***Heart Disease Prediction with Machine Learning Approaches***

***Abstract:***

The heart is the most essential or crucial portion of our body. The heart is used to maintain and conjugate blood in our body. There are a lot of cases in the world related to heart disease. People are leading to death due to heart disease. Various symptoms like chest pain, fasting of heartbeat, and so on are mentioned. The healthcare industry found a large amount of data. This paper gives the idea of predicting heart disease using machine learning algorithms. Here, we will use various machine learning algorithms such as support vector classifier, random forest, knn, naïve Bayes, decision tree, and logistic regression. The algorithms are used on the basis of features and for predicting heart disease. This paper uses different machine learning algorithms to compare the accuracy among them.

***Keywords:***

Coronary artery disease, Decision tree, K nearest neighbor; Machine Learning, Support vector, Accuracy, Logistic Regression, Naïve Bayes.

***Introduction***

Heart disease is a term that damages our health badly. Every year too many people die due to heart disease. Due to the weakening of the heart muscle, heart disease can occur. Heart disease can be defined as the breaking of the heart to pump blood. Coronary artery disease or Coronary heart disease is another term for heart disease. (Coronary artery disease) CAD can arise due to insufficient blood supply to arteries.

The most common indications of heart attack are

• Chest pain.

• Shortness of breath.

• Sweating and Fatigue.

• Nausea, Indigestion, Heartburn, or Stomach pain.

• Pressure in the upper back pain that spreads to the arm.

***Types of heart disease are:***

• Coronary artery disease (CAD).

• Angina pectoris.

• Congestive heart failure.

• Cardiomyopathy.

• Congenital heart disease.

On the basis of the above factors, this paper gives the best try to predict the risk of heart disease. Related to heart disease prediction, a huge amount of work has been done using machine learning algorithms by many authors. The aim of this paper is to achieve better accuracy so that it can predict the chances of a heart attack. The patient risk level is classified using data mining techniques such as K nearest neighbor, Decision tree, Random forest, Support vector classifier, Logistic Regression, and Naïve Bayes. Some risk factors are Age, Sex, Blood pressure, Cholesterol, Chest pain, Heart rate, and so on.

In this paper, the supervised machine learning concept is used for making the predictions. The various machine learning algorithms such as knn, random forest, support vector machine, decision tree, naïve Bayes, and logistic regression are used to make the predictions using the heart disease dataset.

***Methodology***

In this paper, we have used our dataset to apply different machine learning algorithms to identify if a person has heart disease or not. Then, we will handle the missing values in the dataset, visualize the dataset, and observe the accuracy obtained by different machine learning algorithms. The machine learning algorithms used are defined below.

***Data Collection***

In this paper, the dataset is obtained from the Cleveland Heart Disease database at UCI Repository. There are 14 attributes in the dataset.

The description of the dataset is given as follows:

1) Age: describes the age of a person.

2) Sex: describes the sex of a person; 1 for male, 0 for female.

3) Cp: describes the chest pain type in a person ( 1 for angina, 2 for typical angina, 3 for non-angina, 4 for asymptomatic).

4) Trestbps: describes the resting blood pressure.

5) Chol: describes the serum cholesterol.

6) FBS: describes the Fasting Blood Sugar ( 1 for true & 0 for false).

7) Restecg: describes the resting electro-graphic results( 0 for normal, 1 for ST-T wave abnormality, 2 for left ventricular hypertrophy).

8) Thalach: describes the maximum heart rate.

9) Exang: describes exercise-induced angina

10) Oldpeak: describes the depression raised by exercise relative to rest.

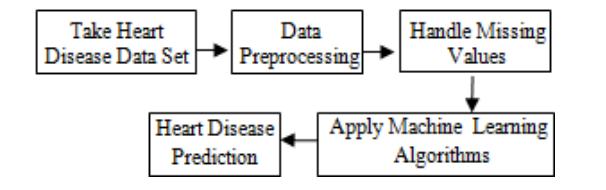
11) Slope: describes the slope of the peak exercise ST segment (1 for up-sloping, 2 for flat, 3 for down-sloping).

12) Ca: describes the number of blood vessels.

13) Thal: describes that feature (3 for normal, 6 for fixed defect, 7 for reversible effect).

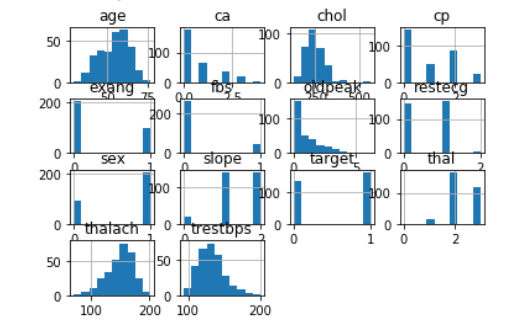
14) Target: describes the target class (0 for no heart disease, 1 2 3 4 for having heart disease).

***Flow Diagram***



**Histogram**

The histogram is the best and easiest way to visualize the data because it only takes a single line of code to make the plots. Let’s take a look at the plots. Before applying any machine learning algorithms we will have to look for categorical variables. The target class is used for describing whether a person is having heart disease or not.

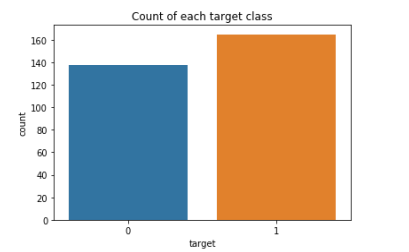


***Exploratory Data Analysis:***

Exploratory Data Analysis (EDA) is an approach to analyze the data sets to describe their main highlights using visual methods. There are many different methods for conducting exploratory data analysis out there, so it can be hard to know what analysis to perform and how to do it properly. EDA, feature selection, and feature engineering are often tied together and are important steps in the machine learning journey.

Bar plot for target class with different features:

It is very important that the dataset we are using should be pre-processed and cleaned. This graph shows the count of each target class.



***Machine Learning Algorithms Logistic Regression:***

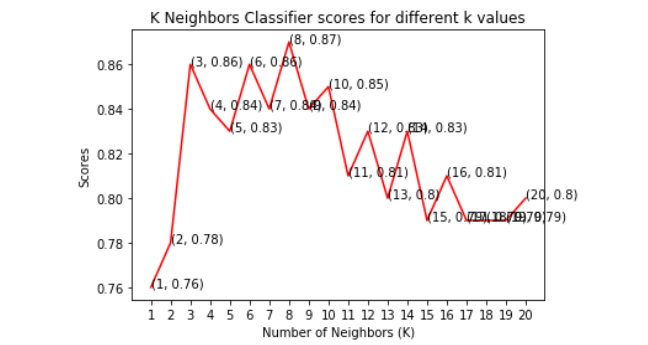
Logistic regression is a supervised learning algorithm used to predict the binary form of a target variable. It is the easiest and simplest algorithm used in machine learning that can be used for various problems such as disease prediction, cancer detection, and so on. In this paper, we achieved an accuracy of 84% by using this model.

***Naïve Bayes Classifier***:

Naive Bayes is a statistical classifier. It is based on Bayes’ theorem. A naïve Bayesian classifier has comparable performance with a decision tree and other selected classifiers. The computation cost can be reduced greatly. It is easy to implement. In this paper, we achieved an accuracy of 80% by using this classifier.

***K Nearest Neighbors Classifier:***

K Nearest Neighbors is a non-parametric method used for classification. It is a lazy learning algorithm where all computation is deferred until classification. It is also an instance-based learning algorithm, where the function is approximated locally. This algorithm is used when the amount of data is large and there are non-linear decision boundaries between classes. KNN explains a categorical value using the majority votes of nearest neighbors. Not only for classification, KNN can be used for function approximation problems.



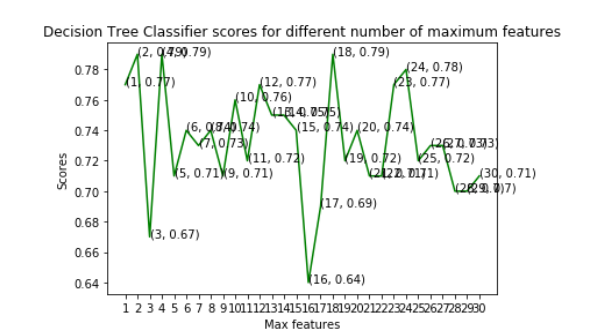
This graph shows that the maximum accuracy achieved by K neighbors classifier is 87%

***Support Vector Classifier:***

SVM (Support Vector Machine) is a supervised machine learning algorithm that can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). This classifier separates data points using a hyper plane with the largest amount of margin. Support vectors are the data points which are closest to the hyper plane. There are several kernels on which the hyper plane can be decided. This paper mainly focuses on four kernels namely linear, polynomial (poly), radial basis function (rbf) and sigmoid. This type of classifier uses less memory because they use a subset of training points in the decision phase.

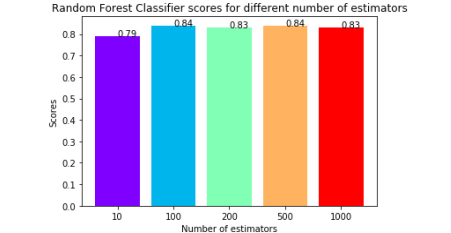
***Decision Tree Classifier***

This classifier falls under the category of supervised learning. It can be used to solve regression and classification problems. We can use this algorithm for issues where we have continuous but also categorical input and target features. It is the most effective machine learning algorithm used for describing the trees in a graphical manner.

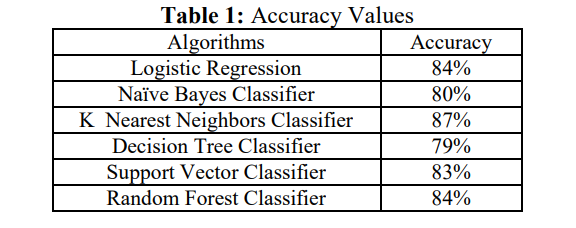


***Random Forest Classifier:***

Random forest is a supervised learning algorithm. It can be used for classification and regression. It is simple and easy to implement. A forest is comprised of trees. This classifier creates decision trees on randomly selected data samples, gets predictions from each tree, and selects the best solution by means of voting. The random forest is composed of multiple decision trees. It creates a forest of trees.



This graph shows that the maximum accuracy is 84% and was obtained for both 100 & 500 tree



***Table 1*** shows that the K Nearest Neighbors Classifier gives the best accuracy with 87% in comparison with the other machine learning algorithms used in this paper. Because KNN algorithm is based on feature similarity and is one of the most famous classification algorithms as of now in the industry simply due to its simplicity and accuracy. K nearest neighbors is a simple algorithm that stores all the accessible cases and classifies new cases based on a similarity measure.

***Conclusion and Future Work***

This paper involves the prediction of the heart disease dataset with proper data processing and implementation of machine learning algorithms. In this paper, we use six machine learning algorithms for prediction. Among all the machine learning algorithms used in this paper, the highest accuracy is achieved by the K Nearest Neighbors Classifier with 87%. This paper shows that machine learning algorithms can be used to predict heart disease easily with different parameters and models. Machine learning is very useful in prediction, solving problems, and other areas. Machine learning is an effective way to solve problems in different areas too.

**TEAM MEMBERS**

DHAIRYA GOEL (E22BCAU0123)

JADHAV CHANDRAMSH PATEL(E22BCAU0006)

***References***

[1] Avinash Golande, Pavan Kumar T, (June 2019): Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering (IRTE), ISN: 2277-3878, Volume-8, Issue-1S4. [2] A. Sahaya Arthy, G.Murugeshwari, (April 2018): A survey on heart disease prediction using data mining techniques. [3] Amita Malav, Kalyani Kadam, (2018): “A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K – Means”, International Journal of Pure and Applied Mathematics. [4] Benjamin EJ et.al, (2018): Heart Disease and Stroke Statistics At-a-Glance. [5] DhafarHamed, Jwan K.Alwan, Mohamed Ibrahim, Mohammad B.Naeem, (march – 2017): “The Utilization of Machine Learning Approaches for Medical Data Classification” in Annual Conference on New Trends in Information & Communications Technology Applications