Machine Learning Project on Heart Disease Prediction

Group Members: Date: (15.Nov.2024)

Muhammad Owais (Team Leader)

Abdul Hanan

Zahid Hussain

Abstract

This project focuses on predicting heart disease using machine learning. The dataset was preprocessed to handle missing values and normalized for consistency. Various machine learning models were tested, and the best-performing model was selected. The project concludes with insights on prediction accuracy and potential applications in healthcare.

DataSet Background

Heart disease is a major global health concern. Early detection can significantly reduce fatal outcomes. This project leverages machine learning to build predictive models for early detection.

Motivation

Using machine learning for healthcare can enhance accuracy, reduce costs, and save lives.

Objectives

The goal of this project is to develop a predictive model that accurately classifies individuals as at-risk or not at-risk for heart disease based on clinical and demographic features. The project will involve analyzing the dataset, selecting important features, and training machine learning models to predict the presence of heart disease.

Data Sources and Preprocessing

Data Sources: The dataset was sourced from Kaggle, containing patient health attributes.

Data Cleaning: Removed unnecessary columns (fbs). Checked and handled missing values. Normalized data to ensure consistency.

Challenges and Solutions

Challenge: Missing values in key columns.

Solution: Filled with median values.

Challenge: Imbalanced dataset.

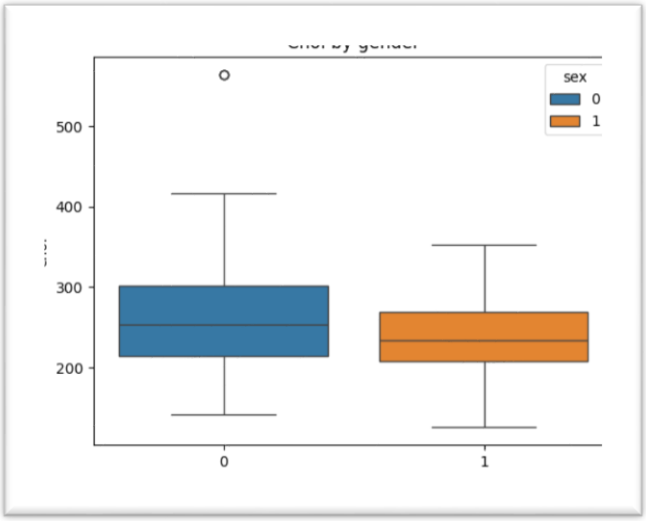
Solution: Applied oversampling techniques.

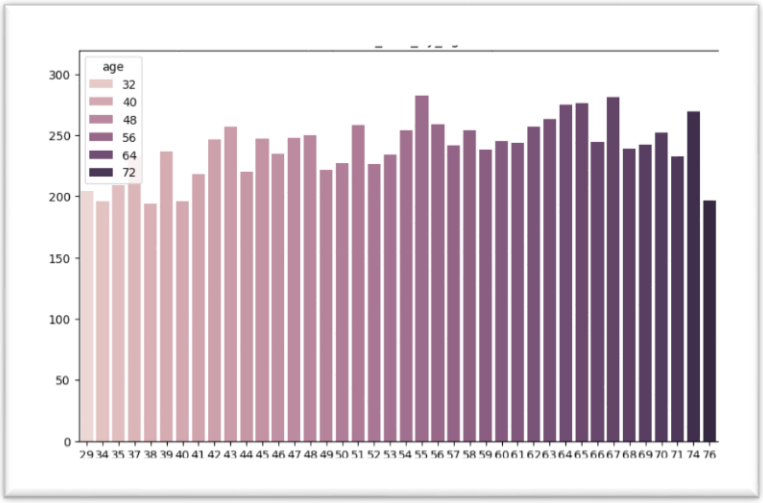
Exploratory Data Analysis

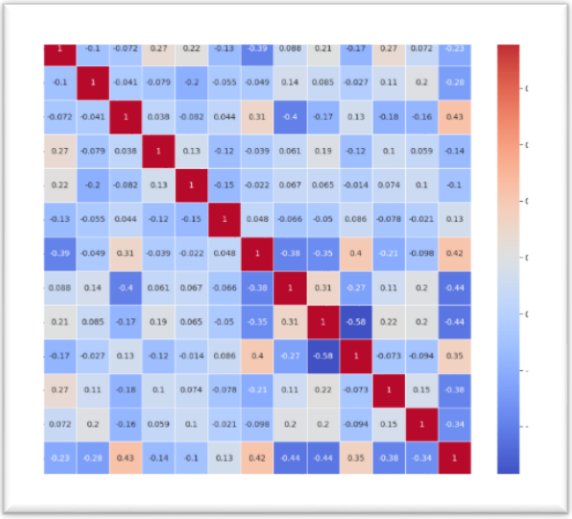
Key Findings: Certain attributes like cholesterol levels and age showed strong correlations with heart disease.

Visualizations: Histograms and scatter plots were used to identify patterns in age, cholesterol, and other attributes. Heatmaps to show correlations between variables.

Insights: Strong correlation observed between chest pain type and disease presence. Older individuals were more prone to heart disease.







Modeling Process

Model Selection: Evaluated Logistic Regression, Random Forest, and Neural Networks.

Model Training: Split data into training (80%) and testing (20%) sets. Hyperparameter tuning was performed using GridSearchCV.

Model Name	Accuracy
Logistic Regression	79.51%
SVM (Support Vector Machine)	97.07%
Decision Tree	97.56%
Random Forest	98.54%
k-Nearest Neighbors (k-NN)	95.61%
Multi-layer Perceptron (MLP)	93.66%

Conclusions

Summary of Findings: Successfully developed a predictive model for heart disease. Random Forest outperformed other models in terms of accuracy.

Implications: The model can be integrated into healthcare systems for early screening.

Limitations:

- The dataset used in this project is relatively small and may not fully represent real-world scenarios.
- The model's performance might vary when deployed on larger or more diverse datasets.

Scope

The scope of a heart disease dataset includes various features collected from patients that can help predict or analyze the presence and severity of heart disease.

Clinical Data: age, gender, blood pressure, cholesterol levels, blood sugar, etc.

Lifestyle Data: Smoking status, alcohol use, physical activity.

Outcome Data: Indications of disease presence or severity, such as diagnosis labels (0 for no heart disease, 1 for presence of disease), type of heart disease, and survival or recovery **rates.**

Future Work

- **Test on larger datasets**: Evaluating the model on a more comprehensive dataset would provide better insights into its real-world performance.

- **Web-based Application**: Developing a user-friendly, web-based application for healthcare professionals to input patient data and receive predictions.