Design Document

### **Goals:**

**Long Term Overall-** To successfully construct a model to accurately predict the age of the Abalone based on their physical characteristics.

**Intermediate Goals:**

* To create a model that produces an accuracy score of above 80%.
* To minimize and avoid possible group attribution bias.
* To complete the project by Jul 25, 2021

### **Data Acquisition and Preparation**

#### **Part 1**

1. Download data set
2. Turn dataset into data frame.
3. Exploratory data analysis(repair data, find missing values, etc.)
4. Define target column(s) as the number of rings (which determine age which we want to predict).
5. Define feature column(s) based on rest of the columns
6. Find non-numerical columns, determine whether to one-hot encode or remove.
7. Train/test split (70% train, 30% test)
8. Create a Gaussian Naïve Bayes classifier model.
9. Fit model to training data.
10. Get predictions for test set target values.
11. Get accuracy using predictions for testing data.

#### **Part 2**

1. Start with state of data frame after part 1 step 3.
2. Remove features and/or change targets for other models.
3. Repeat steps (part 1) 6 through 11.

### 

### **Project Roles**

Kenneth : Program Leader

Schedule to start: Throughout the CapStone

Isaiah : Lead Code Designer

Schedule to start: Throughout Code Development Process

Erika : Note Taker

Schedule to start: Throughout the CapStone

Adam : Auditor

Schedule to start: Once the code has been completed

Bobbi : Presentation Designer

Schedule to start: Once we have all our findings from implementation of the model/code

### **Problem Space**

We want to find out how we can accurately predict the age of abalone based on their physical traits. Traditionally, their age is determined by cutting and staining the shell. Then, counting the rings on the shell. The challenge we need to face is finding what other characteristics can we use to determine the age of the abalone. Not only do we want to find another method to determine the age of abalone, but we want to build a more efficient model that’s just as accurate, if not more than previous methods.

**Motivated Questions**

* What physical traits in abalone change over time?
* What other existential factors will produce bias and how can we mitigate them?
* What model best suits the data provided?
* How can our findings improve future research?