Heart Disease Analysis - Final Report

By Abigail Chutnik and Kcee Landon

Research Questions

1. Can we predict the presence of heart disease based on a patient's symptoms and laboratory results?
    a. We will use the most commonly known features of heart disease and create a model that would determine if a patient has the presence of heart disease based on their proclaimed symptoms and laboratory results obtained. This model may have significance in being utilized by physicians to obtain a diagnosis and offer treatment based on the results as heart disease usually is implicated based on multiple factors.

    Findings: We found that the 5 biggest indicators of heart disease were maximum observed heart rate, whether or not a person had exercise induced angina, ST depression induced by exercise relative to rest, the number of major vessels, and presence of a heart defect called thalassemia, which involves the formation of abnormal hemoglobin.

    Additionally, a DecisionTreeClassifier is able to predict the presence or absence of heart disease based on the 5 biggest indicators with about 98% training accuracy and 73% testing accuracy. We also found that exang, exercise induced angina, and thal, thalassemia, equally contribute the highest feature importance in our model.

2. How does blood pressure range across age groups and sex amongst healthy and diseased patients?
    a. We want to compute blood pressure by grouping healthy and diseased patients and seeing how their blood pressure differs by age in order to see if there are correlations between age, sex, and health to blood pressure.

    Findings: Based on our plots, diseased patients had a higher floor and ceiling for blood pressure levels and blood pressure increased similarly over time for healthy and diseased patients, with the exception of diseased females. Males show a slightly higher blood pressure in healthy patients, but in diseased, females are shown to have higher blood pressures.

3. Which heart disease symptom or attribute is most strongly correlated with the presence of angina in a patient, the most pain-inducing symptom. What about with the presence of exercise-induced angina? (Edited)
    a. By comparing the possible correlations between these symptoms and angina, we can more accurately predict which risk factor leads to a higher rate of resting and exercise-induced angina presence.

Findings: We backtracked and flipped the question to ask which attribute has the strongest correlation to the presence of both normal angina, and exercised induced angina (exang). We found that the attribute thalach (maximum heart rate achieved) had the strongest correlation to both normal resting angina and exercised induced angina.

Motivation and Background

Today, heart disease remains the leading cause of death globally, with approximately 9 million deaths per year reported by WHO. The CDC defines heart disease as several types of heart conditions that impact blood flow within the cardiovascular systems mostly caused by lifestyle factors and medical conditions. Our research questions wish to explore the risk factors for heart disease amongst a subset of the population in Cleveland, Ohio in order answer questions about the causal effects of heart disease.

For our primary research question, we plan to create a machine learning model that could be used to predict rapidly whether a patient has a presence of heart disease based on inputs of their laboratory results, which subsequently could be provided to physicians to aid them in deciding if further testing or treatment is needed. We won't be working directly with physicians to answer our questions but if we decided to send out our results to the real world, these results may be helpful for them. Furthermore, our second question examines how age and sex could influence blood pressure amongst healthy and diseased patients, which could be useful information for when deciding thresholds for high blood pressure. Lastly, our third research question investigates health disease attributes, and the most painful symptoms of heart disease, resting and exercise-induced angina. It is commonly known that these risk factors are correlated with the presence of angina, however we would like to test which one has a stronger correlation, which would provide a better insight for physicians on which risk factor to take into higher consideration when noticing angina in a patient.

Dataset(s)

Below is the Cleveland dataset for Heart Disease data. The Cleveland data set is the most widely used among other datasets for machine learning purposes. The full dataset is also linked below. In order to access the dataset we initially plan to extract the data file from the UCI website under "Download: Data Folder" and convert the cleveland.data file into a csv file using excel just by saving the data as a csv file.

Heart Disease Cleveland UCI | Kaggle
UCI Machine Learning Repository: Heart Disease Data Set

Challenge Goals

1. **Machine Learning**

    a. We believe our project will meet this goal of implementing a machine learning model because our project is based around making accurate predictions in order to answer our research questions. In order to create this model we will train and test the model with our dataset to be able to establish a machine learning tool that can predict if a patient has a presence of heart disease. We were able to find the best features that produced the highest accuracy for classification using new functions that we explored in the scikit-learn library.

2. **New Library**
    a. We ended up not using the Gleam library because that was more web-application based and switched to using plotly for data visualization. We still used matplotlib for some plots but plotly was a new library we dabbled with.

Methodology

For our first research question, we plan to create a machine learning model in order to determine if a patient has a presence of heart disease based on a subset of features taken from our dataset. We are planning to utilize about 5 of the impactful risk factors of heart disease as our features based on a Chi squared test using scikit-learn instead of 5 random factors to create a better model. The prediction that we wish to result from our model would be if the patient is a 0 (no presence of heart disease) or a 1 (presence of heart disease). The number of risk factors that we are including as our features may change based on the success of our model. Our research question will be answered if we obtain successful results from our testing and training analysis of accuracies. We will also establish a decision tree that will track what the program is deciding upon through each risk factor or feature we establish and a graph of the importances of each selected feature in order to analyze our model better to meet the challenge goal.

For our second research question, we are planning to create a visual or graphical representation of the blood pressure of healthy and diseased patients across all ages and with sex. The healthy patients should not have a presence of heart disease, indicated by a 0, and the diseased patients should have a presence of heart disease, indicated by a 1. These groups will be plotted separately by age vs blood pressure and separated by if they have heart disease or not. Since we wish to solely highlight our question's results on our visual representation, we will attempt to use the Plotly library in order to create our graphs and try out its interactive features. Our research question will be answered based on the differences between the results of age vs blood pressure of both groups, healthy and diseased and across sexes.

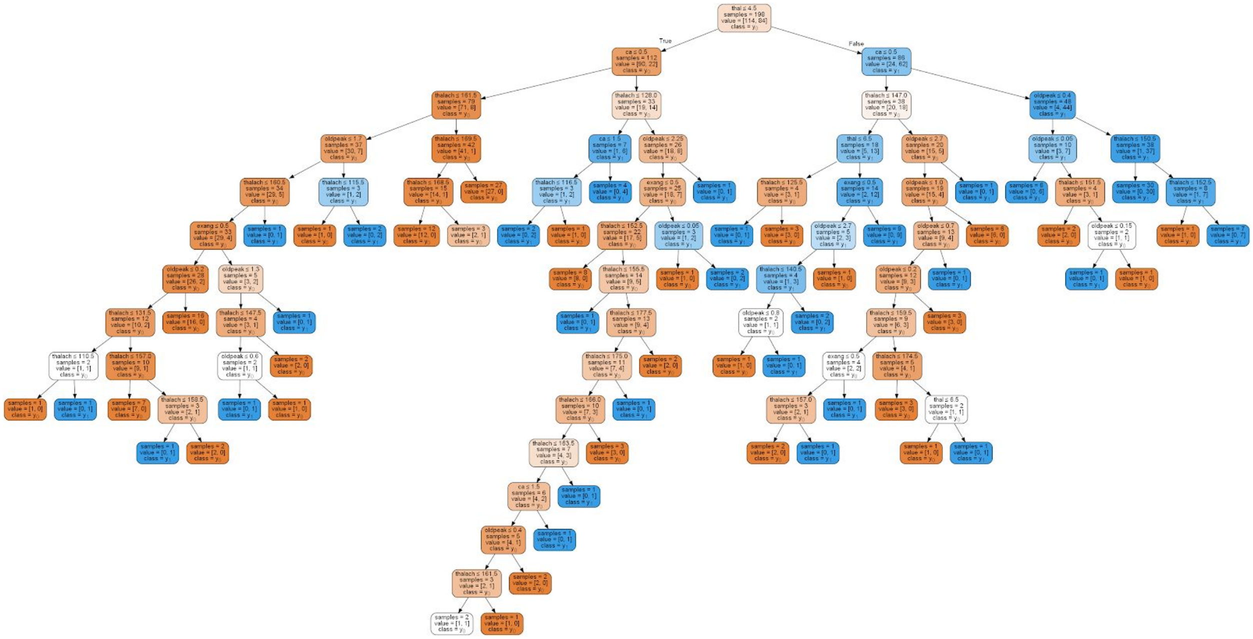
For our third research question, we are planning to perform a statistical analysis for all heart disease attributes with the presence of angina and with the presence of exercise-induced angina to work with the new library a bit more. These statistical results will be graphed either with Plotly as it was the updated library we wanted to use, since Gleam could not work with these results. For a strong correlation, we are specifically looking for a high correlation coefficient of close to positive 1 for both risk factors with angina, however the closest one to +1 will answer our research question.

The data filtering that we did for all three questions generally required us to change some of the values so that they were boolean values. For example, the ‘num’ column in the dataframe determined whether or not a patient had presence of heart disease but the values in the original dataset ranged from 0 to 4. We learned that this just signified the severity of disease. So to make the dataframe fit a classification model, we changed every value that was not 0 to 1 which converted the column to boolean values where 0 meant no heart disease presence and 1 meant there was indeed heart disease presence. Most of our data filtering followed this process of changing the values from ints to booleans or ints to strings, etc.

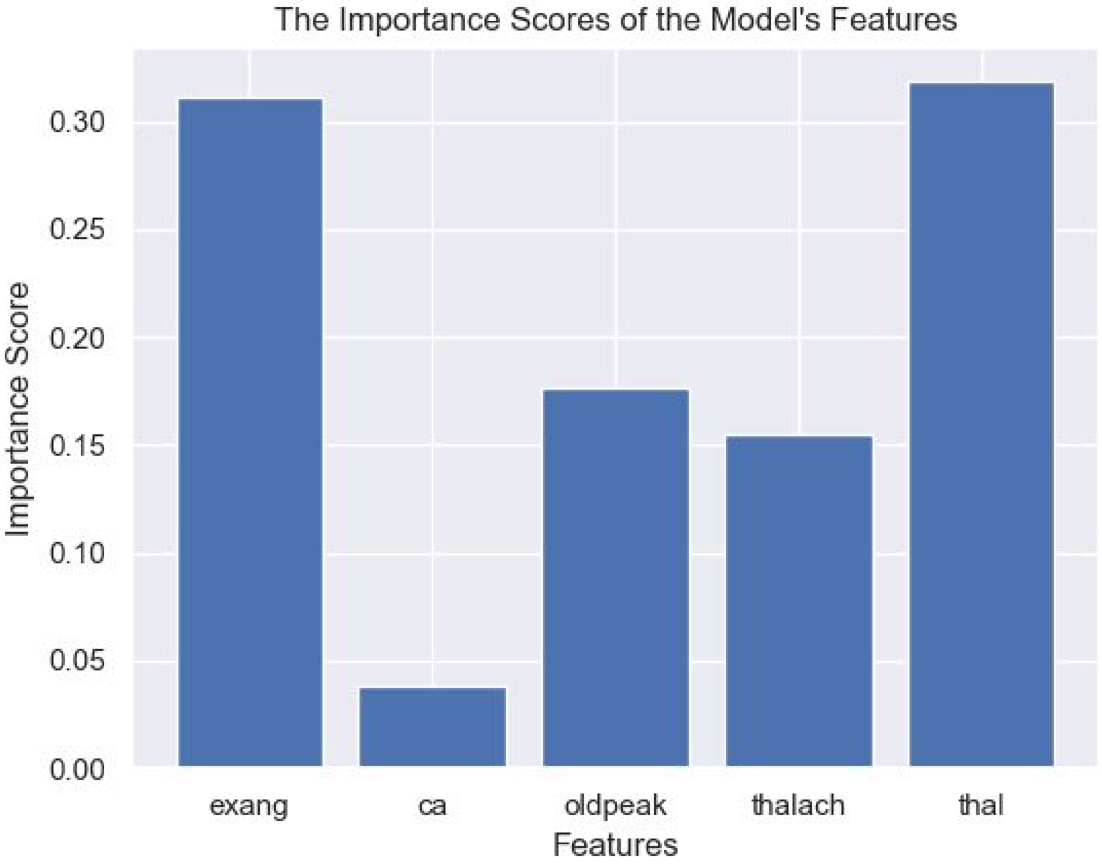
Results:

Research Question 1:

We explored a new feature of the scikit-learn library in order to select the k-best features or symptoms exhibited by a patient that are most strongly related to the presence of heart disease. We used a k value of 5 along with the chi2 function to test whether two variables are related in order to identify the 5 most important features and they are as follows: exang, ca, oldpeak, thalach, and thal. Once we found these features, we used scikit-learn’s feature_importances_ property of the DecisionTreeClassifier model to identify each attribute’s importance score which demonstrates how important it is in predicting the presence of heart disease. Our DecisionTree visualization can be seen below:



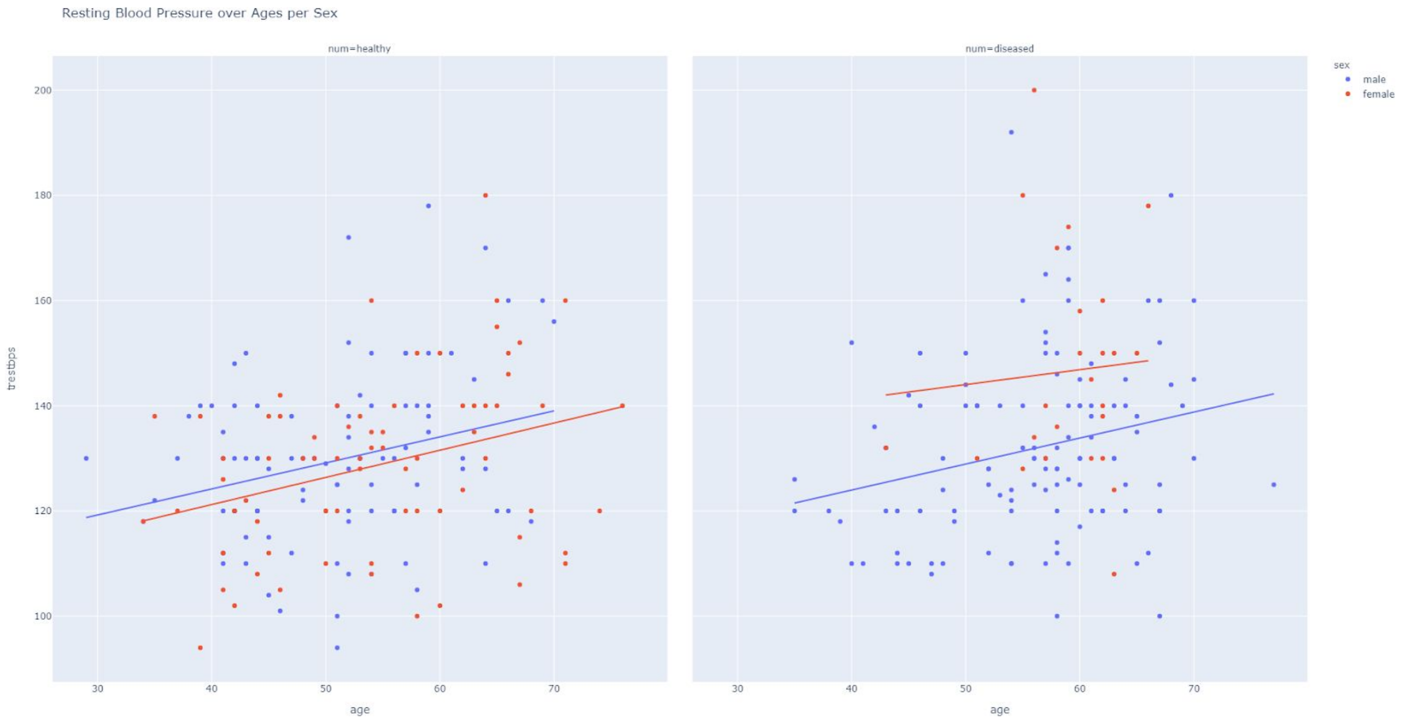
We then plotted the importance scores of each feature in the bar graph shown below:



As seen from our graph, both exang and thal are almost tied for the most important identifiers in predicting the presence of heart disease. This makes sense since exercise contributes to a higher heart rate which in turn can cause a heart attack if a patient is already present with heart disease. Normally, exercise is good for the heart but if someone already has heart disease, they must monitor their activity properly and control their heart rate otherwise cardiac arrest or heart attacks may occur for some individuals. Additionally, thal or thalassemia was a bit unexpected as it is a genetically associated disorder that is not directly associated with health disease. There may be a bias that exists in the dataset to lead to this as many of patients may happen to have a mutation in particular hemoglobin-making genes. This makes sense as it is known that thalassemia occurs most frequently in African American populations and Cleveland, Ohio has a very large African American demographic, about half its population. All in all, our model is was a good predictor of heart disease in the Cleveland, Ohio area as our accuracy score was about 73%.

Research Question 2:
To answer our second research question, we plotted the resting blood pressure across all age ranges in our dataset and separated the two graphs depending on presence of heart disease and sex. We used Plotly to plot our scatter linear regression plot in order to test out a more interactive plotting library.

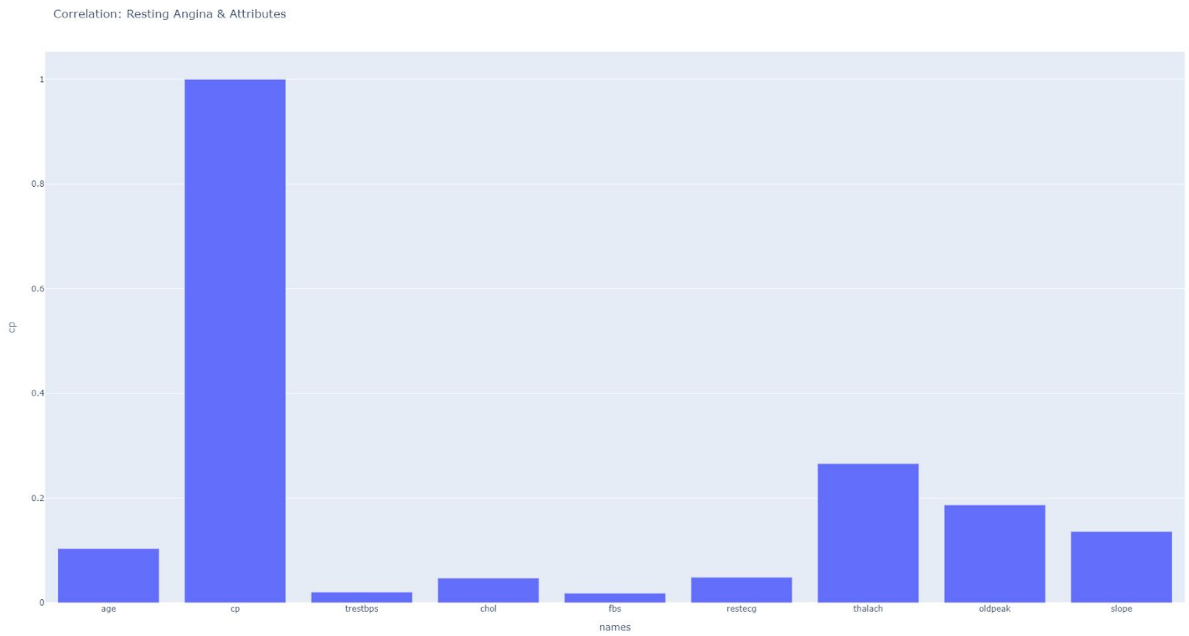
Our results are as follows:



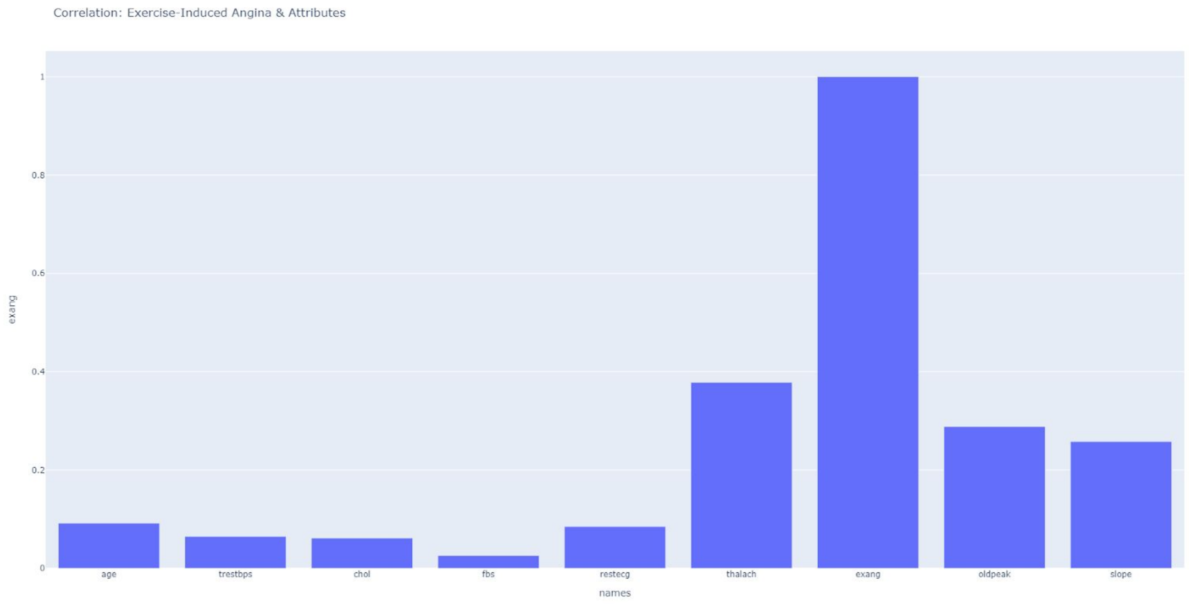
Analyzing this graph, we can see that among diseased patients, women have higher blood pressure compared to men but among healthy patients, men generally have a higher blood pressure across all age ranges. This can be explained due to the fact that there were much less samples of diseased women, which means that for diseased patients, we cannot conclude anything on sex and blood pressure. Also, comparing healthy and diseased patients overall, we can see that both the floor and ceiling for resting blood pressure is slightly higher in diseased patients which signifies that resting blood pressure is generally higher for diseased patients. Furthermore, the trend lines show that blood pressure increases across ages in both healthy and diseased patients. Our trend lines can somewhat predict the average resting blood pressure for a given sex and age for healthy patients. Overall, the trends aren't too highly correlated and it seems as though the data is spread apart quite a bit although we are certain that diseased patients generally have a higher resting blood pressure and older ages, healthy and diseased, have higher blood pressures.

Research Question 3:

For our final research question, we decided to alter our original question and find out which factors have the highest correlation to both exercise induced angina and normal resting angina. The original dataset had resting angina presence split into different groups so we manipulated the column to have boolean values dependent on if there was a presence of angina. No further data processing was needed for exang or exercise induced angina. When plotting each absolute value of the correlation coefficient on the bar graph, we excluded exang for resting angina's bar graph and vice versa. Additionally, we used Plotly in order to try out a more interactive library. For resting angina, our results were as follows:



And for exang, our results are below:



Since cp and exang are the same attributes for their respective graphs, they have a correlation value of 1.0 and so we use this as our control variable just to show that the highest correlation coefficient value is possible. As we can see in both graphs, thalach has the highest correlation value for both resting and exercise induced angina which is similar to what we discovered from our first research question. A higher heart rate (thalach) makes it so that the blood is pumping faster and thus your heart is working harder to get blood pumping throughout your body. Since angina is defined as any pain caused by the narrowing of the arteries which prevents blood from being supplied to the heart muscles, it makes sense that a patient will exhibit angina more often since the heart is pumping blood at a higher rate but is getting partially blocked. In short, everytime blood gets pumped by the heart, angina pain will be experienced and with a higher heart rate or thalach, a patient will experience more pain.

Work Plan:

Our work plan has been divided into five parts:

1. **Cleaning and Filtering Data (approx. 1 hr)**
2. **Creating the ML Model (approx. 1 hr)**
3. **Data Visualizations (approx. 3-4 hrs)**
4. **Debugging (approx. 3-4 hrs)**
5. **Analysis (approx. 2-3 hrs)**

We plan on using version control through a Github repository which is linked below:
https://github.com/klandon2-1736208/heart-disease-project

We plan on splitting up the coding by each working on one research question each, yet the individual working on question 2 will collaborate a bit with the individual working on question 1. We will do the debugging and analysis together in a Zoom call.

Work Plan Evaluation:

We overestimated a lot of the times it took for some of the parts. The data filtering and data visualizations took considerably less time than expected but creating the model was pretty on point with our estimation. The total overall time of debugging might have been close to two hours. Most time was spent on looking into functions for all of the libraries we used and figuring out how to use the code. Our analysis is expected to take about an hour or two.

Testing:

In order to test our code, we mainly used print statements to verify that our pandas dataframes were producing the correct data. We also produced a test file “main_test.py” to additionally verify that our dataframe manipulation for each of our research questions were correct.

Collaboration:

Outside of our group, we did not consult anyone. Unfortunately, Ashina was dealing with personal issues so she was unable to work on the project with us. For outside resources, we used Google to look up documentation for all the code that we used to develop with. Some of the links we used can be found below.

Reproducing the Results:

1. Download the dataset from our repository: https://github.com/klandon2-1736208/heart-disease-project
    a. For more information on column descriptions: open heart-disease(2).txt
    b. Cleveland.csv is the dataset you will be using
2. Install all packages
    a. Graphviz, scikit-learn, ipython, pandas, matplotlib, plotly, numpy
3. Run main.py and get the results in the repository downloaded folder
    a. Data_filtering.py will alter the dataset for each question to be utilized in main.py
    b. Tree.gv.png will save Q1's model
    c. Importance.png - bar graph of feature importance values that will be saved
    d. Q1's accuracy scores and iPython display of the tree is the expected output of our model
    e. Q2 and Q3 plotly graphs will be opened automatically by your web browser
4. Run main_test.py to test if the data filtered properly

Workspace and Links:

1.13. Feature selection — scikit-learn 0.24.1 documentation (scikit-learn.org)
https://www.kaggle.com/jepsds/feature-selection-using-selectkbest?utm_campaign=News&utm_medium=Community&utm_source=DataCamp.com
https://graphviz.org/download/

- Download the 64-bit version for windows and install

On the command line:
"pip install ipython"
"pip install graphviz"

Correlation using Python (thecleverprogrammer.com)
https://git-scm.com/downloads

pip install pandas

```
pip install -U scikit-learn
https://www.statsmodels.org/stable/index.html
```