Wine Class Id

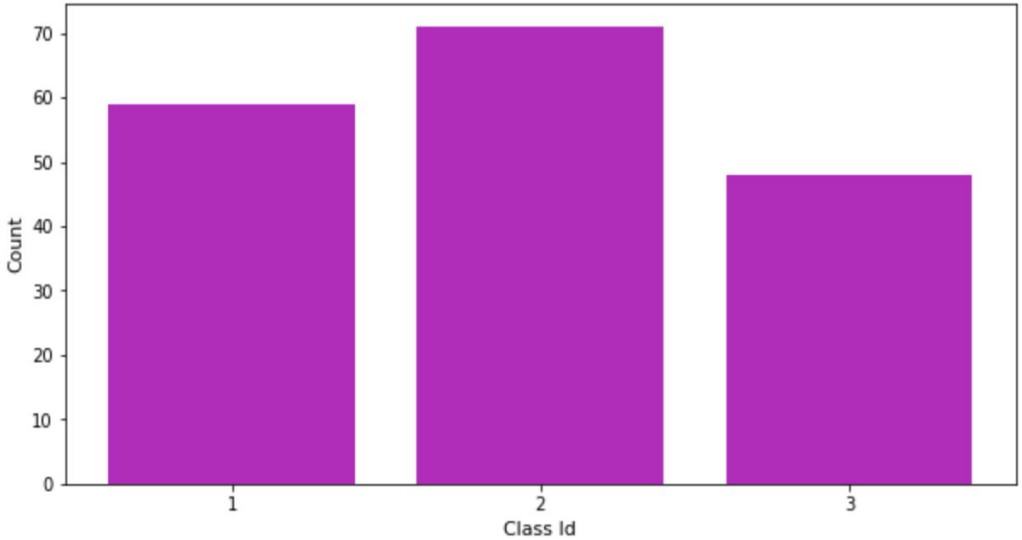
Determining the quantities of 13 constituents found in each of the three different types of wine.

Description of the Dataframe

This dataset has the result of a chemical analysis of wines that are grown in the same country (Italy) but originate from three different cultivars. The dataset consists of 178 Italian wine samples. Each sample has 14 features: one is the cultivar's identifier, and the others are chemical elements. The cultivar is a categorical variable that has 3 values: (Cultivar) 1, (Cultivar) 2, and (Cultivar) 3. The chemical elements are all numerical and continuous. The list of the 14 different features: Class Id, Alcohol, Malic Acid, Ash, Alcalinity Ash, Magnesium, Total Phenols, Flavanoids, Non-Flavanoids Phenols, Proanthocyanins, Color Intensity, Hue, OD280 OD315 DWines, Proline.

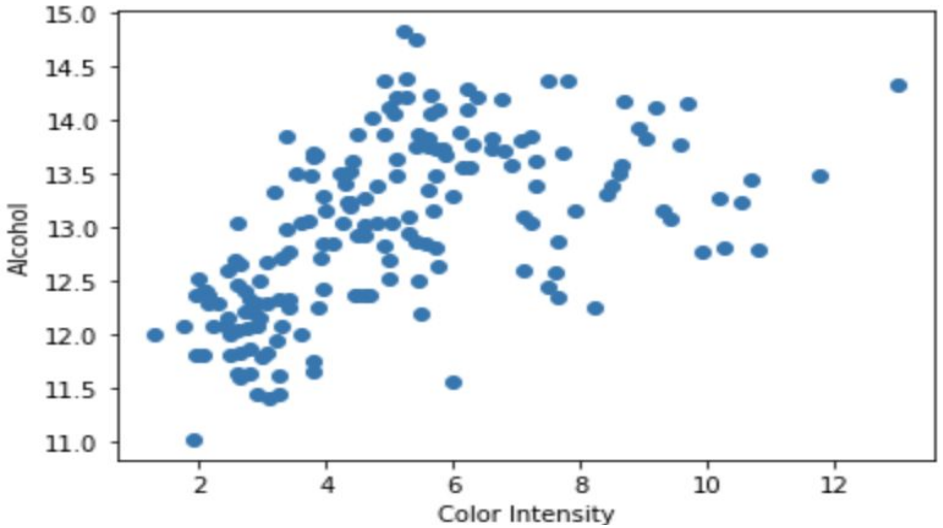
Visual 1

This is a representation of the three different wine cultivars. Cultivar 2 has the majority types of different wines.



Visual 2

This scatter plot is a visualization of the relation with the features: Alcohol and Color Intensity. They have a strong correlation with each other, whilst also showing a positive relation.



Model Evaluation

The models that I made are two different versions of the decision tree model and two different versions of KNN models. The best outcome was with the second KNN model. It had the highest accuracy score, 0 false negatives, and 1 false positive. Having a low amount of false negatives and false positives indicates that the model has a higher chance of predicting correctly. Although it had 15 true positives and 29 false negatives. The true positives were the highest out of the rest of the models, while also having the least amount of false negatives.

Final Model

The model that would perform the best would be the second KNN model. The classification report on the training and testing set is almost perfect. The precision, recall, and f1-score are high on both the testing and training set, all close to 1. This model would do well with the least amount of errors.