J A P GIHAN MARTHRIS

COMSCDS221P-015

Principles of Data Science – group Course Work

## **Tables & Figures**

[Table 1 Wine Attributes Table: Definitions and Insights 5](#_Toc138196798)

[Table 2 .isnull().sum() Output 6](file:///D:\Data%20Science\Principals%20fo%20Data%20Science\Course%20Work\Course%20Work%202\Wine%20Quality%20dataset\Principles%20of%20Data%20Science%20Group%20Course%20Work.docx#_Toc138196799)

[Table 3 info() output 6](#_Toc138196800)

[Table 4 Summary Statistics 7](#_Toc138196801)

[Table 5 Attribute Skewness 10](#_Toc138196802)

[Table 6 Scaled Data Frame 21](#_Toc138196803)

[Table 7 Linear Regresion Table 28](#_Toc138196804)

[Table 8 SVC 29](#_Toc138196805)

[Table 9 Random Forest Classifier 30](#_Toc138196806)

[Table 10 Decision Tree Classifier 31](#_Toc138196807)

[Table 11 Gradient Boosting Classifier 32](#_Toc138196808)

[Figure 1 Map of Vinho Verde, Portugal 3](file:///D:\Data%20Science\Principals%20fo%20Data%20Science\Course%20Work\Course%20Work%202\Wine%20Quality%20dataset\Principles%20of%20Data%20Science%20Group%20Course%20Work.docx#_Toc138196557)

[Figure 3 Histogram Skewness of Attributes 10](#_Toc138196559)

[Figure 4 Box Plot (Outliers Removed) 12](#_Toc138196560)

[Figure 5 Histogram on Count of Quality of Wine 12](#_Toc138196561)

[Figure 6 Correlation Plot 13](#_Toc138196562)

[Figure 7 Bar plots against Quality 15](#_Toc138196563)

[Figure 8 Pie Chart by Grades of the Wine 18](#_Toc138196564)

[Figure 9 Scatter Plot: Quality vs Alcohol 18](#_Toc138196565)

[Figure 10 Scatter Plot: pH vs Fixed Acidity 19](#_Toc138196566)

[Figure 11 Scatter Plot: Citric Acid vs Fixed Acidity 19](#_Toc138196567)

[Figure 12 Scatter Plot: Free Sulfur Dioxide vs Total Sulfur Dioxide 20](#_Toc138196568)

[Figure 13 Scatter Plot: Density vs Alcohol 20](#_Toc138196569)

[Figure 14 Over sampling and under sampling 22](#_Toc138196570)

[Figure 15 Over and Under Sampler 22](#_Toc138196571)

[Figure 16 Quality Value Count 23](#_Toc138196572)

[Figure 17 Split the Data for Training and Testing 24](#_Toc138196573)

[Figure 18 Linear Regression Code Script 25](#_Toc138196574)

[Figure 19 Support Vector Machine Code Script 25](#_Toc138196575)

[Figure 20 Random Forest Code Script 26](#_Toc138196576)

[Figure 21 Decision Tree Code Script 26](#_Toc138196577)

[Figure 22 Gradient Boosting Code Script 27](#_Toc138196578)

[Figure 23 ROC Curves 33](#_Toc138196579)

[Figure 24 Model Save Pickle 34](#_Toc138196580)

[Figure 25 Model Load Pickle 34](#_Toc138196581)

[Figure 26 New Variable for Predictions 35](#_Toc138196582)

[Figure 27 Prediction for New Variable 35](#_Toc138196583)

## **Task 1**

## **Introduction**

The Wine Quality dataset is a prominent data set that is widely used for machine learning. The data set consists of chemical analysis of wine in a region in Portugal named Vinho Verde. The term ‘Vinho Verde’ translates to green wine and has reference to its soft carbonation and low alcohol content and is one of Portugal’s finest wines.

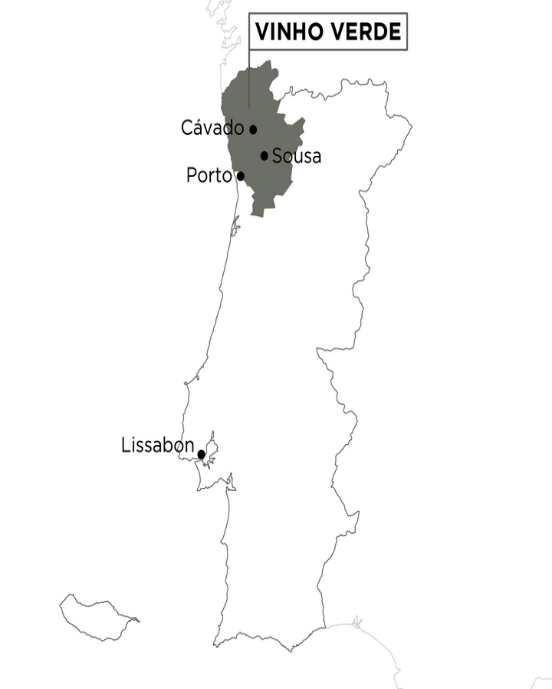
 The data set was created by several chemists from university of Portugal for a research project. The purpose of their research was to classify wine types based on the chemical properties. The objective of this research was to understand the relationship between the chemical composition and its quality. The data set includes 1,599 samples of red wine with 12 attributes. The attributes describe the chemical properties of the red wine. The unique attributes have insights that can be used for further analysis.

Figure 1 Map of Vinho Verde, Portugal

* 1. **Loading the data set and analyzing the attributes.**

To load the data set Pandas library is used and assigned to the variable named ‘df’ data frame. After loading the data set and all its attributes and examining the definition of each attribute and insight for each attribute.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Explanation** | **Insight** |
| Fixed Acidity | The total amount of acids presents in the wine. | Higher acidity wine has a crisper and a refreshing taste. |
| Volatile Acidity | The presence of volatile acids, which includes acetic acid. | Balanced level of volatile acid enhances the aroma and complexity of the wine. |
| Citric Acid | Citric acid is naturally found in fruits and contribute to the overall acidity. | Presence of citric acid adds freshness to the wine. |
| Residual Sugar | The amount of sugar that remains in the wine after fermentation. | Higher residual sugar contributes to sweeter wines and low residual sugar considered dry. |
| Chlorides | Chlorides are salt that can be found in wine. | Contributes to the wines flavor and mouth feel. Higher chlorides result in unpleasant taste of salt. |
| Free Sulfur Dioxide | Helps to preserve the wine and helps in preventing spoilage and oxidation. | This helps to prevent in spoilage and oxidation of the wine. |
| Total Sulfur Dioxide | This is the total amount of free and bound forms of sulfur dioxide in wine. | This is to ensure the preservation of the wine’s freshness its characteristics. |
| Density | Provides the concentration of solids and dissolved substance. | Gives an indication on the richness and body of the wine. |
| pH | pH refers how acidic the wine is from a scale of 0 to 14. | Higher pH value indicates lower acidity and lower pH value indicates higher acidity. |
| Sulphates | Sulphates are added as a preservative. | Helps to prevent oxidation and microbial growth. |
| Alcohol | Alcohol measured as a percentage of the volume in the wine. | This contributes the wine’s mouthfeel and warmth and sensation. |

Table Wine Attributes Table: Definitions and Insights

Based on the above table and its insights to the taste, aroma, and the quality of the wines. Therefore, wines with higher fixed acidity gives a refreshing taste when balanced with volatility acid. Citric acid ads a touch of fresh taste to the wine and the residual sugar helps to find the sweetness of the wine. Chlorides enhance the wines mouthfeel, with a touch of a salty taste. Free sulfur dioxide acts as a preservative for the wine. Density breaks the solids and dissolved substance and indicate the richness and body. pH value represents the wine’s acidity level and the represent the balance of the wine. Sulphates helps to prevent oxidation and microbial growth. The alcohol level helps in the mouthfeel and enhances sensory experience. Therefore, understanding the right balance based on the characteristics help the quality of the wine that is produced.

**1.2 Performing Pre-Processing on the Wine Data Set**

Prior to exploratory data analysis, the data needs to be preprocessed to ensure that the quality of the data set is reliable and effective prior for EDA and data modeling. As the first step prior to preprocessing, identifying if there are any missing values in the data with “df.isnull().sum()” function and understanding the data types. Based on the output there are no missing values, and the dependent variable ‘quality’ is int64 data type and the remaining features are float64 data type.

A screen shot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated

Table 2 .isnull().sum() Output

Table info() output

Since the ‘quality’ attribute is the dependent variable and its important to identify the unique values under quality using the ‘['quality'].unique()’. Based on the output, the unique data types are quantified from a minimum value of 3 to a maximum quality of 8. Therefore, the quality of the wine is measured as per the given breakdown from the lowest to the highest.

3, 4, 5, 6, 7, 8

To further examine the data set using summary statistics of the numerical columns using ‘describe()’ function.

A picture containing screenshot, text, black and white, black

Description automatically generated

Table Summary Statistics

Based on the above given summary statistics under acid levels the ‘fixed acidity’ ranges from 4.6 to 15.9 and ‘volatile acidity’ ranges from 0.12 to 1.58. This represents that the respective acidity level has influence on the taste and quality of the wine. Most importantly, the standard deviation of 1.741096 suggest that most of the wines have a fixed acidity level close to the mean. Sulphate level ranges from 0.33 to 2.0 and helps in preservation of the wine. In terms of the alcohol range the standard deviation of 1.065668 suggest a moderate variation among the wines and suggest there are range of alcohol strengths. In terms of density the standard deviation of 0.001887 portray most of the wines have density values close to the mean and minimal variation.

Thereafter after having a better understanding the meaning and definition of each attribute and the statistical representation of all the observations of the data set attributes. The outliers need to be identified and removed as it will affect the analysis and the outcome of the exploratory data analysis and the data modeling process in measuring the accuracy.

To identify if there are any outliers in the data set, a box plot is visualized with its attributes.

A picture containing text, screenshot, line, plot

Description automatically generated

Figure Box Plot (Original Data frame)

Based on the box plot it is quite evident that there are outliers visible under total sulfur dioxide column. It is important to remove these outliers prior to exploratory data analysis. Furthermore, there could be more outliers in the data set and in order to remove these outliers there are several methods including Z score method, percentile method and IQR method to identify these outliers and remove them accordingly.

Interquartile range method is the most suited method in removing the outliers from this data set is useful in relation to skewed data. Therefore, identifying how the skewness of the data set is. After developing the histogram for each attribute, the skewness is visible.

A picture containing diagram, text, plot, origami

Description automatically generated

Figure Histogram Skewness of Attributes

Based on the histogram the data points which includes the attributes sulphates, chlorides and residual sugar have a positive skewness and is greater than the median.

|  |  |
| --- | --- |
| **Type** | **Skewness** |
| volatile acidity | 0.671593 |
| chlorides | 5.680347 |
| residual sugar | 4.540655 |
| fixed acidity | 0.982751 |
| citric acid | 0.318337 |
| free sulfur dioxide | 1.250567 |
| total sulfur dioxide | 1.515531 |
| density | 0.071288 |
| pH | 0.193683 |
| sulphates | 2.428672 |

Table Attribute Skewness

Since IQR method can easily identify outlies based on the quartiles, the IQR method calculates the quartiles for each column and defines a threshold value for all the outliers and removes them accordingly. Thereafter, the original data set of 1599 row values are reduced to 1194 rows. As a result, IQR method removed 405 outliers in total (Appendix 1.1).

The new data frame after removing the outliers the data frame is assigned to a variable named ‘outliers\_removed’. Thereafter, a new box plot is visualized to determine the removed outliers.

A picture containing text, line, screenshot, plot

Description automatically generated

Figure Box Plot (Outliers Removed)

* 1. **Expletory data analysis**

After performing data preprocessing and removing outliers, performing EDA will provide insights and provide a better understanding on the relationship of the data set. Since the dependent variable is ‘quality’, creating a histogram to understand the distribution of the quality variable.

A picture containing screenshot, line, plot, diagram

Description automatically generated

Figure Histogram on Count of Quality of Wine

The above histogram gives an overall view of how the quality of the wine is distributed and most of it is 5 in quality with 513 observations and secondly as 6 with 498 observations.

To understand an overview how the remaining attributes correlates with each other a correlation plot can help identify relationships between attributes. This helps to identify the patterns of the data set.

A picture containing text, screenshot, square, parallel

Description automatically generated

Figure Correlation Plot

There are several relationships that can be gathered from the correlation plot based on the attributes.

1. **Quality vs Alcohol**

The Target variable has a positive relationship of 0.511 and a negative correlation of -0.352. Therefore, we can assume that the alcohol level has a positive influence on wine quality and higher volatility acidity have a negative impact.

1. **Fixed Acidity vs pH**

These two attributes have the strongest negative correlation of -0.684. Where there could be multicollinearity, or the two variables are both variables are highly correlated. Therefore, making it difficult to identify the individual effect of each variable and needs to be excluded.

1. **Fixed Acidity vs Citric Acid**

Fixed acidity demonstrates a positive correlation against citric acid and a negative correlation against pH. Therefore, with higher fixed acidity have higher levels of citric acid and lower pH values.

1. **Free Sulfur dioxide vs Total Sulfur Dioxide**

These two show a positive correlation of 0.619 and it is important in identifying the impact of sulfur dioxide on wine.

1. **Alcohol vs Density**

Both have a negative correlation of -0.546. Therefore, the as alcohol content increase the density of the wine tends to decrease.

These insights help in developing the model and machine learning tasks. However, further data exploration can be conducted to identify relationships of these attributes based on the quality of the wine. To identify how quality plays a role in the other attributes.A picture containing text, colorfulness, screenshot, diagram

Description automatically generated

Figure Bar plots against Quality

There are several interesting relationships that is visible in the given bar plots. One of the striking findings is that the ‘density’ of the wine is consistent across all quality types. This relationship provides information that the density level of the wine across all the quality levels are consistent. Therefore, the density level does not have any influence on the relationship between quality.

The volatility acidity has a negative slope as the quality level increase, and it is visible that the volatility acidity reduces from 0.7 to 0.4 as the quality of the wine increase. There is no right level of volatility acidity that could determine the overall quality of the wine. However, based on the distribution and observation a level of 0.4 volatility acidity enhances the aroma and complexity of the wine of the best quality of wine.

Based on the bar plot the citric acid also has an influence in the quality of the wine. Where when the average citric acid starts at 0.2 at a lower quality level and progress to higher citric acid levels of 0.4 as the quality increases. Therefore, it is evident that the best quality wine has a higher citric acid level. Where higher level of citric acid contributes to the freshness of the wine.

The residual sugar level is at average value of 2.1. The sugar level only contributes to the sweetness of the wine. However, the best quality wine has the highest level of residual sugar among the rest of wines.

The chlorides level contributes the overall saltiness of the wine and based on the quality level it is evident that the best quality wine has a 0.7 chlorides level. Where, higher levels of chlorides could make the wine unpleasant.

The pH level across the wine qualities have been consistent. Furthermore, based on the bar plots the highest quality of wine which is 8 has the lowest pH level and lower pH value indicates higher acidity level in the best quality of wine.

In terms of sulphates the lower quality wines 3 to 6 has a lower level of sulphate and quality level 7 and 8 has higher levels of sulphates which exceeds 0.7. As sulphates helps to prevent oxidation and microbial growth in the wine, it is evident that best quality wines have higher levels of sulphates to preserve the wine longer.

Furthermore, the alcohol level plays an important role in determining the quality of the wine. Where the lowest quality wine has alcohol level of less than 10 and it increases gradually till the highest quality wine.

It is quite evident how quality is determined based on the attributes and how the level of chemicals can influence the wine and result in the best quality wine. The quality level of 3 being the lowest and 8 being the highest, the numerical values can be broken down to 'Low', 'Medium', 'High' to better understand the quality at overall value. Therefore, the quality levels are broken down to three categories and included to the data frame. This is to carry out further exploratory data analysis. After including the new quality grade values, a pie chart is visualized to understand the distribution of wine quality by each grade.

A picture containing screenshot, circle, graphics, diagram

Description automatically generated

Figure Pie Chart by Grades of the Wine

The above donut chart visualizes the grades of the wines into three main categories from low, medium, and high. Most of the quality of the wine falls under medium level.

Based on the correlation plot, few attributes had high correlations and further investigation is performed to determine the dispersion of the variables.

A picture containing text, line, plot, screenshot

Description automatically generated

Figure Scatter Plot: Quality vs Alcohol

The quality vs alcohol plot shows a clear division that alcohol level being less than 0 is low quality wines.

A picture containing line, diagram, screenshot, plot

Description automatically generated

Figure Scatter Plot: pH vs Fixed Acidity

In terms of pH vs Fixed acidity there is a large portion of data points clustered around the pH value 0 and fixed acidity as 0. As a result of this these are high quality wine in this area.

A picture containing diagram, line, screenshot, plot

Description automatically generated

Figure Scatter Plot: Citric Acid vs Fixed Acidity

The above figure visualizes a high-quality wine in citric acid level greater than 0 and fixed acidity above -1. The low-level quality wines have the lowest fixed acidity and citric acid.

A picture containing font, screenshot, line, text

Description automatically generated

Figure Scatter Plot: Free Sulfur Dioxide vs Total Sulfur Dioxide

In terms of free sulfur dioxide and total sulfur dioxide most of the high quality wines are closely related to negative values in sulfur dioxide and free sulfur dioxide.

A picture containing screenshot, line

Description automatically generated

Figure Scatter Plot: Density vs Alcohol

On the other hand, in terms of density vs alcohol level, most of the high-quality wines have an alcohol level above 0 and density level 0.

Overall, the combination of chemicals such as alcohol, pH, fixed and citric acid and density provide important insights. Therefore, patterns are visible for these attributes.

* 1. **Handling Data Imbalance**

Prior to handling data imbalance in the data set, it is important to scale the data as the variables are not in the same scale and to perform machine learning. This will bring all the variable data points to a similar scale. Furthermore, scaling the data will help to normalize the data points and enhance the model performance. Finally, the data scaling assist to easily compare the importance of different features of the model.

Using the ‘StandardScaler’ from the sklearn library the attributes are scaled, and the new data frame is created and assigned to the variable ‘outlier\_removed’.

A screenshot of a phone

Description automatically generated with low confidence

Table Scaled Data Frame

After scaling the data and after the expletory data analysis, there is a data imbalance between the data set. To explore the imbalance, the oversampling method is used to determine the data imbalance. In order to generate the data imbalance across the attributes ‘imblearn.over\_sampling’ library is used to understand the data imbalance. Under the quality variable the following value counts are determined.

A screen shot of a computer program

Description automatically generated with low confidence

Figure Over sampling and under sampling

After using the Random over sampler library, the following output is generated. Based on the above value counts.

A screen shot of a computer program

Description automatically generated with low confidence

Figure Over and Under Sampler

As there are several methods in handling data imbalance, which includes oversampling, under sampling, Combining oversampling and under sampling, class weighting and other ensemble methods. However, based on the given wine data set and the exploratory data analysis it is quite visible on the figure 5 histogram of the quality.

A picture containing screenshot, line, plot, diagram

Description automatically generated

Figure Quality Value Count

There are many observations for quality 5 and 6 wines and a very portion for the 3 and 8 wine quality. This is predominantly due to the low values count of quality variables that are in 3 and 8 qualities. Therefore “Under sampling method” will reduce further. Therefore, to balance the data set a combination of both oversampling and under sampling is used.

* 1. **Data Split to Training and Testing Sets**

Data split of the data set is crucial to evaluate the model. Where the data set is split into two areas which includes training and testing to evaluate against one another and check the accuracy of the model. Furthermore, it can help in hyperparameter tunning, where the learning rate is tweaked to optimize the model’s performance. In the case of the wine data set the data splitting is crucial to predict the quality of the wine’s accuracy and separation of the tests and provide reliable predictions.

A picture containing text, software, multimedia software, screenshot

Description automatically generated

Figure Split the Data for Training and Testing

**1.6 Applying Machine Learning Algorithms**

To perform machine learning models to the trained and test data set, five machine learning algorithms are selected to check and predict wine quality.

1. Linear Regression
2. Support Vector Machine
3. Random Forest
4. Decision Tree
5. Gradient Boosting

1.6.1 Linear Regression

The linear regression model is most sort after training model that can be applied to this data set. To evaluate the performance of the linear regression model. Which includes metrics mean square error. This will assist to determine how well the model fits the data and its predictive accuracy. Furthermore, this will help find model coefficients, p values and confidence intervals. Where these will provide additional insights into the model’s performance and its features.

A picture containing text, screenshot, font

Description automatically generated

Figure Linear Regression Code Script

1.6.2 Support Vector Machine

Support vector machine is useful for the data set to classify and separate data points of different classes and to fit a regression line with the largest margin to derive predictions on unseen data points. Where, this model has the capability to handle nonlinear relationships between the target variable and the feature. Furthermore, has the capability to separate the different classes in the data set and improve the classification and performance. Finally, SVM has the ability to handle large features which are greater than the number of samples.

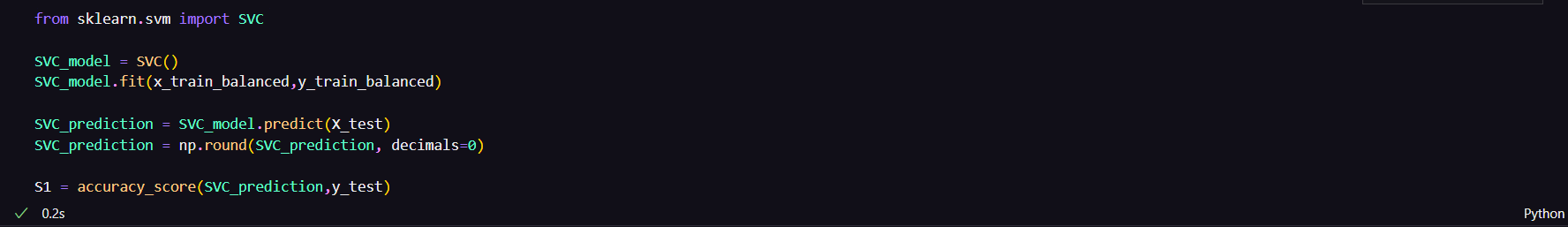


Figure Support Vector Machine Code Script

1.6.3 Random Forest

The objective of selecting this model is to find complex relationships in the data through random feature selection and averaging of multiple predictions. The model can handle the high number of features in the given wine data set. Furthermore, the model provides a measure of feature importance and indicates the relative contribution of each feature to find the most influential factors which are affecting the quality of wine.

A picture containing screenshot

Description automatically generated

Figure Random Forest Code Script

1.6.4 Decision Tree

This flow chart like model helps to predict using its flowchart like model. The model is interpretable with its decisions and easily identify the predictions. This can handle both numerical and categorical data, where this model is quite useful in the decision tree model. Also, decision tree helps in the measure of feature importance. This information is valuable in for feature selection and understanding which feature has the most importance in predictions. A picture containing screenshot, text

Description automatically generated

Figure Decision Tree Code Script

1.6.5 Gradient Boosting

This model is like a decision tree. However, it trains each subsequent model to correct the mistakes made by the previous model and improving the accuracy of the overall model. This is quite useful in nonlinear relationships between the features and target variables. Furthermore, this model is useful in handling imbalance data and missing values. However, the wine data set does not have any missing values. Furthermore, if the data set is imbalanced the model can still perform by adjusting weights of the sample or using class weights. Therefore, it is best to use the data as per the original data set when performing gradient boosting.

A picture containing text, multimedia software, screenshot

Description automatically generated

Figure Gradient Boosting Code Script

**1.7 Evaluating the best performing model.**

Based on the 5 models that were selected, to determine accuracy, precision, recall, F1-score, and ROC curve. The best model can be evaluated based on the given criteria.

1.7.1 Linear Regression

A picture containing text, screenshot, font, number

Description automatically generated

Table Linear Regresion Table

Precision: The model shows low precision across all classes, indicating that it struggles to correctly identify instances of each class.

Recall: The recall values vary across classes, with some classes having zero recall, suggesting the model fails to capture instances of those classes.

F1-score: The overall F1-score is low, indicating poor performance in terms of both precision and recall.

Accuracy: The accuracy is 0.39, suggesting that the model performs poorly in predicting the wine quality.

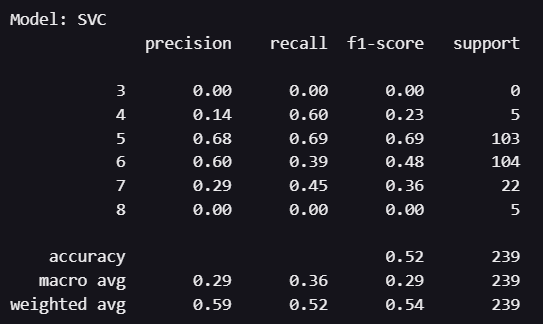
1.7.2 SVC (Support Vector Machine)

Table SVC

Precision: The precision values vary across classes, but the model performs relatively well in identifying class 5 and 6.

Recall: The recall values also vary across classes, with class 5 and 6 having higher recall rates.

F1-score: The F1-scores for class 5 and 6 are relatively higher, indicating better balance between precision and recall for those classes.

Accuracy: The accuracy is 0.52, indicating a moderate performance of the model in predicting the wine quality.

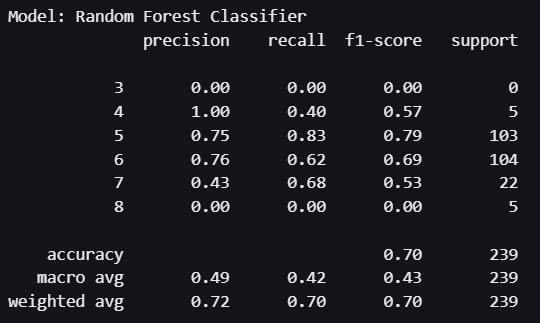
1.7.3 Random Forest Classifier

Table Random Forest Classifier

Precision: The precision is high for class 4 and 5, indicating good identification of instances in those classes. However, precision is low or zero for other classes.

Recall: The recall values are relatively higher for class 5 and 6, suggesting the model performs better in capturing instances of those classes.

F1-score: The F1-scores for class 4 and 5 are high, indicating a good balance between precision and recall for those classes.

Accuracy: The accuracy is 0.70, indicating a relatively good performance of the model in predicting the wine quality.

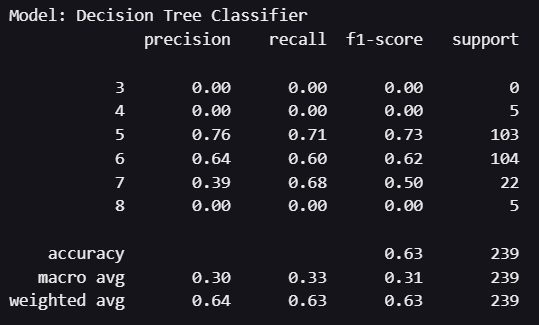
1.7.4 Decision Tree Classifier

Table Decision Tree Classifier

Precision: The precision is low across all classes, indicating the model struggles to identify instances of each class accurately.

Recall: The recall values vary across classes, with class 5 having a relatively higher recall rate.

F1-score: The overall F1-score is moderate, suggesting a fair balance between precision and recall for the dataset.

Accuracy: The accuracy is 0.63, indicating a moderate performance of the model in predicting the wine quality.

A screenshot of a computer screen

Description automatically generated with low confidence1.7.5 Gradient Boosting Classifier

Table Gradient Boosting Classifier

Precision: The precision is relatively higher for class 5, indicating good identification of instances in that class. Precision for other classes is low or zero.

Recall: The recall values vary across classes, with class 5 having the highest recall rate.

F1-score: The F1-scores for class 5 and 6 are moderate, suggesting a reasonable balance between precision and recall for those classes.

Accuracy: The accuracy is 0.63, indicating a moderate performance of the model in predicting the wine quality.

Furthermore, to visualize the models and to identify the area under the curve (AUC) for each model and the highest AUC provides the best model.

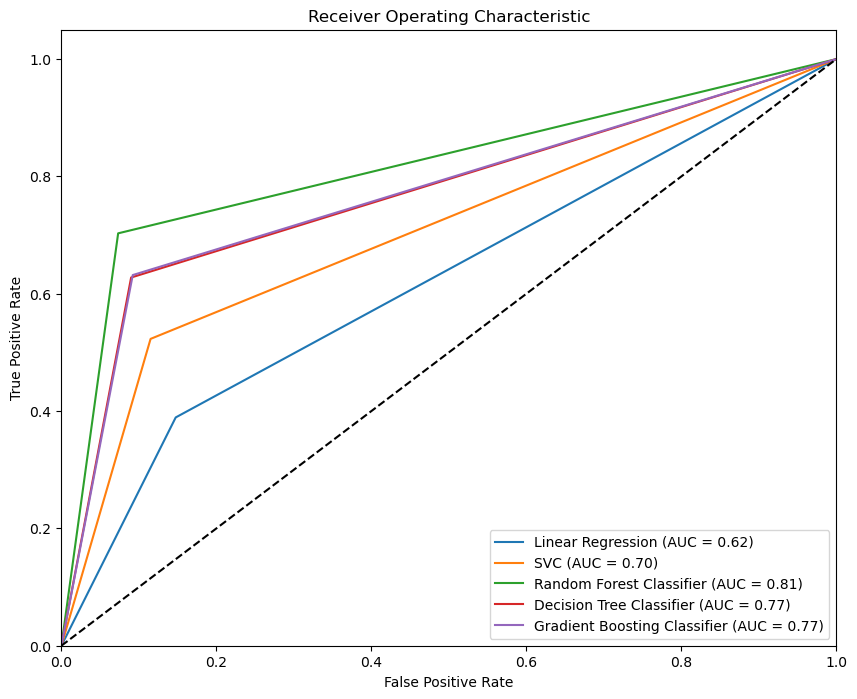


Figure ROC Curves

Based on the above outputs the ‘Random Forest Model’ performs best in comparison to the rest of the models that were developed with an accuracy of 0.70 and AUC value of 0.81. Therefore, based on the two metrics Radom Forest is the winning model.

**1.9 Libraries That Enable Reusability of a Trained Model**

There are several methods of saving models are reusable model. Which includes from Pickle to Tenserflow Saved Model. However, pickle is easiest method and its inbuilt module and can save trained models as binary files. Where this can be applied and used again at any given time.

A picture containing text, screenshot

Description automatically generatedFigure Model Save Pickle

**1.10 Loading the model and giving Predictions on the Model**

After successfully pickling the file and saving, the file can be called again and be used to do predictions.

A picture containing multimedia software, black, screenshot

Description automatically generated

Figure Model Load Pickle

After loading the saved model for random forest the model can be reused to give predictions based on values of each attribute and give a quality score based on the algorithm. To give predictions based on the decision tree the new variables or the test variables are given. Thereafter the data is scaled since, the data was scaled prior to machine learning.

A picture containing text, multimedia software, software, screenshot

Description automatically generated

Figure New Variable for Predictions

After scaling the new variables, the data set can be given for predictions. Therefore, using the model predict function generate the quality of the wine based on the variables given.

A black screen with white text

Description automatically generated with low confidence

Figure Prediction for New Variable

Based on the test\_data\_3 there are values that are represented in the third row of the data set. The same values are used to test the random forest machine learning algorithm. Based on the predictive model the out put is ‘6’. Which is the quality that the variable was first assigned to in the original data set. Therefore, the prediction for the given variable was accurately predicted.

## **References**

Dahal, K., Dahal, J., Banjade, H., & Gaire, S. (2021). Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289. doi: 10.4236/ojs.2021.112015. Retrieved from <https://www.scirp.org/journal/paperinformation.aspx?paperid=107796>