HEART DISEASE PREDICTION

Hari Prabha N

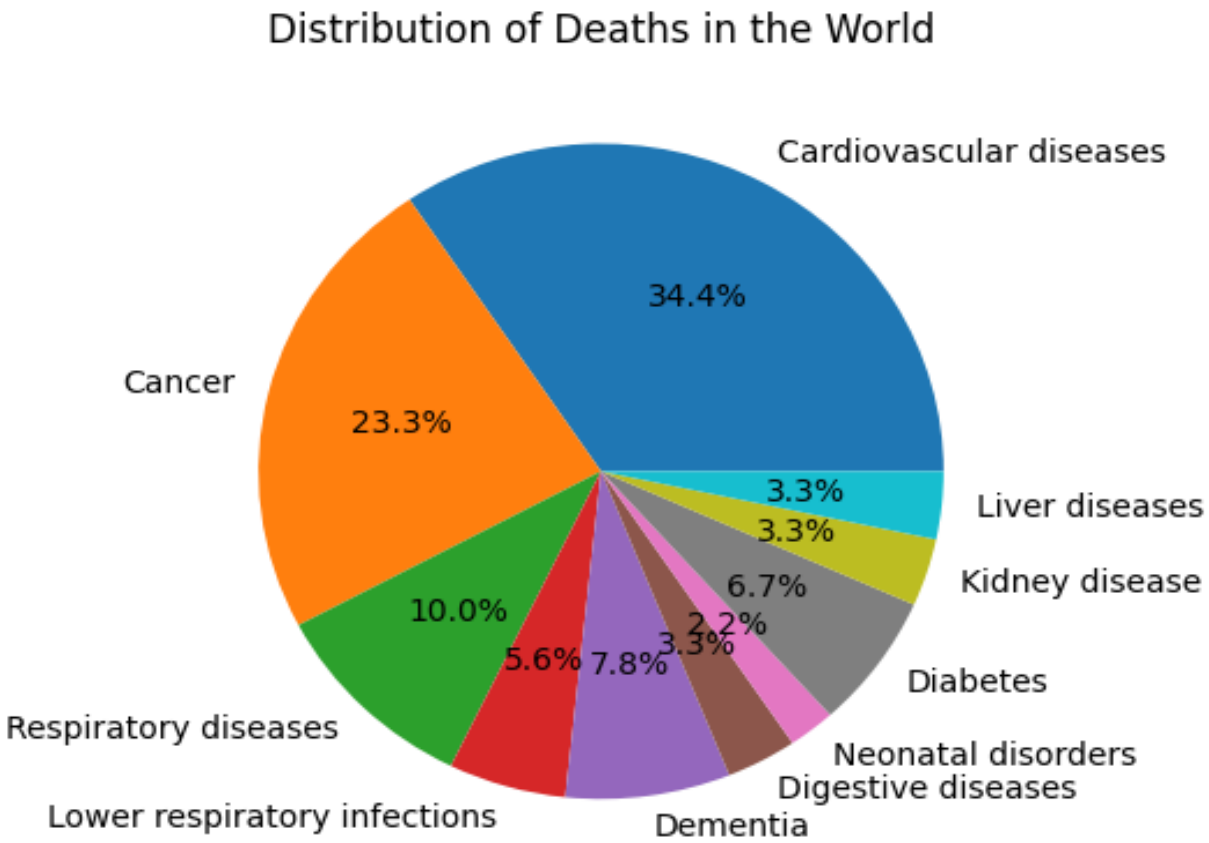
Agenda

1. Introduction
2. Objective
3. Data Summary
4. Cleaning & Preprocess Technique
    4.1 Outliers
5. Machine Learning Algorithm
    5.1 Model Accuracy Comparison
    5.2 Model ROC-AUC Curves Comparison
    5.3 Model Metrics Evaluation

Agenda

6. Model Selection: Random Forests
   - 6.1 Model: Random Forests – ROC-AUC Curve
   - 6.2 Model: Random Forests – Optimization
   - 6.3 Model: Random Forests – Evaluation
7. Model Deployment
8. Importance of Heart Disease Prediction
9. Business Values
10. Future Enhancements
11. Conclusion

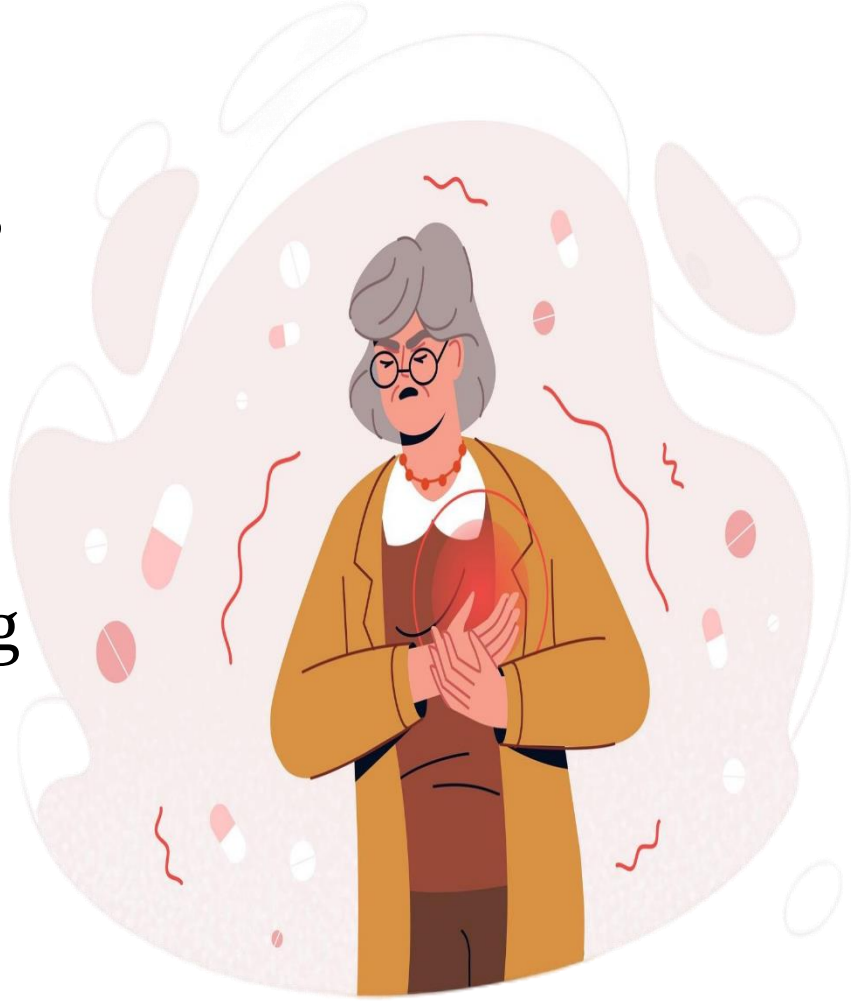
1. Introduction



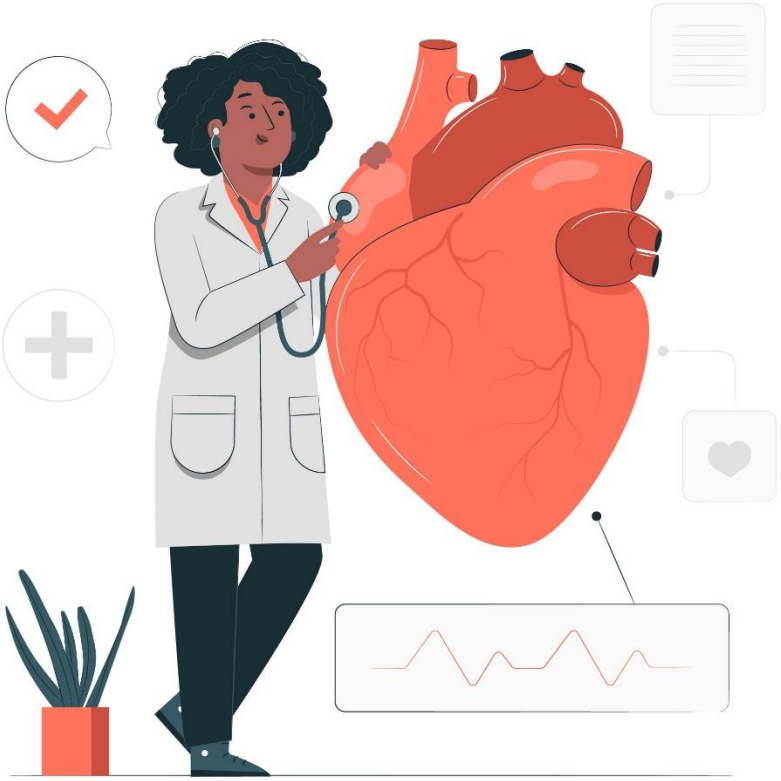
- Heart disease, a major global health issue, affects the heart's structure and function, leading to serious health complications and high mortality rates.

- Heart disease cases are escalating rapidly, making early prediction and diagnosis increasingly critical.

- Heart disease is the leading cause of death worldwide, responsible for 17.9 million deaths annually. In the U.S., it causes 1 in every 5 deaths, totaling around 697,000 annually.



2. Objective

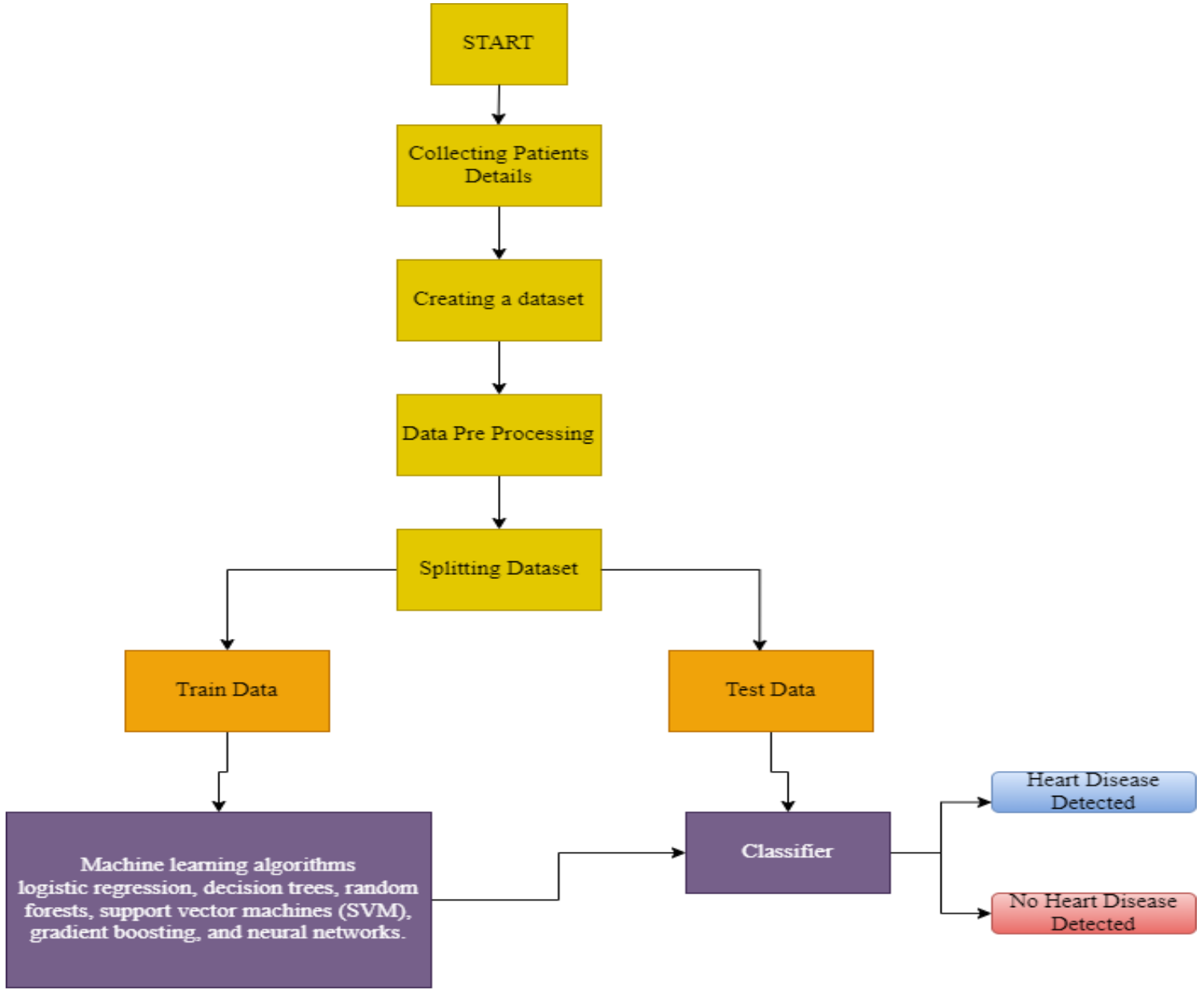


- The primary objective of this project is to build and evaluate machine learning models for predicting the presence of heart disease based on a given set of features such as age, sex, blood pressure, cholesterol levels, and other relevant medical data.

- Early detection and prediction of heart disease are vital. Identifying at-risk individuals early allows for timely interventions, improving patient outcomes and reducing the risk of severe events like heart attacks.

- The goal is to achieve high accuracy while prioritizing and to minimize the risk of misdiagnosing of an individual with heart disease.

Data Flow Diagram

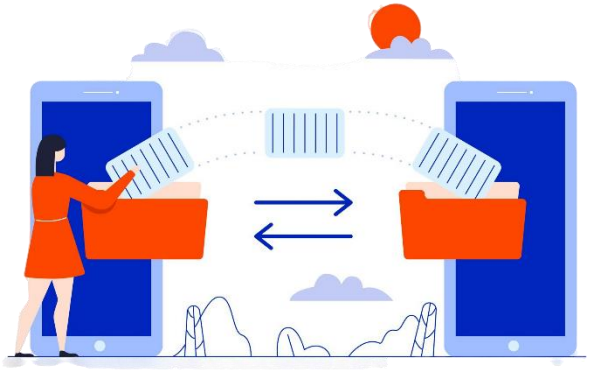


DATA
COLLECTION

DATA
PREPROCESSING

DATA SPLITTING
(Train & Test Data)

MODEL
COMPARISION



MODEL
SELECTION

PREDICTION

3. Data Summary

- The dataset consists of 14 attributes that provide crucial insights into the health profile of individuals at risk of heart disease.
- Each attribute plays a vital role in the predictive models' ability to discern patterns indicative of heart disease.

No. of Columns	Attributes	Descriptions	Type	Values	Normal Values
1	**Age**	The patient's age	Numeric	[29,77]	
2	**Sex**	The gender of the patient	Binary	"Female"=0	
				"Male"=1	
3	Type of Chest Pain **(CP)**	type of chest pain experienced by the individual	Nominal	"Typical Angina"=1	Healthy individuals typically fall into the categories of 2 (Non-anginal pain) or 3 (Asymptomatic).
				"Atypical Angina"=2	
				"Non-angina pain"=3	
				"Asymptomatic"=4	
4	Resting blood pressure **(Trestbps)**	resting blood pressure value in mmHg	Numeric	[94, 200]	Normal: 90-120 mm Hg systolic

No. of Columns	Attributes	Descriptions	Type	Values	Normal Values
5	Serum cholesterol (Chol)	cholesterol levels in mg/dL	Numeric	[126, 564]	Normal: Less than 200 mg/dL Borderline high: 200-239 mg/dL High: 240 mg/dL and above
6	Fasting blood sugar (Fbs)	blood sugar levels	Binary	"No Blood sugar" =0	Normal: Less than 120 mg/dL
				"Blood sugar > 120 mg/dL" =1	Diabetes: 120 mg/dL and above
7	Resting Ecg (Restecg)	Resting electrocardiographic results	Nominal	"Normal"=0	A healthy heart typically shows a result of 0 (Normal).
				"having ST-T wave anomaly" =1	
				"left ventricular hypertrophy" =2	

No. of Columns	Attributes	Descriptions	Type	Values	Normal Values
8	Maximum heart rate **(Thalach)**	maximum heart rate reached by the individual in number of samples	Numeric	[71, 202]	Normal range varies by age and fitness level, but a common formula is 220 minus the individual's age.
9	Exercise-induced angina **(Exang)**	provides information if exercise induces angina	Binary	"Yes"=0	A healthy heart typically shows 0 (No).
				"No"=1	
10	Exercise peak ST segment **(Slope)**	displayed slope value on the ecg machine	Nominal	"upward slope" =0	Healthy individuals typically have an upsloping ST segment (value 0).
				"flat"=1	
				"downward slope" =2	

No. of Columns	Attributes	Descriptions	Type	Values	Normal Values
11	Number of Major Vessels Colored by Fluoroscopy **(ca)**	displays the number of vessels coloured by fluoroscopy	Numeric	[0,3]	0: Typically indicates a healthier state (no major vessels with significant narrowing).
12	St induced depression **(Oldpeak)**	shows the value of exercise-induced ST depression	Numeric	[0, 6.2]	Normal: 0-1 mm Higher values can indicate myocardial ischemia.
13	Thallium Stress Test **(Thal)**	Stress test results	Nominal	"Normal"=0	A healthy heart typically shows a value of 0 (Normal).
				"Fixed defect" =1	
				"Reversible defect" =2	
14	Diagnosis of heart disease **(Target)**	provides information on whether the person has heart disease	Binary	"No"=0	Signs of heart disease
				"Yes"=1	No Signs of heart disease

4. Cleaning & Preprocessing Technique

Before training the models, the dataset undergoes thorough pre-processing to ensure it is clean and ready for analysis.

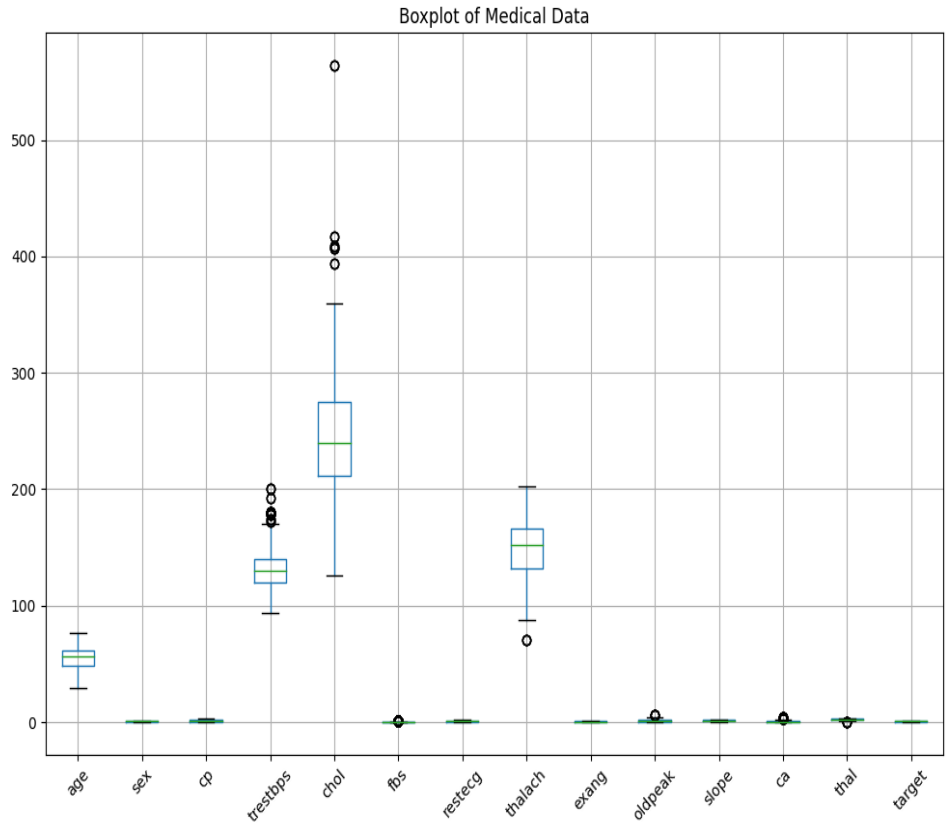
This process begins with

1. **Handling missing data,** though in this case there are no missing values in the dataset.
2. **Scaling,** we have used the StandardScaler method to ensure numerical values are standardized. This prevents any single attribute from dominating the model's learning process.
3. **Outliers,** we have checked the highlighting significant discoveries or trends, and revealing errors or unusual conditions.

4.1 Outliers

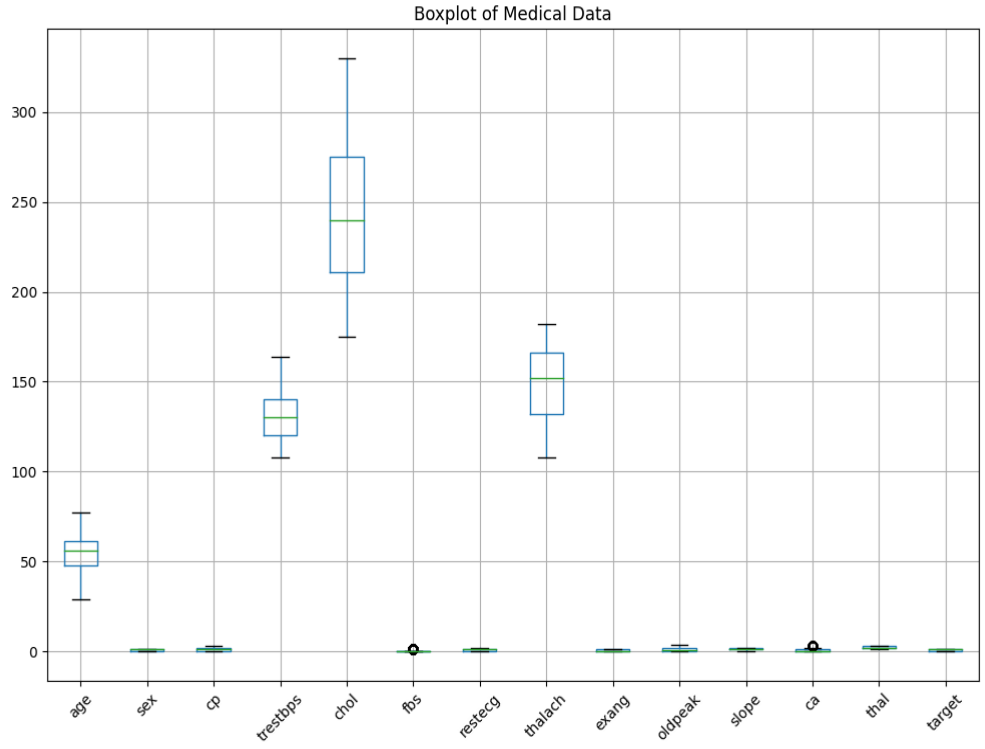
Outliers are present in several attributes, such as

- 'trestbps' (Resting Blood Pressure)
- 'chol' (Cholesterol)
- 'fbs' (Fasting Blood Sugar)
- 'thalach' (Maximum Heart Rate)
- 'oldpeak' (ST Depression)
- 'ca' (Number of Major Vessels Coloured by Fluoroscopy)
- 'thal' (Thalassemia)

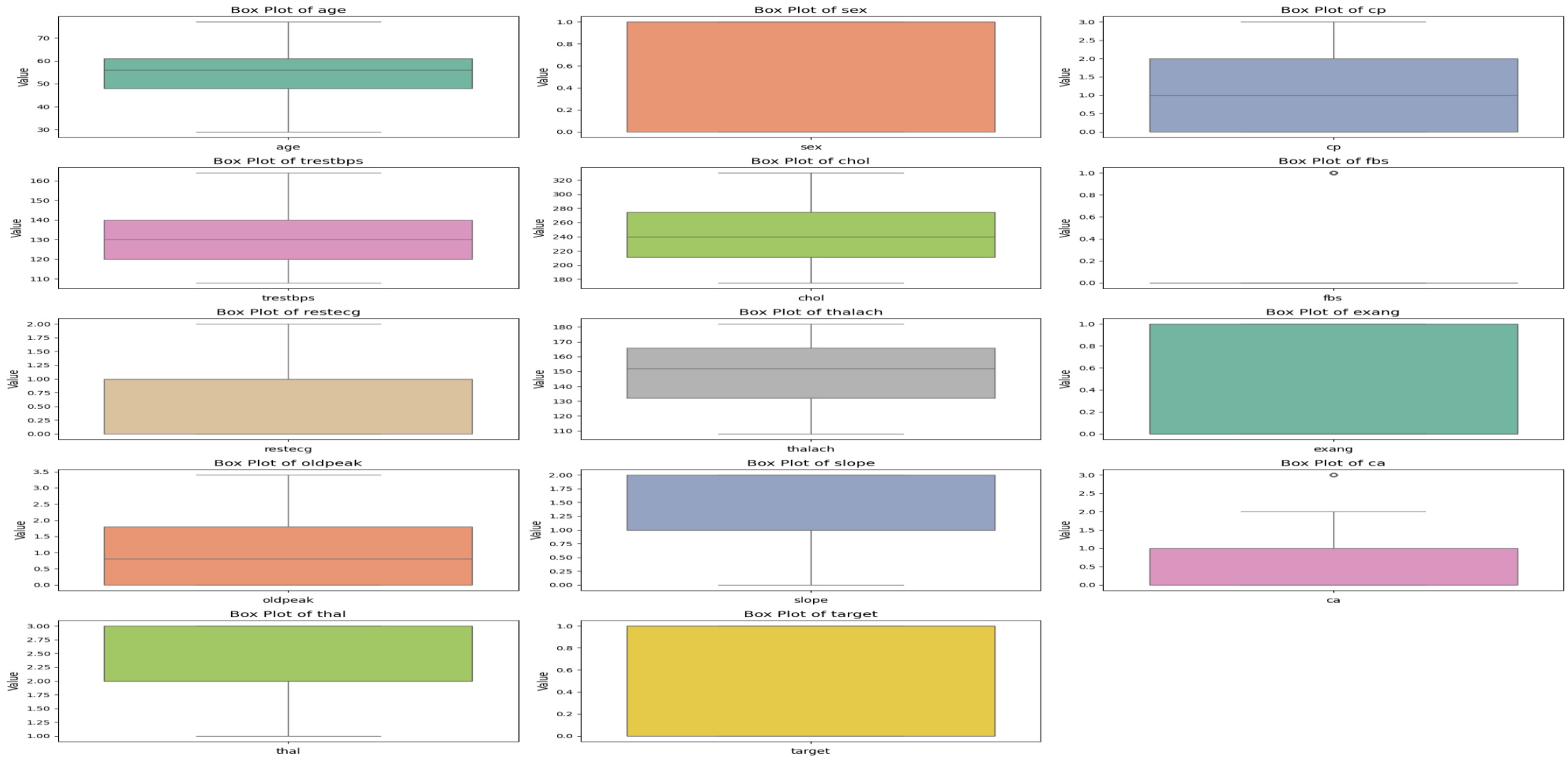


Outliers - Winsorization method

- Winsorization method, will replace extreme values with the nearest values within defined bounds.
- This approach ensures that outliers do not skew the predictive models, maintaining the accuracy of the predictions by preventing overly influential outliers from misleading the analysis.
- Accurate predictions rely on patterns and trends in the data, and managing outliers is crucial to maintaining the reliability of heart disease risk or severity predictions.



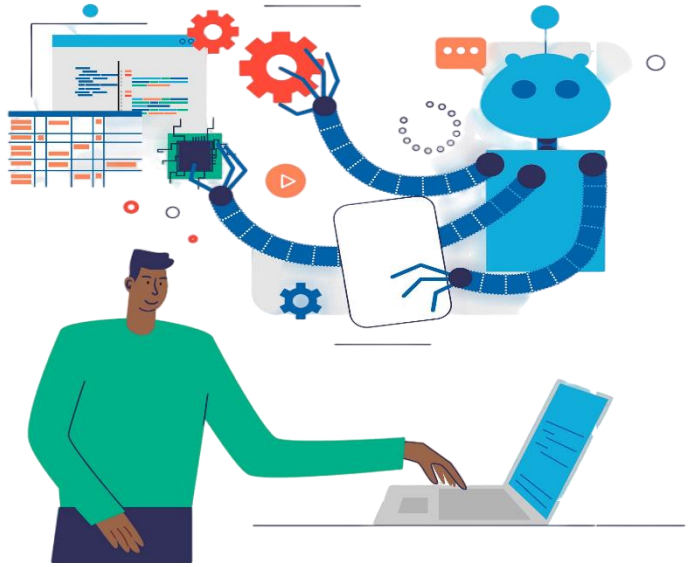
Box Plots of All Features in Heart Disease Dataset



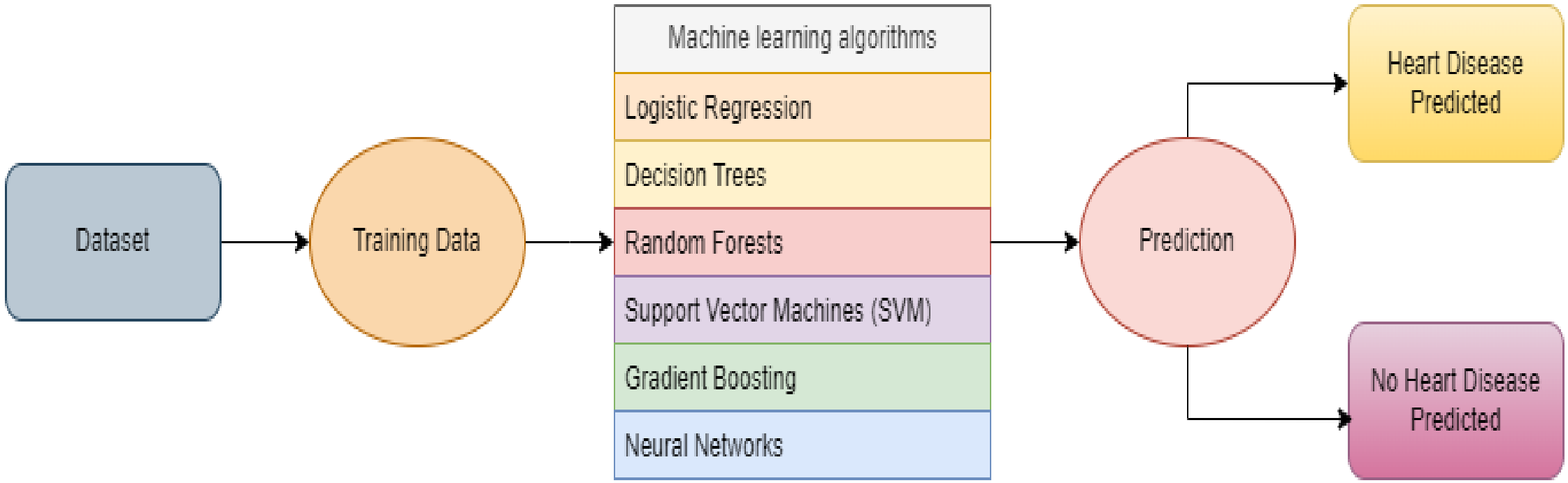
5. Machine Learning Algorithm

I have used the below mentioned models for my Heart Disease Prediction,

1. Logistic Regression,
2. Decision Trees,
3. Random Forests,
4. Support Vector Machines,
5. Gradient Boosting ,
6. Neural Network - MLP

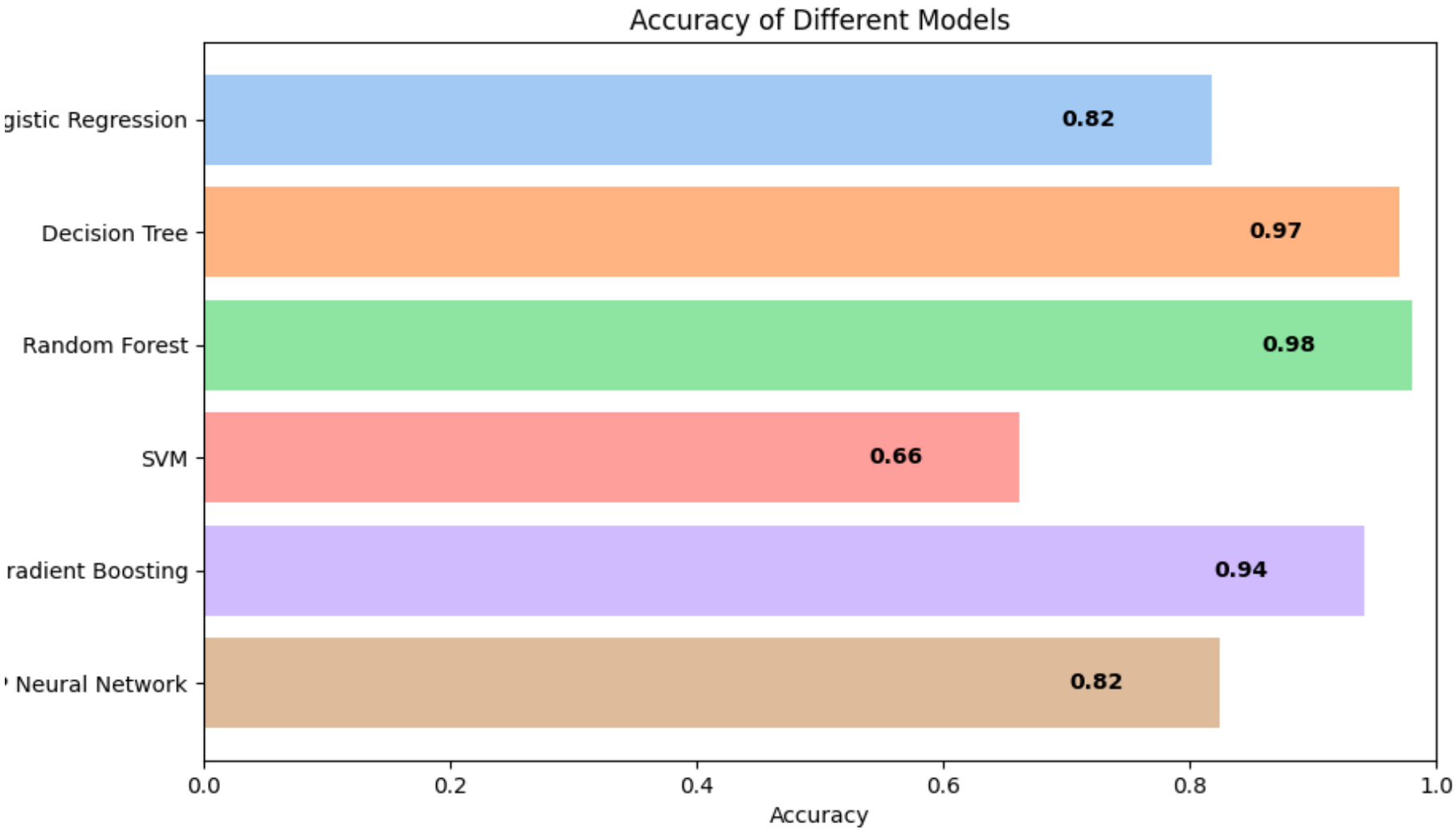


These models are commonly used in predictive analytics and have been shown effective in various healthcare applications such as Diabetes Prediction, Disease Outcome Prediction, and etc.

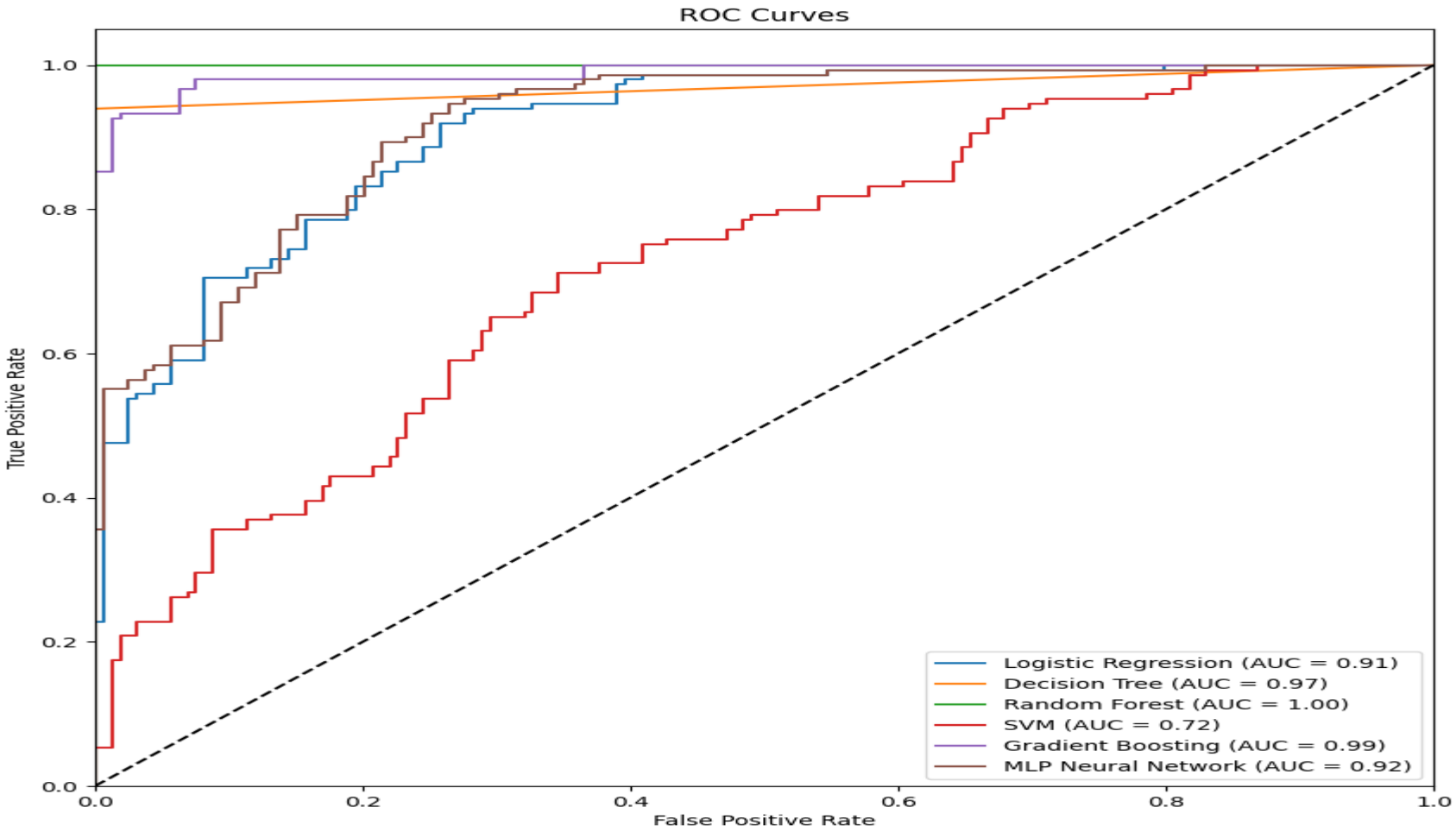

Dataset
Training Data
Machine learning algorithms
Logistic Regression
Decision Trees
Random Forests
Support Vector Machines (SVM)
Gradient Boosting
Neural Networks
Prediction
Heart Disease Predicted
No Heart Disease Predicted

- **Logistic Regression:** Simple and interpretable, often used for binary classification tasks like predicting heart disease.
- **Decision Trees:** Can capture complex interactions and are easy to interpret.
- **Random Forests:** Improve upon decision trees by reducing overfitting and providing robust predictions.
- **Support Vector Machines (SVM):** Effective in high-dimensional spaces, good at handling complex relationships in data.
- **Gradient Boosting:** Builds models sequentially, focusing on correcting errors of previous models, leading to high accuracy.
- **Multi-layer Perceptron (MLP):** Neural networks capable of learning complex patterns, suitable for tasks with non-linear relationships.

5.1]


Accuracy of Different Models
gistic Regression
Decision Tree
Random Forest
SVM
radient Boosting
Neural Network
0.82
0.97
0.98
0.66
0.94
0.82
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy

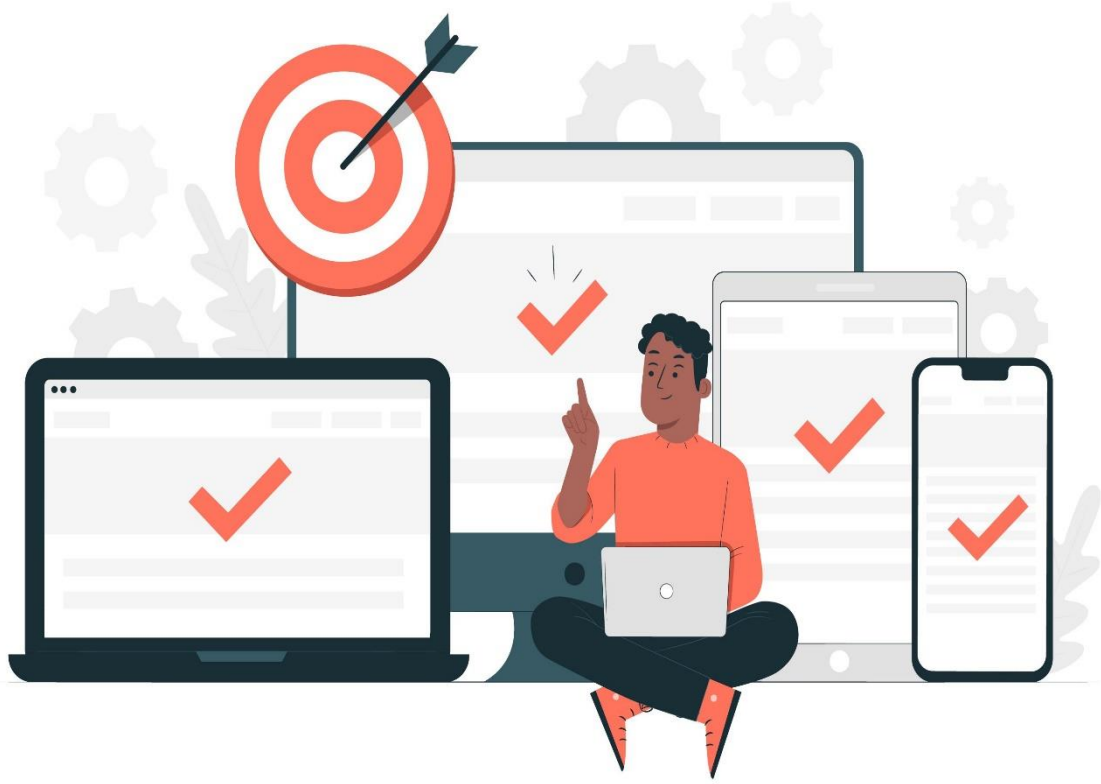
5.2 Model ROC-AUC Curves Comparison



5.3 Model Metrics Evaluation

Model	Class	Precision	Recall	F1-Score	Support
Logistic Regression	0	0.856164384	0.786163522	0.81967	159
	1	0.790123457	0.859060403	0.82315	149
Decision Tree	0	0.946429	1	0.97248	159
	1	1	0.939597	0.96886	149
Random Forest	0	0.963636	1	0.98148	159
	1	1	0.959732	0.97945	149
SVM	0	0.946309	0.886792	0.91558	159
	1	0.886792	0.946309	0.91558	149
Gradient Boosting	0	0.944785	0.968553	0.95652	159
	1	0.965517	0.939597	0.95238	149
Neural Network - MLP	0	0.96319	0.987421	0.97516	159
	1	0.986207	0.959732	0.97279	149

Therefore, **Random Forest** is used as the optimal model for heart disease prediction based on the evaluation metrics.



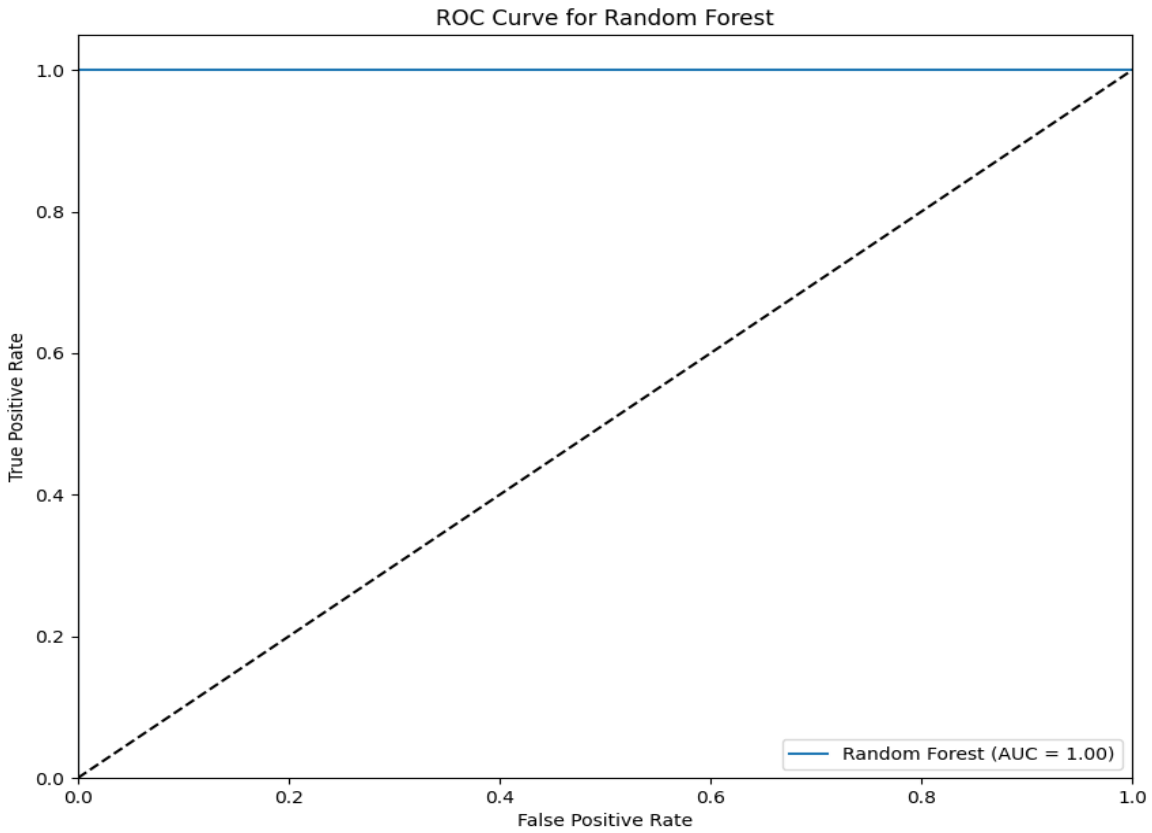
6. Model Selection: Random Forests

- Based on the highest accuracy and F1-score, Random Forest suits to be the best fit for this heart disease prediction task.

- It provides near-perfect accuracy, precision, and recall, indicating reliable predictions across both classes.

- The ROC curve for your Random Forest classifier indicates outstanding performance, with an **AUC of 1.**

6.1 Model: Random Forests - ROC – AUC Curve



This suggests that the model perfectly distinguishes between patients with and without heart disease.

6.2 Model: Random Forests - Optimization

Optimization enhances model accuracy and generalizability by fine-tuning hyperparameters and using techniques like cross-validation.

- **Hyperparameter Tuning**

  Hyperparameters control the learning process and structure of machine learning models. Fine-tuning them can significantly impact performance.

- **Cross-Validation**

  Cross-validation splits the data into multiple folds to ensure the model generalizes well to unseen data. It reduces overfitting and provides reliable performance estimates.

6.2.1 Hyperparameter Tuning

```
# Define the hyperparameters grid for Grid Search
param_grid = {
    'n_estimators': [50, 100, 200, 300],  # Number of trees in the forest
    'max_depth': [None, 10, 20, 30],  # Maximum depth of the trees
    'min_samples_split': [2, 5, 10],  # Minimum number of samples required to split an internal node
    'min_samples_leaf': [1, 2, 4]  # Minimum number of samples required to be at a leaf node
}
```

By defining a grid of these hyperparameters, Grid Search tests each combination to identify the best settings that maximize the model's performance, ensuring higher accuracy and better generalization. This method is essential for fine-tuning the Random Forest to achieve optimal results.

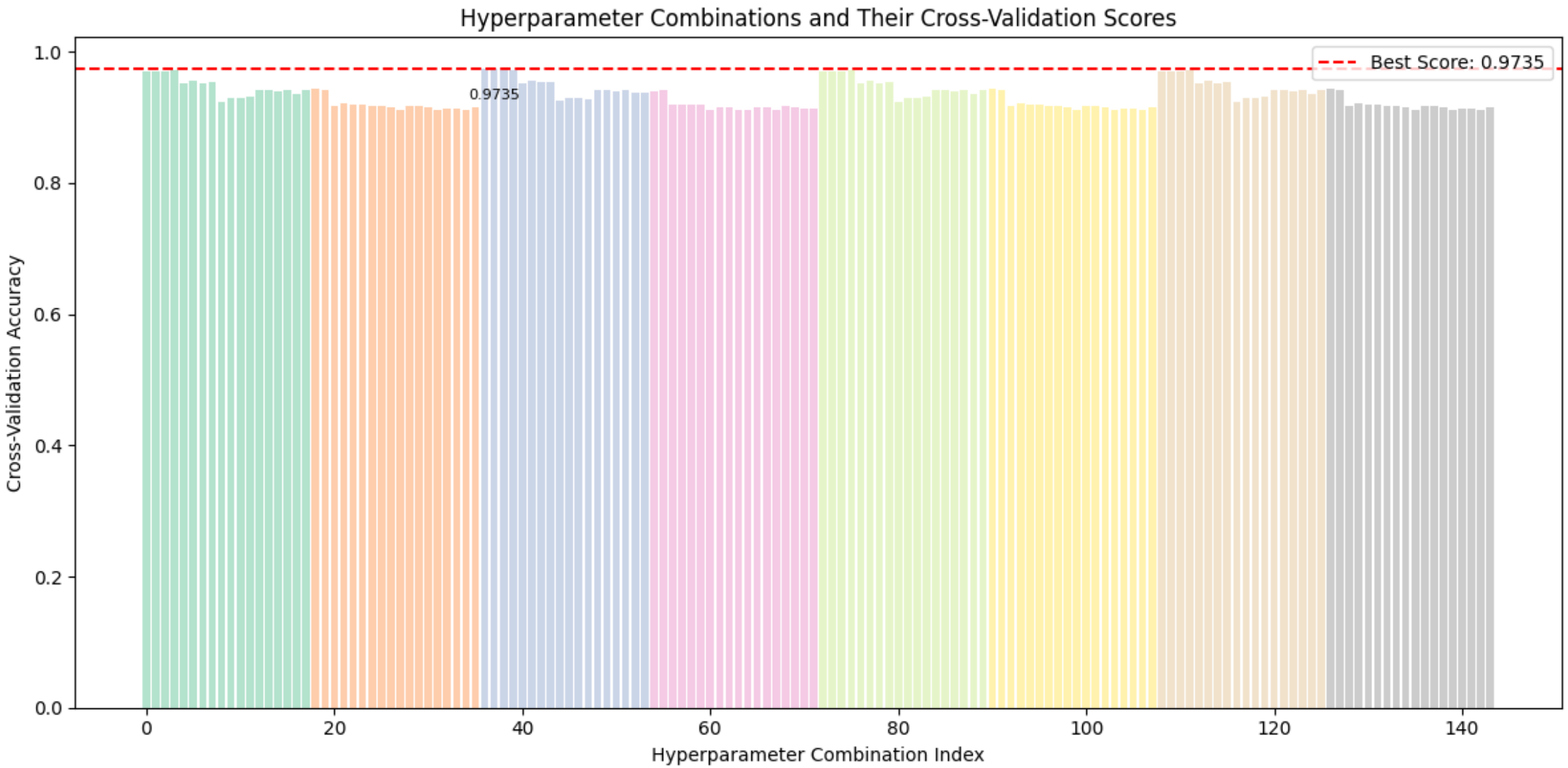
6.2.2 Grid Search with Cross-Validation

```
# Perform Grid Search with cross-validation to find the best hyperparameters
grid_search = GridSearchCV(estimator=rf_classifier, param_grid=param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train, y_train)
```

- Grid Search is used to find the best combination of hyperparameters, using 5-fold cross-validation to ensure generalizability.
- Cross-validation ensures that each combination is tested on multiple subsets of the training data to avoid overfitting and assess the model's generalizability.

Best Parameters and Training the Model

```
# Get the best parameters and the best score
best_params = grid_search.best_params_
best_score = grid_search.best_score_
print(f"Best Parameters: {best_params}")
print(f"Best Cross-validation Accuracy: {best_score}")

# Instantiate the Random Forest classifier with the best hyperparameters
best_rf_classifier = RandomForestClassifier(random_state=42, **best_params)

# Train the model on the training data with the optimized hyperparameters
best_rf_classifier.fit(X_train, y_train)
```

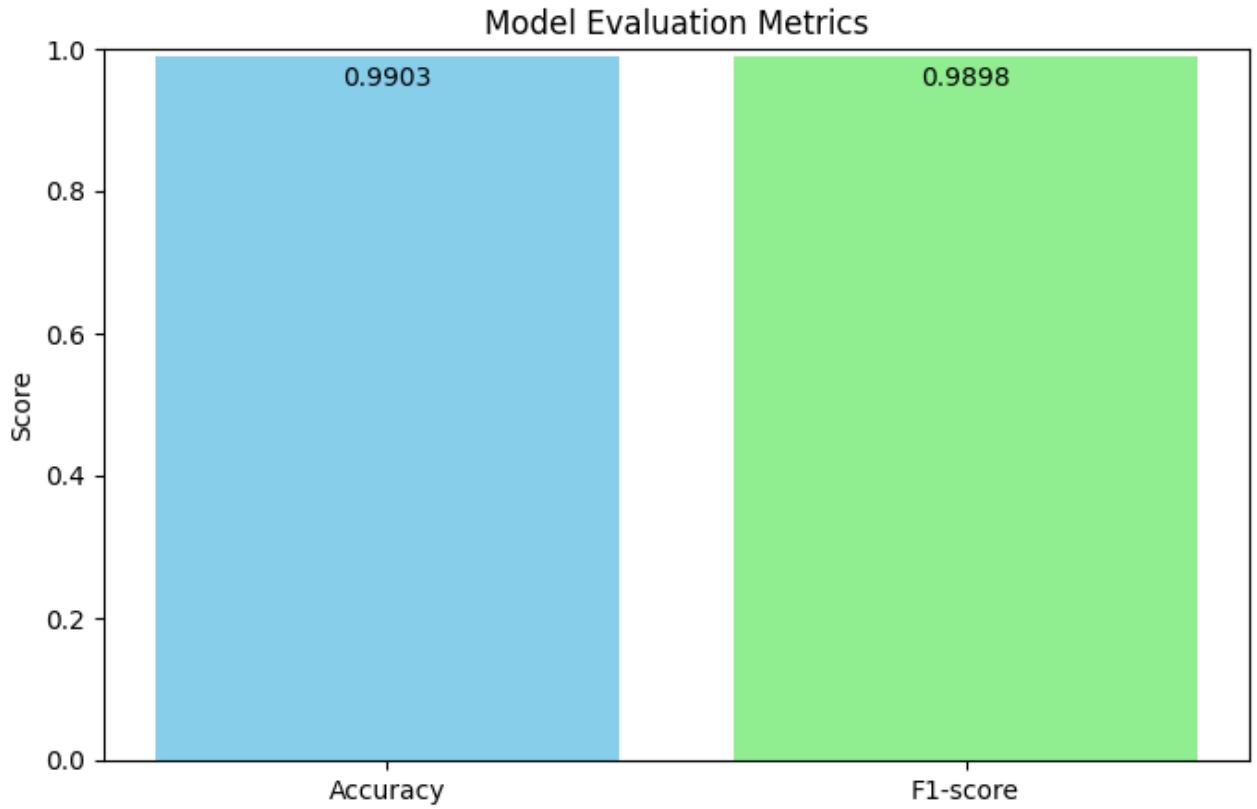

Hyperparameter Combinations and Their Cross-Validation Scores
Best Score: 0.9735
0.9735
Cross-Validation Accuracy
Hyperparameter Combination Index
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
140

6.3 Model: Random Forests - Evaluation

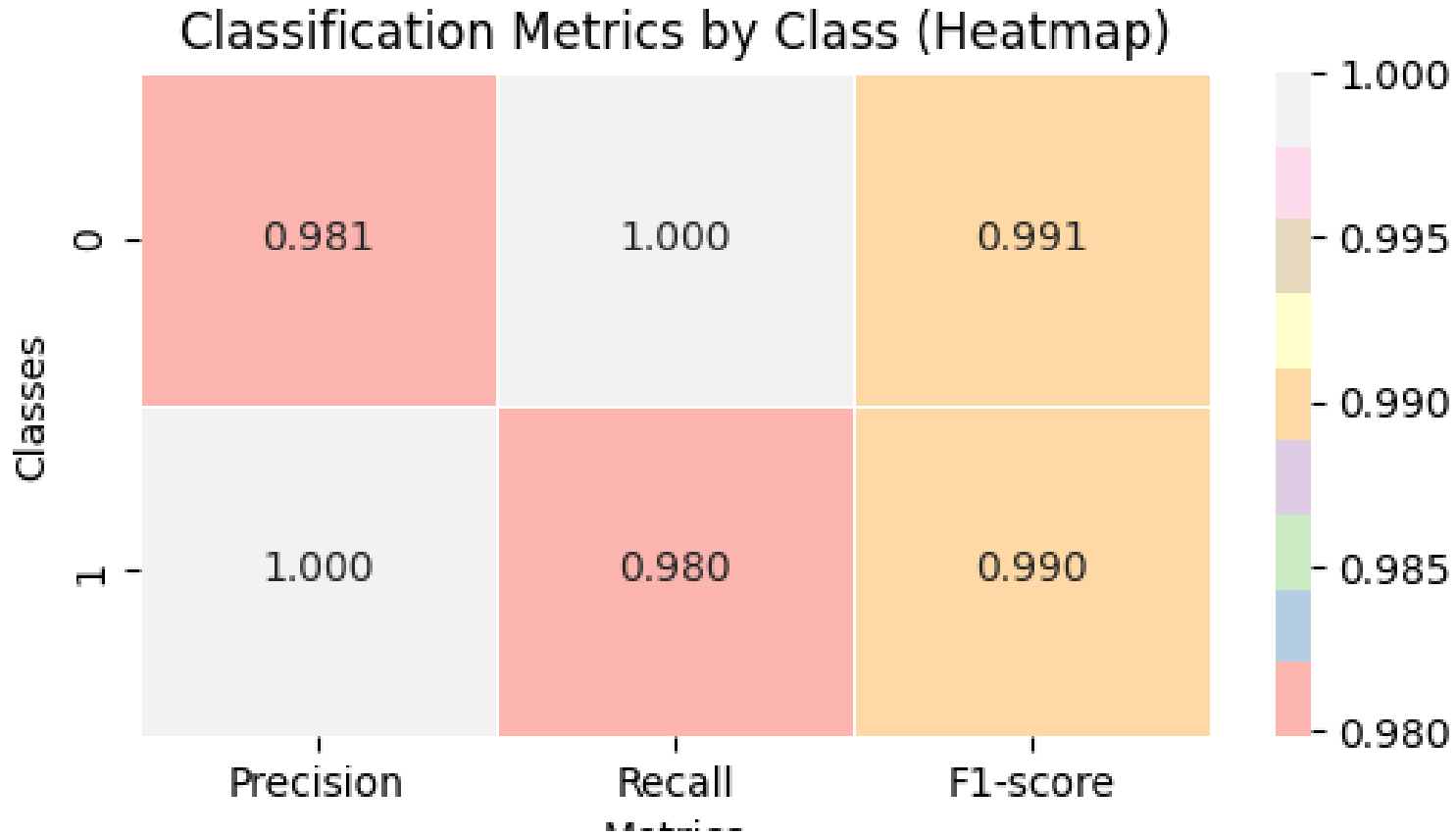
```
# Predictions on the validation set
y_pred = best_rf_classifier.predict(X_test)

# Evaluate the model performance on the validation set
accuracy = accuracy_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
print(f"Validation Accuracy: {accuracy:.4f}")
print(f"Validation F1-score: {f1:.4f}")

# Print classification report for detailed metrics
print(classification_report(y_test, y_pred))
```

- The model's performance is evaluated on the validation set using metrics like accuracy , F1-score & Classification report

Model: Random Forests – Accuracy & F1 Score

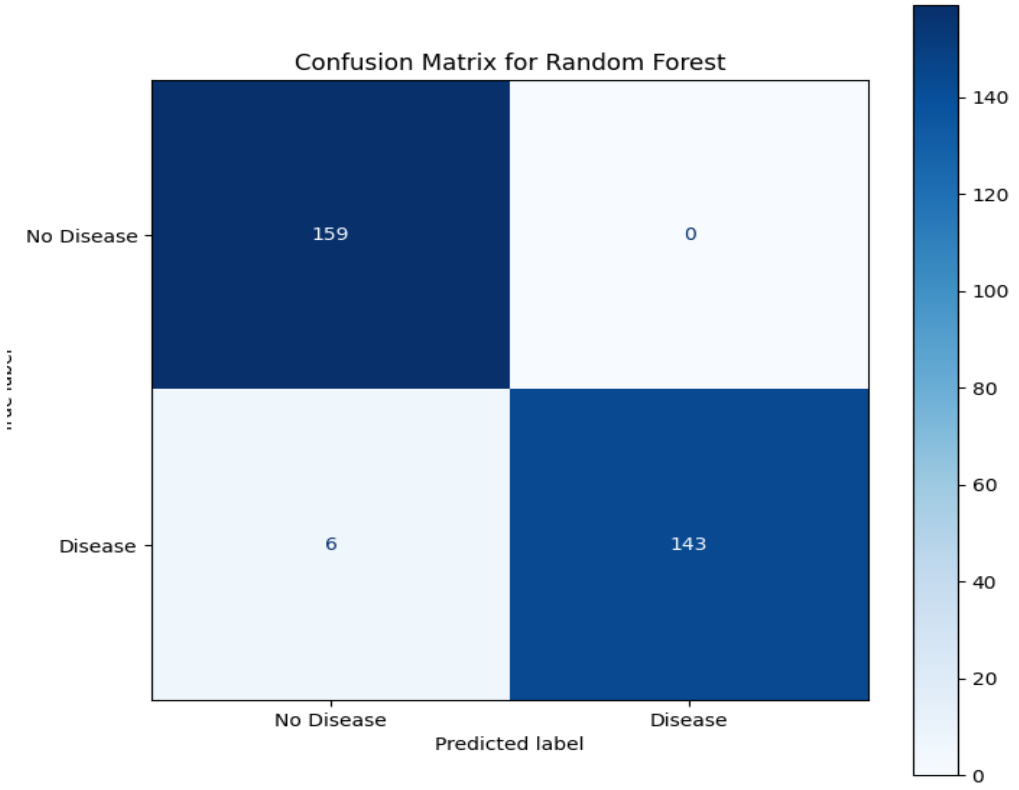


Model: Random Forests – Classification Metrics



Model: Random Forests - Confusion Matrix

- **True Positives (TP):** The value at the top-left corner (**159**) represents the number of correctly classified positive cases. The model correctly predicted these cases as positive.
- **True Negatives (TN):** The value at the bottom-right corner **(143)** represents the number of correctly classified negative cases. The model correctly predicted these cases as negative.
- **False Positives (FP):** The value in the top-right corner **(0)** represents the number of incorrectly classified cases. The model predicted as there is no negative cases as positive **(Type I error).**
- **False Negatives (FN):** The value in the bottom-left corner **(6)** represents the number of missed positive cases. The model predicted these positive cases as negative **(Type II error).**



7. Model Deployment

- After training and evaluation, the best-performing machine learning model for predicting heart disease is selected and prepared for deployment in a web application.

- Deploying machine learning models is critical for real-world applications to make predictive capabilities accessible and impactful.

Deployment in Streamlit

- Streamlit offers a user-friendly interface for deploying models, making it accessible to healthcare professionals and individuals concerned about heart health.

Connecting via LocalTunnel

- After deployment, use LocalTunnel to establish a secure connection to Streamlit, enabling remote access and predictions.

Scenario 1:

Age: 55 years
Sex: Female
Chest Pain Type (cp): 3 (Non-anginal pain)
Resting Blood Pressure (trestbps): 140 mm Hg
Serum Cholesterol (chol): 240 mg/dL
Fasting Blood Sugar (fbs): Yes (> 120 mg/dL)
Resting ECG Results (restecg): 0 (Normal)
Maximum Heart Rate Achieved (thalach): 150 bpm
Exercise Induced Angina (exang): No
ST Depression Induced by Exercise (oldpeak): 1.0
Slope of the Peak Exercise ST Segment (slope): 1 (Flat)
Number of Major Vessels Colored by Fluoroscopy (ca): 0
Thalassemia (thal): 0 (Normal)

Scenario 1 - Result

No Heart Disease Detected

Scenario 2:

Age: 60 years

Sex: Male

Chest Pain Type (cp): 1 (Typical angina)

Resting Blood Pressure (trestbps): 160 mm Hg

Serum Cholesterol (chol): 280 mg/dL

Fasting Blood Sugar (fbs): Yes (> 120 mg/dL)

Resting ECG Results (restecg): 2 (Showing probable or definite left ventricular hypertrophy)

Maximum Heart Rate Achieved (thalach): 130 bpm

Exercise Induced Angina (exang): Yes

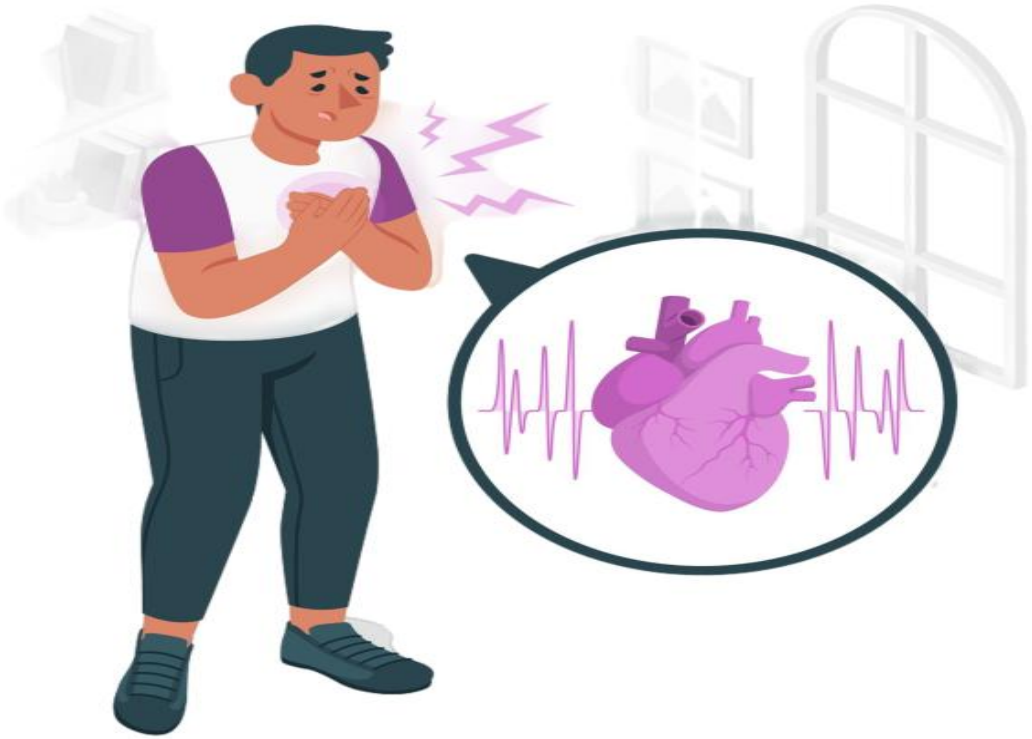
ST Depression Induced by Exercise (oldpeak): 3.0

Slope of the Peak Exercise ST Segment (slope): 0 (Upsloping)

Number of Major Vessels Colored by Fluoroscopy (ca): 2

Thalassemia (thal): 6 (Fixed defect)

Scenario 2 - Result



Heart Disease Detected

8. Importance of Heart Disease Prediction

1. **Enhanced Diagnostic Accuracy:** Our machine learning model leverages comprehensive data to identify early signs of heart disease, leading to more accurate diagnoses.

2. **Timely and Effective Interventions:** Early prediction facilitates prompt and effective treatments, improving patient outcomes.

3. **Cost Reduction:** Implementing predictive models in clinical practice can significantly reduce healthcare costs associated with late-stage diagnosis and treatment.

9. Business Values

1. **Cost Efficiency and Reduction:** Early detection and prevention of heart disease can significantly reduce the costs associated with treatment and management of advanced stages of the disease.

2. **Improved Patient Outcomes:** Accurate prediction models enhance the ability of healthcare providers to diagnose and treat heart disease at its early stages, leading to better patient outcomes and quality of life.

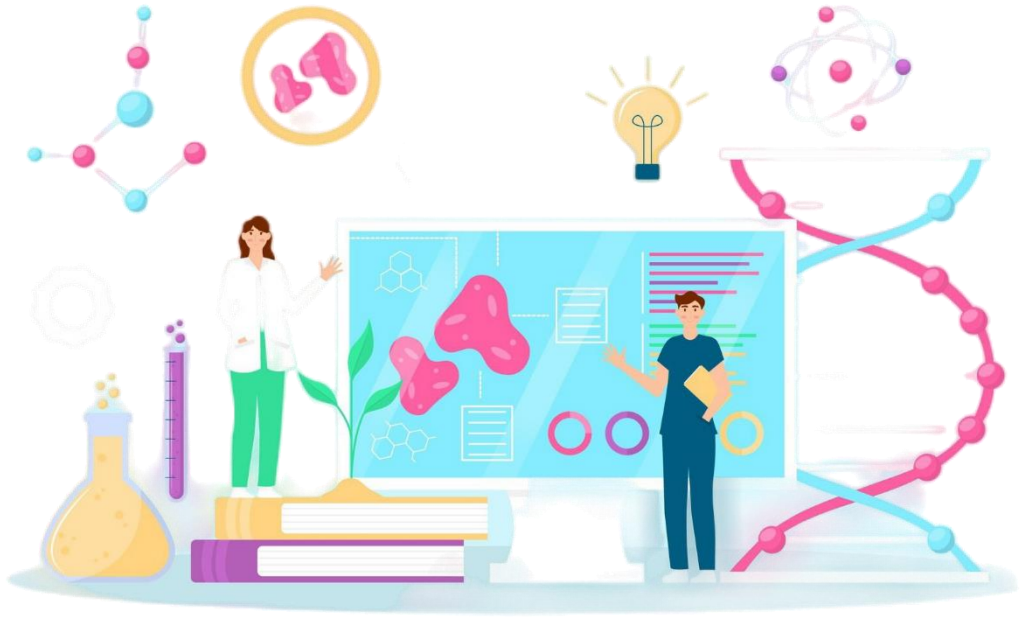
3. **Enhanced Healthcare Services:** Integration of machine learning models into clinical workflows streamlines diagnostic processes, improving efficiency and reducing the burden on healthcare professionals.

10. Future Enhancements

- **Incorporation of Genetic Data:** Enhance prediction accuracy by including genetic markers and family history data to identify individuals at higher risk due to hereditary factors.

- **Multi-Disease Prediction Models:** Expand predictive capabilities to include co-morbid conditions such as diabetes, hypertension, and stroke, providing a holistic view of a patient's cardiovascular health.

- **Advanced Visualization Tools:** Create interactive and intuitive visualization tools that help healthcare professionals and patients understand risk factors and prediction results.



11. Conclusion

- The machine learning model developed through this study offers enhanced accuracy in predicting heart disease, facilitating earlier diagnosis and intervention.

- This can lead to better patient management and improved health outcomes.

"Early diagnosis saves lives. Turning data into actionable insights paves the way for a healthier future."

"Thank you for your attention. Wishing you all a healthy and happy life!"

Thank You