Wine Quality Prediction

Sai Sindhu Thondapu (002785016)
College of Engineering
Northeastern University

Abstract

Wine Quality Prediction Analysis is a Kaggle project that predicts the quality of wine using machine learning. The goal of this project is to use feature-selective methods to analyze the dataset. I chose three algorithms at random and intend to use them all (Logistic Regression, Decision Tree & Random Forest). By focusing on the main features, I was able to extract as much information as possible for the model. This was accomplished through the use of data visualization both individually and in comparison. The data was prepared by filling in the missing values, and sampling was used to balance the data. In this project, I have not only developed the best model for Wine Quality Prediction but also established a predictive system to check whether the wine is "Good" or "Bad".

1. Dataset:

I used Kaggle's Wine Quality dataset to build various classification models to predict whether a particular red wine is "good quality" or not. Each wine in this dataset is given a "quality" score between 0 and 10. For the purpose of this project, I converted the output to a binary output where each wine is either "good quality" (a score of 6 or higher) or not (a score below 6). The quality of a wine is determined by 11 input variables.

	Dataset Characteristics	Associated Tasks	Number of Instances	Number of Attributes
Dataset	Multivariate	Classification, Regression	6497	12

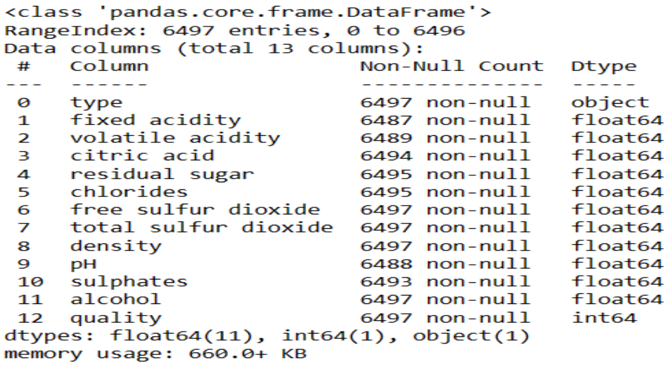
Table 1: Basic feature of the dataset

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality

Table 2: Variables of the dataset

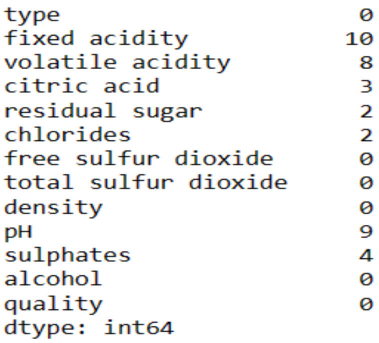
2. Data Analysis:

The dataset has the shape (6497, 13), with 11 columns of input variables and 1 column of targeted variables. I've added another column to categorize the types of wine. This data is analyzed by using visual techniques. According to the dataset information below, there are some missing values in fixed acidity, volatile acid, residual sugar, chlorides, pH, and sulphates. These missing values are imputed by means as the data present in the different columns are continuous values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   type                  6497 non-null   object
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

Figure 1: Datatype Information



```
type                     0
fixed acidity           10
volatile acidity         8
citric acid              3
residual sugar           2
chlorides                2
free sulfur dioxide      0
total sulfur dioxide     0
density                  0
pH                       9
sulphates                4
alcohol                  0
quality                  0
dtype: int64
```

Figure 2: Missing data information

As we can see, the boxplots show the mean, median, and quartile measurements for each variable, as well as the range of values for each variable. We can see a few outliers, and removing them will improve the model, but I haven't removed them in my project because it doesn't make much of a difference in the accuracy. I identified the pattern of each variable by visualizing the histogram plots. As can be seen in Figure 4, there is a significant skewness in Free sulphur dioxide. The log transformation is used to normalize this.

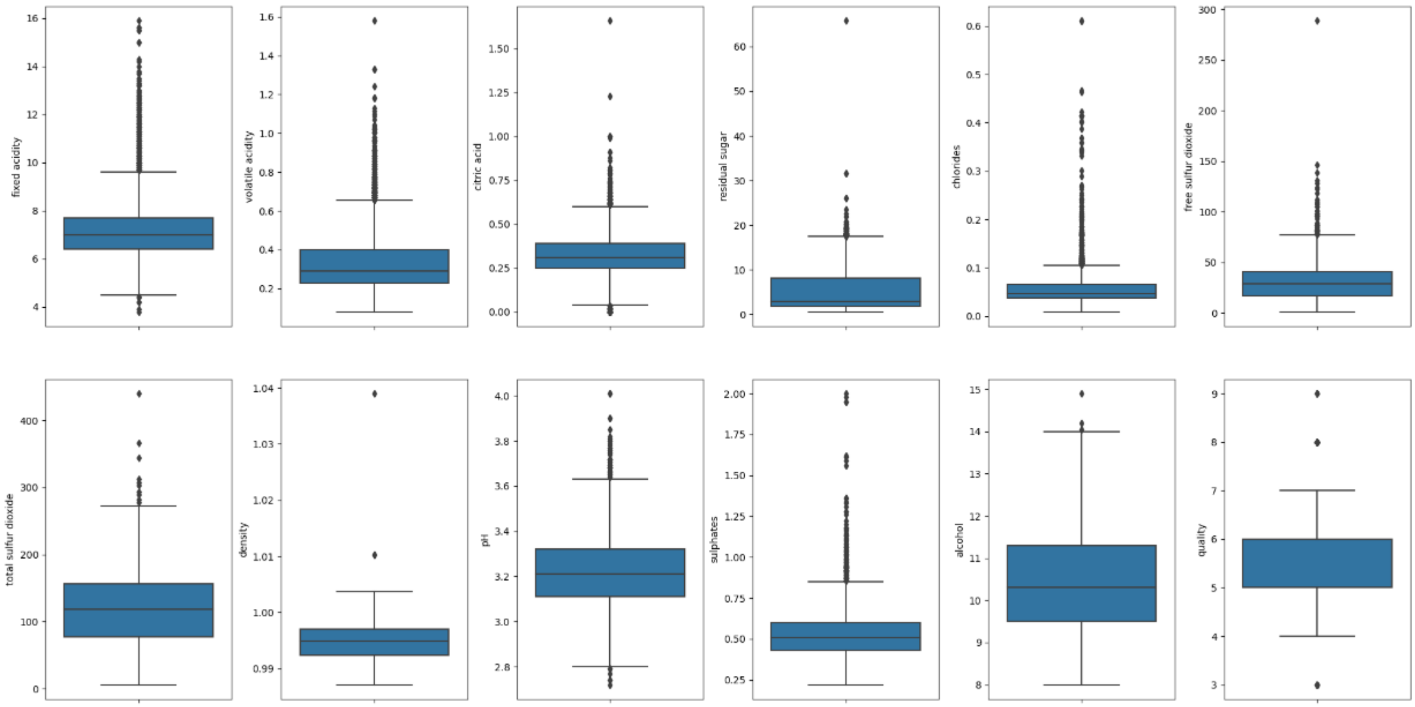


Figure 3: Box plots to analyze the numeric independent variables

Finally, a correlation was established between the variables, and by analyzing the matrix, it is clear that the variables ph, sulphates, citric acid, free sulphur dioxide, and alcohol are positively associated with quality, while the others are negatively associated. The most important variable is alcohol, and as the alcohol content rises, it improves the wine's quality. Density, which has a negative value, is inversely proportional to quality, implying that as density increases, so the quality of the wine decreases.

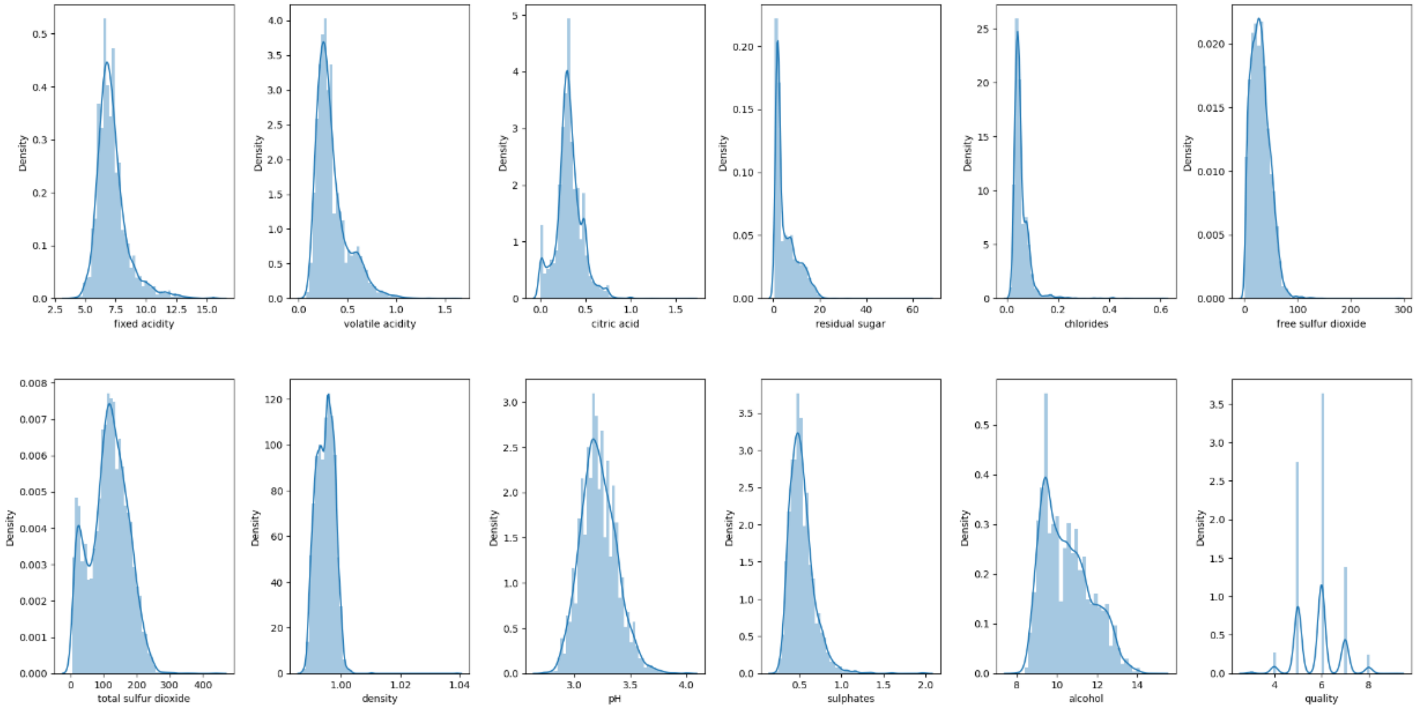


Figure 4: Histogram plots to analyze the numeric independent variables

3. Data Processing and modeling

Before training the model, ensure that the data is balanced. The below Figure 5 shows that the data is imbalanced which means dataset is biased. That is why it is essential to follow a stratified sampling method when splitting the data into the train and test set. First, I replaced the quality with 0 for the values less than 6 and 1 for the values greater than or equal to 6, with 0 indicating "Bad Quality" and 1 indicating "Good Quality." To create a balanced dataset, all classes are over sampled to the upper class, and the data is then split into test and train. I proceed to develop the models and determine which model can accurately predict the quality of red wine. The model is created with Logistic Regression, Decision tree and Random forest to check the accuracy, cross validation and confusion matrix. The below table provides the information about the results.

	Logistic Regression	Decision Tree	Random Forest
Accuracy	73.41	78.22	84.83
Cross Validation	67.22	65.74	70.84
Correctly predicted values	6061	7778	7,914
Misclassification error	26.3%	5.44%	3.79%

Table 3: Model information with balanced data

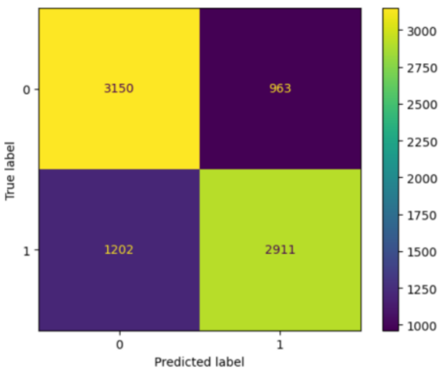


Figure 5: Confusion Matrix of Logistic

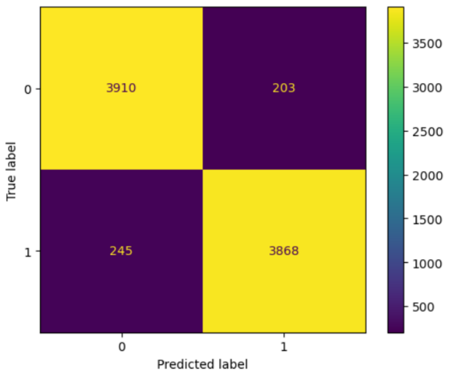


Figure 6: Confusion Matrix of Decision Tree

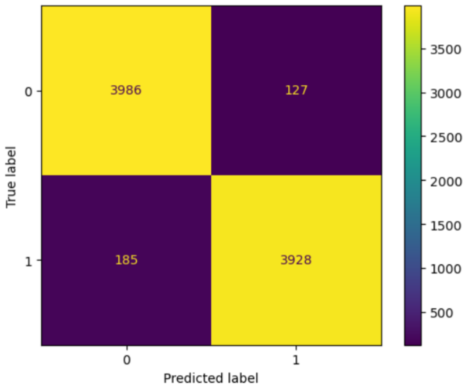


Figure 7: Confusion Matrix of Random Forest

Random forest is a better way to deal with outliers in data and avoid overfitting by aggregating many decision trees to produce an optimal model. There is no need to normalize the data because it is based on a decision tree. According to one study, alcohol is an important factor in determining wine quality, and random forest models are better at predicting high-quality wines. However, because logistic regression is based on an arithmetic function, it performs well even when the numerical features of the test data are far outside the range of the training data.

5. Conclusion

After analyzing the results of the various machine learning algorithms, I concluded that the Random Forest model performed better in predicting the quality of red wine. This model correctly predicted 7,914 values with an accuracy of 84.83% and a cross validation score of 70.85, implying that the model's misclassification error was 3.79%. It is also important to note that the accuracy of models with imbalanced datasets is poor.

6. Acknowledgments:

The learning code is adapted from: GitHub - aswintechguy/Machine-Learning-Projects: This repository contains mini projects in machine learning with notebook files and references there within.

7. Reference Links:

[1] *Prediction of Wine Quality.* (2018). Kaggle
https://www.kaggle.com/code/muammerhuseyinoglu/prediction-of-wine-quality
[2] *Red Wine Quality Prediction*. (2022). Kaggle
Red Wine Quality Prediction | Kaggle
[3] *Logistic Regression Wine Quality (~92%).* (2021). Kaggle
https://www.kaggle.com/code/abolarinbukola/logistic-regression-wine-quality-92