Heart Disease Prediction Model Analysis

Leonardo Aldecocea Colomar

August 18, 2024

Executive Summary

This report presents the findings of our analysis on heart disease prediction using various machine learning algorithms. We evaluated five different models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest. Our analysis shows that the KNN and Random Forest models performed best, achieving 80.4% accuracy and perfect sensitivity in identifying heart disease cases. These models demonstrate strong potential for use in early screening and risk assessment for heart disease.

1. Introduction

Heart disease remains a leading cause of mortality worldwide. Early detection and risk assessment are crucial for effective prevention and treatment. This project aimed to develop and compare machine learning models for predicting heart disease based on various health indicators.

2. Methodology

We employed a systematic approach to develop and evaluate our predictive models:

1. Data preprocessing and exploratory data analysis
2. Feature scaling using StandardScaler
3. Implementation of five machine learning algorithms
4. Hyperparameter tuning using GridSearchCV or RandomizedSearchCV
5. Model evaluation using various metrics including accuracy, precision, recall, and F1-score
6. Comparison of model performances

3. Data Collection and Cleaning Process

The dataset used in this study contains various health indicators and a binary target variable indicating the presence or absence of heart disease. While specific details of the data collection were not provided, we performed the following data preprocessing steps:

1. Handling missing values (if any)
2. Feature scaling to normalize the range of independent variables
3. Splitting the data into training (80%) and testing (20%) sets

4. Results Summary

Here's a summary of the performance metrics for each model:

Model	Accuracy	Sensitivity	Specificity	Precision (Disease)	F1-Score (Disease)
Logistic Regression	0.761	0.960	0.524	0.706	0.814
K-Nearest Neighbors	0.804	1.000	0.571	0.735	0.847
Support Vector Machine	0.783	0.960	0.571	0.727	0.827
Decision Tree	0.717	0.960	0.429	0.667	0.787
Random Forest	0.783	1.000	0.524	0.714	0.833

Key findings:

- KNN and Random Forest models achieved the highest accuracy (80.4%) and perfect sensitivity.
- All models showed high sensitivity (>95%), indicating strong performance in identifying positive cases.

- Specificity was generally lower across all models, suggesting a tendency to overpredict positive cases.
- The Decision Tree model underperformed compared to other algorithms.

5. Conclusion

Our analysis demonstrates that machine learning models, particularly KNN and Random Forest, show promise in predicting heart disease. These models achieved high accuracy and perfect sensitivity, making them valuable tools for initial screening. However, the lower specificity across all models indicates a need for further refinement to reduce false positives.

Recommendations for future work include:

1. Feature engineering to create more predictive variables
2. Collecting additional relevant data, such as family history and lifestyle factors
3. Implementing ensemble methods to potentially improve overall performance
4. Conducting a more in-depth analysis of misclassified cases to identify patterns
5. Consulting with domain experts to validate the model's applicability in clinical settings

While these models show potential as screening tools, it's crucial to remember that they should complement, not replace, clinical judgment in heart disease diagnosis and risk assessment.

This report provides a comprehensive overview of our analysis so far. It can serve as a foundation for further discussion, decision-making, and additional research in the project. Remember to adjust any details based on specific information about your dataset or analysis process that I might not have been aware of.