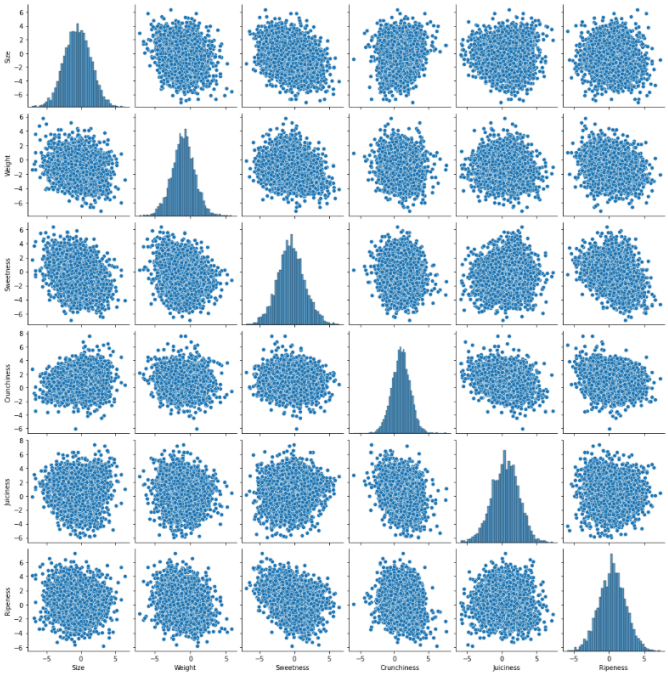
Introduction

My dataset is from Kaggle and has data on just over 4000 different apples. It has information on the size, weight, ripeness, sweetness, crunchiness, juiciness, ripeness, acidity, and a binary quality variable(good or bad). My objective is to determine which characteristics of apples determine the quality of an apple. Other than quality, something I'm aiming to predict, each variable is continuous and is normalized to a mean of 0. Because the variables are already normalized, there is no information on what exactly each variable is measured by, other than the fact that it is continuous. The biggest problem with this dataset is that we don't know anything about how the variables are actually conceptualized.

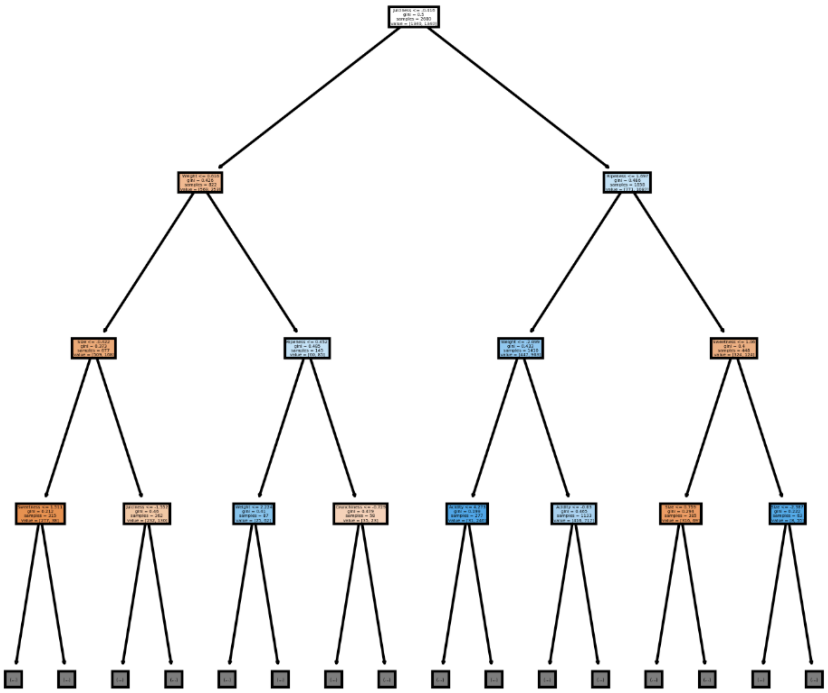


Preprocessing

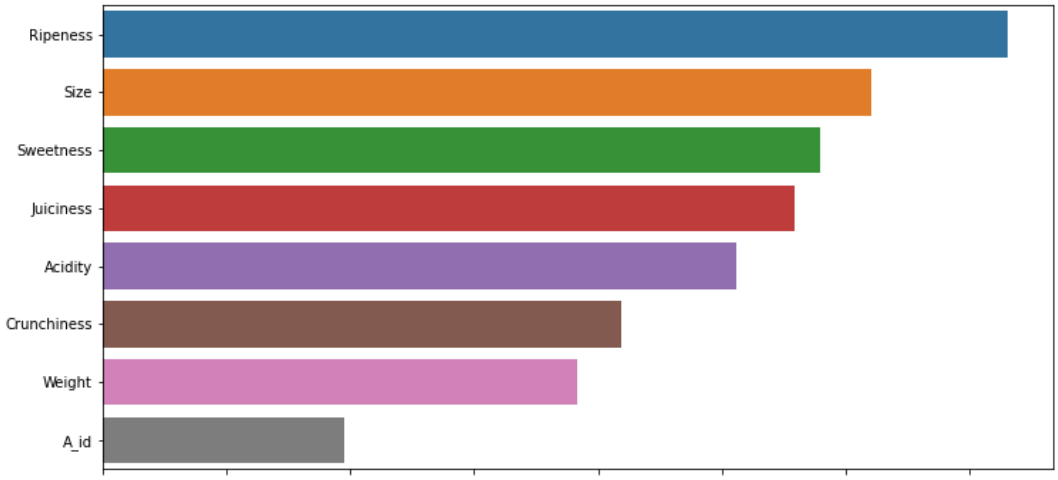
Because the data was already normalized and there were no null values, there wasn't much preprocessing that could have been done. The only cleaning that needed to happen was dropping the last line of data because it was stating who put the dataset together but was formatted with a lot of NaNs that were stopping the models from running. The pairplot to the right shows relationships between each of the variables. I made this to see if there were any variables that have a strong enough relationship that would require me to drop one of them. While some do have weak correlations, none stick out as anything significant.

Modeling and Evaluation

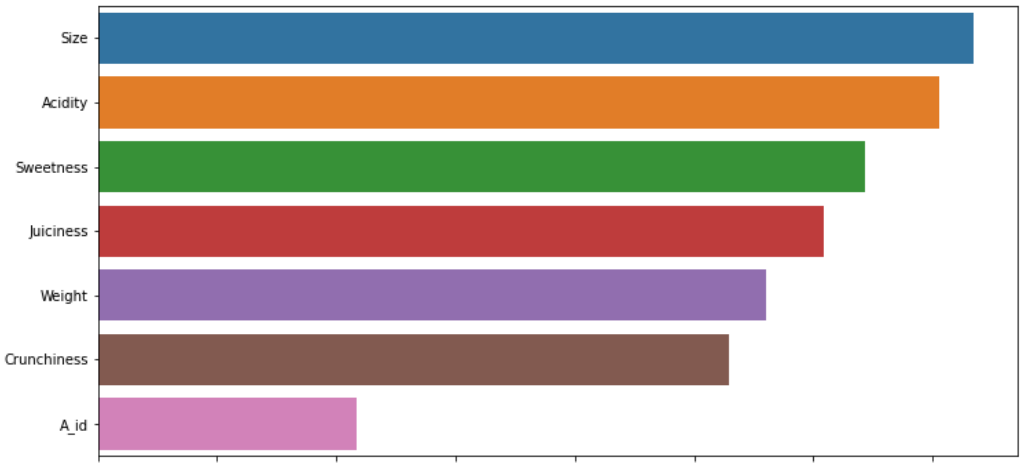
I used a decision tree to classify the apples as either good or bad. I used a decision tree because the dataset I was using had a lot of variables and the importance of each individual variable would be easy to see. Decision trees classify each object by



attempting to find the best place in each variable to split the data in order to have the cleanest, or purest, split in the data. Their depth is determined by how many variables are used before each object is classified.



After running the model a few times with a test/train split of .33/.67, I think my model performed pretty well. The clf scores were around .800 for the most part, and the cross validation scores were just under .800. Precision, recall, and f-1 scores were all around .80 when the model was run. The two most important features of apples when determining quality were ripeness and size. Because the biggest gap between variables was between the top 2, I ran the model again, but without the most important ripeness, the most important variable. We see that the dropoff from one variable to the next is relatively constant. The clf and cross validation scores dropped to around .750, which is expected when we drop the most important variable. The precision, recall, and f-1 scores all dropped to around .70, a bit bigger of a drop.



Conclusion

Through this project, we learned that ripeness was the most important factor in determining apple quality. Assuming this data is real and accurate, it could be used to help apple farmers determine which apples they bring to markets or send to stores. There was no mention of whether or not this data was subjective or objective, knowing if there was human bias involved in the ratings could help us think about not only the sellers, but also the consumers.