Final Project Report: Heart Disease Prediction Using Machine Learning

DSA460

Hardik Polamarasetti

2864643

Contents

[Introduction 3](#_Toc165844641)

[Project Overview 3](#_Toc165844642)

[Data Preprocessing 4](#_Toc165844643)

[Model Development and Evaluation 5](#_Toc165844644)

[Challenges and Solutions 7](#_Toc165844645)

[Conclusion and Future Work 7](#_Toc165844646)

# Introduction

This project focuses on developing a predictive model to identify individuals at risk of heart disease using machine learning techniques. Given the significant global impact of heart diseases, which are major contributors to mortality worldwide, effective and early prediction models are useful for improving patient outcomes through timely medical interventions.

# Project Overview

The heart disease prediction project uses the comprehensive "heart.csv" dataset, which includes a broad spectrum of cardiovascular metrics collected from clinical environments. This dataset incorporates clinically relevant features such as patient demographics, blood pressure, cholesterol levels, and electrocardiogram results, making it a good foundation for predictive modeling.

Data Description

The dataset comprises various attributes directly linked to cardiovascular health, collected under clinical supervision. It includes both numerical and categorical data, such as age, sex, types of chest pain, resting blood pressures, cholesterol levels, fasting blood sugar, electrocardiographic outcomes, and more. The small sample size of the dataset, while sufficient for initial analyses, presents unique challenges and opportunities for model training and validation, as larger datasets could provide more generalized and reliable insights.

Goals of the Project

The primary goal of this project is to utilize data analytics and machine learning to accurately predict the presence of heart disease in patients. The project aims particularly to achieve high accuracy for the positive class to ensure that potential cases of heart disease are not missed, so that patients can be provided reliable support in clinical decision-making processes.

Methodology

The project utilized a variety of tools from the Python ecosystem to handle different aspects of the data analysis and model building process efficiently: Pandas was used for data manipulation and analysis, facilitating data adjustments and preparation for modeling. NumPy was used for high-level mathematical functions and array operations that were necessary for handling data transformations. Matplotlib and Seaborn were utilized for visualizing data and insights. Scikit-learn provided machine learning functionalities, such as data preprocessing, model building, and model evaluation techniques.

A screenshot of a computer program

Description automatically generated

# Data Preprocessing

Data preprocessing involved several steps designed to convert raw data into a clean dataset ready for modeling.

Feature Selection: Features were carefully selected based on their clinical relevance and statistical significance in predicting heart disease.

Encoding and Scaling: Categorical variables were encoded using OneHotEncoder, allowing models to better interpret the data, while continuous variables were scaled using StandardScaler to normalize their range.

**A graph with a line going up

Description automatically generatedA green graph with a green line

Description automatically generated**

# Model Development and Evaluation

The development and evaluation of models involved implementing Decision Trees, Random Forests, and KNN. Each model's performance was assessed using metrics such as accuracy, precision, recall, and F1-score:

Decision Tree Model: Provided a foundational understanding of the features' influences on outcomes. The smaller sample size could lead to overfitting, where the model too closely fits the limited data points available.

Random Forest Model: Offers improvements over the Decision Tree by aggregating results from multiple trees, which results in reducing variance and lowers the risks of overfitting. However, larger sample sizes would further improve the stability and accuracy of this model's predictions.

KNN Model: Achieved the highest accuracy and demonstrated excellent model performance. The effectiveness of KNN can be sensitive to the dataset's size because it relies on neighbors, meaning more data can provide a better approximation.

A screenshot of a report

Description automatically generated

A screenshot of a report

Description automatically generatedA screenshot of a computer

Description automatically generated

# Challenges and Solutions

Handling smaller datasets and overfitting were significant challenges. The lack of samples in the dataset limited the ability for the models to create more reliable and accurate results. 10 fold Cross-validation was used to enhance model generalizability.

# Conclusion and Future Work

The project successfully demonstrated the use of machine learning techniques to predict heart disease effectively, with KNN performing the best in terms of accuracy and balance between sensitivity and specificity. Future work will focus on acquiring larger datasets to improve model accuracy and reliability. Exploring more algorithms and feature engineering techniques will also be used for enhancing predictive performance. Additionally, integrating real-time data analysis could provide quick and real-time predictions to healthcare professionals for their patients.