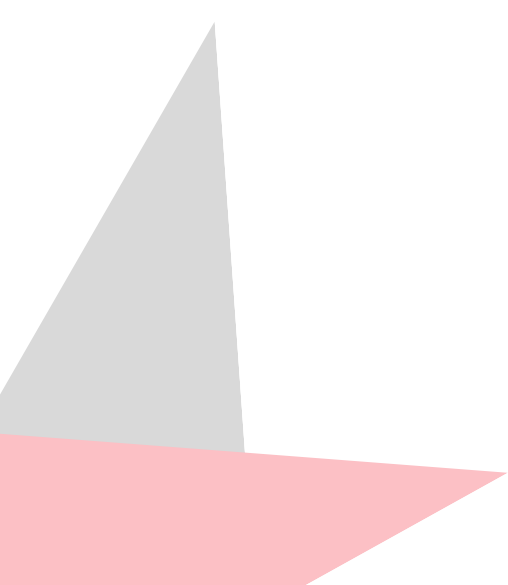




HEART DISEASE PREDICTION



Presented By
AARYA YOGESH PAKHALE
IIT KHARAGPUR

TABLE OF CONTENTS

- 1. INTRODUCTION
- 2. OBJECTIVES
- 3. DATA DESCRIPTION
- 4. MODEL APPROACH
- 5. EXPLORATORY DATA ANALYSIS
- 6. DATA PREPROCESSING
- 7. MODEL TRAINING
 - 7.1 XGBOOST
 - 7.2 RANDOM FOREST
 - 7.3 GRADIENT BOOSTING
- 8. RESULTS
 - 8.1 ERROR MEASUREMENT METRICS

INTRODUCTION

Heart disease remains a leading cause of mortality worldwide, representing a significant public health concern. The ability to accurately predict the likelihood of heart disease in individuals is crucial for timely intervention and prevention strategies.

This report presents the development and evaluation of a heart disease prediction model leveraging machine learning algorithms. The primary objective of this study is to construct a reliable predictive model capable of identifying individuals at high risk of developing heart disease based on their demographic, clinical, and lifestyle factors.

Through the analysis of a comprehensive dataset encompassing various patient attributes, including age, gender, blood pressure, cholesterol levels, and lifestyle habits, this study aims to uncover significant predictors of heart disease and develop a robust predictive framework.

The model's performance will be assessed based on key metrics such as Accuracy, F1 Score, Precision, Recall and Support providing insights into its effectiveness in clinical application.

The findings of this study have the potential to enhance risk stratification strategies and inform personalized healthcare interventions aimed at reducing the burden of heart disease. By harnessing the power of machine learning, we endeavor to contribute to the ongoing efforts in preventive cardiology and improve patient outcomes through early detection and targeted interventions.

OBJECTIVE

The primary objective of this study is to develop and evaluate a predictive model for heart disease, leveraging machine learning techniques. Specifically, the objectives include:

1. **Model Development:** Constructing a predictive model using machine learning algorithms to identify individuals at risk of developing heart disease based on a comprehensive set of demographic, clinical, and lifestyle factors.
2. **Model Evaluation:** Assessing the performance of the developed model using appropriate evaluation metrics such as Accuracy, F1 Score, Precision, Recall and Support.
3. **Clinical Utility:** Evaluating the clinical utility of the predictive model by examining its potential for risk stratification, early detection, and personalized intervention strategies in the context of heart disease management.
4. **Impact Assessment:** Investigating the potential impact of the developed model on healthcare outcomes, including its ability to facilitate timely interventions, reduce healthcare costs, and improve patient quality of life.

DATA DESCRIPTION

Attribute Information:

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholestoral in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest

- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

MODEL APPROACH

PROBLEM STATEMENT

DATASET

EXPLORATORY DATA ANALYSIS

DATA PREPROCESSING

MODEL IMPLEMENTATION

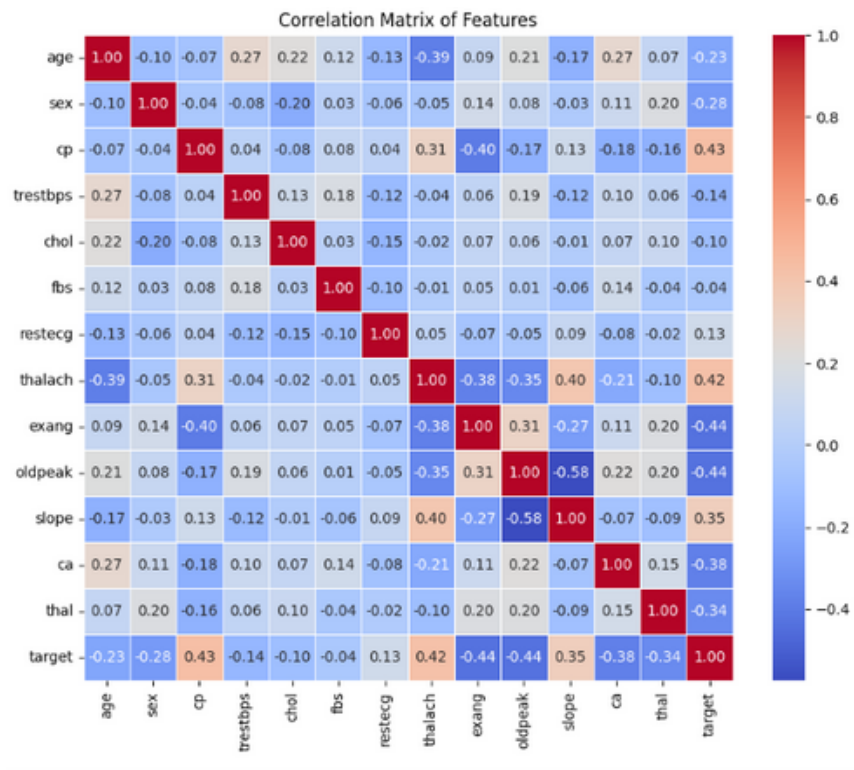
CLASSIFICATION MODELS

XGBOOST

RANDOM
FOREST

GRADIENT
BOOSTING

EXPLORATORY DATA ANALYSIS



This image presents a comprehensive Correlation Matrix showcasing the relationships between various features related to heart disease. At the heart of the matrix is a list of features including age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol levels (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), oldpeak (depression induced by exercise relative to rest), the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), and a blood disorder called thalassemia (thal), all correlated with the target variable, which indicates the presence of heart disease.

Each feature is analyzed for its correlation with the others, denoted by a range of values from -1.0 to 1.0, where 1.0 signifies a perfect positive correlation, -1.0 indicates a perfect negative correlation, and values close to 0 suggest no correlation. Positive values, for instance, between the target and chest pain type (cp, 0.43) suggest that as the severity of chest pain increases, the likelihood of heart disease also increases. On the other hand, negative values like the correlation between the target and maximum heart rate achieved (thalach, -0.44) imply an inverse relationship; as the maximum heart rate increases, the probability of heart disease decreases.

Based on this correlation matrix, we will consider features with a strong positive or negative correlation with the target variable. Features like age, trestbps, chol, thalach and oldpeak could be good candidates based on their correlation values.

Now , I have plot the distribution plot, QQ (Quantile-Quantile) plot and Box plot each of these features to handle the outliers and to compare the practical data with the theoretical data

Distribution Plot:

The Distribution Plot provides information on how the feature data is spread out across different values, indicating the frequency of each feature group within the dataset.

QQ Plot:

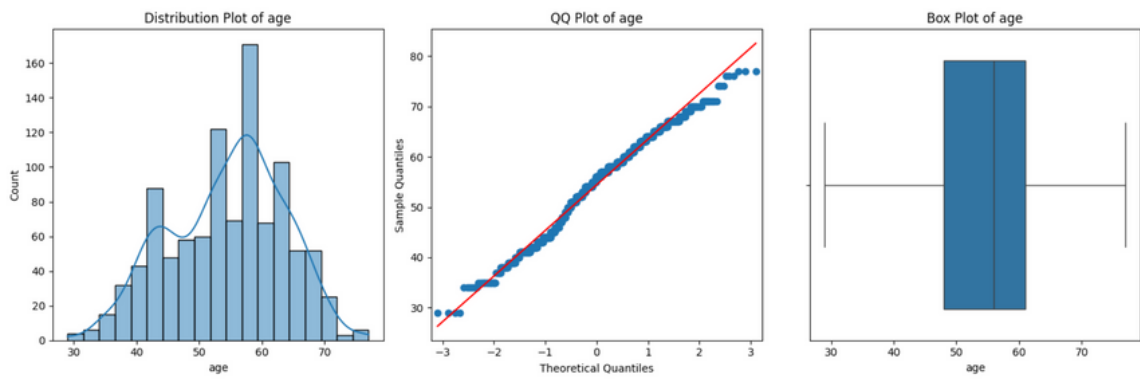
The QQ Plot allows us to compare the actual distribution of feature data to a theoretical distribution, helping us assess if the data follows a specific pattern or if there are deviations from expected values.

Box Plot:

The Box Plot gives us a visual summary of the central tendency (median) and variability (interquartile range) of the feature data, as well as any potential outliers that may exist.

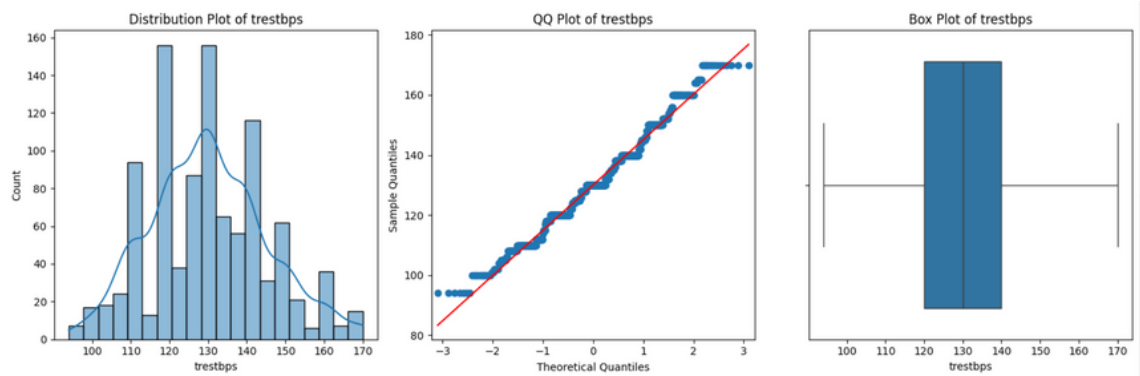
These Graphs were obtained after removing the outliers, graphs obtained before removing the outliers are not shown.

For age,



Age is a critical risk factor for heart disease, with its prevalence increasing as individuals get older. This relationship is influenced by cumulative exposure to risk factors, physiological changes, lifestyle habits, comorbidities, genetic predisposition, and vascular aging. As age increases risk of getting a heart disease increases.

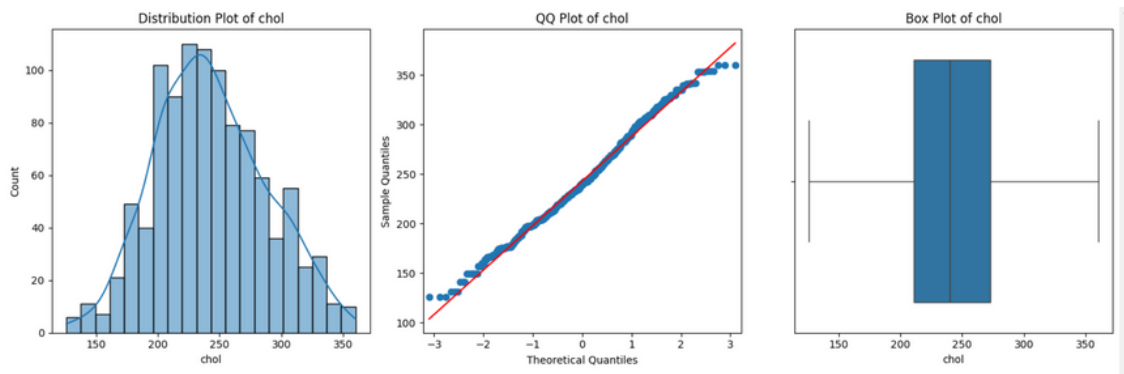
For trestbps,



Trest BPS (resting blood pressure) serves as an important indicator of heart disease risk. Elevated resting blood pressure levels are associated with an increased likelihood of developing cardiovascular conditions such as hypertension, coronary artery disease, and heart failure. Monitoring resting blood pressure levels is crucial for early detection and management of heart disease, as it reflects the strain on the heart and blood vessels during periods of rest.

Figure shows a normally distributed histogram and a even box plot. We can see that the outliers are removed.

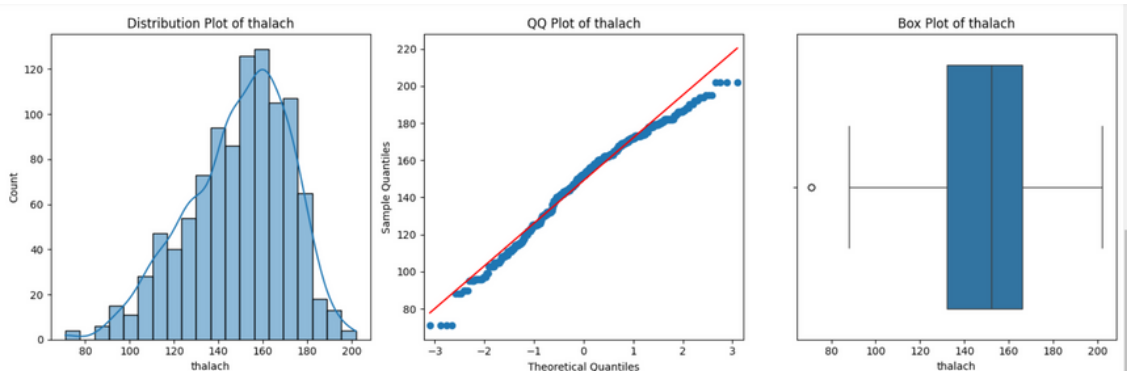
For chol,



Cholesterol levels play a significant role in the development of heart disease. Elevated levels of cholesterol, particularly low-density lipoprotein (LDL) cholesterol, contribute to the buildup of plaque in the arteries, a condition known as atherosclerosis. This narrowing and hardening of the arteries can restrict blood flow to the heart, leading to various cardiovascular conditions such as coronary artery disease, heart attack, and stroke. Lowering LDL cholesterol levels can significantly reduce the risk of heart disease and improve overall cardiovascular health.

This shows a normally distributed histogram and an even box plot.

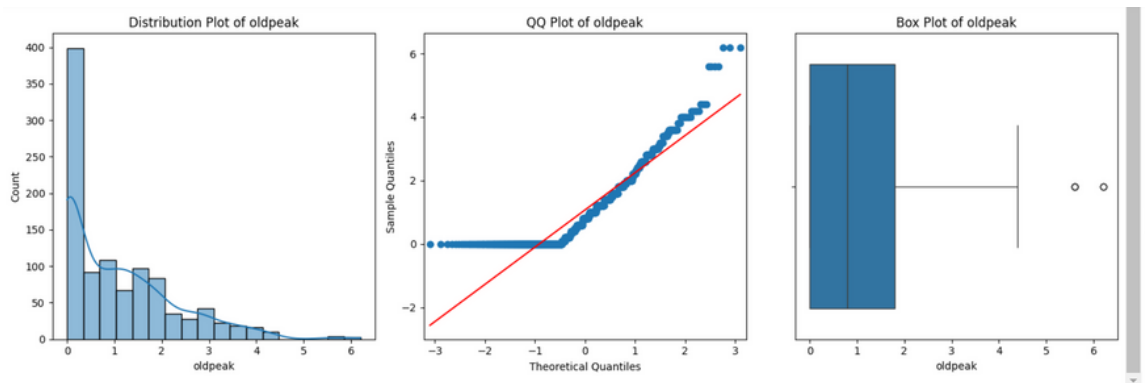
For thalach,



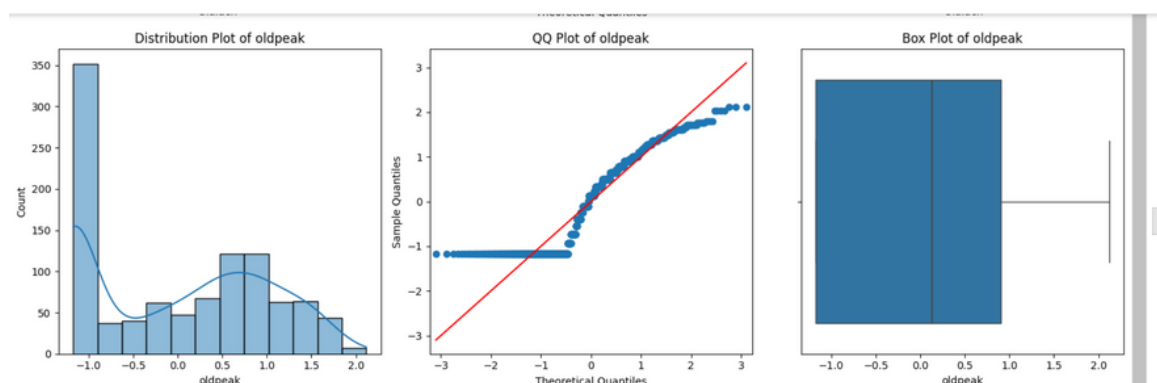
Thalach, or maximum heart rate achieved during exercise, is an important indicator of cardiovascular health. Individuals with a lower maximum heart rate may have reduced cardiovascular fitness, which can increase their risk of developing heart disease. A higher thalach is generally associated with better cardiovascular health and a lower risk of heart disease.

This shows a normally distributed histogram and an even box plot. We can see that the outliers are removed.

For oldpeak,



Oldpeak, also known as ST depression induced by exercise relative to rest, is a measure used during exercise stress testing to assess cardiac function and detect coronary artery disease. Higher levels of oldpeak indicate greater ischemia or reduced blood flow to the heart during physical activity, which can be indicative of underlying heart disease. Individuals with elevated oldpeak values are at increased risk of experiencing cardiovascular events such as heart attack or angina. Monitoring oldpeak levels during exercise stress testing provides valuable diagnostic information and helps guide treatment decisions for individuals at risk of heart disease. The data is right skewed and hence, needs normalization.



Graph after normalization

DATA PREPROCESSING

After assessing the plots, we found that the outliers had to be removed and the Old peak function had to be normalized.

The outliers are removed using the inter-quartile method.

```
1 #Handling Outliers
2
3 # Calculate the IQR for trestbps and chol
4 Q1_trestbps = df['trestbps'].quantile(0.25)
5 Q3_trestbps = df['trestbps'].quantile(0.75)
6 IQR_trestbps = Q3_trestbps - Q1_trestbps
7 upper_bound_trestbps = Q3_trestbps + 1.5 * IQR_trestbps
8
9 Q1_chol = df['chol'].quantile(0.25)
10 Q3_chol = df['chol'].quantile(0.75)
11 IQR_chol = Q3_chol - Q1_chol
12 upper_bound_chol = Q3_chol + 1.5 * IQR_chol
13
14 # Calculate the mean for trestbps and chol excluding upper bound outliers
15 mean_trestbps = df.loc[df['trestbps'] <= upper_bound_trestbps, 'trestbps'].mean()
16 mean_chol = df.loc[df['chol'] <= upper_bound_chol, 'chol'].mean()
17
18 df.loc[df['trestbps'] > upper_bound_trestbps, 'trestbps'] = mean_trestbps
19 df.loc[df['chol'] > upper_bound_chol, 'chol'] = mean_chol
```

The oldpeak feature is normalized using the fit_transform function.

```
1 from sklearn.preprocessing import PowerTransformer
2
3 pt = PowerTransformer()
4
5
6 oldpeak_column_2d = df['oldpeak'].values.reshape(-1, 1)
7
8 # Fit the transformer to the 'oldpeak' column and transform it
9 oldpeak_transformed = pt.fit_transform(oldpeak_column_2d)
10
11 # Convert the transformed column back to a 1D array
12 oldpeak_transformed = oldpeak_transformed.flatten()
13
14 # Replace the original 'oldpeak' column with the transformed values
15 df['oldpeak'] = oldpeak_transformed
```

MODEL TRAINING

1. XGBoost (Extreme Gradient Boosting):

- **Algorithm:** XGBoost is an implementation of gradient boosting machines designed for speed and performance. It builds an ensemble of weak decision trees sequentially, where each tree corrects the errors of the previous ones. It employs a regularization term in the objective function to control overfitting.
- **Key Features:**
 - **Regularization:** XGBoost includes L1 and L2 regularization terms in its objective function to control model complexity and prevent overfitting.
 - **Gradient Boosting:** It uses gradient boosting techniques to optimize the model's performance, iteratively improving predictions by minimizing a differentiable loss function.
 - **Parallel Processing:** XGBoost is designed for efficient parallel processing, making it scalable and suitable for large datasets.
- **Advantages:**
 - High performance and speed
 - Regularization to control overfitting
 - Feature importance estimation
- **Limitations:**
 - Sensitive to hyperparameters
 - Requires tuning for optimal performance
- **Best parameters for XGBoost Classifier:**
{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}
- **Classification Report:** precision recall f1-score support

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.97 | 1.00 | 0.99 | 102 |
| 1 | 1.00 | 0.97 | 0.99 | 103 |
| accuracy | | | 0.99 | 205 |
| macro avg | 0.99 | 0.99 | 0.99 | 205 |
| weighted avg | 0.99 | 0.99 | 0.99 | 205 |

2 . Random Forest:

- **Algorithm:** Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It introduces randomness by training each tree on a random subset of the features and bootstrapped samples of the dataset.
- **Key Features:**
 - Ensemble Learning: Random Forest aggregates predictions from multiple decision trees to make final predictions, reducing the risk of overfitting and improving generalization.
 - Feature Importance: It can provide estimates of feature importance based on how much each feature contributes to reducing impurity across all trees.
 - Robust to Overfitting: Random Forest is less prone to overfitting compared to individual decision trees due to the averaging effect of multiple trees.
- **Advantages:**
 - Robust to overfitting
 - Handles high-dimensional data well
 - Efficient for large datasets
- **Limitations:**
 - Less interpretable compared to single decision trees
 - May not perform as well as gradient boosting for structured data
- **Best parameters for Random Forest:**
`{'max_depth': 7, 'max_features': 'sqrt', 'n_estimators': 100}`
- **Classification Report:**

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.97 | 0.93 | 0.95 | 102 |
| 1 | 0.93 | 0.97 | 0.95 | 103 |
| accuracy | | | 0.95 | 205 |
| macro avg | 0.95 | 0.95 | 0.95 | 205 |
| weighted avg | 0.95 | 0.95 | 0.95 | 205 |

- **Gradient Boosting:**

- **Algorithm:** Gradient Boosting is an ensemble technique that sequentially builds decision trees to minimize a loss function. Unlike Random Forest, which builds trees independently, gradient boosting builds trees sequentially, with each tree learning from the errors made by the previous ones.
- **Key Features:**
 - **Sequential Learning:** Gradient Boosting builds trees sequentially, with each tree focusing on correcting the errors of the previous ones, leading to improved model performance.
 - **Gradient Descent:** It uses gradient descent optimization to minimize the loss function, making small adjustments to the model's predictions at each iteration.
 - **Weak Learners:** Gradient Boosting typically uses shallow decision trees as weak learners, which are combined to form a strong predictive model.
- **Advantages:**
 - High predictive accuracy
 - Handles heterogeneous data types well
 - Generally robust to overfitting
- **Limitations:**
 - Sensitive to hyperparameters
 - Longer training time compared to Random Forest
- **Best parameters for Gradient Boosting:**
{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}
- **Classification Report:**

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.97 | 1.00 | 0.99 | 102 |
| 1 | 1.00 | 0.97 | 0.99 | 103 |
| accuracy | | | 0.99 | 205 |
| macro avg | 0.99 | 0.99 | 0.99 | 205 |
| weighted avg | 0.99 | 0.99 | 0.99 | 205 |

RESULTS

The best model is **XGBoost** or **Gradient boosting**. Both these models give an accuracy of **99 percent** and hence can be deployed safely for real-time use.

ERROR MEASUREMENT

METRICS

Error measurement metrics such as precision, recall, F1 score, and support provide insights into the performance of a model in terms of its ability to correctly classify instances of each class.

- **Precision:** Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive, including both true positives and false positives. It quantifies the accuracy of the positive predictions made by the model.

$$\text{Precision} = (\text{True Positives}) / (\text{True Positives} + \text{False Positives})$$

- **Recall (Sensitivity):** Recall, also known as sensitivity, measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances in the dataset. It quantifies the model's ability to capture all positive instances.

$$\text{Recall} = (\text{True Positives}) / (\text{True Positives} + \text{False Negatives})$$

- **F1 Score:** The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both precision and recall. It gives equal weight to precision and recall and is particularly useful when there is an uneven class distribution or when both precision and recall are important.

$$\text{F1 Score} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

- **Support:** Support refers to the number of actual occurrences of each class in the dataset. It provides context for the precision, recall, and F1 score metrics by indicating the distribution of instances across different classes. Support helps interpret the performance metrics in relation to the dataset's class distribution.