

DEPARMTMENT OF ICT AND NATURAL SCIENCES

AUTOMATION AND INTELLIGENT SYSTEMS

AIS 2101 - INTELLIGENT SYSTEMS

Practical Assignment - Part D

Machine learning

Github Link to project and dataset:

<https://github.com/EliasWR/MachineLearning>

Written by:

Refsdal, Elias Woie

Supervisors:

Anohina-Naumeca, Alla

Singh, Vijander

March, 2022

Contents

List of Figures	ii
List of Tables	ii
1 Intro	1
2 Part 1	2
2.1 Feature description	2
2.2 Snippet of file	2
2.3 Dataset description	3
2.4 Scatter plots	3
2.4.1 1.5.a	3
2.5 1.5.b and 1.5.c Histograms and distribution.	4
2.6 1.5.d	4
2.7 Conclusion	5
3 Part 2 - Unsupervised learning	6
3.1 Hierarchical clustering	6
3.2 DBSCAN	7
4 Part 3 - Supervised Learning	10
4.1 kNN	10
4.1.1 Hyperparameters	10
4.2 Logistic Regression	10
4.2.1 Hyperparameters	11
4.3 Neural Network	11
4.3.1 Hyperparameters	11
4.4 Test and Training datasets	12
4.5 Experiments	12
4.5.1 Neural network	12
4.5.2 kNN	12
4.5.3 Logistic Regression	13
4.5.4 Trained data models	13
4.6 Conclusion	14
References	15

List of Figures

1	Overview of the Orange tool workflow.	1
2	Snippet of dataset	3
3	Scatter plots with the target, quality as a color attribute.	4
4	Histogram and distribution	4
5	Feature statistics	5
6	Example parameters for Hierarchical Clustering	6
7	Hierarchical clustering with adjusted height ratio.	7
8	DBSCAN with different hyperparameters.	8
9	DBSCAN chart	9
10	kNN example param	10
11	Logistic Regression example param	11
12	Neural network example param	11
13	Confusion matrices for the best fit of each model.	14

List of Tables

1	Feature description	2
2	Neural network parameters	12
3	Neural network scores	12
4	kNN parameters	13
5	kNN Scores	13
6	Logistic Regression Parameters	13
7	Logistic Regression scores	13
8	Best fits for each model	13

1 Intro

This task is written by a student at NTNU Ålesund.

This report is about utilizing machine learning on analyzing red wine.

Figure 1 show an overview of the workflow that was utilized in the Orange tool in order to get the results of this report.

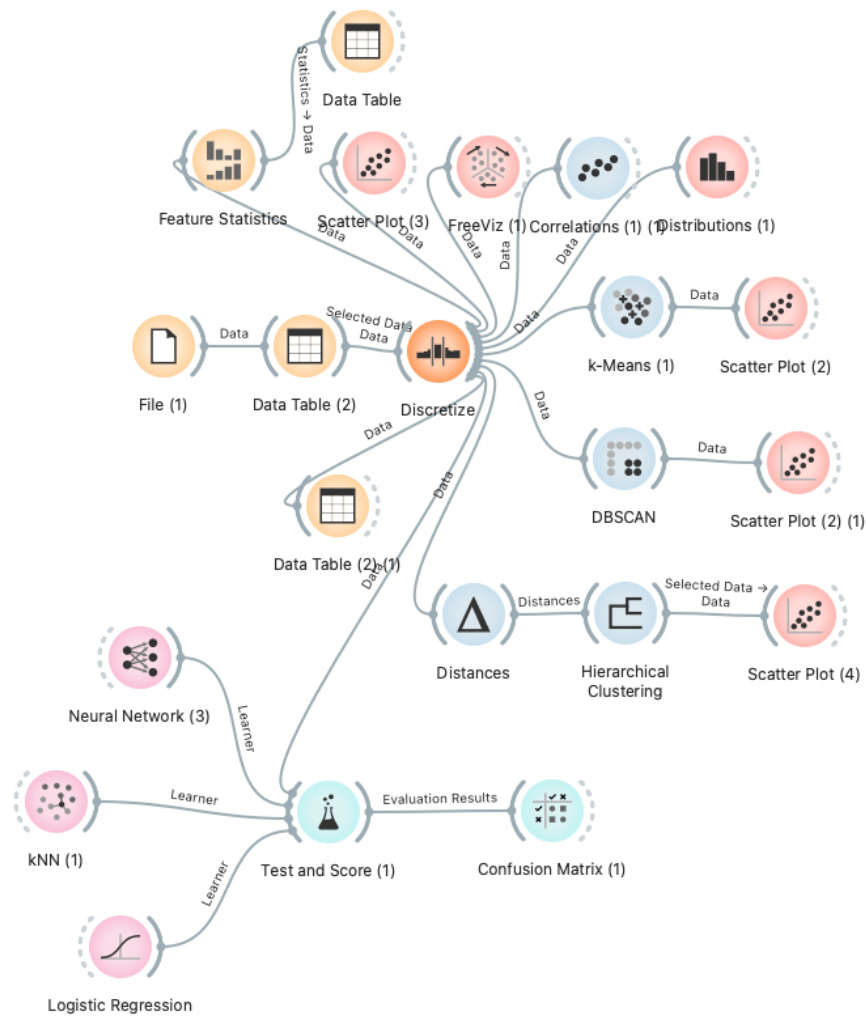


Figure 1: Overview of the Orange tool workflow.

2 Part 1

2.1 Feature description

Index	Feature	Value type	Value range
1	Fixed acidity	Numerical	4.6 - 15.9
2	Volatile acidity	Numerical	0.12 - 1.58
3	Citric acid	Numerical	0 - 1
4	Residual sugar	Numerical	0.9 - 15.5
5	Chlorides	Numerical	0.012 - 0.611
6	Free sulfur dioxide	Numerical	1 - 72
7	Total sulfur dioxide	Numerical	6 - 289
8	Density	Numerical	0.99007 - 1.00369
9	pH	Numerical	2.74 - 4.01
10	Sulphates	Numerical	0.33 - 2
11	Alcohol	Numerical	8.4 - 14.9

Table 1: Feature description

1. Fixed acidity refers to a wines acids that do not evaporate easily. This helps preserve the wine and it contributes to a sour taste.
2. Volatile acidity refers to a wines acids that evaporate easily. This contributes to the sharp and vinegary aroma of the wine.
3. Citric acid is one of the fixed acids. It contributes to the freshness and acidity of the wine.
4. Residual sugar represent the amount of natural grape sugars of the wine. The residual sugar can sweeten and balance out the acidity.
5. Chlorides refers to the salt composure. Chlorides give the wine a salty taste.
6. Free sulfur represent the amount of individual SO_2 molecules. An exaggerated use of SO_2 may result in a unpleasant taste or smell.
7. Total sulfur represent the total amount of SO_2 molecules. Sulfur dioxide is used for preserving and preventing oxydation of the wine.
8. The density refers to the $\frac{Weight}{Volume}$ ratio of the wine. The density can indicate the sweetness and richness of the wine.
9. pH measure the wines acidity or basicity between 0 and 14. The pH is carefully regulated during the making of the wine to ensure the wines quality.
10. Sulphates measure the level of sulfites which is used for wine preservation. Even though too much sulfites can harm the wines quality a moderate amount is neccessary for ensuring the wines stability and quality.
11. Alcohol refers to the ethanol that is produced while fermenting the wine. A higher alcohol can affect a wines taste considerably and higher alcohol generally give a more intense flavor.

2.2 Snippet of file

The below figure 2 show a snippet of the wine quality dataset before aggregation. There are a total of 1599 rows in the dataset and one target column with two classes, and 11 feature columns. The target of wine quality is divided into two classes of lower and higher quality wine. The wine is of higher quality if the quality is graded 6 or above, and anything below is considered lower quality.

	quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
1	5	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
2	5	7.8	0.880	0.00	2.60	0.098	25.0	67.0	0.9968	3.20	0.68	9.8
3	5	7.8	0.760	0.04	2.30	0.092	15.0	54.0	0.997	3.26	0.65	9.8
4	6	11.2	0.280	0.56	1.90	0.075	17.0	60.0	0.998	3.16	0.58	9.8
5	5	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
6	5	7.4	0.660	0.00	1.80	0.075	13.0	40.0	0.9978	3.51	0.56	9.4
7	5	7.9	0.600	0.06	1.60	0.069	15.0	59.0	0.9964	3.30	0.46	9.4
8	7	7.3	0.650	0.00	1.20	0.065	15.0	21.0	0.9946	3.39	0.47	10
9	7	7.8	0.580	0.02	2.00	0.073	9.0	18.0	0.9968	3.36	0.57	9.5
10	5	7.5	0.500	0.36	6.10	0.071	17.0	102.0	0.9978	3.35	0.80	10.5

Figure 2: Snippet of dataset

2.3 Dataset description

”Red Wine Quality” was retrieved from Kaggle.com (LEARNING 2023) and it is written by UCI MACHINE LEARNING. The license can be found in this reference Foundation 2023.

This dataset can separate good and bad wines through its properties. This data set was stumbled upon when searching through Kaggle and it was selected due to its detailed data on a variety of measurable attributes such as density and pH. It checks all the boxes for a machine learning data set. All the values are numerical and can be categorized and targeted after their quality score which is between 1 and 10. I have extrapolated the target to only contain two categories of above and below a quality of 6. When sorting the columns after their target category there are 744 wines that are considered lower quality (Below 6) and 855 wines considered higher quality (Equal to and above 6).

There is one downside to the dataset, the clusters are close to one another and it can therefore be challenging for supervised learning algorithms to determine affiliation to classes.

The values of the data set was of a desirable format and the dataset was complete and no imputing was required. Additional factors listed in the task regarding the dataset were not relevant.

The data set consists of 1599 data points and twelve columns. This includes the target value of quality and 11 additional features.

2.4 Scatter plots

2.4.1 1.5.a

Even though the task said to avoid using the target as a variable of the plot it was included as a colour attribute in order to emphasize the scattering of the classes.

Figure 3a show the data represented with alcohol on the X axis, sulphates on the Y axis and each data point is coloured with their quality. From the plot it can be concluded that very few wines of worse quality (Below 6) have an alcohol percentage above 11.5 %. The sulphate level of the wine do not have a strong correlation to the quality. The background color show that wines of higher quality tend to have a higher alcohol percentage.

Figure 3b show the data represented with fixed acidity on the X axis, pH on the Y axis and each data point is coloured with their quality. The plot show a clear negative correlation of 68 % which means that the values of fixed acidity and pH is closely linked. This is sensible due to pH being a measure for acidity and basicity. From the coloured background it can be interpreted that wines of higher quality tend to maintain a fixed acidity while also having a low pH.

Even though the plots show data groupings they are relatively close to one another. On the other it is likely sufficient for separating the classes.

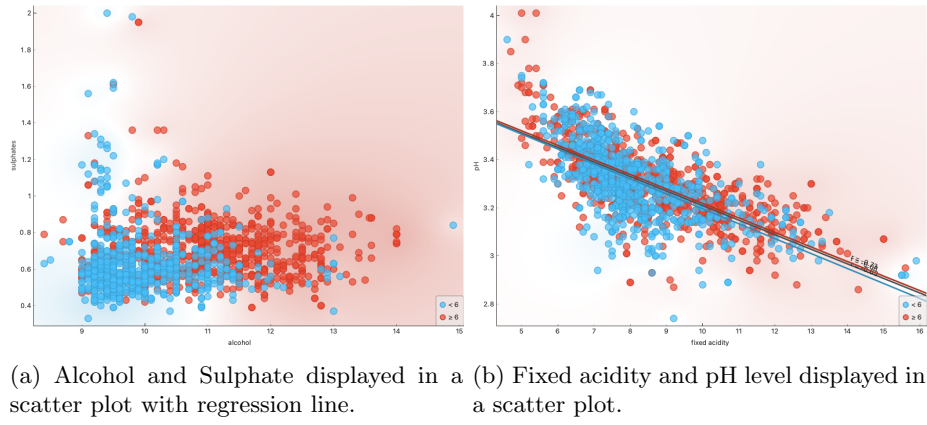


Figure 3: Scatter plots with the target, quality as a color attribute.

2.5 1.5.b and 1.5.c Histograms and distribution.

The features of wine density and pH is displayed in histograms and distribution shown in figure 4. These figures show that the pH has a center of gravity at about 3.3 and most of the wines have a pH more than 3 and less than 3.7. It is interesting to note that the distributions are congruent which means that the pH is not directly related to the quality of the wine. The wines have a center of gravity at about 0.9967 and most of the wines have a density higher than 0.992 and less than 1.001. The distributions show that the wines of lower quality have a lower standard deviation, and thereby dispersion, in it's distribution.

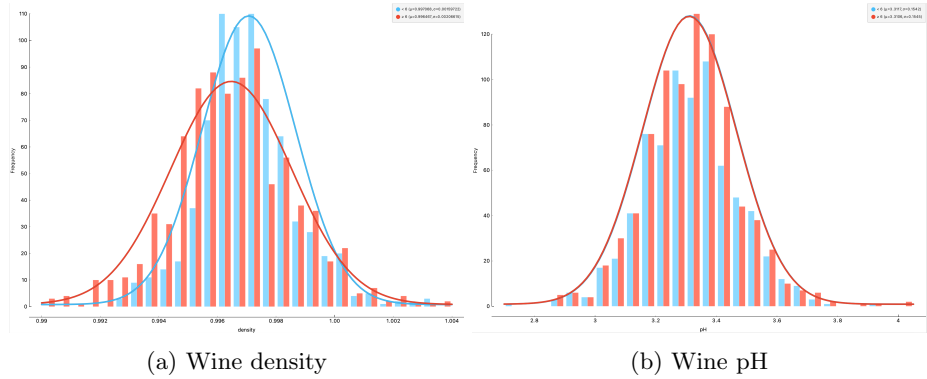


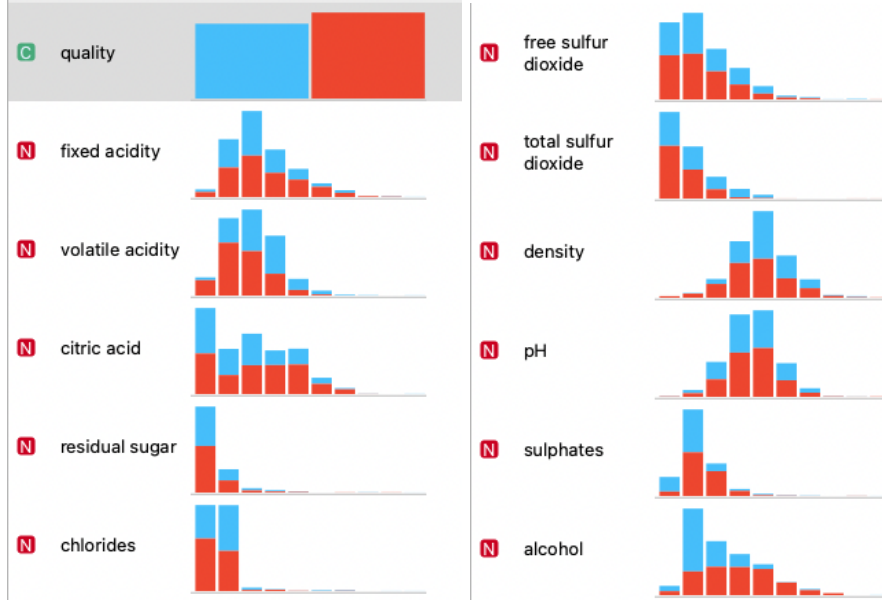
Figure 4: Histogram and distribution

2.6 1.5.d

Figure 5 show the displayed statistics of all the features and the bar charts are coloured after the target classes. The chart reflect the earlier bar charts when compared. It is noteworthy that the median is in the lower $\frac{1}{3}$ of the number range in 9 out of 11 features. This comes to show that the features tend to center around lower values but can reach significantly high values.

	Feature	Mode	Mean	Median	Dispersion	Min.	Max.	Missing
1	fixed acidity	7.2	8.31964	7.9	0.20921	4.6	15.9	0
2	volatile acidity	0.600	0.527821	0.52	0.339137	0.12	1.58	0
3	citric acid	0.00	0.270976	0.26	0.718663	0	1	0
4	residual sugar	2.00	2.53881	2.2	0.555177	0.9	15.5	0
5	chlorides	0.08	0.0874665	0.079	0.537927	0.012	0.611	0
6	free sulfur di...	6.0	15.8749	14	0.658705	1	72	0
7	total sulfur d...	28.0	46.4678	38	0.707695	6	289	0
8	density	0.9972	0.996747	0.99675	0.0018929	0.99007	1.00369	0
9	pH	3.30	3.31111	3.31	0.0466122	2.74	4.01	0
10	sulphates	0.60	0.658149	0.62	0.257471	0.33	2	0
11	alcohol	9.5	10.423	10.2	0.10221	8.4	14.9	0
12	quality	5	5.63602	6	0.143242	3	8	0

(a) Feature specifics



(b) Feature statistics 1

(c) Feature statistics 2

Figure 5: Feature statistics

2.7 Conclusion

The data groupings of the dataset are quite close to one another, but they are still separable and they show clear trends. It would be desirable for the data to be even more separated but the data should be classifiable even though the algorithms may classify the datapoints with a little lower accuracy than desired. Overall this dataset is most certainly suitable for machine learning tasks.

3 Part 2 - Unsupervised learning

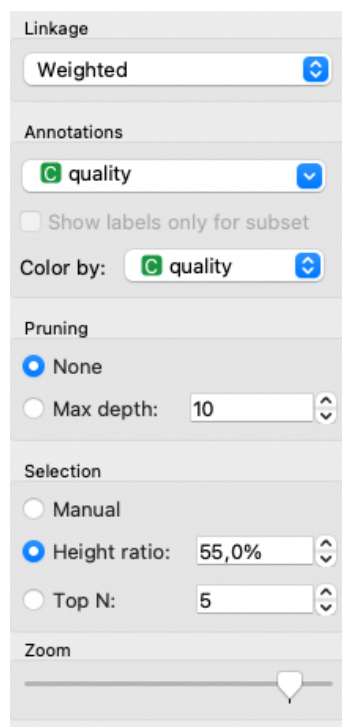
3.1 Hierarchical clustering

For the hierarchical clustering the values were displayed with the same scatterplot axis' of citric acid and fixed acidity with different height ratio for the Hierarchical clustering. The following height ratios were tried out on the plot: 40%, 45% and 55.9 %. Higher or lower height ratios were not informative and inane and the plots in figure 7 were found through trial and error. The linkage of use was Weighted due to the clear distinction of the results.

Figure 7 show three plots with a clear division between the clusters even though they overlap. The division is emphasized by the background colors and the symbol of the datapoints show what class they belong to. Each of the big clusters are clearly overrepresented by one of the classes (either dots or X-marks).

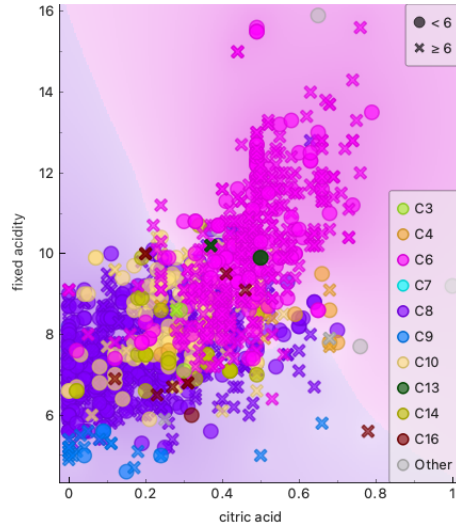
Hyperparameters

- "Linkage" refers to how the algorithm calculates the distance between data points. Different linkages output different clusters.
- "Annotations" refers to how the nodes in the dendrogram are labeled. It also allows for the option "show labels only for subsets" and "color by" which changes how the data is coloured and labeled when it is presented.
- "Pruning" is used for huge dendrograms and only affect the presentation of the dataset and it does not limit the actual clustering.
- "Selection" can be altered in order to change the level of the clustering. It can be chosen in percent, number of clusters or manually.

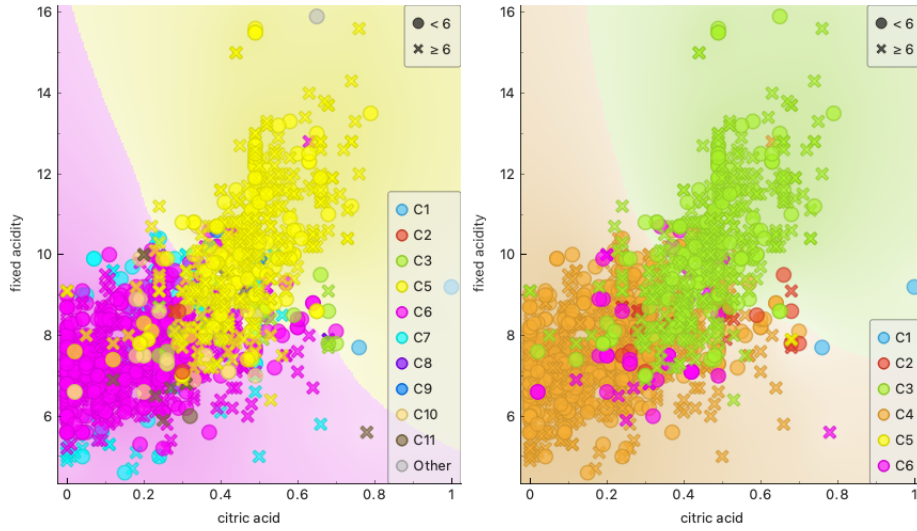


The image shows a vertical panel of controls for hierarchical clustering. It is organized into several sections: 'Linkage' with a dropdown set to 'Weighted'; 'Annotations' with a dropdown set to 'quality', a checkbox for 'Show labels only for subset' which is unchecked, and a 'Color by:' dropdown also set to 'quality'; 'Pruning' with radio buttons for 'None' (selected) and 'Max depth' (set to 10); 'Selection' with radio buttons for 'Manual', 'Height ratio' (set to 55,0% and selected), and 'Top N' (set to 5); and a 'Zoom' section at the bottom with a horizontal slider.

Figure 6: Example parameters for Hierarchical Clustering



(a) Hierarchical clustering at 40% height ratio.



(b) Hierarchical clustering at 45% height ratio. (c) Hierarchical clustering at 55.9% height ratio.

Figure 7: Hierarchical clustering with adjusted height ratio.

3.2 DBSCAN

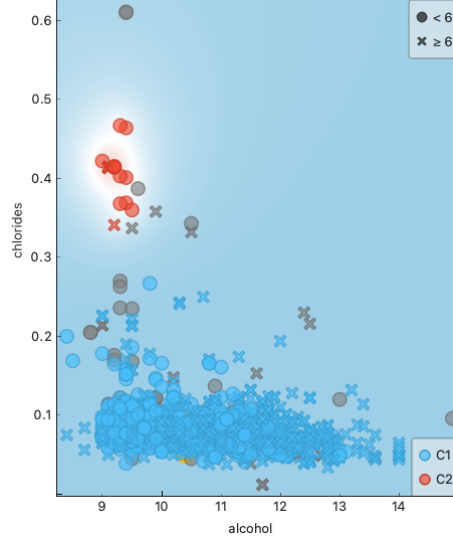
When using DBSCAN I discovered that the parameters had a very fine line where they displayed useful results. The DBSCAN was used for finding outlying clustering in the scatter plots which were highlighted by colouring the clusters. The neighborhood distance was kept at the leftmost break of the chart in figure 9 in order to get sensible results. I decided to try one of each of the three distance metrics, Euclidian, Manhattan and Cosine for each plot.

Figure 8a show an outlying cluster with an alcohol percentage of 9.3 and chloride of 0.4. The wines located in this cluster tend to have lower quality and it is hard to determine the cause of this cluster. But high chloride provide the wine with a salty taste and too much would not be desirable for the sake of the wines taste.

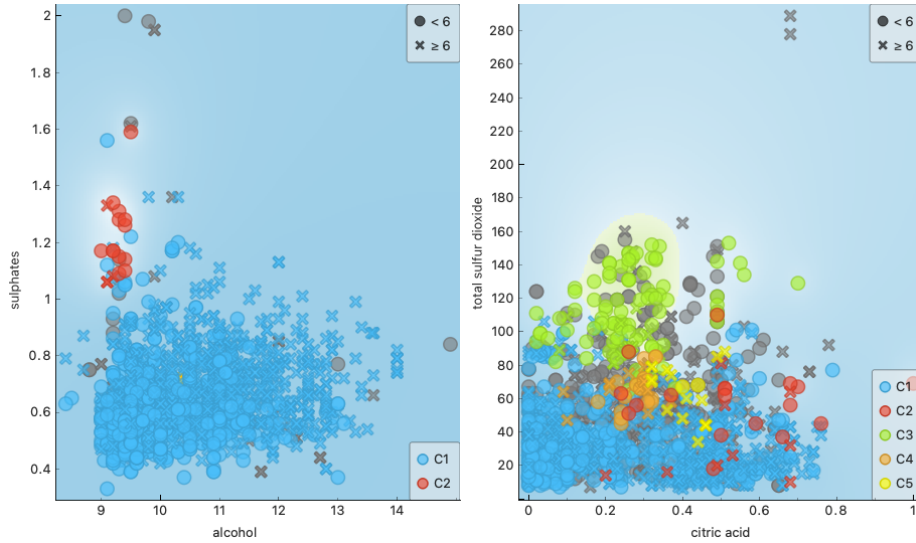
Figure 8b show an outlying cluster in the range 1-1.4 of Sulphates and 9 % of alcohol. This cluster is also over-represented by wines of lower quality as too much Sulphate can harm the quality of the wine. Even though there are wines to the right in the plot of higher quality and an equal level of sulphate they also compensate with a higher alcohol percentage.

Figure 8c was harder to find as the parameters were drastically different when the cosine distance

metric. There is one cluster that stands out in this plot and it is the green data points to the left of the center. This cluster has a total sulfur dioxide of approximately 140 and a citric acid of approximately 0.25. The cluster is also over-represented by wines of lower quality. This could be because too much Sulfur dioxide give the wine an unpleasant taste or smell. It is safe to say that the outlying clusters of the dataset likely are overrepresented by wines of lower quality.



(a) Chloride and alcohol DBSCAN with 10 core point neighbours, 2.51 neighbouring distance and Euclidian distance metric.



(b) Sulphates and Alcohol DBSCAN with 10 core point neighbours, 2.75 neighbouring distance and Manhattan distance metric. (c) Tot Sulfur dioxide and Citric acid DBSCAN with 20 core point neighbours, 0.15 neighbouring distance and Cosine distance metric.

Figure 8: DBSCAN with different hyperparameters.

Hyperparameters

- "Core point neighbours" determine the neighbour points within a specified radius around a core. This limiting factor is used for determining whether data points are part of the same

cluster or not.

- "Neighbourhood distance" sets a maximum limit distance of which two points can be considered neighbours.
- "Distance metric" describes the method used for calculating the distance between the distance between two data points.
- "Normalize features" means that the values are scaled to the same value range.

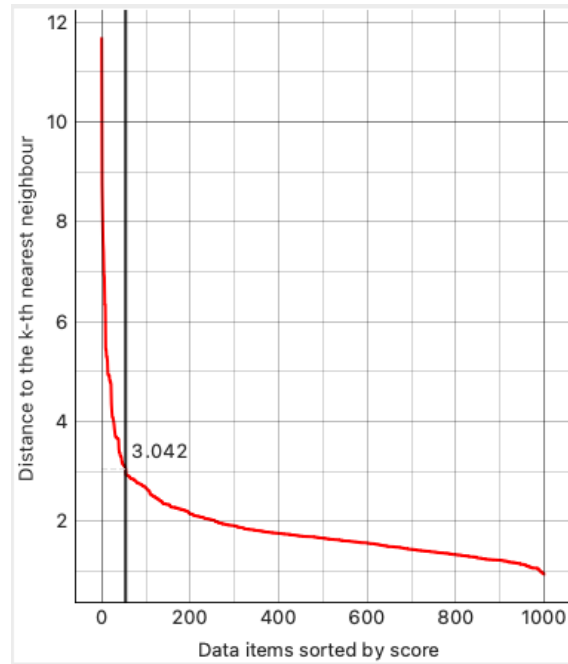


Figure 9: DBSCAN chart

4 Part 3 - Supervised Learning

4.1 kNN

k-nearest neighbours, hereby kNN, utilize a subset of the training data to make predictions. When given a new data point, the algorithm calculates the kNN based on distance calculated through either Euclidian, Manhattan, Chebyshev or Mahalanobis metrics. The algorithm then makes a prediction for the class or value for the new point. The prediction is usually calculated through an average of the kNN. Larger k-values tend to provide a smoother but more biased result while lower k-values tend to increase variance.

I chose to use kNN because this is one of the algorithm we focused on in lectures. Because of the simplicity of it's algorithm I think it was easy to understand how it works. It showed relatively good results after only a few tries.

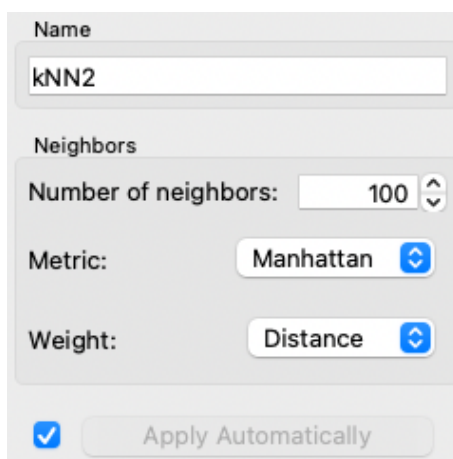


Figure 10: kNN example param

4.1.1 Hyperparameters

- "Number of neighbours" refers to the number of the nearest neighbours that are taken into consideration when calculating the prediction.
- "Metric" determine the way distance between data points are calculated.
- "Weight" refers to the way the algorithm determine the class of a given data point based on it's neighbours. It can either be done through distance or uniform weight.

4.2 Logistic Regression

Logistic Regression is often used for binary classification tasks with for example classification of two classes. The model is based on linearity while it utilize a logistic function to model the output probability. The logistic function is also known as the sigmoid function and this allows the algorithm to map the values to a probability between 0 and 1 of werther the data point is a value of a certain class.

I chose this model because it is known for being especially suitable for binary classification and my problem has exactly two classes in the target.

Name: LR3

Regularization type: Ridge (L2)

Strength: Weak ————— Strong
C=1000

☐ Balance class distribution

☒ Apply Automatically

Figure 11: Logistic Regression example param

4.2.1 Hyperparameters

- "Regularization type" is used for preventing overfitting of a model by applying a punishment to the cost function that is being minimized.
- "Strength" is the factor of which the algorithm affects the models output.
- "Balance class distribution" is a boolean hyperparameter that set the number of instances of each class to an approximately equal amount. It is useful when the classes are imbalanced.

4.3 Neural Network

Neural Network (3)

Name: NN1

Neurons in hidden layers: 100,100

Activation: ReLU

Solver: Adam

Regularization, $\alpha=0.0002$: —————

Maximal number of iterations: 200

☒ Replicable training

Cancel ☒ Apply Automatically

Figure 12: Neural network example param

4.3.1 Hyperparameters

- "Neurons in hidden layers" refers to the number of neurons for each layer of the network that connect the input to the output of the Neural Network. Each layers number of neurons can be set with a number and separated with commas.
- "Activation" decide the function that is applied to the output of the neurons in order to introduce non-linearity to the network.

-
- "Solver" determine the optimization algorithm that is used for finding the best weights or parameters of the neural network model.
 - "Regularization" is used for preventing over-fitting of the model to the training data. This means that the model becomes too complex and use too much noise rather than the pattern for training.
 - "Maximum number of iterations" limits the number of times the training algorithm updates the parameters of the model.
 - "Replicable training" is either true or false and it determine wether the algorithm provide the same result every time or new ones.

4.4 Test and Training datasets

The algorithms were trained by 70% of training data which include 521 data points of lower quality wine and 599 data points of higher quality wine. This means that the testing is done by 30% of the data points which equals 223 wines of lower quality and 256 wines of higher quality. The testing and training is repeated 10 times and the data was chosen by random picking. The data was stratified and divided into their respective classes.

4.5 Experiments

4.5.1 Neural network

Parameters

Name	Neurons in hidden layers	Activation	Solver	Regularization	Max Num of Iterations
NN1	100,100	ReLu	Adam	0.05	200
NN2	100,100	Logistic	Adam	0.05	1000
NN3	20,20,20,20,20	tanh	Adam	0.05	1000

Table 2: Neural network parameters

Score

Name	Area Under Curve	Classification accuracy	F1	Precision	Recall
NN1	0.837	0.778	0.778	0.778	0.778
NN2	0.817	0.742	0.742	0.742	0.742
NN3	0.795	0.756	0.756	0.756	0.756

Table 3: Neural network scores

4.5.2 kNN

Parameters

Name	Num of neighbours	Metric	Weight
kNN1	10	Euclidian	Distance
kNN2	100	Manhattan	Distance
kNN3	50	Chebyshev	Distance

Table 4: kNN parameters

Score

Model	Area Under Curve	Classification accuracy	F1	Precision	Recall
kNN1	0.813	0.714	0.714	0.714	0.714
kNN2	0.842	0.746	0.745	0.747	0.746
kNN3	0.814	0.723	0.722	0.724	0.723

Table 5: kNN Scores

4.5.3 Logistic Regression

Parameters

Name	Regularization type	Strength
LR1	Ridge(L2)	1000
LR2	Lasso(L1)	1
LR3	None	N/A

Table 6: Logistic Regression Parameters

Score

Model	Area Under Curve	Classification accuracy	F1	Precision	Recall
LR1	0.819	0.745	0.745	0.745	0.745
LR2	0.816	0.743	0.743	0.743	0.743
LR3	0.819	0.745	0.745	0.745	0.745

Table 7: Logistic Regression scores

4.5.4 Trained data models

Score

Best performing	Area Under Curve	Classification accuracy	F1	Precision	Recall
NN1	0.837	0.778	0.778	0.778	0.778
LR1	0.819	0.745	0.745	0.745	0.745
kNN2	0.842	0.746	0.745	0.747	0.746

Table 8: Best fits for each model

		Predicted		
		< 6	≥ 6	Σ
Actual	< 6	1687	543	2230
	≥ 6	524	2046	2570
Σ		2211	2589	4800

(a) Neural network confusion matrix.

		Predicted		
		< 6	≥ 6	Σ
Actual	< 6	1504	726	2230
	≥ 6	492	2078	2570
Σ		1996	2804	4800

		Predicted		
		< 6	≥ 6	Σ
Actual	< 6	1631	599	2230
	≥ 6	624	1946	2570
Σ		2255	2545	4800

(b) kNN Confusion matrix.

(c) Logistic Regression Confusion matrix.

Figure 13: Confusion matrices for the best fit of each model.

4.6 Conclusion

Table 8 clearly show that the best fit for the classification task was the neural network with a classification accuracy of 77.8 %. This accuracy is not as high as desired but the reason for this is that the data-points are very close to one another and the clusters are overlapping. Figure 13a show that the neural network has managed to predict 1687 out of 2230 wines of lower quality and it predicted 2046 wines of 2570 wines correctly. Therefore the best algorithm managed to classify 3582 wines out of 4800 wines correctly.

In conclusion this dataset was a good fit for machine learning but it could have been even better if the clustering were more scatter. That would have allowed for the supervised learning to achive even better results.

This task was very interesting but it required alot of work in a very short time. I found it hard to interpret some of the results and I wish I would have recieved more practice before the assignment. Overall I have learned alot from this assignment.

References

- Foundation, Open Knowledge (2023). *Open Data Commons*. URL: <https://opendatacommons.org/licenses/dbcl/1-0/> (visited on Mar. 27, 2023).
- LEARNING, UCI MACHINE (2023). *Red Wine Quality*. URL: <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009> (visited on Mar. 20, 2023).
- Mining, Orange Data (2023). *Orange Data Mining Library*. URL: <https://orange3.readthedocs.io/projects/orange-data-mining-library/en/latest/#reference> (visited on Mar. 27, 2023).
- Refsdal, Elias Woie (2023). *Machine Learning*. URL: <https://github.com/EliasWR/MachineLearning> (visited on Mar. 27, 2023).