**Purpose of Heart Disease Case Study Assignment:**

As a level three assignment, this case study’s primary focus is on the initial “Model”, “Understand” portions of the Data Science Analytics Process (see below). This will allow students to grow in their abilities to gain understanding from what the models they create describe.



Can be expected to be a good assignment in either DASC 2113 Data Principles and Techniques.

The goal of this case study is to determine if there are indicators present in the data that can help determine if one has heart disease. The case study is designed in three separate sections that all lead to help answer that question. The case study assignments are divided with specific skills and learning objectives in mind.

**Case Study Assignment 1**: The first assignment allows for the students to deep dive in exploratory data analysis. Using pandas, seaborn, and matplotlib to describe how the variables interact with one another and interact with having heart disease or not. A major graphing technique used in this assignment is creating bar plots with the np.arrange function to group the variable values. This helps the student recognize that individual whose age falls between 60 and 70 has a higher amount of heart disease individuals in their group than the 30 to 40 age group. At the end of the assignment the student will assess those variables that are common with individuals with heart disease is that they are mostly male, older age, lower heart rate, and they have a distinct old peak value.

**Case Study Assignment 2**: The second assignment tests the statistical significance of each finding made within the first assignment, by utilizing 2-sample Z-Test and 2-Sample T-test.

**Case Study Assignment 3**: The final assignment in the trio introduces machine learning concepts to predict if the indicators can determine if an individual has or will obtain heart disease in the future. Package utilized in this assignment are manly from sklearn and branches of that package. The first predictive measure the student will do is a decision tree analysis, train the model to best fit the data, and evaluate the accuracy of the tree by accuracy scores, classification report, and confusion matrix. The students will also complete a binary logistic regression. This is a good model for the data as the regression requires the dependent variable to be a Boolean. In this dataset, the dependent variable is the heart disease dataset with the Boolean 0 or 1. The student will create the binary logistic regression model, an odds table, and accuracy score of the model.

These assignments would be easily implementable in the course schedule near the end of the semester before finals begin. This would be a good time for it as in incorporates many skills learned throughout the semester into one assignment.

Also, because the curriculum of DASC 2113 does not include predictive measures, there are a few ways that these assignments can be implemented:

1. Assigning all three of the assignments
2. Assigning the first two assignments and the predictive assignment as extra credit assignment
3. If this class has an honors section, assign the first two assignments to all students, then assign the predictive assignment to the honors student.