Akurasi Klasifikasi Kualitas Wine Menggunakan Algoritma Random Forest dengan Min-Max Normalization

Putu Putri Pratiwi[a1], Ida Bagus Made Mahendra[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Udayana, Bali
Jln. Raya Kampus UNUD, Bukit Jimbaran, Kuta Selatan, Badung, 08261, Bali, Indonesia
[1]Putripratiwi720@gmail.com
[2]ibm.mahendra@unud.ac.id

Abstract

In this research, we will discuss the use of the Random Forest algorithm in classifying wine quality using Min-Max normalization. The data obtained will be subjected to data preprocessing and data normalization using Min-Max Normalization which is then applied to the Random Forest algorithm. This algorithm was chosen because it can provide good accuracy for the classification process. Data normalization and preprocessing are needed to produce a classification model with better accuracy. Min-Max normalization is used because it can improve the performance of the Random Forest algorithm in increasing accuracy.

Keywords: *Random Forest, Min-Max Normalization, Accuracy*

1. Pendahuluan

Wine merupakan minuman beralkohol yang berasal dari fermentasi buah anggur. Minuman ini banyak diminati di kalangan masyarakat. Wine biasanya disajikan pada acara-acara tertentu, baik formal maupun tidak formal. Wine juga kerap digunakan untuk menghangatkan badan di cuaca yang dingin. Banyaknya varian wine menyebabkan masyarakat menjadi bingung untuk menentukan wine mana yang akan dikonsumsi berdasarkan kualitas yang diinginkan. Klasifikasi wine dilakukan untuk memudahkan permasalahan tersebut, namun metode klasifikasi yang digunakan harus menghasilkan tingkat akurasi yang baik guna meningkatkan kinerja dan tingkat prediksi dalam klasifikasi.

Adapun beberapa algoritma klasifikasi yang dapat digunakan adalah Algoritma Decision Tree, Random Forest dan Support Vector Machine. Suatu penelitian di tahun 2020 yang berjudul Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah menyatakan bahwa algoritma Random Forest mengahsilkan akurasi terbaik di antara Algoritma Decision Tree dan Support Vector Machine [3]. Penelitian tersebut membahas mengenai perbandingan tingkat akurasi yang dihasilkan antara algoritma Decision Tree, Random Forest dan Support Vector Machine dengan menggunakan metode cross validation [3]. Penelitian lain di tahun 2021 yang berjudul Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi, membahas tentang perbandingan metode normalisasi Min-max normalization, Z-score normalization, dan satu tanpa metode normalisasi data pada algoritma Random Forest menyatakan bahwa Min-max normalization, dapat menghasilkan akurasi terbaik pada klasifikasi deteksi pasien diabetes menggunakan algoritma Random Forest [1]. Dari hasil penelitian tersebut, penulis tertarik untuk melakukan penelitian yang berjudul “Akurasi Klasifikasi Kualitas Wine Menggunakan Algoritma Random Forest Dengan Min-Max Normalization” untuk menemukan akurasi yang lebih tinggi dalam klasifikasi kualitas wine dengan dataset wine.

2. Metode Penelitian

Alur penelitian ini dimulai dari pengunduhan dataset skunder dari laman Kaggle, melakukan preprocessing data yang berupa data cleaning serta normalisasi data menggunakan Min-Max Normalization, kemudian penentuan hasil akurasi menggunakan algoritma Random Forest.

2.1 Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data skunder yang berupa wine classification dataset yang diperoleh dari laman Kaggle.

2.2 Preprocessing Data

Preprocessing data merupakan teknik pengolahan data untuk memudahkan dalam proses data mining [1]. Permasalahan pada data seperti terlalu banyak atribut, nilai data berada di range yang sangat jauh, missing value, ataupun format data yang tidak sesuai, menyebabkan gangguan pada proses data mining [2]. Preprocessing data penting dilakukan guna membuat kualitas data yang baik, termasuk kelengkapan, konsistensi, ketepatan waktu dan meningkatkan hasil akurasi [1]. Adapun preprocessing data yang dilakukan pada penelitian ini adalah Data Cleaning dan Min-Max Normalization.

a. Data Cleaning

Data cleaning merupakan proses menghapus atau mengisi nilai yang kosong untuk seluruh dataset dengan menggunakan rata-rata dari tiap kolom pada nilai yang kosong[1]. Berikut hasil dari proses data cleaning pada dataset wine.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

Gambar 1. Hasil Data Cleaning

b. Min-Max Normalization

Normalisasi data adalah proses membuat skala nilai atribut ke dalam rentang yang lebih kecil dengan bobot yang sama [2]. Min-Max Normalization adalah suatu metode yang melakukan transformasi linear dengan menggunakan nilai minimum dan maksimum yang menghasilkan keseimbangan antara data satu dengan yang lain pada rentang yang sama [4]. Adapun formulasi dalam Min-Max Normalization adalah

$$data(x) = \frac{(x - MinValue) * (MaxRange - MinRange)}{MaxValue - MinValue} + MinRange \quad (1)$$

Keterangan :

Data(x) : data baru dari hasil normalisasi
x : data yang akan dinormalisasi
MinValue : nilai terkecil dari satu kolom baris
MaxValue : nilai terbesar dari satu kolom baris
MinRange : batas nilai terkecil dari normalisasi
MaxRange : batas nilai terbesar dari normalisasi

Berikut hasil normalisasi data menggunakan Min-Max Normalization.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	0.247788	0.397260	0.00	0.068493	0.106845	0.149254	0.098940	0.567548	0.606299	0.137725	0.153846	0.4	0.000000
1	0.283186	0.520548	0.00	0.116438	0.143573	0.358209	0.215548	0.494126	0.362205	0.209581	0.215385	0.4	0.000626
2	0.283186	0.438356	0.04	0.095890	0.133556	0.208955	0.169611	0.508811	0.409449	0.191617	0.215385	0.4	0.001252
3	0.584071	0.109589	0.56	0.068493	0.105175	0.238806	0.190813	0.582232	0.330709	0.149701	0.215385	0.6	0.001879
4	0.247788	0.397260	0.00	0.068493	0.106845	0.149254	0.098940	0.567548	0.606299	0.137725	0.153846	0.4	0.002505

Gambar 2. Hasil Normalisasi Data

2.3 Percobaan pada Algoritma Random Forest

Algoritma Random Forest dikembangkan oleh Leo Bremen, algoritma ini merupakan algoritma klasifikasi yang termasuk ke dalam kelompok Supervised Learning yang terdiri dari lebih satu pohon keputusan yang setiap pohon keputusan dibentuk bergantung pada nilai-nilai vector acak sampel secara independen dan identik didistribusikan yang sama untuk semua pohon. Percobaan akan dilakukan menggunakan bahasa pemrograman Python.

3. Hasil dan Pembahasan

Berdasarkan alur dari penelitian yang dilakukan, dataset yang diunduh dari laman Kaggle akan diproses dengan metode preprocessing data yang berupa data cleaning dan Min-Max Normalization, kemudian data akan diterapkan pada algoritma Random Forest untuk diukur akurasinya. Semua proses yang dilakukan, diuji dalam kode program Python. Berikut merupakan hasil uji coba pada kode program.

```
Akurasi:  0.9868995633187773
Classification Report:
              precision    recall  f1-score   support

           4       1.00      0.83      0.91         6
           5       0.99      1.00      0.99        96
           6       1.00      1.00      1.00        99
           7       0.93      1.00      0.96        26
           8       0.00      0.00      0.00         2

    accuracy                           0.99       229
   macro avg       0.78      0.77      0.77       229
weighted avg       0.98      0.99      0.98       229
```

Gambar 3. Hasil Uji Coba

Dari hasil pengujian, tingkat akurasi yang dihasilkan adalah 0.98, angka tersebut menunjukkan bahwa akurasi yang dihasilkan sangatlah baik. Hasil tersebut juga membuktikan bahwa algoritma Random Forest dapat menghasilkan akurasi yang tinggi. Penggunaan Min-Max Normalization juga dapat memberikan pengaruh dalam peningkatan akurasi yang dihasilkan.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, akurasi yang dihasilkan sebesar 0.98. Jadi dapat disimpulkan bahwa algoritma Random Forest memiliki performa yang bagus untuk menghasilkan nilai akurasi yang tinggi. Min-Max Normalization pada preprocessing data juga berperan terhadap hasil akurasi yang dihasilkan.

Daftar Pustaka

[1] G. A. B. Suryanegara, Adiwijaya, and M. D. Purbolaksono, “Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi,” *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114–122, Feb. 2021, doi: 10.29207/resti. v5i1.2880.

[2] Han, J., Kamber, M., and Pei, J., 2011. Data Mining Concepts and Techniques. (3rd ed.). USA: Morgan Kaufmann.
[3] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, “Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah,” Semin. Nas. Mhs. Ilmu Komput. dan Apl., vol. 2, no. 2, pp. 260–268, 2021.
[4] Suyanto, 2018. Machine Learning Tingkat Dasar dan Lanjut. Bandung: Informatika Bandung.