Python Project Final Report – CDA 501
Hillary Gesel, Shengqi Ye, Pronata Datta, Shun An Chang

Predicting Heart Disease Diagnosis with Several Health Parameters

Abstract:
Heart disease is the leading cause of death for both men and women and people of most racial and ethnic groups in the United States (1). With extensive research, several significant risk factors have been showed to predict and diagnose heart disease which include diet, inactivity, smoking, hypertension, and gender (2). Our group used a Heart Disease Data Set from Kaggle that was a combination of datasets from around the world to predict heart disease based on the predictors in the dataset. Using KNN, Logistic Regression, Support Vectors, and decision trees, we were able to find how accurate different analysis methods were to predict the heart disease classification. Using the machine learning methods, were we able to see different accuracies and the significance of the different predictors for predicting heat disease.
Conclusion:

Introduction:
Cardiovascular Disease (CVD) or heart disease is the leading cause of morbidity and mortality in the United States, according to the Centers for Disease Control and Prevention (1). Many families, including some of our own families have been affected by heart disease. We wanted to evaluate and analyze several pathological parameters that are commonly used to diagnose heart disease. Heart disease is a condition that is developed over time and leads to a significant amount of medical expenses and long term
Using the database found in Kaggle (3), we performed analysis on 11 parameters that can be used to predict a possible heart disease. Using several analysis methods, we were able to see the impact and accuracy of our analytical models and how well one can predict heart disease with the information provided.

Data:
The data was created by combining different datasets already available independently but not combined before, according to the description from Kaggle. In this dataset, 5 heart datasets are combined over 11 common features. The parameters include age, sex, chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic), resting Systolic blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise angina, old peak, ST slope with exercise and the last variable was the response of heart disease. The patient's observations included patients with know heart disease and those who are not diagnosed with heart disease. Our goal is to find patterns and significance with the different predictors to conclude if given values for the predictors will result in a classification of heart disease or healthy.

Models:
For our machine learning models, we used KNN, logistic regression, support vector machines and decision trees for classification. To do these calculations, we uploaded the csv file and wanted to separate the columns into normalized tables (4). We created normalized tables for gender, ExerciseAnina, ChestPainType, RestingECG, ST_Slope. From here, we were able to perform our analysis. After normalization, we added created a full dataset with the normalized values.

Results:
For initial observations of the dataset, we plotted several box plots, histograms with density, and pairs plots to find obvious relationships between variables. Some of the strongest, positive relationships between the variables and the response of heart disease include the presence of an Old

peak, Age, fastingBS. Negative relationships with a diagnosis of heart disease include cholesterol and maxHR, according to this dataset. The pairs plot in cell 21 is color coded by heart disease diagnosis. No diagnosis of heart disease is 0 and heart disease diagnosis is 1. There are several observations that can be seen from this plot including the majority cluster of heart disease patients have a lower maxHR when compared with RestingBP. Moving to our analytical models, we used KNN, Logistic Regression, Decision Trees and Support Vector machines to predict how accurate each model would for predicting a diagnosis of heart disease based on the eleven predictors. From here, the testing and training sets were determined and used for all our models. Based on our analysis of logistic regression, the accuracy of determining the diagnosis of heart disease was 87.16%. For decision trees, the max leaves were 10 with a max depth of 5. For our decision tree, ST_Slope was the main root node, which was followed by chest pain type and cholesterol. Based on the decision tree, the model accuracy for predicting heart disease was 82.3%. KNN was run for 21 values of k. The accuracies are listed for 1-21 respectively: 81.81%, 81.81%, 85.02%, 83.95%, 87.16%, 86.01%, 87.70%, 87.70%, 88.77%, 87.70%, 88.77%, 86.63%, 88.77%, 88.23%, 88.77%, 87.70%, 89.30%, 89.30%, 89.83%, 89.30%, 89.30%.Support vector machines used a linear kernel and resulted in an 82.35% accuracy. The models we created used all variables to predict. As mentioned above and with the benefit of observing the decision tree graph and the scatterplots, bar graphs and strip plots we were able to see the more significant variables including, ST_Slope type UP, Atypical Angina, (ATA), Male sex, the presence of an old Peak and presence of exercise angina. Lastly, we created an interactive plotly graph that shows the distribution of patients diagnosed with heart disease based on age, MaxHR and their ST_Slope. Here, we see that most patients diagnosed with heart disease have a ST_Slope classified as "Flat or Down", while most patients not diagnosed with heart disease have up-sloping ST_Slope. ST segment depression (horizontal or down sloping) is the most reliable indicator of exercise-induced ischemia (5).

Conclusion:

As one of the leading causes of morbidity and mortality in the United States, heart disease affects many Americans and their families. The medical community continues to research and prevent heart disease, however, the need for continued, longitudinal studies are needed to further investigate the causes and predictions of heart disease. The data set used for this project used mostly objective results as a predictor and one subjective predictor, chest pain type. Based on the predictors used and the models we created, our methods for predicting heart disease were quite accurate. The most accurate model was KNN at a k value of 19. This is consistent with the general use of KNN as a good predictor for classification. In comparison to KNN, support vector machines, logistic regression and decision trees were next in accuracy respectively. Interestingly KNN for all k values was more accurate than the decision tree. Future analysis including family history, environmental factors, including smoking, secondhand smoke exposure, diet and alcoholism should be included.

References:

1. Centers for Disease Control and Prevention. "Heart Disease Facts & Statistics." Centers for Disease Control and Prevention, 27 Sept. 2021, www.cdc.gov/heartdisease/facts.htm. Accessed 15 Dec. 2021.
2. World Health Organization. "Cardiovascular Diseases (CVDs)." Who.int, World Health Organization: WHO, 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 15 Dec. 2021.
3. fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved 12/15/2021 from https://www.kaggle.com/fedesoriano/heart-failure-prediction.
4. "Python - How to Count Unique Values in Pandas Column Base on Dictionary Values." Stack Overflow, stackoverflow.com/questions/70369170/how-to-count-unique-values-in-pandas-column-base-on-dictionary-values. Accessed 15 Dec. 2021.
5. Hill J, Timmis A. Exercise tolerance testing. *BMJ*. 2002;324(7345):1084-1087. doi:10.1136/bmj.324.7345.1084