Heart Disease Prediction Report

Austin Dobbins

DSC 680 Applied Data Science

Project 1

**Introduction and Background**

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. About 647,000 Americans die from heart disease each year. This calculates to about 1 in every 4 deaths being caused by heart disease. [1] Many Americans that are at risk for heart disease are not actually aware they are at risk for this disease. Although, those that come from ethnic backgrounds are more at risk to contract this disease, the impact is felt in every community in the United States. To protect Americans from this disease, we must find ways to predict whether a person will contract it. This is the main goal of this project. In this project, we will also attempt to find the most important factors in making this prediction. I believe that the success of this project will lead to a better understanding of heart disease and eventually, will lead us to becoming healthier people.

**Methods**

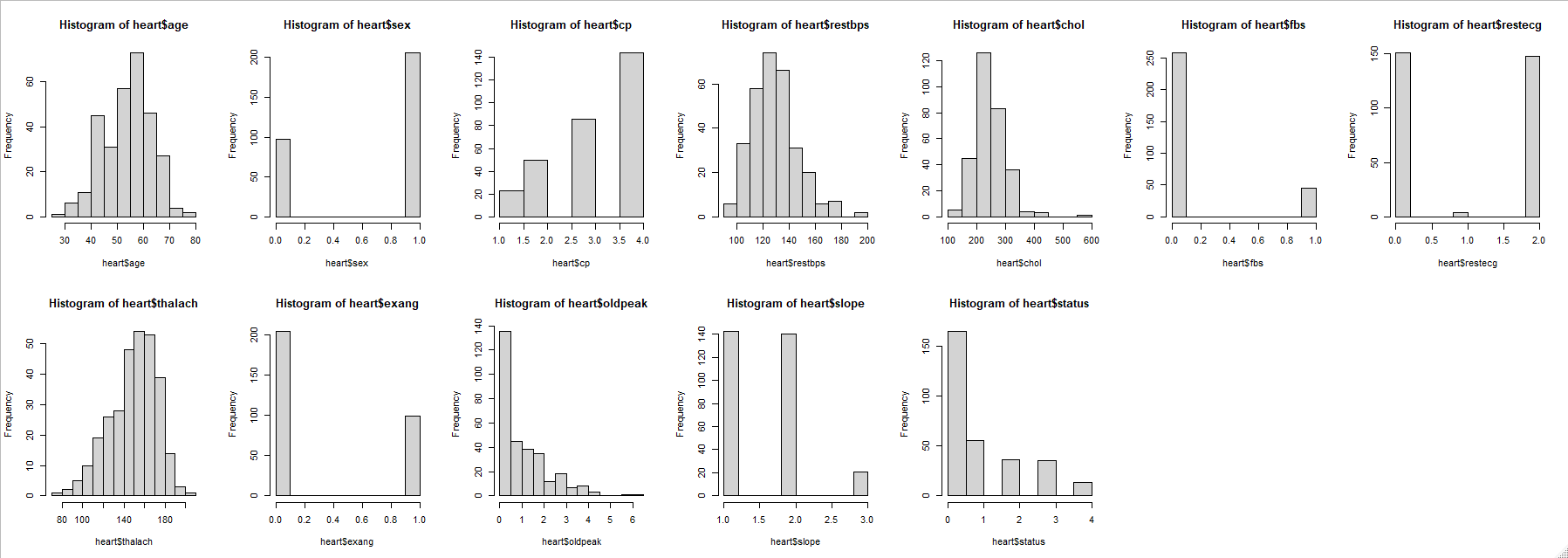
To accomplish this goal, a combination of regression analysis and classification analysis was used. The regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. This provides us with the ability to verify if there is one factor that indicates a heart disease diagnosis more than another. The classification analysis will be used to determine what type of person is more at risk for heart disease than another based on their health factors. Having information on what factors contribute the most to a heart disease diagnosis and what type of person is more at risk for heart disease will provide us with a plethora of information for early detection and eventually early treatment for the disease. Most of this project was completed using RStudio, however, Python was used for some of the data cleaning and exploratory data analysis.

The dataset used for this project comes from the UCI Repository. The Cleveland Heart Disease dataset is very popular among machine learning researchers. It originally contains 76 patient attributes, however, the dataset that is used for most ML projects is a subset of 14 of these attributes.

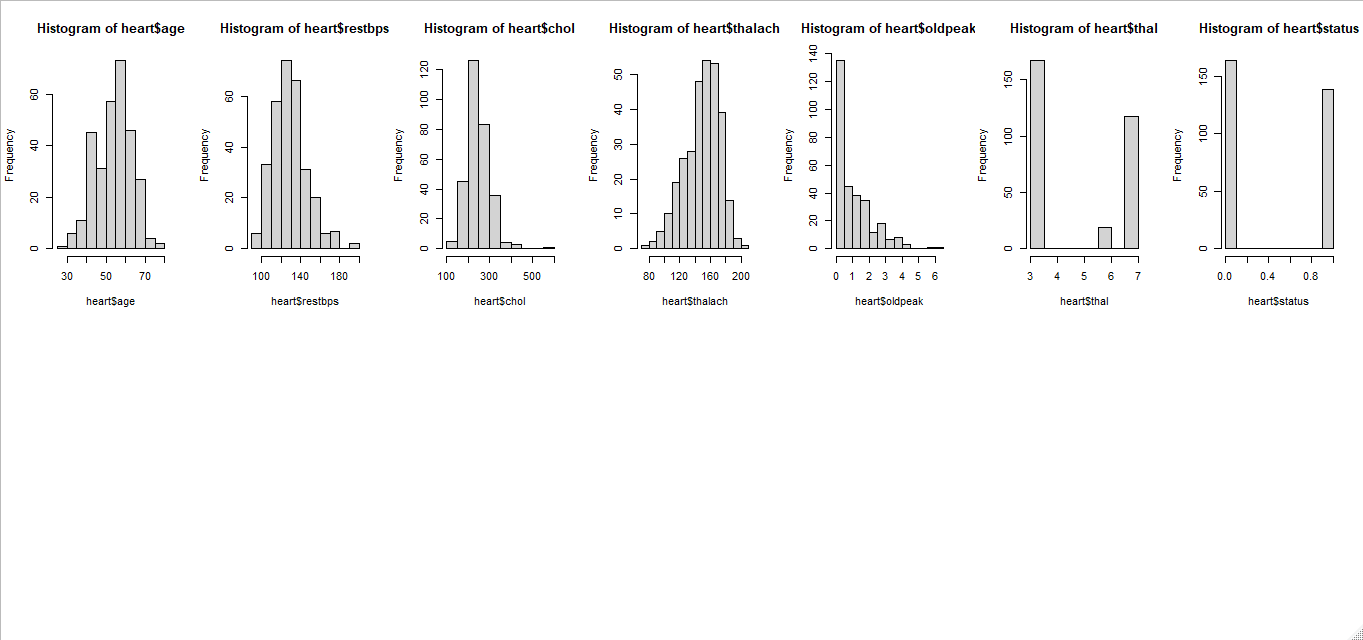
**Data Preparation/Exploratory Data Analysis**

The first step in this project was cleaning and transforming the data. Fortunately, the dataset did not require much cleaning. The first transformation that needed to be done was to change the names of a couple of the attributes to better describe what they are. An example of this is the now “status” attribute was originally called “num”. The status attribute indicates whether there is a heart disease diagnosis for a patient. Zero indicates that a patient does not have heart disease, while 1-4 is a diagnosis of heart disease and its severity, with 4 being the highest. Although this was a good measure for this data set, it does not work for the goal of this project which is simply predicting a heart disease diagnosis and not its severity. The status attribute was changed to have a binomial distribution. This means a 1 would indicate a patient has heart disease and a 0 would indicate the opposite. All patients that had 2-4 as their status were switch to a 1.

The next step of the cleaning and transformation process was to establish the properties of each attribute for model building purposes. Before this was done, all attributes were considered numeric, even though there were a few attributes that are categorical. To alleviate this issue, every attribute that should be numeric was transformed to an integer and attributes that should be categorical were transformed to factors. In the two figures below, you can see the data before and after the changes as well as the distributions of each attribute.



*Fig. 1 Histogram of all Attributes (Before Cleaning)*



*Fig. 2 Histogram of all Attributes (After Cleaning)*

The distribution of each attribute was as expected based on the nature of the dataset. This also contributed to the way outliers were handled in this project. Because we are looking at health attributes and we are using them to predict a disease, outliers will not be thrown out.

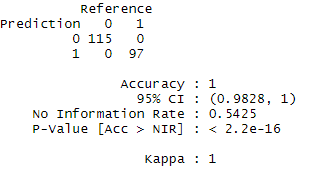
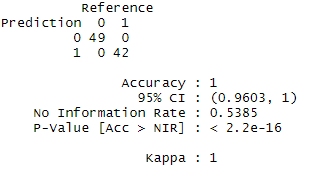
**Model Selection**

Based on the preliminary results of the exploratory data analysis, the models originally selected for the project were Linear Regression, Naïve Bayes Classifier, Random Forest Classifier, and K-Nearest Neighbors. However, this required some changes after more of the exploratory data analysis was completed and the initial models were attempted. Although in theory linear regression would have worked well with this data set because of the correlation coefficients between 12 of the 13 attributes and the target variable, the linear regression model did not produce any useable results.

Due to this fact, the final models selected for this project were the Random Forest Classifier, Naïve Bayes Classifier, and the Support Vector Machine (SVM) Classifier. Naïve Bayes is a probabilistic machine learning model that is used for classification task. The crux of the classifier is based on the Bayes theorem. Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one feature does not affect the other. Hence it is called naive. The Random Forest Classifier like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. • Support Vector Machine abbreviated as SVM can be used for both regression and classification tasks. But it is widely used in classification objectives. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points where N = the number of features.

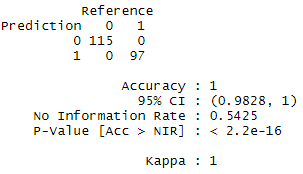
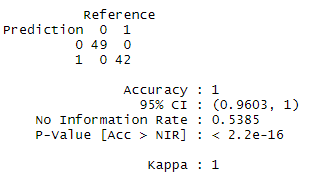
**Model Results and Selection**

I could not have been more pleased with the results of each model created. Each model performed well beyond all expectations. The Random Forest Classification Model was the first to be created and evaluated. The data was split 70%:30% for the training data and the validation data. The model was able to predict heart disease in the training set and validation set with 100% accuracy.

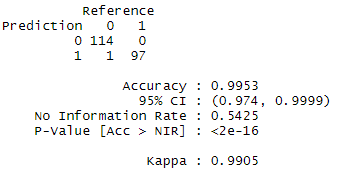
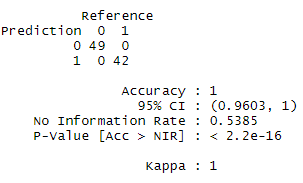
*Fig. 3 Random Forest Classifier Results, Training Left, Validation Right*

The SVM Classification model was also created using a 70%:30% split for the training data and validation data. This model also performed beyond expectation and was able to predict heart disease in the training set and validation set with 100% accuracy.

*Fig. 4 Support Vector Machine Results, Training Left, Validation Right*

The last model created was the Naïve Bayes algorithm. This model did not perform with 100% accuracy in training like the over two models. However, it still performed well above expectations. Like the other two models, it was created using a 70%:30% split for the training data and validation data. It was able to predict the training set with 99.5% accuracy and the validation set with 100% accuracy.

*Fig. 5 Naïve Bayes Results, Training Left, Validation Right*

As previously stated, all models performed well above expectation. They were all easy to implement and provided amazing results. This makes it a bit difficult to select just one model for this project. However, the Random Forest Classification Model is the model that will be selected for this project. Personally, when I choose a model for a project, I am looking for a model that will be able to stay accurate even is data is changed or added. Because of this fact, I believe the Random Forest Classifier is the best choice for this project.

**Conclusion**

Based on the overall results, I would say this is a successful project. We were able to accomplish our main goal of predicting a heart disease diagnosis with impressive results from each model. I believe these results of this project are important and can provide information for early detection and prevention for heart disease. However, I believe that this can only be considered a starting point. I feel that this project would need to be performed on the entire Cleveland Heart Disease dataset before it can be considered significant. Also, only medical history is considered in this dataset. Based on my research, lifestyle habits such as eating habits, exercise habits, and consumption of tobacco products also influence the development of heart disease. To have the most accurate results possible, a dataset containing lifestyle information as well as complete medical history would need to be used.

**References:**

[1]<https://www.cdc.gov/heartdisease/facts.htm#:~:text=Heart%20disease%20is%20the%20leading,1%20in%20every%204%20deaths.>

[2] <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>

[3] <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

[4] <https://www.osplabs.com/insights/the-map-of-healthcare-analytics-domains/>