Supplementary Material

❖ Glossary section

- Definitions of the variables in the dataset based on the description of the dataset in Kaggle [4] and [1]:
  - Age: Patient's age in complete years.
  - Sex: Patient's gender (value 1 represents male and value 0 represents female).
  - Cp: Chest pain type categorized in 4 levels. The first level (value 1) represents a typical angina. The second level (value 2) represents an atypical angina. The third level (value 3) represents a non-anginal pain and finally the fourth level (value 4) is asymptomatic.
  - trestbps: Describes the resting blood pressure (counted in mmHg, upon admission to the hospital)
  - chol: serum cholesterol in mg/dl
  - fbs: blood sugar levels on fasting >120mg/dl (true case is represented by the value 1, false case is represented by value 0)
  - Results of electrocardiogram test while at rest. The results are represented by 3 values: value 0 represents a normal state, value 1 represents abnormality in ST-T wave (T wave inversions and/or ST elevation or depression of > 0.05 mV) and value 2 represents the probable or definite left ventricular hypertrophy by Estes'
  - thalach: The greatest number of beats per minute a heart can reach during exercise.
  - exang: Angina induced by exercise (value 1 represents a true case and value 0 represents a false case).
  - oldpeak: ST depression induced by exercise relative to rest (in mm, achieved by subtracting the lowest ST segment points during exercise and rest)
  - slope: ST segment in terms of the slope during peak exercise (value 1 represents upsloping, value 2 represents flat slope and value 3 represents downsloping).
  - ca: number of major vessels coloured by fluoroscopy.
  - thal: Status of the heart (value 1 represents a normal state of the heart, value 2 represents a fixed defect heart state which means that the heart tissue can't absorb thallium both under stress and in rest and value 3 represents a reversible defect heart state which means that the heart tissue can't absorb thallium only under the exercise portion of the test)
- Angina: Chest pain caused by reduced blood flow to the muscles [3].
- ST Segment: In electrocardiography, the ST segment refers to a specific part of an electrocardiogram that measures the electrical activity of the heart to find heart abnormalities [2].

References for glossary:

[1] S. Mohan, C. Thirumalai, and G. Srivastava, 'Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques', *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[2] 'ST Segment', 2012. Accessed Dec 20, 2023. [online]. Available at: https://www.sciencedirect.com/topics/medicine-and-dentistry/st-segment

[3] 'Overview Angina'. Accessed Dec 20, 2023. [online]. Available at: https://www.nhs.uk/conditions/angina/#:~:text=Angina%20is%20chest%20pain%20caused,of%20these%20more%20serious%20problems .

[4] Heart Attack Dataset (About Dataset). [online]. Available at: https://www.kaggle.com/datasets/pritsheta/heart-attack

❖ Intermediate results

- **Feature selection**: Before feature selection, we created a correlation heatmap that shows the correlation coefficient between the variables and a heatmap that shows the p-value of each correlation coefficient. This will help us identify which variables are strongly correlated to our dependent variable and identify independent variables that have multicollinearity. In the end, we chose the variables that have the strongest, statistically significant correlation with our dependent variable. We made sure all selected features have a correlation coefficient lower than 0.5 between them.

- **Baseline model results:** First, we created and trained 2 models, one for each machine learning algorithm used to evaluate and compare the results before adding hyperparameter optimization. This will help us understand how the optimal hyperparameters that fit to the models improve their performance. The first results showed that the naïve bayes model performs better overall but the decision trees model has higher precision. After adding the hyperparameters we see that the naïve bayes still performs better overall but the decision trees model has higher precision and equal accuracy.

- **Unexpected anomalies in the MATLAB code:** There is a problem with the reproducibility of the code. Although the random seed was set to rng(1) for all models, every time I run the code the best naïve bayes model with hyperparameters (BestNBModel.mat) has slightly different metrics. The metrics don't change significantly and don't affect the overall evaluation and comparison of the 2 optimized models but there are still some differences. This only happens to the naïve bayes best model.

❖ **Relevant implementation details that were not included in the poster**

- **Exploratory data analysis and data preprocessing graphs:**

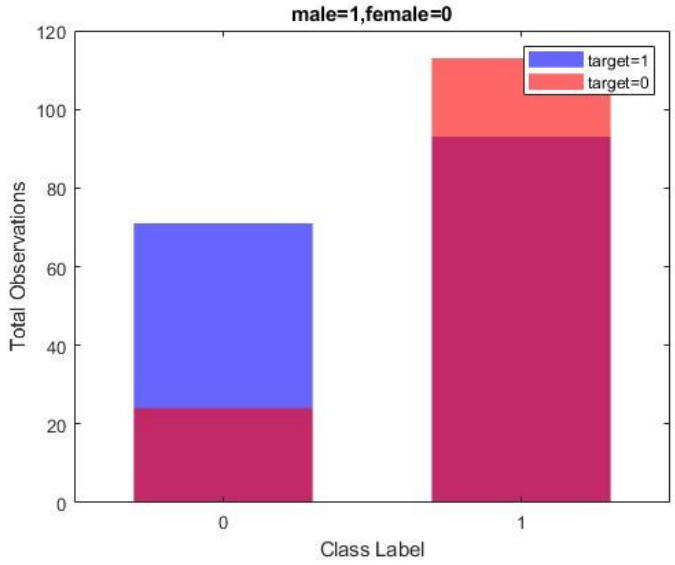

Figure 1. Male and Female patients with and without heart attack

In this dataset, it seems that more females belong to the class of patients with no heart attack and more males belong to the class of patients with heart attack. This indicates that the sex might have a correlation with our dependent variable.

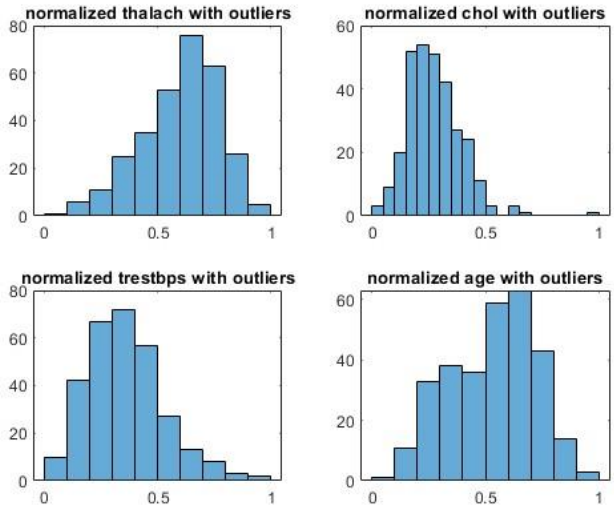

Figure 2. Data Distribution and skewness of variables after normalization and with outliers

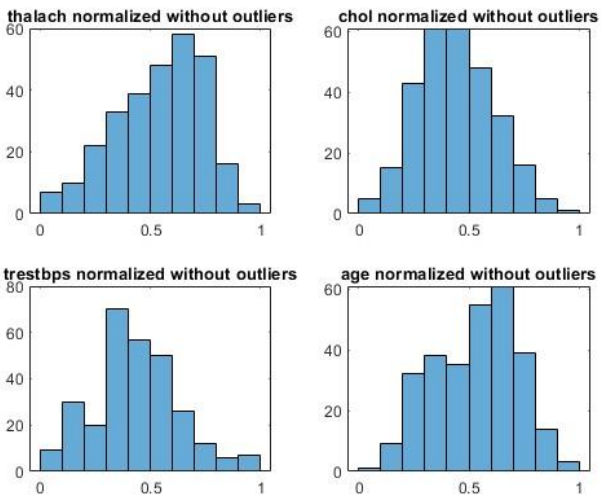

Figure 3. Data Distribution and skewness of variables after normalization and without outliers

ables with d
ical values w

show the distribution and skewness of these variables after being normalized in comparison to before and also before and after removing the outliers. A total of 16 observations were removed as outliers (more than 3 standard

deviations away from the mean). We can see that the variables follow a more normal distribution after removing the outliers in comparison to the distribution they have before. So based on these 2 graphs we decided to completely remove the 16 observations with outliers from the dataset. So from a total of 303 rows the dataset what reduced to 287. Around 5% of the rows were removed which is not a high percentage and shouldn't affect the results.

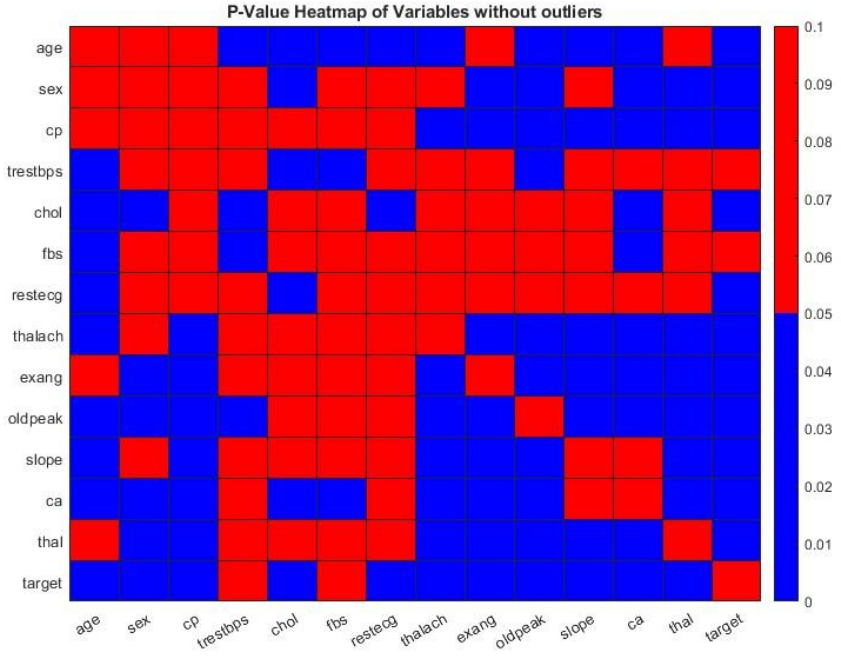

Figure 4. Heatmap of spearman correlation coefficient p-values between the variables

As explained before, feature selection was necessary to identify which variables are more useful for building and training the models and which are irrelevant and might produce false results. After creating a correlation heatmap that shows the relationship between each pair of variables, we also created a p-value heatmap of the correlation coefficients to test the significance of the relationships. As we can see in the colour scale, we set a threshold of 0.05, meaning that relationships with a p-value less than 0.05 indicate that we are more than 95% confident that the observed correlation is significant and not due to random chance. The blue correlations show statistically significant correlations and the red correlations the opposite. So we only chose independent features that have a statistically significant strong correlation with our dependent variable (p-value<0.05) and a low, statistically significant correlation between them.