Emma Mukuri Maka

**Final Report:**

**Optimizing the Prediction Accuracy of Heart Disease Presence**

**Introduction**

Heart disease remains one of the leading causes of mortality worldwide. Accurate prediction of heart disease presence is crucial for early diagnosis and effective treatment. This project aims to optimize the prediction accuracy of heart disease presence by leveraging feature selection techniques and various machine learning algorithms. The dataset used in this project is sourced from the UCI Machine Learning Repository, which contains a mix of numerical and categorical features related to patient health and demographic information.

**Problem Statement**

How can we optimize the prediction accuracy of heart disease presence using machine learning algorithms?

The primary goal of this project is to determine how we can optimize the prediction accuracy of heart disease presence using machine learning algorithms. This involves selecting the best-performing machine learning models to enhance prediction accuracy.

**Data Wrangling & EDA**

With the dataset being mostly clean the only real problem to handle for the data wrangling had been to handle the missing values. So, after loading the necessary libraries, that was done.

The Exploratory Data Analysis (EDA) involves visualizing and summarizing the key characteristics of the dataset. With this, histograms, boxplots, and a correlation matrix were used.

Histograms and box plots were used to visualize the distribution of numerical features. This helped in understanding the spread and central tendency of the data.

A group of colored boxes

Description automatically generated with medium confidenceA group of blue and gray graphs

Description automatically generated with medium confidence

A diagram of a number of numbers

Description automatically generated with medium confidenceA correlation matrix was generated to identify the relationships between features. Features with high correlation were noted as they might influence the model’s performance.

**Modeling**

I decided to try 3 different models for this exercise. The models I used include Random Forest, K-Nearest Neighbors, and Gradient Boosting Classifier.

The model evaluation metrics used and individual results:

* Confusion matrix and accuracy score were used to understand the distribution of true positives, true negatives, false positives, and false negatives, and measure the overall accuracy of the model.

A diagram of a graph

Description automatically generated with medium confidenceFirst, **Gradient Boosting**:

Had a high accuracy for class 0 (27 correct predictions). However, struggles with classes 1 and 3, showing multiple misclassifications.

**A diagram of a confusion matrix

Description automatically generated with medium confidenceK-Nearest Neighbors**:

Had good performance for class 0 (27 correct predictions), but poor performance for classes 1, 2, and 3, with significant misclassifications.

**A diagram of a forest confusion matrix

Description automatically generatedRandom Forest**:

Had the highest accuracy for class 0 (28 correct predictions) and had a balanced performance across other classes, with fewer misclassifications than Gradient Boosting and KNN.

* The other metrics used include the ROC Curve and AUC, to evaluate the model's ability to distinguish between classes.

**Gradient Boosting** and **Random Forest** both had high AUC values for class 0 (0.91), but Random Forest showed better performance across other classes with higher AUC values, indicating better discriminatory power. **KNN** had the lowest AUC values across all classes, indicating it is the least effective in distinguishing between classes.

A graph showing the different types of class

Description automatically generated with medium confidenceA graph of different colored lines

Description automatically generatedA diagram of a receiver operating characteristic

Description automatically generated

*From top to bottom: Gradient Boosting, KNN and Random Forest.*

**Conclusions**

In conclusion to optimize the prediction accuracy of heart disease presence, **Random Forest** is the most effective machine learning algorithm among those tested. It provides the highest accuracy, robust performance across different classes, and superior AUC values. This model's ability to handle both numerical and categorical data and its robustness against overfitting make it well-suited for this classification task. Future work could focus on further hyperparameter tuning and potentially integrating feature selection techniques to enhance the model's performance even further.