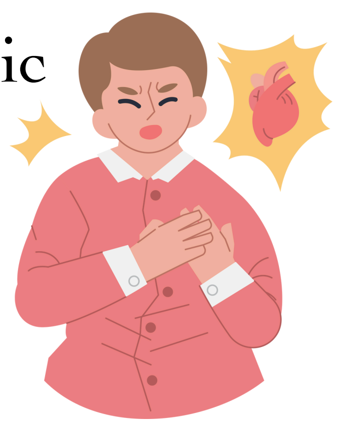
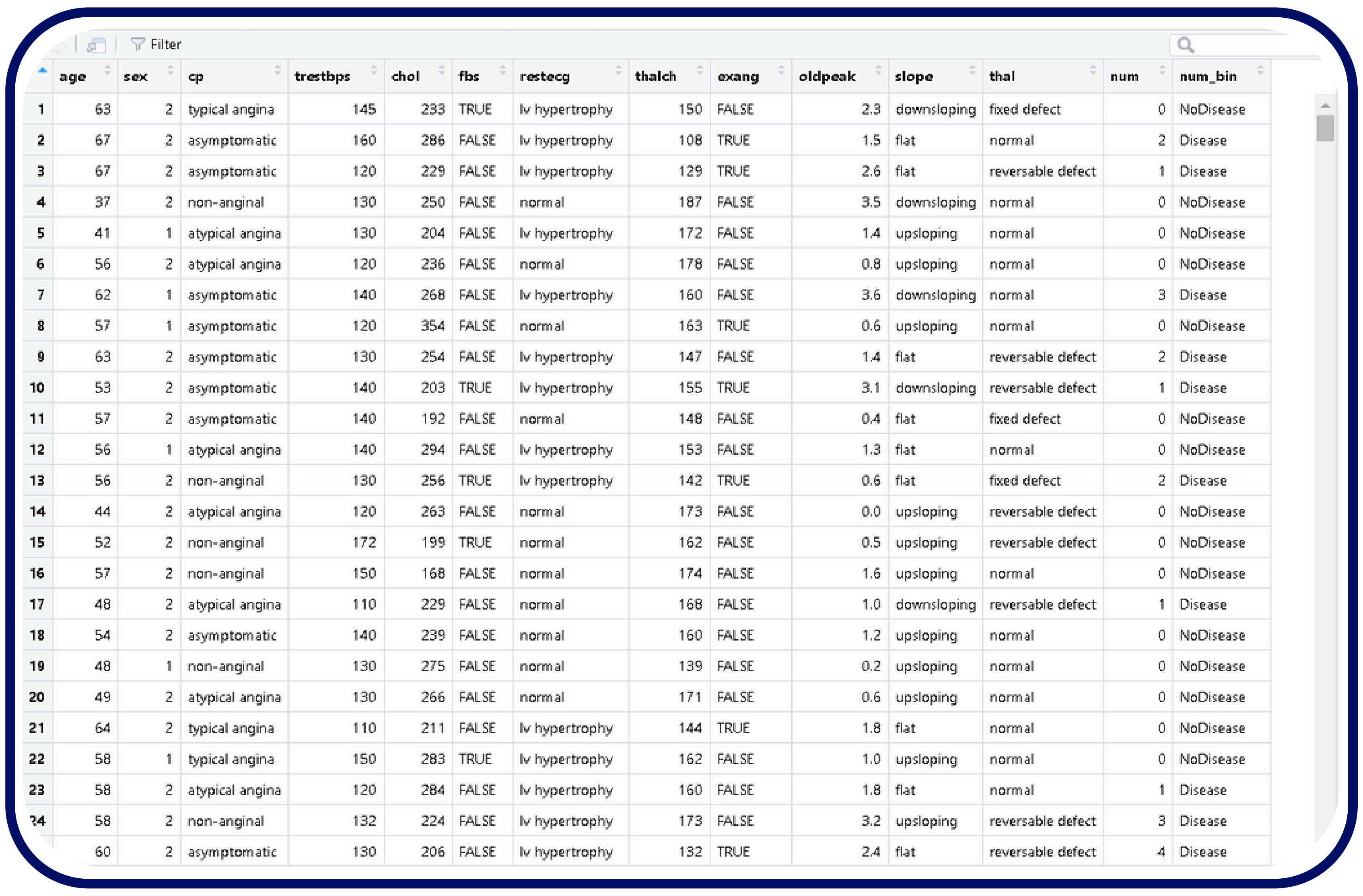
by:Mayar hany

# Heart disease

## Data Overview

This study utilizes the "Heart Disease UCI" dataset, a benchmark dataset in machine learning The data comprises medical records from four sources (Cleveland, Hungary, Switzerland, and Long Beach VA). It combines demographic information (e.g., age, sex) with biometric .measurements to investigate factors influencing heart disease risk





| | age | sex | cp | trestbps | chol | fbs | restecg | thalch | exang | oldpeak | slope | thal | num | num_bin |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 63 | 2 | typical angina | 145 | 233 | TRUE | lv hypertrophy | 150 | FALSE | 2.3 | downsloping | fixed defect | 0 | NoDisease |
| 2 | 67 | 2 | asymptomatic | 160 | 286 | FALSE | lv hypertrophy | 108 | TRUE | 1.5 | flat | normal | 2 | Disease |
| 3 | 67 | 2 | asymptomatic | 120 | 229 | FALSE | lv hypertrophy | 129 | TRUE | 2.6 | flat | reversable defect | 1 | Disease |
| 4 | 37 | 2 | non-anginal | 130 | 250 | FALSE | normal | 187 | FALSE | 3.5 | downsloping | normal | 0 | NoDisease |
| 5 | 41 | 1 | atypical angina | 130 | 204 | FALSE | lv hypertrophy | 172 | FALSE | 1.4 | upsloping | normal | 0 | NoDisease |
| 6 | 56 | 2 | atypical angina | 120 | 236 | FALSE | normal | 178 | FALSE | 0.8 | upsloping | normal | 0 | NoDisease |
| 7 | 62 | 1 | asymptomatic | 140 | 268 | FALSE | lv hypertrophy | 160 | FALSE | 3.6 | downsloping | normal | 3 | Disease |
| 8 | 57 | 1 | asymptomatic | 120 | 354 | FALSE | normal | 163 | TRUE | 0.6 | upsloping | normal | 0 | NoDisease |
| 9 | 63 | 2 | asymptomatic | 130 | 254 | FALSE | lv hypertrophy | 147 | FALSE | 1.4 | flat | reversable defect | 2 | Disease |
| 10 | 53 | 2 | asymptomatic | 140 | 203 | TRUE | lv hypertrophy | 155 | TRUE | 3.1 | downsloping | reversable defect | 1 | Disease |
| 11 | 57 | 2 | asymptomatic | 140 | 192 | FALSE | normal | 148 | FALSE | 0.4 | flat | fixed defect | 0 | NoDisease |
| 12 | 56 | 1 | atypical angina | 140 | 294 | FALSE | lv hypertrophy | 153 | FALSE | 1.3 | flat | normal | 0 | NoDisease |
| 13 | 56 | 2 | non-anginal | 130 | 256 | TRUE | lv hypertrophy | 142 | TRUE | 0.6 | flat | fixed defect | 2 | Disease |
| 14 | 44 | 2 | atypical angina | 120 | 263 | FALSE | normal | 173 | FALSE | 0.0 | upsloping | reversable defect | 0 | NoDisease |
| 15 | 52 | 2 | non-anginal | 172 | 199 | TRUE | normal | 162 | FALSE | 0.5 | upsloping | reversable defect | 0 | NoDisease |
| 16 | 57 | 2 | non-anginal | 150 | 168 | FALSE | normal | 174 | FALSE | 1.6 | upsloping | normal | 0 | NoDisease |
| 17 | 48 | 2 | atypical angina | 110 | 229 | FALSE | normal | 168 | FALSE | 1.0 | downsloping | reversable defect | 1 | Disease |
| 18 | 54 | 2 | asymptomatic | 140 | 239 | FALSE | normal | 160 | FALSE | 1.2 | upsloping | normal | 0 | NoDisease |
| 19 | 48 | 1 | non-anginal | 130 | 275 | FALSE | normal | 139 | FALSE | 0.2 | upsloping | normal | 0 | NoDisease |
| 20 | 49 | 2 | atypical angina | 130 | 266 | FALSE | normal | 171 | FALSE | 0.6 | upsloping | normal | 0 | NoDisease |
| 21 | 64 | 2 | typical angina | 110 | 211 | FALSE | lv hypertrophy | 144 | TRUE | 1.8 | flat | normal | 0 | NoDisease |
| 22 | 58 | 1 | typical angina | 150 | 283 | TRUE | lv hypertrophy | 162 | FALSE | 1.0 | upsloping | normal | 0 | NoDisease |
| 23 | 58 | 2 | atypical angina | 120 | 284 | FALSE | lv hypertrophy | 160 | FALSE | 1.8 | flat | normal | 1 | Disease |
| 24 | 58 | 2 | non-anginal | 132 | 224 | FALSE | lv hypertrophy | 173 | FALSE | 3.2 | upsloping | reversable defect | 3 | Disease |
| | 60 | 2 | asymptomatic | 130 | 206 | FALSE | lv hypertrophy | 132 | TRUE | 2.4 | flat | reversable defect | 4 | Disease |

## Data Dictionary

To understand the dataset, below are the definitions of key variables:

Demographics:

age: Patient's age in years.

sex: Patient's gender (Male / Female).

Symptoms & Vitals:

cp (Chest Pain Type): Categorized into 4 types.

trestbps: Resting blood pressure (mm Hg).

chol: Serum cholesterol (mg/dl).

fbs: Fasting blood sugar (> 120 mg/dl is True).

Cardiac Tests (ECG & Stress):

restecg: Resting electrocardiographic results.

thalch: Maximum heart rate achieved.

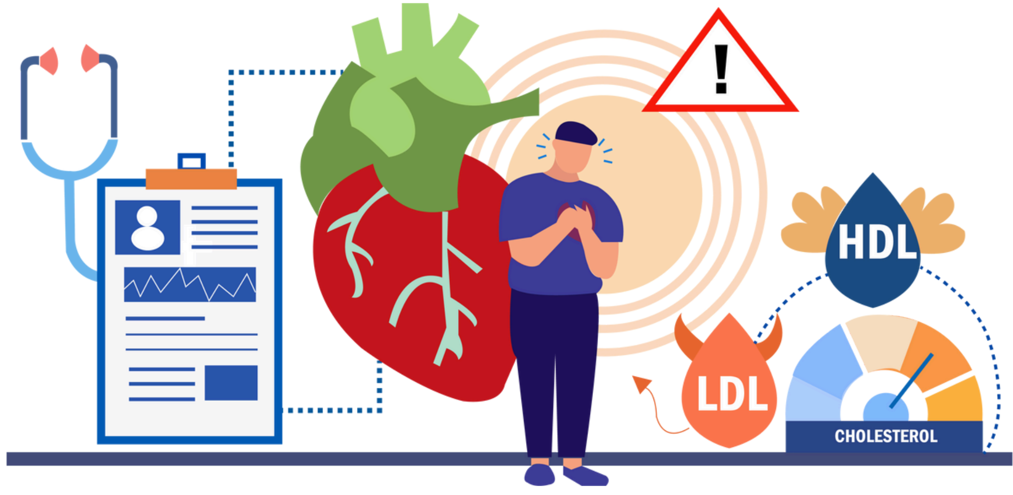
exang: Exercise-induced angina.

oldpeak: ST depression induced by exercise relative to rest.

slope: Slope of the peak exercise ST segment.

Target Variable:

num: Diagnosis of heart disease (0: Healthy, 1-4: Disease severity)



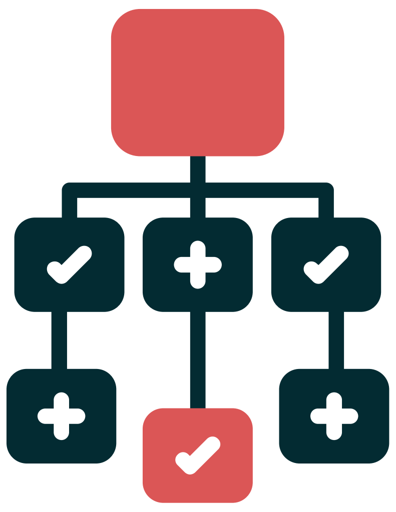
by:Mayar hany

# Heart disease

## Project Objective

The primary goal is to build Predictive Models to diagnose heart disease based on medical attributes using three algorithms:

1. Decision Tree.
2. Random Forest.
3. Logistic Regression.



## Data Cleaning & Preprocessing

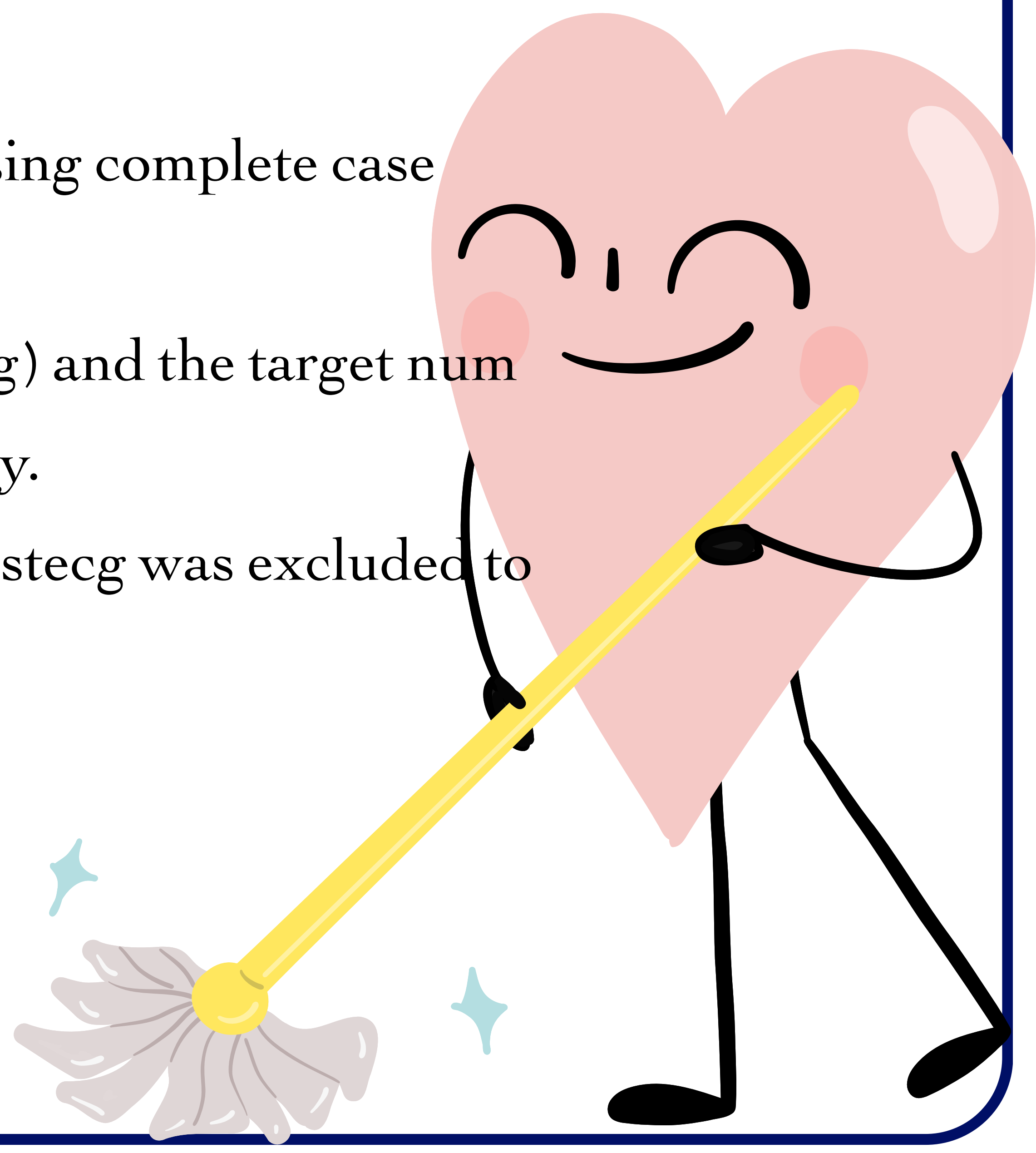
To ensure model quality, the data underwent the following preprocessing steps:

Feature Selection: Dropped dataset and id as they are non-diagnostic, and ca due to a high volume of missing values.

Missing Values: Rows with missing values were removed using complete case analysis (na.omit).

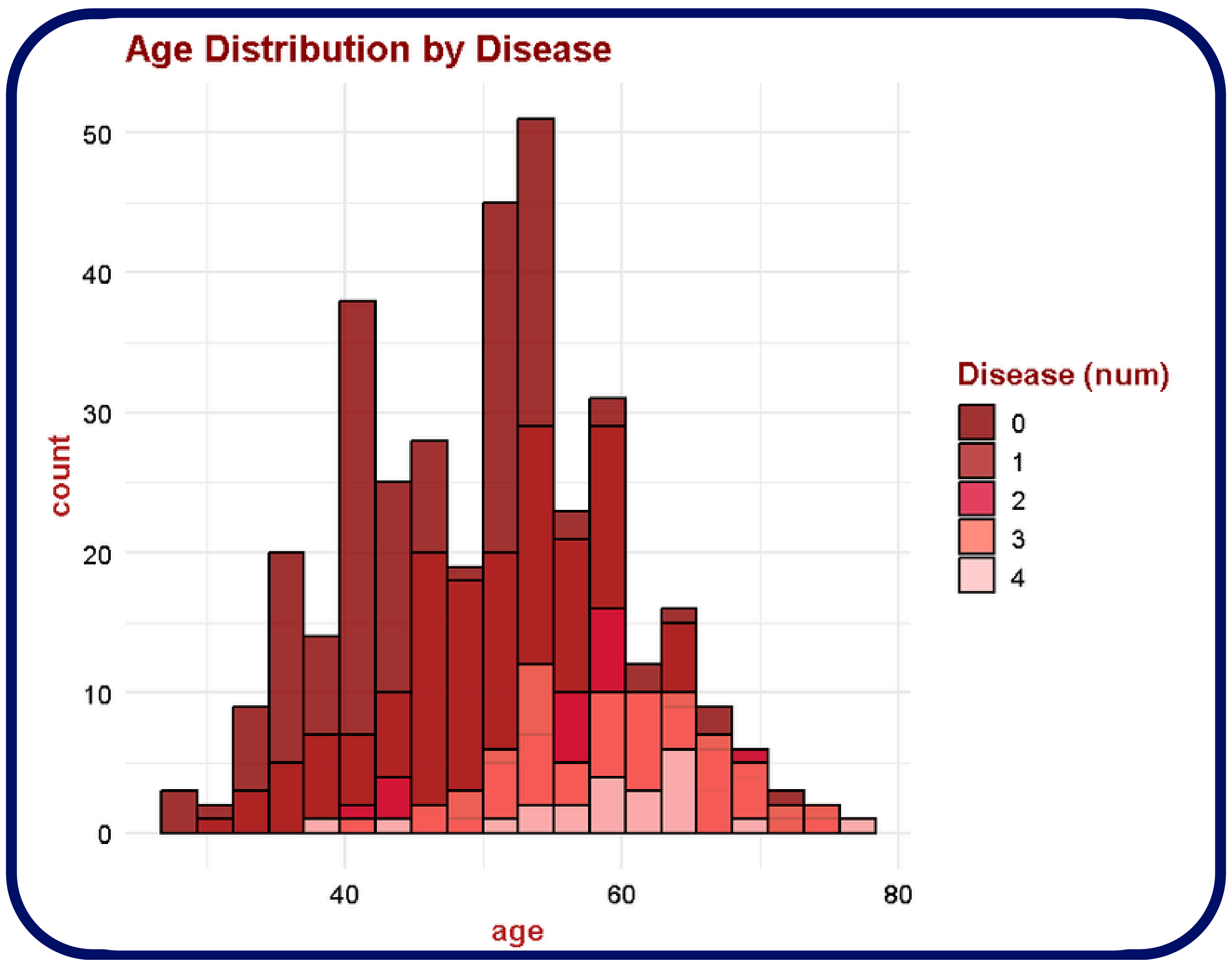
Data Transformation: Categorical variables (e.g., cp, restecg) and the target num were converted to Factors. Gender was encoded numerically.

Handling Rare Classes: The rare class st-t abnormality in restecg was excluded to prevent data splitting errors and ensure model stability.



by:Mayar hany

# Heart disease



## Age Distribution by Disease Severity

What it represents: This histogram displays the frequency distribution of patients' ages, stratified by their heart disease diagnosis status (indicated by different colors).

Interpretation:

X-axis: Represents patient age in years.

Y-axis: Represents the number of patients (frequency).

Color Coding: Differentiates between healthy individuals (lightest color, num=0) and varying severities of heart disease (darker shades, num=1 to 4).

Key Insights:

Peak Risk Age: The highest concentration of heart disease cases is observed in the 55-65 age range, identifying this demographic as the most vulnerable group in the dataset.

Age Factor: Younger individuals (under 40) are predominantly healthy (represented by the light bars), confirming that heart disease prevalence is significantly lower in early adulthood.

Disease Progression: As age increases beyond 60, the proportion of diagnosed cases (darker bars) rises relative to healthy individuals, highlighting age as a significant, non-modifiable risk factor for heart disease.

## Chest Pain Type Distribution

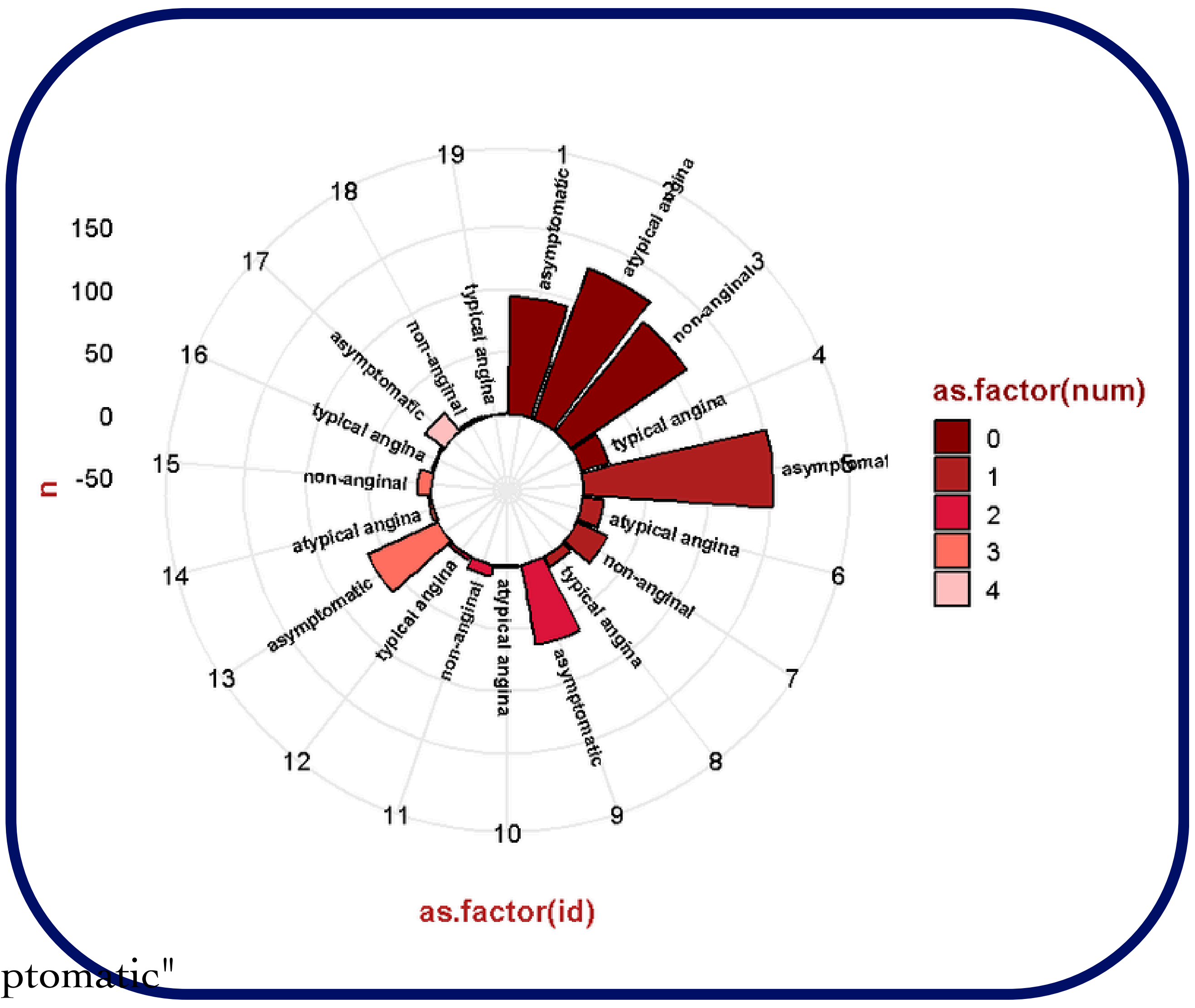
What it represents: This circular bar plot categorizes patients by their reported Chest Pain Type (cp) and visualizes the prevalence of heart disease within each category.

Interpretation:

Segments: The chart is divided into four sectors representing: Typical Angina, Atypical Angina, Non-anginal Pain, and Asymptomatic.

Color Coding: Distinguishes between healthy subjects and those with heart disease severity (num 1-4).

Bar Length: Corresponds to the number of patients in each group.



Key Insights:

The "Asymptomatic" Risk:

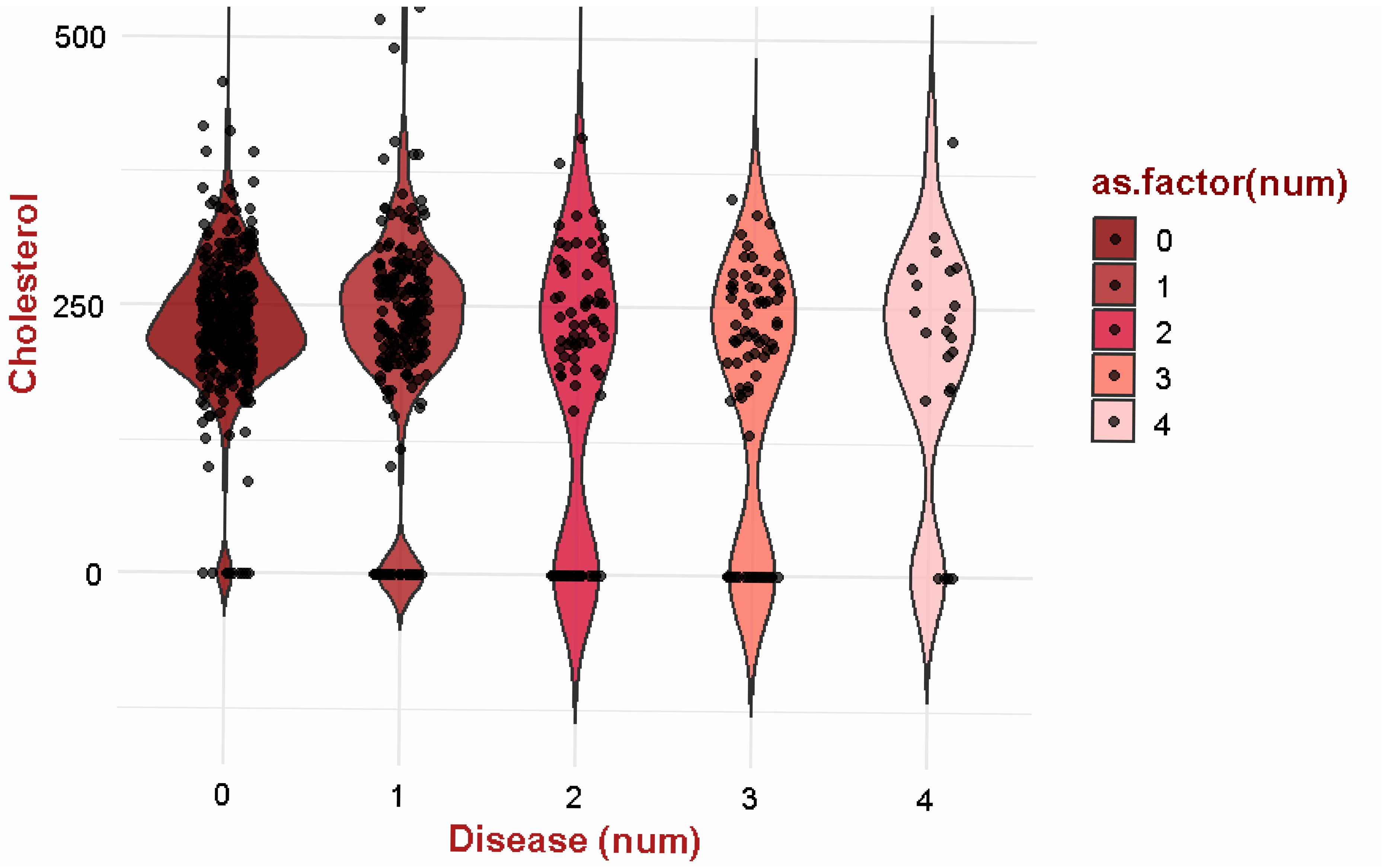
A significant portion of confirmed heart disease cases falls under the "Asymptomatic" category. This highlights the danger of "Silent Ischemia," where patients do not experience classic chest pain despite having severe heart conditions.

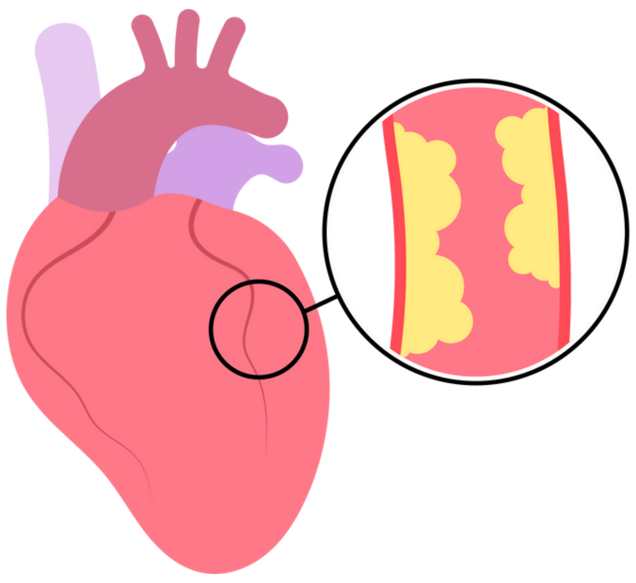
Differentiation: The chart visually separates patients with "Non-anginal pain" (who are largely healthy) from those with true angina, validating the importance of clinical history taking in the initial diagnosis.

by:Mayar hany

# Heart disease

## Cholesterol Levels vs. Heart Disease Severity





What it represents:

This Violin Plot visualizes the distribution and density of serum cholesterol levels (chol) across different stages of heart disease (num 0 to 4).

Interpretation:

Violin Width: Indicates the density of data points.

Wider sections represent the cholesterol levels where the majority of patients are concentrated.

Scattered Points: Represent individual patient records, revealing the spread and potential outliers.

Y-axis: Cholesterol concentration in mg/dl.

Key Insights:

Significant Overlap: The distributions across healthy (0) and diseased groups (1-4) are visually similar, with significant overlap in the 200-300 mg/dl range. This explains why cholesterol appeared lower in the "Feature Importance" chart compared to stress test results; it is not a binary separator in this specific dataset.

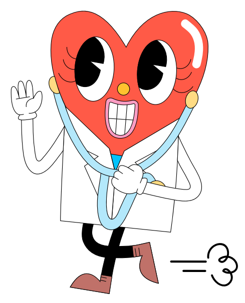
Outliers: Extreme values (cholesterol > 400 mg/dl) are visible as scattered points at the top. While these high levels are clinically dangerous, the chart shows that many patients with heart disease still have "average" cholesterol levels, reinforcing the need for multi-factor diagnosis.

by:Mayar hany

# Heart disease

## Resting Blood Pressure Distribution by Disease Severity

What it represents: This Ridge Plot visualizes the density distribution of Resting Blood Pressure (trestbps) across the different heart disease stages (num 0-4).



Interpretation:

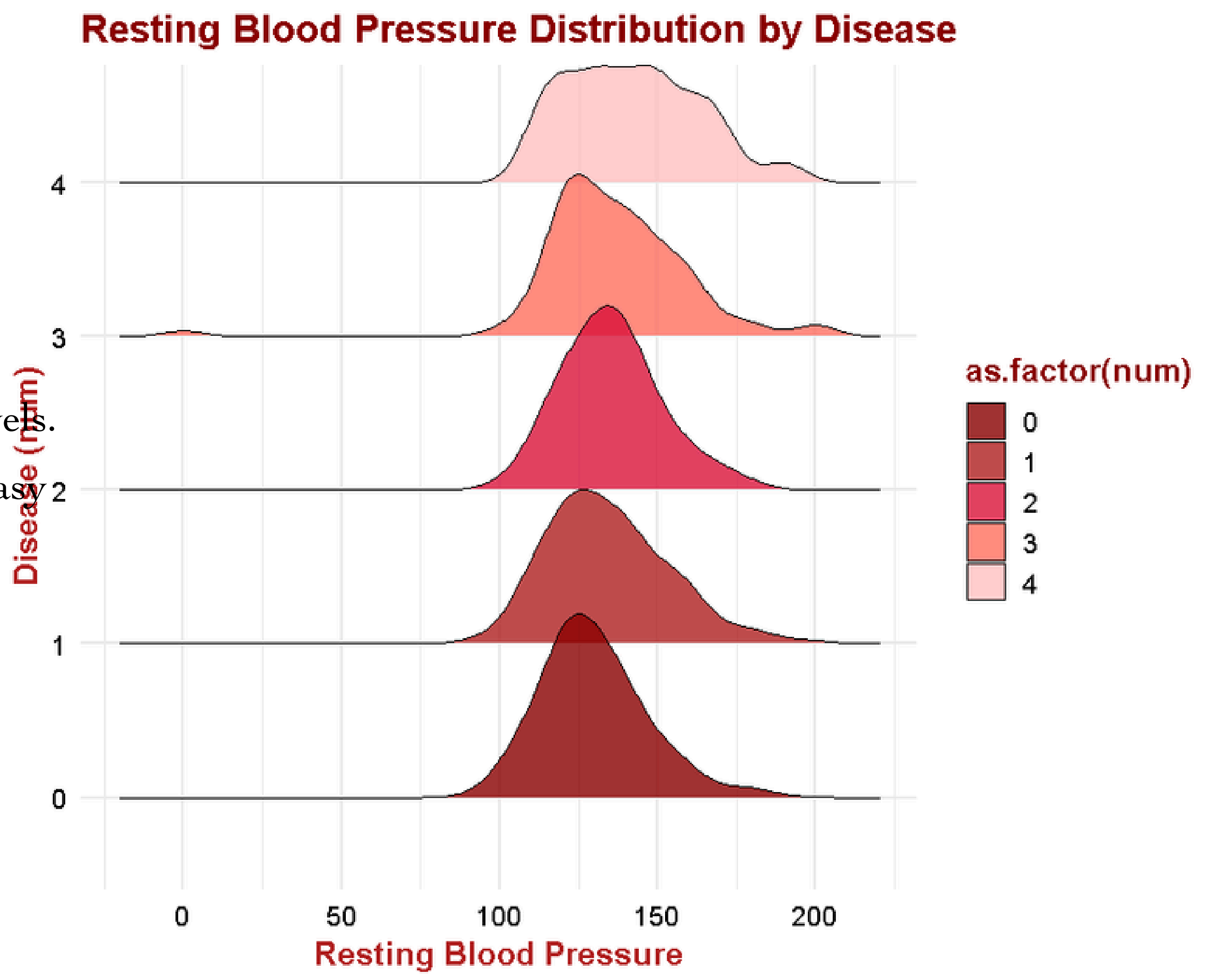
Ridges (Mountains): Each curve represents the concentration of patients at specific blood pressure levels.

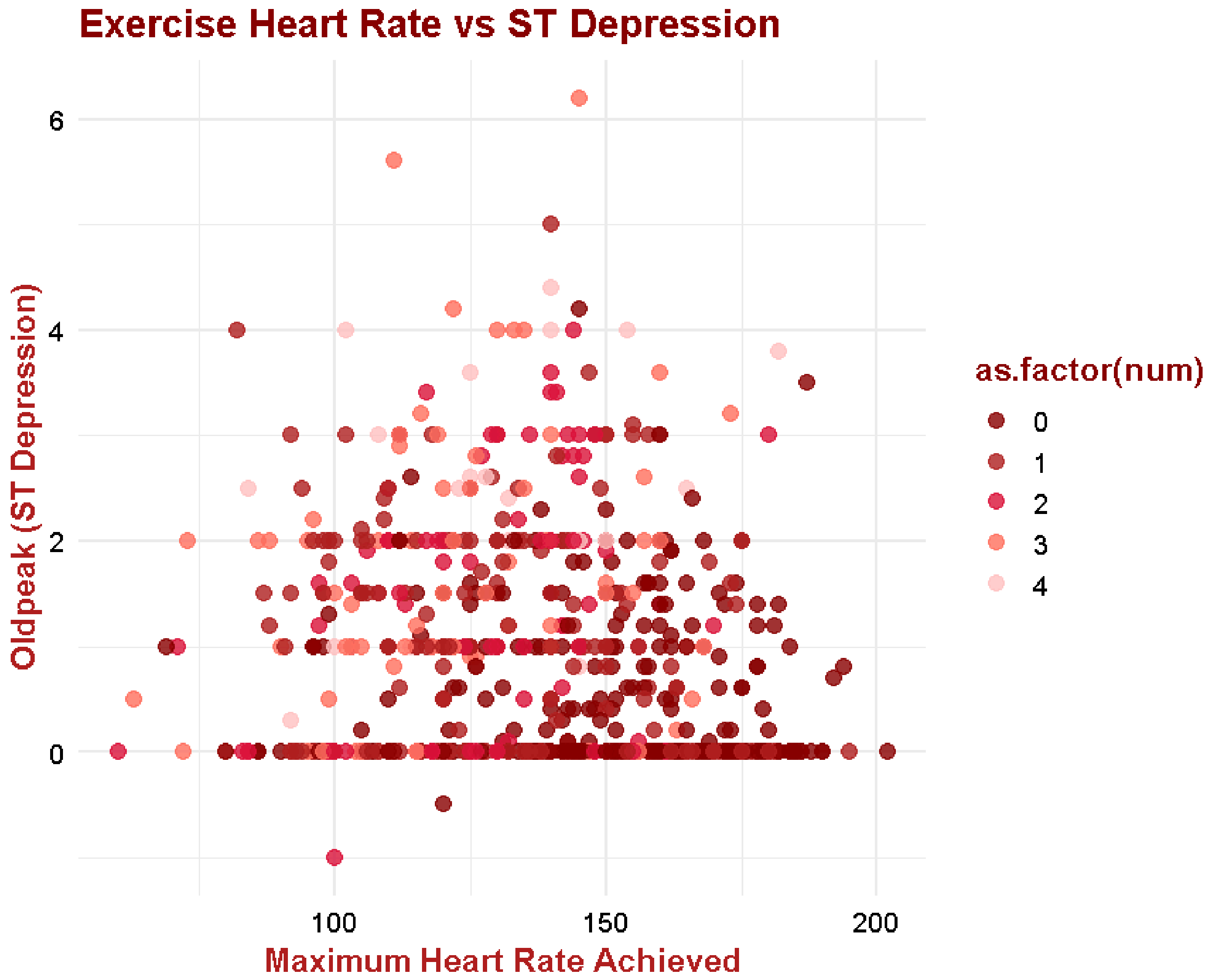
Alignment: The plots are stacked vertically to allow easy comparison of the distribution shapes.

Key Insights:

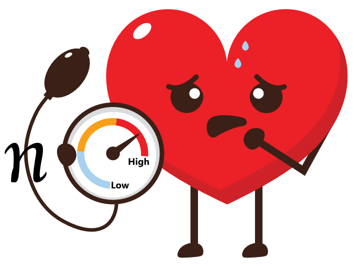
Uniformity: The peaks of the distributions for all groups (healthy and diseased) align closely around the 120-130 mmHg mark.

Weak Discriminator: The significant overlap suggests that Resting Blood Pressure, while a general risk factor, does not distinctly differentiate between the levels of heart disease severity in this specific dataset. It is less predictive than dynamic factors like exercise-induced angina.





## Max Heart Rate vs. ST Depression



What it represents: This Scatter Plot visualizes the correlation between the maximum heart rate achieved during exercise (thalch) and the exercise-induced ST depression (oldpeak).

Interpretation:

- X-axis: Max Heart Rate (thalch). Higher values indicate better cardiac capacity.
- Y-axis: ST Depression (oldpeak). Higher values indicate cardiac distress.
- Clusters: Lighter points represent healthy patients, while darker points represent those with heart disease.

Key Insights:

The Healthy Cluster: Healthy individuals (light dots) predominantly occupy the bottom-right region. They exhibit high heart rates (>150) with minimal ST depression (<1.0), indicating a heart that functions well under stress.

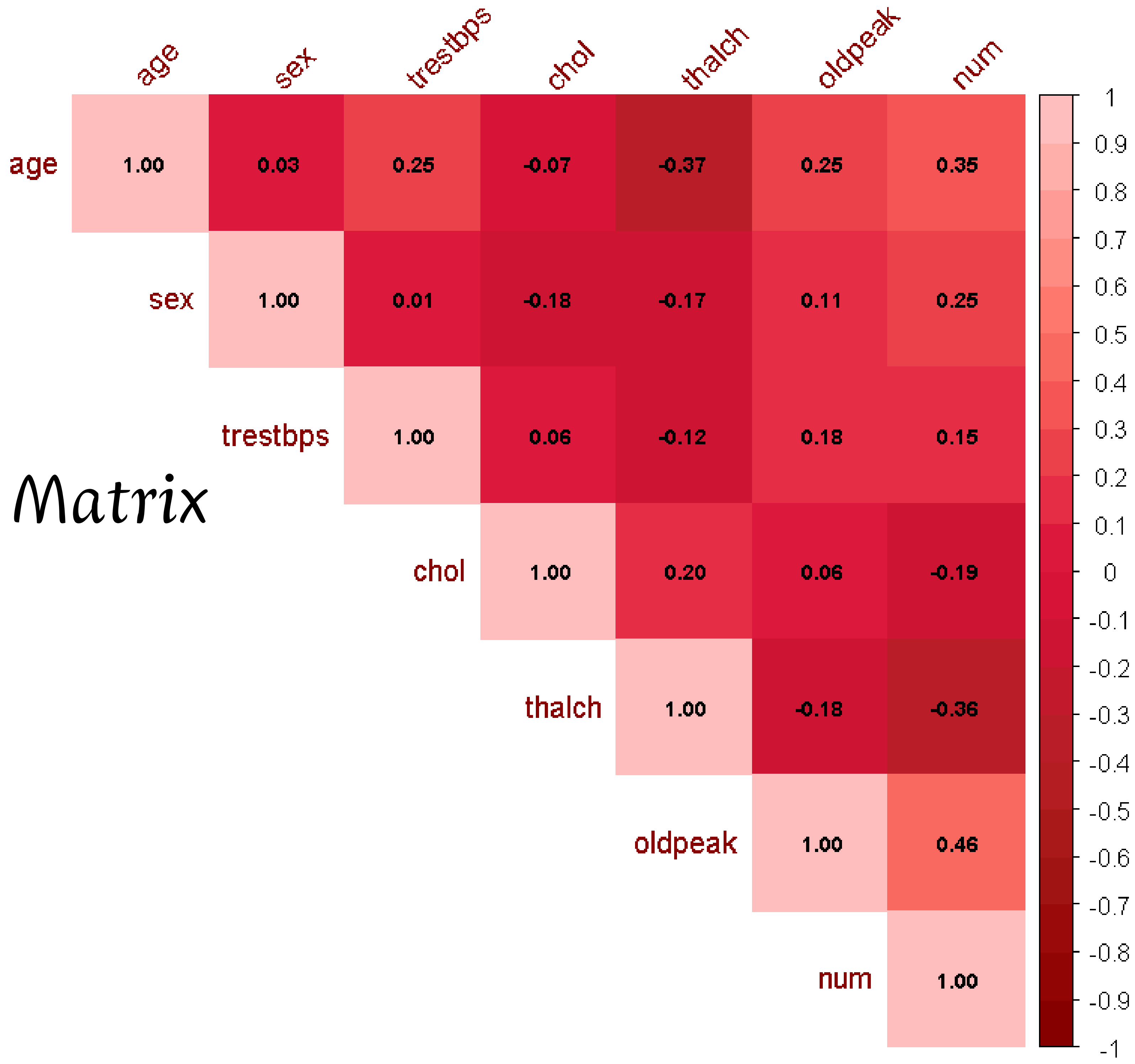
The At-Risk Cluster: Patients with heart disease (dark dots) are scattered towards the top-left. Their hearts struggle to reach high rates and exhibit significant ST depression early on.

Diagnostic Value: The distinct separation between the two groups confirms that the combination of thalch and oldpeak serves as a powerful predictive feature for the machine learning models.

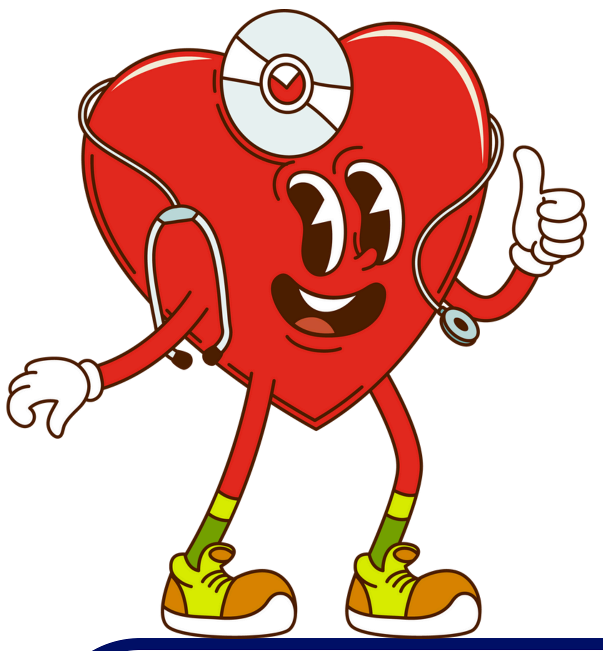
by:Mayar hany

# Heart disease

## Feature Correlation Matrix



|  | age | sex | trestbps | chol | thalch | oldpeak | num |
|---|---|---|---|---|---|---|---|
| age | 1.00 | 0.03 | 0.25 | -0.07 | -0.37 | 0.25 | 0.35 |
| sex |  | 1.00 | 0.01 | -0.18 | -0.17 | 0.11 | 0.25 |
| trestbps |  |  | 1.00 | 0.06 | -0.12 | 0.18 | 0.15 |
| chol |  |  |  | 1.00 | 0.20 | 0.06 | -0.19 |
| thalch |  |  |  |  | 1.00 | -0.18 | -0.36 |
| oldpeak |  |  |  |  |  | 1.00 | 0.46 |
| num |  |  |  |  |  |  | 1.00 |



What it represents: This heatmap visualizes the correlation coefficients between numerical variables, showing how strongly pairs of variables are related.

Interpretation:

Color Scale: Dark Red indicates strong positive correlation, while Blue indicates negative correlation. Lighter colors imply weak or no correlation.

Values: Ranging from -1 to +1, where values closer to 1 indicate a strong linear relationship.

Key Insights:

Disease Indicators: There is a noticeable positive correlation between oldpeak (ST depression) and the target variable num (heart disease), reinforcing oldpeak's role as a key predictor.

Physiological Trends: A positive correlation exists between age and trestbps (blood pressure), aligning with medical knowledge that blood pressure tends to rise with age.

Inverse Relationship: A negative correlation is observed between age and thalch (max heart rate), confirming that maximum cardiac output decreases as patients age.

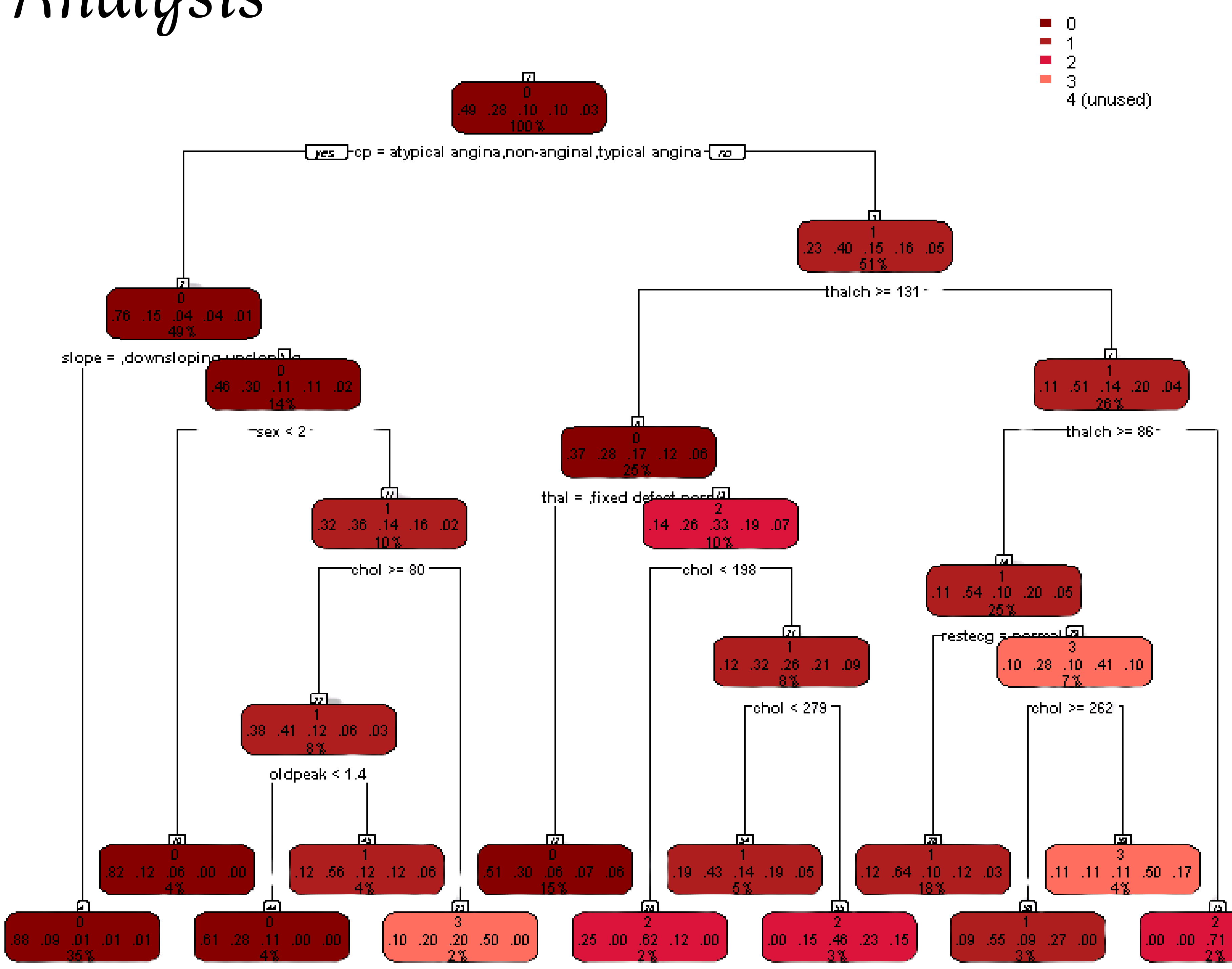
by:Mayar hany

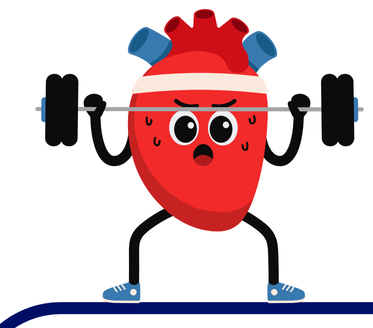
# Heart disease

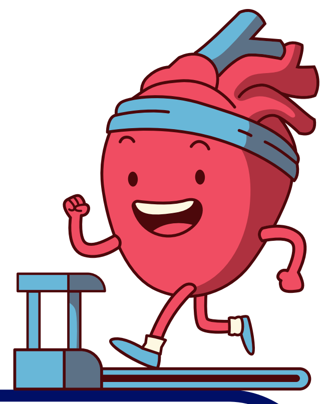
## The First Model

## Decision Tree Analysis

Decision Tree for Heart Disease







Decision Tree Model Results

The Decision Tree model demonstrated strong performance in identifying healthy individuals but faced challenges in distinguishing between the specific severity levels of the disease.

- Confusion Matrix Analysis:
  - High Specificity (Healthy Class 0): The model correctly identified 89 out of 102 healthy individuals. This indicates a low rate of False Positives, meaning the model is reliable at confirming when a patient is healthy.
  - Class Overlap (Severity Levels): While the model successfully flagged patients as "diseased," it struggled to categorize the exact severity. For instance, most patients with Level 2 or 3 heart disease were misclassified as Level 1. This is a common issue in medical datasets where advanced disease cases are fewer than mild cases (Class Imbalance).

Accuracy Metrics:

- Multiclass Accuracy: Approximately 60% (due to misclassification between disease stages 1-4).
- Binary Accuracy (Healthy vs. Sick): When evaluating the model simply on distinguishing between "Healthy" and "Sick" (regardless of severity), the accuracy jumps to approximately 78%, making it a viable tool for initial screening

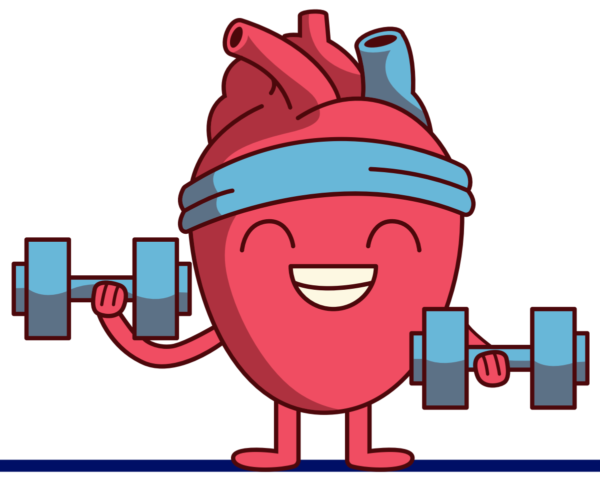
```
         Predicted
Actual   0  1  2  3  4
     0  89 10  1  2  0
     1  22 21  4  3  0
     2   3 11  1  0  0
     3   3 10  2  1  0
     4   0  2  2  0  0
```

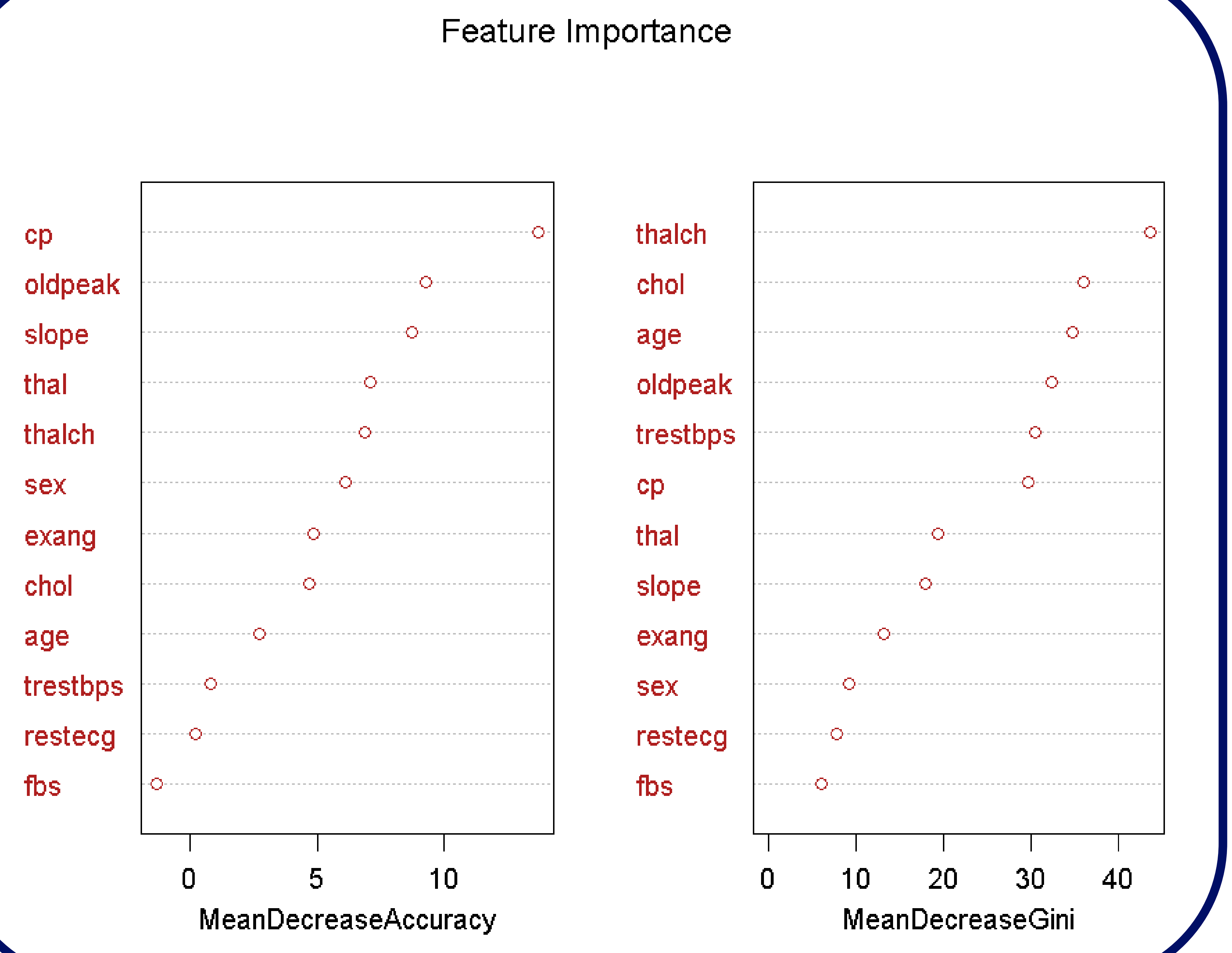
by:Mayar hany

# Heart disease

## The Second Model

## Random Forest Analysis





Random Forest Model Results

The Random Forest model demonstrated improved sensitivity in detecting positive cases compared to the Decision Tree, offering a more robust diagnostic capability.

Confusion Matrix Analysis:

- Improved Detection (Sensitivity): This model outperformed the Decision Tree in identifying Class 1 patients, correctly classifying 29 cases (vs. 21 in the previous model). This higher sensitivity is crucial in medical contexts to minimize "False Negatives" (missed diagnoses).
- Specificty: The model maintained strong performance in identifying healthy individuals (87 correct), with only a marginal increase in False Positives compared to the Decision Tree.
- Severity Classification Challenge: Similar to the previous model, Random Forest struggled to differentiate between advanced disease stages (2, 3, and 4) due to the dataset's class imbalance. Most severe cases were correctly flagged as "sick" but misclassified as Stage 1.

Conclusion: The Random Forest model offers the most balanced performance, providing a safer diagnostic tool by capturing more actual disease cases while maintaining high accuracy for healthy subjects.

```
      Predicted
Actual  0  1  2  3  4
     0 87 15  0  0  0
     1 17 29  0  4  0
     2  1  9  3  2  0
     3  5  8  1  1  1
     4  0  3  0  0  1
```

by:Mayar hany

# Heart disease

## The Third Model

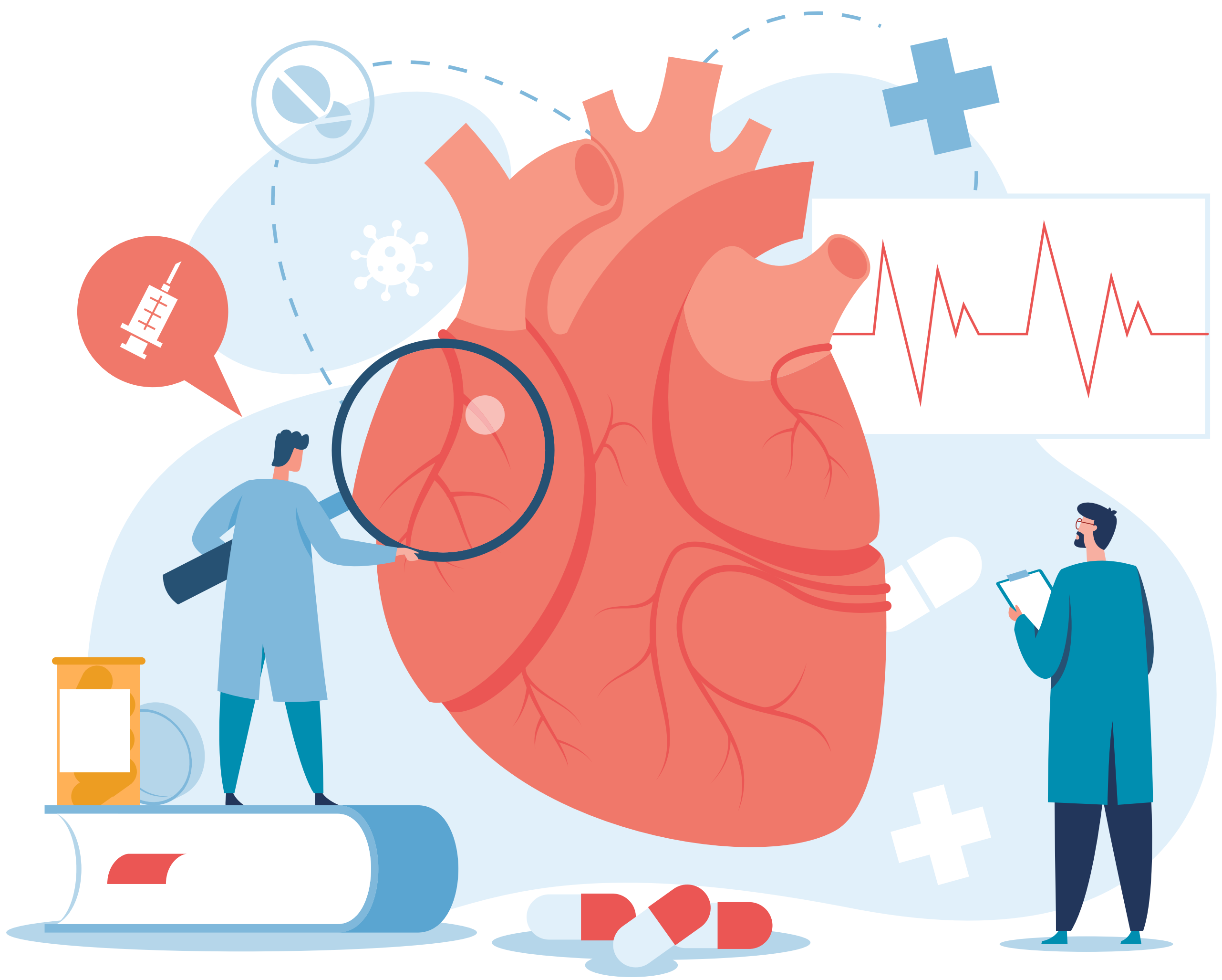
## Logistic Regression

```
        Predicted
Actual    0   1
     0  81  21
     1  15  70
```

Logistic Regression Results

This model delivered robust and balanced performance when treating the problem as a binary classification (Healthy vs. Sick), serving as an excellent baseline.

- Confusion Matrix Analysis:
  - High Predictive Power: The model achieved an overall accuracy of approximately 81%, correctly classifying 151 out of 187 cases.
  - Excellent Sensitivity: It successfully identified 70 out of 85 sick patients (82% sensitivity), making it highly effective for initial screening purposes.
  - Trade-off: In exchange for this high sensitivity, the model showed a slightly higher rate of False Positives (21 healthy individuals misclassified as sick) compared to the tree-based models.



## Conclusion & Recommendations

1. Key Drivers: All models consistently identified Chest Pain Type and ST Depression (Oldpeak) as the most critical predictors of heart disease, outperforming traditional metrics like resting blood pressure and cholesterol.
2. Model Selection:

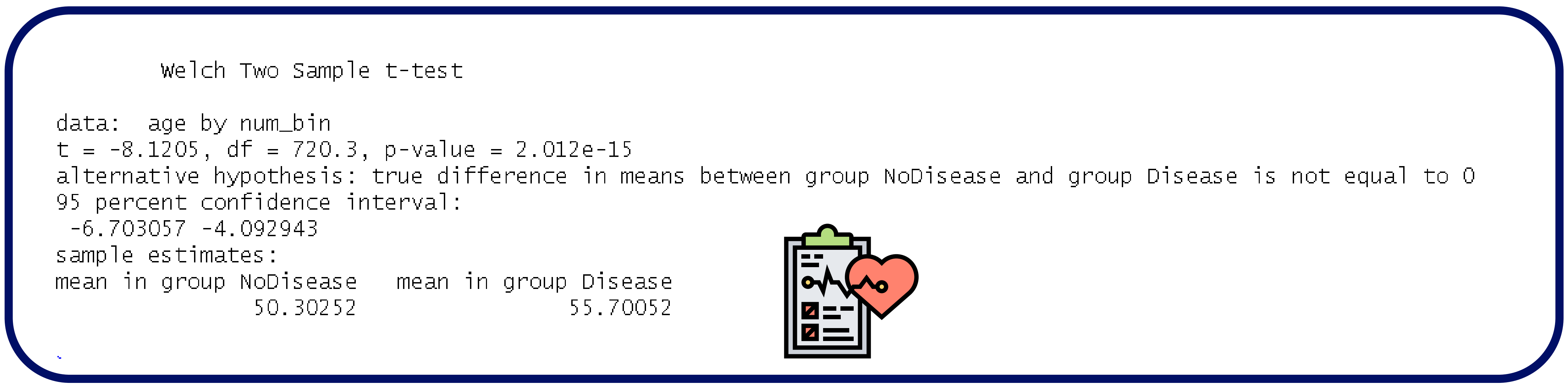
- Random Forest is the winner for predictive accuracy, offering the best balance between sensitivity (detecting the sick) and specificity (clearing the healthy).
- Decision Tree is the winner for interpretability, providing clear, rule-based logic that clinicians can easily follow.

3. Medical Recommendation: It is recommended to deploy the Random Forest model as a decision support tool. Clinicians should pay special attention to "Asymptomatic" patients who exhibit abnormal stress test results, as data indicates this group is at significant risk despite the lack of pain symptoms.

by:Mayar hany

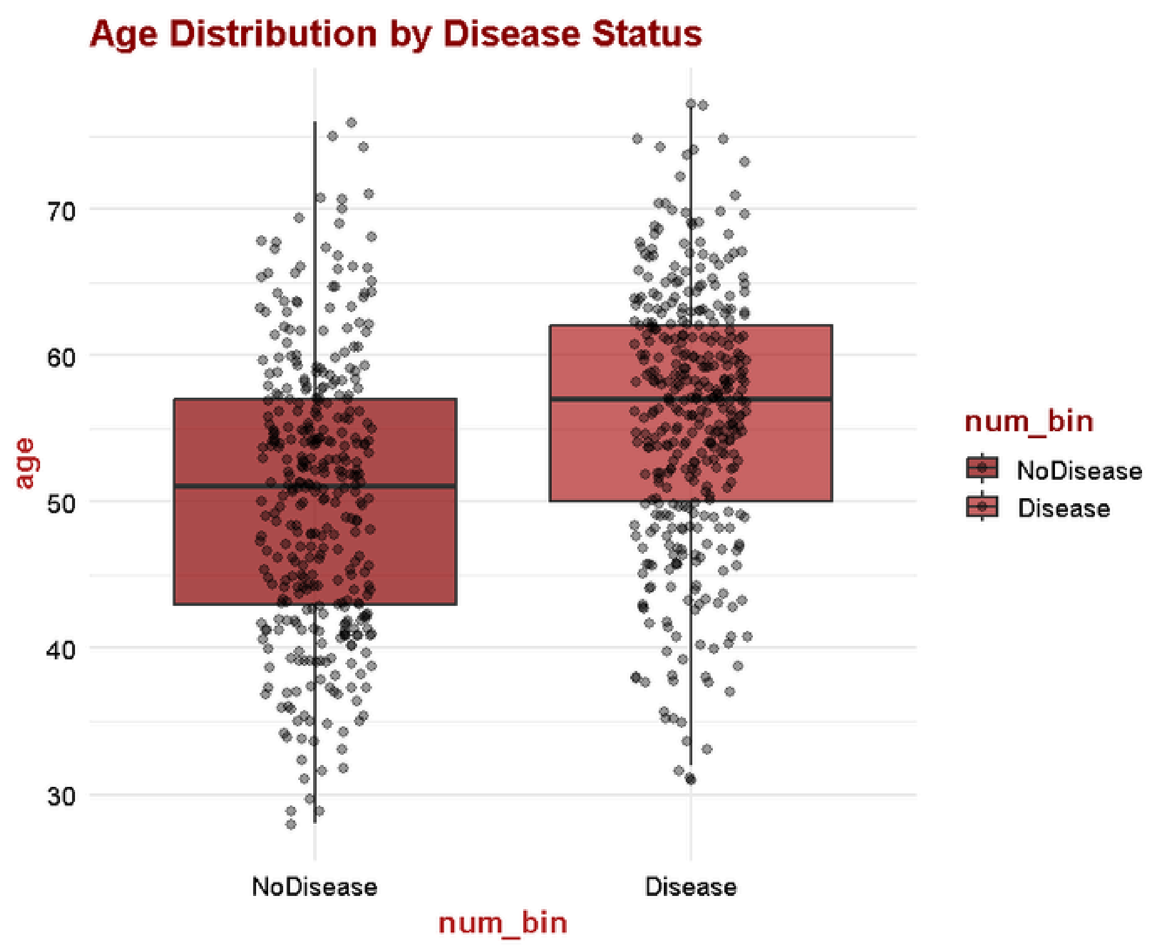
# Heart disease

## Descriptive Statistics



```
        Welch Two Sample t-test

data:  age by num_bin
t = -8.1205, df = 720.3, p-value = 2.012e-15
alternative hypothesis: true difference in means between group NoDisease and group Disease is not equal to 0
95 percent confidence interval:
 -6.703057 -4.092943
sample estimates:
mean in group NoDisease   mean in group Disease 
               50.30252                55.70052 
```


Age Distribution by Disease Status

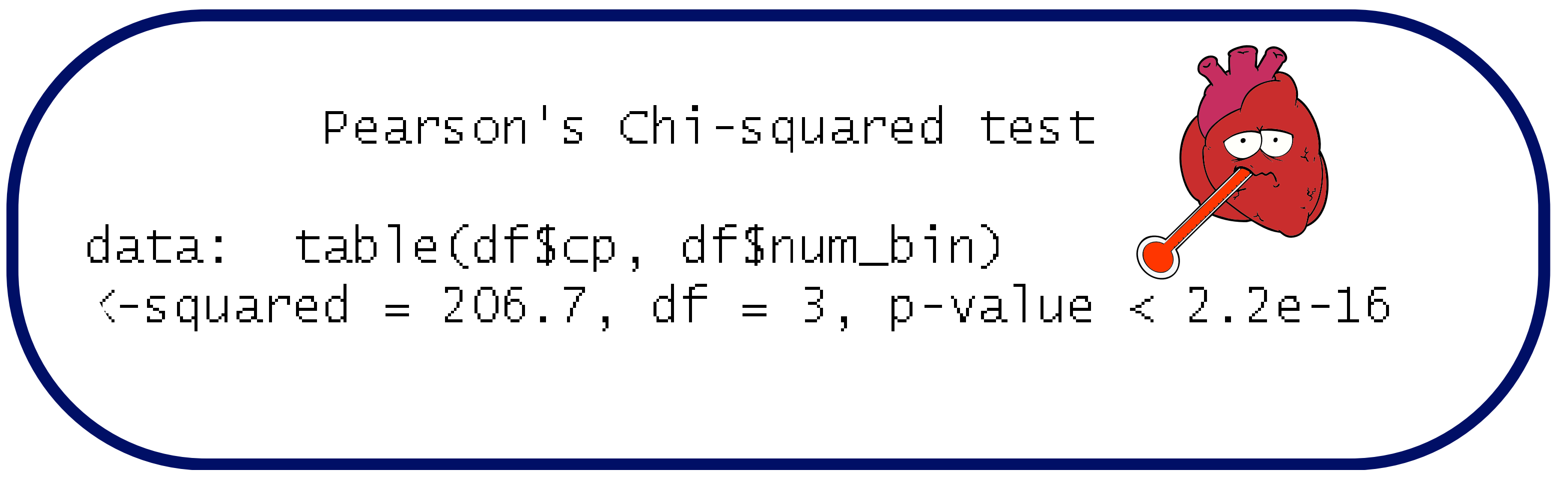
Based on the summary statistics of the dataset, we derive the following key demographic and clinical insights:

1. Demographics (Age):
   - The dataset spans a wide age range, from a minimum of 29 years to a maximum of 77 years.
   - The mean age is 53.5 years, indicating that the study primarily targets middle-aged and elderly individuals, who are the highest-risk demographic for cardiovascular diseases.
2. Clinical Vitals:
   - Resting Blood Pressure (trestbps): The average resting blood pressure is 132 mm Hg, slightly above the ideal threshold, with maximum values reaching hypertensive crisis levels (200 mm Hg).
   - Cholesterol (chol): The mean serum cholesterol is 199 mg/dl. However, extreme outliers are present, with maximum levels reaching 603 mg/dl, indicating severe hypercholesterolemia in some patients.
   - Max Heart Rate (thalch): Ranging from 60 to 202 bpm with a mean of 137, reflecting significant variability in the patients' cardiac capacity.
3. Target Variable Distribution (num):
   - The summary reveals a Class Imbalance. The majority of the dataset consists of healthy individuals (Class 0). The count of patients decreases significantly as the disease severity increases (from Class 1 to 4). This scarcity of data for advanced stages explains the lower sensitivity of the models in classifying severity levels 3 and 4.

## Data Quality Analysis

Prior to modeling, a data quality check was conducted to identify missing entries (NAs). The output reveals critical insights regarding data integrity:

1. High Missingness in ca: The ca (number of major vessels) column has 611 missing values, representing over 65% of the dataset. This justifies the decision to drop this column entirely, as imputation on such a large scale would introduce significant bias.
2. Significant Gaps in thal & slope: The thal column (486 missing) and slope (309 missing) also show substantial data loss. This explains the reduction in sample size after applying the list-wise deletion (na.omit) method.
3. Vital Signs: Variables like resting blood pressure (trestbps: 59 missing) and fasting blood sugar (fbs: 90 missing) showed moderate missingness, which was handled by removing the affected rows to maintain model accuracy.

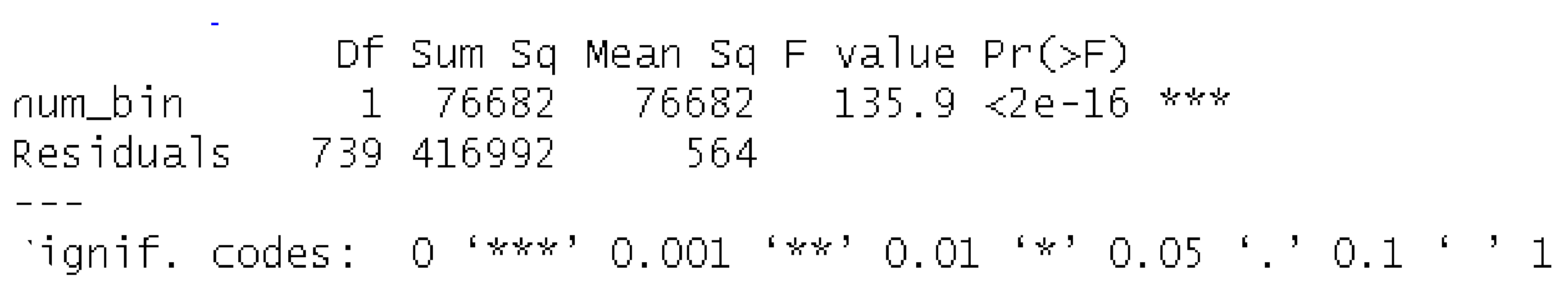


```
        Pearson's Chi-squared test

data:  table(df$cp, df$num_bin)
<-squared = 206.7, df = 3, p-value < 2.2e-16
```

by:Mayar hany

# Heart disease



```
            Df Sum Sq Mean Sq F value Pr(>F)    
num_bin      1  76682   76682   135.9 <2e-16 ***
Residuals  739 416992     564                   
---
`ignif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

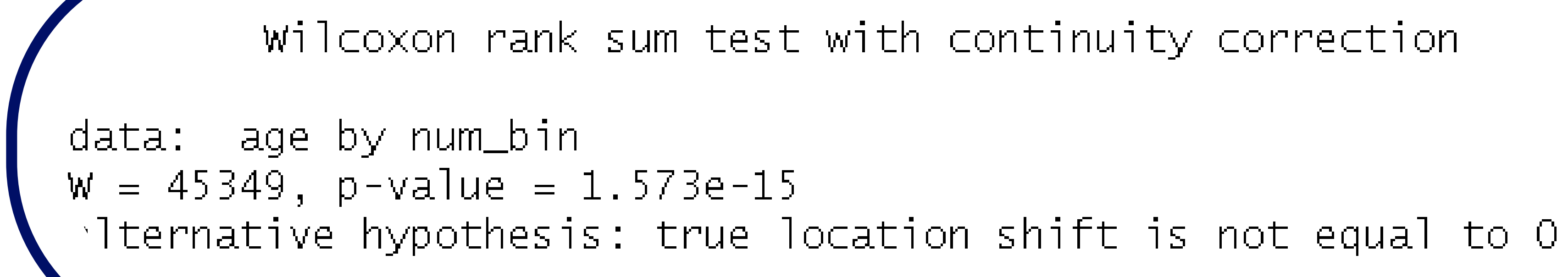
1. Gender Distribution:

- The table shows 194 Females versus 726 Males.
- Insight: This significant disparity highlights a gender bias in the dataset. The model is predominantly trained on male physiology, which may limit its generalizability to female patients.

2. Disease Severity Distribution:

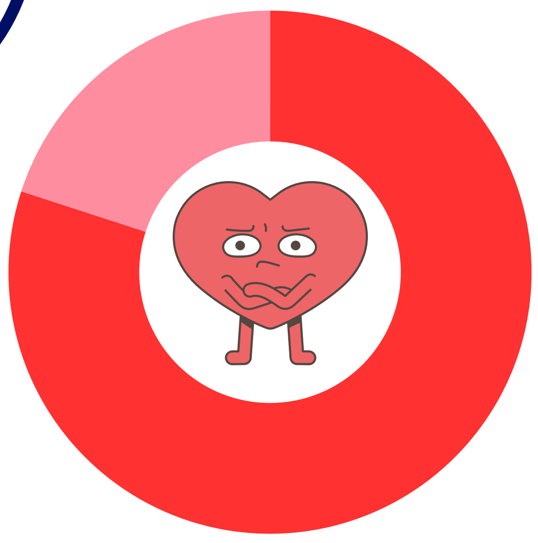
- The largest group is Class 0 (Healthy) with 411 individuals.
- Patient counts decrease as disease severity increases: Class 1 (265) -> Class 2 (109) -> Class 3 (107) -> Class 4 (28).
- Insight: The scarcity of data for Class 4 (Severe) explains the models' struggle to accurately classify advanced disease stages (Class Imbalance problem).

## Proportions Analysis



```
        Wilcoxon rank sum test with continuity correction

data:  age by num_bin
W = 45349, p-value = 1.573e-15
`lternative hypothesis: true location shift is not equal to 0
```



1. Gender Ratio:

- Confirms that Males constitute 78.9% of the dataset, leaving only 21% for Females.

2. Chest Pain Types (cp):

- Key Finding: 53.6% of patients fall into the "Asymptomatic" category.
- In contrast, "Typical Angina" accounts for only 5%.
- Medical Insight: This indicates that over half of the cases present with "Silent Ischemia" (no pain). Relying solely on pain symptoms for diagnosis would miss 50% of patients, validating the need for machine learning models that analyze physiological markers like oldpeak.

3. Thalassemia (thal):

- The distribution is split mainly between "Reversable Defect" (50%) and "Normal" (42%). This balanced variance makes thal a highly effective feature for the Random Forest model to distinguish between healthy and sick individuals.