# Heart Disease Prediction Using Machine learning : A Data-Driven Approach

OMKAR RAJENDRA PATIL
232010045

# (1)INTRODUCTION

## 1)Heart Disease

1)Heart disease is one of the main causes of death around the world.

2)According to the World Health Organization (WHO), about 17 million people die each year from heart-related diseases

## 2)Importance of Early Detection:

1)The number of people getting heart disease is rapidly growing because of factors like getting older, eating unhealthy food, obesity, smoking, and high blood pressure

2)As heart disease continues to increase, it's very important to detect it early and predict it correctly to help reduce the number of deaths and improve patient care.

## 3)Machine Learning (ML):

1)In the past, diagnosing heart disease required tests like angiography, which are expensive and need skilled doctors to interpret the results.

2)But machine learning (ML) provides a simple way to predict heart disease early by analyzing large amounts of data.

# 2)PROBLEM STATEMENT

**1)Traditional Methods Have Limitations**

Old methods sometimes give late or wrong results, which can delay heart disease treatment.

**2)Imbalanced Data Affects Predictions**

If the data has more healthy people, the system may wrongly say a sick person is healthy.

**3)Missing Information in Healthcare Data**

Missing test details or patient history can lead to wrong treatment or poor predictions.

**4)Complexity of Heart Diseases**

Heart diseases depend on many factors, which makes it hard for doctors to diagnose correctly.

**5)Need for New Data-Driven Methods**

We need smarter tools like machine learning to support doctors and improve decisions.

# (3)OVERVIEW OF HEART DISEASE

## (1)Causes of Heart Disease

**1)High Blood Pressure** – When your blood moves through your body with too much force, it can damage your heart.

**2)High Cholesterol** – Too much fat in your blood can block your heart's blood vessels.

**3)Chest Pain (Angina)** – This happens when your heart doesn't get enough blood.

**4)Unhealthy Habits** – Eating junk food, smoking, or not exercising can make heart problems worse.

## 2)Risk Factors

**1)Age** – Older people are at higher risk.

**2)Gender** – Men are more likely to get heart disease earlier than women.

**3)Family History** – If your parents or grandparents had heart problems, you may be at higher risk.

**4)Lifestyle** – Eating too much unhealthy food, not exercising, or being stressed can increase the risk.

# (4)CLEVELAND HEART DISEASE DATASET

1)The Cleveland Heart Disease Dataset is commonly used in heart disease research

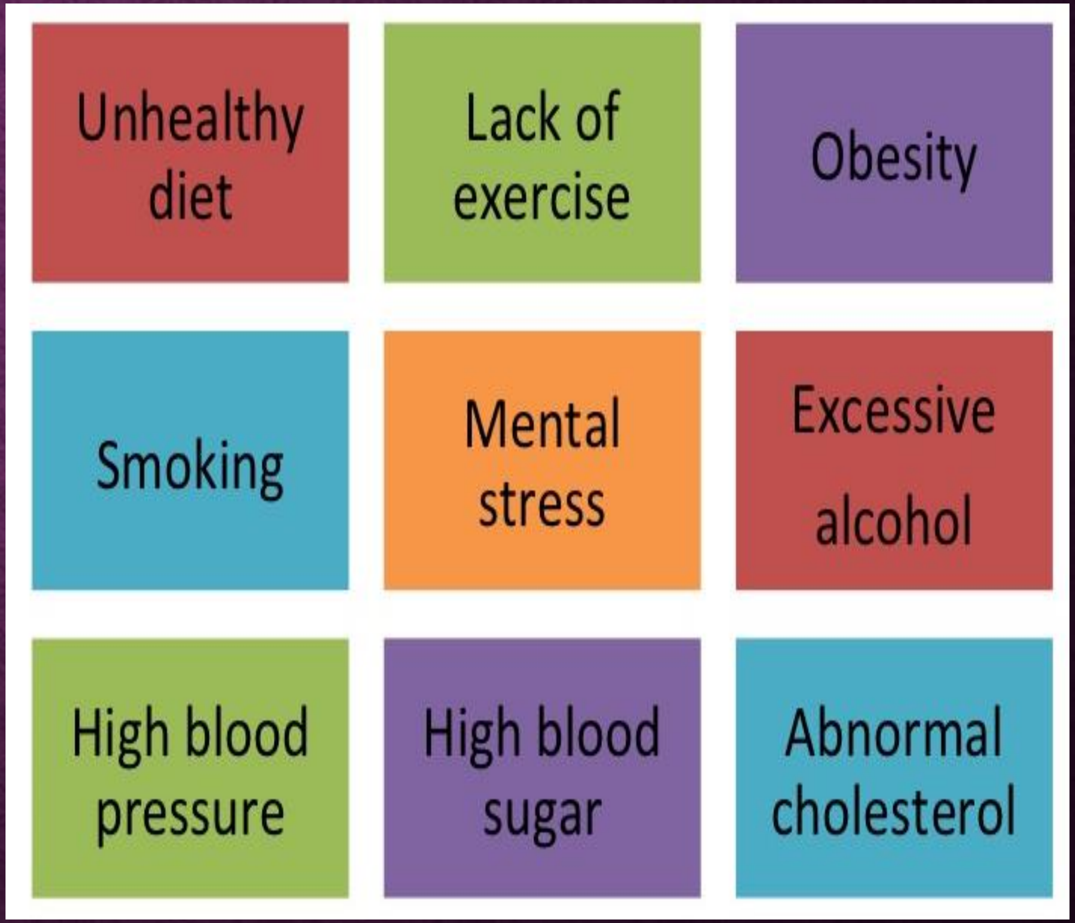
2)It contains data from 920 patients with 76 features.

3)Many studies focus only on 14 important features that are most useful for predicting heart disease.

4)These features include age, gender, cholesterol levels, blood pressure, maximum heart rate, and exercise-induced chest pain etc.

5)These factors give us valuable information about heart health and improve the accuracy of machine learning models.

6)By studying these factors, doctors can identify people who are at high risk and create better prevention plans.



Unhealthy diet

Lack of exercise

Obesity

Smoking

Mental stress

Excessive alcohol

High blood pressure

High blood sugar

Abnormal cholesterol

# 5)WHY PREPROCESS THE DATA

**1)Removes Errors:**
Preprocessing helps fix wrong or missing values in the data.

**2)Improves Accuracy:**
Clean and organized data helps the model make better predictions.

**3)Speeds Up Training:**
Properly prepared data allows the model to learn faster.

**4)Balances Data:**
Preprocessing ensures no feature (like age or salary) dominates the results.

**5)Reduces Noise:**
It removes useless or extra information that can confuse the model.

# 6)DATA PREPROCESSING

## 1)Handling Missing Data:

1)SimpleImputer: Fills missing values with the **average (mean)**, **most common value**, or **zero**.

2)KNNImputer: Looks at similar data points (neighbors) and guesses the missing value based on them.

## 2)Feature Scaling:

**1)Height** may be in centimeters (like 170 cm).

**2)Weight** may be in kilograms (like 65 kg).

3)This adjusts all values to a similar range, ensuring fair comparisons.

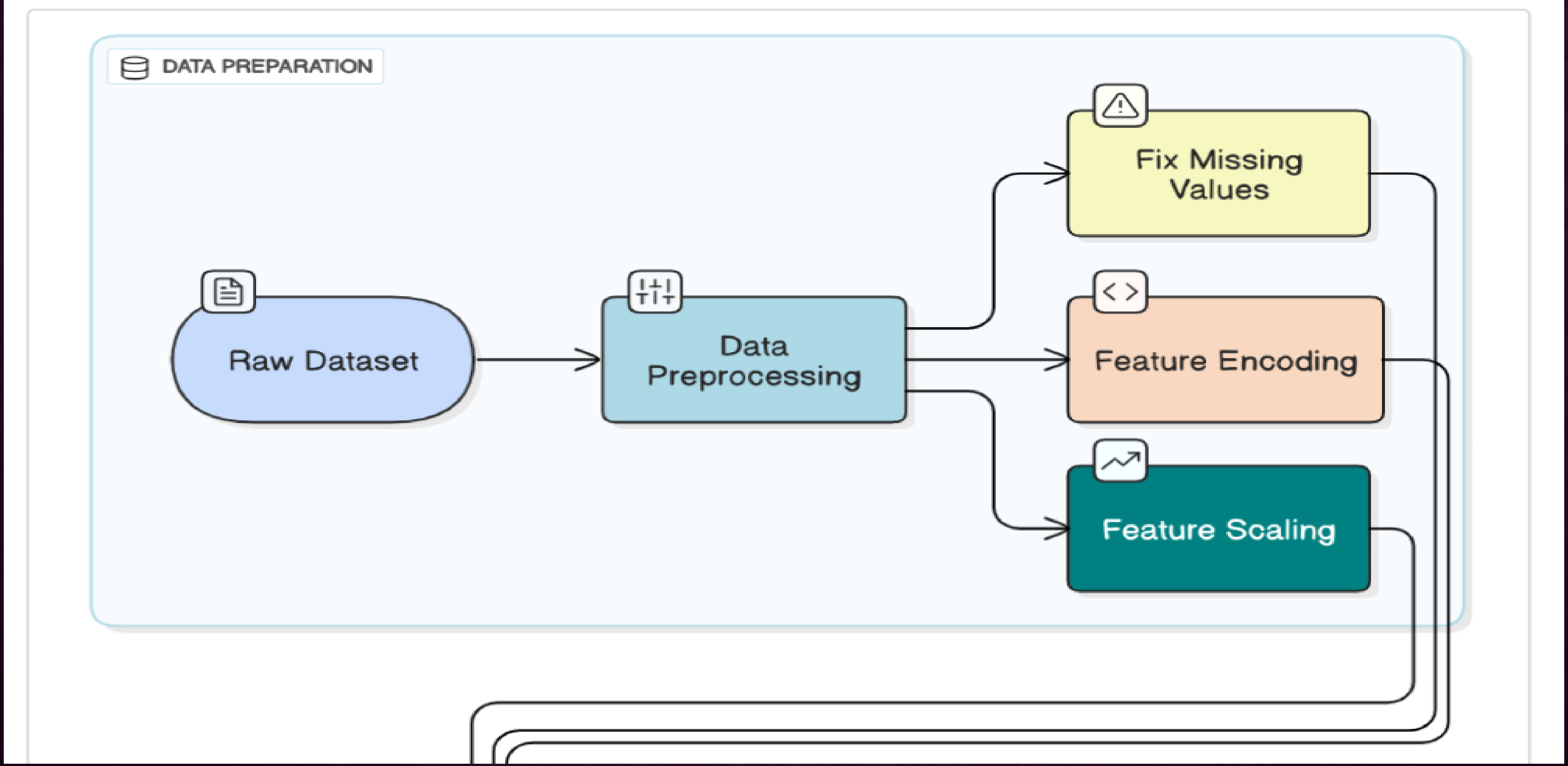
## 3)Data Cleaning:

**1)Incorrect entries** (e.g., someone’s age recorded as 200 years).

**2)Duplicate records** (e.g., the same student’s data appearing twice)

3)Ensuring the dataset is ready for accurate predictions.

# 7)DATA PREPROCESSING WORK FLOW DIAGRAM



# (8)MODEL SELECTION AND TRAINING(1)

**1)Random Forest:**

1)This model uses decision trees to make predictions.

2)This model was good at making predictions without overfitting.

3)This means it gave stable results, even when tested with new data.

4)The model’s 84% accuracy shows it can predict well, even for data it has never seen before.

5)This model is great when you want reliability and simplicity over super-high accuracy.

# (9)MODEL SELECTION AND TRAINING(2)

**1)XGB Classifier :**

1)This model uses gradient boosting to improve accuracy.

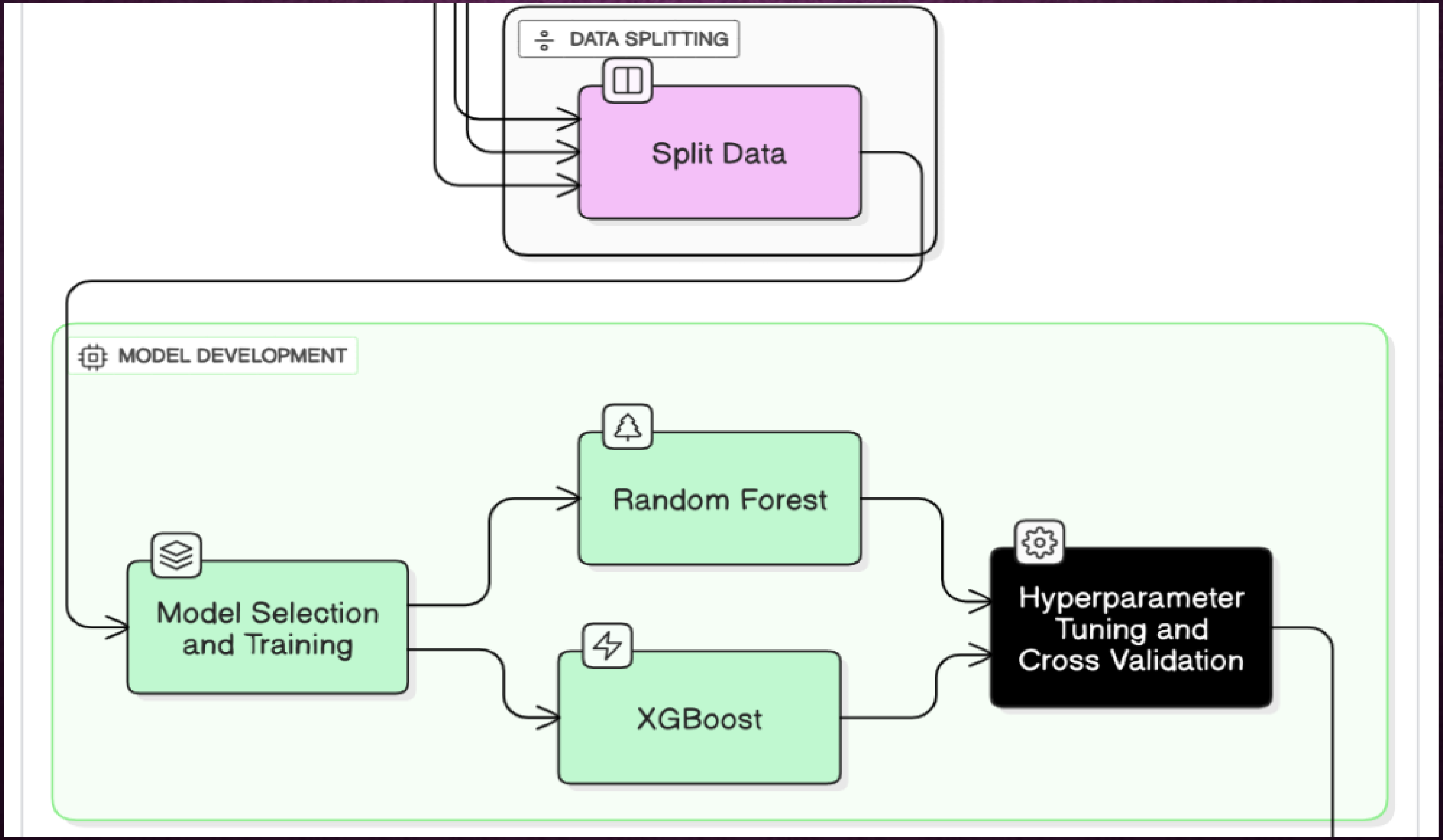
2)The XGB Classifier did slightly better than the Random Forest model, with 86% accuracy.

3)It was better at finding more complex patterns in the data.

4)This made it a good choice when we wanted the model to be as accurate as possible.

5)especially when the patterns in the data are complicated.

# (10)MODEL SELECTION AND TRAINING WORK FLOW DIAGRAM


DATA SPLITTING
Split Data
MODEL DEVELOPMENT
Random Forest
Model Selection and Training
Hyperparameter Tuning and Cross Validation
XGBoost

# 11)MODEL EVALUATION(1)

## 1)Accuracy

**1)**Accuracy tells us how many predictions were correct out of all the predictions made.
2)If a model predicts 100 test cases, and 90 of them are correct, the accuracy is **90%**.

**3)Higher accuracy = Better model** (in simple cases).

## 2)Precision

**1)**Precision shows how many of the **positive predictions** were actually correct.
2)Imagine a heart disease detection model predicted 10 people have heart disease, but only 8 of them were true cases.

3)Precision = **8/10 = 0.8 (or 80%)**

**4)Higher precision = Fewer false alarms**.

## 3)Recall (Sensitivity)

**1)** Recall shows how well the model finds **actual positive cases** (e.g., actual heart disease patients).
2)If there are 20 real heart disease cases and the model found 18 of them, recall = **18/20 = 0.9 (or 90%)**

**3)Higher recall = Fewer missed cases**.

# 12)MODEL EVALUATION(2)

## 4)F1-Score

1)F1-Score is a balance between **precision** and **recall**.

2)It’s useful when both precision and recall are important.

**3)Higher F1-Score = Better overall performance**.

## 5)Confusion Matrix

**1)True Positives (TP):** Correctly predicted positive cases.

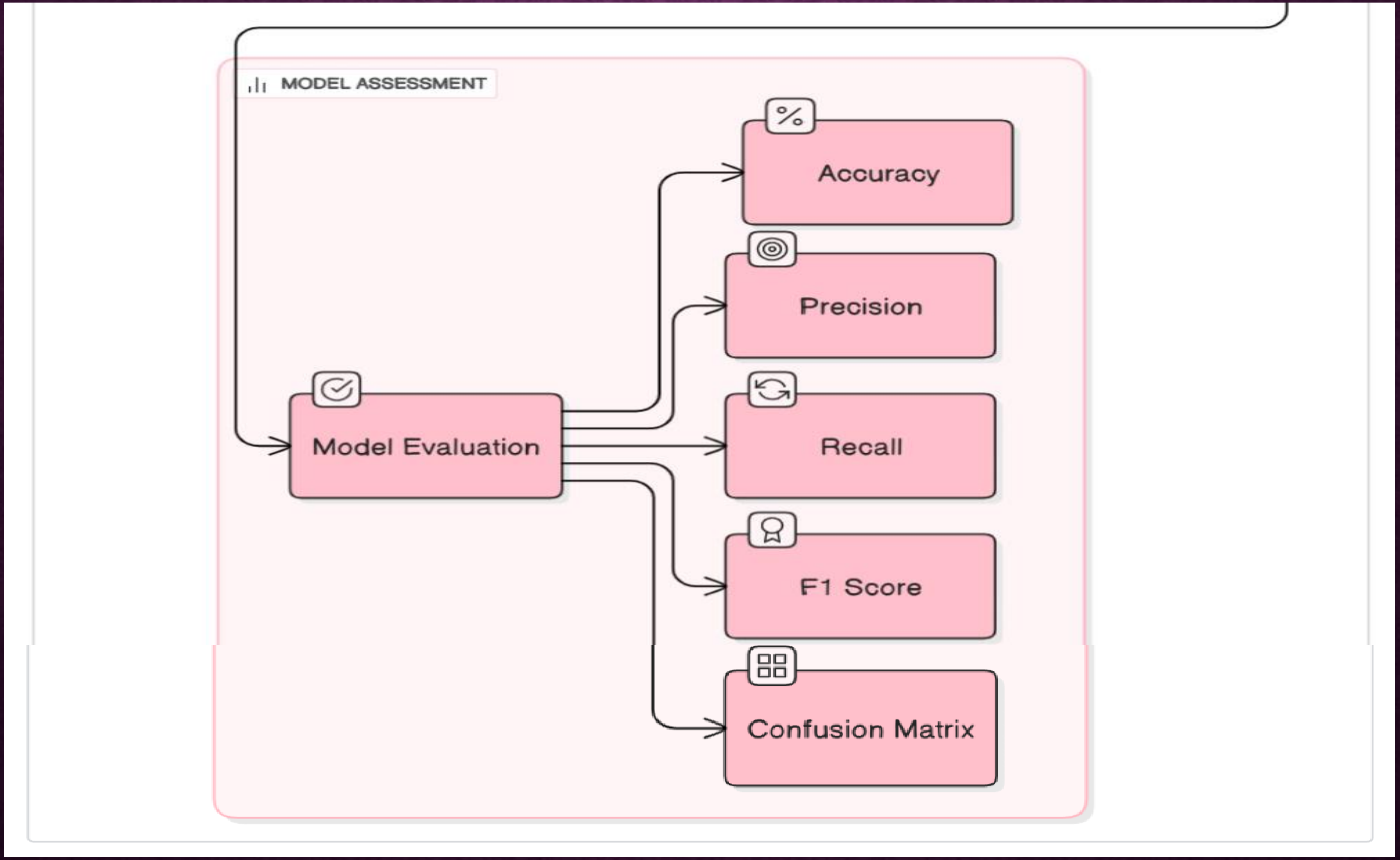
**2)True Negatives (TN):** Correctly predicted negative cases.

**3)False Positives (FP):** Incorrectly predicted positive cases.

**4)False Negatives (FN):** Missed positive cases.

5)The confusion matrix helps you understand exactly where the model is making mistakes.

# 13)MODEL EVALUATION WORK FLOW DIAGRAM



# 14)EXPERIMENTAL RESULTS (1)

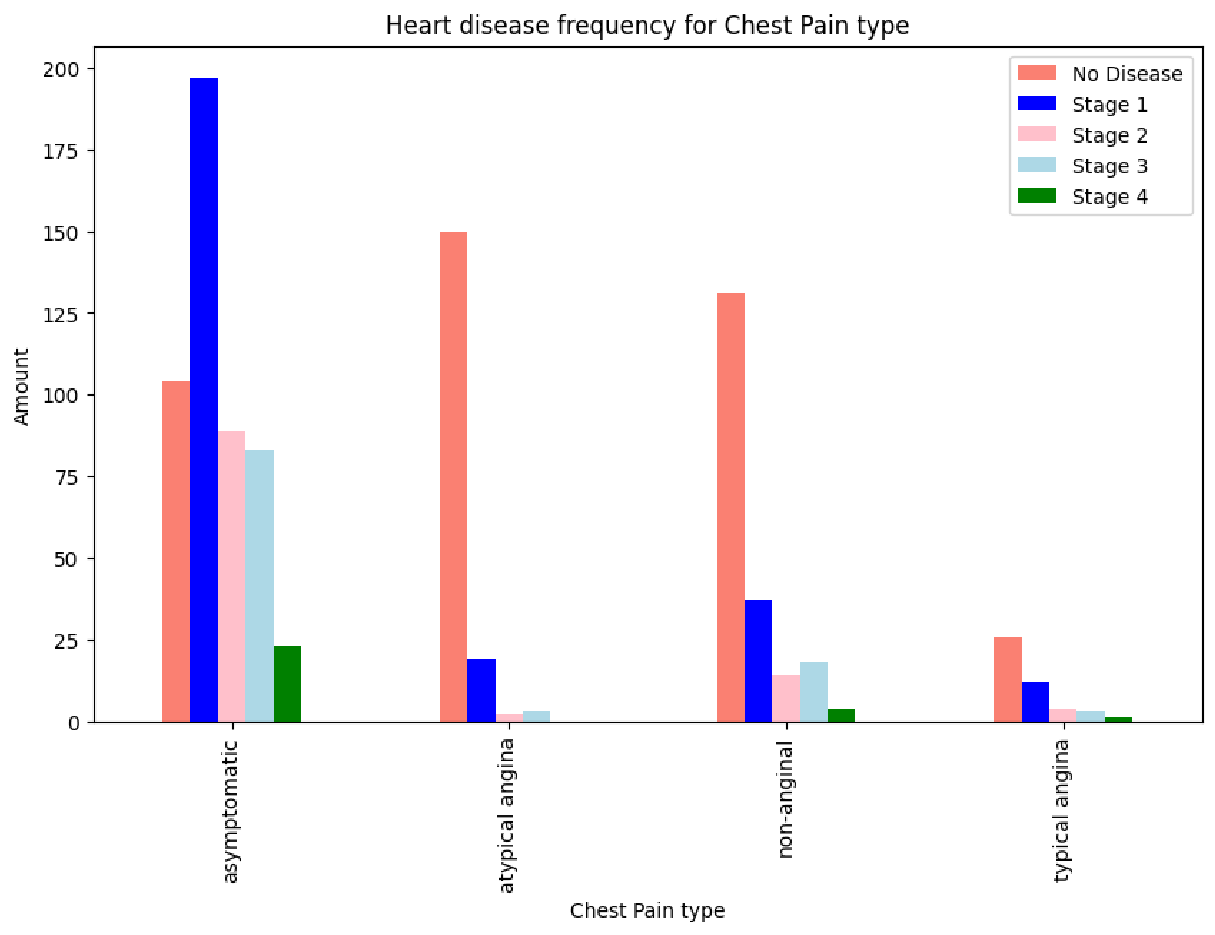


Figure 4: Visualizing the Relationship Between Chest Pain Type and Heart Disease Level

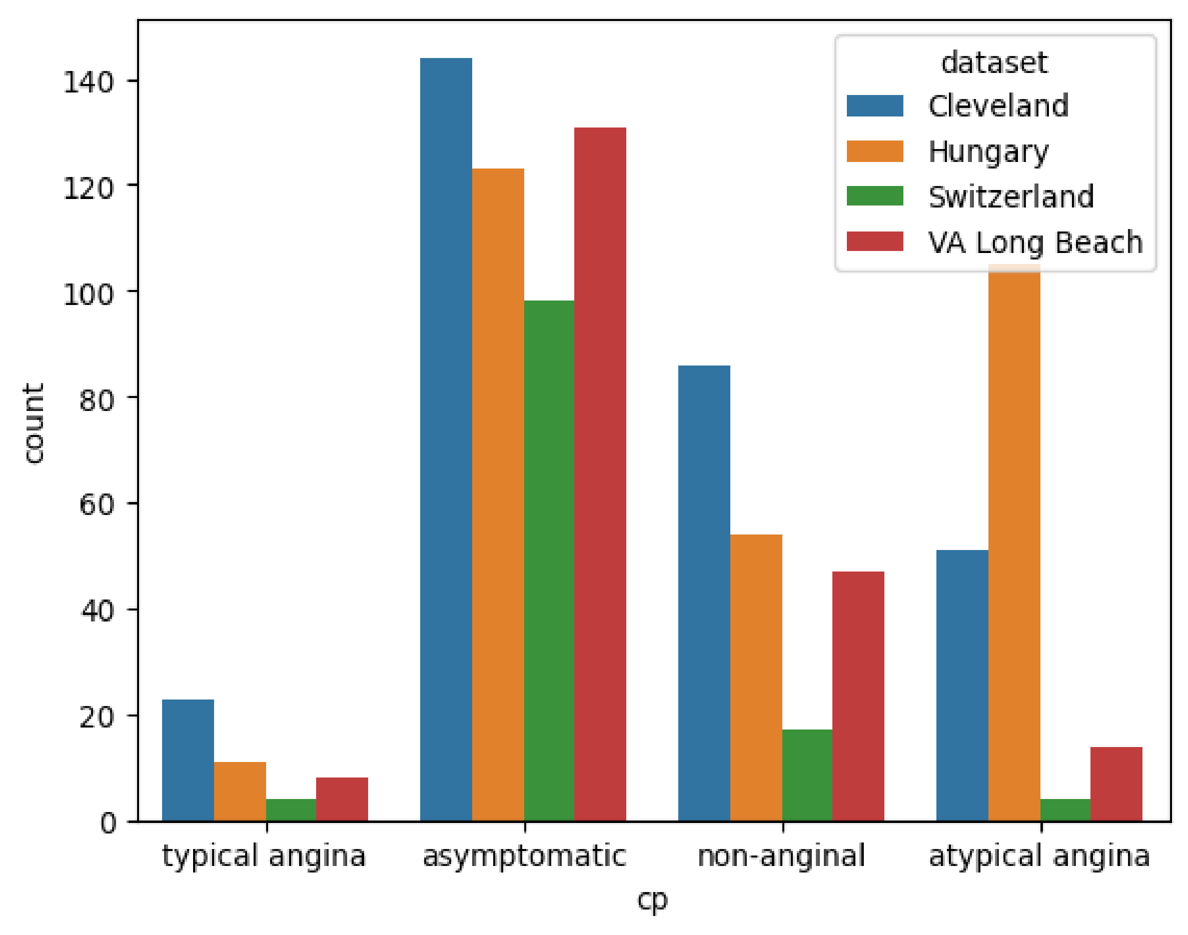


Figure 5: Visualizing Chest Pain Types by Dataset Using Seaborn

# 15)EXPERIMENTAL RESULTS (2)

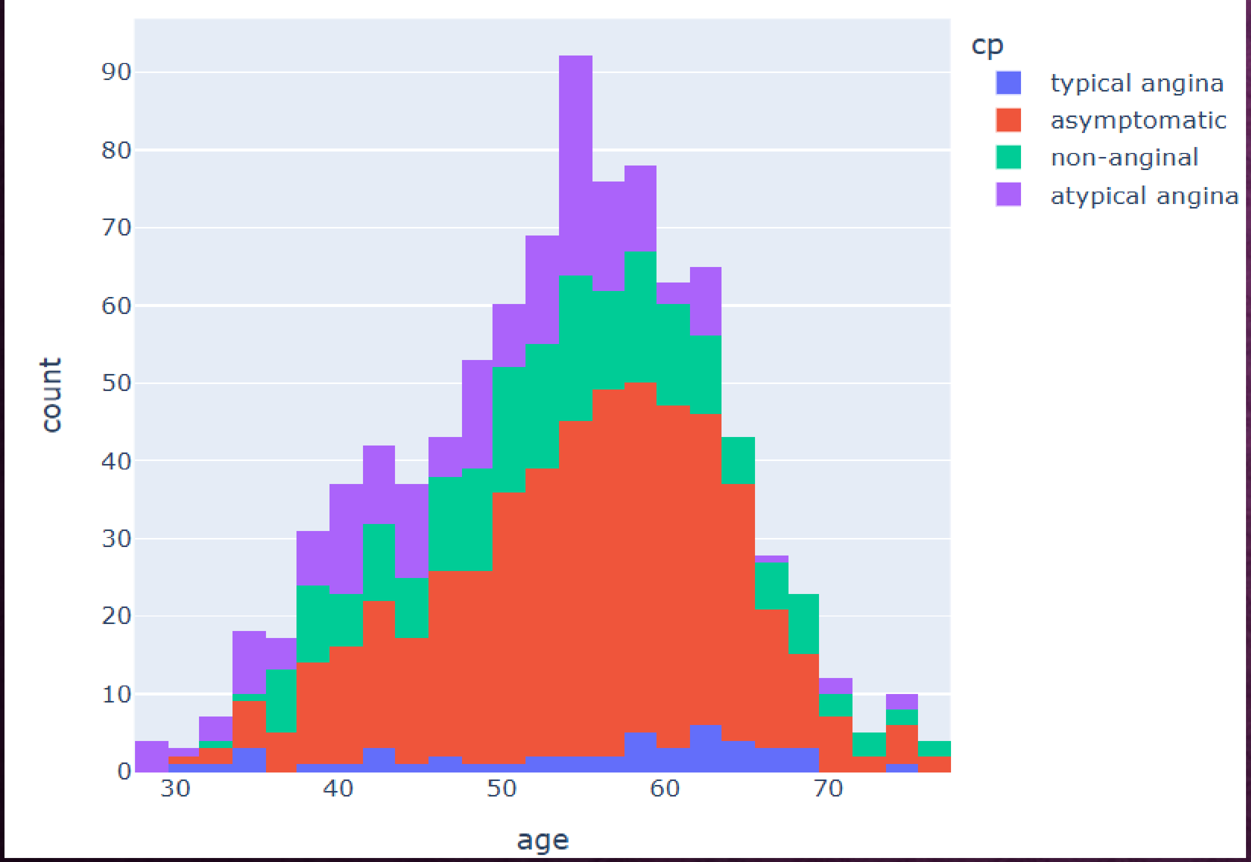


Figure 7: Plotting Age Distribution Grouped by Chest Pain Type using Plotly Histogram

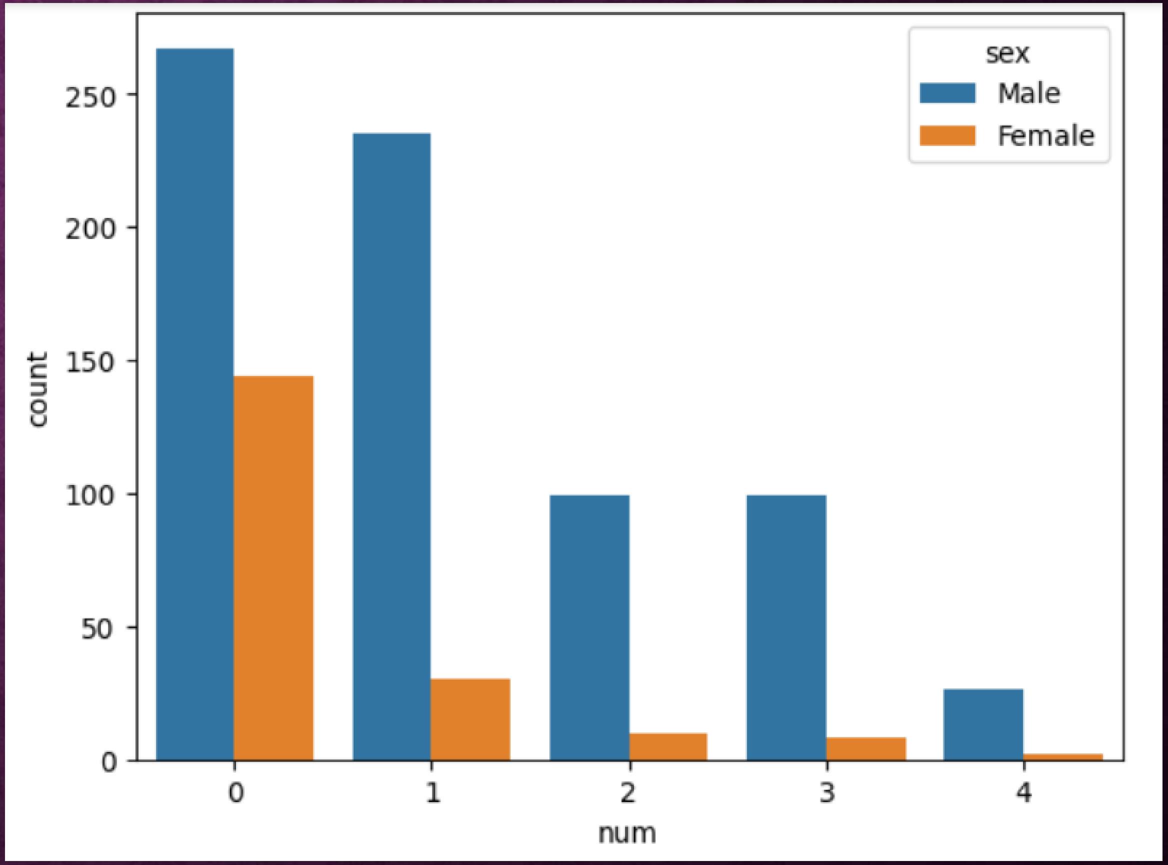


Figure 8: Comparing Gender with Heart Disease Presence

# 16)FINAL RESULTS

**1)Random Forest:**

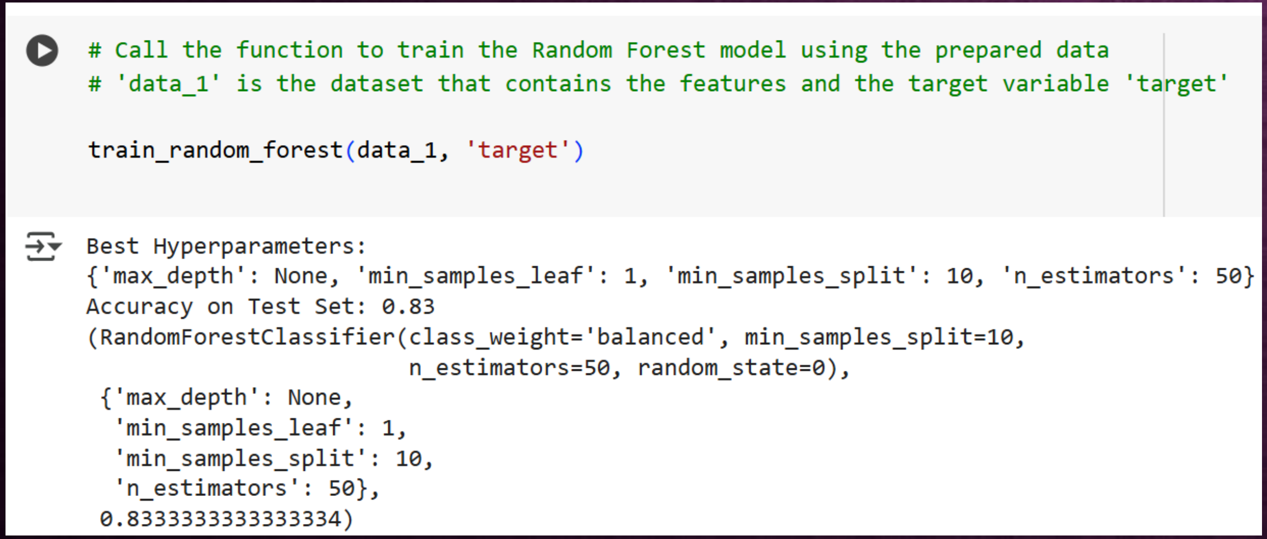
83% accuracy, stable, and good for simple predictions.

**2)XGB:**

83% accuracy, better at complex pattern recognition.

**3)Key Findings:**

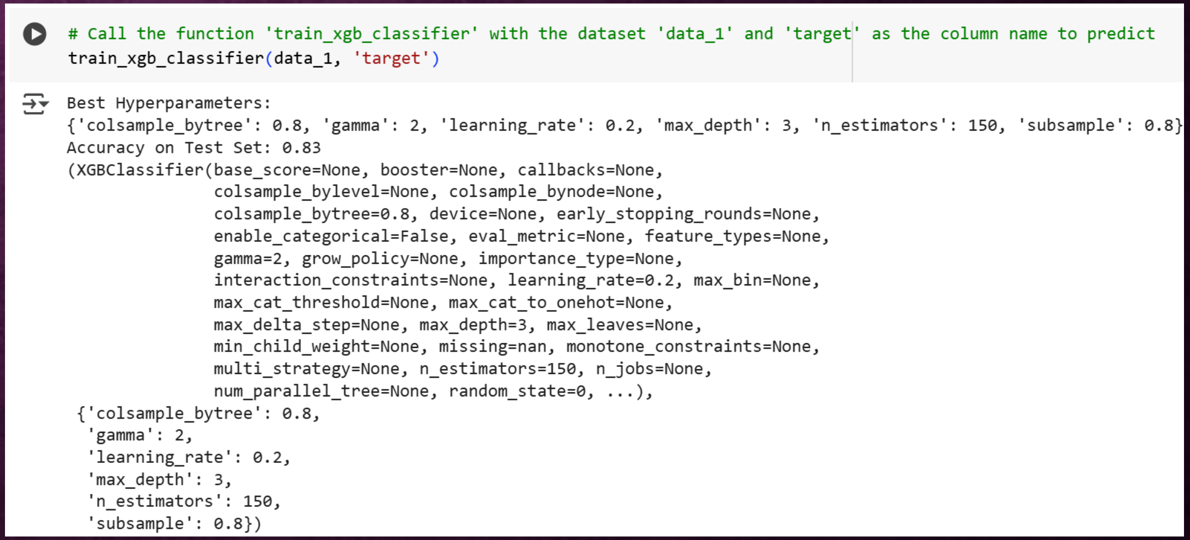
Blood pressure, chest pain type, and age are critical in predictions.



```
# Call the function to train the Random Forest model using the prepared data
# 'data_1' is the dataset that contains the features and the target variable 'target'

train_random_forest(data_1, 'target')
```

```
Best Hyperparameters:
{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 50}
Accuracy on Test Set: 0.83
(RandomForestClassifier(class_weight='balanced', min_samples_split=10,
                        n_estimators=50, random_state=0),
 {'max_depth': None,
  'min_samples_leaf': 1,
  'min_samples_split': 10,
  'n_estimators': 50},
 0.8333333333333334)
```



```
# Call the function 'train_xgb_classifier' with the dataset 'data_1' and 'target' as the column name to predict
train_xgb_classifier(data_1, 'target')
```

```
Best Hyperparameters:
{'colsample_bytree': 0.8, 'gamma': 2, 'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 150, 'subsample': 0.8}
Accuracy on Test Set: 0.83
(XGBClassifier(base_score=None, booster=None, callbacks=None,
               colsample_bylevel=None, colsample_bynode=None,
               colsample_bytree=0.8, device=None, early_stopping_rounds=None,
               enable_categorical=False, eval_metric=None, feature_types=None,
               gamma=2, grow_policy=None, importance_type=None,
               interaction_constraints=None, learning_rate=0.2, max_bin=None,
               max_cat_threshold=None, max_cat_to_onehot=None,
               max_delta_step=None, max_depth=3, max_leaves=None,
               min_child_weight=None, missing=nan, monotone_constraints=None,
               multi_strategy=None, n_estimators=150, n_jobs=None,
               num_parallel_tree=None, random_state=0, ...),
 {'colsample_bytree': 0.8,
  'gamma': 2,
  'learning_rate': 0.2,
  'max_depth': 3,
  'n_estimators': 150,
  'subsample': 0.8})
```

# 17)CHALLENGES

**1)Data Imbalance:**

1)More cases of no heart disease than those with heart disease.
2)Imagine a classroom with **95 healthy students** and **5 sick students**.
3)If a model predicts that **everyone is healthy**, it will still be **95% accurate**, but it completely missed the 5 sick students.
4)This happens when one group (like "no heart disease" cases) is much larger than the other (like "heart disease" cases).

**2)Model Interpretability:**

1)Understanding how the model makes decisions is difficult.
2)Some models, like **Random Forest** or **XGBoost**, are like "black boxes."
3)This means they predict well but don't clearly explain *why* they made a certain prediction.
4)For example, if the model says **"This person has heart disease"**, it's hard to understand which factors (like age, blood pressure, etc.) influenced that decision.

# 18)FUTURE SCOPE

1)We can try more models like Logistic Regression, SVM, or Deep Learning.

2)Combining models using Voting or Stacking can give better results.

3)Adding more patient data will improve accuracy.

4)We can build a web or mobile app for real-time heart risk check.

5)The system can connect with hospitals to give live predictions.

6)We can use Explainable AI to show why a result was given.

# 19)CONCLUSION

1)We used two machine learning models — Random Forest and XGB Classifier — to predict heart disease.

2)Both models gave good results, but XGB was a little better in accuracy.

3)Chest pain type, age, and blood pressure were the most important health factors.

4)These models help doctors find heart problems early and make better decisions.

5)Random Forest is simple and stable, while XGB gives more accurate results.

6)This system can save lives by giving faster and smarter heart disease predictions.