DOI:

A Survey and Analysis on Predicting Heart Disease Using Machine Learning Techniques

Honey Joseph
Dept. Of Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
honeyjoseph@amaljyothi.ac.in

***Abstract*—The early prognosis of cardiovascular diseases can aid in making decisions to lifestyle changes in high-risk patients and in turn reduce their complications. This paper compares the accuracies of different machine learning algorithms on the Cleveland Heart Disease Database in order to present an accurate model of predicting heart disease.**

Keywords—Machine Learning, Classification Techniques, Prediction, Heart Disease

I. Introduction

Heart disease, also known as cardiovascular disease, refers to a group of conditions that affect the heart and blood vessels. It is the leading cause of death worldwide, and includes conditions such as coronary artery disease, heart failure, arrhythmias, and heart valve problems.

Coronary artery disease is the most common type of heart disease and is caused by a build-up of plaque in the arteries that supply blood to the heart. This can lead to chest pain, shortness of breath, heart attack, and other complications.

Heart failure occurs when the heart is unable to pump enough blood to meet the body's needs. It can be caused by a variety of factors, including coronary artery disease, high blood pressure, and diabetes.

Arrhythmias are abnormal heart rhythms that can cause the heart to beat too fast, too slow, or irregularly. They can be caused by a variety of factors, including heart disease, stress, and certain medications.

Heart valve problems occur when one or more of the heart valves does not function properly, which can lead to a variety of complications, including shortness of breath, fatigue, and heart failure.

Treatment for heart disease varies depending on the type and severity of the condition, but may include lifestyle changes, medications, medical procedures, or surgery. Prevention of heart disease involves adopting a healthy lifestyle, including regular exercise, a healthy diet, and avoiding tobacco and excessive alcohol consumption.

Predicting heart disease using machine learning techniques is a growing field in healthcare. With the help of machine learning algorithms, doctors and researchers can identify patients at high risk for heart disease and develop effective prevention and treatment plans.

This paper discussed the results of the current technique and predicted the result for heart disease. Additionally, the experiment results compare the accuracy achieved by these algorithms and evaluated results by various respective authors.

II. LITERATURE REVIEW

H. Benjamin et al. [1], in their work on the "supervised machine learning" concept used to find the predictions of heart disease, have used the following data mining classification algorithms for analysis and prediction, namely,

Naïve Bayes, Random Forest, and Decision Tree. They have proposed by experimental results and proved that Random Forest gives better result performance as compare to Naïve Bayes and Decision tree. In this research work, the dataset is sourced from the data source StatLog for creating heart disease prediction.

Senthilkumar Mohan et al. [2], Hybrid Random Forest, and novel method by using Linear Model (HRFLM) and their goals to finding the important features by using Machine learning's techniques and increase the performance and accuracy for heart disease prediction. Research work core aims to process raw data through different steps and deliver a new respective novel judgment of heart disease prediction. Their prediction model is presented by various combinations of features and numerous recognized classification methods to increase the accuracy performance result. They have done work on many classification models to predict cardiovascular heart disease and compared their accuracy. They have proposed a comparison with HRFLM. Dataset used by UCI ML repository, and their approach claimed an accuracy level of 88.7%.

Mohammad Shafenoor Amin et al. [3], in their work they suggested data mining classification methods for predicting the heart disease result. The proposed testing was used to classify important features by using data mining techniques. The Cleveland dataset was collected from the UCI ML Repository for heart disease prediction. They have used some data mining techniques, namely SVM, DT, K-NN, LR, Naïve Bayes, vote, and Neural network. They have also performed experiments on another dataset using the UCI Statlog data set to identify the verdicts. The maximum accuracy gain results for heart disease diagnostic system can proficiently predict the danger level of heart disease in the future. Their approach claimed maximum accuracy was accomplished by SVM.

III. METHODOLOGY

A. Data Analysis

The objective of data analysis step is to increase the understanding of the problem from the data. There are two approaches to describe a given dataset. Summarizing and Visualizing data.

1) Feature Infornation:

a) Age- age in year

b) sex - sex(1 = male; 0 = female)

c) chest_pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)

d) blood_pressure - resting blood pressure (in mm Hg on admission to the hospital)

e) serum_cholestoral - serum cholesterol in mg/dl

f) fasting_blood_sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

g) electrocardiographic - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)

h) max_heart_rate - maximum heart rate achieved

i) induced_angina - exercise induced angina (1 = yes; 0 = no)

j) ST_depression - ST depression induced by exercise relative to rest

k) slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 =downsloping)

l) no_of_vessels - number of major vessels (0-3) colored by fluoroscopy

m) thal - 3 = normal; 6 = fixed defect; 7 = reversible defect

n) diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50%

o) diameter narrowing; Value 1 = > 50% diameter narrowing).

2) Types of features:

a) Categorical features (Has two or more categories and each value in that feature can be categorized by them): sex, chest_pain

b) Ordinal features (Variables having relative ordering or sorting between the values): fasting_blood_sugar, electrocardiographic,no_of_vessels, thal, diagnosis, , induced_angina, slope.

c) Continuous features (Variable taking values between any two points or between the minimum or maximum values in the feature column): age, max_heart_rate, ST_depression, blood_pressure, serum_cholestoral

B. Architecture Diagram

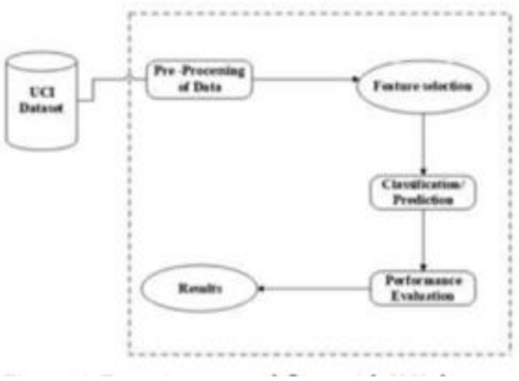


Fig. 1. Experiment workflow

IV. ALGORITHMS AND TECHNIQUES

A. K means Algorithm

When it comes to clustering problems K means is the most straightforward algorithm which can be used. It is used for clustering the data set into k no. of clusters and then find centroid for each cluster. Patient’s data set is very large so to get the output accurately, we divide the data set into 3 clusters. We divided the data set into clusters on the basis of their stress value ,1st cluster contains the people which are normal, 2nd which are stressed and 3rd which are highly stressed .Further k means give us centroid for each cluster.

B. Decision Tree Algorithm

After forming clusters, these clusters are the input for decision tree algorithm. It produces decision rules at the output. Decision tree creates a tree structure to classified data as yes → prone to heart disease or no → not prone to heart disease. For each cluster decision tree classify data as yes or no.

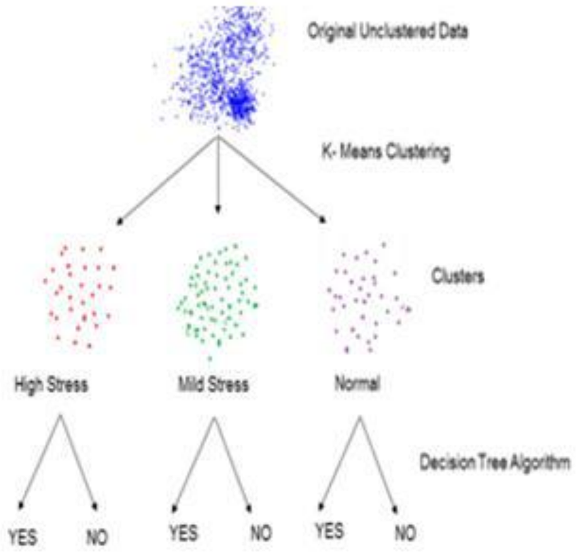


Fig. 2. Decision Tree

C. *Logistic Regression*

Logistic regression is a machine learning (ML) algorithm for classification. In this algorithm, the probabilities describing the attainable outcomes of a single trial are modelled using a logistic function.

D. *Naïve Bayes*

Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as spam filtering and
document classification.

E. *Support Vector Machine*

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. New examples are then mapped into the same space & predicted to which category they belong to base on which side of the gap they fall.

F. *Random Forest*

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls overfitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

VI. METRICS

Accuracy: It is the number of correct predictions made by the model over all kinds of predictions made

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

Accuracy is a good measure when the target variable classes in the data are nearly balanced.

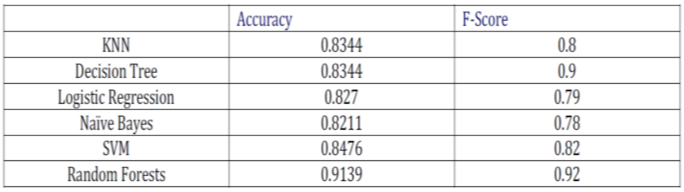
F-score: It is used to measure a test's accuracy and it balances the use of precision and recall to do it. It can provide a realistic measure of test's performance.

$$F1 = 2 \times \frac{Precision*Recall}{Precision+Recall}$$

A. *Prediction*

Out of the chosen algorithms we will start with KNN classification model. We will take a classifier and fit the training data. After that we will predict that by using predict (X_train). Now we will predict the accuracy of the testing data by using accuracy score (y_test, pred) and F-score by importing fbeta_score from sklearn.metrics. By doing so for, the KNN will give us the accuracy of 0.834. We will continue the same procedure on Naïve Bayes, SVM, Decision tree, Logistic Regression and Random Forest. By following the same procedure above that is fitting, predicting and finding the accuracy score and F-score we will get the accuracy score and F-score as below.

TABLE 1. Comparison Table



	Accuracy	F-Score
KNN	0.8344	0.8
Decision Tree	0.8344	0.9
Logistic Regression	0.827	0.79
Naïve Bayes	0.8211	0.78
SVM	0.8476	0.82
Random Forests	0.9139	0.92

From the above reports Random Forest seems to be performing well based on Accuracy and F-Score.

Acknowledgment

The goal of the paper was to compare different machine learning algorithms and predict if a certain person, given various personal characteristics and symptoms, will get heart disease or not. Now the distribution of target value is almost equal, so using standard metrics in further machine learning modelling like accuracy and AUC is justified. It does not come as a surprise that the more complex algorithms like SVM and Random Forests generated better results compared to the basic ones. It is worth to emphasize that in most cases hyper parameter tuning is essential to achieve robust results out of these techniques. By producing decent results, simpler methods proved to be useful as well. Machine learning (ML) has absolutely bright future in medical field. Just imagine a place where heart disease experts are not available. With just basic information about a certain patient's medical history, we may quite accurately predict whether a disease will occur or not.

DOI:

References

[1] M.A.Jabbar, B.L. Deekshatulu and Priti Chandra, 2015. Prediction of heart disease using Random forest and Feature subset selection, AISC SPRINGER, vol 424, pp187-196.

[2] Mr.Santhana Krishnan.J and Dr.Geetha.S, 2019. Prediction of Heart Disease Using Machine Learning Algorithms, (ICIICT) IEEE, 2019. DOI: 10.1109/ICIICT1.2019.8741465.

[3] Cincy Raju, Philipsy E, Siji Chacko, L Padma Suresh, Deepa Rajan S, 2018. A Survey on Predicting Heart Disease using Data Mining Techniques, IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018).DOI 10.1109/ICEDSS.2018.8544333.

[4] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," Artificial Intelligence in Medicine, vol. 23, no. 1, pp. 89–109, 2001. https://doi.org/10.1016/S0933-3657(01)00077-X.

[5] J. Soni et al., "Intelligent and effective heart disease prediction system using weighted associative classifiers," International Journal on Computer Science and Engineering, vol. 3, no. 6, pp. 2385–2392, 2011.

[6] N. Khateeb and M. Usman, "Efficient heart disease prediction system using k-nearest neighbor classification technique," in Proceedings of the International Conference on Big Data and Internet of Thing (BDIOT), New York, NY, USA: ACM, 2017, pp. 21–26. https://doi.org/10.1145/3175684.3175703.

[7] H. Almarabeh and E. Amer, "A study of data mining techniques accuracy for healthcare," International Journal of Computer Applications, vol. 168, no. 3, pp. 12–17, Jun 2017.

[8] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[9] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[10] K. Elissa, "Title of paper if known," unpublished.

[11] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[12] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[13] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.