Assignment 2 Report: Machine Learning Using Trees

Created by: Shraddha Gangaram and Tanvi Nair

Heart Disease Dataset: https://archive.ics.uci.edu/dataset/45/heart+disease

Objective: The objective of this project is to analyze the Heart Disease dataset, perform necessary data preprocessing and exploratory data analysis (EDA), and then build, tune, and evaluate four different classification models to predict the presence of heart disease:

1. **Decision Tree (DT):** A single tree model implementation for classification, used as a baseline.
2. **Random Forest (RF):** A bagging ensemble method that aggregates multiple decision trees to reduce variance.
3. **AdaBoost (AB):** A boosting ensemble method that sequentially builds weak learners to correct previous errors.
4. **XGBoost (XGB):** An optimized, distributed gradient boosting library used for high-performance classification.

Dataset Introduction: The dataset used in this study is the Processed Cleveland Heart Disease dataset from the UCI Machine Learning Repository. It contains 303 observations, each described by 13 clinical attributes. The goal is to predict the presence of heart disease in patients.

Attributes:

- **age:** Age in years
- **sex:** Sex (1 = male; 0 = female)
- **cp:** Chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic)
- **trestbps:** Resting blood pressure (in mm Hg)
- **chol:** Serum cholesterol in mg/dl
- **fbs:** Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- **restecg:** Resting electrocardiographic results (0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy)
- **thalach:** Maximum heart rate achieved
- **exang:** Exercise induced angina (1 = yes; 0 = no)
- **oldpeak:** ST depression induced by exercise relative to rest
- **slope:** The slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
- **ca:** Number of major vessels (0-3) colored by fluoroscopy
- **thal:** Thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect)

Initial Data Structure: A preview of the raw dataset shows the structure and initial values of the attributes upon loading, including the presence of categorical codes and the target variable's initial multi-class format:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4

	slope	ca	thal	target
0	3.0	0.0	6.0	0
1	2.0	3.0	3.0	2
2	2.0	2.0	7.0	1
3	3.0	0.0	3.0	0
4	1.0	0.0	3.0	0

Target Variable: Diagnosis of heart disease. The original values (1-4) were converted to a binary format: 0 for no disease and 1 for the presence of heart disease.

Data Cleansing and Pre-Processing: Before analysis, several preprocessing steps were performed to ensure data quality and model compatibility:

- **Handled Missing Values:** Missing values in the 'ca' and 'thal' columns, originally marked with '?', were imputed using the mode of each respective column.
- **Removed Duplicates:** Redundant rows were dropped from the dataset.
- **Target Variable Conversion:** The original multi-class target variable (0, 1, 2, 3, 4) was successfully converted into a binary format: 0 for no disease and 1 for the presence of heart disease. As shown below, this process resulted in a relatively balanced distribution for the final classification task: 164 instances of 'No Disease' (0) and 139 instances of 'Disease Present' (1).

```
Value counts for the binary target variable:
target
0    164
1    139
Name: count, dtype: int64
```

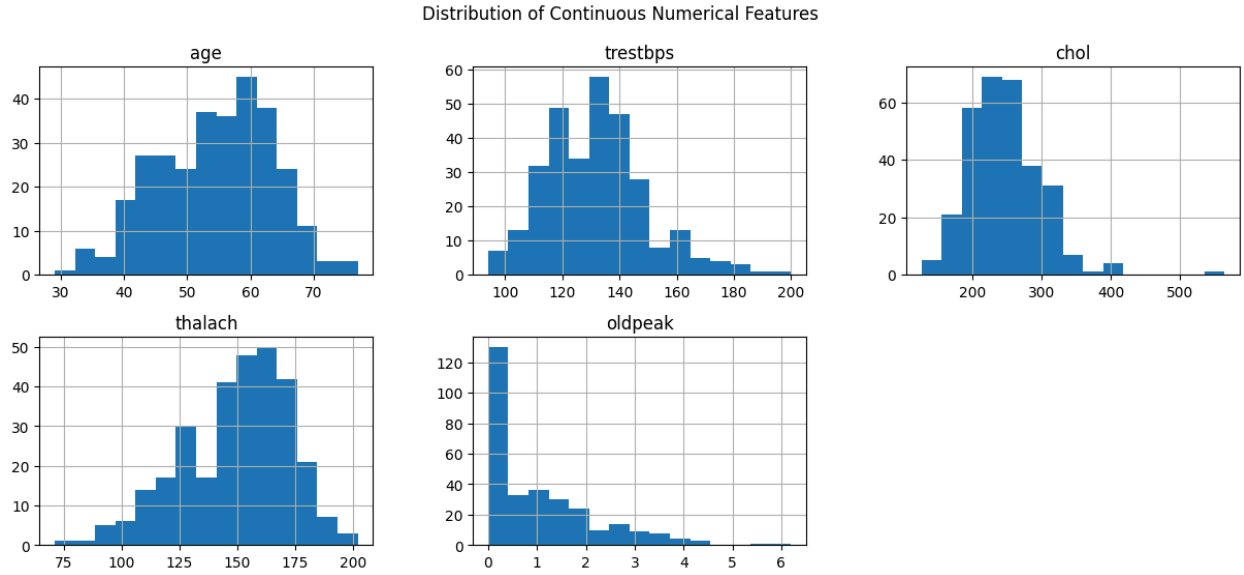
- **One-Hot Encoding:** Categorical variables (sex, cp, fbs, restecg, exang, slope, ca, thal) were converted into numerical format using one-hot encoding.
- **Data Splitting:** The dataset was split into training (80%) and testing (20%) sets.
- **Standardization & Normalization:** Numerical features (age, trestbps, chol, thalach, oldpeak) were standardized. Subsequently, all features in the training and testing sets were L2-normalized.

Attribute Summary: A descriptive summary was generated for all attributes before splitting and normalization.

Statistical Summary of All Attributes:

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.438944	9.038662	29.0	48.0	56.0	61.0	77.0
sex	303.0	0.679868	0.467299	0.0	0.0	1.0	1.0	1.0
cp	303.0	3.158416	0.960126	1.0	3.0	3.0	4.0	4.0
trestbps	303.0	131.689769	17.599748	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.693069	51.776918	126.0	211.0	241.0	275.0	564.0
fbs	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.990099	0.994971	0.0	0.0	1.0	2.0	2.0
thalach	303.0	149.607261	22.875003	71.0	133.5	153.0	166.0	202.0
exang	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
slope	303.0	1.600660	0.616226	1.0	1.0	2.0	2.0	3.0
ca	303.0	0.663366	0.934375	0.0	0.0	0.0	1.0	3.0
thal	303.0	4.722772	1.938383	3.0	3.0	3.0	7.0	7.0
target	303.0	0.458746	0.499120	0.0	0.0	0.0	1.0	1.0

Attribute Distributions: Histograms were created for all continuous numerical features to visualize their distributions.



Histogram Analysis: Based on the histograms provided, the continuous numerical attributes in the Heart Disease dataset are generally not normally distributed. The features age and thalach (maximum heart rate achieved) appear the most symmetric and closest to a normal distribution, though even they show slight skewness. In contrast, trestbps (resting blood pressure) and chol (serum cholesterol) both exhibit a distinct right-skew, meaning the bulk of the observations are clustered at the lower-to-middle values, with a long tail extending toward higher values. The feature oldpeak (ST depression induced by exercise relative to rest) displays the most severe deviation, showing a concentration (or "floor effect") near zero, resulting in a highly right-skewed distribution. The non-normal nature of these attributes is typically attributed to the natural skewness of biological metrics in a population and the sample selection bias inherent in clinical datasets, which tend to over-represent individuals presenting with symptoms, moving the distribution away from a general population mean.

Data Preparation and Feature Selection Outcomes: After cleaning, the final dataset was split into the Training Set (242 samples) and the Testing Set (61 samples). All subsequent modeling and feature selection steps were performed exclusively on the training data. The features underwent two crucial scaling steps: first, the numerical columns were standardized (z-score scaling), and then the entire feature set was subjected to L2 Normalization. This normalization step converts each row into a unit vector, fundamentally altering the magnitude of feature values to ensure uniform scaling before feature selection.

For example, a sample of the normalized training features shows the scaled magnitude of the chosen predictors:

Training features after standardization and normalization:

	age	trestbps	chol	thalach	oldpeak	sex_1.0	cp_2.0	\
132	-0.670167	-0.029744	-0.204023	0.546443	-0.206249	0.236099	0.236099	
202	0.065168	0.263166	-0.670607	0.275745	-0.190318	0.270010	0.000000	
196	0.467158	0.456291	-0.072188	-0.254843	-0.236143	0.299213	0.000000	
75	0.334105	0.454378	0.707607	0.011969	-0.059204	0.000000	0.000000	
176	-0.112004	-0.485102	-0.095092	-0.050157	-0.286418	0.362915	0.000000	

	cp_3.0	cp_4.0	fbs_1.0	restecg_1.0	restecg_2.0	exang_1.0	\
132	0.000000	0.000000	0.000000	0.0	0.236099	0.0	
202	0.270010	0.000000	0.270010	0.0	0.000000	0.0	
196	0.000000	0.000000	0.299213	0.0	0.299213	0.0	
75	0.297959	0.000000	0.000000	0.0	0.297959	0.0	
176	0.000000	0.362915	0.362915	0.0	0.000000	0.0	

	slope_2.0	slope_3.0	ca_1.0	ca_2.0	ca_3.0	thal_6.0	thal_7.0
132	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.000000
202	0.000000	0.0	0.270010	0.0	0.000000	0.0	0.270010
196	0.299213	0.0	0.299213	0.0	0.000000	0.0	0.000000
75	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.000000
176	0.000000	0.0	0.000000	0.0	0.362915	0.0	0.362915

Testing features after standardization and normalization:

	age	trestbps	chol	thalach	oldpeak	sex_1.0	cp_2.0	\
179	-0.075828	-0.048095	0.003013	0.389871	-0.333496	0.381761	0.0	
228	-0.025092	-0.347262	-0.232887	-0.531495	-0.247313	0.283105	0.0	
111	0.048697	-0.148710	0.026018	-0.100831	0.051433	0.370718	0.0	
246	0.123978	-0.627018	-0.085132	0.092854	-0.278486	0.352865	0.0	
60	-0.172917	-0.052039	0.509268	-0.149189	0.057309	0.000000	0.0	

	cp_3.0	cp_4.0	fbs_1.0	restecg_1.0	restecg_2.0	exang_1.0	\
179	0.381761	0.000000	0.381761	0.0	0.381761	0.000000	
228	0.000000	0.283105	0.000000	0.0	0.283105	0.283105	
111	0.000000	0.370718	0.370718	0.0	0.370718	0.370718	
246	0.000000	0.352865	0.000000	0.0	0.000000	0.000000	
60	0.000000	0.413067	0.000000	0.0	0.000000	0.413067	

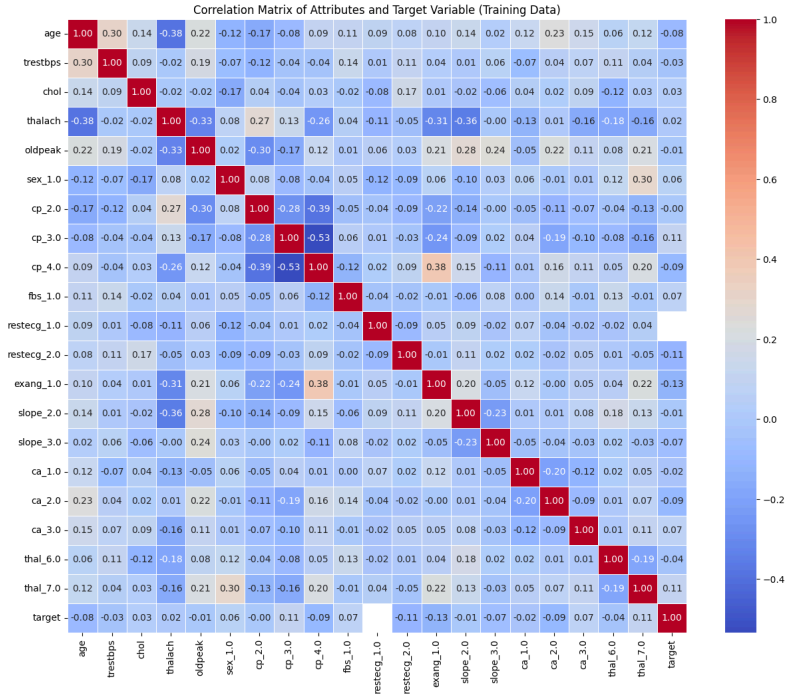
	slope_2.0	slope_3.0	ca_1.0	ca_2.0	ca_3.0	thal_6.0	thal_7.0
179	0.000000	0.0	0.000000	0.0	0.381761	0.0	0.000000
228	0.283105	0.0	0.283105	0.0	0.000000	0.0	0.000000
111	0.370718	0.0	0.370718	0.0	0.000000	0.0	0.000000
246	0.000000	0.0	0.352865	0.0	0.000000	0.0	0.352865
60	0.413067	0.0	0.000000	0.0	0.000000	0.0	0.413067

Correlation of Attributes with Target Variable: The correlation coefficients were calculated on the normalized training data to quantify the linear relationship between each attribute and the binary target variable (Disease Status). This analysis is crucial for guiding feature selection. The

results, ranked by correlation magnitude, confirm that all linear relationships are **weak** following the L2 normalization, with absolute values below 0.14:

```
Correlation with the target variable (Training Data)
target          1.000000
cp_3.0          0.113077
thal_7.0        0.107020
fbs_1.0         0.069512
ca_3.0          0.069252
sex_1.0         0.063513
chol            0.028528
thalach         0.015849
cp_2.0         -0.002083
oldpeak        -0.011053
slope_2.0      -0.011834
ca_1.0         -0.020957
trestbps       -0.029684
thal_6.0       -0.042350
slope_3.0      -0.065714
age            -0.080177
ca_2.0         -0.090906
cp_4.0         -0.091044
restecg_2.0    -0.111719
exang_1.0      -0.130464
restecg_1.0          NaN
```

Correlation Matrix/Heatmap: A color-coded visualization of the correlation matrix, highlighting strong positive and negative relationships was created, which is as follows:



Correlation Matrix/Heatmap Analysis: The correlation matrix provides valuable insights into the dataset's structure, revealing both the strongest feature interactions and the predictive weakness of individual features. Analysis shows that severe multicollinearity (strong correlation between two predictors) isn't a major issue, as most correlations are moderate, with the highest values (≈0.53) logically occurring between the dummy variables of the same category (like different chest pain types). Regarding predictive strength, the most notable finding is the weak overall linear relationship between the attributes and the target variable; all correlation magnitudes are below 0.54. This confirms that no single feature acts as a dominant linear predictor, underscoring the necessity of using complex, non-linear ensemble methods (like Random Forest or XGBoost) to find the combined, interacting signals required for accurate classification.

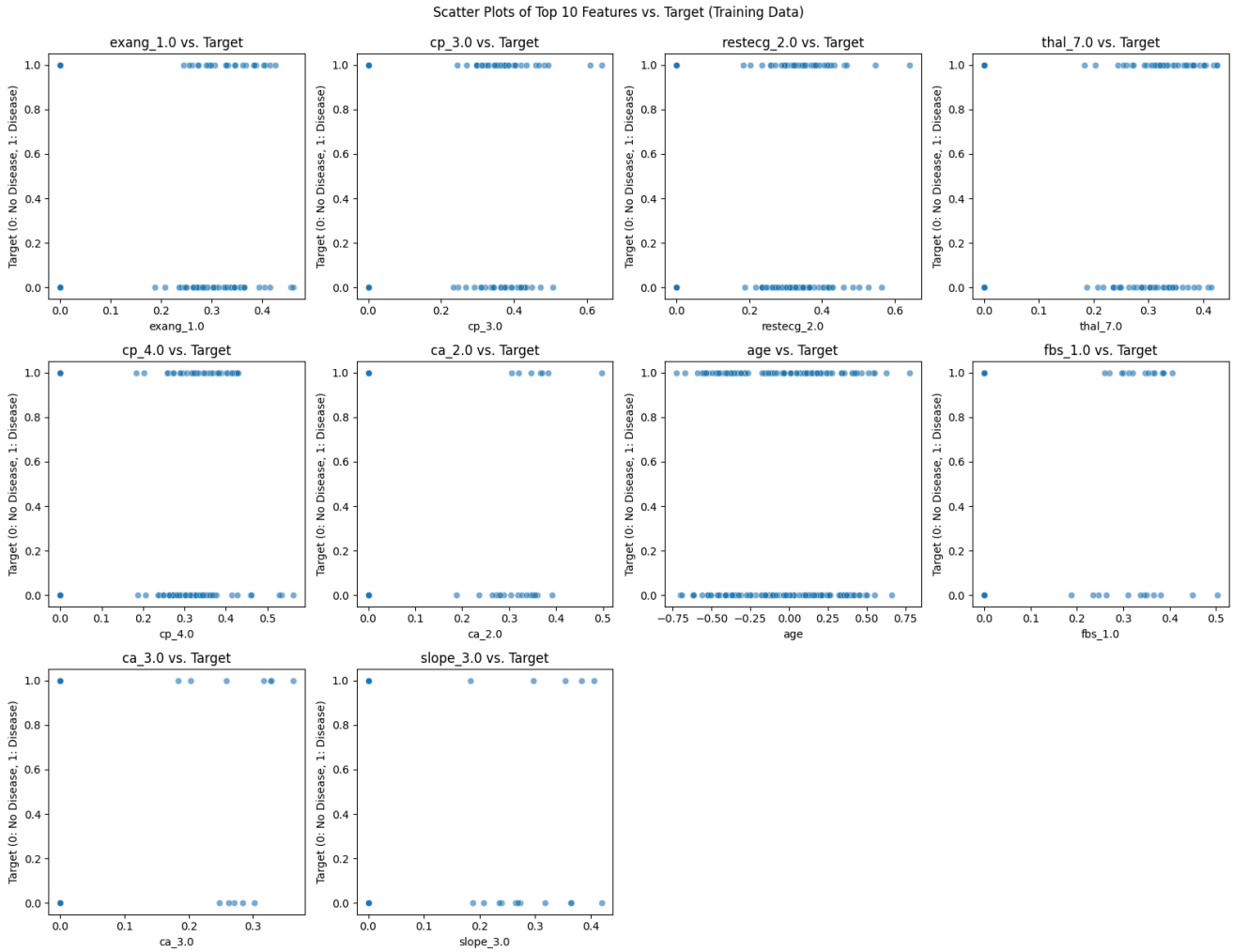
Feature selection: The correlation of each attribute with the target variable was calculated on the normalized training data. The top 10 features with the highest absolute correlation were selected for model building to reduce dimensionality and focus on the most influential predictors.

Selected Features for Modeling:

```
Selected important attributes based on Training Data (Top 10):
['exang_1.0', 'cp_3.0', 'restecg_2.0', 'thal_7.0', 'cp_4.0', 'ca_2.0', 'age', 'fbs_1.0', 'ca_3.0', 'slope_3.0']
```

- exang_1.0 (Exercise Induced Angina)
- cp_3.0 (Non-anginal Pain)
- restecg_2.0 (Left Ventricular Hypertrophy)
- thal_7.0 (Reversible Defect Thalassemia)
- cp_4.0 (Asymptomatic Chest Pain)
- ca_2.0 (2 Major Vessels)
- age
- fbs_1.0 (Fasting Blood Sugar > 120)
- ca_3.0 (3 Major Vessels)
- slope_3.0 (Downsloping ST Segment)

Scatterplot: Furthermore, we generated scatterplots for the top 10 features based on training data correlation, which is as follows:



Scatterplot Analysis: The scatter plots visually confirm the weak predictive power of individual features after the normalization process. Because the target variable is binary (0 or 1), the data points for both 'No Disease' and 'Disease' are heavily overlapped across the entire range of every feature's X-axis values. This extreme overlap is significant because it clearly indicates that no single feature alone possesses the necessary separation power to reliably classify a patient. This observation strongly validates the low numerical correlation coefficients found earlier and underscores the necessity of relying on complex, non-linear ensemble models, such as Random Forest and XGBoost, which are specifically designed to effectively combine these weak, interacting signals for classification.

Final Predictor Set for Modeling: The above selection process resulted in the following optimized input data being fed directly into the four classification models. A sample of the Selected and Normalized Training Features confirms the final state of the predictors:

```
Training features with selected attributes:
     exang_1.0    cp_3.0  restecg_2.0  thal_7.0    cp_4.0  ca_2.0       age  \
132        0.0  0.000000     0.236099  0.000000  0.000000     0.0 -0.670167
202        0.0  0.270010     0.000000  0.270010  0.000000     0.0  0.065168
196        0.0  0.000000     0.299213  0.000000  0.000000     0.0  0.467158
75         0.0  0.297959     0.297959  0.000000  0.000000     0.0  0.334105
176        0.0  0.000000     0.000000  0.362915  0.362915     0.0 -0.112004

      fbs_1.0    ca_3.0  slope_3.0
132  0.000000  0.000000        0.0
202  0.270010  0.000000        0.0
196  0.299213  0.000000        0.0
75   0.000000  0.000000        0.0
176  0.362915  0.362915        0.0
```

```
Testing features with selected attributes:
     exang_1.0    cp_3.0  restecg_2.0  thal_7.0    cp_4.0  ca_2.0       age  \
179   0.000000  0.381761     0.381761  0.000000  0.000000     0.0 -0.075828
228   0.283105  0.000000     0.283105  0.000000  0.283105     0.0 -0.025092
111   0.370718  0.000000     0.370718  0.000000  0.370718     0.0  0.048697
246   0.000000  0.000000     0.000000  0.352865  0.352865     0.0  0.123978
60    0.413067  0.000000     0.000000  0.413067  0.413067     0.0 -0.172917

      fbs_1.0    ca_3.0  slope_3.0
179  0.381761  0.381761        0.0
228  0.000000  0.000000        0.0
111  0.370718  0.000000        0.0
246  0.000000  0.000000        0.0
60   0.000000  0.000000        0.0
```

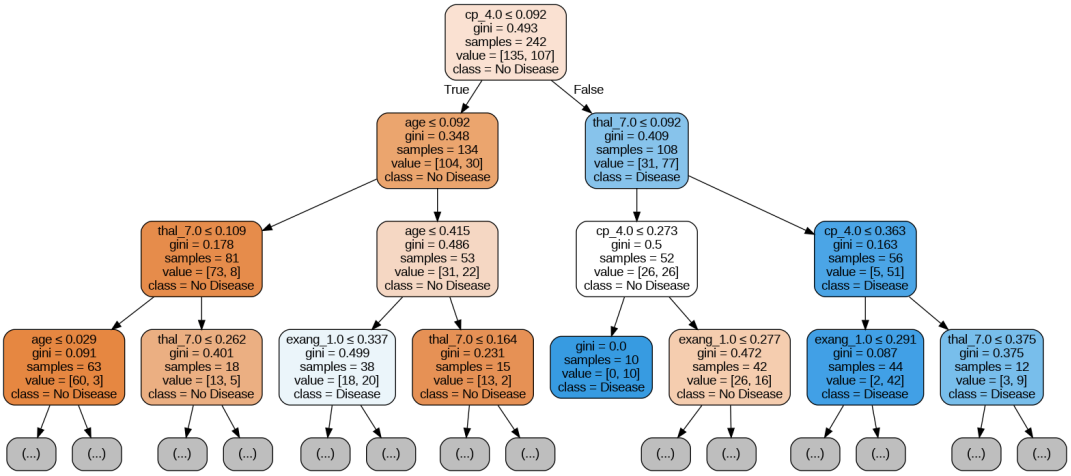
Model Construction and Evaluation: Four classification models were built using the selected features shown above. Furthermore, hyperparameters for each model were tuned using GridSearchCV with 5-fold cross-validation to optimize performance.

Decision Tree Classifier Analysis:

```
Best parameters found for Decision Tree:
{'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 10}

Decision Tree Model Performance:
Accuracy: 0.8033
Precision: 0.8333
Recall: 0.7812
F1-score: 0.8065

Confusion Matrix:
[[24  5]
 [ 7 25]]
```

- **Hyperparameter Analysis:**
    - **Rationale:** A wide range of max_depth, split, and leaf constraints were tested to find the best balance between model complexity and generalization, ensuring the final model was not severely underfit or overfit to the training data.
    - **Best Values Found:** Criterion: gini, Max Depth: None, Min Samples Split: 10, Min Samples Leaf: 2.

    - **Conclusion:** The search successfully identified constraints that balance complexity. The optimal max_depth = None suggests the model required the full complexity available within the other constraints.
- **Metric Analysis:** The Decision Tree achieved an F1-score of 0.8065, showing a fair balance between Precision (0.8333) and Recall (0.7812). Precision is slightly higher, indicating the model is better at ensuring its positive predictions are correct than it is at catching all true positives.
- **Confusion Matrix Analysis:** The matrix [[24 5] [7 25]] shows 24 true negatives (correctly predicted no disease) and 25 true positives (correctly predicted disease). The 7 False Negatives (FN) are the most critical errors in this medical context. The model failed to detect disease in 7 patients.
- To visually inspect the model's high-variance structure, the final, best-tuned tree is presented below:

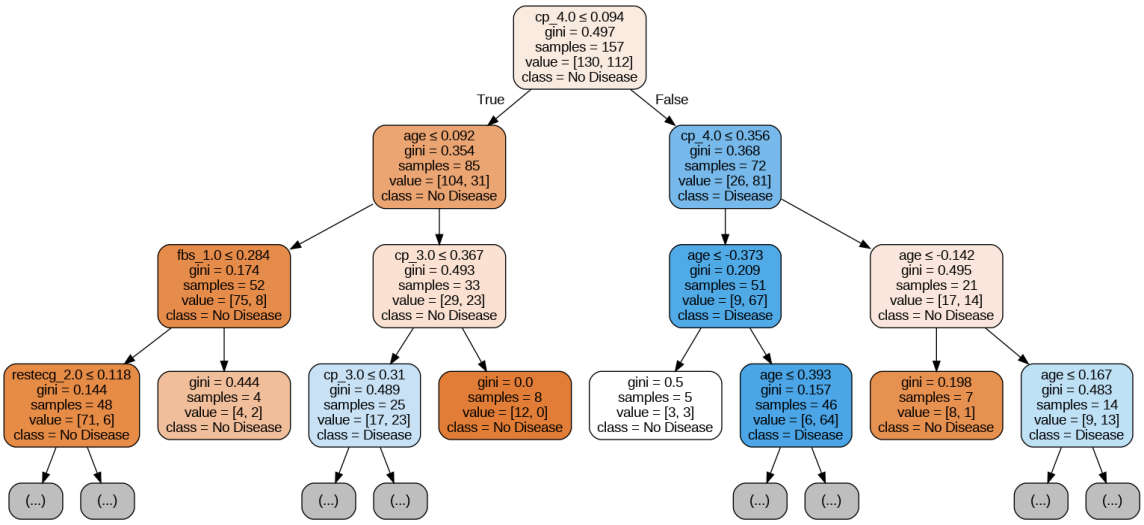


Random Forest Classifier Analysis:

```
Best parameters found for Random Forest:
{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200}

Random Forest Model Performance:
Accuracy: 0.8852
Precision: 0.8788
Recall: 0.9062
F1-score: 0.8923

Confusion Matrix:
[[25  4]
 [ 3 29]]
```

- **Hyperparameter Analysis:**
    - **Rationale:** Focused on 'n_estimators' and 'max_depth' as primary drivers, testing enough trees (up to 200) to stabilize ensemble variance.
    - **Best Values Found:** N Estimators: 200, Max Depth: 10, Min Samples Leaf: 4, Criterion: gini.
    - **Conclusion:** The optimal value for N_Estimators was found at the edge of the search range, but 200 was deemed sufficient for variance reduction, achieving one of the highest F1-scores.

- **Metrics Analysis:** The RF achieved a high F1-score of 0.8923, demonstrating a substantial improvement over the DT model. This improvement stems from a strong balance: Precision (0.8788) and Recall (0.9062). High Recall means the model is very effective at identifying true disease cases.
- **Confusion Matrix Analysis:** The matrix [[25 4] [3 29]] shows low error rates. Crucially, the model reduced False Negatives (FN) to only 3, making it much safer in a medical diagnostic context compared to the DT's 7 FNs.
- The internal complexity of the ensemble's base learners is shown by visualizing the structure of the first estimator:

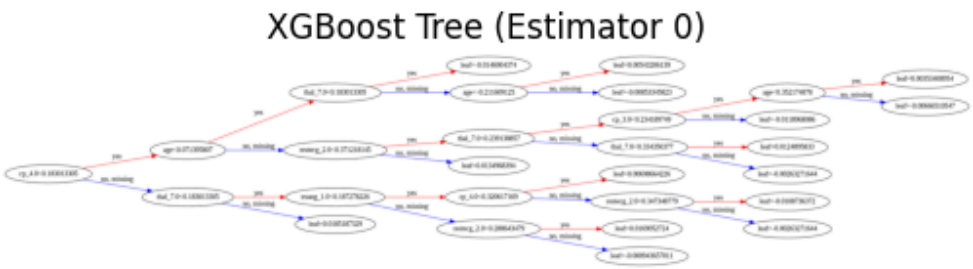


XGBoost Analysis:

```
Best parameters found for XGBoost:
{'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': 7, 'n_estimators': 100, 'subsample': 0.8}

XGBoost Model Performance:
Accuracy: 0.8852
Precision: 0.9032
Recall: 0.8750
F1-score: 0.8889

Confusion Matrix:
[[26  3]
 [ 4 28]]
```

- **Hyperparameter Analysis:**
  - **Rationale:** Performed a wide search on critical parameters to control the bias-variance trade-off.
  - **Best Values Found:** N Estimators: 100, Learning Rate: 0.01, Max Depth: 7, Subsample: 0.8, Colsample_bytree: 1.0, Gamma: 0.
  - **Conclusion:** XGBoost performs best with a very conservative shrinkage (Learning Rate=0.01) combined with moderately deep trees, prioritizing stability and precision.
- **Metrics Analysis:** XGBoost achieved a very high F1-score of 0.8889. It prioritized Precision (0.9032) over Recall (0.8750), suggesting a focus on making highly accurate positive predictions. This priority is a characteristic of regularization in boosting.

- **Confusion Matrix Analysis:** The matrix [[26 3] [4 28]] shows the highest number of True Negatives (26), meaning it is very accurate when predicting ‘no disease’. Its False Negatives (4) are slightly higher than Random Forest, confirming its slightly lower Recall compared to Random Forest.
- For insight into how the boosting process operates, the structure of the first sequentially built tree is displayed here

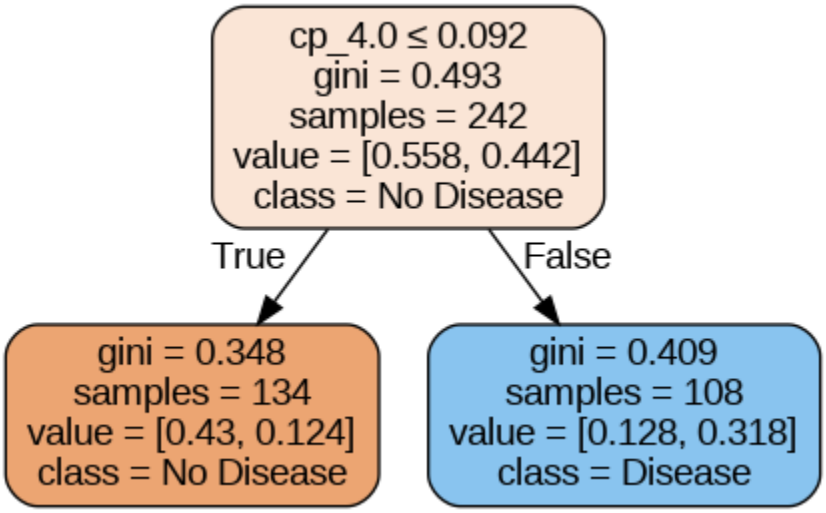


Adaboost Analysis:

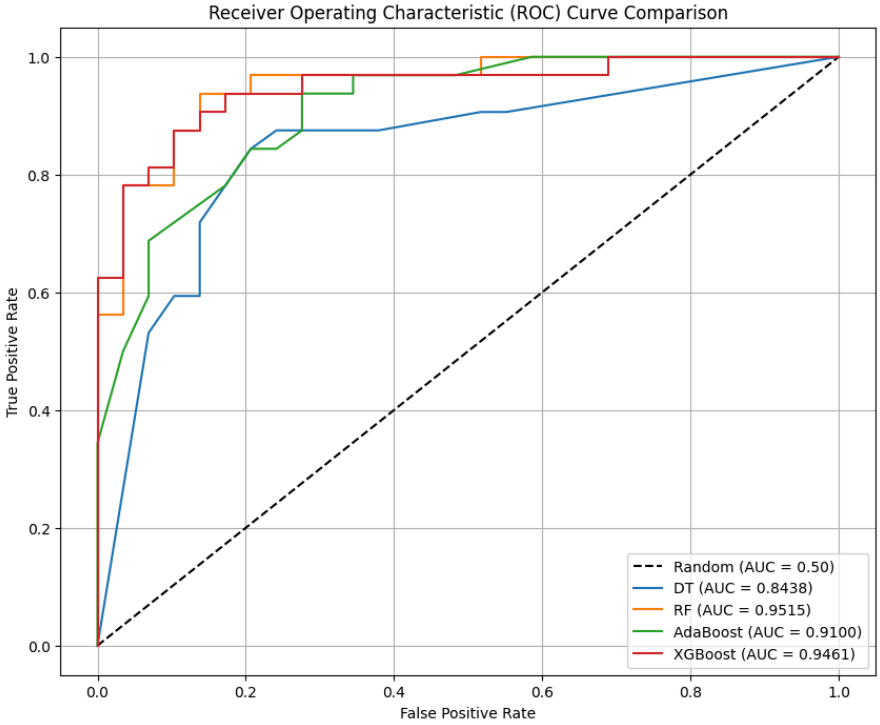
```
Best parameters found for AdaBoost:
{'learning_rate': 0.1, 'n_estimators': 100}

AdaBoost Model Performance:
Accuracy: 0.8033
Precision: 0.7941
Recall: 0.8438
F1-score: 0.8182

Confusion Matrix:
[[22  7]
 [ 5 27]]
```

- **Hyperparameter Analysis:**
    - **Rationale:** Tuned 'n_estimators' and 'learning_rate', fixing the base estimator to a Decision Stump.
    - **Best Values Found:** N Estimators: 100, Learning Rate: 0.1.
    - **Conclusion:** These values maximize the model's ability to correct previous errors without overfitting the training examples too quickly, and the optimal values were well within the tested ranges.
- **Metrics Analysis:** AdaBoost achieved an F1-score of 0.8182, placing its performance just above the single Decision Tree. It shows strong Recall (0.8438) but lower Precision (0.7941), indicating it is more aggressive in predicting disease but incurs more False Positives.
- **Confusion Matrix Analysis:** The matrix [[22 7] [5 27]] shows 7 False Positives (FP), confirming the lower precision, and 5 False Negatives (FN), placing its critical error rate between the DT (7 FN) and the superior ensemble methods (RF/XGB).

- The reliance of the AdaBoost ensemble on shallow learners is confirmed by visualizing its base estimator, which acts as a simple decision stump, as shown below:

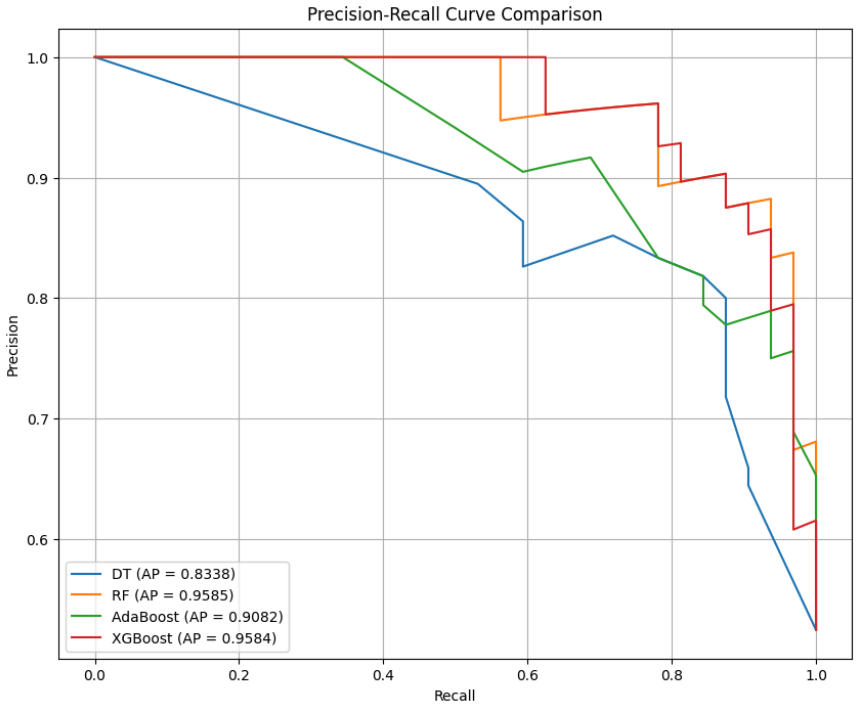


Comparative Analysis: ROC and Precision-Recall curves were generated to compare the models' performance across different classification thresholds.



Receiver Operating Characteristic Curve Analysis: The single Decision Tree (DT, AUC 0.8438) provides the lowest baseline performance. Random Forest (AUC 0.9515) achieves the highest overall discriminatory power across all thresholds, indicating it is the best model at correctly separating positive and negative cases generally. XGBoost (AUC 0.9461) shows

exceptional performance in the low False Positive Rate (FPR) region, suggesting it is highly reliable when minimizing false alarms is the primary concern.



Precision - Recall Curve Analysis: The PR curves show that the ensemble models maintain significantly higher Precision across all levels of Recall than the DT. Random Forest (AP 0.9585) and XGBoost (AP 0.9584) are nearly tied for the most robust performance. Both models maintain Precision above 0.9 even when achieving high Recall (>0.8). This high Average Precision (AP) confirms they are excellent for this diagnostic task, as they minimize both False Positives (high precision) and False Negatives (high recall).

```
--- Final Model Comparison (F1-Scores)
Decision Tree F1: 0.8065
Random Forest F1: 0.8923
AdaBoost F1: 0.8182
XGBoost F1: 0.8889
```

Conclusion: Based on the cumulative evidence from all metrics (F1-score) and the plots (AUC/AP), the Random Forest model is marginally the overall best performer. It achieved the highest AUC (0.9515), the highest F1-score (0.8923), and demonstrates superior stability and balance between critical metrics due to its variance-reducing (bagging) mechanism. It also had the lowest number of dangerous False Negative predictions (3), making it the most reliable and effective choice for this clinical application.