**Heart Disease Predictor**

# Submitted to



Tanya Gupta, Ankna Litoriya, Shivank Pandey

Samrat Ashok Technological Institute, Vidisha

Date: 27/02/2024

Abstract

Day by day the cases of heart diseases are increasing at a rapid rate and it’s very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and K-nearest neighbor to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using K-nearest neighbor and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naïve bayes etc. so a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the ‘.pynb’format.

# ABSTRACT iii

# TABLE OF CONTENTS iv

# LIST OF FIGURES v

# LIST OF TABLES vi

# 1 INTRODUCTION 1

* 1. Brief Research Background 2
  2. Problem Statement 3
  3. Objective 3
  4. Software Requirements Specification 4

# LITERATURE SURVEY 4

# DATA COLLECTION AND ANALYSIS 8

* 1. Python Libraries 9
  2. Supervised Learning 10
  3. Data Set Analysis and Result 11

# METHODOLOGY 14

* 1. Machine Learning Algorithm for Cardiovascular Disease Prediction 14
     1. Decision Trees 14
     2. Naïve Bayes 15

# CODES 23

# DISCUSSIONS OF RESULTS 25

# CONCLUSION AND FUTURE ENHANCEMENT 26

# 8 REFERENCES 28

# INTRODUCTION

*Data is the new science. Big Data holds the answers.” –Pat Gelsinger*

Rising healthcare costs have been a major issue for developed nations. According to CDC, an estimated 859,000 people in the US die from cardiovascular disease or 1 in every 3 deaths. In light of the rapid advancement of biotechnology, and an era of big data generated for healthcare by mainly EHR(electronic health records) in various structures, it is increasingly more important to intelligently use this information to make sense of hidden patterns, detect abnormalities, and predict heart diseases.

Artificial intelligence has certainly made computers smarter. Machine learning which is a subset of artificial intelligence plays an important role in mining large datasets and extracting valuable knowledge from them. Training a machine appropriately with proper train data set, the machine’s algorithm can learn patterns and therefore detect any abnormalities in the initial stage of a disease which can help patients save overall cost and time. This project will examine the opportunities of machine learning and data mining in the healthcare industry especially in heart diseases, how early diagnosis can minimize healthcare costs; and how data generated by EHR can provide insights for medical professionals in terms of detecting abnormalities for potential chronic diseases. We begin by providing a brief research background, followed by the problem statement, research questions, objectives, and the organization of this culminating experience project

Brief Research Background

Heart Disease

Cardiovascular diseases are the dominant cause of cost and disease burden in the world. (Roth et al., 2020). Cardiovascular diseases refer to any disorder in the heart and blood vessels. Major blood vessels that supply to the heart muscles are affected by a heart condition. These blood vessels build up on cholesterol deposits called plaque reducing blood flow to major parts of the body and the heart (Heart disease, 2022). Over time if left untreated, this can lead to stroke, heart attack, or heart failure. Heart diseases are considered silent killers and at times not diagnosed until life-threatening symptoms start to emerge.

Diagnosis of these diseases can include various blood tests, MRIs and CT scans, ECGs, or Holter monitoring. All this medical big data is collected and stored in various databases, which do not provide value on their own, but if integrated and analyzed using Artificial Intelligence, machine learning and data mining techniques it is possible to generate diagnostic information that can lives while minimizing costs.

Machine Learning

In the modern era, humans are experiencing exponential

growth of data like never before. With the availability of online data and inexpensive computational computer power, machine learning algorithms can learn and develop models without human intervention (Jordan & Mitchell, 2015). Machine learning, a subset of artificial intelligence, can collect meaningful knowledge from its training data and automatically improve through exposure without having to be programmed. The machine’s algorithm can be classified into four main types, which are Supervised, Unsupervised, Semi-supervised, and Reinforcement Learning (Sarker, 2021). Supervised Learning can be split into two categories: Classification and Regression. Unsupervised learning can be classified into Clustering and Association (Delua, 2021). Both learning approaches are mainly distinguished by using labeled or unlabeled datasets to anticipate the outcome. Each of these has a distinctive set of guidelines when applied to medical data and effectively using it will help extract vital knowledge (Gupta et al., 2021).

Problem Statement

Modern information technology tools and techniques such as AI, machine learning and data mining could help support healthcare professionals by providing them with the information they need to make decisions that will minimize deaths caused by heart disease at minimal cost. For example, machine learning algorithms can mine large databases to identify frequent patterns that eventually lead to heart disease and death

Objectives

The main objective of this project is to explore how Machine Learning algorithms can be used in the diagnosis of heart disease by building an optimized model that can be used to predict heart diseases.

# LITERATURE REVIEW

Bardhwaj et al., (2017), Shailaja et al., (2018), Sun et al., (2019), and Lee & Yoon, (2017) studied a broad overview of machine learning techniques used in healthcare for various diseases. They provided insights into the potential value of medical big data that can be used for clinical decision support, diagnostics, treatment decisions, fraud detection, and prevention. They briefly summarized the nine-step data mining process along with focusing on why efficient decision support was required by the healthcare system. The results from their experiment showed that machine learning models can be used for the early diagnosis of diseases. Their research is applicable to this project to an extent; however, their research is less focused on the diagnosis of heart diseases. Therefore, we move forward to review the literature that aligns with our project objective which is how machine learning algorithms can be used in the diagnosis of heart disease.

A comprehensive review by Tripoliti et al., (2017) focused on machine learning methodologies evaluating heart failure. They researched severity estimation of heart failure and the prediction of re-hospitalization, mortality, and destabilizations. They performed an extensive study on related works of heart failure.

A study by J. & S., (2019) used two supervised classifiers called Naïve Bayes Classifier and Decision Tree Classifiers to predict heart diseases on a dataset.

Their Decision Tree model predicted the heart disease patients with an accuracy of 91 percent and the Naïve Bayes Classifier had an accuracy of 87 percent.

A study by Kamal kant et al.(2014) proposed a model using the Naïve Bayes algorithm to predict heart diseases. The naïve Bayes algorithm is used to assign no dependency between the features. Their study concluded that the Naïve Bayes algorithm is the most effective for heart disease prediction after that Neural Networks and Decision Trees.

Nidhi Bhatla et al., (2012) used different data mining techniques to predict heart diseases. Their study revealed that the Neural Networks algorithm has performed with higher accuracy than Decision Trees. Their research project included two additional features such as obesity and smoking other than the common attributes.

A review by Rishi Dubey et al., (2015) studied different machine learning algorithms for the prediction of heart disease. Their study concluded that Neural Network is an efficient technique for heart disease prediction. Further adding that this method can also be used to select appropriate treatment.

Ashish Chhabbi et al., (2016) used a dataset collected from UCI repository to perform different data mining techniques to predict heart disease. They applied K-means algorithm and Naïve Bayes and their results revealed that tuning the number of clusters of the k-means algorithm gave better results than the default K-means.

Boshra Baharami et al., (2015) evaluated various classification methods such as, Decision Tree, K-Nearest Neighbors(k-NN), SMO (used to train Support Vector Machines). On their dataset, they used feature selection techniques to only select the important attributes and achieved the highest accuracy of 83.732% with Decision Trees.

Mrudula Gudadhe et al., (2010) studied heart disease classification using a decision support system. The methods they used were Support Vector Machine (SVM) and Artificial Neural Network (ANN). They incorporated a multilayer perceptron neural network (MLPNN) with three layers in their decision support system and revealing that MLPNN can be used for successfully diagnosing heart disease.

Asha Rajkumar et al., (2010) used the classification method based on supervised machine learning to diagnose heart disease. Their dataset was divided into two parts, 20% for testing and 80% for training, and ran the model used Naïve Bayes, Decision list, and K-NN algorithms. The study concluded that Naïve Bayes recorded a lower error ratio and was the most efficient.

Sairabi H. Mujawar et al., (2015) used altered K-means and Naïve Bayes algorithms to predict heart disease. Their Naïve Bayes model resulted in 93% accuracy in predicting heart disease and 89% accuracy when the patient does not have heart disease.

Mustafa et al., (2018) proposed an ensemble approach for better prediction by combining five classifiers. Their work included SVM, ANN, Naïve Bayes,

Regression analysis, and Random Forest. Their goal was to predict and diagnose cardiovascular disease.

Samuel et al., (2017) predicted the risk of heart failure using the Artificial Neural Network (ANN). Their work included fuzzy analytic hierarchy (AHP) to calculate the global weights of features depending on individual contributions. Afterward, the feature contributions were applied to train the ANN classifier to predict the patient’s risk of heart failure.

The literature review reveals emerging and advanced machine learning and data mining algorithms involved in predicting heart diseases. It is evident from the above literature review that data mining algorithms have effectively predicted heart diseases. The trustworthiness of the model for predicting heart diseases with different risk factors is a high concern, however, SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and RandomForest have achieved reliable results in the diagnosis of heart disease (Jan et al., 2018).

# DATA COLLECTION AND ANALYSIS

The Cleveland heart disease dataset used to build the machine learning model in this project was collected from the UCI machine learning repository (Latha & Jeeva, 2019). The dataset has 303 instances and 14 attributes; the dataset's description is given in the table below.

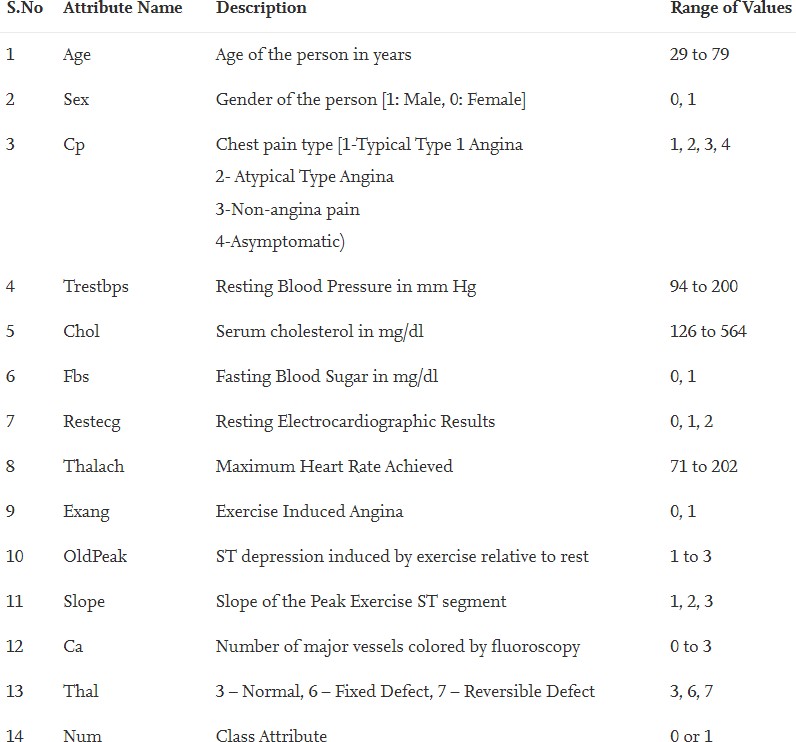


Figure 9. Examination of the Cleveland Dataset (Latha & Jeeva, 2019)

To summarize a few of the attributes in the Cleveland dataset we can conclude that the dataset included patients from the range 29 to 79 age, the male patients were given the value of 1 and the females were given 0. To indicate any sort of heart disease four types were denoted, Type 1 is Typical Angina which is when blood flow to the heart is reduced resulting in chest pain. (Mayo Clinic,2022). Type 2 is Atypical Angina, Type 3 is indicated as Non-Angina pain, and Type 4 was considered Asymptomatic. The fourth feature of the dataset was Trestbps which is the resting blood pressure measure ranging from 94 to 200. The next attribute is Chol ranging from 126 to 564. Fasting blood sugar (Fbs) was denoted as 1 if the blood sugar is below 120mg/dl and 0 if it was above. Thalach is the maximum heart rate achieved ranging from 71 to 202. Exercise- induced Angina (exang) was given the value of 0 if there is no pain and 1 if there is pain. The target or the num attribute is denoted as 1 if the patient is diagnosed with heart disease and 0 for normal patients.

Given the dataset, there are several competing algorithms that can be considered immediately. They are considered in no order and where appropriate, their general relative merits and disadvantages will be considered as pertaining to this dataset. We will not perform a full training regime on the data.

Python and Libraries:Python is a high-level programming language that is

currently widely used for scientific computing. Its interactive nature and powerful libraries such as Scikit learning, NumPy, Matplolib, and Pandas have positively impacted Data Science. Scikit- learn is a comprehensive and open-sourced machine-learning package that includes a collection of efficient machine-learning methods. (Hao & Ho, 2019). This collection of methods includes data transformation, supervised and unsupervised learning, selection, and model evaluation which are important topics related to machine learning. (Hao & Ho, 2019)

Supervised Learning:Supervised learning is mapping between feature variables and correlating target variables implemented by machine learning algorithms, (Hao & Ho, 2019). One of the main conditions of supervised learning is that both the feature and target variable’s labels are known. The labeled datasets are then used to train the machine learning algorithm until it can find patterns between feature and target variables. Once the supervised learning algorithm is finished training on a given dataset to find a pattern and thus build a model, the trained model is then introduced to the testing dataset where labels are intentionally not revealed. The purpose of this is to measure the accuracy the model accomplishes on an unlabeled dataset. In addition, depending on the results the model can be fine-tuned to achieve higher accuracy.

Data Set Analysis and Results

Initially, to build a supervised machine model we are going to take the following steps:

1. Get the Cleveland dataset ready for use
2. Choose the right algorithm for our dataset
3. Fit the model and use it to make predictions on our data
4. Evaluate a model
5. Improve the model by changing its parameters
6. Save and load a trained model

In this project, we used PyCharm which is an Integrated Development Environment primarily used by Python developers for its wide range of essential tools to build our machine learning model. At first, we imported the following library Pandas to read our dataset, following our dataset being uploaded to the environment we created two variables ‘X’ and ‘Y ’to assign all the features of the dataset to ‘X’ except for the column “Target’ which was assigned to Y. After successfully reading and writing variables we move on to select a suitable algorithm, in this case, we selected RandomForest Classifier. We imported RandomForestClassifier from the Scikit-learn package and instantiated it to ‘Clf’. We then split the dataset into training and testing, for our model we selected 75 percent of the dataset to be used for our training purpose and the remaining 25 percent for testing. Using Scikit-learn ‘Model Selection’ method we were able to train, test, and fit our dataset. Upon training, we evaluated our model’s score which was 89.4 percent Accuracy. Our model was able to predict if the patient has heart disease with high accuracy. Furthermore, to determine if tuning hyperparameters yielded a better model, we created a loop to change one of its parameters called ‘N\_estimators’ and increased it by 10 until 50, and recorded the results. However, we found in our case that the default hyperparameters for RandomForest Classifier yielded the highest accuracy.

Table 1. Hyperparameter: N\_Estimator Results

|  |  |
| --- | --- |
| Accuracy | N\_Estimator |
| 81.58% | 10 |
| 88.16% | 20 |
| 87.43% | 30 |
| 85.53% | 40 |
| 89.1% | 50 |

To understand our score table below we will have to first introduce you to the following.

* 1. True Positive (TP): When the model predicted as positive, and the case was positive.
  2. True Negative (TN): At the time when the model predicted the instance as negative, and the case was negative.
  3. False Positive (FP): When the model predicted the as positive, although the case was negative.
  4. False Negative (FN): While the model predicted the case to be negative, but it was positive.

Table 2. Confusion Matrix Score

|  |  |  |  |
| --- | --- | --- | --- |
| Precision | Recall | F1-Score | Support |
| 0.88 | 0.96 | 0.92 | 46 |

Given our introduction to TP, TN, FP, FN we will now try to understand our score table above:

1. Precision: It is defined by as the ratio of ‘True Positive’ to the sum of ‘True Positive’ and ‘False positive’. In other words, it is the accuracy of the positive prediction.

The mathematical formula is TP / (TP + FP)

1. Recall: This is defined as the ratio of ‘True Positives’ to the sum of ‘True Positive’ and ‘False Negative’, it the fraction of positives that were correctly defined.

The mathematical formula is TP / (TP + FN)

1. F1-Score: It is the value of weighted mean of ‘Precision’ and ‘Recall’. This score would address the question of ‘What percent of positive predictions were right?

The mathematical formula is 2\*(Recall\*Precision) / (Recall + Precision)

# Software and Hardware Requirement

Operating System: Window 7 or higher

Network: Wi-Fi Internet or cellular Network

Programming: Python 3.6 and related libraies

Software: Google Collab

Processor: above 500 MHz

Ram: 4GB

Input Device: Standard Keyboard and Mouse

# METHODOLOGY

We introduce the following Machine Learning algorithms used in predicting heart disease; SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and RandomForest, we also list some of the advantages and disadvantages of using these algorithms. Finally, we move forward using RandomForest Classifier algorithm for building our optimized model.

Machine Learning Algorithms for Cardiovascular Disease Prediction

Machine learning has been widely employed in a variety of medical prediction datasets, and cardiovascular disease prediction is the primary among them. In particular, the medical problem of identifying high-risk patients early on is notably important, as cardiovascular incidents are often fatal; clearly, a timely diagnosis, or even better, preventative care, is a worthy goal.

# Decision Trees

One of the most often used forms of supervised learning, a decision tree is a powerful prediction-making/ categorization algorithm that uses previous data to progress from root nodes to decision nodes to leaf nodes. In its most basic form, information is split along branches and ultimately into leaf nodes. The dataset contains independent variables; data pruning without pertinent medical knowledge presents challenges, and a continuous variable decision tree must be employed given the nature of the data. Thus, multiple variables can be considered in this way which is crucial for creating an accurate model. Shown below is an early decision tree model created by Lee Goldman in 1996 (Qamar et al., 1999):

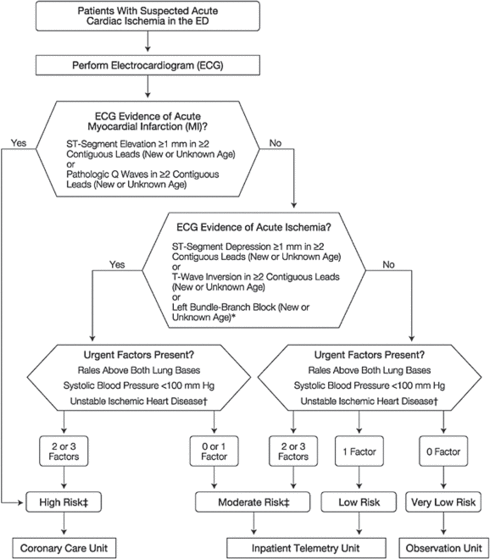


Figure 1. Decision Tree Model (Qamar et al., 1999)

Given the relatively small dataset in the study by Maheswari & Pitchai (2019), good results have been achieved using only decision trees in the past. In particular, the ease of visualization makes it easily understood by non-technical personnel. However, decision trees are not without disadvantages: they are prone to be affected by noise in the data and skew easily on certain datasets.

Naïve Bayes

The Naïve Bayes algorithm is another excellent option for classification problems. It is based on the Bayes theorem (Chauhan, 2022):

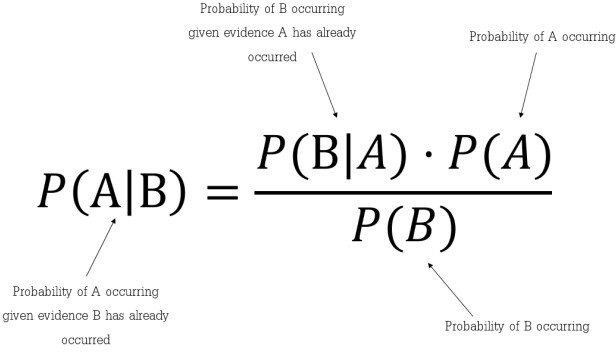


Figure 5. Bayes Theorem Probability (Chauhan, 2022)

The Naïve Bayes theorem is a slight variation on this, as it assumes that each feature is both independent and equal in its contribution to the outcome: *Naïve Bayes Classifier*



It should be noted that this assumption is generally not held to be accurate, however for the purposes of machine learning from a large data set, it is ‘good enough’; thus, the moniker ‘naïve’.

It should also be noted that it is a probabilistic classifier, meaning it predicts based on the probability of an object.

Naïve Bayes has been used in numerous studies to detect cardiovascular diseases (Miranda et al.,2016). Studies indicate that it requires short computational time and achieves good performance with the proviso that a large

training data set is required. Additionally, image data and other esoteric data forms that can be obtained through data mining have not yet yielded good results using Naïve Bayes.

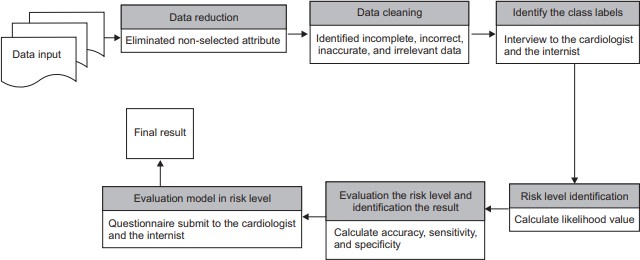
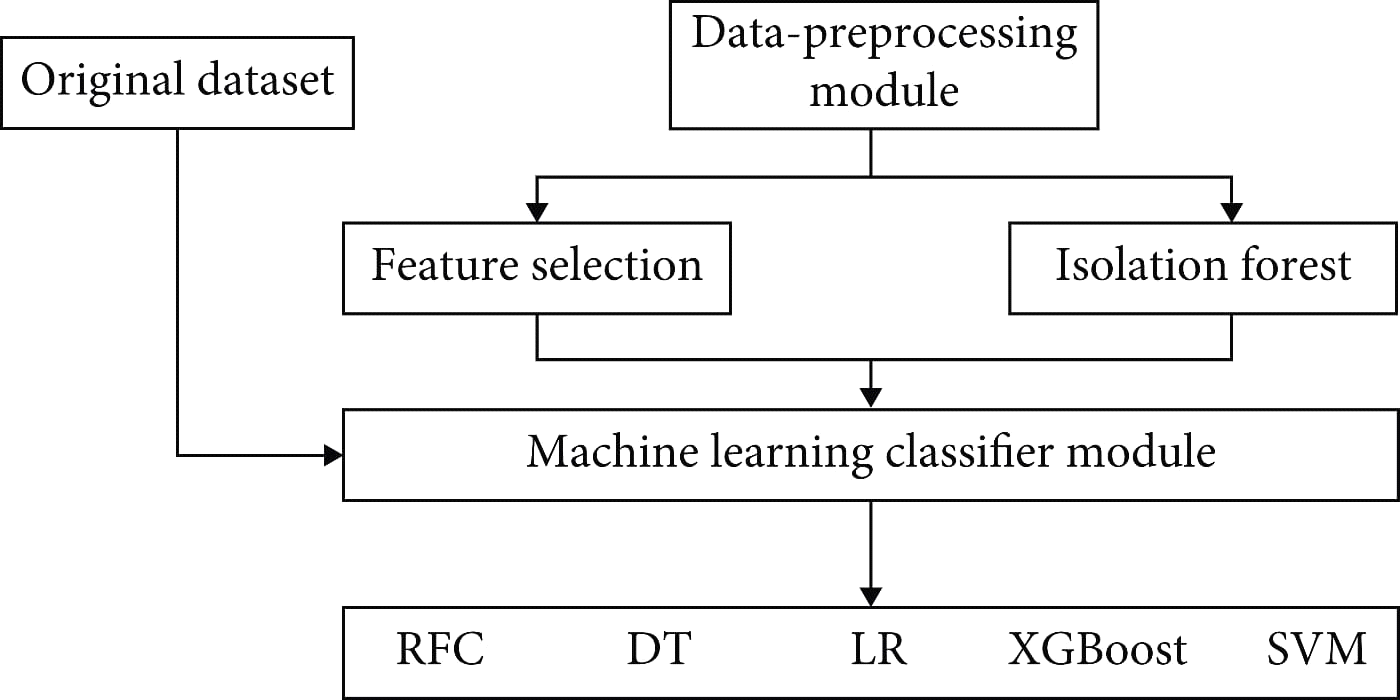


Figure 6. Naïve Bayes Model for Cardiovascular Disease Risk’s Level Detection

(Miranda et al.,2016)

# Machine Learning Classifiers Proposed

The proposed approach was applied to the dataset in which firstly the dataset was properly analyzed and then different machine learning algorithms consisting of linear model selection in which Logistic Regression was used. For focusing on neighbor selection technique KNeighbors Classifier was used, then tree-based technique like DecisionTree Classifier was used, and then a very popular and most popular technique of ensemble methods RandomForest Classifier was used. Also for checking the high dimensionality of the data and handling it, Support Vector Machine was used. Another approach which also works on ensemble method and Decision Tree method combination is XGBoost classifier

# Deep Learning Pseudocode:

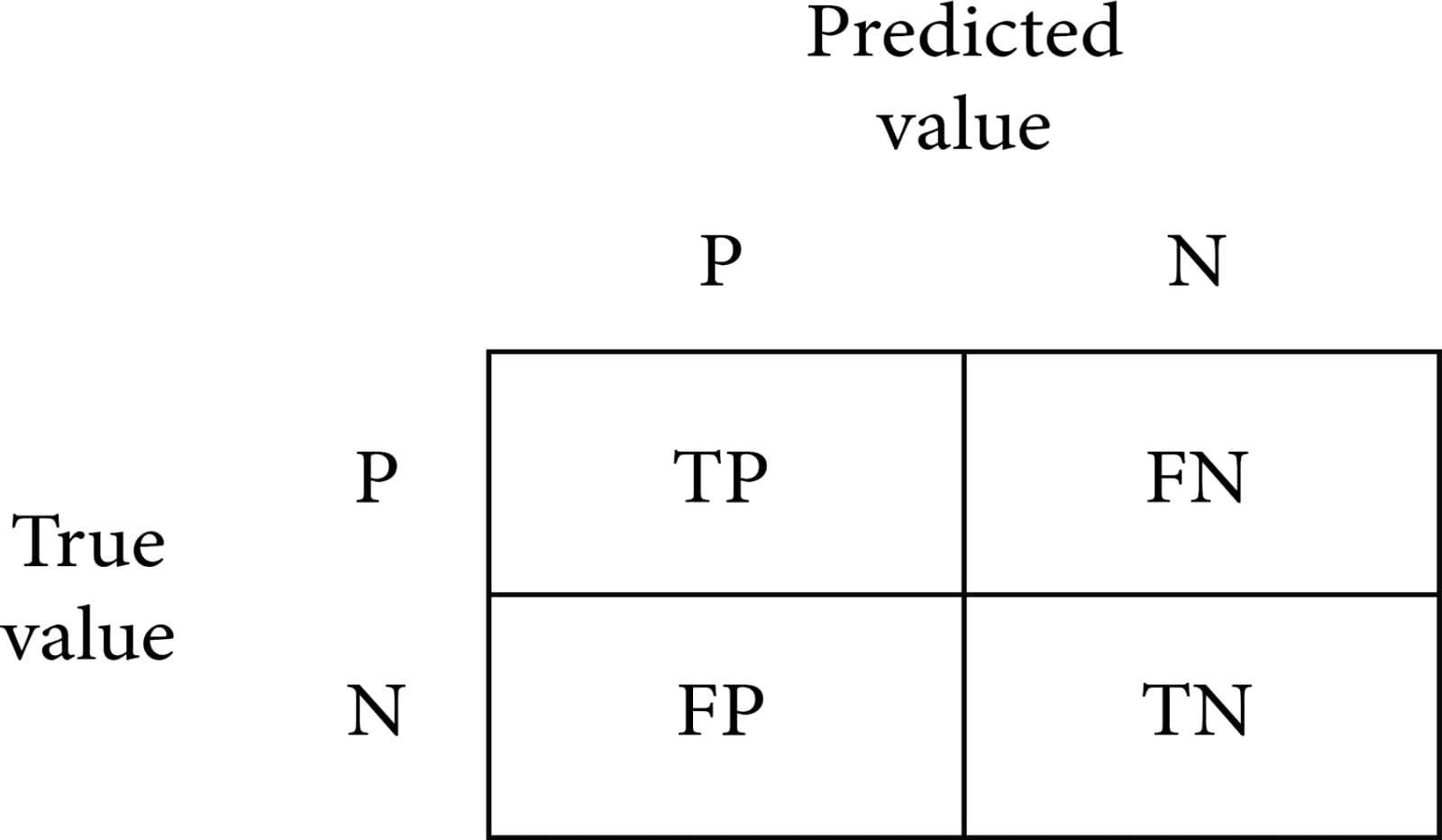
* Dataset of training
* Dataset of testing
* Checking the shape/features of the input
* The procedure of initiating the sequential layer
* Adding dense layers with dropout layers and ReLU activation functions
* Adding a last dense layer with one output and binary activation function
* End repeat
* L (output)
* End procedure

# Deep Learning Proposed:

There are two ways a deep learning approach can be applied. One is using a sequential model and another is a functional deep learning approach. In this particular research, the first one is used. A sequential model with a fully connected dense layer is used, with the flatten and dropout layers to prevent the overfitting and the results are compared of the machine learning and deep learning and variations in the learning including computational time and accuracy can be analyzed and can be seen in the figures further discussed in the Results section.

# Evaluation Process Used:

For the evaluation process, confusion matrix, accuracy score, precision, recall, sensitivity, and F1 score are used. A confusion matrix is a table-like structure in which there are true values and predicted values, called true positive and true negative. It is defined in four parts: the first one is true positive (TP) in which the values are identified as true and, in reality, it was true also. The second one is false positive (FP) in which the values identified are false but are identified as true. The third one is false negative (FN) in which the value was true but was identified as negative. The fourth one is true negative (TN) in which the value was negative and was truly identified as negative.



P=positive, N=negative, TP=true positive, FN=false negative, FP=false positive, TN=true negative.

Then for checking how well a model is performing, an accuracy score is used. It is defined as the true positive values plus true negative values divided by true positive plus true negative plus false positive plus false negative. The formula is

accuracy = TP+TN/TP+TN+FP+FN.

After accuracy there is specificity which is the proportion of true negative cases that were classified as negative; thus, it is a measure of how well a classifier identifies negative cases. It is also known as the true negative rate. The formula is

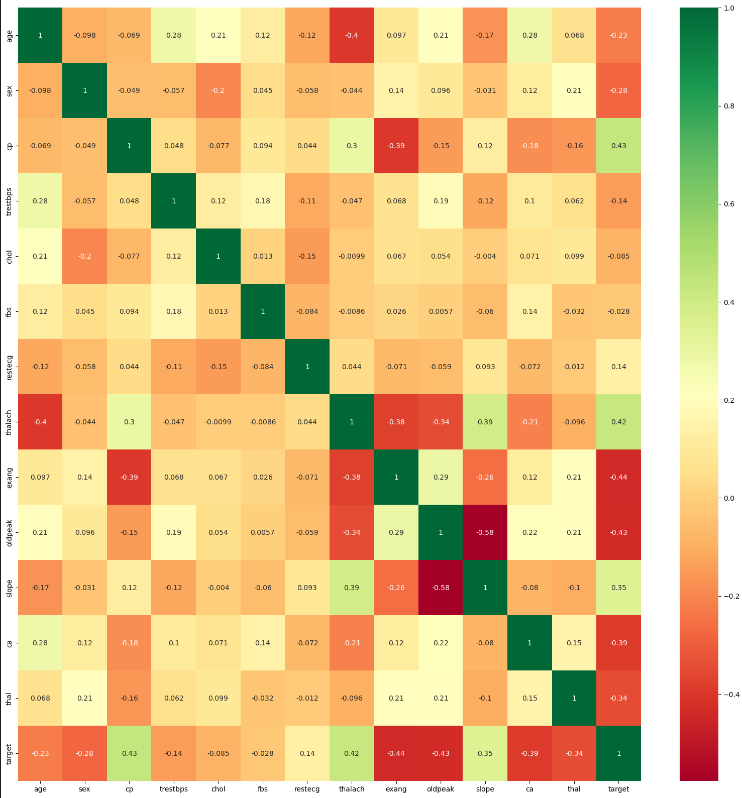
Specificity = TN/TN+FP.

Then there is sensitivity in which the proportion of actual positive cases got predicted as positive (or true positive). Sensitivity is also termed as recall. In other words, an unhealthy person got predicted as unhealthy. The formula is

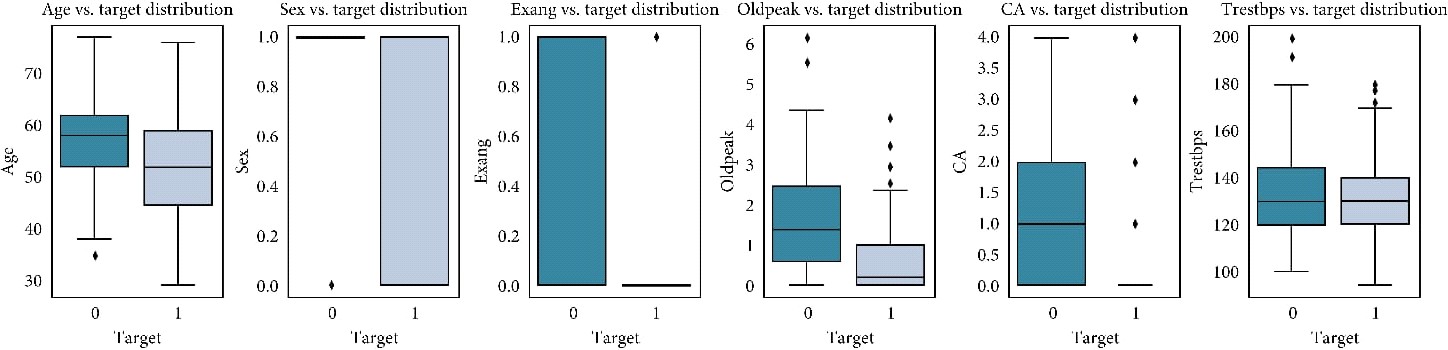
Sensitivity = TP/TP+FN

# Using the First Approach (without Doing Feature Selection and Outliers Detection):

As can be seen in the dataset is not normalized, there is no equal distribution of the target class, it can further be seen when a correlation heatmap is plotted, and there are so many negative values.



So, even if the feature selection is done, still, we have outliers.



By applying the first approach, the accuracy achieved by the Random Forest is 76.7%, Logistic Regression is 83.64%, KNeighbors is 82.27%, Support Vector Machine is 84.09%, Decision Tree is 75.0%, and XGBoost is 70.0%. SVM is having the highest accuracy here which is achieved by using the cross-validation and grid search for finding the best parameters or in other words doing the hyperparameter tuning. Then after machine learning, deep learning is applied by using the sequential model approach. In the model, 128 neurons are used and the activation function used is ReLU, and in the output layer which is a single class prediction problem, the sigmoid activation function is used with loss as binary cross-entropy and gradient descent optimizer as Adam. The accuracy achieved is 76.7%.

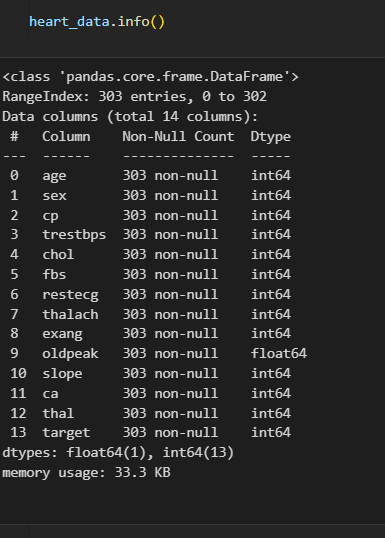
# Using the Second Approach (Doing Feature Selection and No Outliers Detection) :

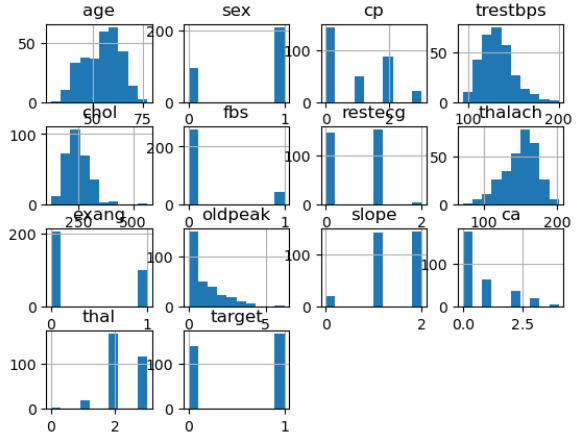
After selecting the features (feature selection) and scaling the data as there are outliers, the robust standard scalar is used; it is used when the dataset is having certain outliers. In the second approach, the accuracy achieved by Random Forest is 88%, the Logistic Regression is 85.9%, KNeighbors is 79.69%, Support Vector Machine is 84.26%, the Decision Tree is 76.35%, and XGBoost is 71.1%. Here the Random Forest is the clear winner with a precision of 88.4% and an F1 score of 86.5%. Then deep learning is applied with the same parameters before and the accuracy achieved is 86.8%, and the evaluation accuracy is 81.9%, which is better than the first approach.

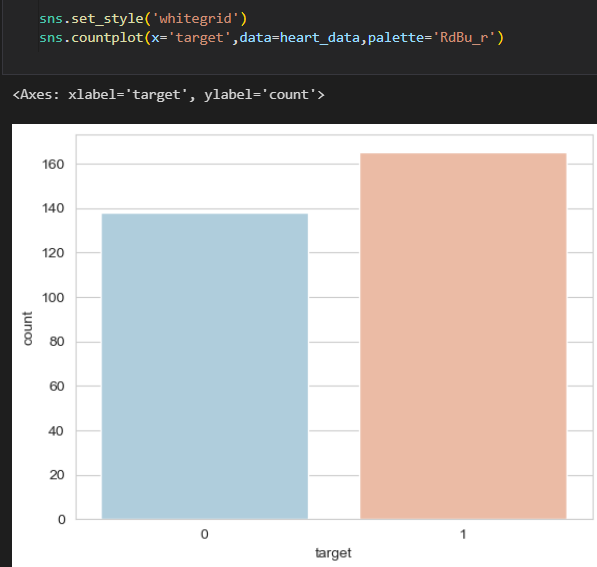
# Using the Third Approach (by Doing Feature Selection and Also Outliers Detection):

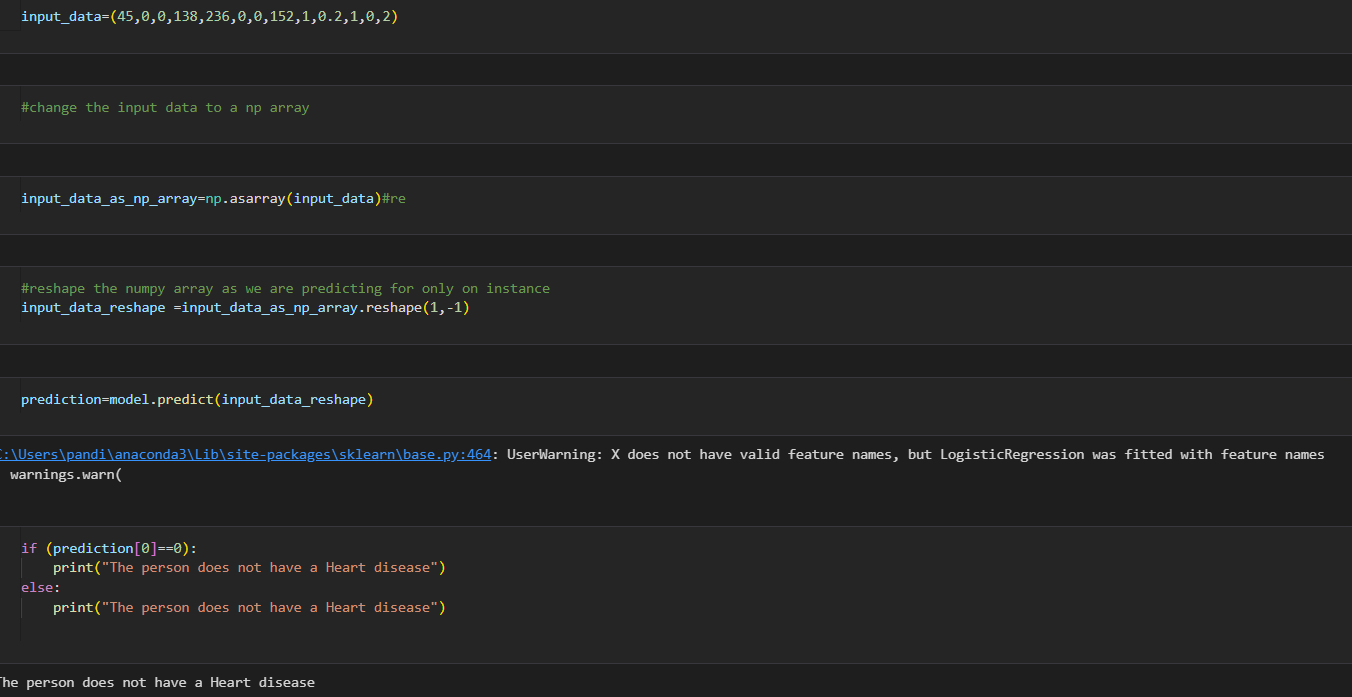
In this approach, the dataset is normalized and the feature selection is done and also the outliers are handled using the Isolation Forest. The correlation comparison can be seen in Figure 10. The accuracy of the Random Forest is 80.3%, Logistic Regression is 83.31%, KNeighbors is 84.86%, Support Vector Machine is 83.29%, Decision Tree is 82.33%, and XGBoost is 71.4%. Here the winner is KNeighbors with a precision of 77.7% and a specificity of 80%. A lot of tips and tricks for selecting different algorithms are shown by Garate-Escamila et al. [38]. Using deep learning in the third approach, the accuracy achieved is 94.2%. So, the maximum accuracy achieved by the machine learning model is KNeighbors ( 83.29%) in the third approach, and, for deep learning, the maximum accuracy achieved is 81.9%. Thus, the conclusion can be drawn here that, for this dataset, the deep learning algorithm achieved 94.2 percent accuracy which is greater than the machine learning models. We also made a comparison with another research of the deep learning by Ramprakash et al. [39] in which they achieved 84% accuracy and Das et al. [33] achieved 92.7 percent accuracy. So our algorithm produced greater accuracy and more promising than other approaches. The comparison of different classifiers of ML and DL

# CODING AND TESTING









**DISCUSSION OF 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 KFold, 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. 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.

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

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

# 

# 

# CONCLUSION AND FUTURE ENHANCEMENTS

Conclusion

Heart disease is a life-threatening disease affecting millions of people around the world every year (Asadi et al., 2021). Hence, early prediction of heart disease can benefit patients and healthcare professionals by providing the information they need to minimize death and reduce costs. Since medical big data has been increasing daily and data storage costs decreasing, machine learning algorithms can play an important part in processing these medical data and predicting diseases.

With the help of the RanfomForest Classifier algorithm, we were able to build a machine-learning model. Our model was trained and tested by a dataset from the UCI repository. The dataset consisted of labeled 303 patients, it included both diagnosed heart disease patients and normal patients. After the model was trained and then tested, we achieved an accuracy of 89.4% with the default hyperparameter. While we tried to tune RandomForest Classifier’s hyperparameter; N\_estimator in the hope of higher accuracy, we noticed that the default resulted in the highest.

We can conclude that machine learning and data mining can play an important role in our healthcare system. Traditionally, diagnosis of the disease was performed by standard procedures and doctor’s intuitions which had limitations and led to costly expenses, but with machine learning models, diagnosis can be done on large datasets cost-effectively.

Areas for Further Study

As we have developed a supervised machine-learning model using the RandomForest algorithm and tuning one of its hyperparameters called ‘N\_Estimator’, in the future this model can be trained and tested using a larger set of data with additional attributes. Additionally, our model holds an opportunity for further research to be performed by modifying different hyperparameters.

# REFERENCES

|  |  |
| --- | --- |
| [1] | A. 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://towardsdatascience.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..](http://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". |