### **Assignment 2 Report**

**Names:**[Kevin Spike](mailto:spik3669@kettering.edu) & Andrew Gabler

#### **Work Done for Sections**

**Section 1**

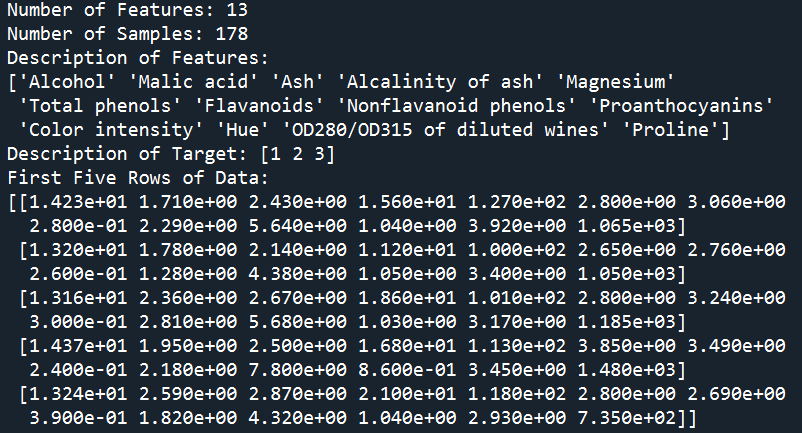
For our assignment we chose to use the wine data set that was given from the three options.

**Section 2**

For part 2 we created a python file that reads the data from the csv file and returns the following. We used Pandas to load a CSV. The load function assumes that the first column is always the feature name since that was the case for the Wine dataset. The Python file for this is *csvreader.py*. This file mostly functions as a library file, per the instructions, however, you can choose to run it independently and it will ask to be supplied the name of a file to read.

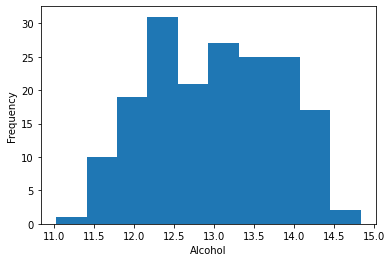
**Section 3**

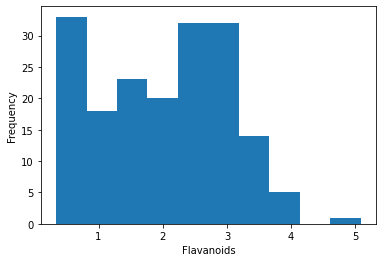
This table shows the required information about the data. This information, as well as the charts, were generated by the *meet\_the\_data.py* script. This was the output.

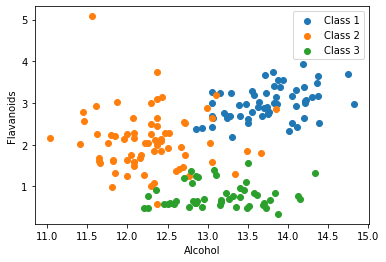


| Number of features | 13 |
| --- | --- |
| Number of samples | 178 |
| Description of features | Alcohol, Malic Acid, Ash, Alcalinity of Ash, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanidins, Color intensity, Hue, OD280/OD315 of diluted windes, Proline |
| Description of target | [1 2 3] |
| First five rows of data | [[1.423e+01 1.710e+00 2.430e+00 1.560e+01 1.270e+02 2.800e+00 3.060e+00  2.800e-01 2.290e+00 5.640e+00 1.040e+00 3.920e+00 1.065e+03]  [1.320e+01 1.780e+00 2.140e+00 1.120e+01 1.000e+02 2.650e+00 2.760e+00  2.600e-01 1.280e+00 4.380e+00 1.050e+00 3.400e+00 1.050e+03]  [1.316e+01 2.360e+00 2.670e+00 1.860e+01 1.010e+02 2.800e+00 3.240e+00  3.000e-01 2.810e+00 5.680e+00 1.030e+00 3.170e+00 1.185e+03]  [1.437e+01 1.950e+00 2.500e+00 1.680e+01 1.130e+02 3.850e+00 3.490e+00  2.400e-01 2.180e+00 7.800e+00 8.600e-01 3.450e+00 1.480e+03]  [1.324e+01 2.590e+00 2.870e+00 2.100e+01 1.180e+02 2.800e+00 2.690e+00  3.900e-01 1.820e+00 4.320e+00 1.040e+00 2.930e+00 7.350e+02]] |

For the features that appeared to be the two most influential to us, we chose Alcohol and Flavonoids. Below are the respective histograms and the scatter plot.

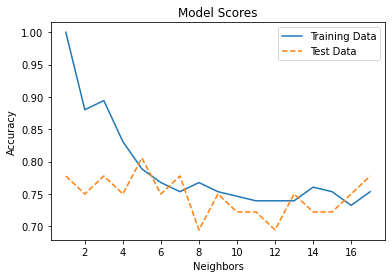






**Section 4**

We used the file *model\_development.py* to train a KNN model using different numbers of neighbour counts to determine the optimal count of neighbors and the best accuracy.



Best KNN Score Was With 5 Neighbors and Score Was 0.8055555555555556

**Section 5**

We used the file *cross\_validation.py* to generate the following table. The script uses stratified k-fold cross validation.

|  | Fold - 1 | Fold - 2 | Fold - 3 | Fold - 4 | Fold - 5 | mean |
| --- | --- | --- | --- | --- | --- | --- |
| Training Accuracy | 0.795775 | 0.809859 | 0.816901 | 0.825175 | 0.762238 | 0.801990 |
| Test Accuracy | 0.666667 | 0.638889 | 0.611111 | 0.685714 | 0.714286 | 0.663333 |

*Is the training and test accuracy in Step 4 validated using cross validation? Why or why not?*

#### **File Manifest**

*cross\_validation.py*

This is the file used to perform k-fold cross validation.

*csvreader.py*

This is the file we use to read the CSVs and use the data for the algorithms.

*meet\_the\_data.py*

This is the section where we do some pre-analysis on the data before we use it.

*model\_development.py*

This is where we develop our initial KNN model.

*report.docx*

Microsoft Word copy of the report.

*report.pdf*

PDF copy of the report.

*wine.data*

The dataset used. Modified for feature names.

*wine.names*

The description of the dataset used.