**INTRODUCTION:**

The purpose of this assignment is to build two models to predict the type of wine and the wine evaluations of the wine experts by studying and analyzing the data on approximately 6,500 commercially available wines.

* Task 1: Predicting the type of wine (Red or white)
* Task 2: Predicting the quality (wine evaluation of wine experts)

These tasks will be accomplished by using several approaches, from these the best model will be selected. The best model from task 1will be used to predict the type of wine and the best model from task 2 will be used to predict the quality of wine.

**Data Exploration:**

* We have a total of 6497 data points with 13 data columns ('fixed\_acidity', 'volatile\_acidity', 'citric\_acid', 'residual\_sugar', 'chlorides', 'free\_sulfur\_dioxide', 'total\_sulfur\_dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'type', 'quality')
* The descriptive statistics of the data are as follows:

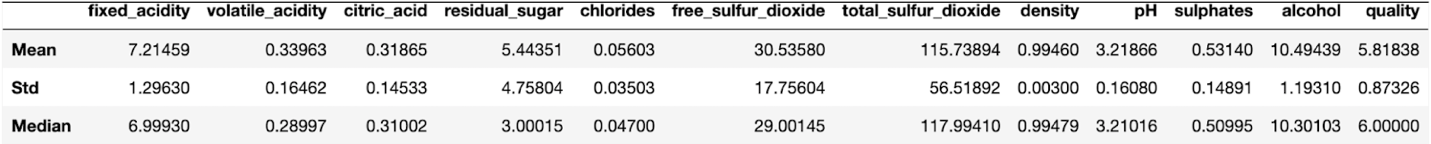


Table 1- Descriptive Statistics

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 1- Wine Type & Quality

* We have created a new feature 'molecular\_sulfur\_dioxide' from an existing feature 'free\_sulfur\_dioxide' using the below formulae:

df['molecular\_sulfur\_dioxide']=df['free\_sulfur\_dioxide'] / (1 + np.power(10, df['pH'] - 1.81))

* In the entire dataset there was just one categorical variable- Type, which has been converted into binary to ensure the functioning of models (Red- 1, White-0).
* **Correlation** is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate) the correlation heatmap for the given data set is as follows:

Chart

Description automatically generated with medium confidence

Figure 2- Correlation Matrix of Wine Type

Correlation heatmap is graphical representation of correlation matrix representing correlation between different variables. The value of correlation can take any values from -1 to 1. Correlation between two random variables or bivariate data does not necessary imply causal relationship

* If the value is 1, it is said to be positive correlation between two variables. This means that when one variable increases, the other variable also increases.
* If the value is -1, it is said to be negative correlation between two variables. This means that when one variable increases, the other variable decreases.
* If the value is 0, there is no correlation between two variables. This means that the variables change in a random manner with respect to each other.

Table

Description automatically generatedTable

Description automatically generated

Table 2- Correlation Table

* The above table is the correlation table from which the correlation heatmap is generated. The table contains the correlation coefficient between the ‘type’ and following:

total\_sulfur\_dioxide -> 0.7

chlorides -> -0.5

fixed\_acidity -> -0.48

volatile\_acidity -> -0.65

molecular\_sulphurdioxide -> 0.47

Which implies that ‘total\_sulfur\_dioxide’ is highly correlated with wine ‘type’ followed by ‘molecular\_sulphurdioxide’ which is also positively correlated. ‘chlorides’,’ fixed\_acidity’ and ‘volatile\_acidity’ are negatively correlated.

* There are only a few null values ('free\_sulfur\_dioxide':6; 'density':3; 'sulphates':15; 'alcohol':10)

**Data Preparation:**

* *Outlier Identification*: An **outlier** is an observation that lies outside the overall pattern of a distribution. Identify outliers is crucial.
* *Outlier Removal/management:* After replacing the outlier values to ‘nan’ (null value) We have a total of 1654 null values which needs to be imputed.
* *Missing (Null) value imputation:* We have used both ‘KNN Imputer’ and ‘Iterative Imputer’

*KNN Imputer*: This **imputer** utilizes the k-Nearest Neighbors method to replace the missing values in the datasets with the mean value from the parameter '

*Iterative Imputer*: This is a Multivariate **imputer** that estimates each feature from all the others.

**Task 1: Predicting the type of wine (Red or white)**

* Before building the models, let us understand the relation between all the data fields and our target variable, wine type.
* **multiple density plot** is a great way to understand the similarities and differences between the groups.

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generated**

**Chart, histogram

Description automatically generated**

Figure 3- Density Plot for Wine Type

* The density plots show each predictor distribution grouped by wine type.
* Features with little or no overlap make better prediction with higher accuracy.
* The distribution of alcohol is very similar for red and white wines, meaning that alcohol is a bad predictor for wine type.
* On the other hand, the distribution of volatile acidity, chlorides and total sulfur dioxide is clearly shifted for both types.
* Sulphates, citric acid, and pH have large overlap.
* We can notice that chlorides and total sulfur dioxide also have distinct peaks for red and white wines.
* These predictors may help distinguish both wine types.
* **Scaling all numerical feature:** Scaling is applied to independent variables or features of data. It basically *helps to normalize the data within a particular range*. Sometimes, it also helps in speeding up the calculations in an algorithm.
* We have used ‘standard scaler’, standard scaler *removes the mean and scales each feature/variable to unit variance*. This operation is performed feature-wise in an independent way.

**Building Models:**

For building the model for classification, the data has been divided into training and testing sets. 75% of data is used for training the models and 25 % is used for testing. Since the data is not time series (independent of time) we divided the data randomly.

1. **Random forest (model 1):**

It is a collection of decision trees, with that said random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit overfitting as well as error due to bias and therefore yield useful results.

1. **KNN (model 2):**

KNN works on a principle **assuming every data point falling in near to each other is falling in the same class**. In other words, it classifies a new data point based on similarity.

1. **Logistic Regression (model 3):**

Logistic regression is a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) that in its basic form uses a [logistic function](https://en.wikipedia.org/wiki/Logistic_function) to model a [binary](https://en.wikipedia.org/wiki/Binary_variable) [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable).

1. **AdaBoost Classifier (model 4):**

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

Chart, funnel chart

Description automatically generated

Figure 4- Important features

The above bar plot shows the importance of all the data types(features) in predicting wine type. We can see that ‘chlorides’, ‘total\_sulfur\_dioxide’ and ‘density’ are the top 3 important and ‘molecular\_sulfur\_dioxide’ is the least important feature in predicting wine type.

1. **Neural Networks (model 5):**

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. Neural networks can adapt to changing input.

* After fitting the data in all the 5 models for training data, then we tested on test/validation data to check for accuracy, the below is table consisting of all the accuracy and AUC measures of test data with respective to all the above 5 models:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AUC |
| Random Forest | 0.992 | 0.9862 |
| KNN | 0.9889 | 0.9817 |
| Logistic Regression | 0.9876 | 0.9808 |
| AdaBoost Classifier | 0.9950 | 0.9925 |
| Neural Networks | 0.9870 | 0.9788 |

**Conclusion:**

Our main goal here is to predict the type of wine, for which we have information on wine types, But the data points on type=’white’ are very less compared to data points type=’red’. So, accuracy is not our best measure to decide on the best model, therefore we are using AUC score.

The higher the AUC, the better the performance of the model at distinguishing between the white and red wines. Therefore, from the above table we can say that **AdaBoost Classifier** is the best model used to predict the wine ‘type’

**Task 2- Wine Quality**

Data Exploration

* We have created a new feature 'molecular\_sulfur\_dioxide' from an existing feature 'free\_sulfur\_dioxide' using the below formulae:

df['molecular\_sulfur\_dioxide']=df['free\_sulfur\_dioxide'] / (1 + np.power(10, df['pH'] - 1.81))

* There are only a few null values ('free\_sulfur\_dioxide':6; 'density':3; 'sulphates':15; 'alcohol':10)
* We have used correlation coefficients to check the correlation between target and other variables present in the dataset.

Chart

Description automatically generated with medium confidence

Figure 5- Correlation Matrix for Wine Quality

Table

Description automatically generated

Table 3- Correlation Table

* The above table is the correlation table from which the correlation heatmap is generated. The table above contains the correlation coefficient between the ‘quality’ and the following:

Alcohol-> 0.44

density -> -0.305

chlorides -> -0.20

Which indicates that the ‘alcohol’ is highly correlated with the ‘quality’ followed by the ‘density’. Alcohol is positively correlated whereas density is negatively correlated.

**Data Preparation:**

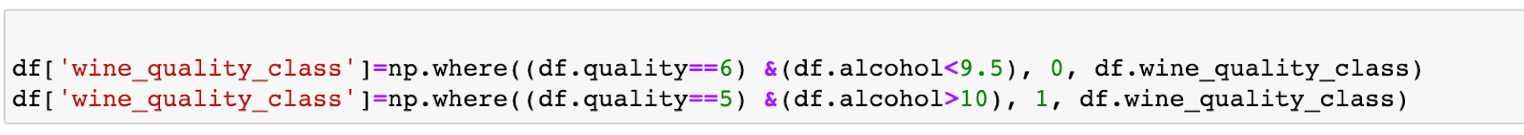
* We have identified few outliers in the data set thus, to clean the dataset, all the outlier values have been replaced by Nan values.
* After replacing the outlier values to Nan (null value) We have a total of 1654 null values which needs to be imputed.
* We have used Iterative Imputer to impute the null values from the dataset.
* We have tried to group the target column Quality into 2 categories

When the quality of wine < 6, then it is categorized as a low quality wine (0).

When the quality of wine > 6, then it is categorized as a high quality wine (1).

Moreover, to ensure the quality of the data, we have tried to filter the Quality column using the Alcohol column which has a highest correlation.

So our target variable is – ‘wine\_quality\_class’ with 2 categories.



After such grouping and filtering, we managed to have 2 quality classes.

Chart, bar chart, histogram

Description automatically generated

Figure 6- Wine Quality Class

* In the entire dataset there was just one categorical variable- Type, which has been converted into binary to ensure the functioning of models (Red- 1, White-0).
* To standardize the dataset, we have used the StandardScaler.

**Predicting the Quality of wine (Low/ High)**

* Before building the models, let us understand the relation between all the data fields and our target variable, wine quality.
* **Density plot** is a great way to understand the similarities and differences between the groups. ​​

Chart

Description automatically generated with low confidence Chart

Description automatically generated

Chart

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart

Description automatically generated A picture containing chart

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Figure 7- Density Plot for Wine Quality

* The density plots show each predictor distribution grouped by wine quality.
* Features with little or no overlap make better prediction with higher accuracy.
* The distribution of alcohol has very less overlap with the quality and hence it’s a good predictor for wine quality.
* Sulphates, free\_sulphur\_dioxide, fixed\_acidity and pH have large overlap.
* We can notice that total sulfur dioxide and type also have distinct peaks for the good and bad quality of wine.

**Building Models:**

For building the model for classification, the data has been divided into training and testing sets. 75% of data is used for training the models and 25 % is used for testing. Since the data is not time series (independent of time) we divided the data randomly.

Chart, bar chart

Description automatically generated

Figure 8- Important Features

The above bar plot shows the importance of all the data types(features) in predicting the quality of wine. We can see that alcohol, density and chlorides are the top 3 important feature and molecular\_sulfur\_dioxide is the least important feature in predicting the quality of wine. The importance of variables are calculated using Random Forest.

After fitting the data in all the 5 models for training set, then we tested on test set to check for accuracy, the below table consists of all the accuracy and AUC measures of test data with respect to all the 5 models:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | AUC |
| Random Forest | 0.9421 | 0.9315 |
| KNN | 0.8830 | 0.8612 |
| Logistic Regression | 0.8953 | 0.8838 |
| XGBOOST | 0.9249 | 0.9139 |
| Neural Network | 0.9033 | 0.8902 |

**Conclusion:**

Our main goal here is to predict the quality of wine, for which we have information, but since the data points are not properly distributed across different types of quality, the accuracy is not our best measure to decide on the best model, therefore we are using AUC score.

The higher the AUC, the better the performance of the model at classifying the quality of the wine. Therefore, from the above table we can say that **Random Forest** with the measure of area under the curve as 0.93 is the best model used to predict the wine quality.

Task 2: **Modeling by classifying target data into 3 classes**

Data Preparation

* Although we were quite satisfied by the results we got from our previous model, we further tried to analyze the data to validate our model and thus have decided to classify the targeted variable Quality into 3 classes.

When the quality of wine < 6, then it is categorized as a low quality wine (0).

When the quality of wine >= 6 and <=7, then it’s as a medium quality wine (1).

When the quality of wine > 7, then it is categorized as a high quality wine (2).

Chart, bar chart

Description automatically generated

Figure 10- Wine Quality Class

**Predicting the Quality of wine (Low/ Medium/High)**

* Before building the models, let us understand the relation between all the data fields and our target variable, wine quality.
* **Density plot** is a great way to understand the similarities and differences between the groups. ​​

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart

Description automatically generated with medium confidence Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Figure 11- Density Plot

* The density plots show each predictor distribution grouped by wine quality.
* Features with little or no overlap make better prediction with higher accuracy.
* The distribution of alcohol has very less overlap with the quality and hence it’s a good predictor for wine quality.
* Density and Alcohol have separate peaks for the quality of wines.
* Sulphates, free\_sulphur\_dioxide, fixed\_acidity,molecular\_sulphur\_dioxide and pH have large overlap.

**Building Models:**

For building the model for classification, the data has been divided into training and testing sets. 75% of data is used for training the models and 25 % is used for testing. Since the data is not time series (independent of time) we divided the data randomly.

Chart, bar chart

Description automatically generated

Figure 12- Important Features

The above bar plot shows the importance of all the data types(features) in predicting the quality of wine. The feature importance is calculated using the Random forest. We can see that alcohol, density and chlorides are the top 3 important feature and molecular\_sulfur\_dioxide is the least important feature in predicting the quality of wine.

After fitting the data in all the 5 models for training data, then we tested on test data to check for accuracy, the below table consist of all the accuracy score of test data with respect to all the 5 models:

|  |  |
| --- | --- |
| Model | Accuracy |
| Random Forest | 0.9169 |
| KNN | 0.8375 |
| Logistic Regression | 0.8621 |
| XGBOOST | 0.8886 |
| Neural Network | 0.8190 |

**Conclusion:**

Here we have used accuracy score to make out judgement on selecting the best model.

Therefore, from the above table we can say that **Random Forest** with an accuracy score of 0.91 is the best model used to predict the quality of the wine.