Utsav Kailash Kothari
Student No: 210100637

Abstract:

This research looks at efficient learning algorithms for estimating the quality of red and white wines. Then, broad trends for each type of wine were looked at. Analysis of the quality distribution revealed that the majority of the wines scored between 5 and 6, with very few wines falling outside of those ranges. White wines also had a reputation for being of superior quality and had a greater alcohol level than red wines. While there was no discernible relationship between quality and sweetness, red and white wines showed a positive correlation between alcohol concentration and quality. Moreover, density showed a significant association with both residual sugar and alcohol.

The classification model with the brilliant f1 and MSE/RMSE scores among the used models, the random forest classifier, showed the most positive findings. Yet, despite somewhat similar performance, logistic regression may be less prone to errors. The regression models might not be as efficient and reliable due to being error-prone. Regression models might not be the best option for this specific result set, though, as certain characteristics have a relatively small number of data points. While using SMOTE to address the problem of unbalanced data, similar problems were discovered when predicting all features.

What Was Done and How:

This research project was conducted using Python modules, particularly Pandas, to perform data aggregation, summarization, and graph plotting based on the given dataset. The analysis focused on examining the correlations between various variables in the dataset to identify the relationships between them. Pearson's correlation coefficient was used to quantify the strength and direction of the correlation. The analysis involved plotting large amounts of data to visualize the relationships between variables. The results showed that there were significant correlations between quality, alcohol content, and residual sugar. Furthermore, various modelling methods, such as linear regression, binary classification, and decision trees, were employed to analyze the data and gain insights into the factors that affect the perceived quality of red and white wines. Despite the limitations posed by the dataset, the findings provide valuable insights into the distribution of wine quality, alcohol content, sweetness, and their effects on the perceived quality of wine.

Quality Distribution and Discritising It:

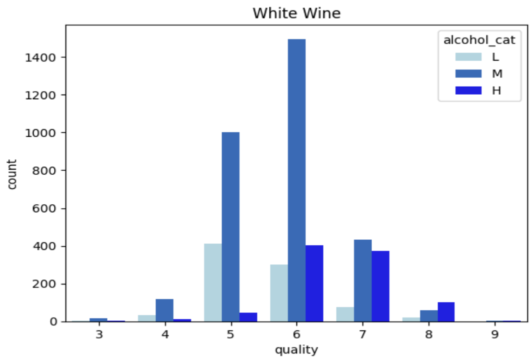
As part of the initial data exploration, one of the key tasks was to understand the distribution of wine quality across samples. This was achieved through the use of countplot to visualize the distribution of quality scores for both red and white wines. A preliminary analysis of the plots revealed that the majority of wines were classified between 5 and 7, regardless of the type of wine.

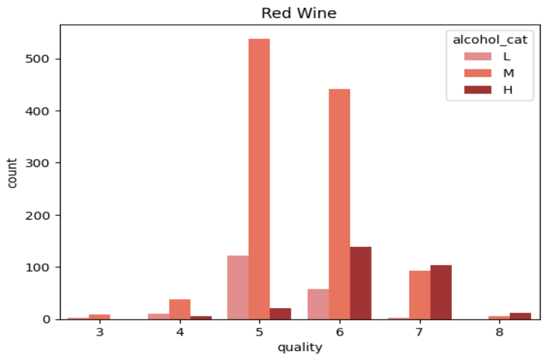
However, when examining the distribution more closely, it became apparent that there were some differences between red and white wines. Specifically, red wines were more heavily concentrated around the 5 and 6 quality marks, with a sharp drop-off in the number of wines with a quality score of 7 or above. In contrast, white wines had a more even distribution across quality scores, with a greater proportion of wines scoring 6, 7, 8, and even 9(which red did not have).

Overall, these initial findings suggest that there may be some important differences in the quality distribution of red and white wines. Further analysis was conducted to explore the factors that may be driving these differences and to develop predictive models for wine quality based on various features of the data.

Quality Against Alcohol Content:

I added additional columns with discretized alcohol content into L, M, and H categories to both data frames using standard deviations to improve the analysis of the data. So first, I looked at the link between alcohol quantity and quality before exploring the new variable. A definite positive connection was seen when the two datasets were combined and plotted against one another, with a Pearson's R-value of 0.44 for white wine and 0.49 for red wine. These results demonstrated great statistical significance. It became clear from looking at the discretized alcohol content that there weren't many wines with a high alcohol content that was below grade 6. Moreover, from grade 4 onward, the representation of wines with low alcohol concentration declined as quality rose. When it came to white wine, the distribution of high-alcohol wines was predominant in quality 7 and above. The analysis of the data was further enhanced by the inclusion of the alcohol discretization variable, which gave a more thorough knowledge of the connection between alcohol concentration and quality further on.

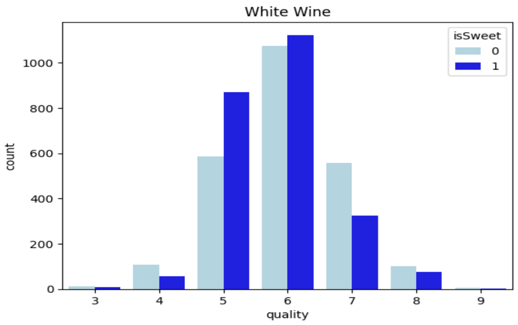


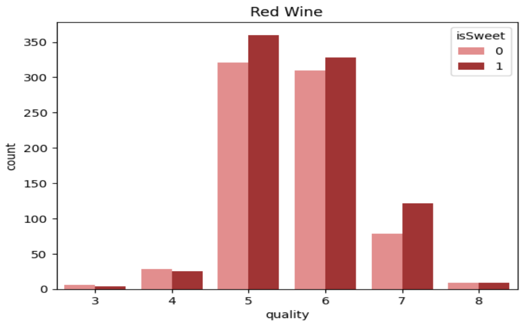


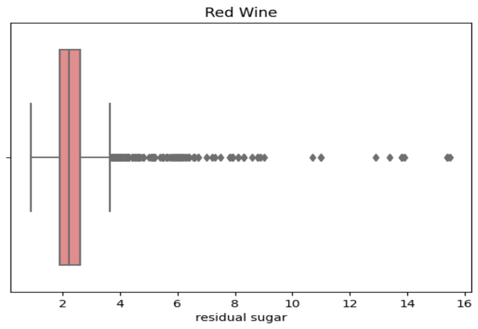
Quality Against Sweetness With Regards to Residual Sugar Variable:

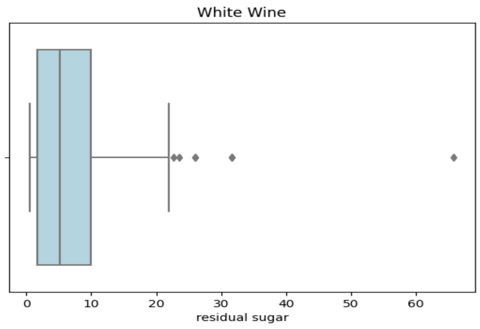
To analyze the distribution of residual sugar in the wines, I used a box plot. This was an appropriate type of graph to use because it shows the distribution of the data and highlights any outliers. Before creating the box plot, I had some prior knowledge that white wines are generally much sweeter than red wines. The results of the box plot were consistent with this, as it was evident that white wines had a wider distribution of residual sugar levels with many more wines at a higher level of residual sugar than red wines. The majority of red wines were clustered around the 2.5 mark, indicating that they were generally drier than white wines. In addition to the box plot, I created histograms for each wine individually to determine if the wine was sweet or not (1 or 0). This helped me to better understand the variance in the data and determine the appropriate thresholds for sweetness. Overall, these visualizations helped to highlight the significant differences between the two types of wines in terms of residual sugar levels.

I discretized the residual sugar variable into either sweet or dry depending on the mean of each dataset to evaluate the association between the residual sugar and wine quality. I initially plotted quality against residual sugar using a bar chart and then used the resulting "isSweet" variable to investigate the effect of residual sugar on quality. Research showed that there was no obvious association between the two factors. A Pearson correlation test, however, revealed that white wine had a strong statistical significance with an R-value of -0.098, indicating that higher-grade white wines often tend to be dryer, whereas red wine exhibited no significant link.





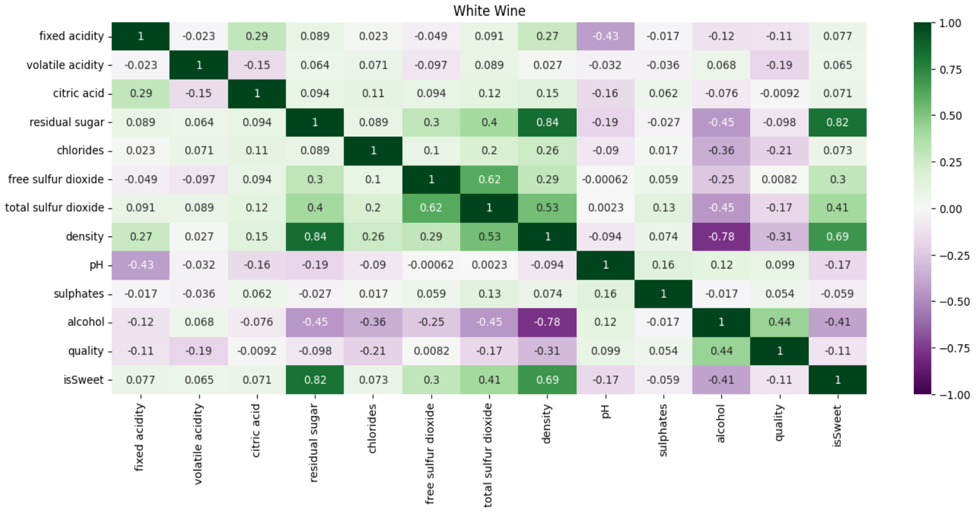


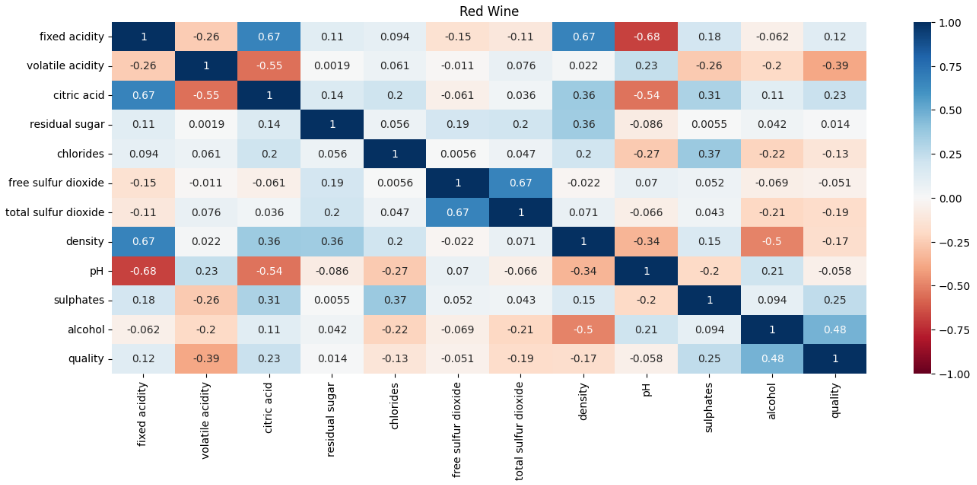


Analysing Correlations:

I used the pairplot function to create scatterplots of all attribute combinations for both white and red wines to better understand the correlations found in the dataset. Each plot also included a regression line to help visualise the correlations. At first, several relations were clearly discernible, including a clear inverse relationship between fixed acidity and pH, alcohol, and total sulphur dioxide in both wines, a very strong positive relationship between residual sugar and density, especially in white wine, and a positive relationship between fixed acidity and density, chlorides, and density. Despite the pairplot's value, it can be difficult to read and easy to overlook some less evident relationships. As a result, a correlation analysis was also developed to offer a more thorough study of the relationships that were already there. The matrix offers a more thorough study of the dataset, although some of the correlations might not be as useful in certain cases.

To see the relationships between various characteristics and wine quality, I used heatmaps from seaborn. Given that the majority of the data were continuous and linearly connected, I concluded that Pearson's correlation coefficient was the best approach to adopt. While Spearman's and Kendall's approaches were also taken into consideration, Pearson's correlation coefficient was selected since it offered a clearer analysis. The Kendall matrix, which had smaller R values than the Spearman's matrix, followed a similar pattern, despite a few minor changes. After that, I looked at Pearson's matrix to determine which characteristics would be most helpful in machine learning. I discovered that the qualities of white and red wine did not correlate in the same ways, necessitating a second analysis. In white wine, there was a strong positive link between alcohol and quality and a significant negative correlation between density and alcohol. For machine learning models, it is essential to comprehend the relationships between attributes and wine quality, and the best correlation approach should be carefully chosen based on the type of data.





Determining Quality:

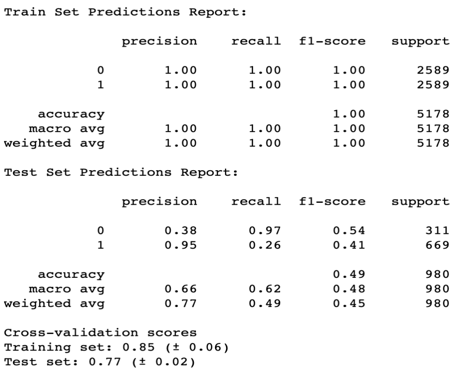
After finishing the first study, I moved on to machine learning and discretized the quality outcome into low and high categories in order to treat it as a classification issue. I first played around with several quality criteria before deciding to choose 6 (LQ < 6 and remaining as HQ i.e. >=6) as the criterion. This choice was made because it was discovered that the LQ to HQ ratio was skewed when 7 was used as the threshold, especially in the case of red wine.

To produce more HQ samples, I also split the data and used SMOTE. The number of HQ samples increased as a result, but the model accuracy remained similar. Instead, using 6 as the threshold yielded higher accuracy when the models were tested on the test set. However, it should be noted that applying this criteria led to the majority of the wines in the data set being labelled as "LQ," which could be a concern. The choice of the threshold, however, is arbitrary and is based on one's definition of high- and low-quality wines. In actuality, it is best to consult wine specialists to establish the proper threshold.

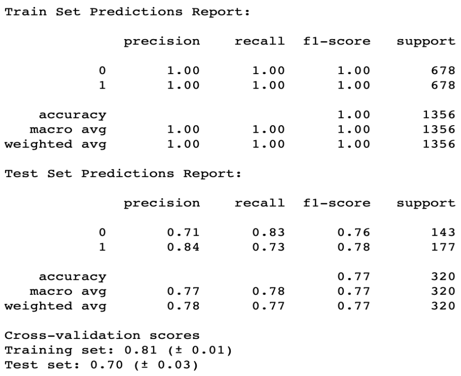
To address the class imbalance in data sets, the SMOTE method creates synthetic samples of the minority class. In this instance, it was used to generate extra high-quality samples in order to balance out the data set. This is crucial to avoid biased models that inaccurately anticipate the minority class.

Many machine learning models were run on the data, including logistic regression, K-Nearest Neighbours Algorithm to see if SMOTE affected the outcomes. This is a crucial stage since it helps to guarantee that the models are neither overfitting nor underfitting and that the findings are more reliable. The f-1 score typically increased after using SMOTE, suggesting improved analysis and results.

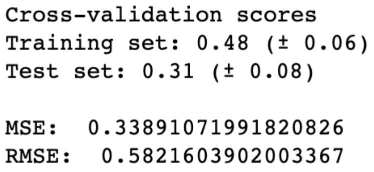
It is intriguing to observe that decision trees and random forests did not significantly improve, which may be because they tend to oversample and incur errors.
Using different models on the data after SMOTE can assist provide more precise and reliable findings. SMOTE was an essential approach for addressing the class imbalance in data sets.



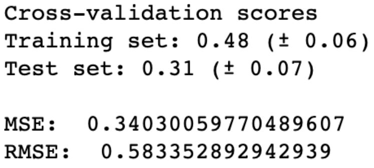
```
Train Set Predictions Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2589
           1       1.00      1.00      1.00      2589

    accuracy                           1.00      5178
   macro avg       1.00      1.00      1.00      5178
weighted avg       1.00      1.00      1.00      5178

Test Set Predictions Report:

              precision    recall  f1-score   support

           0       0.38      0.97      0.54       311
           1       0.95      0.26      0.41       669

    accuracy                           0.49       980
   macro avg       0.66      0.62      0.48       980
weighted avg       0.77      0.49      0.45       980

Cross-validation scores
Training set: 0.85 (± 0.06)
Test set: 0.77 (± 0.02)
```



```
Train Set Predictions Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       678
           1       1.00      1.00      1.00       678

    accuracy                           1.00      1356
   macro avg       1.00      1.00      1.00      1356
weighted avg       1.00      1.00      1.00      1356

Test Set Predictions Report:

              precision    recall  f1-score   support

           0       0.71      0.83      0.76       143
           1       0.84      0.73      0.78       177

    accuracy                           0.77       320
   macro avg       0.77      0.78      0.77       320
weighted avg       0.78      0.77      0.77       320

Cross-validation scores
Training set: 0.81 (± 0.01)
Test set: 0.70 (± 0.03)
```



```
Cross-validation scores
Training set: 0.48 (± 0.06)
Test set: 0.31 (± 0.08)

MSE:  0.33891071991820826
RMSE:  0.5821603902003367
```



```
Cross-validation scores
Training set: 0.48 (± 0.06)
Test set: 0.31 (± 0.07)

MSE:  0.34030059770489607
RMSE:  0.583352892942939
```

Running Models:
Two different types of models were used: Classification and Regression models. Firstly, Classification models where K-Nearest Neighbors (KNN), Decision Tree, Random Forest Tree (RFT), Hyperparameter Tuning, Radial Basis Function, and Support Vector Machine classification models, both linear and poly were chosen. RFT outperformed the other models in tests for a variety of parameters and thresholds. In the white wine test set, it scored 76% accuracy, but just 54% on the red wine test set. The RFT model, however, achieved 100% accuracy in the f1 score on the training set for both wines, indicating the overfitting difficulties. Random forest models are prone to erroneous, which may lead to reduced accuracy when used with less comparable datasets. Because the random forest is a better form of decision tree than the decision tree model, the accuracy of the decision tree model was lower, as predicted. As logistic regression exhibits accuracy that is equivalent to RFT, it could be a useful alternative model to utilise if errors become a major issue.

I used a variety of models for regression analysis, including regression modelling, linear and polygonal logistic regression, radial basis function (RBF), and support vector machine classification models (SVM). Yet, because the quality is at fixed intervals, the regression models performed similarly to classifiers in terms of performance. I used MSE (Mean Square Error) and RMSE to assess the models' accuracy (Root Mean Square Error). RMSE is measured in the same units as the target variable and is the square root of MSE. Finding the effect of greater mistakes on the model's overall accuracy can be achieved through the use of the difference between MSE and RMSE. Here also Random Forest outperformed the other models in MSE and RMSE, earning 34% and 58% accuracy for the two wines, respectively. This comparison can help to determine the best model to use based on the level of accuracy required for the specific analysis.

Conclusions and Evaluation:

Overall, the research offers a thorough examination of a dataset on wine quality using different machine-learning algorithms. The approach adopted is thoroughly explained in the report, including the preprocessing stages and the justifications for selecting particular models. Another asset of the paper is the extensive use of visualisation, which makes the conclusions more approachable and understandable.

The way the findings are presented may use some refinement. Although the report does offer numerical values for accuracy ratings and other metrics, additional context of these values would be beneficial. For instance, how do they stack up against related studies or industry standards? Additionally, some of the inferences made from the findings may be more directly related to the study.

The way the dataset handles class imbalance is another area that needs work. While SMOTE was utilised to solve this issue, it would have been advantageous to investigate additional strategies and evaluate their efficacy. The paper also would have benefitted from a more thorough assessment of the potential effects of class imbalance on the model's accuracy.

The report's extensive investigation of several machine learning models, including classification and regression, is one of its strongest points. It is admirable that a discussion of the advantages and disadvantages of each model and how well they match the study topic was included. The examination is further made more thorough by the use of multiple assessment measures, such as MSE and RMSE.

Overall, the work constitutes an important contribution to the study of machine learning methods for analysing wine quality. The research shows a strong approach and in-depth analysis, yet there are still certain areas that may be improved.

Further Work:

Quality Control: Wine production can benefit from the application of the machine learning models created in this study for quality control. Winemakers can forecast the quality of the wine and take action to enhance it by looking at its chemical and physical features. This can aid in early problem detection during the wine production process and result in higher-quality wine production.

Marketing: This report's classification models can be applied to marketing campaigns. Based on its chemical and physical characteristics, the models can forecast the wine's quality, which may be used to inform buyers about the wine they are purchasing through promotional materials. The models can also assist in identifying the crucial elements that go into the manufacture of high-quality wines, which can then be utilized in the marketing and branding of the wine.

References:

1. Madhavan, S. (2019) Learn classification algorithms using Python and scikit-learn, IBM developer. Available at: https://developer.ibm.com/tutorials/learn-classification-algorithms-using-python-and-scikit-learn/ (Accessed: March 17, 2023).
2. Dua, R., Ghotra, M.S. and Pentreath, N. (no date) Machine learning with spark - second edition, O'Reilly Online Learning. Packt Publishing. Available at: https://www.oreilly.com/library/view/machine-learning-with/9781785889936/669125cc-ce5c-4507-a28e-065ebfda8f86.xhtml#:~:text=MSE%20is%20the%20average%20of,the%20square%20root%20of%20MSE. (Accessed: March 17, 2023).
3. 1. what is the jupyter notebook?¶ (no date) 1. What is the Jupyter Notebook? - Jupyter/IPython Notebook Quick Start Guide 0.1 documentation. Available at: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html#notebook-document (Accessed: March 17, 2023).
4. Users guide# (no date) Users guide - Matplotlib 3.7.1 documentation. Available at: https://matplotlib.org/stable/users/index.html (Accessed: March 17, 2023).
5. User guide and tutorial# (no date) User guide and tutorial - seaborn 0.12.2 documentation. Available at: https://seaborn.pydata.org/tutorial.html (Accessed: March 17, 2023).