**Research Question**

Can we predict the class of wine based on specific features?

**Abstract**

Derived from Institute of Pharmaceutical and Food Analysis and Technologies based in Italy, current dataset represents the chemical analysis of Italian wines which are grown at the same region but by different cultivators. Deploying features of this dataset we try to make out relationships between them and their relative connectivity, which can then be used for classifying an unlabeled wine according to its continuous-valued characteristic features.

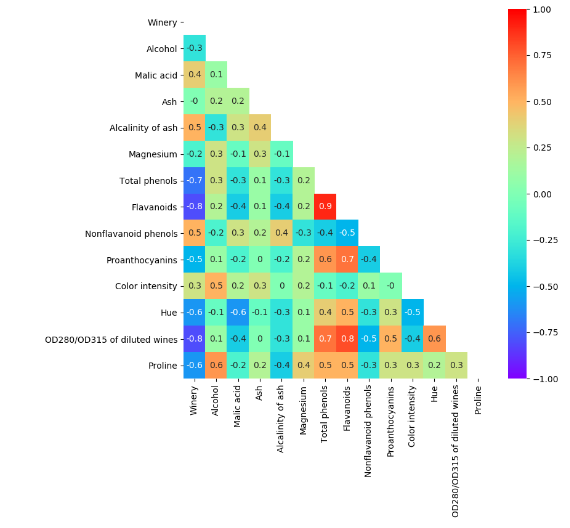
**Introduction**

We will focus on a small (“toy”) wine database which is publicly available at [UCI Wine data](https://archive.ics.uci.edu/ml/datasets/wine) and carries a categorical label for each wine along with several continuous-valued features such as alcohol content, flavor, hue etc.

**Methods**

The method used for modelling this data was the Logistic Regression, and PCA matrix decomposition algorithms built into scikit-learn. Pseudocode (and in particular, the objective function being minimized) can be found [here](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression). This model used for classification problems and logistic function used to describe the possible outcomes of a single trial by the probabilities. This method was chosen because of its simplicity and widespread usage for such cases.

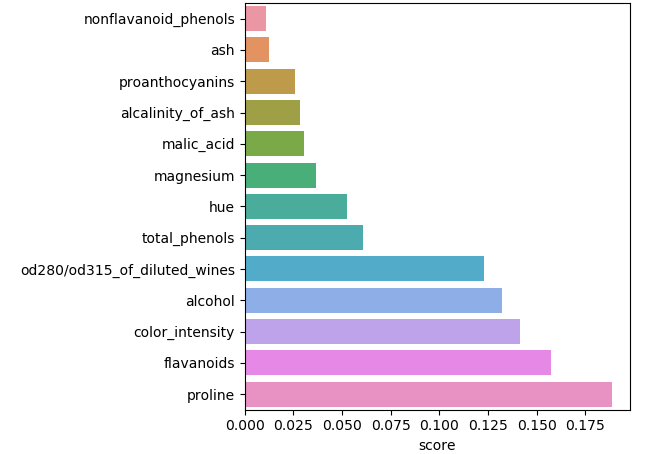
Firstly, the dataset was organized into a pandas data frame this allowed us to pairwise relationships between features and create correlation graph. As result, we can see that for some of the features such as flavanoids and total phenols, it is clear the class distributions have quite different means. Thus, we could expect that even simple models may be able to distinguish the wines.



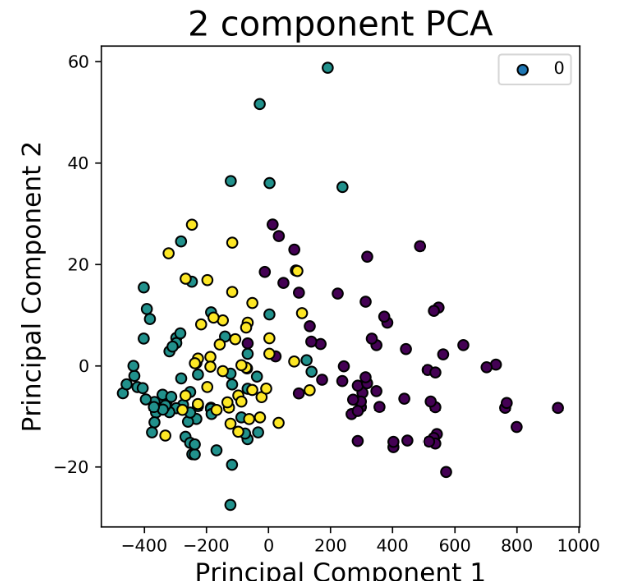
Secondly, we separated the data into a training and test sets, so that we can gauge the real-world applicability of the model.

Next step is a predicting the labels for our test dataset using three metrics to report performance of this model: accuracy score (a number that were correctly classified), confusion matrix (a matrix that tells the number of examples in a matrix based on the predicted and the real classes) and classification report (a precision, recall, and F1 score related to each classes).

Hereafter, we rank the importance of the features in the data, so we can learn which features were more and less important and select just the one's which contribute the most.



Further, we wanted to use PCA for feature selection and we only kept the top 2 components. By default, the transformed data was ordered by the components with the maximum variance (in descending order).



**Results**

The performance of the regressor was a value of 0.93. This shows that 'real-world' models for wine classification that could reliably predict wine classes >93% of the time. The figure below shows the performance on the testing set.



**Discussion**

By our point of view, this method could be improved by eliminating highly correlated features (for example, flavonoids and total\_phenols), setting up threshold in accordance with feature importance algorithm and applying other methods (for instance, KNeighbors Classifier).

**References**

The links referenced were included in my discussion, above.