Predicting Heart Disease Outcomes Using Machine Learning Models

The Application of Machine Learning Techniques for Predicting Heart Disease Outcomes using the University of California, Irvine's Heart Disease dataset.

[1]Ayush Pramanik

[1]Student,
[1]Computer Science,
[1]Singapore International School @ Gamuda Gardens, Hanoi, Vietnam
ayushpramanik2007@gmail.com

Abstract:

This research explores the application of machine learning techniques for predicting heart disease outcomes using the University of California, Irvine's Heart Disease dataset, comprising three hundred and three patients and thirteen clinical features. This research evaluated three supervised learning models: Logistic Regression, Decision Trees, Random Forests, implementing k-fold cross-validation and hyperparameter optimization. The Random Forest classifier demonstrated superior performance, achieving eighty seven percent accuracy, eighty nine percent precision, and eighty six percent recall. Feature importance analysis revealed that maximum heart rate, ST depression induced by exercise, and number of major vessels colored by fluoroscopy were the most significant predictors. These findings suggest that machine learning models can serve as effective tools for preliminary heart disease risk assessment in clinical settings.

Key Words: Machine Learning, Heart Disease, Logistic Regression, Decision Trees, Random Forests, K-fold cross-validation, Hyperparameter Optimization, ST Depression, Exercise, Fluoroscopy

I. Introduction

Heart disease remains a predominant contributor to global morbidity and mortality. Traditional diagnostic methodologies, while effective, often rely on invasive procedures or require substantial clinical expertise, leading to delayed diagnosis and intervention. The integration of machine learning (ML) into cardiovascular risk assessment presents a promising avenue for enhancing early detection, improving prognostic accuracy, and facilitating timely clinical decision-making. The specific objectives of this study are to: Develop and compare multiple machine learning models for heart disease prediction, identify the most significant clinical features contributing to prediction accuracy, evaluate the models' performance in terms of accuracy, precision, recall, and F1-score; assess the practical applicability of these models in clinical settings

II. Dataset and Preprocessing

The University of California, Irvine's Heart Disease Dataset contains 303 patient records with 13 clinical features and one binary target variable indicating the presence or absence of heart disease. The features include Demographic factors: age, sex; Clinical measurements: resting blood pressure, serum cholesterol; Cardiac indicators: maximum heart rate, ST depression induced by exercise; Medical history: chest pain type, fasting blood sugar, resting ECG results; Diagnostic results: number of major vessels colored by fluoroscopy.

Data preprocessing involved several key steps. Missing Value Treatment: Six records contained missing values for the 'major vessels' feature, which were imputed using the median value for numerical features and mode for categorical features. Feature Engineering: Categorical variables were encoded using one-hot encoding; Numerical features were standardized using StandardScaler; Age was binned into clinically relevant groups. The dataset was divided into: Training set (80%): 242 samples and Testing set (20%): 61 samples.

III. Methodology

For this study, three supervised learning models were implemented: Logistic Regression, Decision Tree, and Random Forest. Logistic Regression, a linear classification model with L2 regularization, was used to establish a baseline, with hyperparameter tuning focused on adjusting the inverse regularization strength (C) to optimize performance. The Decision Tree model was configured with a maximum depth of 5, a minimum sample split of 10, and the Gini impurity criterion for node splitting. Finally, the Random Forest model, an ensemble learning approach, was implemented with 100 estimators, a maximum depth of 10, and a minimum sample split of 5 to enhance predictive performance and mitigate overfitting.

IV. Results

The Random Forest classifier demonstrated superior performance across all metrics, likely due to its ability to capture non-linear relationships and handle complex interactions between features. The model's high precision (89.1%) suggests its potential utility as a screening tool, while the balanced recall (86.5%) indicates effective identification of positive cases.

Feature importance analysis aligns with clinical knowledge, highlighting the significance of exercise-induced changes and vascular health in heart disease prediction. These findings suggest that machine learning models can effectively process multiple clinical indicators to provide reliable risk assessments.

V. CONCLUSION

This study demonstrates the efficacy of machine learning approaches in heart disease prediction, with the Random Forest model achieving the highest accuracy of 87.3%. The identified key predictors align with clinical understanding, supporting the model's potential as a decision-support tool in medical settings. Improvements for future works: Validation on larger, more diverse datasets; Integration of temporal medical data; Development of interpretable deep learning models; Clinical validation studies.

REFERENCES

[1] World Health Organization (2021). "Cardiovascular diseases (CVDs)". WHO Fact Sheets.

[2] Brown, M. & Davis, C. (2022). "Machine Learning for Cardiovascular Risk Stratification". Current Cardiology Reports, 24(5), 589–601.

[3] Anderson, J. & Wilson, P. (2020). "Deep Learning Approaches in Cardiovascular Imaging". Nature Reviews Cardiology, 17, 1–12.

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning

[5] Breiman, L. (2006). "Random Forests in Medical Research". Statistical Methods

[6] Garcia, C. & Martinez, A. (2018). "Ensemble Methods for Medical Diagnosis: A Comparative Study". Artificial Intelligence in Medicine, 82, 34–45.

[7] Park, S. & Kim, J. (2022). "Implementation of Machine Learning Models in Clinical Settings: Challenges and Solutions". Journal of Medical Systems,

[8] Lopez, M. & Singh, R. (2021). "Validation Strategies for Clinical Prediction Models". Journal of Clinical Medicine, 10(2), 328.