

Predicting The Quality of Red Wine and White Wine Using Data Mining

Ni Wayan Priscila Yuni Praditya^{1*}, Noor Akhmad Setiawan², Fery Antony³


^{1,3}Department of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia

²Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta, Indonesia

ABSTRACT

In business intelligence or artificial intelligence, data mining is a technique that can classify and cluster data based on the nature and correlation of the data set used. In data mining, several methods can be used, such as C45, K-Means, Apriori Decision Tree, KNN, LSTM, Naive Bayesian, etc. This research utilizes the Decision Tree method which aims to classify the quality of red wine and white wine. The results of this study indicate that the prediction of red wine has a precision of 61.1%, recall of 60.7%, f-measure of 60.3%, and an average accuracy of 60.7% while white wine has a precision of 58.2%, recall of 58.7%, f-measure 58.4%, and 58.7% accuracy. The method used in this study also shows that Decision Tree can outperform other methods such as Lib-SVM, BayesNet, and Multi Perceptron.

Keyword : artificial intelligence; business intelligence; data mining; Prediction; Decision Tree.

 This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Ni Wayan Priscila Yuni Praditya,
Department of Computer Science
Universitas Indo Global Mandiri
Jl. Jendral Sudirman No. 629 Km 4 Palembang, Indonesia
Email : niwayanpris@uigm.ac.id

Article history:

Received : Jun 3rd 2023
Revised : Jul 15th 2023
Accepted : Aug 27th 2023

1. INTRODUCTION

Wine is an alcoholic beverage made from the anaerobic fermentation of grape juice without the presence of O₂ (Hardinata, Okprana, Windarto, & Saputra, 2019). The balance of the natural properties contained in grapes can cause the fruit to be fermented without the addition of sugar, acids, enzymes, or other nutrients. Making wine by fermenting grape juice uses certain yeasts which then contain the sugar in the grapes which will be consumed by yeast and turn it into alcohol. Different types of grapes and yeast strains are used, depending on the type of wine to be produced (Aich, Al-Absi, Hui, & Sain, 2018). In producing wine, the composition used must have a high nutritional content, have high acidity so that it can inhibit the growth of unwanted microbes, the sugar content is high enough and the aroma is delicious, therefore the quality of the wine must be prioritized.

One way that can be used to predict the quality of a wine can be used by classifying data using data mining. Classification aims to predict the class of an object that has not been known before and is measured objectively and subjectively. Data mining is a combination or blend of factual models and machine learning, the term data mining is related to the extraction of learning and models from large datasets (Kumar, Agrawal, & Mandan, 2020). The application of data mining can use several algorithms that can help in the success of this research such as Naïve Bayes (Razan, et al., 2021), Random Forest (Ardiningtyas & Rosa, 2021), Support Vector Machines (SVM) (Atmanegara & Purwa, 2021), Decision Tree (Febriantono, Herasmara, & Pangestu, 2021), K-NN (Prasetyo, 2012), etc.

In this writing, the wine data set used is a collection of white wine and red wine. White wine consisted of 4898 samples and red wine consisted of 1599 samples, each sample from both types of wine consisted of 12 physiochemical variables, namely fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.

The next part of this writing is structured as follows: Part II discusses the literature review related to research. Part III discusses the methodology used. Part IV discusses implementation. Section V is the conclusion.

2. MATERIAL AND METHOD

Explaining The data mining process in this study uses the Wine Quality data set which can be used for several implementations for both wine producers and wine consumers in predicting the quality of wine. contained in each quality of wine for the benefit of health, drinking responsibility, and other interests. The data used in this data mining process is Wine Quality obtained from the UC Irvine Machine Learning Repository (Aeberhard & Forina, 2020). The data obtained has 1599 red wine data and 4898 white wine data and 12 variables consisting of:

1. Fixed acidity: Fixed amount of acid contained in wine, where the acid content does not evaporate easily.
2. Volatile acidity: The amount of volatile acetic acid contained in wine, at too high a concentration level can spoil the taste.
3. Citric acid: The amount of citric acid contained, is useful for adding freshness and taste to wine.
4. Residual sugar: The amount of sugar remaining after fermentation stops.
5. Chlorides: The amount of salt contained in wine.
6. Free sulfur dioxide: The amount of SO₂ content in wine. Where SO₂ is in free form in equilibrium between SO₂ molecules and bisulfite ions, which prevents microbial growth and prevent oxidation in wine.
7. Total sulfur dioxide: Total free and bound forms of SO₂, at low concentrations, SO₂ is largely undetectable in wine.
8. Density: The density of water depends on the percentage of alcohol and sugar content in the wine.
9. pH: The acidity or base level of wine, most wines are on a scale of 3-4 on the pH scale.
10. Sulphates: Wine additive levels that can contribute to SO₂ gas levels, which act as anti-microns and antioxidants.
11. Alcohol: Percent alcohol content in wine.
12. Quality: Output variable (based on sensory data, the value is between 0-10).

```
wine.shape
(1599, 12)

wine.dtypes
fixed acidity      float64
volatile acidity   float64
citric acid        float64
residual sugar     float64
chlorides          float64
free sulfur dioxide float64
total sulfur dioxide float64
density           float64
pH               float64
sulphates         float64
alcohol           float64
quality           int64
dtype: object
```

Figure 1. Amount and type of data

Table 1. Means and ranges of the physicochemical data in the UCI Wine Quality Data Set

Attribute	Red wine	White wine
	Mean (Range)	Mean (Range)
Fixed acidity	8.3 (4.6 - 15.9)	6.9 (3.8 - 14.2)
Volatile acidity	0.5 (0.1 - 1.6)	0.3 (0.1 - 1.1)

Citric acid	0.3 (0.0 - 1.0)	0.3 (0.0 - 1.7)
Residual sugar	2.5 (0.9 - 15.5)	6.4 (0.6 - 65.8)
Chlorides	0.08 (0.01 - 0.61)	0.05 (0.01 - 0.35)
Free sulfur dioxide	14 (1 - 72)	35 (2 - 289)
Total sulfur dioxide	46 (6 - 289)	138 (9 - 440)
Density	0.996 (0.990 - 1.004)	0.994 (0.987 - 1.039)
pH	3.3 (2.7 - 4.0)	3.1 (2.7 - 3.8)
Sulphates	0.7 (0.3 - 2.0)	0.5 (0.2 - 1.1)
Alcohol	10.4 (8.4 - 14.9)	10.4 (8.0 - 14.2)

The table above shows the scope of physiochemical data, each wine in this data set has also been evaluated by at least three people. Figure 2 below will show the importance of each physiochemical data item in the UCI Repository.

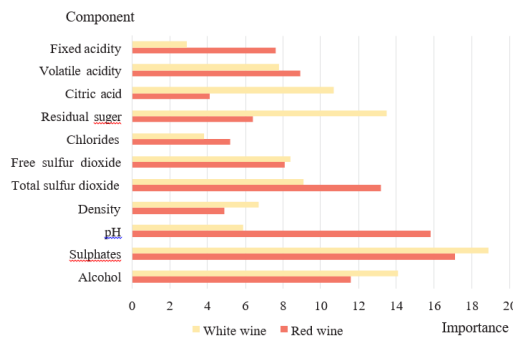


Figure 2. The importance of physiochemical indicators

The Wine Quality data set which has been in the form of classification data will be tested for its classification accuracy before mining. This accuracy test is carried out using the classification model on Scikitlearn.

```
models=[LogisticRegression(),LinearSVC(),SVC(kernel='rbf'),KNeighborsClassifier(),
DecisionTreeClassifier(),GradientBoostingClassifier(),GaussianNB()]
model_names=['LogisticRegression','LinearSVM','rbfSVM','KNearestNeighbors',
['DecisionTree','GradientBoostingClassifier','GaussianNB']
```

Figure 3. Coding classification correlation test

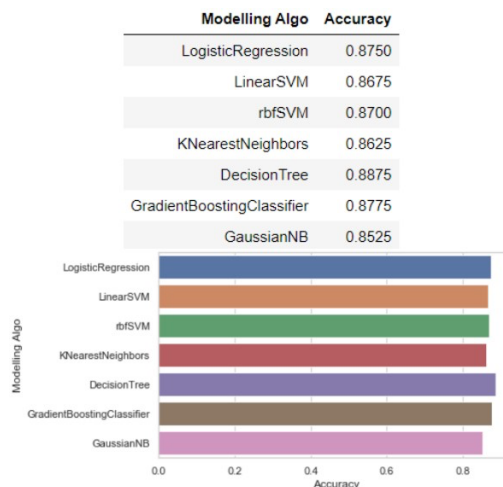


Figure 4. Classification accuracy test results

Based on the classification accuracy test that has been carried out, it is known that the most superior accuracy model is the decision tree with an accuracy rate of 88.75%. A decision tree or decision tree is a simple structure that can be used as a classifier. In a decision tree, each internal (non-leaf) node represents an attribute variable and each branch represents a state of this variable. Each of the three leaves represents the expected value of the class of variables to be predicted, an important aspect of the procedure for building this decision tree is the split criterion or separation of criteria including the criteria for creating a new branch and the stop criterion or the last criterion, the criteria used to stop branching. The decision tree is made using a set of data that is used as a training dataset or learning data.

3. RESULTS AND DISCUSSION

A. Decision tree

The predictive model that will be used in this study uses the decision tree method, based on the physiochemistry of wine, and uses it to predict taste. The decision tree is built with recursive subdivisions, namely selecting each of the most influential attributes from the research sample at each node and separating the set of instances into subsets with high and low values of these attributes, each of these subsets becomes a subdivision. Once the subdivisions are complete, all files are distributed across the leaf nodes (Supriyadi, Gata, Maulidah, & Fauzi, 2020).

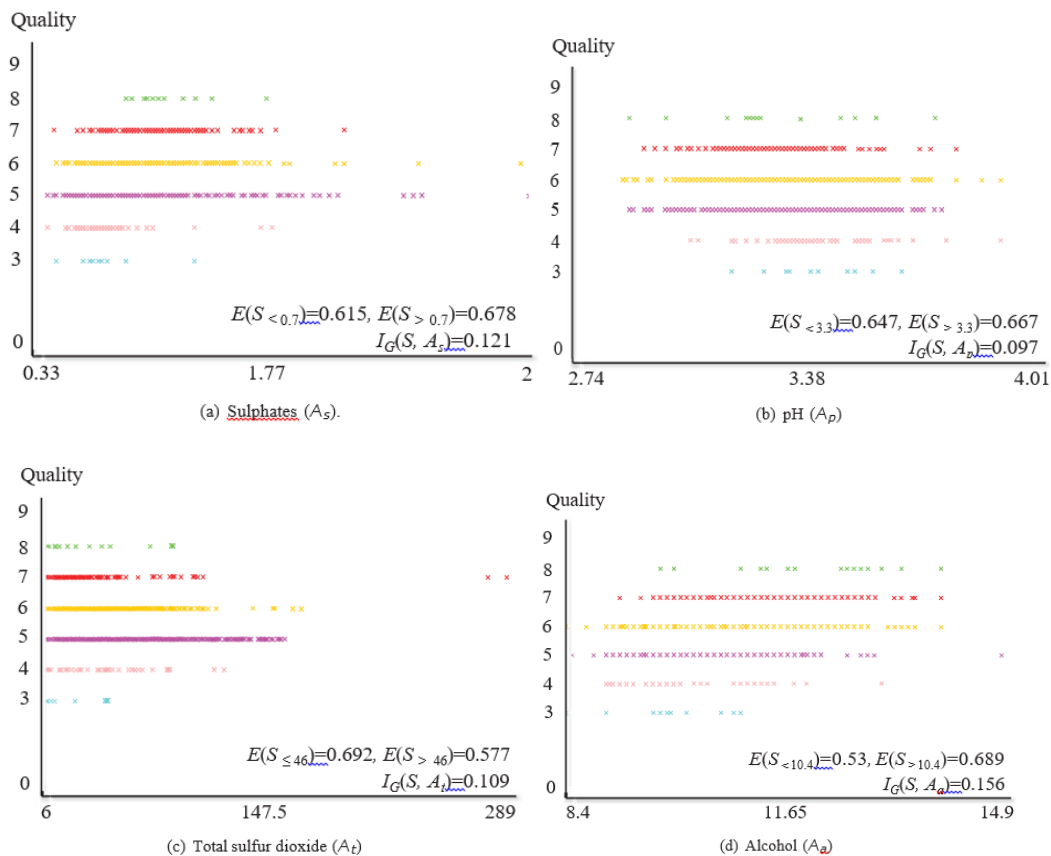


Figure 5. Subset with the high and low values of the attributes in the decision tree, (a) Sulphates, (b) pH, (c) Total sulphur dioxide, (d) Alcohol

To select influential attributes, C4.5 uses to gain information considered as entropy expressed as $\sum_{i=1}^n (-p_i \log_2 p_i)$ where S denotes a dataset with 11 attributes, I denotes a degree of preference, and p_i denotes the proportion of S to a degree i . If all instances belong to the same class, the value is $E(S) = 0$, if all belong to a different class, then $E(S) = 1$. We can proceed to express the effectiveness of an attribute as information gain IG , as follows:

$$I_G(S, A) = E(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} E(S_v),$$

Where $V(A)$ denotes the set of all possible values of attribute A , and S_v is a subset of S where the attribute value is v . The most influential attribute on a node is that with the highest information gain.

In this study, entropy comparisons were made to describe the approximation and obtain information from the four main physiochemical indicators above of the quality of red wine, and the results are shown in Figure 2. First, the entropy $E(S)$ of all instances was calculated, and it was found to be 0.754. Then the information obtained from each attribute is calculated using Eq. Next sets v to the average value of that attribute in this group of instances. This shows that alcohol content is the most significant determinant of the quality of red wine.

B. Correlation for each variable

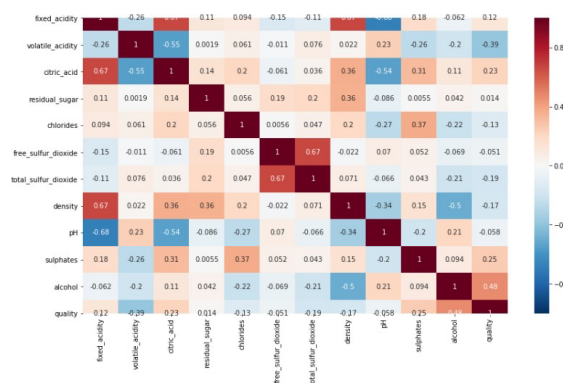


Figure 6. Correlation for each variable

Based on the table above, it can be seen that:

1. Quality has a (+) positive relationship between alcohol.
2. Quality has a weak (-) negative relationship between volatile_acidity.
3. Quality has almost no relationship between residual_sugar, free_sulfur_dioxide, and pH. (Corr = ~ 0).
4. Alcohol has a (+) positive relationship between quality and weak pH.
5. Alcohol has a (-) negative relationship between densities.
6. Alcohol has almost no relationship between fixed_acidity, residual_sugar, free_sulfur_dioxide, sulfate.
7. Volatile_acidity has a weak positive relationship (+) between pH.
8. Volatile_acidity has a strong negative relationship (-) between citric_acid.
9. Volatile_acidity has a weak (-) negative relationship between fixed_acidity and sulfate.
10. Volatile_acidity has almost no relationship between residual_sugar, chloride, free_sulfur_dioxide, total_sulfur_dioxide, density.
11. Density has a (+) positive relationship between fixed_acidity.
12. Density has (-) negative relationship between densities
13. Density has almost no relationship between volatile_acidity, free_sulfur_dioxide, total_sulfur_dioxide.
14. Citric_acid has (+) a positive relationship between fixed_acidity.
15. Citric_acid has (-) negative relationship between volatile_acidity, pH.
16. Citric_acid has almost no relationship between residual_sugar, free_sulfur_dioxide, total_sulfur_dioxide.

Information obtained from this collaboration can make the wine industry composition efficient in its production process.

C. Linear regression of several variables on alcohol

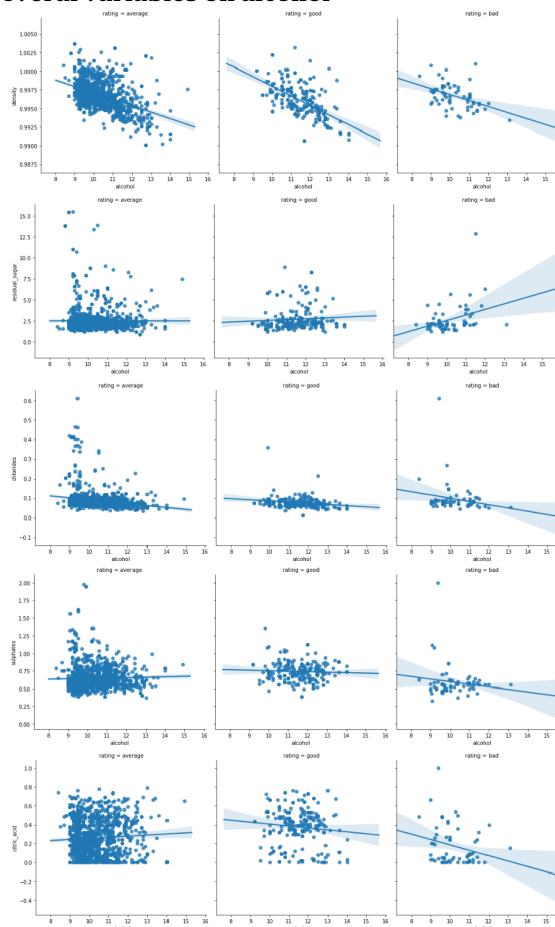


Figure 7. Linear regression

From the linear regression above, the wine industry can also make production efficient. For example, to make high-quality wine, producers can decrease the density by adding alcohol, increase residual sugars by adding alcohol, or decrease the levels of chloride, sulphate, and citric acid by adding alcohol.

D. Result

Based on the prediction results regarding the taste preferences of wine, the quality measured is 3-8 for red wine and 4-8 for white wine. The red wine prediction has a precision of 61.1%, recall of 60.7%, f-measure of 60.3%, and an average accuracy of 60.7%. White wine has 58.2% precision, 58.7% recall, 58.4% f-measure, and 58.7% accuracy.

Table 2. Accuracy predicted taste preferences

Quality	White wine			Red wine		
	Precision (%)	Recall (%)	F-Me (%)	Precision (%)	Recall (%)	F-Me (%)
3	7.7	10.0	8.70	-	-	-
4	24.4	20.8	22.4	27.3	23.9	25.5
5	48.2	72.2	70.2	60.0	61.9	60.9
6	57.9	57.7	57.8	63.9	64.6	64.3
7	55.7	48.7	52.0	52.8	51.9	52.4
8	10.0	60.7	60.3	36.7	31.4	33.8
Avg	61.1	60.7	60.3	58.2	58.7	58.5
Ac (%)	60.7			58.7		

4. CONCLUSION

Based on the implementation results using the decision tree method and the dataset obtained from the UCI Repository in this study, it can be concluded that the decision tree method can be used to predict the quality of wine, both red wine and white wine, so that producers and consumers can easily determine the quality of the wine.

REFERENCES

- Aeberhard, S., & Forina, M. (2020, February 22). *Wine Dataset*. Retrieved from UC Irvine Machine Learning Repository: <https://archive.ics.uci.edu/dataset/109/wine>
- Aich, S., Al-Absi, A. A., Hui, K. L., & Sain, M. (2018). Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques. *ICACT Transactions on Advanced Communications Technology*, 7(3), 1122-1127.
- Ardiningtyas, Y. E., & Rosa, P. H. (2021). Analisis Balancing Data Untuk Meningkatkan Akurasi Dalam Klasifikasi. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi* (pp. 24-28). Yogyakarta: Universitas Sanata Dharma.
- Atmanegara, E., & Purwa, T. (2021). Hybrid Support Vector Machine and Logistic Regression for Multiclass Classification: A Case Study on Wine Dataset. *Indonesian Journal of Data Science*, 1-7.
- Croce, R., Malegori, C., Oliveri, P., Medici, I., & Cavaglioni, A. (2020). Prediction of quality parameters in straw wine by means of FT-IR spectroscopy combined with multivariate data processing. *Food Chemistry*. Italy.
- Er, Y., & Atasoy, A. (2016). The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. *International Journal of Intelligent Systems and Applications in Engineering*. Turkey.
- Febriantono, M. A., Herasmara, R., & Pangestu, G. (2021). Cost Sensitive Tree dan Naive Bayes pada Klasifikasi Multiclass. *Jurnal Informatika Polinema*, 57-64.
- Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, (pp. 305-312). Kurukhshetra.
- Hardinata, J. T., Okprana, H., Windarto, A. P., & Saputra, W. (2019). Analisis Laju Pembelajaran dalam. *Jurna Sains Komputer & Informatika (J-SAKTI)*, 3 No 2, 422-432.
- Khakim, E. N., Hermawan, A., & Avianto, D. (2023). Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine. *JIKO (Jurnal Informatika dan Komputer)*, 2, 158-166.
- Kumar, S., Agrawal, K., & Mandan, N. (2020). Red Wine Quality Prediction Using Machine Learning Techniques. *International Conference on Computer Communication and Informatics*. Coimbatore, India.
- Lee, S., Park, J., & Kang, K. (2015). Assessing wine quality using a decision tree. *IEEE International Symposium on Systems Engineering (ISSE)*. Rome, Italy: IEEE.
- Prasetyo, E. (2012). K-Support Vector Nearest Neighbor Untuk Klasifikasi Berbasis K-NN. *Jurnal Sesindo Unpad*, 245-250.
- Radosavljević, D. (2019). A Data Mining Approach to Wine Quality Prediction. *International Scientific Conference*. Gabrovo.
- Razan, I. M., Fatchan, M., Agam, R., Suryani, G., Asykari, R. H., & Darma, A. (2021). Studi Perbandingan Metode Naive Bayes dan Linear Discriminant Analysis Untuk Permasalahan Klasifikasi. *Jurnal Rekayasa ELEktro Sriwijaya*, 181-185.
- Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah. *Jurnal Ilmiah Ekonomi dan Bisnis*, 67-75.