Concepts and Technologies of AI 5CS037

Classification Analysis Report

Uni ID: 2408485

Submitted by: Aryush Khatri

Lecturer: Ms. Sunita Parajuli

Tutor: Mr. Siman Giri

Submitted on: 2025/02/11

Abstract

This report aims to predict the presence of heart disease using classification techniques. The dataset used for this analysis is heart.csv, which includes patient health records and various risk factors. The process involves several steps: conducting Exploratory Data Analysis (EDA), building models using Random Forest and Gradient Boosting classifiers, optimizing hyper-parameters, and selecting relevant features. This report aims to predict the presence of heart disease using classification techniques. The dataset used for this analysis is heart.csv, which includes patient health records and various risk factors. The process involves several steps: conducting Exploratory Data Analysis (EDA), building models using Random Forest and Gradient Boosting classifiers, optimizing hyper-parameters, and selecting relevant features.

The evaluation of the developed models included the assessment of their accuracy combined with precision along with recall and F1-score measurements. The Random Forest Classifier model demonstrated 89% accuracy together with 86% accuracy from the Gradient Boosting model. Random Forest produced superior values for all performance metrics including recall and precision and f1-score than Gradient Boosting. Before selection, I chose Random Forest over the other performance metrics because it maintained the strongest balance throughout all capacity evaluation metrics.

Random Forest outperformed all other classification models by reaching the highest levels of accuracy and f1-score and recall which makes it the preferred choice for predicting heart disease. The slightly inferior performance of Gradient Boosting makes it less appropriate for medical diagnostics that put high priority on positive case detection. This assessment reveals cholesterol levels and rest blood pressure together with age act as significant predictors for heart disease. The model achieved better results after hyperparameter tuning because decision thresholds received optimized values. Feature selection proved useful for both decreasing disorders in data and enhancing model generality

Table of Contents

1. Introduction

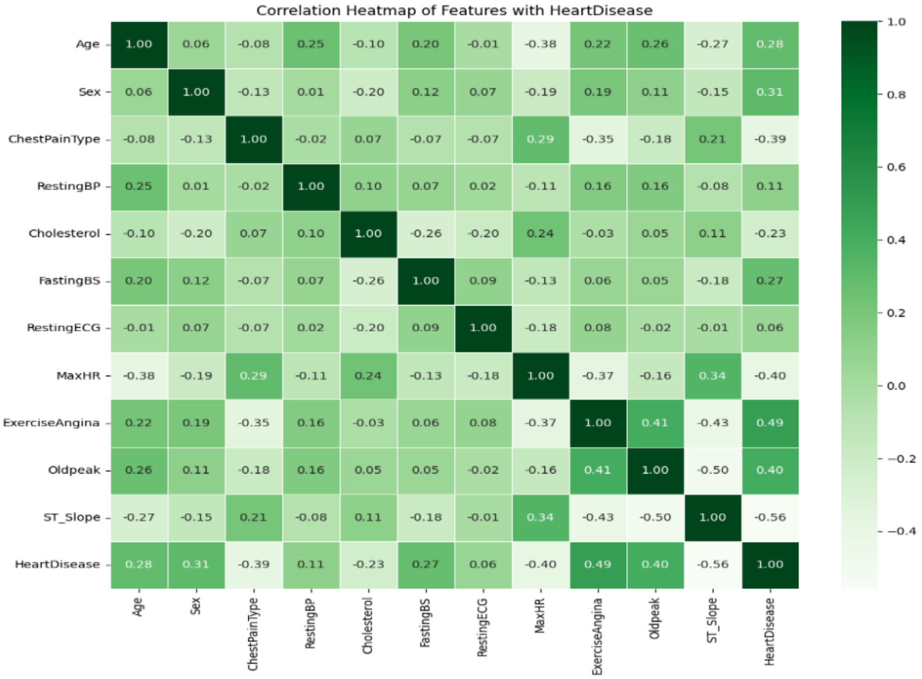
This project aims to determine heart disease infection in patients through analysis of their medical characteristics. Prediction of heart disease becomes essential because it aids early detection as well as preventive healthcare initiatives.

The research utilizes the heart.csv dataset, which can be obtained from Kaggle. This dataset contains both numerical and categorical health features that describe patients' data. It aligns with the objectives of the United Nations Sustainable Development Goals (UNSDG) by promoting good health and well-being.

The purpose of this evaluation is to develop a classification model that predicts heart disease emergence from medical data.

2. Methodology

- No missing values in my dataset.
- No duplicate values found.
- Categorical variables were encoded appropriately using Label Encoder.



This heatmap illustrates the mathematical relation between heart disease diagnosis and other features within the dataset. ST_Slope with a strength of -0.56 and ExerciseAngina at 0.49 together with Oldpeak exerting 0.40 strength indicate their prominence in predicting heart disease incidence. The likelihood of heart disease diminishes with rising values of ST_Slope and MaxHR but increases with ExerciseAngina and FastingBS values.

For this classification task I have chosen following models:

- Random Forest Classifier
- Gradient Boosting Classifier

Steps to build the model:

- Splitting the dataset into training and testing sets.
- Training the models using relevant features.
- Performing feature selection to improve model performance.
- Optimizing hyperparameters using GridSearchCV.

Hyperparameter tuning was performed using GridSearchCV, resulting in the identification of the optimal parameters:

Optimal Parameters for Random Forest:

- n_estimators: 100
- max_depth: 10
- min_samples_split: 5

Optimal Parameters for Gradient Boosting:

- learning_rate: 0.1
- n_estimators: 100
- max_depth: 3

From SelectFromModel I did feature selection, identifying the most significant features for each model:

- For Random Forest Classifier**:** The model initially had 11 features, which were reduced to 6 after selection. The selected features were:
    - Age
    - ChestPainType
    - Cholesterol
    - MaxHR
    - Oldpeak
    - ST_Slope
- For Gradient Boosting Classifier**:** Features were selected based on the median importance threshold. The selected features were:
    - ChestPainType

- Cholesterol
- MaxHR
- ExerciseAngina
- Oldpeak
- ST_Slope

3. Conclusion

The results show model performance

The initial classification report using Random Forest:

accuracy = 89%

	precision	Recall	F1-score
0	0.85	0.89	0.87
1	0.92	0.90	0.91

The initial classification report using Gradient Boosting:

accuracy = 89%

	precision	Recall	F1-score
0	0.79	0.88	0.84
1	0.91	0.84	0.88

The Random Forest Classifier was the most effective at predicting heart disease, achieving: Accuracy: 89% and its other metric were also better so it was used to rebuild the final model using best parameters.

The project faced several challenges, including:

- Selecting the right models for this dataset
- Finding the best parameters using GridSearchCV was computationally time-consuming.

The model's performance can be improved by doing as follows:

- Optimizing feature selection techniques
- Increasing dataset size

4. Discussion

The models were evaluated using accuracy, precision, recall, and F1-score. The results indicated that the Random Forest Classifier outperformed Gradient Boosting in terms of overall accuracy, making it the most effective model for predicting heart disease.

These techniques were applied effectively.:

- Hyperparameter Tuning**:**
    - Optimizing Random Forest Classifier with GridSearchCV slightly decreased accuracy from 89% to 88%.
- Feature Selection**:**
    - Reducing the number of features from 11 to 6 in Random Forest Classifier resulted in:
        - Improved model efficiency without significant loss of accuracy.
        - Selected Features: Age, ChestPainType, Cholesterol, MaxHR, Oldpeak, ST_Slope
    - ➢ Gradient Boosting Classifier selected:
        - ChestPainType, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, ST_Slope

Overall, these techniques enhanced model performance, reduced overfitting, and improved interpretability while maintaining high predictive power.

The chosen features and models performed as expected, providing insights into:

- Key Features for Prediction**:**
    - Both models identified ChestPainType, Cholesterol, MaxHR, Oldpeak, and ST_Slope as critical indicators of heart disease.
    - ExerciseAngina was additionally selected by Gradient Boosting, indicating its importance in detecting heart-related issues.

- Model Effectiveness:
    - The Random Forest Classifier was the most effective model, achieving an accuracy of 89%. This success is likely due to its capability to manage feature importance and reduce overfitting through ensemble learning..
    - Gradient Boosting Classifier performed well but was slightly less accurate, potentially due to its sensitivity to hyperparameters.

Overall, the findings confirm that machine learning models can effectively predict heart disease using a well-selected set of medical attributes**.**

- Dataset Size Constraints**:**
    - The dataset contains 918 samples, which may not be sufficient for a highly generalized model. A larger dataset could improve performance and robustness.
- Feature Dependencies:
    - Certain attributes, including Cholesterol and MaxHR, might exhibit intricate interactions that the models do not completely capture. More sophisticated methods such as deep learning might delve deeper into these connections.
- Model Interpretability:
    - While Random Forest performed well, it operates as a black-box model, making it harder to interpret is more complex compared to simpler models, such as Logistic Regression.

These limitations suggest areas for improvement in future research and modeling efforts.

Future research could explore:

• Evaluating further classification algorithms

- Applying deep learning techniques

- Enlarging the dataset to enhance generalization