**Investigating the Physicochemical attributes to determine the overall quality of wine and classification of wine according to its distinguished type.**

Mithushan Soundararajah

National Institute of Business Management (NIC)

Colombo, Sri Lanka

**Abstract**

A delicate drink like wine which is typically reserved for special occasions, has a complex process of fermentation which makes the grapefruit into an expensive delicate beverage widely known as wine. Overall quality of wine does matter during consumption, wine with the best quality tends to achieve lower shelf time and drive towards higher revenue. This research study explores the physicochemical attributes of wine and its influence on determining the overall quality of wine and it distinguishes the type of wine. This research study also explores the possibility of predicting the alcohol concentration level in wine using the physicochemical properties of wine.

The research study uses machine learning algorithms to build necessary models to contest the objectives of this research study, to make accurate predictions of wine quality, wine type and alcohol concentration level in wine using the wine’s physicochemical properties. The first objective, which deals with a multiclass classification problem, performed well in making predictions of the overall quality of wine using the Random Forest classifier model, where the model showed promising performance in accurately predicting the overall quality of wine using the wine’s physicochemical properties. The second objective of this research study which deals with binary classification problem, out of all the models constructed, Random Forest classifier stood out with a high accuracy score of correctly distinguishing wine type (red or white) using the wine’s physicochemical properties. The final objective which deals with the regression problem, in predicting the alcohol concentration level in wine using the physicochemical attributes. A multiple linear regression model was used to make the prediction of alcohol concentration level in wine.

This study includes the machine learning workflow framework used to test out the derived objectives, all of the findings of the machine learning model were clearly presented along with its evaluations, this study succeeded in deriving a conclusion to all three research objectives.

**Table of Contents**

[Chapter 1 4](#_Toc152006793)

[Introduction 4](#_Toc152006794)

[1.1 Background 4](#_Toc152006795)

[1.2 Research problem 4](#_Toc152006796)

[1.3 Research Questions 5](#_Toc152006797)

[1.4 Research objectives 5](#_Toc152006798)

[1.5 Expected Limitation 6](#_Toc152006799)

[Chapter 2 6](#_Toc152006800)

[Literature Review 6](#_Toc152006801)

[2.1 Theoretical explanation about the key words in the topic 6](#_Toc152006802)

[2.2 Findings by other researchers 6](#_Toc152006803)

[2.3 Tables for variables, their definitions and source 9](#_Toc152006804)

[Chapter 3 11](#_Toc152006805)

[Methodology 11](#_Toc152006806)

[3.1 Population, sample and sampling techniques 11](#_Toc152006807)

[3.2 Types of data to be collected and data sources 11](#_Toc152006808)

[3.3 Conceptual framework 11](#_Toc152006809)

[3.4 Operationalization table 14](#_Toc152006810)

[3.5 Method of data analysis 15](#_Toc152006811)

[Chapter 4 16](#_Toc152006812)

[Data analysis 16](#_Toc152006813)

[4.1 Data preprocessing and descriptive data analysis 16](#_Toc152006814)

[4.2 exploring the physicochemical variables and its impact on predicting the overall wine quality 21](#_Toc152006815)

[4.3 Exploring the Physicochemical factors influence and predictability in distinguishing different types of wine (red wine and white wine) 27](#_Toc152006816)

[4.4 Exploration of physicochemical influence on alcohol concentration level in wine 33](#_Toc152006817)

[Chapter 5 36](#_Toc152006818)

[Discussion, recommendations, and conclusion 36](#_Toc152006819)

[Reference 40](#_Toc152006820)

**List of tables**

[Table 1: Key word explanation 6](#_Toc152006848)

[Table 2: Variable definition table 9](#_Toc152006849)

**List of figures**

[Figure 1: Independent and dependent features 13](#_Toc152006915)

[Figure 2: Machine learning workflow 14](#_Toc152006916)

[Figure 3: null values summation 18](#_Toc152006917)

[Figure 4: fixed acidity boxplot 18](#_Toc152006918)

[Figure 5: volatile acidity boxplot 19](#_Toc152006919)

[Figure 6: free sulfur dioxide boxplot 19](#_Toc152006920)

[Figure 7: residual suagr boxplot 20](#_Toc152006921)

[Figure 8: chlorides boxplot 20](#_Toc152006922)

[Figure 9: citric acid boxplot 21](#_Toc152006923)

[Figure 10: numerical features five summary statistics 22](#_Toc152006924)

[Figure 11: class balance prior to SMOTE 22](#_Toc152006925)

[Figure 12: class balance post SMOTE 23](#_Toc152006926)

[Figure 13: support vector classifier HP 24](#_Toc152006927)

[Figure 14: Logistic Regression hp 24](#_Toc152006928)

[Figure 15: Random forest classifier hp 24](#_Toc152006929)

[Figure 16: KNN hp 24](#_Toc152006930)

[Figure 17: objective 1 phase 1 machine learning model performance 25](#_Toc152006931)

[Figure 18: variance threshold value for objective 1 25](#_Toc152006932)

[Figure 19: objective 1 phase 2 machine learning model performance 26](#_Toc152006933)

[Figure 20: objective 1 feature importance decision tree 26](#_Toc152006934)

[Figure 21: Random Forest classifier confusion matrix 27](#_Toc152006935)

[Figure 22: Random Forest classifier F1 classification report 27](#_Toc152006936)

[Figure 23: Random Forest classifier ROC 28](#_Toc152006937)

[Figure 24: Distribution of wine type prior to SMOTE 29](#_Toc152006938)

[Figure 25: numerical mapping of target variables 29](#_Toc152006939)

[Figure 26: wine type distribution post SMOTE 30](#_Toc152006940)

[Figure 27: objective 2 Phase 1 model result 30](#_Toc152006941)

[Figure 28: feature variance objective 2 31](#_Toc152006942)

[Figure 29: objective 2 feature importance score 31](#_Toc152006943)

[Figure 30: objective 2 phase 2 model result 32](#_Toc152006944)

[Figure 31: confusion matrix for wine type prediction 32](#_Toc152006945)

[Figure 32: classification report for wine type prediction 33](#_Toc152006946)

[Figure 33: ROC for wine type prediction 33](#_Toc152006947)

[Figure 34: objective 3 cross validation scores 34](#_Toc152006948)

[Figure 35: objective 3 key metrics prior to feature selection 35](#_Toc152006949)

[Figure 36: decision tree regressor feature importance 35](#_Toc152006950)

[Figure 37: objective 3 cross validation score post feature selection 36](#_Toc152006951)

[Figure 38: objective 3 key metrics post feature selection 36](#_Toc152006952)

[Figure 39: Actual vs predicted multi linear regression 37](#_Toc152006953)

# Chapter 1

# Introduction

## Background

This research study is conducted to investigate the Physicochemical attributes of wine to determine its overall quality, whether those attributes can be used in predicting the different types of wine made available in the consumer market for final consumption. Quality wine absorbs more consumers towards them, this why when it comes to the wine market overall quality of wine matters. The quality of wine is typically evaluated using its aging potential, stylistic purity and ranking by experts. Ranking by experts may not be an ideal quality assessment method since, it takes their preference of wine when it comes to assessing the overall quality of the wine, the relationship between physicochemical and human sensory is a complex process. Every alcoholic and nonalcoholic beverage is based on chemical factors which completes the drink, psychochemical factors are influential metrics used along with human sensory factors to predict the overall quality of the wine. Physicochemical tests are conducted in a controlled environment along with the help of machines to determine the characteristics of wine, that’s why understanding the chemical factors and their influence on assessing the overall quality and type of wine is necessary to make business decision and to safeguard the brands reputation by increasing the overall quality of wine with the help of physicochemical attributes (Cortez et al., 2009).

## Research problem

Wine production is resource intensive business, it requires a significant investment of financial resources to sustain in the highly competitive consumer market. Wine production should be handled with utmost care, the smallest deviation from chemical attributes of the wine could alter the overall quality wine. Maintaining the best possible wine quality is ideal to have a competitive advantage against the other wine producers. This research paper focuses on the physicochemical variables affecting the overall quality of wine, can those chemical factors be used effectively in predicting the overall quality of wine and its distinguished type. Accurate prediction of overall quality of wine using physicochemical attributes could help the company financially by bringing in more profit from more sales by delivering the best quality wine to the consumers in the market.

## Research Questions

1. Does Physicochemical attributes of wine have a substantial influence on the overall quality of the wine?
2. Does wine quality and Physicochemical factors of wine have an influence on determining the respective wine type (red or white)?
3. Is the level of alcohol concentration in wine directly affected by its Physicochemical factors?

## Research objectives

1. Investigating the relationship between physicochemical variables and the overall quality of the wine and identifying whether the quality of the wine can be determined using classification study on its Physiochemical.
2. Analyzing whether the Physicochemical variables can be used in distinguishing the different types of wine ( red wine or white wine), evaluating the possibility of classifying the respective wine types in association with the Physicochemical factors.
3. Exploring the Physicochemical factors of wine to distinguish any positive and negative relationship between the chemical factors and alcohol concentration of the wine.

## Expected Limitation

This research study is entirely based off on secondary data set collected from UCI machine learning repository, the accuracy of data cannot be validated only author of the data set knows the intention of creating such data set. Using secondary could be challenging when it comes to understanding the variables mentioned in the data set.

# Chapter 2

# Literature Review

### Theoretical explanation about the key words in the topic

Table 1: Key word explanation

|  |  |  |
| --- | --- | --- |
| **Key word** | **Explanation** | **Reference** |
| Acid | Acid classifies as any substance that reacts in water solution and produces a sour taste. | (Acid 2023) |
| microorganisms | Organism that is microscopic. Bacteria, fungi, and archaea these organism falls under category of microorganism. | (What are microorganisms? 2010) |

### Findings by other researchers

An article retrieved from the University of California, Davis viticulture and oenology faculty, explores the chemical attributes and their contributions to overall wine making process. The article states that citric acid aid wine production in multiple ways. Citric acid is known for its ability to add sour tase to the food and beverages, it also widely used as natural preservatives in food or drink production. Critic acid in wine production is used to boost the acidity level of wine, it also gives a ‘fresh’ flavour to the wine, which can increase the overall quality of wine. Along with its perks citric acid has certain disadvantages associated with it, citric acid has the ability to increase the growth of unwanted microbes (Hakim, 2018).

A journal article published about the investigation carried to explore the chloride concentration in red wines: influence of terroir and grape type. The journal article explores soil property, grapefruit attributes according to the cultivated region to explore the chloride level of grapes from different region. It has been found that grapes cultivated in Australia and Argentina contains high levels of chloride. The journal article points out that wine flavour is highly influenced by the chloride ion presence, a high level of chloride concentration can form a salty taste in the wine, which directly affects the overall quality of wine ready for consumption (Coli et al., 2015).

A journal article on prediction of wine quality based on different wine types, used to computational process in predicting the overall quality of wine based on the different types of wine. This study shows that computational analytics such as machine learning techniques can be used in accurately predict the overall quality of wine according to different wine types. The study states that support vector classifier machine learning algorithm performed better with accuracy of 95% in assessing the quality of wine according to different wine types (Aich et al., 2019).

A study conducted on analysing different techniques for reducing the alcohol levels in wine, explores the alcohol and its effect on wine production. The percentage of alcohol concentration in wine is determined by the sugar content of the grapefruit. wine in alcohol is important to keep the microorganisms is tact and stop from overgrowing, the study states high levels of alcohol concentration can affect certain aromatic volatile compounds, high alcohol concentrated wine could bring burning sensation to consumers, and it also can increase the bitterness of the wine, which can drastically affect the overall quality of the wine (Ozturk & Anli, 2014).

A research journal article written by Manisha Koranga, Richa Pandey, Mayurika Joshi, Manisha Kumar on using different machine learning algorithms to predict the overall quality of white wine, explores different machine learning algorithms which capable enough to handle classification of wine quality based on Physicochemical factors. Out of all the classification algorithms used in the research study, Random Forest classifier stands out by maintaining high accuracy scores throughout different evaluation metrics, random forest classier has outperformed all other classification machine learning algorithm with the accuracy score of 99% in the making prediction on overall quality of white wine (Koranga et al., 2021).

This research, which is published through the International Journal of Environmental Research and Public Health, seeks out the influence of the assimilation grape skin and stem extracts to red wine during storage in order to minimize the reliance on sulfur dioxide (SO2) as well as enhance red wine quality. Casquete and her team have guided the study, rooted in the antioxidant and antimicrobial properties of bioactive extracts which are directly derived from various grape by-products. The research finds that extracts from grape stems have marked overall higher concentrations of total phenolic compounds and greater in vitro antioxidant activity, after evaluating six types of monovarietal wines. Both stem and skin extracts reflected vigorous antimicrobial influence against pathogenic bacteria. Additionally, skin extracts set out superior antioxidant and antimicrobial activities in the wines produced. The findings underscore the feasibility of utilizing by-products of wine production to lead the way toward the redundancy of SO2, and healthier red wine production chain engagement while following the principles of a sustainable economy while minimizing the environmental footprint of the wine industry (Casquete et al., 2021).

A study led by Richard Gawel, Steven C. Van Sluyter, Paul A. Smith, and Elizabeth J. Waters, published research that examines the convoluted interaction between phenolics, pH, and alcohol levels in the white wine sector’s perception. This examination involves the extraction of phenolics from white wines and reintroducing them into wines that are in the range of pH and ethanol applications in a realistic approach. Notably, the addition of phenolics into wine at pH 3.3 significantly marked that it heightens astringency of wine, contrary to the unpresented similar impact at pH 3.0. Higher phenolic levels generally boost bitterness and viscosity, upon varying influences based on the phenolic source. Wines made with larger phenolics are recognized for higher perceived heat, especially when low-alcohol variants are taken into consideration. The study reflects the idea of an additive relationship between phenolic content and alcohol concentration, marking out the direct contribution of alcohol to astringency and bitterness in white wines. In summary, the research indicates that the sensory classifiers that directly stem from white wine phenolics are more foregrounded in wines with lower alcohol levels (Gawel et al., 2013).

A research study which was conducted by R. Gawel and P.W. Godden at The Australian Wine Research Institute, the main target was to evaluate the consistency of experienced wine tasters in marking the quality scores to red and white wines. Data over a 15-year period, with the participation of 571 seasoned wine tasters, their ability to consistently score wines was computed through various metric systems. While most tasters manifest statistically significant consistency of quality marking, there was considerable deviation in individual abilities. As a factor, tasters marked more consistent in scoring red wines compared to white wines overall. The study underscored that combining scores from a small-scale team of tasters, a practice derived from the Australian wine show system significantly improved the consistency of the right quality predicting process. The findings signify the necessity of conducting replicate tastings for evaluating wine quality, as individual repeatability cannot be guaranteed for their quality assurance (GAWEL & GODDEN, 2008)

### Tables for variables, their definitions and source

Table 2: Variable definition table

|  |  |  |
| --- | --- | --- |
| **Variables** | **Definition** | **Source** |
| Fixed acidity | Measures the acid levels of tartaric, malic, citric, and succinic. | (Mor et al., 2022) |
| Volatile acidity | This variable contains the measurement of low molecular weight and fatty acid presence in wine. |
| Citric acid | Contains the measurement level of natural acid presence in wine typically used as preservatives. |
| Residual sugar | Contains the measurement level of sugar leftover form grapefruit after wine alcoholic fermentation process, it is typically measured in grams per litre. |
| Chlorides | Contains the measurement level of wine’s salty flavour. |
| Sulfur dioxide | Includes two measurements of sulfure dioxide in wine, free and total sulfure dioxide presence in wine, its type of salts added during the wine making process it is well known for its dual properties acting as both anti-oxidation and preservatives. |
| Density | Measures the mass per unit volume of wine, it contains the final reading of alcohol density in wine once the fermentation process in completed. |
| pH | Ph level used to indicate the specific acidity of the wine. |
| Sulphate | It’s a naturally found chemical component in variety of food, typically used as preservatives in food production, it is used in wine making process to maintain the freshness of the wine. |
| Quality | From a scale of 1 to 10, 1 being low quality and 10 being the best quality |

# Chapter 3

# Methodology

#### Population, sample and sampling techniques

The secondary data set used in this wine quality research study is retrieved from UCI machine learning repository website (Cortez et al., 2009). The data set contains physicochemical values of red and white variants of Portuguese “Vinho Verde” wine. The data set omitted certain information about grape type, wine brand and market selling price of the wine due to some privacy reasons. The data downloaded from the website had two separate csv files containing the necessary information on red and white wines. For simplicity purposes, both files were merged together by creating an additional column with name ‘type’ which contains information about the specific wine type (red and white). The data contains a total row of 6498 and 13 columns.

#### Types of data to be collected and data sources

The secondary data frame used in this research study has been collected using various scientific method and techniques to accurately retrieve the physicochemical factors of the wine and store them in csv file format by the author of the data set, majority of the data set contains numerical variable expect for the target variable columns

Categorical data- overall quality of the wine- categorical variable (rating from 1 to 10 where 1 being worst quality and 10 being the best possible quality). Wine type contains binary categorical data (red and white).

#### Conceptual framework

A diagram of wine quality

Description automatically generated

Figure 1: Independent and dependent features

The above conceptual framework points to the variables and their types. The purpose of the above conceptual framework diagram is to explore the relationship between the variables and for easy understanding of the main research objective, the diagram places the target variable in the middle and surrounding the target variable along with its features which will be used in this research study to make predictions and build the necessary model to understand the relationship, between features and target variables, to accurately make prediction of the wine quality and type.

A diagram of a machine learning model

Description automatically generated

Figure 2: Machine learning workflow

The above diagram is designed with the intention of presenting the workflow of the methodology in flow diagram for easy understanding, building machine learning models with fine tuning is a complex process. This research paper is aiming to build several machine learning models and fine tune them to get the best possible accuracy. Once the initial phase of the machine learning model is built and complete, the second phase will start by using variance threshold selection; certain features will be eliminated in order to perform dimensional reduction and rerunning the entire selected Machine learning models and fine tuning them to get the possible accuracy. Once phase two is completed accuracy scores from phase one and phase two will be compared in order to make a final selection on which phase and which machine learning model along with its hyperparameters performed better in making accurate predictions.

#### Operationalization table

|  |  |  |
| --- | --- | --- |
| **Variable** | **Indicators** | **Measures** |
| Fixed acidity | Observed fixed acidity value of wine | Numerical (from lab test) |
| Volatile acidity | Observed volatile acidity value in wine | Numerical (from lab test) |
| Citric acid | Observed citric acid value in wine | Numerical (from lab test) |
| Residual sugar | Observed residual sugar level in wine | Numerical (from lab test) |
| Chlorides | Observed chlorides level in wine | Numerical (from lab test) |
| Sulfur dioxide | Observed Sulfur dioxide level in wine | Numerical (from lab test) |
| Density | Observed density level in wine | Numerical (from lab test) |
| pH | Observed pH level in wine | Numerical (from lab test) |
| Sulphate | Observed sulphate level in wine | Numerical (from lab test) |
| alcohol | Observed alcohol level in wine | Numerical (from lab test) |
| Quality | Rated quality of the win | Categorical ( from sensory observation)  Rating from 1 to 10   1. Unacceptable 2. Poor 3. Below average 4. average 5. above average 6. good 7. very good 8. excellent 9. outstanding 10. perfect |
| Type | Distinguish wine type | Categorical (binary)  Red  White |

#### Method of data analysis

The second objective which explores the possibility of classifying the wine quality with respective wine types (red and white) along with its physicochemical factors. Since the target variable takes the form of binary output “red” and “white”, to test out the established research objective several classification machine learning models which support binary classification will be used. This research study will be conducted using python programming language, and popular machine learning library Scikit learning. Necessary data preprocessing steps will be done to eliminate null values to transform the data set to be used in machine learning model building.

The first objective explores the physicochemical variables and its impact on predicting the overall wine quality. Building various classification machine learning models along with predefined parameters of the respective model will be used in achieving the first objective of this research study.

* Multinomial logistic regression
* Support vector machine classifier
* Randomforest classifier
* K nearest neighbor classifier

The above-mentioned models will be used to support the first objective of this research study.

The second objective which explores the possibility of classifying the wine quality with respective wine types (red and white) along with its physicochemical factors. Since the target variable takes the form of binary output “red” and “white”, in order to test out the established research objective several classification machine learning models which support binary classification will be used.

* Logistic regression
* Support vector machine
* Randomforest classifier
* K nearest neighbor classifier

The third objective which explores the influence of the physicochemical factors and its influence on the presence of alcohol concentration level in wine, in order to identify any possible relationship between the physicochemical factors and alcohol concentration level, a well-known machine learning algorithm, linear regression analysis will be used to build a suitable machine learning model to fulfill the third objective of this research study.

* Multiple Linear regression

# Chapter 4

# Data analysis

##### Data preprocessing and descriptive data analysis

The combined two csv files on red and white wine recorded about 6000 plus data points on physicochemical factors and rating results about overall quality of the wine. Due to the computational constrains in building multiple machine learning models and fine tuning them in order to increase the overall accuracy of the model, data set has been randomly sampled by bringing down the data entries to 800, this was carefully down using python libraries, where the random sample is stratified to make sure, during the sampling process each class target variable is included.

Once the necessary sampling process is done, we have initiated a null value check to identify any missing values, fortunately data set did not consist of any missing values, once the missing value check process is complete, using python “info” function, data type check was conducted to identify any potential data type mismatch. From the data type function, it has been found that one of the target columns ‘qualities’ has been recorded as integer type. after the conversion of columns into required data types, using sns box plot outliers’ presence detection is conducted.

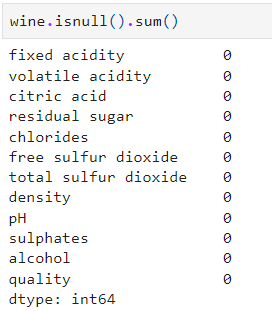


Figure 3: null values summation

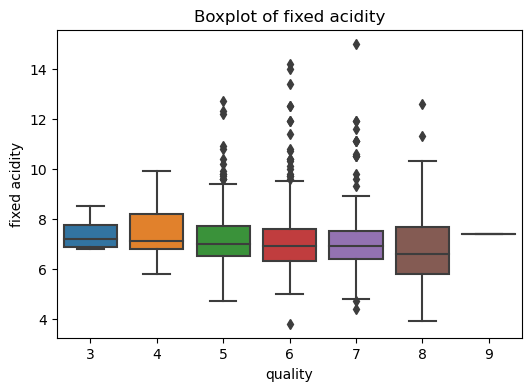


Figure 4: fixed acidity boxplot

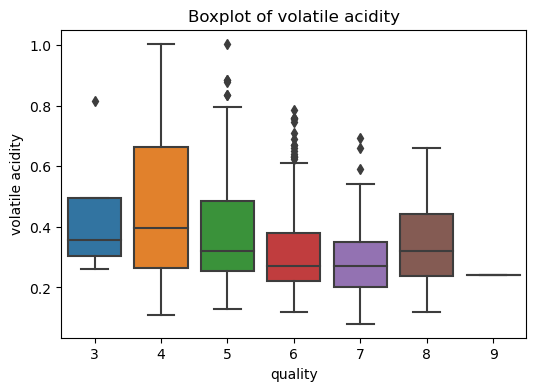


Figure 5: volatile acidity boxplot

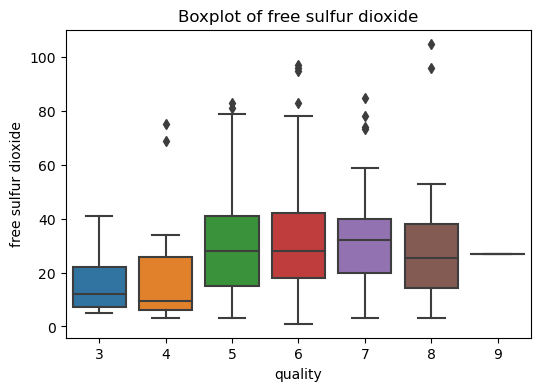


Figure 6: free sulfur dioxide boxplot

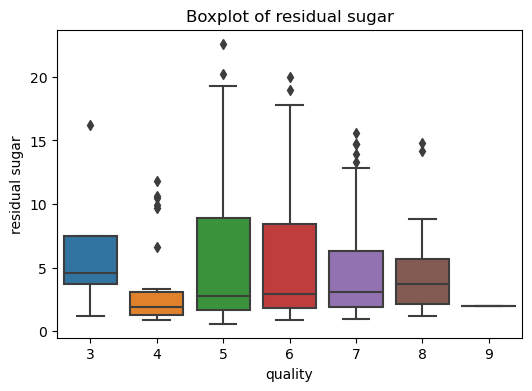


Figure 7: residual suagr boxplot

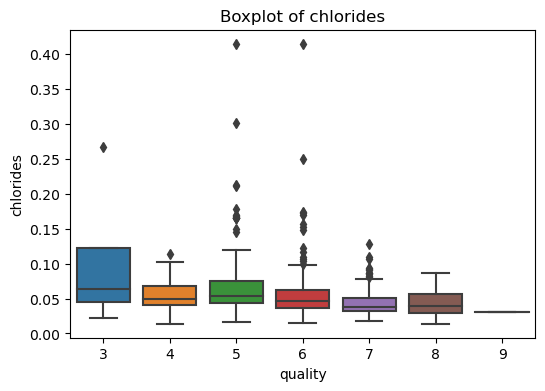


Figure 8: chlorides boxplot

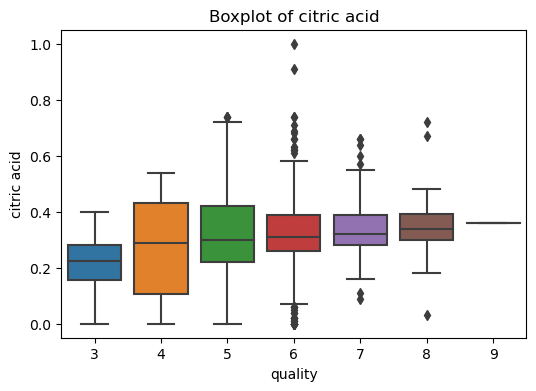


Figure 9: citric acid boxplot

Almost all numerical feature variables ( fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol) showed the presence of outliers, to eliminate the outlier presence, we used Inter quartile range method to calculate the upper and lower threshold and remove the outliers using the predetermined threshold values.

Using sns pair plot the overall distribution and relationship between data points was explored and it has been found that almost all the feature variables did not have a normally distributed curve, it was either left and right skewed. Once the elimination process is complete using python describe function, five summary statistics for the numerical feature variables were obtained

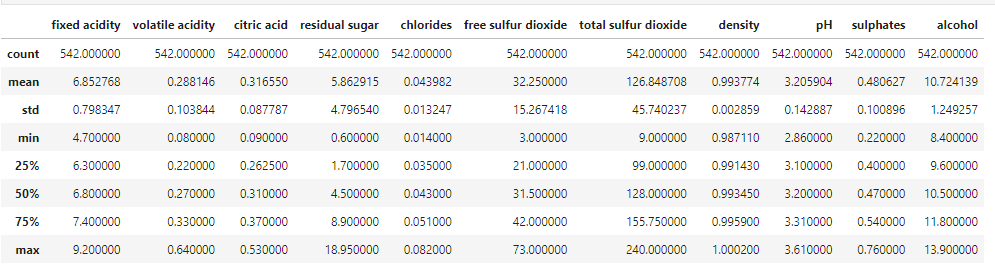


Figure 10: numerical features five summary statistics

The above figure contains statistical information about the all the numerical feature variables, it has been found that, variable ‘free sulfur dioxide’ and ‘total sulfur dioxide’ recorded the highest level of standard deviation from the mean.

##### exploring the physicochemical variables and its impact on predicting the overall wine quality

The first objective of this research study deals with a multiclass classification type, as mentioned in the conceptual framework various machine learning classification models will be built, and fine tuning by adjusting the hyper parameters of those models to identify the best model to support the first objective of this research study. Since it is a classification problem, prior to any model building it is a good practice to check for the imbalance nature of the target variable.

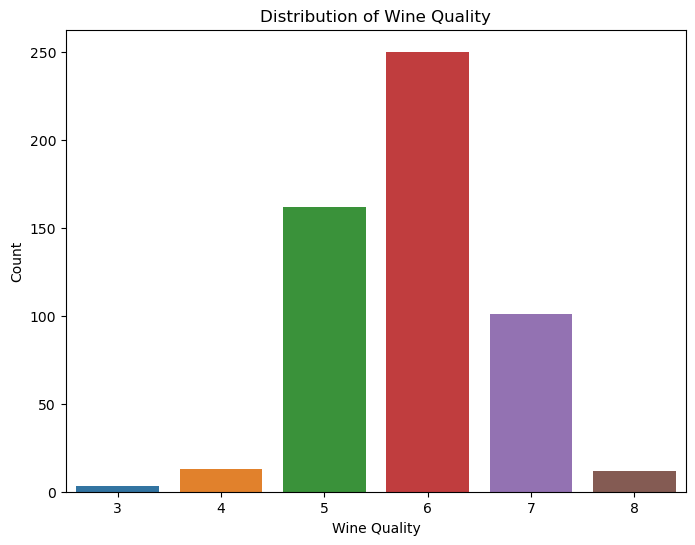


Figure 11: class balance prior to SMOTE

It has been found that, the wine quality rating classes are highly imbalanced with ‘6’ rating containing the highest level of entries, building classification model with imbalanced data will significantly affect the validity of the model.

For a multiclass imbalance, Synthetic Minority Oversampling Technique (SMOTE) is ideal model, where it typically used to oversample the minority class, SMOTE oversamples the minority class in way which it doesn’t affect the overall integrity of the data set significantly, it uses the existing samples and synthesis it to produce new sample to match according to match the number data points in the majority class (Brownlee, 2021).

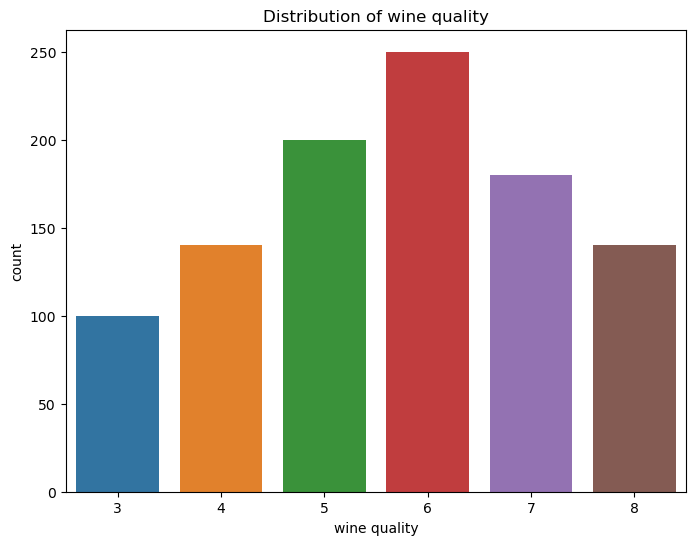


Figure 12: class balance post SMOTE

The above count plot shows the distribution among the target variable after the process of SMOTE oversampling of minority class. Random level of over sample has been used, instead of using the default parameters, where the SMOTE algorithm oversamples the minority class to match the number of instances of the highest majority class. Assigning random expected level of oversampling is done in order to prevent the model from overfitting.

Once the balancing phase is completed, the first phase of model building started with splitting the data set into training and testing where 20% of the entire data set has been allocated for the testing and rest for the training purpose, to assess the split of the test and train, fivefold cross validation method will be used throughout the machine learning model fitting.

All four machine learning classification models were built with different hyperparameters which will be explored by iterating the specific model to run through all possible combinations of hyperparameters. Once all the combination fitted and complete using python libraries, only the parameter combination with the highest accuracy score is retrieved to compares with other classification models, to select the one with best accuracy in predicting the overall quality of wine by exploring the physicochemical features.

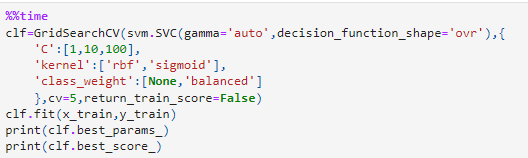


Figure 13: support vector classifier HP

A computer code with text

Description automatically generated

Figure 14: Logistic Regression hp

A screen shot of a computer code

Description automatically generated

Figure 15: Random forest classifier hp

A computer code with red and green text

Description automatically generated

Figure 16: KNN hp

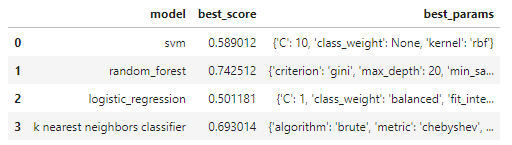


Figure 17: objective 1 phase 1 machine learning model performance

The above figure shows the best parameter combinations used to obtain the best possible accuracy score for that relevant classification model, out of all four-model random forest classifier performed best with the accuracy score of 74%.

For the second phase in the process of machine learning workflow same four classification model with same multiple combination of hyperparameter values will be used along with the process of dimensional reduction, where features deemed to be less important form the model will be eliminated using ‘variance’ threshold’ method, to analyze the model’s performance with selected features. Variance threshold method works by eliminating the features with zero variance among them by default, the zero threshold can be set to a value desired by the user based on the type of data used (Gupta, 2023). The purpose of eliminating the features with zero variance is, to reduce the computational complexity of the machine learning models since zero variance features does not contribute or contain any useful information. Zero variance features have instance of constant values, which just increases the computational complexity of machine learning model rather being useful in making predictions.

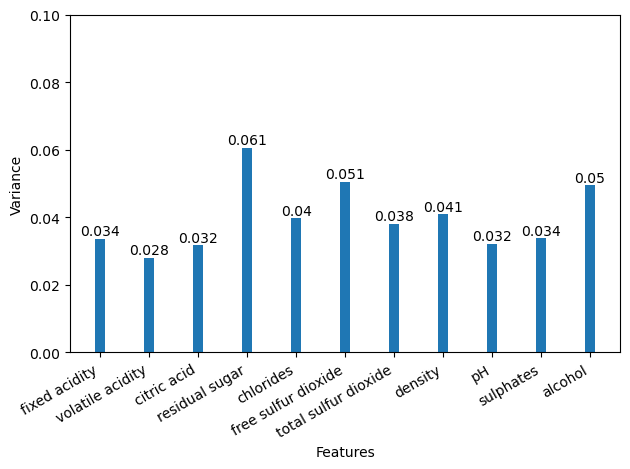


Figure 18: variance threshold value for objective 1

The independent variable used in this research shows that all the features have some level of variance, there is not a significant difference between the highest and lowest variance level of the features. Setting the lowest threshold as 0.035 and eliminating all the features with variance lower than the 0.035. By setting the lowest threshold value features, ‘fixed acidity’, ‘volatile acidity’, ‘citric acid’, ‘pH’ and ‘sulphates’ will be removed, and all other features will be used in model fitting process.

A screenshot of a computer

Description automatically generated

Figure 19: objective 1 phase 2 machine learning model performance

In the phase 2 model performed poorly after feature selection compared to phase 1 where models were built without omitting any features. This is quite understandable since all the independent variables had some sort of variation among them, which shows that every feature contributes to predicting the overall quality of the wine.

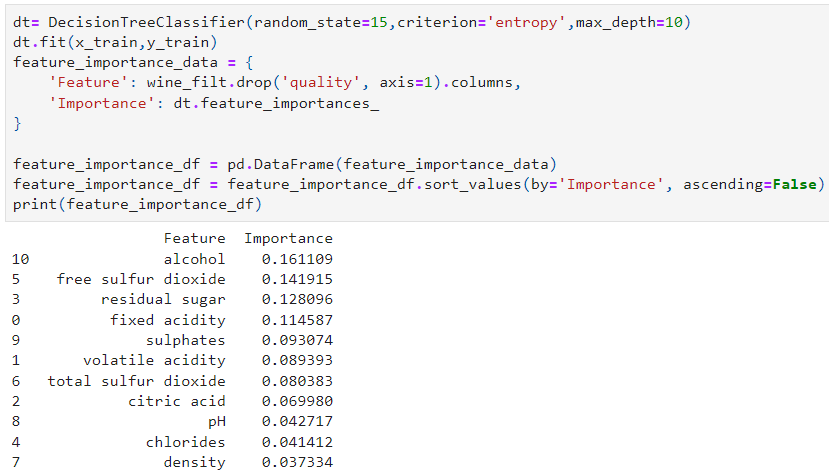


Figure 20: objective 1 feature importance decision tree

Decision tree classifier, which is used to identify important features, reaffirms that feature elimination would not increase the overall performance of the model since the difference between the highest value and lowest value is minimal and there isn’t a significant difference among the features importance.

As for final conclusion random forest classifier model with all featuers performed better compared to the other models with the accuracy score of 74% with the hyperparameter combinations of ['’n\_estimators’: 100, ‘criterion’:gini, ‘max\_depth’:20, ‘min\_samples\_leaf’:1, ‘min\_samples\_split’:2 ], several evaluation metrics (confusion matrix, f1 classification report, ROC curve) were explored to gain a in-depth analysis of the random forest classifier’s model performance.

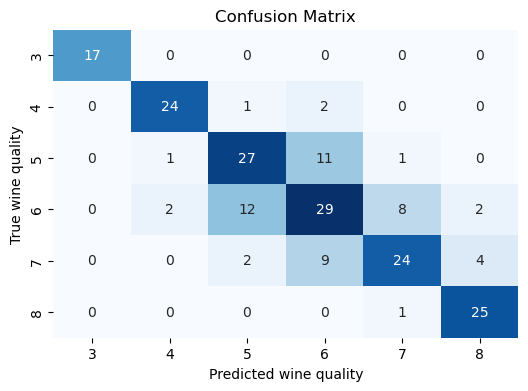


Figure 21: Random Forest classifier confusion matrix

The above confusion matrix for random forest classifier shows the model performed better at predicting the true outcomes, which shows the model is capable enough to establish relationship between the physicochemical variables and predicting the overall wine quality accurately. The diagonal line of the confusion matrix shows the true predictions.

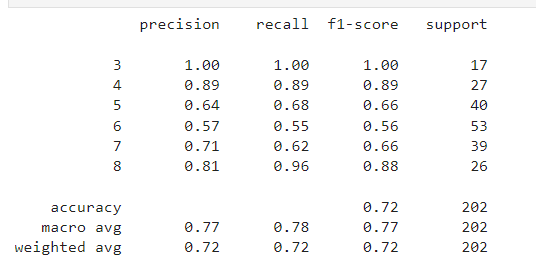


Figure 22: Random Forest classifier F1 classification report

By the exploring the classification report of the random forest classifier, it is quite clear that model has achieved a balance between the precision and recall with the overall weighted average f1-score of 0.72, which explains the model’s ability to predict the overall quality of the wine.

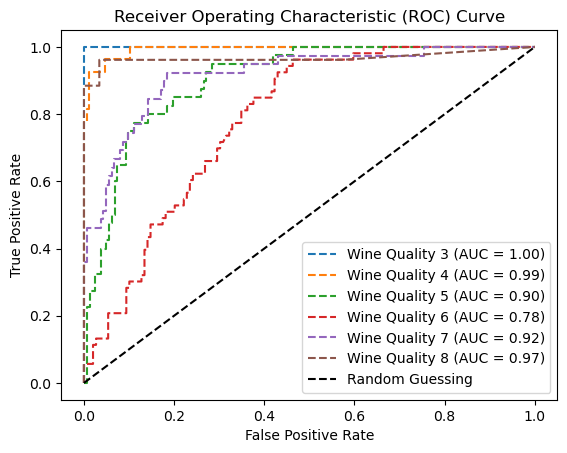


Figure 23: Random Forest classifier ROC

The roc curve further confirms the model’s ability to predict the quality of wine accurately with the help of physicochemical features, the actual prediction line of each individual class is above the base line (0.5).

##### Exploring the Physicochemical factors influence and predictability in distinguishing different types of wine (red wine and white wine)

The second objective follows the similar pattern mentioned in the figure 2 describing the machine learning workflow process, which will be used throughout this research study, second objective deal with the wine type as the target variable. Wine type consists of two entries, red and white wine, since there are only two entries, second objective falls under the preview of binary classification. The pre-processed data frame used in the first objective with stratified random sampling, has been used in the second objective. Data set was reduced, to tackle computational constraints arises from running multiple types of machine learning models.

Right before initializing the phase one of machine learning workflow, target variable used in this objective was visualized using the bar graph to check the imbalance nature of the output variable.

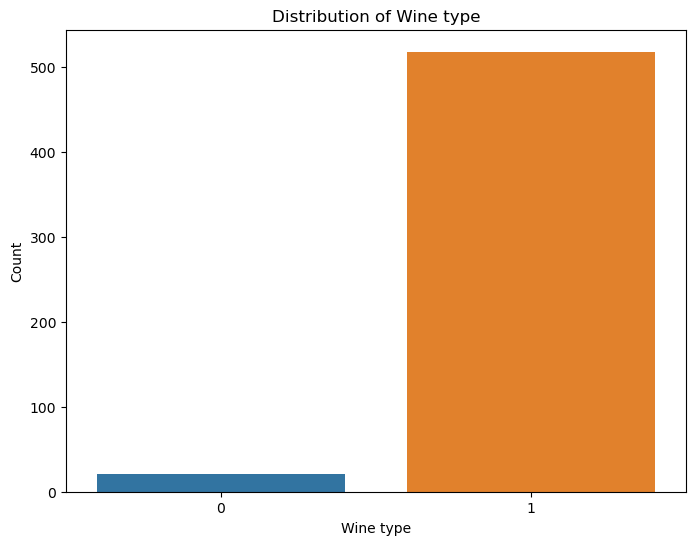


Figure 24: Distribution of wine type prior to SMOTE

From the figure 24 it is quite clear that data entries are highly imbalanced. The difference between the proportion of red wine and white wine is significantly high. Running the data as it is, without any imbalance treatment would lead to model predicting only one specific type of wine in this case 1 (1 = white wine). The entries of red and white have been label encoded to map to a numerical value.

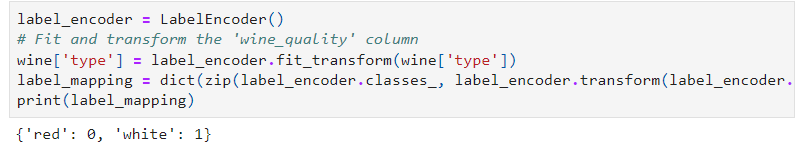


Figure 25: numerical mapping of target variables

To tackle the problem associated with the target variable imbalance, SMOTE algorithm was used to oversample the minority class (red wine) to match the 75% of total entries recorded in the majority class (white wine).

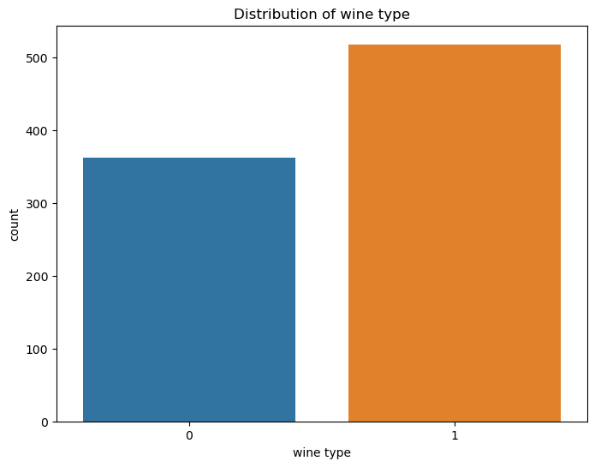


Figure 26: wine type distribution post SMOTE

One the data set is synthetically oversampled, the data set has been split to training and testing set, where 20% of entire data set has been allocated for testing purposes and rest 80% has been allocated for training purposes. The split will be done using five-fold cross validation method to ensure that model performance is accurately measured by using five different folds to validate the model’s performance. The phase one of the machine learning workflow will run, all four machine learning algorithms (Support vector classifier, random forest classifier, logistic regression, K nearest neighbor) with all the feature variables. The models are given different hyperparameters to find the optimal combination of hyperparameters to fine tune the model.

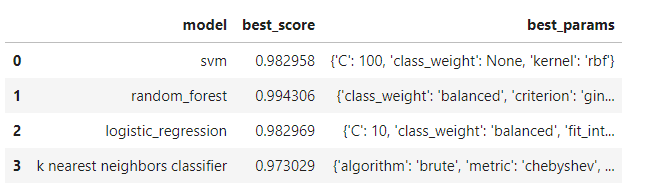


Figure 27: objective 2 Phase 1 model result

The phase one result table shows that random forest with the hyperparameters ( ‘class\_weight’: balanced, ‘criterion’: gini, ‘max\_depth’:30, ‘min\_samples\_leaf’: 2, ‘min\_samples\_split’: 2, ‘n\_estimators’ 100) performed well with the accuracy score of 99%.

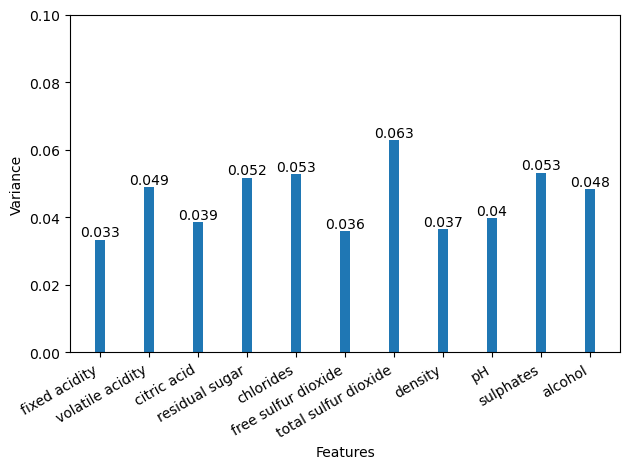


Figure 28: feature variance objective 2

As for the phase two of machine learning workflow, to eliminate less important features and rerunning the model with the same hyperparameters combination, the proposed variance threshold method indicates all the numerical features have some sort of variation among them and the difference between the variation among the highest and lowest feature variance is significantly small, same as objective 1, there by using decision tree classifier the obtained the feature importance score.

A screenshot of a computer

Description automatically generated

Figure 29: objective 2 feature importance score

The decision tree classifier model shows there are several independent variables with feature importance value of 0.00. The phase of two of the objectives two started with eliminating the features with value 0.00 (fixed acidity, volatile acidity, citric acid, residual sugar, free sulfur dioxide, pH), and fitting the machine learning model with same combination of hyper parameters to obtain the accuracy scores of the different models and how it performs with selected highly important features.

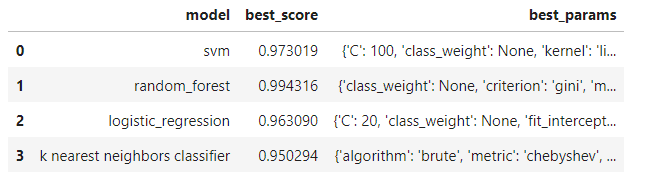


Figure 30: objective 2 phase 2 model result

Eliminating less important features haven’t had any significant contribution towards the model predicting the type of wine with respect to its physicochemical attributes. Certain models like support vector classifier, logistic regression and k nearest neighbors performed poorly compared to model’s performance, prior to feature selection process. Random forest classifier performed similar to the model prior to feature selection process, with the same level of accuracy of 99 %.

As for conclusion for the second objective random forest classifier has been selected, to gain an in-depth analysis of the model, several evaluation metrics such as confusion matrix, classification report and roc curve analysis will be conducted to gain a better understanding of the random forest classifier model’s overall performance.

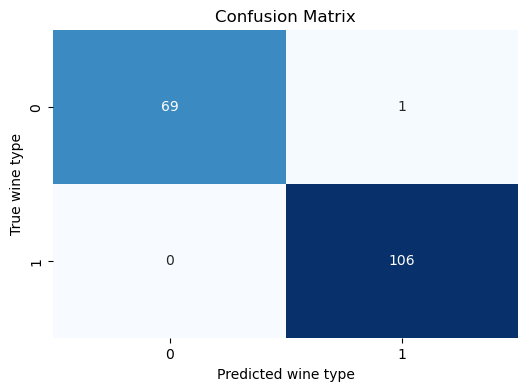


Figure 31: confusion matrix for wine type prediction

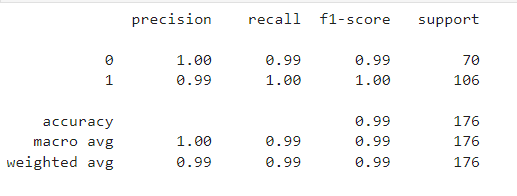


Figure 32: classification report for wine type prediction

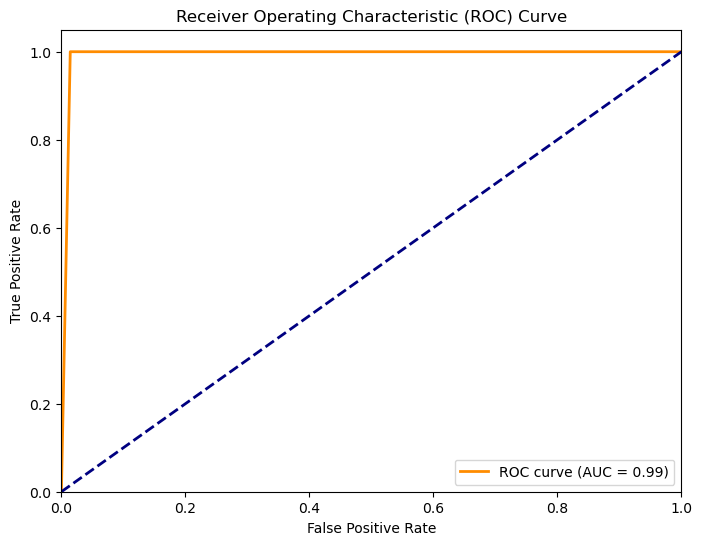


Figure 33: ROC for wine type prediction

The confusion matrix further confirms model’s ability to predict the high number of true positive (white wine) and true negative (red wine) outcome compared to false negative and false positive outcome. The F1 classification report indicates that the model has achieved balance in precision and recall with a weighted average F1 score of 0.99, which signifies the model’s ability to determine the actual wine type based off on the physicochemical factors. The ROC curve shows the model has true positive rate of 0.99 which way above the base line (0.5).

##### Exploration of physicochemical influence on alcohol concentration level in wine

The third objective follows a machine learning workflow process, which includes a feature selection and cross-validation that will be used throughout this research study. The third objective deals with the alcohol concentration amount as the target variable. since this alcohol amount is a numeric figure that contains numeric continuous entries, this third objective falls under the modelling of multiple linear regression.

Before the process of feature selection, initial model was built with all the features. The number of folds is set to 5 to perform a 5-fold cross-validation across the data features, used in the multiple linear regression model. For each fold and obtaining a R mean squared error, mean absolute error, and r-squared value.

A close-up of a number

Description automatically generated

Figure 34: objective 3 cross validation scores

As the output for training accuracy and testing data accuracy separately, training set validations have obtained an RMSE score at each fold that ranges from 0.40860 to 0.42273, which only indicates a slight difference among each. Test sets of each fold show a similar range of RMSE score from 0.40492 to 0.45912 indicating there is only a 0.05 difference at maximum. Respectively, MAE ( mean absolute error) ranges from approximately 0.3142 to 0.3486 across different folds for the training set and approximately 0.3142 to 0.3486 for the test set indicating a 0.03 difference between minimum and maximum errors. Finally, for each fold the linear regression has generated an r-squared score that ranges throughout starting from 0.87879 up to 0.88701 on training and 0.85549 to 0.88773 on the test set of data accordingly.

A screenshot of a computer

Description automatically generated

Figure 35: objective 3 key metrics prior to feature selection

Lower RSME and MAE, means difference between the actual value and predicted values are relatively small, this indicates that model’s ability to make accurate prediction on wine’s alcohol concentration level. On average, the absolute variation between predicted model values and actual values is relatively small, thereby mean accuracy scores were obtained for determined evaluation metrics for multiple linear regression model. The R-squared values represent the idea that the model describes about 88% of the variance in the target variable which is alcohol on both training and testing datasets overall. Since higher R-squared values suggest a good fit of the model to the data it suggests that this model can be used to make good predictions.

A table of information with numbers and letters

Description automatically generated with medium confidence

Figure 36: decision tree regressor feature importance

After analysing the data and its overall accuracy measurements without any feature selections, as the next step starter here explains the feature selections using DesitionTreeRegresser a well-known method for selecting the importance of numeric figures for building regressions. According to the results and importance of each attribute putting a threshold of 0.01 eliminated four variables with low importance based on DesitionTreeRegresser. The rest of the variables which are density, residual sugar, fixed acidity, total sulphur dioxide, pH, and sulphates carried out for further examination following the same prior steps.

A group of numbers on a white background

Description automatically generated

Figure 37: objective 3 cross validation score post feature selection

After feature selection and running through cross-validation again to comply with the previous steps, there was not a significant change in all the accuracy measurements compared to the results which are obtained before. It is clear that small changes of varying ranges in each had made a slight difference but in a negative manner to expand the error measure towards a larger range from non-feature eliminated results, the cross validation score range for RMSE and MAE has seen its range increased compared to score ranges obtained prior to feature selection. However r-squared values for training and testing data set has seen in slight reduction with only selected feature variables.

A white background with black numbers

Description automatically generated

Figure 38: objective 3 key metrics post feature selection

Furthermore, analysing the mean values of each fold’s results compared with the non-feature eliminated method there was a significant difference in results towards a negative direction marking the increment of each error on both training and test data following RMSE from overall 0.41 to a new 0.43 and MAE 0.32 to a new 0.33 indicating that the feature selection causes more errors than first approach with all features included. In addition to that these results also highlight the R-score of 0.87 overall which has a decrement of 0.01 compared to the previous overall R-score of 0.88 another way of saying the new results failed to have an impact on making better results.

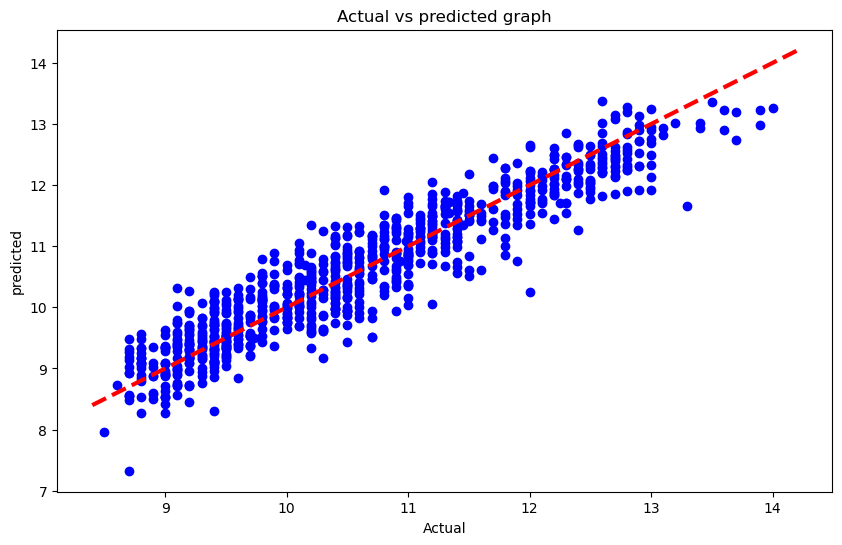


Figure 39: Actual vs predicted multi linear regression

# 

# Chapter 5

# Discussion, recommendations, and conclusion

###### Discussion

This research study explored the application of machine learning algorithms to predict the overall quality and type of wine using the physicochemical attributes of wine, the first objective explores the possibility of using different types of machine learning algorithms to predict the overall quality of red and white wines. The quality of the wine has been recorded in a rating format from 1 to 10 where 1 will be the lowest quality and 10 will be the best quality. Since it deals with multiple classes, machine learning algorithms which can handle multiclass classification has been used (Logistic regression, SVM classification, random forest classification, KNN classification) out of all four-classification algorithm, Random Forest classifier outperformed with the accuracy score of 74% in predicting the overall quality of wine using its physicochemical attributes. The result of the objective one, can be corroborated with several other research which have been listed in the findings by other researcher’s section. This reassures that a random forest classifier is the optimal machine learning algorithm which can handle a multiclass classification problem in predicting the overall quality of the grape wine.

The second objective of this research study explores the possibility of analysing the physicochemical attributes of wine and using those attributes in determining the respective wine type (red or white). The second objective deals with a binary classification problem (red and white), to explore the possibility of predicting the wine type, different types of machine learning algorithms which have the capability to handle binary classification were used. The second objective result showed some similarities between the first objective result, Random Forest classifier with the accuracy score of 99% showed the potential of predicting the wine type using its physicochemical attributes.

The third objective of this research study explores the possibility of predicting the alcohol concentration level of wine with its physicochemical properties. To validate the research objective, a multiple linear regression model was used to derive the conclusion for the third objective of this research study. During phase one model building there was good accuracy of 88% overall on combination of both testing and training data, during phase two of model building, with only selected features the overall accuracy of the multiple linear regression model dropped to 87%. Overall both models indicate a good accuracy score, which implies that model is accurate enough to precisely predict the alcohol concentration level in wine using its physicochemical attributes.

###### Recommendation

Wine making process is delicate and it requires utmost care and maintaining of production quality at optimum levels to make sure the wine produced to the consumer market is within the range of best quality. Wine is the go-to beverage reserved for special occasions. In order for a wine brand to stand out from the rest of the competitors it needs to have the best quality wine for a reasonable price. The overall quality of wine varies based on many factors but one of the well-known common factors are the physicochemical attributes of wine, these physicochemical attributes could be altered during the production process to adjust the overall quality of wine according to customer’s likings or based on the quality type which has the lowest shelf period in the highly competitive consumer market. Typically, the quality of wine is derived from its physicochemical attributes which is combined with a human sensory test. Human sensory test is a timely process and human to human preference changes. It does not have consistent scale when it comes to ranking or determining the overall quality of wine produced, a research study conducted by R. Gawel and P.W. Godden, highlights that human sensory wine quality assessment requires replicate tasting of wine since human sensory scores cannot ensure repeatability, the score keeps changing if the individual tastes the same wine multiple times for quality assessment. Unlike human sensory tests, physicochemical factors are consistent, and it can be easily retrieved from wine, using industry acceptable lab testing equipment. Automation and incorporation of machine learning in predicting the overall quality of wine based on the collected data of wine’s physicochemical levels to predict the overall quality of wine, could benefit the manufacturing firms to produce best quality wines by giving deep understanding on factors affecting the overall quality of wine. Use of machine learning in the production quality testing environment, could be financially beneficial for the company to gain competitive advantage over its competitors.

###### Conclusion

This research report was formulated to explore the physicochemical properties of wine and use those chemical properties to determine the overall quality of wine, type of wine (red wine or white wine), and alcohol concentration level in wine. Several machine learning algorithms were used to validate the objectives of this research study. Use of machine learning models were to explore the possibility of predicting the target variables using wine’s physicochemical properties. Out of all the models used in this research study Random forest classification algorithm performed well in predicting the overall quality of wine with an accuracy score of 74% and the second objective which focuses on determining the discrete type of wine (red or white), out of all the model random forest classifier performed well with an accuracy score of 99% in predicting the type of wine using it physicochemical properties. This shows that machine learning algorithms could be used in wine production in making accurate predictions on overall quality of wine, which could benefit the production company in the long run by incorporating machine learning techniques in their production process. The model used to determine the alcohol concentration in wine, multiple linear regression model performed well with an accuracy score of 88% in predicting the alcohol concentration level in wine using the rest of the physicochemical attributes of the wine. This shows machine learning models could be used in the wine production process to train the model to identify pattern and relationship between the features and target variable, which can be further refined to make accurate predictions on determining the overall quality, type, and alcohol concentration level in wine.

# Reference

Acid (2023) Encyclopedia Britannica. Available at: https://www.britannica.com/science/acid (Accessed: 13 November 2023).

Aich, S. et al. (2019) ‘Prediction of quality for different type of wine based on different feature sets using supervised machine learning techniques’, 2019 21st International Conference on Advanced Communication Technology (ICACT) [Preprint]. doi:10.23919/icact.2019.8702017.

Brownlee, J. (2021) Smote for imbalanced classification with python, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (Accessed: 17 November 2023).

Cortez, P. *et al.* (2009) ‘Modeling wine preferences by data mining from physicochemical properties’, *Decision Support Systems*, 47(4), pp. 547–553. doi:10.1016/j.dss.2009.05.016.

Casquete, R. et al. (2021) ‘Addition of grape skin and stems extracts in wines during the storage to reduce the sulfur dioxide: Impact on Red Wine Quality’, International Journal of Environmental Research and Public Health, 18(5), p. 2783. doi:10.3390/ijerph18052783.

Coli, M.S. et al. (2015) ‘Chloride concentration in red wines: Influence of terroir and grape type’, Food Science and Technology (Campinas), 35(1), pp. 95–99. doi:10.1590/1678-457x.6493.

Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. <https://doi.org/10.24432/C56S3T>.

Gawel, R. et al. (2013) ‘Effect of ph and alcohol on perception of phenolic character in white wine’, American Journal of Enology and Viticulture, 64(4), pp. 425–429. doi:10.5344/ajev.2013.13016.

Gupta, A. (2023) *Feature selection techniques in Machine Learning (updated 2023)*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/#:~:text=The%20variance%20threshold%20is%20a,same%20value%20in%20all%20samples. (Accessed: 17 November 2023).

GAWEL, R., and GODDEN, P.W., 2008. Evaluation of the consistency of wine quality assessments from expert wine tasters. Australian Journal of Grape and Wine Research, 14(1), pp.1–8.

Hakim, S. (2018) Citric acid, Viticulture and Enology. Available at: https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid (Accessed: 13 November 2023).

Koranga, M.,, Pandey, R.,, Joshi, M., and Kumar, M., 2021. Analysis of white wine using machine learning algorithms. *Materials Today: Proceedings*, 46, pp.11087–11093.

Ozturk, B. and Anli, E. (2014) ‘Different techniques for reducing alcohol levels in wine: A Review’, BIO Web of Conferences, 3, p. 02012. doi:10.1051/bioconf/20140302012.

Mor, N.S. *et al.* (2022) ‘Wine quality and type prediction from physicochemical properties using neural networks for Machine Learning: A Free Software for winemakers and customers.’, *agriRxiv*, 2022. doi:10.31220/agrirxiv.2022.00125.

What are microorganisms? (2010) University of Bergen. Available at: https://www.uib.no/en/geobio/56846/what-are-microorganisms#:~:text=Technically%20a%20microorganism%20or%20microbe,generally%20classified%20as%20non%2Dliving. (Accessed: 14 November 2023).