Data Science Project: 1st Edition

Fikri Abdillah

June 10, 2024

Contents

Chapter 1 Introduction of the Wine

Wine is a grape drink that has gone through a fermentation process. European people usually use wine not only for drinking, but also as a cooking ingredient. Moreover, these wines are auctioned to collectors at auction houses at fantastic prices due to the quality and rarity of these drinks. However, the wine produced by the factory does not have both of these things. Therefore, this project aims to analyze and predict the quality of the wine produced by the factory.

☞ **Remark**
The results of this analysis are purely based on the results of a search by the author and have not been verified by a wine expert.

1.1 Goals

The goal of this project is to predict wine quality while providing brief knowledge to the public who are still unfamiliar with wine.

1.2 Data Description

This data was taken from the website UCI machine learning Repository. This wine dataset has about 1000 indices and 12 columns with **quality** as labels and 11 columns as features. The following is a table containing brief information on this data:

1.3 Tools

The software used in this project is as follows:

1.3.1 Python

Python is high-level programming languange which designed to wide range of application and supported by large number of libraries.

☞ **Remark**
This project was carried out using Python version 3.7.

1.3.2 Scikit Learn

Scikit Learn is a library that has quite extensive features. This library can import various machine learning models, supervised and unsupervised such as Random Forest, Decision Tree, Principle Component Analysis, and so on.

Activity 1.1
In this project, Scikit Learn will be utilized for preprocessing, data pipeline, and predicting wine quality.

No.	Columns	Role	Dtype	Mean	Median	Min	Std
1	quality	label	integer	5.65	6	3	0.805
2	citric acid	feature	float	0.26	0.25	0	0.196
3	volatile acidity	feature	float	0.53	0.52	0.12	0.179
4	sulphates	feature	float	0.65	0.62	0.33	0.170
5	free sulfur dioxide	feature	float	15.6	13	1.0	10.25
6	total sulfur dioxide	feature	float	45.9	37	6.0	32.78
7	density	feature	float	0.99	0.99	0.99	0.001
8	fixed acidity	feature	float	8.31	7.9	4.6	1.74
9	residual sugar	feature	float	2.53	2.20	0.9	1.35
10	chlorides	feature	float	0.086	0.079	0.01	0.047
11	pH	feature	float	3.31	3.31	2.74	0.156

Table 1.1. Informasi isi data

☞ **Remark**
The Scikit Learn used is version 1.5.0

1.3.3 Pandas

Pandas (Panel Data) is one of the python libraries commonly used to analyze, clean, explore, and manipulate data. This library is fast, powerful, flexible, and open-source. The limitation of Pandas is that it is not able to import very large data (big data).

Activity 1.2
In this project, Pandas is tasked with importing, creating data descriptions, and searching for 'null' values in the data.

☞ **Remark**
This project use version 2.2.2. Pandas.

1.3.4 Flowchart

The flowchart in the figure 1.1 is the work flow (*workflow*) in this project.

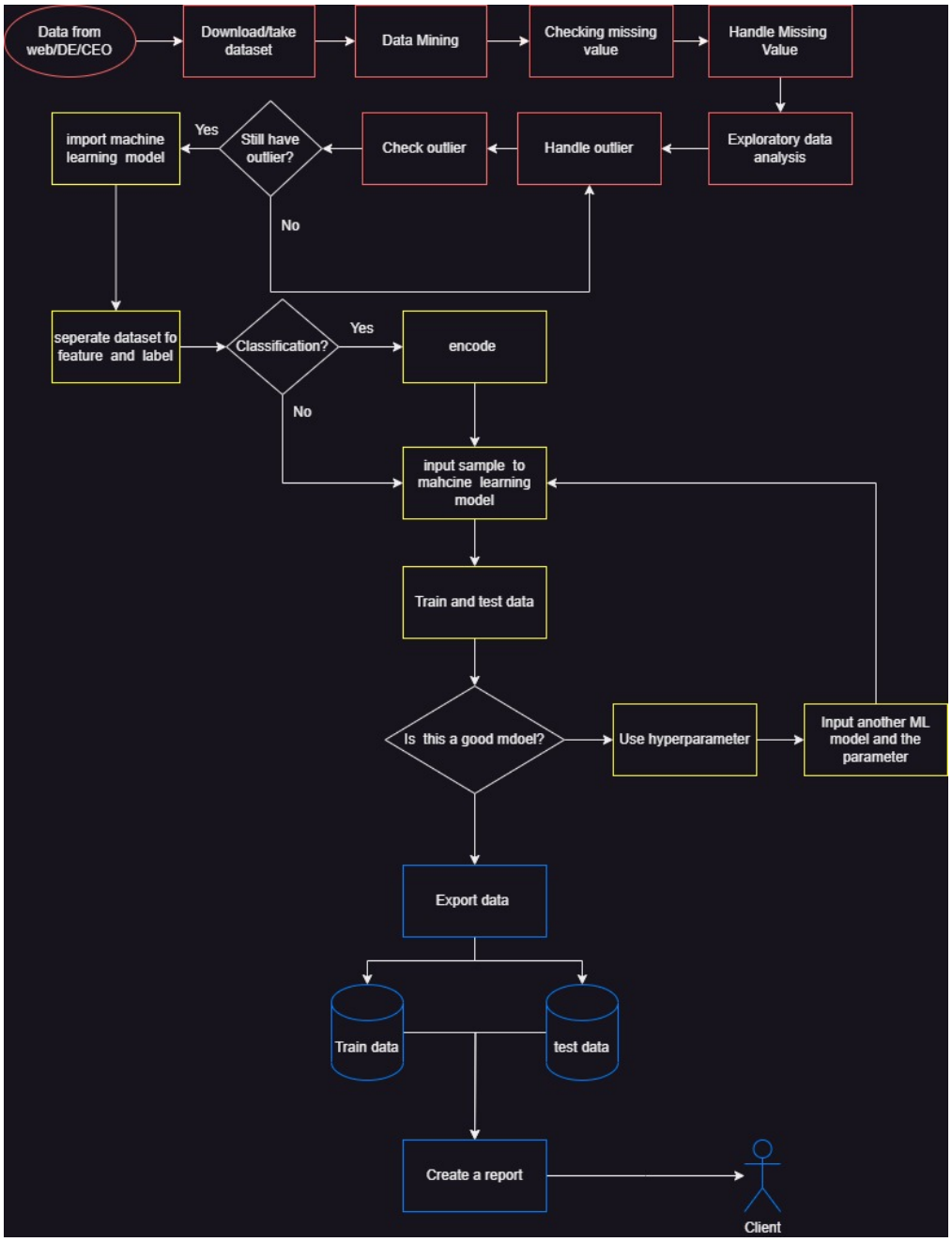

Figure 1.1. Flowchart of the Project

Chapter 2 Exploratory Data Analysis of the Wine

This chapter discusses the findings obtained from the data analysis results and provides a brief explanation of the data content.

2.1 Result and Findings

This session discusses what was found after carrying out *Exploratory Data Analysis*. Seaborn and Matplotlib are the two most frequently used python libraries for visualizing graphs in this project. Since there are many features in this data, we will discuss the features that have high correlation coefficient scores because, statistically, these features have a strong relationship to the label. Then, the features that have the opposite scores will be discussed briefly afterward.

Correlation, in statistics, is a measure and strength of the relationship between 2 variables. This project uses Pearson correlation to determine the magnitude 'Quality' influences on other feature with a value range of -1 to 1. Based on the Figure. 2.1, there are 4 features that have a large correlation score with the 'Quality' label.

Definition 2.1

Pearson Correlation is statistical measure that used to find out strength of the relationship between two variables which represented by its coefficient *r*. The coefficient can be expressed as

$$r = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where x_i is the value of i-th index in the feature, $\bar{x}$ is the feature mean, y_i is the value of the label at i-th index, dan $\bar{y}$ is the mean of the label.

2.1.1 Alcohol

Alcohol in wine plays a crucial role in terms of taste, aroma, preservation, and the experience it offers to the connoisseur. This compound gives texture, mouthfeel, and color thickness to wine. The higher the content, the more intense the color.

Definition 2.2

Alcohol is a byproduct formed from fermentation. Yeast 'eat' the the sugar from grape and then produces alcohol and carbon dioxide.

☞ Remark

Alcohol content is calculated based on the percentage of ABV (Alcohol By Volume) or the ratio of alcohol content to the total volume of wine. The ABV range in wine is from 5.5% to 17% (Grainger, 2021).

The 'alcohol' column in the dataset shows the alcohol content in wine formed from the grape fermentation process. *Feature* have positive correlation to 'quality'. This indicates that the higher the alcohol

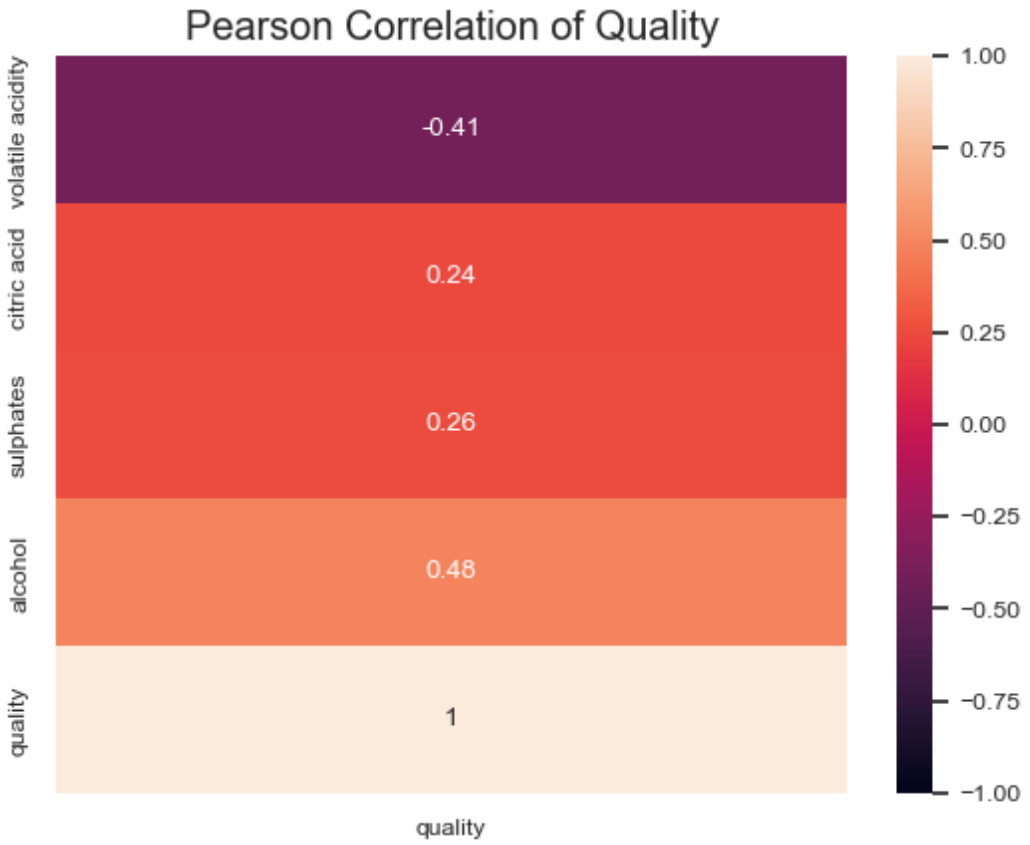

Figure 2.1. The highest correlation coefficient between quality and the other columns

content in wine, the better its quality. Figure 2.2 shows alcohol levels recorded in the dataset have a range of approximately 9% to 12%. If we assume the volume of the wine samples tested is the same, then wines with quality numbers 7 and 8 have a slightly higher alcohol content than the quality numbers below them.

2.1.2 Volatile Acidity

The 'volatile acidity' column is a column that has a fairly large coefficient score after alcohol. This feature represents the volatile acid content contained in the wine. This acid is gaseous, which produces a certain smell due to the collection of acids mixed in the wine.

Definition 2.3

Volatile (gaseous) acids are a group of acids that can be detected by inhaling them, because these acids are in the form of vapor.

In fact, one of the acids in wine, acetic acid, is also found in vinegar. Therefore, if the lid of a wine bottle is opened, it will produce a smell that is almost similar to vinegar (Molly, 2022). The Federal Tax and Trade Bureau, one of the countries with the highest wine consumption, has regulations regarding the limits of volatile acid levels in wine, namely 1.4 g/L for red wine and 1.2 g/L for white wine (Jackson, 2014). Figure 2.3 illustrates volatile acid contents found in each wine quality. The higher the quality, the lower the acid.

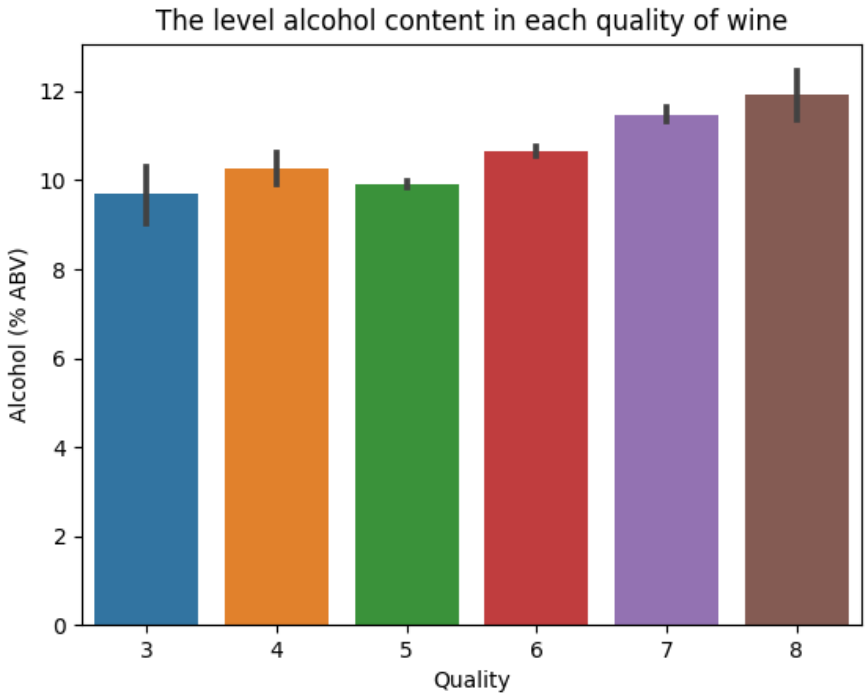

Figure 2.2. Plot of Alcohol level for each quality of Wine

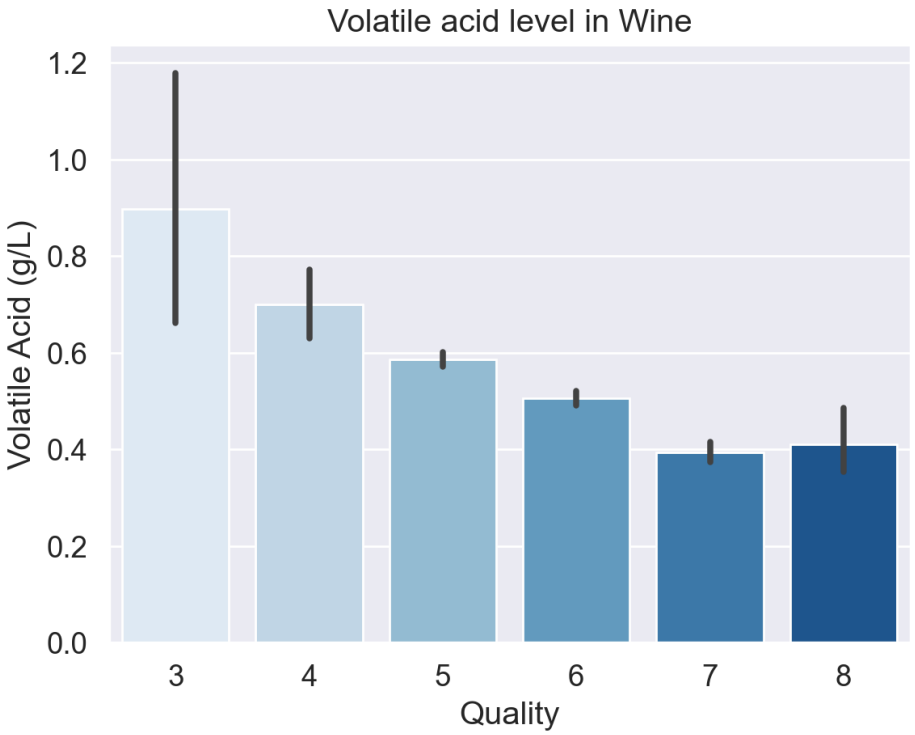

Figure 2.3. Volatile (gaseous) acid level in each Wine Quality

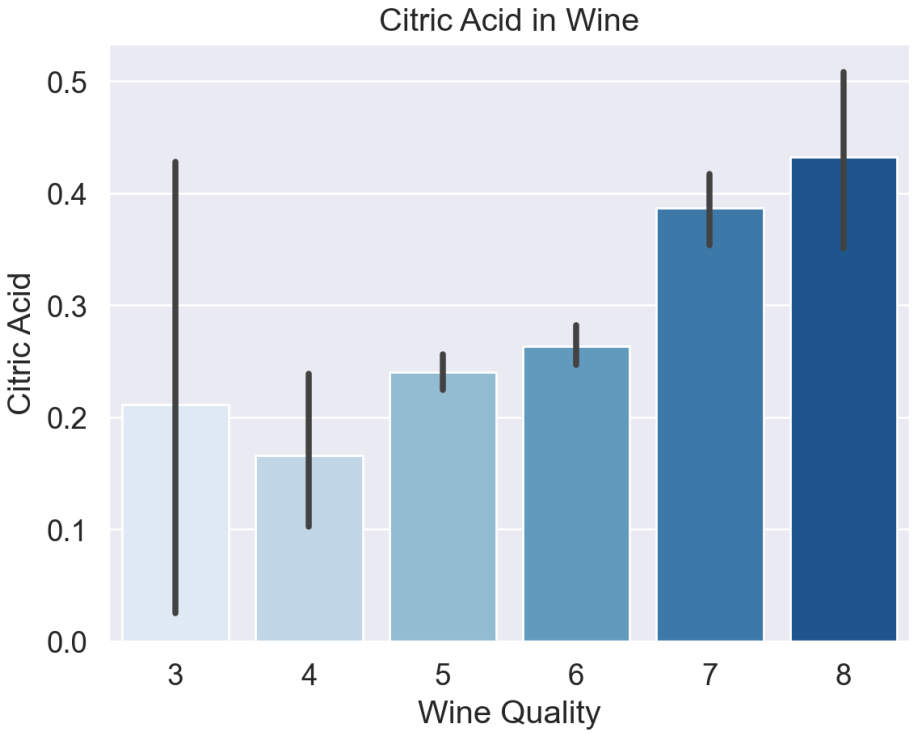

Figure 2.4. Citric Acid for each Wine Quality

☞ Remark
Volatile acid in wine is expressed in grams/liter or g/L (there is 1 gram of volatile acid in 1 liter of wine).

2.1.3 Citric Acid

Features that also need attention besides the two previously mentioned are citric acid and sulfates because their Pearson coefficient scores are quite close.

Definition 2.4
Citric acid is a weak organic acid that is usually used as a natural preservative or additive to food or drinks to enhance the sour taste, while sulphates (sulfites) are preservatives to maintain the freshness and taste of the wine. (European Food Information Council, 2018).

Citric acid is also found in wine so that the drink can extend its shelf life or just to add a sour and 'fresh' taste. However, this natural preservative can also create *microbial instability*.

Definition 2.5
Microbial Instability refers to the presence and activity of unwanted microbes happened in wine, such as bacteria, fungi, or yeast, which can spoil the wine.

If we refer to the barplot in Figure 2.4, the citric acid content increases with increasing quality number. However, Wine with quality number 3 has a higher citric acid content than wine with quality number 4. This is an anomaly because the data in quality 3 has *outliers*, so the mean value of the barplot is higher than quality 4.

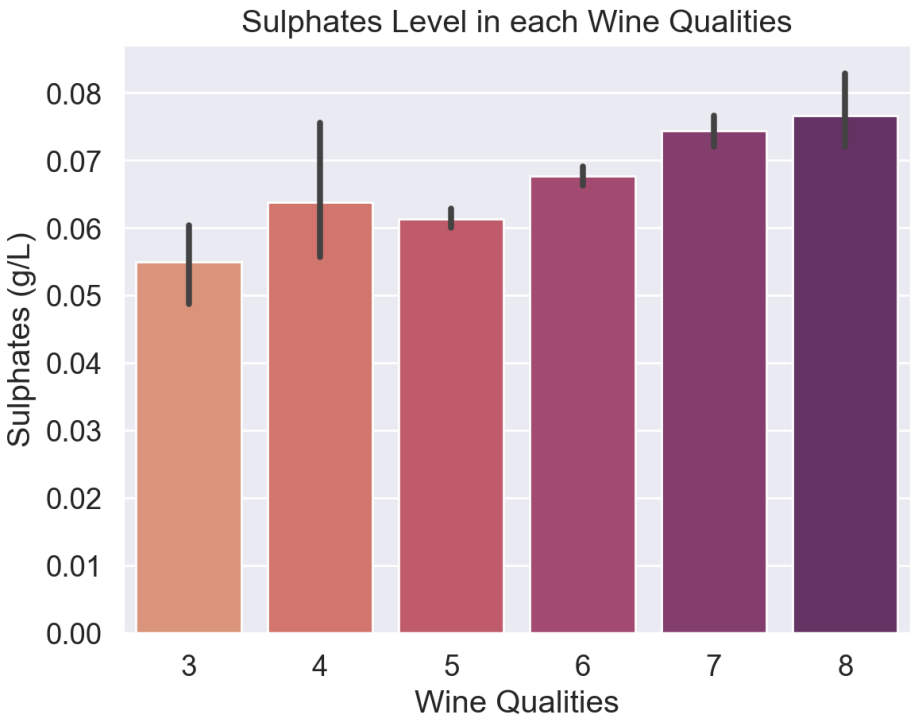

Figure 2.5. Sulphates in Wine

2.1.4 Sulphates

'Sulphates' has a correlation coefficient score that is almost as large as 'citric acid'. The Citric acid in wine aims not only to extend the shelf life and give a fresh aroma to the wine, but also to produce *microbial instability*. Therefore, the sulphates or sulphites in wine create stability so that the wine remains durable. Sulfites are formed naturally during the fermentation process. However, winemaker also add it if necessary.

If you look at the Figure 2.5, the sulfit content in the data ranges from 0.05 g/L to 0.075 g/L. According to *Winefolly*, several countries such as US have made regulations regarding the sulfite content in wine, namely maximum allowed is 0.35 g/L.

☞ Remark
If you look at the description data table 1.1, the mean value for sulfite (sulphates) feature is 0.66. This value is too large when compared to the wine regulation in several countries. The data source does not provide the information regarding the unit they used.

2.1.5 Another Features

After discussing several features that have large Pearson coefficient values, this section will discuss one of the other 7 features that has a small correlation score to the 'quality' label.

The 'residual sugar' column is a column that represents the remaining sugar content of the grapes after the fermentation process is complete. According to Winefolly, Wine is divided into several types based on its sugar content, namely:

1. Bone Dry: 0 to 1 gr/L
2. Dry: 1 to 17 gr/L

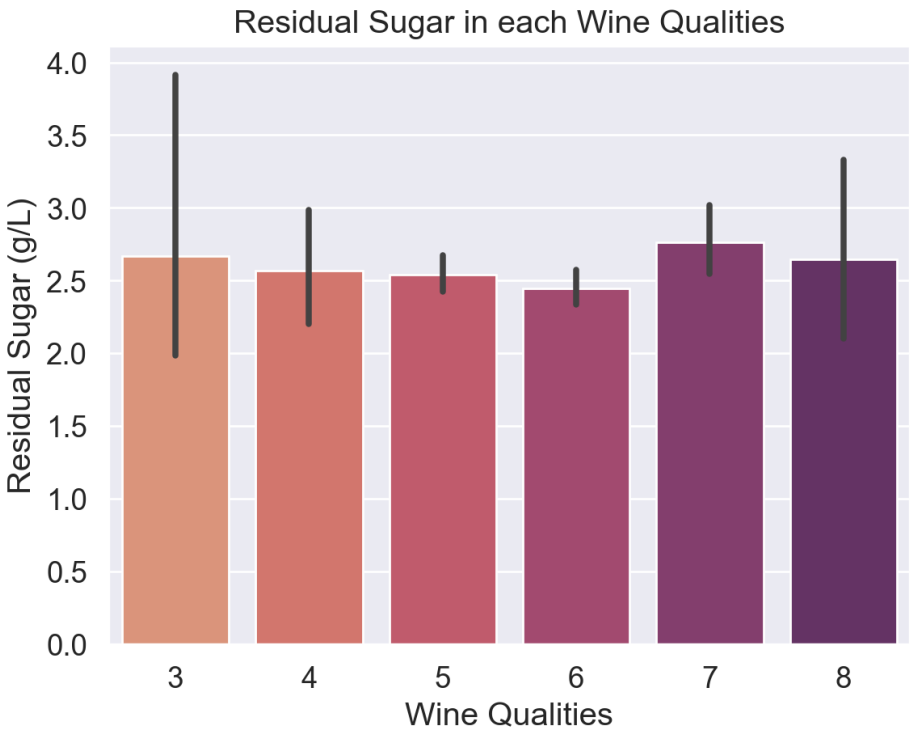

Figure 2.6. Residual Sugar in Wine

Figure 2.7. Barolo, Italy (source: wine-searcher)

3. Off-Dry: 18 to 35 gr/L

4. Medium Sweet: 35 to 120 gr/L

5. Sweet: 120 gr/L or more

Figure 2.6 shows average sugar content in each quality of wine is quite close, about 2.5 g/L which indicates that all wine in this dataset are dry wines.

Generality 2.1

After briefly discussing the features founds in wine, we can conclude that, there are 4 important features that influence the quality number of the wine, namely volatile acidity, citric acid, sulphates, and alcohol.

2.2 Trivia

Barolo is the name of a *comune* (municipality) located in northwest Italy and is the place where Italian red wine is produced, made from Nebbiolo grapes. These grapes are known to have high levels of acidity and high levels of *Tannin* (Rendoni, 2020). *Tannin* is the chemical compound that gives a sharp bitter taste and dry in the throat. Wines from Barolo not only have sharp tannins, but strong acidity and a high alcohol content (14.5 to 15 percent).

The process of making red wine from Nebbiolo grapes takes a long time and is very careful. The grapes, which are picked every October, are put into *stainless steel* or cement barrels and then undergo a fermentation process. After that, the wine is aged for at least 38 months with 18 months in oak barrels, although some wine producers choose to store it longer (VinePair, 2022).

2.3 Reference

- Grainger, K 2021, Wine Faults and Flaws: A Practical Guide, Wiley, Broadway.
- Jacson, R. S 2014, Wine Science: Principle and Applications (4th edition). Academic Press, Amsterdam.
- Bartowsky, E.J., Henschke, P.A. 2, 2004, The 'buttery' attribute of wine-diacetyl-desirability, spoilage and beyond. Int. J. of Food Microbiology
- Zoecklein, B., Fugelsang, K. C., Gump, B. H., Nury, F. S. 1995. Wine Analysis and Production, Springer, New York.
- Puckette, M., 2019, What is Residual Sugar in Wine?, accessed 10 June 2024, <https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/>.
- Kelly, M., Gardner. D., 2022, Volatile Acidity in Wine, The Pennsylvania State University, accessed 12 June 2024, <https://extension.psu.edu/volatile-acidity-in-wine>.
- Rendoni, V., 2020, Nebbiolo in the Nutshell, accessed 20 Aug 2024, <https://winefolly.com/deep-dive/nebbiolo-in-a-nutshell/>
- VinePair Staff, 2022, A Guide to Barolo Wines, accessed 1 September 2024, <https://vinepair.com/wine-101/barolo-wine-guide/>.

Chapter 3 Wine Quality Prediction and Result

After importing and exploring the data, the next step is to perform feature engineering and predict the wine quality using machine learning. Moreover, in this chapter, we will discuss the technical aspects behind the experiment. The type of machine learning used is supervised machine learning with a classification type. The Scikit-Learn library will often be used in this stage.

3.1 Feature Engineering

3.1.1 Label Modification

The Feature Engineering stage is an important stage in machine learning to simplify features and improve model accuracy. Regardless of the data or architecture, bad features make for bad prediction models. The crucial thing before doing feature engineering is to understand the context of the data being processed.

Definition 3.1

Feature engineering is the process of selecting, manipulating, and transforming raw data into usable features.

This step must be carried out because there are several features in this data that need to be adjusted. One of the features in the wine data, **sulphates**, has a content value that is inconsistent with currently produced wines (see Table 1.2). Therefore, this feature needs to be adjusted to some actual information such as books, research papers, or special websites that discuss specific topics.

Because the machine learning used is supervised machine learning in classification, maka kolom yang sebagai label, the **quality** column, will turn into 0 for the quality score below 6.5, and 1 for otherwise. There are several ways to change it, one of which is using Label Encoder. This technique can change categorical data (which usually in string type) to numeric (because machine learning can only process numeric labels), or numeric to numeric. The table 3.1 shows the label changes in the **quality** column.

Quality	Changed into
3	0
4	0
5	0
6	0
7	1
8	1

Table 3.1. Tranformation of the label

3.1.2 Data Separation

Feature-label separation needs to be done so that machine learning can perform calculations (train) and predictions on data. In this experiment, the data splitted into two parts, the *train* data to train the machine learning, and the *test* data to predict.

3.1.3 Feature Scaling

After encoding the label, The next step is to perform *scaling* on the features.If you look at the Table 1.2, these features have different scales, some even have three digits. Hal tersebut bisa mempengaruhi keakuratan model. Therefore, *scaling* needs to be done even though this is not mandatory. In this experiment, the standardization used is *Standard Scaler*.

Definition 3.2

Standard Scaler is a scaling method based on mean data with the formula

$$X_{new} = \frac{X_i - \mu}{\sigma}$$

with X_{new} represents after-scaling features, X_i is the i-th index of the feature, μ is the mean of the feature, and σ is standard deviation.

3.1.4 Import Model

Definition 3.3

Decision tree is a machine learning model that uses an algorithm that forms a tree to create decisions.

Then, we *import* the model that will be used in this experiment, namely the decision tree. This model consists of several parts such as:

- *Root Node*: The main part and becomes the beginning of the formation of the tree structure. This part represents the entire dataset and the initial decision to be made.
- *internal node*: a subsection formed from decisions taken in the root node.
- *leaf node*: the very last part of the decision tree that represents the final decision or prediction.
- *branch*: connecting each node.

The Figure 3.1 illustrates the example of simple application of decision tree in daily life. If you want to buy a car, there are several things you should pay attention to, one of which is the color. Blue is the desired color for the car (root node). If there is (or vice versa), then what year model (or color alternative) is desired (interal node) and so on until deciding whether to buy it or not (leaf node).

The advantages of the Decision Tree model are as follows:

1. This model does not need complicate data preparation.
2. Decision Tree can handle the null data.
3. This model can handle non-linear relationships between features and labels in the data.

Namun, model ini juga memiliki beberapa kekurangan:

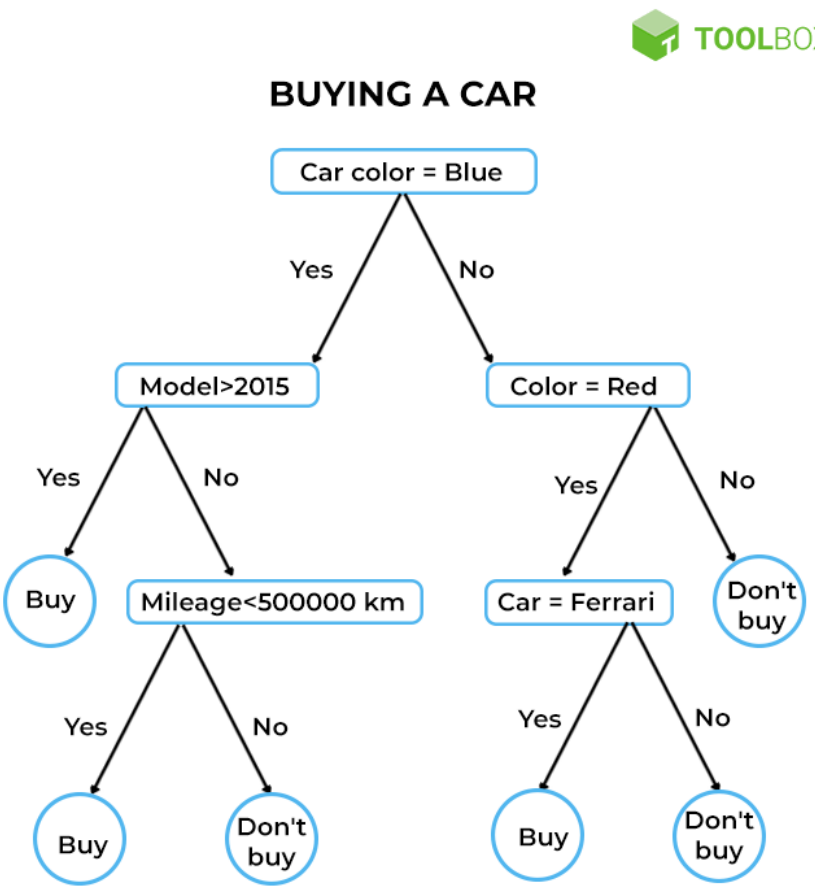

Figure 3.1. Decision Tree when buy a car (Source:Spiceworks)

1. This model has quite high instability, which means that small changes in the data can affect the results or decisions taken.
2. This model is also susceptible to overfitting, which is when a model can produce accurate predictions on training data, but that accuracy dropped when given new data. If the tree gets deeper, the possibility of overfitting will rise.

Activity 3.1

This research will use the Decision Tree and Random Forest models.

The *Decision Tree* model is a machine learning algorithm that forms a tree to determine decisions. If this algorithm is used repeatedly, it will form trees where each tree produces different decisions (outputs). In order to produce a definite decision, the *Random Forest* model collects the results from each of these trees to see the average or majority output. This model has some advantages:

1. Produces better prediction accuracy because it utilizes the accumulation of *decision tree* models that have different data subsets.
2. This model is resistant to noise such as outliers or null value.
3. This model does not require performing data preprocessing first such as data imputation or removing outliers because each decision tree has a random subset of inputs and the *decision tree* itself has the advantage of handling null values.

Of all the advantages mentioned, this model is also not without its disadvantages:

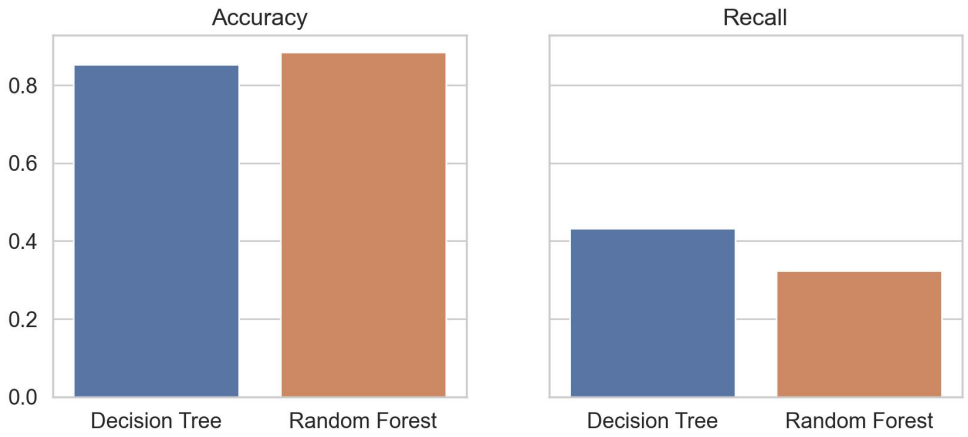

Figure 3.2. Accuracy and Recall of the Model

1. This model uses many trees to make decisions. However, the *random forest* model is a model that requires a large amount of space and quite a long processing time when used for large data sizes.
2. This model integrates multiple outcomes or decisions from the *decision tree* so it is difficult to understand the logic behind each prediction. Therefore, the *random forest* model is also called a "black box" model.

3.2 Result

Definition 3.4

Metrics are quantitative benchmarks used to assess the effectiveness of a *machine learning* model.

In order to know whether the model used is good or not, data scientists use metrics. There are many types of metrics that can be used, such as accuracy, precision, recall, *Area Under Curve*, f1 score, Mean Absolute Error (MEA), Mean Square Error and others.

Activity 3.2

This experiment uses the metrics of accuracy, recall, ROC, and AUC.

Definition 3.5

Accuracy is a metric that measures how precise the predictions produced by a machine learning model are. The mathematical expression that describes accuracy is

$$accuracy = \frac{CorrectPrediction}{AllPrediction} = \frac{TN + TP}{TN + TP + FP + FN}.$$

where TN, TP, FP, FN denotes True Negative, True Positive, False Positive, False Negative, respectively.

True Positive (TP) is the number of **correct** predictions by *machine learning* that correspond to the actual (correct), while True Negative (TN) is the number of **incorrect** predictions by *machine learning* that correspond to the actual (incorrect). For Instance, The wine tester predicted that the wine he was drinking was of good quality, and it turned out to be true (TP), and vice versa (TN). False Positive (FP) is the number of **correct** predictions by *machine learning* that do not correspond to the actual (wrong)

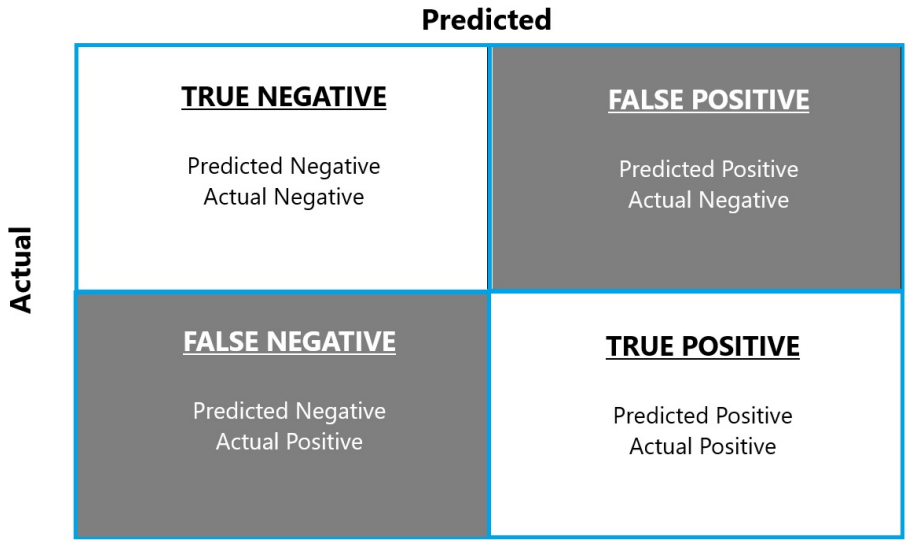

Figure 3.3. Confusion Matrix

ones, while False Negative (FN) represents the number of **incorrect** predictions by *machine learning* that do not correspond to the actual (right) ones. For example, the customer predicted that the wine is the best based on the price they saw, but in fact the wine is not the best quality wine (FP), or vice versa (FN). EThese four terms can be combined in the *confusion matrix* that can be found in Figure 3.3.

In addition to accuracy, recall metrics are also used in this experiment. This metric utilizes all actual positive samples after the machine learning model has been tested.

Definition 3.6

Recall is a benchmark metric for how often a machine learning model correctly guesses a True Positive (TP) from all actual positive samples in the dataset. The equation that describes recall is

$$recall = \frac{TP}{FN + TP}.$$

The accuracy scores of the 2 machine learning models can be seen in Figure **??**. Both models have almost the same accuracy scores. The Random Forest model can differentiate between low and high quality wines accurately, but still lack in terms of how much it correctly guesses high quality wines from the overall high quality wines in the data.

This experiment also utilizes the Receiving-Operating Characteristic Curve (ROC) graph to determine the performance of the classification model used. Receiving-Operating Characteristic Curve (ROC) is a graph that shows the True Positive rate (TP Rate) and the False Positive rate (FP Rate). TP Rate and FP Rate, respectively, can be calculated using the equations

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The performance of a machine learning model is not only assessed from the ROC curve, but also from the Area Under Curve (AUC). The model performance can be considered good if the AUC value is more than 0.5. If the value is less than or equal to this limit, then the model used is not yet able to

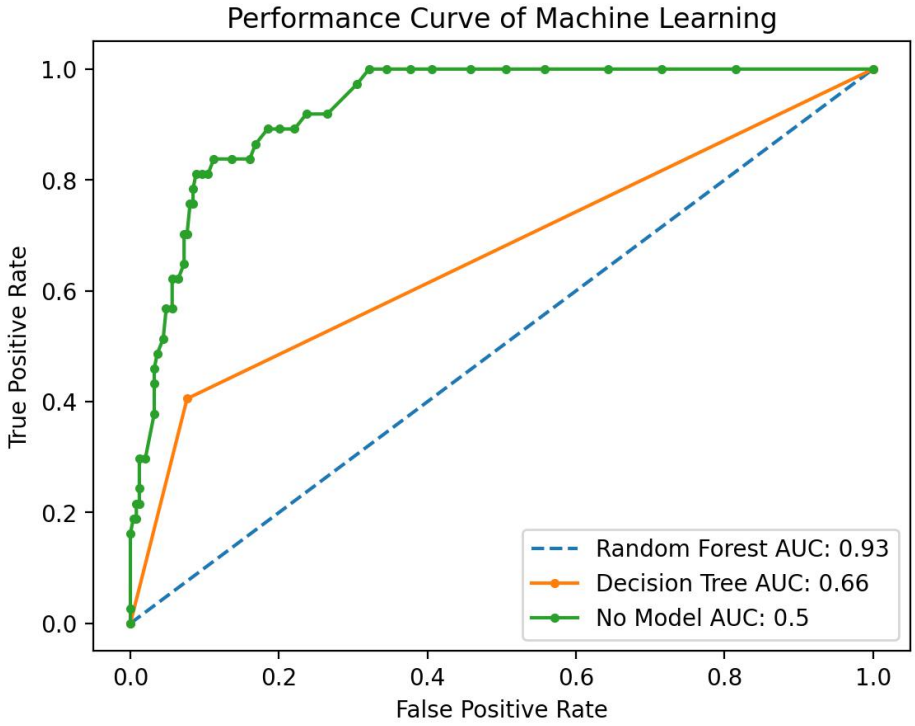

Figure 3.4. Performance Curve of the Models

differentiate the quality of the wine. Based on the curve in the image 3.4, the *Decision Tree Classifier* and *Random Forest Classifier* models show different performances. If the model displays a 90-degree angle at the top left end of the graph, then the AUC score displayed will be higher. In terms of AUC score, *Random Forest* gets a score of around 0.9.

Definition 3.7

Area Under Curve is the area value under the ROC curve.

From several selected benchmarks and without adding any parameters to the two models, it can be concluded that the performance of the *Random Forest* model is better than the *Decision Tree* model.

3.3 Reference

- Hillier, W., 2023, What is a Decision Tree and How is used?, accessed 3 September 2024, <https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/>.
- GeeksforGeeks team., 2024, Decision Tree, accessed 15 July 2024, <https://www.geeksforgeeks.org/decision-tree/>.
- EvidentlyAI team, 2024, Accuracy vs Precision vs Recall, accessed 3 September 2024, <https://www.evidentlyai.com/classification-metrics/>.
- IBM team, 2022, What is a Decision Tree?, IBM, accessed 15 July 2024, <https://www.ibm.com/topics/decision-trees>.
- Anggreany, M.S., Confusion Matrix, BINUS University, accessed 29 August 2024, <https://socs.binus.ac.id/2020/11/01/confusion-matrix/>.
- Brownlee, J., 2020, How to use StandardScaler and MinMaxScaler Transforms in Python, accessed 27 Aug 2024,

<https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/>.

- Narkhede, S, 2018, Understanding AUC-ROC Curve, accessed 29 August 2024, <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>.
- GeeksforGeeks Team, 2024, AUC ROC Curve in Machine Learning, accessed 29 August 2024, <https://www.geeksforgeeks.org/auc-roc-curve/>.
- Google Developers Team, 2024, Classification: ROC and AUC, Google. Inc, accessed 28 August 2024, <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>.

Chapter 4 Recommendation and Limitations

This chapter contains several recommendations from two perspectives, producers and consumers.

4.1 Recommendation

4.1.1 Producers

Producers are parties who produce wine, such as factories or home industries. After conducting data analysis, the recommendations that can be given to increase profits are as follows:

- Consumers buy low-quality wine because it is affordable and available in many retailers or supermarkets. In addition, consumers do not only buy wine to drink, but also to use as a cooking ingredient. For these reasons, **low quality wine production is the best option**. In addition, low quality wine does not need to pay attention to the four important compositions discussed previously. The most important aspect of wine that needs to be considered in production is simply the shelf life of the wine. The element *Sulphates* in wine plays an important role in this.

4.1.2 Consumers

Consumers are the ones who utilize wine. Based on the data that has been analyzed, the recommendations that can be given to consumers are as follows:

- If consumers want to use wine as a cooking ingredient (a substitute for vinegar), it is recommended to choose wine that has a high volatile acid content.
- If consumers want to try wine for the first time, it is recommended to choose a low alcohol content so as not to get drunk when drinking it because each person has different alcohol tolerance.
- If the consumer is a collector of the best quality wines, it is advisable to look for wines that have low levels of volatile acids.

4.2 Limitations

There are some outliers that are still not handled, even after making predictions using machine learning, because the dataset size is too small. If these outliers are handled, there will be a significant deviation in the accuracy of the machine learning model predictions. Moreover, this dataset is a small sample that does not represent wines around the world. Machine learning yang digunakan