ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Research Article

Open Access

Prediction of Heart Disease Using Voted Perceptron

Safia Naveed S

Women Scientist, Under WISE KIRAN IPR, TIFAC, DST, India

ABSTRACT

Heart Disease is the most dominating disease which is taking a large number of deaths every year. A report from WHO in 2016 portrayed that every year at least 17 million people die of heart disease. This number is gradually increasing day by day and WHO estimated that this death toll will reach the summit of 75 million by 2030. Despite having modern technology and health care system predicting heart disease is still beyond limitations. As the Machine Learning algorithm is a vital source predicting data from available data sets we have used a machine learning approach to predict heart disease. We have collected data from the UCI repository. In our study, we have used Random Forest, Zero R, Voted Perceptron, K star classifier. We have got the best result through the Random Forest classifier with an accuracy of 97.69.

***Corresponding author**

Safia Naveed S, Women Scientist, Under WISE KIRAN IPR, TIFAC, DST, India. E-mail: sweetysafia@gmail.com

Received: December 15, 2022; **Accepted:** December 21, 2022; **Published:** December 30, 2022

Keywords: Heart Disease, Prediction, Machine Learning, Kstar, Random Forest, Voted Perceptron, Zero R

Introduction

Cardiovascular disease which is termed heart disease is the number 1 cause of death in the whole world taking at least 17 million people's death every year [1]. Though at least three quarters of death has occurred in the low and middle-income country rate of death is also alarming in developed countries. According to the Center for Disease Control and Prevention, at least 25 percent of death occurred due to heart disease in the USA. This situation is also depicted in another country of different ages, races, classes, etc. Though medical science has progressed tremendously all over the world but preventing different types of heart disease is yet to possible. In Bangladesh, from 1986 to 2006 death from heart disease increase at least 3527% whereas the death of dysentery and respiratory infection reduced by 79% and 86% [2]. The most alarming matter of heart disease is that most of the people are suffering from heart disease at most productive years of their life stated that in India 50% percent of heart disease occurred before 50 years whereas at least 25 percent faces the same disease before 40 years. As the low-income countries are lacking basic health care facilities they don't get proper guidelines about heart disease which causes death, as well as huge costs in medication, lead every family towards poverty. Though heart disease is creating a catastrophic moment to both patient and health authority still vital challenge is to predict and detect its presence in the human body despite having different techniques [3,4]. That's why to decrease the death rate and protect every family from economic vulnerability prediction of heart disease is an important factor which will ultimately help policymakers to take appropriate step anent heart disease. Machine learning is a useful instrument to conclude a huge number of data regarding health, technology, business, etc. It can assist in increasing access and analysis of health care facilities in developing countries. The instrument of the decision tree, naïve bays, support vector machine can be used in predicting heart disease which will be more efficient than other techniques [5]. Therefore, in our study, we have used machine learning algorithms to predict heart disease.

Literature Survey

Heart disease is the most important issue and a common problem in the total world. Thousands of people died of heart disease every year. For this reason, many researchers are trying to predict this cardiovascular disease which is a critical challenge in the area of clinical data analysis. In this paper, had proposed a unique method to predict heart disease by using machine learning techniques [6]. This prediction model was done with different combinations of features which were known as classification techniques. They had used several classification methods like data pre-processing, feature selection and reduction, different classification modeling, decision trees, language model, support vector model, and random forest. Finally, with the hybrid random forest with a linear model (HRFLM) they had able to show the accuracy level of 88.7% through the prediction model for heart disease. Heart disease is one of the leading causes of death nowadays. As it is a complex task sometimes predicting the heart attack is more difficult for medical practitioners because of less knowledge and experience. Not only that sometimes the health sector hides some information that is needed for making decisions. Showed a model to predict heart disease [7]. There are different data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net were used in this research for predicting heart attacks. Last of all the research result showed the prediction accuracy which was 99%. And this research also showed that data mining enabled the health sector to predict patterns in the dataset. From an investigation, constructed a model which could detect the symptoms which will be helpful to prevent heat stroke at an early age, and day by day its increasing rate had

been developed [8]. They proposed an application that would use to show the symptoms like age, sex, pulse rate, etc, and would able to predict heart disease. They had used machine learning algorithm neural networks to find the best accuracy of heart disease. Due to many reasons, heart disease is increasing rapidly. Although different health care centers and doctors collect data daily as they don't use machine learning and pattern machine techniques it is reducing their predictability. For this reason, showed a prediction model [9]. In this paper, they had collected data and attributes for the UCI repository. By using this data, they had tried to predict heart disease. For this development,they had usedseveral techniques in Artificial Neural Network(ANN). They had shown accuracy such as 94.7% for ANN but 97.7% accuracy rate for Principle Component Analysis(PCA). Collected the information for prediction from the UCI repository [10]. 1025 Instances with14 attributes dataset were used for this prediction model. After accomplishing this research, they had proposed a model and analyzed classification accuracy, precision, and sensitivity by four tree-based classification algorithms like M5P, random Tree, and Reduced Error Pruning with the Random forest ensemble method. After the feature selection of the heart patient's dataset, all the prediction algorithms were used. They had used three features-based algorithms like Pearson Correlation, Recursive Features Elimination, and Lasso Regularization. Three experimental setups were used to finish this analysis. Pearson Correlation on M5P, random Tree, Reduced Error Pruning, and Random forest ensemble method was applied for the first experiment. In the second experiment, Recursive Features Elimination and application on the above four tree-based algorithms were used. And for the third experiment Lasso Regularization and applied on as above tree-based algorithms were used. After completing this experiment, they had analyzed and calculated classification accuracy, precision, and sensitivity. Finally, they were capable to show the best accuracy that was 99% and it is conducted by feature selection methods Pearson correlation and Lasso Regularization with random forest ensemble method.

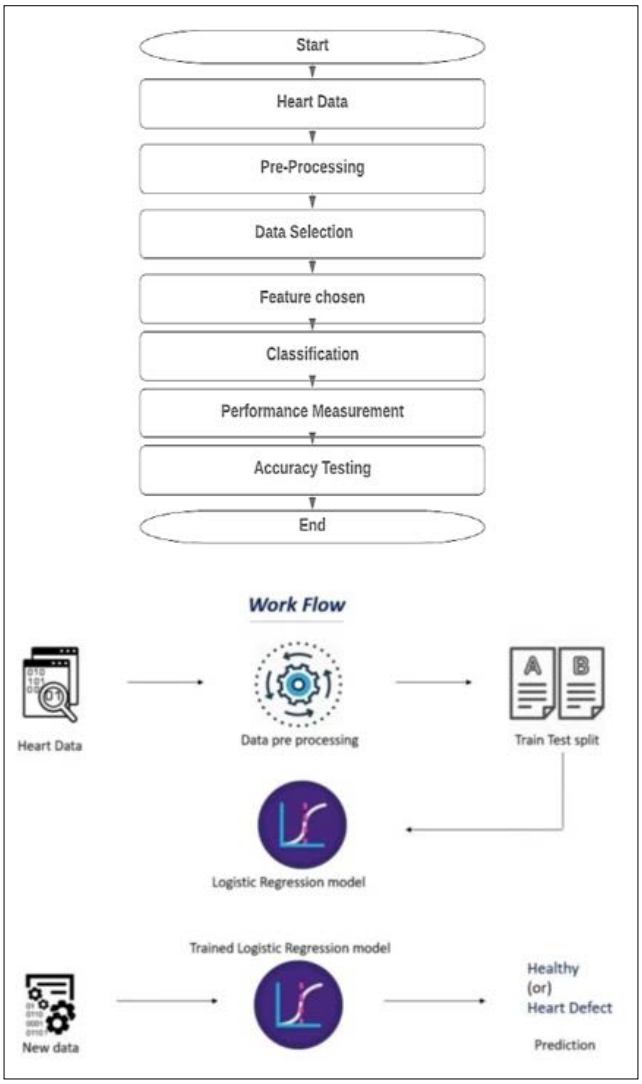
Optimization Techniques

Suggested few optimization techniques such as Particle Swarm Optimization (PSO), Discrete Particle Swarm Optimization (DPSO) and Fractional Order Discrete Particle Swarm Optimization (FODPSO) Techniques based onwhich best or optimum features from the face of the chauffeur can be selected to denote his drowsiness [11].

Also suggested that the region of interest can be captured and the same region can be inspected thoroughly in terms of pixel values, evaluating the degree of noise present, analysing the position of boundaries to study the region of interest, analyse the fluctuating intensities across the forehead region of the face and also consider the edge effects of the eyes [12].

Methodology

Our proposed methodology showing in Fig.1 the flow chart. The steps by steps approach are discussed. We use the UCI data repository for our machine learning approach. Data set managing, collecting the data set features, pre-process the data set, choosing the feature, classify the instances, measure the performance of the classifiers, compare the accuracy and last the result is acquired. Four machine learning techniques are applied to examine the accuracy rate for our heart data set. Evaluating the performance, we tuned for improve the accuracy rate. Later in that, the confusion matrix for each machine learning technique has been visualized for the validity of the experimental model. After preprocessing the data set and cleaning the data we use nine attributes that make sense for our experiment. There were fourteen attributes but did not take all of those to grant as some of the attributes did not make any sense at all.



Random Forest

It is a supervised machine learning algorithm, it makes decision trees, it highly depends on the trained data set. It is a learning model for getting the results. Plotting trees randomly for getting the highest and stable predictions.

Algorithm (RandomForest):

Training Set :D=(a1,b1)….(an,bn) Feature is :F
Number of trees:N Function<-RANDOM(D,F)
H<-0
for i from D(ai,bi)….to N do D(i)<-Sample from D
Hk<-RANDOMLEARN(D(i),F)
End

ZeroR

It is the least difficult order strategy which depends on the objective and overlooks all indicators. ZeroR classifier predicts the larger part classification (class). Even though there is no consistency power in ZeroR, it helps decide a pattern execution as a benchmark for other order strategies.

Voted Perceptron

The voted perceptron method is based on the perceptron algorithm of Rosenblatt and Frank [17]. The calculation exploits information that is directly distinct with huge edges. This strategy is less complex to execute, and considerably more proficient as far as calculation time when contrasted with Vapnik's SVM. The calculation can likewise be utilized in high dimensional spaces utilizing piece capacities.

K-Star

It is used for finding the depth in the field or this case the depth in the accuracy. Model-based upon two training sets, predicts the values in the nearest first looking mode. It works instantly, super-fast classification for the training sets.

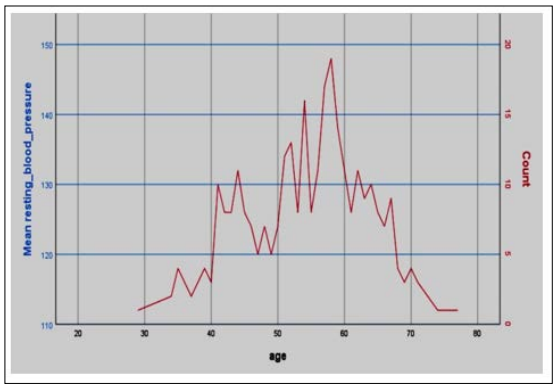


Figure 1: Scatter Point Among Age and Maximum Heart Rate

Results

The experimental results were obtained through different analyses. There were test data set which was roughly 30% and the training consisted with 70%, a total of 100% data is converged with the experiment. Table showing the different accuracy from the test data set.

Table 1: Test Dataset Results

Classifier Name	Accuracy (%)
Random Forest	96.703
ZeroR	90.10
Voted Perceptron	94.20
Kstar	97.80

Below Table showing the training result of our data set.

Table 2: Training Data Set Results

Classifier Name	Accuracy (%)
Random Forest	97.69
ZeroR	85.14
Voted Perceptron	94.39
Kstar	94.05

The result from test data and training data varies. Fig clearly showing that there is a relation between age and maximum heart rate. Heart rate increase with age. Table shows the confusion matrix of each classifier. The confusion matrix results were acquired from the training data set as the test data set worked with only 30% of the total data set. After the confusion matrix result, we can surely say that the result of Random Forest is higher than the rest of the four algorithms. The true positive rate of Random Forest is 0.967 and the false positive is 0.300. With having the roc and PRC area in both cases is 1.

Age-High Blood Pressure Relation showing a relation between age and high blood pressure. Our finding showed there's a relation between age and high blood pressure. People are likely to develop high pressure in their 40s or 50s. So better to follow up with a personal physician after the late 30s.

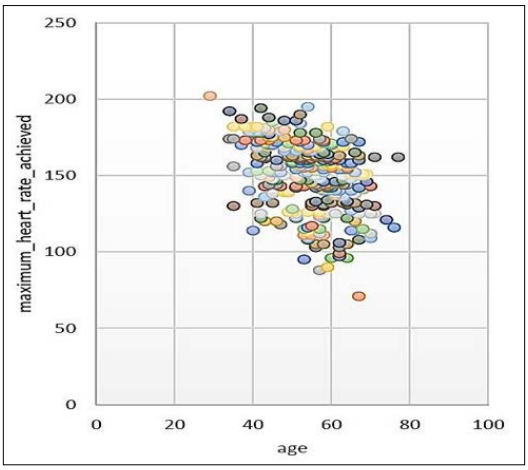


Figure 2: Confusion Matrix of the Classifiers

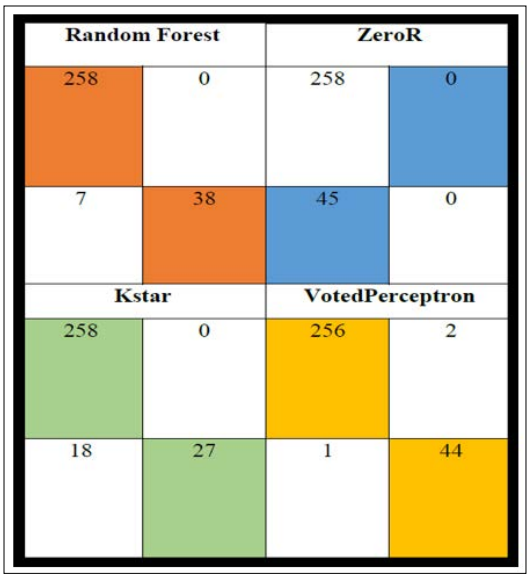
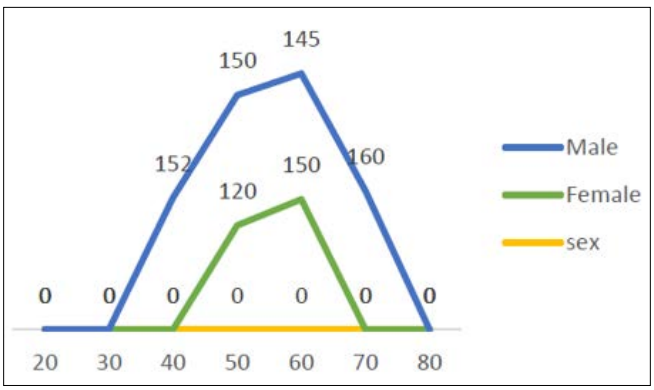


Figure 3: Resting Blood Sugar and The Relation with Age and Sex



Showing a visualization between sex, age, and resting blood sugar. The resting blood sugar is on the higher side if the age increase and vice versa.

Conclusion

Heart disease is a crucial health problem all over the world. A proper and scientific prediction approach can mitigate the loss of heart disease. we have constructed a technique to predict heart disease by using machine learning algorithms. In our study, we have used Random Forest, Kstar, Zeror, Voted Perceptron classifier. The accuracy we have got from our study that Random Forest classifier is 97.69%, ZeroR is 85.14%, Voted Perceptron is 94.39%, Kstar is 94.05%. Among the classifier we have used in our study, we have found that the Random Forest classifier has produced the best result with an accuracy of 97.69 percent. In our future work, we will use a more accurate data set to get better results with more technological and scientific knowledge in another branch of medical science.

Limitation and Future Work

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research / work need to be performed for the future. Would like to make use of testing different discretization techniques, multiple classifier voting technique and different decision tree types namely information gain and gain ratio. Willing to explore different rules such as association rule, logistic regression and clustering algorithms.

Source Code: Importing the Dependencies
import numpy as np import pandas as pd
from sklearn.model_selection import train_test_split from sklearn.linear_model import LogisticRegression from sklearn.metrics import accuracy_score

Data Collection and Processing
loading the csv data to a Pandas DataFrame heart_data = pd.read_csv('/content/data.csv') # print first 5 rows of the dataset heart_data.head()
print last 5 rows of the dataset heart_data.tail()
number of rows and columns in the dataset heart_data.shape
getting some info about the data heart_data.info()
checking for missing values heart_data.isnull().sum()
statistical measures about the data heart_data.describe()
checking the distribution of Target Variable heart_data['target'].value_counts()
1 --> Defective Heart 0 --> Healthy Heart

Splitting the Features and Target

X = heart_data.drop(columns='target', axis=1) Y = heart_data['target']
print(X) print(Y)

Splitting the Data into Training data & Test Data

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2) print(X.shape, X_train.shape, X_test.shape)

Model Training

Logistic Regression

model = LogisticRegression()

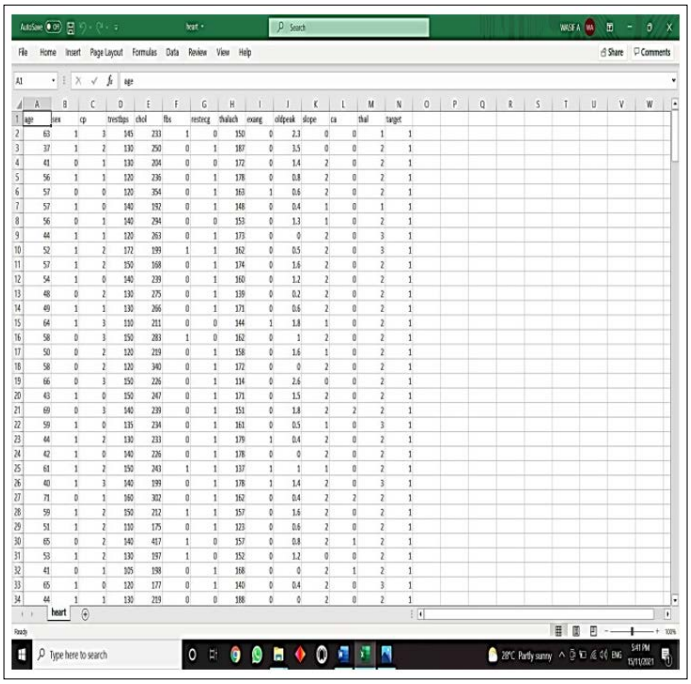
training the LogisticRegression model wth Training data model.fit(X_train, Y_train)

Model Evaluation

Accuracy Score

accuracy on training data X_train_prediction = model.predict(X_train) training_data_accuracy = accuracy_score(X_train_prediction, Y_train) print('Accuracy on Training data : ', training_data_accuracy)

# accuracy on test data
X_test_prediction = model.predict(X_test) test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy on Test data : ', test_data_accuracy)
Building a Predictive System
input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)
# change the input data to a numpy array input_data_as_numpy_array= np.asarray(input_data)
# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
prediction = model.predict(input_data_reshaped) print(prediction)
if (prediction[0]== 0):
print('The Person does not have a Heart Disease') else:
print('The Person has Heart Disease')csv file



References

1. World Health Organization/health-topics/cardiovascular-diseases.
2. Ahsan Karar Z, Alam N, Kim Streatfield P (2009) Epidemiological transition in rural Bangladesh 1986–2006 Global health action. 2: 1904.
3. Mittal R A (2017) Increasing heart attacks in young Indians the Times of India.
4. M A Jabbar, P Chandra, B L Deekshatulu (2012) Prediction of risk score for heart disease using associative classification and hybrid feature subset selection. Int Conf Intell Syst des Appl ISDA 628-634.
5. Sharma H, Rizvi M A (2017) Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication 5: 99-104.
6. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques IEEE Access 7: 81542-81554.
7. Masethe H D, Masethe M A (2014) Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science 2: 22-24.
8. Gavhane A, Kokkula G, Pandya I, Devadkar K (2018) Prediction of heart disease using machine learning. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) IEEE 1275-1278.

9. Awan S M, Riaz M U, Khan A G (2018) Prediction of heart disease using artificial neural network. VFAST Transactions on Software Engineering 6: 51-61.
10. Yadav D C, Pal Saurabh (2020) Prediction of heart disease using feature selection and random forest ensemble method International Journal of Pharmaceutical Research 12.
11. Safia Naveed S, Geetha G, Leninisha S (2020) Early Diabetes Discovery From Tongue Image. The Computer Journal 65: 237-250.
12. Safia Naveed S, Gurunathan Geetha (2019) Intelligent Diabetes Detection System based on Tongue Datasets Current Medical Imaging Reviews 15: 672-678.