ML 1- Yilmaz
2024-2025

Name of the group members (max 2 people): Aryan Gadre, Adarsh Bharadwaj
If you are working alone, please state it clearly.
If you are working in groups, only one group member should submit the proposal.

Must haves:

- Link to dataset
- Information about your data
    - Meaning of attributes
    - Dimension
    - Number of instances
    - How many missing values
    - Is it uniform? Skewed?
    - What are the class distributions?
- What are you classifying/predicting?
- How will this be useful?
- Plans for preprocessing

Don't just insert your answers in the above, make it look like you're writing the portions of the final report. See the final report sample I posted at Schoology.
As soon as you are ready with your proposal, please see me, so that I can approve it.

Also, please add your project to the spreadsheet I will provide under Q1 Project Folder.

Link to Dataset

Purpose

According to the CDC, every year about 695,000 people in the country die from heart disease, equivalent to 1 out of 5 deaths in the United States. Out of all of those deaths, approximately 1 out of 3 deaths could have been prevented with earlier diagnosis. This dataset contains heart disease data from the 5 most popular independent heart disease datasets.

- Hungarian
- Cleveland
- Switzerland
- Long Beach VA
- Statlog (Heart) Data Set

About The Dataset

The dataset contains information about heart disease, with various features used for predicting heart conditions. This dataset is curated by combining 5 popular heart disease datasets about individuals related to heart disease.

Dataset Summary:
- File Name: heart.csv
- Number of Instances (Rows): 1190
- Number of Non-Class Attributes (Columns): 11
- Missing Values: None
- Class Attribute: "target" represents whether the individual has heart disease (1) or not (0).

Attributes (Columns):
1. age: Age of the patient
2. sex: Gender of the patient (1 = Male, 0 = Female)
3. chest pain type: Chest pain type (4 values)
4. resting bp s: Resting blood pressure (in mmHg)
5. cholesterol: Serum cholesterol in mg/dl
6. fasting blood sugar: Fasting blood sugar > 120 mg/dl (1 = True, 0 = False)
7. resting ecg: Resting electrocardiographic results (values 0, 1, 2)
8. max heart rate: Maximum heart rate achieved
9. exercise angina: Exercise induced angina (1 = Yes, 0 = No)
10. oldpeak: ST depression induced by exercise relative to rest
11. ST slope: The slope of the peak exercise ST segment

Class Attribute (Target):
Target Distribution: The dataset contains two classes:
- Class 0: No heart disease (47.1% of the instances)
- Class 1: Heart disease (52.9% of the instances)

Data Characteristics:
- Missing Values: No missing values.
- Data Distribution: The dataset has a slightly imbalanced class distribution, with a skew towards individuals diagnosed with heart disease (52.9%).

Preprocessing

1. Normalization
To ensure that our models don’t overestimate the impact of one specific attribute, we intend to perform min-max normalization to transpose all values on a scale from 0 to 1. This process would have to be done for the attributes: Age, resting bp s, cholesterol, max heart rate, oldpeak, and st slope.

2. Missing and Disguised Values

As mentioned earlier the dataset does not have any missing values but it has some disguised missing values. For resting heart rate there are a couple of values that have 0 listed as the resting heart rate. As this is not possible it can be deduced that the values that are 0 are disguised missing values. We will use the mean of the remaining values to replace those values.

3. Train-test split

To split the data into training and testing sets we will randomly select 85% of instances to use for the training set and then use the remaining 15% for the testing set. It is appropriate to randomly split because the distribution of the classes is close to even, so there is highly unlikely to be a major difference in the class distribution of the test set and the training set. This split will result in 1012 instances in the training set and 178 instances in the testing set.

4. Further Considerations

Depending on the accuracy of our models or if our models take far too long to train we will likely do further preprocessing such as PCA or correlation analysis to reduce the amount of data we will need to train on and to reduce the amount of noise in our dataset.