Anastasia Psarou

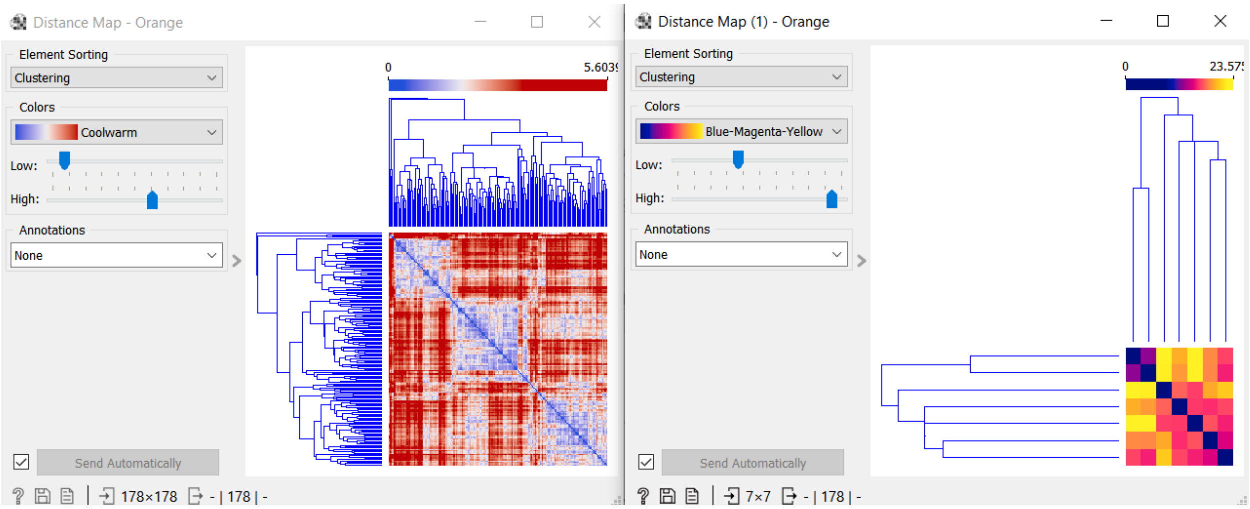
Lab 1. The simplest classifier and data representation

a) Select a dataset (preferably containing numerical values as features and categorical feature as class numbers).

b) Use **select columns** widget to control data structure.

c) 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.



Distance Map - Orange

Element Sorting: Clustering

Colors: Coolwarm

Annotations: None

☑ Send Automatically

178×178 | - | 178 | -

Distance Map (1) - Orange

Element Sorting: Clustering

Colors: Blue-Magenta-Yellow

Annotations: None

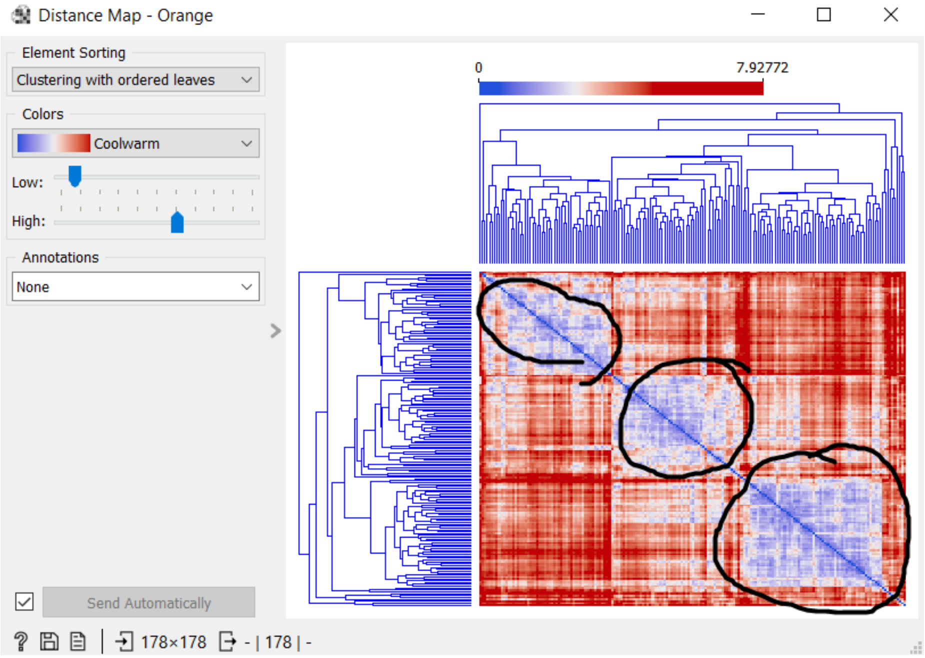
☑ Send Automatically

7×7 | - | 178 | -

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

d) 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.



Distance Map - Orange

Element Sorting: Clustering with ordered leaves

Colors: Coolwarm

Low:

High:

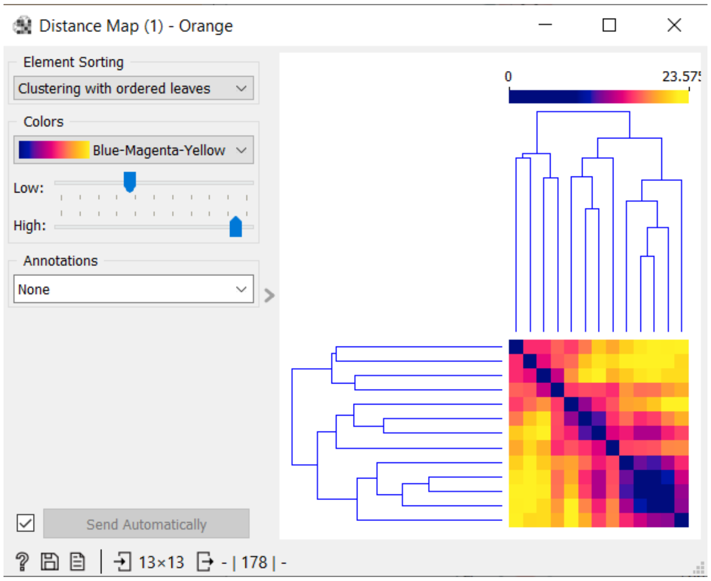
Annotations: None

☑ Send Automatically

178×178 - | 178 | -

e) For which definition of distance, the classes are visible the most clearly?

- Eucleide



Distance Map (1) - Orange

Element Sorting: Clustering with ordered leaves

Colors: Blue-Magenta-Yellow

Low:

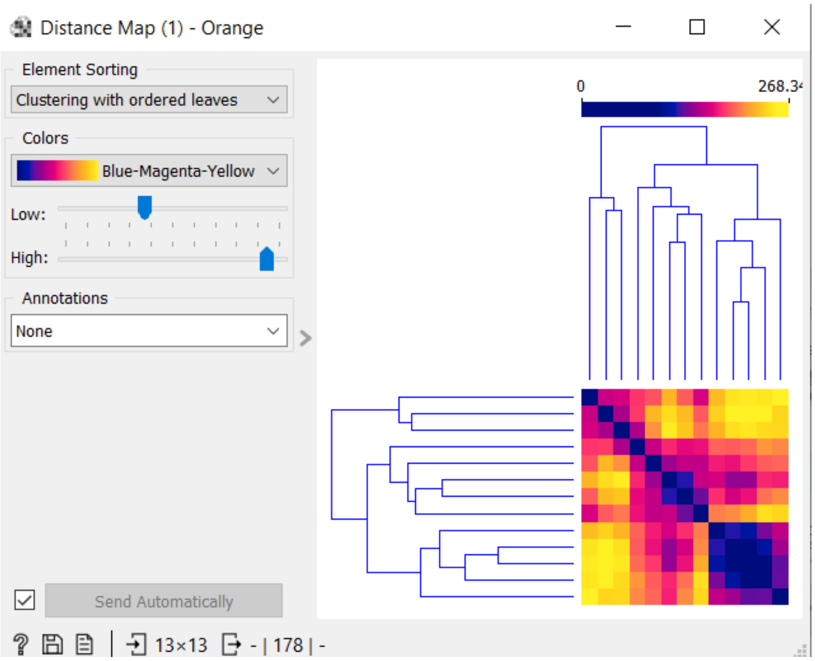
High:

Annotations: None

☑ Send Automatically

13×13 - | 178 | -

- Manhatan


Distance Map (1) - Orange

Element Sorting: Clustering with ordered leaves

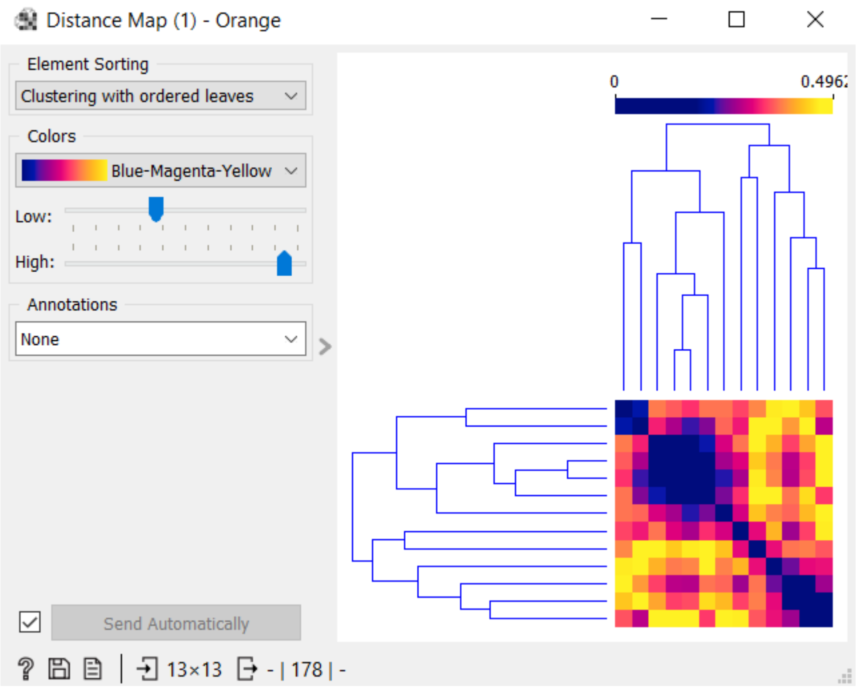
Colors: Blue-Magenta-Yellow

Annotations: None

☑ Send Automatically

13×13 - | 178 | -

- Absolute Spearman


Distance Map (1) - Orange

Element Sorting: Clustering with ordered leaves

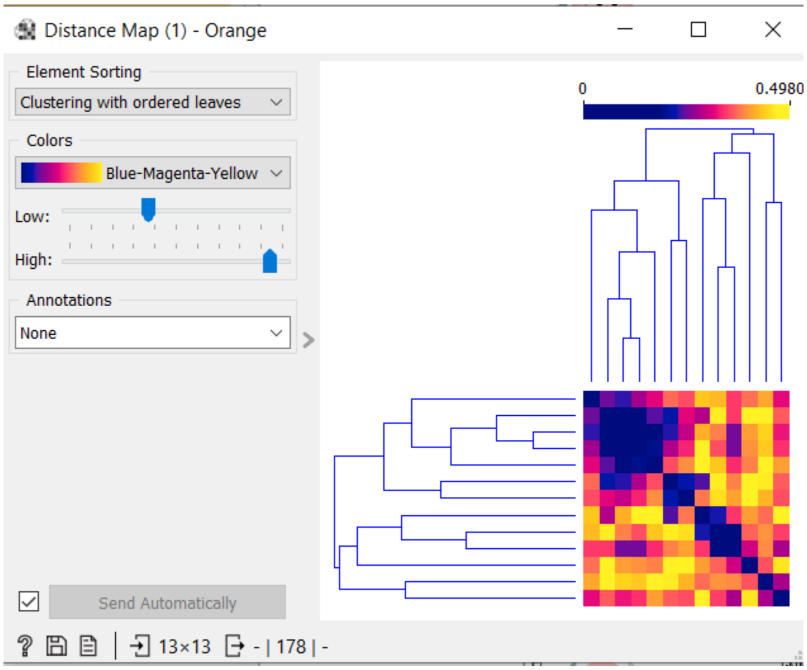
Colors: Blue-Magenta-Yellow

Annotations: None

☑ Send Automatically

13×13 - | 178 | -

- Absolute Pearson



Distance Map (1) - Orange

Element Sorting

Clustering with ordered leaves

Colors

Blue-Magenta-Yellow

Low:

High:

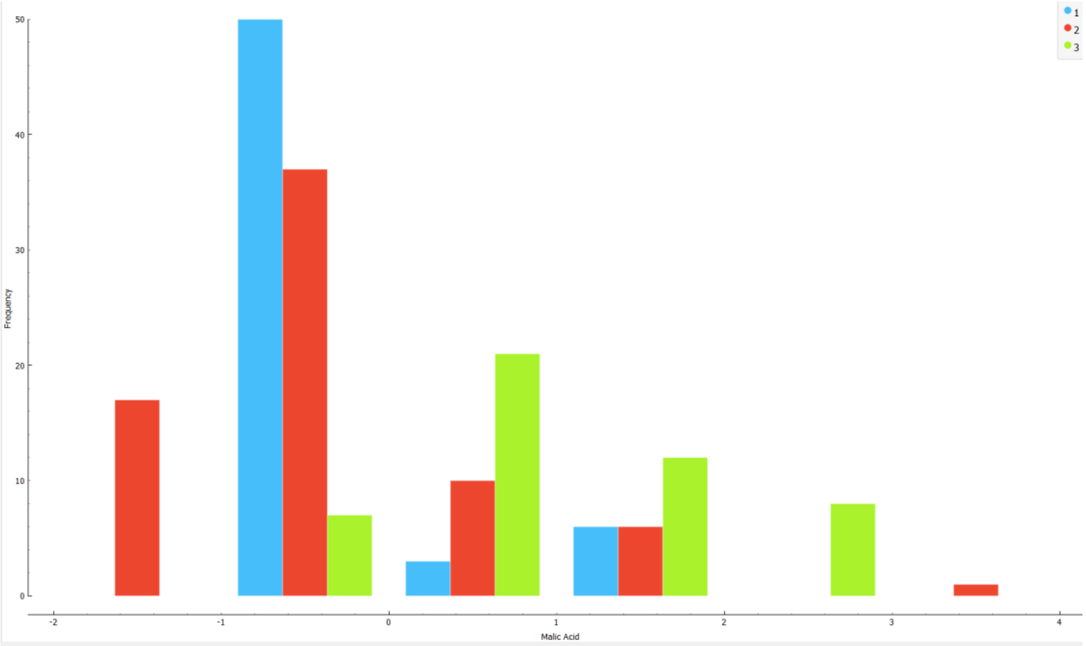
Annotations

None

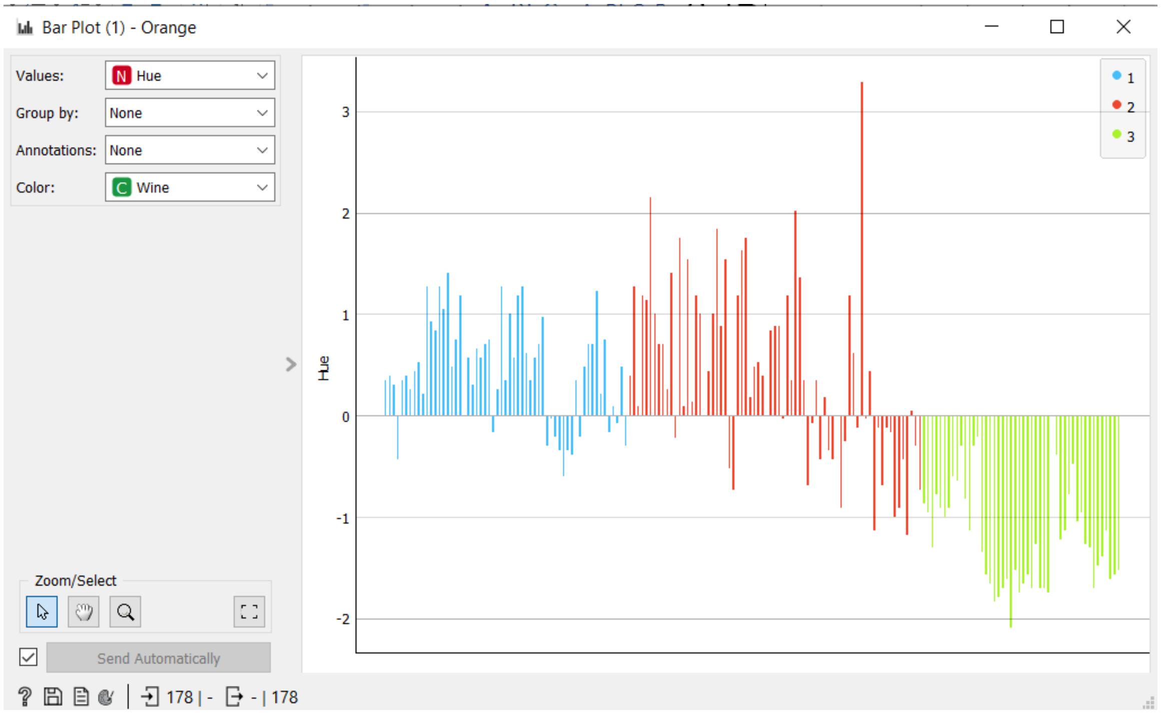
☑ Send Automatically

13×13 | - | 178 | -

f) 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.



Bar Plot (1) - Orange

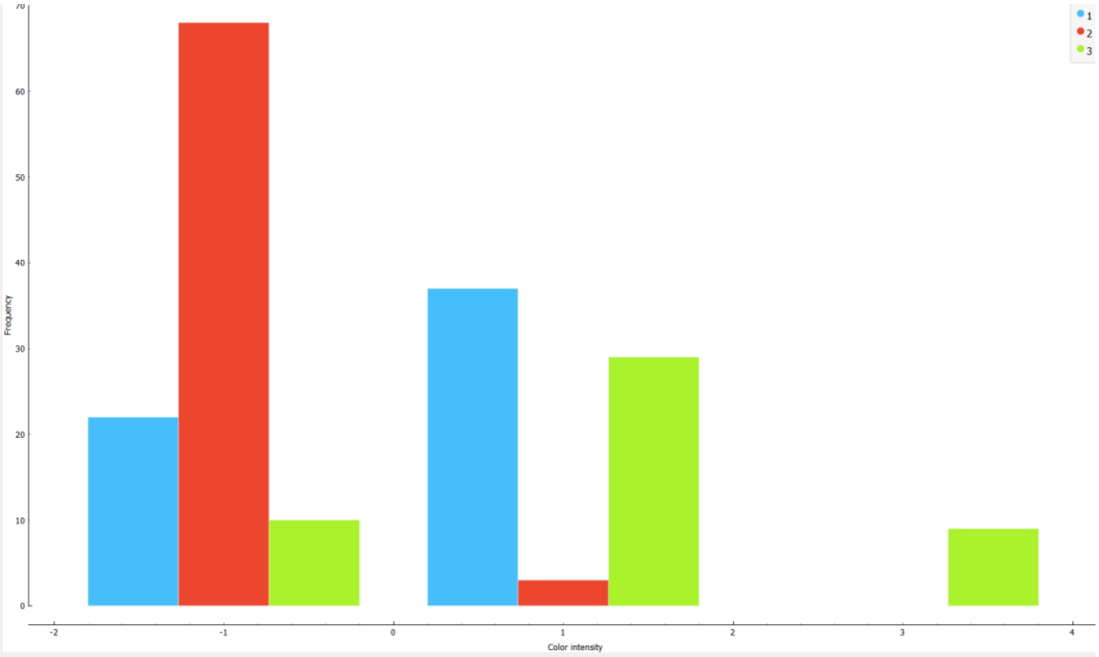
Values: N Hue

Group by: None

Annotations: None

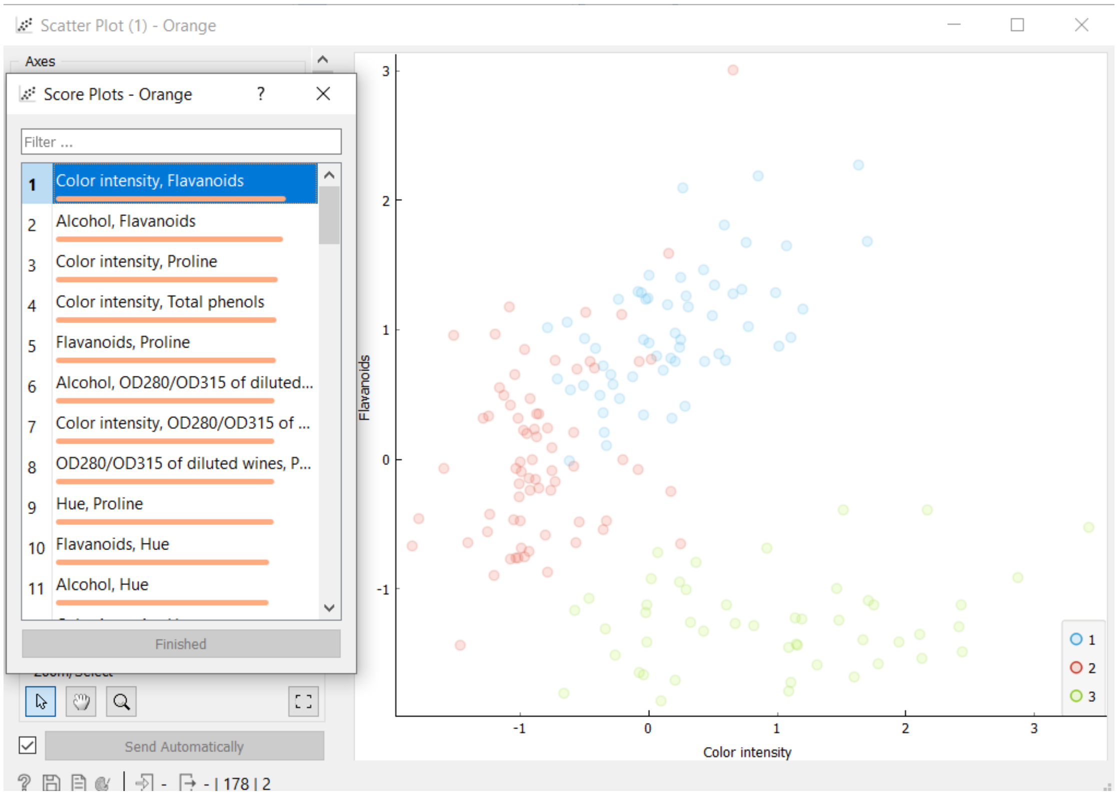
Color: C Wine

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.

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



1 Color intensity, Flavanoids
2 Alcohol, Flavanoids
3 Color intensity, Proline
4 Color intensity, Total phenols
5 Flavanoids, Proline
6 Alcohol, OD280/OD315 of diluted...
7 Color intensity, OD280/OD315 of ...
8 OD280/OD315 of diluted wines, P...
9 Hue, Proline
10 Flavanoids, Hue
11 Alcohol, Hue

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.

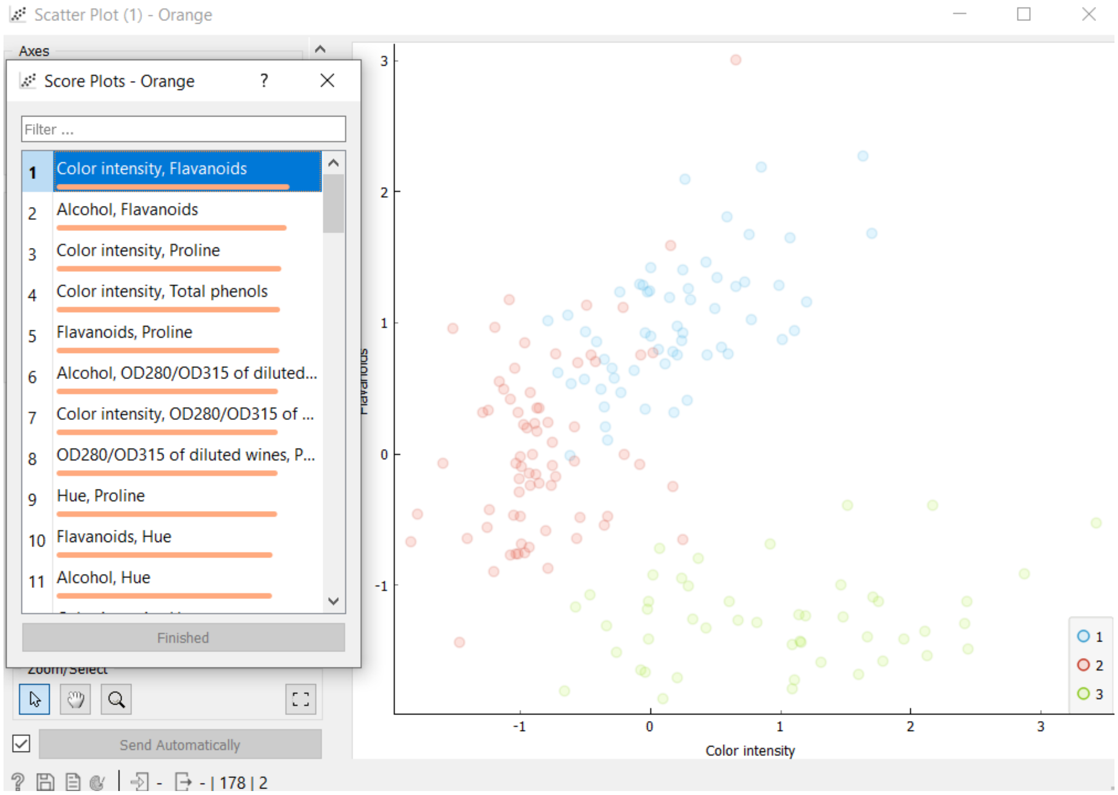
h) Can you use **distance map** for finding the most contrastive features.

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

i) Find k and metrics to obtain the best k-NN classification result.

j) Which result you obtain selecting the 3 best features.

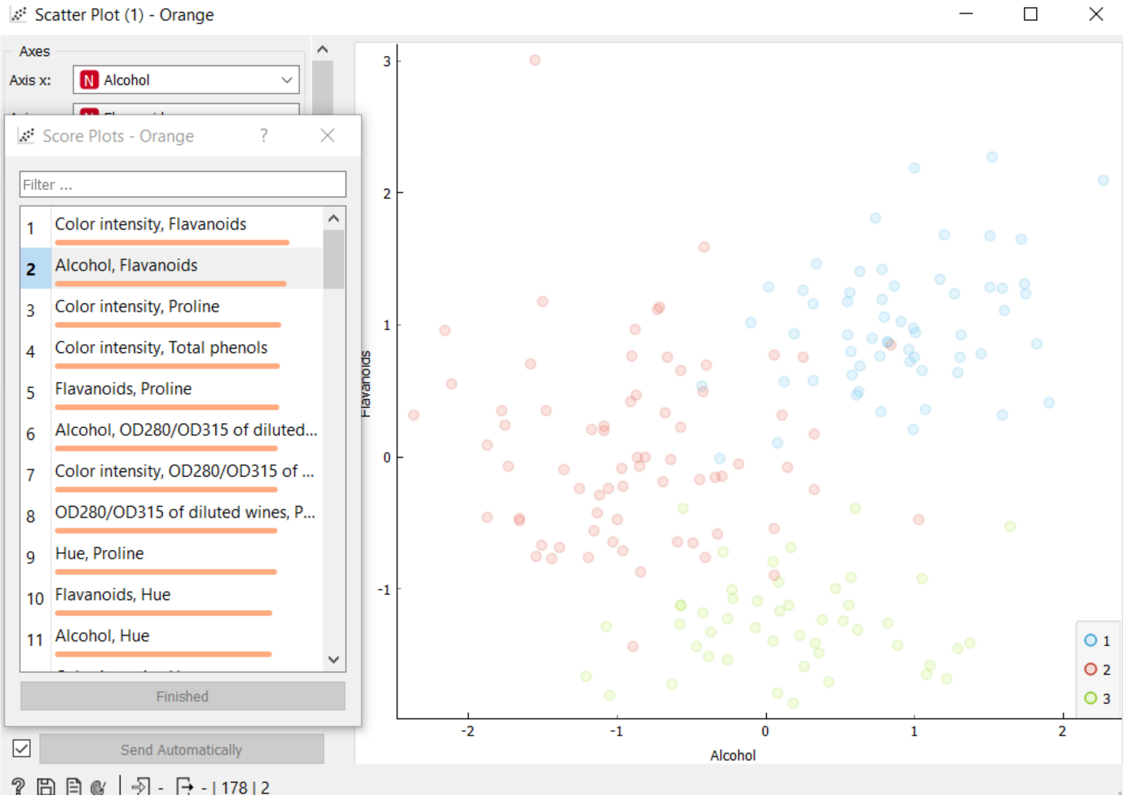
These represent the 3 best features and we can see that there are very few overlappings.



Scatter Plot (1) - Orange

Score Plots - Orange

1 Color intensity, Flavanoids
2 Alcohol, Flavanoids
3 Color intensity, Proline
4 Color intensity, Total phenols
5 Flavanoids, Proline
6 Alcohol, OD280/OD315 of diluted...
7 Color intensity, OD280/OD315 of ...
8 OD280/OD315 of diluted wines, P...
9 Hue, Proline
10 Flavanoids, Hue
11 Alcohol, Hue

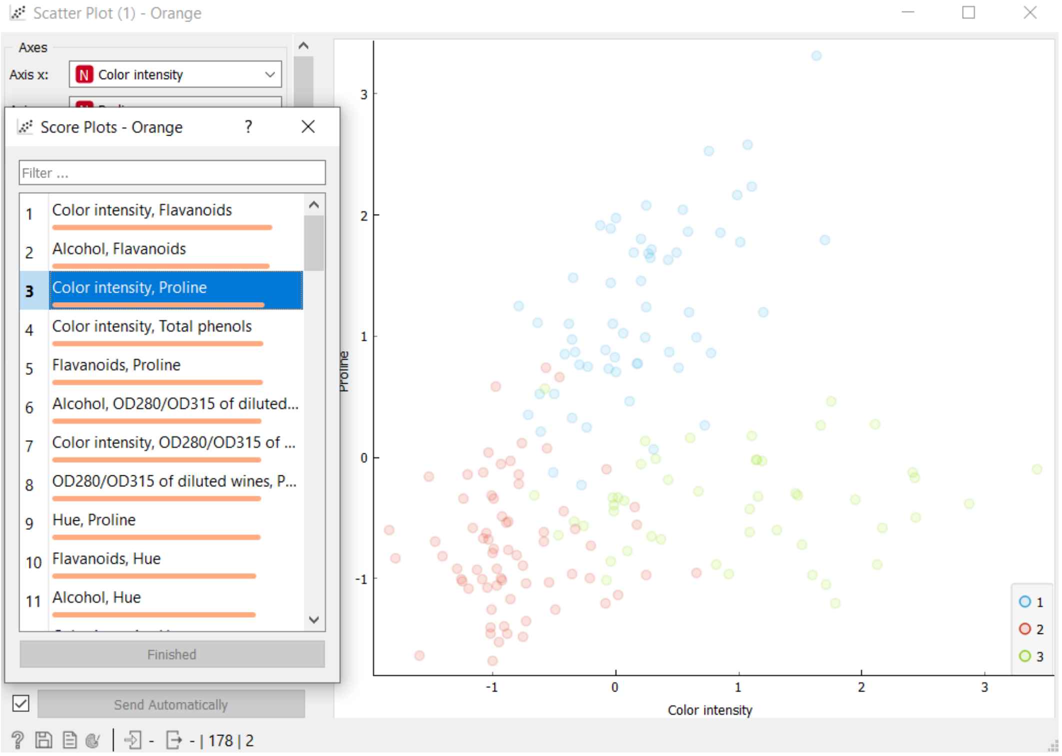


Scatter Plot (1) - Orange

Axes
Axis x: Alcohol

Score Plots - Orange

1 Color intensity, Flavanoids
2 Alcohol, Flavanoids
3 Color intensity, Proline
4 Color intensity, Total phenols
5 Flavanoids, Proline
6 Alcohol, OD280/OD315 of diluted...
7 Color intensity, OD280/OD315 of ...
8 OD280/OD315 of diluted wines, P...
9 Hue, Proline
10 Flavanoids, Hue
11 Alcohol, Hue

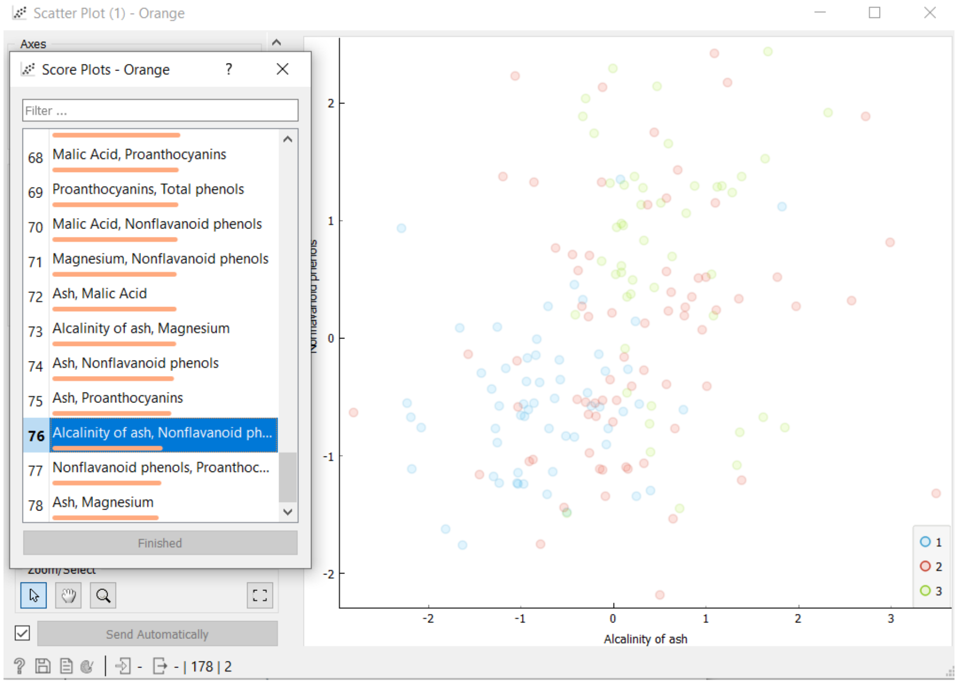


Axis x: Color intensity

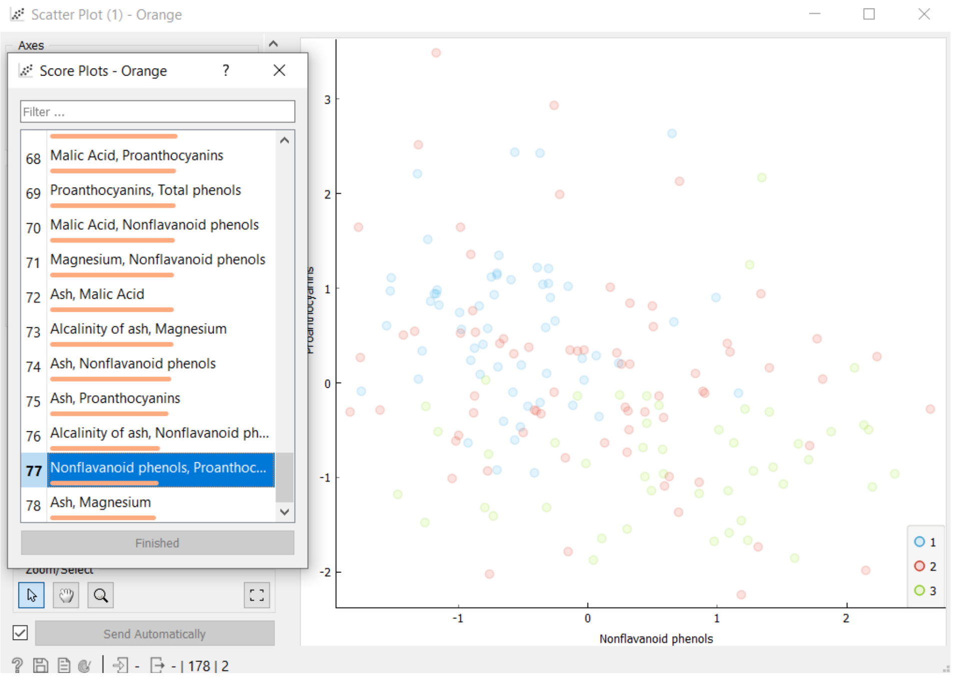
1 Color intensity, Flavanoids
2 Alcohol, Flavanoids
3 Color intensity, Proline
4 Color intensity, Total phenols
5 Flavanoids, Proline
6 Alcohol, OD280/OD315 of diluted...
7 Color intensity, OD280/OD315 of ...
8 OD280/OD315 of diluted wines, P...
9 Hue, Proline
10 Flavanoids, Hue
11 Alcohol, Hue

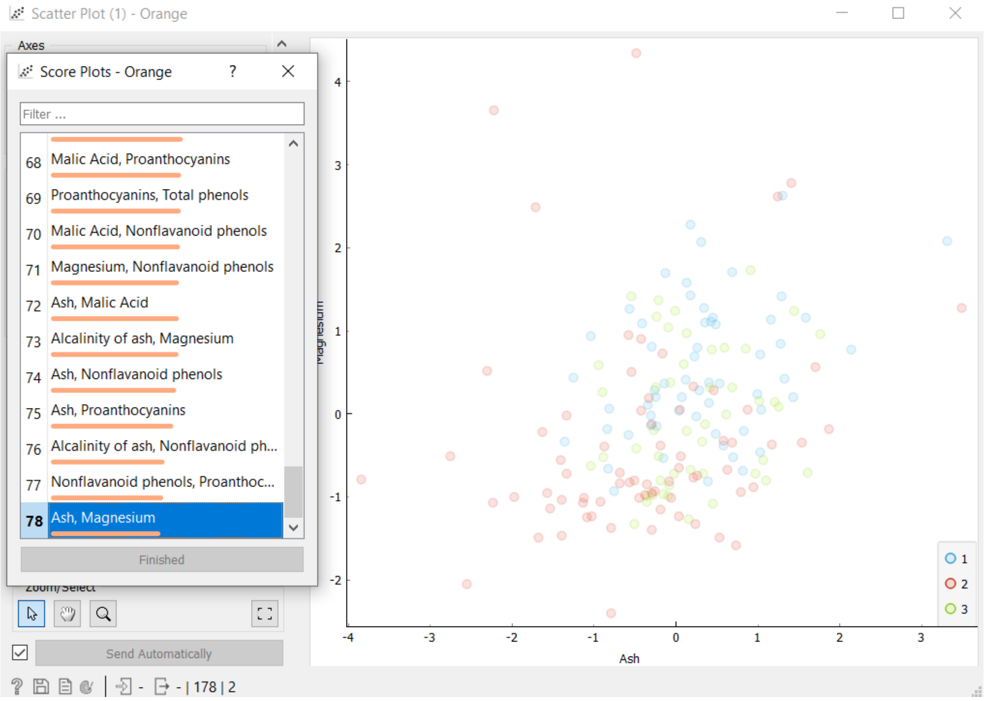
k) Which result you obtain selecting the 3 worst features.

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

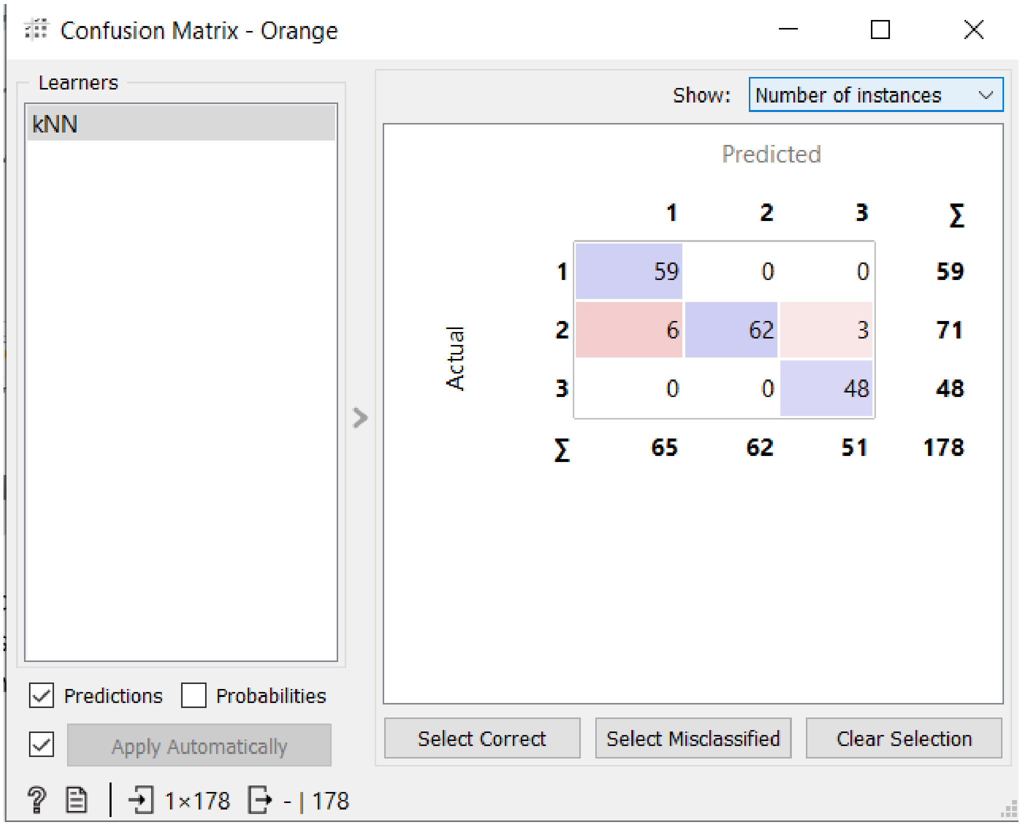


68 Malic Acid, Proanthocyanins
69 Proanthocyanins, Total phenols
70 Malic Acid, Nonflavanoid phenols
71 Magnesium, Nonflavanoid phenols
72 Ash, Malic Acid
73 Alcalinity of ash, Magnesium
74 Ash, Nonflavanoid phenols
75 Ash, Proanthocyanins
76 Alcalinity of ash, Nonflavanoid ph...
77 Nonflavanoid phenols, Proanthoc...
78 Ash, Magnesium


Scatter Plot (1) - Orange
Axes
Score Plots - Orange
?
Filter ...
68 Malic Acid, Proanthocyanins
69 Proanthocyanins, Total phenols
70 Malic Acid, Nonflavanoid phenols
71 Magnesium, Nonflavanoid phenols
72 Ash, Malic Acid
73 Alcalinity of ash, Magnesium
74 Ash, Nonflavanoid phenols
75 Ash, Proanthocyanins
76 Alcalinity of ash, Nonflavanoid ph...
77 Nonflavanoid phenols, Proanthoc...
78 Ash, Magnesium
Finished
Zoom/Select
Send Automatically
- | 178 | 2
Proanthocyanins
3
2
1
0
-1
-2
-1
0
1
2
Nonflavanoid phenols
1
2
3


Scatter Plot (1) - Orange
Axes
Score Plots - Orange
?
Filter ...
68 Malic Acid, Proanthocyanins
69 Proanthocyanins, Total phenols
70 Malic Acid, Nonflavanoid phenols
71 Magnesium, Nonflavanoid phenols
72 Ash, Malic Acid
73 Alcalinity of ash, Magnesium
74 Ash, Nonflavanoid phenols
75 Ash, Proanthocyanins
76 Alcalinity of ash, Nonflavanoid ph...
77 Nonflavanoid phenols, Proanthoc...
78 Ash, Magnesium
Finished
Zoom/Select
Send Automatically
- | 178 | 2
Magnesium
4
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
Ash
1
2
3

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



Show: Number of instances

Predicted

Actual

	1	2	3	∑
1	59	0	0	59
2	6	62	3	71
3	0	0	48	48
∑	65	62	51	178

☑ Predictions □ Probabilities

☑ Apply Automatically

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.