



Modeling Red Wine Quality Based on Physicochemical Tests: A Data Mining Approach

Maeve Eunicia^{1*}, Richie Skyszygfrid², Tiara Vitri³, Vicky Caren⁴
Hospitality Management Major, Pelita Harapan University Medan, Indonesia

ABSTRACT : Classification of the quality of red wine is done in the hope of making it easier to assess the quality of red wine. Data used for this research is the wine quality data set with 4898 number of instances, obtained from UCI machine learning repository. Classification of the quality of red wine this study was carried out by comparing the three algorithms of data mining, that is random forest, naive bayes and generalized linear model. From the results of this study comparing the three algorithms, the generalized linear model showed the highest accuracy among the other algorithms. It was tested with a generalized linear model with 68.75% accuracy results, this algorithm is ideal for classifying the quality of red wine. In addition, a secondary random forest gives 67.81% accuracy results, while Naive Bayes gives 61.25% accuracy results. Studies conducted to classify the quality of red wine based on its composition use a generalized linear model for the optimal algorithm.

Keywords: wine, random forest, naive bayes, generalized linear model, data mining, descriptive, predictive, prescriptive analysis

Submitted: 06-05-2022; Revised: 15-05-2022; Accepted: 26-05-2022

INTRODUCTION

Wine has been enjoyed in many households and populations all over the world. It is considered as a luxury good as it is often expensive and mostly enjoyed by the upper class population, however it has gradually expanded its reach to a wide range of customers from every population (Cortez et al., 2009). Wine is generally a complex drink to evaluate. Only wine connoisseurs and wine experts can truly know a wine's quality or what notes of flavors the wine possesses. First drinkers would tend to say that they can only taste bitterness when sipping on their first glass of wine, not conscious of what qualities the wine actually has. Therefore it is relatively complex to determine the quality of wine, and therefore requires exemplary knowledge, experience and taste buds to determine its quality.

This proves to be useful for the growth of the wine industry. The wine industry strives to improve the quality of its wine products and so especially due to the complexity of determining the quality of wine and the formula that comes behind the best quality of wine, lots of scientific tests need to be made to discover which variables and how much of the variables will yield the best outcome for the quality of wine. Using these tests and the data we get from them, we can conduct data mining and predict the outcome of the quality of the wine and also look into the requirements that are needed to obtain a specific quality of wine. The wine quality assessment usually uses physicochemical tests along with sensory tests (tasting). Physicochemical tests involve testing the alcohol, chlorides, pH, density levels of the wine among many others, and sensory tests involve tasting of the wine and evaluating its quality which comes with rating it from 0 (very bad) to 10 (excellent) that are conducted by wine experts (Cortez et al., 2009).

The purpose of this journal is to use data mining to determine the quality of red wine in a psychochemical test. The data mining methods that we use are random forest, naive bayes, and generalized linear models. With that, we compare the 3 algorithms and get which method has the highest accuracy among the three algorithms.

LITERATURE REVIEW

Introduction to Wine

Wine is an alcoholic beverage that is usually made from grapes that are fermented. The main factor in some of the differences in the taste of wine is the different types of yeast and wine. There are also some types of wine that are not made with fermented grapes but rather fermented other plants such as rice wine and other fruit wines such as plums, cherries, pomegranates, raisins, and elderberries.

Many terms are used to describe wine and its culture. One of them is the chateau, we must be familiar with this term. Chateau is a French term for winery, it's a common term found in many French wine bottles, it refers to a winemaker. Wine has been produced for thousands of years. The earliest evidence of wine is from ancient China, Persia, and Italy (Li, 2018). Today, the

five countries with the largest wine-producing regions are Italy, Spain, France, the United States, and China(J., 2007).

Wine Quality

Wine consumption has increased rapidly in recent years not only for recreational purposes, but also for its beneficial effects on the human heart. Today, all industries are applying new technologies, implementing new methods, maximizing production and streamlining the entire process. These processes become more and more expensive over time, and their requirements also increase. Although they have a different purpose than wine, they use almost the same chemicals, but they need to be evaluated for the type of chemicals they use, so we apply these methods to verification. Wine quality can be measured using either physicochemical tests or sensory tests. The former test can be performed without human intervention, while the latter can be performed under the supervision of a human expert. Producers use this model's predictions to improve wine quality, and certification bodies to have a better understanding of the essentials of quality and give consumers the opportunity to make purchasing decisions(et al., 2021).

Data Mining

Data mining is then used to process these datasets obtained from the tests to produce trendlines, visualizations of the data for us to observe and analyze. It in turn also optimizes the datasets to produce predictions of outcomes, which we can compare to the real outcome to analyze the model's accuracy. Furthermore, it provides us with the data requirements that are needed for us to obtain a certain specific outcome. There are a lot of data mining algorithms which we can use to process these datasets, and each has their own uses and advantages as well as disadvantages, such as,

1. Random Forest

Random Forest is an ensemble monitored machine learning technique. Machine learning techniques have been in the area of Data mining application. Random forest It has great potential to become a popular technique Because its performance was for future classifiers Turned out to be comparable to the ensemble technique Sag and boost. Therefore, a detailed study of Existing work related to Random Forest contributes to this Accelerate machine learning research (Kullarni& Sinha, 2013).

2. Generalized Linear Model

Generalized Linear Model has correlation with linear regression model. It assumes that the conditional expectation Y is equal to a linear combination X and that Y is a member of the exponential family. Generalized linear model is determined by the distribution of Y and the link function(Gentle et al., 2012).

3. Naive Bayes

Naïve Bayes is a simple data classification algorithm. This algorithm calculates a set of probabilities by calculating the frequency and combination of values in the given data. The probability of a particular feature in the data appearing as an individual of the probability sequence is obtained by

calculating the frequency of each feature value in the class from the training data set.(Wibawa et al., 2019).

Wine Tasting

Wine tasting is a sensual inspection and evaluation of wine. Wine contains many compounds similar to or identical to those found in fruits, vegetables and spices. The sweetness of wine is related to the acidity present in the wine and is determined by the amount of sugar remaining in the fermented wine. For example, dry wine has very little sugar left. Some wine labels recommend opening the bottle and "breathing" the wine for a few hours before serving, while others recommend drinking immediately. Decantation (the process of pouring wine into a special container just to breathe) has been controversial among wine lovers. In addition to ventilation, decanting with a filter can remove the bitterness that has accumulated in the wine. Precipitates are more common in old bottles, but aeration can benefit young wines(HUGH JOHNSON and JANCIS ROBINSON, 2013).

As we know there are many types of wine and usually wine is always paired with food. The most common is red wine with steak. Wines have different phenolic compounds, and different wines have great variability in these compounds. For example, resveratrol is found primarily in white wines as well as red wines, but its antioxidant properties have many health benefits. Wine is often the ingredient when marinating meat. Marinated meat is thought to enhance the flavor and the alcohol and acidity of the wine can soften the meat(Blackhurst et al., 2011).Alcohol fermentation involves continuous solid-liquid extraction from grape skins and seeds. Therefore, the fermentation temperature has a great influence on the final composition of the resulting wine(Ariel Massera, Mariela Assof, Santiago Sari, Ivan Ciklis, Laura Mercado, Viviana Jofre, 2021).

Wine Consumption

According to the latest World Health Organization data, wine consumption data from a list of countries by alcohol consumption, measured in liters of pure ethyl alcohol consumed per person in a particular year. The methodology includes people over the age of 15(World Health Organization, n.d.). Approximately 40% of people over the legal drinking age consider themselves "wine drinkers". This is higher than all other alcoholic beverages combined (34%) and those who do not drink at all (26%) (Thach & Feb, 2020).

Health Effects of Wine

1.Short Term

Wine contains ethyl alcohol, an intoxicating chemical found in beer and spirits. Different levels of alcohol in the human body have different effects on humans. The effectiveness of wine depends, among other things, on the amount consumed, the duration of consumption, the alcohol content of the wine, and the amount of food consumed. Drinking enough to have a blood alcohol level (BAC) of 0.03% to 0.12% usually improves overall mood, increases self-confidence and sociability, reduces anxiety, causes hot flushes, and makes

judgments. Power and motor coordination are reduced. Use 0.09% to 0.25 BAC-lethargy, sedation, imbalance, vision impairment. Use 0.18% to 0.30 BAC-deep confusion, speech disorders (eg, obscure speech), stagger, dizziness, and vomiting. BAC of 0.25% to 0.40 causes drowsiness, unconsciousness, anterograde amnesia, vomiting and can die from respiratory depression and vomiting inhalation during unconsciousness. A BAC of 0.35% to 0.80 causes coma, life-threatening respiratory depression, and potentially fatal alcoholism. Driving a car or machine drunk increases the risk of an accident, and many countries have enacted drunk driving laws. Wine can quickly cause positive emotions such as Relaxation and comfort. The context and quality of the wine can also affect mood and emotions(Danner et al., 2016).

2.Long Term

The main active ingredient of wine is alcohol, and therefore, the health effects of alcohol apply to wine. The World Health Organization (WHO) Alcohol Health Goal is to reduce the health burden of harmful alcohol consumption, thereby saving lives, reducing illness and preventing injuries. Dangerous and harmful alcohol consumption is a global major cause of death, illness and injury. For drinkers through health effects such as alcoholism, cirrhosis, cancer and injuries. For others, such as drunk driving and dangerous acts such as violence, or through the effects of drinking on the development of the fetus and children. Harmful alcohol use causes about 2.5 million deaths and 2.25 million lives each year, taking into account the estimated beneficial effects of low alcohol intake on some diseases in some population groups. Harmful drinking can also be very costly to the community and society. Alcohol intake is estimated to cause 20% to 50% of cirrhosis, epilepsy, poisoning, road accidents, violence and various types of cancer. Major alcohol-related diseases and injury categories: cancer, cardiovascular disease, fetal alcohol syndrome and premature infant complications, diabetes(World Health Organization, n.d.).

Storage

The effect of temperature on the chemical composition of red wine must be important. This is due to improper storage. It shortens shelf life, reduces the sensual quality of wine and can have a significant impact on the aging reaction(Mattivi et al., 2015)(Butzke C. E.Vogt E. E.Chacón-Rodríguez L., 2012). The ideal storage temperature for wines is between 15-20°C. In fact, this temperature is not always used during the storage or transportation of wine, as it is exposed to higher fluctuating temperatures(Robinson et al., 2010). The humidity of the environment in which the wine is stored also plays a very important role in the quality of the wine. Therefore, low humidity can cause the cork to dry and deform. When the cork shrinks, cracks, or loosens, excess air enters the bottle and comes into contact with the wine to accelerate the oxygen changes. On the other hand, high humidity (80% or higher) increases the risk of mold on the cork. In addition, wine should not be exposed to excessive light that can cause certain unwanted odors. In addition, the wine cellar should be designed to control all parameters that affect aging(Chung et al., 2008).

DATA AND METHODOLOGY

Table 1. Data Attribute Descriptions

Attribute	Description	Unit	Role
Fixed acidity	Levels of tartaric acid in wine	g/dm ³	Input variables
Volatile acidity	Levels of acetic acid in wine	g/dm ³	Input variables
Citric acid	Citric acid levels in wine	g/dm ³	Input variables
Residual sugar	Levels of residual sugar in wine	g/dm ³	Input variables
Chlorides	Levels of sodium chloride in wine	g/dm ³	Input variables
Free sulfur dioxide	Levels of free sulfur dioxide in wine	mg/dm ³	Input variables
Total sulfur dioxide	Portion of free sulfur dioxide plus the portion that is bound to other chemicals in the wine such as aldehydes, pigments, or sugars	mg/dm ³	Input variables
Density	Density of wine	g/cm ³	Input variables
pH	pH levels of wine	Range: 0 - 14	Input variables
Sulphates	Levels of potassium sulphate in wine	g/dm ³	Input variables
Alcohol	Alcohol concentration in wine	% vol.	Input variables
Quality	Wine quality (target attribute)	Range: 0 (very bad) to 10 (excellent)	Label

Note. Data are from UCI machine learning repository wine quality data set and Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez>

A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

METHODOLOGY

Descriptive analytics is performed if the research goal is involved in improving the application of a leader's knowledge, understanding and survey. The phrase for research is extremely extreme. The exact word of the goal is related to research design and statistics. Qualitative or quantitative. Qualitative variables are classified and characterized and Z. Gender, socio-economic position, pain plain, treatment group, etc. Quantitative variables, on the other hand, are measurable, continuous, and numerical. Age, height, weight, pain score, etc(Hussain, 2014).

Predictive analytics is a term used primarily in statistical and analytical techniques. Predictive analytics is used to predict future events. Analyze current and historical data and use statistical, data mining, machine learning, and artificial intelligence techniques to predict the future(Chauhan & Kaur, 2015). Integrate information technology, business modeling processes, and management to make predictions about the future. By successfully applying predictive analytics, companies can properly use big data for profit. This helps organizations become proactive, think positively, and predict trends and behaviors based on data. The term comes from statistics, machine learning, database technology, and optimization technology. It has its roots in classic statistics. Predictive analytics models can be used to predict the behavior of future events and variables. Scores are primarily assigned by predictive analytics models. The higher the score, the more likely it is that an event will occur, and the lower the score, the less likely it is that an event will occur(Kumar & L., 2018).

Prescriptive analytics is a set of mathematical techniques to determine complex targets, requirements, and limitations to improve business outcomes. This approach determines various alternatives and guides based on results drawn from descriptive analytics and predictive analytics.(Gim et al., 2018)

Data Mining

Random Forest

Random Forest is an effective prediction tool. For the great law The number they do not over fit. When you inject the right kind of randomness, They are accurate classifiers and regressions. Moreover, The strength of individual predictors and their correlations provide information about Random forest predictive power. It embodies the theoretical value of intensity and correlation. Forests produce results that compete with boosting and adaptive bagging, but they don't gradually change the training set. Their accuracy shows that they are taking action to reduce prejudice. In Breiman's experiment with Random Forest, bagging was Random feature selection.

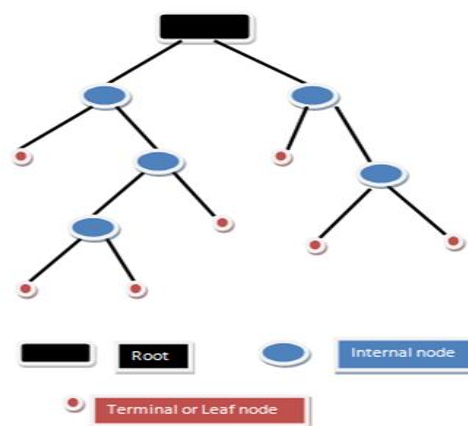


Fig 1. Decision Tree

Each new training set is drawn and replaced. From the original training set. Then grow the tree using the new training set Random feature selection. Growing trees are not pruned. There are two reasons to use a pouch. The first is the use of bagging. Using the random feature seems to improve accuracy. The second is this You can use bagging to give a running estimate of generalization error (PE *). In addition to the ensemble of wood combinations, strength and correlation. Tree Forest cannot break into a simple interpretation of its mechanism walk(Breiman, 2016). The grouping of data in the decision tree is based on the value of the specified data attribute. Decision trees are created from pre-classified data. The division into classes is determined by the characteristics that optimally divide the data. Data elements are split according to their value feature. This process is applied recursively to each shared subset of data items. If all the data items in the current subset belong to the same class, the process will terminate(Ali et al., 2012).

Naive Bayes

The naive Bayes algorithm is a simple probabilistic algorithm that calculates a probabilistic group by calculating the combination and frequency of a data. by using the bayes theorem, the algorithm assumes all of the attributes to be independent even though there is class variable value(Yasar & Saritas, 2019). The simplicity of naive bayes makes an efficient methodology, which is

Naive Bayes techniques attractive and suitable for many domains. However, for some researchers, they see this methodology as a weak approach despite its popularity, simplicity, and ease of use. because the Naive Bayes classification has a drawback. Naive Bayes is based on the assumption of independence attributes, but in most cases, this assumption does not match or correspond to reality. Naive Bayes makes redundant, irrelevant, interactive and noise-contaminated features equal status with other important features which reduces accuracy of the classification.(Chen et al., 2021).

Although a system based on Naive Bayes is not able to use two or more pieces of evidence at once, because the independence assumptions are very strong, but used in the appropriate domain, it will result in fast training, fast data analysis and decision making, and direct interpretation of results.(Xhemali et al., 2009).

Naive bayes theorem :

$$P(H | X) = P(X | H) \cdot P(H) / P(X)$$

X : Data with unknown class

H : Hypothesis data is a specific class

P(H | X) : Probability of hypothesis H based on condition X (posteriori probability)

P(H) : Hypothesis probability H (prior probability)

P(X | H) : Probability of X based on condition on hypothesis H

P(X) : Probability X

In explaining the Naive Bayes theorem, the classification process requires a number of instructions in determining which class is suitable for the analyzed sample (Syaputri&Irwandi, 2020).

$$P(H | X) = P(X | H) \cdot P(H) / P(X)$$

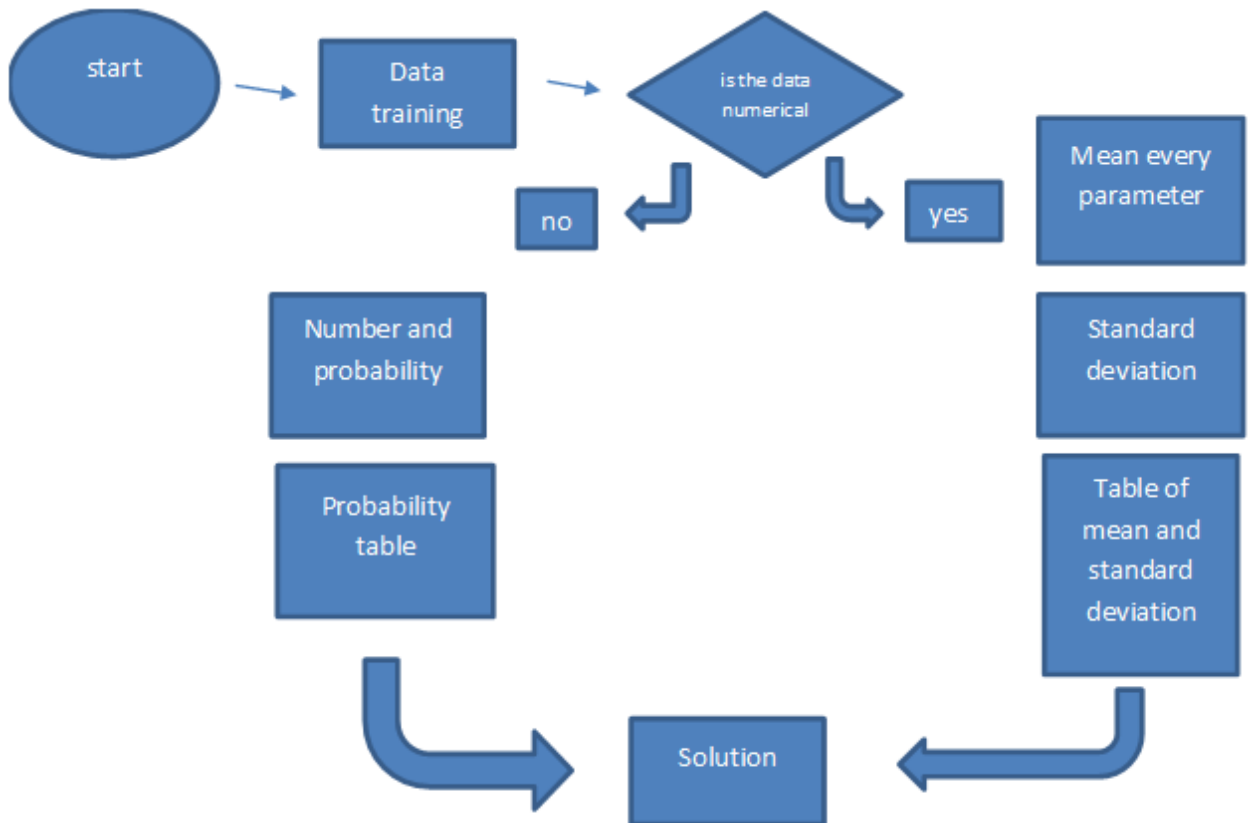


Fig 2. Flow chart of naïve bayes

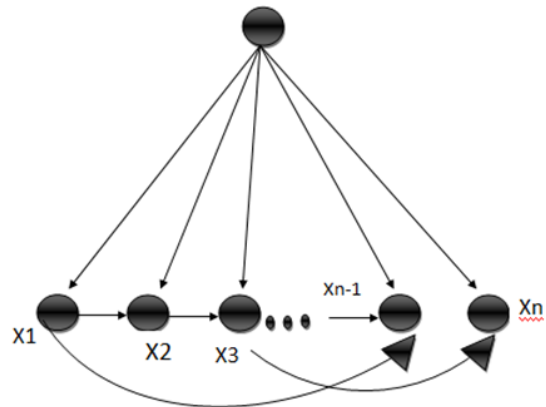


Fig 3. Example of Tree Augmented Naive Bayes

Generalized Linear Model

Generalized linear models extend the concept of the well understood linear regression model. The linear model assumes that the conditional expectation of Y (the dependant or response variable) is equal to a linear combination X , i.e.

The generalized linear model is determined by 2 components:

- the distribution of Y .
- the link function.

In order to define the generalized linear model methodology as a specific class of nonlinear models, we assume that the distribution of Y is a member of the exponential family(Gentle et al., 2012).

As seen in the data mining sequence process in Figure 4 below, we start by feeding the red wine quality dataset into the system. The next step that we do is to discretize the quality attribute data into several classes, with the details as shown below in Table 2. The classes are an expansion of the supposed range of the label attribute, which is 0(very bad) - 10(very good). This will then transform the numerical quality attributes to nominal.

Table 2. Discretize (Discretize by User Specification) Parameter List: classes

class names	upper limit
very bad	0.9
Bad	2.9
below average	4.9
Average	5.9
above average	7.9
Good	8.9
very good	10.0

After discretizing the nominal quality attributes to classes, we set the quality attribute as our label role using the set role operator, then we use stratified sampling to build random subsets and ensure that the class distribution in the subsets is the same as in the whole dataset. In this case, we take a relative sample of 0.2 or 20% of the whole dataset. The next step is to perform a dimensionality reduction method by using the weight by information gain operator to weigh and select the input variables by the level of information gain. Afterwards using cross validation we divide the dataset into 5 subsets of equal size, then retain a single subset as the test data set or as the input of the training subprocess and the remaining 4 subsets are used as training data set or input of the training subprocess. The validation process is repeated 5 times with each of the 5 subsets used exactly once as the test data. The 5 results from the 5 iterations are then averaged to produce a single estimation. This is done to avoid overfitting.

We apply each of the 3 algorithms we are comparing in the training subprocess. The application of the model and the performance of the model is

measured during the testing phase. After all this is done, we use the filter example range operator to set the example range we want to predict, selecting data number 5 as the first and the last example. We then select the most appropriate and important parameters of each algorithm in the optimize parameters operator, then finally insert the prescriptive analytics operator to acquire the optimal inputs that we need for our desired outcome.

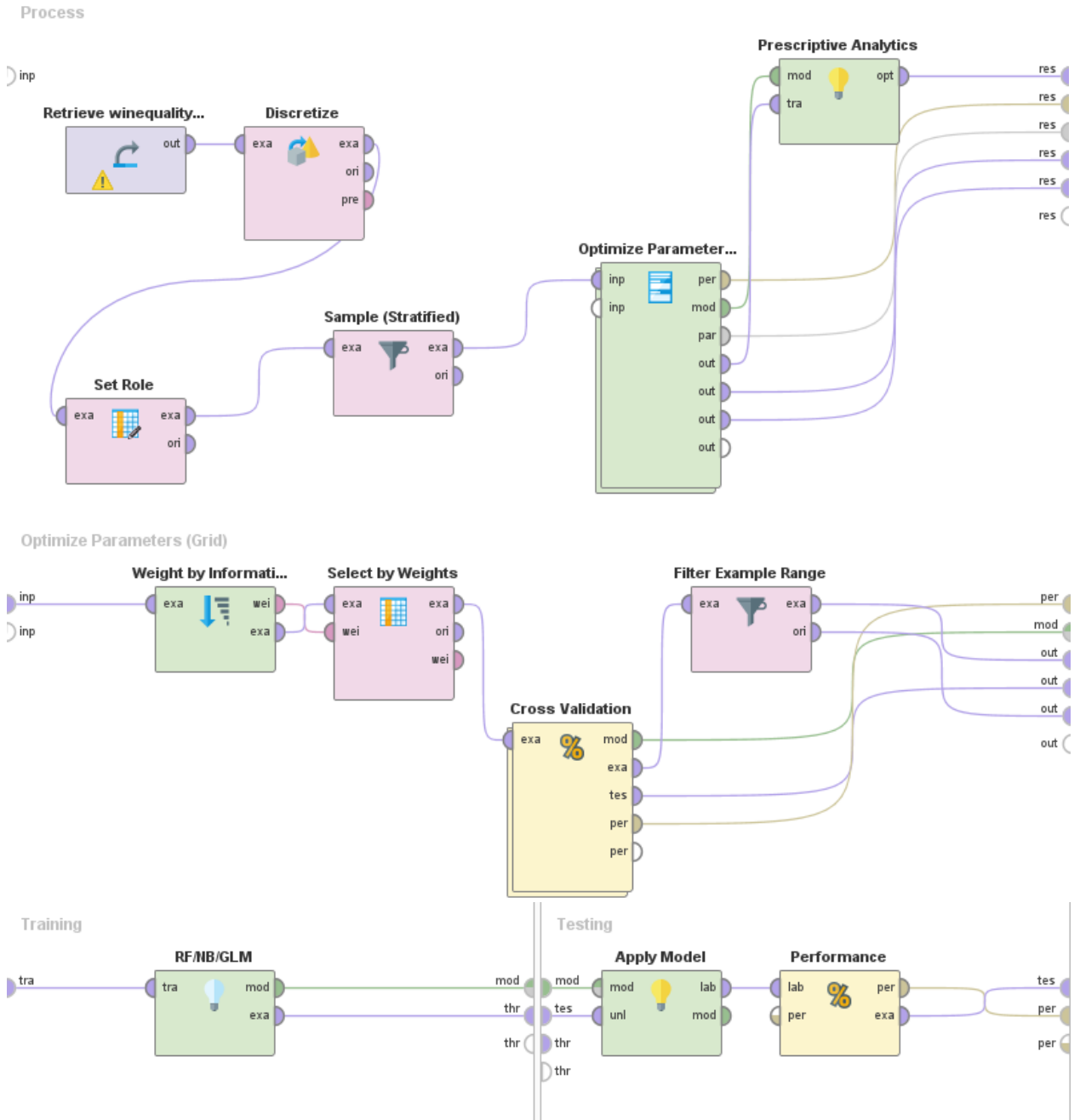
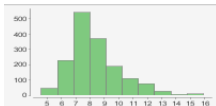
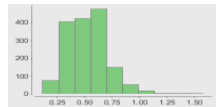
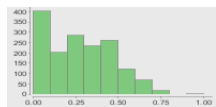
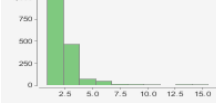
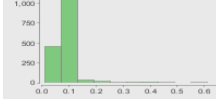
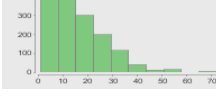




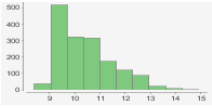
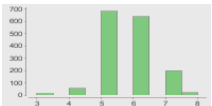


Fig. 4 Data Mining Sequence using Rapidminer

DISCUSSION AND RESULTS

Table 3. The Physicochemical Data Statistics for Red Wine

No	Attributes (units)	Data Type	Visualization	Min	Max	Mean	Deviation
1	fixed acidity (g/dm ³)	Real		4.600	15.900	8.320	1.741
2	volatile acidity (g/dm ³)	Real		0.120	1.580	.0528	0.179
3	citric acid (g/dm ³)	Real		0	1	0.271	0.195
4	residual sugar (g/dm ³)	Real		0.900	15.500	2.539	1.410
5	chlorides (g/dm ³)	Real		0.012	0.611	0.087	0.047
6	free sulfur dioxide (mg/dm ³)	Integer		1	72	15,876	10.462
7	total sulfur dioxide (mg/dm ³)	Integer		6	289	46.468	32.896
8	density (g/dm ³)	Real		0.990	1.004	0.997	0.002
9	pH (range 0-14)	Real		2.740	4.010	3.311	0.154
10	sulphates (g/dm ³)	Real		0.330	2	0.658	0.170

11	alcohol (%vol)	Real		8.400	14.900	10.423	1.066
12	quality	Integer		3	8	5,636	0.808

Predictive Analysis

1. Random Forest Algorithm

We first analyze the results of the model process using a random forest algorithm. In this model, we optimized the number of trees parameter of random forest. By using the optimal parameters, the model obtained 67.8% accuracy with a standard deviation of 5.91%. We can see from Table 3 below that the algorithm had to do 11 iterations before achieving the specified level of accuracy.

Table 4. Grid Search Results of Random Forest

iteration	Random Forest.number_of_trees	Accuracy
1	1	0.569
2	11	0.616
3	21	0.641
4	31	0.675
6	51	0.656
5	41	0.628
7	60	0.647
9	80	0.672
10	90	0.647
8	70	0.644
11	100	0.678

We analyze the model performance by looking at Table 4 below. There are 217 out of 320 samples tested that are precise predictions, which is 67.8% out of total samples. The class above average has the most precise predictions with 142 predicted as above average correctly out of the 167 supposed true above average class samples. This implies that the model is robust in predicting the above average level of quality of wine, with its quality rating in numerical values ranging from 6 - 7.9. We cannot actually say that it isn't robust enough to predict the highest wine quality class that is 'very good' as the mistake lies on the dataset sample itself, not possessing any quality value that is classified as 'very good' therefore we cannot predict the 'very good' quality level of red wine.

Table 5. The confusion matrix of Random Forest

	true very bad	true bad	true below average	true average	true above average	true good	true very good	class precision
pred. very bad	0	0	0	0	0	0	0	0.00%
pred. bad	0	0	0	0	0	0	0	0.00%
pred. below average	0	0	5	3	0	0	0	62.50%
pred. average	0	0	5	70	25	0	0	70.00%
pred. above average	0	0	3	63	142	4	0	66.98%
pred. good	0	0	0	0	0	0	0	0.00%
pred. very good	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	38.46%	51.47%	85.03%	0.00%	0.00%	

2. Naive Bayes Algorithm

Next, using the Naive Bayes algorithm, we optimize the laplace correction parameter of Naive Bayes. By using the optimal parameter, we obtained an accuracy of 61.25% or rounded up into 61.3% with a standard deviation of 6.57%. This is shown in Table 5 below, where there are two iterations namely true and false, implying the true predictions and the false predictions.

Table 6. Grid Search Results of Naive Bayes

iteration	Naïve Bayes.laplace_correction	Accuracy
1	true	0.613
2	false	0.606

We then also analyze the model performance by looking at Table 6 below. There are 196 out of 320 precise predictions, which is lower than that predicted using the Random Forest algorithm. The same thing happens here where the most correctly predicted class is the above average class, with 121 precise predictions out of 164 supposed true above averages. This portion is also less than that of the Random Forest model performance, concluding that the Naive Bayes model is less robust and less precise than the Random Forest model.

Table 7. The confusion matrix of Naïve Bayes

	true very bad	true bad	true below average	true average	true above average	true good	true very good	class precision
pred. very bad	0	0	0	0	0	0	0	0.00%
pred. bad	0	0	0	0	0	0	0	0.00%
pred. below average	0	0	8	16	12	0	0	22.22%
pred. average	0	0	3	67	31	0	0	66.34%
pred. above average	0	0	1	52	121	4	0	67.98%
pred. good	0	0	1	1	3	0	0	0.00%
pred. very good	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	61.54%	49.26%	72.46%	0.00%	0.00%	

3. Generalized Linear Model Algorithm

Finally, using the Generalized Linear Model algorithm, we optimize the family parameter of AUTO or multinomial to fit the algorithm to our nominal label attribute. As a result, we obtained an accuracy of 68.75% with a standard deviation of 4.84%, the highest we've observed among the three algorithms that we have tested. We can also find in Table 8 below that 220 out of 320 are predicted precisely which is 68.75% of the total sample, and the highest precision of prediction lies on the prediction of the above average quality class, with 126 correct predictions out of a total of 167 supposed samples of above average.

Table 8. The confusion matrix of Generalized Linear Model

	true very bad	true bad	true below average	true average	true above average	true good	true very good	class precision
pred. very bad	0	0	0	0	0	0	0	0.00%
pred. bad	0	0	0	0	0	0	0	0.00%
pred. below average	0	0	5	3	2	0	0	50.00%
pred. average	0	0	5	89	39	0	0	66.92%
pred. above average	0	0	3	44	126	4	0	71.19%
pred. good	0	0	0	0	0	0	0	0.00%
pred. very good	0	0	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	38.46%	65.44%	75.45%	0.00%	0.00%	

4. Algorithm Comparisons

In Table 9 we provide an overview of the accuracy, time processing and class precision of every algorithm we have tested, and we can analyze that the longer the processing time is, the more accurate the model is, and we can observe that the Generalized Linear Model algorithm has the highest accuracy as well as the longest processing time. It also produces the most favorable results with the highest precision percentage of 71.19% for the 'above average' quality class, our highest quality rating in our tests so far since the sample data of classes 'good' and 'very good' are very scarce.

Table 9. Algorithm Comparisons

Algorithms	Accuracy	Time Processing	Precision (pred. very bad)	Precision (pred. bad)	Precision (pred. below average)
Random Forest	67.81%	6 seconds	0.00%	0.00%	62.50%
Naïve Bayes	61.25%	1 second	0.00%	0.00%	22.22%
GLM	68.75%	9 seconds	0.00%	0.00%	50.00%

Precision (pred. average)	Precision (pred. above average)	Precision (pred. good)	Precision (pred. very good)
70.00%	66.98%	0.00%	0.00%
66.34%	67.98%	0.00%	0.00%
66.92%	71.19%	0.00%	0.00%

Prescriptive Analysis

Table 10. Prescriptive Analytics using Random Forest

row no	prediction	Conf(very bad)	Conf(bad)	Conf(below average)	Conf(average)	Conf(above average)	Conf(good)	Conf(very good)	chlorides
1	below average	0	0	0.882	0.098	0.021	0.000	0	0.080

alcohol	density	free sulfur	fixed acidity	citric acid	pH	total sulfur)	volatile acidity	residual sugar	sulfates
---------	---------	-------------	---------------	-------------	----	---------------	------------------	----------------	----------

							y)		
9.400	0.997	14	7.100	0	3.470	35	0.710	1.900	0.550

Table 11. Prescriptive Analytics using Naive Bayes

row no	prediction	Conf(very bad)	Conf(bad)	Conf(below average)	Conf(average)	Conf(above average)	Conf(good)	Conf(very good)	chlorides
1	average	0	0	0.016	0.495	0.487	0.002	0	0.080

alcohol	density	free sulfur	fixed acidity	citric acid	pH	total sulfur)	volatile acidity)	residual sugar	sulphates
9.400	0.997	14	7.100	0	3.470	35	0.710	1.900	0.550

Table 12. Prescriptive Analytics using Generalized Linear Model

row no	prediction	Conf(very bad)	Conf(bad)	Conf(below average)	Conf(average)	Conf(above average)	Conf(good)	Conf(very good)	chlorides
1	average	0	0	0.159	0.637	0.204	0.000	0	0.080

alcohol	density	free sulfur	fixed acidity	citric acid	pH	total sulfur)	volatile acidity)	residual sugar	sulphates
9.400	0.997	14	7.100	0	3.470	35	0.710	1.900	0.550

Hence, the requirements needed to obtain sample number 5's red wine quality of random forest (below average), naive bayes (average) and GLM (average). We can conclude that there are several contents needed in the use of the three methods, which is chlorides, alcohol, density, free sulfur, fixed acidity, pH level, total sulfur dioxide, volatile acidity, residual sugar, and sulfate. the amount of chloride needed is 0.080g/dm, alcohol is 9.400%vol, density at 0.977g/dm³, Free sulfur dioxide as much as 14 mg/dm³, Fixed acidity at 7.100g/dm³, No citric acid, pH level at 3.470, Total sulfur dioxide at 35 mg/dm³, Volatile acidity at 0.710g/dm³, Residual sugar as much as 1.900g/dm³, Sulfates as much as 0.550g/dm³.

CONCLUSION

This study is about a data mining approach to predict the quality of the red wine base on physicochemical tests. The quality is described in numerical ranges from 0(very bad) - 10(very good), in which we discretize it to 7 classes to transform the numerical quality attributes to nominal. We used three algorithms of data mining on predicting the quality of the red wine, such as random forest, generalized linear model, and naives bayes. In finding the accuracy of random forest, we do 11 iterations to reach a specified level of accuracy, which is 67.8% of accuracy (217 wine quality is predicted true out of 320). The second one is naïve bayes, in naïve bayes we got 61.3% accuracy (196 wine quality is predicted true out of 320). The algorithm with the highest accuracy is on a generalized linear model with 68.75% accuracy (220 wine quality is predicted true out of 320). In using this algorithm, almost all the average wine quality is within the average range, with 13 wines below average, 136 wines on average, and 167 wines above average. Wine that has good quality is only 4 wines. Therefore, the total wine analyzed is 320 wines. The highest percentage of predictions is in wine that have above average quality in the generalized linear model algorithm which has an accuracy of 71.19% (126 true out of 177).

REFERENCES

- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *IJCSI International Journal of Computer Science Issues*, 9(5), 272–278.
- Ariel Massera, Mariela Assof, Santiago Sari, Ivan Ciklis, Laura Mercado, Viviana Jofre, M. C. (2021). *Effect of low temperature fermentation on the yeast-derived volatile aroma composition and sensory profile in Merlot wines*. 142.
<https://www.sciencedirect.com/science/article/abs/pii/S002364382100222X?via%3Dihub>
- Blackhurst, D., Pietersen, R., & Marais, D. (2011). Marinating beef with South African red wine may protect against lipid peroxidation during cooking. *African Journal of Food Science*, January.
- Breiman, L. (2016). A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset. *Statistics Department*

- University of California Berkeley, CA 94720, 7(6), 1–33.*
<https://doi.org/10.14569/ijacs.2016.070603>
- Butzke C. E. Vogt E. E. Chacón-Rodríguez L. (2012). *Effects of heat exposure on wine quality during transport and storage.*
<https://doi.org/10.1080/09571264.2011.646254>
- Chauhan, R., & Kaur, H. (2015). Predictive Analytics and Data Mining. *Business Intelligence, June*, 359–374. <https://doi.org/10.4018/978-1-4666-9562-7.ch019>
- Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *Eurasip Journal on Advances in Signal Processing, 2021(1)*. <https://doi.org/10.1186/s13634-021-00742-6>
- Chung, H., Son, J., Park, E., Kim, E., & Lim, S. (2008). *Journal of Food Composition and Analysis Effect of vibration and storage on some physico-chemical properties of a commercial red wine.* 21, 655–659.
<https://doi.org/10.1016/j.jfca.2008.07.004>
- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47(4)*, 547–553.
<https://doi.org/10.1016/j.dss.2009.05.016>
- Danner, L., Scientific, T. C., Ristic, R., & Johnson, T. (2016). *Context and wine quality effects on consumers ' mood , emotions , liking and willingness to pay for Australian Shiraz wines Context and wine quality effects on consumers ' mood , emotions , liking and willingness to pay for Australian Shiraz wines.* November 2018. <https://doi.org/10.1016/j.foodres.2016.08.006>
- Gentle, J. E., Härdle, W. K., & Mori, Y. (2012). Handbook of computational statistics: Concepts and methods: Second Edition. *Handbook of Computational Statistics: Concepts and Methods: Second Edition, February 2004*, 1–1192. <https://doi.org/10.1007/978-3-642-21551-3>
- Gim, J., Lee, S., & Joo, W. (2018). A study of prescriptive analysis framework for human care services based on CKAN cloud. *Journal of Sensors, 2018*. <https://doi.org/10.1155/2018/6167385>
- Gupta, M., & C, V. (2021). A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality. *International Journal of Recent Technology and Engineering (IJRTE), 10(1)*, 314–319.
<https://doi.org/10.35940/ijrte.a5854.0510121>
- HUGH JOHNSON and JANCIS ROBINSON, M. B. (2013). *The World Atlas of Wine, 7th edition.* <https://doi.org/10.1017/jwe.2013.39>
- Hussain, M. (2014). Descriptive statistics - presenting your results I Descriptive statistics - presenting your results I Qualitative Variable Analysis : Types of Statistical Analyses : Descriptive Statistical Analysis : *Journal of the Pakistan Medical Association, July 2012*, 8–11.
- J., R. J. R. J. H. R. (2007). *Vintage: The Story of Wine.*
<https://doi.org/10.2307/4612162>
- Kullarni, V. Y., & Sinha, P. K. (2013). Random Forest Classifier: A Survey and Future Research Directions. *International Journal of Advanced Computing, 36(1)*, 1144–1156.

- Kumar, V., & L., M. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, 182(1), 31–37. <https://doi.org/10.5120/ijca2018917434>
- Li, H. et al. A. (2018). *www.econstor.eu*. 178. <https://doi.org/10.1016/j.wep.2018.10.002>
- Mattivi, F., Arapitsas, P., & Perenzoni, D. (2015). *Influence of Storage Conditions on the Composition of Influence of Storage Conditions on the Composition of Red Wines*. August 2014. <https://doi.org/10.1021/bk-2015-1203.ch003>
- Robinson, A. L., Mueller, M., Heymann, H., Ebeler, S. E., Boss, P. K., Solomon, P. S., & Trengove, R. D. (2010). *Effect of Simulated Shipping Conditions on Sensory Attributes and Volatile Composition of Commercial White and Red Wines*. 3, 337–347.
- Syaputri, A. W., & Irwandi, E. (2020). *Na e Ba e A g i h f C a ifcai S de Maj S eciai a i f*. 1(1), 1–5.
- Thach, L., & Feb, M. W. (2020). *Statistics on US Wine Market for 2019*. i, 1–8.
- Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. *International Journal of Recent Contributions from Engineering, Science & IT (IJES)*, 7(2), 91. <https://doi.org/10.3991/ijes.v7i2.10659>
- World Health Organization. (n.d.). *Global status report on alcohol and health*. https://www.who.int/substance_abuse/publications/global_alcohol_report/msbgsruprofiles.pdf
- Xhemali, D., J. Hinde, C., & G. Stone, R. (2009). Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science*, 4(1), 16–23. <http://cogprints.org/6708/>
- Yasar, A., & Saritas, M. M. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, 7(2), 88–91. <https://doi.org/10.18201/ijisae.2019252786>