MACHINE LEARNING CASE STUDY

Title: Predicting Heart Disease Risk with Machine Learning

MUHAMMAD KAZIM - 22067489

OBJECTIVES

- Heart disease is the leading cause of death globally, claiming the lives of over 18 million people annually.
- Early detection and prevention are crucial for managing this health concern.
- This project aims to analyze the factors influencing heart disease and develop a classification model to identify individuals at high risk of developing the condition.

INTRODUCTION TO THE DATASET

- The dataset selected for this analysis, "heart.csv," is sourced from Kaggle, a prominent platform for data science and machine learning resources.
- https://www.kaggle.com/datasets/arezaei81/heartcsv
- This dataset offers a valuable resource for studying heart disease, comprising a comprehensive collection of medical attributes associated with the condition. With a focus on predictive modeling, the dataset includes features such as age, gender, cholesterol levels, and exercise-induced angina, which might be influential factors in determining heart disease risk.
- By exploring the relationships between these attributes, we aim to gain a deeper understanding of the underlying patterns and correlations that contribute to heart disease. This exploration is vital for developing effective predictive models and advancing our knowledge in cardiovascular health.

EXPLORATORY DATA ANALYSIS

DATASET OVERVIEW

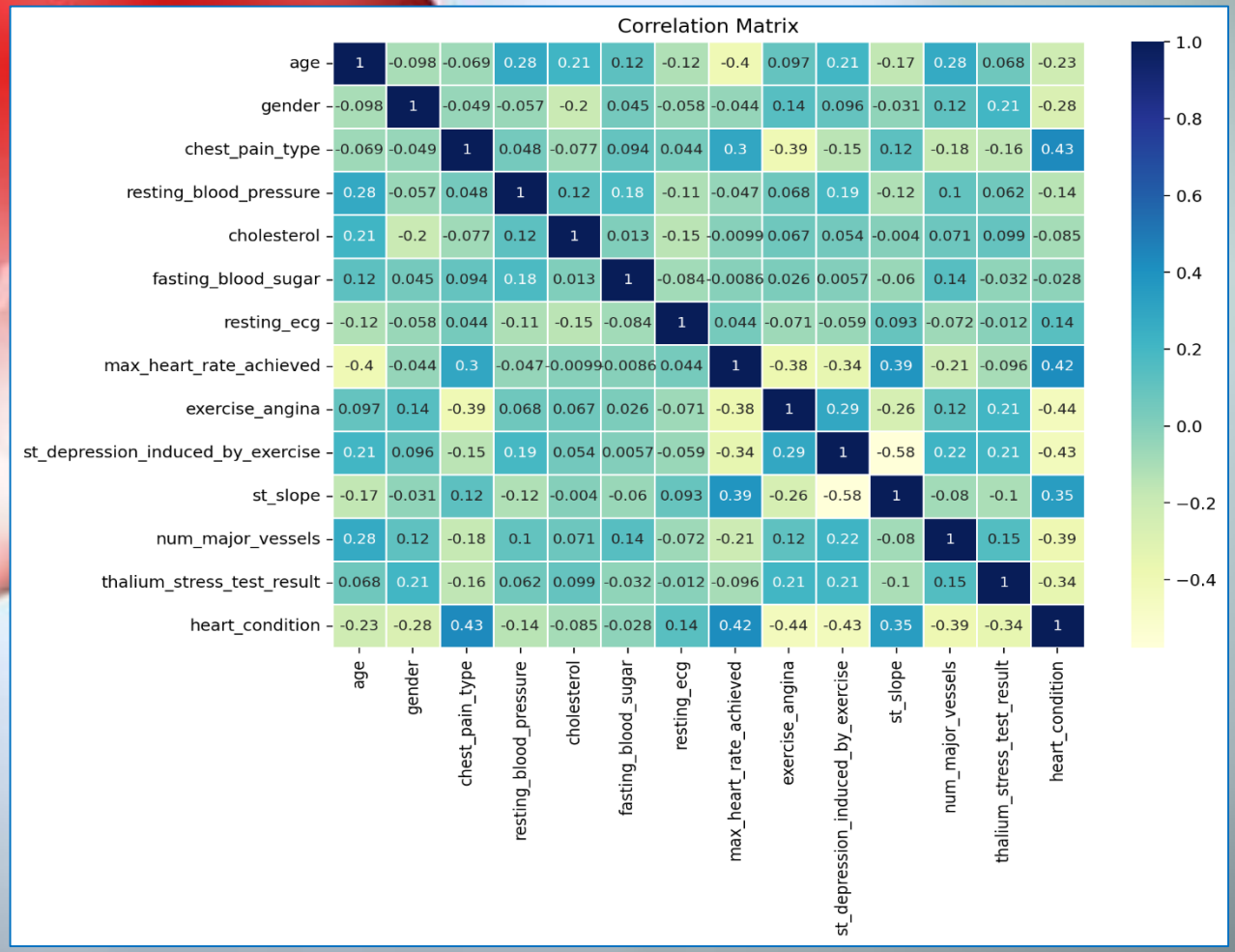
Size: The dataset contains a total of 303 entries, representing individual patients.

Attributes: The data is structured into 14 columns, each corresponding to a specific attribute or health-related test result of the patients.

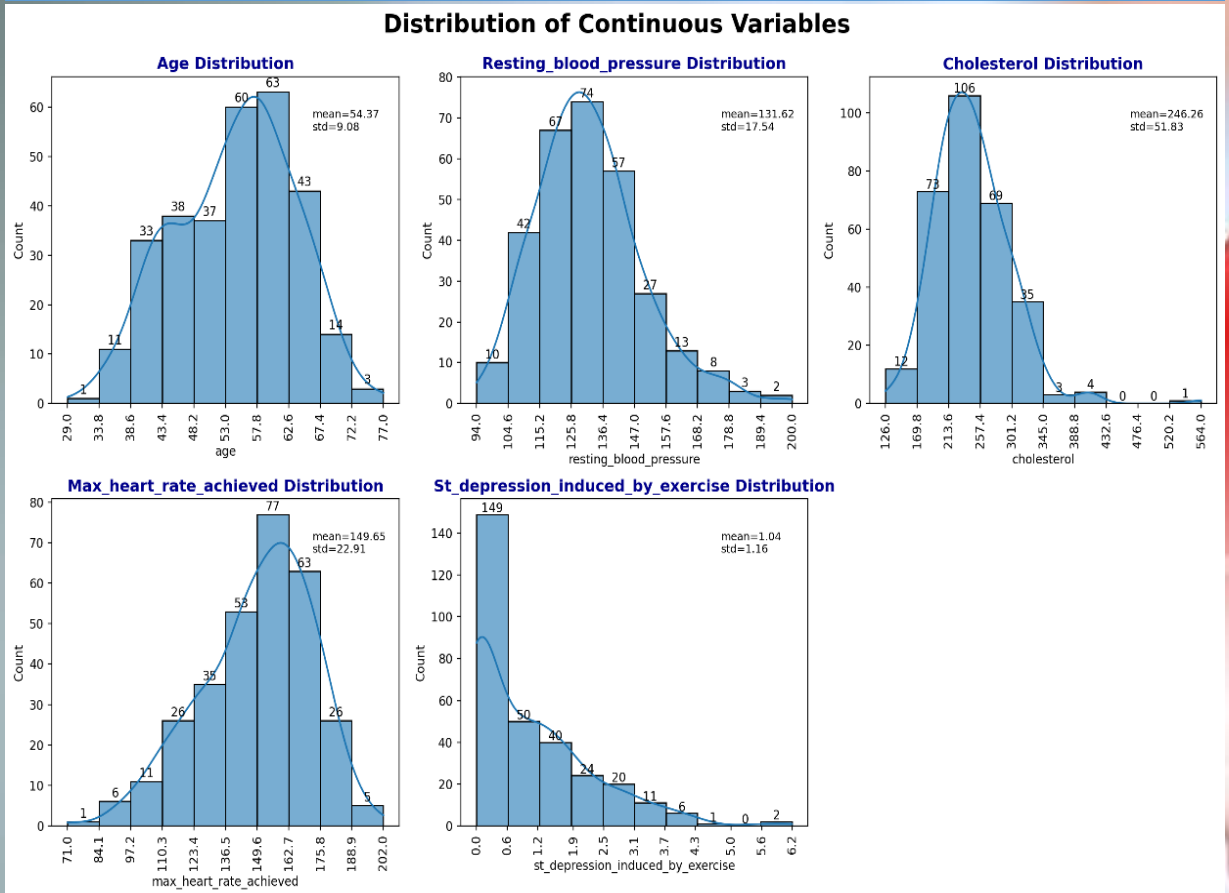
Data Types: The majority of features (13 out of 14) are stored as integers (int64 data type), indicating categorical data values.

Missing Values: Notably, there are no missing values in the dataset. All 303 entries have complete information for each feature.

In the **correlation Matric**, it can be observed The thallium test assists in assessing the risk of heart disease.



UNIVARIATE ANALYSIS FOR NUMERICAL DATA



We examined histograms representing the numerical data (continuous attributes) and cross-referenced them with our existing knowledge of the dataset.

All observations appear typical, without any anomalies or unexpected figures.

UNIVARIATE ANALYSIS FOR CATEGORICAL DATA

"Typical Angina" is the most common chest pain type, making up almost half the data. 85% of patients have normal fasting blood sugar levels.

A large portion (67%) of patients don't experience exercise-induced angina.

These are the two most common categories for peak exercise slope.

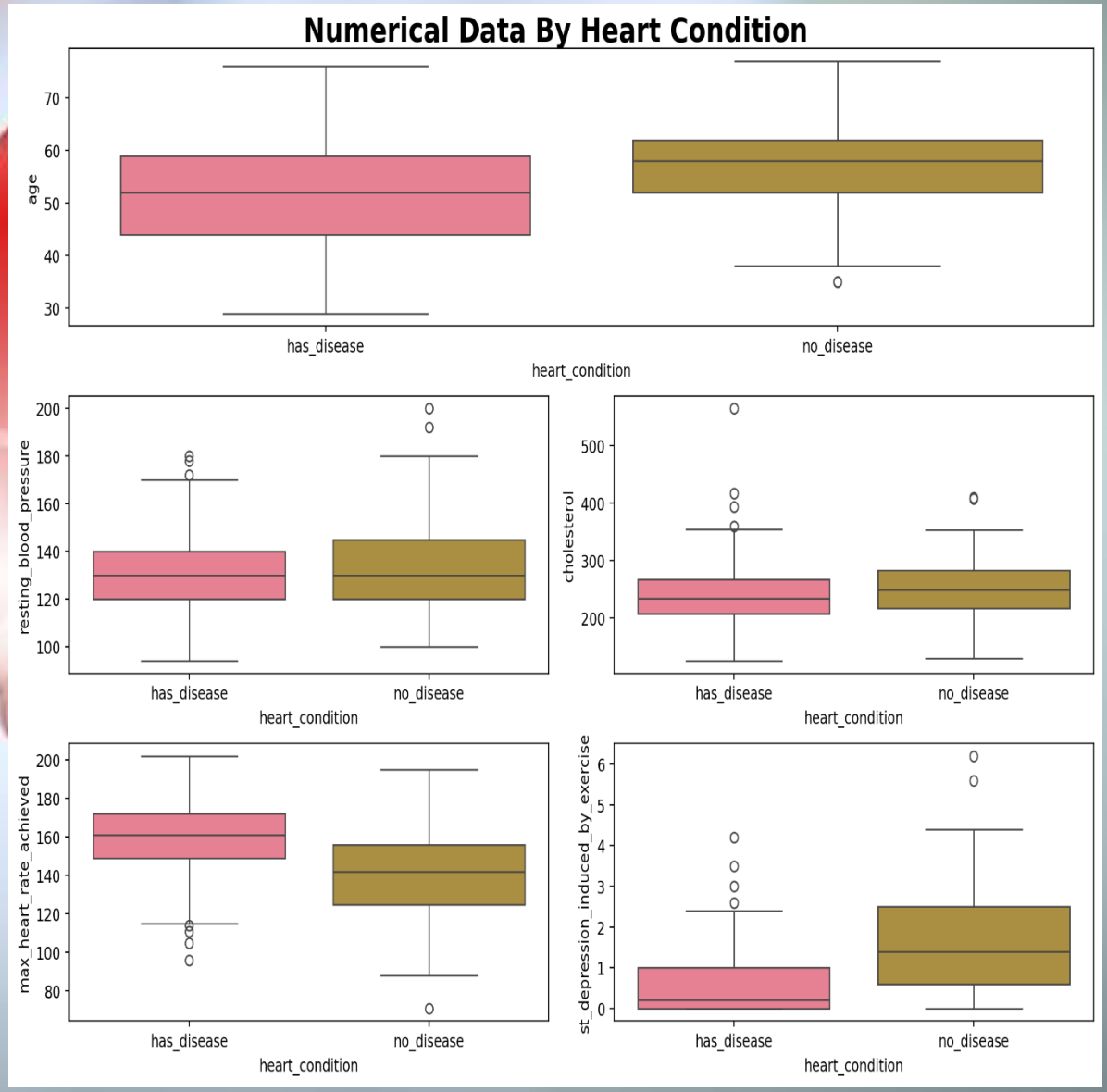
BIVARIATE ANALYSIS FOR NUMERICAL DATA

High blood pressure and cholesterol are both linked to a higher risk of heart disease, although cholesterol's effect seems weaker. Interestingly, there's one person with normal cholesterol but no heart disease, suggesting other factors might be at play.

Maximum heart rate might be less indicative than expected. It's possible that exercise intensity during testing influenced heart rate - people with normal ECGs during exercise might have been pushed harder.

ST depression is a strong indicator of heart disease risk - the more depressed the ST segment, the higher the risk.

Age is also a factor - older people are generally more likely to have heart disease



BIVARIATE ANALYSIS FOR CATEGORICAL DATA

Men are at a much higher risk of heart disease compared to women.

The type of chest pain reported isn't very reliable: "Asymptomatic" chest pain (no pain) is actually the most common type among those with heart disease.

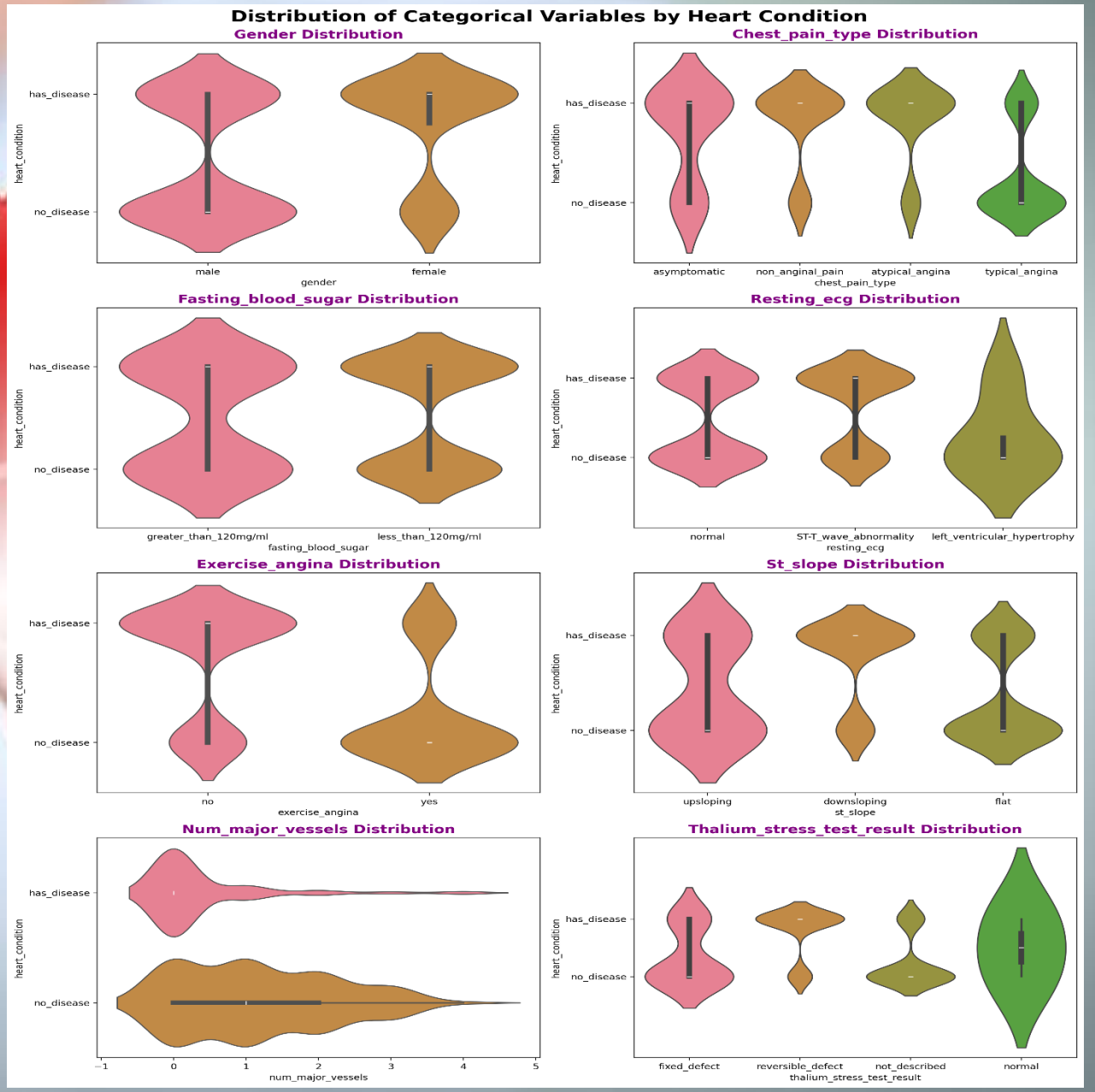
Blood sugar levels don't seem to be directly linked to heart disease.

A normal resting ECG is a good sign, but abnormal ST-T waves increase the risk of heart disease by three times.

Exercise-induced angina is a strong indicator, patients with exercise angina are nearly three times more likely to have heart disease. Additionally, a flat slope during peak exercise test also suggests a higher risk.

The number of major vessels doesn't seem to be a strong indicator, but having none is a positive sign for not having heart disease.

An abnormal thallium test result is a clear warning sign of heart disease.



DATA PREPROCESSING

Outlier Detection

- We identify outliers in numerical features using techniques like Interquartile Range (IQR). Following were the outliers
- Age: 0
- resting_blood_pressure: 9
- cholesterol : 5
- max_heart_rate_achieved: 1
- st_depression_induced_by_exercise: 5

Categorical Encoding

- We use one-hot encoding to transform these categories into numerical features suitable for machine learning models. This allows the model to understand the relationships between these categories and heart disease

Data Splitting

- Splitting the data in two acts: features (X) like age and blood pressure, and the target (y) - has heart disease or not. This data gets split again: training (X_train, y_train) teaches the model, testing (X_test, y_test) checks how well it learned on unseen data.
- X_train shape: (242, 22)
- X_test shape: (61, 22)
- y_train shape: (242,)
- y_test shape: (61,)

THE ML ARCHITECTURE AND PARAMETER

We tackled a classification problem in this analysis. Our goal was to predict whether a patient has heart disease based on various features in the dataset.

Decision Tree Classifier: We initiated our analysis with the Decision Tree Classifier, valued for its interpretability. Decision trees offer a transparent insight into feature importance and their impact on predictions.

Random Forest Classifier: To enhance predictive accuracy and reduce variance, we turned to the Random Forest Classifier. By aggregating multiple decision trees, this ensemble method typically yields superior performance compared to individual trees.

Gradient Boosting Classifier: Leveraging its capability to achieve high accuracy, we incorporated the Gradient Boosting Classifier into our model selection. This technique sequentially constructs an ensemble of models, each refining predictions by focusing on the errors of its predecessors

- CLASSIFICATION REPORT OF DECISSION TREE CLASSIFIER

```
Training Accuracy: 1.0
Testing Accuracy: 0.7540983606557377
Classification Report:
              precision    recall  f1-score   support

  has_disease       0.81      0.69      0.75        32
   no_disease       0.71      0.83      0.76        29

    accuracy                           0.75        61
   macro avg       0.76      0.76      0.75        61
weighted avg       0.76      0.75      0.75        61
```

- CLASSIFICATION REPORT FOR RANDOM FORREST CLASSIFIER

```
Training Accuracy (Random Forest): 1.0
Testing Accuracy (Random Forest): 0.8360655737704918
Classification Report (Random Forest):
              precision    recall  f1-score   support

  has_disease       0.87      0.81      0.84        32
   no_disease       0.81      0.86      0.83        29

    accuracy                           0.84        61
   macro avg       0.84      0.84      0.84        61
weighted avg       0.84      0.84      0.84        61
```

- CLASSIFICATION REPORT FOR GRADIENT BOOSTING CLASSIFIER

```
Training Accuracy (Gradient Boosting): 0.9958677685950413
Testing Accuracy (Gradient Boosting): 0.8524590163934426
Classification Report (Gradient Boosting):
              precision    recall  f1-score   support

  has_disease       0.90      0.81      0.85        32
   no_disease       0.81      0.90      0.85        29

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.86      0.85      0.85        61
```

HYPERPARAMETER TUNING

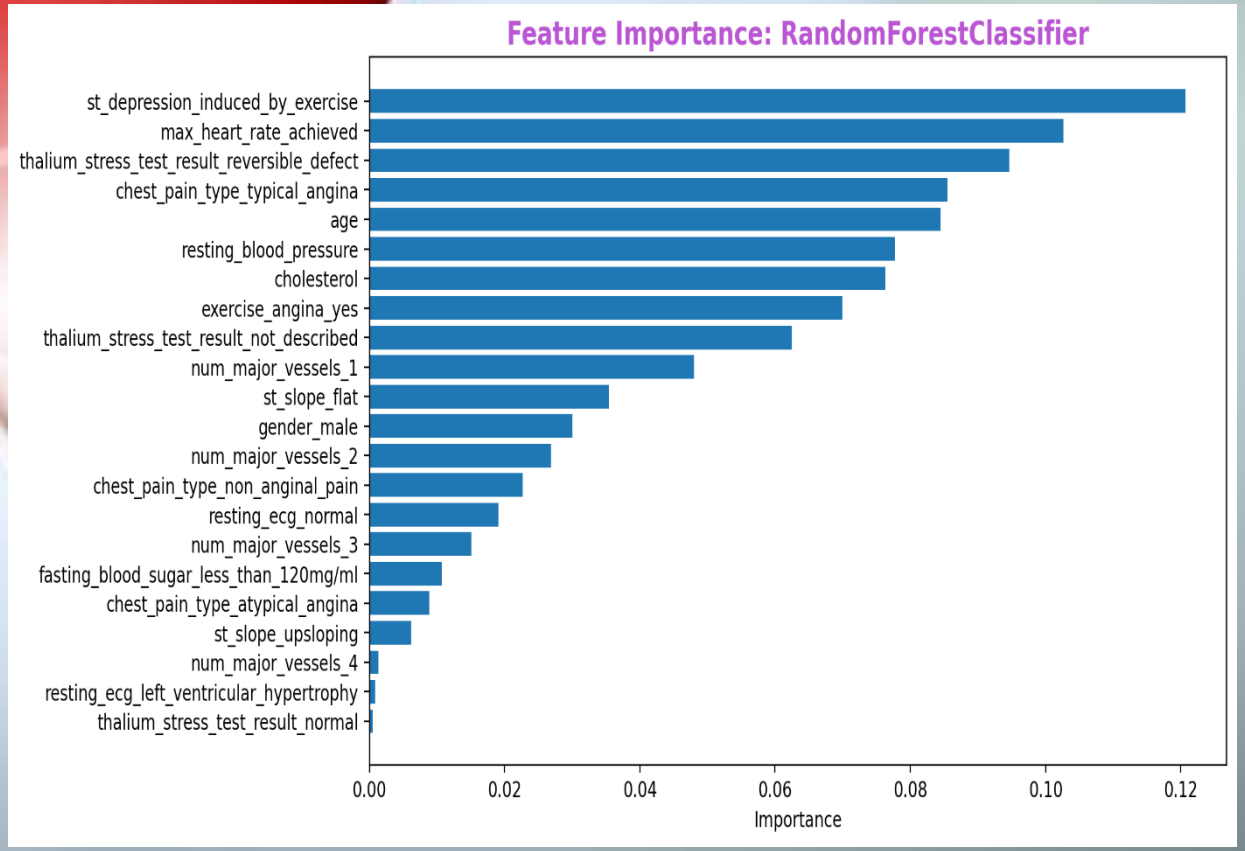
- • To fine-tune the decision tree, we used hyperparameter tuning. This method explores various parameter combinations and evaluates the model's performance for each.
- • We focused on adjusting key parameters that influence the tree's complexity and structure:
- max_depth: Limits tree growth, preventing overfitting with overly complex decision boundaries.
- min_samples_split: Sets the minimum samples required to split a node, impacting the balance between overfitting and underfitting.
- min_samples_leaf: Determines the minimum samples allowed in a leaf node, helping prevent overly specific decision rules.

```
Best Parameters: {'max_depth': 7, 'min_samples_leaf': 1, 'min_samples_split': 10}
Testing Accuracy (Tuned Model): 0.7868852459016393
Testing Accuracy (Tuned Model): 0.7868852459016393
Classification Report (Tuned Model):
              precision    recall  f1-score   support

  has_disease       0.85      0.72      0.78        32
   no_disease       0.74      0.86      0.79        29

    accuracy                           0.79        61
   macro avg       0.79      0.79      0.79        61
weighted avg       0.80      0.79      0.79        61
```

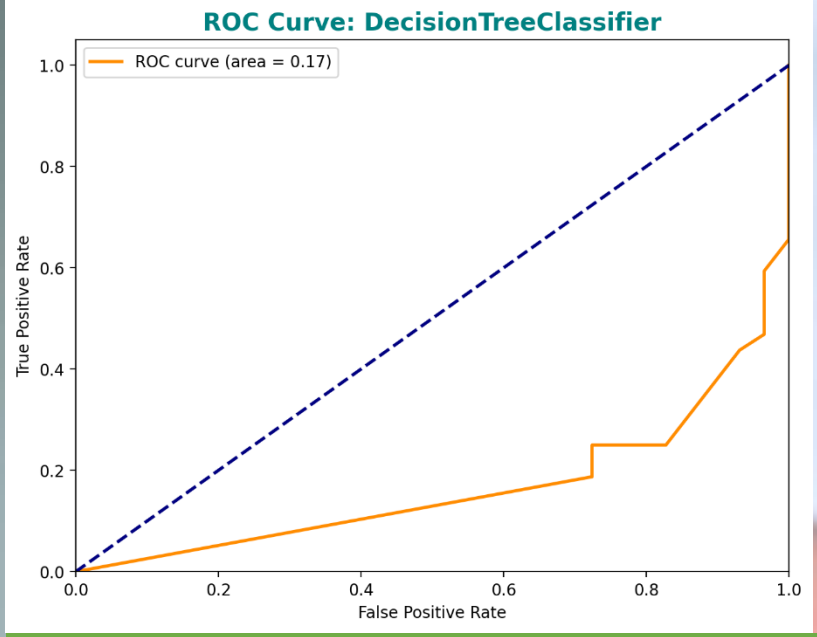
RESULTS AND EVALUATION

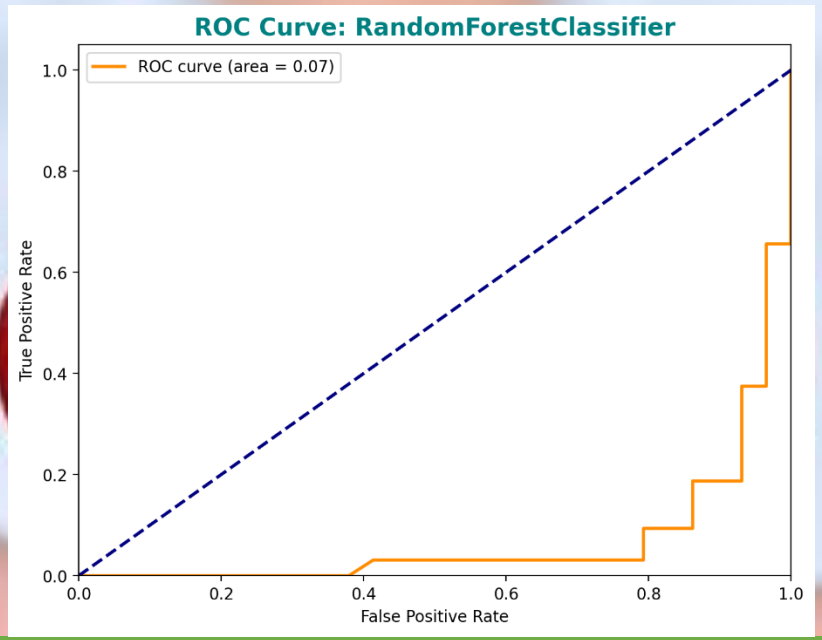
Initial analysis suggests that the Random Forest Classifier performed best among untuned models, followed by the Gradient Boosting Classifier.

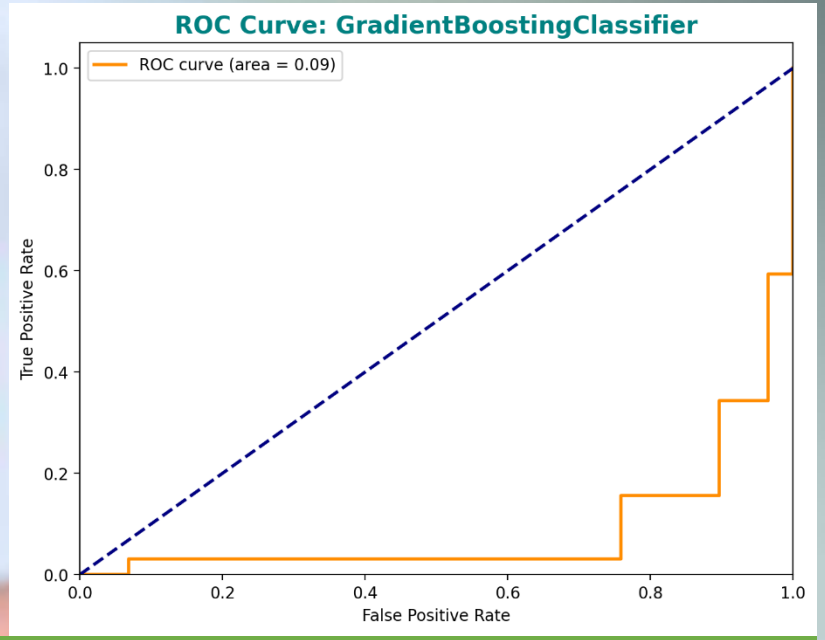
After applying the tuning techniques in decision tree, it can be seen that model is not overfitting.



LIMITATION OF THE MODELS







- The analysis revealed challenges in accurately predicting heart disease. All three models (decision tree, random forest, and gradient boosting) achieved low AUC scores on the ROC curves.There are a couple of reasons why this might be happening:
- Model limitations: Decision trees are known to be susceptible to overfitting, particularly with intricate datasets.
- Hyperparameter tuning: Random forests and gradient boosting might benefit from further refinement of their hyperparameters or exploration of alternative model architectures to achieve optimal performance.

POTENTIAL AREAS OF IMPROVEMENTS

- Hyperparameter Tuning: We can delve deeper into fine-tuning the hyperparameters specifically for Random Forest and Gradient Boosting models. This optimization process can potentially unlock their full potential and enhance their effectiveness in predicting heart disease.
- Exploring Alternative Models: Since the current models exhibited limitations, we can investigate the use of different machine learning algorithms. Support Vector Machines (SVMs) and deep learning models are promising avenues to explore. Their capabilities might be better suited to handle the complexities of this heart disease prediction task.