Fundamentals of Data Science

Project Proposal 2

Tanner Patrom

For my second proposal I would like to analyze the quality of a banana. The dataset can be found at <https://www.kaggle.com/datasets/l3llff/banana>. This would be a binary classification, observing the output as being either good or bad based on the following metrics within the data set: the size, the weight, sweetness, softness, ripeness, and acidity of the fruit, along with the amount of time passed from the harvesting of the fruit. This would be a supervised learning model as I am using labeled data to train and make decisions based on prior knowledge. I could use three different models to try and predict the quality. The first would be linear regression. I think it would be affective as the data seems to have a linear relationship between the information collected and the outcome of the banana. The second would be decision tree. I think that this might notice if the information collected is not necessarily linear to the quality of the banana, it could capture more complex interactions. The third would be a random forest. Being a more complex supervised model it may be able to be more accurate and handle extreme values better than the other models.