JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

IMPLEMENTASI CORRELATION MATRIX PADA KLASIFIKASI DATASET WINE

Erfin Nur Rohma Khakim[1*], **Arief Hermawan**[2], dan **Donny Avianto**[3]
[1,2]Magister Teknologi Informasi, Universitas Teknologi Yogyakarta
[3]Informatika, Universitas Teknologi Yogyakarta
Email: erfinnrk@gmail.com[1], ariefdb@uty.ac.id[2] donny@uty.ac.id[3],

Abstrak

Wine merupakah minuman mengandung alkohol yang juga terdiri dari banyak sekali kandungan yang bermacam-macam yang dapat mempengaruhi kualitas wine. Begitu banyaknya jenis wine ini membuat masyarakat kesulitan untuk memilah jenis-jenis dan kualitas wine. Penelitian ini menggunakan metode klasifikasi untuk menentukan kualitas wine agar mengurangi peran pakar wine atau mempermudah pakar dalam melabeli jenis-jenis dan kualitas wine. Algoritma klasifikasi data mining yang digunakan dalam penelitian ini adalah k-NN. Untuk meningkatkan nilai akurasi dari klasifikasi algoritma k-NN digunakan metode normalisasi dan seleksi fitur. Normalisasi yang digunakan adalah normalisasi Z-sccore dan Min-max, sedangkan seleksi fitur yang digunakan untuk lebih meningkatkan akurasi adalah correlation matrix. Hasil dari penelitian ini membutktikan bahwa penggunaan seleksi fitur correlation matrix mampu meningkatkan nilai akurasi pada normalisasi Z-score dari 73,75% menjadi 75,62% dan pada normalisasi Min-max mampu meningkatkan akurasi dari 68,12% menjadi 71,25%. Sehingga bisa disimpulkan bahwa seleksi fitur correlation matrix dapat digunakan untuk meningkatkan nilai akurasi agar lebih tinggi dan penelitian menjadi lebih akurat.
Kata Kunci: *Klasifikasi, k-NN, Seleksi fitur, Correlation matrix, Normalisasi*

Abstract

Wine is an alcoholic drink with various ingredients that can affect the quality of the wine. So many types of wine make it difficult for people to sort out the types and quality of the wine. This study uses a classification method to determine the quality of wine to reduce the role of wine experts or make it easier for experts to label the types and quality of the wine. The data mining classification algorithm used in this study is k-NN. To increase the accuracy value of the k-NN classification algorithm, normalization and feature selection methods are used. The normalization used are Z-score and Min-max normalization, and the feature selection used to improve accuracy is the correlation matrix. This study's results proved that using a feature selection correlation matrix could increase the accuracy value in Z-score normalization from 73.75% to 75.62%, and in Min-max normalization can increase accuracy from 68.12% to 71.25%. So it can be concluded that the correlation matrix feature selection can be used to increase the accuracy value and make it higher, and the study will be more accurate.
Kata Kunci: *Klasifikasi, k-NN, Seleksi fitur, Correlation matrix, Normalisasi*

I. PENDAHULUAN

Wine adalah jenis minuman yang cukup populer di luar negeri. Wine merupakah salah satu minuman yang mengandung alkohol dan menjadi produk yang sangat dibatasi dan diatur di Indonesia. Meskipun demikian, wine bukan berarti tidak bermanfaat sama sekali, ada beberapa kandungan dalam wine yang bermanfaat bagi kesehatan jika dikonsumsi tidak secara berlebihan. Beberapa manfaat wine bagi kesehatan antara lain untuk mencegah penyakit kanker, memelihara kesehatan otak dan dungsi memori, menajaga kesehatan jantung, menjaga kesehatan gigi dan mulut serta menurunkan kadar gula darah [1]. Wine dibuat dari fermentasi buah anggur atau beberapa terbuat dari buah lain yang difermentasikan sehingga kaya akan antioksidan [2].

Wine terdiri dari banyak sekali kandungan yang bermacam-macam. Satu jenis wine memiliki kandungan yang berbeda dengan jenis yang lain. Tentu saja jumlah dari kandungan-kandungan di setiap jenis ini mempengaruhi kualitas dari minuman wine. Beberapa kandungan dalam wine yang mempengaruhi kualitasnya di anatara lain ada kandungan *citric acid*, besaran pH, kandungan *fixed acidity* dan lain-lain. Selain kandungannya sebagai hasil akhir setelah wine jadi, kualitas wine juga dapat ditentukan berdasarkan komposisi bahan baku dan fermentasinya selama proses pengolahan dari buah-buahan.

Mengingat begitu banyak jenis wine di dunia ini, maka kualitasnya pun juga sangat beragam. Begitu banyaknya jenis wine ini membuat masyarakat kesulitan untuk memilah jenis-jenis dan kualitas wine. Di berbagai wilayah di dunia ini banyak para penikmat wine yang menjadi pakar. Pakar inilah yang bertugas melabeli wine berdasarkan jenis dan kualitasnya. Metode klasifikasi dapat diterapkan sebagai salah satu cara untuk mengurangi peran pakar wine atau mempermudah pakar dalam melabeli jenis-jenis dan kualitas wine.

Proses klasifikasi menggunakan algoritma data mining dilakukan tidak secara manual, tetapi melalui teknik melihat variable dan kelas dari data yang sudah ada. Proses klasifikasi dapat memprediksi data baru yang belum diketahui kelasnya berdasarkan model dari data yang sudah ada. Klasifikasi terdiri dari tiga tahap, yaitu pembangunan model, penerapan model, dan evaluasi . Dalam penelitian ini, dataset wine dengan variable dan kelas yang pernah ditentukan oleh pakar digunakan untuk membangun model, kemudian model akan diterapkan terhadap data wine baru yang belum ada kelasnya, setelahnya, hasil klasifikasi data baru dibandingkan dengan pembagian kelas kualitas wine dari para pakar untuk membandingkan pembangunan dan penerapan model.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

Metode klasifikasi k-NN diterapkan dalam penelitian ini untuk mengolah dataset wine. Setelah dilakukan klasifikasi dengan algoritma k-NN maka hasil dari klasifikasi ini akan diuji dan dievaluasi untuk mendapatkan nilai akurasi. Penelitian sebelumnya yang pernah untuk mengklasifikasi dataset wine adalah penelitian dengan judul Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma k-NN [3]. Penelitian tersebut mengklasifikasikan dataset wine menggunakan algoritma k-NN. Hanya saja, penelitian ini berfokus pada penggunaan normalisasi data untuk meningkatkan akurasi klasifikasi. Penelitian tersebut membandingkan beberapa metode normalisasi seperti *Min-max*, *Z-score*, dan *Decimal scaling* dengan *cross validation* untuk mendapatkan hasil akurasi yang terbaik. Berdasarkan percobaan yang telah dilakukan, didapatkan nilai akurasi tertinggi pada nilai *cross validation* k=1. Pada k=1 normalisasi *Min-max* dihasilkan nilai akurasi 63,10%, untuk normalisasi *Z-score* dihasilkan nilai akurasi 65,92%, dan untuk normalisasi *Decimal scaling* dihasilkan nilai akurasi 65,85%. Berdasarkan hasil tersebut dapat disimpulkan bahwa normalisasi *Z-score* dapat menghasilkan nilai akurasi tertinggi.

Penelitian lain yang membahas klasifikasi dataset wine adalah penelitian judul Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah [4]. Penelitian kedua ini fokus pada perbandingan performa dari beberapa algoritma klasifikasi. Algoritma klasifikasi yang digunakan dalam penelitian ini antara lain *Decision Tree*, *Random Forest* dan *Support Vector Machine* dengan *cross validation*. Hasil penelitian tersebut menjabarkan bahwa *Decision Tree* memiliki nilai akurasi sebesar 70,31%, dengan nilai *Area Under Curve* (AUC) 70,00% dan F1 Score 72,93%. Algoritma *Random Forest* menunjukan hasil akurasi sebesar 74,68%, dengan nilai *Area Under Curve* (AUC) 74,68% dan F1 Score 74,92%, sedangkan *Support Vector Machine* menunjulan hasil akurasi 65%, dengan nilai *Area Under Curve* (AUC) 63,73% dan F1 Score 70,83%. Sehingga kesimpulan dari penelitian terbut adalah bahwa algoritma klasifikasi *Random Forest* menghasilkan performa yang tertinggi dibandingkan *Decision Tree* dan *Support Vector Machine*.

Dua penelitian terdahulu yang terlah dijabarkan sebelumnya menunjukkan bahwa hasil akurasi terhadap teknik *data mining* klasifikasi masih dibawah 75%. Penelitian ini mencoba untuk meningkatkan performa akurasi dari klasifikasi dataset wine dengan menggunakan *feature selection*. *Feature selection* dengan *correlation matrix* digunakan untuk meningkatkan akurasi dibandingkan dengan penelitian-penelitian sebelumnya. *Correlation matrix* digunakan untuk mencari atribut-atribut dalam dataset yang berkorelasi dan bisa dihilangkan.

Data mining merupakan sebuah proses menggali informasi dari sekumpulan data yang sudah ada. Informasi yang berhasil digali ini dapat dijadikan sebagai pendukung keputusan dalam sebuah bisnis. Teknik data mining dapat digunakan untuk memecahkan masalah dengan menganalisis data yang telah ada. *Data mining* adalah analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara berbeda dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data [5]. Dalam definisi lain, data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [6].

Salah satu teknik *data mining* adalah klasifikasi. Klasifikasi adalah suatu Teknik data mining yang memodelkan data yang telah ada kelasnya kemudian melakukan pengujian model terhadap data yang belum ada kelasnya. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada [7]. Klasifikasi merupakan Teknik dalam data mining yang masuk dalam kategori supervised, yaitu dilakukan pada data yang telah memiliki label datau kelas.

Penelitian ini menggunakan algoritma *k-Nearest Neighbor* yang merupakan salah satu metode klasifikasi. *k-Nearest Neighbor* (k-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut [8]. Hal ini memudahkan dalam pencarian klasifikasi atau penentuan prediksi dari sebuah variabel. K-NN merupakan algoritma klasifikasi yang fokus pada jarak satu objek dengan tetangga terdekatnya. Ada beberapa rumus yang dapat digunakan dalam menghitung jarak objek terhadap tetangga terdekatnya. Euclidien dan *Cosine similarity* adalah rumus yang paling sering digunakan. Persamaan (1) adalah perhitungan jarak dengan rumus euclidien [8].

$$D(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_1 - x_2)^2} \tag{1}$$

dimana D adalah jarak terdekat, x_1 adalah sampel data atau *data training*, x_2 adalah data uji atau *data testing*, n adalah jumlah atribut setiap kasus dan i adalah atribut individu dari 1 sampai n. Persamaan (2) adalah perhitungan jarak dengan rumus *Cosine similarity* [9].

$$cos(i,k) = \frac{\sum_k (d_i d_k)}{\sqrt{\sum_k d_{ik}^2 \sqrt{\sum_k d_{jk}^2}}} \tag{2}$$

dimana $\sum_k (d_i d_k)$ adalah vektor dari produk i dan k, $\sum_k d_{ik}^2$ adalah panjang dari vektor i, $\sum_k d_{jk}^2$ adalah panjang dari vektor j, i adalah data uji ke-i dan j adalah data latih ke-j.

Preprocesing data adalah proses untuk mengubah data mentah menjadi lebih teratur agar Ketika dilakukan teknik data mining tingkat akurasinya lebih tinggi. *Preprocesing* dilakukan untuk membuat data menjadi lebih bersih. Adapun tujuan utama dalam *preprocessing* data ini ialah sebagai berikut [10]: Pembersihan Data, yaitu mengisi nilai yang hilang, menghaluskan *noise* data, mengidentifikasi dan menghapus *outlier* serta menyelesaikan inkonsistensi. Normalisasi adalah salah satu teknik preprocesing

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

untuk menghapus *outlier*. Dua jenis normalisasi yang paling serig digunakan adalah *Z-score* dan *Min-max*. Normalisasi *Z-score* atau dikenal dengan standarisasi merupakan teknik yang mana nilai pada atribut akan dinormalisasi berdasarkan *mean* dan standar deviasi [11]. Persamaan (3) adalah rumus normalisasi *Z-score* [12].

$$Z = \frac{x - \bar{x}}{\sigma} \tag{3}$$

dimana x adalah nilai yang diamati, $\bar{x}$ adalah nilai rata-rata dan σ adalah standar deviasi. Sedangkan normalisasi *Min-max* adalah metode yang mengubah sebuah kumpulan data menjadi skala mulai dari 0 (*min*) hingga 1 (*max*) [11]. Persamaan (4) adalah rumus normalisasi *Min-max* [13].

$$X = \frac{MinRange + (X - MinValue)(MaxRange - MinRange)}{MaxValue - MinValue} \tag{4}$$

dimana $MinRange$ adalah nilai konversi terkecil yang ditentukan, $MaxRange$ adalah nilai konversi terbesar yang ditentukan, $MinValue$ adalah nilai terkecil pada atribut yang dibanddingkan dan $MaxValue$ adalah nilai terbesar pada atribut yang dibandingkan.

Salah satu cara untuk meningkatkan performa dari hasil klasifikasi algoritma k-NN adalah melakukan seleksi fitur. Seleksi fitur berfungsi untuk menentukan suatu kelas pada nilai target dengan cara mengurangi jumlah fitur yang tidak relevan serta mengurangi dimensi data untuk meningkatkan performa sistem, efisiensi dan meningkatkan akurasi [14]. Seleksi fitur yang digunakan dalam penelitian ini adalah *Correlation matrix*. Prinsip dari *Correlation matrix* adalah mencari hubungan antar atribut yang berkorelasi paling erat, kemudian menghilangkan salah satu dari atribut yang saling berkorelasi tersebut.

Evaluasi dan pengujian dari performa algoritma k-NN dengan *Correlation matrix* diukur dengan nilai akurasi. Dengan mencari nilai tertinggi dari akurasi maka diperoleh performa terbaik dalam sebuah algoritma. Persamaan (5) adalah rumus akurasi [15].

$$Accuration = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

dimana TP adalah *true positif*, TN adalah *true negative*, FP adalah *false positif* dan FN adalah *false Negative*.

Rapidminer adalah perangkat lunak yang bersifat terbuka (*open source*). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi [16]. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *Rapidminer* merupakan perangkat lunak untuk melakukan *data mining* yang paling mudah karena berbasis grafis dengan metode *drag* operator.

II. Metode Penelitian

A. Sumber Data Penelitian

Sumber data yang digunakan dalam penelitian ini adalah data sekunder dari UCI *Machine Learning*. Dataset wine terdiri 1599 baris. Atribut dari dataset wine berjumlah 12 yang terdiri dari 11 atribut reguler dan satu *output* atribut atau kelas atau label. Sebelas atribut reguler tersebut adalah *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, pH, *sulphates* dan *alcohol*. Sedangkan satu atribut yang menjadi kelas atau labelnya adalah atribut *quality*. Adapun metode pengumpulan data yang digunakan adalah studi Pustaka.

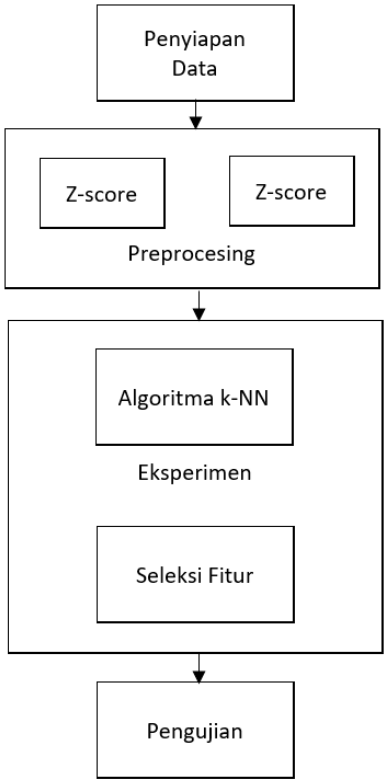
B. Metode Penelitian

Metode penelitian yang digunakan adalah metode penelitian terapan deskriptif. Metode ini bertujuan untuk melakukan penerapan, melakukan pengujian serta melakukan evaluasi terhadap suatu teori algoritma dalam proses memecahkan suatu permasalahan. Penelitian ini bertujuan untuk membandingkan klasifikasi dataset wine hanya dengan menggunakan algoritma k-NN dan klasifikasi dataset wine menggunakan analisis korelasi atribut sebelum diterapkan algoritma k-NN.

C. Tahapan Penelitian

Tahapan penelitian ini digambarkan dalam sebuah diagram pada Gambar 1 sebagai berikut.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771



Gambar 1: Diagram Tahapan Penelitian

Tahapan pertama yang dilakukan dalam penelitian ini adalah penyiapan data. Data diambil dari UCI *Machine Learning*. Wine memiliki 1599 baris data. Atribut dari dataset wine berjumlah 12 yang terdiri dari 11 atribut reguler dan satu output atribut atau kelas atau label. Atribut yang ada dalam dataset wine ditunjukkan pada Tabel I sebagai berikut.

Tabel I: Atribut Dataset Wine

No	Nama Atribut	Role Atribut
1	*fixed acidity*	Atribut
2	*volatile acidity*	Atribut
3	*citric acid*	Atribut
4	*residual sugar*	Atribut
5	*chlorides*	Atribut
6	*free sulfur dioxide*	Atribut
7	*total sulfur dioxide*	Atribut
8	*density*	Atribut
9	pH	Atribut
10	*sulphates*	Atribut
11	*alcohol*	Atribut
12	*quality*	Kelas/Label

Pada Table I diperlihatkan bahwa dataset wine terdiri dari 11 atribut reguler tersebut adalah *fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates* dan *alcohol*. Sedangkan satu atribut kelas adalah *quality*.

Tahapan kedua adalah *preprocesing* data. Pada tahapan inilah digunakan metode normalisasi *Z-score* dan normalisasi *Min-max*. Normalisasi *Z-score* digunakan untuk mengubah nilai data dengan nilai rata-rata atau deviasi. Sedangkan normalisasi *Min-max* akan merubah data ke dalam rentang nilai 0.0 – 1.0. Tujuan dari normalisasi ini adalah untuk menghilangkan *outlier* data.

Tahapan ketiga adalah tahapan inti yaitu eksperimen. Setelah dilakukan normalisasi, maka dilakukan analisis korelasi. Metode yang digunakan dalam analisis korelasi ini menggunakan *Correlation matrix*. Dengan *correlation matrix*, dapat diketahui nilai korelasi antar atribut dalam dataset wine. Nilai korelasi terbesar dari dua atribut atau lebih akan menjadi penentu atribut yang akan dihilangkan. Setelah menghilangkan beberapa atribut yang berkorelasi sangat erat dengan atribut lain, maka langkah selanjutnya adalah melakukan pembagian data sebelum menerapkan algoritma k-NN. Pembagian data dimaksudkan untuk membagi data latih dan data uji. Pembagian data diujicoba dalam beberapa kombinasi. Setelah data dibagi ke dalam data latih

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

dan data uji maka diterapkan algoritma k-NN untuk klasifikasi dan dicoba dalam beberapa nilai K untuk mengetahui hasil terbaik.

Tahap terakhir dari penelitian ini adalah pengujian. Setelah diterapkan klasifikasi dengan algoritma k-NN maka dilakukan pengujian untuk mengetahu performa dari dari setiap kombinasi percobaan. Pengujian performa dicoba dengan dua macam normalisasi, nilai split data yang berbeda dan dengan nilai K yang juga berbeda-beda.

III. Hasil dan Pembahasan

A. Penyiapan Data

Dataset wine yang diambil dari UCI *Machine Learning* memiliki 1599 baris data dengan 11 atribut reguler dan satu atribut sebagai kelas atau label. Dataset wine ditampilkan dalam *operator retrieve* pada *rapidminer* seperti ditunjukkan pada Gambar 2 sebagai berikut.

Row No.	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density
1	5	7.400	0.700	0	1.900	0.076	11	34	0.998
2	5	7.800	0.880	0	2.600	0.098	25	67	0.997
3	5	7.800	0.760	0.040	2.300	0.092	15	54	0.997
4	6	11.200	0.280	0.560	1.900	0.075	17	60	0.998
5	5	7.400	0.700	0	1.900	0.076	11	34	0.998
6	5	7.400	0.660	0	1.800	0.075	13	40	0.998
7	5	7.900	0.600	0.060	1.600	0.069	15	59	0.996
8	7	7.300	0.650	0	1.200	0.065	15	21	0.995
9	7	7.800	0.580	0.020	2	0.073	9	18	0.997
10	5	7.500	0.500	0.360	6.100	0.071	17	102	0.998
11	5	6.700	0.580	0.080	1.800	0.097	15	65	0.996
12	5	7.500	0.500	0.360	6.100	0.071	17	102	0.998
13	5	5.600	0.615	0	1.600	0.089	16	59	0.994
14	5	7.800	0.610	0.290	1.600	0.114	9	29	0.997
15	5	8.900	0.620	0.180	3.800	0.176	52	145	0.999

ExampleSet (1,599 examples, 1 special attribute, 11 regular attributes)

Gambar 2: Dataset wine

Gambar 2 tersebut memperlihatkan dataset wine dalam *tools rapidminer*. Pada bagian bawah terlihat jelas bahwa informasi jumlah baris dan atribut dataset wine.

B. Preprocesing Data

Pada tahapan *preprocesing* mulai dilakukan normalisasi data. Normalisasi dilakukan dengan dua metode yaitu normalisasi *Z-score* dan normalisasi *Min-max*. Normalisasi ini menyebabkan dataset wine berubah nilai sesuai dengan tujuan dari normalisasi. Hasil normalisasi menggunakan metode *Z-score* ditunjukkan deperti pada Gambar 3 sebagai berikut.

Row No.	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density
1	5	-0.528	-0.053	-1.391	-0.453	-0.244	-0.466	-0.379	-0.128
2	5	-0.298	-0.050	-1.391	0.043	0.224	0.872	0.624	-0.128
3	5	-0.298	-0.052	-1.186	-0.169	0.096	-0.084	0.229	-0.128
4	6	1.654	-0.060	1.484	-0.453	-0.265	0.107	0.411	-0.128
5	5	-0.528	-0.053	-1.391	-0.453	-0.244	-0.466	-0.379	-0.128
6	5	-0.528	-0.054	-1.391	-0.524	-0.265	-0.275	-0.197	-0.128
7	5	-0.241	-0.055	-1.083	-0.666	-0.392	-0.084	0.381	-0.128
8	7	-0.586	-0.054	-1.391	-0.950	-0.477	-0.084	-0.774	-0.128
9	7	-0.298	-0.055	-1.288	-0.382	-0.307	-0.657	-0.865	-0.128
10	5	-0.471	-0.056	0.457	2.526	-0.350	0.107	1.688	-0.128
11	5	-0.930	-0.055	-0.980	-0.524	0.203	-0.084	0.563	-0.128
12	5	-0.471	-0.056	0.457	2.526	-0.350	0.107	1.688	-0.128
13	5	-1.562	-0.054	-1.391	-0.666	0.033	0.012	0.381	-0.128
14	5	-0.298	-0.054	0.098	-0.666	0.564	-0.657	-0.531	-0.128
15	5	0.333	-0.054	-0.467	0.895	1.881	3.453	2.995	-0.128

ExampleSet (1,599 examples, 1 special attribute, 11 regular attributes)

Gambar 3: Hasil Normalisasi Z-score

Pada Gambar 3 ditunjukkan nilai isi dari dataset wine berubah. Nilai tersebut telah disesuai dengan menghitung nilai rata-rata dan standar deviasi sesuai dengan rumus normalisasi *Z-score*. Adapun hasil dari normalisasi *Min-max* ditunjukkan pada Gambar 4 sebagai berikut.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

Row No.	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density
1	5	0.248	0.000	0	0.068	0.107	0.141	0.099	0.000
2	5	0.283	0.001	0	0.116	0.144	0.338	0.216	0.000
3	5	0.283	0.001	0.040	0.096	0.134	0.197	0.170	0.000
4	6	0.584	0.000	0.560	0.068	0.105	0.225	0.191	0.000
5	5	0.248	0.000	0	0.068	0.107	0.141	0.099	0.000
6	5	0.248	0.000	0	0.062	0.105	0.169	0.120	0.000
7	5	0.292	0.000	0.060	0.048	0.095	0.197	0.187	0.000
8	7	0.239	0.000	0	0.021	0.088	0.197	0.053	0.000
9	7	0.283	0.000	0.020	0.075	0.102	0.113	0.042	0.000
10	5	0.257	0.000	0.360	0.356	0.098	0.225	0.339	0.000
11	5	0.186	0.000	0.080	0.062	0.142	0.197	0.208	0.000
12	5	0.257	0.000	0.360	0.356	0.098	0.225	0.339	0.000
13	5	0.088	0.000	0	0.048	0.129	0.211	0.187	0.000
14	5	0.283	0.000	0.290	0.048	0.170	0.113	0.081	0.000
15	5	0.381	0.000	0.180	0.199	0.274	0.718	0.491	0.000

ExampleSet (1,599 examples, 1 special attribute, 11 regular attributes)

Gambar 4: Hasil Normalisasi Min-max

Pada Gambar 4 juga ditunjukkan bahwa nilai dari dataset wine telah berbeda dibandingkan dengan data awal. Hal ini terjadi setelah nilai dari data set diubah ke dalam rentang nilai 0.0 sampai dengan 1.0 sesuai dengan rumus normalisasi *Min-max*.

C. *Eksperimen*

Eksperimen dilakukan dalam dua tahap. Tahap pertama adalah melakukan seleksi fitur dan tahap kedua melakukan klasifikasi. Tahapan seleksi fitu pada penelitian ini dilakukan dengan menerapkan *Correlation matrix*. Hasil dari matrik dataset wine ditunjukkan pada Gambar 5 sebagai berikut.

Attribut...	fixed ac...	volatile ...	citric ac...	residual...	chlorides	free sul...	total sul...	density	pH	sulphat...	alcohol
fixed aci...	1	-0.002	0.672	0.115	0.094	-0.154	-0.113	0.266	-0.683	0.183	-0.022
volatile a...	-0.002	1	-0.033	0.045	0.001	-0.030	-0.014	-0.007	0.025	-0.041	-0.002
citric acid	0.672	-0.033	1	0.144	0.204	-0.061	0.036	0.091	-0.542	0.313	-0.015
residual ...	0.115	0.045	0.144	1	0.056	0.187	0.203	0.301	-0.086	0.006	0.048
chlorides	0.094	0.001	0.204	0.056	1	0.006	0.047	0.074	-0.265	0.371	-0.005
free sulf...	-0.154	-0.030	-0.061	0.187	0.006	1	0.668	0.093	0.070	0.052	0.005
total sulf...	-0.113	-0.014	0.036	0.203	0.047	0.668	1	0.039	-0.067	0.043	0.041
density	0.266	-0.007	0.091	0.301	0.074	0.093	0.039	1	-0.127	0.049	-0.006
pH	-0.683	0.025	-0.542	-0.086	-0.265	0.070	-0.067	-0.127	1	-0.197	0.005
sulphates	0.183	-0.041	0.313	0.006	0.371	0.052	0.043	0.049	-0.197	1	-0.005
alcohol	-0.022	-0.002	-0.015	0.048	-0.005	0.005	0.041	-0.006	0.005	-0.005	1

Gambar 5: Correlation Matrix

Dari matriks pada Gambar 5 tersebut dapat dilihat nilai korelasi satu atribut terhadap atribut yang lain. Kedalaman warna dalam matriks juga menunjukkan nilai korelasinya. Semakin dalam warnanya maka semakin kuat hubungan korelasi atributnya. Ada dua nilai korelasi cukup tinggi yang ditemukan seperti ditunjukkan pada Tabel II sebagai berikut.

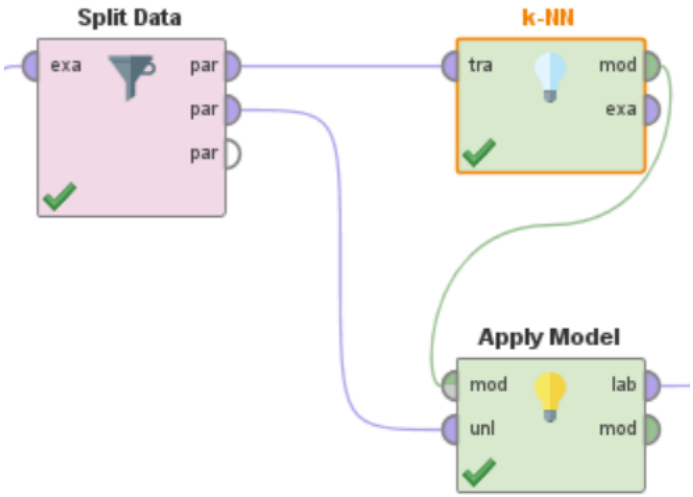
Tabel II: Nilai Korelasi Tertinggi

Atribut 1	Atribut 2	Nilai Korelasi
Fixed acidity	*Citric acid*	0,672
Total sulfur dioxide	*Free sulfur dioxide*	0,668

Tabel II menunjukkan nilai korelasi antara atribut *fixed acidity* dengan *citric acid* sebesar 0,672 dan *total sulfur dioxide* dengan *free sulfur dioxide* sebesar 0,668. Hal tersebut menunjukkan bahwa nilai korelasi hanya berkisar pada angka 0,6 yang berarti bahwa setiap atribut dalam dataset wine sudah cukup unik. Namun begitu, dua nilai korelasi tertinggi dapat dijadikan dasar percobaan untuk menghilangkan salah satu atributnya.

Berdasarkan analisis korelasi dengan menggunakan *Correlation matrix* sebelumnya, maka langkah selanjutnya adalah melakukan menghapus atribut. Penghapusan atribut dilakukan pada empat atribut seperti yang ditunjukkan pada Tabel II secara bergantian untuk mendapatkan nilai akurasi tertinggi. Setelah atribut dihapus, maka dilakukan Langkah *split* data. *Split* data dilakukan untuk membagi data ke dalam data latih dan data uji. Data dibagi 90% untuk data latih dan 10% untuk data uji. Setelah data dilakukan pembagian maka dilanjutkan dengan proses klasifikasi menggunakan algoritma k-NN seperti ditunjukkan pada Gambar 6 sebagai berikut.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

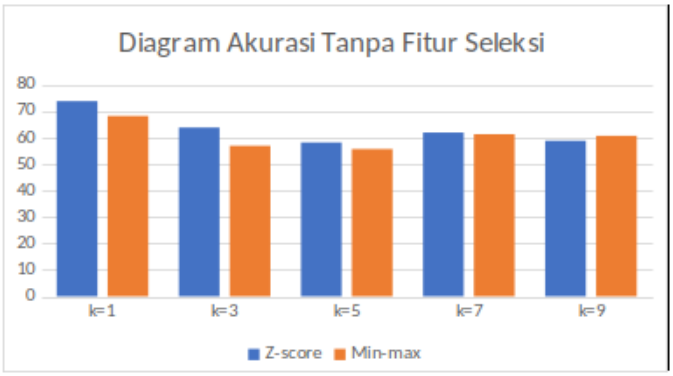


Gambar 6: Proses Split Data dan k-NN

Dari Gambar 6 tersebut, dapat dilihat bahwa pembagian pertama dari data sebanyak 90% sebagai data latih yang digunakan untuk memodelkan data menggunakan algoritma k-NN. Sedangkan *split* data kedua sebanyak 10% sebagai data uji yang digunakan untuk menguji model menggunakan *operator apply model*.

D. Pengujian

Pengujian dilakukan untuk melihat nilai akurasi dari eksperimen yang telah dilakukan. Hasil eksperimen menggunakan metode normalisasi *Z-score* dan *Min-max* tanpa menghilangkan atribut sama sekali menghasilkan nilai akurasi seperti yang ditunjukkan pada Gambar 7 sebagai berikut.



Gambar 7: Nilai akurasi tanpa fitur seleksi

Pada Gambar 7 menunjukkan bahwa normalisasi *Z-score* menghasilkan nilai akurasi tertinggi pada K=1 yaitu sebesar 73,75%. Sedangkan normalisasi *Min-max* menghasilkan nilai akurasi tertinggi pada K=1 juga yaitu sebesar 68,12%.

Hasil dari eksperimen menggunakan metode normalisasi *Z-score* dengan algoritma klasifikasi k-NN dan sudah menggunakan analisis korelasi untuk menghilangkan atribut ditunjukkan pada Tabel 3 sebagai berikut.

Tabel III: Nilai Akurasi dengan Normalisasi Z-score

Variabel	**Nilai Akurasi**				
	k=1	k=3	k=5	k=7	k=9
Tanpa fixed acidity	71,25	64,38	58,75	56,25	55,00
Tanpa citric acid	75,62	65,62	61,88	62,50	58,75
Tanpa total sulfur dioxide	67,50	58,13	54,37	58,13	55,62
Tanpa free sulfur dioxide	71,25	63,75	60,62	58,75	55,00
Tanpa fixed acidity & total sulfur dioxide	66,25	63,12	57,50	59,38	60,62
Tanpa fixed acidity & free sulfur dioxide	68,75	65,00	57,50	53,12	54,37
Tanpa citric acid & total sulfur dioxide	65,62	55,62	53,75	56,88	55,00
Tanpa citric acid & free sulfur dioxide	67,50	63,12	58,75	58,13	57,50

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 1, Februari 2023, hal. 158–166
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i1.771

Dari Tabel III dapat diketahui bahwa dengan normalisasi *Z-score* maka didapatkan nilai akurasi tertinggi adalah pada K=1 dengan mengurangi *variable citric acid*. Eksperimen dengan menghilangkan atribut ini mampu menaikkan nilai akurasi dari 73,75% menjadi 75,62%. Sedangkan hasil dari eksperimen menggunakan metode normalisasi *Min-max* dengan algoritma klasifikasi k-NN dan sudah menggunakan analisis korelasi untuk menghilangkan atribut ditunjukkan pada Tabel IV sebagai berikut.

Tabel IV: Nilai Akurasi dengan Normalisasi Min-max

Variabel	**Nilai Akurasi**				
	k=1	k=3	k=5	k=7	k=9
Tanpa fixed acidity	71,25	58,75	59,38	57,507	56,88
Tanpa citric acid	66,25	58,18	57,50	55,62	56,25
Tanpa total sulfur dioxide	63,75	60,00	56,88	57,50	51,88
Tanpa free sulfur dioxide	66,88	63,12	55,00	56,25	56,88
Tanpa fixed acidity & total sulfur dioxide	63,12	56,25	56,88	55,00	55,62
Tanpa fixed acidity & free sulfur dioxide	70,00	60,00	56,25	55,62	56,25
Tanpa citric acid & total sulfur dioxide	61,25	57,50	57,50	56,88	60,00
Tanpa citric acid & free sulfur dioxide	66,25	59,38	60,62	56,88	55,25

Dari Tabel IV dapat diketahui bahwa dengan normalisasi *Min-max* maka didapatkan nilai akurasi tertinggi adalah pada K=1 dengan mengurangi *variable fixed acidity*. Eksperimen dengan menghilangkan atribut ini mampu menaikkan nilai akurasi dari 68,12% menjadi 71,25%.

Eksperimen yang dilakukan dalam penelitian ini sejalan dengan beberapa penelitian sebelumnya yang menggunakan analisis korelasi untuk meningkatkan performa dari klasifikasi. Penelitian pertama yang menggunakan analisis korelasi adalah penelitian berjudul Uji Komparasi Algoritma *Naïve Bayes* dan *Decision Tree Classification* Menggunakan Covid-19 Dataset [17]. Dalam penelitian tersebut menggunakan *correlation matrix* dan *feature importance* untuk meningkatkan niali akurasi. *Correlation matrix* menunjukkan nilai *total cases* adalah fitur tertinggi dalam korelasi dengan fitur lainnya. Hal ini diperkuat dengan *feature importance* yang dihasilkan, sehingga dapat digunakan sebagai perhitungan dalam kedua algoritma untuk menghasilkan keakuratan yang lebih baik.

Penelitian kedua adalah penelitian dengan judul Implementasi Korelasi untuk Seleksi Fitur pada Klasifikasi Jamur Beracun Menggunakan Jaringan Syaraf Tiruan [18]. Dalam penelitian tersebut menggunakan analisis korelasi yaitu menghilangkan atribut yang berkorelasi lebih dari 0,8. Dengan dihilangkannya 2 variabel tersebut, terjadi peningkatan keakuratan klasifikasi dari 97,97%, menjadi 99,02%.

Penelitian ketiga adalah penelitian berjudul Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan *Correlation Matrix With Heatmap* [19]. Penelitian ini menggunakan analisis korelasi antara atribut dengan kelas *output*. Analisis korelasi menggunakan *correlation matrix with heatmap* untuk menunjukkan semakin dalam warna maka semakin atribut tersbut tidak berkorelasi dengan kelas *output*. Sehingga untuk melakukan klasifikasi waktu kelulusan Mahasiswa cukup menggunakan 9 atribut yang terseleksi agar bisa mendapatkan hasil akurasi yang maksimal.

Penelitian terakhir yang membahas mengenai implementasi analisis korelasi untuk meningkatkan akurasi adalah penelitian berjudul Komparasi Akurasi Metode Correlated Naive Bayes Classifier Dan Naïve Bayes Classifier Untuk Diagnosis Penyakit Diabetes [20]. Penelitian ini melakukan komparasi metode *Correlated-Naive Bayes Classifier* dan *Naive Bayes Classifier* untuk mendapatkan akurasi terbaik. Berdasarkan pengujian yang telah dilakukan menunjukkan bahwa metode *Correlated Naive Bayes Classifier* (CNBC) memperoleh akurasi terbaik dibandingkan dengan metode *Naive Bayes Classifier* (NBC) untuk Dataset Pima indian Diabetes.

IV. Simpulan

Berdasarkan penelitian yang telah dilakukan maka dapat disimpulkan bahwa untuk normalisasi *Z-score*, akurasi dari dataset wine akan mencapai nilai tertinggi jika diklasifikasikan tanpa atribut *citric acid* yaitu sebesar 75,62% lebih tinggi jika tanpa menghilangkan atribut *citric acid* yang hanya menghasilkan nilai akurasi sebebsar 73,75%. Sedangkan untuk normalisasi *Min-max*, akurasi akan mencapai nilai tertinggi jika diklasifikasi tanpa atribut *fixed acidity* yaitu sebesar 71,25% lebih tinggi jika tanpa menghilangkan atribut *fixed acidity* yang hanya menghasilkan nilai akurasi sebebsar 68,12%. Sehingga secara umum, akurasi tertinggi pada klasifikasi dataset wine menggunakan k-NN didapatkan dengan metode normalisasi *Z-score* tanpa atribut *citric acid*.

Pustaka

[1] A. Putri, "Inilah Manfaat Red Wine bagi Kesehatan," Unair, 2021. http://ners.unair.ac.id/site/index.php/news-fkp-unair/30-lihat/1291-inilah-manfaat-red-wine-bagi-kesehatan

[2] M. Jenar, "Mengenal 5 Jenis Wine dan Harganya, Pemula Boleh Baca!," Vantage, 2021. https://www.vantage.id/mengenal-jenis-wine-dan-harganya-pemula-boleh-baca/

[3] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," Comput. Eng. Sci. Syst. J., vol. 4, no. 1, pp. 78–82, 2019, doi: 10.24114/cess.v4i1.11458.

[4] 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.
[5] D. P. Utomo and Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," J. Media Inform. Budidarma, vol. 4, no. 2, pp. 437–444, 2020, doi: 10.30865/mib.v4i2.2080.
[6] Fitria, "Perbandingan Algoritma Naive Bayes Validasi 2 dan 3 Pada Klasifikasi Keluarga Miskin di Kabupaten Banjar," J. Phasti, vol. 05, no. April, pp. 8–14, 2019.
[7] E. Fitriani, "Perbandingan Algoritma C4.5 dan Naive Bayes untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan," J. Sist. Inf., vol. 9, no. 1, pp. 103–115, 2019.
[8] Y. I. Kurniawan and T. I. Barokah, "Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan K-Nearest Neighbor," J. Ilm. Matrik, vol. 22, no. 1, pp. 73–82, 2020, doi: 10.33557/jurnalmatrik.v22i1.843.
[9] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia," J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 4, pp. 427–434, 2018, doi: 10.25126/jtiik.201854773.
[10] Alfarisi, "Data Preprocessing - Konsep Pembelajaran Data Mining," Steemit, 2017. https://steemit.com/education/alfarisi/data-preprocessing-konsep-pembelajaran-data-mining
[11] Trivusi, "Normalisasi Data: Pengertian, Tujuan dan Metodenya," Trivusi, 2022. https://www.trivusi.web.id/2022/09/normalisasi-data.html#:∼:text=Normalisasi min-max biasanya memungkinkan,tidak memperlakukan outlier dengan baik.
[12] R. G. Whendasmoro and Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN," JURIKOM (Jurnal Ris. Komputer), vol. 9, no. 4, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i4.4526.
[13] H. E. Wahanani, M. H. P. Swari, and F. A. Akbar, "Case based Reasoning Prediksi Waktu Studi Mahasiswa Menggunakan Metode Euclidean Distance dan Normalisasi Min-Max," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 6, pp. 1279–1288, 2020, doi: 10.25126/jtiik.2020763880.
[14] Sulandri, A. Basuki, and F. A. Bachtiar, "Metode Deteksi Intrusi Menggunakan Algoritme Extreme Learning Machine dengan Correlation-based Feature Selection," J. Teknol. Inf. dan Ilmu Komput., vol. 8, no. 1, pp. 103–110, 2021, doi: 10.25126/jtiik.0813358.
[15] E. N. R. Khakim, "Perbandingan Algoritma Klasifikasi Data Kesejahteraan Sosial Kabupaten Bantul," Process. J. Ilm. Sist. Informasi, Teknol. Inf. dan Sist. Komput., vol. 17, no. 2, pp. 91–100, 2022.
[16] A. Umar, "Rapidminer, Definisi dan Fitur-fiturnya," 2021. https://www.abdumar.com/2021/03/rapidminer-definisi-dan-fitur-fiturnya.html?m=1
[17] H. Sastypratiwi, Yulianti, and H. Muhardi, "Uji Komparasi Algoritma Naïve Bayes dan Decision Tree Classification Menggunakan Covid-19 Dataset," J. Edukasi dan Penelit. Inform., vol. 8, no. 1, pp. 1–6, 2022.
[18] A. Hermawan and A. P. Wibowo, "Implementasi Korelasi untuk Seleksi Fitur pada Klasifikasi Jamur Beracun Menggunakan Jaringan Syaraf Tiruan," INTEK J. Inform. Dan ..., vol. 5, no. 1, pp. 63–67, 2022, [Online]. Available: http://jurnal.umpwr.ac.id/index.php/intek/article/view/1973
[19] Amiruddin and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," Jambura J. Electr. Electron. Eng., vol. 4, no. 2, pp. 169–174, 2022, doi: 10.37905/jjeee.v4i2.14403.
[20] Hairani, G. S. Nugraha, N. Abdillan M, and M. Innuddin, "Komparasi Akurasi Metode Correlated Naive Bayes dan Naive Bayes Classifier untuk Diagnosis Penyakit Diabetes," InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan), vol. 3, no. 1, pp. 6–11, 2018.

