Paper Title:

Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques

Paper Link:

https://www.hindawi.com/journals/bmri/2023/6864343/

1 Summary ;

Almost 17.9 million people are losing their lives due to cardiovascular disease, which is 32% of total death throughout the world. It is a global concern nowadays. However, it is a matter of joy that the mortality rate due to heart disease can be reduced by early treatment, for which early-stage detection is a crucial issue. This study is aimed at building a potential machine learning model to predict heart disease in early stage employing several feature selection techniques to identify significant features. Three different approaches were applied for feature selection such as chi-square, ANOVA, and mutual information, and the selected feature subsets were denoted as SF1, SF2, and SF3, respectively. Then, six different machine learning models such as logistic regression (C1), support vector machine (C2), K-nearest neighbor (C3), random forest (C4), Naive Bayes (C5), and decision tree (C6) were applied to find the most optimistic model along with the best-fit feature subset. Finally, we found that random forest provided the most optimistic performance for SF3 feature subsets with 94.51% accuracy, 94.87% sensitivity, 94.23% specificity, 94.95 area under ROC curve (AURC), and 0.31 log loss. The performance of the applied model along with selected features indicates that the proposed model is highly potential for clinical use to predict heart disease in the early stages with low cost and less time.

1.1 Motivation

The sequential steps involved in predicting heart disease using machine learning.This can be a flowchart or diagram showing the process from data collection to algorithm application and results interpretation.

1.2 Contribution

Applied feature selection algorithms to determine feature importance.Created three different selected feature (SF) sets. The dataset into training and testing sets used 70% of the data as a training set and the remaining as a test set. Trained six different machine learning algorithms using the 70% test data.

Evaluated algorithm performance to select the one with the highest accuracy.The algorithms used for training (LR, SVM,KNN, RF, NB, DT).

1.3 Methodology

Python 3.8 was chosen for its accessibility and ease of rapid algorithm testing. UCI Cleveland Dataset utilized for predicting heart disease.303 patient records with 13 features each.Binary classification: heart patients or normal cases. For Data Pre-processing post-collection, identified and corrected 4 incorrect records on NMV and 2 on TS.Replacement of incorrect values with optimal values.Applied StandardScaler to standardize features (mean 0, variance 1). ANOVA F value used in test to measure similarity and reduce high-dimensional data. Here , these Machine Learning Algorithms Logistic Regression (LR), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT). Accuracy is assessed using confusion matrices which are N × N matrices. Sensitivity measures the proportion of true positive cases relative to all actual positive cases. Specificity calculates the proportion of true negative cases relative to all actual negative cases.Area Under the Receiver Operating Characteristic (AUROC) assesses the performance of a classification model using True Positive. Log Loss is a classification loss function used to evaluate the performance of machine learning algorithms.

1.4 Conclusion

In this study, the main focus was on enhancing the predictive capabilities of heart disease detection through the implementation of various feature selection techniques and the application of six distinct machine learning algorithms. The research involved the identification of highly significant features crucial for accurate prediction. Among the algorithms, SVM and LR emerged as particularly best performers. However, it's important to note that the dataset used for this analysis was relatively limited in size impacting the potential of the predictive model. To achieve a more robust and reliable outcome future experiments are planned intending to leverage larger real-world patient datasets. Furthermore, the exploration of more advanced algorithms, including deep learning is on the agenda along with an emphasis on refining feature selection techniques. The ultimate goal is to continually enhance the accuracy and efficiency of algorithms in the critical domain of heart disease prediction.

2 Limitations

2.1 First Limitation:

One limitation of the study is the reliance on the quality and correctness of the dataset. The authors mentioned that they replaced incorrect records with optimal values. However, the process of identifying and replacing incorrect values may introduce errors or biases into the dataset. If the initial data collection had inaccuracies or if the replacement process introduced new inaccuracies, it could impact the reliability of the results. Additionally, the authors did not provide details on how they identified and determined the optimal values for the incorrect records, leaving room for ambiguity in the data preprocessing step.

2.2 Second Limitation ;

The study employed three different feature selection algorithms (ANOVA, Chi-Square, Mutual Information) to identify important features for predicting heart disease. While feature selection is crucial for improving model performance, the choice of algorithms introduces a potential limitation. Different algorithms may yield different results, and the authors did not conduct a comprehensive analysis or comparison of various feature selection techniques. The lack of a thorough exploration of the impact of different feature selection methods on model performance limits the generalizability of the findings.

3 Synthesis:

Implementing more rigorous data quality assurance measures during the initial data collection phase and providing a detailed description of the process can enhance the reliability of the dataset. Additionally, conducting sensitivity analyses to evaluate the impact of potential errors in the dataset on the results would be beneficial.Future studies should explore a broader range of feature selection algorithms and provide a comparative analysis of their performance. This could involve using additional algorithms and considering ensemble methods to combine the strengths of multiple techniques. A systematic evaluation of the stability and consistency of selected features across different algorithms would contribute to a more robust feature selection process.