CESS Program

Fall 2024

Introduction to Machine Learning

CSE 381

**Heart Disease Prediction**

Submitted To:

Dr. Mahmoud Khalil

Eng. Engy Ahmed

Team Members

|  |  |
| --- | --- |
| Siveen Said Sayed Ahmed | 21P0149 |
| Mariam Mahmoud Abdou | 21P0214 |
| Malak Mohamed Salem | 21P0277 |

# 1. Introduction

This project presents a comprehensive study on predicting heart disease using machine learning techniques. It aims to analyze a dataset containing various medical and physiological parameters to predict the presence of heart disease. The project encompasses data preprocessing, exploratory data analysis, feature engineering, model selection, evaluation, and reasoning behind every step.

# 2. Dataset Description

## Features

The dataset contains information on various attributes representing patient data:

* **Age**
* **Sex**: "M" for male, "F" for female.
* **ChestPainType**: Describing the type of chest pain experienced.

("TA" for typical angina, "ATA" for atypical angina, "NAP" for non-anginal pain, "ASY" for asymptomatic). [1]

* **RestingBP**: Representing the resting blood pressure (in mmHg).
* **Cholesterol**: Representing the serum cholesterol level (in mg/dL).
* **FastingBS**: Indicating if fasting blood sugar is greater than 120 mg/Dl

(1 = Yes, 0 = No).

* **RestingECG**: Describing the results of a resting electrocardiogram (ECG) test, which measures the electrical activity of the heart.

("Normal," "ST," "LVH" for left ventricular hypertrophy). [2]

* **MaxHR**: Representing the maximum heart rate achieved.
* **ExerciseAngina**: Indicating if exercise-induced angina is present.

(1 = Yes, 0 = No).

* **Oldpeak**: Representing the ST depression induced by exercise relative to rest (often used to measure ischemia).
* **ST\_Slope**: Representing the slope of the peak exercise ST segment.

("Up," "Flat," "Down").

* **HeartDisease**: Indicating the presence of heart disease. (1 = Yes, 0 = No).

## Features Nature

Categorical features: Sex, ChestPainType, RestingECG, ST\_Slope.

Numerical features: Age, RestingBP, Cholesterol, FastingBS, MaxHR, ExerciseAngina, Oldpeak.

Target: HeartDisease.

# 3. Data Preprocessing

## 3.1 Handling Range of Values

For each feature in the dataset, its range was examined to identify potential anomalies. If a value was found to be outside the expected range, it was either considered a missing value or corrected based on logical inference. This ensured the data quality was upheld for further analysis. A detailed breakdown of feature ranges and identified issues has been documented separately to maintain transparency and guide preprocessing steps.

### Range of values

* **Age** Range: 0 to 120 years.
* **Sex** Values: ['M', 'F'].
* **ChestPainType** Values: ['ATA', 'NAP', 'ASY', 'TA'].
* **RestingBP (Resting Blood Pressure)** Range: 0 to 250 mmHg.

Note: Values below 90 or above 200 are clinically abnormal.

* **Cholesterol**: Range: 0 to 600 mg/dL.

Note: Values above 300 are rare and usually indicate a serious condition.

* **FastingBS (Fasting Blood Sugar)** Values: [0, 1].
* **RestingECG** Values: ['Normal', 'ST', 'LVH'].
* **MaxHR** Range: 0 to 220 bpm.

Note: Typical values range from 60 bpm (resting) to around 200 bpm (during exercise).

* **ExerciseAngina** Values: [0, 1].
* **Oldpeak** Range: 0 to 10.0 mm.

Note: Values above 5.0 are rare and clinically significant.

* **ST\_Slope** Values: ['Up', 'Flat', 'Down']

### Issues in data

* Max observed value in **Cholesterol** is 603. Since only one value is outside the expected range, it might be plausible and will be addressed during outlier detection.
* Negative values observed in **Oldpeak**. Since it cannot be negative, the negative values are considered entry errors and are converted to positive values that it is inside the absolute range.

## 3.2 Missing Values

The dataset initially had no missing values; however, outliers were identified and replaced with null values. These null values were subsequently handled using the KNN-imputer.

## 3.3 Handling Outliers

Outlier detection was performed for both categorical and numerical data to ensure data quality and reliability.

### Numerical data

The Z-score method was applied to detect outliers. This method identifies values that deviate significantly from the mean, typically using a threshold of |Z| > 3.

These outliers were added to a list for further handling, such as replacement, removal, or adjustment based on their impact on the analysis and model performance.

The maximum number of outliers detected in a column was 7 out of 918 rows. Since this is a small proportion, and they are meaningful and provide important insights (as rare but valid occurrences in medical datasets), they will be retained in the dataset.

### Categorical data

Infrequent categories were identified in each feature based on a threshold of 5% of the dataset. Categories present less frequently than the threshold were considered potential outliers and added to a list for further analysis. Handling these categories typically involves decisions such as grouping them into an "Other" class, retaining them if they represent valid rare occurrences, or treating them as noise based on their significance. However, no outliers were detected, as all values were frequent and met the defined threshold criteria.

## 3.3 Encoding Categorical Variables

Since machine learning algorithms generally require numerical input, categorical variables needed to be converted into a numeric format through encoding. This transformation ensures the data is suitable for the algorithms.

For this purpose, **Label Encoding** was applied, which assigns a unique numerical value to each category within a variable.

## 3.3 Feature Scaling

Machine learning algorithms, especially those that rely on distance-based metrics, perform better when features are on a similar scale. Feature scaling helps prevent features with larger numerical values from disproportionately influencing the model and improves convergence during training, ultimately leading to better model performance.  
For this, continuous variables were scaled using **Standard Scaler**. Standardization ensures uniformity by transforming the data to have a mean of 0 and a standard deviation of 1.

# 4. Model Development

The data was split into training and testing sets using an 80-20 split. Models were trained using cross-validation to ensure robustness.

## 4.1 Model Training

Several machine learning algorithms were considered, including:

* Logistic Regression
* Random Forest
* Support Vector Machine (SVM)
* XGBoost

## 4.2 Model Evaluation

Evaluation metrics included:

* Accuracy
* Precision
* Recall
* F1-Score
* Area Under the Receiver Operating Characteristic (ROC-AUC)

Random Forest emerged as the best-performing model with an accuracy of [value] and an ROC-AUC score of [value].

# 7. Results and Discussion

## 7.1 Key Findings

The Random Forest model effectively predicted heart disease with high accuracy and recall.

* Features like Age, Maximum Heart Rate, and ST Depression were the most significant predictors.

## 7.2 Limitations

* The dataset size was relatively small, which might affect generalizability.
* Imbalanced classes required oversampling techniques like SMOTE.

## 7.3 Future Work

* Collecting a larger and more diverse dataset.
* Incorporating deep learning techniques for better accuracy.
* Enhancing feature engineering with domain-specific metrics.

# 8. Conclusion

This study demonstrated the potential of machine learning in predicting heart disease. By leveraging patient data, the Random Forest model provided reliable predictions, aiding early diagnosis and intervention. Future improvements can further enhance the model’s applicability in clinical settings.

# 9. Appendices

[1] **ChestPainType** values representing different types of chest pain, as follows:

**ATA (Atypical Angina)**: This refers to chest pain that is not typical of angina. It can be caused by a variety of factors and might not follow the usual pattern of angina pain.

**NAP (Non-Anginal Pain)**: This type of chest pain is not related to heart disease. It might arise from other conditions like digestive issues or musculoskeletal problems.

**ASY (Asymptomatic)**: This refers to a lack of symptoms. Individuals with asymptomatic chest pain may not feel any discomfort, even if they have underlying heart disease.

**TA (Typical Angina)**: This refers to chest pain caused by restricted blood flow to the heart, often triggered by physical exertion or stress. It is a common symptom of coronary artery disease.

[2] **RestingECG** column values:

**Normal**: The electrocardiogram shows normal heart rhythm with no signs of ischemia or abnormal electrical activity.

**ST**: This indicates ST-segment depression or elevation, which is often associated with myocardial ischemia (reduced blood flow to the heart).

**LVH (Left Ventricular Hypertrophy)**: refers to thickening of the heart's left ventricle, which can occur because of high blood pressure or other heart conditions.