Heart Disease Data Analysis & Prediction Model

Overview

- Heart disease is a leading cause of mortality globally, emphasizing the importance of early detection.
- The dataset includes features like age, sex, chest pain type, cholesterol levels, blood pressure, and other clinical indicators.
- Data analysis and machine learning techniques are used to identify patterns and predict heart disease risk.
- The project aims to improve healthcare decision-making and patient outcomes.

Methodology

Data Preprocessing:

- Handle missing values, outliers, and inconsistent data.
- Normalize or standardize continuous variables.
- Encode categorical variables using label encoding or one-hot encoding.

Exploratory Data Analysis (EDA):

- Visualize data distributions and relationships.
- Identify trends and patterns in the dataset.
- Correlation analysis to determine feature importance.

Feature Selection:

- Use statistical methods or algorithms like Recursive Feature Elimination (RFE) to select the most relevant features.

Model Building:

- Train machine learning models like Logistic Regression, Random Forest, and XGBoost.

Model Evaluation:

- Assess models using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Deployment:

- Deploy the model as a web application or integrate it into a healthcare system for real-time predictions.

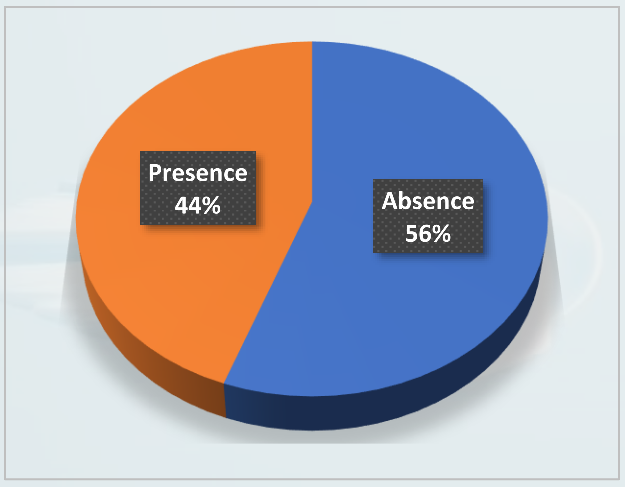
Heart Disease Distribution

Insights:

- The number of individuals without heart disease (**150**) is higher than those with heart disease (**120**).
- This suggests a slightly lower prevalence of heart disease in the dataset.

Distribution Summary:

- Approximately 44.4% (120 out of 270) of the dataset represents individuals with heart disease.
- The remaining 55.6% (150 out of 270) represents individuals without heart disease.



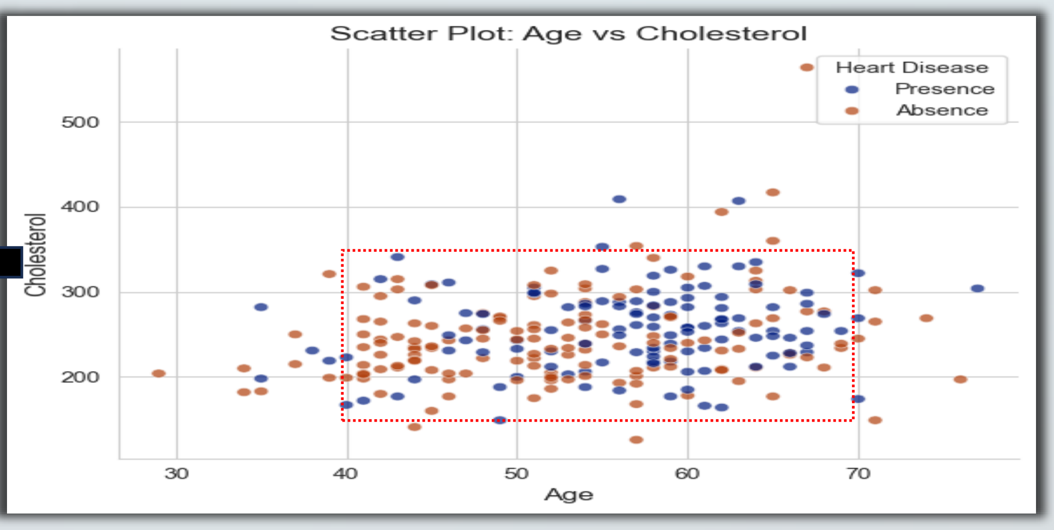
SMOTE Classification Results

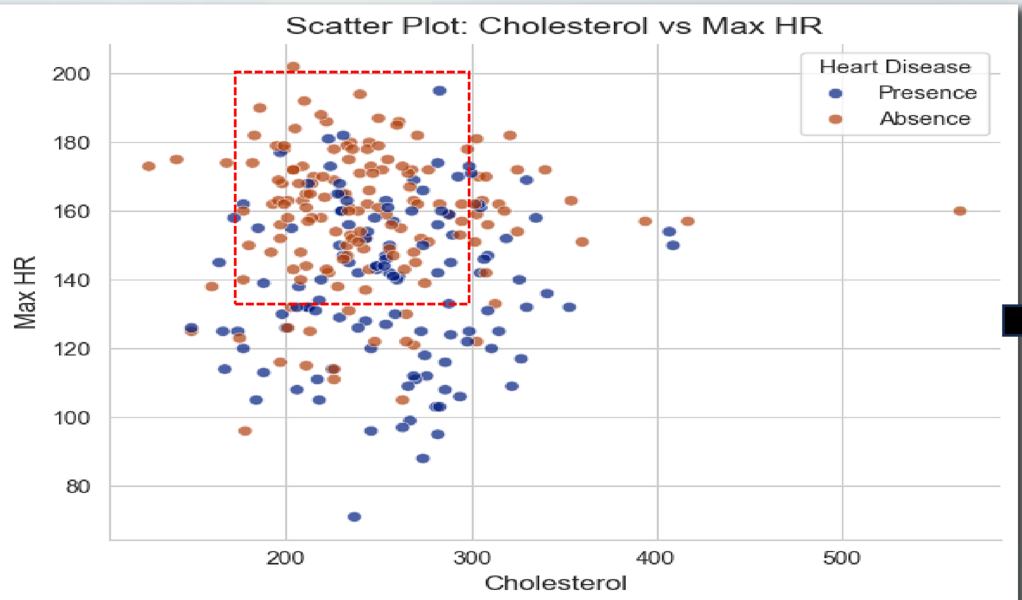
- The model performs well overall, achieving high precision, recall, and F1-scores for both classes.
- SMOTE effectively balances the dataset, improving the model's ability to handle the minority class ("Presence").

```
Accuracy with SMOTE: 0.8888888888888888
Classification Report:
              precision    recall  f1- score   support

    Absence      0.89      0.94      0.91        33
    Presence     0.89      0.81      0.85        21

    accuracy                         0.89        54
   macro avg     0.89      0.87      0.88        54
weighted avg     0.89      0.89      0.89        54
```

Age vs Cholesterol Distribution

- Cholesterol levels vary widely across all age groups but appear to cluster between **200 and 300** mg/dL for most individuals.
- There are a few outliers with cholesterol levels exceeding **400 mg/dL**.
- Most data points appear to be concentrated in the age range of **40 to 70 years**.
- There is significant overlap between the cholesterol levels of those with and without heart disease. This suggests that cholesterol alone may not be a strong indicator of heart disease when age is considered.



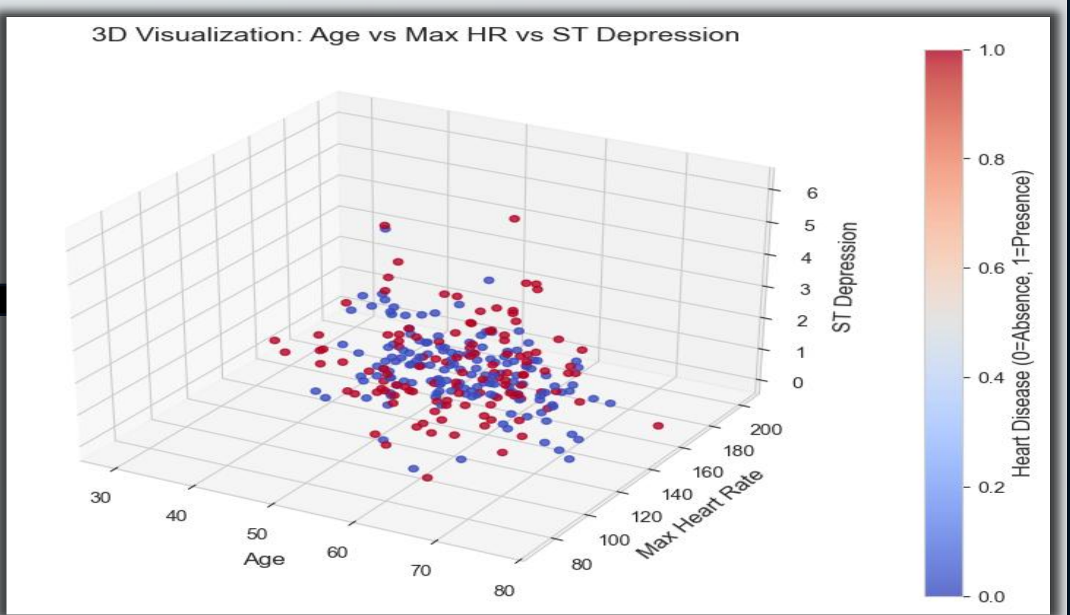


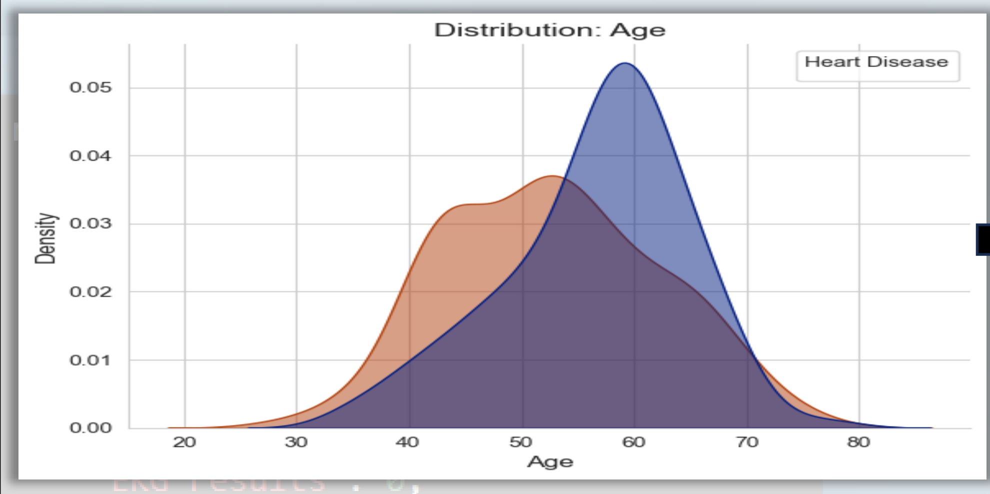
Cholesterol vs Max HR Distribution

- Cholesterol levels for both groups are primarily concentrated between **200 and 300 mg/dL**.
- There are a few individuals with cholesterol levels exceeding **400 mg/dL**, but these are rare outliers.
- Maximum heart rate (Max HR) values are generally between **80 and 200 bpm** for both groups.
- There is a noticeable cluster of data points around **Max HR = 140–180 bpm**, indicating that most individuals achieve their heart rate in this range.
- A few outliers exist where individuals with high cholesterol and relatively high Max HR values are present in both groups.

Relationship between Age, Max Heart Rate, ST Depression, and Heart Disease

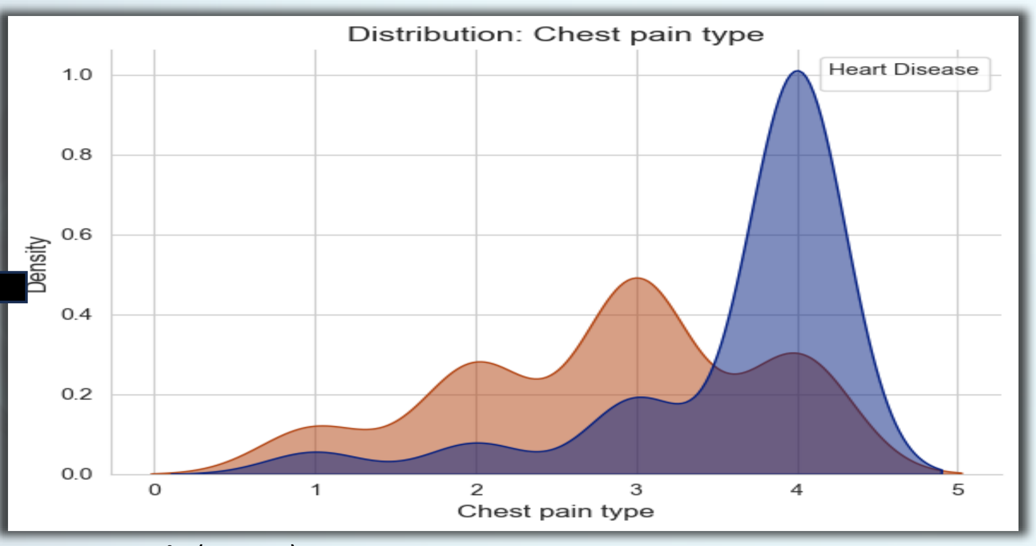
- **Age vs. Max Heart Rate**: As age increases, maximum heart rate generally decreases.
- **ST Depression vs. Heart Disease**: Higher ST depression values are strongly linked to a higher prevalence of heart disease.
- **Age vs. ST Depression**: ST depression slightly increases with age, but the trend is not very strong.
- **Max Heart Rate vs. ST Depression**: No clear relationship is observed between maximum heart rate and ST depression.





Distribution of Age and Heart Disease

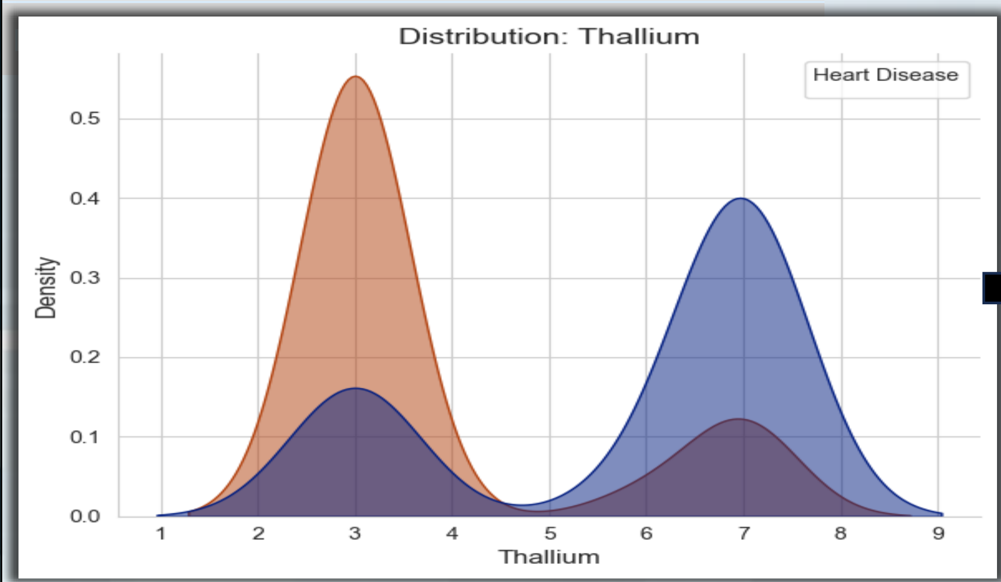
- Overall age distribution is **right-skewed**, with a **peak around 60 years**.
- Age distribution for individuals with heart disease (**blue line**) is also **right-skewed**, peaking slightly earlier, around **55-60 years**.
- The blue line overlaps with the orange line at **younger ages**, indicating a **lower prevalence of heart disease** in this group.
- As age increases, the blue line becomes more prominent, showing an **increased risk of heart disease** in the middle-aged and older population.
- **Heart disease risk increases with age**, particularly after **55 years**.



Distribution of Chest Pain Type and Heart Disease

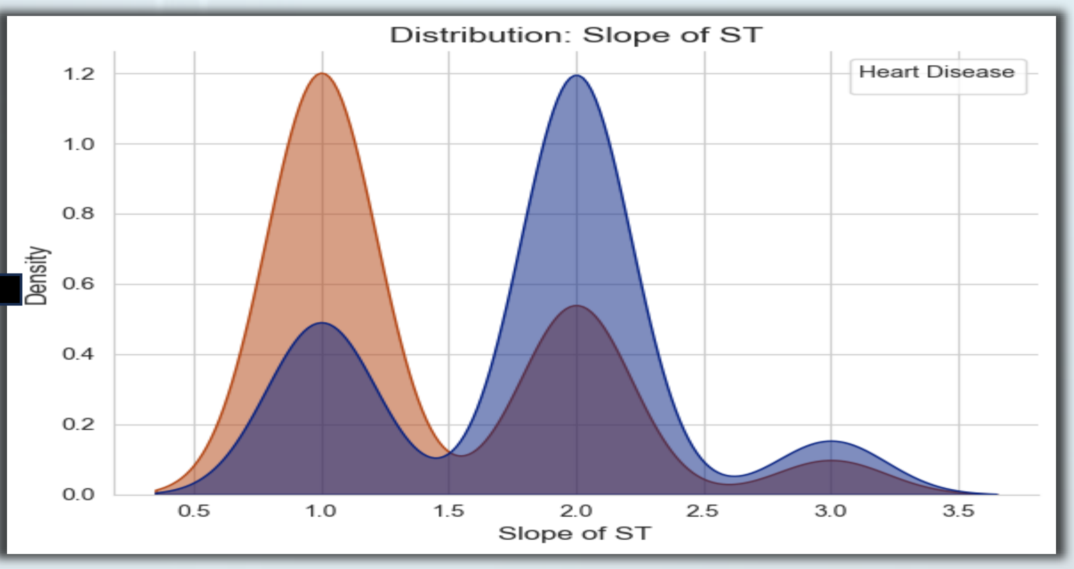
- The overall distribution is **multi-modal**, with peaks around chest pain types **2, 3, and 4**.
- The distribution for heart disease cases is also **multi-modal** but shows a **distinct peak at type 4**, indicating a higher likelihood of heart disease.
- The blue line overlaps with the orange line for chest pain types **2 and 3**, suggesting a **lower prevalence of heart disease** for these types.
- For chest pain type **4**, the blue line is significantly higher than the orange line, indicating a **strong correlation** with heart disease.
- Chest pain type **4** is a **strong indicator** of heart disease compared to other types.

NOTE:- Typical Angina (Type 1)| **Atypical Angina** (Type 2) **| Non-Anginal Pain** (Type 3) **| Asymptomatic** (Type 4).



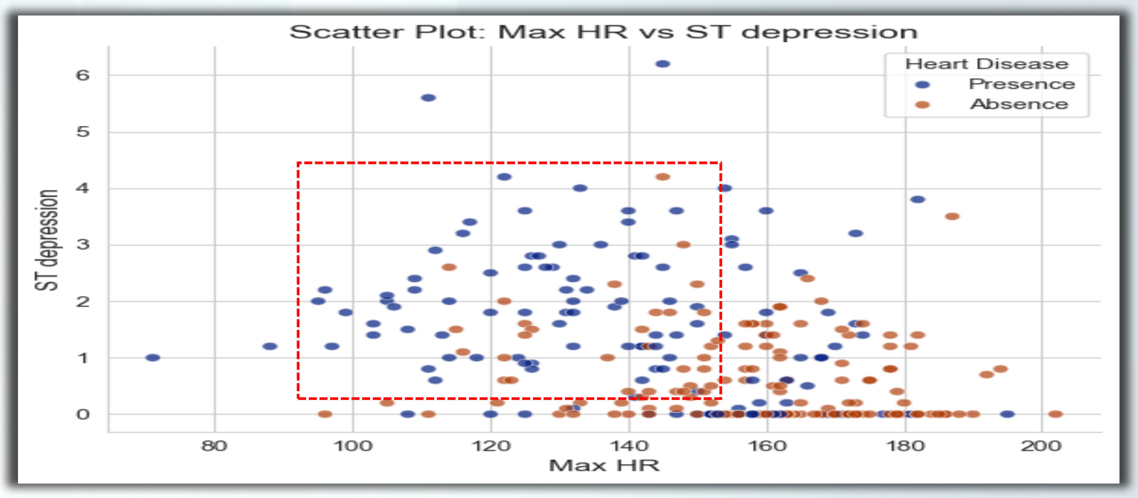
Distribution of Thallium and Heart Disease

- The distribution is bimodal, with peaks around 3 and 7.
- For individuals with heart disease (**blue line**), the distribution is also **bimodal**.
- A more pronounced peak is observed around **7**, suggesting a higher likelihood of heart disease.
- The blue line overlaps with the orange line at **lower thallium values**, indicating a **lower prevalence** of heart disease in this range.
- For **higher thallium values (around 7)**, the blue line rises significantly above the orange line.
- This highlights a **strong association** between higher thallium values and heart disease.
- Thallium values closer to **7** are indicative of a higher risk of heart disease, while lower values show less association.



Distribution of Slope of ST and Heart Disease

- The slope of ST shows a **multi-modal distribution** with peaks at **1, 2, and 3**.
- For individuals with **heart disease** (blue line), the distribution is also multi-modal but has a **distinct peak at 2**.
- A **slope of ST = 2** is strongly associated with heart disease, as indicated by the higher blue line at this value.
- For **slope values = 1 and 3**, the blue line overlaps with the orange line, indicating a **lower likelihood** of heart disease.
- A **slope of 2** is a potential risk factor for heart disease, while slopes of 1 and 3 show less association.



Max HR vs ST Depression Distribution

- Max HR values range from **80 to 200 bpm**, with most data points clustered between **120 and 160 bpm.**
- ST depression values range from **0 to 6** but are concentrated below **2** for most individuals.
- There is a visible trend indicating that individuals with **lower Max HR and higher ST depression** are more likely to have heart disease.
- While there is overlap in the data, the combination of **low Max HR (<140 bpm)** and **ST depression (>1)** shows a stronger association with heart disease.
- Individuals with **Max HR >160 bpm** and **ST depression close to 0** are mostly without heart disease.

Heart Disease Prediction Model

```
joblib.dump(model, 'heart_disease_model.pkl')
print("Model saved as 'heart_disease_model.pkl'")
```

```
Model saved as 'heart_disease_model.pkl'
```

Model Testing

```
new_input = {
    'Age': 55,
    'Sex': 1,
    'Chest pain type': 3,
    'BP': 140,
    'Cholesterol': 250,
    'FBS over 120': 1,
    'EKG results': 0,
    'Max HR': 150,
    'Exercise angina': 1,
    'ST depression': 2.5,
    'Slope of ST': 2,
    'Number of vessels fluro': 0,
    'Thallium': 5  }
result = predict_heart_disease(new_input)
print("Predicted Heart Disease:",
      "Yes" if result == 1 else "No")
```

```
Predicted Heart Disease: No
```

Presented By:- Kumar Siddharth http://www.linkedin.com/in/kumar-siddharth-3858aa199%20%20