HeartWise

Predicting Tomorrow's Heart Health

Final Project Presentation - Data Driven Knowledge Discovery

Team 01 - Namitha Bhaskar, Dhairya Dutt

Background

Understanding Heart Disease:

- Heart disease is a leading cause of death globally, accounting for significant morbidity and mortality each year.

- Early diagnosis and prevention are crucial to reducing the global burden of cardiovascular diseases.

About the Dataset

- The dataset includes 14 key attributes associated with heart disease, such as age, sex, cholesterol levels, blood pressure, and more.
- Originally derived from the Cleveland Heart Disease database

Multivariate Dataset:

- The dataset combines numerical, categorical, and binary variables, making it ideal for testing a variety of predictive modeling techniques.
- Contains medical attributes that directly correlate with the likelihood of heart disease.

Current Challenges:

- Timely and accurate prediction of heart disease remains a challenge due to the complex relationships between various medical factors.

Motivation

- Significance of Early Detection
- Personalized Healthcare
- Research Opportunity
- Data-Driven Approach
- Real-World Impact

Preprocessing

Objective of Preprocessing:

To handle missing values and ensure data consistency and completeness before building predictive models.

How did we handle missing values?

Imputation using **Miss Forest Imputer**

How this works?

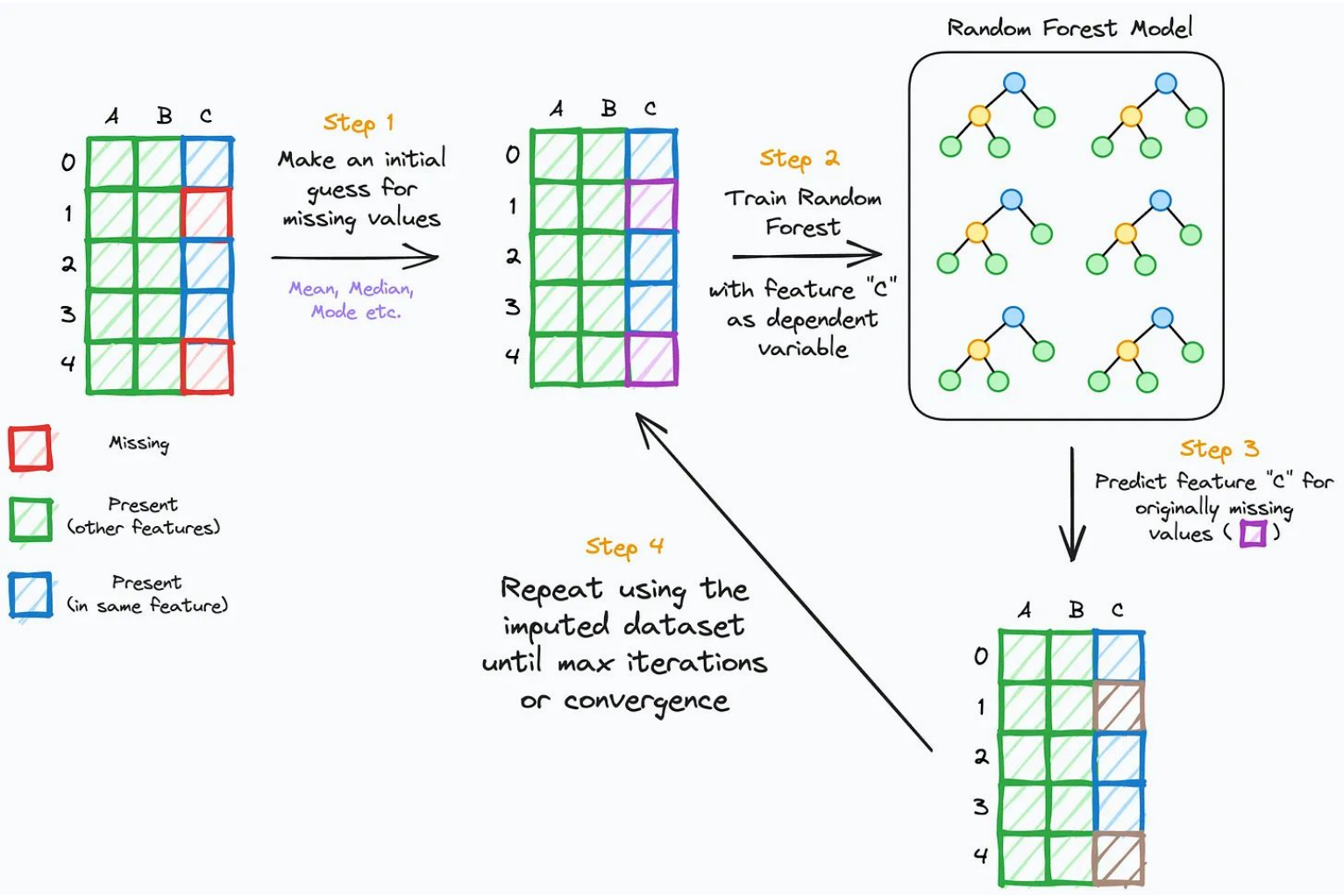
Step 1: To begin, impute the missing feature with a random guess — Mean, Median, etc.

Step 2: Model the missing feature using Random Forest.

Step 3: Impute ONLY originally missing values using Random Forest's prediction.

Step 4: Back to Step 2. Use the imputed dataset from Step 3 to train the next Random Forest model.

Step 5: Repeat until convergence (or max iterations).


A B C
0
1
2
3
4
Missing
Present (other features)
Present (in same feature)
Step 1
Make an initial guess for missing values
Mean, Median, Mode etc.
A B C
0
1
2
3
4
Step 2
Train Random Forest with feature "C" as dependent variable
Random Forest Model
Step 3
Predict feature "C" for originally missing values ( )
A B C
0
1
2
3
4
Step 4
Repeat using the imputed dataset until max iterations or convergence

What we did?

Based on Miss Forest Imputer, we did a custom implementation:

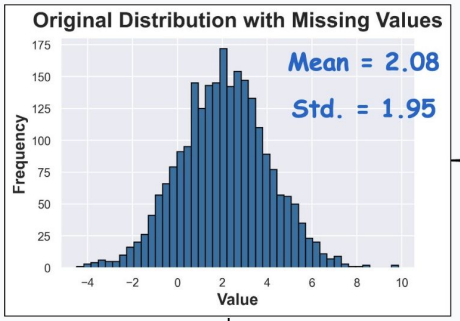
Custom Implementation

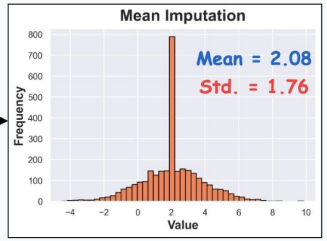
Used **RandomForestClassifier** for categorical data and **RandomForestRegressor** for continuous data.

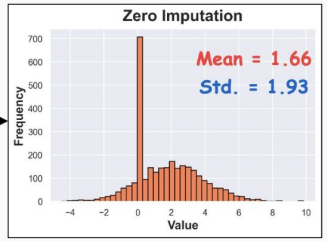
Missing values in each column are imputed independently by training a **Random Forest model** on the non-missing data.

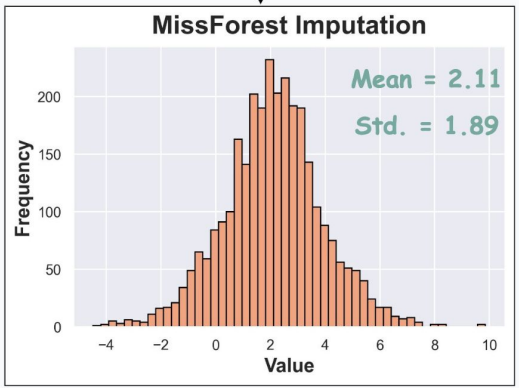
Iterative Imputer (IterativeImputer) is used to address dependencies among features

Avoid Filling Missing Values With Zero or Mean



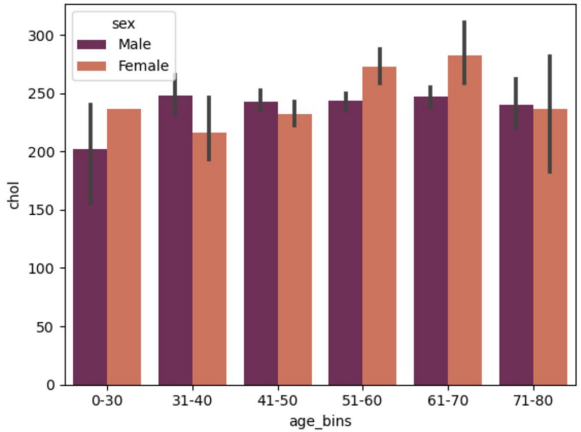


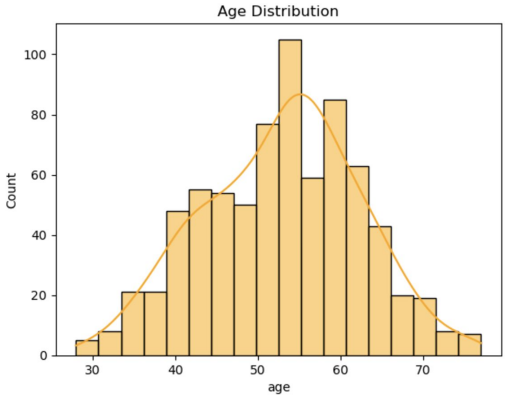


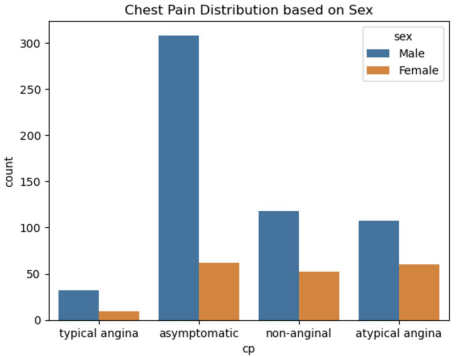


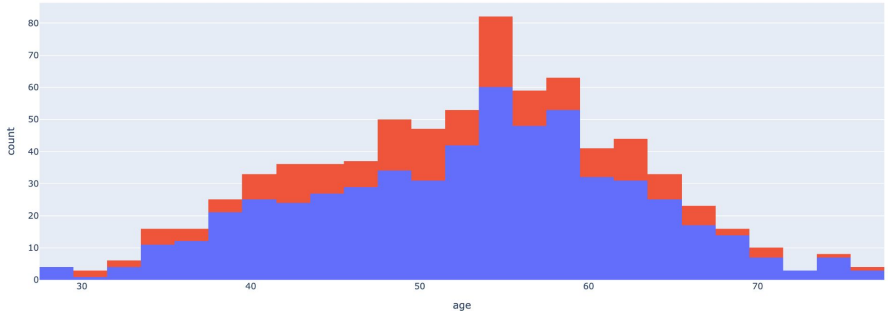
blog.DailyDoseofDS.com

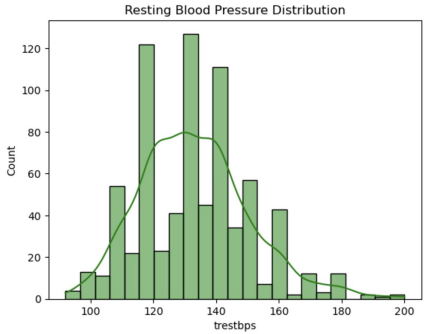
Exploratory Data Analysis

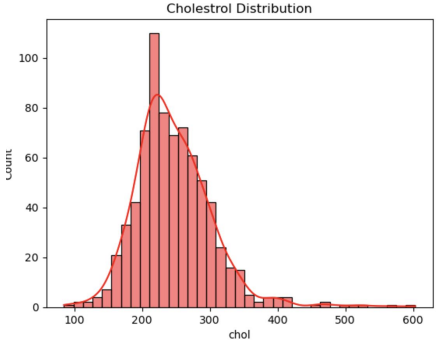


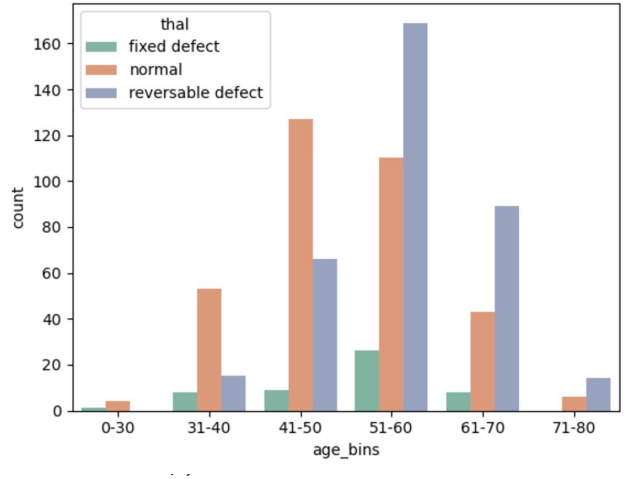


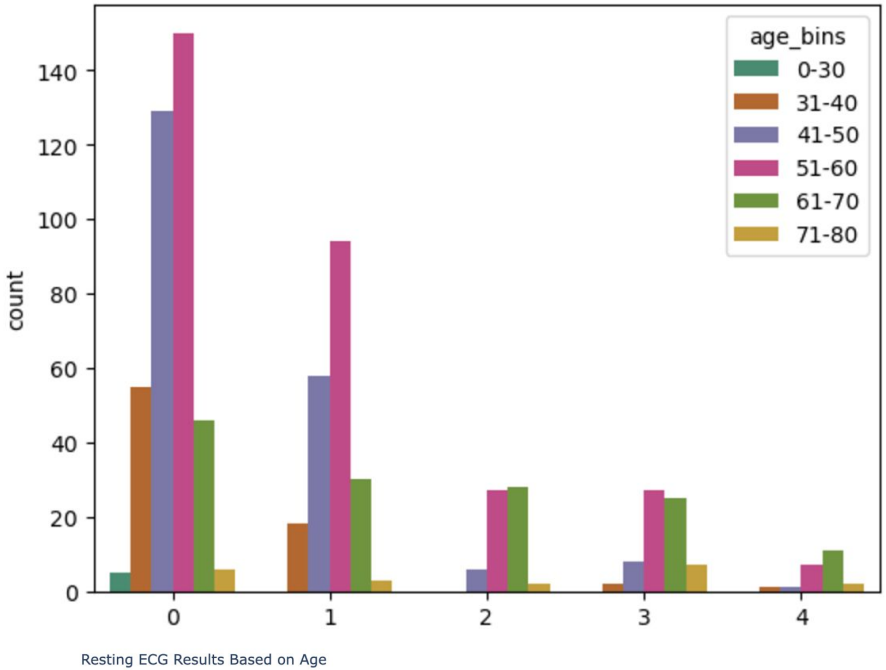


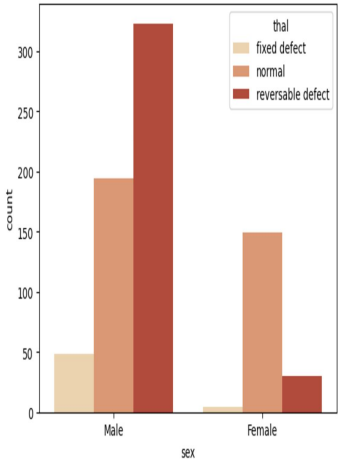


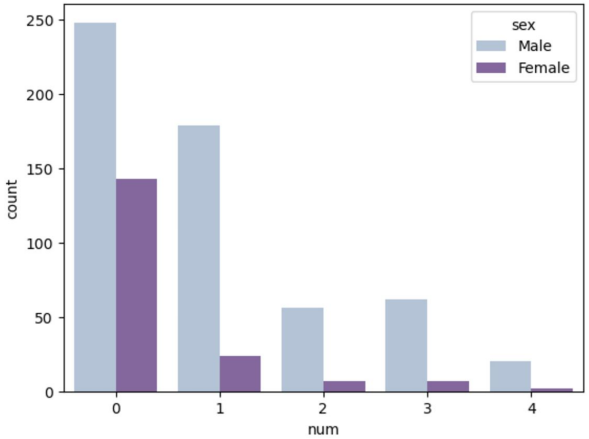




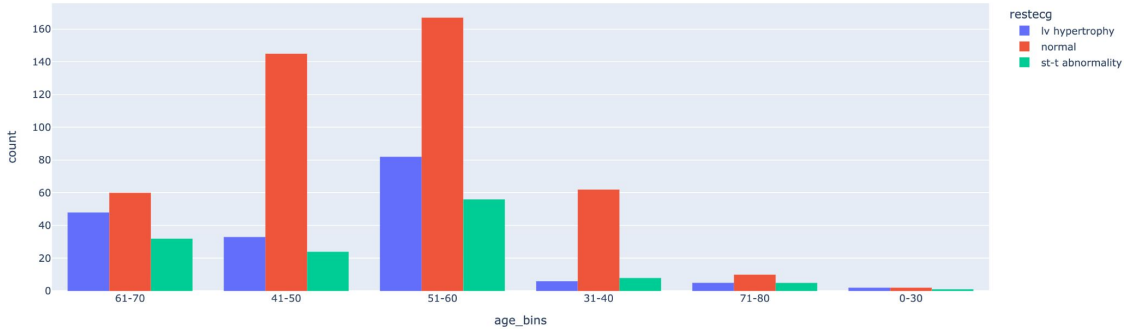








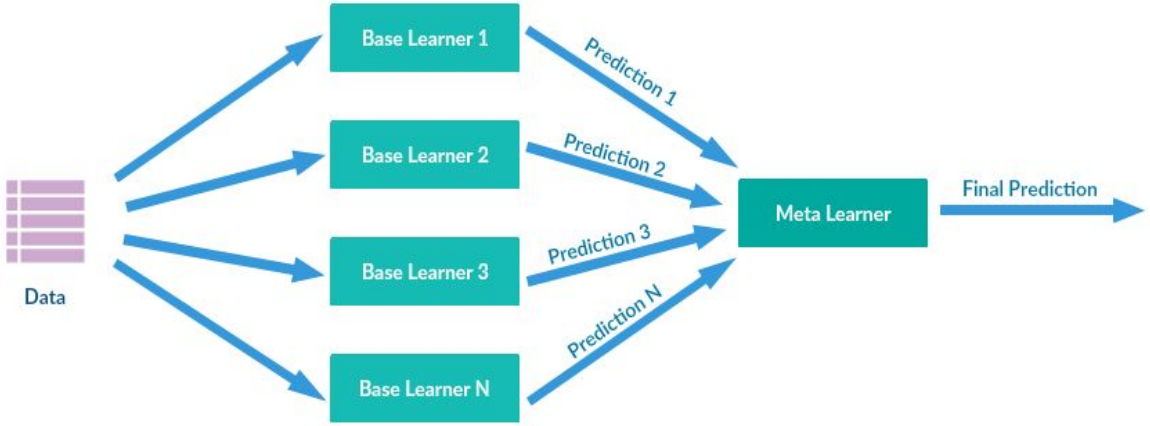
Resting ECG Results Based on Age



Model Training

Models	Val Accuracy	Macro Avg F1 score	Weighted Avg F1 score
Random Forest	0.823	0.84	0.84
Gradient Boosting	0.814	0.80	0.80
Logistic Regression	0.807	0.81	0.81
KNN	0.828	0.82	0.82
XG Boost	0.811	0.83	0.83

Stacking Classifier




```
Training Stacking Classifier...
Stacking Classifier - Test Accuracy: 0.8342245989304813
Stacking Classifier - Confusion Matrix:
 [[82 19]
 [12 74]]
Stacking Classifier - Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.81      0.84       101
           1       0.80      0.86      0.83        86

    accuracy                           0.83       187
   macro avg       0.83      0.84      0.83       187
weighted avg       0.84      0.83      0.83       187
```

Base Models - Random Forest, Gradient Boosting

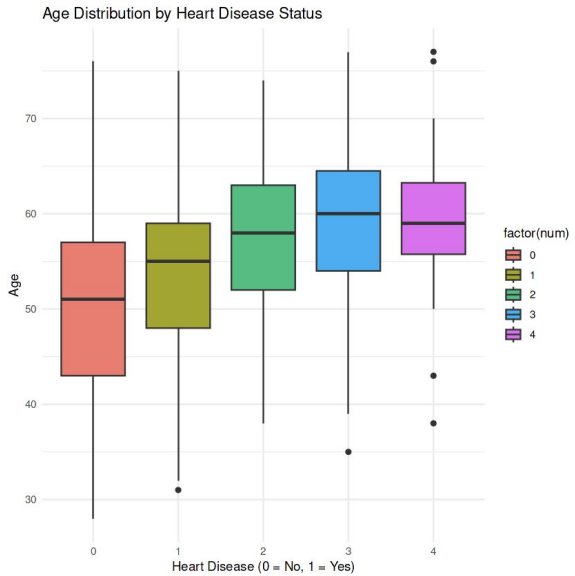
Final estimator - XG Boost

Some Inferences !!

Age Distribution by Heart Disease Status

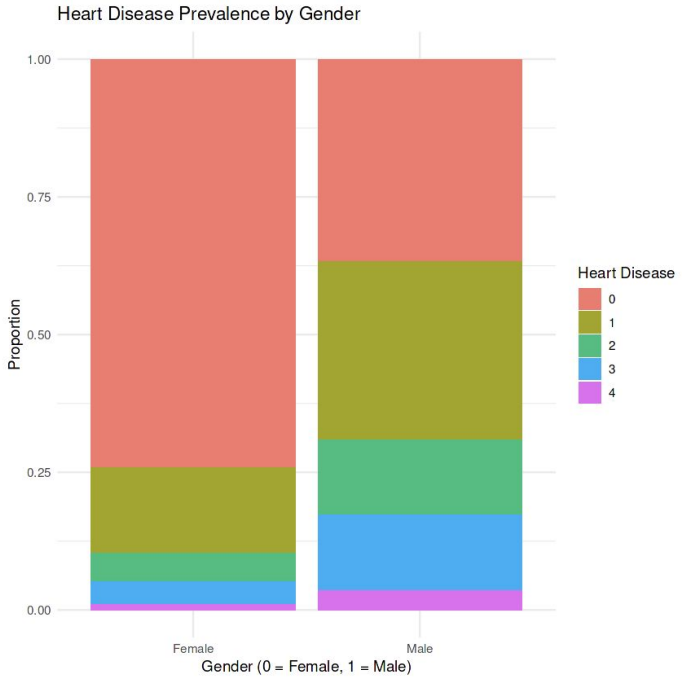
- Interpretation: This box plot shows the age distribution across different levels of heart disease severity.

  We observe that patients with **higher heart disease severity** tend to be **older,** with a slight increase in median age as severity rises. Additionally, the interquartile range widens slightly with severity, indicating more variability in age among patients with more advanced heart disease. This supports our hypothesis that age is a contributing factor to heart disease risk.



Heart Disease Prevalence by Gender

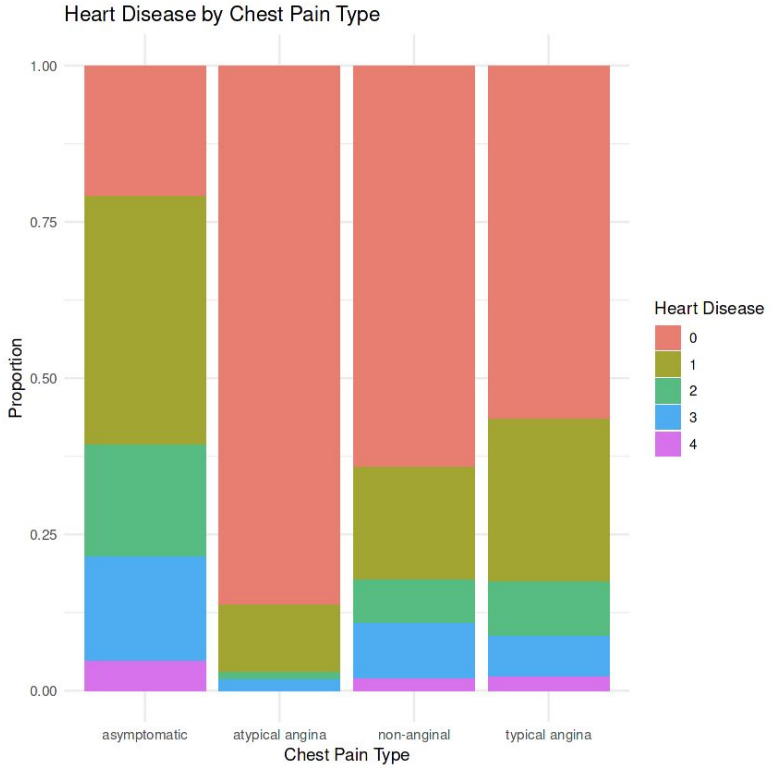
- **Interpretation**: This bar plot illustrates the proportion of heart disease severity levels across male and female patients. **Males** have a **higher prevalence** of severe **heart disease** compared to females, which aligns with our hypothesis that men are more prone to heart disease. This could be due to lifestyle, hormonal differences, or varying risk profiles between genders.



Heart Disease by Chest Pain Type

- **Interpretation**: This plot shows the distribution of heart disease severity across different chest pain types.

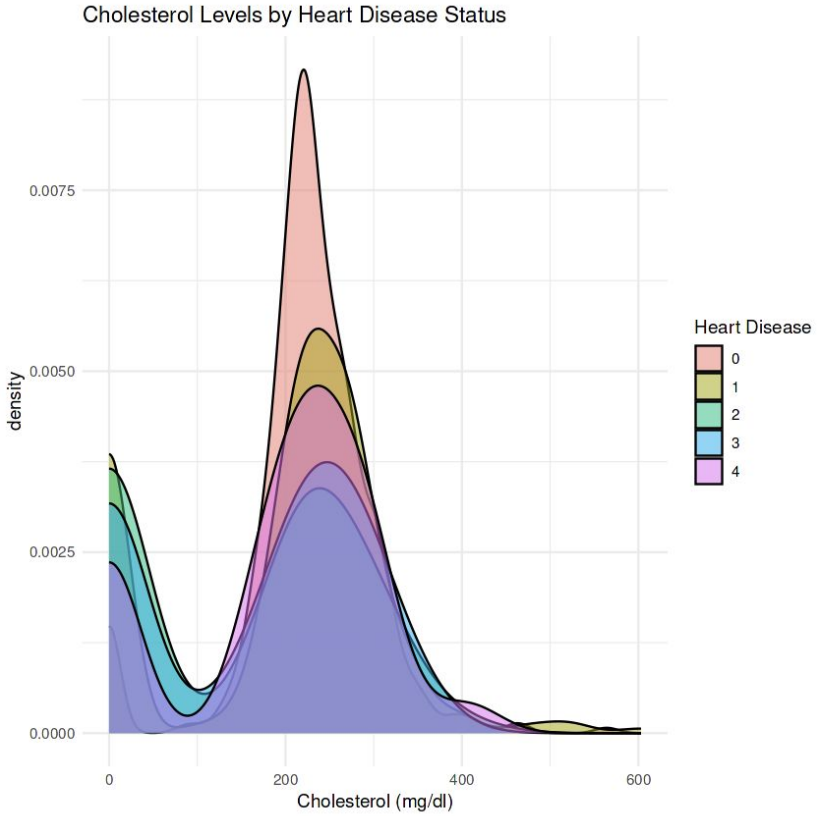
  Patients with **asymptomatic chest pain** are observed to have a **higher likelihood of severe heart disease,** supporting the hypothesis that asymptomatic chest pain could indicate more advanced heart conditions. This insight suggests that certain chest pain types may signal different levels of risk.



Cholesterol Levels by Heart Disease Status

- **Interpretation**: The density plot illustrates cholesterol levels across heart disease severities.

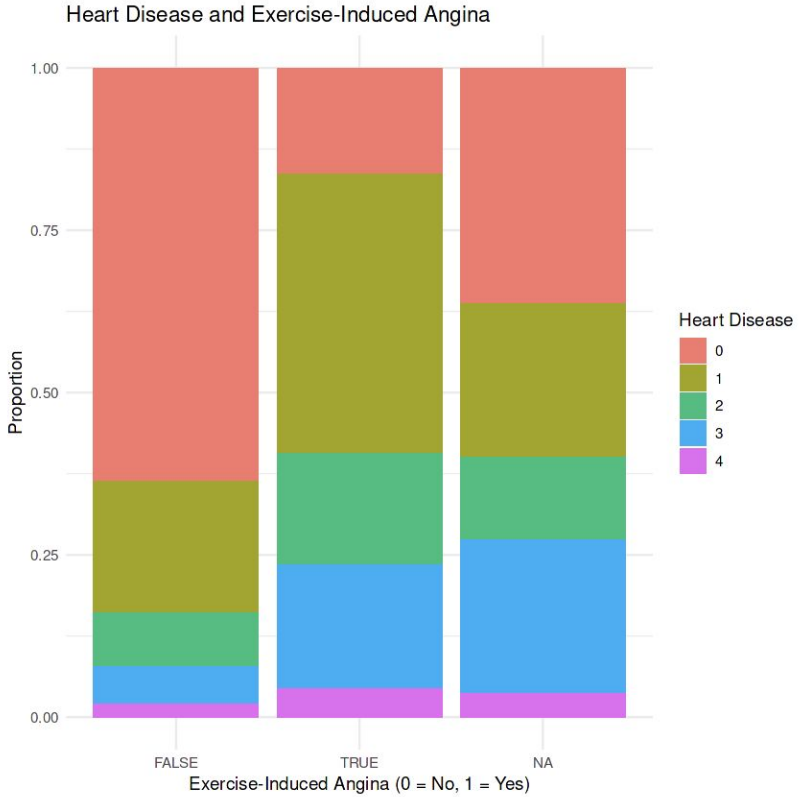
  Higher cholesterol levels are more prevalent in patients with **severe heart disease,** which supports our hypothesis that elevated cholesterol is associated with heart disease risk. Some outliers with extremely high cholesterol levels may warrant further investigation, as they could represent cases with additional risk factors.

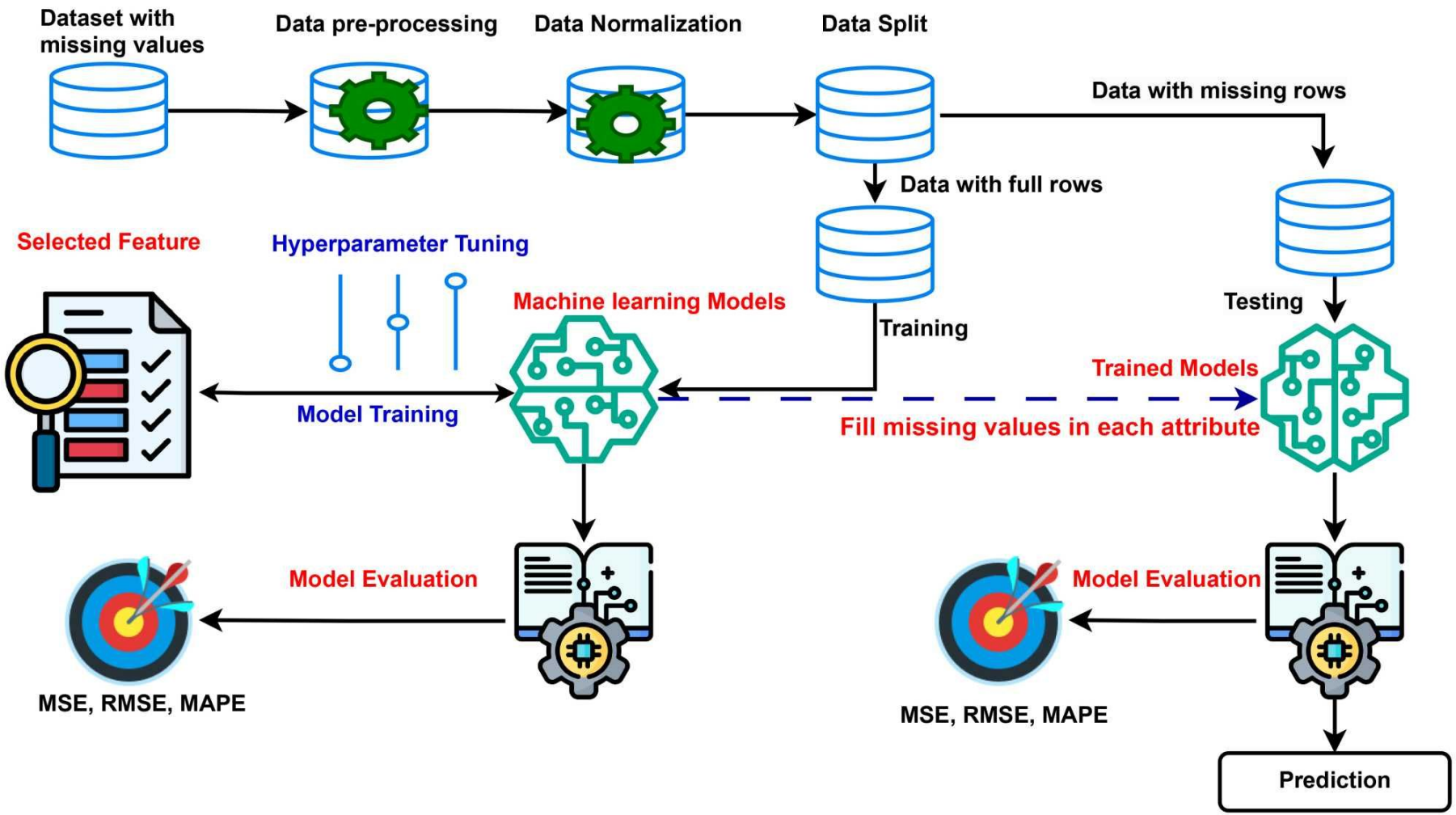


Heart Disease and Exercise-Induced Angina

- **Interpretation**: This bar plot shows the prevalence of heart disease severity among patients with and without exercise-induced angina.

  Patients with **exercise-induced angina** tend to show **higher heart disease severity**, aligning with our hypothesis that exercise-induced angina could be an indicator of restricted blood flow and potential heart issues. However, the trend here is less distinct, suggesting additional factors might also play a role.




Dataset with missing values
Data pre-processing
Data Normalization
Data Split
Data with missing rows
Data with full rows
Selected Feature
Hyperparameter Tuning
Machine learning Models
Training
Testing
Trained Models
Model Training
Fill missing values in each attribute
Model Evaluation
Model Evaluation
MSE, RMSE, MAPE
MSE, RMSE, MAPE
Prediction


THANK YOU
ANY QUESTION?