Wine Quality Prediction

CSE523: Machine Learning

Anshi Shah	Rahi Shah	Kenil Shah	Yesha Dhivar
AU2040087	AU2040070	AU2040111	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 such as decision tree, KNN classification, Support Vector Machine model 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, Logistic Regression, Hypothesis Testing and ANOVA, KNN-Classification and SVM methods. The results show the accuracy and the errors in each of the methods for both white and red wine dataset.

IV. Implementation

A. *Dataset*

The dataset consists of samples of vinho verde red and white wine. This dataset is taken from UCI Machine Learning repository. There are 11 features excluding the ‘quality’ feature, that are fundamental parameters of wine. There are around 1600 samples of red wine and 4900 samples of white wine.

B. *Problem Statement*

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.

C. *Pre-processing Steps*

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. The data was also normalized before training on specific model. The standard MinMax Scaler was used to normalize the data.

D. *Exploratory Data Analysis*

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:

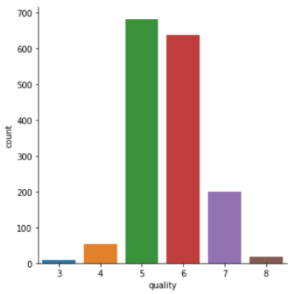

Figure 1: Quality: Red Wine

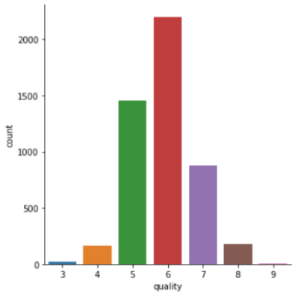

Figure 2: Quality: White Wine

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.

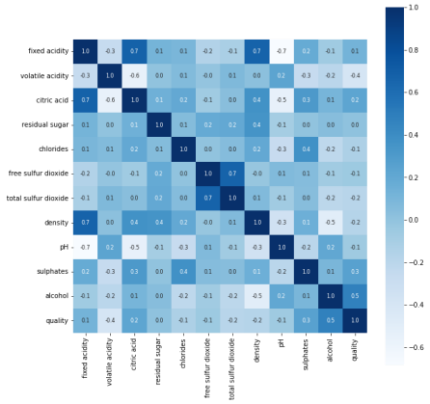

Figure 3: Covariance Matrix for Red Wine

A pair plot was also plotted for all the 11 features In which the scatter plots are displayed in a grid format, with the diagonal of the grid showing the distribution of each individual feature (in this case, displayed as a kernel density estimate). The scatter plots in the upper triangle of the grid show the scatter plots for pairs of features, with different colors indicating the two classes (0 or 1) of the target variable ('quality_c').

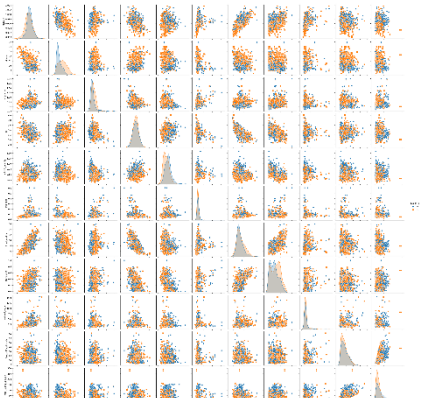

Figure 4: Pair Plot

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.

E. Model Training

Before training the models, the quality column is recoded from '0-9' to '0' and '1' based on predefined mapping. It changes the multiclass classification to binary classification.

Different models are then trained on the processed dataset. 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. Next a KNN model is then trained on the training data using the scikit-learn KNeighborsClassifier class. Next, a random forest classifier is then trained on the training data using the scikit-learn RandomForestClassifier class. A SVM model is also trained on the dataset with a linear kernel. Grid Search CV was used to find the appropriate parameter for the SVM model, but due to higher processing power need, it was not successful to show the results. Confusion matrix, model plots and accuracy score for every model are then compared to find the most accurate prediction and eventually the model and the features used.

V. RESULTS

A. Logistic Regression

The accuracy of the model is around 76% for red wine and 75% for white wine.
Features used: sulphates, alcohol, volatile acidity

```
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 5: Logistic Regression Results

B. Decision Tree

The accuracy of the model is around 70% for red wine and 71% for white wine.
Parameters used: residual sugar, alcohol

```
Red Wine                          White Wine

Confusion Matrix:                 Confusion Matrix:
 [[109  51]                        [[ 387  254]
 [ 45 115]]                        [ 298 1021]]
Accuracy: 0.7                     Accuracy: 0.7183673469387755
Score: 0.7                        Score: 0.7183673469387755
RMSE: 0.5477225575051661          RMSE: 0.5306907320287632
```

Figure 6: Decision Tree Results

C. *K-Nearest Neighbours*

The accuracy of the model is 70% for red wine for k-value 25 and it is 72% for white wine for k-value 21.

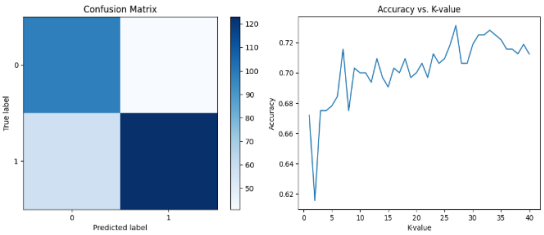

Figure 7: KNN Results for Red Wine

D. *Random Forest*

The accuracy of the model is 80% percent for red wine and 82% for white wine. The predictors that are most important for the model are ‘alcohol’ and ‘volatile acidity.’

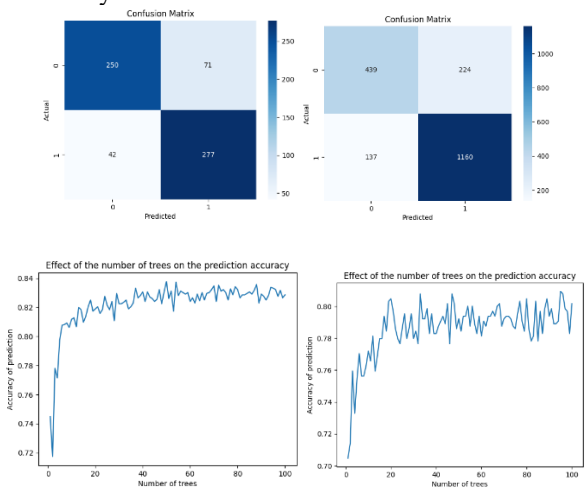

Figure 8: Random Forest Results

E. *Support Vector Machine*

The accuracy of the model is 71% for red wine and 75% for white wine.
The kernel used here is a linear kernel.

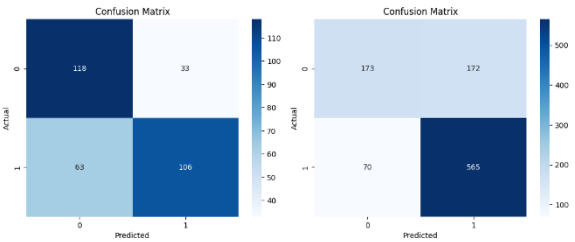

Figure 9: SVM Results

In conclusion, the Random Forest Classifier shows the best results, with the accuracy score of 80% and 82%.

When the importance score of the predictors for this model was printed, the ‘alcohol’, ‘sulphates’ and ‘volatile acidity’ turns out to be the most important parameter for predicting the wine quality.

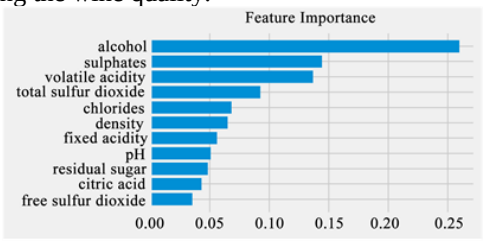

Figure 10: Feature Importance

The most important parameter i.e., alcohol makes perfect sense because it affects the taste, texture, and structure of the wine itself and not just how you feel after drinking. The sulphates, which are by definition somewhat associated with the first property, are the second most significant feature. The least significant feature, which is also noted from the plot, is the free sulphur dioxide. It is an indicator of how much sulphur dioxide, or SO2, is used during every step of the process of producing wine to stop oxidation and microbial development.

VI. Conclusion

This project showed how several statistical analyses may be utilized to analyze the parameters in the available dataset to assess the quality of the wine. Prior to production, it is possible to forecast the wine's quality using several analyses. Our research demonstrates that Random Forest, when compared to other ML models, predicts wine quality the best. This research demonstrates a different method that might be used to determine wine quality, making it a suitable place to start when identifying the factors that affect wine quality.

VII. Bibliography

[1] U. o. M. G. P. Paulo Cortez, "UCI Machine Learning Repository," 2009. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[2] A. C. F. A. T. M. J. R. Paulo Cortez, "Science Direct," November 2009. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377?via%3Dihub#preview-section-references.

[3] K. A. N. Sunny Kumar, "IEEE Xplore," June 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9104095.