Heart Disease Prediction Using Machine Learning

Introduction

This project leverages machine learning to predict coronary heart disease based on various clinical parameters. The dataset, sourced from Kaggle, provides comprehensive medical data which is key for developing solid predictive models.

Dataset Overview

The dataset consists of 918 entries, each described by 11 features, such as Age, Sex, Chest Pain Type, Blood Pressure, Cholesterol levels, and more. The target variable is HeartDisease, indicating the presence (1) or absence (0) of heart disease.

Key Features

Categorical: Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope
Numerical: Age, RestingBP, Cholesterol, MaxHR, Oldpeak

Data Preprocessing

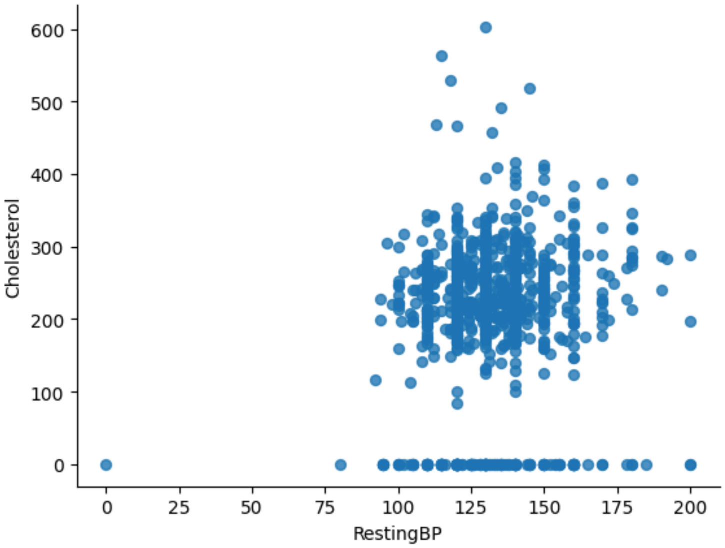
Data Splitting: The dataset was split into training and test sets to ensure robust model evaluation. (80/20 split). Further splitting into a validation set wasn't needed as that's done automatically when running the SVM and Random Forest algorithms (CV=5)

The data was first inspected for missing values, with no null entries found. Some outliers might be present (box plots can help with detecting them) but further investigation by a cooperating medical expert could help in that regard. Removing them could potentially improve the model.

Encoding: Categorical variables ('Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_Slope') were transformed using OneHotEncoder.

Scaling: Numerical features were scaled using StandardScaler. (Particularly useful for SVM)

Imputation: For the 'RestingBP' and 'Cholesterol' features, non-positive(<=0) values were deemed incorrect. 1 was detected in RestingBP (which could be removed since it's not a big part of the dataset) and 172 were 0 for Cholesterol. All 173 of these false values were imputed with the mean of the rest of the valid data.



Support Vector Machine (SVM)

Hyperparameters Tuned: Regularization parameter (C) and kernel type.
Grid Search: Employed to find the optimal parameters with accuracy as the scoring metric.

Results:
Best parameters:
C=1, kernel='rbf'.

Cross-validation accuracy: 0.86
High, indicating good generalization on training data.

Test accuracy: 0.86
Also high, Indicates that the model generalizes well to new, unseen data.

Random Forest Classifier

Hyperparameters Tuned:
Number of trees (n_estimators)
Number of features considered for splits (max_features).

Results:
Best parameters:
300 trees, max_features='auto'

Cross-validation accuracy: 0.85
Comparable to SVM.

Test accuracy: 0.89
Higher than SVM, showcasing better performance and generalization.

Comparative Analysis

The Random Forest model outperformed the SVM in terms of test set accuracy. This suggests that for this particular dataset, Random Forest is more effective, likely due to its ability to handle non-linear relationships and feature interactions more effectively than SVM.

Conclusion and Recommendations

The project successfully demonstrates the application of SVM and Random Forest to predict heart disease. Random Forest, with its superior performance, is recommended for further development and deployment. Future work could explore more complex ensemble methods and deep learning approaches to further enhance model performance.

Additional metrics like precision, recall, and F1-score should be considered to fully assess the model's clinical applicability, especially to minimize false negatives in disease prediction.