Advanced Production and Industrial Engineering
R.M. Singari and P.K. Kankar (Eds.)

doi:10.3233/ATDE220723

Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction

Sanjay Patidar[1], Deepak Kumar[2], Dheeraj Rukwal[3]
[1]Department of Software Engineering, Delhi Technological University,
sanjaypatidar@dtu.ac.in
[2]Department of Software Engineering, Delhi Technological University,
deepak.ky.delhi@gmail.com
[3]Department of Software Engineering, Delhi Technological University,
dheerajrukwal7224@gmail.com

Abstract. Heart disease has become a common cause of death worldwide in recent years. People's way of living changes, dietary habits, office working cultures, and other factors have all played a role in this worrisome problem around the world. The best way to stop this disease is to develop a method that will detect early symptoms and hence save more lives. With the help Machine Learning (ML) algorithms, researchers can predict the likelihood of developing cardiovascular disease in people who are at risk. It is critical to develop a precise and dependable technique to have early disease prediction by automating the task and therefore achieving efficient disease management. Several academics have described their efforts to develop the best feasible technique for predicting heart disease in previous publications. The goal of this study is to compare alternative algorithms for predicting cardiac disease. The results of important data mining techniques are presented in this work, which can be utilized to construct a highly efficient and accurate prediction model that will aid doctors in minimizing the number of people killed by heart disease. This study compares the metrics for prediction of heart disease for 6 ML algorithms which are "Logistic Regression" (LR), "Decision Tree" (DT), "Random Forest" (RF), "Support Vector Machine" (SVM), "Gaussian Naïve Bayes" (GNB) and "k-Nearest Neighbor" (kNN).

Keywords. Heart Disease, Machine Learning, Prediction, Classification, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Gaussian Naïve Bayes, k-Nearest Neighbour

1. Introduction

The human heart is one of the most important organs in the human body. It's the device that circulates oxygen-rich blood to different parts of the body. The heart works 24 into 7 to ensure that all other organs get the right amount of oxygen-rich blood, and any interference in its functions would have an adverse effect on other organs' appropriate functioning, which can be catastrophic. Heart disease or cardiovascular disease, is a dangerous medical disorder caused by the heart's failure to perform its circulation functions properly.

If a patient ignores the disease's early symptoms, which appear to be warning signs, the patient will have no time to recover and will eventually die on the spot. A heart attack is the medical term for this. It occurs because the purpose of the arteries is to give oxygen-rich blood to the heart, but plaque forms as a result of fatty and other substances, which disrupts the functioning of a normal artery and converts it into a narrowed coronary artery.

As a result, blood flow might be slowed or entirely stopped. Controllable risk factors and uncontrolled risk factors are the two types of risk variables that cause coronary artery disease. Diabetes, smoking, obesity or overweight, cholesterol, hypertension, less physical activities are all controllable risk factors. Age, sex, previous medical conditions and history are all uncontrollable risk factors. In the last decade, heart disease is the top cause for the death of people worldwide According to a WHO report, about 17.9 million people die each year as a result of cardiovascular disorders, with coronary heart disease and brain stroke accounting for 80% of these deaths [1]. A variety of laboratory tests and imaging examinations can be used to identify cardiovascular disease. However, the patient's medical and family history, risk factors, and physical examination are the most important aspects of diagnosis. We can synchronize the results and predict the existence of disease from findings and processes using statistical data. Doctors can make better decisions with the help of automation and deep learning.

2. Literature Review

S. Musfiq Ali et.al conducted research on the Cleveland database, with 10-fold cross validation and achieved a highest accuracy of 91.2% for GNB [5].

In 2021, A. Kondababu et. al.[6] used the Cleveland dataset to study comparative analysis and found the HRFLM technique, a combination of Random Forest(RM) and Linear Method(LM) to have the highest accuracy.Rohit Bharti et. al.[7] used a dataset with 13 features, preprocessed data using Isolation Forest, and found that KNeighbors classifier performed the best.

Sfruti Sarah et. al.[8] compared various models for Heart Disease Prediction and found that LR had the maximum accuracy of 85.25%.Lubana Riyaz et. al.[9] performed a survey of various ML algorithms and their performances for heart disease prediction and found that highest average prediction accuracy was achieved by ANN with 86.91% and the lowest with C4.5 decision tree with 74.0%.Xiao-Yan Gao et. al.[10] found that bagging ensemble learning algorithm with Decision Tree and Principle component analysis feature extraction method performed the best.

Abdullah et. al developed a Data Mining model to increase the accuracy for heart Disease Prediction, the model was based on RF classifier [11]. Sonam Nikhar et al.[12] used Cleveland Dataset with 303 instances and 19 attributes with GNB, DT technique. They also discovered that Decision Tree has a higher accuracy than the Nave Bayes Classifier. Ravindra Yadav et al.[13] deployed ML approach for the Cardio Vascular Disease Prediction Survey, which included the DT, GNB , Neural Networks , Deep Learning and SVM. The decision tree's conclusion is generated using ID3, CART Cpercent.0,CYT , and J48.

Devansh Shah et.al,[14] used Cleveland database with 303 instances and 14 attributes for heart disease prediction. K-NN algorithm had the highest accuracy. Archana Singh et al. [15] employed the Cleveland dataset resulting in the following results: Linear Regression: 78 percent, DT: 79 percent, SVM: 83 percent, and K-NN: 87 percent accuracy.

3. Proposed Solution

The key components include Data Collection, Data preprocessing, Data splitting and Performance Evaluation.

3.1. Data Collection

The dataset that was used in this study is available at kaggle[20] . The dataset dates to 1988 and consists a combination of four datasets which are from Long Beach V, Switzerland, Hungary and Cleveland. In total it contains 75 Attributes, excluding the predicted attribute. All the researches use a subset of 14 of them. The "target" field is the predicted attribute making a total of 76 columns. The dataset consists of 1025 patients with 713 male and 312 females. The description of attributes is given in Table 1.

Table 1. Description of the Attributes

S. No	Attribute	Description	Values
1	age	Patient's age in years	Value is continuous in range [29-77]
2	sex	Patient's sex	1 : male 0 : female
3	cp	Type of chest pain	0 : asymptomatic 1 : atypical angina 2 : non-angina pain 3 : typical angina
4	trestbps	Resting blood pressure of patient (mm Hg, noted on the time at which patient was admitted to the hospital)	Value is continuous in range [94-200]
5	chol	Patient's serum cholesterol measurement in mg/dl	Value is continuous in range [126-564]
6	fbs	Fasting blood sugar of patient	1 if fbs > 120 mg/dl Else 0
7	restecg	Resting electrocardiographic results	0 : normal 1 : having abnormal _ST_T wave 2 : left ventricular hypertrophy
8	thalach	Maximum heart rate achieved by patient	Value is continuous in range [71-202] bpm
9	exang	Exercise included angina	0 denotes no 1 denotes yes
10	oldpeak	Depression in ST brought by exercise that is relative to rest	Value is continuous in range [0-6.2]
11	slope	ST segment's peak exercise slope.	0 : down sloping 1 : flat 2 : up sloping
12	ca	Count of major vessels that have been colored with fluoroscopy	0-4 value
13	thal	thalassemia value, a type of blood disorder	0 : Null 1: represents a defect that is fixed. In this condition in some parts of the heart there is no blood flow 2 : blood flow is normal 3: defect is reversible.
14	target	Is the heart disease present	0 : No 1 : Yes

3.2. Data Pre-processing

The collected dataset contained no missing values. The dataset consists of 5 attributes with continuous numerical values which are age, trestbps, chol, thalach and oldpeak. There are also 8 categorical columns excluding the "target" column. These are exang, fbs, sex, ca, cp, restecg, thal, and slope. Out of these 8 categorical columns 3 are binary categorical values which are exang, fbs and sex. The rest are ordinal categorical variables.

One-Hot encoding was done on these variables. This means converting categorical data into numerical form. In one hot encoding a set of binary variables in a particular order represent the integer encoded variables.

Table 2. Before One Hot Encoding

Index in Dataset	Restecg
0	1
1	0
6	2

Table 3. After One Hot Encoding

Index in dataset	restecg_0	restecg_1	restecg_2
0	0	1	0
1	1	0	0
6	0	0	1

For example the "restecg" variable has three possible values 0, 1 or 2, basically three categories are present. 3 binary variables would be needed to depict the categories. A "1" value is used for that particular category and "0" for the other categories in a particular row. For example Table 2 would be converted into Table 3.

Our dataset was scaled using StandardScaler. Numerical input variables scaled to the normal (standard) ranges improve the performance of several machine learning methods. It is done by first subtraction of the mean and then division by the standard deviation to every variable. After the following operations have been performed the Standard Deviation becomes one and the mean equals to zero.

4. Results and Analysis

The comparison of performance is displayed in Table 4. Metrics that have been used are accuracy, precision, recall/sensitivity/ F1 Score and Harmonic mean. Random Forest Classifier has performed the best in Accuracy, Precision, Recall and F-1 Score. Gaussian Naïve Bayes has performed the worst in all four metrics. Data was pre-processed using one-hot encoding which makes the data more usable and expressive and it can be rescaled easily.

Fig 1 depicts the ROC-AUC curve of the various classifiers. The more the area under the curve the better it is. The value ranges from 0 to 1, more closely to 1 the better the algorithm has performed. It is visible from the graph that the red dotted line which denotes the Random Forest has the maximum area, which means RF has performed the best.

Table 4. Comparison of Metrics of Different Classifiers

S. No.	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-1 Score (%)
1	Logistic Regression	82.92	78.33	91.26	84.30
2	Decision Tree	95.12	94.28	96.11	95.19
3	Random Forest	98.53	100	97.08	98.52
4	Support Vector Machine	87.31	84.07	92.23	87.96
5	Gaussian Naïve Bayes	74.63	68.34	92.23	78.51
6	K- Nearest Neighbors	81.95	77.96	89.32	83.25

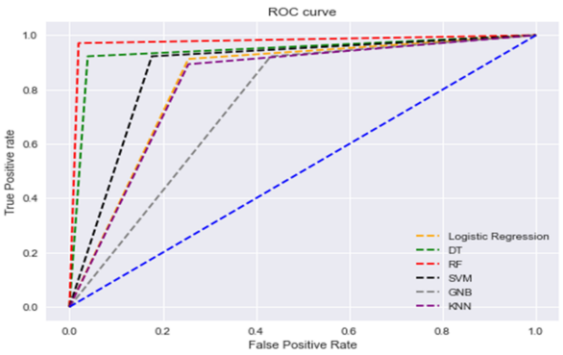

Figure 1. ROC-AUC for each Classifier

5. Conclusion

The goal of this research was to examine the performance of different supervised machine learning algorithms for predicting heart disease. which were "Logistic Regression" (LR), "Decision Tree" (DT), "Random Forest" (RF), "Support Vector Machine" (SVM), "Gaussian Naïve Bayes" (GNB) and "k-Nearest Neighbor" (kNN). Many prior researches on the same topic were analyzed. Data was preprocessed using one-hot encoding and then split into testing and training data. Models were trained and various metrics were drawn. In this study it was found out that random Forest algorithm performed the best with 98.53% accuracy and GNB the worst with 74.63% accuracy. It was mainly due to one-hot encoding that was done which helped Random Forest in making better informed decisions. Further work that can be done in this area is increasing the accuracy using hyper-parameter tuning, feature selection and ensemble methods.

6. References

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clinical epidemiology. 2011;3:67.

[2] Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random forest. InInternational Conference on Information Computing and Applications 2012 Sep 14 (pp. 246-252). Springer, Berlin, Heidelberg.

[3] Mythili T, Mukherji D, Padalia N, Naidu A. A heart disease prediction model using SVM-decision trees-logistic regression (SDL). International Journal of Computer Applications. 2013 Jan 1;68(16).

[4] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Computer Science. 2020 Nov;1(6):1-6.

[5] Ali M, Khan MD, Imran MA, Siddiki M. Heart disease prediction using machine learning algorithms (Doctoral dissertation, BRAC University).
[6] Kondababu A, Siddhartha V, Kumar BB, Penumutchi B. A comparative study on machine learning based heart disease prediction. Materials Today: Proceedings. 2021 Feb 19.
[7] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience. 2021 Jul 1;2021.
[8] Sarah S, Gourisaria MK, Khare S, Das H. Heart Disease Prediction Using Core Machine Learning Techniques—A Comparative Study. InAdvances in Data and Information Sciences 2022 (pp. 247-260). Springer, Singapore.
[9] Riyaz L, Butt MA, Zaman M, Ayob O. Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. InInternational Conference on Innovative Computing and Communications 2022 (pp. 81-94). Springer, Singapore.
[10] Gao XY, Amin Ali A, Shaban Hassan H, Anwar EM. Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. Complexity. 2021 Feb 10;2021.
[11] Abdullah AS, Rajalaxmi R. A data mining model for predicting the coronary heart disease using random forest classifier. InInternational Conference in Recent Trends in Computational Methods, Communication and Controls 2012 Apr (pp. 22-25).
[12] Lafta R, Zhang J, Tao X, Li Y, Tseng VS. An intelligent recommender system based on short-term risk prediction for heart disease patients. In2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 2015 Dec 6 (Vol. 3, pp. 102-105). IEEE.
[13] Hasan R. Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. InITM Web of Conferences 2021 (Vol. 40, p. 03007). EDP Sciences.
[14] Shah D. Heart Disease Prediction using Machine Learning Techniques Springer Nature Singapore Pte Ltd, 2020.
[15] Singh B, Prabhakar Tiwari SN, Singh RP, Vishwakarma M, Patel DK, Kumar A, Pratap A, Singh SP, Mishra S, Raj R, Lohia P. SN Paper ID. InInternational Conference on Electrical and Electronics Engineering (ICE3) 2020 Feb (Vol. 14, p. 15).
[16] Heart Disease Dataset: Heart Disease Dataset | Kaggle