JOURNAL SHIFT
VOL 3 NO 2 (2023)
ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

Prediksi Kualitas Red Wine dan White Wine Menggunakan Data Mining

Ni Wayan Priscila Yuni Praditya[1*]
[1]Prodi Sistem Komputer Fakultas Ilmu Komputer, Universitas Indo Global Mandiri
[1]niwayanpris@uigm.ac.id,

Informasi Artikel

Article historys:

Diterima 15 Mei, 2023
Revisi 21 Mei, 2023
Publish 30 Jun, 2023

Kata Kunci:

Data Mining
K-Means
Prediction
Decision Tree
Business Intelligence

ABSTRACT

Data mining is a technique used in business intelligence or artificial intelligence capable of classifying and clustering data based on the nature and correlation of the data sets used. The methods commonly used in data mining are C45, K-Means, Apriori Decision Tree, KNN, LSTM, Naive Bayesian, etc. In this study, the method used is 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 the Decision Tree can outperform other previously applied methods, namely Lib-SVM, BayesNet, and Multi Perceptron.

Koresponden Author:

Ni Wayan Priscila Yuni Praditya,
Jurusan Sistem Komputer,
Universitas Indo Global Mandiri,
Jl. Jendral Sudirman No. 629 Km 4 Palembang, Indonesia
Email: niwayanpris@uigm.ac.id

1. PENDAHULUAN

Wine merupakan minuman beralkohol yang terbuat dari hasil fermentasi anaerob jus buah anggur tanpa kehadiran O2 [1]. Keseimbangan sifat alami yang yang terkandung pada buah anggur dapat menyebabkan buah tersebut difermentasi tanpa penambahan gula, asam, enzyme, maupun nutrisi lain. Pembuatan wine dengan cara fermentasi jus buah anggur ini menggunakan khamir tertentu yang kemudian kandungan gula yang ada pada buah anggur tersebut akan dikonsumsi oleh yeast (ragi) dan mengubahnya menjadi alcohol. Jenis anggur yang berbeda dan strain khamir yang digunakan, tergantung pada jenis wine yang akan di produksi [2]. Dalam memproduksi wine, komposisi yang digunakan harus mempunyai kandungan nutrisi tinggi, mempunyai keasaman yang tinggi sehingga dapat menghambat pertumbuhan mikroba yang tidak diinginkan, kandungan gula cukup tinggi dan aroma yang sedap, oleh itu qualitas wine harus diutamakan.

Salah satu cara yang dapat digunakan untuk memprediksi kualitas dari sebuah wine dapat digunakan dengan cara klasifikasi data menggunakan data mining. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang belum diketahui sebelumnya dan diukur secara obyektif dan subyektif. Data mining adalah gabungan atau perpaduan model faktual dan machine learning,

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

istilah data mining ini berkaitan dengan ekstraksi pembelajaran dan model dari dataset yang besar [3]. Penerapan sebuah data mining dapat menggunakan beberapa algoritma yang dapat membantu dalam kesuksesan penelitian ini seperti Naïve Bayes, Random Forest, Support Vector Machines (SVM), Decision Tree, K-NN, dll.

Dalam penulisan ini, kumpulan data wine yang digunakan adalah koleksi white wine dan red wine. White wine terdiri dari 4898 sampel dan red wine terdiri dari 1599 sampel, setiap sampel dari kedua jenis wine terdiri dari 12 variabel fisiokimia, yaitu fixed acidity, volatile acidity, citric acid, redisual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, dan quality.

Bagian selanjutnya dari penulisan ini disusun dengan struktur sebagai berikut: Bagian II membahas mengenai tinjauan pustaka terkait penelitian. Bagian III membahas mengenai metedologi yang digunakan. Bagian IV membahas mengenai implementasi. Bagian V adalah kesimpulan.

2. METODE PENELITIAN/ALGORITMA

Dalam proses data mining pada penelitian ini menggunakan data set Wine Quality yang dapat digunakan untuk beberapa implementasi baik untuk produsen wine maupun konsumen wine dalam memprediksi kualitas wine.industri wine dapat melihat bagaimana kualitas wine, dan dari hasil data mining ini dapat membantu konsumen untuk mengetahui kandungan yang ada dalam tiap kualitas wine untuk kepentingan kesehatan, drinking responsibility, maupun kepentingan lainnya.

Data yang digunakan dalam dalam proses data mining ini adalah Wine Quality yang didapatkan dari UC Irvine Machine Learning Repository. Data yang didapatkan memiliki 1599 data red wine dan 4898 data white wine dan 12 variabel yang terdiri dari:

1. Fixed acidity : Jumlah asam tetap yang terkandung didalam wine , dimana kandungan asam tersebut tidak mudah menguap.
2. Volatile acidity : Jumlah asam asetat mudah menguap yang terkandung dalam wine, dimana pada tingkat konsentrasi terlalu tinggi dapat merusak rasa.
3. Citric acid : Jumlah asam sitrat yang terkandung, berguna untuk menambah kesegaran dan rasa pada wine.
4. Residual sugar : Jumlah gula yang tersisa setelah fermentasi berhenti.
5. Chlorides : Jumlah garam yang terkandung di dalam wine.
6. Free sulfur dioxide : Jumlah kandungan SO2 yang ada dalam wine. Dimana SO2 berbentuk bebas dalam kesetimbangan antara molekul SO2 dan ion bisulfit, yang mencegah pertumbuhan mikroba dan mencegah terjadinya oksidasi pada wine.
7. Total sulfur dioxide : Jumlah bentuk bebas dan terikat SO2, dalam konsentrasi rendah, SO2 sebagian besar tidak terdeteksi dalam wine.
8. Density : Kerapatan air yang tergantung pada persen alkohol dan kadar gula dalam wine.
9. pH : Tingkat asam atau basa wine, kebanyakan wine berada di skala 3-4 pada skala pH.
10. Sulphates : Kadar aditif wine yang dapat berkontribusi pada kadar gas SO2, yang bertindak sebagai anti mikrona dan antioksidan.
11. Alcohol : Persen kandungan alkohol dalam wine.
12. Quality : Variabel output (berdasarkan data sensorik, nilainya antara 0-10).

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

```
wine.shape
```

```
(1599, 12)
```

```
wine.dtypes
```

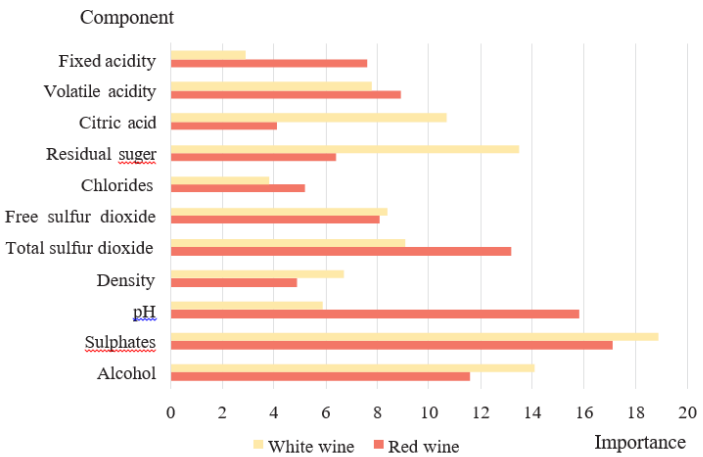
```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates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

Gambar 1. Jumlah dan Jenis Data

Tabel 1. Means and Ranges of The Physicochemical Data In The UCI Wine Quality Data Set

Attribute	Red wineMean (Range)	White wine 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)

Tabel diatas menunjukkan ruang lingkup data fisiokimia, setiap wine dalam kumpulan data ini juga telah dievaluasi minimal lebih dari tiga orang. Gambar 2 dibawah ini akan menunjukkan pentingnya setiap item data fisiokimia di UCI Repository [11].



Gambar 2. Pentingnya Indikator Fisiokimia

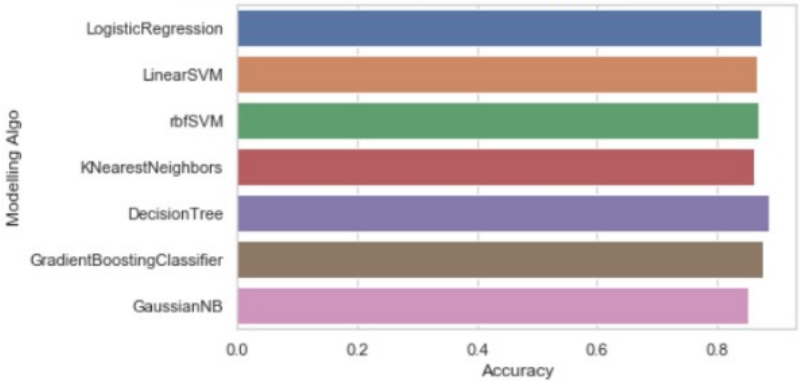
ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

Data set Wine Quality yang telah berbentuk data klasifikasi akan diuji akurasi pengklasifikasiannya sebelum dilakukan mining. Uji akurasi ini dilakukan dengan menggunakan model klasifikasi pada Sckitlearn.

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

Gambar 2 koding uji kolerasi klasifikasi

Modelling Algo	Accuracy
LogisticRegression	0.8750
LinearSVM	0.8675
rbfSVM	0.8700
KNearestNeighbors	0.8625
DecisionTree	0.8875
GradientBoostingClassifier	0.8775
GaussianNB	0.8525



Gambar 3. Hasil Uji Akurasi Klasifikasi

Berdasarkan uji akurasi klasifikasi yang telah dilakukan, diketahui bahwa model akurasi yang paling unggul adalah decision tree dengan tingkat akurasi 88,75%. Decision tree atau pohon keputusan adalah struktur sederhana yang dapat digunakan sebagai pengklasifikasian. Dalam pohon keputusan masing-masing node internal (non-leaf) mempresentasikan sebuah variable atribut dan masing-masing cabang mempresentasikan satu keadaan dari variable ini. Masing-masing dari tiga leaf mempresentasikan nilai yang diharapkan dari kelas variable yang akan diprediksi, aspek penting dalam prosedur untuk membangun decision tree ini adalah split criterion atau pemisahan kriteria termasuk ktiteria untuk membuat cabang baru dan stop criterion atau kriteria terakhir, kriteria yang digunakan untuk menghentikan pencabangan. Decision tree dibuat menggunakan himpunan dari data yang digunakan sebagai training dataset atau data pembelajan.

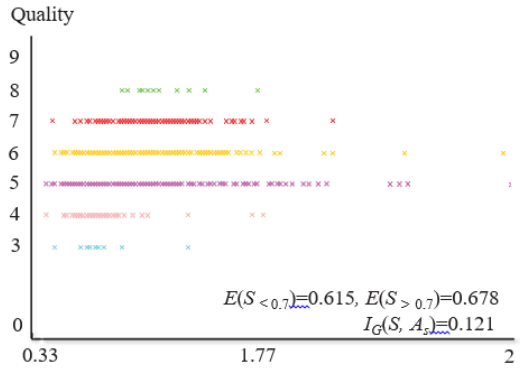
3. ANALISIS EKSPERIMEN

Decision tree

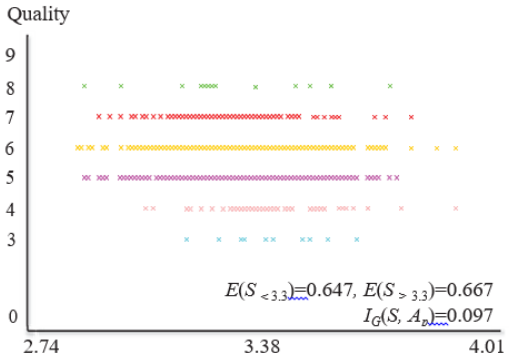
Model prediktif yang akan digunakan dalam penelitian ini menggunakan metode decision tree, berdasarkan fisiokimia wine dan menggunakannya untuk memprediksi rasa. Decision tree dibangun dengan subdivisi rekursif yaitu memilih masing-masing atribut yang paling berpengaruh dari contoh penelitian di masing-masing node dan memisahkan set instance tersebut menjadi subset

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

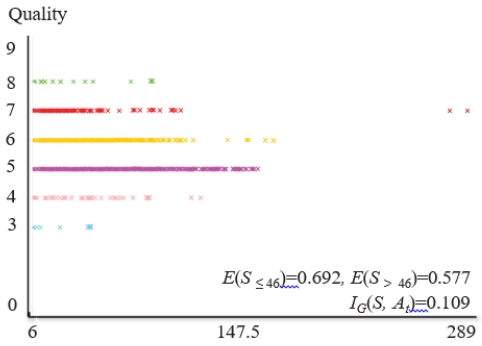
dengan tinggi dan rendah nilai dari atribut tersebut, masing-masing subset ini menjadi subdivisi. Setelah subdivisi selesai, semua file didistribusikan di seluruh node leaf.



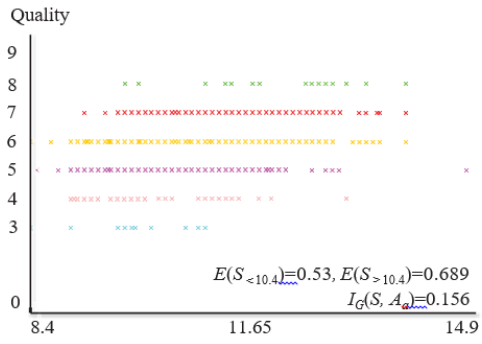
(a) Sulphates (A_s).



(b) pH (A_p)



(c) Total sulfur dioxide (A_t)



(d) Alcohol (A_a)

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

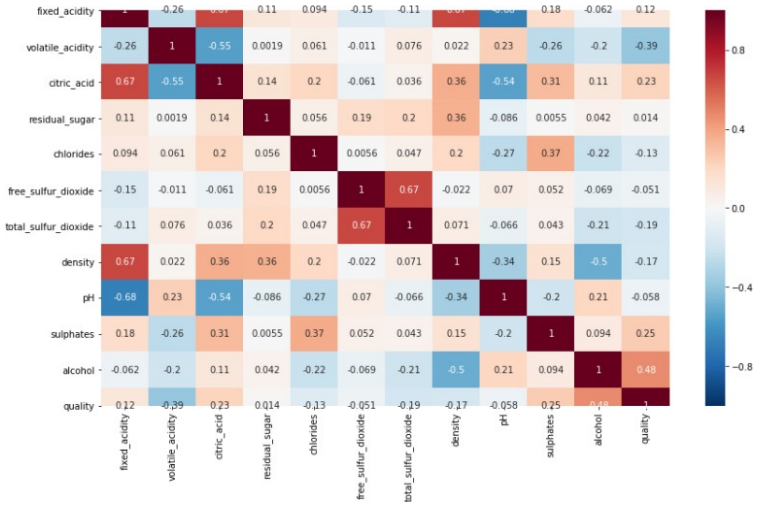
Untuk memilih antribut yang berpengaruh, C4.5 menggunakan informasi gain dianggap sebagai entropi yang dinyatakan sebagai $\sum_{i=1}^{n}(-pi\ \log 2pi)$ dimana S menunjukkan sebuah dataset dengan 11 atribut, I menunjukkan derajat preferensi, dan pi menunjukkan proporsi S dengan derajat i. Jika semua instance milik kelas yang sama, file nilai $E(S) = 0$, jika semuanya termasuk dalam kelas yang berbeda, maka $E(S) = 1$. Kita bisa melanjutkan untuk mengungkapkan keefektifan sebuah atribut sebagai perolehan informasi $I\,G$, sebagai berikut:

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

di mana $V(A)$ menunjukkan himpunan semua kemungkinan nilai dari atribut A, dan $S\,v$ adalah himpunan bagian dari S di mana nilai atributnya adalah v. Atribut paling berpengaruh pada sebuah node adalah bahwa dengan perolehan informasi tertinggi.

Dalam penelitian ini, pembandingan entropi dilakukan untuk mengambarkan pendekan dan memperoleh informasi dari empat indicator fisiokimia utama diatas dari kualitas red wine dan hasilnya menunjukkan pada Gambar 2. Pertama, entropi $E(S)$ dari semua instance telah dihitung, dan ditemukan menjadi 0,754. Kemudian perolehan informasi dari masing-masing atribut itu dihitung menggunakan Persamaan. Selanjutnya menetapkan v ke nilai rata-rata atribut itu dalam grup instance ini. Hal ini menunjukkan bahwa kadar alcohol merupakan penentu paling signifikan mengenai kualitas red wine.

Korelasi untuk tiap variabel



Gambar 8 Kolerasi antar variabel

Berdasarkan tabel gambar diatas dapat dilihat bahwa:

- Kualitas memiliki (+) hubungan positif antara alkohol.
- Kualitas memiliki (-) hubungan lemah negatif antara volitile_acidicity.
- Kualitas hampir tidak memiliki hubungan antara residual_sugar, free_sulfur_dioxide, dan pH. (Corr = ~ 0).
- Alkohol memiliki (+) hubungan positif antara kualitas dan pH lemah.
- Alkohol memiliki (-) hubungan negatif antara kepadatan.
- Alkohol hampir tidak memiliki hubungan antara fixed_acidicity, residual_sugar, free_sulfur_dioxide, sulfat.
- Volitile_acidicity memiliki hubungan positif yang lemah (+) antara pH.
- Volitile_acidicity memiliki hubungan negatif yang kuat (-) antara citric_acid.
- Volitile_acidicity memiliki hubungan negatif yang lemah (-) antara fixed_acidicity dan sulfat.
- Volitile_acidicity hampir tidak memiliki hubungan antara residual_sugar, klorida,

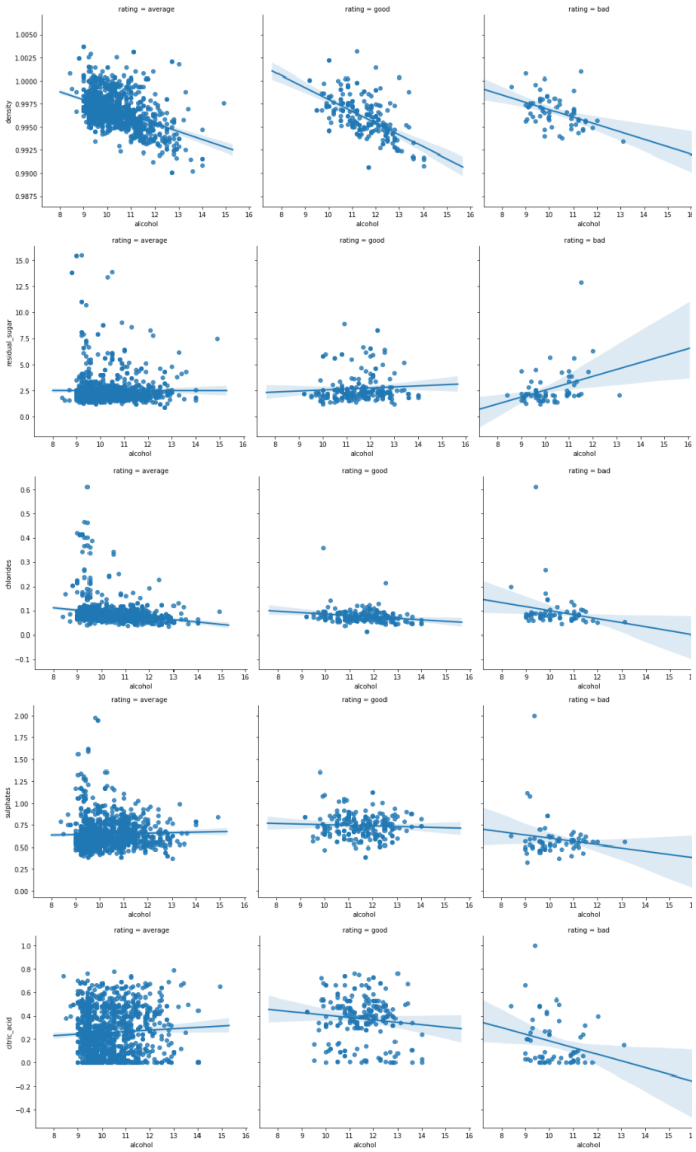
ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

free_sulfur_dioxide, total_sulfur_dioxide, kepadatan.

- Densitas memiliki (+) hubungan positif antara fixed_acidicity.
- Densitas memiliki (-) hubungan negatif antara densitas
- Densitas hampir tidak memiliki hubungan antara volitile_acidicity, free_sulfur_dioxide, total_sulfur_dioxide.
- Citric_acid memiliki (+) hubungan positif antara fixed_acidicity.
- Citric_acid memiliki (-) hubungan negatif antara volitile_acidicity, pH.
- Citric_acid hampirtidak memiliki hubungan antara residual_sugar, free_sulfur_dioxide, total_sulfur_dioxide.

Informasi yang didapat dari kolerasi tersebut indutri wine dapat melakukan efesiensi komposisi dalam proses produksinya.

a. Regresi linear dari beberapa variabel terhadap alcohol



Gambar 9. Regresi Linier

Dari regresi linear diatas, industri *wine* juga dapat melakukan efisiensi dalam melakukan produksi. Contohnya untuk membuat *wine* dengan kualitas tinggi, produsen dapat menurunkan

ISSN 2776-8961 (ONLINE)
2808-3385 (PRINT)

density dengan menambahkan alkohol, meningkatkan gula residu dengan menambahkan alkohol, atau menurunkan kadar klorida, sufat, dan asam citric dengan menambahkan alkohol.

Hasil

Berdasarkan dari hasil prediksi mengenai preferensi rasa sebuah wine, kualitas yang diukur adalah 3-8 untuk red wine dan 4-8 untuk white wine. Prediksi red wine memiliki precision 61,1%, recall 60,7%, f-measure 60,3%, dan accuracy rata-rata 60,7%. Sedangkan white wine memiliki precision 58,2%, recall 58,7%, f-measure 58,4%, dan akurasi 58,7%.

Tabel 2. Accuracy Predicted Taste Preferences

Quality	White wine			Red wine		
	Precision (%)	Recall (%)	F-Me (%)	Precision (%)	Recall (%)	F-Me (%)
3	07.07	10.00	0,38194444	–	–	–
4	24.04.00	20.08	22.04	27.03.00	23.09	25.05.00
5	48.02.00	72.02.00	70.02.00	60.00.00	61.09.00	60.09.00
6	57.09.00	57.07.00	57.08.00	63.09.00	64.06.00	64.03.00
7	55.07.00	48.07.00	52.00.00	52.08.00	51.09.00	52.04.00
8	10.00	60.07.00	60.03.00	36.07.00	31.04.00	33.08.00
Avg	61.01.00	60.07.00	60.03.00	58.02.00	58.07.00	58.05.00
Ac (%)	60.07.00			58.07.00		

4. KESIMPULAN

Berdasarkan dari hasil implementasi dengan menggunakan metode decision tree dan dataset yang di dapat dari UCI Repository pada penelitian ini, dapat disimpulkan bahwa metode decision tree dapat digunakan untuk mempresiksi kualitas wine baik red wine maupun white wine, sehingga prosuden dan konsumen dapat dengan mudah mengetahui kualitas dari wine tersebut.

DAFTAR PUSTAKA

[1] J. T. Hardinata, H. Okprana, A. P. Windarto and W. Saputra, "Analisis Laju Pembelajaran dalam," Jurna Sains Komputer & Informatika (J-SAKTI), vol. 3 No 2, pp. 422-432, 2019.

[2] S. Aich, A. A. Al-Absi, K. L. Hui and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques," ICACT Transactions on Advanced Communications Technology, vol. 7, no. 3, pp. 1122-1127, 2018.

[3] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," in International Conference on Computer Communication and Informatics, Coimbatore, India, 2020.

[4] S. Lee,, J. Park and . K. Kang, "Assessing wine quality using a decision tree," in IEEE International Symposium on Systems Engineering (ISSE), Rome, Italy, 2015.

[5] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," in Procedia Computer Science, Kurukhshetra, 2018.

[6] Y. Er and A. Atasoy, "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities," in International Journal of Intelligent Systems and Applications in Engineering, Turkey, 2016.

[7] R. Supriyadi, W. Gata, N. Maulidah and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," Jurnal Ilmiah Ekonomi dan Bisnis, vol. 13 No.2, pp. 67-75, 2020.

[8] D. Radosavljević, "A Data Mining Approach to Wine Quality Prediction," in International Scientific Conference, Gabrovo, 2019.

[9] R. Croce, C. Malegori, P. Oliveri, . I. Medici and A. Cavaglioni, "Prediction of quality parameters in straw wine by means of FT-IR spectroscopy combined with multivariate data processing," in Food Chemistry, Italy, 2020.

[10] E. N. R. Khakim, A. Hermawan and D. Avianto, "Implementasi Correlation Matrix Pada Klasifikasi Dataset Wine," JIKO (Jurnal Informatika dan Komputer), vol. 2, pp. 158-166, 2023.