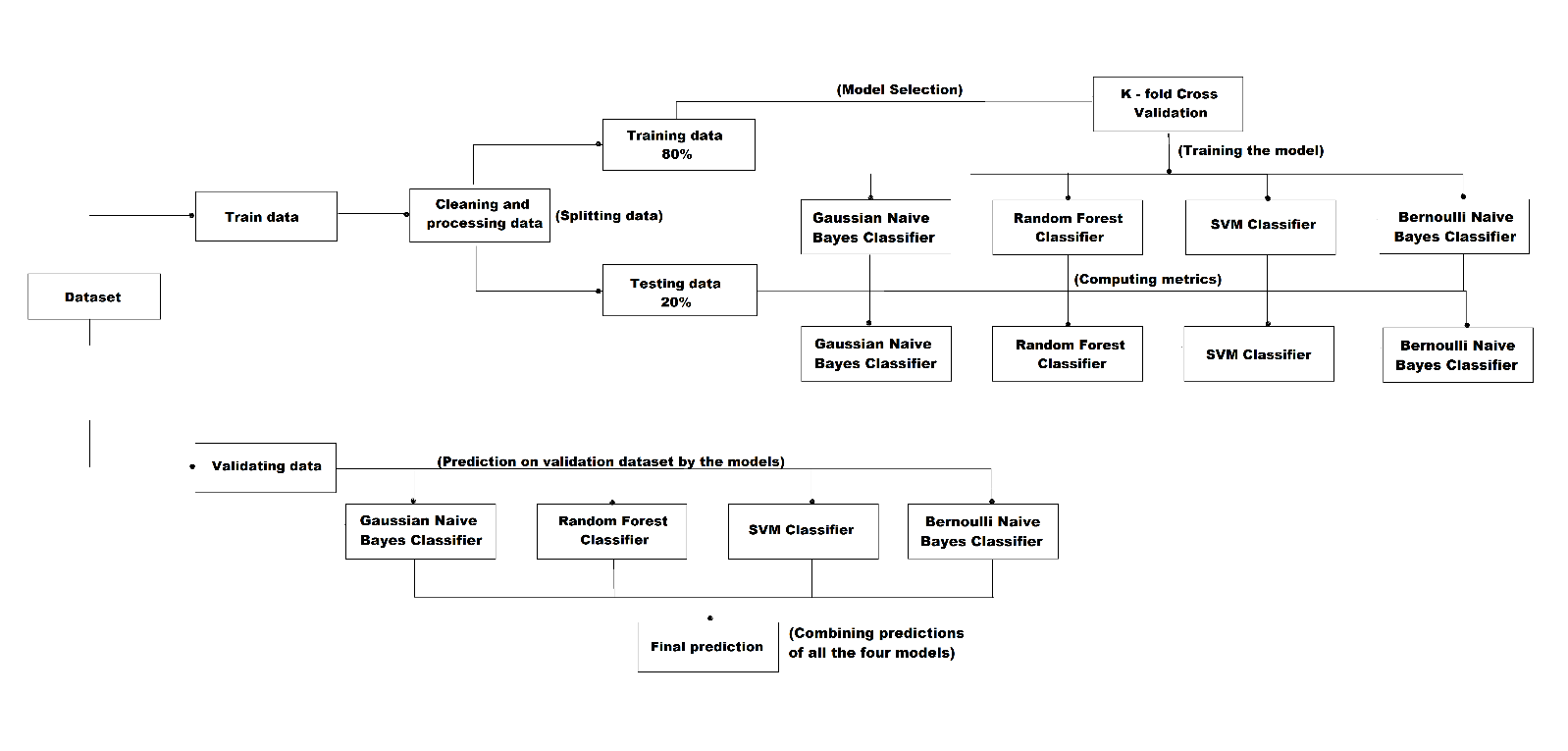
**Model Building**

* K-Fold Cross-Validation: K-Fold cross-validation is one of the cross-validation techniques in which the whole dataset is split into k number of subsets, also known as folds, then training of the model is performed on the k-1 subsets and the remaining one subset is used to evaluate the model performance.
* Support Vector Classifier: Support Vector Classifier is a discriminative classifier i.e., when given a labelled training data, the algorithm tries to find an optimal hyperplane that accurately separates the samples into different categories in hyperspace.
* Gaussian Naive Bayes Classifier: It is a probabilistic machine learning algorithm that internally uses Bayes Theorem to classify the data points.
* Random Forest Classifier: Random Forest is an ensemble learning-based supervised machine learning classification algorithm that internally uses multiple decision trees to make the classification. In a random forest classifier, all the internal decision trees are weak learners, the outputs of these weak decision trees are combined i.e., mode of all the predictions is as the final prediction.

****

* Bernoulli Naïve Bayes Classifier: is another useful naïve Bayes model. The assumption in this model is that the features are binary (0s and 1s) in nature.

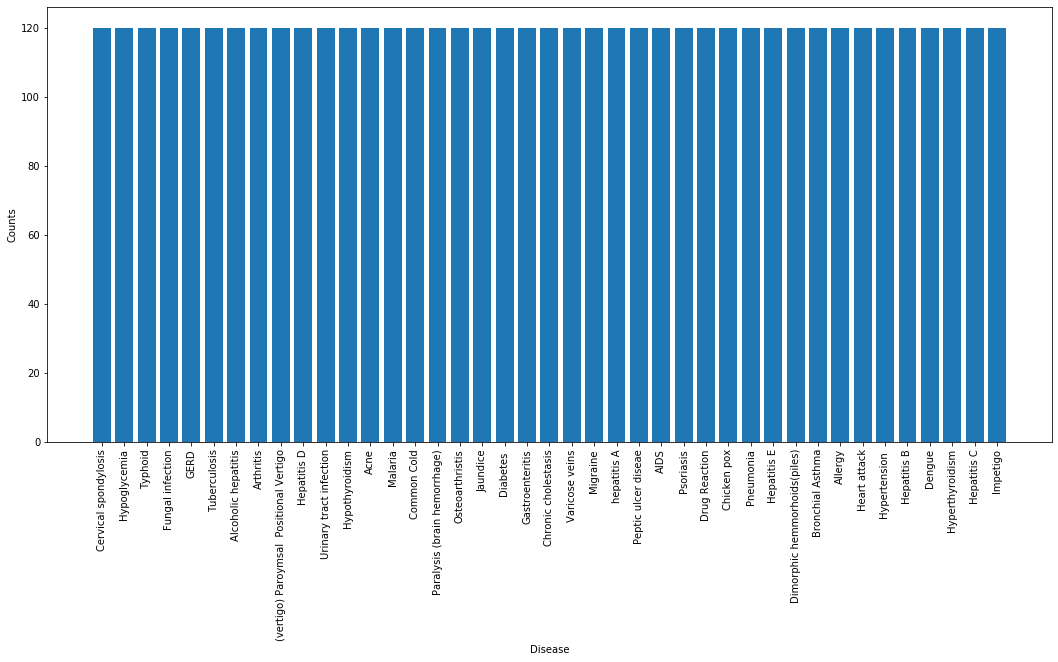
**Flow of Model**



**Performance measure:**

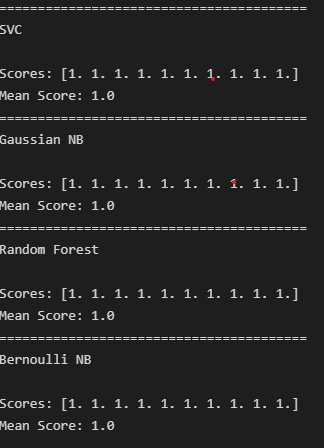
Reading the Train data set to check if the dataset is balanced or not by using:

**import** matplotlib.pyplot as plt





Using K-Fold Cross-Validation for model selection to test its accuracy by shuffling the data and providing test scores, mean score for each model:



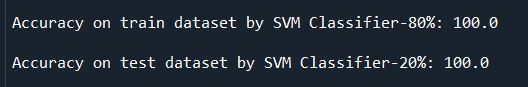


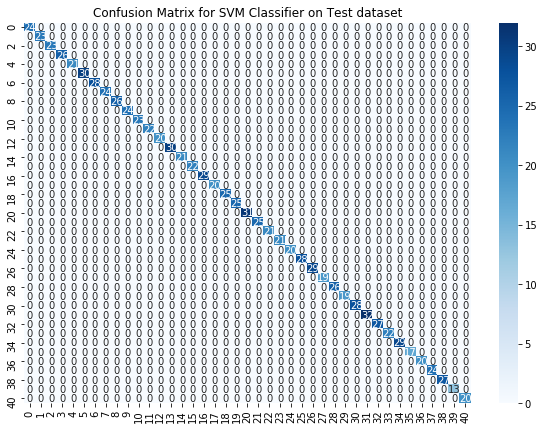
Training all the three models with the train dataset and checking the quality and measuring performance of each model using confusion matrix:

Accuracy of each model for both training-80% and testing-20% dataset from Train dataset & Confusion matrix:

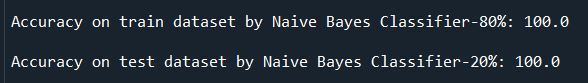
* To get the accuracy of the Model we use a built-in function: **from** sklearn.metrics **import** accuracy\_score, confusion\_matrix

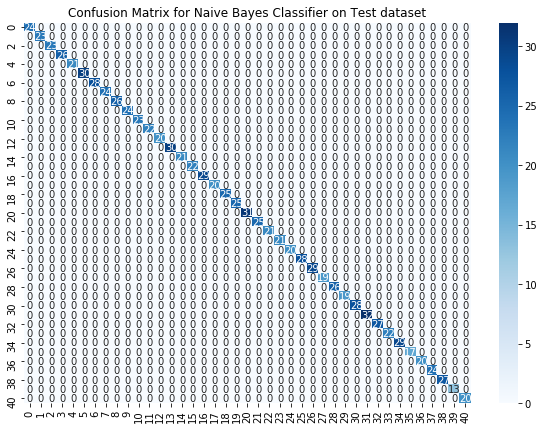
Support Vector Classifier’s accuracy & Confusion matrix:



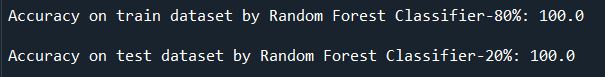


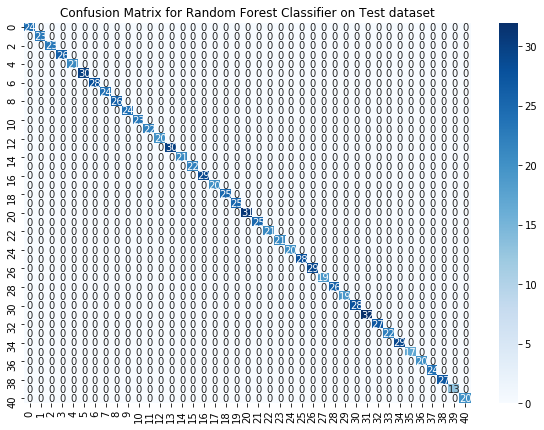
Naive Bayes Classifier’s accuracy & Confusion matrix:





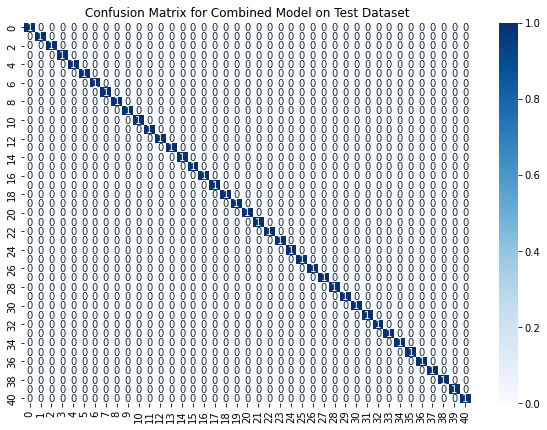
Random Forest Classifier’s accuracy & Confusion matrix:





Fitting the model on whole data and validating on the Test dataset and the Confusion matrix for final prediction (mode of all three models):





**Proposed Results**

Splitting the Train data for training (80%) and testing (20%) the model:

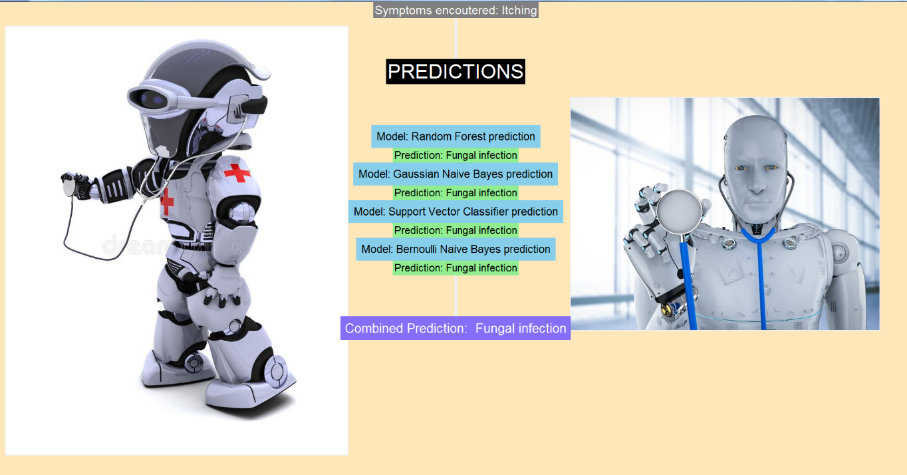


* 3936 is the number of training data and 984 is the number of testing data from Train dataset.

Acquired symptoms has to be given as the input for the Model to run:



Final Result where input is taken and the illness is predicted:

****

**Source**

Train dataset: [Disease Prediction Using Machine Learning | Kaggle](https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning?select=Training.csv)

Test dataset: [Disease Prediction Using Machine Learning | Kaggle](https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning?select=Testing.csv)