PREDICTION OF RED WINE QUALITY

COSC 2670 – Practical Data Science
Assignment 2: Data Modelling

Tenzing Sangay Bhutia (s3680446) and Qianwei Yang (s3716296)
s3680446@student.rmit.edu.au s3716296@student.rmit.edu

Monday, May 27, 2019.

Table of Contents

COSC 2670 – Practical Data Science
Assignment 2: Data Modelling

Tenzing Sangay Bhutia (s3680446) and Qianwei Yang (s3716296)
s3680446@student.rmit.edu.au s3716296@student.rmit.edu

Monday, May 27, 2019.

ABSTRACT

The determination of quality of wine have always been mysterious question in the history of the world because it is not absolute. It totally depends on who's judging. The combined opinion of so-called wine experts normally labels the quality of wine as being good, average or bad. The implementation of machine learning algorithm and domain knowledge together, can now successfully help us solve this enigma to clearly classify the quality of red wine. To help us solve this question we acquired the dataset from the machine learning repository and went forward with the objective and hypothesis below.
The objective of the project is to predict the quality of Red Wine and to investigate to the following hypothesis:

1. The quality of low alcoholic red wine is better than that of high alcoholic red wine.
2. How does acidity and density effect the quality of red wine.

DATASET

The dataset was obtained from Vinho Verde wine samples, from the north of Portugal, available in the UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Wine+Quality?ref=datanews.io%5D]. Due to privacy and logistic issues, only physiochemical and sensory variable are available (there are no data about the brand or the price). The dataset includes 1599 observations and 12 variables. The variables include are:

- Fixed Acidity: refers to how well acidity balances out the sweetness and bitterness of the wine.
- Volatile Acidity: measure of wine volatile acidity.
- Critic Acid: Weak Organic Acid
- Residual Sugar: Sweetness of Wine
- Chlorides
- Free Sulphur Dioxide
- Total Sulphur Dioxide
- Density
- Sulphates
- Alcohol
- Quality (Score between 0 to 10)

METHODOLOGY

1. **Retrieving Data:** We start by loading all the relevant packages necessary for our analysis. Then we import the raw dataset into our Jupyter workspace. As our source file is of the type csv, we use the read csv function in pandas to help us with the importation and store it to a variable name Wine. We then label the target variable 'quality' which is actually a score between 1 to 10 to good, average and bad by using the qcut function of pandas. We perform this to change so as to make our problem a classification problem. We then use the head and shape function to check whether the loaded dataset is equivalent to the data in the source CSV file. The datatypes of the variables were also observed for further awareness of the dataset.

2. **Data Preprocessing or Data Preparation:** In the data preparation we do an elaborate check and focus on the content of the variables. We primarily focus on typos, data entry error, missing values, inconsistencies and transformation. After an elaborate check we realized that the dataset was free from error however there were a lot of duplicate values, hence we used the drop duplicate function to remove all similar observations. The number of observations was then removed to 1359. After exploring the target variable, we find a huge imbalance in the target variable hence we perform the under-sampling method to handle the imbalance. Under-Sampling is the process of randomly deleting some of the observation from the majority class in order to match the number of the minority class. We then again replace the datatype and the label of the target variable to increase the significance of the target variable and to enable modelling with K-Nearest Neighbours. We finally change the label to numeric for example, 1 = 'poor', 2 ='average', 3 = 'good'.

3. **Data Exploration and Hypothesis Check:** So far, we have only taken a quick glance at the data to get a general understanding of the kind of data we are manipulating. Now we dive into exploring the data and gain deeper visual insights for our prediction and investigation of hypothesis. Our exploration involves:
    - **Explore Each Column**: We start this process by getting a deep idea of each distribution of features by plotting a density graph and a histogram.

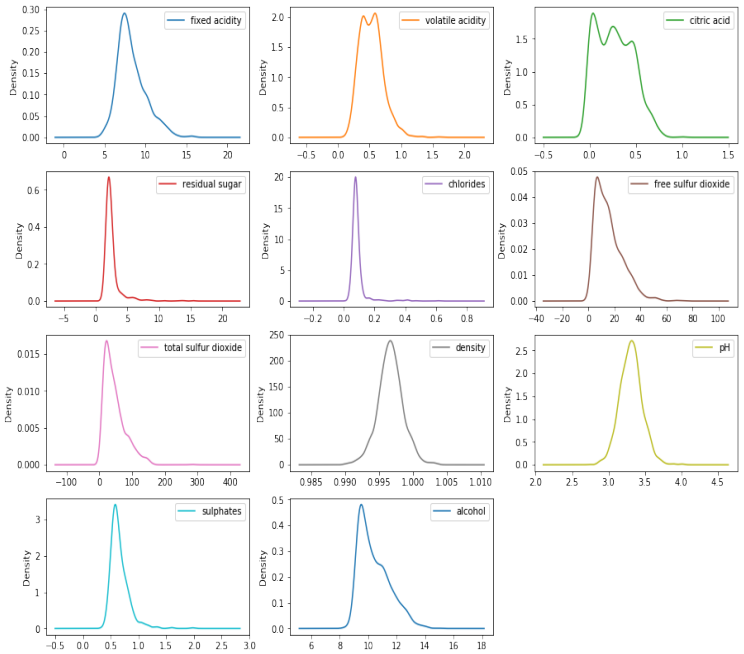


Fig -1: Density Graph of all the features

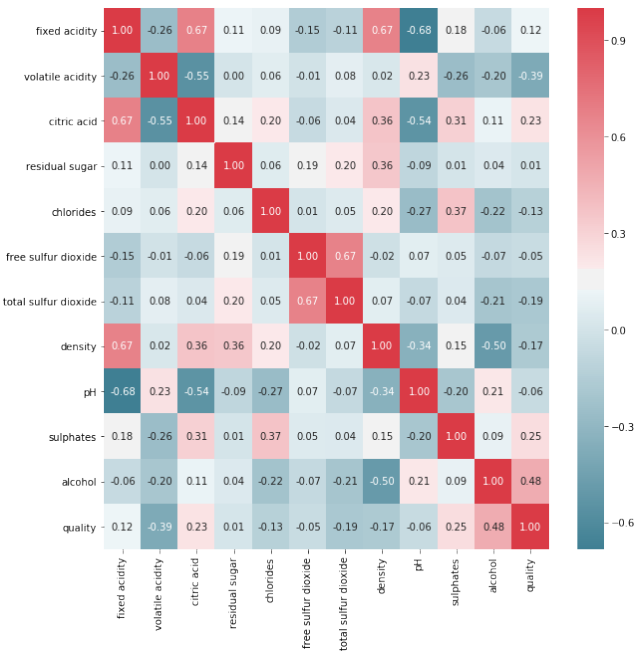


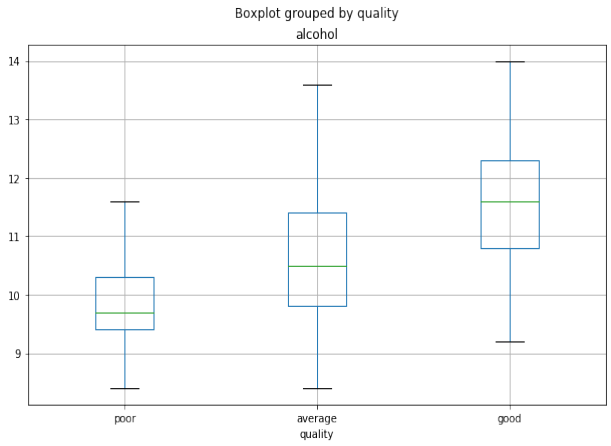
Fig-2: Scatter Matrix of the Features in the Dataset

From the density graph and the histogram, we see that majority of the features are positively skewed however pH and density are normally distributed.

- **Relationship between Attributes:** To explore the relationship between attributes we use the Correlation matrix which gives us a good diagrammatic visualization of the relation of the entire features with their correlation. Correlation coefficient ranges from-1 to 1. -1 indicates a strong negative correlation and +1 indicates a strong positive correlation and coefficients close to 0 indicate no linear correlation.

- **Investigation of Hypothesis:**
  "The Quality of Low Alcoholic Red Wine is Better than that of High Alcoholic Red Wine."
  We use a box plot of alcohol grouped by quality to confirm the hypothesis. From the boxplot, we see that the quality if wine increases with the increase in alcohol content hence we reject the hypothesis and confirm statistical insignificance. This finding can be very good indicators for consumers to detect good quality wine.

- **Investigation of Research Question:**
  "How does Acidity and Density effect the quality of Red Wine?"
  We go about the research question by plotting a scatter by taking into consideration all the features necessary namely Acidity, Density and Quality. In the observation of the scatter plot we see a strong positive correlation between acidity and density which is also confirmed by the scatter matrix however, we see so clusters being formed in the graph which signifies no difference in quality resulting that the two variables on quality are fruitless.



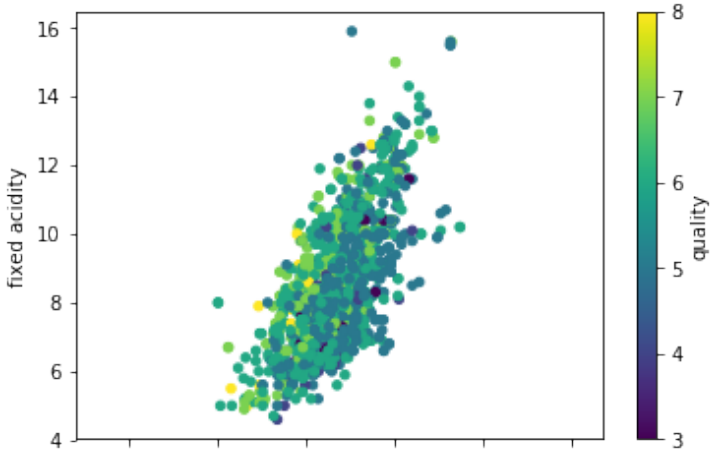


Fig 3: Boxplot of Alcohol content grouped by Quality.

Fig 4: Scatter of Fixed Acidity and Density by Quality

- **Exploring Relationship Between the Target Variable and the Features:** The last exploration we perform is to look at how the features are affecting each level in the target variable so as to get a better understanding of the distribution with respect to each level. The observation resulted in the following conclusions:
  - The quality of wine is better if the alcohol content is more.
  - Good Quality wine normally have less volatile acidity.
  - The quality of wine is better if the amount of Critic Acid is more.
  - The quality of wine is inversely proportional to Volatile Acidity.

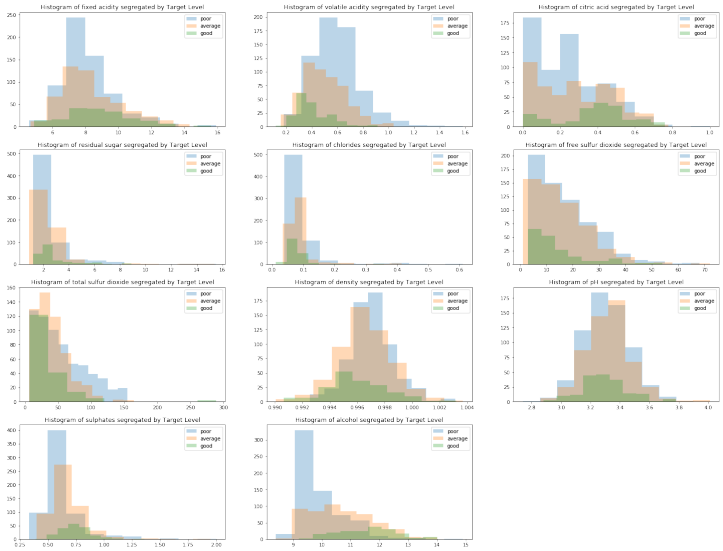


Fig 5: Relationship between target variables and the features

4. **Data Modelling:** As we have gained very important insights from our data exploration process we now move to data modelling where we need to basically start with model and variable selection, followed by model execution and lastly model diagnostic and model comparison. As we have the target variable 'quality', we consider our problem as a supervised machine learning problem. We start our data modelling process by normalising our features by min-max scaler to change the values of the numeric columns in the dataset to a common scale, without distorting difference in the range of target values.

   - **Feature Selection:** Features with high positive and negative correlation to the target variable will be taken into consideration. As we already explored the effect of features to the target variable in the data exploration stage with the help of correlation matrix. The feature that were selected are **fixed acidity, volatile acidity, citric acid, residual sugar, sulphates, alcohol.**

- **Hyperparameter Tuning:** A hyperparameter is a parameter whose value is set before the learning process begins. Hyperparameter tuning is choosing a set of optimal hyperparameter for a learning algorithm. In the case of K, we use the cross-validation technique to get the best value of K for the model. In the cross validation model the loop repeats the over different K and gives us the accuracy of the best K. In our case the best value was K=17. The Grid Search Cross-validation method was also used to get the rest of the parameter such as the weights, leaf size, etc… as given below.

```
GridSearchCV(cv=10, error_score='raise-deprecating',
       estimator=KNeighborsClassifier(algorithm='auto', leaf_size
=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=29, p=2,
           weights='uniform'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 2
7, 28, 29]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='w
arn',
       scoring='accuracy', verbose=0)

0.6014492753623188

KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='mink
owski',
           metric_params=None, n_jobs=None, n_neighbors=17, p=2,
           weights='uniform')

17
```

Fig 5: Grid Search CV Output

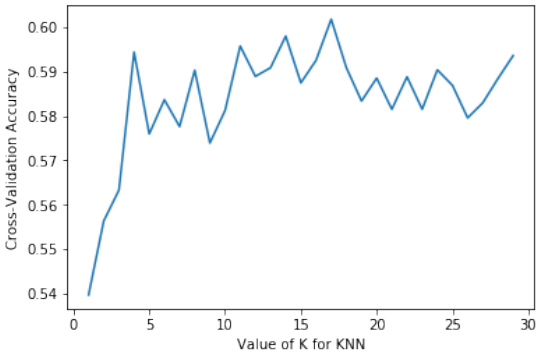


Fig 6: C. V. Accuracy v/s K. {Determination of Best K}

These features and hyperparameters were then tested on split test train ratio 50:50, 60:40 and 80:20. K-NN and decision tree algorithm were used to get the prediction of quality and the results of best model for KNN and Decision Tree was selected and compared to finally choose the best model.

RESULTS

As mentioned earlier the modelling was tested on 3 different train: test ratio using the K-Nearest Neighbours and Decision Tree algorithm. Explained below is the best KNN and Decision Tree model.

- **Decision Tree**: A decision tree is a flowchart like structure. The path from the root to leaf represent classification rules. The gini index was used to split the node to the most homogenous sub-nodes. Parameter were provided to prevent overfitting.
  In the case of our study, for decision tree the best model was found in the split 50:50. The algorithm had a confidence score of 55.79% which is basically the confidence of the model. The confusion matrix of the model is provided below. The classification accuracy is simply the number of correct predictions divided by all predictions or a ratio of correct predictions to total predictions.By using cross validation score method, the average accuracy for the decision is around 55%, and the standard deviation for the decision tree is about 0.06. And the classification error is among 45%.The classification accuracy is nearly same as the mean of accuracy

	precision	recall	f1-score	support
1	0.62	0.66	0.64	86
2	0.42	0.51	0.46	90
3	0.68	0.51	0.58	100
micro avg	0.56	0.56	0.56	276
macro avg	0.57	0.56	0.56	276
weighted avg	0.58	0.56	0.56	276

Fig 7:Classification Report

```
[[57 27  2]
 [22 46 22]
 [13 36 51]]
```

Fig 8 : Confusion Matrix

Precision- out of all examples for which we have predicted y = 1, how many are actually belonging to class 1
Recall - out of all examples actually belonging to class y = 1, how many have we predicted to be of class y = 1?
F1-Score - A number that can single-handedly be used to compare two model performances (a higher F1 score implies a better performance)

From above, there is 54% accuracy labeled your model as poor which actually belonged to poor class. 42% and 68% for 'average' quality class and 'good' quality class respectively.
For recall there is among 55% accuracy that we predicted to be poor quality class that actually belonged to poor class. 52% and 57% accuracy for average quality class and goo quality class respectively.
And for F1-score, there is 51%, represent an average of precision and recall for the poor class. And 62% for the good quality class. The higher F1 score implies a better performance.

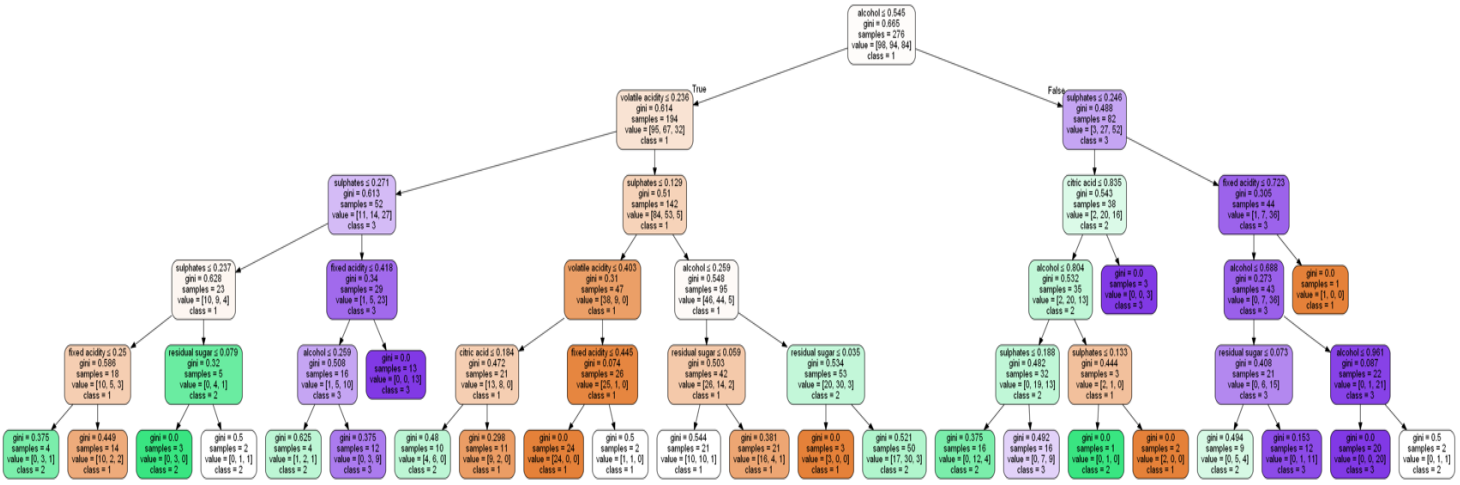


Fig 9: Decision Tree of Red Wine

We observed that our model first used alcohol concentration to make decisions. Also, note that if the alcohol content is high and the sulfate content is high, then the quality of the wine is likely to be "good."On the other hand, a wine with low alcohol and high volatile acidity is likely to be "poor" in quality.

- **K –Nearest Neighbours:** KNN is the simple yet powerful algorithm to categorize observation into classes. KNN considers various parameters such as no of neighbours, weights metric and p.
  In the case of our study the best split for K-NN was the 80-20 split. We achieved an accuracy of 61.26% with a classification error rate of about 38.7%.

	precision	recall	f1-score	support
1	0.54	0.81	0.65	26
2	0.50	0.28	0.36	39
3	0.72	0.78	0.75	46
micro avg	0.61	0.61	0.61	111
macro avg	0.59	0.62	0.59	111
weighted avg	0.60	0.61	0.59	111

Fig 10:Classification Report

```
[[21  5  0]
 [14 11 14]
 [ 4  6 36]]
```

Fig 11: Confusion Matrix

From above, there is 54% accuracy labeled your model as poor which actually belonged to poor class. 58% and 73% for 'average' quality class and 'good' quality class respectively.

For recall there is among 85% accuracy that we predicted to be poor quality class that actually belonged to poor class. 28% and 78% accuracy for average quality class and goo quality class respectively.

And form F1-score, there is 64%, represent an average of precision and recall for the poor class. And 76% for the good quality class. The higher F1 score implies a better performance.

Final Model: From the above two model taking into consideration the classification report and confusion matrix, we conclude that the best model is the K-nearest Neighbour which has an accuracy of 61.26%.

DISCUSSION

Form the above analysis of the red wine dataset with 1599 observations and 13 variables. The target feature was quality of wine which initially ranges from 1 to 10. However, we consider the problem as a classification problem by converting the target features into 3 ordinal values {good, bad, average}. The objective was to predict the quality of wine and to check a number of hypothesis. After observing different data distribution and matrix of the variables we proved that the hypothesis stating that low quality wine is better in quality was proved statistically insignificant.
The research question about how acidity and density effect quality was also answered by seeing now significant effect on the quality. One of the major struggles of the of the analysis was that the data was unbalanced as most of the values in quality existed in the average quality. The under - sampling method done to control this issue also limited the resources of the analysis.
From the decision tree algorithm which used Gini index to split the node also showed that wine with high alcohol and sulphate content normally tend to be better however low alcohol content and high volatile acidity normally tend to be of poor quality.

CONCLUSION

In this project as we have only limited ourselves with two algorithms for supervised classification problem we have been limited to find the best model for the prediction. This project can be taken further now by using new modelling techniques and by getting a new dataset where the data is distributed equally for the levels.

Finally, we conclude the project by listing the main results:

- High quality wine has more alcohol content.
- Density and Acidity has no apparent effect on alcohol
- K-NN with K =17 is the best model for this classification.

REFERENCES

- http://archive.ics.uci.edu/ml/datasets/Wine+Quality?ref=datanews.io%5D].
- Practical Data Science Notes – by Prof. Yongli Ren
- https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/