**K-NN classifier with stratified cross-validation**

**1-**

We selected the wine data as it was more interesting to us personally to understand how the chemical composition of a wine differentiated it form other ones from a similar region. We felt that predicting the region of a wine based on the chemical structure is a simpler model than predicting the danger level of tumors based on the size and other factors.

**2-**

We create a load function to load all aspects of the dataset into separate sections. Namely an array containing the predictors we will sue for finding the nearest neighbors in future sections, one with the correct predictions, which will be used in regression later, the names of these features are predictors for display purposes and the name of the file, for reference if needed.

**3-**

Here we show the first five samples of the dataset. We have 13 features which all describe a particularity of the wine (columns 1 to 13). Overall, we hope to see that these parameters can predict which one of these three cultivars (our 14th parameter, which ranges from 1-3, representing the three different wine cultivars in tour dataset) the wine came from. We can see a rough variation range of our predictors in these five initial rows.

Table

Description automatically generated

Table : First five rows of data in our dataset; shows roughly how the data varies between data point to data point. Since it is ordered, it only reflects wines from cultivar 1.0

In the following histogram we explore the distribution of alcohol percentage over our samples. We can see something akin to a normal distribution in the alcoholic percentages. It would be beneficial to graph each individual cultivar’s alcohol percentage against its occurrences to see the distribution of each of the different cultivars.

Chart, histogram

Description automatically generated

In the following graph we graph the two most influential features against one another. We plot alcohol content against malic acid content and plot the different colors to match the different cultivars accordingly to visualize clustering.

Chart, scatter chart

Description automatically generated

**4-**

Here we notice the relationship between the number of neighbors and the accuracy presented by the training and testing sets versus the model.

We found an optimum value of k = 3 neighbors, and an inverse relationship between the number of neighbors and the training set accuracy which seemed to stabilize at about 7 neighbors. We also found an unpredictable relationship between the number of neighbors and the testing set accuracy that requires further study.

Chart, line chart

Description automatically generated

**5-**When we run cross validation, we find around a 70% accuracy across the test and training sets, with the training set maintaining a higher accuracy on average, which reflects the findings in section 4 of this paper.

A picture containing table

Description automatically generated