Heart Disease Prediction System Using Machine Learning Algorithms

Govind Raut
Department of Technology
Savitribai Phule Pune University
Email: govindraut1106@gmail.com

***Abstract*—Heart disease remains a leading cause of death worldwide, making an early and accurate diagnosis crucial. This study presents a machine learning-based prediction system that uses health parameters to identify potential heart disease cases. The system applies logistic regression, support vector machines (SVM), and random forest classifiers to the UCI Heart Disease dataset. The comparative evaluation shows that the random forest model outperforms others in accuracy.**

***Index Terms*—Heart Disease, Machine Learning, Logistic Regression, Support Vector Machine, Random Forest, Prediction System**

I. Introduction

Heart disease is a global health issue that affects millions of people each year. Traditional diagnosis involves costly and time-consuming medical tests. Machine learning (ML) can provide a data-driven solution to predict heart disease with high accuracy and reduced cost. This paper develops and evaluates a predictive model using standard health features such as age, cholesterol, blood pressure, and more.

II. Dataset and Preprocessing

The data set used is the UCI Heart Disease Dataset, which contains 303 samples and 14 attributes. Data preprocessing involved:

- Importing required libraries using Pandas, NumPy, and visualization tools such as Matplotlib and Seaborn.
- Check for missing values and ensure data consistency.
- Feature selection and target identification.
- Normalization using `StandardScaler` for better model performance.

III. Methodology

Three classification algorithms were used:

A. Logistic Regression

A simple and interpretable model suitable for binary classification tasks. It works by estimating probabilities using a logistic function.

B. Support Vector Machine (SVM)

SVM finds an optimal hyperplane that best separates the data into classes. Kernel functions were used to handle non-linear boundaries.

C. Random Forest Classifier

An ensemble method based on decision trees. It aggregates multiple decision trees and outputs the majority class, improving accuracy and reducing overfitting.

IV. Model Training and Evaluation

The data set was divided into 80% training sets and 20% testing sets. The accuracy, confusion matrix, and classification report were used as evaluation metrics.

- **Logistic Regression:** Accuracy $\approx$ 85%
- **SVM:** Accuracy $\approx$ 88%
- **Random Forest:** Accuracy $\approx$ 91%

V. Results and Discussion

The trained models were compared using their respective accuracies. The results show that Random Forest outperforms the others in terms of predictive performance.

Random Forest's superior accuracy is attributed to its ensemble learning capability, which reduces overfitting and increases generalization.

VI. Conclusion

This project demonstrates that machine learning models, particularly Random Forest, can effectively predict heart disease. Future work could involve deep learning models and larger datasets to improve robustness and generalization.

Acknowledgment

The author thanks the Department of Data Science, Savitribai Phule Pune University, for their guidance and support throughout the project.

References

[1] UCI Machine Learning Repository: Heart Disease Dataset. https://archive.ics.uci.edu/ml/datasets/heart+Disease
[2] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2011.
[3] A. Geron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow," O'Reilly, 2019.