DISEASE PREDICTION

1.Definition

1.1 Introduction

Nowadays, people face various diseases due to environmental conditions and their living habits. So the prediction of disease at an earlier stage becomes an important task. If the patient is not very serious and just wants to know the type of disease, this helps. Nowadays doctors are adopting many scientific technologies and methodology for both identification and diagnosing diseases. The successful treatment is always attributed to the right and accurate diagnosis. Doctors may sometimes fail to take accurate decisions while diagnosing the disease of the patient, hence this disease prediction using predictive algorithms plays an important role in assisting such cases to get accurate results. As we all know in the competitive environment of economic development, mankind has involved so much that he/she is not concerned about health. Many of the diseases when not predicted in the early stage turn fatal. Due to increased data growth in the medical and healthcare field, the accurate analysis of medical data there is an urge for early patient care.

Advancement in healthcare aids humans to process huge and complex medical datasets and then analyze them into clinical insights. Instead of diagnosis, when a disease prediction is implemented using certain predictive algorithms then healthcare can be made smart. Some cases can occur when the early diagnosis of a disease is not within reach. Hence disease prediction can be effectively implemented. As widely said “Prevention is better than cure”, prediction of diseases and epidemic outbreaks would lead to early prevention of an occurrence of a disease. Predictive analysis with the help of efficient multiple algorithms helps to predict the disease more correctly and help treat patients.

1.2 Problem Statement

Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. As we can observe that, due to big data progress in biomedical and healthcare communities, early disease recognition, patient care, and community services are not efficiently provided. Moreover, the quality of medical data is incomplete. Also, different regions exhibit unique appearances of certain regional diseases, which may result in weakening the prediction of disease outbreaks. So we need to develop a system or we could say there is a need for the immediate

medical provision which would incorporate the symptoms collected from multisensory devices and other medical data and store them into a healthcare dataset.

The main focus of our project is to develop a model for early disease prediction and smart healthcare.

1.3 Project Purpose

The main purpose of this project is to predict the disease of a patient using all the general information and the symptoms with utmost accuracy. Using this information, we'll compare this data with our previous datasets of the patients and predict the disease he/she has been through. If the prediction of disease is done in the early stages and all other necessary measures, the disease can be cured else may turn into fatal.If the health industry adopts this project,the work of doctors can be reduced and they can easily predict the disease of the patient. The general purpose of this Disease Prediction is to provide predictions for various general occurring diseases that when unchecked or sometimes ignored can turn into fatal diseases and cause a lot of problems to the patient. The health industry is information rich and knowledge poor and it is a vast industry in which a lot of work needs to be done.So, with the help of these algorithms, techniques, and methodologies,we have implemented this project which helps the people in need.

1.4 Performance metric

The metric that we will use for our problem is K-fold accuracy score. Accuracy is defined as the number of samples correctly classified to the total number of samples.

2.Algorithms and Techniques

Four supervised learning approaches are selected for this problem for developing model.These algorithms are chosen such that their approaches are fundamentally different from each other, so that we can cover a wide spectrum of possible approaches. The algorithms and techniques used are as follows:

1.Naive Bayes
2.Random Forest
3.Decision Tree

4.Logistic Regression

To learn the best model the algorithm used is as follows:
Cross-validation using K-fold

3.Methodology

3.1 Data Preprocessing

Data Preprocessing is necessary to bring data into good shape, remove any abnormalities, or modify any characteristics, before feeding it into an algorithm.
The following points are noted regarding our data:

1.The dataset does not have any missing value
2. All the input features have different ranges, so we need to perform normalization to scale all features in the range of 0-1.

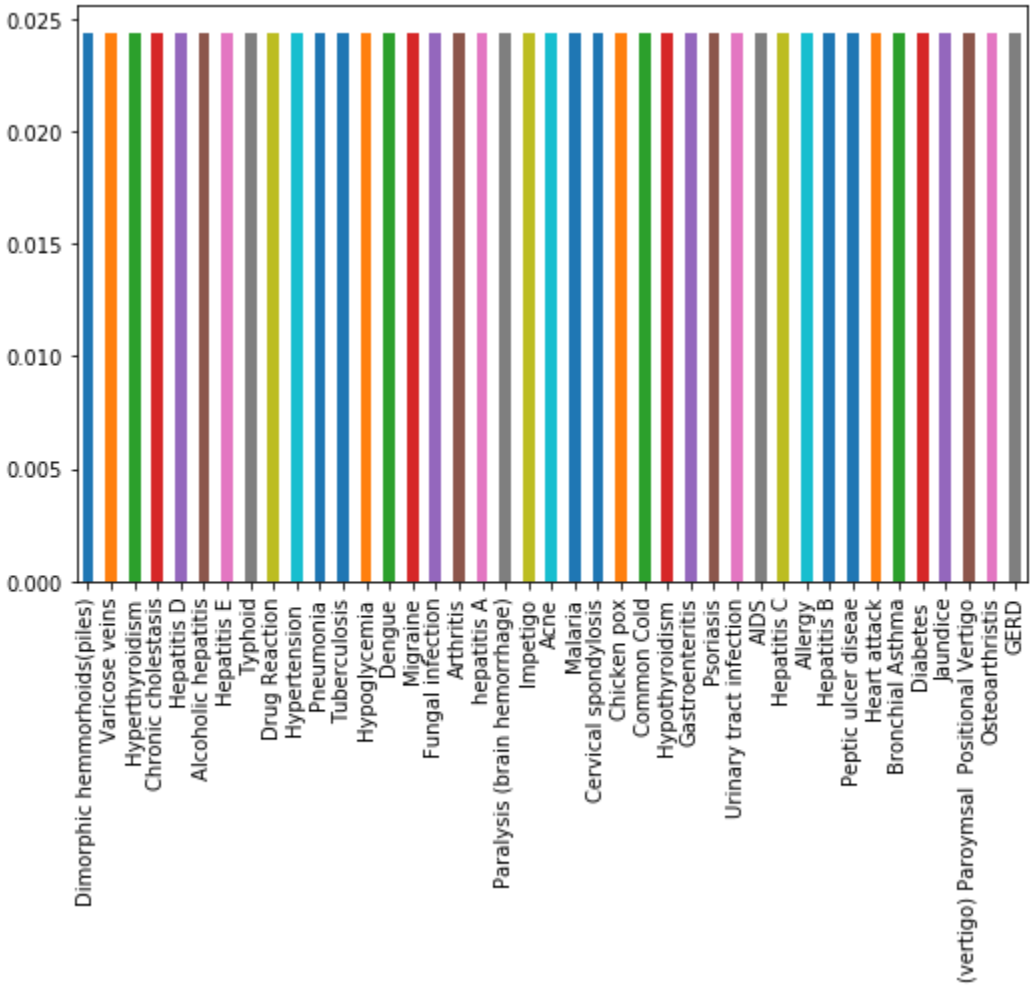
3.1.1 Data Normalization

From analyzing the data, we identify that we need to perform scaling on the features to bring them down in the range of 0 to 1. This way, the classifier will treat all features equally.
The following formula is used to normalize the data:

$$Xnorm = (Xcurrent - Xmin)/(Xmax - Xmin)$$

We perform scaling so that any supervised learning algorithm is not biased towards any feature.

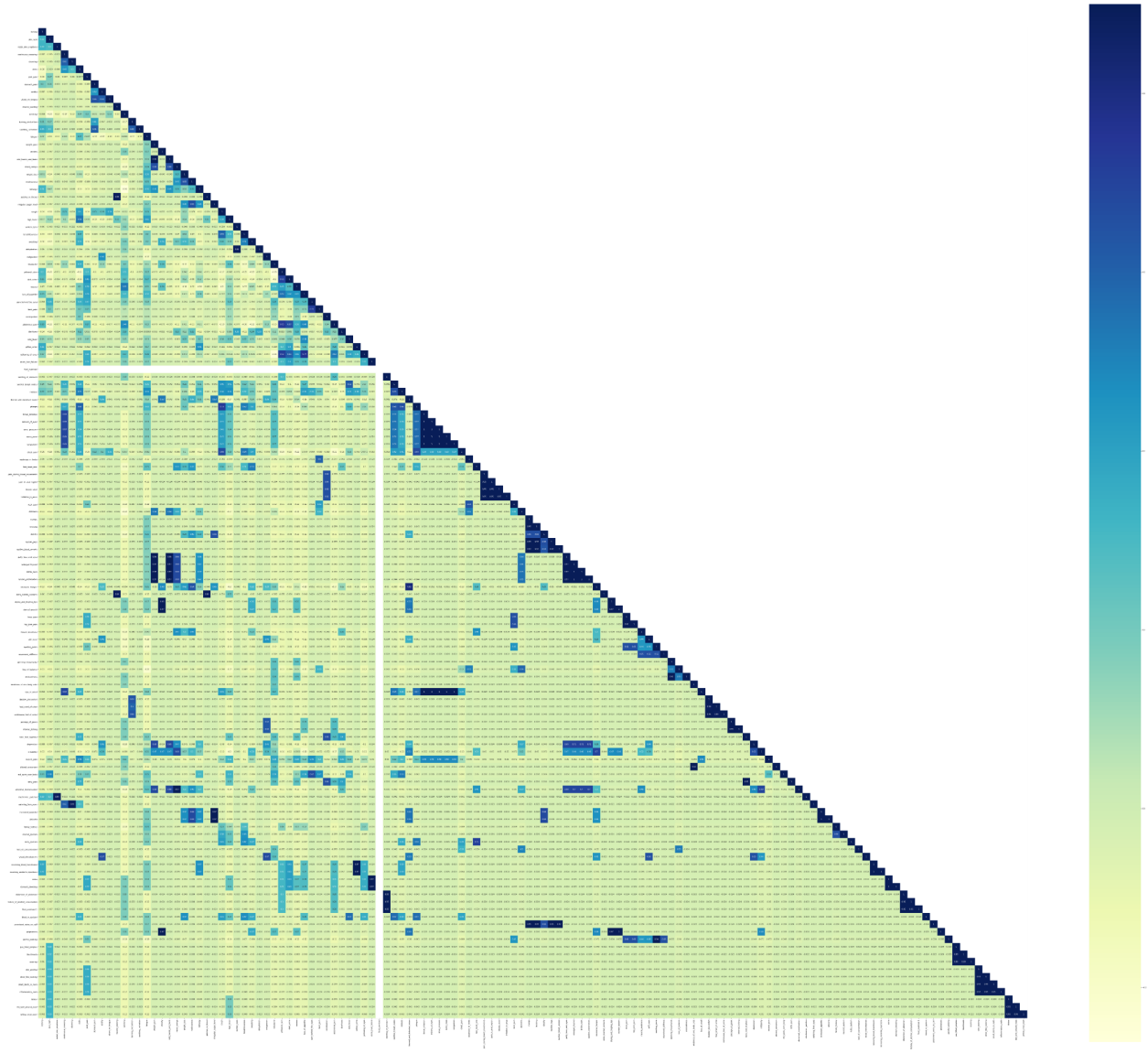
PLOT:



3.2 Correlation

Correlation analysis is used to quantify the degree to which two variables are related. Through the correlation analysis, you evaluate the correlation coefficient that tells you how much one variable changes when the other one does.

PLOT:



3.3 Implementation

3.3.1 Splitting data

The reason we perform splitting is to test the trained model over samples which it has never seen before. This way, we make sure that the model has extracted classification patterns from the training samples and has not memorized them. For the given problem, to split our data, we use a function train_test_split available in the cross_validation module of sklearn library.

This function does two tasks that are important in the current context.

1. It shuffles the dataset so that both training and test sets have nearly equal numbers of samples from both classes.

2. After shuffling, it performs the split.

We can specify what fraction of the total data to be included in the training or test set. After splitting, the function returns four lists, comprising input features of the training set, input features of the test set, target labels for the training set, and target labels for the test set. We split such that the training set has 67% of samples, while the test set has 33% of samples.

3.3.2 Cross-Validation using K-fold

We define a method called 'train_split' that takes as input the following parameters: learner, sample_size, X_train, y_train, X_test, y_test. It returns the accuracy score on training and test sets. The function fits the 'learner' on the training data of size defined by 'sample_size', and computes the time spent on training.
We train the four classifiers that we have chosen, each with varying values of K as 2,4,6,8,10 so that it can be seen how performance varies with varying sizes of train sets.
The K-Fold accuracy score result is as follows:

Training:

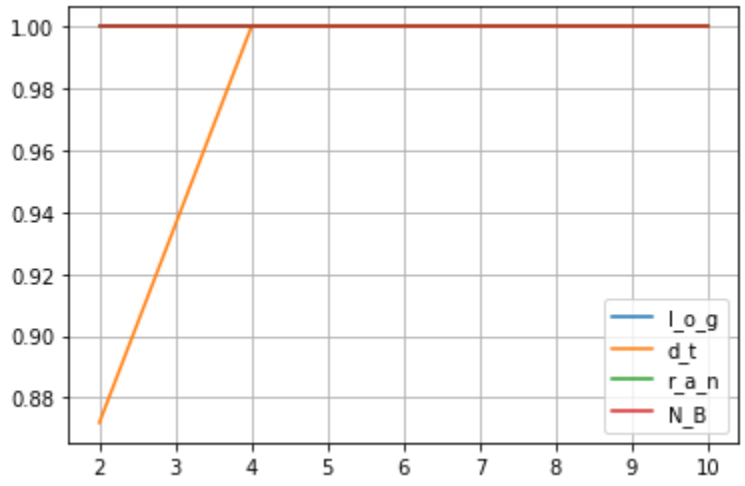
Model	**k=2**	**k=4**	**k=6**	**k=8**	**k=10**
'log'	1.0	1.0	1.0	1.0	1.0
'dt'	1.0	1.0	1.0	1.0	1.0
'ran'	1.0	1.0	1.0	1.0	1.0
'NB'	1.0	1.0	1.0	1.0	1.0

Testing:

Model	**k=2**	**k=4**	**k=6**	**k=8**	**k=10**
'log'	1.0	1.0	1.0	1.0	1.0
'dt'	1.0	0.869512	1.0	1.0	1.0
'ran'	1.0	0.998780	1.0	1.0	1.0

‘NB’	1.0	1.0	1.0	1.0	1.0

PLOT:



4. Results

After consideration of four classifier models for the given problem, viz., Naive Bayes,Decision tree,Random Forest,Logistic Regression. We evaluated the performance of four classifiers using accuracy metric .All the models are validated using cross-validation with K-fold test by varying the sizes of test data and the accuracy scores were obtained.Based on the accuracy scores on training and testing data,we have observed that logistic regression is the best suitable model.

5.Sample output of disease prediction:

7: chills

10: acidity

16: fatigue

27: high_fever

36: nausea

40: constipation

41: abdominal_pain

96: toxic_look_(typhos)

102: belly_pain

Enter your name :abc
Enter your age:20
Enter the Serial no.s in which your Symptoms exist: 7 10 16 27 36 40 41 96 102

Name of the infection = Typhoid , confidence score of : = 93.46987614886676 %
Name = abc , Age : = 20

5.Conclusion:

- This project-Disease prediction is to provide the users prediction for various and generally occurring diseases that when unchecked or sometimes ignored can turn into fatal disease.

- It also plays a major role if adopted by a healthcare sector in order to reduce the amount of work of doctors and can easily predict the disease of patients.