Theodore Koby-Hercsky

March 24, 2022

DSC630-T301 Predictive Analytics (2225-1)

Dr. Brett Werner

**Course Project: Milestone 2 – Data Selection and Project Proposal**

The Data set I have selected for my project is an Excel CSV file that is on heart failure that is broken up into twelve variables. Which consist of the patients age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG (electrocardiogram), maximum heart rate, exercise induced angina, old peak, ST slope, and heart disease. I would like to use this data set to predict the likelihood of patients experiencing heart failure with the variables given. As I have a congenital heart condition and have had multiple open-heart surgeries including a pacemaker which makes this topic furthermore important to me.

Regarding my project I am proposing that I will be able to run and evaluate multiple models and may even include other data sets if applicable throughout the course of this class. Which I am proposing that the findings will show that some values in variables from the data set will show a strong correlation between heart failure while others will show a perfectly healthy patient. As for example some may say a male is more likely to suffer from heart failure compared to a female, but we will let the data decide and determine the true probability. But this data set will take our research further as some variables such as cholesterol and Chest pain types that include TA for typical angina, ATA for atypical angina, NAP for non-anginal pain, and last ASY for asymptomatic. A patient that is having non-anginal pain and has high cholesterol can be proposed to have a higher likelihood of having heart failure compared to an individual with low cholesterol and no pain.

**What types of model or models do you plan to use and why?**

A model that I plan on using is the Gradient Boosted Model as this model will create a prediction model that is made up of decision trees that are created one tree at a time. This model uses techniques that are considered machine learning that is for the most part boosted. Once a decision tree is created it will fix errors made from previous trees. Another model I plan on using is the Random Forest Classifier as this is one of the most common algorithms that is used. While also being a classifier that can handle large sets of data that is comprised of an assortment of decision trees. As these trees are dependent on the value random vector sample and are created to the largest degree achievable. When using Random Forest for prediction the classifier will use random subsets of the training data which is known as bagging that can have multiple samples to generate an overall average. The reasoning for using Random Forest is to generate a solid learner compared to the Gradient Boosted Model as these will be two great models to compare.

**How do you plan to evaluate your results?**

The results from the Gradient Boosted Model will be evaluated by creating a plot that shows the variables with the highest value for predicting heart failure. As for the most part I will most likely use the training set data when creating the Gradient Boosted Model to evaluate the variables and the connection, they display with heart failure. I will evaluate the Random Forest Classifier by reviewing the accuracy, area under the curve, and the F1-score to determine the performance of the model regarding the different metrics.

**What do you hope to learn?**

When creating this report, I hope to learn new techniques and different models and or classifiers. While also learning which variables have the highest likelihood of predicting heart failure for the patients when using the Gradient Boosted Model. As for the Random Forest Model I hope to learn that the accuracy along with the F1-score to be high with a low standard deviation. On another note, I would also like to expand my knowledge on creating visualizations from my model outputs.

**Assess any risks with your proposal.**

A risk that can be associated with the proposal of this project can be that the change in values within a variable could have no positive or negative effect on a patient’s chance for heart failure. While one might say another risk with this proposal could be that the prediction of a variable could be incorrect meaning that values, we thought had an effect could be completely opposite of what was thought. Another risk I might face could be that the data I chose might not be compatible with other data sets I might find and want to use as the class goes on. If that happens then I will need to determine what other data sets I could use and if I would even need additional data

**Identify a contingency plan if your original project plan does not work out.**

In contingency if my original project plan does not work out, I would like to create a Logistic Regression as this also is a very popular regression that is used in classification. The use of this regression can estimate the mean and median of the data which is great as it can attempt to prevent overfitting. On another note, the use of a Logistic Regression does have a disadvantage as some may say that the results can be complex and misunderstood to some viewers. While I would also like to create a Decision Tree in my contingency plan as this is a straightforward algorithm that is easy to interpret. As a Decision Tree will produce an accuracy score along with a F1 score that can be used to compare against the results from the Logistic regression. While also the use of a Decision Tree can help when creating a Random Forest if no issues arise in my original plan.

**Reference**

PARTHASARATHY, SRIRAM. “Top 5 Predictive Analytics Models and Algorithms: Logi Analytics Blog.” *Logi Analytics*, 12 Feb. 2021, https://www.logianalytics.com/predictive-analytics/predictive-algorithms-and-models/.