Assignment No.: _____

HELP University
university of achievers

Assignment Cover Sheet

Student Information (For group assignment, please state names of all members)		Grade/Marks
Name	**ID**	
Chin Zheng Yin	B2101086	

Module/Subject Information		Office Acknowledgement
Module/Subject Code	BDA205	
Module/Subject Name	Data Mining and Visualization	
Lecturer/Tutor/Facilitator	Ts. Dr Yong Yoke Leng	
Due Date	25th November 2024	
Assignment Title/Topic	Final Assessment	
Intake (where applicable)	Semester 3, 2024	
Word Count	n/a	**Date/Time**

Declaration

- I/We have read and understood the Programme Handbook that explains on **plagiarism**, and I/we testify that, unless otherwise acknowledged, the work submitted herein is entirely my/our own.
- I/We declare that no part of this assignment has been written for me/us by any other person(s) except where such collaboration has been authorized by the lecturer concerned.
- I/We authorize the University to test any work submitted by me/us, using text comparison software, for instances of plagiarism. I/We understand this will involve the University or its contractors copying my/our work and storing it on a database to be used in future to test work submitted by others.

Note: 1) The attachment of this statement on any electronically submitted assignments will be deemed to have the same authority as a signed statement.
2) The Group Leader signs the declaration on behalf of all members.

Signature: Chin Zheng Yin	Date: 2024, November 25
E-mail: B2101086@helplive.edu.my	

Assignment No.: _____

Feedback/Comments*
Main Strengths
Main Weaknesses
Suggestions for improvement

	Student acknowledge feedback/comments
Grader’s signature	Student’s signature:
Date:	Date:

Note:

1) A soft and hard copy of the assignment shall be submitted.
2) The signed copy of the assignment cover sheet shall be retained by the marker.
3) If the Turnitin report is required, students have to submit it with the assignment. However, departments may allow students up to **THREE** (3) working days after submission of the assignment to submit the Turnitin report. The assignment shall only be marked upon the submission of the Turnitin report.

*Use additional sheets if required.

Table of Contents

1.0 Introduction

1.1 Overview

According to Latessa et al. (2023), the wine industry has become a significant economic pillar in some countries. Thus, there is no doubt that this industry would be very competitive. To remain their market position and boost sales, wine production companies or independent winemakers have to focus on producing high-quality wine. Naturally, it is crucial for them to know the chemical properties that would directly influence the wine quality and wine type. With the knowledge of producing high-quality wine based on the wine type, production can be more effective as well. Overall, it ensures consistent quality for each production of wine.

However, it would be challenging if wine quality is determined based on human taste as it is such a subjective matter. Therefore, organizations start to use technologies to predict the quality of wine and wine type based on the relevant chemical properties. Several data sets are made available on websites for organizations to use and analyze. With the dataset, machine learning models can be trained to predict wine qualities or wine types. This way, winemakers can have a better idea on the quality of their production.

This report focuses on addressing two problems related to the wine industry using data mining techniques such as Random Forest, Neural Network and Gradient Boosting. This first problem would be predicting the quality of red wine and white wine while the second problem would be classifying red wine and white wine. The tasks would be carried out using the same set of chemical properties. Upon evaluating the performance of each data mining model, additional findings would be uncovered as well.

1.2 Data Description

The dataset used is provided in UCI machine learning repository. The link for the raw dataset would be provided in Appendix A. The dataset provided basically covers the chemical properties of the red and white wine and also the type of wine. In the start, there are two separate datasets: red wine dataset and white wine dataset. To make data exploration easier, both of the datasets are combined using a new attribute called “type”. This process in done using the program R Studio. In the end, there is a total of 6,497 rows in the dataset, having 13 variables to differentiate between them. 12 variables would be numerical variables while 1 variable would be the categorical variable. Table 1 shows all the variables involved in this dataset.

Table 1

Variables Involved in Wine Quality Dataset

Variables	Type	Description
fixed.acidity	Numerical	Amount of fixed acid in the wine. This includes acids such as tartaric acids, malic

		acids and lactic acid. These acids are the most durable in the wine, giving the wine its distinct flavor.
volatile.acidity	Numerical	Amount of volatile acid, mainly acetic acid in the wine. It is mostly gaseous, giving the wine a strong and unpleasant smell if the level is too high.
citric.acid	Numerical	Amount of citric acid in the wine. It helps to increase the acidity of the wine. It is also a form of preservative to prevent unwanted bacteria from growing inside. Ultimately, it keeps the wine fresh.
residual.sugar	Numerical	Amount of sugar remaining after fermentation. It makes the wine to have a sweeter taste.
chlorides	Numerical	Number of chlorides in the wine. It adds the saltiness taste to the wine, affecting the taste and texture of it.
free.sulfur.dioxide	Numerical	Amount of free sulfur dioxide in the wine. This would mean the number of remaining sulfur dioxide that can help in preserving the wine.
total.sulfur.dioxide	Numerical	Total amount of sulfur dioxide in the wine. It acts an antioxidant and also a preservative in the wine.
density	Numerical	Density of the wine.
pH	Numerical	The pH value of the wine. The lower the pH value, the more acidic the wine
sulphates	Numerical	Amount of sulphates in the wine. It acts as preservative to prevent bacteria and yeast from growing in the wine.
alcohol	Numerical	Alcohol level of the wine.
quality	Numerical	Rank of wine quality from 3 to 9.
type	Categorical	Type of wine (red wine or white wine)

Note. Table 1 shows all the variables and their respective meaning.

Each variable stated in Table 1 can significantly relate to the business problem. For starters, variables such as “fixed.acidity”, “volatile.acidity” and “pH” can let wine production companies to analyze the acidity level of each wine and see how it affects the taste of the wine. In the end, the taste of the wine automatically determines the quality of it.

Variables like “alcohol”, “sulphates” and “sugar” can allow wine production companies to adjust to their wine’s fermentation process. Different fermentation process would produce different qualities of wine.

Lastly, the “quality” and “type” variables would be the target variables in the dataset. Based on the chemical properties stated, it helps to determine whether the quality of the wine in high or low and it can help to classify whether it would be white wine or red wine. With that, wineries can improve their production and making sure every wine that they produce have high qualities.

2.0 Problem Formulation

There are two problems that can be formulated through this dataset using data mining analysis. Table 2 and Table 3 shows the problems defined, their objects and other references from research papers regarding on the same issues.

Table 2

Problem 1 Definition, Objectives and Literature Review

Problem 1	Predicting quality of red wine and white wine based on their chemical properties
Definition	It is important for wineries to predict the quality of their own wine based on the chemical contents in it. Instead of carrying out human taste testing which is highly subjective, the chemical properties would be more direct and straightforward when it comes to determining the quality of the wine itself. Using the dataset, we can see how different chemical properties would affect the quality of specific types of wine as well.
Objectives	The main objectives of this problem would be: 1. To identify the most significant chemical feature that would determine the quality of wine 2. To predict the score of wine quality based on all the chemical features 3. To evaluate the model’s performance

	4. To provide insights to wineries so that they can better improve their wine quality
Literature Review	According o Bhardwaj et al. (2022), most wine businesses would need a wine quality certification to remain their reputation in the industry. To further determine the quality of wine, machine learning models can be used. In this modern era, advanced technology have the ability to make correct interpretations to help wine businesses to determine the quality of wine. With that, predictions of wine quality based on the chemical characteristics are carried out using machine learning techniques such as Random Forest and Nave Bayes. In this paper, there is a limited number of raw data available. Thus, a small dataset might lead to bias or inaccurate predictions in the future. Therefore, in this analysis report, the raw data is much larger than the data mentioned in the literature paper. With that, it would help to build on to the analysis of predicting wine quality using machine learning techniques. On top of that, Dahal et al. (2021) also carried out research on the same issue. Dahal et al. (2021) implemented machine learning algorithms such as Ridge Regression, Gradient Boosting Regressor and Artificial Neural Network (ANN) to predict the wine quality. As a conclusion in the analysis, it is stated that Gradient Boosting Regressor would be the best to predict the wine quality. On top of that, it is also highlighted that outliers will greatly affect the prediction. Thus, this matter is something to be taken note of during the analysis paper. With that, the model suggested in this paper can be considered when carrying out this analysis.

Note. Table 2 shows the overall summary of the problem of predicting wine qualities mainly based on their chemical properties.

Table 3

Problem 2 Definition, Objectives and Literature Review

Problem	Classifying type of wine based on their chemical properties
Definition	Instead of looking at labels of the wine bottle, it is good if wineries can differentiate between red wine and white wine. The wineries can use the chemical properties in the wine itself to classify whether it is a red wine or white wine. In the case of where wine labels might be wrong or inconsistent, having a classification model to safely predict the type of wine would be useful as well.
Objectives	The main objective of this problem would be: 1. To examine the difference in chemical properties of red wine and white wine. 2. To build a classification model to categorize the type of wine accurately 3. To evaluate the model's accuracy 4. To implement model so that wine can be categorized automatically during the production or quality control phase.
Literature Review	In Er & Atasoy (2016) research, they have also implemented machine learning models to classify red wine and white wine based on their chemical properties. There are three techniques used in their data mining analysis which are k-Nearest Neighborhood (k-NN) Classifiers, random forest (RF) and support vector machines. Out of these algorithms, it is shown that the RF model used would be the most efficient and accurate classifying model. Therefore, RF algorithms would be highly considered to be used for this analysis. Nevertheless, it would be fair to explore other models as well. Other than that, Perez-Magariño et al. (2004) also tried to classifies Spanish rose wines using Artifical Neural Networks (ANN). He stated that ANN technique is suitable as it can handle complexed problems. It is not affected by the imbalances in the sample and does not restrict to any types of data. However, though being very flexible, ANN requires more computing skills and its results would be harder to interpret. Thus, in this case, the classification of red wine and white wine can also consider using

	ANN while taking the drawbacks into consideration at the same time. With that, there will be less restriction.

Note. Table 3 shows the overall summary of the problem of classifying whether it is a red wine or white wine based on its chemical properties.

3.0 Methodology

3.1 Methodology chosen

The methodology used in this data mining analysis would be **SEMMA**. SEMMA stands for Sample, Explore, Modify, Model and Assess. The reason behind choosing this methodology is because SAS tools, mainly SAS Studio and SAS Enterprise Miner would be used throughout this entire data mining analysis. Since SEMMA methodology works best with SAS environment, the data mining analysis process would be much easier. On top of that, using this methodology allows quick transitions between different SAS tools, creating a more user-friendly environment. Other than that, SEMMA focuses on using the sample to build model, preventing underfitting and overfitting. Thus, this methodology would be ideal for this data mining analysis as it would be testing different machine learning models throughout the analysis and then evaluate them along the way. Lastly, SEMMA would help in streamlining the automation process within SAS platforms. This allows the data mining process to be much easier and faster. With that, the problems can be answered in a more efficient way.

3.2 Steps taken in each phase

Overall, SEMMA consists of 5 phases: sample, explore, modify, model and assess. This subsection will explain each phase and the steps that would be taken.

3.2.1 Sample

This phase primarily focuses on extracting the sample data that would be used to carry out the data mining analysis. This sample data would be the representative data to build the model and to be explored. In order to do this, Table 4 shows each step taken and its significance in the analysis.

Table 4

Steps Taken in Sample Phase

Steps	Justification
Formulate problem	The problem formulated would have to be defined clearly with its objectives. This is to make sure that we would not go out of scope as the data mining analysis proceeds. In this case, it would be predicting the wine quality and classifying the type of wine.
Extract sample data	A sample set of data from the dataset would be extracted. This is to make the dataset to be more manageable. Fortunately, SAS Enterprise Miner has a tool to extract representative data using either random sampling procedure or stratified sampling. With that, a more proper representation can be extracted from the raw data.
Verify sample data quality	This is to make sure that the sample data extracted fully reflects the overall dataset. The distribution of important datasets such as “quality” and “type” would be examined to ensure that they are the same as the overall dataset.

Note. Table 4 shows steps that would be taken in the Sample phase of SEMMA methodology, together with each of their justifications.

3.2.2 Explore

The explore phase would focus on carrying out exploratory data analysis (EDA). This phase is important so that we can understand the patterns and relationship of each variable in the sample data. Table 5 shows the steps taken in this phase with their explanation.

Table 5

Steps Taken in Explore Phase

Steps	Justification
Univariate analysis	This will be able to obtain the key statistics of each variable in the dataset. This includes the mean, range and standard deviation. With that, the distribution of the data points in each variable can be known so that necessary adjustments can be made.

Multivariate analysis	This analysis is helpful to see the relationship between different variables. Correlation coefficient of between two variables can be calculated to understand the type of relationship and the strength of the relationship of these variables.
Identify noisy data	Noisy data includes missing data, inconsistent data and outliers. Upon identifying, specific methods will be implemented to handle these data so that the analysis can be carried out smoothly later on.
Create visualization	With visualizations, data relationship and correlation between different variables can be viewed easily. For example, scatter plots can be created to clearly see the direct relationship between two variables. On the other hand, box plots can be used to see the distribution of data point of a specific variable

Note. Table 5 shows steps that would be taken in the Explore phase of SEMMA methodology, together with each of their justifications.

3.2.3 Modify

In this phase, we would use the result in the EDA process to clean and modify the sample data. This phase would emphasize on handling the noisy data identified in the Explore phase and also partitioning the sample data. Table 6 shows the detailed steps taken in this phase including their respective justifications.

Table 6

Steps Taken in Modify Phase

Steps	Justification
Handle missing data	This step is to make sure that there are no missing data present in the dataset, resulting a skew afterwards. To handle this, the missing data can either be handled through imputation, replacement, or deletion. At the end of the day, there should be no missing data left in the dataset.
Encoding variables	Some machine learning only works with numbers. Thus, the categorical data can be changed to numbers

	so that the models can smoothly process these variables.
Transform necessary variables	Normalizing and scaling can be carried out on the continuous variables in the sample data so that it increase the accuracy of the machine learning model.
Handle outliers	Outliers might cause bias in the machine learning model. Therefore, it is essential to address them. For one, if that specific data is insignificant, it can be removed. If the outliers are important and should be kept, they would be transformed to fit the model better.
Selecting key features	The key features should be the key factors that would affect the prediction of wine quality and classification of wine type. Based on the correlation analysis in the EDA process, these features can be determined.
Partition sample data	The sample data would be partitioned into two sets: training and validation. With that, the training data set would be used to train the model while the validation data set would be used to validate the effectiveness of the model. Overall, it prevents overfitting the machine learning model.

Note. Table 6 shows steps that would be taken in the Modify phase of SEMMA methodology, together with each of their justifications.

3.2.4 Model

In the Model phase, this is where actual machine learning models would be implemented and tested on their performance. Overall, the models implemented would be used to answer the formulated problems. Table 7 shows the steps undergone in this phase that is revolving around the machine learning models chosen.

Table 7

Steps Taken in Model Phase

Steps	Justification
Select data mining models	The data mining models would fully algin with the problems formulated. It is considered based on the result that is desired after the data mining analysis. For prediction, random forests and Gradient Boosting

	Regressor can be used; for classification, decision trees and neural networks can be used.
Train models	Train the machine learning models using the partitioned training data set. After that, their performance can be assessed using the validation data set.
Tuning hyperparameters	This step allows us to explore the accuracy of the data mining model. For example, the depth of the RANDOM can be adjusted to create a more detailed and accurate split among nodes.
Compare the models	After the models are trained and validated, the models will be used to compare between one another. It would be the comparison between the accuracy, precision and other metrics.

Note. Table 7 shows steps that would be taken in the Model phase of SEMMA methodology, together with each of their justifications.

3.2.5 Assess

In this phase, the models selected would be evaluated and be considered for deployment. It would conclude each efficiency of each model and determine whether the model can be used in the future. Table 8 shows the steps in this phase and how they are conducted.

Table 8

Steps Taken in Assess Phase

Steps	Justification
Evaluate model performance	The model performance would be evaluated using several metrics such as misclassification rate, RASE and MAE. This is so that we can compare the exact performance statistics of each data mining model.
Compare data sets	The performance of the data mining model towards each data set is evaluated as well. This includes performance when using training data and validation data.
Sensitivity analysis	This steps mainly mentions about the feature variables that are more sensitive during model training. These

	variables would affect the prediction or the classification at a certain level.
Report	This step would summarize the overall performance of each model, including the recommendations for future similar analysis.

Note. Table 8 shows steps that would be taken in the Assess phase of SEMMA methodology, together with each of their justifications.

3.3 Data Mining Models Selected

In this section, the data mining models selected would be listed down. Overall, it is decided that Neural Network, Random Forest and Gradient Boosting would be used. Each model is further discussed in their respective subsections. For the metrics, same metrics would be used throughout three models. For Problem 1, ROC index, misclassification rate and RASE value will be used; for Problem 2, the accuracy, precision, recall and F1-score will be used.

3.3.1 Neural Network

Neural Network is a type of machine learning model that is inspired by the structure and functioning of the human brain. From Han et al. (2018), they stated that just like a human, an input signal is received, goes through processing and then produces an output. In this case, the output will only be produced when the input signal exceeds a designated threshold. There is a function called activation function that would be triggered and activated that would process the input signal and produce the final output. Choi et al. (2020) also mentioned that Neural Network models usually consist of an input layer, an output layer and multiple hidden layers in between the input layer and output layer. The input layer would receive the input nodes. Then, the hidden layers would consist of multiple layers, all performing computations and extracting hidden patterns from the data. The output layer would then produce the final output of the model. In this case, the output would either the predicted quality of wine or the type of wine.

Without a doubt, Neural Networks models can address the problems in this data mining analysis. The reason is because they are suitable for predicting both binary and nominal target variables (Banoula, 2024). Thus, it is flexible and can carry out the tasks required to answer Problem 1 and Problem 2. The models are able to learn subtle, non-linear pattern dependencies between features and quality. With the hidden layers, they can detect complexed relationships in the dataset and then capture the subtle differences in each target class. Therefore, when it comes to predicting quality score of red wine and white wine, Neural Networks models can capture a deeper relationship between chemical properties of the wine and the quality. On top of that, Neural Network models can be very powerful in classification tasks. Thus, classifying the type of wine would not be an issue during the data mining process.

3.3.2 Random Forest

Random Forest models are mainly made up from multiple decision trees. Multiple subsets of the sample data would be used to train specific decision trees. Other than that, from the each training data sample, another small sample called out-of-bag (OOB) sample will be extracted to further evaluate the performance of the model trained (*What is random forest?, n.d.)*. Feature bagging is also carried out to randomize the sample data as well. In this case, not all feature variables will be used to train all the decision trees in the model. In the end, to produce a final prediction, the majority vote among each decision is chosen.

This model is chosen because it is an ensemble method, eliminating the risk of overfitting or underfitting. Random forest model can calculate the importance of each feature variable in determining the final prediction of the target variable. This gives additional insights to the stakeholders on which chemical properties matters the most in predicting the wine quality and the type of wine. To address the problem of predicting the wine quality, the final prediction of the wine quality would be derived through the average prediction of all decision trees. On the other hand, the majority voting of a type of wine would be the final classification for the wine type.

3.3.3 Gradient Boosting

Gradient Boosting models are also an ensembled algorithm which is made up from multiple weak learners (usually decision trees). The thing different about this model would be the way it trains itself. Gradient Boosting models trains each weak learner sequentially and not concurrently (Belyadi & Haghighat, 2021). Each new weak learner would be compared with the actual values from the dataset. Then, the errors are corrected in the next weak learner produced. In other words, each new decision tree tries to reduce the errors made by the combined ensemble up to that point. With that, the specific loss function can be minimized with its efficient iterative learning approach. It

This model is chosen to address the problem due to its high learning ability. By solving the errors in each iteration, the accuracy and the performance of the model would be high. On top of that, it can capture complexed relationships at the same time. For Problem 1, this model can capture the errors made when predicting the quality of red wine and white wine. With that, the nuanced relationship can be known by the model, allowing to perform better. As for Problem 2, using Gradient Boosting models to classify binary target variable would also be high effective. The model enhances its predictions iteratively and outperforms other machine learning models that only has one architecture.

4.0 Results and Discussion

4.1 Data Preprocessing

The first three phases the SEMMA methodology: Sample, Explore Modify would be carrying out data preprocessing tasks. For starters, the dataset will be explored and then modified as a whole. This is to ensure data consistency for the overall dataset. Then, only a sample dataset will be extracted. In that case, a clean sample data set is assured.

4.1.1 Data Import

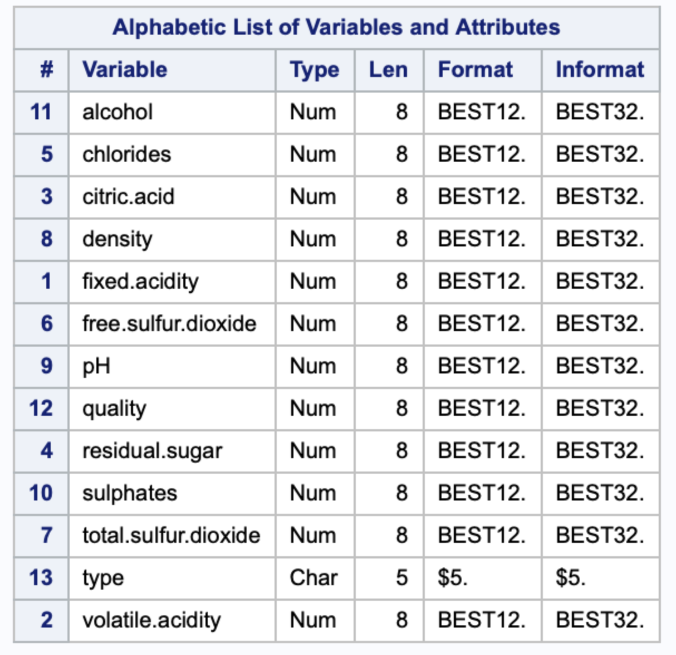
Before exploring the dataset, it is important to know that there are initially two separate datasets: red wine and white wine dataset. Both attributes in the dataset are exactly the same. Throughout this entire data mining analysis, these two datasets and also a combined dataset are required. To combine the two datasets, a new attribute "type" is used to merge the two datasets in R Studio. The combined dataset, together with the separate datasets are all imported to SAS Studio as a folder.

4.1.2 Data Exploration

After importing the datasets to SAS Studio, the first thing would be making sure that the data types are correct and suitable for analysis. The combined dataset would be used to see the overall data types of each variable. Figure 1 shows the final results of the data types in the combined dataset.

Figure 1

List of Table Attributes and Their Types



Alphabetic List of Variables and Attributes

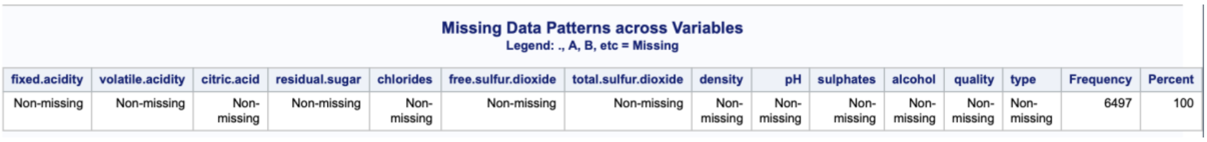
#	Variable	Type	Len	Format	Informat
11	alcohol	Num	8	BEST12.	BEST32.
5	chlorides	Num	8	BEST12.	BEST32.
3	citric.acid	Num	8	BEST12.	BEST32.
8	density	Num	8	BEST12.	BEST32.
1	fixed.acidity	Num	8	BEST12.	BEST32.
6	free.sulfur.dioxide	Num	8	BEST12.	BEST32.
9	pH	Num	8	BEST12.	BEST32.
12	quality	Num	8	BEST12.	BEST32.
4	residual.sugar	Num	8	BEST12.	BEST32.
10	sulphates	Num	8	BEST12.	BEST32.
7	total.sulfur.dioxide	Num	8	BEST12.	BEST32.
13	type	Char	5	$5.	$5.
2	volatile.acidity	Num	8	BEST12.	BEST32.

Note. Figure 1 depicts the list of attributes in the combined dataset. There are 12 variables in total, 11 of the variables are numeric whereas 1 of the variables is categorical data. The type of the attributes are all suitable and so no modifications are needed.

Secondly, missing values in the entire dataset are also checked. If there are any missing values, they should be handled through imputations or removal. Figure 2 shows the final result of missing data in each variable.

Figure 2

Number of Missing Data in the Dataset



Missing Data Patterns across Variables
Legend: ., A, B, etc = Missing

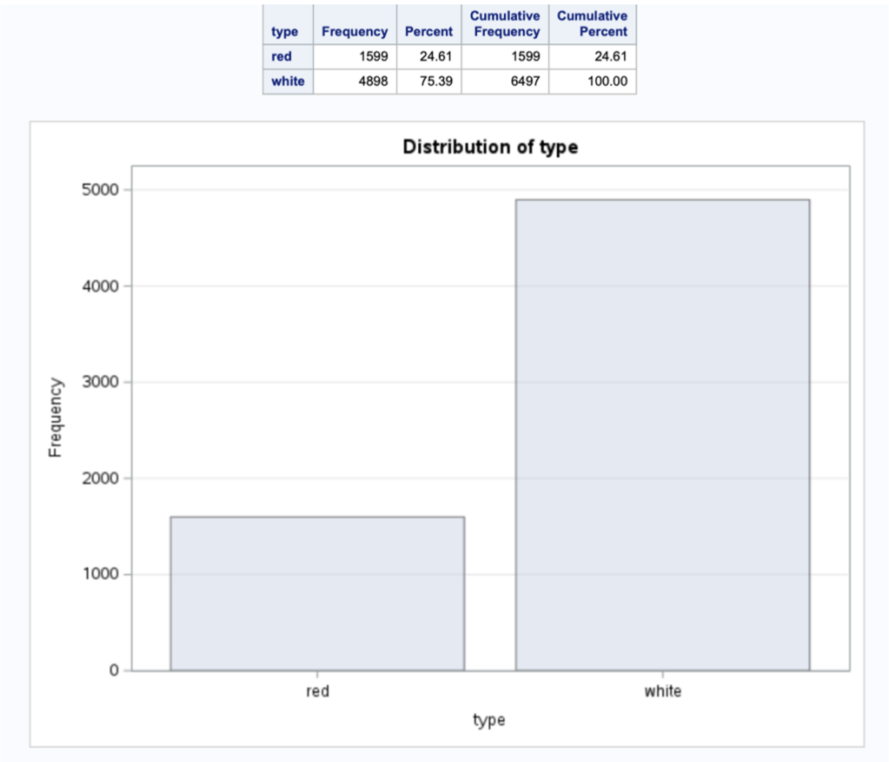
fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	type	Frequency	Percent
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	6497	100

Note. Figure 2 depicts the number of missing data in the whole dataset. As a result, it is shown that there are no missing data. Therefore, the dataset can proceed to be analyzed on other aspects.

Another task that is carried out is to analyze the distribution of the dataset. For starters, since there are two wine types in the dataset, the distribution of wine type is analyzed. Figure 3 shows the histogram and the descriptive table of the distribution of wine type in the dataset.

Figure 3

Wine Type Distribution



type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
red	1599	24.61	1599	24.61
white	4898	75.39	6497	100.00

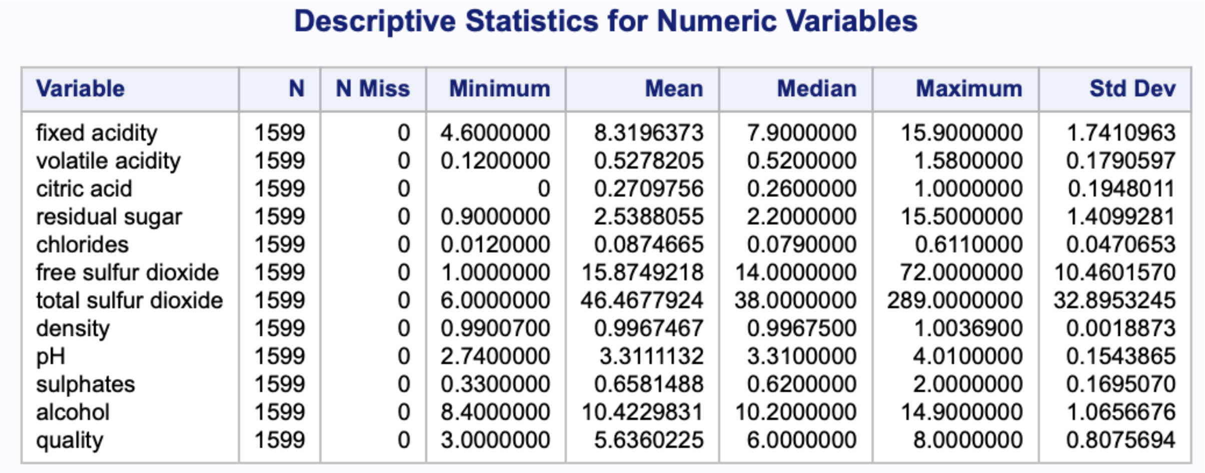
1

Note. Figure 3 depicts the distribution of the different wine types in the dataset. It is shown that there a more data about white wine compared to red wine. There is a total of 1599 rows about red wine whereas there is a total of 4898 rows in the dataset that is categorized as white wine.

After exploring the combined dataset, it is also important to view the descriptive statistics of the numerical values in each of the dataset. Thus, the descriptive statistics of the red wine, white wine and the combined dataset would be shown in Figure 4 to Figure 6 respectively.

Figure 4

Descriptive Statistics of the Numerical Variables in the Red Wine Dataset



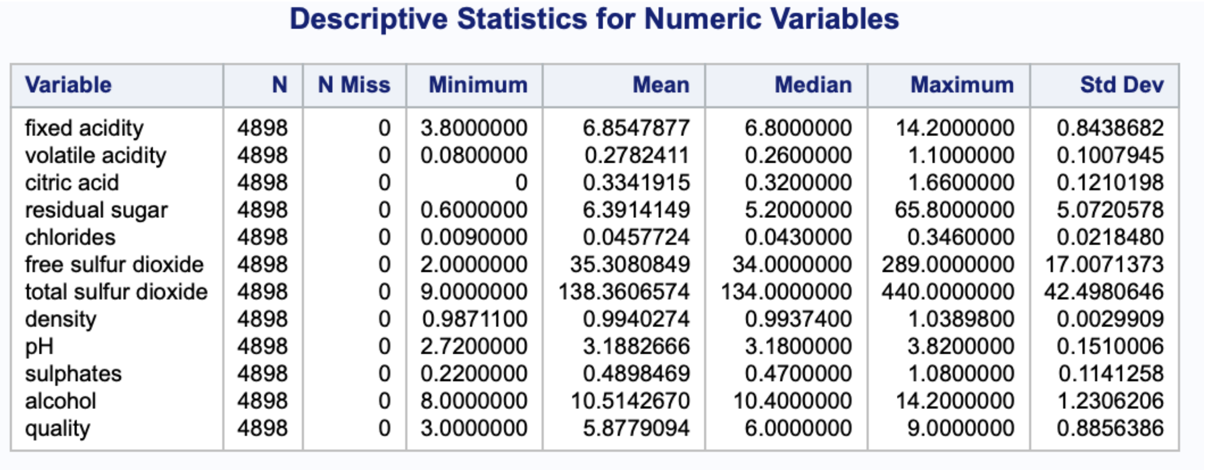
Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
fixed acidity	1599	0	4.6000000	8.3196373	7.9000000	15.9000000	1.7410963
volatile acidity	1599	0	0.1200000	0.5278205	0.5200000	1.5800000	0.1790597
citric acid	1599	0	0	0.2709756	0.2600000	1.0000000	0.1948011
residual sugar	1599	0	0.9000000	2.5388055	2.2000000	15.5000000	1.4099281
chlorides	1599	0	0.0120000	0.0874665	0.0790000	0.6110000	0.0470653
free sulfur dioxide	1599	0	1.0000000	15.8749218	14.0000000	72.0000000	10.4601570
total sulfur dioxide	1599	0	6.0000000	46.4677924	38.0000000	289.0000000	32.8953245
density	1599	0	0.9900700	0.9967467	0.9967500	1.0036900	0.0018873
pH	1599	0	2.7400000	3.3111132	3.3100000	4.0100000	0.1543865
sulphates	1599	0	0.3300000	0.6581488	0.6200000	2.0000000	0.1695070
alcohol	1599	0	8.4000000	10.4229831	10.2000000	14.9000000	1.0656676
quality	1599	0	3.0000000	5.6360225	6.0000000	8.0000000	0.8075694

Note. Figure 4 depicts the table showing the descriptive statistics of the numerical variables in the red wine dataset.

Figure 5

Descriptive Statistics of the Numerical Variables in the White Wine Dataset



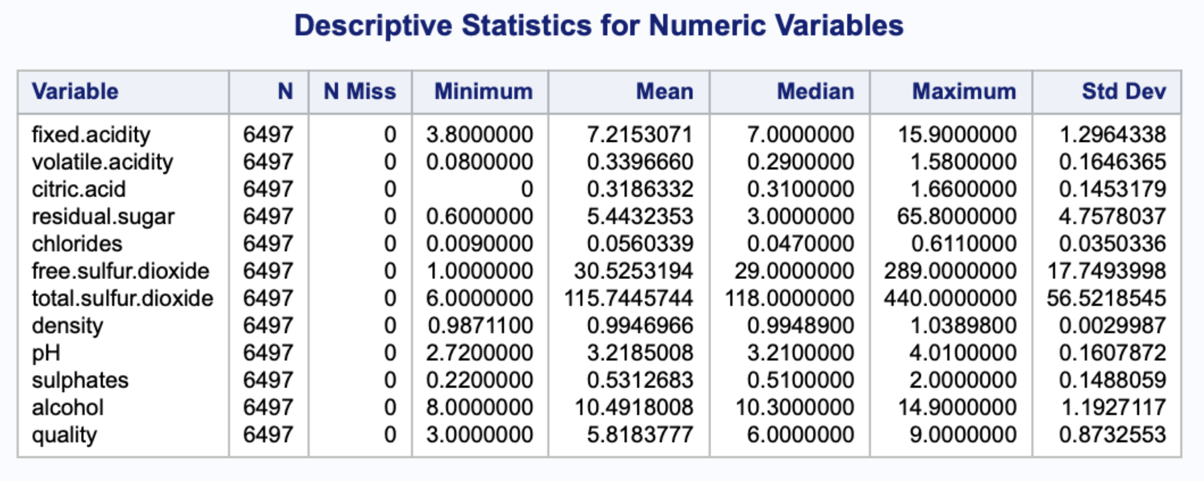
Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
fixed acidity	4898	0	3.8000000	6.8547877	6.8000000	14.2000000	0.8438682
volatile acidity	4898	0	0.0800000	0.2782411	0.2600000	1.1000000	0.1007945
citric acid	4898	0	0	0.3341915	0.3200000	1.6600000	0.1210198
residual sugar	4898	0	0.6000000	6.3914149	5.2000000	65.8000000	5.0720578
chlorides	4898	0	0.0090000	0.0457724	0.0430000	0.3460000	0.0218480
free sulfur dioxide	4898	0	2.0000000	35.3080849	34.0000000	289.0000000	17.0071373
total sulfur dioxide	4898	0	9.0000000	138.3606574	134.0000000	440.0000000	42.4980646
density	4898	0	0.9871100	0.9940274	0.9937400	1.0389800	0.0029909
pH	4898	0	2.7200000	3.1882666	3.1800000	3.8200000	0.1510006
sulphates	4898	0	0.2200000	0.4898469	0.4700000	1.0800000	0.1141258
alcohol	4898	0	8.0000000	10.5142670	10.4000000	14.2000000	1.2306206
quality	4898	0	3.0000000	5.8779094	6.0000000	9.0000000	0.8856386

Note. Figure 5 depicts the table showing theCR descriptive statistics of the numerical variables in the white wine dataset.

Figure 6

Descriptive Statistics of the Numerical Variables in the Combined Dataset



Descriptive Statistics for Numeric Variables

Variable	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
fixed.acidity	6497	0	3.8000000	7.2153071	7.0000000	15.9000000	1.2964338
volatile.acidity	6497	0	0.0800000	0.3396660	0.2900000	1.5800000	0.1646365
citric.acid	6497	0	0	0.3186332	0.3100000	1.6600000	0.1453179
residual.sugar	6497	0	0.6000000	5.4432353	3.0000000	65.8000000	4.7578037
chlorides	6497	0	0.0090000	0.0560339	0.0470000	0.6110000	0.0350336
free.sulfur.dioxide	6497	0	1.0000000	30.5253194	29.0000000	289.0000000	17.7493998
total.sulfur.dioxide	6497	0	6.0000000	115.7445744	118.0000000	440.0000000	56.5218545
density	6497	0	0.9871100	0.9946966	0.9948900	1.0389800	0.0029987
pH	6497	0	2.7200000	3.2185008	3.2100000	4.0100000	0.1607872
sulphates	6497	0	0.2200000	0.5312683	0.5100000	2.0000000	0.1488059
alcohol	6497	0	8.0000000	10.4918008	10.3000000	14.9000000	1.1927117
quality	6497	0	3.0000000	5.8183777	6.0000000	9.0000000	0.8732553

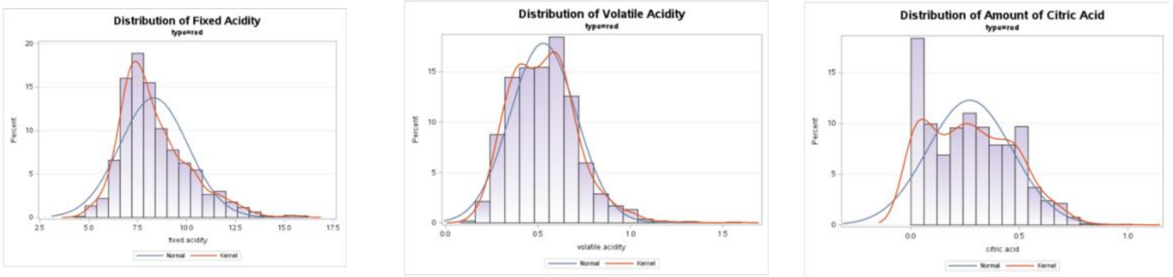
Note. Figure 6 depicts the table showing the descriptive statistics of the numerical variables in the combined wine dataset.

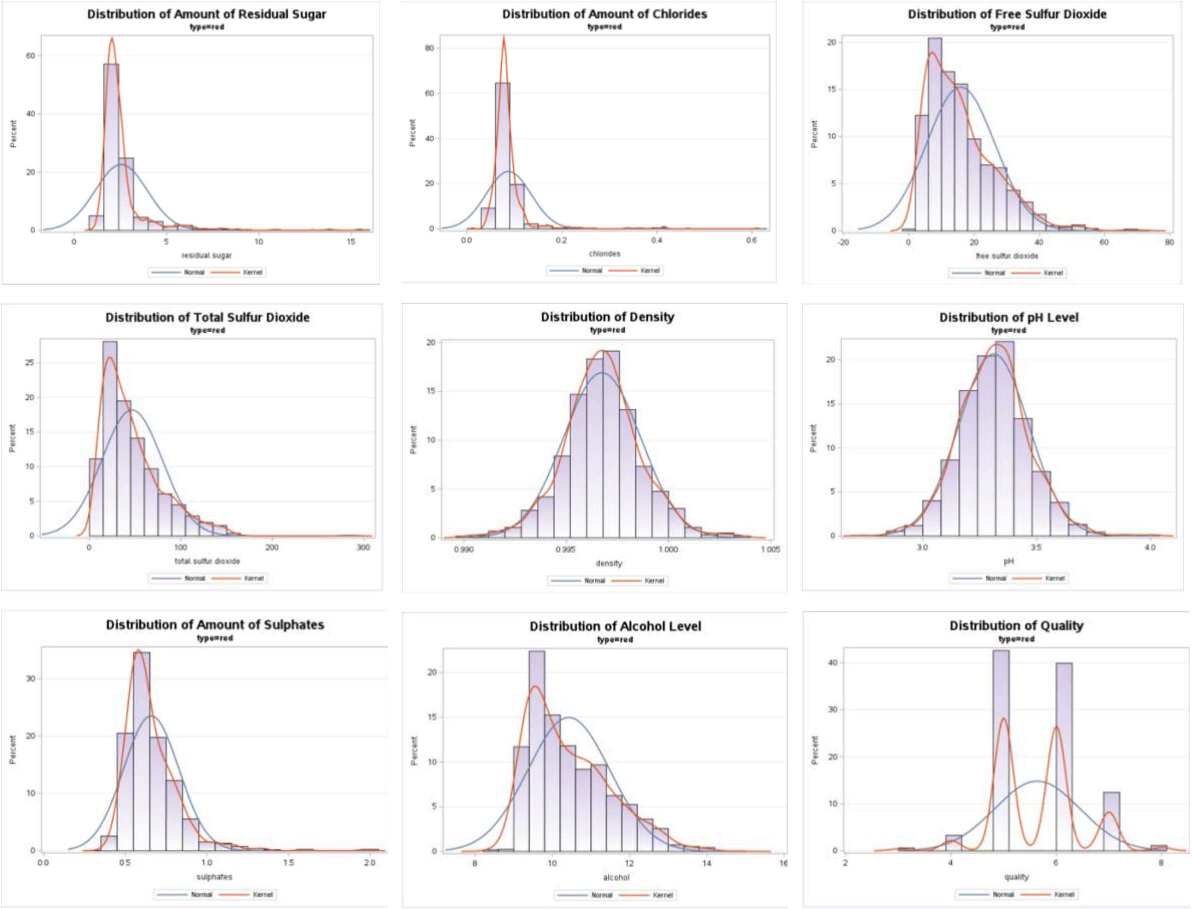
Based on the standard deviation, the variables "free.sulfur.dioxide" and "total.sulfur.dioxide" in the three datasets are the highest. This would mean that data points might be the most inconsistent in the variables. This should be something that must be taken note of later in the analysis.

The distribution of each variable are also examined. Since the dataset combines both red wine and white wine. The variables for each type of wine would be examined separately. Then, the distribution of the combined dataset would be examined. This would allow us to make decisions on whether the imbalance should be handled later in the phase. Figure 7 first shows the histograms for the distribution of variables of red wine, followed by Figure 8 showing histograms for the distribution of variables of white wine, then Figure 9 showing the histograms for the distribution of variables of both wines combined. In the histograms, two density curves are added to further view the distribution. The blue curve is the normal whereas the orange curve is the kernel.

Figure 7

Histograms for the Distribution of Variables for Red Wine Dataset





Distribution of Amount of Residual Sugar

Distribution of Amount of Chlorides

Distribution of Free Sulfur Dioxide

Distribution of Total Sulfur Dioxide

Distribution of Density

Distribution of pH Level

Distribution of Amount of Sulphates

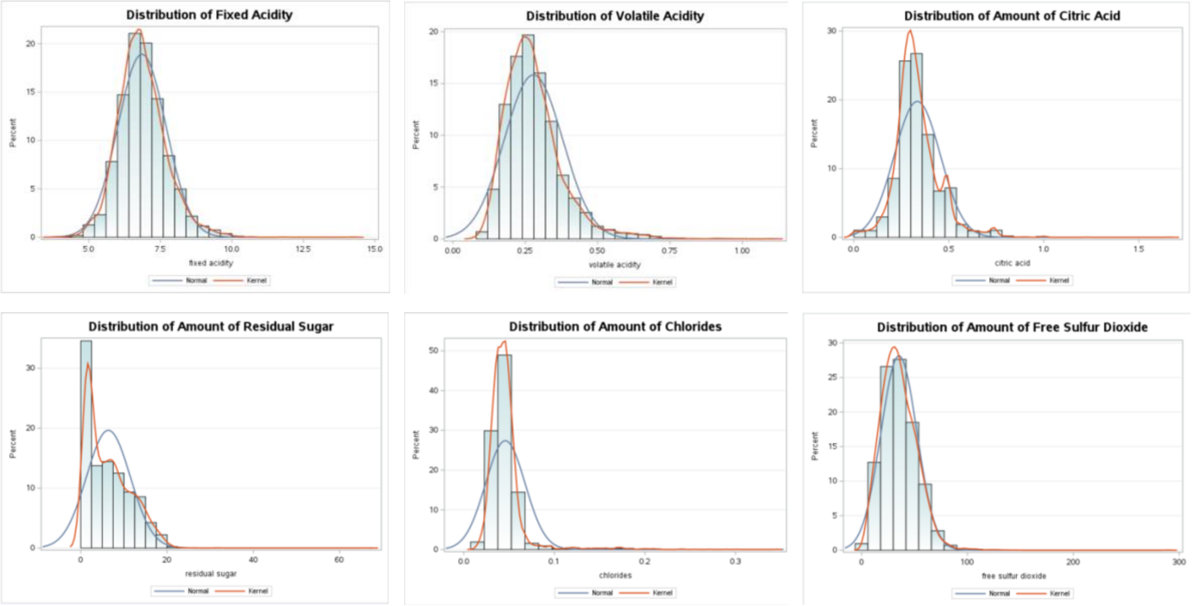
Distribution of Alcohol Level

Distribution of Quality

Note. Figure 7 shows the histograms plotted to view the distribution of all the variables for red wine.

Figure 8

Histograms for the Distribution of Variables for White Wine Dataset



Distribution of Fixed Acidity

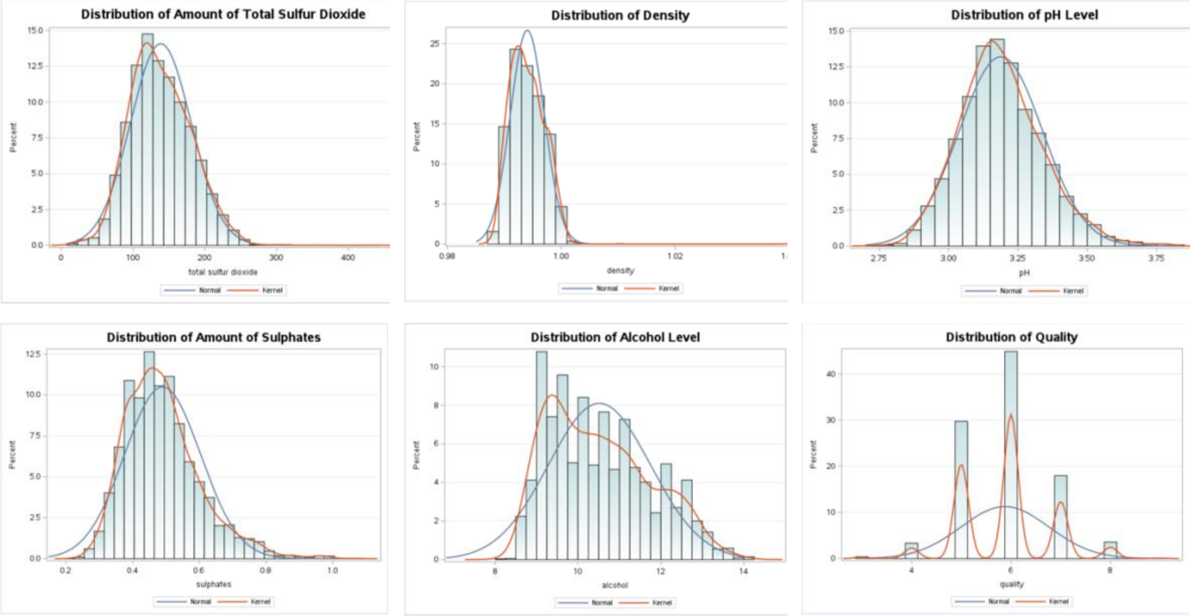
Distribution of Volatile Acidity

Distribution of Amount of Citric Acid

Distribution of Amount of Residual Sugar

Distribution of Amount of Chlorides

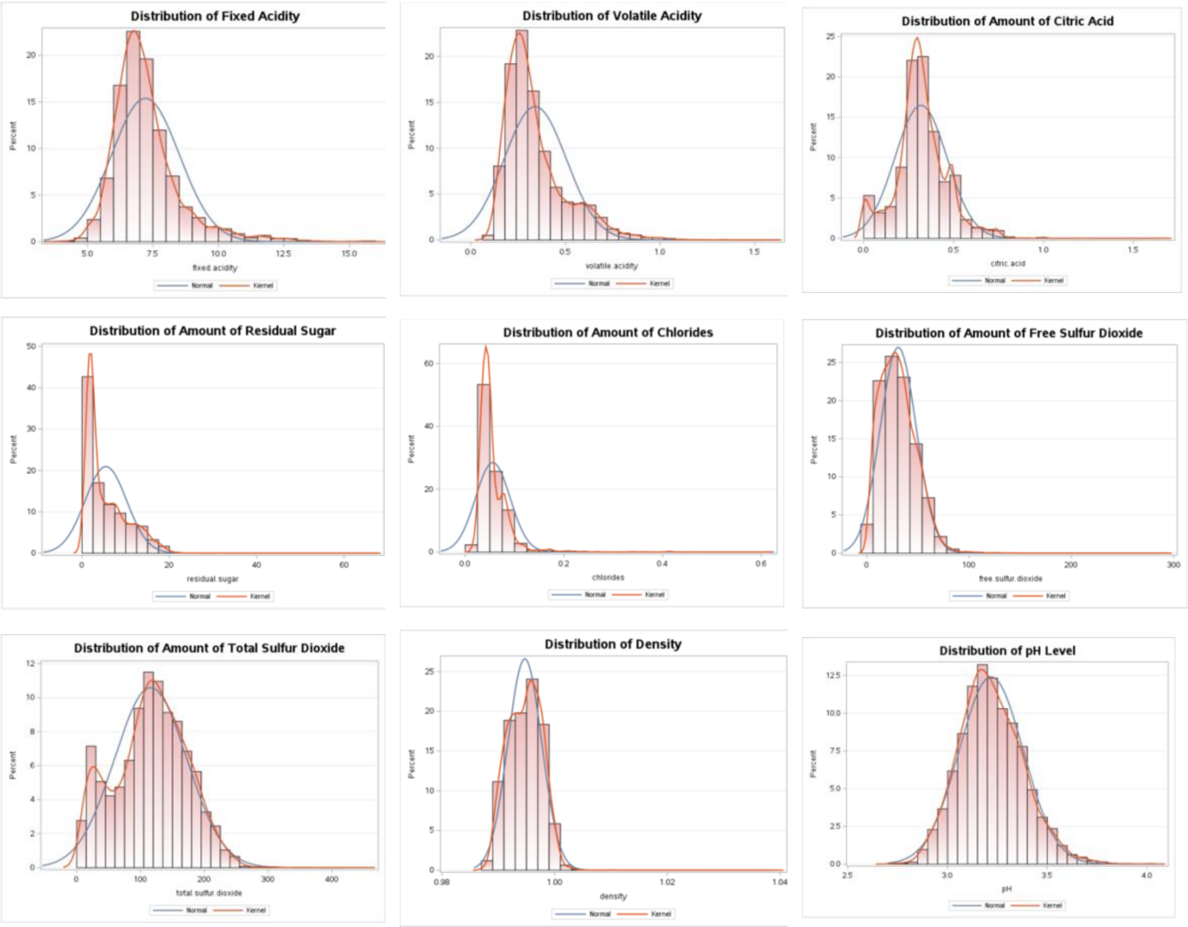
Distribution of Amount of Free Sulfur Dioxide

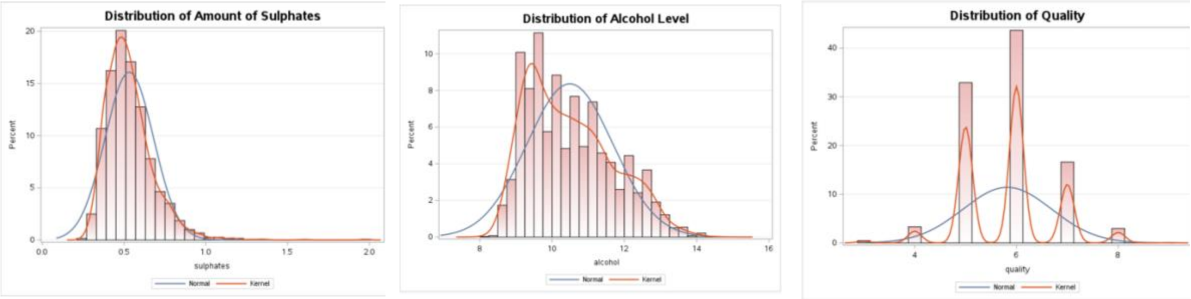


Note. Figure 8 shows the histograms plotted to view the distribution of all the variables for white wine.

Figure 9

Histograms for the Distribution of Variables for Red and White Wine Combined Dataset





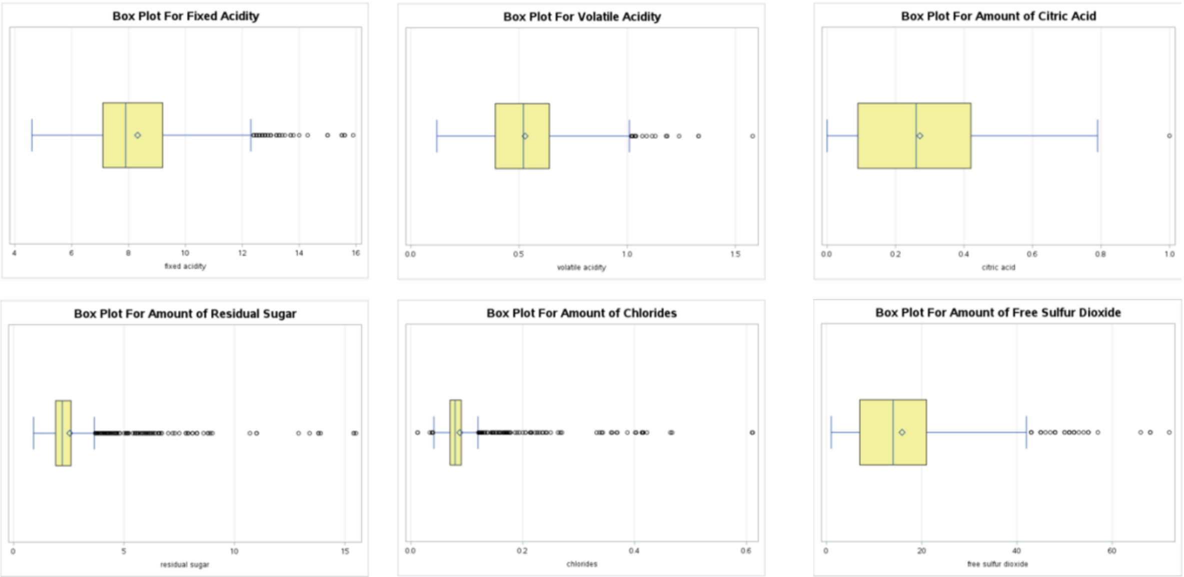
Note. Figure 9 shows the histograms plotted to view the distribution of all the variables for both red wine and white wine combined.

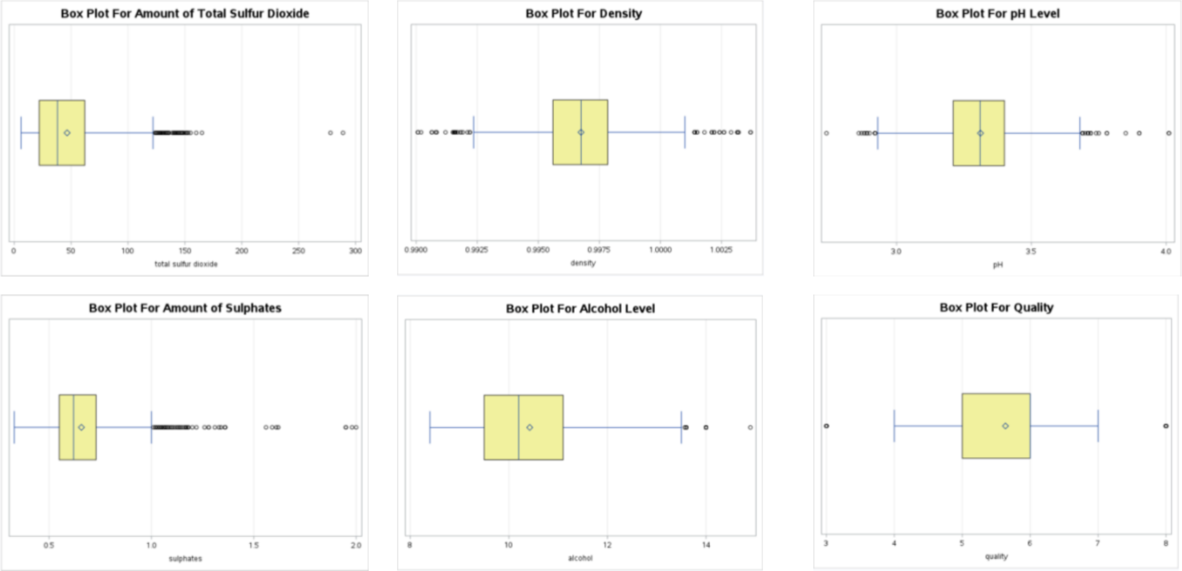
Based on the histograms, it is seen that for red wine and white wine datasets, the data points in each variable are relatively normally distributed except for the variable “residual.sugar” and “chlorides”. This is something that should be taken note of later in the analysis. There is also an imbalance in the target variable “quality”. However, this could be easily explained as quality is considered a nominal data. Therefore, the density curve would be as such.

Other than distribution, box plots are also created to clearly visualize the outliers present in the dataset. The box plots are created through three categories: box plots for red wine, box plots for white wine and box plots for both red wine and white wine. Figure 10 to Figure 12 would show all the box plots for each category respectively.

Figure 10

Box Plots for All Variables in the Red Wine Dataset

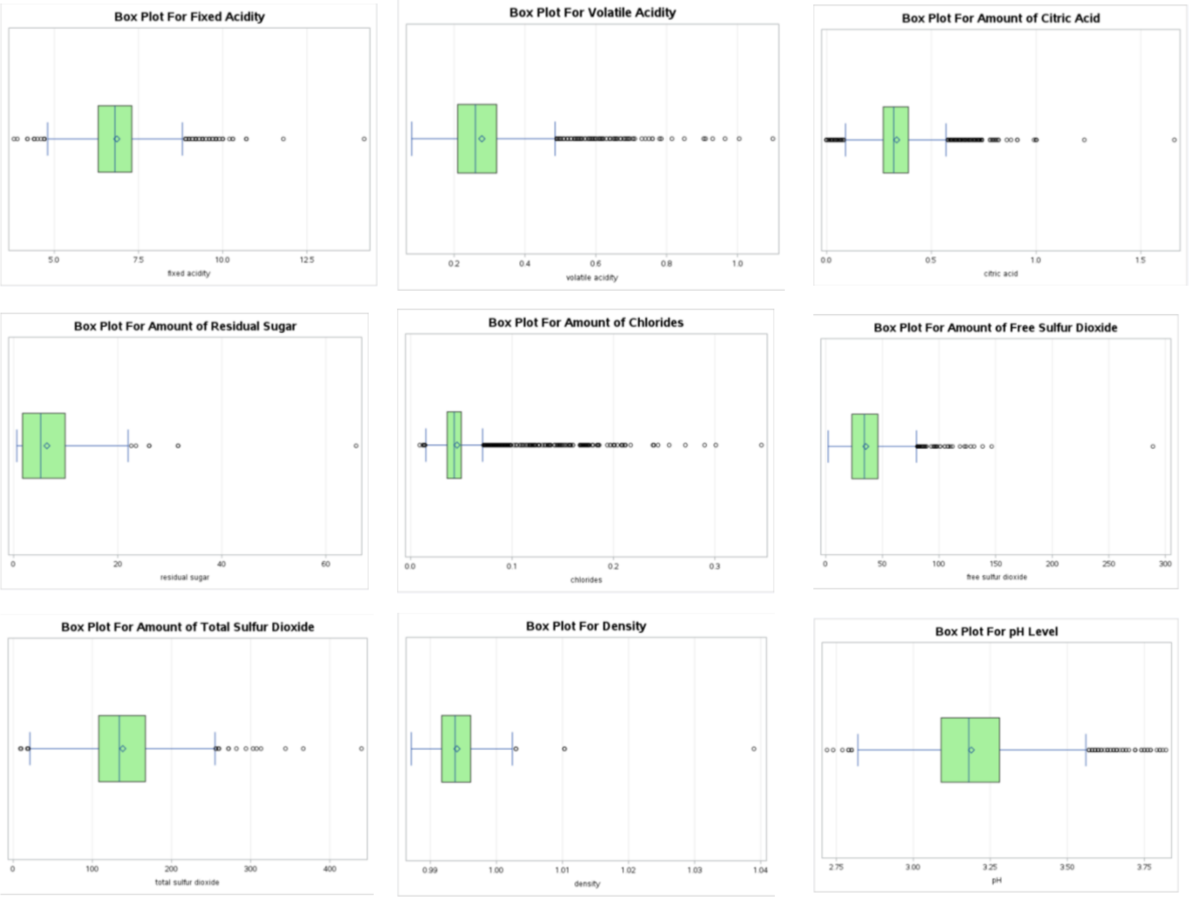


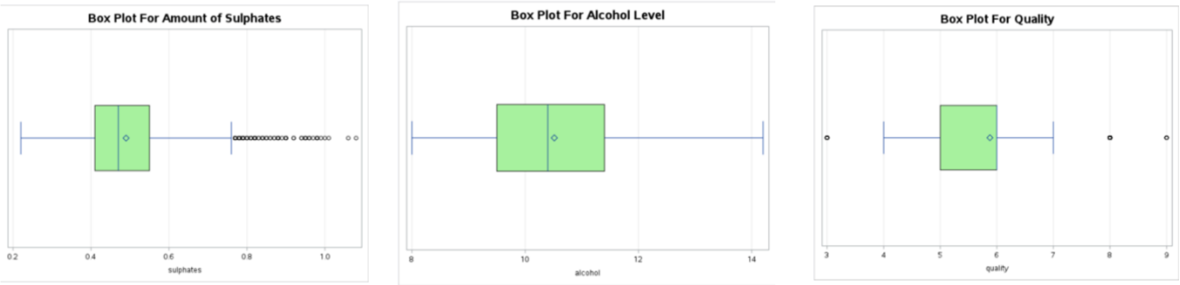


Note. Figure 10 shows the box plots for the numerical variables in the red wine dataset.

Figure 11

Box Plots for All Variables in the White Wine Dataset

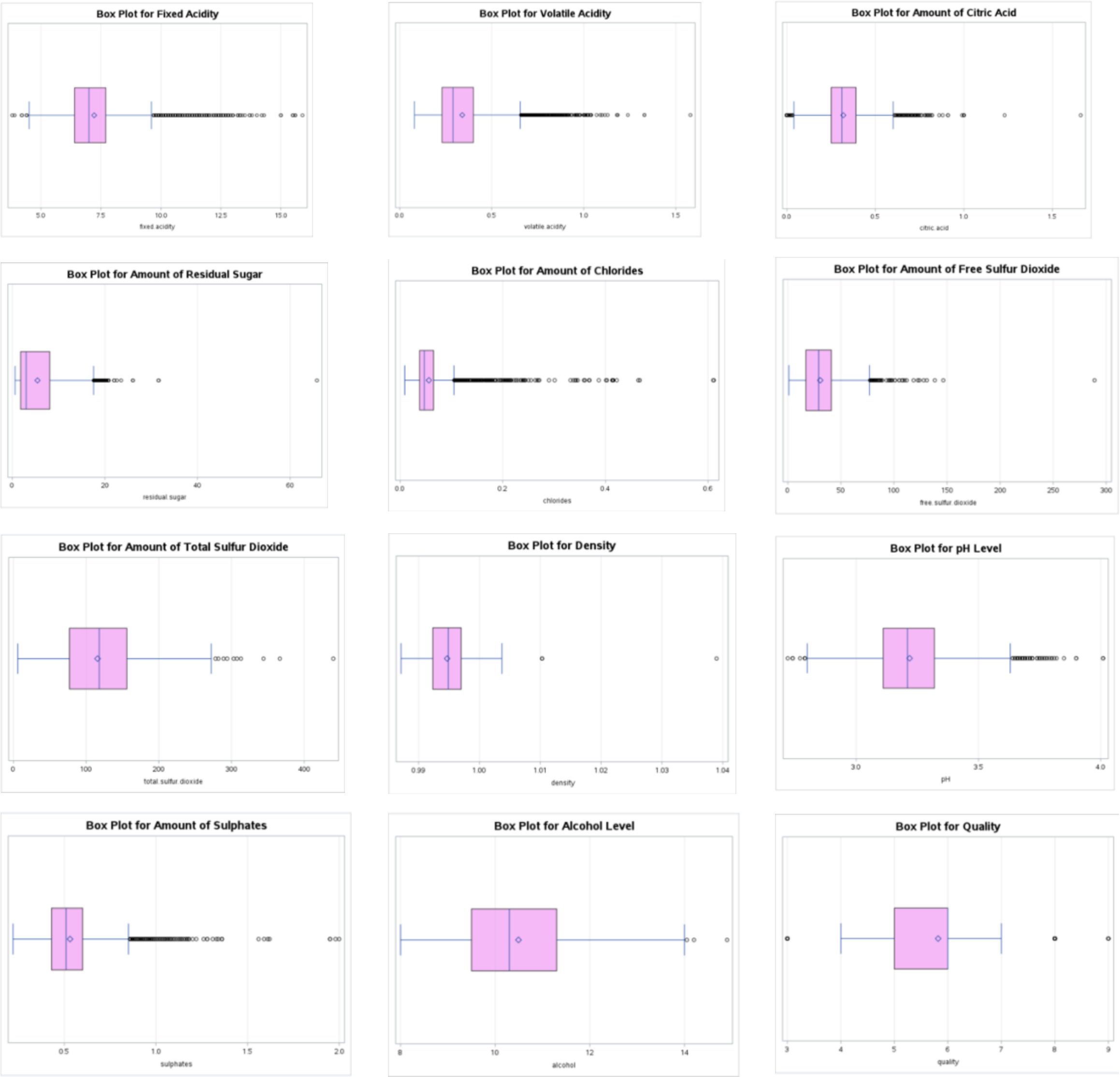




Note. Figure 11 shows the box plots for the numerical variables in the white wine dataset.

Figure 12

Box Plots for All Variables in the Red Wine and White Wine Combined Dataset



Note. Figure 12 shows the box plots for the numerical variables in the combined dataset of red wine and white wine.

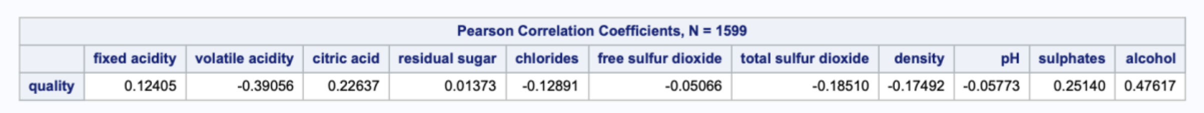
Based on the box plots, except for the variable "alcohol", most of the variables contain outliers and are somewhat skewed. Therefore, the extreme values would be considered to be removed from the dataset if these outliers contain measurement errors. Nevertheless, the outliers would not be removed yet until we are certain that they causes measurement errors in the machine learning models.

4.1.2 Relationship Between Variables

Relationship between the feature variables and the target variable is also important and should be examined. This is to uncover the patterns and trends between different variables. The correlation coefficient for between the feature variables (fixed acidity, sulphates, etc.) with the target variable (quality) is calculated. In this case, Pearson's Method will be used in this calculation. The correlation coefficient are calculated using separate datasets and also the combined dataset. Figure 13 to Figure 15 shows the final calculation of the Pearson's correlation coefficient between different numerical variables in the red wine dataset, white wine dataset and the combined dataset respectively.

Figure 13

Pearson's Correlation Coefficient Between Feature Variables and the Quality Variable in the Red Wine Dataset



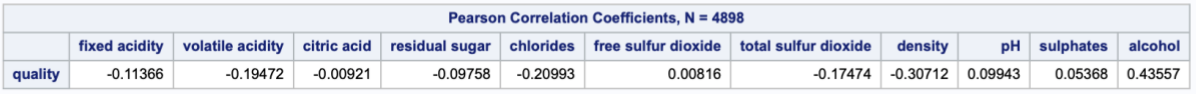
Pearson Correlation Coefficients, N = 1599

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
quality	0.12405	-0.39056	0.22637	0.01373	-0.12891	-0.05066	-0.18510	-0.17492	-0.05773	0.25140	0.47617

Note. Figure 13 shows the value of Pearson's Correlation Coefficient for each feature variable, correlating with the target variable "quality" in the red wine dataset

Figure 14

Pearson's Correlation Coefficient Between Feature Variables and the Quality Variable in the White Wine Dataset



Pearson Correlation Coefficients, N = 4898

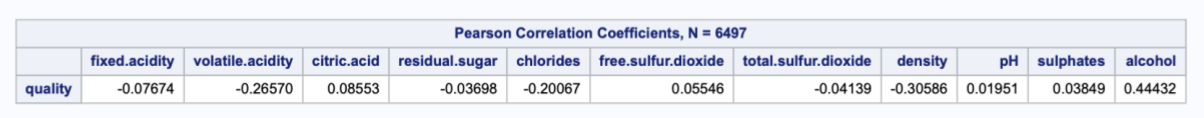
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
quality	-0.11366	-0.19472	-0.00921	-0.09758	-0.20993	0.00816	-0.17474	-0.30712	0.09943	0.05368	0.43557

Note. Figure 14 shows the value of Pearson's Correlation Coefficient for each feature variable, correlating with the target variable "quality" in the white wine dataset

As shown in Figure 13 and Figure 14, there is a slight difference between the relationships of the variables between red wine and white wine. The variables such as fixed acidity, citric acid, residual sugar, free sulfur dioxide and pH have a positive relationship in the red wine dataset whereas they have a negative relationship in the white wine dataset. Thus, it is shown here that different wine types indeed would affect the value of the variables.

Figure 15

Pearson's Correlation Coefficient Between Feature Variables and the Quality Variable in the Combined Dataset



Pearson Correlation Coefficients, N = 6497

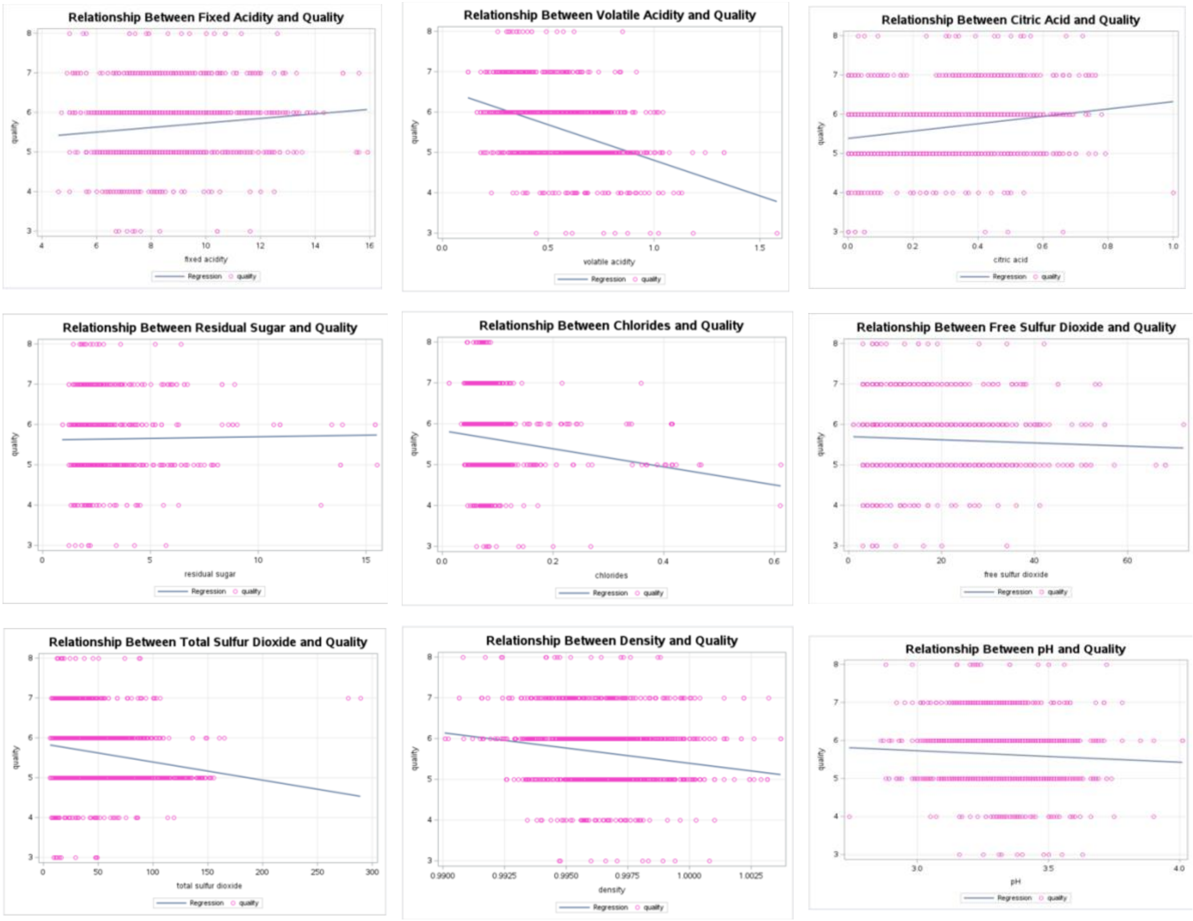
	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol
quality	-0.07674	-0.26570	0.08553	-0.03698	-0.20067	0.05546	-0.04139	-0.30586	0.01951	0.03849	0.44432

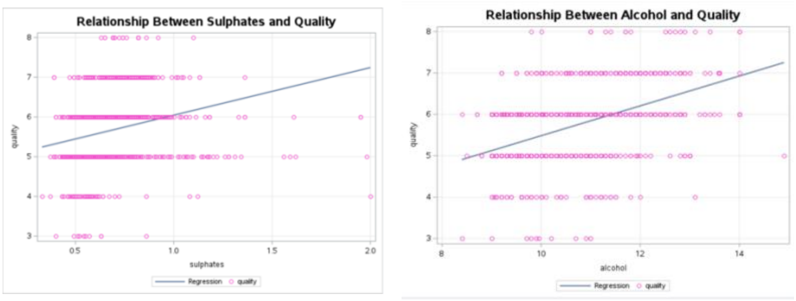
Note. Figure 15 shows the overall value of Pearson's Correlation Coefficient for each feature variable, correlating with the target variable "quality". The relationship is more similar to the relationships in the white wine dataset as there are more instances in the white wine compared to the red wine dataset.

Other than calculating the correlation coefficient, scatter plots are also created to clearly visualize the relationship of each variable with the quality of the wine. Figure 16 to Figure 18 shows the scatter plots created to visualize the relationship between each feature and the target variable in the red wine dataset, white wine dataset and also the combined dataset.

Figure 16

Scatter Plots Between Quality of Wine and Other Variables in the Red Wine Dataset





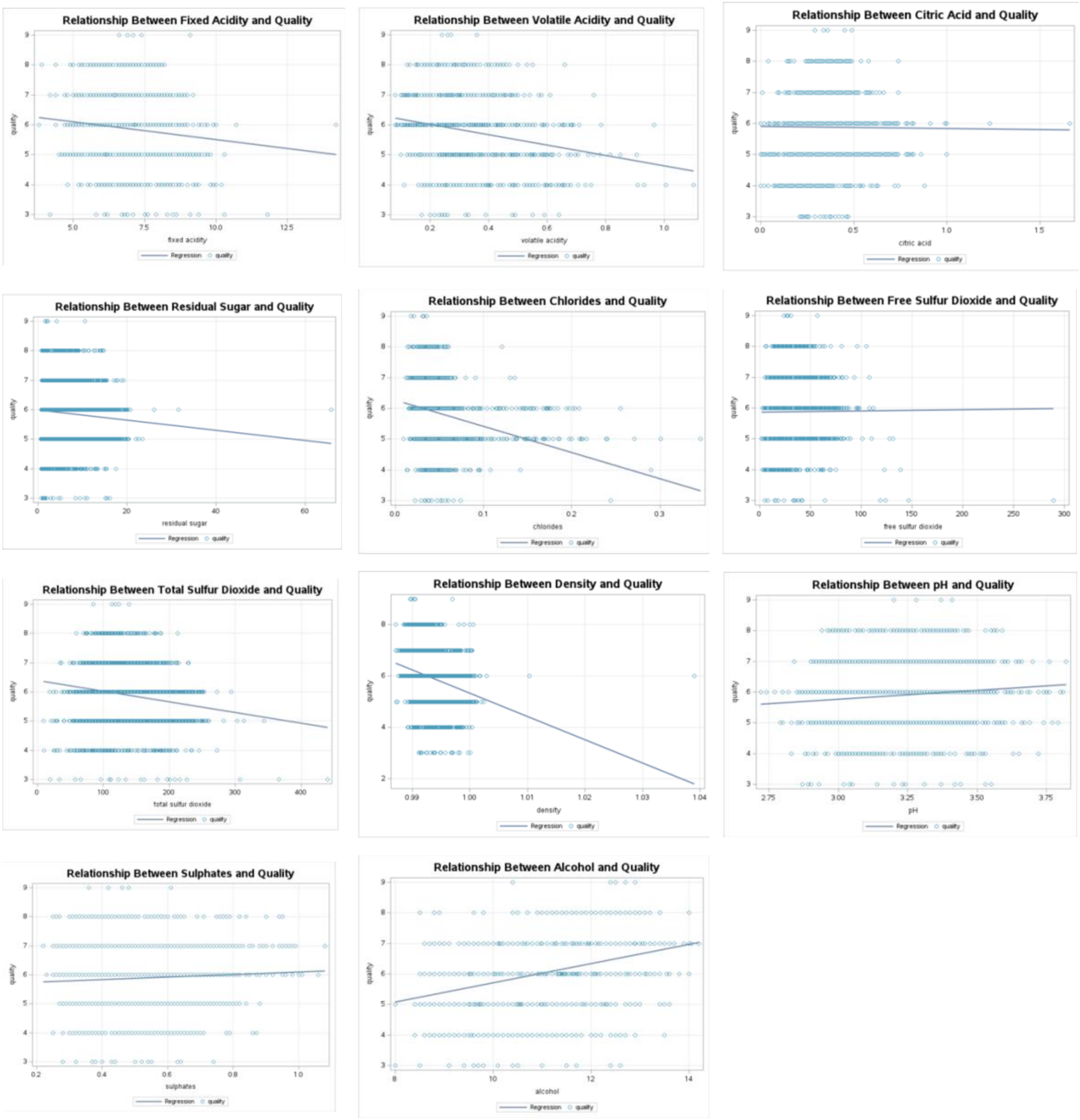
Relationship Between Sulphates and Quality

Relationship Between Alcohol and Quality

Note. Figure 16 shows the scatter plot between each variable with quality as the depending variable in the red wine dataset. Each variable has a significant relationship with the red wine quality, except for free sulfur dioxide, residual sugar and pH.

Figure 17

Scatter Plots Between Quality of Wine and Other Variables in the White Wine Dataset



Relationship Between Fixed Acidity and Quality

Relationship Between Volatile Acidity and Quality

Relationship Between Citric Acid and Quality

Relationship Between Residual Sugar and Quality

Relationship Between Chlorides and Quality

Relationship Between Free Sulfur Dioxide and Quality

Relationship Between Total Sulfur Dioxide and Quality

Relationship Between Density and Quality

Relationship Between pH and Quality

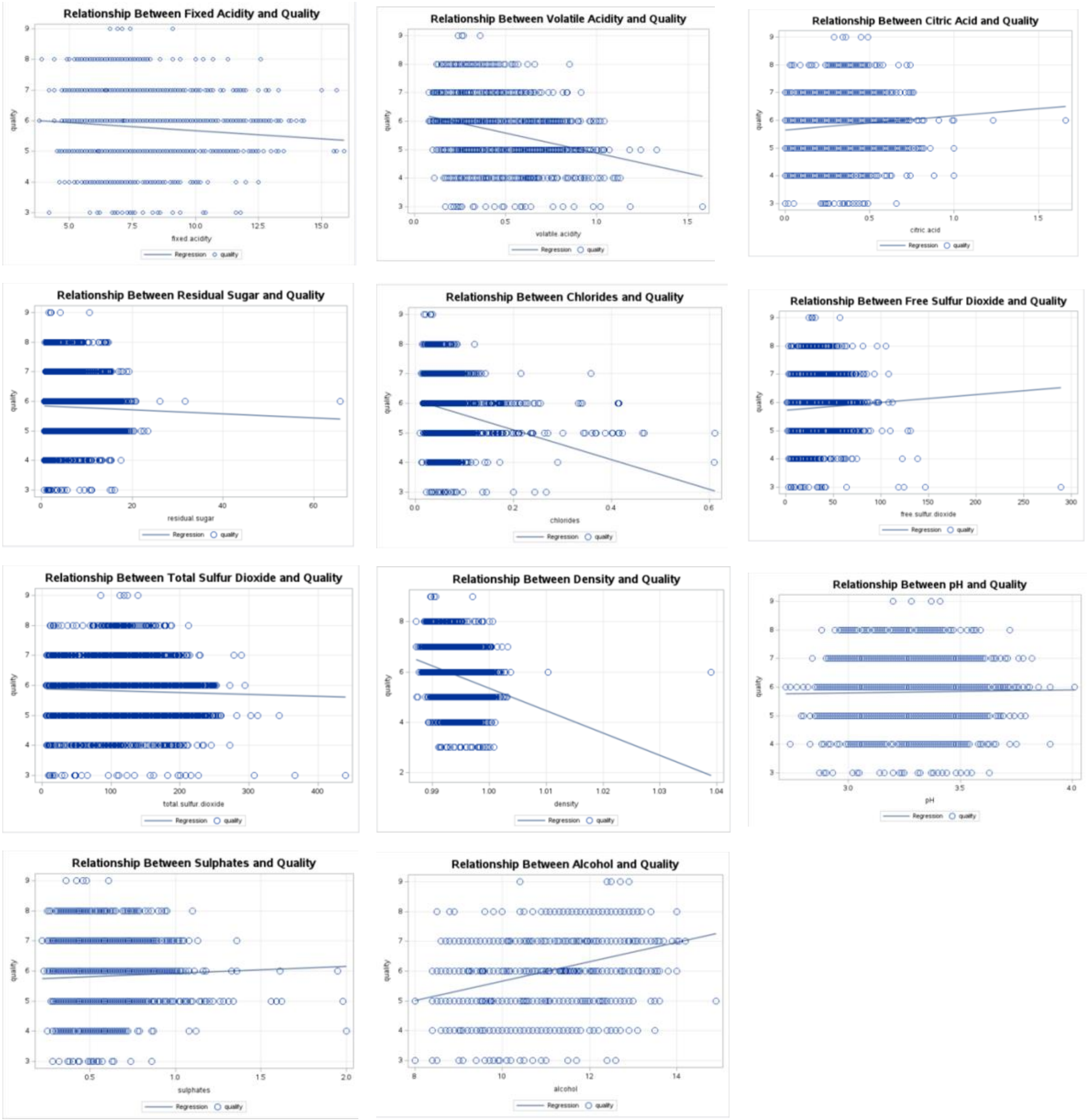
Relationship Between Sulphates and Quality

Relationship Between Alcohol and Quality

Note. Figure 17 shows the scatter plot between each variable with quality as the depending variable in the white wine dataset. Except for citric acid, free sulfur dioxide and sulphates, the other variables have shown a certain type of relationship with the quality of white wine.

Figure 18

Scatter Plots Between Quality of Wine and Other Variables in the Combined Dataset



Note. Figure 18 shows the scatter plot between each variable with quality as the depending variable in the combined dataset. From the scatter plots, all variables have a certain amount of relationship with the quality of wine. However, pH and sulphates would have the least significance between the quality of wine.

4.1.3 Data Modification

Based on the descriptive statistics of the dataset, it is seen that the variables are widely spread. For example, the values for the variable free sulfur dioxide is a lot larger than the values of other variables. When training machine learning models, they are extremely sensitive to these large-scale values. If they are not modified, these variables might dominate the performance of the machine learning models. Therefore, feature scaling is carried out. In this case standardization is used towards the dataset. The formula for standardization is as follows:

$$z = \frac{x - mean}{std}$$

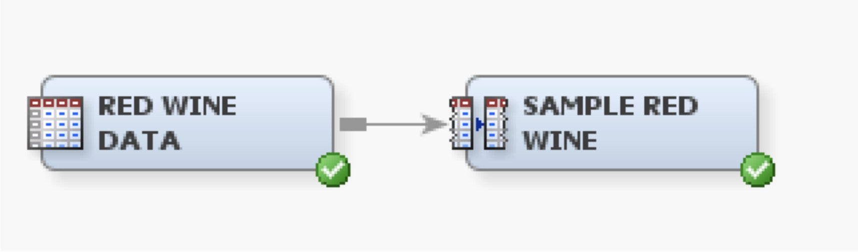
Other than standardization, the categorical data in the dataset is encoded. This is because most machine learning models work better with numerical variables. With that, the categorical variable in the dataset: “type” is encoded. Since the variable “type” only has two level, the values would be in binary form. Red wine will be encoded as 0 while white wine will be encoded as 1.

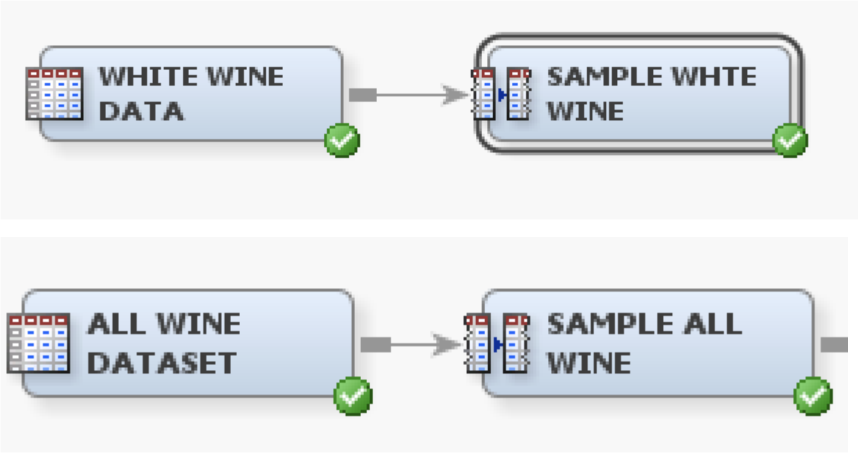
4.1.4 Sample Extraction

After that, a sample dataset is extracted. This sample extraction would be done in SAS Enterprise Miner as shown in Figure 19. After the data is successfully transformed and modified in SAS Studio, the folder is connected to SAS Enterprise Miner as a new library. With that, new data sources can be added in SAS Enterprise Miner easily. For Problem 1, red wine data and white wine data would be separated and used to train the same machine learning model so that the quality can be predicted more specifically for either red wine or white wine. For Problem 2, the combined dataset would be used carry out classification tasks. All three of the datasets would extract 20% of the data as a sample. After sampling, 320 rows of red wine data would be used; 980 rows of white wine data would be used; 1300 rows of data with combined wine types would be used. With that, it prevents overfitting and decreases the processing time when training the machine learning models.

Figure 19

Sample Extraction of Red Wine Data, White Wine Data and Combined Data in SAS Enterprise Miner





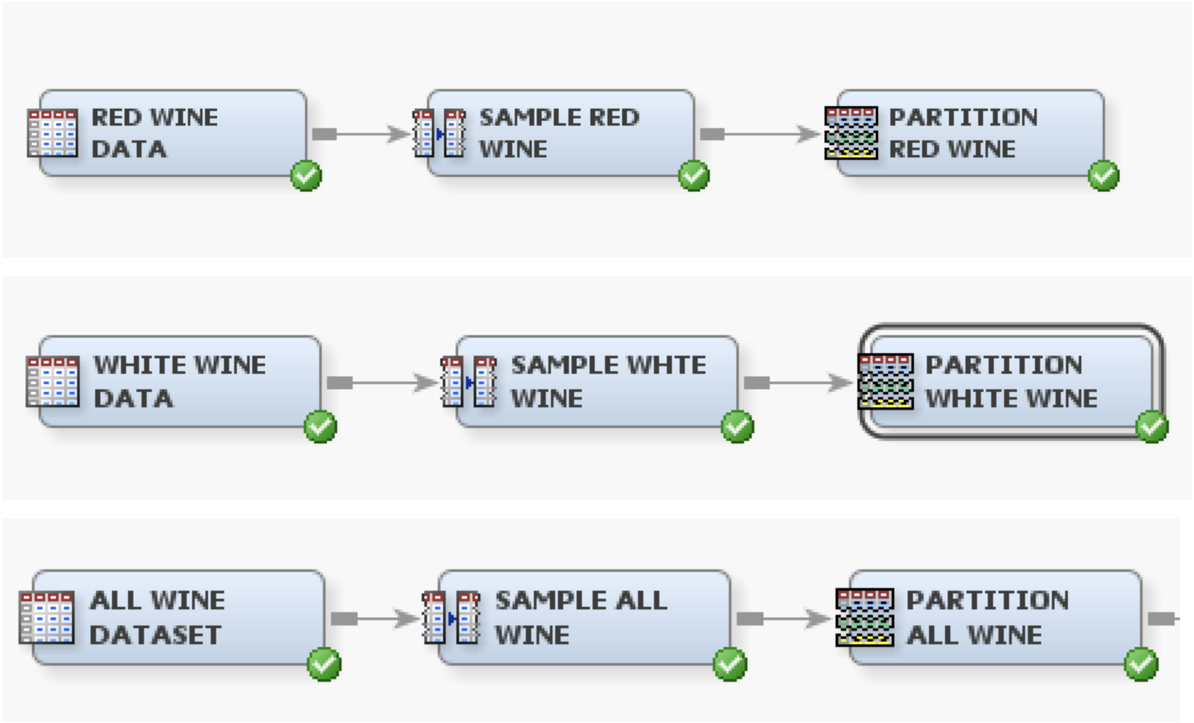
Note. Figure 19 shows the nodes connected in SAS Enterprise Miner to extract a sample set from the red wine dataset, white wine dataset and the combined dataset.

4.1.5 Sample Partitioning

After extracting a sample data set, the sample data is partitioned into training data, validation data and testing data. The training data is basically used to train the machine learning model; validation data is used to validate whether the model is accurate and then testing data is used to actually test the performance of the machine learning model trained. In this case, sample data will be partitioned into 70% of training data, 20% validation data and 10% testing data. For final model evaluation, the metrics for testing data would be used. Figure 20 shows the nodes connected in SAS Enterprise Miner to perform this step on the red wine data, white wine data and combined data.

Figure 20

Partitioning Data for Red Wine Data, White Wine Data and Combined Data in SAS Enterprise Miner



Note. Figure 20 shows the nodes connected in SAS Enterprise Miner to partition the red wine dataset, white wine dataset and combined dataset. The final percentage of partition would be 70% for training, 20% for validation and 10% for testing.

4.2 Data Mining Techniques

In this section, it covers the two final phases in the SEMMA methodology: Model and Assess. Firstly, same models with different hyperparameters are trained and compared among themselves. This is so that the model of hyperparameter that gives the best performance can be chosen. With that, then only the best model parameters can be used to compare with the other types of machine learning models. Overall, for these two phases, SAS Enterprise would be used.

4.2.1 Training Models using Different Hyperparameters

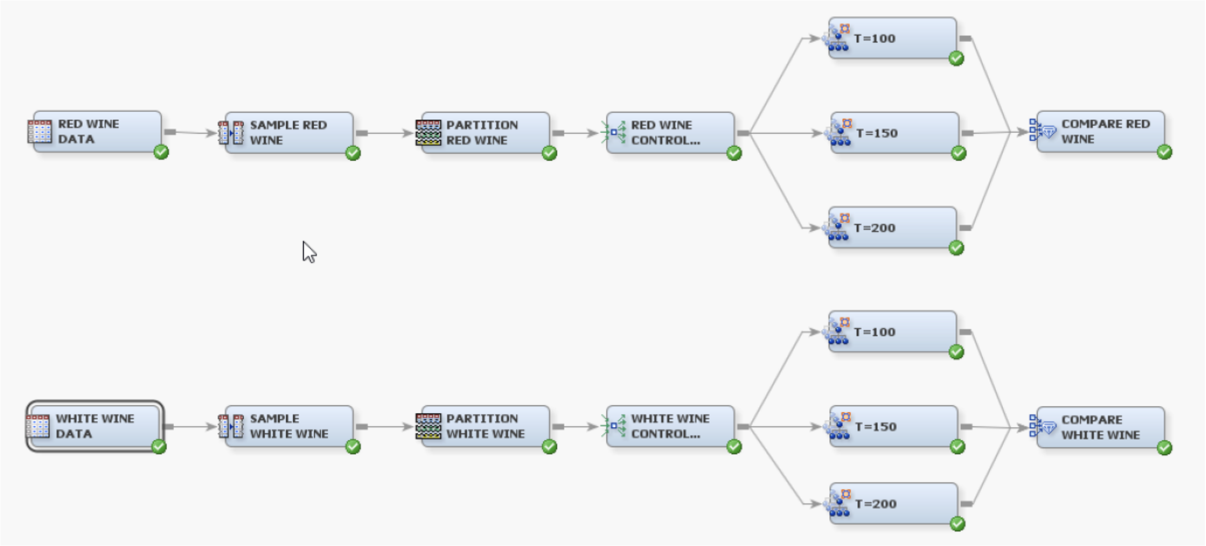
The hyperparameter of each model are tuned to determine the hyperparameter that gives trains the best machine learning model. For Random Forest model, the main hyperparameter tuned is the maximum number of trees; for Neural Network model, the hyperparameter tuned is the maximum number of iterations; for Gradient Boosting model, the hyperparameters tuned are number of weak learners (N iterations) and maximum depth. Each weak learner would aim to correct the previous mistakes made by the previous learner, improving the performance. This subsection would discuss about the hyperparameter tuning process for both Problem 1 and Problem 2.

4.2.1.1 Problem 1. For Problem 1, the quality of red wine and white wine are predicted using the machine learning models of different hyperparameters. In the end, the hyperparameter that gives the model the best result would be chosen. To achieve this, a control point node in SAS Enterprise Miner is added between the data nodes and the machine learning model nodes. Then, in the end, a model comparison node is added to directly see which model's hyperparameter should be chosen.

For Random Forest model, Figure 21 first shows the nodes connected in SAS Enterprise Miner to compare the performance of Random Forest models of different maximum number of trees.

Figure 21

Training Different Random Forest Models Using Different Hyperparameters in SAS Enterprise Miner for Problem 1

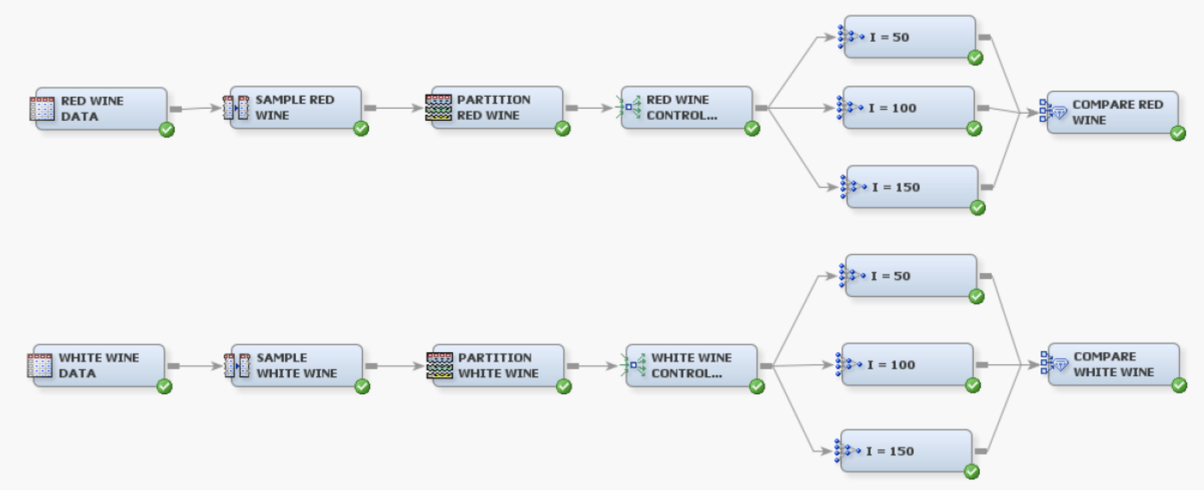


Note. Figure 21 shows the nodes connected to carry out internal Random Forest model training with different hyperparameters using red wine data and white wine data. Each Random Forest node is labelled with their respective maximum number of trees (T). The result of the model comparison node for red wine data and white wine data is shown in Table 9 and Table 10 respectively.

The parameters for Neural Network are also tuned to make sure that the best Neural Network model is chosen to compare with other types of machine learning models. Figure 22 shows the nodes connected in SAS Enterprise Miner to test the performance of the Neural Network models with different parameters.

Figure 22

Training Different Neural Network Models Using Different Hyperparameters in SAS Enterprise Miner for Problem 1

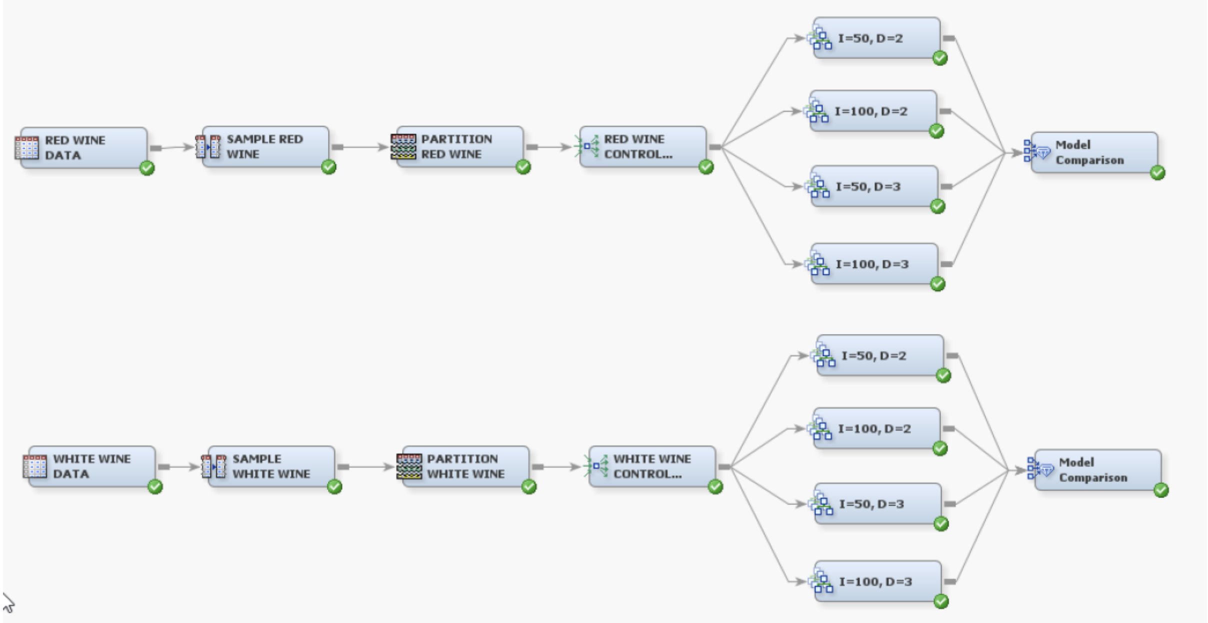


Note. Figure 22 shows the nodes connected to carry out internal Neural Network model training with different hyperparameters for red wine data and white wine data. Each Neural Network node is labelled with their respective hyperparameter. “I” stands for the maximum number of iterations. The result of the model comparison node for red wine data and white wine data is shown in Table 11 and Table 12 respectively.

Lastly, the hyperparameters for Gradient Boosting model are also modified to achieve optimum performance. Figure 23 shows the nodes connected in SAS Enterprise Miner to carry out this process.

Figure 23

Training Different Gradient Boosting Models Using Different Hyperparameters in SAS Enterprise Miner for Problem 1

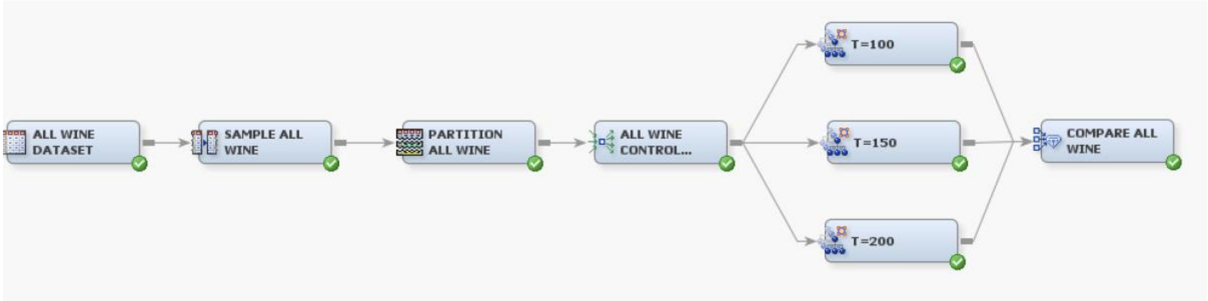


Note. Figure 23 shows the nodes connected to carry out internal Gradient Boosting model training with different hyperparameters for red wine data and white wine data. Each Gradient Boosting node is labelled with their respective hyperparameter. "I" stands for the number of weak learners while "D" stands for maximum depth. The result of the model comparison node for red wine data and white wine data is shown in Table 13 and Table 14 respectively.

4.2.1.2 Problem 2. For Problem 2, the same hyperparameters are tuned to see which model can classify the type of wine the best. Figure 24 shows the different Random Forest models of different number of iterations (T) implemented in SAS Enterprise Miner.

Figure 24

Training Different Random Forest Models Using Different Hyperparameters in SAS Enterprise Miner for Problem 2

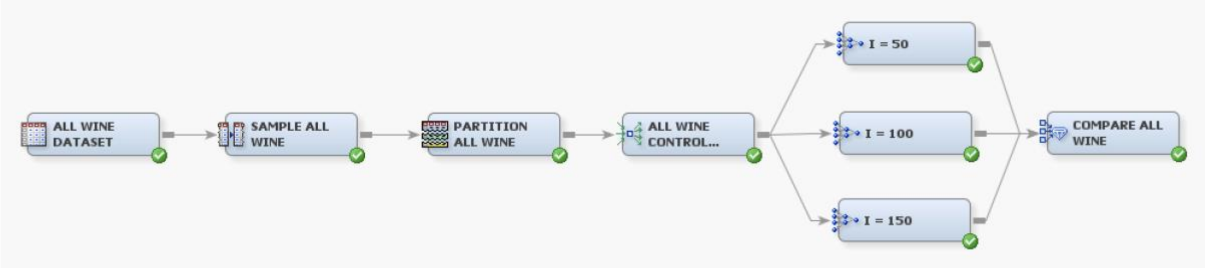


Note. Figure 24 shows the nodes connected to carry out internal Random Forest model training with different hyperparameters the combined data. Each Random Forest node is labelled with their respective hyperparameter. "T" stands for the maximum number of trees. The result of the model comparison that shows the performance of each model is shown in Table 15.

As for Neural Network model, the maximum number of iterations are tuned to see which gives the best performance. Figure 25 shows the relevant nodes in SAS Enterprise Miner to compare different Neural Network models of different maximum number of iterations.

Figure 25

Training Different Neural Network Models Using Different Hyperparameters in SAS Enterprise Miner for Problem 2

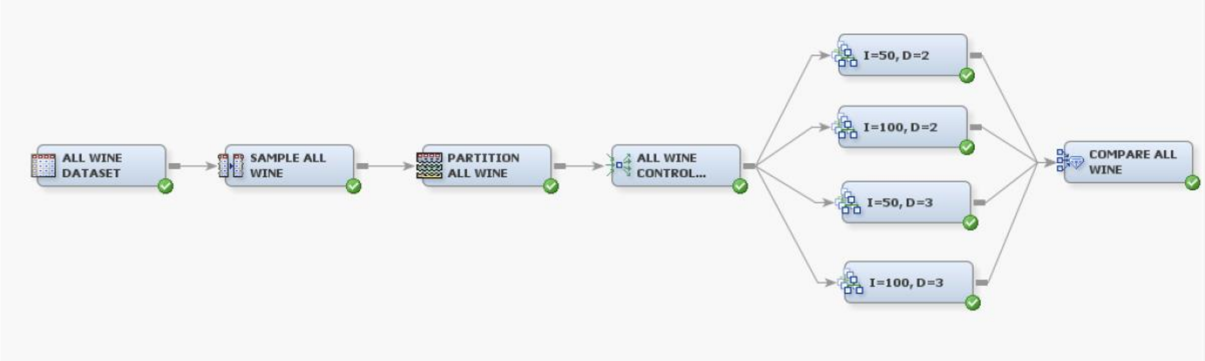


Note. Figure 25 shows the nodes connected to carry out internal Neural Network model training with different hyperparameters using the combined dataset. Each Neural Network node is labelled with their respective hyperparameter. "I" stands for the maximum number of iterations. The performance of each model is shown in Table 16.

For Gradient Boosting models, hyperparameters such as the number of weak learners and the maximum depth are tuned. Figure 26 shows the relevant nodes in SAS Enterprise Miner to compare between the performance of the Gradient Boosting models with different hyperparameters.

Figure 26

Training Different Gradient Boosting Models Using Different Hyperparameters in SAS Enterprise Miner for Problem 2



Note. Figure 26 shows the nodes connected to carry out internal Gradient Boosting model training with different hyperparameters using the combined dataset. Each Gradient Boosting node is labelled with their respective hyperparameter. "I" stands for the number of weak learners while "D" stands for maximum depth. The result for the performance of each model is shown in Table 17.

4.2.2 Result for Models of Different Hyperparameter

After running the nodes in SAS Enterprise Miner, the performance of each model are analyzed using certain metric based on the problem. Therefore, this subsection seeks to uncover all the results of each model comparison node.

4.2.1.1 Problem 1. For Problem 1, the target variable (quality) is a nominal data. With that, classification tasks to predict nominal data is carried out. In that case, the metrics used to compare the models would be misclassification rate, Receiver Operating Characteristic (ROC) index and Root Average Squared Error (RASE). Misclassification Rate is suitable as it directly tells us the rate of wrong predictions in the model. It highlights the models' overall accuracy when predicting the quality of red wine and white wine. Other than that, ROC index measures the models' ability to distinguish between classes. The higher the value of ROC index, the more capable the model is to separate between each quality class. Lastly, the RASE value can be used to see the squared difference between the predicted value and the actual value. In short, RASE value lets us see how far off the prediction might be from the actual quality score. Thus, these three metrics would be mainly used to measure models' performance for Problem 1.

For starters, Table 9 and Table 10 shows the results of the Random Forest models for red wine data and white wine data respectively. The maximum number of trees are set to 100, 150 and 200.

Table 9

Result of Random Forest Models for Red Wine Data

Maximum Number of Trees (T)	ROC Index	Misclassification Rate	RASE
100	0.757	0.389	0.300
150	0.714	0.389	0.301
200	0.657	0.361	0.302

Note. Table 9 shows the result of the performance of each Random Forest model using different maximum number of trees on red wine data.

Table 10

Result of Random Forest Models for White Wine Data

Maximum Number of Trees (T)	ROC Index	Misclassification Rate	RASE
100	0.885	0.429	0.283
150	0.923	0.419	0.281
200	0.928	0.429	0.281

Note. Table 10 shows the result of the performance of each Random Forest model using different maximum number of trees on white wine data.

From Table 9, it is shown that the Random Forest model with a maximum of 100 trees has the highest ROC index and lowest RASE value when predicting the quality of red wine. This indicates that with a maximum of 100 trees, the model is the most efficient in separating each quality class and its prediction is the least furthest from the actual predictions. However, the misclassification rate slightly higher. Nonetheless, it still makes the model the best when predicting the quality of red wine.

As for Table 10, the Random Forest model with a maximum of 200 trees has the highest ROC index and lowest RASE value. Its misclassification rate, though the highest, is still not that far from the model with the lowest misclassification rate. Naturally, it means that this model performs better when predicting the quality of white wine. With that being said, the difference of the result might be because of the difference size between red wine dataset and white wine dataset. Thus, the hyperparameters of the models when predicting the quality of red wine and white wine would naturally be different. As a result, when predicting quality of red wine using Random Forest model, the parameter of maximum 100 trees would be used; when predicting the quality of white wine, the model with maximum 200 trees would be used to compare between different types models.

Secondly, Table 11 and Table 12 would show the results of the Neural Network models for red wine data and white wine data respectively. For this model, the maximum number of iterations are tuned to 50, 100 and 150.

Table 11

Result of Neural Network Models for Red Wine Data

Maximum Number of Iterations (T)	ROC Index	Misclassification Rate	RASE
50	0.103	0.472	0.320
100	0.103	0.472	0.320
150	0.103	0.472	0.320

Note. Table 11 shows the result of the performance of each Neural Network model using different maximum number of iterations on red wine data.

Table 12

Result of Neural Network Models for White Wine Data

Maximum Number of Iterations (I)	ROC Index	Misclassification Rate	RASE
50	0.981	0.448	0.286
100	0.981	0.448	0. 286
150	0.981	0.448	0. 286

Note. Table 12 shows the result of the performance of each Neural Network model using different maximum number of iterations on white wine data.

Based on Table 11, the ROC index, misclassification rate and RASE value are all the same throughout the models even though their maximum number of iterations are different. This would mean that the hyperparameters do not change the model's performance much when predicting the quality of red wine. This also can be said to Table 12 in which it shows the same ROC index, misclassification rate and RASE value for the model when predicting the quality of white wine. In that case, a simpler model can be chosen. This can speed up the processing speed and maintain the performance at the same time. With that, the Neural Network model with maximum 50 iterations would be chosen to compare between other types of machine learning models.

Lastly, Table 13 and Table 14 shows the results of the performance of the Gradient Boosting models when predicting the quality of red wine and white wine respectively. The number of weak learners is tuned to 50 or 100 while the maximum depth is tuned to 2 or 3.

Table 13

Result of Gradient Boosting Models for Red Wine Data

Number of Weak Learners	Maximum Depth (D)	ROC Index	Misclassification Rate	RASE
50	2	0.457	0.500	0.323
50	3	0.200	0.528	0.329
100	2	0.457	0.500	0.323
100	3	0.200	0.528	0.329

Note. Table 13 shows the result of the performance of each Gradient Boosting model using different number of weak learners and maximum depths on red wine data.

Table 14

Result of Gradient Boosting Models for White Wine Data

Number of Weak Learners	Maximum Depth (D)	ROC Index	Misclassification Rate	RASE
50	2	0.100	0.486	0.316
50	3	0.087	0.457	0.294
100	2	0.100	0.486	0.316
100	3	0.087	0.457	0.294

Note. Table 14 shows the result of the performance of each Gradient Boosting model using different number of weak learners and maximum depths on white wine data.

From Table 13, the Gradient Boosting model with 50 weak learner and a maximum depth of 2 has the highest ROC index and lowest RASE value. This indicates that overall, this model performs better than the others even though its misclassification rate is slightly lower. It is able to predict the quality of red wine most accurately. From Table 14, the model with the same hyperparameter also has the highest ROC index and lowest RASE value, indicating the it also has the highest performance when predicting quality of white wine. Even if the misclassification rate is slightly higher, the model is still better as an overall. Therefore, the Gradient Boosting model with 50 weak learner and a maximum depth of 2 would be used to predict quality of red wine and white wine and to compare with other machine learning models.

4.2.1.1 Problem 2. For Problem 2, since it is classification of binary target variable (type), metrics such as accuracy, precision, recall and F1-score would be used to evaluate the model. Accuracy would give us the percentage of the correctly predicted results, no matter it is a positive or negative. Precision would focus on the quality of positive predictions. It tells us the effectiveness of the model in classifying the correct positives. In this case, positives would be the 1 in the dataset, which is white wine. Recall, also known as sensitivity, would also focus on the models' ability to identify all the true positives. It shows how many positive instances can the model detect. Lastly, F1-score would calculate the harmonic mean of precision and recall. It balances the two metrics and prevents bias even when there is an imbalanced class distribution. The formulas for these metrics would be shown below:

$$Accuracy = \frac{True\ Postive + True\ Negative}{Total\ Number\ of\ Predictions}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 - score = 2\ .\ \frac{Precision\ .Recall}{Precision + Recall}$$

For Random Forest model, as shown in Figure 24, the hyperparameter tuned would be the maximum number of trees. Table 15 shows the performance of the Random Forest models with maximum number of 100 trees, 150 trees, and 200 trees.

Table 15

Result of Random Forest Models for Combined Dataset

Maximum Number of Trees (T)	Accuracy	Precision	Recall	F1-score
100	0.9844	0.9749	0.9898	0.9823
150	0.9844	0.9749	0.9898	0.9823
200	0.9844	0.9749	0.9898	0.9823

Note. Table 15 shows the result of the performance of each Random Forest model trained using different maximum number of trees on the combined dataset.

From Table 15, it is shown that the performance for all hyperparameters are the same. In fact, the accuracy, precision, recall and F1-scores are considered very high in each model. Thus, it would be safe to choose either model to represent the best Random Forest model. In that case, to reduce computational power, the Random Forest model with a maximum of 100 trees can be used to compare between different types of models later in the analysis.

As for Neural Network models, it is mentioned that the maximum number of iterations would be the main hyperparameter tuned. Table 16 shows the performance of each model with a maximum of 50 iterations, 100 iterations and 150 iterations.

Table 16

Result of Neural Network Models for Combined Dataset

Maximum Number of Iterations (I)	Accuracy	Precision	Recall	F1-score
50	0.9807	1.0000	0.9745	0.9870
100	0. 9807	1.0000	0.9745	0.9870
150	0. 9807	1.0000	0.9745	0.9870

Note. Table 16 shows the result of the performance of each Neural Network model trained using different parameters on the combined dataset.

Based on Table 16, accuracy, precision, recall and F1-score of the models are all the same despite their different hyperparameter. In that case, a simpler Neural Network model can be chosen and it would still give the same result. Therefore, the Neural Network model with a maximum of 50 iterations will be used to represent the overall Neural Network model. Even though it has lesser iterations, it still produces a relatively high-performance score.

Lastly, for Gradient Boosting model, the number of weak learners and maximum depth are tuned. Table 17 shows the overall result for each version of the model.

Table 17

Result of Gradient Boosting Models for Combined Dataset

Number of Iterations (I)	Maximum Depth (D)	Accuracy	Precision	Recall	F1-score
50	2	0.9731	0.9748	0.9898	0.9822
50	3	0.9769	0.9798	0.9898	0.9848
100	2	0.9731	0. 9748	0.9898	0.9822
100	3	0.9846	0.9898	0.9898	0.9898

Note. Table 17 shows the result of the performance of each Gradient Boosting model trained using different parameters on the combined dataset.

In Table 17, the Gradient Boosting model with 100 iterations and a maximum depth of 3 produces the highest accuracy, precision, recall and F1-score. This indicates that it has the highest performance among all the Gradient Boosting model. With that, this model would be chosen to represent the overall Gradient Boosting Model.

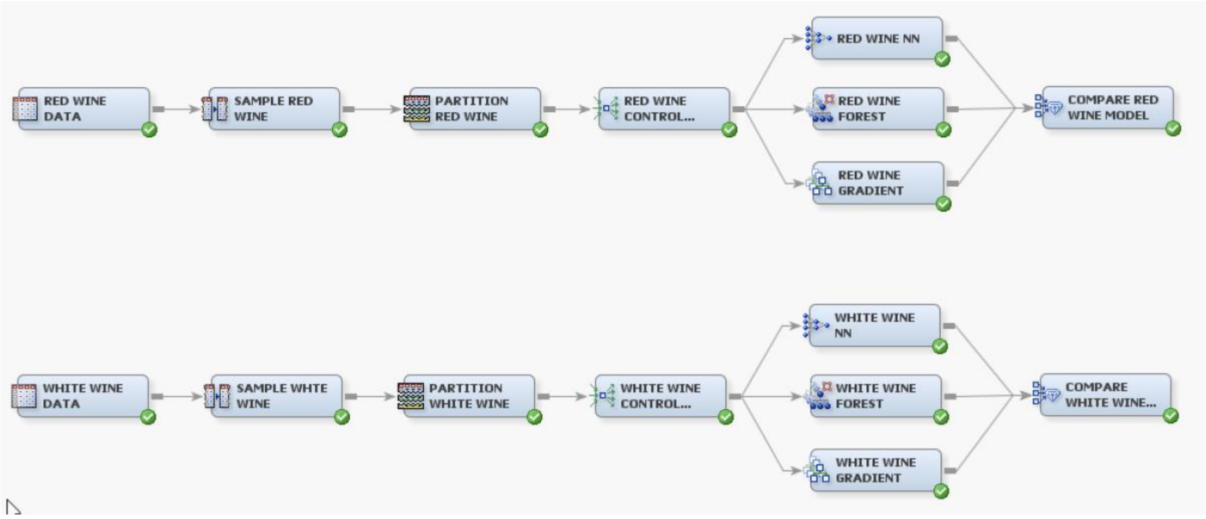
4.2.3 Compare Between Data Mining Models

After determining the best of each machine learning model, they are now used to compare between different types of the models. A new diagram in SAS Enterprise Miner is created to carry out this task. A control point node is added between the data partition node and the three data mining model nodes. In the end, a model comparison node is added to select the model with the best performance. In this subsection, model comparison will be carried out based on each problem.

4.2.2.1 Problem 1. For Problem 1, prediction of quality of red wine and white wine would be carried out separately. The models and their own hyperparameters used would be based on the previous section. Figure 27 shows the nodes connected in SAS Enterprise Miner to compare the performance between Random Forest model, Neural Network model and Gradient Boosting model.

Figure 27

Comparing Machine Learning Models for Problem 1

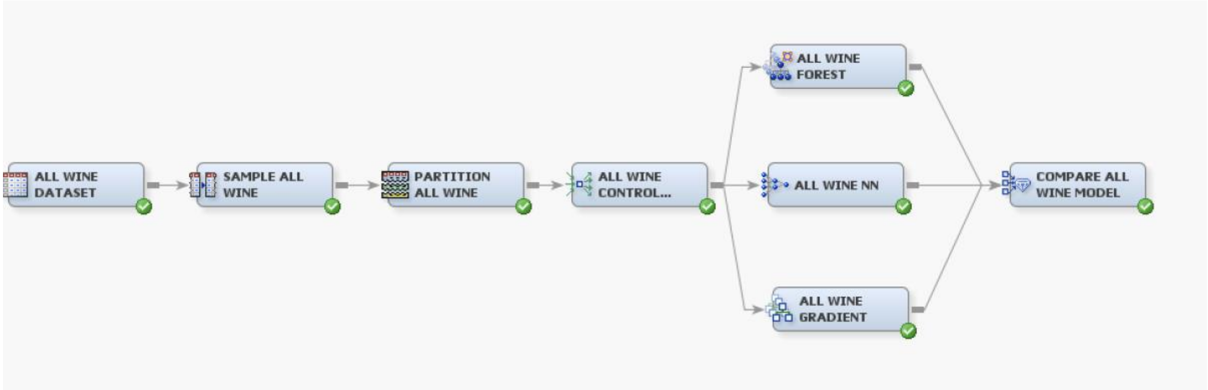


Note. Figure 27 depicts the nodes connected in SAS Enterprise Miner to compare the three machine learning models selected for Problem 1. Each node is labelled with the respective model's name and their target data. The results for the model comparison node for red wine data and white wine data are shown in Table 18 and Table 19 respectively.

4.2.2.1 Problem 2. For Problem 2, the combined dataset with both red wine data and white wine data would be used to train the machine learning models. The hyperparameter used for each model would be based on the section earlier. In the end, the three machine learning models would be compared to see their performance on the classification task. Figure 28 shows the nodes connected in SAS Enterprise Miner to achieve this task.

Figure 28

Comparing Machine Learning Models for Problem 2



Note. Figure 28 depicts the nodes connected in SAS Enterprise Miner to compare the three machine learning models selected for Problem 2. Table 20 shows the result of each machine learning model performance on the classification of red wine and white wine.

4.2.4 Result of Different Data Mining Models

4.2.4.1 Problem 1. Table 18 and Table 19 shows the performance of each type of data mining models when predicting quality of red wine and white wine respectively. An additional processing speed, in seconds, is added to see which machine learning model takes the shortest time to carry out the prediction tasks.

Table 18

Results of Model Comparison Node for Red Wine Data

Model	ROC Index	Misclassification Rate	RASE	Speed (s)
Random Forest	0.757	0.389	0.300	6.320
Neural Network	0.800	0.389	0. 323	3.120
Gradient Boosting	0.457	0.500	0. 323	4.950

Note. Table 18 shows the result of the performance of each type of machine learning model used to predict quality of red wine.

Table 19

Results of Model Comparison Node for White Wine Data

Model	ROC Index	Misclassification Rate	RASE	Speed (s)
Random Forest	0.710	0.480	0.290	12.750
Neural Network	0.950	0.460	0. 300	3.890
Gradient Boosting	0.640	0.490	0. 310	5.810

Note. Table 19 shows the result of the performance of each type of machine learning model used to predict quality of white wine.

From both Table 18 and Table 19, Neural Network performs when predicting the quality of red wine and white wine. It has the highest ROC index. This would mean that Neural Network model is able to separate the quality classes more efficiently than the other models. Also, the model has the lowest misclassification rate, proving that the overall accuracy in predicting quality of both red wine and white wine. Even though the RASE value of Neural Network model is not the lowest, it still does not differ much with the RASE value of Random Forest and Gradient Boosting models. Lastly, in terms of processing speed. Neural Network model takes a lesser time to train and predict the quality of red wine. In that case, looking at the overall performance, Neural Network would be the most suitable model to predict the quality of red wine and white wine.

4.2.4.2 Problem 2. Table 20 shows the performance of each type of data mining models when classifying between red wine and white wine.

Table 20

Results of Model Comparison Node for Wine Classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Speed (s)
Random Forest	97.31	97.49	98.99	98.23	5.270
Neural Network	98.08	100.00	97.44	98.70	3.880
Gradient Boosting	97.31	97.49	98.99	98.23	5.000

Note. Table 20 shows the of results of the performance of each machine learning model used to classify red wine and white wine.

Based on the results shown in Table 20, Neural Network model has the best accuracy, precision and F1-score. Even though it does not have the highest recall score, it still has a very high efficiency when classifying the type of wine. Other than that, the running speed for Neural Network model is also the shortest compared to the other three machine learning models.With that, it is safe to say that Neural Network model is the best data mining model to classify red wine and white wine.

4.3 Findings of each model

Other than performance, each data mining model also provide other valuable insights for stakeholders. This subsection would discuss the findings of each model based on the problems formulated.

4.3.1 Random Forest Model

In general, the Random Forest would display feature importance when carrying out prediction tasks. To analyze the importance of each feature, the number of splitting rules would be examined. The number of splitting rules basically indicates the number of times each tree splits at the specific feature variable. The more times the feature variable is used to split a tree, the more important it is. Thus, in other words, the higher the number of splitting rules, the more important the variable.

4.3.1.1 Problem 1. For Problem 1, each importance of the chemical properties to predict the quality of red wine and white wine are calculated. Table 21 and Table 22 shows the number of splitting rules in the Random Forest models when predicting quality of red wine and white wine.

Table 21

Variable Importance When Predicting Quality of Red Wine

Variables	Number of Splitting Rules
volatile_acidity	732
total_sulfur_dioxide	588
sulphates	493
fixed_acDidity	473
density	463
pH	412
free_sulfur_dioxide	411
residual_sugar	401
alcohol	362
citric_acid	308
chlorides	252

Note. Table 21 shows each feature variable in the red wine dataset and their respective number of splitting rules in the Random Forest model trained.

Table 22

Variable Importance When Predicting Quality of White Wine

Variables	Number of Splitting Rules
volatile_acidity	2510
total_sulfur_dioxide	2356
residual_sugar	1832
pH	1768
sulphates	1728
free_sulfur_dioxide	1660
density	1542
fixed_acidity	1537
citric_acid	1248
alcohol	1239
chlorides	1129

Note. Table 22 shows each feature variable in the white wine dataset and their respective number of splitting rules in the Random Forest model trained.

From Table 21 and Table 22, it shows that volatile acidity has the greatest number of splits in the Random Forest models when predicting quality of red wine and white wine. In that case, it is safe to say that volatile acidity is an important aspect when it comes to determining the quality of red wine and white wine. With this information, the stakeholders can know full well how to improve the quality of the red wine and white wine, which is by adjusting the amount of volatile acidity. As seen back in the correlation coefficient between volatile acidity and quality. The negative value indicates that the lower the volatile acidity, the higher the quality. Therefore, stakeholders can take note on this and keep volatile acidity of red wine and white wine to a minimum. On top of that, they can prioritize the other critical chemical properties that would directly influence the quality of red wine and white wine, allowing them to produce red wines and white wines with high quality.

4.3.1.2 Problem 2. As for Problem 2, the importance of variables to classify the type of wine can be found through Random Forest model. Table 23 shows the number of splitting rules of each feature variable in the combined dataset when training the model.

Table 23

Variable Importance When Classifying Type of Wine

Variables	Number of Splitting Rules
total_sulfur_dioxide	358
volatile_acidity	290
chlorides	230
free_sulfur_dioxide	206
fixed_acidity	197
sulphates	183
pH	177
density	164
citric_acid	146
residual_sugar	132
alcohol	83
quality	43

Note. Table 23 shows each feature variable in the white wine dataset and their respective number of splitting rules in the Random Forest model trained.

From Table 23, the most important and significant feature variable when classifying between red wine and white wine is total sulfur dioxide as it has the greatest number of splitting rules in the Random Forest model. In other words, this would mean that the total amount of sulfur dioxide can somehow best determine whether the wine is a red wine or white wine. With that, stakeholders are able to know which unique chemical property of the wine can be used to distinguish between red wine and white wine. Then, they can take note of this when it comes to quality control, marketing activities and also wine making process.

4.3.1.3 How Organizations Benefit. In conclusion, getting hold of variable importance through Random Forest model can be beneficial for the stakeholders. For Problem 1, stakeholders like winemakers or wine production companies can use this information to produce wine with finer quality by adjusting the amount of each chemical property in the wine. If the volatile acidity has a high number of splitting rules, it would naturally mean it is a critical factor in determining quality of red wine and white wine. Thus, stakeholders can control the volatile acidity of the wine to ensure high quality wine production. By focusing on the impactful variables, cost can be reduced due to better resource allocation. Lastly, it also help in marketing the wine. Stakeholders can highlight the important chemical properties that determines the quality of wine to customers. With data-backed insights, customers would be keener to buy the wine since it has a higher quality.

For Problem 2, the number of splitting rules tells the stakeholders about which chemical property is more critical in distinguishing between red wine and white wine. They stakeholders can implement this into their own automated classification system. This not only prevents human error but also saves time. On top of that, the system of stakeholders can be further optimized by simplifying the model. By selecting the relevant chemical properties, the system can still produce accurate classifications even if lesser feature variables. Ultimately, it reduces the cost of the system as well. Lastly, it allows stakeholders to have better quality checks. For example, if stakeholders know that the total sulfur dioxide can best classify red wine and white wine. The quality control team can carry out target quality checks, making sure that the wine production is accurate and consistent.

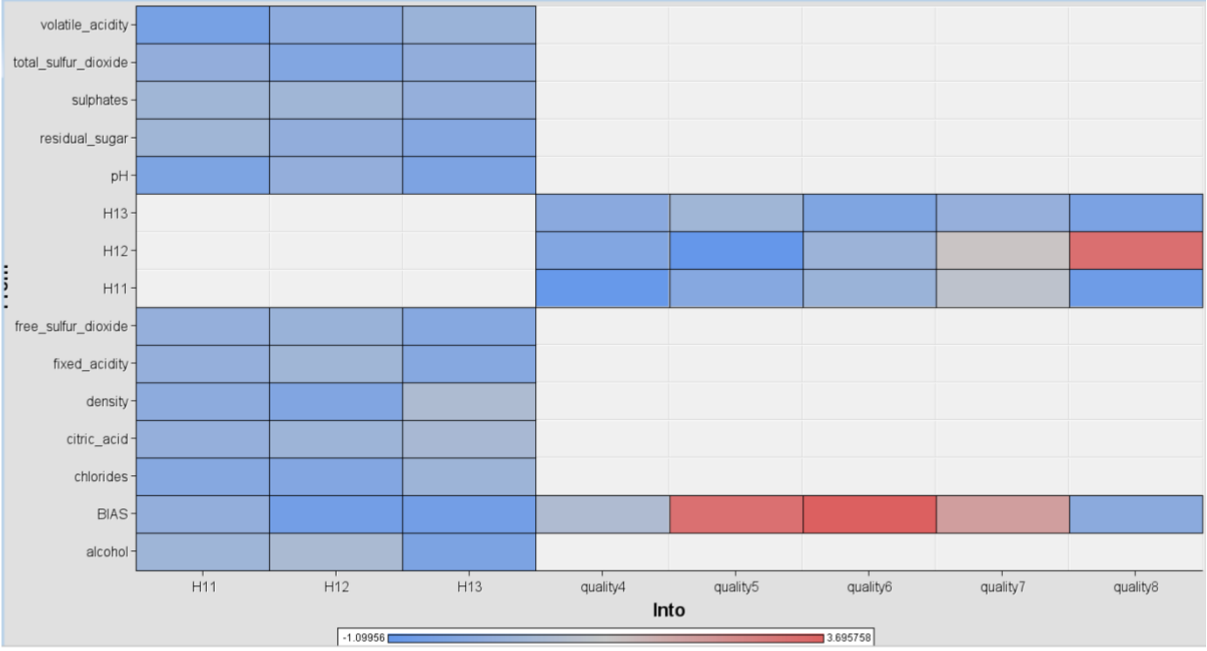
4.3.2 Neural Network Model

During analysis, it would be harder to interpret Neural Network model as it has a black-box nature. In that case, one finding that is given would be the weight plot of the model. However, it is important for stakeholders to know that despite the high accuracy of this model, it would be harder to come out with the exact findings. All the results shown, though could be interpreted, are only the potential results. Due to many hidden layers in the model, there are still many findings that cannot be found and be interpreted.

4.3.2.1 Problem 1. For Problem 1, the weight plot of the model when predicting quality of red wine and white wine are shown in Figure 29 and Figure 30 below.

Figure 29

Weight Plot of Neural Network Model When Predicting Quality of Red Wine



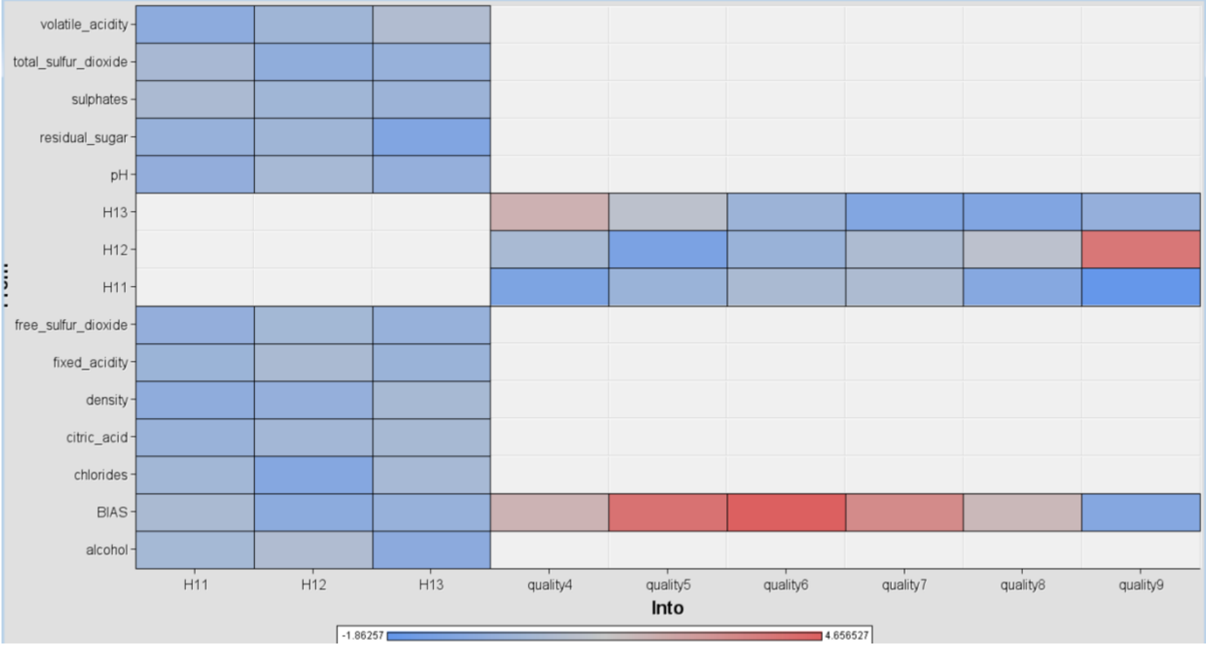
Note. Figure 29 depicts the weight plot of the Neural Network model trained to predict the quality of red wine. On the x-axis, it shows the feature variables used to predict the quality. The labels “H11”, “H12” and “H13” would represent the hidden layers in the model. For y-axis, the labels “quality4”, “quality5”, “quality6”, “quality7” and “quality8” are different classes in the target variable. There is also a BIAS label in the x-axis which is another parameter that shows the baseline adjustments for predicting each quality class. The weight can be differentiated using the color of the nodes. A bright blue shade shows that there is negative influence to the target variable whereas a bright red shade shows a positive influence with the target variable.

From Figure 29, there are three hidden layers shown: H11, H12 and H13. In each hidden layer, the weight of each feature variable and an additional BIAS column in shown. From the BIAS row, we can see that the hidden layer H11 has the lightest shade of blue, indicating a higher activation rate than H12 and H13. This would mean that H11 would be more sensitive and requires weaker input signals to activate. Thus, this specific hidden layer would be analyzed more in depth. From hidden layer H11, it can be seen that volatile acidity has the brightest shade of blue. This would mean that volatile acidity has the strongest negative weight among all the feature variables. This mean volatile acidity has the strongest negative relationship with quality of red wine. This also strengthens the findings from Random Forest model.

On the other hand, in the BIAS row, we can see that the output “quality5”, “quality6” and “quality7” has a red shade, especially “quality6”. This means the model would usually predict the red wine with a quality of 6 even there are no distinct traits from the red wine. This may be because the red wine in the dataset mostly have quality of 5, 6 and 7, making the model to have a tendency to predict these quality levels.

Figure 30

Weight Plot of Neural Network Model When Predicting Quality of White Wine



Note. Figure 30 depicts the weight plot of the Neural Network model trained to predict the quality of white wine. On the x-axis, it shows the feature variables used to predict the quality. The labels “H11”, “H12” and “H13” would represent the hidden layers in the model. For y-axis, the labels “quality4”, “quality5”, “quality6”, “quality7”, “quality8” and “quality9” are different classes in the target variable. There is also a BIAS label in the x-axis which is another parameter that shows the baseline adjustments for predicting each quality class. The weight can be differentiated using the color of the nodes. The brighter the blue shade, the lower the weight; the brighter the red shade, the higher the weight.

From Figure 30, there are also three hidden layers shown: H11, H12 and H13. It can be seen that H11 has the lightest shade of blue in the BIAS row out of all the hidden layers. This would mean that H11 would be more likely to be activated as it has a higher weight. Even with weaker input signals, H11 would most likely be activated first, then followed by H13 and H12. Thus, the weight of each feature variable in H11 and be further examined. It can be seen that volatile acidity has the brightest shade of blue in H11. Therefore, the significance of volatile acidity is higher when predicting the quality of white wine. In fact, with the shade of blue, it would also indicate that there is a negative influence towards the quality of white wine. In other words, this means that the lower the volatile acidity, the higher the quality of red wine.

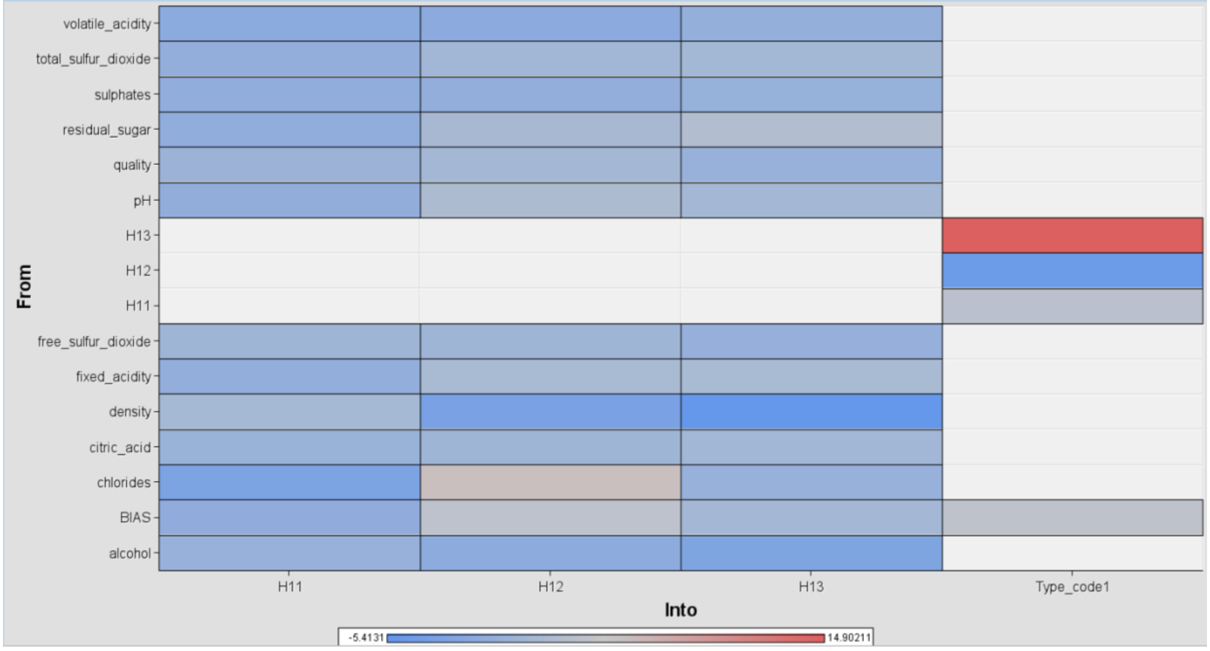
Another finding could also be obtained through the BIAS row. We can see that the shade of red is the brightest in the outputs “quality5”, “quality6” and “quality7”. This would mean that this model has the tendency to predict the white wine with a quality score of 5, 6 or 7. In short, when there is no strong input signals for the model to predict, the model would classify the white wine as quality of 5, 6 or 7. In this case, since the shade of red is especially the brightest at “quality6”, this would mean that the model tend to classify the white wine with a quality of 6 even there are no clear signs that the white wine has the chemical properties that matches a quality 6 white wine.

With these weight plots, we can see the chemical properties in the wine that would potentially influence their quality. It indicates the connections of each feature variable with the nodes that would predict the wine quality. With that, it gives an additional insight about the relationship between each feature variables with the quality of red wine and white wine.

4.3.2.1 Problem 2. For Problem 2, a weight plot can also be obtained to see the weights assigned to each feature variables in the hidden layer. Figure 31 shows the weight plot of the Neural Network model trained to classify the type of wine.

Figure 31

Weight Plot of Neural Network Model When Classifying the Type of Wine



Note. Figure 31 depicts the weight plot of the Neural Network model trained to predict the type of wine. On the x-axis, it shows the feature variables used to predict the quality. The labels “H11”, “H12” and “H13” would represent the hidden layers in the model. For y-axis, the label “Type_code1” is one of the classes in the target variable. There is also a BIAS label in the x-axis which is another parameter that shows the baseline adjustments for predicting each quality class. The weight can be differentiated using the color of the nodes. The color ranges from blue to red, blue being the lowest while red being the highest weight.

From Figure 31, the BIAS row shows that the hidden layer H12 has a shade closer to red. This would indicate that H12 would have a higher weight compared to H11 and H13. Thus, H12 would be more likely to be activated and requires weaker input signals. As we can see, when H12 is activated, chlorides has a relatively light shade of red. This indicates that chlorides has the largest weight among the feature variables, indicating that it has a strong positive relationship with the type of wine. This means the higher the number of chlorides, the more likely the wine would be code 1, which is white wine. From the Random Forest model, we can also see that chlorides is placed high in terms of variable importance as well.

Another thing that stood out in the weight plot would the bright red in H13 when predicting the wine type as white wine. This would mean that in the hidden layer H13. It is very likely for it to have an output as white wine. Even if there is a weak trait, the hidden layer would still tend to produce an output of predicting the type of wine as white wine.

This weight plot will enlighten us more about the chemical properties that have the most potential to influence the type of wine. In other words, it tells us which feature that is more likely to be a strong indicator between red wine and white wine. With a higher weight connecting to the nodes in the layer, it would mean that the input might have more significance than the other inputs.

4.3.2.3 How Organizations Benefit. For Problem 1, by viewing the weight plot, stakeholders can understand the influential chemical properties that would potentially affect the quality of red wine and white wine. With that, winemakers can optimize the production of these wines by adjusting the important chemical properties. Other than that, it also helps in quality control. Stakeholders can use this information as benchmark when checking the quality of red wine and white wine. As a result, it would naturally increase the standard of the wines produced.

Lastly, for Problem 2, it also beneficial for stakeholders to know what kind of chemical property can best distinguish red wine and white wine. With this knowledge, stakeholders can ensure consistent production of their red wine and white wine by taking the potential indicators into consideration. On top that, the stakeholders can market their wines better. Targeted spromotions can be carried to sell the red wine and white wines based on the distinctive chemical properties. Nonetheless, interpretability had always been one of the drawbacks for this model. Stakeholders cannot be extremely confident when obtaining the findings of this model.

4.3.3 Gradient Boosting Model

On the other hand, Gradient Boosting is able to view the exact importance of each feature variable when predicting the target variable in general. Since Gradient Boosting model is also made up from sequential decision trees, we can view the number of splitting rules for each feature variable to see how many time the variable is used to split the tree. It tells us which feature is frequently used to split and also how complexed the whole model is. In addition, validation importance can also be analyzed through this model. It indicates the feature variables' contribution to the accuracy of prediction tasks through the validation dataset. Unlike number of splitting rules, it can directly tell us which feature variable are most reliable and valuable to make more accurate predictions. This subsection would uncover the number splitting rules and the validation importance given by the Gradient Boosting models for Problem 1 and Problem 2.

4.3.3.1 Problem 1. Table 24 and Table 25 shows the calculated number of splitting rules and validation importance of each chemical properties when predicting quality of red wine and white wine respectively.

Table 24

Number of Splitting Rules of Each Feature Variables When Predicting Quality of Red Wine

Variables	Number of Splitting Rules	Validation Importance
alcohol	36	0.872
sulphates	24	1.000
volatile_acidity	23	0.635
fixed_acidity	10	0.207
chlorides	10	0.308
total_sulfur_dioxide	10	0.669
density	9	0.470
residual_sugar	9	0.000
pH	6	0.000
free_sulfur_dioxide	6	0.015
citric_acid	5	0.228

Note. Table 24 shows the number of splitting rules for each chemical property of red wine when predicting its quality using Gradient Boosting model.

Table 25

Number of Splitting Each Feature Variables When Predicting Quality of White Wine

Variables	Number of Splitting Rules	Validation Importance
alcohol	36	1.000
sulphates	24	0.205
volatile_acidity	23	0.835
chlorides	10	0.385
fixed_acidity	10	0.000
total_sulfur_dioxide	10	0.015
residual_sugar	9	0.085
density	9	0.000
free_sulfur_dioxide	6	0.459
pH	6	0.000
citric_acid	5	0.000

Note. Table 25 shows the number of splitting rules for each chemical property of white wine when predicting its quality using Gradient Boosting model.

From Table 24 and Table 25, we can see that alcohol is mostly used to split the decision trees in the Gradient Boosting model trained. This tells us that the variable alcohol has contributed the most when training the model. Though it may not mean, that it is the most important chemical property used to predict the quality of red wine and white wine, alcohol still shows significance in this prediction. On the other hand, we can directly see the importance of the chemical properties when predicting the quality of red wine and white wine in the validation dataset through validation importance. The validation importance of the chemical properties of red wine and white wine are different. In red wine, sulphates is shown to be the most important chemical property when predicting the quality of red wine, followed by alcohol. Meanwhile, alcohol remains the most important variable to predict the quality of white wine, followed by volatile acidity.

4.3.3.2 Problem 2. For Problem 2, number of splitting rules and validation importance are also analyzed. With that, we are able to see which chemical property is mainly used to split the decision trees to classify the type of wine and which chemical property is the most important one in the classification task. Table 26 shows the calculated number of splitting rules and validation important of each chemical properties when classifying red wine and white wine.

Table 26

Number of Splitting Rules of Each Feature Variables When Classifying Red Wine and White Wine

Variables	Number of Splitting Rules	Validation Importance
total_sulfur_dioxide	37	1.000
chlorides	21	0.760
volatile_acidity	20	0.464
sulphates	7	0.056
density	4	0.110
residual_sugar	4	0.070
fixed_acidity	4	0.006
alcohol	3	0.000
pH	2	0.000
citric_acid	0	0.000
free_sulfur_dioxide	0	0.000
quality	0	0.000

Note. Table 26 shows the number of splitting rules and validation importance of each respective chemical properties for the Gradient Boosting model trained to classify the type of wine.

In Table 26, the chemical property total sulfur dioxide has the greatest number of splitting rules among the other chemical properties. This means that it has contributed the most when it comes to splitting the decision trees in the model, allowing it to be the most relevant of them all. Additionally, total sulfur dioxide also has the highest validation importance. Naturally, it tells us not only it contributes the most in training the model, it is also the most importance chemical property used to classify the type of wine.

4.3.3.3 How Organizations Benefit. For Problem 1, validation importance can help stakeholders to have an additional insight on which chemical property would significantly affect the quality of red wine or white wine. By knowing the more important property, stakeholders like wine production companies or winemakers can use it to optimize the fermentation processes to maintain the quality of red wine. Other than, they can also optimize resource allocation. For instance, stakeholders purchase better materials to control the acidity of red wine or white wine, since acidity is seen having a greater importance with the quality. As for number of splitting rules, stakeholders can know the feature variables that are significantly contributing in training the model. For future models, feature selection can be carried out. This prevents overfitting and ensures scalability for the stakeholders s well.

For Problem 2, understanding validation importance allow stakeholders to also obtain actionable insights. If the alcohol level can best distinguish red wine and white wine, stakeholders can adjust the fermentation and distillation process to produce a more accurate type of wine that they desire. In the end, the wine produced can be marketed correctly as well. On top of that, it also optimizes the storage management. Knowing that alcohol has a higher validation importance, red wine with higher alcohol can be stored under optimal condition for aging. Besides that, by understanding the number of splitting rules, stakeholders can use it to optimize their automated sorting systems with only the important variables. That way, it provides a simpler decision-making process. With a simpler system, it automatically saves the cost for the stakeholders.

5.0 Conclusion and Recommendations

5.1 Summary of Key Findings

In a nutshell, out of the three data mining models which are Random Forest model, Neural Network model and Gradient Boosting model, Neural Network is seen to be the best. For Problem 1: predicting quality of red wine and white wine, it achieves the lowest misclassification rate while whereas for Problem 2, it has the highest accuracy. It would be best for stakeholders to choose this model and implement it into their own automated predicting system to achieve the best quality. On top of that, Neural Network model takes the shortest runtime, allowing accurate predictions to be made within seconds. However, one drawback of this model would that it does not give much concrete insights like Random Forest model and Gradient Boosting model. This is because of its black-box nature. If stakeholders would like a more explainable model, Random Forest can be implemented. Though its performance is not as high as Neural Network, it also has a rather high accuracy and it also shows the feature importance as well. Nonetheless, if stakeholders would like to solely carry out prediction tasks, Neural Network models would still be the best choice.

When it comes to variable importance, volatile acidity is seen to have the greatest importance when predicting the quality of red wine and white wine. It has the greatest number of splitting rules in Random Forest model and the greatest weight in Neural Network model. This tells stakeholders that they can use this chemical property to optimize the production quality. As for classifying the type of wine, total sulfur dioxide is seen to be the most relevant indicator. With this information about the significance of each chemical property, stakeholders can use it to optimize their red wine and white wine production process and also improve their own automated predicting systems.

5.2 Recommendations

Overall, it is recommended for stakeholders to always make data-driven decisions. Therefore, red wine and white wine dataset have to be constantly updated. This way, the data mining models in the system can be refined and produce better predictions and classifications. On top of that, stakeholders can also constantly monitor the feature variable that have the most impact on predictions as it might change as time goes by. With that, the chemical properties can then be adjusted to produce better quality red wine and white wine.

Moreover, as technology advances and machine learning models become an important tool for predictions in all industries, stakeholders can start training or hiring more staff. The staff in the organizations have to be knowledgeable on the models applied and should know how to interpret each model. To do this, stakeholders can organize training sessions for the staff to let them understand the findings obtained by the machine learning models in the system. With a more informed team, better decision can be made when it comes to wine production and also wine marketing.

Lastly, without a doubt, stakeholders should start investing on predictive analytics tools to further analyze wine data. User-friendly softwares such as SAS Enterprise Miner, SAS Viya or Power BI can be purchased by the stakeholders to carry out predictive analytics. These tools not only can make accurate predictions, they also provide additional actionable insights for stakeholders. With that, the systems can have optimum performance and the organizations can make better decisions to boost their sales. In short, it is able to achieve operational efficiency.

5.3 Overall effectiveness of data mining processes

The data mining process in this project is considered relatively effective. For one, the data preprocessing phase went smoothly. The dataset is explored thoroughly and relevant data handling is done. Data modifications such as encoding and standardization are done properly before the dataset is used to train the machine learning models. This ensures that the data in the red wine dataset, white wine dataset and combined dataset is accurate and generalizable, allowing better predictions.

On top of that, in terms of model selection, it is also done effectively as all the models selected are flexible and can process both binary and nominal target variables. The models chosen each have their main strengths as well. Random Forest model has an easy interpretability; Neural Network model is able capture more complexed patterns; Gradient Boosting model can make fine-tuned predictions. With that, the findings of each model are very useful for the stakeholders too.

When it comes to training and evaluating the data mining models, suitable metrics are also used. For nominal target variable (quality), ROC index, misclassification rate and RASE value are used. This can capture not only the accuracy of the model but also the error rates. As for binary target variable (type), accuracy, precision, recall and F1-score of the model are compared between one another. It directly tells the stakeholders about the performance of each model when classifying red wine and white wine. On top of that, other insights such as the feature importance and run time speed are also evaluated to obtain more information from the model. This means the models are explainable, allowing stakeholders to use them to make decisions.

5.4 Areas of Improvement

There is no doubt that there are still areas for improvement in this data mining process. Firstly, we can further improve the data quality in the data preprocessing process. Class imbalances can be addressed and handled using techniques such as Synthetic Minority Oversampling Technique (SMOTE) or undersampling. On top of that, feature engineering can also be carried to produce a new relevant feature variable that would train the model in a better way. Also, outliers should be transformed or removed immediately if it is found that they are causing extreme errors or bias in the models. With that, then only a dataset with better quality can be produced. A better dataset would ultimately improve the model performance.

In addition, hyperparameter tuning can also be carried out in a better way. The environment used in the project: SAS Enterprise Miner and users cannot directly carry out advanced automated hyperparameter tuning techniques such as Grid Search or Bayesian Optimization. Thus, the hyperparameter tuning process was done manually instead. To solve this, other more flexible environments such as Python can be used to carry out this technique. Then, model training can still be conducted using SAS Enterprise Miner using the best hyperparameter possible. As a result, it would allow better predictions from the models.

References

Banoula, M. (2023, August 13). *What is Neural Network: Overview, applications and advantages.* Simplilearn. https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-neural-network

Belyadi, H. & Haghighat, A. (2021). *Machine learning guide for oil and gas using Python: A step-by-step breakdown with data, algorithms, codes and application.* Elsevier. https://www.sciencedirect.com/topics/computer-science/gradient-boosting#chapters-articles

Bhardwaj, P., Tiwari, P., Olejar, K. Jr., Parr, W. & Kulasiri, D. (2022). A machine learning application in wine quality prediction. *Machine Learning with Applications, 8*. https://www.sciencedirect.com/science/article/pii/S266682702200007X#b32

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F. & Campbell, J. P. (2020). Introduction to machine learning, Neural Networks, and deep learning. *Translational Vision Science & Technology, 9*(2), 14. https://pmc.ncbi.nlm.nih.gov/articles/PMC7347027/

Dahal, K. R., Dahal, J. N., Banjade, H. & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics, 11*(2). https://www.scirp.org/journal/paperinformation?paperid=107796

Er. Y. & Atasoy, A. (2016). The classification of white wine and red wine according to their physicochemical qualities. *Intelligent Systems and Applications in Engineering, 4*(1), 23-26. https://dergipark.org.tr/en/download/article-file/232398

Han, S.H., Kim, K. W., Kim, S. Y. & Youn, Y. C. (2018). Artificial Neural Network: Understanding the basic concepts without mathematics. *Dement Neurocogn Disord, 17*(3), 83-89, https://pmc.ncbi.nlm.nih.gov/articles/PMC6428006/

Latessa, S. H., Hanley, L. & Tao, W. (2023). Characteristics and practical treatment technologies of winery wastewater: A review of wastewater management at small wineries. *Journal of Environmental Management, 34*. https://www.sciencedirect.com/science/article/abs/pii/S0301479723011313#preview-section-abstract

Perez-Magariño, S., Heras-Ortega, M., José, M. L. G. & Boger, Z. (2004). Comparative study of artificial neural network and multivariate methods to classify Spanish DO rose wines. *Talanta, 62*(5), 983-990. https://www.sciencedirect.com/science/article/abs/pii/S0039914003006155

What is random forest?. (n.d.). IBM. https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

Appendix A

Link to online dataset from UCI Machine Learning Repository website:
https://archive.ics.uci.edu/dataset/186/wine+quality