جامعة أم القرى
UMM AL-QURA UNIVERSITY

# Data Analysis 2 (Task1)

## Analysis of cardiovascular health indicators using classification algorithms for patient diagnosis


| Name | ID |
|---|---|
| Joud Ahmad Al-huthaly | 444002970 |
| Nehal Hamed Al-zahrani | 444001073 |

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS UMM AL-QURA UNIVERSITY

# Table of content

# 1. Introduction

The dataset under examination is a classic resource used in the field of medical diagnostics, particularly for predicting heart disease. Originating from a collection of databases compiled in 1988, this dataset includes records from Cleveland, Hungary, Switzerland, and Long Beach V. It features a total of 76 attributes, of which 14 key attributes are commonly used in practice for analysis and prediction. This dataset is used to build predictive models and perform statistical analyses aimed at identifying patterns and relationships that may indicate the presence of heart disease. The goal is to utilize various machine learning and statistical techniques to classify patients based on their attributes and assess the effectiveness of different methods in predicting heart disease.

In the following sections, we will explore the dataset through Exploratory Data Analysis (EDA), and apply several classification techniques including Gaussian Naive Bayes, Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). Each method will be evaluated to determine its performance and suitability for predicting heart disease, offering insights into how well these techniques can classify and understand the patterns within this medical data.

# 2. Exploratory Data Analysis

Attribute to the dataset:

**age**: Age of the patient (integer)
**sex**: Gender (1 = male, 0 = female)
**cp**: Chest pain type (categorical, encoded as integers)
**trestbps**: Resting blood pressure (integer)
**chol**: Serum cholesterol in mg/dl (integer)
**fbs**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
**restecg**: Resting electrocardiographic results (categorical, encoded as integers)
**thalach**: Maximum heart rate achieved (integer)
**exang**: Exercise-induced angina (1 = yes, 0 = no)
**oldpeak**: ST depression induced by exercise relative to rest (float)
**slope**: Slope of the peak exercise ST segment (categorical, encoded as integers)
**ca**: Number of major vessels colored by fluoroscopy (integer)
**thal**: Thalassemia (categorical, encoded as integers)
**target**: Target variable indicating the presence of heart disease (1 = disease, 0 = no disease)

# The Dataset

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1.0 | 2 | 2 | 3 | 0 |
| **1** | 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| **2** | 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| **3** | 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0.0 | 2 | 1 | 3 | 0 |
| **4** | 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |
| **5** | 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1.0 | 1 | 0 | 2 | 1 |
| **6** | 58 | 1 | 0 | 114 | 318 | 0 | 2 | 140 | 0 | 4.4 | 0 | 3 | 1 | 0 |
| **7** | 55 | 1 | 0 | 160 | 289 | 0 | 0 | 145 | 1 | 0.8 | 1 | 1 | 3 | 0 |
| **8** | 46 | 1 | 0 | 120 | 249 | 0 | 0 | 144 | 0 | 0.8 | 2 | 0 | 3 | 0 |
| **9** | 54 | 1 | 0 | 122 | 286 | 0 | 0 | 116 | 1 | 3.2 | 1 | 2 | 2 | 0 |

```
[ ] heart.shape
```

```
(1025, 14)
```

This code displays the number of columns and rows in the dataset

. • Number of columns: 14

• Number of rows: 1025

We used the "describe" function to explain the values such as the mean, median, minimum, and maximum values. This function provides an overview of the available variables' data distribution and basic statistics.

```
[ ] heart.describe()
```

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1025.000000 | 1025.000000 | 1025.000000 | 1025.000000 | 1025.00000 | 1025.000000 | 1025.000000 | 1025.000000 | 1025.000000 | 1025.000000 | 1025.000000 | 1025.000000 | 1025.000000 | 102 |
| mean | 54.434146 | 0.695610 | 0.942439 | 131.611707 | 246.00000 | 0.149268 | 0.529756 | 149.114146 | 0.336585 | 1.071512 | 1.385366 | 0.754146 | 2.323902 | |
| std | 9.072290 | 0.460373 | 1.029641 | 17.516718 | 51.59251 | 0.356527 | 0.527878 | 23.005724 | 0.472772 | 1.175053 | 0.617755 | 1.030798 | 0.620660 | |
| min | 29.000000 | 0.000000 | 0.000000 | 94.000000 | 126.00000 | 0.000000 | 0.000000 | 71.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | |
| 25% | 48.000000 | 0.000000 | 0.000000 | 120.000000 | 211.00000 | 0.000000 | 0.000000 | 132.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 2.000000 | |
| 50% | 56.000000 | 1.000000 | 1.000000 | 130.000000 | 240.00000 | 0.000000 | 1.000000 | 152.000000 | 0.000000 | 0.800000 | 1.000000 | 0.000000 | 2.000000 | |
| 75% | 61.000000 | 1.000000 | 2.000000 | 140.000000 | 275.00000 | 0.000000 | 1.000000 | 166.000000 | 1.000000 | 1.800000 | 2.000000 | 1.000000 | 3.000000 | |
| max | 77.000000 | 1.000000 | 3.000000 | 200.000000 | 564.00000 | 1.000000 | 2.000000 | 202.000000 | 1.000000 | 6.200000 | 2.000000 | 4.000000 | 3.000000 | |

data quality was improved by checking null values using the (isnull().sum()) function.

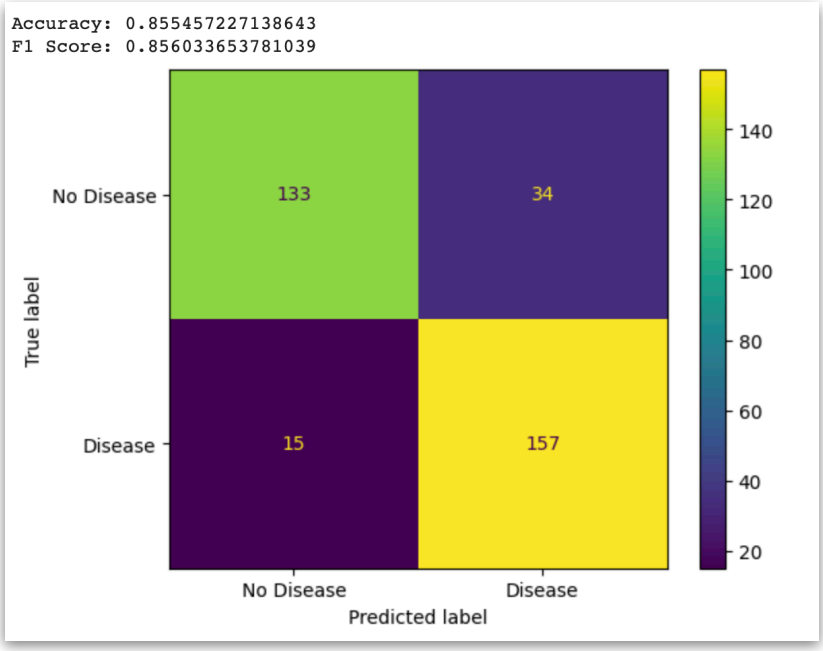
```
[ ] heart.isnull().sum()
```

| | 0 |
|---|---|
| age | 0 |
| sex | 0 |
| chest_pain_type | 0 |
| resting_blood_pressure | 0 |
| cholesterol | 0 |
| fasting_blood_sugar | 0 |
| resting_electrocardiogram | 0 |
| max_heart_rate_achieved | 0 |
| exercise_induced_angina | 0 |
| st_depression | 0 |
| st_slope | 0 |
| num_major_vessels | 0 |
| thalassemia | 0 |
| target | 0 |

**dtype:** int64

# 3. Gaussian Naive Bayes

The Gaussian Naive Bayes algorithm was used because it is the best fit, since most of the features used in the model are continuous




The code separates features from the target column (a person has heart disease), then splits the data into training (67%) and testing (33%). The model is built using the Gaussian Naive Bayes algorithm, and the model accuracy was 85%. The model misclassified 34 people without the disease as infected, and 15 people with the disease as uninfected

# 4. Logistic Regression

This code aims to use logistic regression to analyze data on heart patients and predict the probability of developing the disease based on age. A model is trained on data including age and injury outcome (target). After training the model, it is used to estimate the probability of infection for a 46-year-old person, with a score of 1, indicating a high probability of contracting the disease.

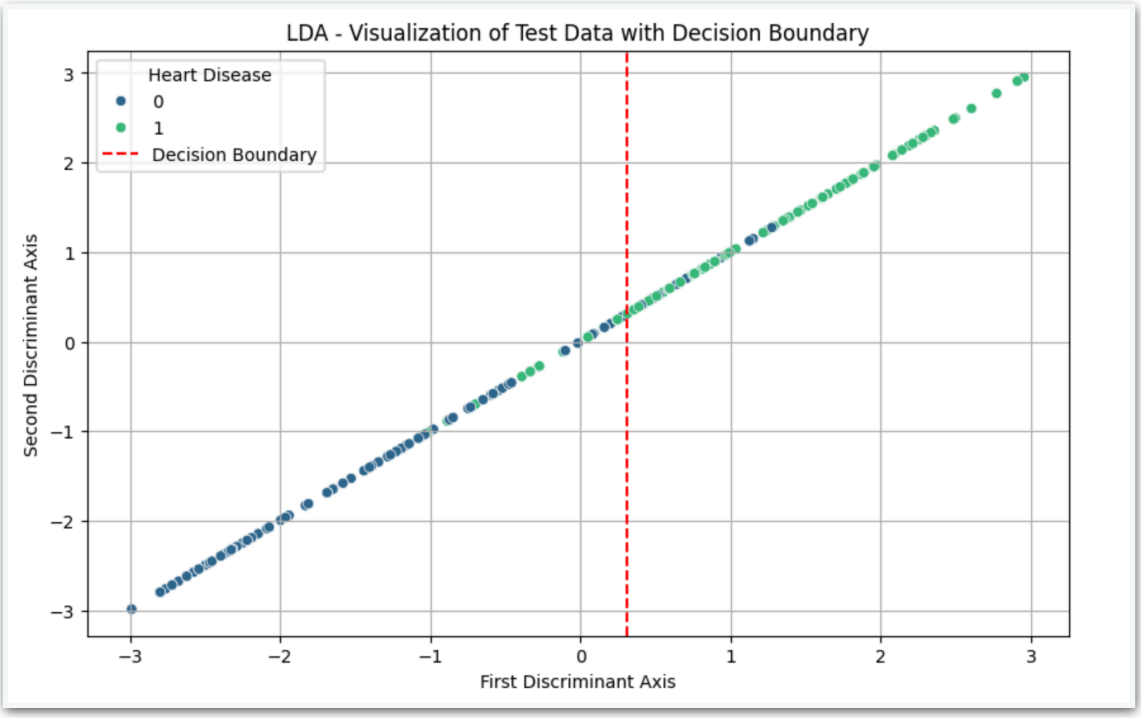
```
predicted: [1]
[[0.54656748]
 [0.53340285]
 [0.31708524]
 ...
 [0.61107051]
 [0.57268454]
 [0.52019154]]
```

# 5. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a technique used for classification that projects data onto a line to maximize the separation between different classes while minimizing the variance within each class. It seeks to find a linear combination of features that best distinguishes between the predefined classes in the data.

The LDA model shows **good performance** in predicting heart disease, with an accuracy of **84.18%**. Key insights:

- It is better at identifying patients with heart disease (**recall of 0.90**) compared to identifying those without it (**recall of 0.78**).
- The model has a **high precision** for both classes (0.88 for no disease and 0.81 for disease), meaning it makes fewer false positive errors.



# 6. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) is a classification technique that estimates class-specific probability distributions, allowing for different variances between classes. It finds decision boundaries as quadratic surfaces to best separate the classes.

The Quadratic Discriminant Analysis (QDA) model achieves an accuracy of **82.14%**. Key findings:

- The model performs slightly better at identifying patients **without heart disease** (precision of 0.87) than those with heart disease (precision of 0.78).
- The **false positive rate** is relatively higher, with 36 false positives (no disease misclassified as disease) and 19 false negatives (disease misclassified as no disease).