Assignment 1 - Big Data Technologies

Alexandra Russell - 201882769

5th of November 2018

This report is the coursework for the CS982 course, and has a word count of 3102 words. The code that accompanies this report can be found in the Code folder within the zip folder that constituted the hand in. The code was written in the Spyder development environment and run on python 2.7.9. The packages used in the code are the following:

- pandas
- numpy
- matplotlib
- seaborn
- sklearn
- random

1 Introduction to the Dataset

The dataset used in this work is the Red Wine Quality dataset which can be found on both the Kaggle (*Kaggle - Red Wine Quality* 2017) and UCI Machine Learning Repository (*UCI Machine Learning Repository - Red Wine Quality* 2017) web sites. The dataset was created for the purpose of research conducted into a "data mining approach to predict human wine taste preferences" (P. Cortez et al. 2009). It contains information about 1566 red wine samples in the form of 11 different chemical measurments of the wines and a quality rating based on blind taste testing. The wine samples are all "Vinho Verde" wines, meaning that they all come from the same north western region of Portugal. Whilst they were all made in the same region they have not all been made by the same producers or using the same grapes. The dataset does not provide producer or grape variety information so this report focuses solely on the chemical properties of the samples.

The different attributes of the wine, and the units they are measured in, are the following;

- Fixed acidity (g/dm^3)
- Volatile acidity (g/dm^3)
- Citric acid (g/dm^3)
- Residual sugar(g/dm^3)
- Chlorides(g/dm^3)

- Free sulfur dioxide (mg/dm^3)
- Total sulfur dioxide (mg/dm^3)
- Density (g/cm^3)
- pH
- Sulphates (g/dm^3)
- Alcohol (vol.%)

The quality of the wine is a categorical numerical value between 0 and 10 that was determined by choosing the median value from at least 3 professional wine tester ratings. No variables in the dataset have any null values.

2 Identification and description of the key challenges and problems to be addressed

I chose this dataset initially because I suspected the range of numerical values that the chemical properties hold would present interesting challenges and I saw potential for the application of some form of machine learning technique given the quality ratings the wine samples have been given. Indeed, the more I looked into the dataset, the more questions and challenges arose. In this report I interest myself in a few of them, whilst outlining the limits of the methods I employed and the dataset itself.

Firstly, as I had initially thought, the quality ratings in the dataset beg the question "can this dataset be used to predict the quality of a wine based on it's chemical properties?". Section 4 of this report explores ways to build a classifier to answer just that question.
Secondly, I wondered if there was anyway to classify these wines according to their chemical properties? Are their wines in the dataset that hold significantly different chemical profiles? An attempt to group the wines in such a way is addressed in section 5 of this report using an unsupervised learning technique.

The main challenge presented by both of these questions is that of determining which of the chemical properties are significant. Intuitively, not all of the attributes will be useful in determining the quality of wine or splitting the wines into distinct groups. Addressing this challenge involves understanding

the properties of these attributes within the context of the dataset, a challenge addressed in section 3 of this report. Once we understand more about these variables we must only used those that are useful to whatever model is being implemented. Without doing so, there is a risk that a model is just modeling noise rather than extracting insight from the values.

The other significant challenge is that the dataset does not have equal amounts of each quality of wine. There are far more medium quality wines which will undoubtedly affect the models as they are being developed.

3 Summary Statistics

3.1 The Wine attributes

The 11 chemical properties of the wines are all stored as continuous variables and the quality is a categorical one. To understand the continuous variables, Table 1 displays some of their key statistical properties. We can note that whilst a lot of the attributes are measured using the same units they hold very different values. For example, chlorides and fixed acidity are both measured in g/dm^3, but the values taken by the wines for these characteristics are very different. Additionally, not all the variables use the same units. This is not particularly surprising, however this will be kept in mind when developing any models to avoid bigger values dominating the model, and will need to be addressed when attempting to compare attributes.

Table 1: Descriptive statistics of the continuous variables

	mean	standard deviation	min	max
fixed acidity (g/dm^3)	8.32	1.74	4.6	15.9
volatile acidity (g/dm^3)	0.52	0.18	0.12	1.58
citric acid (g/dm^3)	0.27	0.19	0.0	1.0
residual sugar (g/dm^3)	2.54	1.41	0.9	15.5
chlorides (g/dm^3)	0.087	0.047	0.012	0.611
free sulfur dioxide (mg/dm^3)	15.87	10.46	1.0	72.0
total sulfur dioxide (mg/dm^3)	46.47	32.89	6.0	289.0
density(g/cm^3)	0.99	0.0019	0.99	1.004
pH	3.31	0.15	2.74	4.01
sulphates(g/dm^3)	0.66	0.17	0.33	2.0
alcohol (vol.%)	10.42	1.07	8.4	14.9

To understand the spread of different values within the chemical proper-

ties we can look at the boxplots in figure 1. Notably, some variables have reduced interquartile ranges with significant outlying points. This can be seen in particular for chlorides and residual sugar variables. For these variables, we can see that most of the values they take are similar, but there are a few outliers with very different values. Figure 2 is the density plot for residual sugar, and we can see the broadly normal distribution with the outliers at the extreme ends of the plot. Chlorides and residual sugar may be the variables to most obviously present this behaviour, however most of the variables have a significant number of outliers. It is important to note that since I do not have an understanding of what constitutes 'normal' values for these types of chemical properties, I am assuming that our samples are representative of the wine in that region. Additionally, when I say that the values are mostly similar, this is relative the the full spread of all the possible values.

We can compare the boxplots by eye, but it is useful to have some numerical values to compare and for this we can look at the interquartile ranges (ie: the range that exists between the extreme values of the central 50% of the data, represented on the box plot by the extremities of the box). However due to the variables holding ranges of different values and using different units of measurement we cannot compare the actual interquartile ranges and must use scaled ones. Once they have been scaled, a larger scaled interquartile range for one chemical property compared to another means the first property have a relatively larger range of diverse values. Table 2 shows these scaled interquartile ranges and we can see that alcohol is the chemical property with the proportionally biggest range, and chlorides has the smallest. In

practice, these smaller ranges mean that there are more extreme outliers that are dominating the total of the values for these attributes. When it comes to modeling, these attributes could be very useful if the wines presenting the outlying values all have something in common. However if they do not, then the attributes could be useless since most of them have very similar values, and splitting them into groups will be difficult.

Table 2: Interquartile ranges of scaled variables

	Interquartile range
fixed acididty (g/dm^3)	0,18
volatile acididty (g/dm^3)	0,17
citric acid (g/dm^3)	0,33
residual sugar (g/dm^3)	0,048
chlorides (g/dm^3)	0,033
free sulfur dioxide (mg/dm^3)	0,19
total sulfur dioxide (mg/dm^3)	0,14
density(g/cm^3)	0,16
pH	0,15
sulphates (g/dm^3)	0,11
alcohol (vol.%)	0,25

After looking at the values that these variables contain, we can ask ourselves how they interact together, bearing in mind that one of our aims is to predict the wine qualities. Figure 3 is a heatmap indicating the correlation between different variables in the dataset. Some values are unsurprising, such as the positive correlation between total sulfur dioxide and free sulfur dioxide, and the negative correlation between pH and fixed acidity. When thinking towards predicting the quality of the wine, we can note the positive correlation between quality and alcohol content, and negative correlation between all three of volatile acidity, total sulfur dioxide and density and the

quality of the wine. It will be interesting to see if these correlations play a part when it comes to building a classifier for the wine quality.

3.2 Wine Quality

As touched upon previously, there are more medium quality wines than very high or very low quality ones. Figure 4 is a bar chart showing the number of wines according to their quality value and the high proportion of medium quality wines is obvious. In fact, 1,319 wines have a quality of 5 or 6 (82 percent). This has the potential to become a problem when building a classifier, since the weighting towards certain quality ratings might bias the model.

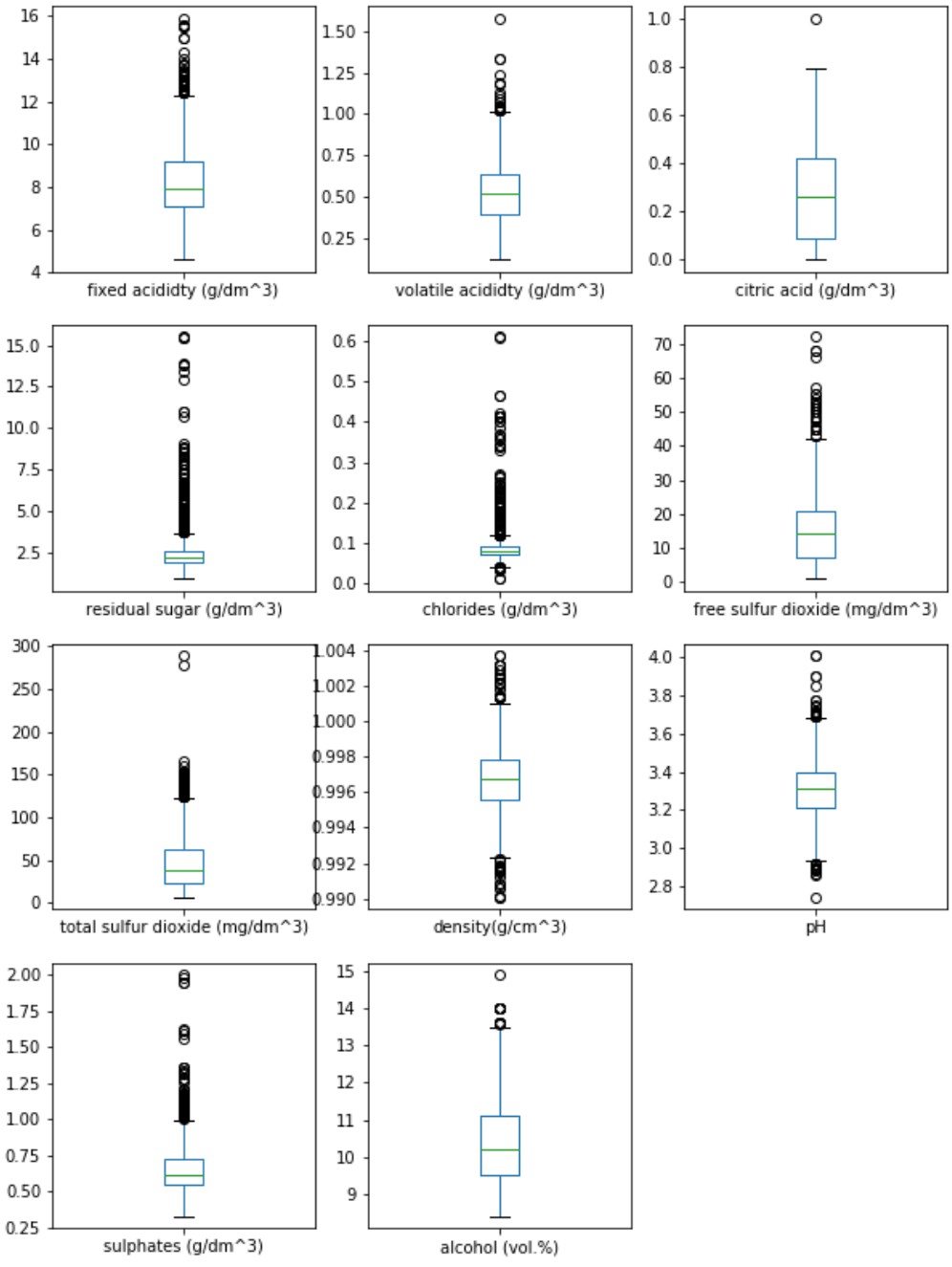

Figure 1: Boxplots of the continuous variables

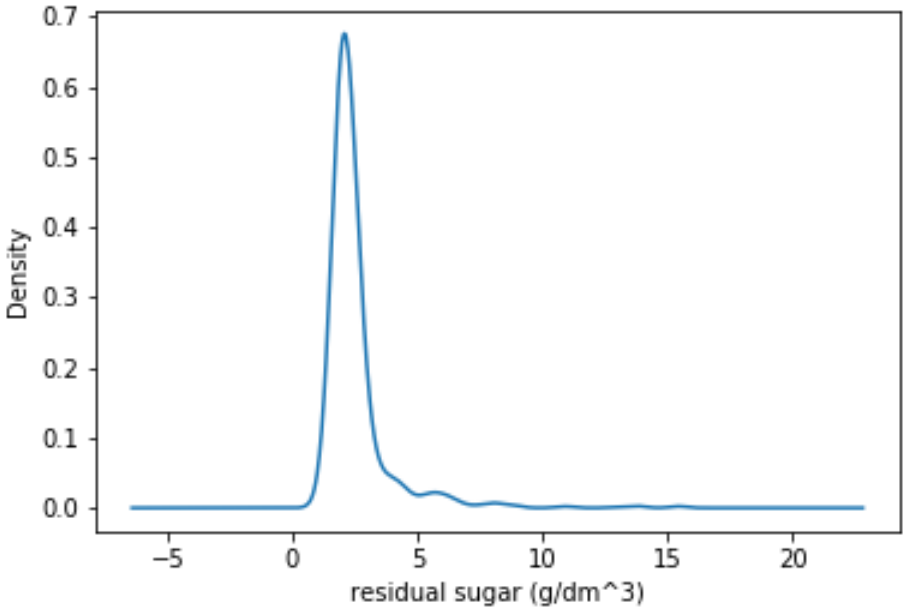

Figure 2: Density of residual sugar values

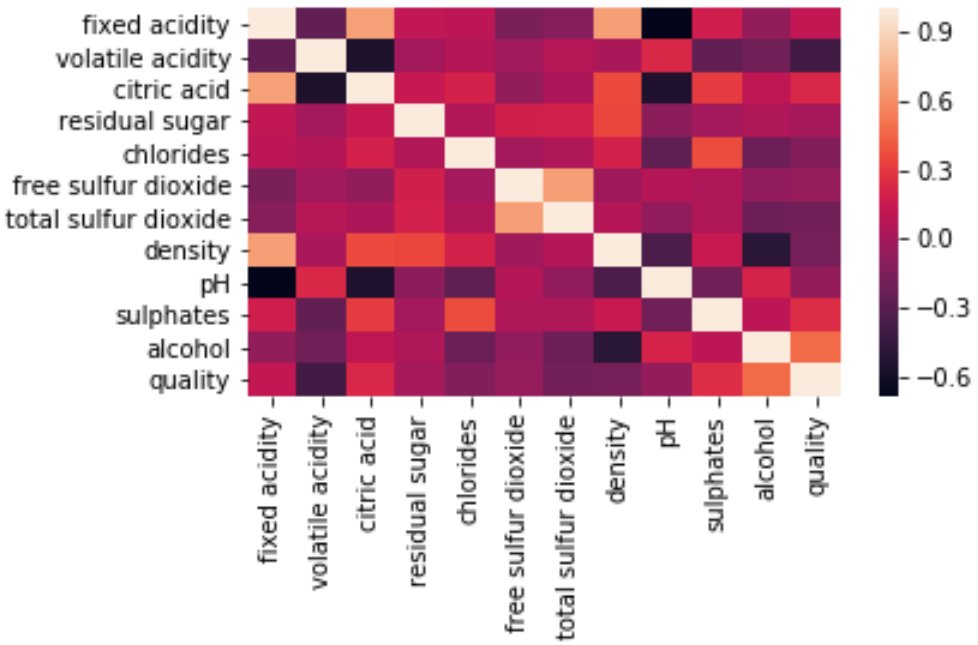

Figure 3: Heatmap of the correlation between variables

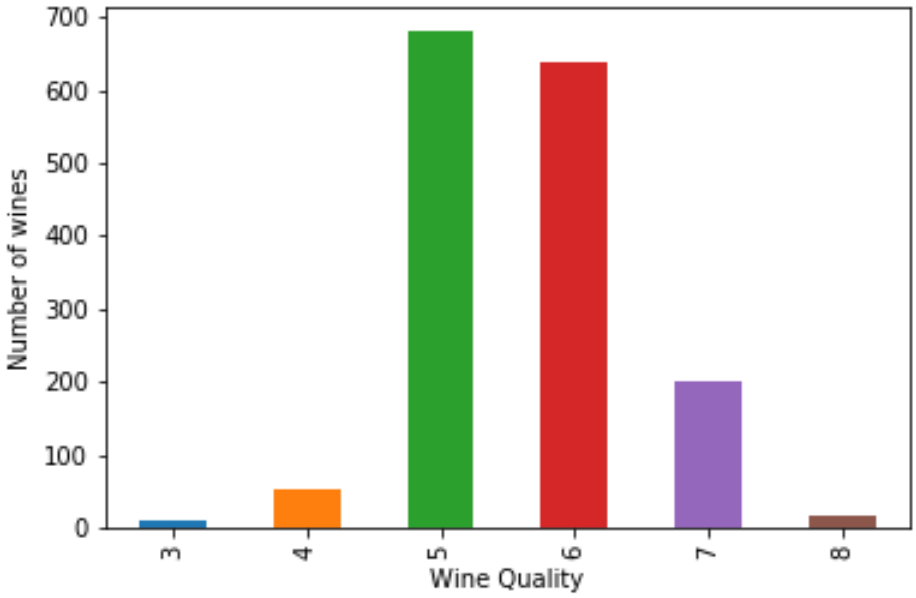

Figure 4: Number of wines according to their quality

4 Description, rationale, application and findings of one supervised method

This red wine dataset would appear to lend itself well to a supervised learning technique to predict the quality of a wine sample based on it's chemical properties. Given that the quality is a categorical variable, the learning technique must be a classifier, and I decided to try a decision tree classifier. It was chosen for it's simplicity and its aptitude for handling continuous input variables (EMC Education Services (2015)) and is implemented using the Scikit Learn packages, the details of which can be found it one of their publications (Pedregosa et al. (2011)). Throughout this report the algorithms are trained using 80% of the data and tested using the other 20%.

Initially the decision tree method used all of the attributes in the dataset and the average accuracy score over 10 tests was 0.61. Whilst not a terrible score, we can wonder if there are superfluous attributes being included in the model that could be removed to get a better score. To determine the best attributes to include I used a genetic algorithm to explore the effectiveness of including different combinations of attributes.

A genetic algorithm is a "population based model that uses selection and recombination operators to generate sample points in a search space" (Whitley 1994) and I developed one for the purposes of 'evolving' the best combination of attributes. The genetic algorithm implemented takes a population of 60 combinations of attributes and creates 40 generations, storing the best com-

bination for each generation. The best combination is determined using the average accuracy scores over 10 tests of decision trees created using that combination of attributes. The genetic algorithm uses tournament selection to determine the parents in each generation and simple mutation and crossover functions.

The best combination of attributes the genetic algorithm found were 'volatile acidity', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'density', 'pH', 'sulphates' and 'alcohol'. Somewhat predictably, all four of the attributes that we noted as being correlated with quality in section 3 (alcohol, volatile acidity, total sulfur dioxide and density) are included in this list. The accuracy score of the decision tree model using those attributes is 0.64. This is slightly better than when the model used all the attributes, and to further analyse the results the classification report (Table 3) and the confusion matrix (Figure 5) were generated. The classification report shows low precision and recall on very high and very low quality ratings and the confusion matrix also shows that many high and low quality wines are being incorrectly classified. This is most likely due to the fact that there is little data to train these quality values on, as few wines hold very high or very low quality values as we saw in section 3 of this report.

With this in mind, I tried changing what we are trying to predict to avoid this problem. I decided to try predicting simply if a wine was 'good' or 'bad' rather than predicting the actual quality rating (this technique is also suggested in the description of the dataset on the Kaggle website). To this end,

Table 3: Decision tree classification report

quality	precision	recall	f1-score	support
3	1.00	0.25	0.40	4
4	0.09	0.11	0.10	9
5	0.71	0.71	0.71	133
6	0.66	0.63	0.64	137
7	0.61	0.76	0.68	33
8	0.00	0.00	0.00	4
avg / total	0.65	0.65	0.65	320

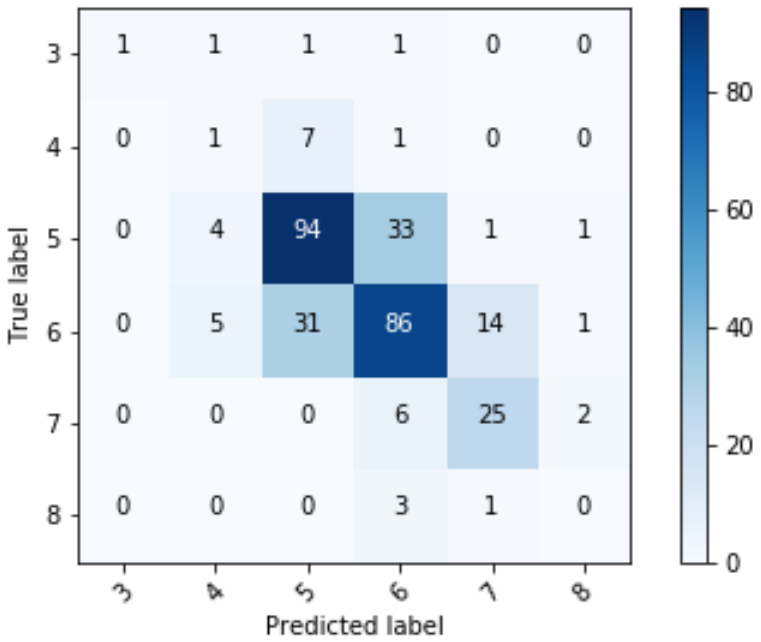

Figure 5: Decision tree confusion matrix

I added a 'binary quality' variable to the dataset which takes a value of 1 if the wine is of 'good' quality (meaning a quality rating of 6 or above) or a value of 0 if the wine is 'bad' (a quality rating of 5 or below). This cutoff

point was chosen because it gives a similar number of wines in each category, and didn't seem an unreasonable place to divide them.

I ran the genetic algorithm once again, but predicting this new 'binary quality' category. It determined that the best attributes to use were 'volatile acidity', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'sulphates' and 'alcohol'. It produced an average accuracy score of 0.78. Once again, we generated the classification table and the confusion matrix. The results are better than when predicting the quality ratings with fewer incorrect classifications and higher precision and recall scores.

Table 4: Decision tree classification report (two outcomes)

quality	precision	recall	f1-score	support
bad	0.77	0.78	0.77	148
good	0.81	0.80	0.80	172
avg / total	0.79	0.79	0.79	320

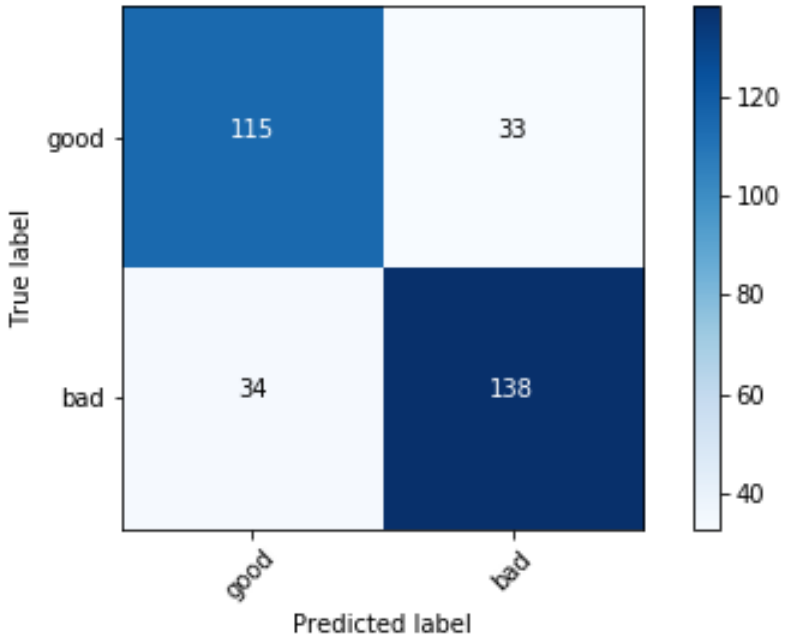

Figure 6: Decision tree confusion matrix (two outcomes)

5 Description, rationale, application and findings of one unsupervised method

When I initially thought about grouping wines into clusters I though that perhaps they would group according to their quality. For this reason the clustering method chosen was K means clustering since it can handle clusters of varying sizes (EMC Education Services (2015)). The initial implementation produces 7 clusters, taking scaled data and all of the variables. To measure the clustering I used the silhouette score for the model and the completeness and homogeneity scores for the 'quality' variable. These values were 0.17, 0.07 and 0.12 respectively.

Similarly to when implementing the supervised method, this not an excellent result and we can ask ourselves if selecting fewer variables could improve our results. Once again, a genetic algorithm was used to determine the attributes to use. It is the same implementation as the one used in the supervised method section but the measure of the solution was the sum of the completeness and homogeneity scores. It determined that the best attributes to use to cluster were volatile acidity, total sulfur dioxide and alcohol, and the model gave a completeness score of 0.11 and a homogeneity score of 0.17. This is an improvement on the initial attempt using all the attributes but still not very good at all regardless of what kind of clusters you want.

Upon reflection, it became clear that clustering the wines to try and group them by quality is not the best use of a clustering mechanism. The supervised method is more apt for this task and a clustering method could be used for a different task. We can ask ourselves if there are specific groups of wines with significant different chemical components. To do this we still use K means clustering and the normalised values, but this time we only see if we get significant clusters by using the silhouette score. We don't know how many clusters we want to get so I tested the different numbers of clusters between 3 and 10 using the silhouette score as the measure of effectiveness. The results are recorded in figure 7. None of the scores are particularly high, however the score for 6 clusters is the highest at just above 0.21, so this seems like a good place to start.

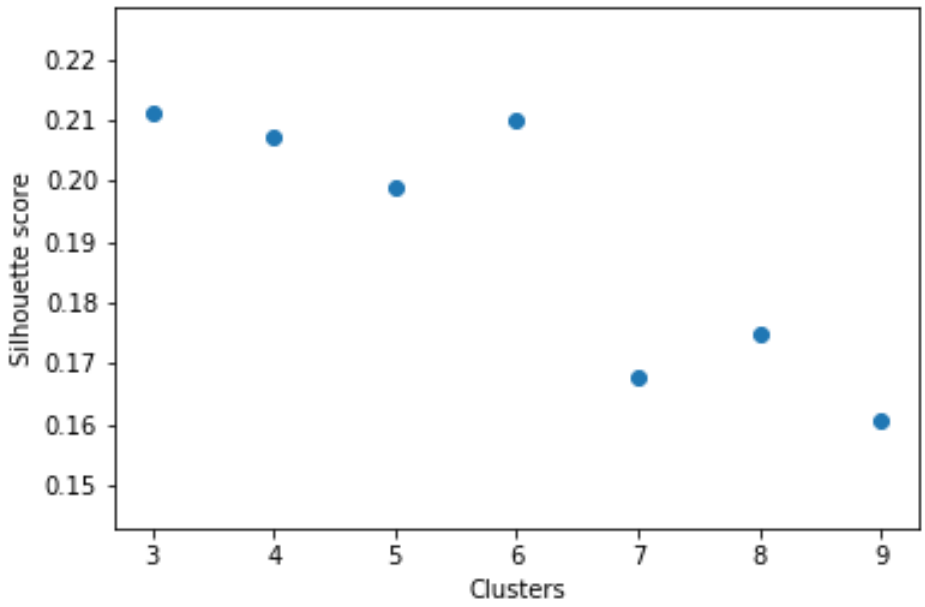

Figure 7: Silhouette scores according to number of clusters

Once again our attention turns to attribute selection. The problem here is the more we reduce the number of attributes we are using to cluster, the higher the silhouette score will be. This is because it is more likely to produce distinct clusters when you are only looking at one variable. So to group wines that have distinct chemical profiles I decided to set a minimum of three attributes to be looked at, and the genetic algorithm would pick the three that separate the wine into the most distinct 6 clusters. It determined that the best attributes to use were residual sugar, chlorides and alcohol, and that using them gave a silhouette score of 0.48. Table 5 shows the completeness and homogeneity scores of each of those 3 attributes that have been used in

the model. We can see that homogeneity is fairly low for all the variables, but the completeness score as better. This indicates that the clusters that have been created group similar scores together well but also include a lot of 'incorrect' values. To help visualise the clusters, figures 8 and 9 show the groups created according to different combinations of the attributes we were using to cluster them. We can see clearly how well the clustering works for alcohol however the other two attributes don't quite have the same distinct clusters.

We can note that two of the variables selected for the clustering are residual sugars and chlorides, the two variables we picked up on as having a lot of outlying values in section 3. Figure 9 in particular shows that the clustering is grouping those more extreme values but struggles with the more dense group of similar values.

Table 5: Completeness and homogeneity of K means clustering

	completeness	homogeneity
residual sugar	0.25	0.092
chlorides	0.25	0.069
alcohol	0.79	0.27

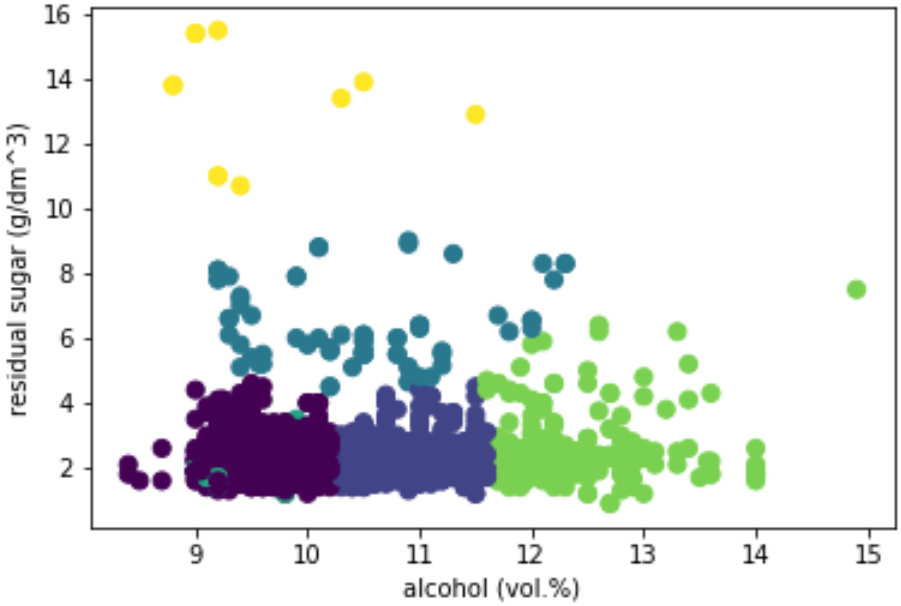

Figure 8: Clusters according to alcohol and residual sugar

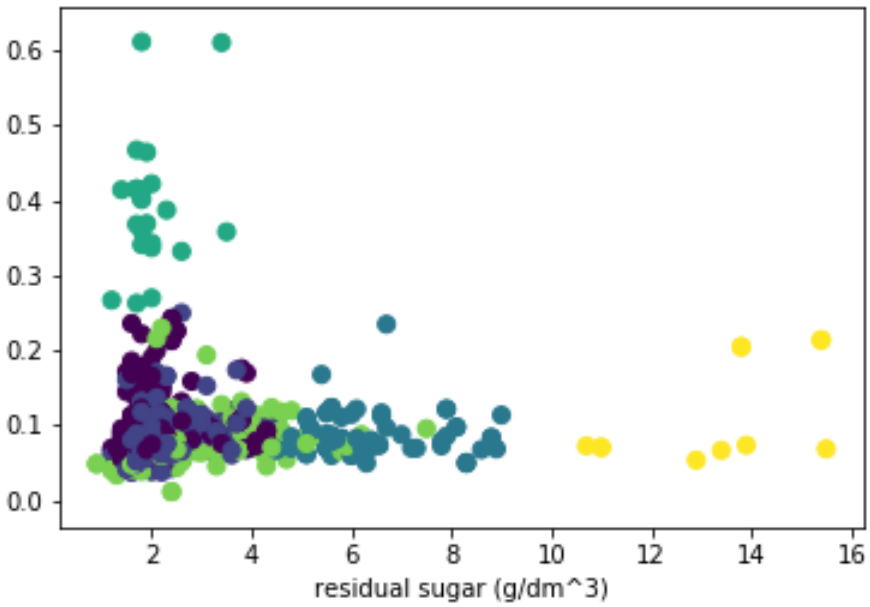

Figure 9: Clusters according to residual sugar and chlorides

6 Reflection on methods used for analysis

Concerning the supervised method, I am happy with my attempts to create a classifier. I am happy with the exploration of the parameters used in the decision tree and the outcome of the classifier is acceptable, even if there are still multiple miss-classifications. It is at least a better model that guessing a class at random. Further work could be done into investigating other kinds of classification algorithms. Given the binary outcomes that were generated it might be possible to use a Naive Bayes algorithm.

When it comes to the unsupervised method, I am less happy with the outcomes. I am not convinced that the final classification is a particularly good way to group wine. The main problem really is that I do not have a very good understanding of the effect of the chemicals on the wine itself, and learning enough about it would be outside the scope of this work. However a greater understanding of this would allow a more comprehensive analysis of the effectiveness of the groups. When grouping wines the main outcome we would want really is groups of wines that taste the same, and I have no idea if residual sugar really plays that big of a role. Additionally I assume that the ranges of chemical values present in the dataset are significant, however it's possible that actually all the different values for a particular chemical are basically the same or will have the same effects. With this in mind, further work could be done into understanding the chemicals that this dataset reports on. However, I also considered that after having selected the variables for the clustering method it would have been good to test more cluster sizes, but I ran out of time.

In conclusion, this dataset can be used effectively to create a classification model to predict wine quality. However it is less apt for creating a clustering mechanism, in particular when not accompanied with knowledge about the chemical constitutions of wine and it's effects.

References

EMC Education Services (2015), *Data Science Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, Wiley, Indianapolis.

Kaggle - Red Wine Quality (2017), https://www.kaggle.com/piyushgoyal443/red-wine-dataset. Accessed: 2018-11-01.

P. Cortez, P., Cerdeira, A., Almeida, F., Matos, T. & Reis, J. (2009), 'Modeling wine preferences by data mining from physicochemical properties', *In Decision Support Systems* **Volume 47(4)**, 547 – 553.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

UCI Machine Learning Repository - Red Wine Quality (2017), https://archive.ics.uci.edu/ml/datasets/wine+quality. Accessed: 2018-11-01.

Whitley, D. (1994), 'A genetic algorithm tutorial', *Statistics and computing* **4**(2), 65–85.