Prediction of Heart Disease Using Classification Algorithms

Authors: Masethe, H. D., & Masethe, M. A. (2014, October)

Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science (Vol. 2, pp. 22-24).

Introduction

Heart disease continues to be one of the most prevalent causes of mortality globally, emphasizing the need for accurate and timely prediction methods. Leveraging classification algorithms in machine learning has become a promising approach to improving diagnostic efficiency and early detection.

This project builds on the study presented in the *Proceedings of the World Congress on Engineering and Computer Science,* which explored the application of classification algorithms to predict heart disease. By analyzing key health parameters such as cholesterol levels, blood pressure, and lifestyle factors, this project aims to develop a predictive model that supports healthcare professionals in identifying high-risk patients effectively.

Methodology

1. Data Collection and Preprocessing:

- **Data Sources:**
  Heart disease datasets obtained from repositories such as the UCI Machine Learning Repository.

- **Data Cleaning:**
  Missing values were handled using imputation techniques, and irrelevant features were eliminated to enhance model performance.

- **Feature Selection:**
  Key predictors such as age, cholesterol, blood pressure, heart rate, and smoking habits were selected using Recursive Feature Elimination (RFE).

2. Classification Algorithms Implemented:

1. **Logistic Regression:**
   Utilized for its simplicity and effectiveness in binary classification tasks.

2. **Random Forest:**
   A powerful ensemble method capable of handling non-linear relationships.

3. **Support Vector Machines (SVM):**
   Applied for its robust performance in high-dimensional spaces.

4. **k-Nearest Neighbors (k-NN):**
   A straightforward algorithm used for comparative analysis.

3. Model Evaluation:

- **Evaluation Metrics:**
  Accuracy, precision, recall, F1 score, and ROC-AUC were calculated to measure model performance.

- **Validation:**
  The models were trained using a stratified 10-fold cross-validation technique to ensure reliability and reduce bias.

Conclusion

The results of the study demonstrated that Random Forest and SVM outperformed other algorithms in terms of prediction accuracy and robustness, achieving high scores across all evaluation metrics. Logistic Regression proved to be a strong baseline model, while k-NN was effective for smaller, well-distributed datasets.

This project highlights the significance of machine learning in healthcare and its potential to revolutionize heart disease prediction. Future work involves incorporating deep learning models and real-world clinical data for further advancements.