Predicting Heart Failure with Various Biological Markers

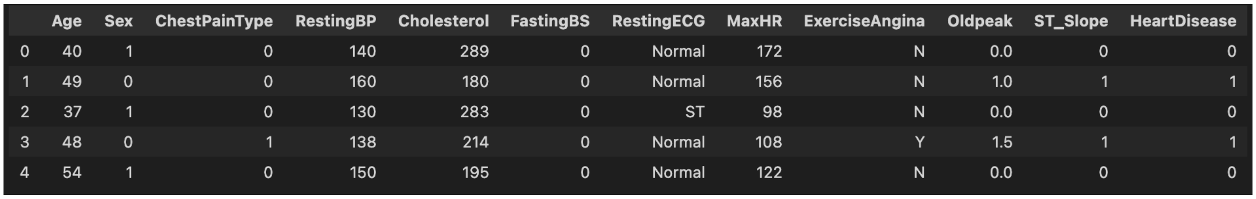
By: Justin Callahan, Pedro Arrizon, Jennice Herrera, Doeun Lee, and Abhinav Mugunda

Introduction:

Heart failure is a progressive condition in which the heart is unable to pump enough blood to meet the body's oxygen demand. Cardiovascular disease causes 17.9 million deaths per year, making it the leading cause of death globally. Additionally, heart disease cost the United States $363 billion in 2016 due to health care services, medicine, and lost productivity. The motivation behind this project is to improve the accuracy of heart diagnoses. Statistical models can bring a lot of value to the field of health care, and cardiology specifically. Their use could allow health professionals to more accurately select which health factors to prioritize when diagnosing patients. By understanding the impact that various biological markers have on an individual's likelihood of developing heart failure, preventative measures can be more effective.

Data:

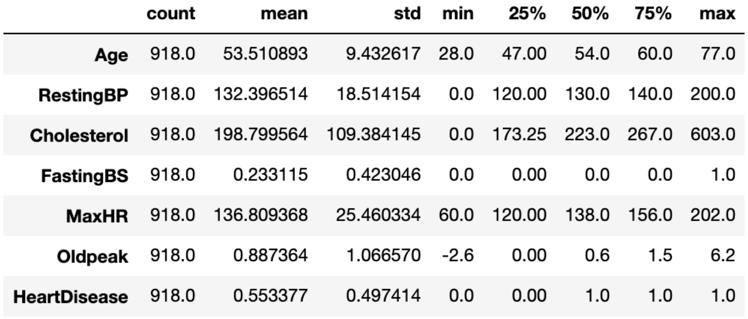
The dataset our group used is one that was downloaded from Kaggle. It contains real data documenting factors related to heart health compiled from datasets put out by hospitals in Virginia, Cleveland, Hungary, and Switzerland. There were originally 1,190 observations in the dataset, but 272 duplicates were removed, leaving 918 observations in the final data we analyzed. The data needed several transformations in order to make it easier to analyze. Sex was turned into a binary variable, where 1 is male and 0 is female. ChestPainType was turned into a binary variable, where asymptomatic is 1, and the three types of symptomatic pain were 0. ST_Slope was turned into a binary variable where 1 is flat, and 0 is sloped. Apart from turning some categorical variables into binary variables, the data did not need to be cleaned, as the data set provided on Kaggle had already been cleaned.



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	1	0	140	289	0	Normal	172	N	0.0	0	0
1	49	0	0	160	180	0	Normal	156	N	1.0	1	1
2	37	1	0	130	283	0	ST	98	N	0.0	0	0
3	48	0	1	138	214	0	Normal	108	Y	1.5	1	1
4	54	1	0	150	195	0	Normal	122	N	0.0	0	0

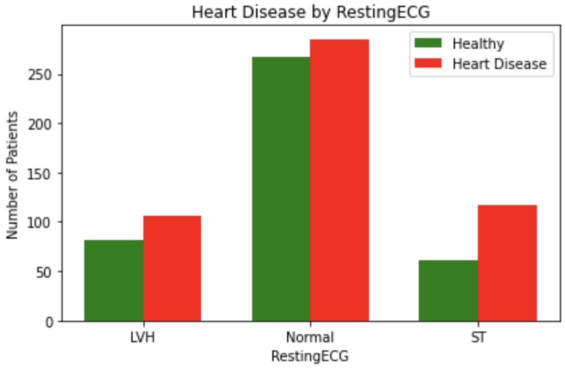
The figure above shows the first five rows of data after the transformations. The first eleven variables are the predictor variables, while the last is the target variable. Most of the variables are fairly self explanatory, however FastingBS stands for fasting blood sugar, and it is a binary variable where it is 1 if over 120 mg/dl.

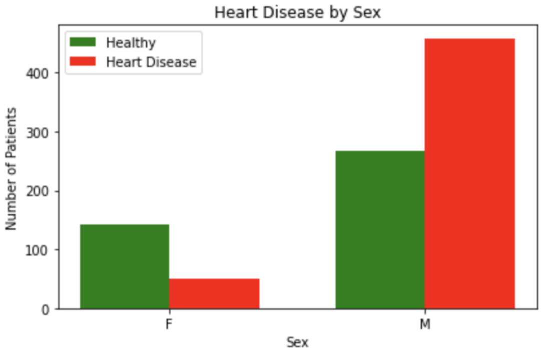
Exploratory Analysis:

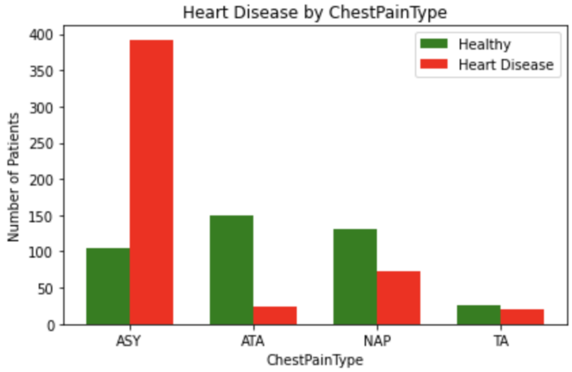


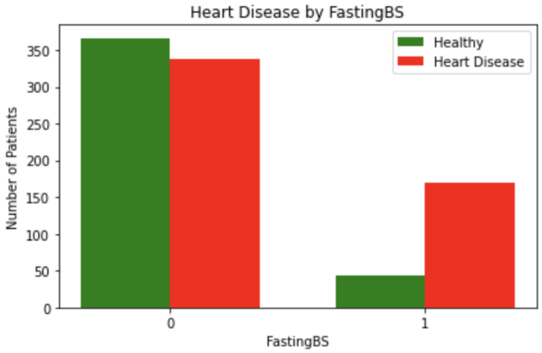
	count	mean	std	min	25%	50%	75%	max
Age	918.0	53.510893	9.432617	28.0	47.00	54.0	60.0	77.0
RestingBP	918.0	132.396514	18.514154	0.0	120.00	130.0	140.0	200.0
Cholesterol	918.0	198.799564	109.384145	0.0	173.25	223.0	267.0	603.0
FastingBS	918.0	0.233115	0.423046	0.0	0.00	0.0	0.0	1.0
MaxHR	918.0	136.809368	25.460334	60.0	120.00	138.0	156.0	202.0
Oldpeak	918.0	0.887364	1.066570	-2.6	0.00	0.6	1.5	6.2
HeartDisease	918.0	0.553377	0.497414	0.0	0.00	1.0	1.0	1.0

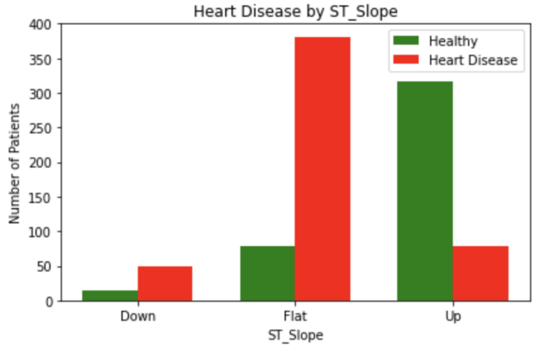
The mean age in this data is 53.5, as opposed to 29.6 in the world or 38.5 in the United States. The average resting blood pressure is 132.4, as opposed to ~125 for the average 53 year old. It is also important to note that 55% of people in this dataset developed heart disease. Overall, the people in this study are older and have more health risk factors than the general populace.

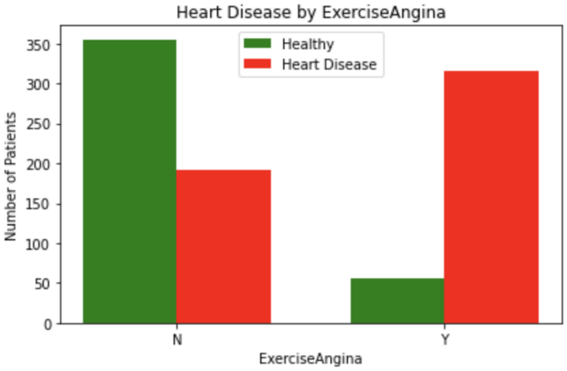




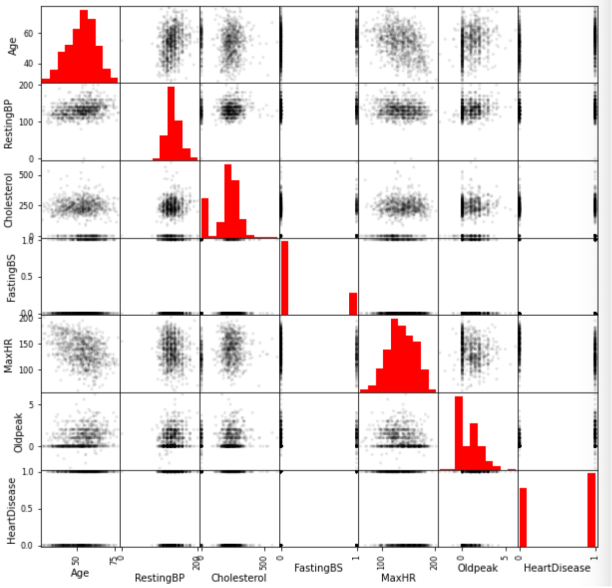








These charts show some of the main trends and possible predictors of heart disease in the dataset. The effect of asymptomatic pain against the symptomatic ChestPainType categories can be seen above. ST_Slope also has a large difference in Flat against an upward sloped line. Another interesting variable we noted was whether the patient experienced Exercise Angina. While there were a number of patients with heart disease who didn't experience it, an overwhelming majority of people who did experience angina ended up having some kind of heart disease. Looking at these differences improved our ability to transform the variables accurately in order to make a cleaner model.



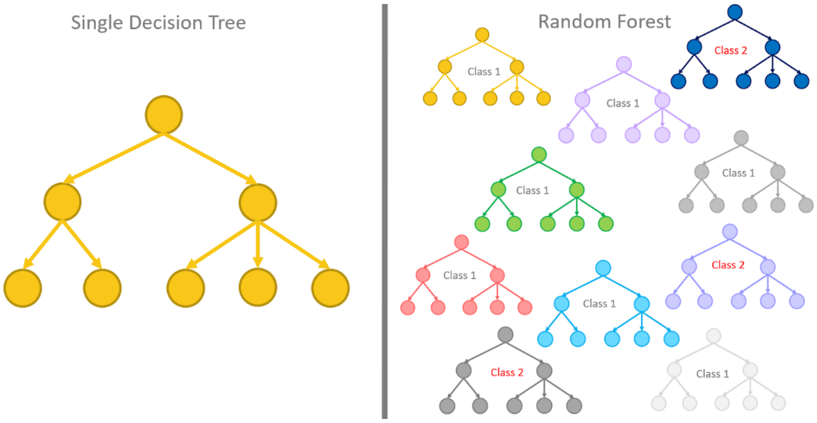
Inspecting the data before plugging it into the model is one of the most important parts of data science. This allows us to see possible trends, relationships, or any clear errors that may exist in the data. For example, we can see how variables such as Age and MaxHR are normally distributed, but the Oldpeak variable is skewed right, which needs to be taken into account.

Hypotheses:

Based off of the charts above, we predict that a sloped ST segment will have high influence in the patient being classified into the heart problem bin. We also predict that high resting blood pressure will have a strong predictive effect on heart problems.

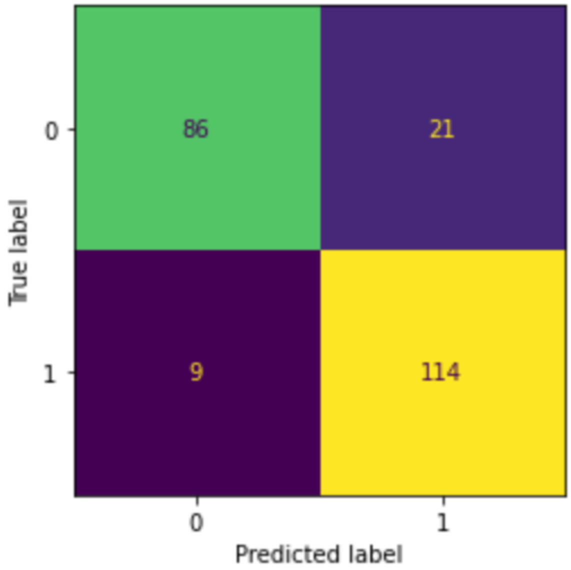
Modeling:

Our problem was a classification problem. The task was to assign the patients into one of two categories, developing heart problems or not developing heart problems, based on the variables given.

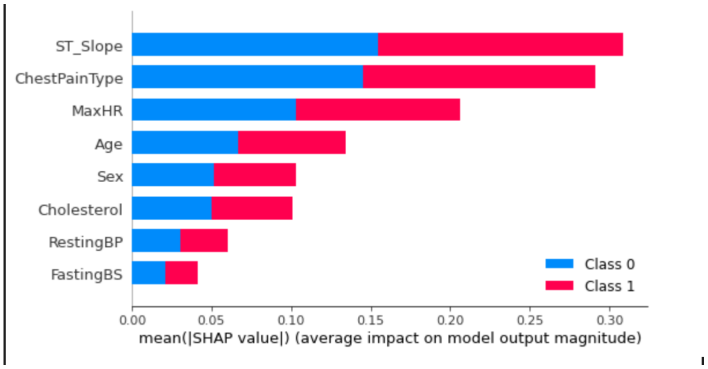


In order to pursue the classification goal, our group decided on using a random forest model, which is a collection of decision tree models. Decision trees work by splitting the predictor feature space into sections. At each split, the model makes a decision in order to make the best model classes. This can lead to problems because these models heavily lean towards only certain features and neglect others. To combat this, we used a random forest. This model creates multiple overgrown trees and uses bootstrapping to help create a low variance model. We chose the random forest model over typical decision trees because regular decision trees are trained to take the best split to maximize the entropy gain(greedy split). A random forest uses a random set of features at each split to decorrelate the features with each other allowing us to account for all predictor features. For the parameters for our model we set up the model to take only 3 predictor features out of our subset of 8 predictor features for each tree. We decided to not prune any of our bootstrapped trees to allow us to limit the amount of bias we introduce into the model, which we further explain in limitations.

Discussion:



Our test data was 230 out of the 918 observations. Of these, there were 200 correct predictions, 21 type 1 errors, and 9 type 2 errors. The accuracy score of .8696 is greater than .7, so this test seems to be successful. On the matrix above, 0 represents negative and 1 represents positive.

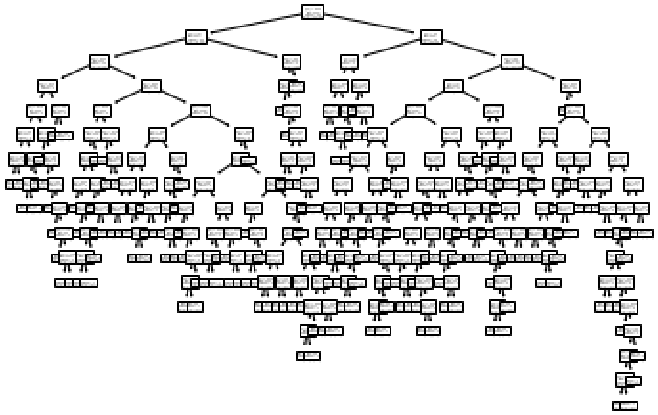


Interpreting from the shap value plot, ST_slope and chest pain type were the most important predictors of heart failure, however these two variables were the ones that we manually changed into binary variables. Max Heart Rate and Age were also fairly significant predictors.

Limitations:

For the analysis, we needed to assume independence and randomness of the data collected even though those conditions were not met. Firstly, the data was not randomly chosen. The data had to be collected at the patient's consent from the hospitals. This introduces selection bias, as certain patients would be more likely to share their information than others. Secondly, the data was not independent as it was collected from hospitals. People from hospitals often suffer from other complications or are sick in general, so the data may not be representative of the general population. One example of this can be seen in there being a much higher proportion of males in the dataset than in the general population. Another issue with translating the data from the survey to the general population is that the data was only collected from select countries in Europe and select states in the United States. This must be noted before attempting to extrapolate to the rest of the world, as each country has differing foods, traditions, and lifestyles. This lends itself to different health levels in each country, including the frequency of heart problems.

When building any model it's important to take into consideration how much of your own bias you are adding into the model and how it affects it. Changing multiple categorical variables into binary values introduces a lot of bias into the data in order to improve the accuracy of the model. Our lack of medical expertise worked against us here, as we could not confidently assert that these changes were non-material. Because of the bias that was introduced, there is lots of model variance on other potential data sets. In order to progress to future studies, it would be important to bring in someone who is knowledgeable in the field, and could provide insight to help manipulate the predictor features more accurately.



Another possible limitation is that the random forest model is hard to understand up front, and requires interpretation before delivering the results to other groups. It is important to understand the purpose which the model will serve. If the model is too hard for doctors or hospital personnel to understand, they may make life-altering decisions based on incomplete or misleading data.

Conclusion:

It is important to be knowledgeable in the relevant subject when making modeling decisions. Knowing what each variable means and how they interact is key to making usable data. Due to our lack of experience in the field, we chose a model that took choosing factors out of our hand, the random forest model. However, while analyzing the results, it must be noted that relying on only one method is not the most reliable method. The model exists in order to give us information to make decisions with, not to make the decisions for us. Our main finding is that the slope of the ST segment and the type of chest pain are the most important predictors of heart problems, out of the variables given. Further studies of these variables would be necessary in order to determine if our treatment of them was correct, and to what degree they actually affect the populace.

Acknowledgement:

Justin Callahan - Created and presented the data slides on the presentation, compiled the information from the presentation into the essay, and formatted it. Created website.

Pedro Arrizon - Created statistical method models, interpreted and discussed models within group, guided presentation and Final Report.

Jennice Herrera - Helped discuss medical background and terminology, helped guide feature selection within the models using domain experience.

Doeun Lee - Interpreted the resulting graphs, created and presented the result and limitations slides on the presentation, and fixed the final report.

Abhinav Mugunda - Proof read the final report, provided input into project decisions, injected levity into proceedings

Bibliography:

Divya Jacob, Pharm. D. “What Is the Normal Blood Pressure Range? Chart, Low, Normal & High.” *MedicineNet*, MedicineNet, 4 Jan. 2021, https://www.medicinenet.com/blood_pressure_chart_reading_by_age/article.htm.

Fedesoriano. “Heart Failure Prediction Dataset.” *Kaggle*, 10 Sep. 2021, https://www.kaggle.com/fedesoriano/heart-failure-prediction.

"Heart Disease Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 27 Sep. 2021, https://www.cdc.gov/heartdisease/facts.htm.

Silipo, Rosaria. "From a Single Decision Tree to a Random Forest." *Medium*, Towards Data Science, 8 Oct. 2019, https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147.