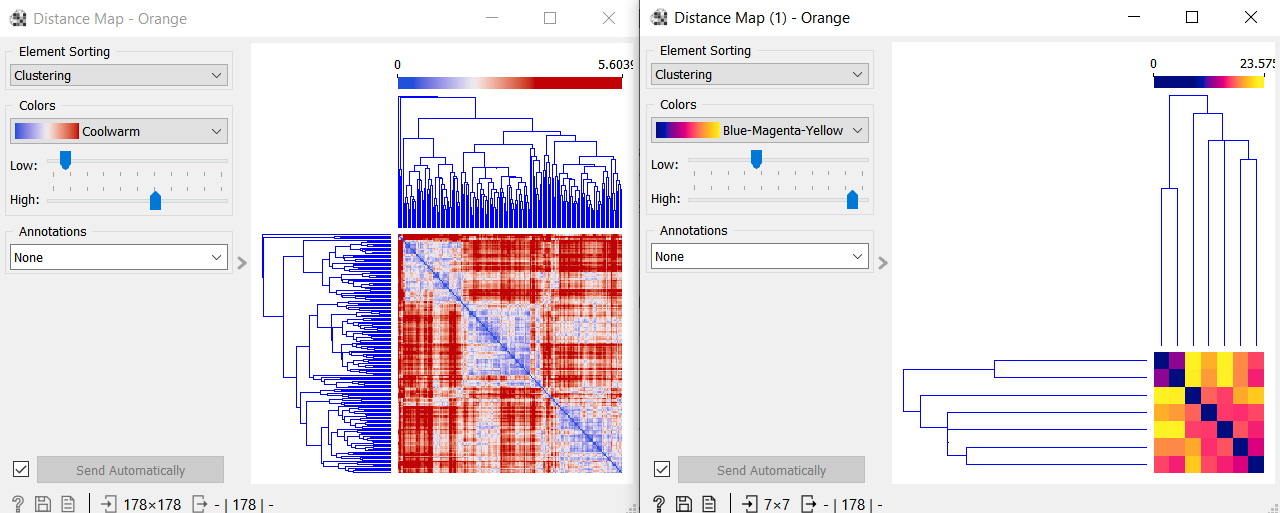
**Anastasia Psarou**

**Lab 1. The simplest classifier and data representation**

1. Select a dataset (preferably containing numerical values as features and categorical feature as class numbers).
2. Use **select columns** widget to control data structure.
3. Use **preprocessing** widget for data standardization. Why should we standardize data? Give an argument basing on **distances** and **distance map** widgets.

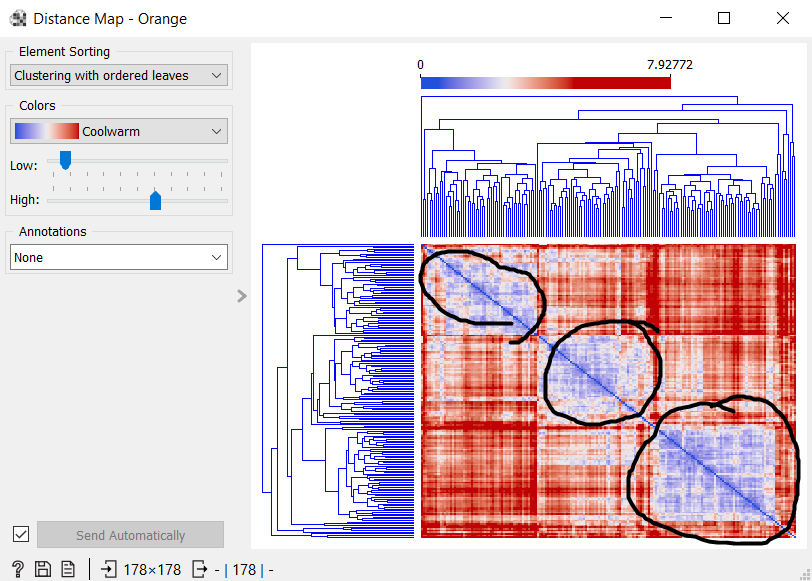
We should standardize data as it helps us make sure that data is internally consistent and each data type has the same content and format. Also, standardization contributes to the tracking of data that is not easy to compare otherwise.



Left is the distance matrix without the standardization and right is the distance matrix after the standardization.

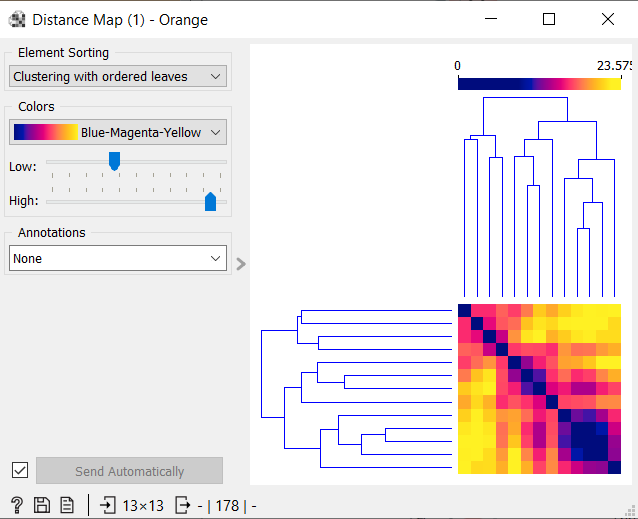
1. Is it possible to see classes using **distance map** widget???

Yes, it is possible. From this picture we can conclude that there are 3 clusters. The line in the middle represents the distance of each cluster from itself. So, we can see that there are 3 classes from where data are close to each other.

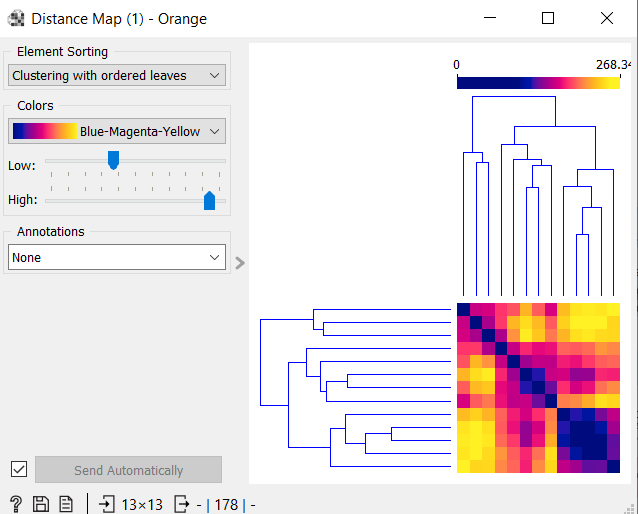


1. For which definition of distance, the classes are visible the most clearly?

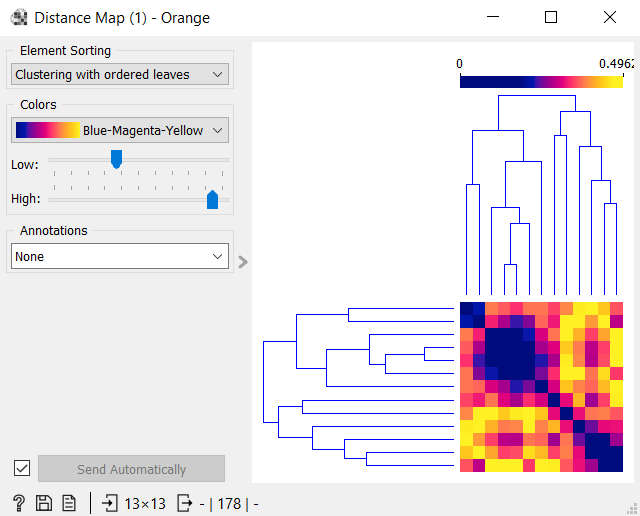
* Eucleide



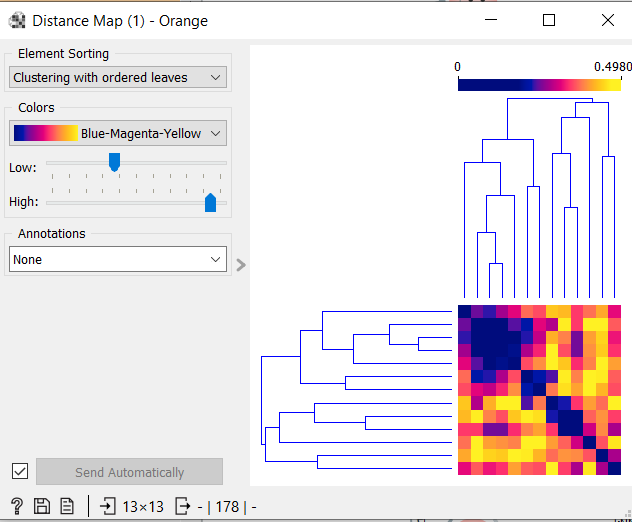
* Manhatan



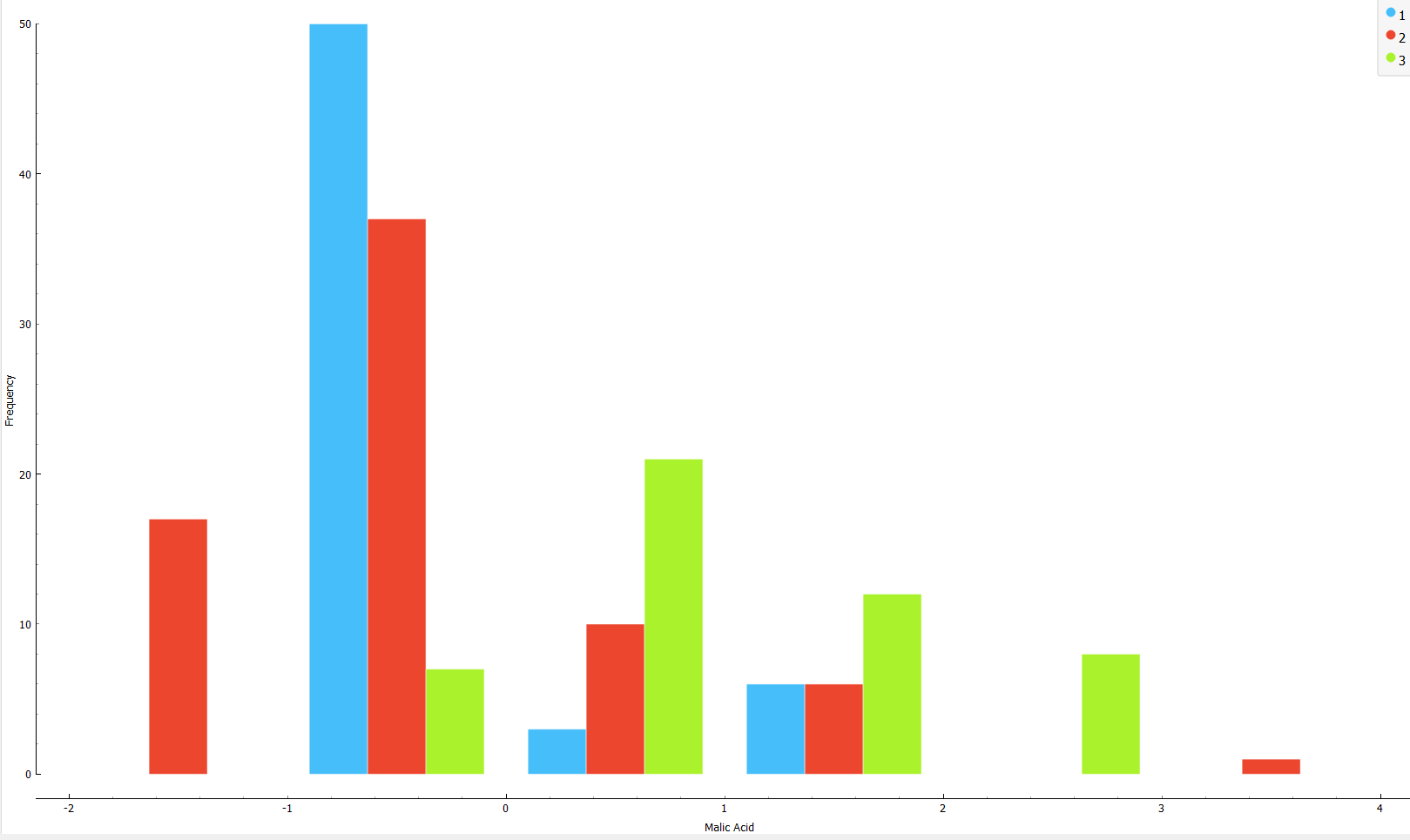
* Absolute Spearman



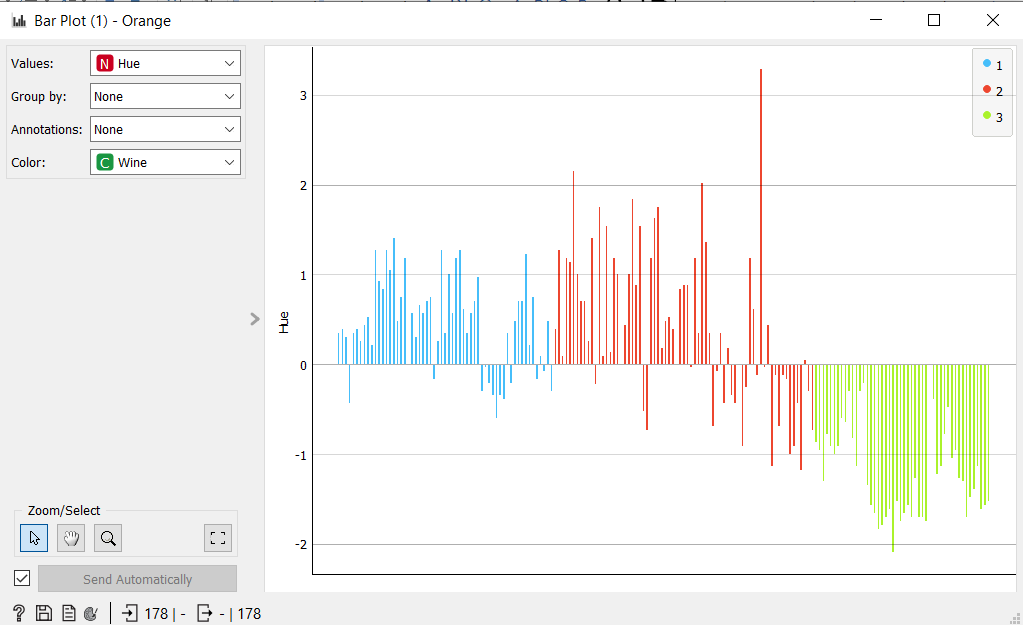
* Absolute Pearson



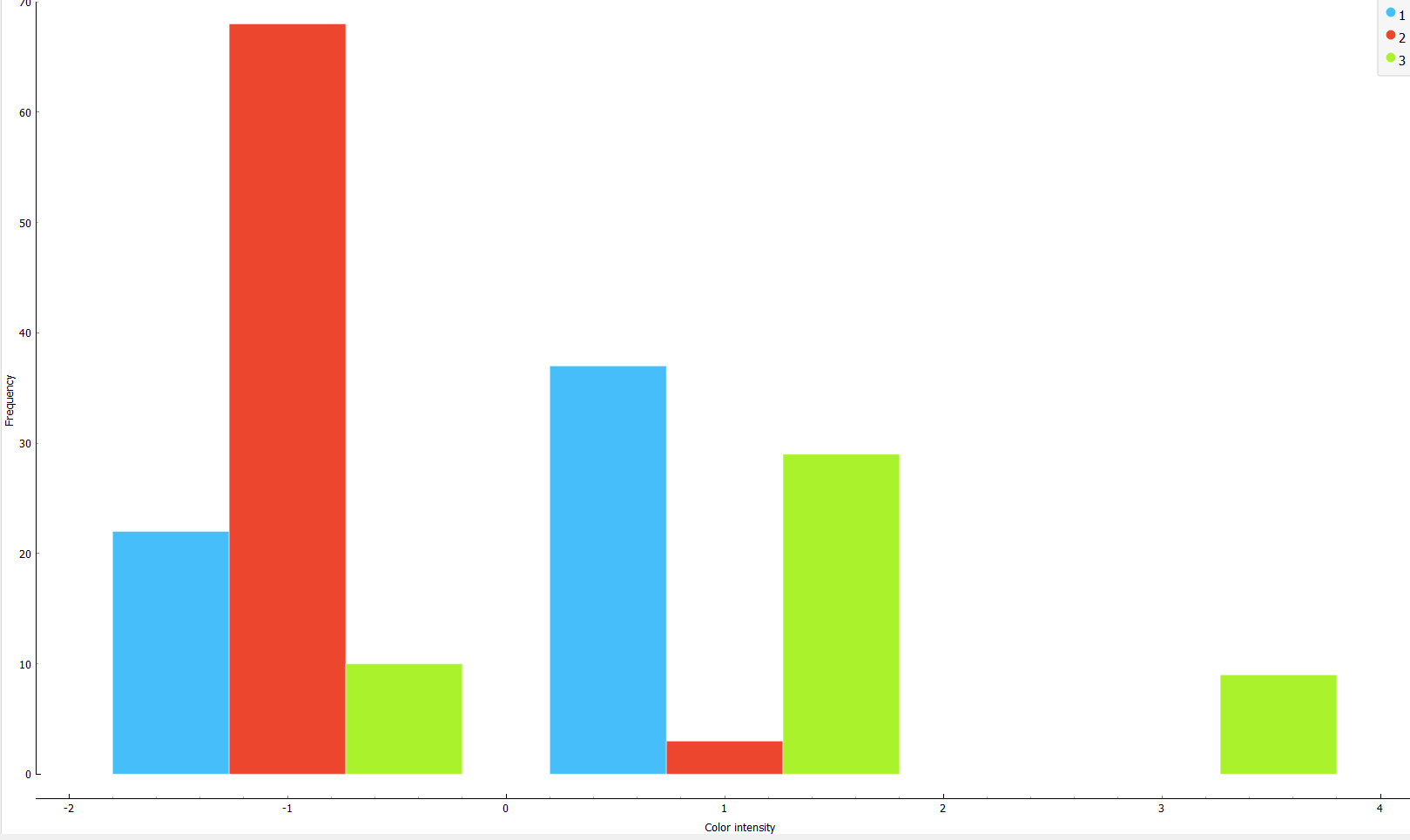
1. Using feature visualization widgets: **bar plot, violin plot, distributions**, define features which diversifies the most all the possible pairs of classes (i.e. 1-2, 2-3, 1-3 for three classes)



For example, in this distribution graph we can notice that between class 1 and 3 there is a lot of differentiation in the Molic Acid feature.

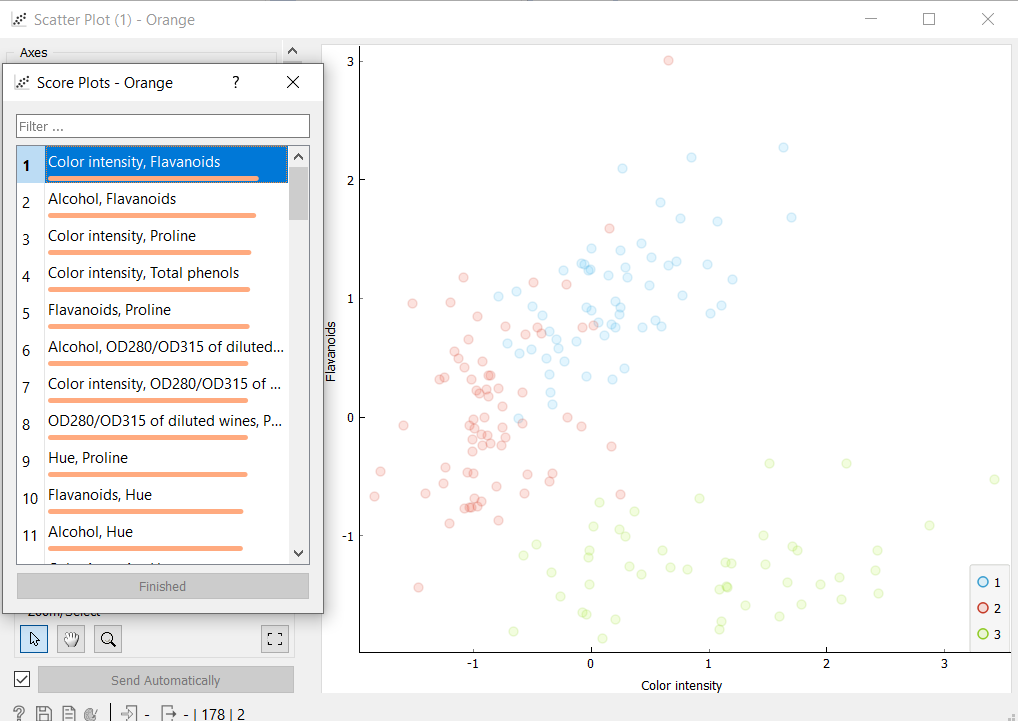


Also, in this bar plot we can see a differentiation between class 2 and 3 in the Hue feature.



Finally, between classes 1 and 2 there is a great differentiation for the Color intensity feature.

1. Using **scatter plot** widget find the two most contrastive features. Explain basing on the plots.



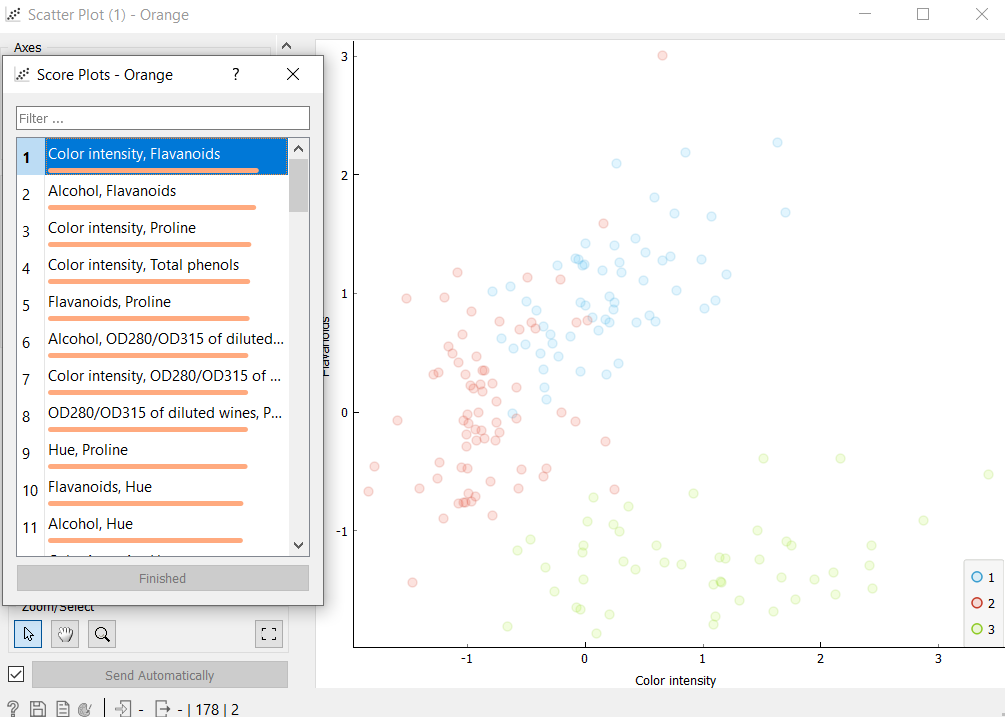
From the scatter plot we can conclude that the two most contrastive features are Color intensity, Flavonoids. From their plot we can also observe that there are few overlappings. We can notice it also from the graph on the right.

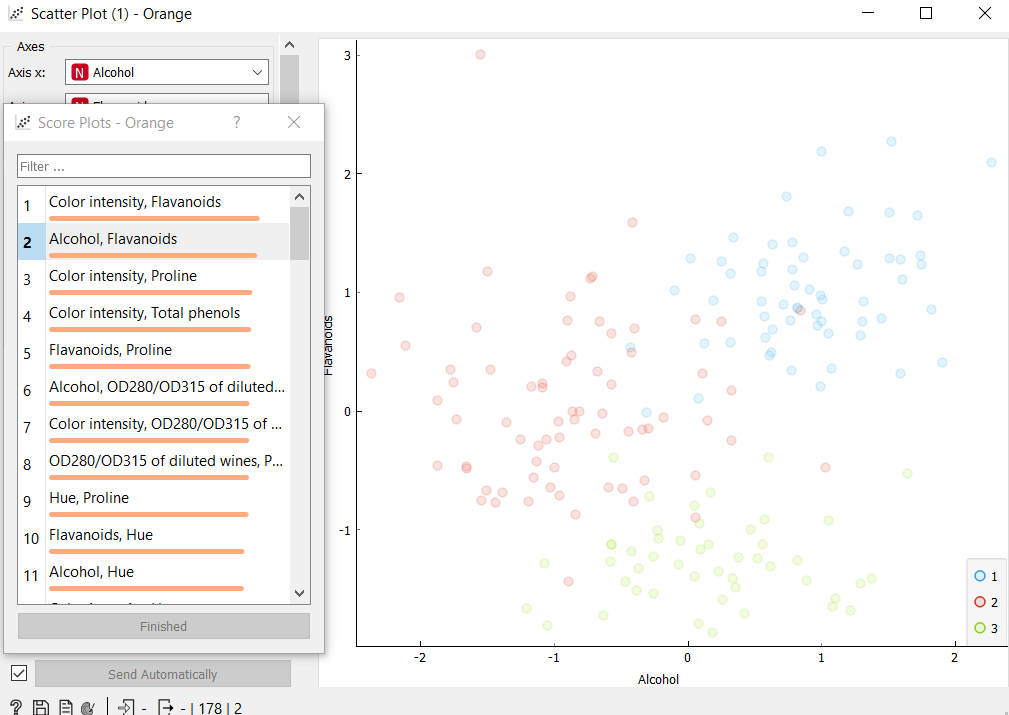
1. Can you use **distance map** for finding the most contrastive features.

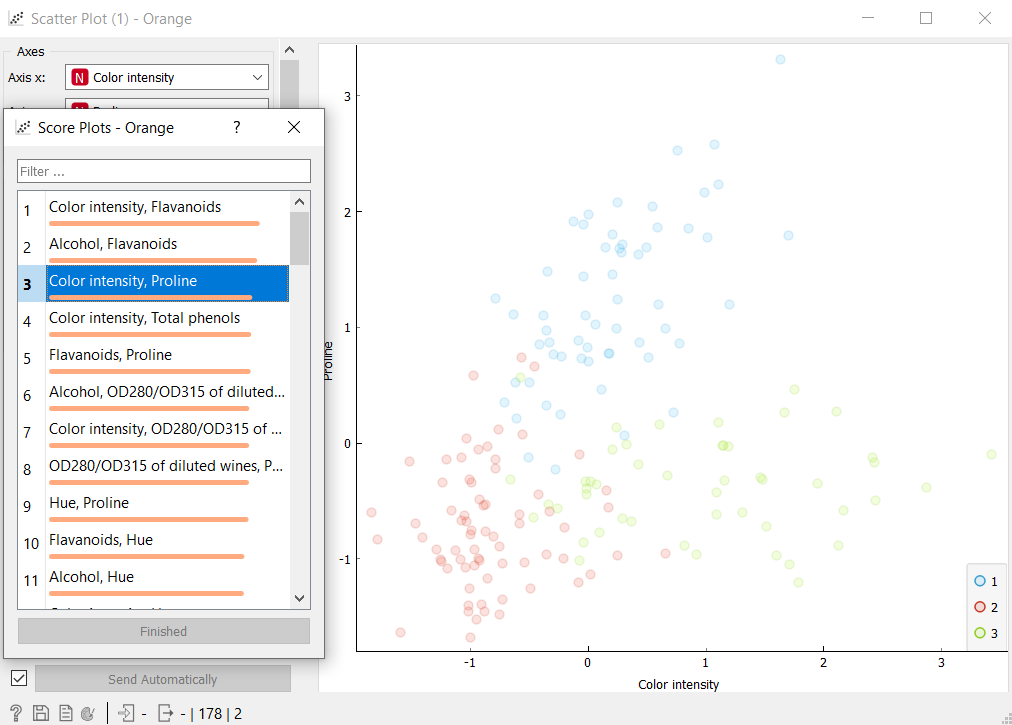
Yes this can be achieved by taking into consideration the distance between the lines of the plot.

1. Find k and metrics to obtain the best k-NN classification result.
2. Which result you obtain selecting the 3 best features.

These represent the 3 best features and we can see that there are very few overlappings.

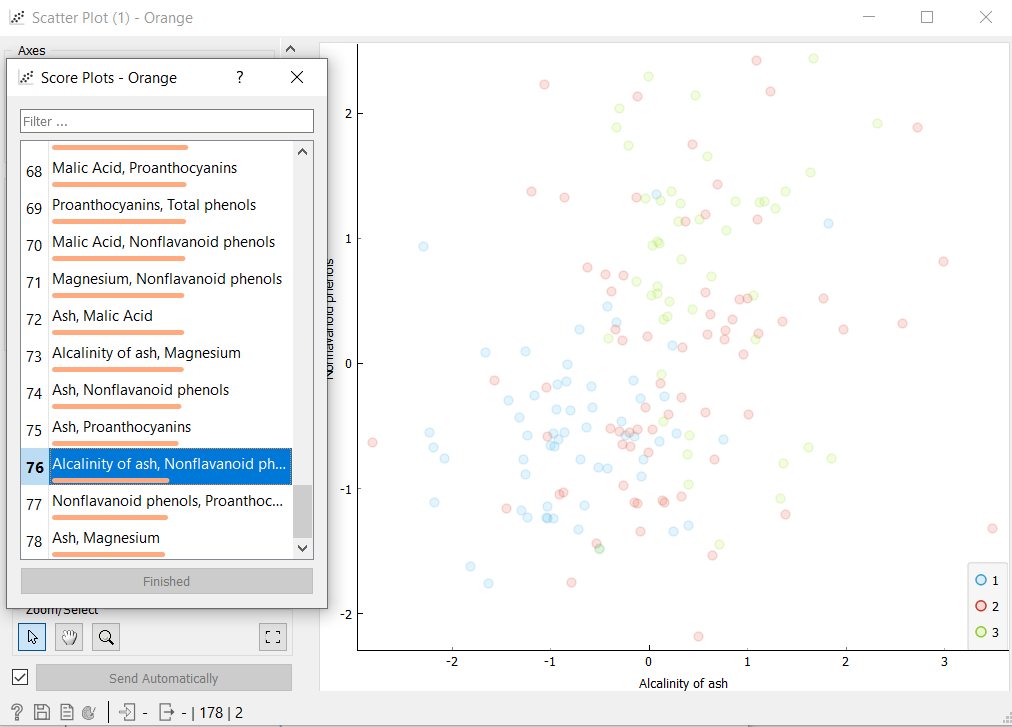


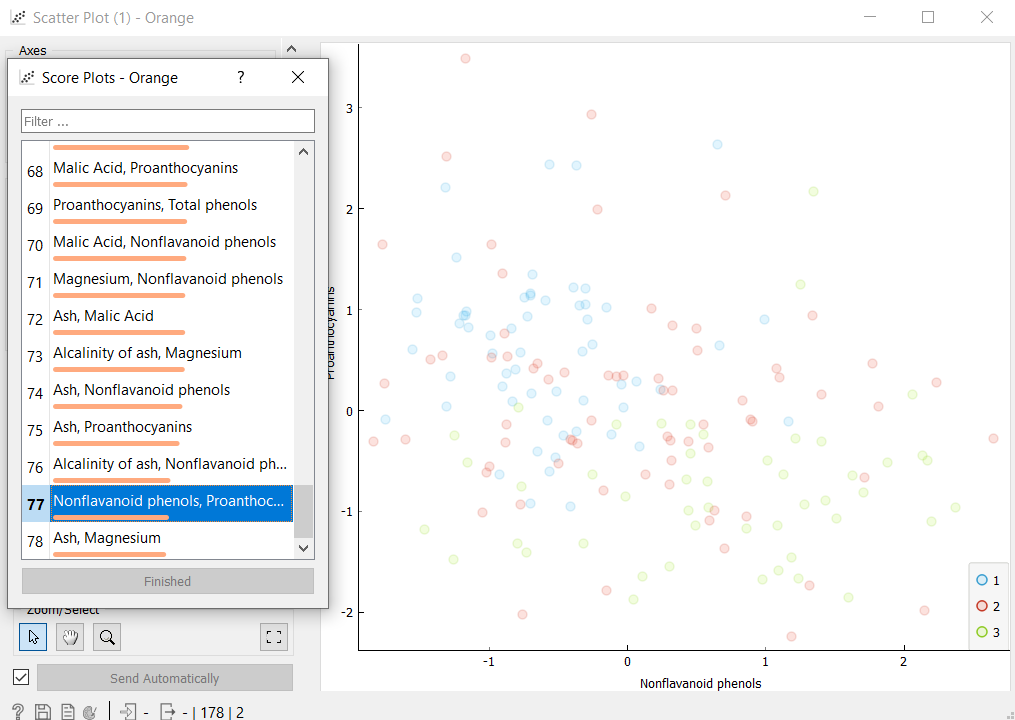


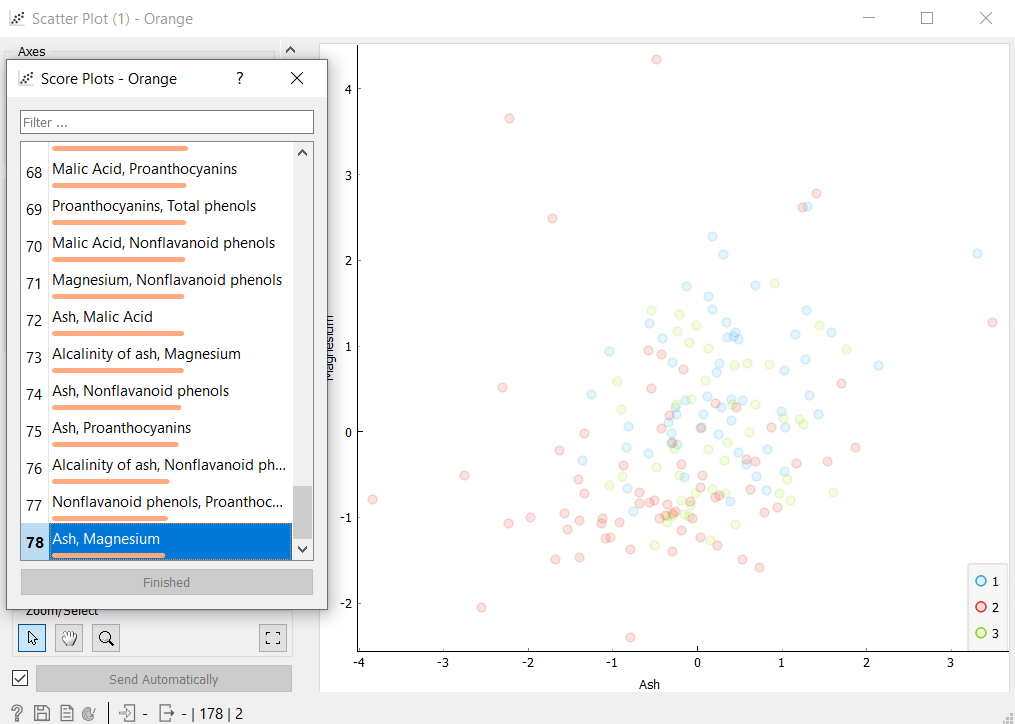


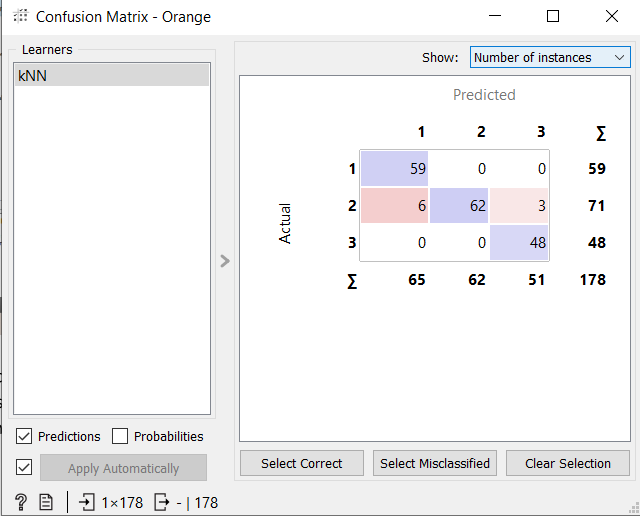
1. Which result you obtain selecting the 3 worst features.

These are the worst features because there are lot of overlappings compared to the previous.







1. Use **confusion matrix** widget for estimating which classes are the closest.  
   

From the confusion matrix we can draw the conclusion that class 1 is closer to class 2 as the percentage of misscategorization is the closest and the pair of 1, 3 is the most distant without misscategorizations.