To: Dr. Mitch Cochran
From: Team 3
Date: November 17, 2022
Subject: Data Mining Projects Report on Red Wine Quality Prediction

The purpose of this memo is to summarize the main findings of the data mining project. You can find our detailed code via this Google Colab Link. Here are the members of Team 3:

- Nguyen Ha My – 11196110
- Nguyen Thao Hau – 11196425
- Hoang Nguyen Long – 11196386

1. Organizational Understanding

We are playing the role of a wine manufacturer searching for the greatest-tasting red wine. We want to apply machine learning models in the production process to ensure that every wine bottle has the highest quality. This will in turn reduce the cost of each wine bottle as less human labor is required for monitoring of quality. Therefore, this analysis aims to predict wine quality given some essential attributes of wine.

We have to deal with the Red Wine Quality Dataset, which consists of 1599 observations and 12 attributes (11 prediction attributes and 1 label of wine quality). Most importantly, the analysis aims to seek the answers to three main questions:

- What are the attributes that potentially have strong effects on the wine quality?
- What is the flavor profile for "good" wine quality in terms of acids, flavors, and preservatives?
- With the predefined chemical elements of red wine products, what will be the corresponding quality levels?

2. Methods and findings summary

Exploratory Data Analysis was carried out to answer the first 2 questions. Regarding the answer to the first question, the main method was correlation analysis to anticipate the influences of 11 chemical components on wine quality based on Pearson's correlation coefficients. Generally, the top 3 chemical components that influence red wine quality most significantly in descending order are Alcohol, Volatile Acid, and Sulphates.

About the answer to the second question, we applied descriptive statistics findings to rule out the profile of high-quality wine. Practically, the wine manufacturer should not go overflow with chemicals that are positively correlated with good wine quality. Thus, the appropriate range for each chemical is provided. For instance, a good red wine should not have excessive volatile acids, residual sugar, and chlorides. Sulfates and other preservative attributes should be added within the health safe regulation.

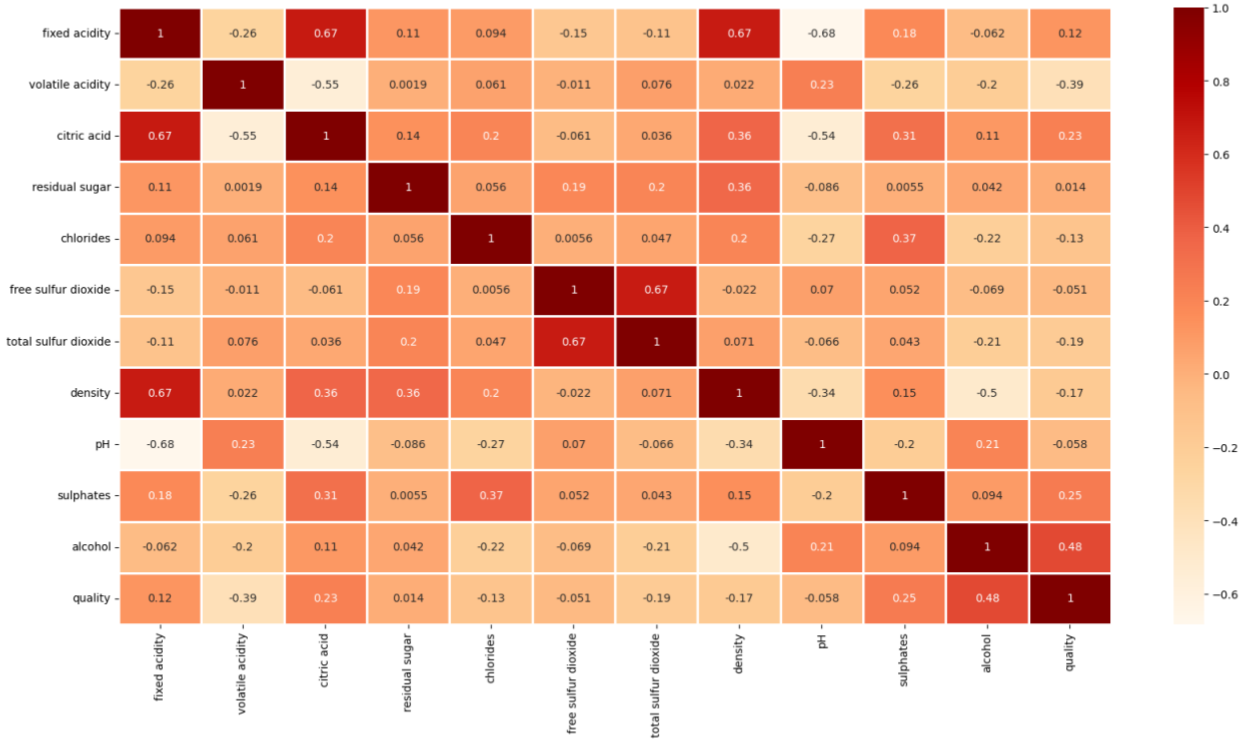
To answer the third question, we applied four different prediction models (1) Random Forest, (2) Logistics Regression, (3) Support Vector Machine, and (4) KNN. By comparing their accuracies, the best model was found as Random Forest Classifier with 90% accuracy. Thus, this could be applied in the actual manufacturing process in the hope to increase quality and reduce cost.

The detailed analysis to come up with the above findings will be presented in the following part of the memo.

3. Detailed Analysis

a. The relationships between chemical elements and red wine quality

First, about the main chemical components that potentially affect good wine quality, we applied Exploratory Data Analysis by calculating and visualizing the correlation matrix and Pearson correlation coefficient to figure out this issue. We decide to choose the top 3 most influential factors of wine quality depending on the levels of relationship between each independent variable and the predictor – red wine quality. The results are illustrated below:



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

- About the strong relationship:
  - Alcohol: strong positive correlation with quality (Pearson correlation = 0.48)
- About the moderate relationship:
  - Volatile Acid: moderate negative correlation with quality (Pearson correlation = -0.39)
  - Sulphates: moderate positive correlation with quality (Pearson correlation = 0.25)
  - Citric Acid: moderate positive correlation with quality (Pearson correlation = 0.23)
- About the weak relationship: the rest of independent variables have a weak relationship with red wine quality (the Pearson correlation range from -0.19 to 0.12)

Looking at the correlation matrix, we chose top 3 chemical components that influence the most significant on the red wine quality are Alcohol, Volatile Acid, and Sulphates. It can be seen that each of the 3 chemical components represents different criteria of the red wine: acidity, flavor, and preservative. These initial findings act as guidelines for the wine research team to pay more attention to adjusting the chemical components in all three criteria to improve wine quality.

b. The detailed profile of "good" red wine

We applied descriptive statistics findings to rule out the profile of high-quality wine. Practically, the wine manufacturer should not go overflow with chemicals that are positively correlated with good wine quality. For instance, Alcohol has the strongest positive correlation with red wine quality, but it does not mean that the proportion of alcohol content should be overly used in the wine. Thus, the appropriate range for each chemical is provided.

The original dataset has divided the label "quality" by 8 levels of red wine quality. Here, as we focus on the best-tasting wine, we only rule out the profile of higher wine quality of level 7 and level 8.

Criteria	Chemical Components	Range	Explanation
Acid	Fixed Acidity	Between 8 and 13g/L	High-quality wines have higher acidity levels.
	Volatile Acidity	Between 0.3 and 0.5g/L	The acid is associated with the smell and taste of vinegar that creates unpleasant characteristics of wine. Higher wine quality only has a small amount of volatile acidity.
	Citric Acids	Between 0.3 and 0.6g/L	Adding citric acid gives the wine "freshness" and supplements the fermentation process. Citric Acids are more commonly found in higher wine quality.
	PH level	Between 3.2 to 3.3	Allow the wine to have a crisper and tart taste than a low acidic wine which has a smoother rounder feeling.

Flavors	Residual Sugar	Below 2.5g/L	For high-quality wine, Residual sugar should be kept at a low level. Residual sugar may trigger re-fermentation in the bottle affecting its stability. Microbes may feed on the sugars left in the wine and generate unwanted flavors and gasses.
	Chlorides	Between 75 and 80 mg/L	Chlorides are the amount of salt in the wine. This should be kept under a controllable amount.
	Alcohol Level	Between 10.4 and 13.4%	A wine with a higher alcohol content will have a fuller, richer body, while a lower-level alcohol wine will taste lighter and more delicate on the palate
	Density	Around 0.996435 g/cc	The density of the wine is inversely proportional to the quality. Enhancing the quality of wine by the addition of sugar, alcohol, and other ingredients can cause the density of the liquid to drop.
Preservatives	Sulfur Dioxides	Between 275 and 290 mg/L	The European Union established a maximum permitted level of total SO2 in wine varying from 150 to 500 mg/L, which is dependent upon the sugar level of the product.
	Sulphates	Maintain at a level of 50 mg/L	Winemakers often use sulfites to help to minimize oxidation in wine and maintain its freshness. This in turn improves its taste and appearance.

c. The prediction of red wine quality based on machine learning models

To apply the machine learning models with high precision, the process of data preparation was carried out:

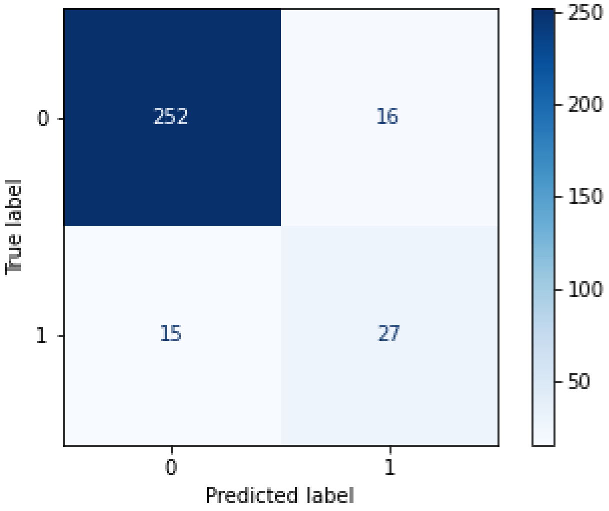
- Based on distribution plots, we found that most independent/ predictive attributes are right-skewed. To partly deal with the abnormality of the data, we applied the box cox algorithms.
- Based on box plot visualizations and descriptive statistics results, we successfully identified 51 outliers. As a solution, we decided to drop these outliers and keep a total of 1548 observations for machine learning modeling.
- Luckily, there were no missing values in this case.

Here, as we want to focus on predicting higher quality wine, we have encoded the 8 levels of quality into only 2 categories of high-quality wine (quality level >= 7) and low-quality wine. However, we encountered the problem of an imbalanced dataset with 1345 observations belonging to the class of 0 and only 203 observations belonging to the class of 1. As a solution, we apply the oversampling method by using SMOTE resampling to achieve the balance between the 2 classes for better prediction results.

With cleaned data, the train-test split was applied, with 80% train and 20% test. Four different prediction models were applied: (1) Random Forest, (2) Logistics Regression, (3) Support Vector Machine, and (4) KNN. The accuracy of the four models is presented in the figure below:

Rank	Score	ML Models
1	0.900000	Random Forest Classifier
2	0.867742	SVC
3	0.796774	KNN
4	0.793548	Logistics Regression

Comparing the four models, the best model was found as Random Forest Classifier with 90% accuracy. To better understand the prediction results, the confusion matrix under the Random Forest Classifier algorithm is also presented below. Class precision for 0 is 94% while that for 1 is 64%. Low class precision of 1 is due to the imbalance in the testing data which was predominated with the class 0. However, this imbalance was not a significant problem for our model since the imbalance of the training dataset was solved by SMOTE oversampling.



As a result, 90% accuracy for the Random Forest Classifier is reliable and the model could be applied in the actual manufacturing process in the hope to increase quality and reduce cost.

4. Business Values of this Project

Despite its small scope, we believe that this project has helped us rule out business values, which could be of great importance for further real-world analysis. These values are summarized into 2 parts of tangible and intangible values.

a. Tangible elements

- By figuring out how to make great-tasting wines, our company can increase revenue and profitability, creating more value for our stakeholders.
- Satisfying customers with an exquisite palette can help our company increase its market share domestically and internationally.

b. Intangible elements

- Maintaining high-quality products at a consistent level will help us to gain the loyalty of new and indecisive customers. Moreover, it will also enhance our customer retention capabilities.
- Improving our product's quality can gradually improve our brand recognition, differentiating our product from other competitors.

5. Limitations and future improvements

First, During the EDA process, we only focus on examining the one-way relationship between each independent variable and wine quality. Determining causality is never perfect in the real world, and risks can happen without careful experimental design. To check the reverse relationship between wine quality and 11 independent variables, working more with the research team and applying some experimental, statistical, and research design techniques for finding evidence toward causal relationships: e.g., randomization, controlled experiments and predictive models with multiple variables are some suggestions for our further research.

Second, the proposed machine learning models might be suitable only for small manufacturers instead of monitoring the quality of the luxurious types of wine. For wine aging from 30 years and over, it is more reasonable to rely on professional wine tasters. This is because most of the chemical components that exist in red wine are usually in very small quantities, some of which can only be discovered by years of wine-tasting experiences.

Thirdly, to enhance the business implications of cutting costs, additional attributes related to manufacturing costs should also need to be gathered. This in turn supports better accuracy and makes the results more reliable in business contexts.