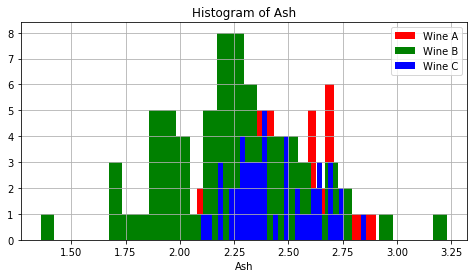
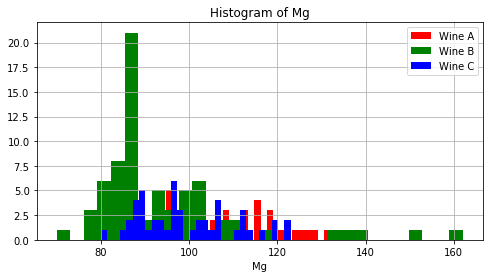
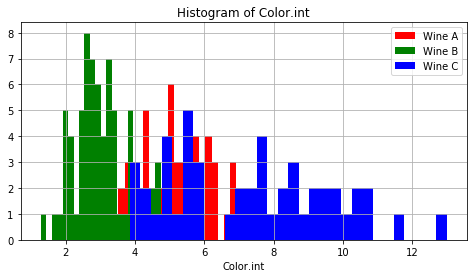
1. Mutual information. We are going to investigate the use of the mutual information criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a typical classifier.

First of all, you should select a suitable subset of variables and plot those with higher mutual information. Are you able to distinguish the three types of wine?







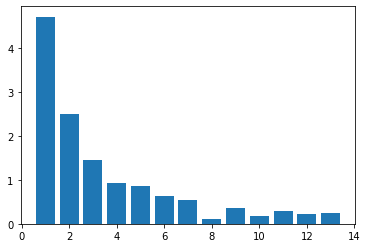
As seen by the plot of the histograms of Ash, Mg and Color.int agains the classes, it is visually difficult to distinguish between the 3.

2. Chi-Square. Repeat the selection of variables with Chi-Square method. Do you get the same results as with the previous one?

No, because chi-square can only be used with categorical data.

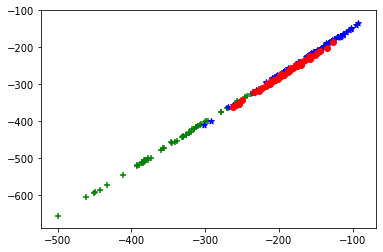
3. Principal Components Analysis (PCA). Now we are going to work with PCA as a method for dimensionality reduction. The Principal Component Analysis (PCA) was independently proposed by Karl Pearson (1901) and Harold Hotelling (1933) to turn a set of possibly correlated variables into a smaller set of uncorrelated variables. The idea is that a high-dimensional dataset is often described by correlated variables and therefore only a few meaningful dimensions account for most of the information. The PCA method finds those directions in the original dataset that account for the greatest variance in data, also called the principal components.

1. PCA without normalization:
   1. Calculate the eigenvalues and plot them. How many components do you need to explain 90% of the total variance?



Considering the plot and calculating the contribution of every component, we’d need at least 8 components to express 90% of the data.

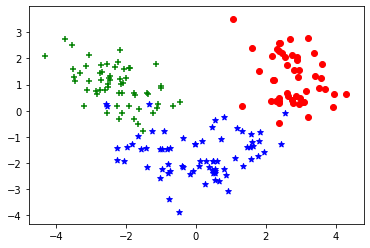
* 1. Plot the two first components, are the resulting clusters clearly separated?



No, they are all mixed.

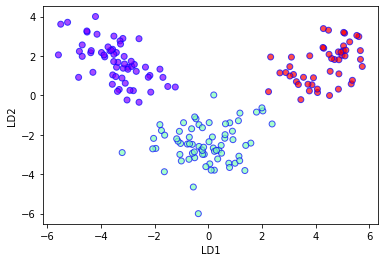
1. PCA with normalization: Repeat the two previous steps but in this case scaling the input to zero mean and unit variance N(0,1), it is also called z-scores. What do you see now? In our dataset, why does PCA without normalization perform poor?

The eigenvalues are the same, but the plot is completely different. The classes are well defined in this one.



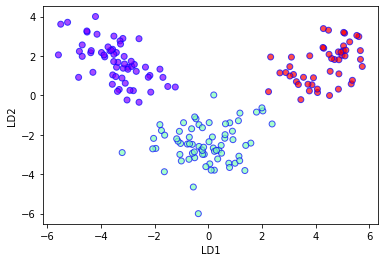
4. Linear Discriminant Analysis (LDA). What we aim for is a projection that maintains the maximum discriminative power of a given dataset, so a method should make use of class labels (if they are known a priori). The Linear Discriminant Analysis, invented by R. A. Fisher (1936), does so by maximizing the between-class scatter, while minimizing the within-class scatter at the same time.

1. LDA without normalization: Calculate the two components (C-1) and plot them, are the resulting clusters clearly separated?



Yes, they are clearly separated.

1. LDA with normalization: Repeat the previous step, what do you see now? Which is the difference with the previous one?



Normalized or not, the separation is the same. Therefore, LDA is indifferent to normalization.

5. Logistic Regression and Model Evaluation

From the wine dataset, we are going to study the importance (or not) of reducing dimensionality. We are going to apply logistic regression as a predictive model and see the influence of increasing the number of predictor variables. In the wine dataset the dependent variable is not binary, therefore we need to perform a study two to two classes: 1 vs. 2, 1 vs. 3 and 2 vs. 3

Considering the cases as 1 vs 2, 1 vs 3 and 2 vs 3, respectively, we got:

Using the full features: 0.97, 1.0 and 0.94 of accuracy.

Using PCA with 2 components: 0.90, 1.0 and 0.97 of accuracy.

Using PCA with 4 components: 0.97, 1.0 and 0.94 of accuracy.

Using PCA with 6 components: 0.92, 1.0 and 0.94 of accuracy.

Using LDA: 1.0, 1.0 and 1.0 of accuracy.

We can conclude, from the data, that PCA either has the same result of induces loss of accuracy, while LDA improves accuracy, compared to the full features case. Since we cannot create a rule for every dataset, we should always test and see what fits best.