International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Heart Disease Prediction with Machine Learning Approaches

Megha Kamboj

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

Abstract: ***Heart is the most essential or crucial portion of our body. Heart is used to maintain and conjugate blood in our body. There are a lot of cases in the world related to heart diseases. People are leading to death due to heart disease. Various symptoms like chest pain, fasting of heartbeat and so on are mentioned. The health care industries 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 the heart disease. This paper uses different machine learning algorithms for comparing the accuracy among them.***

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

1. Introduction

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

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 an arm.

Types of heart disease are:

- Coronary artery disease (CAD).
- Angina pectoris.
- Congestive heart failure.
- Cardiomyopathy.
- Congenital heart disease.

On the basis of 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 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 heart disease dataset.

2. Literature Review

Amandeep Kaur [1] compared various algorithms such as artificial neural network, K-nearest neighbor, Naïve Bayes, Support vector machine on heart disease prediction.

J Thomas, R Theresa Princy [2] made use of K nearest neighbor algorithm, neural network, naïve Bayes and decision tree for heart disease prediction. They used data mining techniques to detect the heart disease risk rate.

Monika Gandhi et. Al, [3] used Naïve Bayes, Decision Tree and neural network algorithms and analyzed the medical dataset. There are a huge number of features involved. So, there is need to reduce the number of features. This can be done by feature selection. On doing this, they say that time is reduced.

Ramandeep Kaur, Er. Prabhsharn Kaur [4] have showed that the heart disease data contains unnecessary, duplicate information. This has to be pre-processed.

Sonam Nikhar et. Al, [5] has built up the paper titled as Prediction of Heart Disease Using Machine Learning Algorithms using decision tree classifier and naïve bayes.

Mr. Santhana Krishnan. J and Dr. Geetha. S, [6] has written paper that predicts heart disease for male patient using classification techniques.

3. Methodology

In this paper, we have used our dataset for applying different machine learning algorithms for identifying 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.

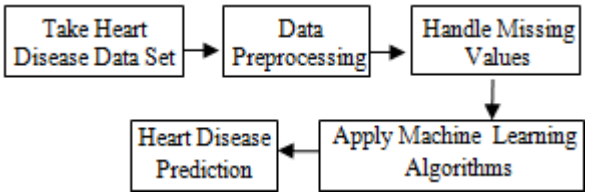
Data Collection

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 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 a 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 the 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 thal 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



4. Results and Discussion

Correlation Matrix

Let's see the correlation matrix of features. From this graph, we can observe that some features are highly correlated and some are not.

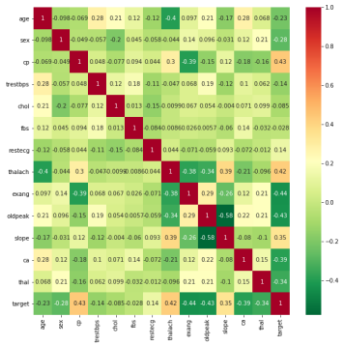


Figure 1: This figure shows the correlation matrix

Histogram

The histogram is best and easy 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.

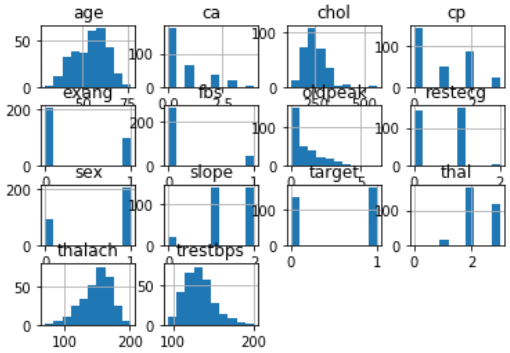


Figure 2: This figure shows the histogram.

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 to 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.

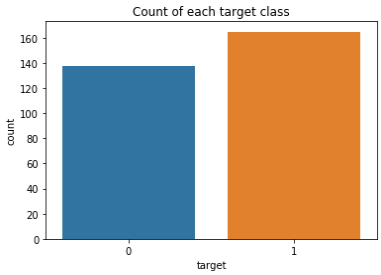


Figure 3: Target versus Count Feature.

The above graph shows the distribution of target versus count class that is used to predict the total number of heart disease whether someone has heart disease or not (0 = no heart disease, 1 = having heart disease).

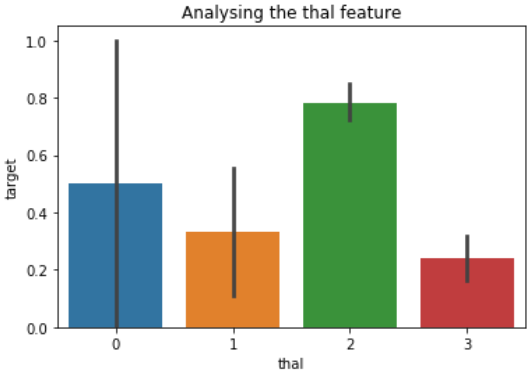


Figure 4: Target versus Thal Feature.

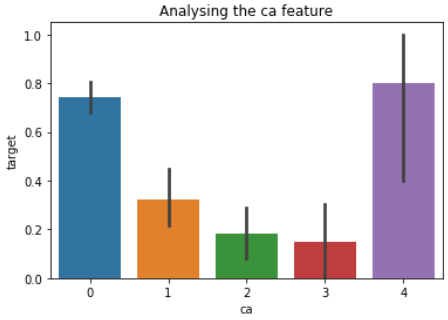


Figure 5: Target versus Ca Feature.

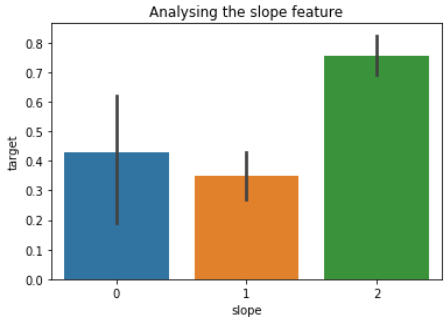


Figure 6: Target versus Slope Feature.

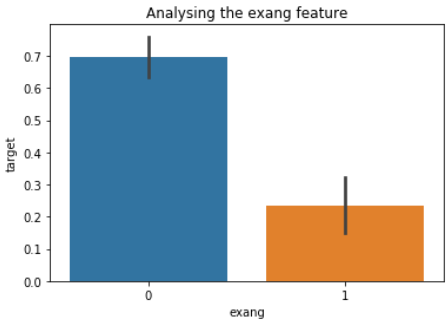


Figure 7: Target versus Exang Feature.

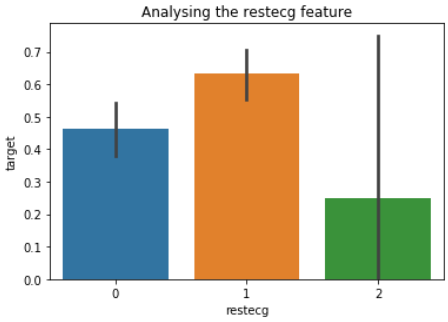


Figure 8: Target versus Restecg Feature.

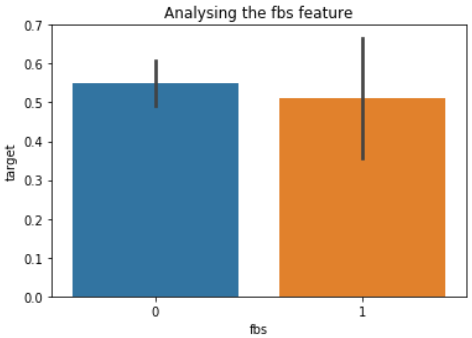


Figure 9: Target versus Fbs Feature.

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 the 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 decision tree and other selected classifiers. The computation cost can be reduced greatly. It is easy to implement. In this paper, we achieved the accuracy of 80% by using this classifier.

K Nearest Neighbors Classifier:
K Nearest Neighbors is a non parametric method used for classification. It is 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 problem.

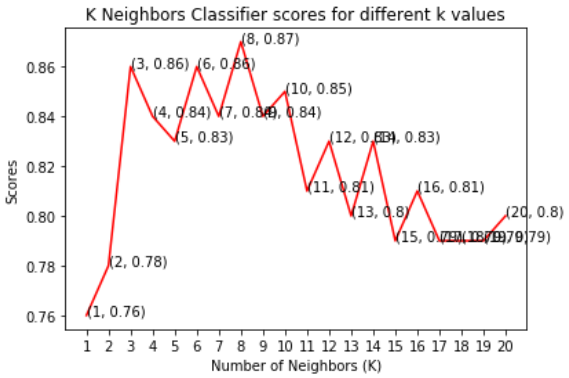


Figure 10: This figure shows the K Neighbors Classifier scores

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 which 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.

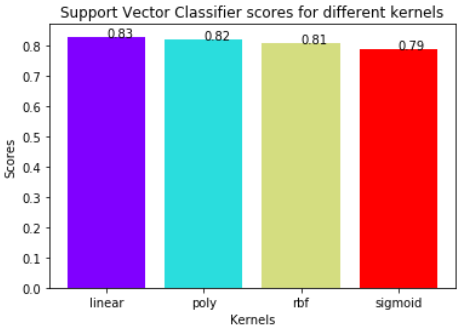


Figure 11: This figure shows the Support Vector Classifier scores.

This graph shows that the linear kernel is having the highest accuracy of 83% by using this dataset.

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.

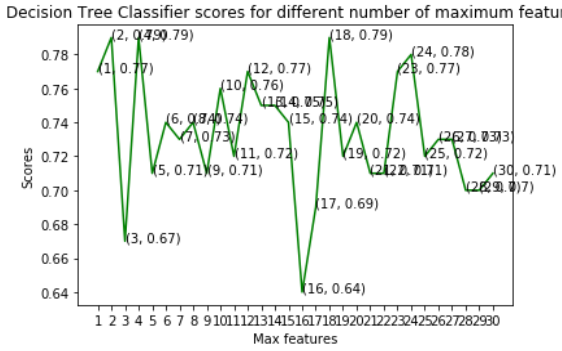


Figure 12: This figure shows the Decision Tree Classifier scores

This graph shows the line graph from which we observed that the maximum accuracy is 79% and is obtained by number of maximum features (2, 4, 18).

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 prediction from each tree and selects the best solution by means of voting. The random forest composed of multiple decision trees. It creates a forest of trees.

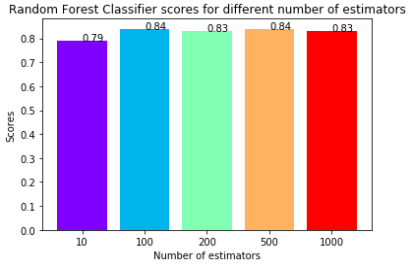


Figure 14: This figure shows the Random Forest Classifier scores.

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

Table 1: Accuracy Values

Algorithms	Accuracy
Logistic Regression	84%
Naïve Bayes Classifier	80%
K Nearest Neighbors Classifier	87%
Decision Tree Classifier	79%
Support Vector Classifier	83%
Random Forest Classifier	84%

Table 1 shows that 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.

5. Conclusion and Future Work

This paper involves prediction of the heart disease dataset with proper data processing and implementation of machine learning algorithms. In this paper, we uses six machine learning algorithms for prediction.

Among all the machine learning algorithms used in this paper, the highest accuracy is achieved by K Nearest Neighbors Classifier with 87%. This paper shows that the machine learning algorithms can be used to predict the 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 the problems in different areas too.

6. Acknowledgement

I have completed this work under the mentorship of Dr. Pankaj Agarwal (Professor & Head) & Ms. Sapna Yadav (Assistant Professor), Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad. I am doing an online summer internship on Machine Learning where I have learnt the various Machine Learning Algorithms from both of my mentors as Course Instructors. This work is been assigned as project assignments to us.

I would like to express my special thanks to both of my mentors for inspiring us to complete the work & write this

paper. Without their active guidance, help, cooperation & encouragement, I would not have been able to write this paper. I am very thankful for their guidance and help on completion of this paper.

I would like to express my gratitude to "IMS Engineering College" for giving me this great opportunity. I would also like to express my special thanks to my parents and my family members, who has always supported me morally as well as economically.

Thanking You.

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.
[6] Himanshu Sharma, M A Rizvi, (August 2017): Prediction of Heart Disease Using Machine Learning Algorithms: A Survey.
[7] I Ketut Agung Enriko, Muhammad Suryanegara, Dadang Gunawan al, (June 2018): "Heart Disease Diagnosis System with k – Nearest Neighbors Method Using Real Clinical Medical Records", 4th International Conference.
[8] Lakshmanarao, Y. Swathi, P.Sri Sai Sundareswar, (November 2019): Machine Learning Techniques For Heart Disease Prediction, International Journal Of Scientific & Technology Research Volume 8, Issue 11.
[9] Monika Gandhi, Shailendra Narayanan Singh, (2015): Predictions in heart diseases using techniques of data mining.
[10] M. S. Amin, Y. K. Chiam, K. D. Varathan, (Mar.2019): Identication of significant features and data mining techniques in predicting heart disease, Telematics Inform., vol. 36, pp. 8293.
[11] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, (2019): Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7.
[12] V. V. Ramalingam, Ayantan Dandapath, M Karthik Raja, (2018): heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology (IJET), 7 (2.8) 684-687.

Author Profile

Megha Kamboj is B.Tech 3rd year student in the Department of Computer Science & Engineering at IMS Engineering College, Ghaziabad, Uttar Pradesh, India. She is interested in Python and Machine Learning. She has done data analyst internship from Edulyt India. She has successfully worked on four different projects provided by Edulyt India which are "Website Blocker with Python, Online Dictionary with Python, Credit Card Approval Prediction With Machine Learning & Loan Approval Prediction With Machine Learning".