Title of Final Project: Heart Failure Predictions

Section: 52745

Group Number: 22

Student: Vaishnavi Sathiyamoorthy

UT EID: vs25229

Student: Medha Nalamada

UT EID: mrn789

Student: Alex Hohmann

UT EID: ajh5399

Student: Saivachan Ponnapolli

UT EID: sp48347

Date: 11/06/2023

Goal (or Thesis)

Cardiovascular disease is the leading cause of death around the world. If we are able to predict whether an individual will have cardiovascular disease, it can save numerous lives. This dataset consists of 11 factors that potentially predict heart failure: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, old peak, and the slope of peak exercise. Using statistics, visualization, and machine learning, this project aims to determine if these factors can accurately predict whether an individual will have cardiovascular disease.

Visualization Tools

Our design exploration will combine various tools to effectively visualize the relationship between our target variable and the predictors. We will use four primary visualization tools: Plotly, Matplotlib, Seaborn, and Altair. We decided to use a range of different tools to explore multiple interactive abilities, including sliders, hover, click selection, etc. These different interactions with varying color schemes according to the tools make the visualizations much more appealing and tell a more compelling narrative. It adds variety and explains our arguments concisely without being repetitive. With these different tools, we aim to present numerous visualizations that substantiate the arguments written below. We will then choose the most relevant variables after this exploration and use them to fit machine learning models to see how well we can predict heart disease.

Argument Layout

- Individuals who are older and have higher resting blood pressure have a higher risk of heart disease.
- Individuals who are older and have lower maximum heart rates have a higher risk of heart disease.
- Those with typical angina have a higher risk for heart disease.
- Those who are older and have a higher old peak slope have a higher risk for heart disease.
- Individuals who exhibit ExerciseAngina have a higher risk for heart disease across both sexes and all age groups.
- Individuals that have a Upwards ST_Slope have a lower Max HR, flat ST_Slope has a medium max HR, and down ST_Slope has the highest Max HR

Visual Layout

For each argument, the visual layout has a very intentional and explanatory meaning to help understand our dataset better, in order to make more accurate statistical predictions for the presence of heart disease.

Introductory Plots:

The purpose of the introductory plots is to present the most basic relationships and justify the arguments we have chosen. These include:

- Matplotlib grid of 6 categorical variables: This grid will show the distribution of heart disease within each categorical value for all our categorical variables

- Seaborn correlation heatmap: The heatmap will show the correlations of all our numerical variables

Argumentative Plots:

Following the introductory plots, we will proceed to investigate each argument individually and visually narrate the relationships we discover.

Argument 1: RestingBP, Age, and HeartDisease
Package: Altair
Approach: Two side-by-side stacked bar charts
Interactive feature: Selection on the left bar chart will change the distribution of the right bar chart.
Color Scheme: Stacked bar charts with two colors: blue for 0 (no heart disease); orange for 1 (heart disease is present)
This visualization is two combined stacked bar charts to show the distribution of resting blood pressure (right) for each age group (left). Upon selecting a bar for age on the left plot, the right plot will display a distribution of the resting blood pressure for that age group, with the bars being stacked to observe the presence of heart disease. The contrasting colors for the presence of heart disease makes the difference amongst distributions immediately visible. We will observe to see if those who are older tend to have heart disease and within each age group, those with higher blood pressure have heart disease.

Argument 2: MaxHR, Age, and HeartDisease
Package: Plotly
Approach: Scatter Plot
Interactive feature: Hover with details about MaxHR, Age, and HeartDisease for each point
Color Scheme: Viridis: purple for 0 (no heart disease); yellow for 1 (heart disease is present)

We will explore the relationship between Age and MaxHR to expand on the introductory correlation matrix; we want to visualize the correlation. Hovering over the data points will provide the exact maximum heart rate. We expect that the overall plot will show the negative correlation between Age and MaxHR, intuitively confirming that people’s heart rates will reduce as they get older. The scatter points will also show the concentration of colored points to imply wherever there is a higher concentration of yellow or purple dots. We chose different colors to add variety to the visualizations. We anticipate seeing that as age increases and MaxHR decreases, heart disease becomes more prevalent. This will confirm our argument that heart disease is more likely to affect older individuals with lower maximum heart rates.

Argument 3: Oldpeak, Age, and HeartDisease

Package: Altair

Approach: Scatter Plot & Bar Chart

Interactive feature: Selection range based on click window in scatter plot will alter the bar chart distributions

Color Scheme: Bar chart and scatter points follow blue for 0 (no heart disease) and orange for 1 (heart disease is present)

The plot will be a combined scatter and bar plot to explain the relationship between Oldpeak and HeartDisease with color indicating the presence of heart disease; we chose the same colors as our first bar chart and also anticipate using these colors for all following bar charts to provide consistency and solidify the distinction between having heart disease and not having heart disease for viewers. Upon selecting a range in the top scatter plot, we will see the count of people with heart disease in the bottom bar chart for a given range of Oldpeak values. We anticipate seeing that as we select values for Oldpeak higher than 0, the count of people with heart disease is increasing.

Argument 4: ChestPainType and HeartDisease

Package: Altair

Approach: Stacked Bar Chart

Interactive feature: Hover over bar chart to show exact count of patients with/without heart disease for each chest pain type

Color Scheme: Bar charts follow blue for 0 (no heart disease) and orange for 1 (heart disease is present)

We also want to see the variation of heart disease with different chest pain types with a simple stacked bar plot in Altair with color indicating the presence of heart disease. Most cases are asymptomatic, and hovering over the chart will reveal the exact count of people with or without heart disease. The results will also connect to the number of entries per chest pain type, but the separation of the stack will be interesting to visualize. For example, although there are fewest of ChestPainType TA, the split is almost 50/50, which is important for our analysis and can be used later to make compelling inferences.

Argument 5: ExerciseAngina, MaxHR and HeartDisease by Gender
Package: Altair
Approach: Scatter Plot
Interactive feature: Slider with ages under the charts; the slider can be dragged across to see the different distributions between sexes at each age for MaxHR with presence of Exercise-Induced angina and without Exercise-induced angina
Color Scheme: Scatter points follow blue for 0 (no heart disease) and orange for 1 (heart disease is present)
The variable ExerciseAngina is also of heavy interest. It indicates whether chest pain occurs during exercise specifically, and can thus also be compared to heart rate. Hence, this Altair plot explores if cardiovascular disease was exercise-based for this dataset and also separates for gender to see if there is any relevance. The plot will show a combined chart to compare between the presence of ExerciseAngina and the distribution of heart rate values for each age for males and females. Interactively, the slider will control the specific age and we will be able to see the differences in the scatter plot for each gender, with the points being color-coded orange for the presence of heart disease and blue for the absence. Our prediction with these visualizations are that exercise angina will associate with heart disease prevalence across heart rates. However, we also hypothesize that where there is no exercise induced angina, there will still be a prevalence of heart disease.

Argument 6: ST_Slope compared to MaxHR
Package: Plotly
Approach: Histogram
Interactive feature: Hover around the graph and get information such as the MaxHR for the section, the count of MaxHR in the range of that section, the color of the graph, and the slope accordingly

Color Scheme: Histograms follow blue for an upward ST_Slope, red for a flat ST_Slope, and green for a downward ST_Slope

We also wanted to test the impact of ST_Slope on the Max Heart Rate through a simple stacked histogram on Plotly, laden with different colors to indicate the various different slope sizes of Upward, Downward, and Flat, all put into the range of the Max HR. Our prediction is that an upward ST_Slope should lead to less of a MaxHR, but a downward ST_Slope would lead to more of a MaxHR rate because of the overall trends in HR. The colors are made in order to check the difference of the stack, so that a difference can be seen between the blue, red, and green graphs accordingly. The results will also connect to the number of overall ranges for MaxHR per ST_Slope, which is important for our analysis and will be used later for strong inferences.

Conclusion: Machine Learning

Package: sklearn

Approach: Random Forest, Logistic Regression, MLP Classification

We will use the most relevant variables based on the visualizations to predict heart disease. Three models will be trained using random forest, logistic regression, and MLP classification. A confusion matrix will be made to determine the number of true positives, false positives, true negatives, and false negatives. Precision and accuracy will be used as metrics for the models. The goal is to show that Age, Resting BP, Max HR, Old peak, Sex, ChestPainType, RestingECG, exercise angina, and ST slope can predict whether an individual has heart disease.