Data Modelling to Investigate the Quality of Wine

Practical Data Science (COSC 2670)

Assignment 2: Data Modelling and Presentation

Ancy Rex

S3760058

Master of Data Science **(MC267)**

RMIT University

Email: s3760058@student.rmit.edu.au

Pooja Suresh

S3749775

Master of Data Science **(MC267)**

RMIT University

Email: s3749775@student.rmit.edu.au

Date of Report: 2 June 2019

TABLE OF CONTENTS

Contents

EXECUTIVE SUMMARY

Wine being one of the most elegant drinks, can be differentiated by quality. Experts in wine tasting agree that the wine quality can depend on many other features including chemical balances and amount of sugar and water. The aim of this report was to investigate which type (Red/ White) wine has higher quality. In order to carry out the investigation, each of the physicochemical attributes were explored to understand how they affect the quality of wine. An exploration was also carried out to understand the relationship each of these attributes had amongst each other and how this ultimately affects the quality of the wine. The classification models of K-Nearest Neighbour and Decision Tree were used to model the dataset in hand and obtain the desired results.

From the exploration, it was observed that the Red Wine variant of the Portuguese "Vinho Verde" wine tend to have better quality as compared to the White Wine variant of the same. The reports conclude by indicating that the physiochemical aspects of the red wine variant make it better in quality, thus indicating that the red wine variant is better than the white wine variant. In order to make the white wine variant also better in quantity, it is recommended that the other aspects affecting the quality of a wine must be fine-tuned.

INTRODUCTION

Wine is an alcoholic beverage, often served at restaurants and fine dining functions. It is considered to be an item of luxury. Wine is generally classified as red and white wine. Red wine is manufactured from dark coloured grapes and white wines from yellow or green. Wines mature with age and so the oldest of wines tend to be of better quality and taste (Dhalia, N. 2018). According to experts in wine tasting, not only age but many other physicochemical attributes must be considered to determine the quality of wine. Attributes such as the water content, residual sugar or sulphates can help identify where the wine stands in terms of quality.

The aim of this investigation is to understand the type of wine that has a better quality. Wine tasting has become an antique process. Having physicochemical attributes such as fixed acidity, citric acid, volatile acidity, density, pH values and so on can help understand better how these attributes differ in the normal wines and the premium high-quality ones. The dataset obtained contains 12 different attributes, with each attribute indicating various properties of the wine such as the pH, density (water content), sulphates, chlorides to name a few of them. This report will discuss how each of these attributes were explored and observed to affect the final quality of the wine.

METHODOLOGY

DATA COLLECTION

The Wine Quality Dataset from the UCI Repository was chosen to carry out the data modelling. The repository consisted of 2 datasets – one for Red and one for White wine. The datasets were merged together and a 13th column, “type” was included, to indicate which type of wine the instance belonged to. The data in the dataset was then checked for its type using the dtypes function. It was found that all the attributes in the dataset were of object type, making them categorical values, while in the description of the dataset the attribute values were described to be real-time values. The attributes containing such real-time values were then converted to numeric type to carry out the exploration and modelling. The “quality” attribute of the dataset consisted of values ranging from 1 to 10, with 1 being the lowest quality of wine and 10 being the highest. It was observed that though the dataset contained instances with a quality of 1 or 2, they were very few in number. These values would ultimately affect the modelling as the data labels would be too many in number to carry out the modelling and get the desired accuracy. Thus, for the ease of exploration and to get better results with modelling, the data labels in the quality column were grouped as below:

- Values 1-5 indicate **low quality** (grouped as 5)
- Values 6 & 7 indicate **medium quality**
- Values 8-10 indicate **high quality** (grouped as 8)

DATA PREPARATION

The data preparation task is essential for the analysis as it saves time spent unnecessarily in fixing errors and strange values (Ren, 2019). The data preparation process for the Wine Quality dataset was done by first cleansing the data. The Wine Quality dataset was cleansed by taking into consideration the following factors:

- NULL,NaN and missing values:

  The Wine Quality dataset was first checked for missing values, which is nothing but a data value that is not stored for the variable of interest. Missing data present various problems including incorrect inferences from the data and biases in the estimation of parameters (J Anesthesiol, 2013). In order to determine the number of such values that are present in each attribute in the dataset, the isnull().sum() function was used, which shows the number of missing or null values present in each column. Luckily, none of the attributes of the Wine Quality dataset contained any such values.

- Redundant whitespaces:

  Redundant whitespaces are extra spaces that maybe present either at the beginning or end of the data. Such whitespaces need to be removed to ensure that they do not cause any mismatch. The redundant whitespaces in the Wine Quality dataset were handled and stripped using the strip() function.

- Data Entry Errors (Typos & Capital Letter Mismatch):

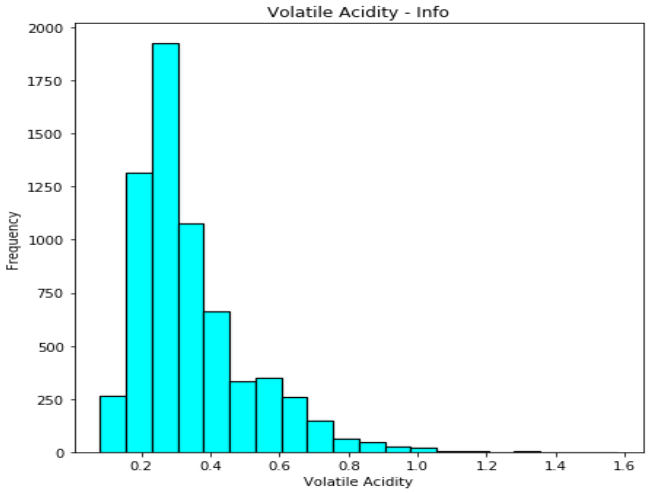
  Data collection and data entry are prone to errors. These errors could either be man-made or caused by machines (Ren, 2019). The Wine Quality dataset was checked for two types of data entry errors, namely, typos and capital letter mismatches. It was observed by listing out the columns using the value_counts() function that none of the columns in the dataset contained the above mentioned data entry errors.

- Sanity Checks:

  The data in the dataset may not always contain values that are of relevance to the analysis or could be values that may not be correct. These could include physically or theoretically impossible values (Ren, 2019). As all the attributes in the Wine Quality dataset are numerical values, sanity checks for impossible values were carried out on each of the 12 attributes and was found to have no impossible or insane values. An outlier check was also carried out on all the attributes and it was found that none of the attributes contained any observed outliers.

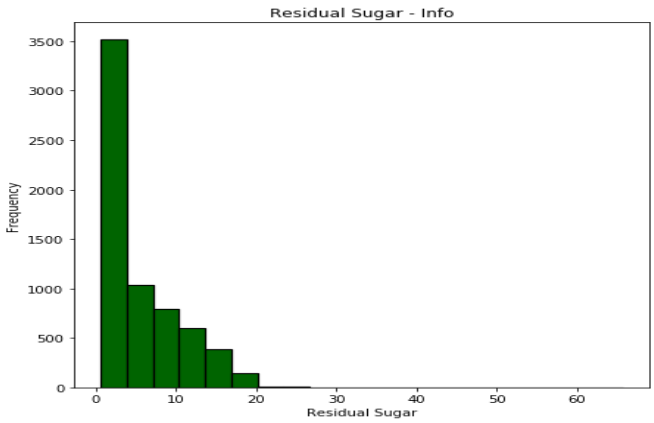
DATA EXPLORATION

Data exploration is the initial step in data analysis, where users explore a large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest (Sisense, n.d.). The aim of data exploration is to get a deep understanding of the data (Ren, 2019). The various physiochemical aspects of the dataset were explored individually, along with other properties and also with the quality column. The graphs depicted below describe few of the interesting ones observed from the exploration process



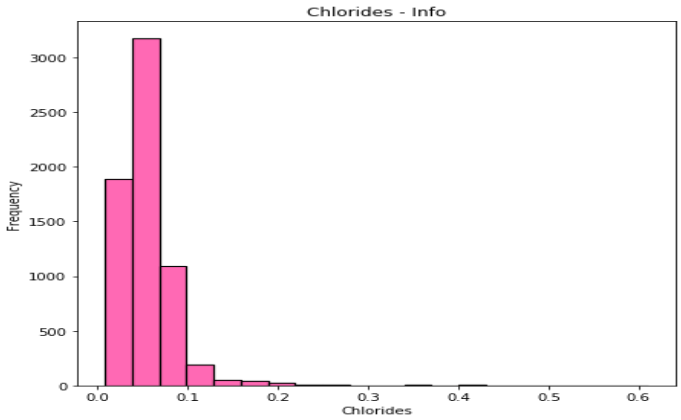
Volatile Acidity

Volatile acidity is the amount of acetic acid present in wine. If wine contain more amount of acetic acid then the wine would taste and smell like vinegar. From the graph, it can be observed that the volatile acidity has a maximum frequency of values at 0.3. The graph is also skewed towards the right.



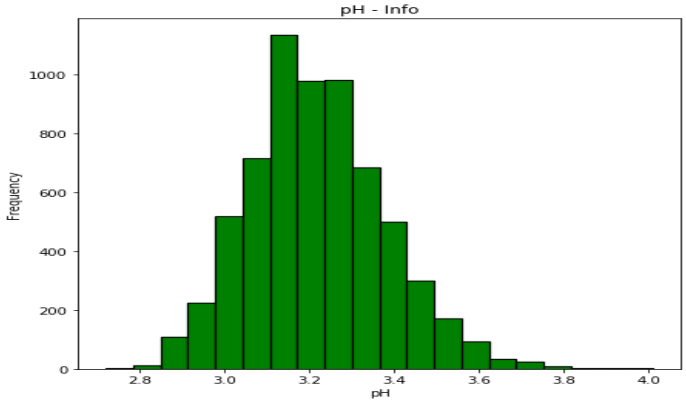
Residual Sugar

Residual sugar is the amount of sugar remaining after fermentation stops. Most wines tend not to have less than 1 grams/litre. Wines that have more than 45 grams/litre are considered too sweet. The graph depicted is skewed to the right and has the highest frequency of 3500 for the value between 0 and 3.



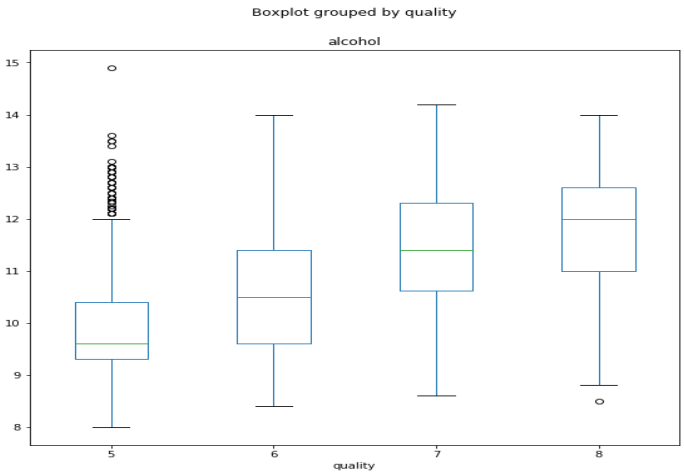
Chlorides

Chlorides are the amount of salt present in the wine. Experts claim that salt amount can help develop taste, colour and aroma. The graph is skewed towards the right.



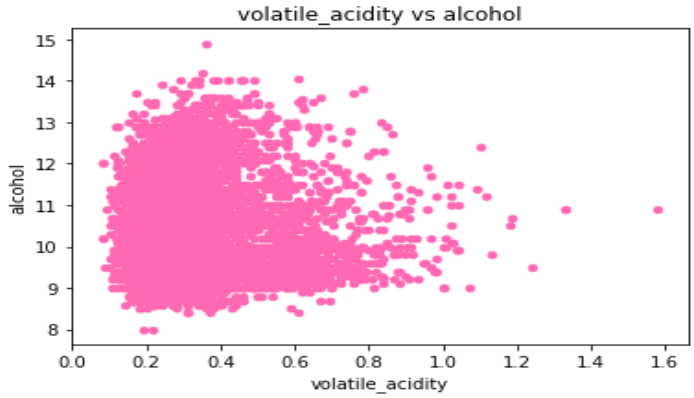
pH

pH levels tell us if the wine is acidic and basic. The pH scale differs from 1 to 14 where 0 is very acidic and 14 is very basic. Most wines tend to have pH value of 3-4. This graph is a symmetric and also depicts that a maximum number of wines have a pH between 3.0 and 3.2, indicating a normal pH.



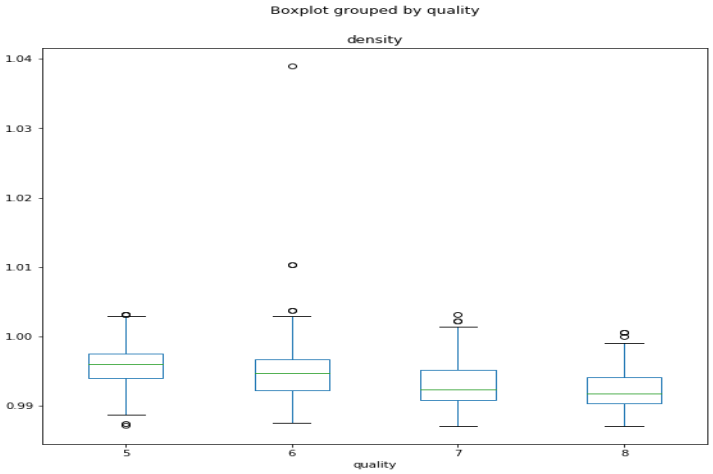
Alcohol vs Quality

Alcohol is the content or percentage of alcohol present in wine. To understand how the alcohol content differs according to each quality of wine, they were plotted against each other using a boxplot. It was observed that the mean of the higher quality of wine tend to be higher than that of lower quality wine, thus proving that alcohol content is greater in higher quality wine.



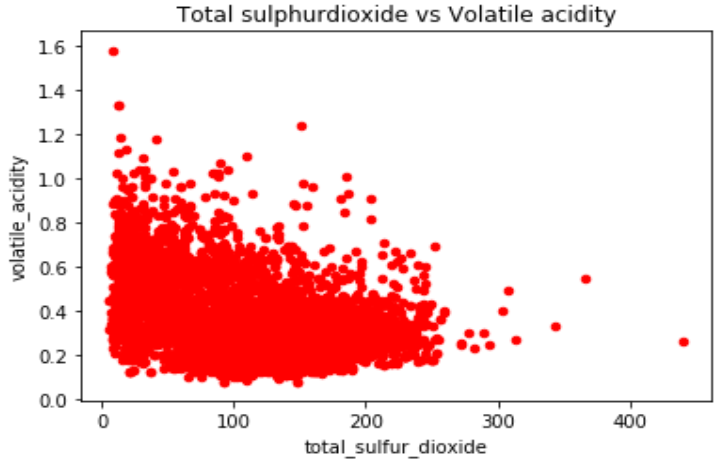
Volatile Acidity vs Alcohol

Volatile acidity is the amount of acetic acid in wine. If volatile acidity increases, then the wine would taste and smell more like vinegar. By plotting two attributes such as volatile acidity and alcohol helped indicate that when volatile acidity increases, the alcohol content decreases, thus proving that when the wine that taste more like vinegar contains lesser content of alcohol.



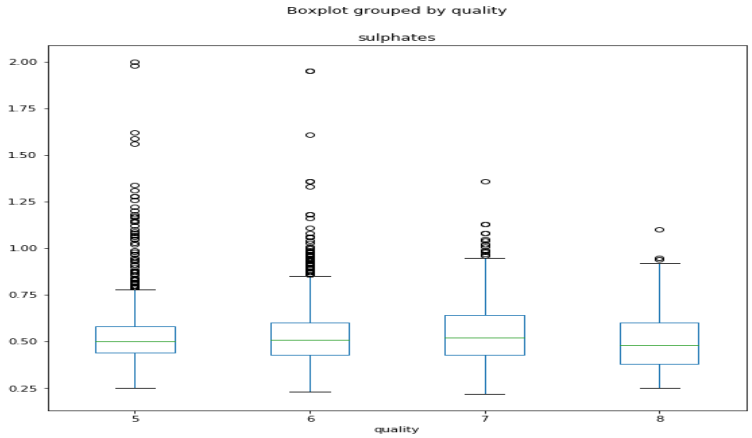
Density vs Quality

The water content present in the wine was represented in numerical form by the density attribute. To understand the water content present in low quality and high-quality wines, the density and quality attributes were plotted against each other using a box-plot. Using the graph, it was observed that the mean of the lower quality wines tends to have higher density values, proving that higher quality wines have lower water content than lower quality wines.



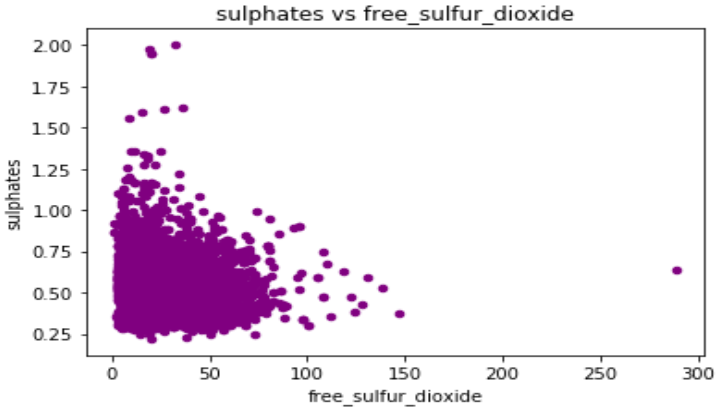
Volatile Acidity vs Total Sulphur Dioxide

Total sulphur-dioxide is the amount of free and bound forms of SO2. SO2 when in greater quantity can be detected in the nose and taste of wine. In lower concentrations they go undetected. When the amount of acetic acid was compared to that of SO2 content present in wine, it was observed that when the SO2 present in the wine increases it reduces the amount of volatile acidity.



Sulphates vs Quality

The sulphate content present in the wine contributes to SO2 levels which acts as the antimicrobial and antioxidant. They reduce the amount of fermentation that takes place in a wine. The quality and sulphate attributes were plotted against each other using a box-plot and it was observed that when sulphate content increases quality decreases. It can thus be concluded that these attributes are inversely proportional.



Free Sulphur Dioxide vs Sulphates

Free sulphur-dioxide is SO2 and bisulphites in free form. It helps prevent microbial growth and oxidation of wine. While Sulphate is the wine additive which acts as antioxidant and antimicrobial. While comparing them together we understand that when the free sulphur-dioxide increases in amount the sulphate amount decreases. Thus, the two attributes are inversely proportional.

DATA MODELLING

ENGINEERING FEATURES AND SELECTING A MODEL:

The first step towards data modelling is the selection of relevant features and choosing a model that best suits the data in hand. Feature Engineering is an essential part of the modelling process as the data maybe scattered among different sources and some transformation would be needed on the inputs in order to yield results with greater impact. The Wine Quality dataset also presented a similar situation, where the dataset was divided based on the type of wine as Red and White. The two datasets were combined and a new categorical column, “type” was included, which indicated the type of wine. The next step in the modelling process was Selection of a Model. In order to train the dataset and make it a good predictor, the Classification model was chosen to work on the dataset. Classification is a Supervised Learning Technique, used to identify which set of categories a new observation belongs to by analysing the dataset with the help of an already existent set of trained observations. This technique was chosen for the Wine Quality dataset as the dataset already consisted of a “quality” column that has predicted values of quality 5(least quality) to 8(highest quality) for each of the various combinations of the physiochemical properties of wine.

TRAINING THE MODEL:

Training of the model was done by selecting two of the most popular classifiers used, the K-Nearest Neighbour classifier and the Decision Tree classifier. Each of these classifiers were trained with three different sets of Train – Test Split (50-50,60-40,80-20) and the results were recorded.

K- Nearest Neighbour with 50%- 50% Train – Test Split:

The 50%-50% train-test set was trained by testing a range of values and combinations for the parameters:

- K value: The k value indicates the number of neighbours. The K value was tested for a range of values between 3 and 25. The ideal value of k that resulted in the best accuracy, precision and recall value for the above train-test set was 19.

- Weights: Weights indicate the distance between the neighbours. For the investigation, both the uniform and distance value of the weight parameter were tested and the distance parameter proved to have the best accuracy, precision and recall value for the above train-test set.

- p Value: The p value is the power parameter for the Minkowski metric. The p value for the analysis was tested for values between 1 and 50, and for the given train-test set, the p value of 1 resulted in the best accuracy, precision and recall value for the various combinations of other parameters.

- Observation: Based on the chosen parameters, it was observed that the classifier predicted the values of the different class labels in the target column with an accuracy of 66%. The weighted average of precision was observed to be 60% with the class label 8 having the maximum precision of 97% and the class label 7 having the minimum precision of 56%.

K- Nearest Neighbour with 60%- 40% Train – Test Split:

The 60%-40% train-test set was trained by testing a range of values and combinations for the parameters:

- K value: The K value was tested for a range of values between 3 and 25. The ideal value of k that resulted in the best accuracy, precision and recall value for the above train-test set was 15.

- Weights: For the investigation, both the uniform and distance value of the weight parameter were tested and the distance parameter proved to have the best accuracy, precision and recall value for the above train-test set.

- p Value: For the analysis, the p value for the analysis was tested for values between 1 and 50, and for the given train-test set, the p value of 1 resulted in the best accuracy, precision and recall value for the various combinations of other parameters.

- Observation: Based on the chosen parameters, it was observed that the classifier predicted the values of the different class labels in the target column with an accuracy of 61%. The weighted average of precision was observed to be 62% with the class label 8 having the maximum precision of 96% and the class label 6 having the minimum precision of 59%.

K- Nearest Neighbour with 80%- 20% Train – Test Split:

The 80%-20% train-test set was trained by testing a range of values and combinations for the parameters:

- K value: The K value was tested for a range of values between 3 and 25. The ideal value of k that resulted in the best accuracy, precision and recall value for the above train-test set was 15.

- Weights: For the investigation, both the uniform and distance value of the weight parameter were tested and the distance parameter proved to have the best accuracy, precision and recall value for the above train-test set.

- p Value: For the analysis, the p value for the analysis was tested for values between 1 and 50, and for the given train-test set, the p value of 1 resulted in the best accuracy, precision and recall value for the various combinations of other parameters.

- Observation: Based on the chosen parameters, it was observed that the classifier predicted the values of the different class labels in the target column with an accuracy of 64%. The weighted average of precision was observed to be 65% with the class label 8 having the maximum precision of 94% and the class label 7 having the minimum precision of 62%.

Feature Engineering - Use of Hill Climbing Technique for K-Nearest Neighbour Classifier:

The Hill Climbing Technique can help enable the K- Nearest Neighbour classifier perform better by filtering out the features in the given dataset based on the score obtained from using different combination of features. The technique selects the number of features to be used based on the highest score obtained. The Hill Climbing technique was used for 3 train-test split (50-50,60-40 &80-20) of the Wine Quality dataset. The 3 splits were each tested for different values of K, ranging from 1 to 25 and the results were observed as given below in the results section.

Decision Tree with 50%-50% Train-Test Split:

The 50%-50% train-test set was trained by testing a range of values and combinations for the parameters:

- Criterion: It is the function to measure the quality of split. The default criterion is 'Gini'. The gini value is used to create binary splits, while entropy is used for the information gain. For the investigation, entropy was used as it provides better accuracy, precision and recall value.

- Random State: As the name suggests, is used for initialising the internal random number generator, which will decide the splitting of data into train and test indices. The ideal value used was 10 as it provided better precision, accuracy and recall value.

The other chosen parameters were test size, max depth, min samples split, max features and min samples leaf which was not tuned as the given default values predicted the different class labels in the target column with an accuracy of 60%. The weighted average of precision was observed to be 60% with the class label 5 having the maximum precision of 69% and the class label 8 having the minimum precision of 37%.

Decision Tree with 60%-40% Train-Test Split:

The 60%-40% train-test set was trained by testing a range of values and combinations for the parameters:

- Criterion: Entropy was used as it provides better accuracy, precision and recall value.

- Random State: The ideal value used was 10 as it provides better precision, accuracy and recall value.

When considering a 60%-40% train-test split, the other parameters with default values predicted the different class labels in the target column with an accuracy of 62%. The weighted average of precision was observed to be 63% with the class label 5 having the maximum precision of 69% and the class label 8 having the minimum precision of 35%.

Decision Tree with 80%-20% Train-Test Split:

The 80%-20% train-test set was trained by testing a range of values and combinations for the parameters:

- Criterion: Entropy was used as it provides better accuracy, precision and recall value.

- Random State: The ideal value used was 7 as it provided with better precision, accuracy and recall value.

When considering an 80%-20% train test split, the other parameters with default values predicted the different class labels in the target column with an accuracy of 66%. The weighted average of precision was observed to be 67% with the class label 5 having the maximum precision of 72% and the class label 8 having the minimum precision of 38%.

RESULTS

- **Observations and Inferences from K-Nearest Neighbour Classifier:**

  From the various splits of train-test data, it is observed that the 80-20 split appears to give results that are of good precision as well as accuracy. Though the above classifier gives a good precision for the class label 8 which is the highest quality of wine, it cannot be considered as the better model because the class label 8 has the least support data among all the available class labels.

- **Observations and Inferences from K-Nearest Neighbour Classifier with Feature Engineering:**

  The table below gives a comparison of the classifier before and after feature engineering:

Train - Test Split	K-Value	Precision of Class Label 8		Weighted Average of Precision	
		Before Feature Engineering	After Feature Engineering	Before Feature Engineering	After Feature Engineering
50-50	19	97%	97%	60%	64%
60-40	15	96%	84%	62%	64%
80-20	15	94%	83%	65%	69%

From the comparison table it can be observed that after feature engineering, the weighted average of the precision scores increases for all the splits, however, the precision score of the class label 8 keeps decreasing. The 80-20 split of the classifier appears to produce a good weighted average as well as precision score for class label 8 when compared to the other class labels that have higher support data. Observing the confusion matrix for this model, we see that the class label 5 and 6 with more support data give a precision of 76% and 66% respectively. Thus, the 80-20 train-test split of the K-Nearest Neighbour Classifier with Feature Engineering gives better results as compared to other variations of the same model.

- **Observations and Inferences from Decision Tree:**

  The table below depicts the results obtained from Decision Tree Classifier using various train-test splits:

Train - Test Split	Accuracy		Precision			
	Gini	Entropy	Gini [Class Label Low]	Gini [Class Label High]	Entropy [Class Label Low]	Entropy [Class Label High]
50-50	60%	59%	32%	68%	37%	69%
60-40	62.10%	62%	35%	70%	36%	70%
80-20	66.30%	66%	38%	71%	38%	72%

From the table, it can be observed that the criterion Gini gives better accuracy as compared to the Entropy criterion for all splits of train-test data, however there is a slight variation in the precision scores of the class labels while using Entropy. The precision scores are higher while using Entropy as those compared to the ones obtained by using Gini. In accordance with the support data provided for each of the class labels, the Decision Tree Classifier with 80-20 Train-Test split appears to predict the results more accurately for the class labels with higher support data as compared to those with lower support data.

DISCUSSION

- The main aim of this experiment was to determine which type of wine has the better quality.
- The quality class labels range from 3 to 8, with 3 being the lowest quality and 8 being the highest.
- From the results of the K-Nearest neighbour classifier, it is clear that the class label 8 is predicted better and has a higher precision as compared to the other class labels. This model can be useful in achieving the research goal mentioned above. Having said that, it is also essential to make note of the fact that class label 8 has the least amount of support data as compared to the other class labels for every variation of the train-test split considered.
- Thus, it is evident that though the K-Nearest Neighbour classifier has a better precision as compared to the Decision Tree Classifier and also answers the research goal, it cannot be considered as it does not provide accurate results when it has to predict the other class labels.
- The Decision Tree classifier on the other hand has a lower precision value for the class label 8 when a lower amount of support data is available and a higher precision rate for the class labels that comparatively have larger support data.
- The results of the Decision Tree classifier with 80-20 train-test split and a criterion of Entropy result in 70% precision for the class label with the highest amount of support data and 33% precision for the class label with the lowest amount of support data, thus proving that the classifier is capable of predicting the results more accurately as compared to the K-Nearest Neighbour classifier.
- Thus, it can be inferred that the Decision Tree Classifier with 80-20 train-test split is the better Classification technique for the given dataset.
- However, applying this model to the research goal (unseen data) would not yield the desired results as the model is able to predict only low and moderate quality wines accurately.
- If there was access to more support data for the values with class label 8 (high quality), the above model could be trained to predict the results of high-quality wine more accurately and ultimately prove the research goal.

CONCLUSION

The experiments carried out on the Wine Quality dataset suggest that Decision Tree Classifier with 80-20 train-test split provides more accurate and efficient results as compared to the other variants of the same classifier as well as the K-Nearest Neighbour classifier. As there is an availability bias in the amount of data that is in hand to support the prediction of high-quality wines, this model cannot be considered to effectively predict our research goal. In order to achieve our research goal, it is imperative to have more data for the class label indicating high quality wines. This would enable the model to be trained better to predict the class label 8.

The results of this experiment can however be applied to a different research goal, such as identifying the type of wines that have low or medium quality. They could also be applied to predict the ideal combination of physiochemical properties that would yield the better quality of wine.

REFERENCES

1. Dhalia, N. (2018). The Effect Of Physicochemical On The Wine Quality. Data Analytics with Manegerial Application Internship. [online] Uttarakhand. Available at: https://rstudio-pubs-static.s3.amazonaws.com/354404_c136afece292494593e4632ec8a2d65c.html [Accessed 1 Jun. 2019].
2. Lee, S., Park, J. and Kang, K. (2015). Assessing wine quality using a decision tree. IEEE International Symposium on Systems Engineering. [online] Available at: https://ieeexplore.ieee.org/document/7302752/citations#citations [Accessed 1 Jun. 2019].
3. Rashid, M. (2015). Wine Quality Exploration. [online] Available at: http://rstudio-pubs-static.s3.amazonaws.com/80458_5000e31f84df449099a872ccf40747b7.html [Accessed 1 Jun. 2019].
4. Ren, Y. (2019). Data Curation - 2019r.
5. Sisense. (n.d.). Data Exploration. [online] Available at: https://www.sisense.com/glossary/data-exploration/.
6. J Anesthesiol, K. (2013). [ebook] The prevention and handling of the missing data. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/.

Dataset Reference:

7. Archive.ics.uci.edu. (2019). UCI Machine Learning Repository: Wine Quality Data Set. [online] Available at: https://archive.ics.uci.edu/ml/datasets/wine+quality.