Detection of heart disease with machine learning methods

İsmail Dereevli
Department of Computer Enginnering
Adana, Turkey
idereevli@gmail.com

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes. Early diagnosis and treatment play a critical role in reducing the impact of heart diseases.In this study, we aim to obtain an ML model that can predict heart disease with the highest possible performance using the heart disease dataset. Statlog heart disease dataset from the open-access UCI database was used in this study. KNN, Decision Trees, Naive Bayes, Support Vector Machines, Random Forest, and Logistic Regression machine learning algorithms were applied to the data sets. Standardization and normalization were preferred as data preprocessing steps. Accuracy values of the highest success rates in the tests were obtained using the Support Vector Machine with 87% for the dataset. Following this, Logistic Regression achieved an accuracy of 87%, followed by the kNN algorithm with 87% accuracy, the Naive Bayes algorithm with 85% accuracy, the Random Forest algorithm with 79% accuracy, and the Decision Tree algorithm with 62% accuracy, respectively.

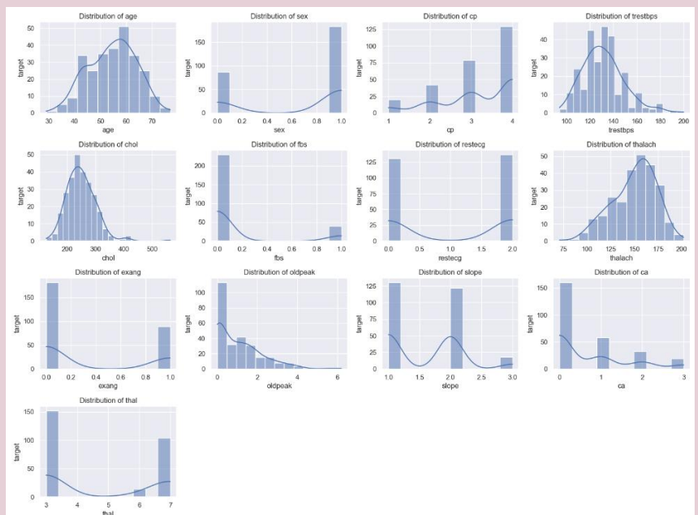

Figure 1. Distribution of observation attributes in the Statlog dataset

Methods

Dataset Description: The Statlog heart disease dataset from the UCI database consists of a collection of attributes representing various clinical factors and a target variable indicating the presence or absence of heart disease. It comprises instances of patients with features such as age, sex, cholesterol levels, blood pressure, and electrocardiogram measurements.

Data Preprocessing: Before training the machine learning models, the dataset underwent preprocessing steps to ensure uniformity and enhance model performance. Standardization and normalization techniques were applied to scale the numerical features to a common range and mitigate the influence of outliers.

Machine Learning Algorithms: Six machine learning algorithms were employed for heart disease prediction:

K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies instances based on the majority class of their k nearest neighbors.

Decision Trees: A tree-based model that partitions the feature space into hierarchical structures to make predictions based on simple decision rules.

Naive Bayes: A probabilistic classifier that applies Bayes' theorem with strong independence assumptions between features.

Support Vector Machines (SVM): A supervised learning model that constructs hyperplanes in a high-dimensional space to separate instances into different classes. Random Forest: An ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and robustness.

Logistic Regression: A linear model that estimates the probability of binary outcomes using a logistic function.

Model Training and Evaluation: Each machine learning algorithm was trained on a portion of the dataset and accuracy was chosen as the primary evaluation metric.

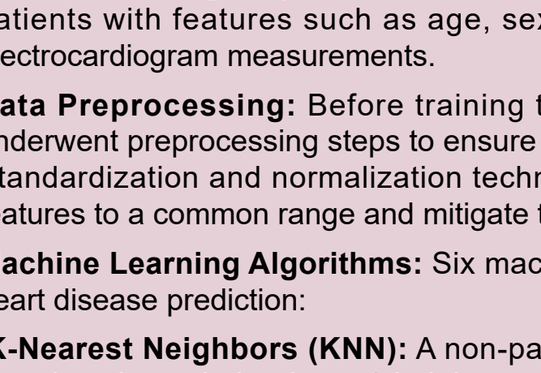

Figure 2. Distribution of Healthy and Heart Disease Individuals

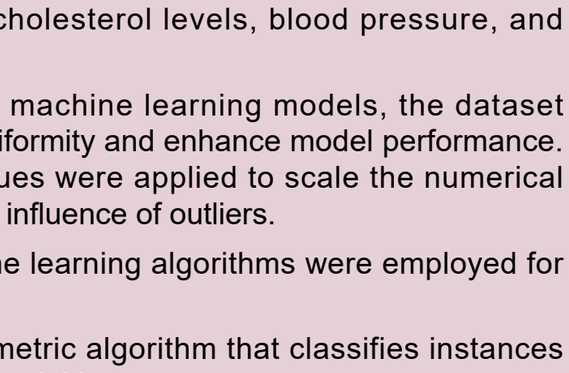

Figure 3. Chest Pain Type and Disease

Conclusion

This study aimed to develop machine learning models for predicting heart disease using the Statlog heart disease dataset. Through the application of various algorithms including KNN, Decision Trees, Naive Bayes, Support Vector Machines, Random Forest, and Logistic Regression, we sought to identify the most effective approach for early detection of cardiovascular conditions.

Our results demonstrate that Support Vector Machines, Logistic Regression, and KNN achieved the highest accuracy rates, all around 87%. This indicates the potential of machine learning in aiding healthcare professionals with early diagnosis and intervention, which is crucial for reducing the burden of cardiovascular diseases worldwide.

However, it's important to acknowledge the limitations of our study, including the reliance on a single dataset and the need for further validation on diverse datasets. Additionally, while our models performed well in terms of accuracy, other metrics such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) should be considered for a more comprehensive evaluation.

Future research should focus on refining these models, incorporating additional clinical features, and exploring advanced machine-learning techniques to improve predictive accuracy and generalization to diverse populations. Ultimately, the development of accurate and reliable predictive models can significantly contribute to early detection and management of heart diseases, thus reducing mortality rates and improving patient outcomes.

Model	Model Specificity Score	Model Sensitivity Score	Model Accuracy Score	Model AUC Score
SVM without Scaling	93.103448	80.000000	87.040000	91.448276
KNN Max k with Standardization	96.551724	56.000000	87.037037	87.448276
LogisticRegression with Standardization	89.655172	80.000000	87.037037	92.137931
Naive Bayes with Standardization	89.655172	80.000000	85.190000	87.448276
Naive Bayes without Scaling	89.655172	80.000000	85.190000	87.448276
SVM with Standardization	93.103448	76.000000	85.190000	91.034483
Naive Bayes with Normalization	89.655172	80.000000	85.190000	87.448276
KNN Max k with Normalization	96.551724	56.000000	85.185185	79.517241
LogisticRegression without Scaling	89.655172	80.000000	85.185185	91.724138
LogisticRegression with Normalization	89.655172	80.000000	85.185185	92.689655
SVM with Normalization	89.655172	76.000000	83.330000	91.862069
Random Forest with Normalization	86.206897	72.000000	79.629630	91.172414
Random Forest without Scaling	86.206897	72.000000	79.629630	91.172414
Random Forest with Standardization	86.206897	72.000000	79.629630	91.172414
KNN Max k without Scaling	96.551724	56.000000	70.370370	60.482759
KNN without Scaling	79.310345	48.000000	64.810000	69.655172
Decision Tree with Standardization	58.620690	64.000000	61.111111	61.310345
Decision Tree without Scaling	62.068966	60.000000	61.111111	61.034483
Decision Tree with Normalization	55.172414	64.000000	59.259259	59.586207

Figure 4. Performance comparison

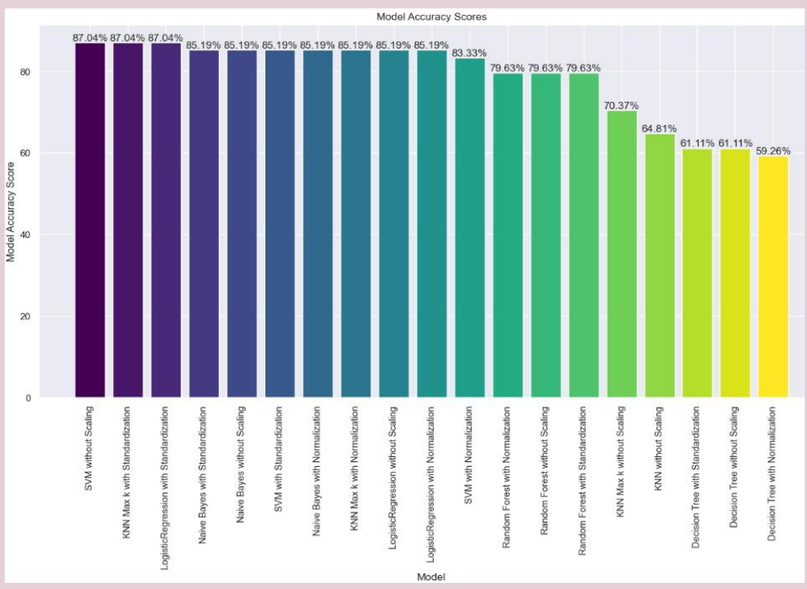

Figure 5. Model Accuracy

References

World Health Organization . World Health Statistics 2021. World Health Organization; Geneva, Switzerland: 2021.

Pisner DA, Schnyer DM. Support vector machine. Machine Learning: Elsevier; 2020. p.101-21. 19.

Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. wiley interdisciplinary reviews: data mining and knowledge discovery. 2019;9:1301. 20. Iwendi C, Bashir AK,

Peshkar A, et al. COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in Public Health. 2020;8:357

UCI.(2019).Statlog(Heart)DataSet.https://archive.ics.uci.edu/ml/datasets/statlog+(heart)Retrieved(EriĢimfromTarihi:17.04.2024)