Wine Quality Prediction

CSE523: Machine Learning

Anshi Shah AU2040087

Rahi Shah AU2040070

Kenil Shah AU2040111

Yesha Dhivar AU2040215

***Abstract*— Consumers and producers both value the quality of wine, and wine quality is used to be assessed only after production, advancements in technology and the abundance of data. This has led to the use of machine learning techniques for evaluating wine quality during the development stage. The goal of this work is to use machine learning to identify the most significant parameters that control wine quality and predict wine quality based on those parameters.**

***Keywords*— Wine Quality, Machine Learning, Correlation, Features, Red Wine, White Wine**

I. Introduction

Globally, wine is the most commonly consumed beverage, and its societal significance is highly regarded. The quality of wine is crucial for both consumers and producers in the current competitive market as it directly impacts revenue. However, determining quality based on personal taste can be challenging. To address this, manufacturers have incorporated technology in the development phase to assess wine quality using various devices, saving time and money while accumulating significant data on production parameters. In recent years, machine learning techniques have been successfully applied to analyse this data and optimize the parameters that influence wine quality. This approach enables manufacturers to fine-tune the quality of their wine and even create new brands with distinct tastes. Therefore, it is vital to analyse the fundamental parameters that determine wine quality.

II. Literature Survey

The work by Sunny Kumar and all shows that support vector machine model gives the most accurate results, outperforming the other models tested. Here, the support vector machine algorithm performs better than Random Forest Algorithm and then comes the Naïve Bayes Algorithm.

The work by Paulo Corte and all have made use of different techniques like Random Forest, Support Vector Machine and Naïve Bayes. The best out of these is decided by the performance over training set and test set.

III. Research Methodology

For the research, the data was retrieved from UCI Machine Learning repository. The dataset consists of 1599 instances with 12 variables such as fixed acidity, citrus acid, volatile acidity, residual sugar, chlorides, thickness, free and absolute sulphur dioxide, pH, alcohol, and sulphates. The quality rating ranges from poor (3) to excellent (8).

To evaluate the performance of the machine learning algorithms, a confusion matrix is calculated, which is a table that shows how well the classification model predicts the outcomes. The confusion matrix is used to calculate relevant performance measures such as accuracy and precision.

Different machine learning algorithms are compared based on the accuracy predicted on this dataset. The algorithms used in the research include Decision Trees and Logistic Regression methods. The results show the accuracy and the errors in each of the methods for both white and red wine dataset.

IV. Implementation

An analysis was performed on both the red and white wine datasets. Data cleaning was performed and the statistical measures were observed to find the outliers and trends of the data.

Catplots were used to visualize the spread of the ‘quality’ column of the dataset. The quality was ranked from ‘3’-worse to ‘8’-best. The plots are as shown below:

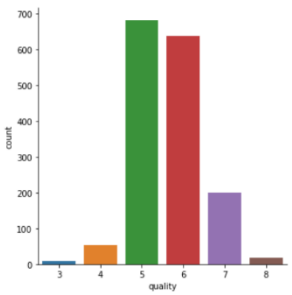

Figure 1: Quality: Red Wine

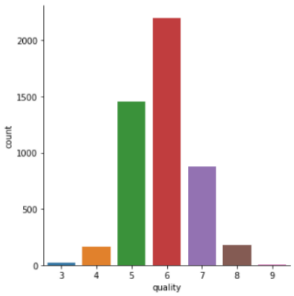

Figure 1: Quality: White Wine

Next, the proportionality of all other features with respect to the ‘quality’ column was measured. Feature engineering was performed on the dataset. Feature selection was done based on the data analysis performed and the correlation matrix. The results of the correlation matrix shows that sulphates, alcohol, volatile acidity, and density are more influencing the quality of the wine compared to other features.

Here shown is the correlation matrix for red wine dataset. A similar matrix is also made for white wine data. Based on this, features are selected for training the model.

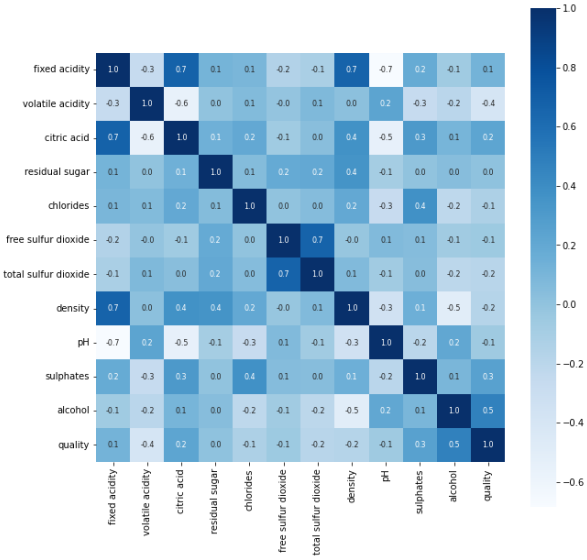

Figure 3: Covariance Matrix for Red Wine

After feature selection, logistic regression model and decision tree model was formed based on the conclusions derived from the same. The values of the quality were recoded from 3-8 to '0' or '1' to indicate the 'Good' or 'Bad' quality wine. The training set contained 80% of the data and the models were trained on it.

In logistic regression the 'sulphates' and 'alcohol' are the predictors used and the recoded 'quality_c' variable as the target.

A logistic regression model is then trained on the training data using the scikit-learn LogisticRegression class. The function then uses the trained model to make predictions on the testing data and prints the confusion matrix and accuracy score of the model. Next, a decision tree classifier model is then trained on the training data using the scikit-learn DecisionTreeClassifier class. The function then uses the trained model to make predictions on the testing data and prints the confusion matrix and accuracy score of the model. Both the models also print the model score and the root mean squared error (RMSE) of the predictions.

V. Results and Conclusion

The logistic regression model estimates the probability of the outcome variable 'quality_c' based on the values of the predictor variables (sulphates and alcohol). The model coefficients are estimated using the training data and are used to make predictions on the testing data.

The accuracy of the model is around 76% for red wine and 75% for white wine.

```
Red Wine

Confusion Matrix:
 [[118  39]
 [ 37 126]]
Accuracy: 0.7625
Score: 0.7625
RMSE: 0.48733971724044817
```

```
White Wine

Confusion Matrix:
 [[185 154]
 [ 86 555]]
Accuracy: 0.7551020408163265
Score: 0.7551020408163265
RMSE: 0.4948716593053935
```

Figure 2: Logistic Regression Results

The code for decision tree model performs a decision tree classification analysis on a red wine dataset. First, it recodes the 'quality' variable into two groups (0 and 1) based on a predefined mapping. Then, it splits the dataset into training and testing sets and uses the 'residual sugar' and 'alcohol' variables as predictors and the recoded 'quality_c' variable as the target.

The accuracy of the model is around 70% for red wine and 71% for white wine.

```
Red Wine

Confusion Matrix:
 [[109  51]
 [ 45 115]]
Accuracy: 0.7
Score: 0.7
RMSE: 0.5477225575051661
```

```
White Wine

Confusion Matrix:
 [[ 387  254]
 [ 298 1021]]
Accuracy: 0.7183673469387755
Score: 0.7183673469387755
RMSE: 0.5306907320287632
```

Figure 3: Decision Tree Results

The confusion matrix in the results shows the number of true positives, true negatives, false positives, and false negatives. The accuracy score represents the proportion of correct predictions (i.e., the number of true positives and true negatives divided by the total number of observations). The model score provides a measure of how well the model fits the testing data, while the RMSE measures the difference between the predicted values and the actual values.

VI. Bibliography

Paulo Cortez, A. C. (2009, November). *Science Direct*. Retrieved from Modeling wine preferences by data mining from physicochemical properties: https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377?via%3Dihub#preview-section-references

Paulo Cortez, U. o. (2009). *UCI Machine Learning Repository*. Retrieved from Wine Quality Dataset: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Sunny Kumar, K. A. (2020, June). *IEEE Xplore*. Retrieved from Red Wine Quality Prediction Using Machine Learning Techniques: https://ieeexplore.ieee.org/abstract/document/9104095