# Heart Disease Prediction Using Machine Learning Techniques

**Vajja Nagavallika**

Department of Computer Science and Engineering, Andhra University College of Engineering (A)
Andhra University, Visakhapatnam - 530003, India
*319206415026[at]andhrauniversity.edu.in*

**Abstract:** ***Heart disease is one of the most significant causes of mortality in today's world. Heart disease proves to be the leading cause of death for both men and women. This affects the human life very badly. The diagnosis of heart disease in most cases depends on a complex combination and huge volume of clinical and pathological data. Machine learning has been shown to be effective assisting in making decisions and predictions from the large quantity of data produced by the health care industry. In this paper, various traditional machine learning algorithms that aims in improving the accuracy of heart disease prediction has been applied. In heart diseases, accurate diagnosis is primary. But, the traditional approaches are inadequate for accurate prediction and diagnosis. In order to apply deep learning technique very large datasets are required which are not available in medical and clinical research. To address this issue, surrogate data is generated from Framingham dataset. The generated synthetic dataset is utilized with traditional machine learning algorithms. The predicted results show that there is an improvement in classification accuracy. The generated synthetic dataset plays a vital role to improve the classification prediction particularly when dealing with sensitive data.***

**Keywords:** Machine Learning, Logistic regression, Random Forest Algorithm, Naïve Bayes Classification, Confusion Matrix, Precision, Recall, Accuracy

## 1. Introduction

The heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all the tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like brain and kidneys suffers, if the heart stops working altogether, death occurs within minutes. The heart disease has been considered as one of the complex and life deadliest human diseases in the world. Life itself is completely dependent on the efficient operation of heart. Symptoms of heart disease include shortness of breath, weakness of physical body, swollen feet and fatigue and it is discussed in [1]. The heart disease diagnosis and treatment are very complex, especially in the developing countries, due to the rare availability of diagnostic apparatus and other resources which affect proper prediction and treatment of heart patients. This makes heart disease a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors.

The invasive–based techniques to the diagnose of heart disease are based on the analysis of the patient's medical history, physical examination report and analysis of concerned symptoms by medical experts. Often there is a delay in the diagnosis due to human errors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Machine learning helps computers to learn and act accordingly. It helps the computer to learn the complex model and predict the data and also has the ability to calculate complex mathematics on big data. The machine learning based heart disease predicting systems will be precise and will reduce therisk. The value of machine learning technology is recognized well in health care industry which has large pool of data. It helps medical experts to predict the disease and lead to improvise the treatment. Machine learning predictive models such as decision tree, k - nearest neighbor, logistic regression, random forest, support vector machine are utilized to predict whether a person is having heart disease or not.

In this project we have developed a model which classifies if patient will have heart disease in ten years or not based on various features (i. e. potential risk factors that can cause heart disease) using logistic regression. In this proposed system, use logistic regression (classification) algorithm. By using sklearn library to calculate score. Finally analysing the results by the help of Comparing Models and Confusion Matrix.

## 2. Literature Survey

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

Authors Yan, Zheng et al.2003; Andreeva 2006; Das, Turkoglu et al.2009; [1] proposed Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies

Authors Sitar - Taut, Zdrenghea et al.2009; Raj Kumar and Reena2010; Srinivas Rani et al.2010 [2] proposed on multiple databases of patients from around the world. One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies.

Authors Sitair - Taut et al. [3]proposedIn particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Used the Weka tool to investigate applying Naive Bayes and J48 Decision Trees for the detection of coronary heart disease. Tu et al. used the bagging algorithm in the Weka tool and compared it with J4.8 Decision Tree in the diagnosis of heart disease.

Authors S. Vijiyaraniet. al. [4] proposed In year 2013, performed a work, An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyses the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset

[5]In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10 - fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent.

[6]Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbors Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random Forest Classifier: 84%.

## 3. Methodology

### 3.1 Data Collection

The dataset is publicly available on the Kaggle Website at which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python

Since the dataset consists of 4240 observations with 388 missing data and 644 observations to be risked for heart disease, two different experiments were performed for data preparation. First, we checked by dropping the missing data, leaving with only 3751 data and only 572 observations risked for heart disease. This leads to reduced number of the observations providing irrelevant training to our model. So, we progressed with imputation of data with the mean value of the observations and scaling them using Simple Imputer and Standard Scaler modules of Sklearn.

### 3.2 Proposed System

The aim of proposed model is to classify people having heart disease or not and finding the accuracy level of machine learning algorithm. The popular Machine Learning Classifiers like Logistic Regression and Random Forest were used for classification task. In this proposed system, use logistic regression (classification) algorithm. By using sklearn library to calculate score. Finally analysing the results by the help of Comparing Models and Confusion Matrix. From the data we are having, it should be classified into different structured data based on the features of the patient heart. From the availability of the data, we have to create a model which predicts the patient disease using logistic regression algorithm. First, we have to import the datasets. Read the datasets, the data should contain different variables like age, gender, cp (chest pain), slope, and target. The data should be explored so that the information is verified. Create a temporary variable and also build a model for logistic regression. Here, we use sigmoid function which helps in the graphical representation of the classified data. By using logistic regression, the accuracy rate increases.

## 4. Results

When performing various methods of feature selection, testing it was found that backward elimination gave us the best results among others. The various methods tried were Backward Elimination with and without K - Fold, Recursive Feature Elimination with Cross Validation. The accuracy that was seen in them ranged around 85% with 85.5% being maximum.

Though both methods gave similar accuracy but it was seen that in Backward Elimination we found that the number of misclassifications of True Negative was more and it was observed that the accuracy had more variance compared to RFEV.

### 4.1 Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e. g. one class is commonly mislabelled as the other.

Key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the number of errors made.

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

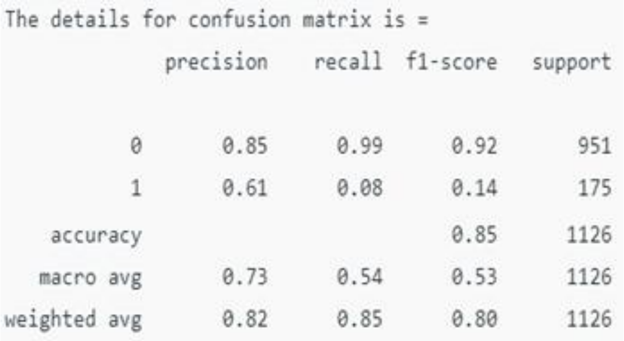
It allows easy identification of confusion between classes e. g. one class is commonly mislabelled as the other. The key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the number of errors made.

**Table 1:** Confusion Matrix Obtained after training the data (feature selection by backward elimination)

| | |
|---|---|
| TP=3569 | FP=27 |
| FN=599 | TN=45 |

**Table 2:** Confusion Matrix Obtained after training the data (feature selection by RFECV method)

| | |
|---|---|
| TP=3582 | FP=14 |
| FN=600 | TN=44 |



```
The details for confusion matrix is =
              precision    recall  f1-score   support

           0       0.85      0.99      0.92       951
           1       0.61      0.08      0.14       175

    accuracy                           0.85      1126
   macro avg       0.73      0.54      0.53      1126
weighted avg       0.82      0.85      0.80      1126
```

**Figure 1:** Details obtained for confusion matrix

### 4.2 Accuracy

Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right. Formally, accuracy has the following definition:

The accuracy is calculated as:

$$\frac{TP + TN}{(TP + TN + FP + FN)}$$

Where,

- True Positive (TP) =Observation is positive, and is predicted to be positive.
- False Negative (FN) = Observation is positive, but is predicted negative.
- True Negative (TN) = Observation is negative, and is predicted to be negative.
- False Positive (FP) =Observation is negative, but is predicted positive

The obtained accuracy during training the data after feature selection using backward elimination was 86 % and during testing was 83%.

The obtained accuracy during training the data after feature selection using REFCV method was 86 % and during testing was 85 %.

### 4.3 Recall

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN). Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

The obtained recall during training the data after feature selection using backward elimination was and during testing was 0.99.

The obtained recall during training the data after feature selection using REFCV method was 1.00 and during testing was 0.99.

### 4.4 Precision

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP). Precision is calculated as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

The obtained precision during training the data after feature selection using backward elimination was 0.86 and during testing was 0.84.

The obtained precision during training the data after feature selection using REFCV method and during testing was 0.86.

The precision of Backward Elimination and RFEV are 84% and 86% respectively. And the recalls are 0.99 and 1 respectively. The precision and recall also shows that the number of misclassifications is less in RFECV than in Backward Elimination

The accuracy score obtained for Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques

**Table 3:** Comparison between the feature selection models after training and testing through Logistic Regression model

| Evaluation Metrics | Backward Elimination | RFECV |
|---|---|---|
| Accuracy | 83% | 85% |
| Recall | 0.99 | 0.99 |
| Precision | 0.84 | 0.86 |

## 5. Conclusion and Future Scope

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset.

The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% for prediction of heart disease. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. This project resolved the feature selection i.e. backward elimination and RFECV behind the models and successfully predict the heart disease, with 85% accuracy.

In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

At some point in future, the machine learning model will make use of a larger training dataset, possibly more than a million different data points maintained in electronic health record system. Although it would be a huge leap in terms of computational power and software sophistication but a system that will work on artificial intelligence might allow the medical practitioner to decide the best suited treatment for the concerned patient as soon as possible [6]. A software API can be developed to enable health websites and apps to provide access to the patients free of cost. The probability prediction would be performed with zero or virtually no delay in processing

## References

[1] H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction, " 2018.

[2] M. I. K.,. A. I.,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[3] K. Bhanot, "towarddatascience. com, " 13 Feb 2019. [Online]. Available: https: //towards data science. com/predicting - presence - of - heart - diseases - using – machine learning - 36f00f3edb2c. [Accessed 2 March 2020].

[4] [Online]. Available: https: //www.kaggle. com/ronitf/heart - disease - uci#heart. csv. . [Accessed 05 December 2019].

[5] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".

[6] With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieve optimum performance than KStar, Multilayer perceptron and J48 techniques using k - fold cross validation. The accuracy performance achieved by those algorithms are still not satisfactory. So that if the performance of accuracy is improved more to give batter decision to diagnosis disease.

[7] In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10 - fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent.

[8] Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbours Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random Forest Classifier: 84%.

[9] Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies (Yan, Zheng et al.2003; Andreeva 2006; Das, Turkoglu et al.2009;

[10] Sitar - Taut, Zdrenghea et al.2009; Raj Kumar and Reena 2010; Srinivas Rani et al.2010) on multiple databases of patients from around the world. One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies.

[11] In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Sitair - Taut et al. used the Weka tool to investigate applying Naive Bayes and J48 Decision Trees for the detection of coronary heart disease. Tu et al. used the bagging algorithm in the Weka tool and compared it with J4.8 Decision Tree in the diagnosis of heart disease.

[12] In year 2013, S. Vijiyaraniet. al. performed a work, An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyses the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.