# OSDA Lazy Fca

Md Shamim Alamgir

December 2024

## 1 Why I Select This Dataset

- It is a well-known dataset for classification problems.
- It contains a mix of numerical and categorical features.
- It is suitable for evaluating the performance of Lazy FCA and other models.
- **Target variable:** Presence of heart disease, which is a binary classification task.

## EDA

### Basic Information

The dataset comprises 303 entries and 14 columns. No missing values were detected.

### Dataset Head

The first few rows of the dataset were examined to gain an initial understanding of the data structure.

### Descriptive Statistics

Key descriptive statistics, including the mean, median, and standard deviation, were calculated for each variable to summarize the data.

### Visualizations

- A bar chart was generated to visualize the distribution of the target variable.
- Histograms were created for all numerical features to understand their respective distributions.

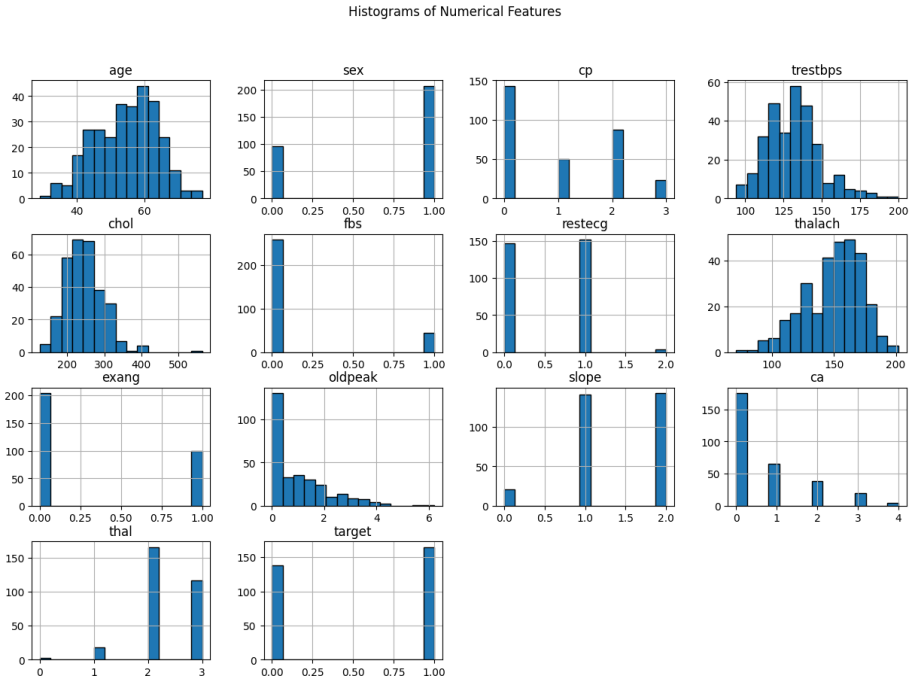
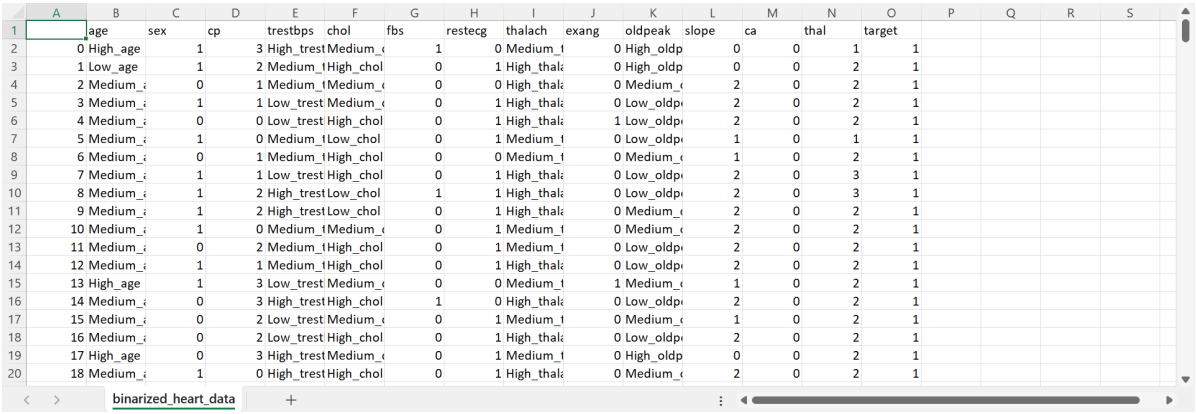

Histograms of Numerical Features
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
target

Figure 1: EDA

## Continuous Variable Analysis and Thresholds

We will analyze the following continuous variables:

1. **Age:**
   - Thresholds: [<40, 40–60, ¿60]
   - Rationale: These thresholds categorize patients into young (<40), middle-aged (40–60), and older (¿60) groups, aligning with established risk stratification for heart diseases.
2. **Resting Blood Pressure (trestbps):**
   - Thresholds: [<120, 120–139, ≥140]
   - Rationale: Based on hypertension guidelines:
     - <120: Normal
     - 120–139: Elevated
     - ≥140: High blood pressure (hypertension)
3. **Cholesterol (chol):**
   - Thresholds: [<200, 200–239, ≥240]
   - Rationale: Categorized as:
     - <200: Desirable
     - 200–239: Borderline high
     - ≥240: High (linked to increased heart disease risk)
4. **Maximum Heart Rate Achieved (thalach):**
   - Thresholds: [<100, 100–160, ¿160]
   - Rationale: Categorized based on exercise performance:
     - <100: Poor exercise capacity
     - 100–160: Normal range for most individuals
     - >160: Excellent capacity (less risk)
5. **ST Depression Induced by Exercise (oldpeak):**
   - Thresholds: [<1, 1–2, ¿2]
   - Rationale: Linked to ischemia severity:
     - <1: Normal/mild
     - 1–2: Moderate
     - >2: Severe risk



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target | | | | |
| 2 | 0 | High_age | 1 | 3 | High_trest | Medium_( | 1 | 0 | Medium_t | 0 | High_oldp | 0 | 0 | 1 | 1 | | | | |
| 3 | 1 | Low_age | 1 | 2 | Medium_t | High_chol | 0 | 1 | High_thal | 0 | High_oldp | 0 | 0 | 2 | 1 | | | | |
| 4 | 2 | Medium_ | 0 | 1 | Medium_t | Medium_( | 0 | 0 | High_thal | 0 | Medium_( | 2 | 0 | 2 | 1 | | | | |
| 5 | 3 | Medium_ | 1 | 1 | Low_trest | Medium_( | 0 | 1 | High_thal | 0 | Low_oldp | 2 | 0 | 2 | 1 | | | | |
| 6 | 4 | Medium_ | 0 | 0 | Low_trest | High_chol | 0 | 1 | High_thal | 1 | Low_oldp | 2 | 0 | 2 | 1 | | | | |
| 7 | 5 | Medium_ | 1 | 0 | Medium_t | Low_chol | 0 | 1 | Medium_t | 0 | Low_oldp | 1 | 0 | 1 | 1 | | | | |
| 8 | 6 | Medium_ | 0 | 1 | Medium_t | High_chol | 0 | 0 | Medium_t | 0 | Medium_( | 1 | 0 | 2 | 1 | | | | |
| 9 | 7 | Medium_ | 1 | 1 | Low_trest | High_chol | 0 | 1 | High_thal | 0 | Low_oldp | 2 | 0 | 3 | 1 | | | | |
| 10 | 8 | Medium_ | 1 | 2 | High_trest | Low_chol | 1 | 1 | High_thal | 0 | Low_oldp | 2 | 0 | 3 | 1 | | | | |
| 11 | 9 | Medium_ | 1 | 2 | High_trest | Low_chol | 0 | 1 | High_thal | 0 | Medium_( | 2 | 0 | 2 | 1 | | | | |
| 12 | 10 | Medium_ | 1 | 0 | Medium_t | Medium_( | 0 | 1 | Medium_t | 0 | Medium_( | 2 | 0 | 2 | 1 | | | | |
| 13 | 11 | Medium_ | 0 | 2 | Medium_t | High_chol | 0 | 1 | Medium_t | 0 | Low_oldp | 2 | 0 | 2 | 1 | | | | |
| 14 | 12 | Medium_ | 1 | 1 | Medium_t | High_chol | 0 | 1 | High_thal | 0 | Low_oldp | 2 | 0 | 2 | 1 | | | | |
| 15 | 13 | High_age | 1 | 3 | Low_trest | Medium_( | 0 | 0 | Medium_t | 1 | Medium_( | 1 | 0 | 2 | 1 | | | | |
| 16 | 14 | Medium_ | 0 | 3 | High_trest | High_chol | 1 | 0 | High_thal | 0 | Low_oldp | 2 | 0 | 2 | 1 | | | | |
| 17 | 15 | Medium_ | 0 | 2 | Low_trest | Medium_( | 0 | 1 | Medium_t | 0 | Medium_( | 1 | 0 | 2 | 1 | | | | |
| 18 | 16 | Medium_ | 0 | 2 | Low_trest | High_chol | 0 | 1 | High_thal | 0 | Low_oldp | 2 | 0 | 2 | 1 | | | | |
| 19 | 17 | High_age | 0 | 3 | High_trest | Medium_( | 0 | 1 | Medium_t | 0 | High_oldp | 0 | 0 | 2 | 1 | | | | |
| 20 | 18 | Medium_ | 1 | 0 | High_trest | High_chol | 0 | 1 | High_thal | 0 | Medium_( | 2 | 0 | 2 | 1 | | | | |

binarized_heart_data

Figure 2: Binarized Data Set

# Binarized Heart Dataset

# Formal Concept Analysis (FCA) and Lazy FCA

1. **Transform the Binarized Dataset:**
   - The discretized dataset will be converted into a binary context table.
   - Rows in this table will represent individual patients (objects).
   - Columns will represent the discretized attributes (e.g., age group, cholesterol level).
   - Cells will indicate the presence (1) or absence (0) of an attribute for each patient.
2. **Compute Formal Concepts:**
   - Formal concepts will be identified from the binary context.
   - Each concept consists of an extent (set of patients) and an intent (set of attributes).
   - The extent includes all patients sharing the attributes in the intent.
   - The intent comprises all attributes common to the patients in the extent.
3. **Apply the Lazy FCA Framework:**
   - Instead of generating the complete concept lattice, a lazy FCA approach will be used.
   - This approach constructs the concept lattice on-demand, based on specific queries.
   - This is more efficient for large datasets and targeted exploration.

## Lazy FCA for Classification and Pattern Extraction

Beyond exploratory analysis, Lazy FCA can be utilized for classification and pattern extraction. Here's how:

1. **Binary Decision Function:**
    - A binary decision function determines if an object (e.g., a patient) belongs to a specific formal concept.
    - This is achieved by checking if the object possesses all the attributes in the concept's intent.
2. **Classifier:**
    - A classifier can be constructed using the concept lattice and a suitable classification strategy.
    - This classifier predicts the class of an object (e.g., presence or absence of heart disease) based on its attribute values and its position within the concept lattice.
3. **Pattern Extraction:**
    - FCA facilitates the extraction of patterns, represented as concept intents.
    - These patterns uniquely define certain extents (groups of objects with shared characteristics).
    - For instance, we can identify attribute combinations that characterize patients with a high risk of heart disease.

## Model Accuracies

The following table presents the accuracies achieved by different classification models on the heart disease dataset:

| **Model** | **Accuracy (%)** |
|---|---|
| Random Forest | 82.42 |
| Logistic Regression | 81.32 |
| Decision Tree | 73.63 |
| SVM | 70.33 |
| Naive Bayes | 83.52 |
| Lazy FCA | 54.46 |

Table 1: Classification accuracies of various models.

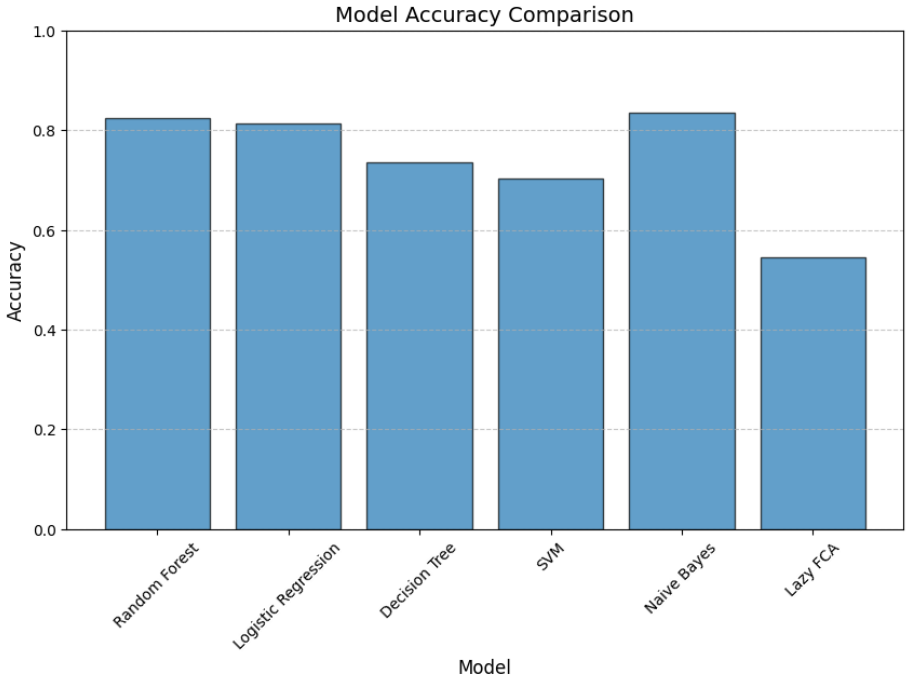



Figure 3: Comparison