DATA ANALYSIS REPORT

Index

VALUES

The dataset is characterized by fourteen different variables:

- age
- sex
- cp: this column indicates the different types of chest pain. In total, there are four different types of pain. A cp of zero indicated that the chest pain is correlated with a “typical angina”, a cp of one corresponds to a “atypical angina”, a cp of two indicated a pain not correlated with the angina, while a cp of three indicates a pain that is asymptomatic.
- trtbps: this column indicates the resting blood pressure in mm/Hg.
- chol: this column indicates the cholesterol in mg/dl.
- fbs: this column indicates when the fasting blood sugar is higher than 120 mg/dl. If it is higher, the value is one, otherwise it is zero.
- restecg: the column indicates the value for the resting electrocardiographic results. We get three different values based on the result. A value of zero if the result is normal, a value of one if the person has a ST or T wave abnormality, a value of two if the result shows a probable or definite left ventricular hypertrophy.
- thalachh: this column indicates the maximum heart rate achieved.
- exon: this column indicates the angina induced by doing exercise. There are two possibilities: one for yes and zero for no.
- oldpeak: this column indicates ST depression caused by activity in comparison to rest.
- slp: this column indicates the slope of the peak exercise ST segment. A value of zero means to have a better heart rate with exercise (this is uncommon), a value of one means that there is minimal change (it is the typical healthy heart), a value of two means that there are signs of an unhealthy heart.
- caa: this column indicates the number of major vessels, and it can have a value going from zero to four.
- thall: this column indicates a blood disorder called thalassemia. Based on the thalium stress result we get three different values. One if it is normal, two if there is a fixed defect and three if it is a irreversible defect.
- output: this column is our target variable. A value of one indicates that there is a higher possibility of having a heart attack, while a value of zero indicates a smaller possibility of having a heart attack.

EXPLORATORY DATA ANALYSIS

Before delving into exploratory data analysis (EDA), it is essential to conduct preliminary data cleaning. The dataset indicates zero missing values and columns, with no repetitions, comprising 9 discrete and 5 continuous variables. With 165 individuals having a high

chance of a heart attack and 138 less likely, we can say that the dataset is reasonably balanced.

For the EDA we decided to focus on univariate and bivariate analysis.
Univariate analysis allows us to comprehend the distribution of individual variables in isolation. This includes examining the central tendency, spread, and shape of each variable's distribution. Bivariate analysis on the other hand, helps to uncover associations, dependencies, or correlations between two variables, providing insights into how they interact with each other.

Univariate Analysis

For the examination of variable distributions, categorical variables were assessed through bar charts while continuous through histograms.

In the realm of continuous variables, a normal distribution is apparent across all, except for "oldpeak," showcasing left skewness and "thalachh" exhibiting right skewness. Olpeak skewness may be attributed to a minimum value at 0.00, potentially influencing the leftward skew. A concentration of lower values, evident in a mean of 1.04 and a median of 0.80, contributes to it as well, signifying a prevalence of these lower values.

The variable "thalachh" exhibits right skewness, influenced by physiological limits on the maximum heart rate and a notable portion with lower physical activity levels. A natural upper limit creates a right tail in the distribution, emphasizing the rightward skewness.

Turning to categorical variables, a distinct gender imbalance is observed, with 207 male individuals and 96 females. Analyzing the distribution of chest pain types reveals that Typical angina (type 0) and non-anginal pain (type 2) are more prevalent, suggesting diverse experiences of chest pain among individuals.
Examining Fasting Blood Sugar (fbs), most individuals register a fasting blood sugar level below 120 mg/dl (coded as 0), implying a lower prevalence of high fasting blood sugar in the dataset. This observation reflects the overall health status of the study population.

Regarding Resting Electrocardiographic Results (restecg), Type 1, indicating ST-T wave abnormality, is most prevalent, focusing on Exercise Induced Angina (exng) the majority of the data pool did not experience exercise-induced angina.
Moreover Number of Major Vessels Colored by Fluoroscopy (caa), the majority have 0 or 1 major vessels colored, with decreasing frequencies for higher counts, while analyzing Thalassemia (thall), Type 2 emerges as the most common.

Bivariate Analysis

For the bivariate analysis we decided to plot all variables against the target variable output equal to 1 that represents a higher chance of heart attack.

For categorical variables we identified that most heart attack cases occur in males, and a significant majority of individuals exhibit Non Anginal Chest Pain Type. Furthermore, a notable proportion of individuals that are more associated with heart attack do not experience exercise-induced angina, this is most likely due to a sedentary lifestyle. In the dataset, individuals displaying ST-T wave abnormality in resting electrocardiogram results are more likely to experience heart attack.

Additionally, among those with a higher chance of heart attack, a substantial portion has a ST slope that is downward, this indicates signs of an unhealthy heart. In regards to the number of Major Vessels colored by fluoroscopy (ca), those with ca equal to 0 (indicating no major vessels colored) are more likely to have heart attack as increased blood movement is considered beneficial.

Focusing on continuous variables we identified that individuals in the middle age range (40 to 60 years) exhibit a higher likelihood of experiencing a heart attack.
The positive correlation of maximum heart rate achieved and the negative correlation of previous peak with the likelihood of heart attacks imply that a higher maximum heart rate reflects increased cardiac demand, while a lower previous peak mght indicate compromised cardiac function. Resting Blood Pressure (trtbps) shows a weak or slightly negative correlation with the occurrence of a heart attack while there were no relevant finding regarding the relationship between cholesterol and heart attack

CORRELATION

To compute the correlation between variables we used the aid of a correlation matrix. The most relevant findings in regards to a positive correlation with the output are:

- cp (Chest Pain Type): There is a strong positive correlation. This suggests that as the severity of chest pain type increases, the likelihood of a heart attack also increases.
- thalachh (Maximum Heart Rate Achieved): A strong positive correlation implies that individuals with a higher maximum heart rate achieved during exercise are more likely to experience a heart attack.
- slp (Slope of the Peak Exercise ST Segment): A positive correlation suggests that a specific pattern in the ST segment during exercise is associated with a higher probability of a heart attack.

Focusing on the negative correlations with output we have:

- exng (Exercise Induced Angina): A negative correlation indicates that the absence of exercise-induced angina is associated with a higher likelihood of a heart attack.
- oldpeak (ST Depression Induced by Exercise): The negative correlation suggests an inverse relationship. Lower ST depression values during exercise are associated with a higher probability of a heart attack.

- caa (Number of Major Vessels Colored by Fluoroscopy): A negative correlation implies that a higher count of major vessels colored is associated with a lower likelihood of a heart attack.
- sex: The negative correlation with 'output' indicates that being male is associated with a higher likelihood of a heart attack.

When looking at relationships between variables excluding the output we identify a moderate positive correlation between maximum heart rate and the slope of the peak exercise ST segment, this means that individuals with a higher maximum heart rate are more likely to exhibit signs of an unhealthy heart during peak exercise. The positive correlation between oldpeak (previous peak) and exercise-induced angina shows that individuals with lower values of oldpeak are more likely to experience exercise-induced angina.

Additionally we see that previous peak has a strong negative correlation with slope of the peak exercise ST segment, this implies that individuals with a lower previous peak are more likely to have a downsloping ST segment during exercise.

PRINCIPAL COMPONENT ANALYSIS (PCA)

For the principal component analysis we decided to remove the output variable and then scale the variables to standardize the data, ensuring that all variables contribute equally to the PCA.

We decided to choose the number of components to explain almost 80% of variance which is 8 that computes a cumulative variance of 79.5%.

The importance of each principal component (PC) was assessed based on standard deviation, proportion of variance and cumulative proportion. The first principal component (PC1) has the highest variance, contributing to 21.25% of the total variance in the data, each successive component contributes to a decreasing proportion of variance.

Further exploration, including clustering and predictive modeling will be later provided.

HIERARCHICAL CLUSTERING

Applying hierarchical clustering used the aid of the dendrogram. The dendrogram visually represents the hierarchy of clusters based on the similarity of principal component scores; it can be cut at a specific height to form a certain number of clusters. In this case, we chose to cut the tree at a height corresponding to 2 clusters: cluster 1 contains 260 observations while cluster 2 contains 43.

Given this relevant dissimilarity in observations we can assume that the data doesn't have a well-defined structure with distinct clusters. This assumption is further explored in k-means and fuzzy c-means analysis.

K-MEANS AND FUZZY C-MEANS

After the hierarchical clustering, we have decided to check if through the k-means analysis, it was possible to better identify the two clusters that are not really well visible from the dendrogram.

Unfortunately, it was not possible to achieve the desired goal since when we plotted the two clusters, we immediately saw how they were one over the other. We concluded that all the observations are just part of one big cluster.

To be even more sure, we have decided to apply the C-Mean with the fuzzy logic in order to understand the probability of belonging to one cluster or the other. Looking at the membership, we had the confirmation that the observations were part of just one single cluster. For each single observation, there was a probability of 50% belonging to one cluster and 50% of belonging to the other cluster.

REGRESSION

For our regression model it was chosen logistic regression instead of linear regression. Logistic regression is a method used for modeling the outcome of a binary dependent variable based on one or more predictor variables. It is appropriate here because it allows you to estimate the probability that a given input point belongs to one of the two categories, which is essential when the response variable is categorical and not continuous.

In our analysis, it was critical to convert the 'output' variable into a factor using the command heart$output <- as.factor(heart$output). This step was essential because our dependent variable is binary, representing two possible outcomes. Logistic regression is specifically designed for categorical outcomes and converting the variable into a factor ensures that R correctly interprets it as such. This conversion facilitates the application of logistic function modeling, allowing us to estimate the probability of occurrence of one outcome over the other. It also aids in the clear interpretation of results, as factors explicitly denote the categories within our dataset.

The first logistic regression model included a comprehensive set of predictors, aiming to capture all potential influences on the 'output' variable. Upon analyzing the model's output, we noted that not all predictors were statistically significant. Therefore, we sought to refine our model by selecting predictors that demonstrated a significant relationship with the dependent variable, thus streamlining the model and potentially enhancing its predictive accuracy.

Refinement and Analysis

In refining our model, we adopted a targeted approach, selecting variables based on their statistical significance from the initial regression results. This led to the construction of a second logistic regression model, logreg2, which included the variables sex, cp, thalachh, exng, oldpeak, caa, and thall. The output from logreg2 confirmed the relevance of the selected predictors. Notably, all included variables showed a significant relationship with the binary outcome at our chosen level of significance.

The normal Q-Q plot of standardized residuals from logreg2 did not align perfectly with the theoretical quantiles, indicating slight deviations from normality. However, in the context of logistic regression, the assumption of normality in residuals is not as stringent as it is for linear regression. The central goal is to model the probability of the binary outcome correctly, and logistic regression is robust to this type of deviation, particularly with large sample sizes.

We assessed multicollinearity through the Variance Inflation Factor (VIF) for each predictor in logreg2. VIF values close to 1 for all predictors suggest that multicollinearity is not a concern for our model. This is crucial as multicollinearity can obscure the individual contribution of predictors to the model and inflate standard errors, thereby undermining the reliability of the coefficient estimates.

The residual plots for logreg2 provided a visual diagnostic of the model fit for individual predictors. The lack of discernible patterns in the scatter plots for most variables indicates that the linear relationship assumption between the log odds of the predictors and the outcome is reasonable.

NAIVE BAYES CLASSIFIER

In order to predict the likelihood of heart disease based on the factors extracted from the regression we used a Naive Bayes Classifier. The classifier was trained on a dataset containing essential features related to heart health. To prepare the data for modeling, categorical variables were converted into factors. This step ensured that the classifier could effectively utilize the information encoded in these variables.

The Naive Bayes classifier was selected for its simplicity and efficiency in handling categorical data. The model was trained on a diverse set of features, including sex, chest pain type, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, number of major vessels colored by fluoroscopy, and thallium stress test result.

To assess the model's generalization ability, the dataset was divided into an 80% training set and a 20% testing set. This split allowed for evaluating the classifier on unseen data.

The performance of the Naive Bayes classifier was assessed using a confusion matrix. This matrix provided insights into the classifier's ability to correctly classify instances of heart disease and non-heart disease.

The accuracy of the Naive Bayes classifier on the test set was determined by calculating the proportion of correctly classified instances out of the total predictions. The accuracy obtained was above 85%.

The Naive Bayes classifier demonstrated promising performance in predicting heart disease based on the selected features. The accuracy of 85% suggests that the model has the potential to be a valuable tool for identifying individuals at risk of heart disease.

DECISION TREE

A decision tree is a useful interpretable model that allows us to explain how different areas in the input space correspond to different outcomes. It has been really useful to implement a decision tree in our analysis since it better described how certain attributes were affecting the target variable, especially from a visual point of view.

The logistic regression model highlighted which variables had a significant relationship with the binary outcome. These variables respectively are: sex, cp, thalachh, exng, oldpeak, caa, and thall. From this starting point, we have graphed the decision tree that has an overall accuracy of 80%. By plotting the decision tree, we saw how certain attributes, such as the sex, exng, oldpeak, were not represented indicating how these attributes were not important enough (a decision tree assigns importance to a certain feature based on its ability to split the data effectively).

The root of the tree is chest pain with value zero, so a chest pain due to angina. When the cp value is zero, we saw how 49% of observations have a low possibility of having a heart attack, while when cp is different from zero the percentage of having a heart attack is equal to 51%. In the 49% of observations, the attributes considered for the construction of the decision tree have been the caa (number of major vessels) and the thall (blood disorder called thalassemia), while in the remaining 51% the attributes considered are again the thall and the thalachh (maximum heart rate achieved).

CONCLUSIONS

In this heart disease prediction analysis, we examined a dataset with fourteen variables, encompassing demographic and medical indicators. Exploratory Data Analysis (EDA) unveiled insights such as gender imbalance, prevalent chest pain types, and associations with a higher likelihood of a heart attack. Correlation analysis highlighted significant positive correlations for chest pain type, maximum heart rate achieved, and slope of the peak exercise ST segment, while negative correlations were observed for exercise-induced angina, ST depression induced by exercise, number of major vessels colored, and male

gender. Principal Component Analysis (PCA) identified eight components explaining 79.5% of variance. Clustering methods (hierarchical, k-means, fuzzy c-means) suggested a lack of well-defined clusters in the data. Logistic regression identified significant predictors, and the resulting Naive Bayes classifier demonstrated promising accuracy above 85% in predicting heart disease based on selected features. A decision tree model provided interpretability and showcased attribute importance, revealing key factors influencing the likelihood of a heart attack.

Challenges

Throughout the data analysis process, several challenges were encountered that impacted the thoroughness and precision of our findings. One notable difficulty was the presence of imbalances in certain categorical variables, such as gender, which skewed the interpretation of results. The dataset exhibited a significant gender imbalance with 207 males and 96 females, potentially influencing patterns and correlations related to heart attack likelihood. Additionally, the dataset's hierarchical clustering revealed a lack of well-defined clusters, suggesting inherent complexities in the underlying structure of the data. This ambiguity hindered the effectiveness of hierarchical clustering and subsequent k-means and fuzzy c-means analyses, leading to difficulties in identifying distinct groups. In the regression analysis, a challenge emerged as the distribution of standardized residuals deviated from the normality assumption, as observed in the normal Q-Q plot. While logistic regression is known for its resilience to such deviations, recognizing these departures underscores the importance of interpreting results with proper judgment.