Medicon Engineering Themes
Volume 7 Issue 6 December 2024
Article Type: Research Article
ISSN: 2834-7218

Predicting Heart Disease Using Machine Learning: A Comparative Analysis of Classification Models

Shreyas CS* and DN Sujatha

Department of Computer Applications, B.M.S. College of Engineering, Bangalore, India

***Corresponding Author:** Shreyas C.S., Department of Computer Applications, B.M.S. College of Engineering, Bangalore, India.

Received: October 27, 2024; **Published:** December 03, 2024

Abstract

This research paper explores the application of machine learning techniques to predict heart disease, a leading cause of deaths worldwide. By utilizing various classification algorithms, including Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbours, and Support Vector Machines, we aim to identify the most effective model for accurate prediction. The study uses the dataset from Cleveland Heart Disease dataset, with data pre-processing steps such as handling missing values, scaling, and feature selection. Model performance is compared and used with metrics like accuracy, precision, recall, F1-score. The findings indicate that Random Forest algorithm outperforms other models, which provides insights for healthcare professionals in early diagnosis and preventive measures. The results also highlight the potential of machine learning to enhance clinical decision-making and improving patient outcomes.

Keywords: Classification Algorithms; Decision Trees; Healthcare Analytics; Heart Disease Prediction; K-Nearest Neighbour; Logistic Regression; Machine Learning; Random Forest; Support Vector Machines

Introduction

Heart diseases or cardiovascular diseases remain a leading cause of diseases and deaths worldwide. Early and accurate prediction of heart disease is necessary for effective intervention and management, potentially saving lives and reducing healthcare costs since traditional diagnostic methods face limitations in terms of accessibility, cost, and timeliness.

This study aims to evaluate and compare the performance of several machine learning algorithms in predicting the occurrence of heart disease. The study aims to compare Logistic Regression, Random Forest, Naive Bayes, K-Nearest Neighbour, and Decision Trees to identify the most effective algorithm for predicting heart diseases.

In this paper, Section 2 describes the literature survey, Section 3 explains the methodology, Section 4 elaborates on the results and Section 5 concludes the work.

Literature Review

Heart disease remains a leading cause of death worldwide, prompting significant research into predictive modelling to improve diagnosis and treatment outcomes. Machine learning and data mining techniques have been extensively studied and applied to predict the likelihood of heart disease, leveraging various clinical and demographic data.

Risk Factors and Predictive Models

Heart disease has a range of conditions including coronary artery disease, arrhythmias, heart failure, etc. The major risk factors for causing heart diseases include diabetes, obesity, unhealthy diet, excessive consumption of alcohol, physical inactivity, hypertension, and high cholesterol.

Machine Learning Algorithms

The following are a brief description of various machine learning algorithms:

- *Random Forest*: This method reduces the risks of over fitting and provides robust predictions by creating an ensemble of decision trees [1-6]. It has shown effectiveness in handling highdimensional data in heart disease prediction tasks.
- *Support Vector Machine*: It is used to separate different classes. It is useful for high-dimensional data and has been used to predict heart disease by mapping input features to higher-dimensional spaces [1, 2, 4, 6].
- *Logistic Regression*: This is commonly used as in logistic regression models; the probability of a binary outcome is based on one or more predictable variables. Logistic regression has been effectively used in predicting heart disease by evaluating risk factors such as age, blood pressure, and cholesterol levels, etc. [2, 3, 6].
- *Naive Bayes*: This algorithm applies Bayes' theorem which assumes the independence between features. It is observed from the literature survey [4, 6] that it has been extensively for heart disease prediction, especially when dealing with large datasets since it has high computational efficiency.
- *Neural Networks*: Neural networks and multi-layer perceptrons in particular, have been used for their ability to model complex relationships between input features and the target variable [4-6]. It used by many researchers to predict heart disease by learning from training data through multiple hidden layers.

Feature Selection and Data Processing

Models which are effective in prediction often depend on the quality of the features selected. Techniques like Principal Component Analysis [4] has been used to reduce dimensionality thereby enhance the model's performance.

Methodology

In the proposed work, Logistic Regression, Random Forest, Naïve Bayes, K-Nearest Neighbour and Decision Trees are implement on Cleveland Data Set. A brief description of the algorithms is as follows:

Logistic Regression

Logistic regression can be utilized for heart disease predictions. Performance metrics can be analysed, and model parameters received can be altered to improve its accuracy.

Random Forest

Random Forest can be implemented with the initial default parameters and then it can be optimized by adjusting hyper parameters such as the number of trees and the depth of tree to enhance model robustness and accuracy.

Naive Bayes

Naive Bayes algorithm can be applied to handle large datasets efficiently by experimenting with different data pre-processing techniques to optimize model performance.

K-Nearest Neighbour (KNN)

KNN can be used to test with n neighbour values to identify the optimal number for best performance.

Results acquired can be used to refine the model predictions.

Decision Trees

Decision Trees are explored both with and without pruning techniques to understand their impact on model accuracy and to prevent over fitting.

The results of the work are depicted using visualization techniques like bar charts and confusion matrix. This demonstrates the performance of the model and false negative rates, thereby enabling a clearer assessment.

Dataset Description

The Table 1 shows a sample data set. The Data set contains a number of important health metrics and indicators that help in predicting the likelihood of heart disease. This includes age, sex: male or female, and chest pain type: from typical angina to asymptomatic. It also keeps a record of resting blood pressure, cholesterol levels, and whether a person has high fasting blood sugar. Added to this are the results of resting Electrocardiogram (ECG), maximum heart rate achieved, and whether exercise-induced angina is also included. It measures ST depression induced by exercise, the slope of the peak exercise ST segment, the number of major vessels coloured by fluoroscopy, and the presence of thalassemia. The target variable indicates if an instance has heart disease or otherwise.

	age	*sex*	*cp*	*trestbps*	*chol*	*fbs*	*restecg*	*thalach*	*exang*	*oldpeak*	*slope*	*Ca*	*thal*	*target*
254	59	1	3	160	273	0	0	125	0	0.0	2	0	2	0
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
172	58	1	1	120	284	0	0	160	0	1.8	1	0	2	0
74	43	0	2	122	213	0	1	165	0	0.2	1	0	2	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1

Table 1: Sample Data from dataset.

Continuous features such as resting blood pressure, cholesterol level, maximum heart rate, and ST depression are explored by showing their distribution and scatter plots, outlining the variability and effect on heart disease.

Correlation Analysis

Correlation analysis provides an assessment of the strength of the relationships between variables that are measured numerically. For this analysis, some key numeric variables were chosen: age, resting blood pressure, cholesterol, maximum heart rate, ST depression, and the number of major vessels.

A correlation matrix has been generated and visualized through the use of a heatmap by using the Clevland data Set. Figure 1 shows the extent to which each pair of variables is related, outlining patterns and dependencies. The heatmap depicts the view for the relationship among these continuous variables.

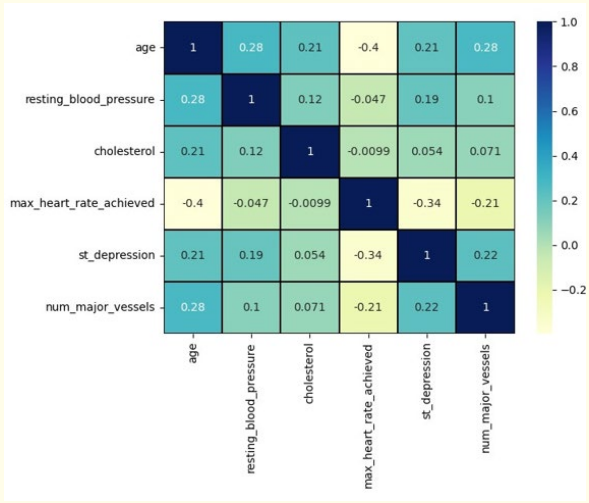

Figure 1: Correlation Analysis.

Dataset Splitting and Model Training

To test the various algorithms of classification, the dataset is split into a training subset and a testing subset. The training data is then further broken down into features, which is X_train, and its corresponding target values, which is Y_train.

For testing data, it is also divided accordingly into features, X_test, and target values, Y_test.

The shapes of these datasets are:

- Training features: 242 records, 13 features.
- Testing features: 61 records, 13 features.
- Training target values: 242 records.
- Testing target values: 61 records.

Modelling and Predicting with Machine Learning

Here, various classification algorithms are tested and their performance metrics are evaluated in terms of accuracy. In this paper, a general function for model training and evaluation is applied.

Logistic Regression Analysis

Logistic Regression is one of the algorithms tested for predicting heart disease. The model is trained on the training dataset and evaluated on the test dataset.

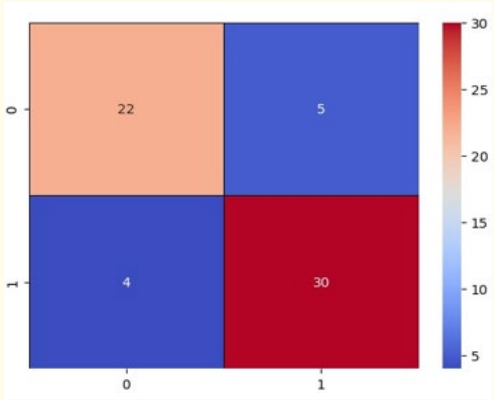

Figure 2: Confusion Matrix for Logistic Regression.

Figure 2 shows the confusion matrix. The performance of model and evaluation metrics are as follows:

Model Performance:

- Accuracy: 85.25% on the test set.
- Train Accuracy: 84.71%.
- Test Accuracy: 85.25%.

Evaluation Metrics:

- Precision: 85.71%.
- Recall: 88.24%.
- F-Score: 86.96%.

Random Forest Analysis

The Random Forest model, using 100 trees, aims to predict heart disease with high accuracy.

Model Performance:

- Accuracy: 88.52% on the test set.
- Training Accuracy: 100%.
- Test Accuracy: 88.5%.

Impact of Pruning:

Pruning the trees to a maximum depth of 3 results in:

- Training Accuracy: 87.6%.
- Test Accuracy: 86.9%.

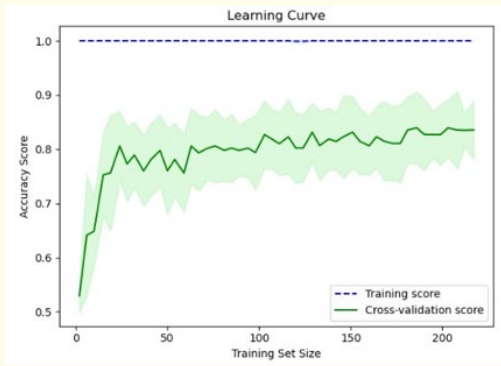

Figure 3: Learning Curve for Random Forest accuracy.

Figure 3 shows the accuracy of Learning Curve for random forest: The learning curve will show how model accuracy changes with the size of the training set. This demonstrates that the model fits appropriately.

Evaluation Metrics:

- Precision: 86.11%.
- Recall: 91.18%.
- F-Score: 88.57%.

Naive Bayes Analysis

The Naive Bayes model, specifically the Gaussian Naive Bayes variant, is evaluated for predicting heart disease.

Heart diseases can be predicted using the Naive Bayes model, specially using the Gaussian Naive Bayes variant.

The results of Model Performance are as follows:

- Train Accuracy: 83.47%.
- Test Accuracy: 85.25%.

The Evaluation Metrics for the same are:

- Accuracy Score: 85.25%.
- Precision: 83.78%.
- Recall: 91.18%.
- F-Score: 87.32%.

The performance of the Naive Bayes model was quite good, achieving an accuracy of 85.25%, an extremely high recall rate, and a balanced F-score. The confusion matrix [Figure 4] highlights the distribution of predictions.

The False Negative Rate obtained is 8.82%.

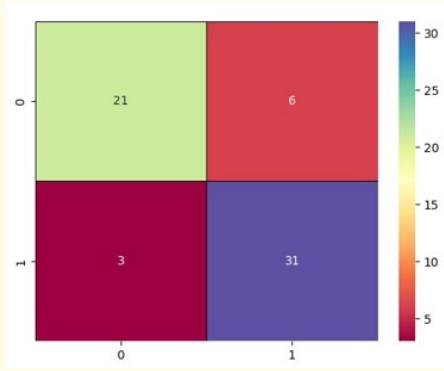

Figure 4: Confusion Matrix for Naive Bayes.

K-Nearest Neighbor (KNN) Analysis

Another simple instance-based learning approach for classification and regression is the K-Nearest Neighbor algorithm.

The Performance Metrics are as follows:

- Train Accuracy: 71.90%.
- Test Accuracy: 68.85%.

Optimal Parameter Search

To find the best n neighbors parameter, multiple values were tested, and their corresponding accuracies are recorded in Table 2.

n_neighbors	*Train Accuracy*	*Test Accuracy*
1	100.00%	52.46%
2	79.75%	59.02%
3	78.10%	63.93%
4	76.03%	63.93%
5	78.10%	63.93%
6	74.38%	65.57%
7	72.31%	67.21%
8	71.90%	68.85%
9	73.14%	67.21%

Table 2: Accuracy with different value of *n*.

The optimal value of n_neighbors was determined to be 8, balancing between overfitting (high training accuracy) and underfitting (low test accuracy).

False Negative Rate (FNR): 32.35%: FNR is the percentage of real positive cases—heart disease—which was misclassified as no heart disease.

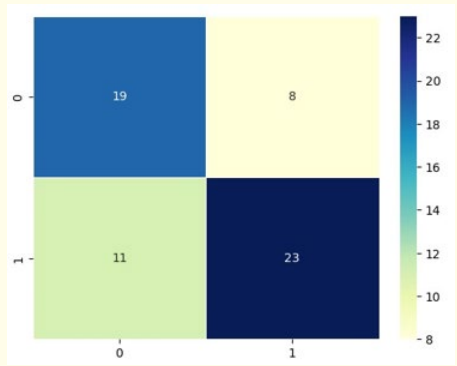

Figure 5: Confusion Matrix for K Nearest Neighbour.

The following are the observations:

- *Bias–Variance Tradeoff*: The train accuracy of 71.90% and test accuracy of 68.85% suggest that the model generalizes fairly with no severe overfitting or underfitting.
- *Performance*: The model has good precision but only fair recall, meaning that it is classifying some of the heart diseases as other types.
- *Parameter Tuning*: The parameters indicate the importance of choosing the right value of n_neighbors to optimize the model's performance.

Decision Tree Analysis

One of the most flexible and interpretable models that can be fitted for both classification and regression problems is the Decision Tree algorithm.

Here, a classifier with a maximum depth of 3 is trained and compared against the heart disease dataset. The performance will also be observed for both training and test data to check for over fitting and generalization.

Over fitting and Pre-Pruning

The first Decision Tree model, without any limit on the depth, perfectly classified the examples in the training set but did worse on those in the test set, which is indicative of over fitting:

- Train Accuracy without max_depth: 100.00%.
- Test Accuracy without max_depth: 78.70%.

This avoided overfitting, which limited the maximum depth of the tree to 3 and improved performance on the test set:

- Train Accuracy (max_depth=3): 84.30%.
- Test Accuracy (max_depth=3): 82.00%.

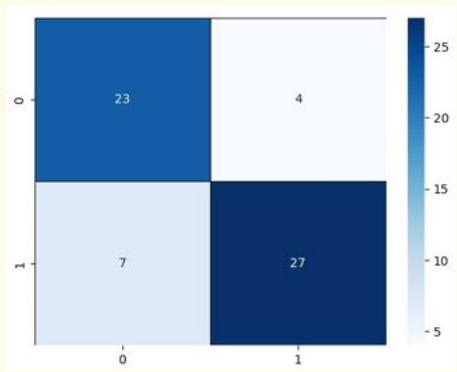

Figure 6: Confusion Matrix for Decision Tree.

The confusion matrix in Figure 6 shows how well a model is performing based on the counts of true positives, true negatives, false positives, and false negatives predictions. The details are as follows:

- True Negatives (TN): 23.
- False Positives (FP): 4.
- False Negatives (FN): 7
- True Positives (TP): 27.

False Negative Rate: 20.59% - This measures the proportion of actual positive cases of heart disease misclassified as no heart disease.

Results

For all of the machine learning models, accuracy based on testing sets and the F1 score were calculated.

This helps in understanding how each algorithm performs in classifying heart disease correctly.

Accuracy Scores

The accuracy scores for every algorithm were calculated and structured into a summary DataFrame. The accuracy is the proportion of correctly classified cases to the total number of instances as represented in Figure 7.

Table 3 presents the accuracy scores of all the algorithms:

F1 Scores

The F1 score is the harmonic mean of precision and recall, therefore being a balance between both metrics. It's especially useful when working with models on datasets where there is class imbalance.

Table 4 has F1 scores for each model.

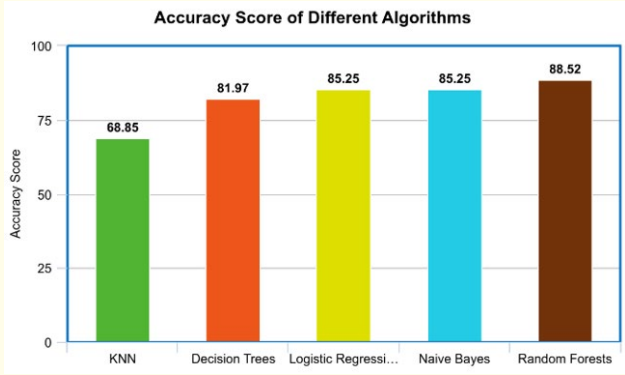

Figure 7: Accuracy Scores for the 5 compared algorithms.

Algorithm	***Accuracy***
KNN	68.85%
Decision Trees	81.97%
Logistic Regression	85.25%
Naive Bayes	85.25%
Random Forests	88.52%

Table 3: Summary of Accuracy Scores.

Algorithm	***F1 Score***
Logistic Regression	0.869
Random Forest	0.885
Naive Bayes	0.873
K-Nearest Neighbors	0.707
Decision Tree	0.830

Table 4: F1 Scores.

Conclusion

Several insightful conclusions can be drawn from the evaluation of various machine learning models for heart disease prediction. Heart diseases are highly rated as one of the top death causes in the world. An early prediction system is quite important in ensuring timely intervention and treatment, and hence a better outcome for the patients.

Therefore, selecting a reliable model for heart disease prediction will be a great help to health professionals.

Implications for Heart Disease Prediction

It will be possible to enable early and more effective interventions to save lives and reduce healthcare costs if heart disease can be accurately predicted. This research effort looks into models that offer variety in terms of options both for health professionals and researchers.

Random Forest: This has very high accuracy and is balanced in performance, making it very proper in a clinical setting where accuracy is paramount. It is capable of handling large datasets with a considerable number of features, making it very suitable for most complex medical data.

Naive Bayes and Logistic Regression: These methods provide more robust alternatives, particularly when computational simplicity and interpretability really matter. They can be implemented very easily and will yield robust predictions.

Decision Tree: A bit less accurate than the earlier models, this model has an intuitive structure that is very useful to explain the decision-making process to health professionals and patients.

K Nearest Neighbour: Less effective compared to other methods on this data set, but still useful in particular contexts where the data has very local patterns or in ensemble methods to improve overall performance.

Hence, Random Forest is the best model to predict heart disease using this dataset because it provides the optimal trade-off between accuracy and balanced precision and recall. More good options are Naive Bayes and Logistic Regression. Not very good ones are KNN and Decision Trees, but of course in certain cases they could be useful according to the context and requirements.

Early and accurate heart disease prediction can make a huge difference in patient outcomes; such models are, therefore, very important tools toward assisting in this aim.

Acknowledgements

We would like to express our gratitude to everyone who supported us during the development of this paper. We have no conflicts of interest to declare, and no specific funding or grants were received for this work.

References

1. Malik Zaibunnisa LH., et al. "Heart Disease Prediction Using Artificial Intelligence". International Journal of Engineering Research & Technology (IJERT) (2021).
2. Chang Victor., et al. "An Artificial Intelligence Model for Heart Disease Detection Using Machine Learning Algorithms". ScienceDirect (2022).
3. Jindal Harshit., et al. "Heart Disease Prediction Using Machine Learning Algorithms". ICCRDA (2020).
4. Boukhatem Chaimaa, Heba Yahia Youssef and Ali Bou Nassif. "Heart Disease Prediction Using Machine Learning". Advances in Science and Engineering Technology International Conferences (ASET) (2022).
5. Bhatt Chintan M., et al. "Effective Heart Disease Prediction Using Machine Learning Techniques". MPDI (2023).
6. Bagadi Kalapraveen., et al. "Cardiovascular Disease Prediction Using Machine Learning Algorithms". IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) (2023).
7. "Heart Disease Cleveland Dataset". Kaggle.

Volume 7 Issue 6 December 2024