Assignment Part D – AIS2101Intelligent systems

Lasse Raaum

Dataset used: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

Github:

Part 1:

Description of the dataset:

This datasets is related to red variants of the Portuguese "Vinho Verde" wine.The dataset describes the amount of various chemicals present in wine and their effect on it's quality. The datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

Your task is to predict the quality of wine using the given data.

This dataset contains the following:
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
12 - quality (score between 0 and 10)

License:

Author:

M Yasser H

The dataset was collected by downloading it from “Kaggle.com”

Description of the content of the dataset:

There are 13 features and 1143 instances in this dataset.

Removed the ID feature, because it doesn’t tell us anything. It just counts from 1-1143.

Description of each feature in the dataset:

Name	Role	Value Type	Range	Meaning
Fixed acidity	Feature	Numeric	4.6 – 15.9	
Volatile acidity	Feature	Numeric	0.12 – 1.58	
Citric acid	Feature	Numeric	0 – 1	
Residual sugar	Feature	Numeric	0.9 – 15.5	
Chlorides	Feature	Numeric	0.012 – 0.611	
Free sulfur dioxide	Feature	Numeric	1 – 68	
Total sulfur dioxide	Feature	Numeric	6 – 289	
Density	Feature	Numeric	0.990 – 1.003	Density
pH	Feature	Numeric	2.74 – 4.01	pH – value
Sulphates	Feature	Numeric	0.33 – 2	
Alcohol	Feature	Numeric	8.4 – 14.9	% alcohol
Quality	**Target**	**Numeric**	**3 - 8**	**Given rating**

Classes:

With our target being Quality, that makes the number of classes 5 (3-8 range). 3 being the lowest quality and 8 being the highest.

Plots:

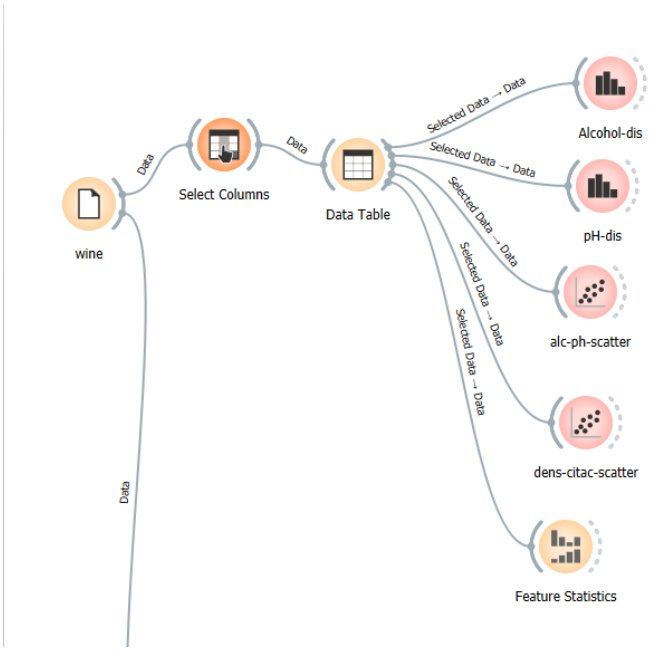
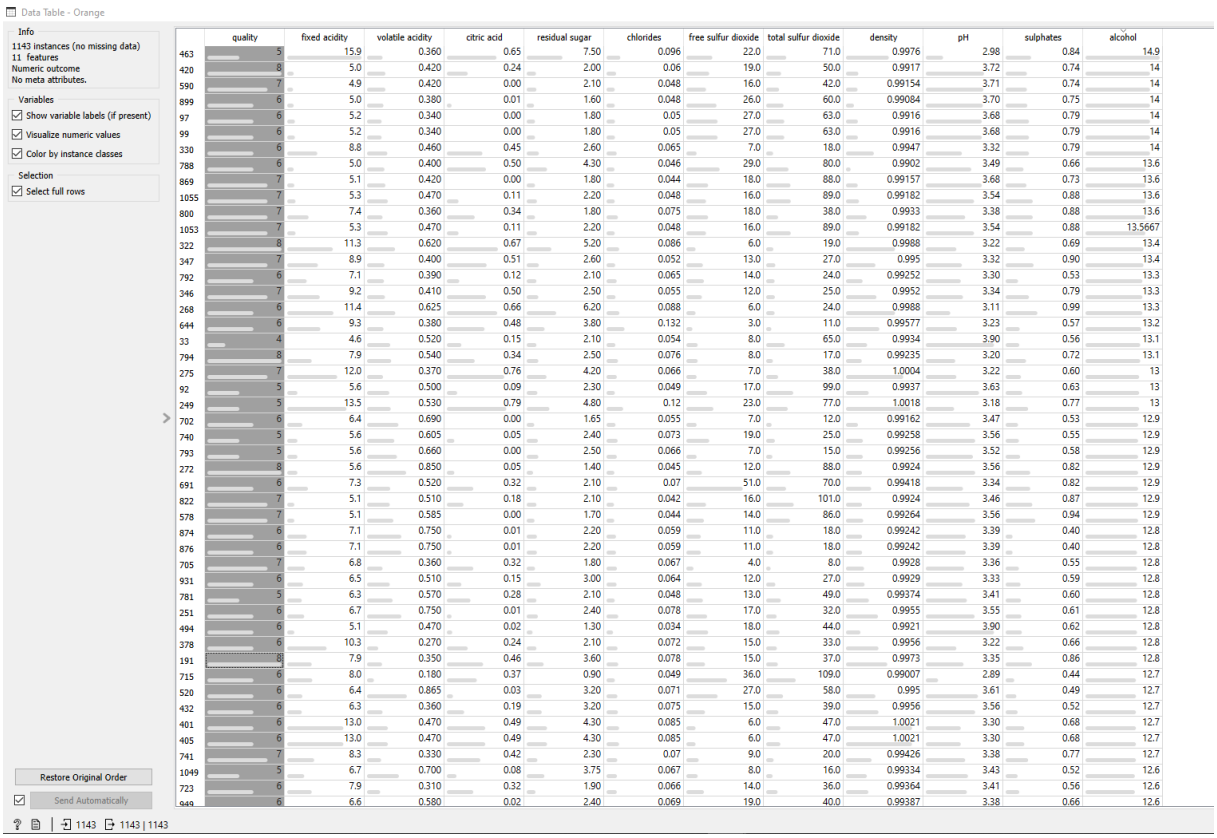


Figure 1 - Part 1 structure



	quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
463	5	15.9	0.360	0.65	7.50	0.096	22.0	71.0	0.9976	2.98	0.84	14.9
420	8	5.0	0.420	0.24	2.00	0.06	19.0	50.0	0.9917	3.72	0.74	14
590	7	4.9	0.420	0.00	2.10	0.048	16.0	42.0	0.99154	3.71	0.74	14
899	6	5.0	0.380	0.01	1.60	0.048	26.0	60.0	0.99084	3.70	0.75	14
97	6	5.2	0.340	0.00	1.80	0.05	27.0	63.0	0.9916	3.68	0.79	14
99	6	5.2	0.340	0.00	1.80	0.05	27.0	63.0	0.9916	3.68	0.79	14
330	6	8.8	0.460	0.45	2.60	0.065	7.0	18.0	0.9947	3.32	0.79	14
788	6	5.0	0.400	0.50	4.30	0.046	29.0	80.0	0.9902	3.49	0.66	13.6
869	7	5.1	0.420	0.00	1.80	0.044	18.0	88.0	0.99157	3.68	0.73	13.6
1055	7	5.3	0.470	0.11	2.20	0.048	16.0	89.0	0.99182	3.54	0.88	13.6
800	7	7.4	0.360	0.34	1.80	0.075	18.0	38.0	0.9933	3.38	0.88	13.6
1053	7	5.3	0.470	0.11	2.20	0.048	16.0	89.0	0.99182	3.54	0.88	13.5667
322	8	11.3	0.620	0.67	5.20	0.086	6.0	19.0	0.9988	3.22	0.69	13.4
347	7	8.9	0.400	0.51	2.60	0.052	13.0	27.0	0.995	3.32	0.90	13.4
792	6	7.1	0.390	0.12	2.10	0.065	14.0	24.0	0.99252	3.30	0.53	13.3
346	7	9.2	0.410	0.50	2.50	0.055	12.0	25.0	0.9952	3.34	0.79	13.3
268	6	11.4	0.625	0.66	6.20	0.088	6.0	24.0	0.9988	3.11	0.99	13.3
644	6	9.3	0.380	0.48	3.80	0.132	3.0	11.0	0.99577	3.23	0.57	13.2
33	4	4.6	0.520	0.15	2.10	0.054	8.0	65.0	0.9934	3.90	0.56	13.1
794	8	7.9	0.540	0.34	2.50	0.076	8.0	17.0	0.99235	3.20	0.72	13.1
275	7	12.0	0.370	0.76	4.20	0.066	7.0	38.0	1.0004	3.22	0.60	13
92	5	5.6	0.500	0.09	2.30	0.049	17.0	99.0	0.9937	3.63	0.63	13
249	5	13.5	0.530	0.79	4.80	0.12	23.0	77.0	1.0018	3.18	0.77	13
702	6	6.4	0.690	0.00	1.65	0.055	7.0	12.0	0.99162	3.47	0.53	12.9
740	5	5.6	0.605	0.05	2.40	0.073	19.0	25.0	0.99258	3.56	0.55	12.9
793	5	5.6	0.660	0.00	2.50	0.066	7.0	15.0	0.99256	3.52	0.58	12.9
272	8	5.6	0.850	0.05	1.40	0.045	12.0	88.0	0.9924	3.56	0.82	12.9
691	6	7.3	0.520	0.32	2.10	0.07	51.0	70.0	0.99418	3.34	0.82	12.9
822	7	5.1	0.510	0.18	2.10	0.042	16.0	101.0	0.9924	3.46	0.87	12.9
578	7	5.1	0.585	0.00	1.70	0.044	14.0	86.0	0.99264	3.56	0.94	12.9
874	6	7.1	0.750	0.01	2.20	0.059	11.0	18.0	0.99242	3.39	0.40	12.8
876	6	7.1	0.750	0.01	2.20	0.059	11.0	18.0	0.99242	3.39	0.40	12.8
705	7	6.8	0.360	0.32	1.80	0.067	4.0	8.0	0.9928	3.36	0.55	12.8
931	6	6.5	0.510	0.15	3.00	0.064	12.0	27.0	0.9929	3.33	0.59	12.8
781	5	6.3	0.570	0.28	2.10	0.048	13.0	49.0	0.99374	3.41	0.60	12.8
251	6	6.7	0.750	0.01	2.40	0.078	17.0	32.0	0.9955	3.55	0.61	12.8
494	6	5.1	0.470	0.02	1.30	0.034	18.0	44.0	0.9921	3.90	0.62	12.8
378	6	10.3	0.270	0.24	2.10	0.072	15.0	33.0	0.9956	3.22	0.66	12.8
191	8	7.9	0.350	0.46	3.60	0.078	15.0	37.0	0.9973	3.35	0.86	12.8
715	6	8.0	0.180	0.37	0.90	0.049	36.0	109.0	0.99007	2.89	0.44	12.7
520	6	6.4	0.865	0.03	3.20	0.071	27.0	58.0	0.995	3.61	0.49	12.7
432	6	6.3	0.360	0.19	3.20	0.075	15.0	39.0	0.9956	3.56	0.52	12.7
401	6	13.0	0.470	0.49	4.30	0.085	6.0	47.0	1.0021	3.30	0.68	12.7
405	6	13.0	0.470	0.49	4.30	0.085	6.0	47.0	1.0021	3.30	0.68	12.7
741	7	8.3	0.330	0.42	2.30	0.07	9.0	20.0	0.99426	3.38	0.77	12.7
1049	5	6.7	0.700	0.08	3.75	0.067	8.0	16.0	0.99334	3.43	0.52	12.6
723	6	7.9	0.310	0.32	1.90	0.066	14.0	36.0	0.99364	3.41	0.56	12.6
949	6	6.6	0.580	0.02	2.40	0.069	19.0	40.0	0.99387	3.38	0.66	12.6

Figure 2 – DataTable

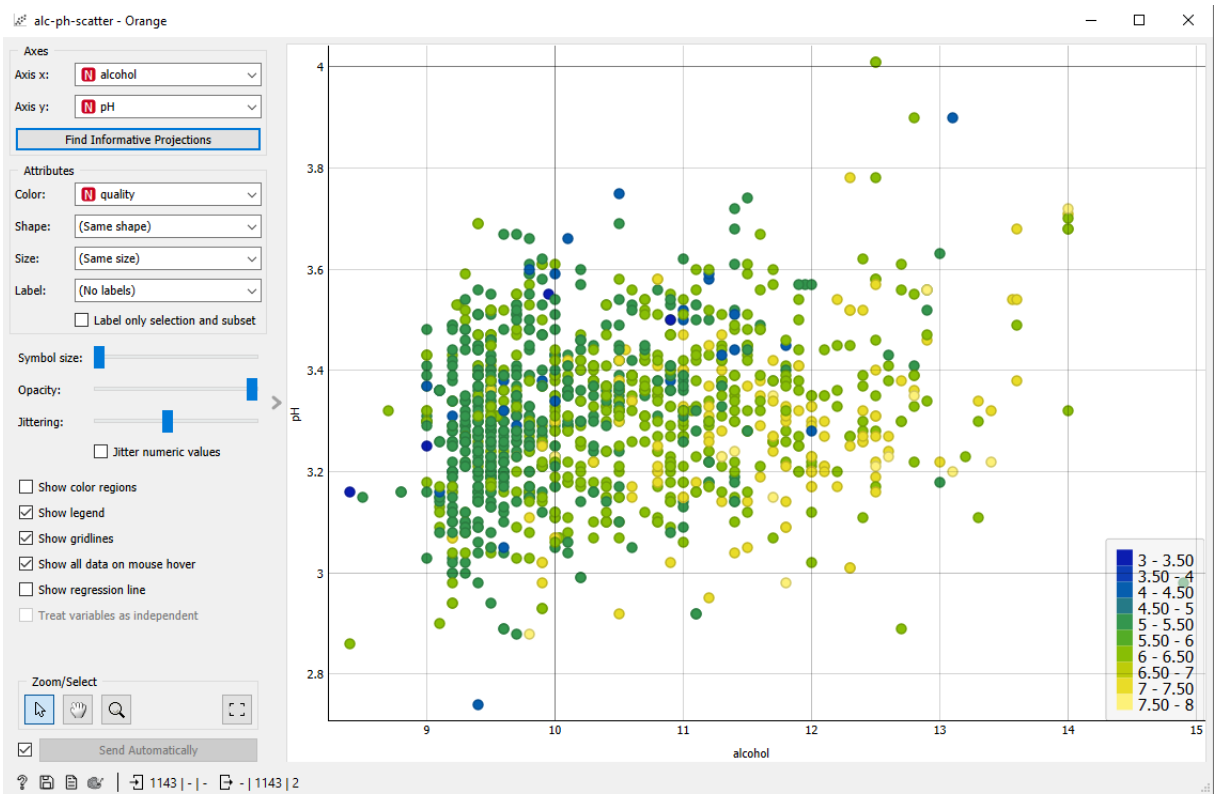


Figure 3 - Scatter - Alcohol & pH

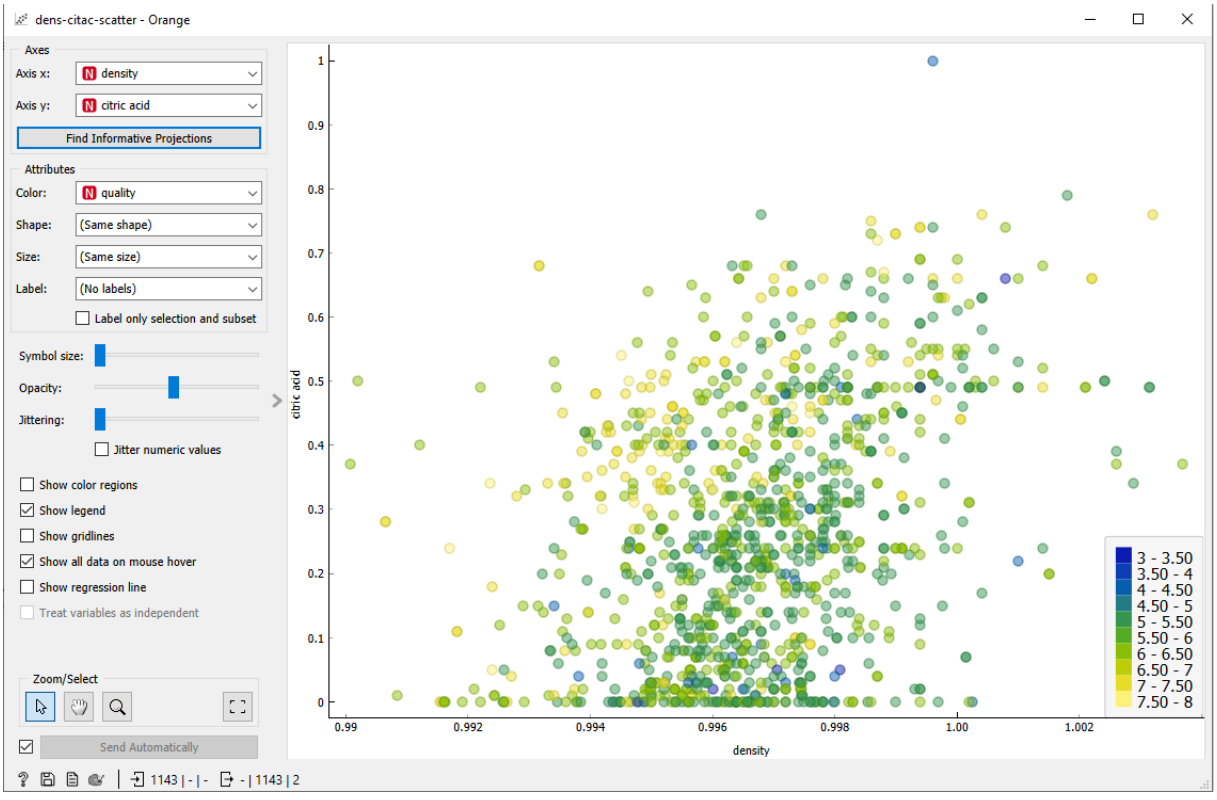


Figure 4 - Scatter - Density & Citric acid

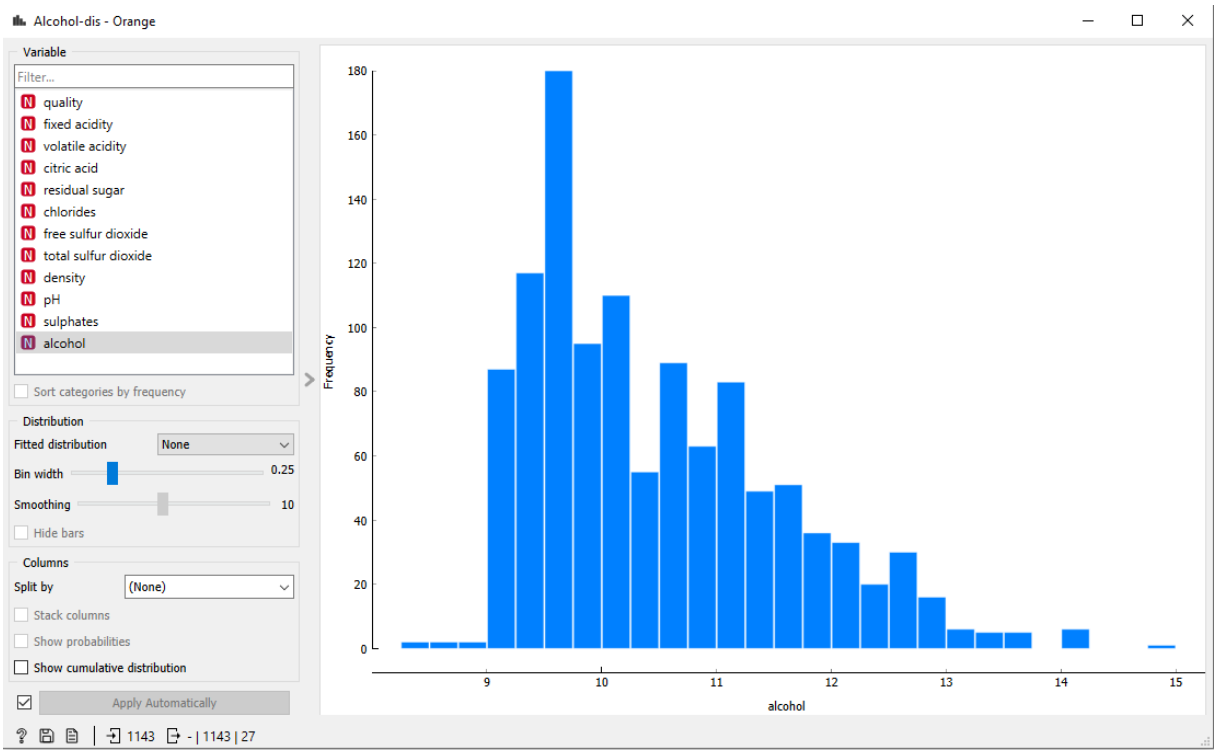


Figure 5 - Distribution – Alcohol

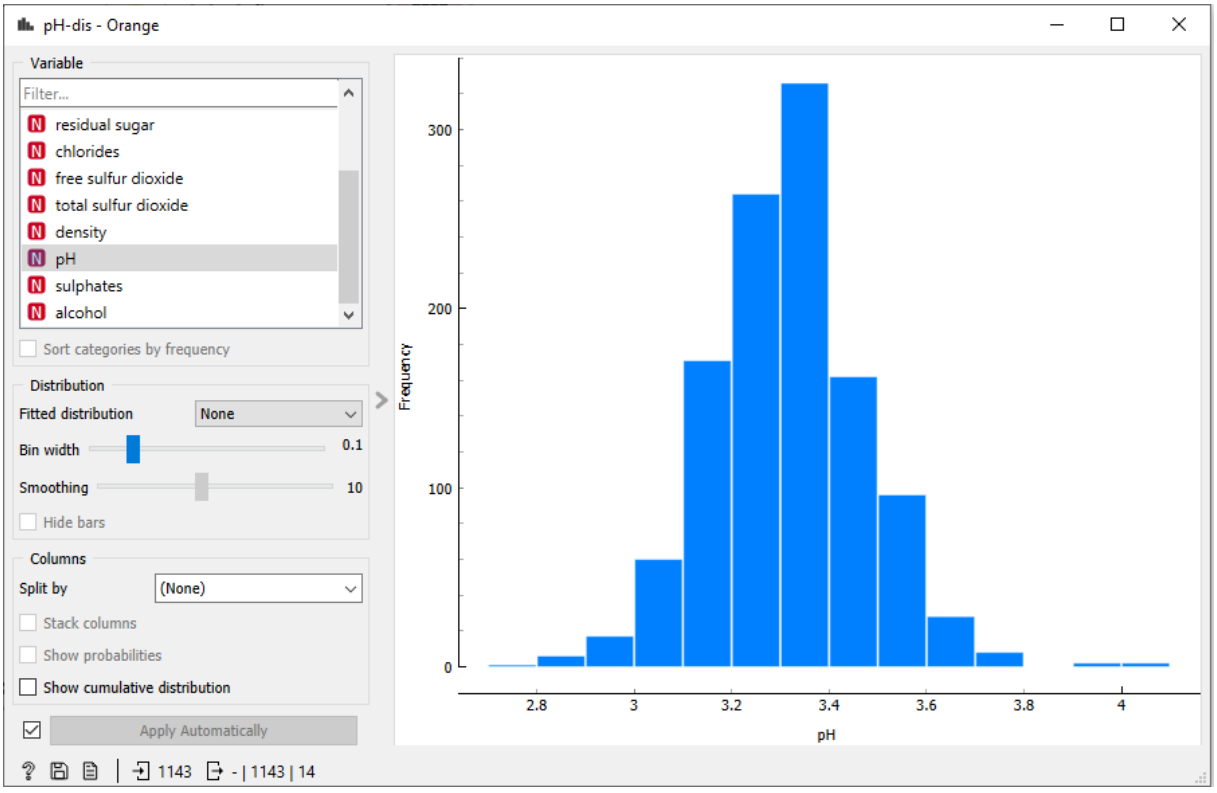


Figure 6 - Distribution – pH

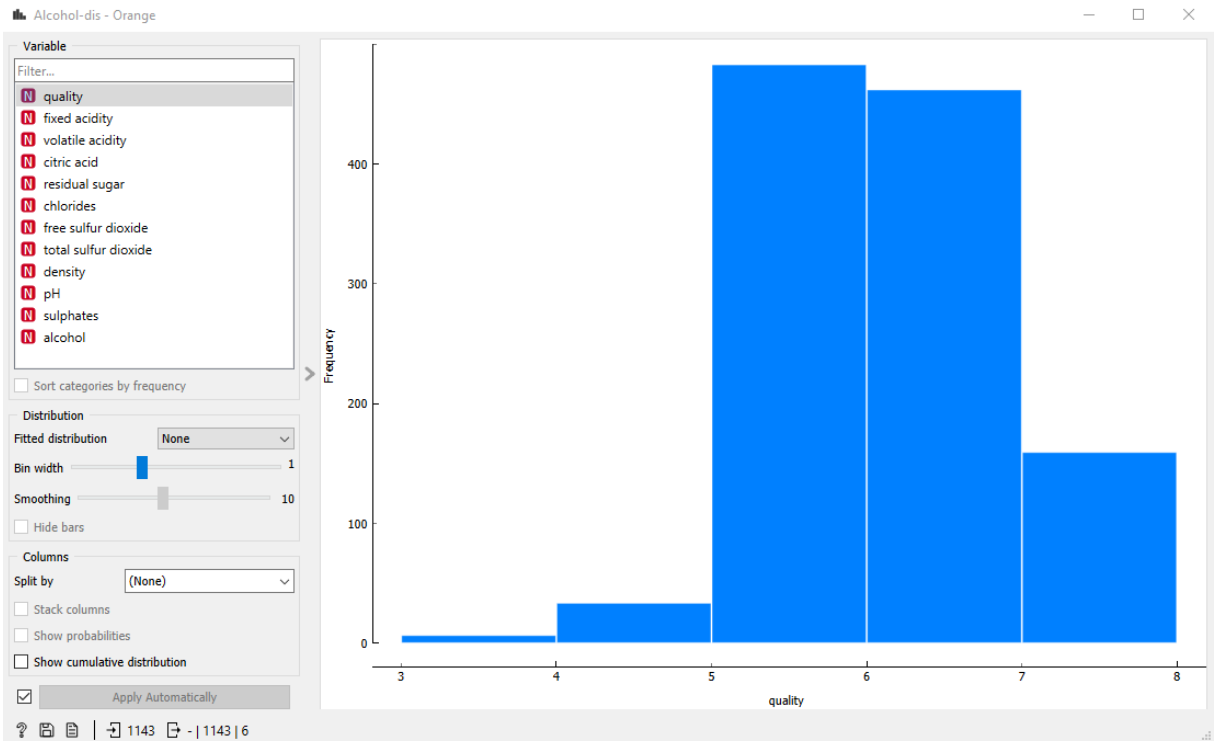


Figure 7 - Distribution showing not balanced classes.

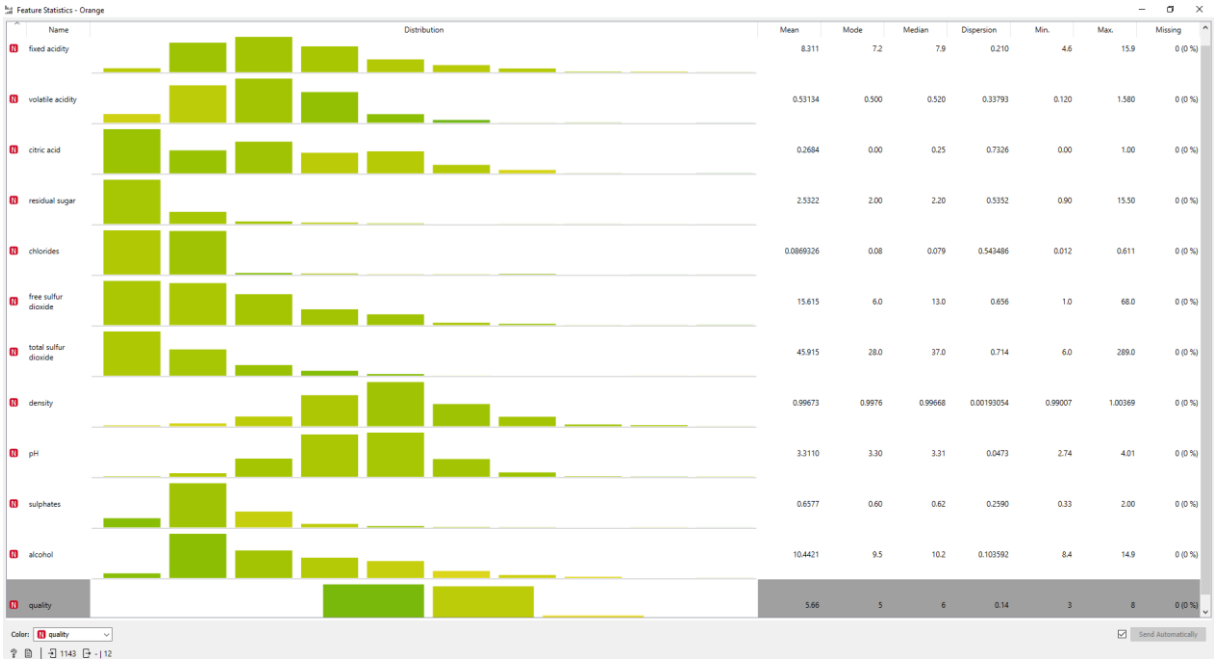


Figure 8 - Statistics of data

Conclusion of scatter plots, histograms and distributions:

From the scatterplots we can see that the quality of a wine often correlates with a higher citric acid and a lower density. Higher quality wine is colored in yellow, while lower quality wine is green. The histograms are showing us for example how many wines containing the different amounts of percentage alcohol. We can see that none of the plots above are balanced.

Conclusion coming from analysis of statistical calculations:

In the figure showing statistics of data above, we can see the distribution of the different features. The numbers on the right can easily be seen has a correlation with the plots on the left. For example in the “fixed acidity” at the top, we have a minimum at 8.4 and a maximum at 14.9 Looking at the plot we can also see that the mean is at 10.4, which is almost at the very left in the plot, which makes sense because the minimum is 8.4 which is a lot closer to 10.4 than 14.9.

Part 2:

K-Means:

The K-Means algorithm is a clustering algorithm used to group similar objects intro clusters based on their attributes. In Orange the algorithm works by partitioning the data into k clusters. It starts by randomly selecting points from the data as the initial cluster center. Then each point is assigned to the nearest center based on the distance from the center.

DBSCAN:

The DBSCAN algorithm is also a clustering algorithm. In Orange the algorithm works by defining a neighborhood around each data point based on distance. Points that fall in the neighborhood are considered part of the same cluster. If a point does not have enough neighbors within its distance, it is considered a outlier or noise.

Plots:

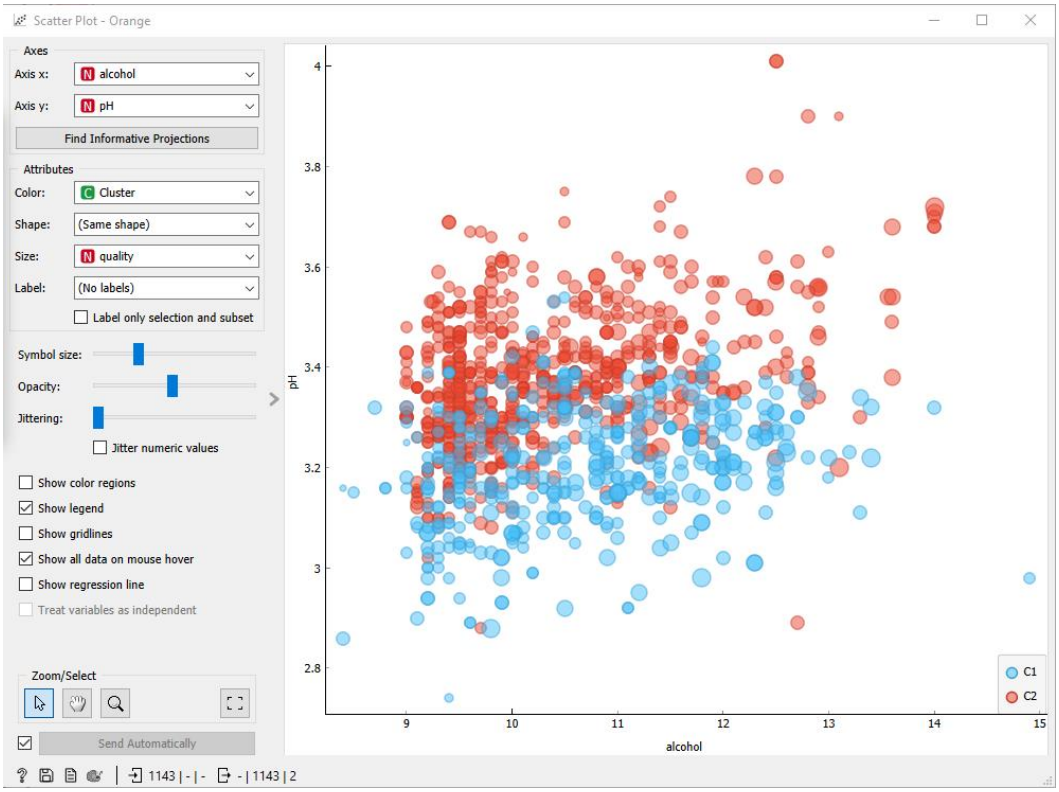


Figure 9 - K-Means - 2 k

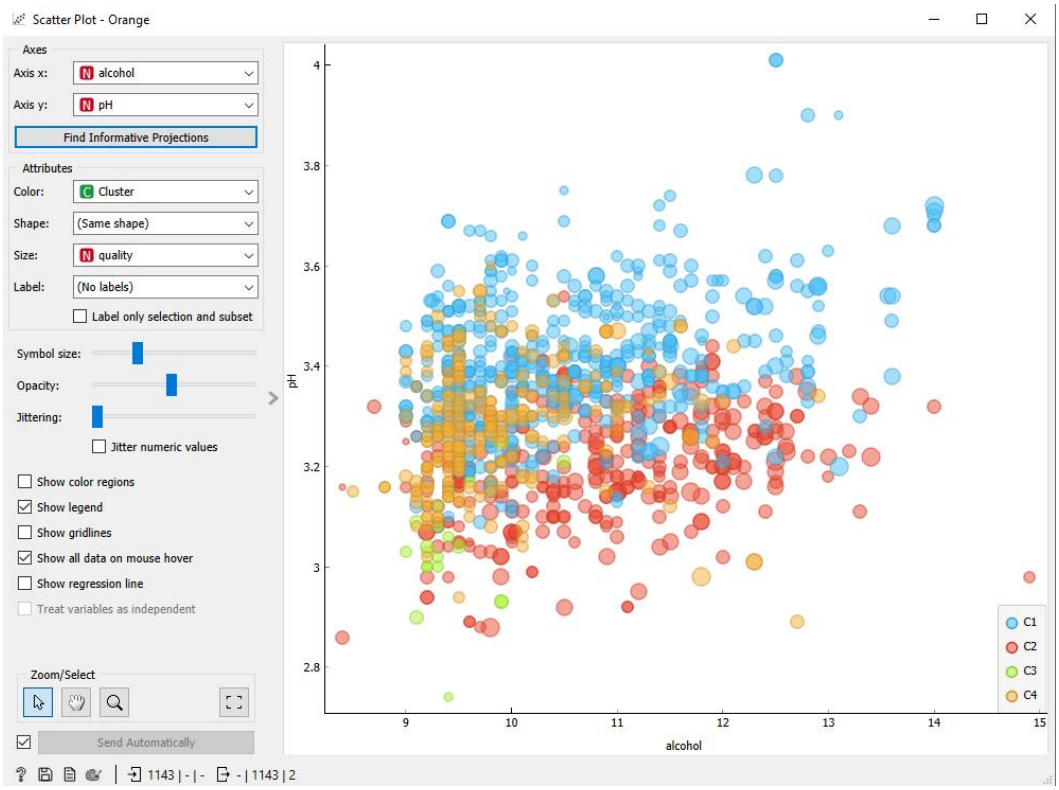


Figure 10- K-Means - 4 k

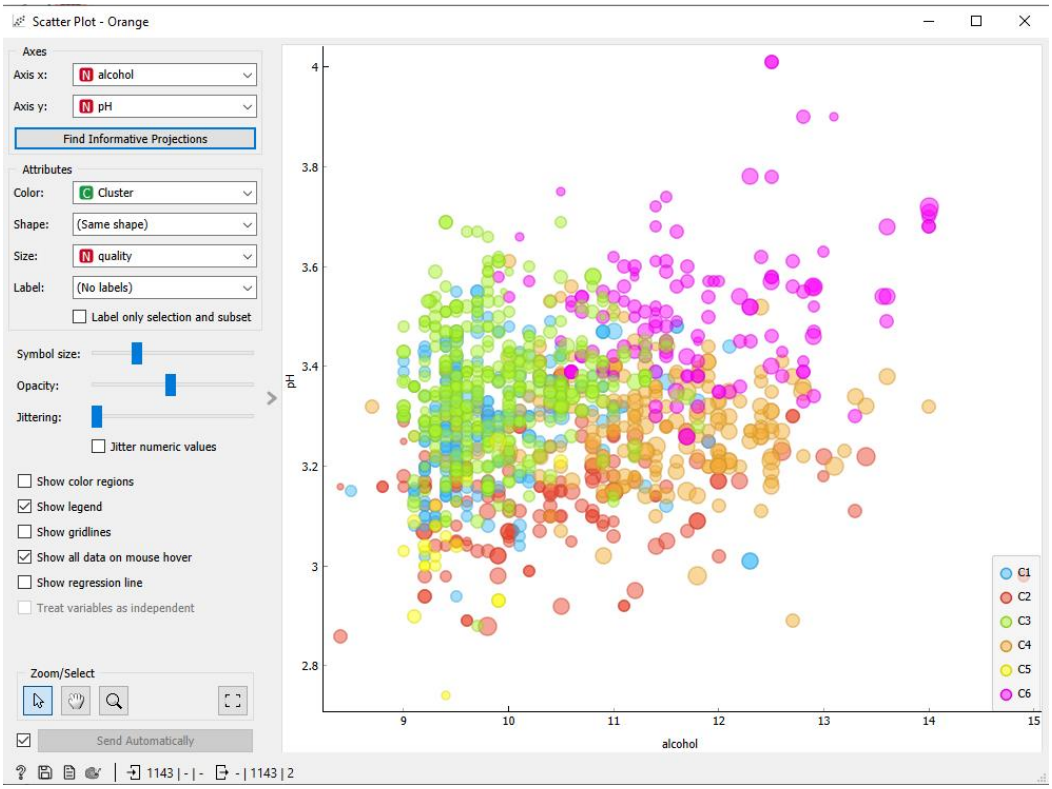


Figure 11- K-Means - 6 k

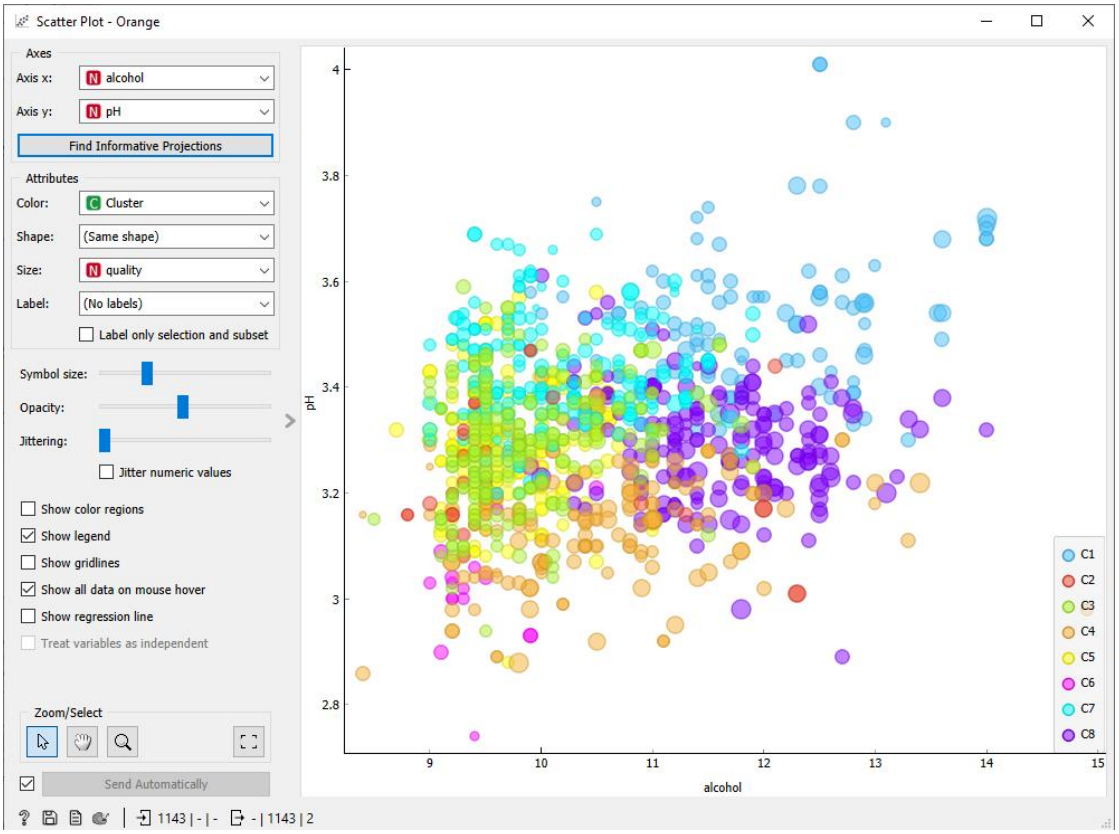


Figure 12- K-Means - 8 k

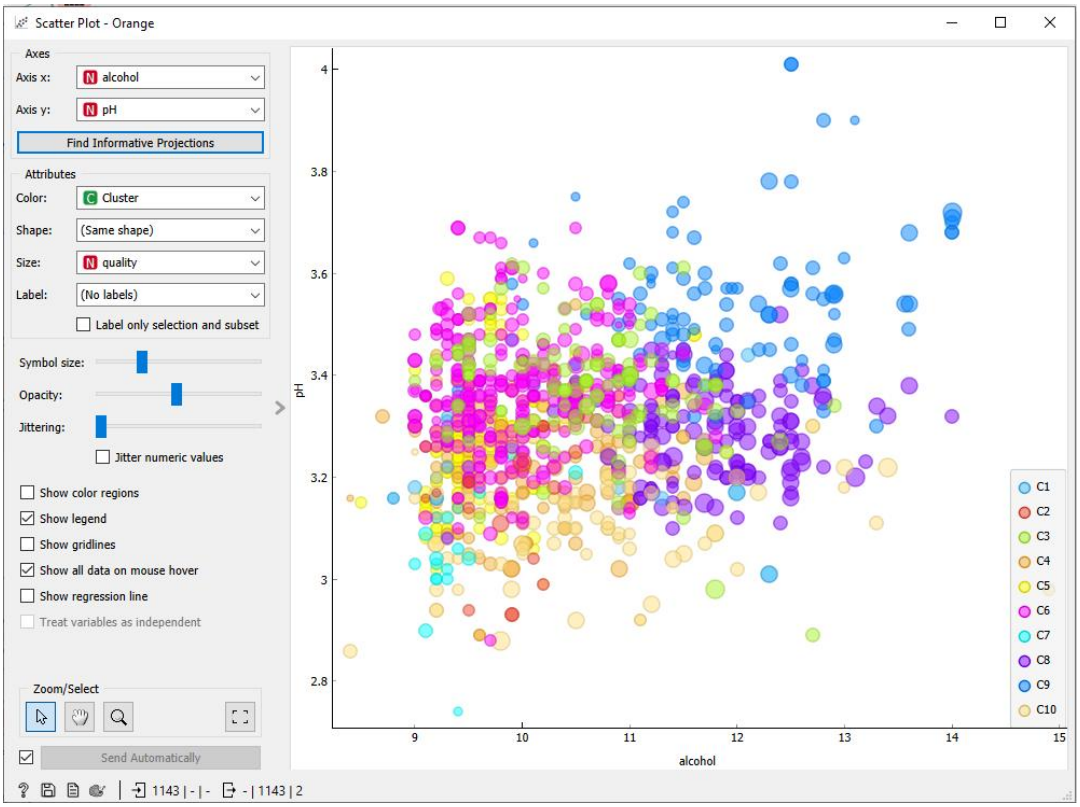


Figure 13- K-Means - 10 k

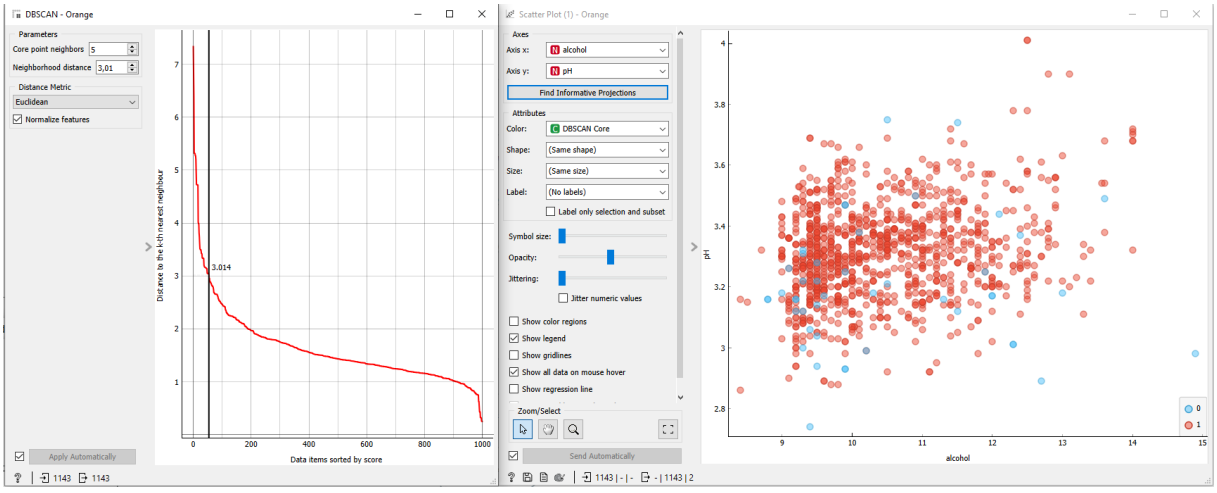


Figure 14 - DBSCAN - Core point Neighbors 5, distance 3

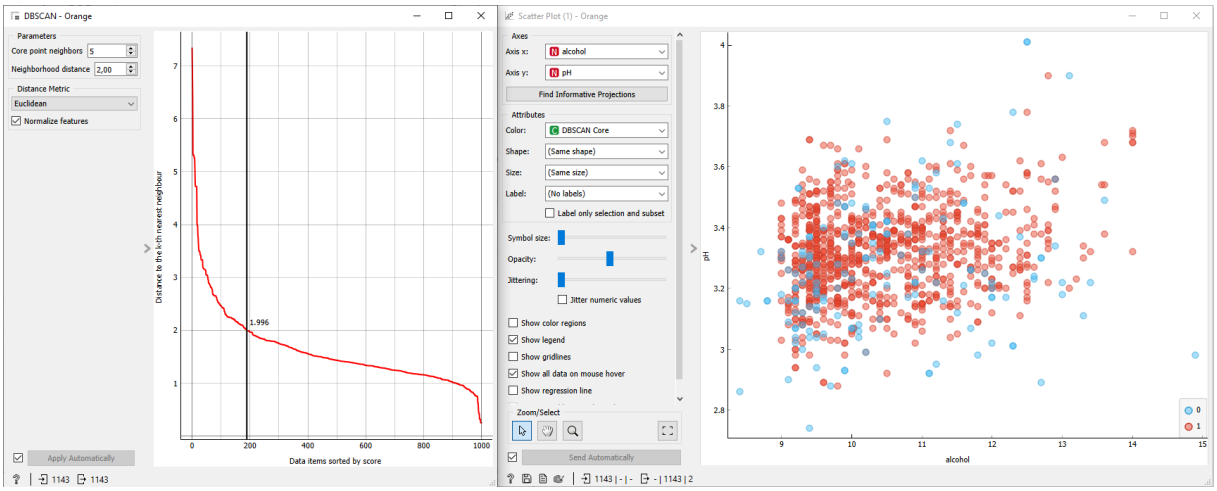


Figure 15- DBSCAN - Core point Neighbors 5, distance 2

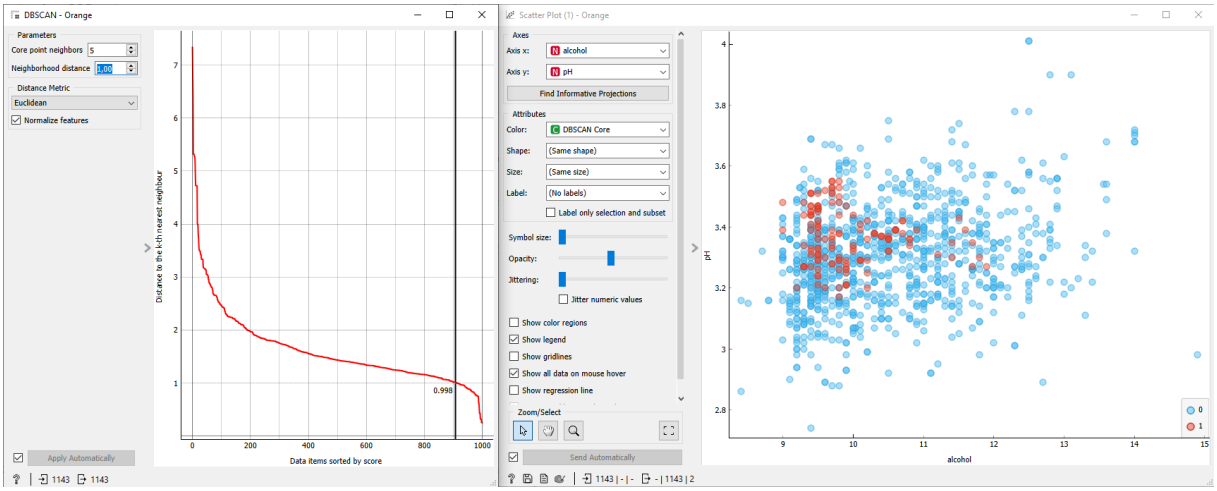


Figure 16 - DBSCAN - Core point Neighbors 5, distance 1

Conclusion:

In the K-Means plots it has been tried with different values, giving different plots. We can see that the colors represent clusters because the same colors are in thee same area. Using more k-values we can see more clusters appearing.

In the DBSCAN plots it has been tried with different distances, giving different plots. The smaller the neighborhood distance, the smaller the cluster becomes. The data is given a value of either 1 or 0, where 1 means its part of the neighborhood and 0 means it's an outlier/noise.

Part 3:

Performance:

Looking at the predictions of the wine quality it looks like kNN was the closest to guessing the real wine quality value.

Motivation:

Hyperparameters:

Table:

Algorithm	**Hyperparameter & Value**
Neural Network	Neurons in hidden layers: 100
	Activation: ReLu
Linear Regression	Ridge regression
	Alpha: 0.0001
kNN	Number of neighbors: 5
	Metric: Euclidean
	Weight: Distance

Plots:

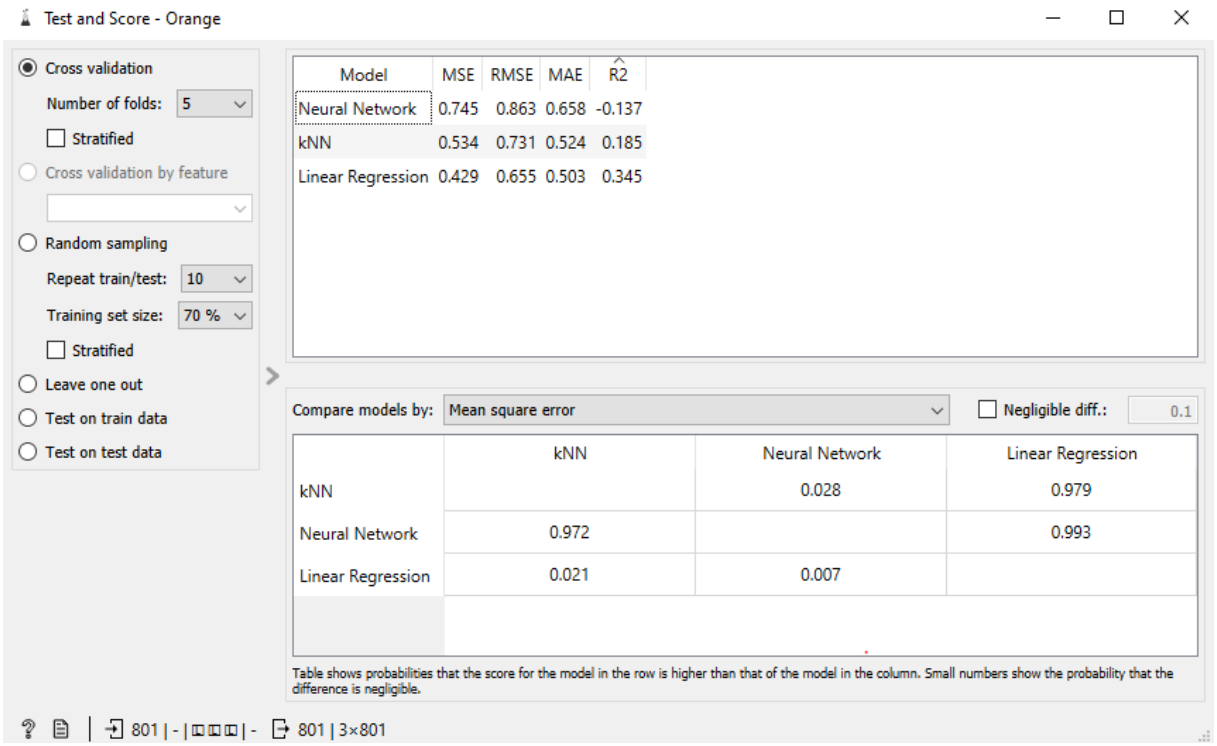


Figure 17 - Test and score results

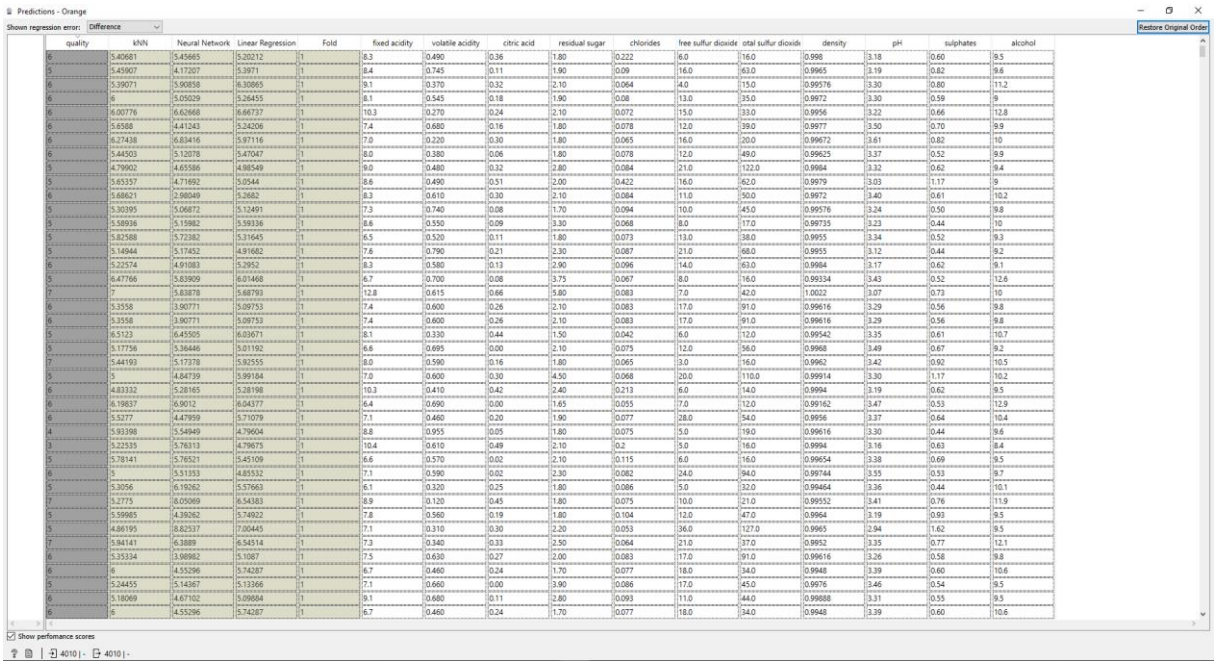


Figure 18 - Predictions

Conclusion:

I think this assignment was very hard and it was way too much to learn in Orange. Did my best but I must be honest and say that some stuff within this report is still confusing to me.

Part 3 – supervised learning, was extremely confusing which is why its as short as it is.

Sources:

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

https://www.youtube.com/@OrangeDataMining (Tutorials)

Different files from BlackBoard.

Entire Orange tool workflow: