Assignment Part D

Diagram, map

Description automatically generated

# Part 1

For assignment D I have chose the Wine Quality Data Set(the red wine one) Paulo Cortez. The sources for the dataset can be viewed at the bottom of the report. The data set is a multivariate set. The data set has 12 attributes and can be viewed as both classification or regression tasks. The wine used for data are the red wines from the Portuguese Vinho Verde wine and the wine seems to have been collected from their company. The wine was then subjected to psychochemical tests to the get the input variables. The data set has 1599 data objects. There is only one class in this dataset and that is the class that evaluates quality of the wine. This class also contains 7 data objects valued 3-8.

There are 11 features and one output variable/class

|  |  |  |
| --- | --- | --- |
| Feature representation | Meaning | Range |
| Fixed acidity | The amount of low volatile organic acids in the wine | [4.6, 15.9] |
| Volatile acidity | The amount of volatile or gaseuos acids in the wine | [0.12, 1.58] |
| Citric acid | The amount of citric acid added to the wine to give it flavor or freshness | [0, 1] |
| Residual sugar | The amount of leftover sugar after the wine is done fermenting | [0.9, 15.5] |
| Chlorides | How much salt is in the wine | [0.012, 0.611] |
| Free sulfur dioxide | The amount of free form SO” in the wine | [1, 72] |
| Total sulfur dioxide | The total amount of SO2 in the wine | [6, 289] |
| Density | The density of the wine | [0.99007, 1.00369] |
| pH | The pH level of the wine | [2.74, 4.01] |
| Sulphates | The amount of SO in the wine | [0.33, 2] |
| Alcohol | The alcohol % of the wine | [8.4, 14,9] |
| Quality(class | The measured quality of the wine | [3, 8] |

Heres the excel file

Graphical user interface, application, table

Description automatically generated

Heres the imported table

Table

Description automatically generated

Here are the scatterplots

Chart, scatter chart

Description automatically generated

Here in scatterplot one we can see that as the % of alcohol increases the volatile acidity tends to decrease. Chart, scatter chart

Description automatically generated

And Here in scatterplot two we can see that as the amount of citric acid increases the volatile acidity also tends to decrease

Here are the histograms showing separation by class:

The first 2 shows the alcohol percentage and the frequency of the data objects from each of the subclasses and the second one does the same for volatile acids. The last histogram shows the alcohol percentage and the frequency of the data objects from each of the subclasses Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Here’s the statistics

Chart

Description automatically generated

Conclusion

Both the scatterplots and the histograms show that the data set is not balanced. We can clearly see that the wine qualities 5, 6 and 7 category has a lot more data objects than the others. So much that if we were to move wine quality 6 and 7 there would still be an abundance of wine quality 5 compared to the other wine qualities. The same can also be said for wine quality 6 and wine quality 7.

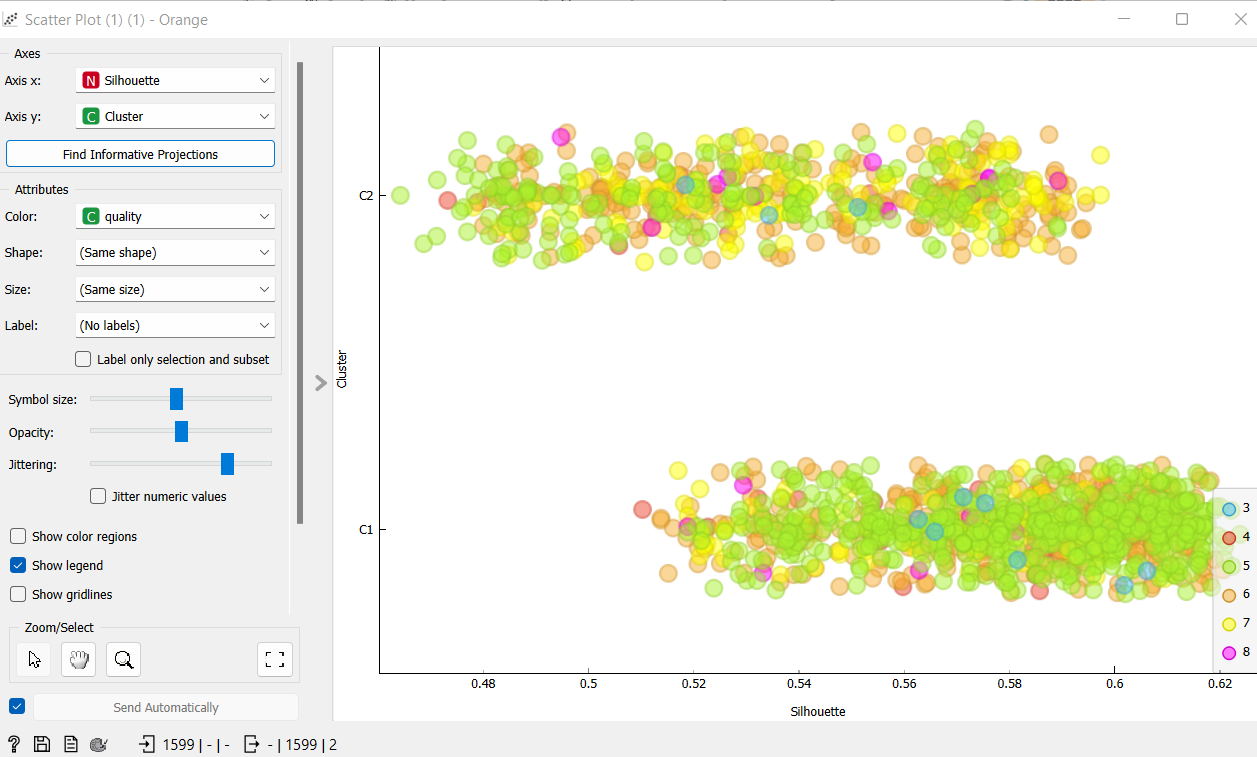
There is no clear data structure, with the exception of outliers all the plots are close by each other in the cluster. We can see that wine quality 5 could be a group since they all mostly reside beside each other on one side of the cluster

# Part 2

K-means algorithm

For this algorithm I did 5 test which will be shown below I gave them all different k values, which are the number of clusters the algorithm should generate. In the orange program I used to from method to get the silhouette score, where x to y represents the cluster range you want the silhouette score to be calculated for.

Graphical user interface

Description automatically generated 

K-means experiment 1

The hyper parameters used were 2 to 8. Since we know that the closer the silhouette score is to 1 the better separated the clusters are we can conclude that these clusters are not particularly well separated. Though this specific experiment has the highest silhouette score which is 0.202

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated A picture containing chart

Description automatically generated

K-means experiment 2

The hyper parameters used were 5 to 20. Since we know that the closer the silhouette score is to 1 the better separated the clusters are we can conclude that these clusters are not particularly well separated. This experiment’s silhouette score is 0.177

Graphical user interface, application

Description automatically generatedGraphical user interface

Description automatically generated

A picture containing chart

Description automatically generated

K-means experiment 3

The hyper parameters used were 4 to 15. Since we know that the closer the silhouette score is to 1 the better separated the clusters are we can conclude that these clusters are not particularly well separated. This experiment’s silhouette score is 0.192

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

Chart

Description automatically generated

K-means experiment 4

The hyper parameters used were 7 to 30. Since we know that the closer the silhouette score is to 1 the better separated the clusters are we can conclude that these clusters are not particularly well separated. This experiment’s silhouette score is 0.175. This experiment also has the lowest silhouette score of 0.120

Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

Chart, diagram

Description automatically generated with medium confidence

K-means experiment 5

The hyper parameters used were 10 to 20. Since we know that the closer the silhouette score is to 1 the better separated the clusters are we can conclude that these clusters are not particularly well separated. This experiment’s silhouette score is 0.143

DBSCAN

For this algorithm I did 3 test which will be shown below. I gave them all different min\_sample values and maximum distance. The min samples hyperparameter is the minimum number f data objects that needs to be located together for the data object to be considered a core data object. The maximum distance hyperparameter is the largest distance to data objects can have while still being considered neighbours. In the orange program these values are called core point neighbours and neighbourhood distance. I also tried changing the distance metric

Chart, line chart

Description automatically generated

In this experiment the min samples is 6 and the maximum distance is 3.06. The distance metric is Euclidean which refers to the distance between 2 points

Chart, line chart

Description automatically generated

In this experiment the min samples is 7 and the maximum distance is 4.87. The distance metric is Manhatten which refers to the sum of absolute differences of all attributes

Chart, line chart

Description automatically generated

In this experiment the min samples is 10 and the maximum distance is 1.01 . The distance metric is Euclidean which refers to the distance between 2 points.

Conclusion

Based on the low/close to 0 silhouette scores of the experiments I’ve concluded that the classes in this dataset are poorly separable

# Part 3

I chose logistic regression and K-Nearest Neighbor algorithm. Logistic regression is a linear regression model that gives us a real valued out between [-inf, inf]. It is used to predict a numerical value and in the case of this assignment it will be used to predict the discrete values for the wine’s quality. The KNN algorithm looks for distance or closeness between the data objects. I’ve gone with the KNN algorithm which is for categorical values. This version counts the number of differing features in 2 data objects. I have chosen these two algorithms because they can do both regression tasks and classification tasks.

The hyperparameters available in orange are number of neighbours, distance metric and weight for the KNN algorithm and regularization type and strength for logistic regression. In the test set there are 1120 data objects which is 70% of the entire dataset.

Graphical user interface, text, application

Description automatically generated

For these experiments I will be focusing on the precision and the classification accuracy

KNN Experiments

Graphical user interface, application

Description automatically generated

Here I sat the number of neighbors to 5 and the distance metric to Euclidean. As this is the model with the highest classification accuracy and precision it is the one I will be using.

Graphical user interface, application

Description automatically generated

Here I sat the number of neighbors to 25 and the distance metric to Manhatten

Graphical user interface, application

Description automatically generated

Here I sat the number of neighbors to 50 and the distance metric to Chebyshev

Logistic regression

Graphical user interface, application

Description automatically generated

Here I sat the strength to 1 and the regularization type to Lasso L1

Graphical user interface, application

Description automatically generated

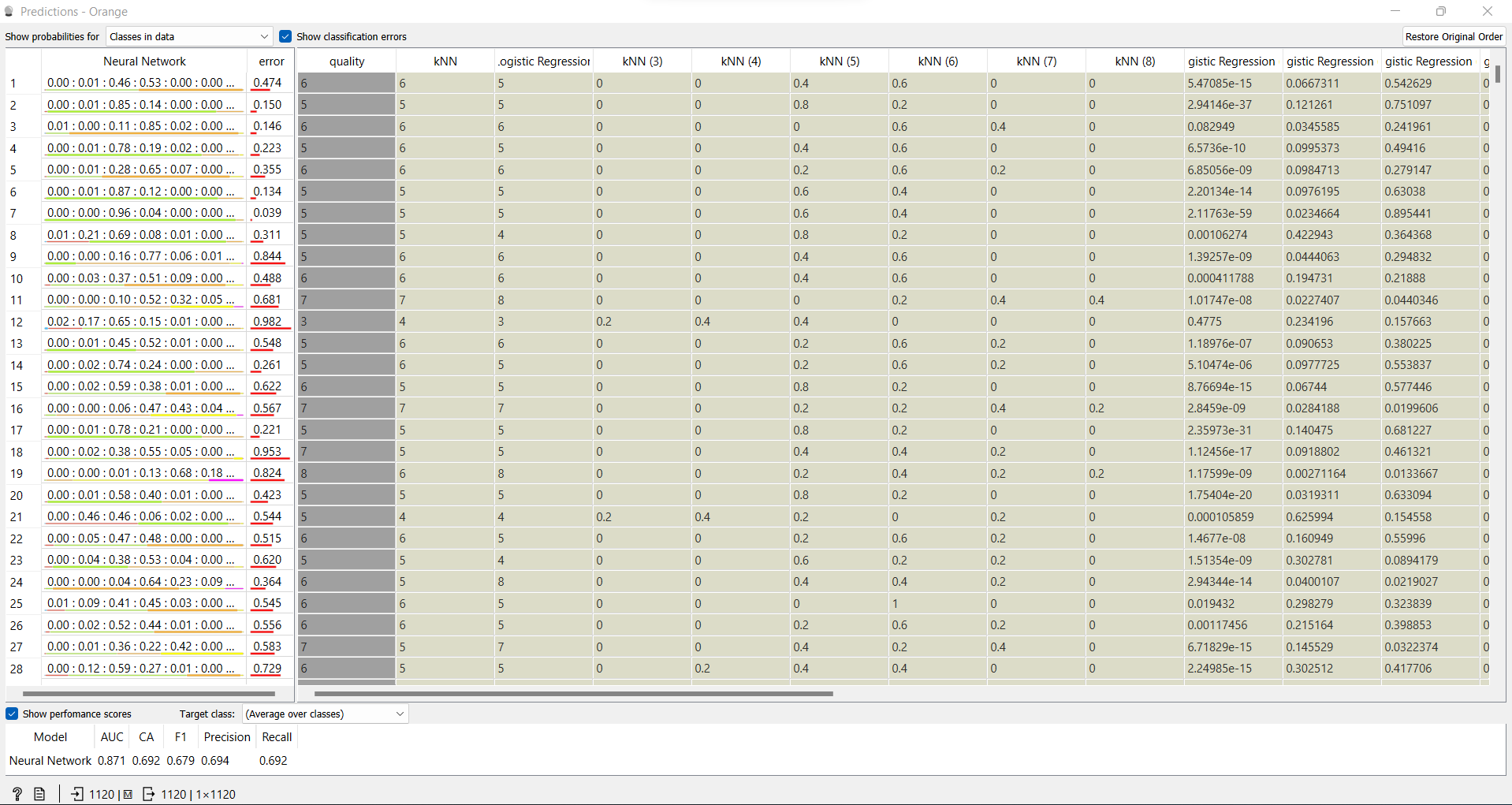
Here I sat the strength to 5 and the regularization type to Lasso L1. I’ve chosen to use this model since it has the highest classification accuracy.

Graphical user interface, application

Description automatically generated

Here I sat the strength to 0.4 and the regularization type to Lasso L1

Here is the neural network presented using the orange prediction tool.



My conclusion is that the model predicts 70% of the classification task which is okay, but it should be better. Perhaps by training the model more

# Sources

Link to: <https://archive.ics.uci.edu/ml/datasets/wine+quality>

Link to the chosen data set: <https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv>

Citation request as requested by the website: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.  
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.