PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN

Darnisa Azzahra Nasution[1], Hidayah Husnul Khotimah[2], Nurul Chamidah[3]
[1,2,3]*Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta*
Jl. Rs. Fatmawati, Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia
[1]*darnisaazzahran@gmail.com,* [2]*hidayahhk@gmail.com,*[3]*nurul.chamidah@upnvj.ac.id*

Abstrak—**Rentang nilai yang tidak seimbang pada setiap atribut dapat mempengaruhi kualitas hasil data mining. Untuk itu diperlukan adanya praproses data. Praproses ini diharapkan dapat meningkatkatkan keakuratan hasil dari pengklasifikasian dataset wine. Metode praproses yang digunakan adalah transformasi data dengan normalisasi. Ada tiga cara yang dilakukan dalam transformasi data dengan normalisasi, yaitu min-max normalization, *z*-score normalization, dan decimal scaling. Data yang telah diproses dari setiap metode normalisasi akan dibandingan untuk melihat hasil akurasi terbaik klasifikasi dengan menggunakan algoritama K-NN. K yang digunakan dalam perbandingan adalah 1, 3, 5, 7, 9, 11. Sebelum dilakukan pengklasifikasian dataset wine yang telah dinormalisasi dibagi menjadi data uji dan data latih dengan k-fold cross validation. Pembagian data menggunakan k sama dengan 10. Hasil pengujian klasifikasi dengan algoritma K-NN menunjukkan, bahwa akurasi terbaik terletak pada dataset wine yang telah dinormalisasi menggunakan metode min-max normalization dengan K = 1 sebesar 65,92%. Rata-rata yang diperoleh, yaitu 59,68%.**

Kata Kunci— **Normalisasi, K-fold cross validation, K-NN.**

I. PENDAHULUAN

Wine merupakan hasil fermentasi anaerob (tanpa kehadiran O2) dari juice buah anggur berupa minuman beralkohol, oleh khamir. Wine adalah minuman populer yang sangat banyak peminatnya terutama di luar negri. Tidak hanya sebagai penikmat, tetapi sebagian orang yang sering mengkonsumsi berbagai jenis wine berkembang menjadi pakar wine. Pakar wine bertugas untuk melakukan pelabelan terhadap jenis-jenis wine. Maka dari itu dapat dilakukan pengklasifikasian pada data wine untuk mengurangi peran pakar dalam pelebelannya [1].

Praproses merupakan sebuah tahap awal yang harus dilakukan pada data mining. Tujuan praproses dalam data mining adalah untuk mempersiapkan data mentah sebelum dilakukan proses lain. Praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Praproses juga dilakukan untuk mendapatkan hasil yang lebih akurat, pengurangan waktu perhitungan untuk large scale problem, dan membuat nilai data menjadi lebih kecil tanpa merubah informasi yang didalamnya. Praproses data dapat berupa data cleaning, data integration, data reduction, dan data transformation [2].

Pada beberapa dataset terdapat rentang nilai yang berbeda disetiap atribut. Perbedaan rentang nilai pada setiap atribut menyebabkan tidak berfungsinya atribut yang memiliki nilai jauh lebih kecil dibandingkan dengan atribut-atribut lainnya. Oleh karena itu, diperlukan adanya transformasi data dengan normalisasi untuk menyamakan rentang nilai pada setiap atribut dengan skala tertentu. Agar dapat menghasilkan data mining yang lebih baik. Transformasi data dengan normalisasi dapat dilakukan dengan beberapa cara, yaitu min-max normalization, z-score normalization, decimal scaling, sigmoid, dan softmax.

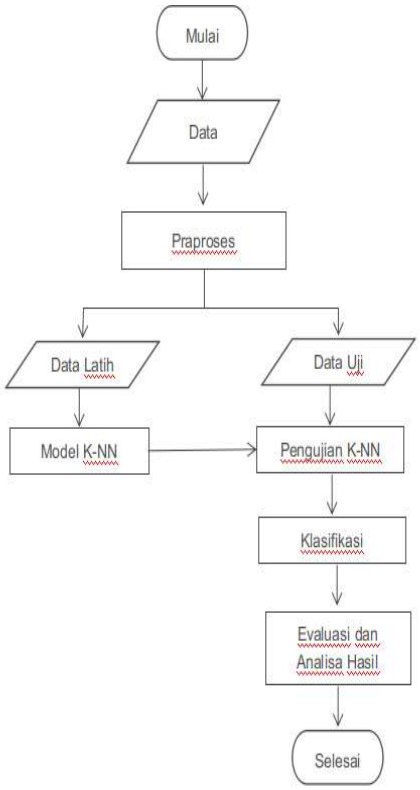
Klasifikasi meupakan salah satu tahap penting dalam data mining. Klasifikasi adalah pengelompokkan data atau objek baru ke dalam kelas atau label berdasarkan atribut-atribut tertentu [9]. Teknik dari klasifikasi adalah dengan melihat variabel dari kelompok data yang sudah ada. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang tidak diketahui sebelumnya. Klasifikasi terdiri dari tiga tahap, yaitu pembangunan model, penerapan model, dan evaluasi. Pembangunan model adalah membangun model menggunakan data latih yang telah memiliki atribut dan kelas. Kemudian, data-data tersebut diterapkan untuk menentukan kelas dari data atau objek yang baru. Setelah itu, data dievaluasi untuk melihat tingkat akurasi dari pembangunan dan penerapan model terhadap data baru [10]. Proses klasifikasi terdiri dari dua fase, yaitu fase training dan fase testing. Fase training adalah fase di mana data digunakan untuk membangun sebuah model sedangkan fase testing adalah pengujian model yang telah dibuat dengan data lainnya untuk mengetahui akurasi dari model tersebut [3].

Penelitian yang berhubungan dengan pengklasifikasian wine pernah dilakukan sebelumnya yang pertama diperoleh hasil akurasi sebesar 68,75%[8]. Kemudian yang kedua pengklasifikasian menggunakan algoritma k-Nearest Neighbor (K-NN)

dengan menerapkan metode k-fold cross validation (k = 3) dalam pembagian data menghasilkan akurasi lebih baik, yaitu sebesar 72.97% [4].
Berdasarkan indikator diatas, maka kami akan melakukan penelitian pengaruh transformasi data dengan metode normalisasi untuk hasil akurasi pada klasifikasi menggunakan algoritma K-NN.

II. METODOLOGI PENELITIAN

Beberapa tahap penelitian dituangkan dalam diagram sebagai berikut :



Gbr 1. Flowchart

A. *Dataset Penelitian*

Dataset yang digunakan adalah dataset wine diambil dari UCI Machine Learning. Dataset wine berjumlah 1599 data. Dataset tersebut memiliki sebelas atribut dan satu *output attribute* berupa kelas dengan rentang 0-10. Rentang tersebut ditampilkan dalam bentuk huruf, yaitu z=0, a=1, b=2, c=3, d=4, e=5, f=6, g=7, h=8, i=9, j=10.

B. *Praproses*

Tahap praproses dilakukan sebagai tahap awal dan tahap penting dalam penelitian. Metode yang digunakan dalam penilitian ini adalah tranformasi data menggunakan normalisasi. Normalisasi adalah proses penskalaan nilai atribut dari data sehingga bisa terletak pada rentang tertentu [5]. Berikut beberapa tahap normalisasi yang dilakukan :

1) Min-Max Normalization: *Min-Max normalization* merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses [6]. Metode ini dapat menggunakan rumus sebagai berikut :

$$normalized(x) = \frac{minRange + (x - minValue)(maxRange - minRange)}{maxValue - minValue}$$

Gbr.2 Persamaan 1

2) Z-score Normalization: Z-score normalization merupakan metode normalisasi berdasarkan mean (nilai rata-rata) dan standard deviation (deviasi standar) dari data. Metode ini sangat berguna jika tidak diketahui nilai aktual minimum dan maksimum dari data. Rumus yang digunakan sebagai berikut :

$$nilaibaru = \frac{nilailama - mean}{stdev}$$

Gbr.3 Persamaan 2

3) Decimal Scaling Normalization: Decimal scaling merupakan metode normalisasi dengan menggerakkan nilai desimal dari data ke arah yang diinginkan. Formula yang digunakan sebagai berikut :

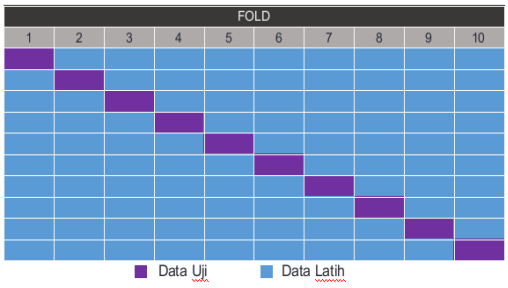
$$nilaibaru = \frac{nilailama}{10^{i}}$$

Gbr.4 Persamaan 3

C. *Klasifikasi*

Tahap klasifikasi merupakan tahap untuk mengklasifikasikan kualitas pada dataset wine. Tahap klasifikasi sebagai berikut :

1) K-fold Cross Validation: Cross validation adalah teknik validasi model untuk menilai keakuratan hasil analisis. Data yang sudah di praproses dilakukan cross validation dengan membagi data menjadi data latih dan data uji untuk proses klasifikasi [7]. Pembagian data dilakukan menggunakan k-fold cross validation dengan nilai k sama dengan 10.



Gbr. 5 Ilustrasi 10-fold *cross validation*

2) K-Nearest Neighbor: Setelah pembagian data uji dan data latih, dilanjutkan proses klasifikasi dengan menggunakan K-NN. Konsep dasar dari K-NN adalah mencari jarak terdekat antara data yang akan dievaluasi dengan k tetangga terdekatnya. Nilai dari jarak antara data uji dengan data latih diurutkan dari nilai terendah. Proses pengurutan tersebut dilakukan untuk memilih jarak minimum sebanyak K buah. Nilai k yang digunakan dalam penelitian ini, yaitu 3, 5, 7, dan 11. Perhitungan dilakukan dengan persamaan sebagai berikut :

$$d(x,y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$$

Gbr.6 Persamaan 4

D. Evaluasi

Tahap evaluasi dilakukan dengan cara menganalisa akurasi dataset wine. Perhitungan akurasi dilakukan dengan cara membagi jumlah data uji yang benar(true positive dan true negative) dengan jumlah data uji keseluruhan kemudian dikalikan dengan 100%.

$$akurasi=\frac{true\ positive+true\ negative}{true\ positive+true\ negative+false\ positive+false\ negative}$$

Gbr.7 Persamaan 5

III. HASIL DAN PEMBAHASAN

Tahap ini merupakan penguraian hasil penelitian yang diperoleh beserta penjelasannya.

A. Praproses

Dalam tahap ini dilakukan praproses terhadap dataset wine. Ada beberapa tahap praproses yang dilakukan, diantaranya *min-max normalization, z-score normalization, decimal scaling.*

TABEL I
DATASET WINE

No.	Fixed acidity	volatile acidity	Citric acid	….	quality
1	7.4	0.7	0	….	e
2	7.8	0.88	0	….	e
3	7.8	0.76	0.04	….	e
4	11.2	0.28	0.56	….	f
…..	…….	…….	…….....	….	…….
1599	6	0.31	0.47	….	f

1) Min-Max Normalization: Pada tabel I dapat dilihat nilai asli dari dataset wine sebelum di praproses. Banyak data yang memiliki rentang berbeda sehingga harus dilakukan normalisasi. Dataset tersebut ditranformasi menggunakan metode *min-max normalization* dengan mengolah nilai minimum dan maksimum dari setiap atribut. Rentang yang digunakan dalam metode ini adalah 0-1. Rumus pada persamaan 1 dapat digunakan untuk melakukan normalisasi dengan metode *min-max normalization*. Hasil pengubahan nilai dengan metode ini dapat dilihat pada tabel 2. Nilai yang dihasilkan setelah pengolahan memiliki rentang nilai yang seimbang.

TABEL II
DATASET WINE SETELAH DILAKUKAN *MIN-MAX NORMALIZATION*

No.	Fixed acidity	volatile acidity	Citric acid	….	quality
1	0.247788	0.39726	0	….	e
2	0.283186	0.520548	0	….	e
3	0.283186	0.438356	0.04	….	e
4	0.584071	0.109589	0.56	….	f
…..	…….	…….	……...	….	…….
1599	0.283186	0.130137	0.47	….	f

2) Z-score Normalization: Dataset wine asli yang dapat dilihat pada tabel I akan dilakukan transformasi ulang dengan metode yang berbeda. Metode selanjutnya yang digunakan adalah z-score normalization. Rumus yang digunakan dalam metode ini dapat dilihat pada persamaan 2. Z-score normalization dilakukan dengan mengolah mean dan standar deviasi dari nilai-nilai atributnya. Hasil transformasi dari metode ini dapat dilihat pada tabel III.

TABEL III
DATASET WINE SETELAH DILAKUKAN *Z-SCORE NORMALIZATION*

No.	Fixed acidity	volatile acidity	Citric acid	….	quality
1	-0.528194	0.961576	-1.391037	….	e
2	-0.298454	1.966827	-1.391037	….	e
3	-0.298454	1.29666	-1.185699	….	e
4	-1.332285	-1.384011	1.483689	….	f
…..	…….	…….	……...	….	…….
1599	-1.332285	-1.216469	1.02168	….	f

3) Decimal Scaling: Untuk perbandingan selanjutnya data wine akan di praproses ulang dengan metode transformasi data yaitu decimal scaling. Tabel IV merupakan hasil dari normalisasi dengan metode decimal scaling.

TABEL IV
DATASET WINE SETELAH DILAKUKAN DECIMAL SCALING

No.	Fixed acidity	volatile acidity	Citric acid	…	quality
1	0.074	0.07	0	……	e
2	0.078	0.088	0	……	e
3	0.078	0.076	0.04	……	e
4	0.112	0.028	0.56	……	f
…..	…….	…….	…… …..	……	…… ….
1599	0.06	0.031	0.47	……	f

Dalam menghitung nilai baru tersebut digunakan rumus pada persamaan 3. Dapat dilihat perubahan yang terjadi, nilai yang dihasilkan setiap atribut memiliki rentang yang tidak terlalu jauh.

B. Klasifikasi

Setelah praproses data dan sebelum dilakukan klasifikasi, data dibagi menjadi data uji dan data latih terlebih dahulu. Pembagian data dilakukan dengan *k-fold cross validation.* K yang digunakan pada k-fold sama dengan 10. Tahap selanjutnya adalah melakukan klasifikasi terhadap dataset wine yang telah di praproses menggunakan algoritma K-NN. K yang digunakan, yaitu 3, 5, 7, 11. Penghitungan jarak pada algoritma ini dapat dilihat pada persamaan 4. Berikut hasil tabel penelitian.

TABEL V
HASIL AKURASI DENGAN METODE NORMALISASI MENGGUNAKAN K-NN

K-NN	Metode Normalisasi		
	Decimal scaling	Min-max normalization	Z-score normalization
K = 1	63,10%	65,92%	65,85%
K = 3	52,47%	59,35%	59,22%
K = 5	53,22%	57.41%	56,60%
K = 7	50,47%	58,03%	57,54%
K = 9	51,66%	58,66%	57,60%
K = 11	51,47%	58.72%	57,85%
Mean	53,73%	59,68%	59,11%

Tahap akhir pada proses ini adalah penghitungan akurasi menggunakan rumus pada persamaan 5. Hasil akurasi dapat dilihat pada tabel V. Perbandingan antara akurasi hasil metode *min-max normalization*, *z-score normalization*, *decimal scaling* menunjukkan bahwa akurasi tertinggi terletak pada data yang dipraproses menggunakan metode *min-max normalization* dengan rata-rata akurasi sebesar 59,68%.

IV. PENUTUP

Dari hasil penelitian yang sudah dilakukan didapatkan kesimpulan, yaitu

- Akurasi tertinggi terletak pada dataset wine yang menggunakan metode min-max normalization pada tahap praprosesnya dengan K = 1 sebesar 65,92%.
- Akurasi dengan rata-rata tertinggi adalah 59,68% yaitu menggunakan metode min-max normalization.
- Akurasi terendah terdapat pada dataset dengan metode decimal scaling, yaitu dengan rata-rata 53,73%.
- Pemilihan metode praproses data pada data mining mempengaruhi akurasi dari hasil klasifikasi data.
- Penelitian ini menemukan bahwa akurasi menggunakan praproses data dengan metode normalisasi tidak lebih baik dari akurasi penelitian sebelumnya yang dilakukan oleh Arandika et al. 2014 dengan tingkat akurasi sebesar 68,75% dan penelitian Saputra&Siahaan. 2007 sebesar 72,97%.

REFERENSI

[1] P. A. Minum, "Non alcohol."

[2] K. Saputra and A. P. U. Siahaan, "Klasifikasi Data Minuman Wine Menggunakan Algoritma K-Nearest Neighbor," pp. 2–4, 2007.

[3] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[4] A. C. Imanda, N. Hidayat, and M. T. Furqon, "Klasifikasi Kelompok Varietas Unggul Padi Menggunakan Modified K-Nearest Neighbor," vol. 2, no. 8, pp. 2392–2399, 2018.

[5] P. Studi, T. Informatika, J. T. Informatika, F. Sains, D. A. N. Teknologi, and U. S. Dharma, "Deteksi Outlier Pada Data Campuran Numerik Dan Kategorikal Menggunakan Algoritma Enhanced Class Outlier Distance Based (Ecodb) Algoritma Enhanced Class Outlier Distance Based (Ecodb)."

[6] T. T. Hanifa, S. Al-faraby, F. Informatika, and U. Telkom,"Analisis Churn Prediction pada Data Pelanggan PT . Telekomunikasi dengan Logistic Regression dan Underbagging," vol. 4, no. 2, pp. 3210–3225, 2017.

[7] R. E. Putri, Suparti, and R. Rahmawati, "Perbandingan Metode Klasifikasi Naïve Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012," *J. Gaussian*, vol. 3, pp. 831–838, 2014.

[8] Arandika A, Mardji, Cholisson I. Implementasi Algoritma K-Neaerest Neighbor (K-NN) Untuk Klasifikasi Data Wine. Jurnal Mahasiswa PTIIK UB.Volume 4, Number 12. 2014.

[9] Septianto, Ryan Hendy. 2015. Diagnosa Penyakit Tanaman Kopi Arabika dengan Metode Modified K-Nearest Neighbor (MK-NN). Skripsi. Universitas Brawijaya, Malang.

[10] Kumalasari, Noviana Ayu. 2014. Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Menentukan Tingkat Resiko Penyakit Lemak Darah (Profil Lipid). Skripsi. Universitas Brawijaya, Malang.