**Faculty of Engineering & Technology**

**Electrical & Computer Engineering Department**

**MACHINE LEARNING AND DATA SCIENCE**

**ENCS5341**

**Assigment\_3**

**Prepared by:**  Salwa Fayyad   1200430

                       Sahar Fayyad    1190119

**Instructor**:      Dr. Yazan Abu Farha

**Section**:            2

**Date**:                20.1.2023

Table of Contents

[**Introduction** 1](#_Toc156673323)

# **Introduction**

In this project, we aim to explore and analyze a comprehensive dataset focusing on predicting heart disease. The dataset comprises various clinical and demographic features: Age, Sex, Chest Pain Type, Resting Blood Pressure (RestingBP), Cholesterol level, etc, and the presence of Heart Disease (target).

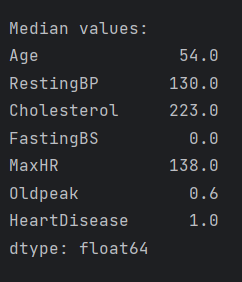
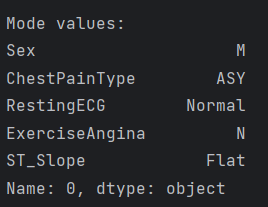
The dataset comprises diverse patient profiles, ranging from a 36-year-old male with atypical angina and a normal cholesterol level to a 60-year-old male with asymptomatic chest pain and high cholesterol. These attributes are essential in understanding the patterns and risk factors associated with heart disease.

In the initial phase of our analysis, we implemented a Nearest Neighbor baseline model to understand the basic patterns and correlations within the data. This model, chosen for its simplicity and effectiveness in capturing the nearest similarities in multidimensional data, was evaluated with both k=1 and k=3 to observe the variation in performance.

For our project on predicting heart disease, we used two different machine learning methods: Support Vector Machine (SVM) and Random Forest. These methods are good for sorting data into categories, like figuring out if someone has heart disease or not. SVM is great for working with complex data and finding patterns that aren't obvious. Random Forest is good for dealing with a lot of data without getting confused. We used both methods to see which one is better at identifying the signs of heart disease. This helps doctors figure out who is at risk. Our project shows how these smart computer methods can be really useful in health care.

To evaluate the performance of these models, we used specific evaluation metrics. These metrics are essential for understanding how well our models work in predicting heart disease. The key metrics we utilized include: Accuracy tells us how often the models make correct predictions. Precision measures the accuracy of the models in identifying positive cases, while Recall assesses their ability to find all actual cases of heart disease. The F1 Score combines precision and recall to provide a balanced view of model performance. Together, these metrics provide a comprehensive assessment of our models' performance in predicting heart disease.

# **Preprocessing dataset**

The EDA (Exploratory Data Analysis) function provided is designed to perform a comprehensive analysis of a dataset, with a focus on both quantitative and qualitative factors such as mean, median, and mode for numeric and non-numeric columns respectively. As shown in the figures below:

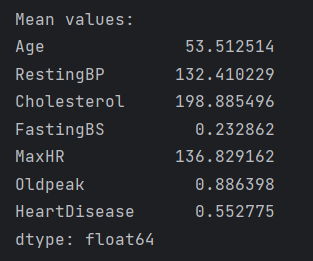


Figure 1: Quantitive measure

The function primarily targets datasets with a mix of numerical and categorical features related to heart disease and visualize them using histograms for numerical features and bar plots for categorical features.

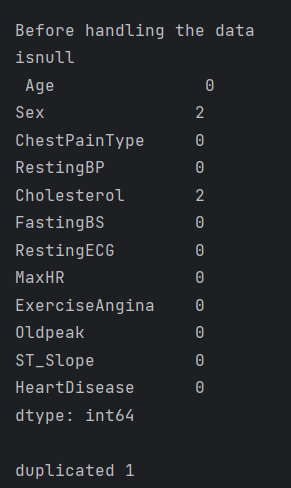
 By utilizing median and mode, the function ensures that there are no missing values and duplicates by filling in the missing ones and eliminating duplicates. With this comprehensive approach, the dataset is thoroughly understood and preprocessed, leaving it ready for further analysis or modeling.

Figure 2: Before and after handling data

Typically, the median is less susceptible to the presence of outliers (compared to the mean), I opted to fill the missing data using the median method in the numerical columns.

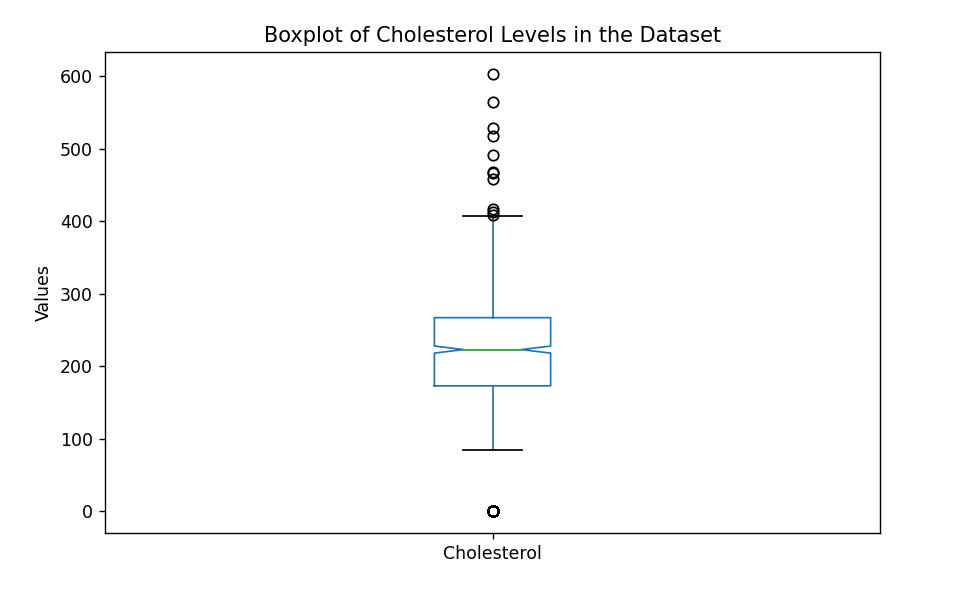
For non-numeric columns, it calculates and displays the mode, which is the most frequently occurring value in each column.

Figure 3: Boxplot of Cholestrol levels in the dataset

