

Prediction of Heart Disease using Machine Learning

Ankit Singh

Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh, India

ABSTRACT

Article Info

Volume 6, Issue 4

Page Number: 150-166

Publication Issue :

July-August-2020

Article History

Accepted : 10 July 2020

Published : 22 July 2020

Cardiovascular Disease is the leading cause of death (Approximately, 17 million people every year) in the all the area of the world. Prediction of heart disease is the critical challenge in the area of the clinical data analysis. The objective of paper is to build the model for predicting the Heart Disease using various machine learning classification algorithm. Classification is a powerful machine learning technique that is commonly used for prediction. Some of the classification algorithm are Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest Classifier, KNN. This paper investigate which algorithm is used for the improving the accuracy in the prediction of heart disease. And, a comparative analysis on the accuracy and mean squared error is to done for predicting the best model. The result of the study indicates that KNN algorithm is effective in predicting the model with the accuracy of the 85.71% and having a very low mean squared error.

Keywords : Heart Disease, Machine Learning, Classification Algorithm, CVD (Cardiovascular Disease), Support Vector Machine

I. INTRODUCTION

The main objective is the prediction of the heart disease using the machine learning technique. Heart Disease describe the extent of circumstances of your heart. According to World Health Organization, the CVDs are number 1 death globally: more people die annually from CVDs then from any other causes, and an estimate 17.9 million people died from CVDs in 2016, representing 31% of all global deaths [1]. Over one crore annual death is reported in India and CVDs causes 20.3% death in men and 16.9% death in women [2].

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical

inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood glucose, raise blood lipids, and overweight and obesity and can lead to a heart attack, stroke, heart failure and other complaints [1]. There are many different types of heart attack :Acute Coronary Syndrome(ACS), STEMI(ST-elevation myocardial infraction), NSTEMI (Non-ST-elevation myocardial infraction), Myocardial Infraction (MI), Coronary Occlusion [2]. So, to determine the odds of heart disease is bit difficult based on the risk factors. Due to such constraints, scientist have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Hence, our objective is the prediction of heart disease by applying the Machine Learning algorithm on the patient's dataset.

II. Literature Survey

Several related works are conducted on the patient's datasets using Machine Learning technique. Many of them are show a good classification result:

Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna [3] has given a paper named 'Prediction of Heart Disease using Machine Learning Algorithm'. This paper signifies that, during the small dataset Naïve Bayes gives the accurate result and when the dataset is large decision tree gives the accurate results.

Senthilkumar Mohan, Chandrasegar Thirimalai, and Gautam Srivastava [4] has given a paper named 'PREDICTION OF HEART DISEASE USING HYBRID MACHINE LEARNING TECHNIQUE'. This paper produce a enhanced performance level with an accuracy of 88.7% through the hybrid random forest with the linear model (HRLFM).

Sellappan Palaniyappan, Rafiah Awang [5] made use of decision tree Naïve Bayes, Decision tree, Artificial Neural Networks to build Intelligent Heart Disease Prediction Systems (IHDPS). To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms. By providing effective treatments, it also helps to reduce treatment costs. Discovery of hidden patterns and relationships often has gone unexploited. Advanced data mining techniques helped remedy this situation.

Benjamin EJ et.al [6] says that there are seven key factors for heart disease such as smoking, physical inactivity, nutrition, obesity, cholesterol, diabetes and high blood pressure. They also discussed the statistics of heart disease including stroke and cardiovascular disease.

C.Beulah Christalin Latha, S.Carolin Jeeva [7] has given a paper named 'Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques'. This paper study indicates that the ensemble technique such as boosting and bagging, are effective in improving the prediction accuracy of weak classifiers, and exhibit satisfactory performance in identifying risk of heart disease. A maximum increase of 7% accuracy for weak classifiers was achieved with the help of ensemble classification.

Animesh Hazra et.al, [8] discussed in detail the cardiovascular disease and different symptoms of heart attack. The different types of classification and clustering algorithms and tools were used. V.Krishnaiah, G.Narsimha, N.Subhash Chandra [9] presented an analysis using data mining. The analysis showed that using different techniques and taking different number of attributes gives different accuracies for predicting heart diseases

III. Methodology:

3.1 Description of Dataset

The Dataset has been taken from the Kaggle named 'Heart Disease UCI'. The Dataset contain 14 attributes and 303 instances. There are 8 nominal and 6 numeric attributes. Description of attributes given in Table 1:

Table 1 : Description of Dataset, category and Ranges

<i>Attributes</i>	<i>Explanation</i>	<i>Category</i>	<i>Ranges</i>
<i>age</i>	Age of the Patient	Numeric	29 to 79
<i>sex</i>	Woman or Man	Nominal	0,1
<i>cp</i>	Chest Pain Type	Nominal	0,1,2,3
<i>trestbps</i>	Resting blood sugar in mm Hg	Numeric	94 to 200
<i>chol</i>	Serum Cholesterol in mg/dL	Numeric	126 to 564
<i>lbs</i>	Fasting blood sugar in mg/dl (> 120 mg/dl result in True else False)	Nominal	0,1
<i>restecg</i>	Resting electrocardiographic results	Nominal	0,1,2
<i>thalach</i>	Maximum heart rate achieved	Numeric	71 to 202
<i>exang</i>	Exercise Induces Angina	Nominal	0,1
<i>oldpeak</i>	ST depression induced by exercise relative to rest	Numeric	1 to 3
<i>ca</i>	No. of major vessel (0 – 3) coloured by flurosopy	Numeric	0 to 3
<i>Slope</i>	Slope of peak exercise ST segment	Nominal	1,2,3`
<i>thal</i>	3 – Normal, 6 – Fixed Defect,7 – Reversible Defect	Nominal	3,6,7
<i>Target</i>	Having Disease or Not	Nominal	0,1

Patient are categorized as Male and Female represented by 1 and 0 and within the age of 29 years to 79 years. Four types of chest pain can be considered as indicative of heart disease. (1): Angina is caused by reduced blood flow to the heart muscle. (2): Angina is a chest pain that occurs during mental or emotional stress. (3): Non-angina chest pain may because due to various reasons and may not often be due to actual heart disease. (4): Asymptomatic, may not be a symptom of heart disease

The next feature is *trestbps* indicate the resting blood sugar, *chol* indicate the cholesterol level, *lbs* is a fasting blood sugar, if *lbs* > 120 mg/dL represent 1 otherwise 0. *restecg* indicate resting electrocardiographic result. *thalach* indicate maximum heart rate, *exang* indicate the exercise induces angina which categorized as 0 and 1, *oldpeak* is the ST depression induced by exercise relative to rest, *ca* is no. of vessel colored by flurosopy, *slope* indicate slope of the peak exercise segment , *thal* is the duration of exercise test in minutes.

- The dataset consist of 303 instances and 14 features each.
- “Outcome” of the feature we are going to predict 0 means No heart disease, 1 means heart disease

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age          303 non-null int64
sex          303 non-null int64
cp           303 non-null int64
trestbps     303 non-null int64
chol         303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB

```

- There are no null values in the dataset (no-empty or missing value).
- The dataset contains only int and float values.

The histogram of the dataset is represented as:

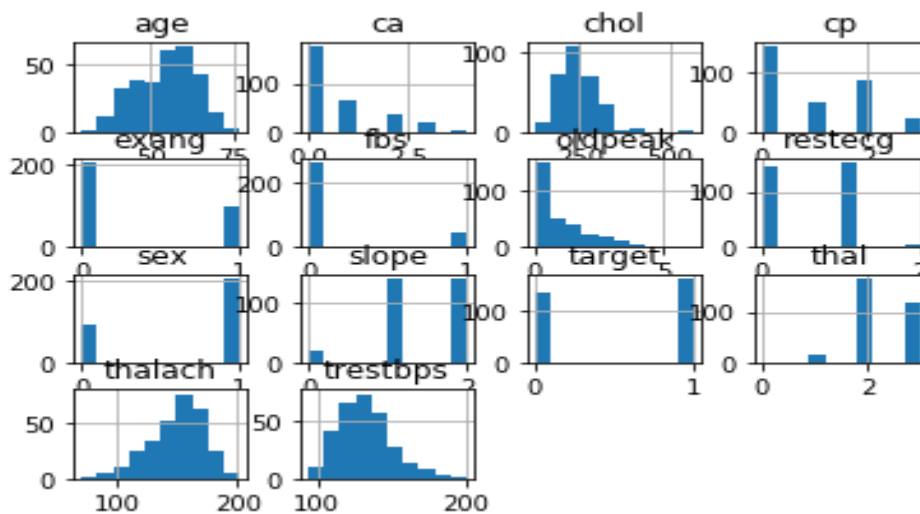


Figure 1: Histogram of Dataset

In the Histogram, it shows that feature and labels are distributed along different ranges, which further confirms the need for scaling. Next, wherever you see the discrete bars, it basically means that each of these is actually a categorical variable.

Now, People age who is suffering from disease or not:

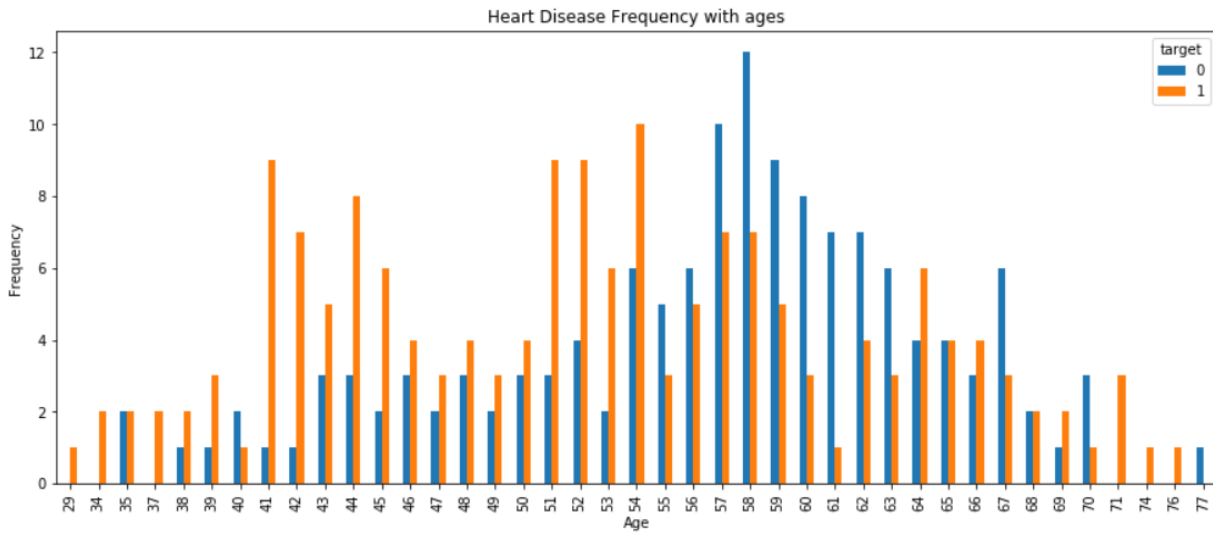


Figure 2 : Frequency of Heart Disease with Age

In figure 2, the bar graph represented at the particular age what is the frequency of people influenced with the heart disease or not, it shows that age between 35 to 55 are generally more influenced with the heart disease orange bar (represent '1') rises high between 35 to 55, while people between the range of 55 to 65 are less influenced with the heart disease, as purple bar (shows 0) is rises high between that range.

Also, Bar Plot for target (the no. of people who is affected from heart disease or not) is represented as:

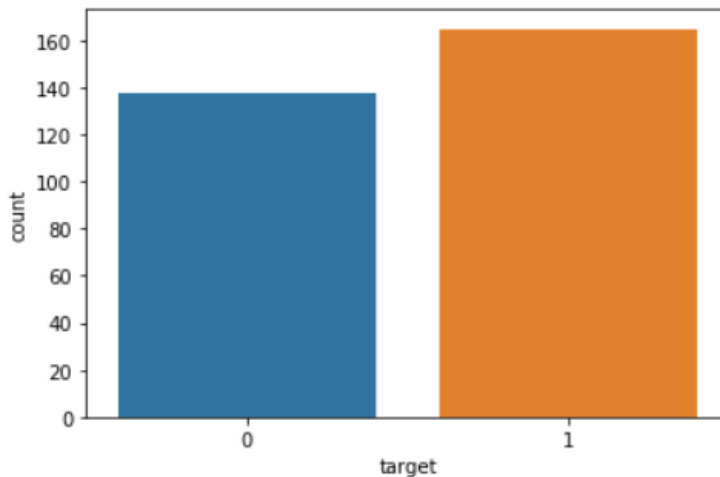


Figure 3: Count plot of No. of person affected with Heart Disease

In figure 3, bar graph shows that the no. of people influenced with heart disease are 165 while which is not influenced with the heart disease are 138. So that, the affected people are more in comparison of non-affected people which represented in the count plot.

The following graph represented Age vs Maximum Heart Rate

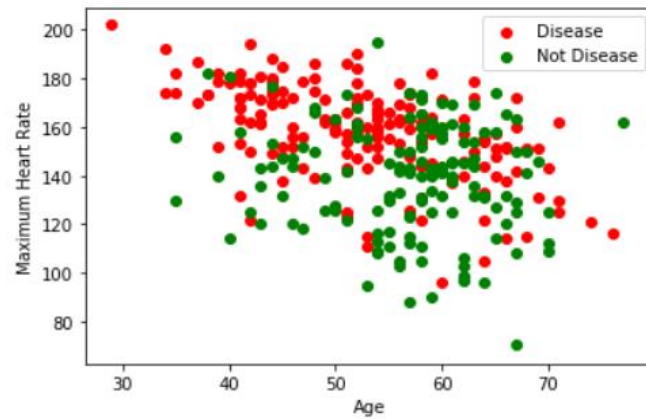


Figure 4: Maximum Heart Rate vs Age having heart disease or not

Figure 4, justify that the people who have a maximum heart rate are generally influenced with the heart disease, in the scatter plot it shows generally, if maximum heart rate > 150 and within the age of 35 and 55, there are more red dots represent the heart disease, the people have a more chances of heart disease. So, the heart rate is the most important factor of influencing the hart disease.

3.2 Proposed Work

The following dataset is carried through the data pre-processing then through the feature and extraction and then training and testing set went through the Classification algorithm. The following flowchart is given as :

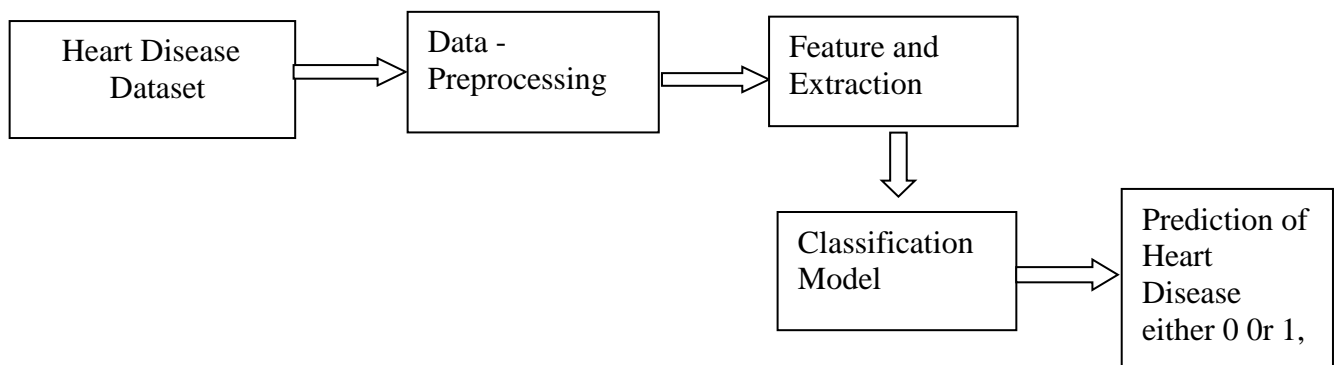


Figure 5: Flow chart of Proposed Model

Here, in this model, 6 different machine learning classification algorithms is used, which is used to predicting the model of Heart Disease.

3.3 Machine Learning Algorithm

3.3.1 Logistic Regression

Logistic Regression is Supervised Learning Algorithm. It is a statistical model that, it is in basic form uses a logistic function to model a binary dependent variable. Generally, Logistic Regression is used when dependent variable(target) is categorical which is of three types:

1. Binary Logistic Regression
2. Multinomial Logistic Regression (three or more categories without ordering)
3. Ordinal Logistic Regression (three or more categories with order)

It is fast and relatively uncomplicated. It models the probability that response falls into specific category. Activation function (which is used to fit the model) like sigmoid function, tanh function, ReLU function or Leaky ReLU function which is used to solve the model of Logistic Regression.

3.3.2 Decision Tree Classifier

It is a Supervised Learning Algorithm where data is continuously split according to the certain parameter. Decision Tree Classifier creates a classification model by building a decision tree. There are two main types of

Decision Tree Classifier:

1. Classification Tree (Categorical)
2. Regression Tree (Continuous Data Type)

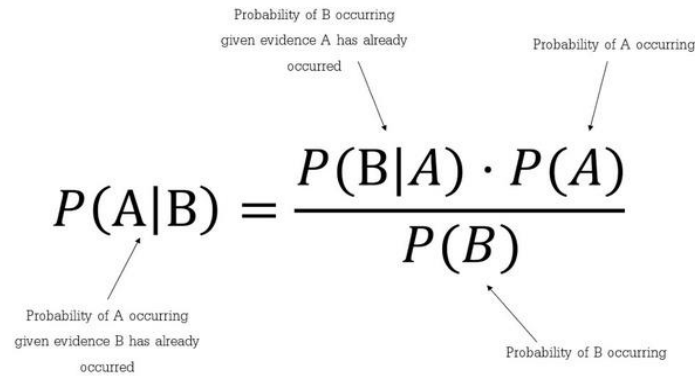
The goal of using a Decision Tree is to create a training model that can used to predict the class or value of the target variable by learning simple Decision rules inferred from the prior data (training data).

In Decision Tree, for predicting a class label for a record we start from the root of the tree. We Compare the values of the root attributes with the record's attributes, on the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Decision Tree are built using a heuristic called recursive partitioning. This approach is commonly known as divide and conquer because it splits the data into subsets, and so on until the process stops when the algorithm determines the data within the subsets are sufficiently homogeneous, or another stopping criterion have been met [8].

3.3.3 Naïve Bayes

Naïve Bayes Classifier is one of the simple and most effective classification algorithms which helps in building the fast machine learning model that can make quick prediction. It is a probabilistic classifier, which means it predicts in the basis of the probability of the object which uses the Bayes Theorem stated as:

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


Naïve Bayes is a naïve because it makes the assumption that features of the measurement are independent of each other. Naïve Bayes is easy to build and particularly useful for vary large dataset.

3.3.4 Support Vector Machine

Support Vector Machine is a supervised learning models associated with the learning algorithm that analyze data used for classification and regression analysis. The objective of Support vector machine is to find the hyperplane of N-Dimensional Space that distinctly classifies the data points. Support Vector are simply coordinate of the individual information. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/line).

3.3.5 k-Nearest Neighbors (KNN)

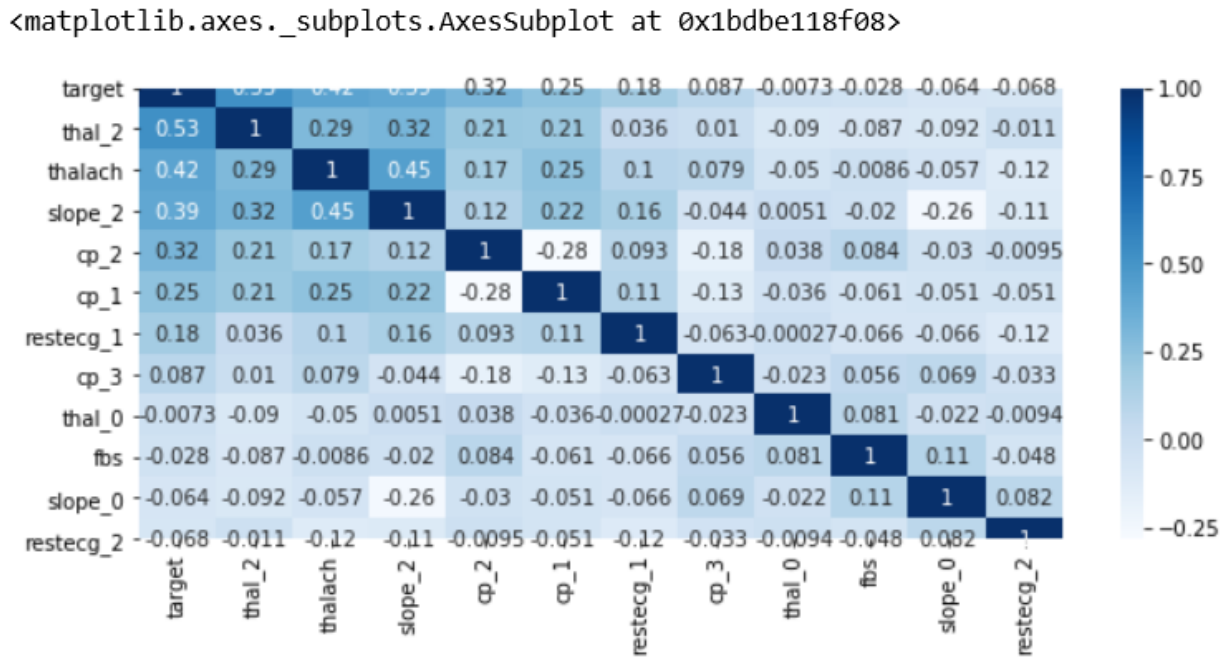
k-Nearest Neighbors algorithm is simple, easy to implement supervised-learning algorithm, that can used to solve both classification and regression problem. It uses the 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

3.3.6 Random Forest

Random Forest is a set of decision tree from randomly selected subset of training set. It aggregates the votes from different decision tree to decide the final class of the object. Random Forest Algorithm works in the two stages, one is the random forest creation and other is to make a prediction from the random forest classifier created in the first stage.

IV. Observation

4.1 Correlation Matrix



From the correlation matrix, we conclude that there are no features has a very high correlation with our target. Some of the features have a negative correlation with the outcome value and some have positive value.

4.2 Confusion Matrix

A Confusion Matrix is a performance measurement of the classification model where output can be of two or more classes. It is a table of the different combination of actual value and predicted value.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Confusion Matrix for Logistic Regression: [[34 14]

[4 39]

Representing TP: 34, TN: 39, FP: 14, FN: 4.

Confusion Matrix for Decision Tree Classifier: [[31 17]

[5 38]

Representing TP: 31, TN: 38, FP: 17, FN: 5.

Confusion Matrix for Naïve Bayes: [[14 34]

[0 43]

Representing TP: 14, TN: 43, FP: 34, FN: 0.

Confusion Matrix for Support Vector Machine: [[33 15]

[2 41]

Representing TP: 33, TN: 41, FP: 15, FN: 2.

Confusion Matrix for Random Forest Classifier: [[35 13]

[4 39]

Representing TP: 35, TN: 39, FP: 13, FN: 4.

Confusion Matrix for KNN: [[41 7]

[6 37]

Representing TP: 41, TN: 37, FP: 7, FN: 6.

4.3 Accuracy Comparison

Accuracy is the ratio of total no. of correct prediction to the total no. of input samples. It can be calculated using the following equation:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Table 2 shows the accuracy value for our machine learning Algorithm:

Table 2: Accuracy Values

Algorithm	Accuracy
Logistic Regression	80.22%
Decision Tree Classifier	75.82%
Naïve Bayes	62.63%
Support Vector Machine	81.31%
Random Forest Classifier	84.61%
KNN	85.71%

In table 2, the accuracy of predicting the model with the KNN algorithm is best in the comparison to all other algorithm which are used to predict the model.

4.4 Mean Squared Error

Mean Squared Error of an estimator measures the average of the squares of the errors i.e., the average squared difference between the estimated value and the actual value

Table 3 shows the MSE of the machine learning algorithm:

Table 3: Mean Squared Error

Algorithm	Mean Squared Error
Logistic Regression	0.197802
Decision Tree Classifier	0.241758
Naïve Bayes	0.373626
Support Vector Machine	0.186813
Random Forest Classifier	0.186813
KNN	0.142758

In table 3, the minimum MSE is of KNN algorithm (which tells that there is very less difference in prediction value and actual value) in comparison to all other algorithm.

4.5 Recall

It calculates how many of the actual positives our model capture through labeling it as Positives (True positives). Recall shall be model metric we used to select our best model when there is a high cost associated with False Negative. It is also known as Sensitivity. It can be calculated as:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Table 4 shows the Recall values of the machine learning algorithm applied in the Dataset.

Table 4: Recall Values

Algorithm	Not Caused (0)	Caused (1)	Average
Logistic Regression	0.71	0.91	0.80
Decision Tree Classifier	0.62	0.84	0.73
Naïve Bayes	0.29	1.00	0.64
Support Vector Machine	0.69	0.95	0.82
Random Forest Classifier	0.77	0.93	0.85
KNN	0.85	0.86	0.86

Graphical representation is given as :

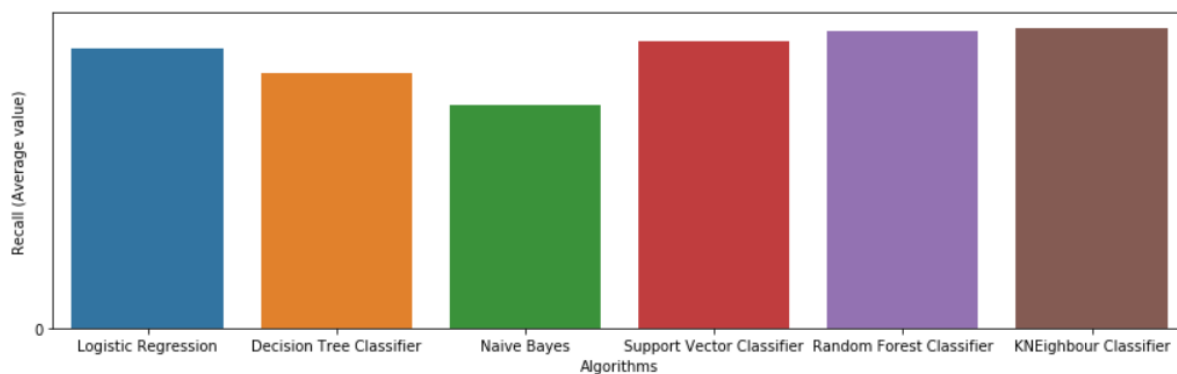


Figure 6 : Recall Value of the Model

In table 4 and in the figure 6, the KNN algorithm recall value is 0.86 i.e., it correctly identifies 86% of all heart disease patient, also the other algorithm which correctly identifies the model with 85% is Random Forest classifier, and at the third place which correctly identifies the heart disease patient with 82% is Support Vector Machine, while the worst recall value is 0.64 of Naïve Bayes i.e., the correctness of our model with Naïve Bayes is very low .

4.6 Precision

Precision is the rate of both True Positives and True Negative that has been identified as true positives. This shows how well the classifier handles the positives observations but does not say much about the negative ones. It calculated as:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Table 5 shows the Precision values of the machine learning algorithm in the Dataset

Table 5 : Precision Values

Algorithm	Not Caused (0)	Caused (1)	Average
Logistic Regression	0.89	0.74	0.82
Decision Tree Classifier	0.81	0.67	0.74
Naïve Bayes	1.00	0.56	0.79
Support Vector Machine	0.94	0.73	0.84
Random Forest Classifier	0.93	0.78	0.65
KNN	0.87	0.84	0.86

Graphical representation of the table 5 is given as:

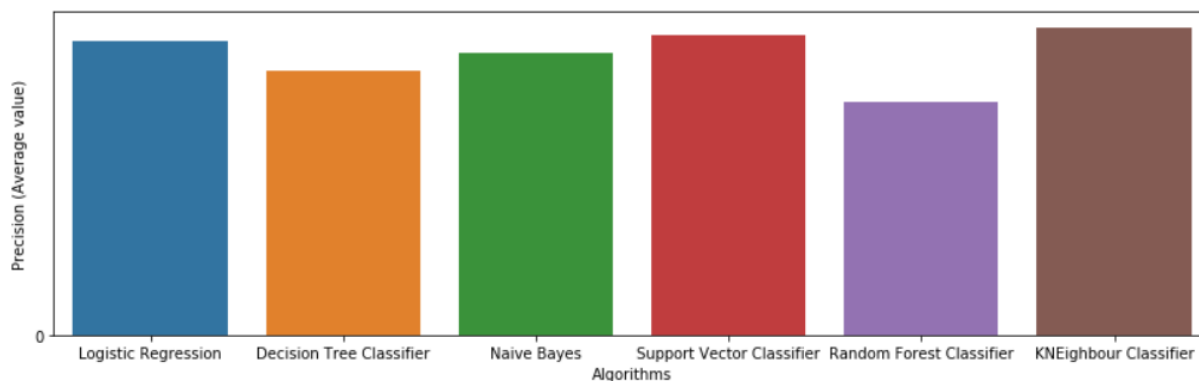


Figure 7: Precision Value of the Model

In table 5 and in the figure 7, best precision value is 0.86 of KNN algorithm, which defines that when we predict the heart disease, it is correct 86% of the time, and the second algorithm which predict with the best precision with 0.84 is the Support Vector machine predict the heart disease correctly 84% of the time while worst precision is 0.65 of Random Forest Classifier.

4.7 F1-Score

F1-Score is a measure of test's accuracy, As, it considers both the precision and recall of the test to compute the score. Therefore, this score takes both false positives and false negative into account. It calculated as follow:

$$F1 = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Table 6 shows the F1 Score after the implementation on the dataset with the various machine learning algorithm is give as:

Table 6: F1-Score

Algorithm	Not Caused (0)	Caused (1)	Average
Logistic Regression	0.79	0.81	0.80
Decision Tree Classifier	0.71	0.74	0.73
Naïve Bayes	0.45	0.72	0.63
Support Vector Machine	0.80	0.83	0.81
Random Forest Classifier	0.84	0.85	0.85
KNN	0.86	0.85	0.86

Graphical Representation of the table 6 is given as:

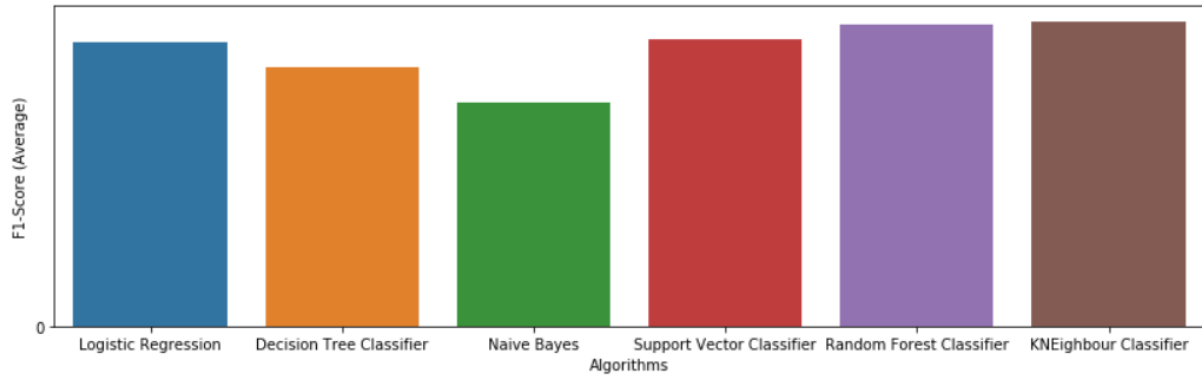


Figure 8: Precision Value of the Model

We know that, a F1-Score is perfect when it is close to 1, while the model is total failure when it is 0. So in table 6 and in figure 8, KNN algorithm has the best F1-Score with 0.86, and it decreases when we change algorithm i.e., Random Forest Classifier and then Support Vector Machine, while the lowest F1-Score is 0.63 of Naïve Bayes.

V. Result

I had implemented the model using 6 machine learning classification algorithms. The Comparative Analysis of the classification algorithm shows that the Classification Accuracy, Recall, Precision, F1-Score. Table 7 shows that the KNN algorithm shows the most significant result comparison to others. Table 7 is represented as:

Table 7: Comparative Analysis

Algorithm	Accuracy	Precision	Recall	F1 - Score
Logistic Regression	80.22%	0.82	0.80	0.80
Decision Tree Classifier	75.82%	0.74	0.73	0.73
Naïve Bayes	62.63%	0.79	0.64	0.63
Support Vector Machine	81.31%	0.84	0.82	0.81
Random Forest Classifier	84.61%	0.65	0.85	0.85
KNN	85.7%	0.86	0.86	0.86

Graphical Representation of the accuracy is given in Figure 2.

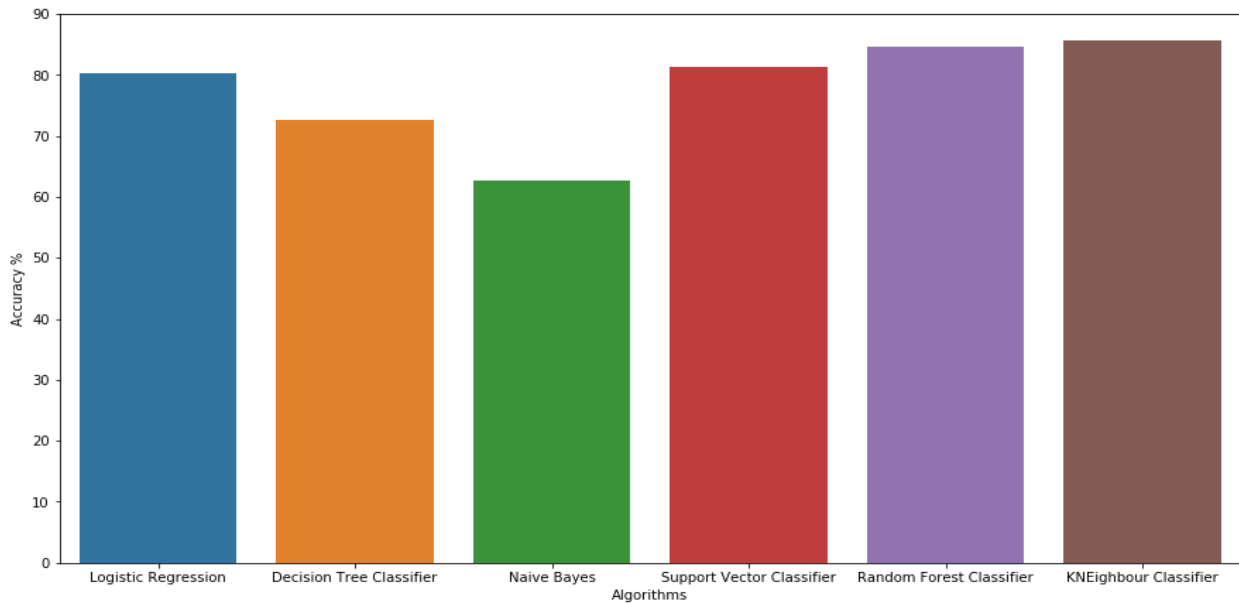


Figure 9 : Graphical Representation of Accuracy Comparison

Figure 9, shows that KNeighbour Classifier gives the most accurate prediction with 85.7%, while the second classification algorithm with the best accuracy (84.61%) is the Random Forest Classifier, and the third most accurate prediction with 81.61% is Support Vector Machine while, the worst accuracy is (62.63%) of Naïve Bayes.

Also, the mean squared error of the model is represented as follow:

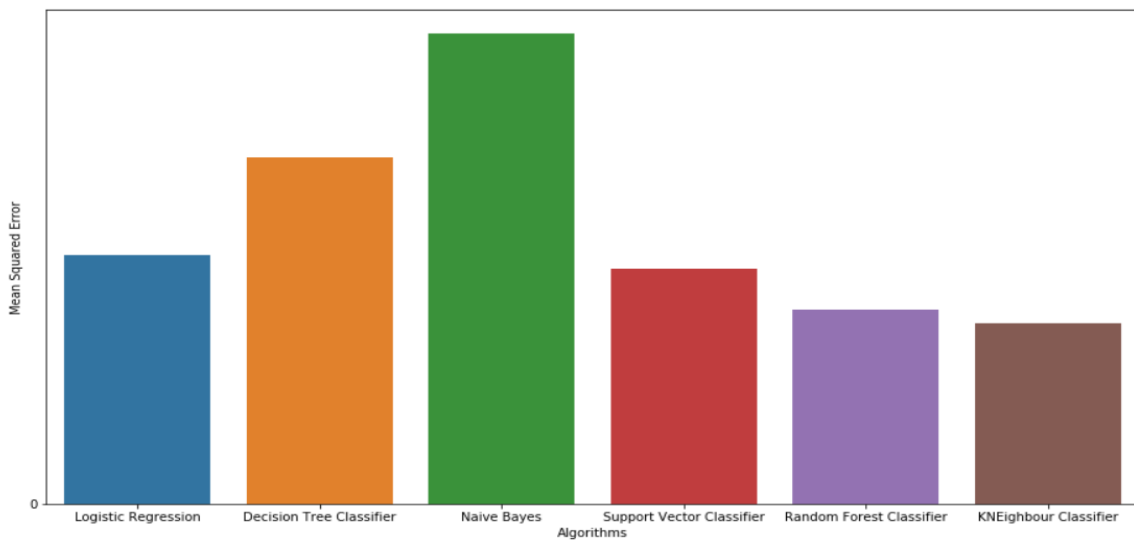


Figure 10 : Graphical Representation of Mean Squared Error

Here, the classification algorithm with minimum MSE is a KNN algorithm and it followed with random forest classifier and Support Vector machine while very high MSE is of Naïve Bayes in comparison to all others. The limitation is the no. of samples used training and test

VI. Conclusion and Future Work

Heart Disease prediction is challenging and very important in medical field. Here, various Machine Learning Algorithm is used to predict on the raw data and trying to predict a novel prediction towards heart disease. However, here dataset is limited with respect to the no. of samples of training and testing. So, this is most critical prediction that's why prediction should be carried out with the larger dataset to make the accuracy better.

The future work of the research can be performed with the more feature selection like Age, lipid levels, obesity, lack of activity and stress can all contribute to blocked arteries, preventing blood flow to the heart, so that the performance of the prediction is better and the prediction should perform with diverse mixture of machine learning technique. Also, if the bagging, boosting, or ensemble method can be applied to the data set and the result can be compared and improved.

VII. Acknowledgement

I have completed this work under the mentorship of Dr. Pankaj Agrawal (Professor & head) & Ms. Sapna Yadav (Assistant Professor), Department of Computer Science and Engineering at IMS Engineering College, Ghaziabad. I am doing a online summer internship on Machine Learning where I have learnt various machine learning algorithm from both of my mentors as a course instructors. This work has been assigned as a project assignment to us.

I would like to express my special thanks to both of the mentors for inspiring us to complete work and write paper. Without their active guidance, help cooperation & encouragement, I would not my headway in writing this paper. I am extremely

thankful for their valuable guidance and support on completion of this paper.

I extend by gratitude to "IMS Engineering College" for giving me this opportunity. I also acknowledge with a deep sense of reverence, my gratitude towards my parents and members of my family, who always supported me morally as well as economically.

Any omission in this brief acknowledgement does not mean lack of gratitude.

VIII. REFERENCES

- [1]. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1
- [2]. <https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks>
- [3]. Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna Computer Science Engineering, K L E F, Guntur, India , Prediction of Heart Disease Using Machine Learning Algorithms (2018), www.sciencepubco.com/index.php/IJET
- [4]. Senthilkumar Mohan, Chandrasegar Thirimalai, and Gautam Srivastava, School of Information Technology and Engineering, VIT University, Vellore, India, Department of Mathematics and Computer Science, Brandon University, Brandon, Canada ,Research Center for Interneural Computing, China Medical University, Taichung, Taiwan , PREDICTION OF HEART DISEASE USING HYBRID MACHINE LEARNING TECHNIQUE', in June 19,2019.
- [5]. Sellappan Palaniyappan, Rafiah Awang Intelligent heart disease prediction using data mining techniques in August 2008.
- [6]. Benjamin EJ et.al Heart Disease and Stroke Statistics 2018 At-a-Glance (2018)
- [7]. C.Beulah Christalin Latha, S.Carolin Jeeva, Karunya Institute of Technology and Sciences,

- India, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, (2019), www.elsevier.com/locate/imu
- [8]. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review in 2017.
- [9]. V.Krishnaiah, G.Narsimha, N.Subhash Chandra Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review in 2016.

Cite this article as :

Ankit Singh, "Prediction of Heart Disease using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 4, pp.150-166, July-August-2020. Available at doi : <https://doi.org/10.32628/CSEIT206415>
Journal URL : <http://ijsrcseit.com/CSEIT206415>