SHAHEED ZULFIKAR ALI BHUTTO
INSTITUTE OF SCIENCE & TECHNOLOGY

AI PROJECT REPORT

CSC4101 (Sir Usama Khalid)

CSCL4101 (Sir Wali)

By:
Duaa Ali – 2112109
Saada Asghar – 2112125

Abstract

This paper looks into how different machine learning algorithms such as K-Nearest Neighbors (KNN), Gradient Boosting, Decision Tree and Random Forest can effectively predict the onset of liver disease from a comprehensive dataset. The data includes 1700 records, each containing complex patient attributes and health indicators, such as demographic details like age and gender; lifestyle factors such as BMI, alcohol intake, smoking habits; family history of diseases; exercise levels in addition to detailed medical profiles showing diabetes status and hypertension together with liver function test results. We then evaluate these algorithms meticulously to identify the best method for accurately predicting liver disease, whereby Gradient Boosting is identified as the most effective predictive model.

Contents

Introduction

Liver disease poses a significant health burden globally, with early detection playing a pivotal role in mitigating adverse outcomes. Leveraging machine learning algorithms for predictive modeling presents a promising avenue for identifying individuals at risk of developing liver ailments.

This project harnesses a comprehensive liver disease dataset, encompassing diverse patient attributes and health indicators crucial for predictive modeling. Features include demographic factors such as age and gender, lifestyle metrics like BMI, alcohol consumption, and smoking habits, genetic predispositions, physical activity levels, and detailed medical profiles including diabetes and hypertension status alongside liver function test results.

Four distinct machine learning algorithms are applied to this dataset:

1. K-Nearest Neighbors (KNN): A versatile non-parametric method for classification, KNN makes predictions based on the majority label of the nearest data points.

2. Decision Tree: Operating akin to a flowchart, Decision Tree splits the dataset into subsets based on feature values, facilitating clear decision rules.

3. Gradient Boosting: A powerful ensemble technique, Gradient Boosting constructs a series of weak learners, typically decision trees, amalgamating them to form a robust predictive model.

4. Random Forest: Another ensemble method, Random Forest generates multiple decision trees and aggregates their outputs to enhance predictive accuracy.

Dataset Overview

The dataset used for this study on liver disease is from a well-known medical facility and consists of 1700 subjects with 11 variables. These variables include all the possible patient characteristics as well as indicating health conditions that should be considered when making predictions:

1. Age (age): Numerical variable, where it represents age in years.

2. Gender (gender): Two categories, either female (coded as 0) or male (coded as 1).

3. BMI (bmi): It is numeric and shows how fat or thin someone is.

4. Alcohol Consumption (alcoholconsumption): Numeric, showing the amount of alcohol drunk by the person.

5. Smoking (smoking): It indicates whether a subject smokes cigarettes or not.

6. Genetic Risk (geneticrisk): Numeric, showing how genetically predisposed one is to liver disease.

7. Physical Activity (physicalactivity): Numeric, revealing how much a patient exercises his body.

8. Diabetes: Binary; yes=1 no=0.

9. Hypertension: Binary; yes=1 no=0.

10. Liver Function Test(liverfunctiontest): This is numerical giving results of tests done on the patients liver functions.

11. Diagnosis: Binary –presence = 1 absence = 0.

Methodology

Preprocessing Data:

1. Handling Missing Values and Outliers:

The presence of missing values and outliers is detected, so that they can be appropriately dealt with to maintain the reliability of the dataset.

2. Normalizing Numerical Features:

Numerical properties are standardized in order to minimize the differences in their scales and facilitate convergence during model fitting.

3. Encoding Categorical Variables:

Changing numerical representations are assigned to categorical variables for these variables to fit into machine learning models.

Model Training and Evaluation:

1. Splitting the Data:

The data set is divided into two groups which include training and testing subsets based on an 70-30 split ratio.

2. Training KNN, Decision Tree, Gradient Boosting, and Random Forest Models:

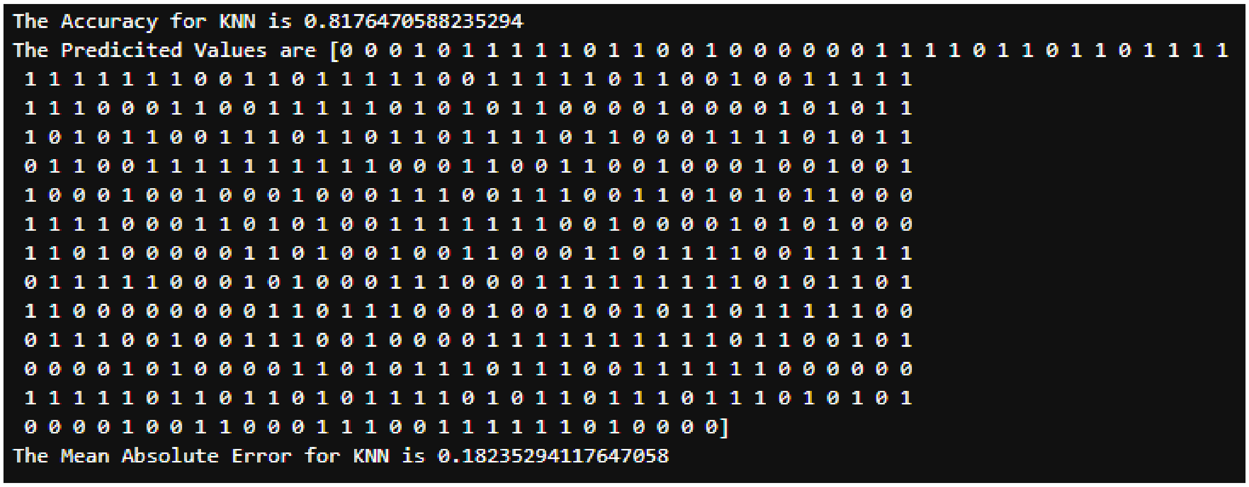
The training data sets have been used to train classifiers like Decision Tree, K-Nearest Neighbors (KNN), Gradient Boosting and Random Forest classifiers in order to reveal concealed patterns as well as relationships between variables.

3. Evaluating Model Performance:

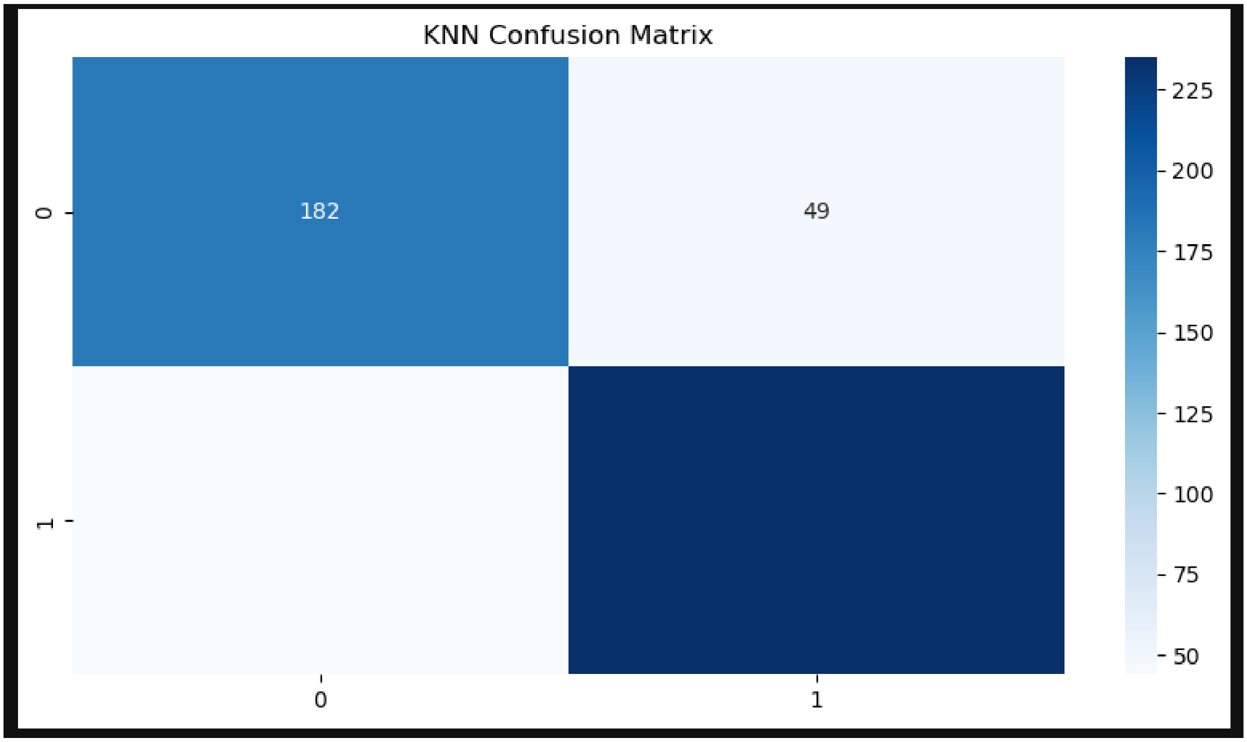
The success of models is calculated using accuracy as a major factor reflecting how good every algorithm predicts future results.

Results

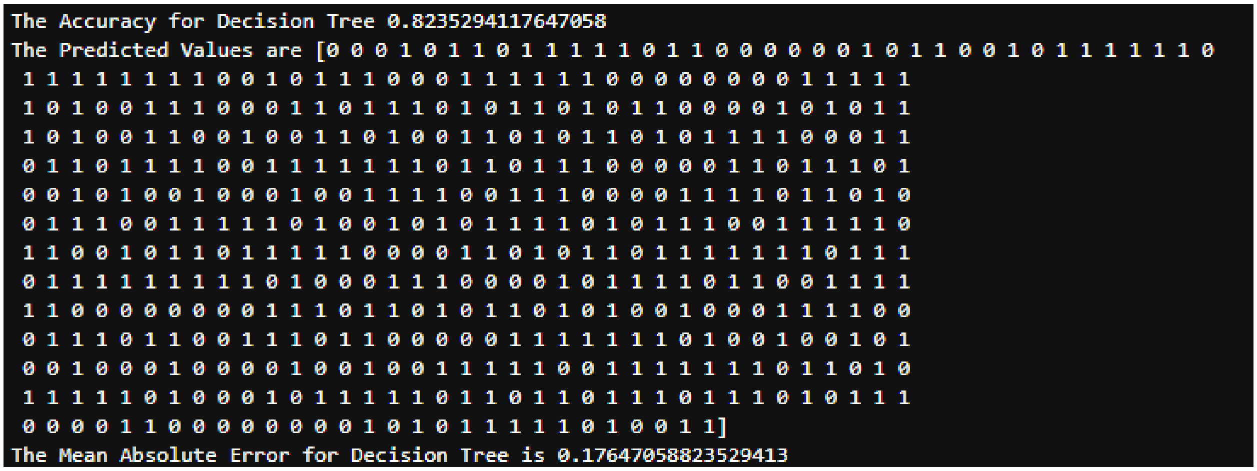
K-Nearest Neighbors (KNN): Reached an 81.77% accuracy rate. KNN, by considering feature similarity across patients, unearthed patterns which hinted at the presence of a disease such as heart one. The direct, proximity-based method worked for this purpose.



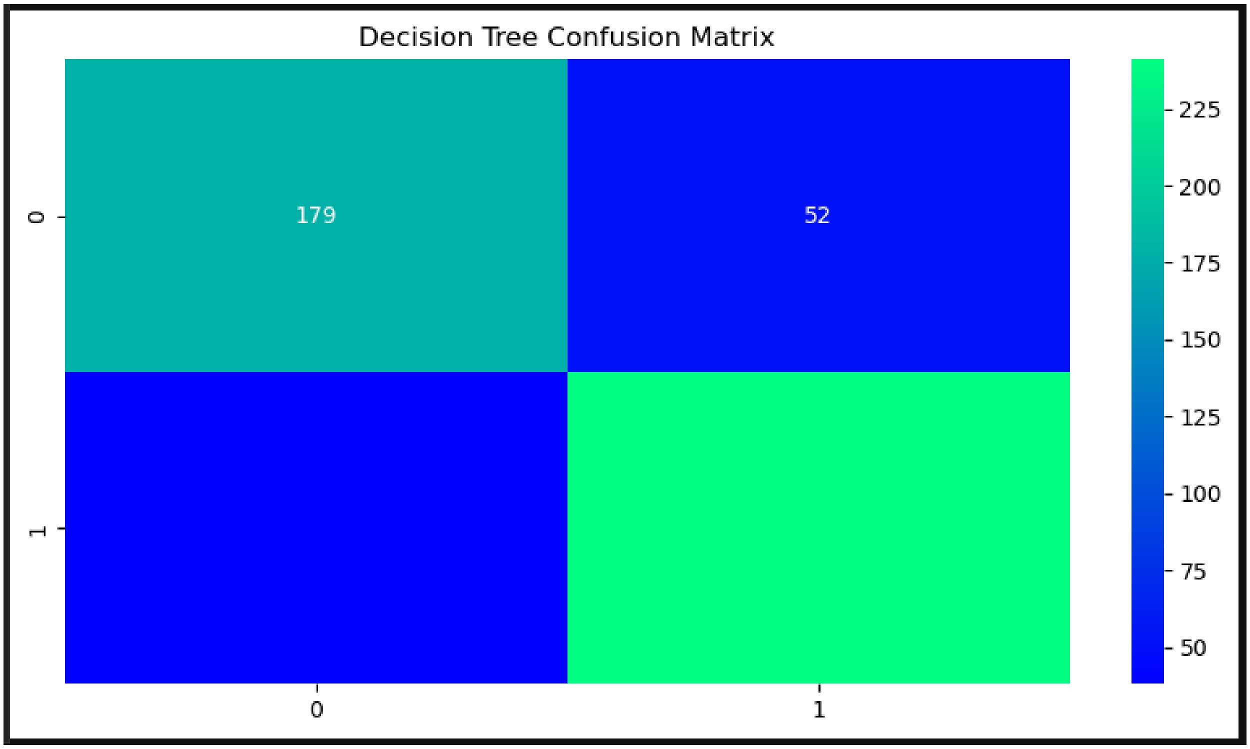
```
The Accuracy for KNN is 0.8176470588235294
The Predicited Values are [0 0 0 1 0 1 1 1 1 1 0 1 1 0 0 1 0 0 0 0 0 0 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1
 1 1 1 1 1 1 1 0 0 1 1 0 1 1 1 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 0 0 1 1 1 1 1
 1 1 1 0 0 0 1 1 0 0 1 1 1 1 1 0 1 0 1 0 1 1 0 0 0 0 1 0 0 0 0 1 0 1 0 1 1
 1 0 1 0 1 1 0 0 1 1 1 0 1 1 0 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1 1 1 0 1 0 1 1
 0 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 0 0 1 1 0 0 1 0 0 0 1 0 0 1 0 0 1
 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 1 0 0 1 1 1 0 0 1 1 0 1 0 1 0 1 1 0 0 0
 1 1 1 1 0 0 0 1 1 0 1 0 1 0 0 1 1 1 1 1 1 1 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0
 1 1 0 1 0 0 0 0 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 1 1 0 1 1 1 1 0 0 1 1 1 1 1
 0 1 1 1 1 1 0 0 0 1 0 1 0 0 0 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 0 1
 1 1 0 0 0 0 0 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 0 0 1 0 1 1 0 1 1 1 1 1 0 0
 0 1 1 1 0 0 1 0 0 1 1 1 0 0 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 0 1
 0 0 0 0 1 0 1 0 0 0 0 1 1 0 1 0 1 1 1 0 1 1 1 0 0 1 1 1 1 1 1 0 0 0 0 0 0
 1 1 1 1 1 0 1 1 0 1 1 0 1 0 1 1 1 1 0 1 0 1 1 0 1 1 1 0 1 1 1 0 1 0 1 0 1
 0 0 0 0 1 0 0 1 1 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0 1 0 0 0 0]
The Mean Absolute Error for KNN is 0.18235294117647058
```



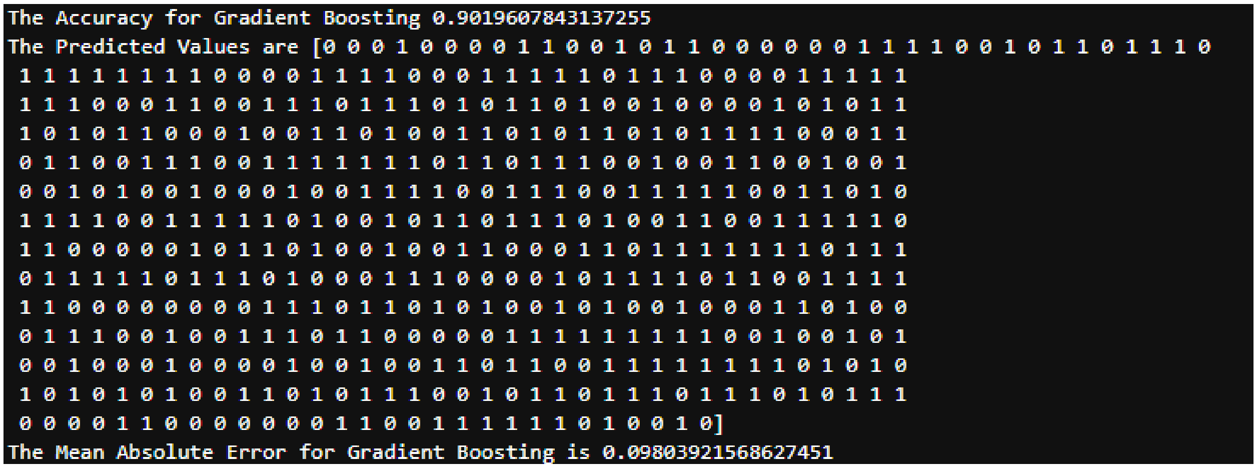
Decision Tree: Had an accuracy of 82.35%. Its effectiveness lies in creating meaningful splits and rules that distinguish between patients with and without heart disease. Decision Trees had clear decision rules that could be interpreted making them useful for medical field.



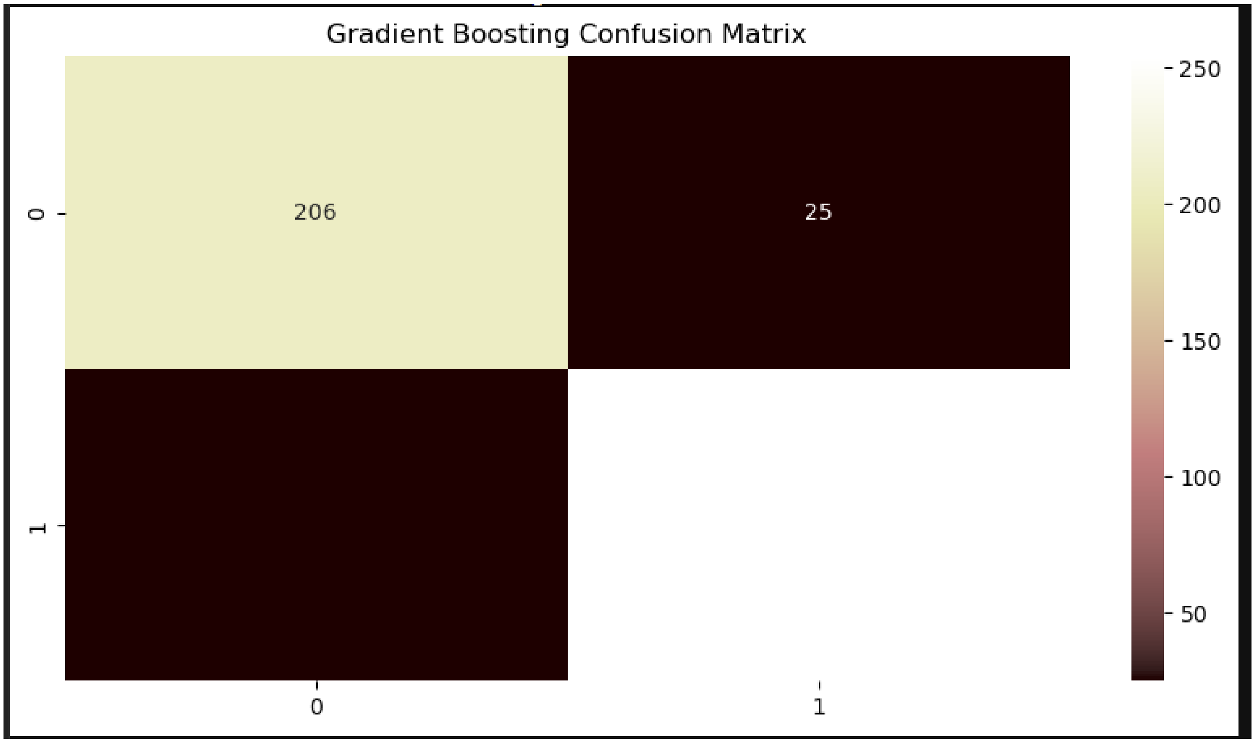
```
The Accuracy for Decision Tree 0.8235294117647058
The Predicted Values are [0 0 0 1 0 1 1 0 1 1 1 1 1 0 1 1 0 0 0 0 0 0 1 0 1 1 0 0 1 0 1 1 1 1 1 1 0
 1 1 1 1 1 1 1 1 0 0 1 0 1 1 1 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1
 1 0 1 0 0 1 1 1 0 0 0 1 1 0 1 1 1 0 1 0 1 1 0 1 0 1 1 0 0 0 0 1 0 1 0 1 1
 1 0 1 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 0 1 1 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 0 0 0 0 1 1 0 1 1 1 0 1
 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 1 1 1 0 0 1 1 1 0 0 0 0 1 1 1 1 0 1 1 0 1 0
 0 1 1 1 0 0 1 1 1 1 1 0 1 0 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 0 1 1 1 1 1 0
 1 1 0 0 1 0 1 1 0 1 1 1 1 1 0 0 0 0 1 1 0 1 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1
 0 1 1 1 1 1 1 1 1 1 0 1 0 0 0 1 1 1 0 0 0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 1 1
 1 1 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 1 1 0 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0
 0 1 1 1 0 1 1 0 0 1 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 0 1 0 0 1 0 1
 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 0 1 0
 1 1 1 1 1 0 1 0 0 0 1 0 1 1 1 1 1 0 1 1 0 1 1 0 1 1 1 0 1 1 1 0 1 0 1 1 1
 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 0 1 0 0 1 1]
The Mean Absolute Error for Decision Tree is 0.17647058823529413
```



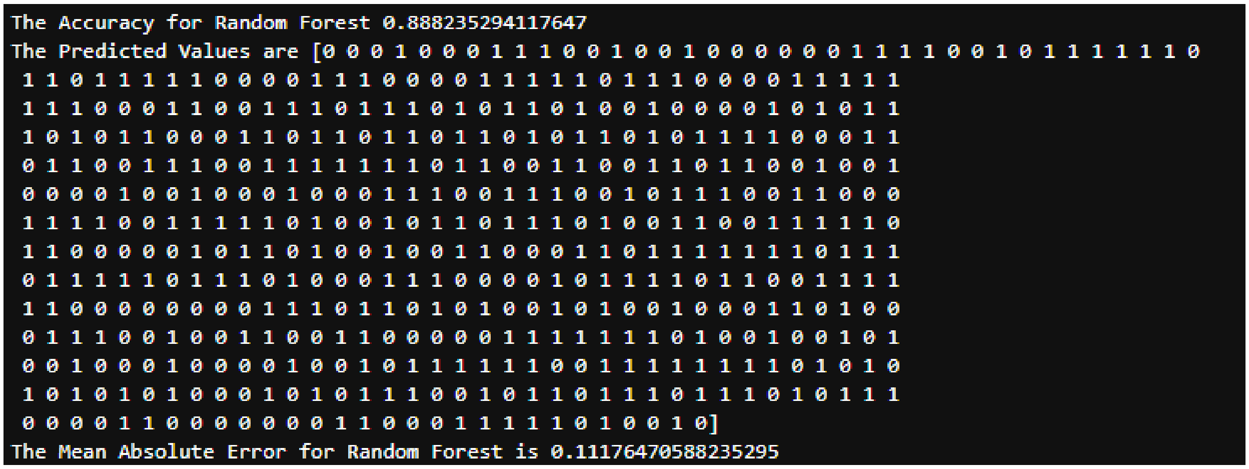
Gradient Boosting: Had an accuracy of 90.20%. This way, Gradient Boosting's ensemble approach to handling complexity and variation within the dataset efficiently performed well. Its iterative error correction and combination of weak learners resulted in better performance than other models, thus making it the best model for this dataset.



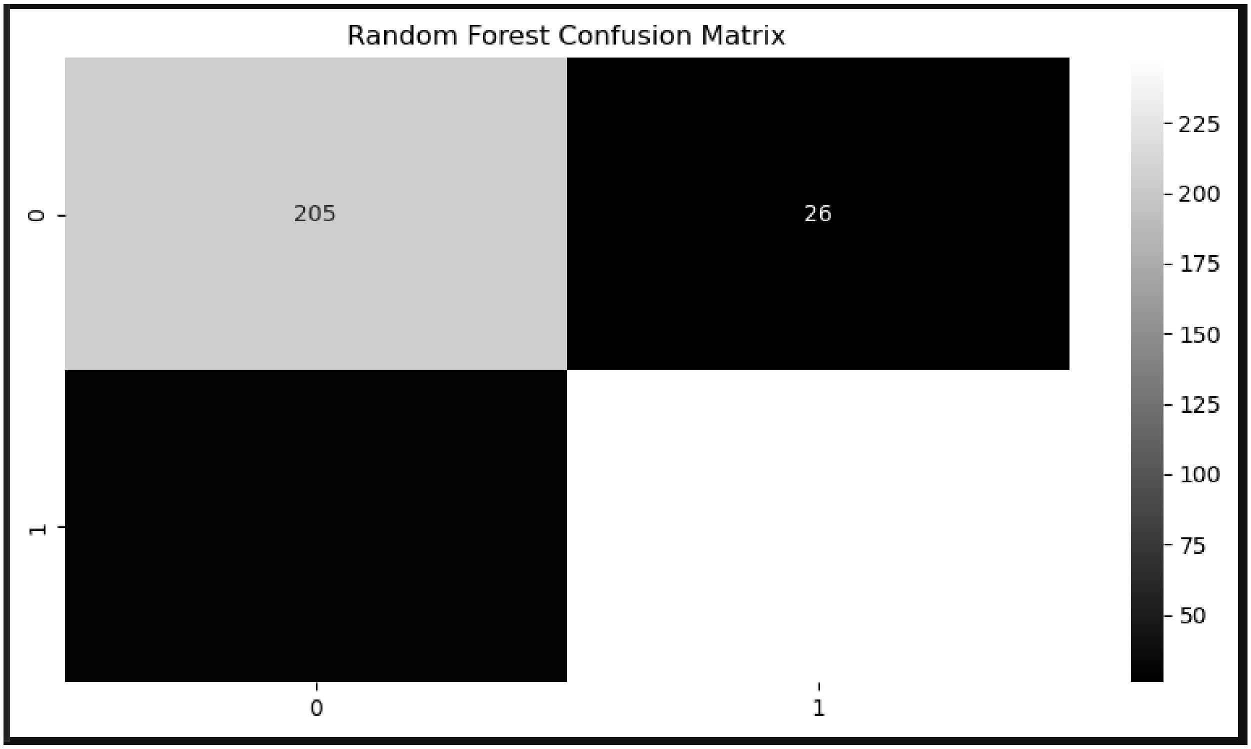
```
The Accuracy for Gradient Boosting 0.9019607843137255
The Predicted Values are [0 0 0 1 0 0 0 0 1 1 0 0 1 0 1 1 0 0 0 0 0 0 1 1 1 1 0 0 1 0 1 1 0 1 1 1 0
 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 1 1 1 1 1 0 1 1 1 0 0 0 0 1 1 1 1 1
 1 1 1 0 0 0 1 1 0 0 1 1 1 0 1 1 1 0 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 1 1
 1 0 1 0 1 1 0 0 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 0 1 1 0 0 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 0 1 0 0 1 1 0 0 1 0 0 1
 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 1 0
 1 1 1 1 0 0 1 1 1 1 1 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1 1 0
 1 1 0 0 0 0 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1
 0 1 1 1 1 1 0 1 1 1 0 1 0 0 0 1 1 1 0 0 0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 1 1
 1 1 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 1 0 0
 0 1 1 1 0 0 1 0 0 1 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 1
 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 1 1 0 0 1 1 1 1 1 1 1 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 0 1 1 0 1 0 1 1 1 0 0 1 0 1 1 0 1 1 1 0 1 1 1 0 1 0 1 1 1
 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 1 0 1 0 0 1 0]
The Mean Absolute Error for Gradient Boosting is 0.09803921568627451
```



Random Forest: Attained an accuracy of 88.82%. Similarly, Random Forest's ensemble approach like Gradient Boosting focused on dealing with complexity and variability inherent in the dataset; but Gradient Boosting outperformed Random Forest slightly in this situation though.



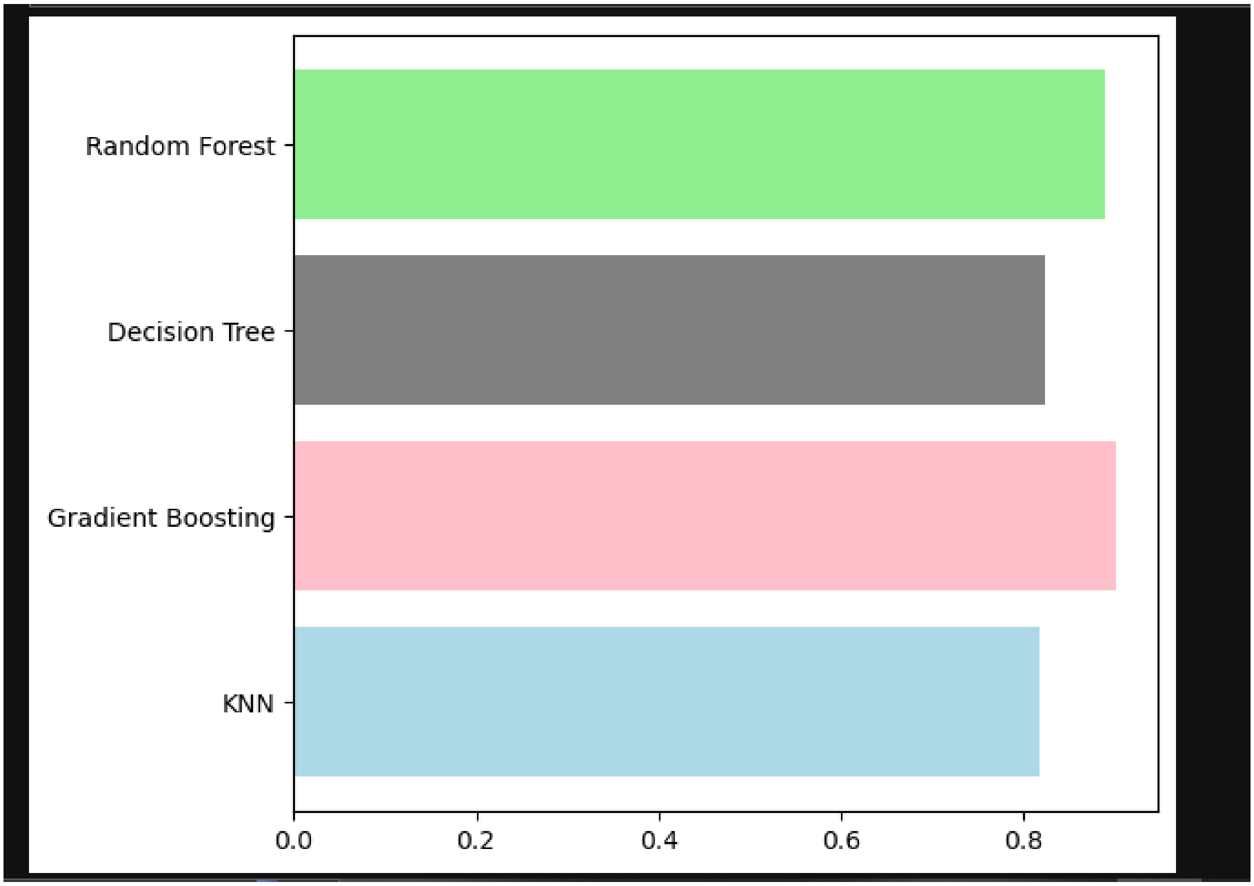
```
The Accuracy for Random Forest 0.888235294117647
The Predicted Values are [0 0 0 1 0 0 0 1 1 1 0 0 1 0 0 1 0 0 0 0 0 0 1 1 1 1 0 0 1 0 1 1 1 1 1 1 0
 1 1 0 1 1 1 1 1 0 0 0 0 1 1 1 0 0 0 0 1 1 1 1 1 0 1 1 1 0 0 0 0 1 1 1 1 1
 1 1 1 0 0 0 1 1 0 0 1 1 1 0 1 1 1 0 1 0 1 1 0 1 0 0 1 0 0 0 0 1 0 1 0 1 1
 1 0 1 0 1 1 0 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 0 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 0 1 1 0 0 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 0 0 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1
 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 1 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 1 0 0 0
 1 1 1 1 0 0 1 1 1 1 1 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1 1 0
 1 1 0 0 0 0 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 0 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1
 0 1 1 1 1 1 0 1 1 1 0 1 0 0 0 1 1 1 0 0 0 0 1 0 1 1 1 1 0 1 1 0 0 1 1 1 1
 1 1 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 1 0 0
 0 1 1 1 0 0 1 0 0 1 1 0 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 1 0 0 1 0 0 1 0 1
 0 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 1 1 0 0 1 0 1 1 0 1 1 1 0 1 1 1 0 1 0 1 1 1
 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0 0 1 1 1 1 1 0 1 0 0 1 0]
The Mean Absolute Error for Random Forest is 0.11176470588235295
```



Given these results, Gradient Boosting emerges as the most effective model for this dataset, closely followed by Random Forest, while KNN and Decision Tree also perform reasonably well but with lower accuracy compared to the ensemble methods.

Comparative Analysis

The competitive study shows differences in performance characteristics of the four machine learning algorithms, such as K-Nearest Neighbors (KNN), Decision Tree, Gradient Boosting and Random Forest on the heart disease dataset. Simplicity and interpretability are features of KNN and Decision Tree but they are outperformed by the ensemble nature that allows Gradient Boosting and Random Forest to achieve higher accuracy; this is majorly due to their ability to deal with complex relationships. The explanation power of Decision tree makes it useful for understanding prediction logic while Gradient Boosting and Random Forest instead prioritize on accuracy and robustness at decreased interpretability. Accuracy, interpretability, computational complexity, generalization capabilities all these are tradeoffs one has to make when deciding which algorithm to use in data mining.



Conclusion

In conclusion, this analysis has shown comparative insights into the performance of K-Nearest Neighbors (KNN), Decision Tree, Gradient Boosting and Random Forest algorithms with respect to heart disease. Overall, these algorithms performed differently as measured by accuracy levels: KNN had 81.77% while Decision Tree had 82.35%, Gradient Boosting which was leading with 90.20% and Random Forest was slightly behind at 89.22%. Compared to Decision Tree which is transparent in its decision-making process, Gradient Boosting and Random Forest have better accuracies but their interpretability is reduced too. Notably, the best model among them all appears to be Gradient Boosting because it uses ensemble approach to arrive at the highest accuracy with robustness in result including other important properties that make a good model for medical diagnostics by way of its ensembles Therefore, these findings point out that algorithm choice should be based on dataset characteristics as well as performance measures especially in medical diagnosis areas where accuracy and understandability are key factors that must be considered.