JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 3, SEPTEMBER, 2024, Pp. 178 – 191
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada *Dataset* Penyakit

Petronilia Palinggik Allorerung [(1)*], Angdy Erna [(2)], Muhammad Bagussahrir [(3)], Samsu Alam [(3)]
[1,3,4] Teknik Informatika, Universitas Dipa Makassar, Makassar, Indonesia
[2] Sistem Informasi, Universitas Dipa Makassar, Makassar, Indonesia
e-mail : {petroniliaallorerung,bagussahrir}@gmail.com, {angdy,alam}@undipa.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 28 Februari 2024, direvisi 14 Agustus 2024, diterima 15 Agustus 2024, dan dipublikasikan 25 September 2024.

Abstract

This study investigates four normalization methods (Min-Max, Z-Score, Decimal Scaling, MaxAbs) across prostate, kidney, and heart disease datasets for K-Nearest Neighbor (K-NN) classification. Imbalanced feature scales can hinder K-NN performance, making normalization crucial. Results show that Decimal Scaling achieves 90.00% accuracy in prostate cancer, while Min-Max and Z-Score yield 97.50% in kidney disease. MaxAbs performs well with 96.25% accuracy in kidney disease. In heart disease, Min-Max and MaxAbs attain accuracies of 82.93% and 81.95%, respectively. These findings suggest Decimal Scaling suits datasets with few instances, limited features, and normal distribution. Min-Max and MaxAbs are better for datasets with numerous instances and non-normal distribution. Z-Score fits datasets with a wide range of feature numbers and near-normal distribution. This study aids in selecting the appropriate normalization method based on dataset characteristics to enhance K-NN classification accuracy in disease diagnosis. The experiments involve datasets with different attributes, continuous and categorical data, and binary classification. Data conditions such as the number of instances, the number of features, and data distribution affect the performance of normalization and classification.

Keywords: Data Normalization, Disease, Min-Max, Z-Score, Decimal Scaling, MaxAbs, K-Nearest Neighbor

Abstrak

Penelitian ini menginvestigasi empat metode normalisasi (*Min-Max, Z-Score, Decimal Scaling, MaxAbs*) pada *dataset* kanker prostat, ginjal, dan jantung untuk klasifikasi K-Nearest Neighbor (K-NN). Skala fitur yang tidak seimbang dapat menghambat kinerja K-NN, sehingga normalisasi data menjadi penting. Hasil penelitian menunjukkan bahwa *Decimal Scaling* mencapai akurasi tertinggi sebesar 90,00% pada penyakit kanker prostat, sementara *Min-Max* maupun *Z-Score* memberikan akurasi tertinggi sebesar 97,50% pada penyakit ginjal. *MaxAbs* juga tampil baik dengan akurasi 96,25% pada penyakit ginjal. Pada penyakit jantung, *Min-Max* dan *MaxAbs* mencapai akurasi masing-masing sebesar 82,93% dan 81,95%. Temuan ini menyimpulkan bahwa *Decimal Scaling* secara umum cocok untuk *dataset* dengan jumlah *instance* yang sedikit, jumlah fitur terbatas, dan berdistribusi normal. *Min-Max* dan *MaxAbs* cenderung lebih sesuai untuk *dataset* dengan jumlah *instance* yang banyak, fitur yang banyak, dan berdistribusi tidak normal. *Z-Score* cocok untuk *dataset* dengan jumlah fitur yang relatif besar atau kecil dan cocok untuk *dataset* berdistribusi normal atau mendekati normal. Penelitian ini membantu menentukan metode normalisasi yang sesuai dengan karakteristik *dataset* untuk meningkatkan akurasi model klasifikasi K-NN dalam mendiagnosis penyakit. Eksperimen menggunakan tiga *dataset* penyakit dengan atribut yang berbeda-beda, jenis data kontinu dan kategorikal, serta klasifikasi biner. Kondisi data seperti jumlah *instance*, jumlah fitur, dan distribusi data mempengaruhi performa normalisasi dan klasifikasi.

Kata Kunci: Normalisasi Data, Penyakit, *Min-Max*, *Z-Score*, *Decimal Scaling*, *MaxAbs*, K-Nearest Neighbor

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 3, SEPTEMBER, 2024: 178 – 191

1. PENDAHULUAN

Data transformation adalah proses pengubahan data menjadi bentuk yang cocok untuk proses eksplorasi data (Marlina & Bakri, 2021). Salah satu cara transformasi data adalah normalisasi data. Normalisasi data ialah metode menskalakan ulang data menjadi lebih kecil (Ambarwari et al., 2020). Penggunaan normalisasi data sangat diperlukan jika terdapat perbedaan skala antar fitur dalam *dataset*. Konsekuensi jika tidak melakukan normalisasi data adalah menghasilkan nilai akurasi yang kecil dan menyebabkan fitur berskala rendah tidak berpengaruh saat mengimplementasikan algoritma *data mining* yang melibatkan pengukuran jarak, atau dengan kata lain hasil analisis *data mining* hanya bergantung pada fitur berskala tinggi, padahal fitur berskala rendah juga memiliki peranan yang sama pentingnya (Kusnaidi et al., 2022).

Penggunaan normalisasi data sangat penting dilakukan apabila menggunakan metode *data mining* yang melibatkan pengukuran jarak seperti K-Nearest Neighbor (K-NN). K-NN ialah salah satu *method* yang didasarkan pada mayoritas tetangga terdekatnya yang dihasilkan dari proses perhitungan jarak euclidean. Jika jarak nilai setiap fitur sangat besar, maka perhitungan euclidean *distance* kurang maksimal. Perhitungan *distance* yang kurang maksimal memberikan hasil yang kurang tepat dan menunjukkan bahwa *dataset* yang gunakan tidak berkualitas (Whendasmoro & Joseph, 2022).

Penerapan normalisasi data tidak digunakan pada semua *dataset*, melainkan hanya pada *dataset* yang memiliki perbedaan skala pada fiturnya. Salah satu *dataset* yang memiliki karakteristik tersebut adalah *dataset* penyakit. *Dataset* penyakit adalah salah satu contoh data yang sering dipakai dalam studi *data mining*. *Dataset* penyakit sendiri umumnya memiliki nilai numerik, di mana terdapat perbedaan skala pada fiturnya, seperti fitur tinggi badan dan berat badan. Perbedaan skala antar fitur menyebabkan fitur dengan nilai yang lebih kecil tidak bermanfaat dibandingkan dengan fitur lainnya. Oleh karenanya, perlu dilakukan normalisasi data untuk menyeimbangkan skala setiap fitur pada *dataset* ke rasio yang lebih kecil. Normalisasi data dapat dilakukan dengan metode yang umumnya digunakan oleh para peneliti seperti *Min-Max*, *Z-Score*, *Decimal Scaling* (Pagan et al., 2023) dan *MaxAbs* (Permana & Salisah, 2022).

Penelitian tentang penggunaan normalisasi data sudah pernah dilakukan pada penelitian prediksi penyakit diabetes dengan dua normalisasi yakni *Min-Max* dan *Z-Score* dengan algoritma K-NN dan didapat bahwa normalisasi *Min-Max* memperoleh akurasi tertinggi (Sholeh et al., 2022). Adapula penelitian yang membandingkan tiga normalisasi yakni *Min-Max, Z-Score,* dan *Decimal Scaling* untuk klasifikasi *wine* dengan metode K-NN dan didapat bahwa akurasi tertinggi ada pada *Min-Max* (Chandra et al., 2022). Pada penelitian klasifikasi status gizi balita dilakukan perbandingan dua normalisasi data yaitu *Z-Score* dan *Min-Max* dengan metode K-NN dan didapat bahwa metode *Z-Score* memiliki akurasi tertinggi (HS et al., 2023). Dari penelitian sebelumnya, dapat dilihat bahwa *Min-Max* merupakan metode normalisasi data terbaik. Akan tetapi, penelitian tersebut masih membandingkan dua metode normalisasi data, sedangkan perbandingan dengan metode *Min-Max, Z-Score, Decimal Scaling*, dan *MaxAbs* belum dilakukan, sehingga belum diketahui apakah *MaxAbs* memiliki akurasi lebih tinggi atau tidak dari *Min-Max.* Selain itu*,* terdapat pula penelitian oleh Pagan et al. (2023) yang membandingkan tiga normalisasi (*Min-Max, Z-Score*, dan *Decimal Scaling*) untuk sepuluh *dataset* yang dipilih oleh peneliti berdasarkan keragaman ukuran, kompleksitas, dan dimensi fiturnya. Hasil penelitian ini menunjukkan ada enam *dataset* yang menghasilkan *Min-Max* sebagai metode terbaik bahkan mayoritas menghasilkan akurasi di atas 85%, sedangkan empat *dataset* lainnya menghasilkan *Z-Score* sebagai metode terbaik dan mayoritas menghasilkan akurasi di atas 97%. Hal ini menunjukkan bahwa penemuan akurasi tertinggi melalui metode normalisasi yang sama disebabkan oleh kemiripan karakteristiknya, namun penelitian ini tidak menjelaskan secara pasti jumlah ukuran, kompleksitas, dan dimensi fitur tiap *dataset*. Oleh karena itu, metode normalisasi data terbaik yang dihasilkan belum bisa dijadikan acuan untuk setiap kondisi *dataset*.

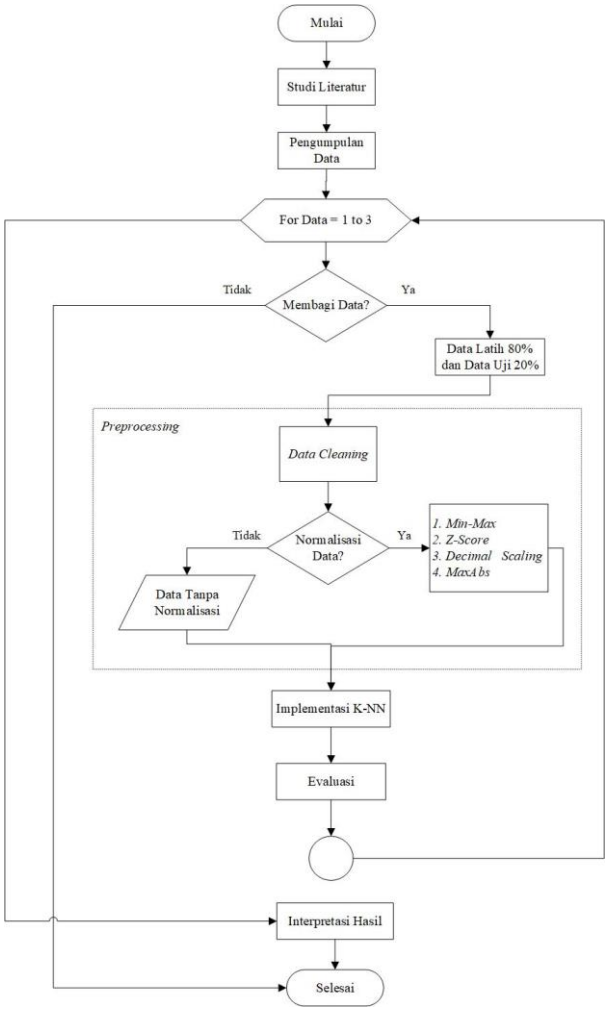
Berdasarkan penjabaran di atas, maka penulis akan mengeksplorasi empat metode normalisasi data yakni *Min-Max Normalization*, *Z-Score Normalization*, *Decimal Scaling Normalization*, dan

MaxAbs Normalization dengan algoritma K-NN. Metode tersebut digunakan dalam mengklasifikasi tiga *dataset* penyakit yang diperoleh dari *website* Kaggle. Tujuannya adalah untuk mengidentifikasi metode normalisasi terbaik sesuai dengan karakteristik data tersebut.

2. METODE PENELITIAN

2.1 Tahap Penelitian

Tahap-tahap yang akan dilaksanakan divisualisasikan oleh Gambar 1. *Flowchart* ini menggambarkan proses penelitian untuk menentukan metode normalisasi terbaik dalam meningkatkan akurasi klasifikasi penyakit menggunakan K-Nearest Neighbor *(*K-NN*)*. Proses dimulai dengan studi literatur dan pengumpulan data, diikuti dengan *looping* untuk tiga *dataset* penyakit (kanker prostat, ginjal, dan jantung). Data kemudian dibagi menjadi 80% data latih dan 20% data uji. Langkah berikutnya adalah *preprocessing*, termasuk *data cleaning* dan normalisasi data menggunakan metode *Min-Max, Z-Score, Decimal Scaling*, dan *MaxAbs*. Setelah itu, algoritma K-NN diterapkan pada data yang sudah diproses, dan kinerja model dievaluasi. Hasil evaluasi kemudian diinterpretasikan untuk mendapatkan kesimpulan akhir sebelum proses penelitian dinyatakan selesai.



Gambar 1 *Flowchart* Tahap Penelitian

2.2 Pengumpulan Data

Data yang dikumpulkan bersumber dari *website* Kaggle yaitu *dataset* penyakit kanker prostat sebanyak 100 data, penyakit ginjal sebanyak 400 data, dan penyakit jantung sebanyak 1025 data. Adapun struktur ketiga *dataset* tersebut ditunjukkan oleh Tabel 1 sampai 3. Pemilihan ketiga *dataset* ini didasarkan pada beberapa alasan yang komprehensif seperti berikut:

1) **Keberagaman Ukuran *Dataset***: Ketiga *dataset* dipilih karena mewakili variasi dalam ukuran *dataset*. Penggunaan ukuran *dataset* yang bervariasi penting dalam mengevaluasi kinerja algoritma pembelajaran mesin karena menurut penelitian oleh Pagan et al. (2023) yang menggunakan sepuluh *dataset* ditemukan bahwa keberagaman ukuran *dataset* dapat mempengaruhi kemampuan algoritma untuk akurasi prediksi.
2) **Variasi dalam Atribut**: Setiap *dataset* memiliki jumlah dan jenis atribut yang berbeda, yang memungkinkan untuk menguji metode normalisasi dalam konteks yang berbeda. Variasi ini penting karena berdasarkan penggunaan jumlah dan jenis fitur yang beragam pada penelitian Pagan et al. (2023), ditemukan bahwa variasi dalam atribut dapat mempengaruhi kinerja algoritma pembelajaran mesin dan normalisasi.
3) **Kelas Biner**: Ketiga *dataset* memiliki kelas biner (dua kelas) yang konsisten. Ini penting karena klasifikasi biner adalah salah satu jenis klasifikasi yang paling umum dalam analisis kesehatan, dan menurut Riaz et al. (2022), klasifikasi biner lebih sederhana dan hampir 60% penelitian membahas klasifikasi biner. Oleh karena itu, klasifikasi biner menjadi langkah awal yang penting sebelum melakukan klasifikasi dengan lebih banyak kelas.
4) **Skala Atribut yang Beragam**: Setiap *dataset* menggunakan skala yang berbeda untuk atribut mereka, mulai dari atribut kontinu hingga atribut diskrit. Variasi skala ini memungkinkan penelitian untuk menguji efektivitas berbagai metode normalisasi yang berbeda dalam menangani skala atribut yang berbeda, yang menurut Jain et al. (2000), dapat signifikan dalam mempengaruhi kinerja algoritma pembelajaran mesin.
5) **Signifikansi Klinis**: Pemilihan *dataset* penyakit kanker prostat, ginjal, dan jantung didasarkan pada prevalensi dan signifikansi klinis penyakit tersebut. Ketiganya adalah penyakit serius dengan dampak besar pada kesehatan masyarakat, sehingga penelitian yang meningkatkan diagnosis mereka sangat berharga. Menurut data World Health Organization yang dirilis tahun 2020, penyebab kematian utama di dunia pada tahun 2019 adalah penyakit, diantaranya kardiovaskular dan kanker, sehingga membuat penelitian ini sangat relevan.

Dengan mempertimbangkan faktor-faktor di atas, ketiga *dataset* ini memberikan basis yang komprehensif untuk mengevaluasi dan membandingkan performa metode normalisasi dalam klasifikasi K-NN. Selain itu, variasi dalam jumlah data, jumlah atribut, dan skala atribut memastikan bahwa temuan penelitian ini akan lebih umum dan aplikatif dalam konteks yang lebih luas.

Tabel 1 Struktur Data Penyakit Kanker Prostat

Fitur	Nilai
Radius	9 – 25
Texture	11 – 27
Perimeter	52 – 172
Area	202 – 1878
Smoothness	0,070 – 0,143
Compactness	0,038 – 0,345
Symmetry	0,135 – 0,304
Fractal Dimension	0,053 – 0,097
Class	*Malignant dan Benign*

Tabel 2 Struktur Data Penyakit Ginjal

Fitur	Nilai
Blood Pressure	50 – 180
Specific Gravity	1.005 – 1.025
Albumin	0 – 5
Sugar	0 – 5
Red Blood Cell	0 dan 1
Blood Urea	1.500 – 391
Serum Creatinine	0.400 – 76
Sodium	4.500 – 163
Potassium	2.500 – 47
Hemoglobin	3.100 – 17.800
White Blood Cell Count	2200 – 26400
Red Blood Cell Count	2.100 – 8
Hypertension	0 dan 1
Class	0 dan 1

Tabel 3 Struktur Data Penyakit Jantung

Fitur	Nilai
Age	29 – 77
Sex	0 dan 1
Chest Pain Type	0 – 3
Resting Blood Pressure	94 – 200
Serum Cholesterol	126 – 564
Fasting Blood Sugar	0 dan 1
Resting Electrocardiographic Result	0 – 2
Maximum Heart Rate Achieved	71 – 202
Exercise Induced Angina	0 dan 1
Oldpeak	0 – 6.200
Slope	0 – 2
Number of Major Vessels	0 – 4
Thalassemia	0 – 3
Class	0 dan 1

2.3 Teknik Normalisasi

2.3.1 *Min-Max Normalization*

Min-Max adalah jenis normalisasi yang mentransformasi linier pada *original data* (Chandra et al., 2022). Skala yang dihasilkan berada di antara 0 hingga 1. Rumus normalisasi *Min-Max* ditunjukkan pada Pers. (1). Di mana x_i adalah data yang dinormalisasi, x' menunjukkan hasil normalisasi, $min(x)$ merupakan data terkecil suatu fitur, dan $max(x)$ adalah data terbesar suatu fitur.

$$x' = \frac{x_i - min(x)}{max(x) - min(x)} \quad (1)$$

2.3.2 *Z-Score Normalization*

Z-Score ialah jenis normalisasi yang didasarkan pada *mean* dan standar deviasi (Chandra et al., 2022). *Z-Score* mengubah data asli menjadi skala yang lebih kecil, tetapi tidak memiliki skala yang tetap. Pes. (2) menunjukkan rumus normalisasi *Z-Score*. Di mana x_i merupakan data yang

dinormalisasi, x' menunjukkan hasil normalisasi, $mean(x)$ adalah nilai rata-rata suatu fitur, dan $std(x)$ merupakan standar deviasi suatu fitur.

$$x' = \frac{x_i - mean(x)}{std(x)} \quad (2)$$

2.3.3 *Decimal Scaling Normalization*

Decimal Scaling adalah jenis normalisasi yang menggeser nilai desimal data (Chandra et al., 2022). Skala yang dihasilkan *Decimal Scaling* berada di antara 0 hingga 1. Rumus normalisasi *Decimal Scaling* dituliskan pada Pers. (3). Di mana x' adalah hasil normalisasi, x_i merupakan data yang dinormalisasi, dan j menunjukkan jumlah digit dari data terbesar pada tabel.

$$x' = \frac{x_i}{10^j} \quad (3)$$

2.3.4 *MaxAbs Normalization*

MaxAbs merupakan jenis normalisasi yang melakukan pembagian seluruh data dengan nilai absolut maksimum (Permana & Salisah, 2022). Data asli yang dinormalisasi dengan *MaxAbs* biasanya berada pada skala -1 hingga 1. Rumus normalisasi *MaxAbs* ditunjukkan pada Pers. (4). Di mana x_i merupakan data yang dinormalisasi, x' menunjukkan hasil normalisasi n, dan $max(x)$ adalah data terbesar suatu fitur.

$$x' = \frac{x_i}{|max(x)|} \quad (4)$$

2.4 Algoritma K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) ialah *supervised learning method* yang digunakan dalam mengklasifikasikan objek baru berdasarkan *class* terbanyak pada tetangga terdekatnya. Berikut langkah-langkah dalam mengimplementasikan K-NN (Cahyanti et al., 2020):

1) Menentukan nilai K yang terdiri dari dua cara, yaitu menggunakan bilangan ganjil yakni 1,3,5,7, dan 9 atau menerapkan rumus pada Pers. (5). Di mana N adalah banyaknya data latih dan k adalah jumlah tetangga terdekat.

$$k = \sqrt{N} \quad (5)$$

2) Hitung *distance* antara *data testing* dan semua *data training*. Perhitungan tersebut menerapkan rumus Euclidean Distance pada Pers. (6). Di mana p_i merupakan *data training*, q_i adalah *data testing*, i menunjukkan data variabel, dan n adalah banyak data.

$$Euclidean = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (6)$$

3) Urutkan hasil perhitungan jarak dari terkecil ke terbesar.
4) Menentukan tetangga terdekat berdasarkan jarak terpendek hingga nilai K.
5) Menentukan *class* berdasarkan tetangga terdekat.
6) Menemukan *class* mayoritas dari tetangga terdekat dan menetapkannya sebagai *class data testing*.

2.5 Teknik Pengujian

2.5.1 Uji Normalitas

Studi ini menerapkan uji normalitas secara statistik dengan Kolmogorov-Smirnov menggunakan aplikasi SPSS. Pengujian ini memakai nilai statistik uji dan nilai tabel Kolmogorov-Smirnov. Uji Kolmogorov-Smirnov yang tersedia di SPSS memungkinkan peneliti untuk dengan mudah membandingkan distribusi sampel dengan distribusi teoretis (normal) tanpa perlu melakukan pemrograman yang kompleks. Hasil pengujian tiap *dataset* tersebut dijelaskan oleh Tabel 4. Berdasarkan Tabel 4 maka dapat diketahui bahwa *dataset* penyakit kanker prostat berdistribusi normal karena nilai statistik uji < nilai tabel *K-S*, sebaliknya *dataset* penyakit ginjal dan *dataset* penyakit jantung tidak berdistribusi normal.

Tabel 4 Hasil Uji Normalitas

Dataset	**Nilai Statistik Uji**	**Nilai Tabel *K-S***
Penyakit Kanker Prostat	0,077	0,13581
Penyakit Ginjal	0,079	0,06790
Penyakit Jantung	0,066	0,04242

Menurut Singh & Singh (2020), normalisasi sangat penting untuk memastikan bahwa setiap fitur memiliki kontribusi yang sama sehingga dapat meningkatkan kualitas data dan kinerja algoritma. Data yang tidak berdistribusi normal dapat menyebabkan model K-NN menjadi lebih bias atau memiliki *variance* yang lebih tinggi. Jika distribusi data tidak terdistribusi secara merata, K-NN dapat cenderung *overfitting* pada daerah tertentu dari data atau *underfitting* pada daerah lain. Ini dapat mengurangi kemampuan generalisasi model. Normalisasi membantu menangani *outlier* yang mana menurut Barus & Sutarman (2023), kehadiran *outlier* juga dapat menyebabkan distribusi data yang tidak normal sehingga menciptakan bias dan menurunkan performa algoritma *machine learning*.

2.5.2 Uji Kinerja Algoritma

Akurasi menunjukkan persentase data yang berhasil diklasifikasikan dengan benar. Nilai akurasi dihasilkan melalui rumus pada Pers. (7). AUC (*Area Under the ROC Curve*) menunjukkan seberapa baik pola yang terbentuk dalam memprediksi kelas dengan tepat. Nilai AUC dibagi menjadi lima kelompok yang ditunjukkan pada Tabel 5.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (7)$$

Tabel 5 Nilai AUC

Nilai	**Klasifikasi**
0,90 – 1,00	Sangat Baik
0,80 – 0,90	Baik
0,70 – 0,80	Cukup
0,60 – 0,70	Buruk
0,50 – 0,60	Salah

3. HASIL DAN PEMBAHASAN

3.1 Membagi Data (*Split Data)*

Spliting data menjadi data latih dan data uji diterapkan untuk ketiga *dataset* penyakit dengan persentase 80% sebagai data latih dan 20% sebagai data uji. Proses pembagian ini dikerjakan oleh RapidMiner secara default sehingga tidak menyebabkan perbedaan hasil ketika dilakukan analisis secara berulang-ulang. Hasil *split data* ini dijelaskan oleh Tabel 6.

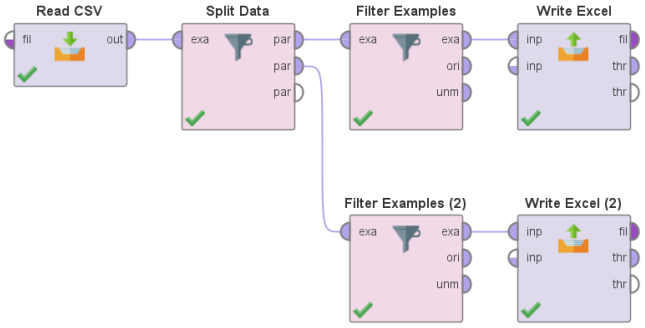
Tabel 6 Hasil Split Data

Dataset	Total Data	Jumlah Data Latih	Jumlah Data Uji
Penyakit Kanker Prostat	100	80	20
Penyakit Ginjal	400	320	80
Penyakit Jantung	1025	820	205

3.2 *Preprocessing*

Preprocessing terbagi menjadi dua yakni *cleaning data* dan normalisasi data. *Cleaning data* ialah proses menghapus data tidak lengkap (*missing value)*. Normalisasi data hanya dilakukan pada fitur yang memiliki rentang yang lebih besar dari fitur lainnya yaitu pada fitur di luar rentang 0 – 1. Berdasarkan struktur *dataset* penyakit kanker prostat pada Tabel 1, maka akan dilakukan normalisasi pada empat fitur pertama. Sedangkan berdasarkan struktur *dataset* penyakit ginjal pada Tabel 2 akan dilakukan normalisasi pada semua fitur kecuali fitur *Red Blood Cell* dan *Hypertension.* Sementara untuk *dataset* penyakit jantung, dilihat dari struktur datanya pada Tabel 3 maka akan dilakukan normalisasi pada semua fitur kecuali fitur *Sex, Fasting Blood Sugar,* dan *Exercise Induced Angina*. Berikut adalah rangkaian *preprocessing* dengan menggunakan beberapa operator yang ada pada RapidMiner.

1) Tanpa Normalisasi



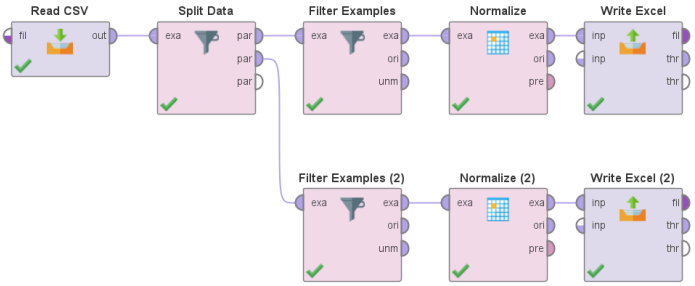
Gambar 2 *Preprocessing* Tanpa Normalisasi

Gambar 2 menunjukkan tahap *preprocessing* tanpa menggunakan normalisasi dimulai dengan membaca *dataset* penyakit dalam format *csv* menggunakan operator *Read CSV.* Setelah itu, dilakukan proses pembagian *dataset* dengan memanfaatkan operator *Split Data.* Hasil dari pembagian data tersebut kemudian dibersihkan untuk menghilangkan data yang tidak lengkap (*missing value*) menggunakan operator *Filter Example*. Karena ketiga *dataset* yang digunakan lengkap, maka jumlah data sebelum dan setelah dibersihkan tetap sama. Hasil dari tahap *preprocessing* ini selanjutnya disimpan ke dalam format *excel* menggunakan operator *Write Excel*. Jika *dataset* memiliki skala yang konsisten atau perbedaan rentang nilai antar atribut yang tidak berbeda jauh, tanpa normalisasi dapat menjadi pilihan yang baik (Permana & Salisah, 2022). Data yang digunakan juga tetap dalam bentuk aslinya sehingga tidak ada risiko informasi yang hilang atau terdistorsi.

2) *Min-Max*

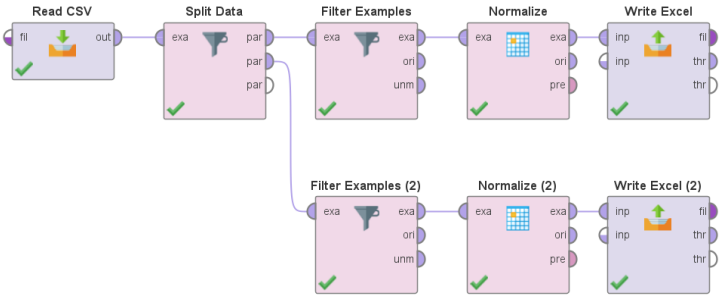
Gambar 3 menunjukkan tahap *preprocessing* menggunakan metode normalisasi *Min-Max* yang mana tahapan awalnya sama dengan *preprocessing* tanpa normalisasi yaitu membaca *dataset*, membagi *dataset*, dan *cleaning dataset*. Setelah menyelesaikan tahap tersebut, dilakukanlah normalisasi data dengan metode *Min-Max* menggunakan operator *Normalize* dan mengatur parameternya yaitu *attribute filter type* menjadi *subset* sehingga dapat memilih fitur yang akan dinormalisasi dengan cara mengisi nama fitur pada bagian *Select Attributes*, dalam hal ini adalah fitur dengan nilai di luar rentang 0 – 1. Selain itu, digunakan pula parameter *method* untuk memilih

metode normalisasi data yaitu *range transformation* untuk *Min-Max*. Hasil dari tahap *preprocessing* ini selanjutnya disimpan ke dalam format *excel* menggunakan operator *Write Excel*. Metode normalisasi *Min-Max* menjadikan semua fitur dalam rentang yang sama, biasanya antara 0 dan 1, sehingga memudahkan komparasi antar fitur (Permana & Salisah, 2022).



Gambar 3 *Preprocessing* dengan *Min-Max*

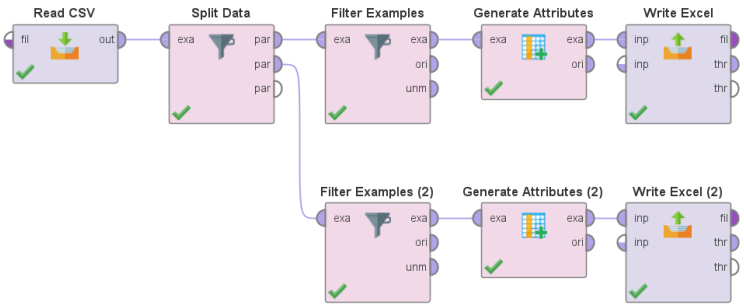
3) *Z-Score*



Gambar 4 *Preprocessing* dengan *Z-Score*

Gambar 4 menunjukkan tahap *preprocessing* menggunakan metode normalisasi *Z-Score* yang mana tahap awalnya hingga akhir hampir sama dengan metode normalisasi *Min-Max* pada Gambar 3. Perbedaan keduanya terletak pada pengaturan operator *Normalize* pada bagian parameter *method. Min-Max* menggunakan *method range transformation* sedangkan *Z-Score* menggunakan *method Z-transformation*. Metode normalisasi *Z-score* dapat menanggulangi masalah *outlier* dengan cara mengukur berapa banyak standar deviasi suatu data poin dari *mean*.

4) *Decimal Scaling*

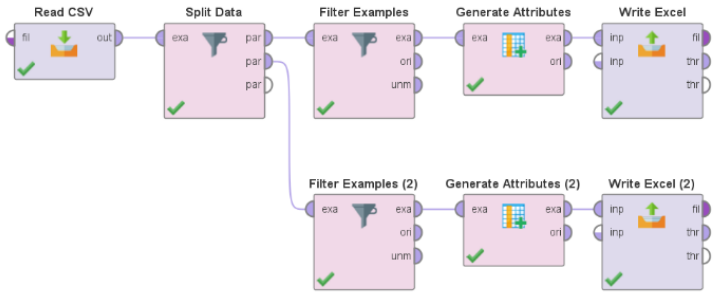


Gambar 5 *Preprocessing* dengan *Decimal Scaling*

Gambar 5 menunjukkan tahap *preprocessing* menggunakan metode normalisasi *Decimal Scaling* yang mana tahapan awalnya sama dengan *preprocessing* tanpa normalisasi yaitu membaca

dataset, membagi *dataset*, dan *cleaning dataset*. Setelah menyelesaikan tahap tersebut, dilakukanlah normalisasi data dengan metode *Decimal Scaling* dengan cara menuliskan rumus dari metode tersebut menggunakan operator *Generate Attributes*. Hasil dari tahap *preprocessing* ini selanjutnya disimpan ke dalam format *excel* menggunakan operator *Write Excel*. *Decimal Scaling* dilakukan dengan membagi nilai data berdasarkan kelipatan 10. Teknik ini cocok untuk data dengan rentang suatu nilai adalah antara 0 dan 1 sedangkan nilai lain pada rentang 0 dan 1000 (Kusnaidi et al., 2022).

5) *MaxAbs*



Gambar 6 *Preprocessing* dengan *MaxAbs*

Gambar 6 menunjukkan tahap *preprocessing* menggunakan metode normalisasi *MaxAbs* yang mana tahap awalnya hingga akhir hampir sama dengan metode normalisasi *Decimal Scaling* pada Gambar 5. Perbedaan keduanya terletak pada penulisan rumus menggunakan operator *Generate Attributes.* Skala data teknik *MaxAbs* berdasarkan nilai absolut maksimum, mempertahankan *sparsity* pada data *sparse* (Permana & Salisah, 2022). Teknik ini sangat efektif untuk data yang mengandung nilai negatif dan positif dengan skala besar.

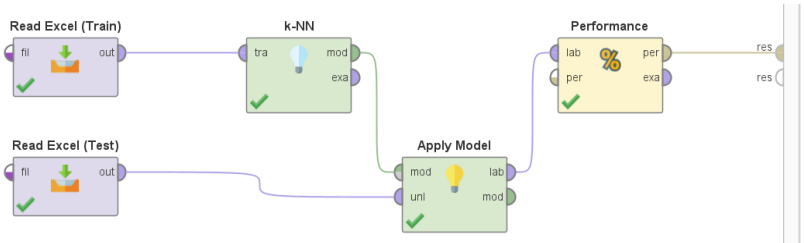
3.3 Implementasi K-NN

Pada tahap ini, algoritma K-Nearest Neighbor (K-NN) diimplementasikan berdasarkan hasil *preprocessing* data. Sebelum melanjutkan ke tahap implementasi, dilakukan terlebih dahulu penentuan nilai k yang optimal untuk masing-masing dataset. Penentuan nilai k ini menggunakan rumus yang ditunjukkan pada Pers. (5). Pemilihan nilai k sangat penting karena mempengaruhi performa dan akurasi dari model K-NN dalam proses klasifikasi atau identifikasi yang dilakukan pada dataset yang telah diproses sebelumnya.

$$K_{(Penyakit\ Kanker\ Prostat)} = \sqrt{80} = 8.94 = 9$$

$$K_{(Penyakit\ Ginjal)} = \sqrt{320} = 17.88 = 18$$

$$K_{(Penyakit\ Jantung)} = \sqrt{820} = 28.63 = 29$$



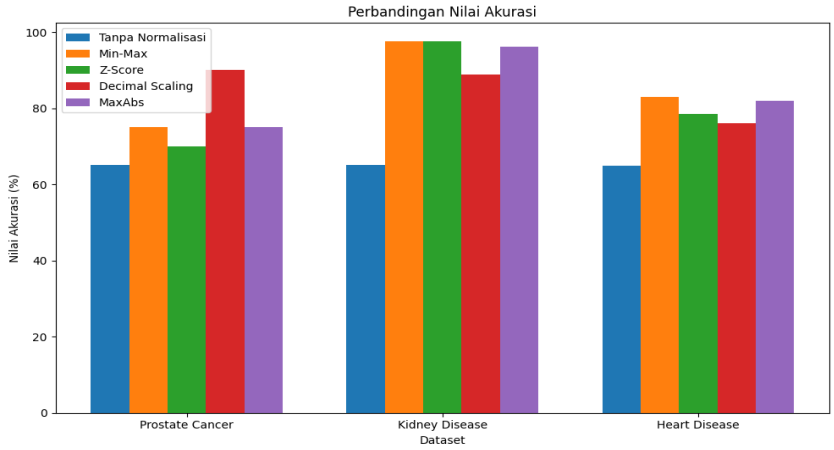
Gambar 7 Proses Implementasi K-NN

Gambar 7 menunjukkan proses implementasi metode K-NN yang dimulai dengan membaca data latih dan data uji hasil normalisasi dan tanpa normalisasi dalam format *excel* menggunakan operator *Read Excel.* Setelah itu, diterapkan K-Nearest Neighbor melalui operator K-NN dan mengatur parameternya yaitu *k* untuk menentukan jumlah tetangga terdekat. Penentuan nilai *k* untuk setiap *dataset* tergantung dari jumlah data latihnya, sehingga pada penelitian ini dimanfaatkan suatu rumus yaitu $\sqrt{N}$. Selain itu, digunakan pula parameter *measure types* untuk memilih metode pengukuran jarak berdasarkan jenis datanya, dalam hal ini digunakan data numerik sehingga dipilih opsi *NumericalMeasures* dan *EuclideanDistance*. Selanjutnya digunakan operator *Apply Model* untuk menerapkan model metode K-NN dari data latih agar melakukan pengklasifikasian pada data uji. Proses klasifikasi dengan metode K-NN ini kemudian di evaluasi untuk mengetahui kinerjanya menggunakan operator *Performance*.

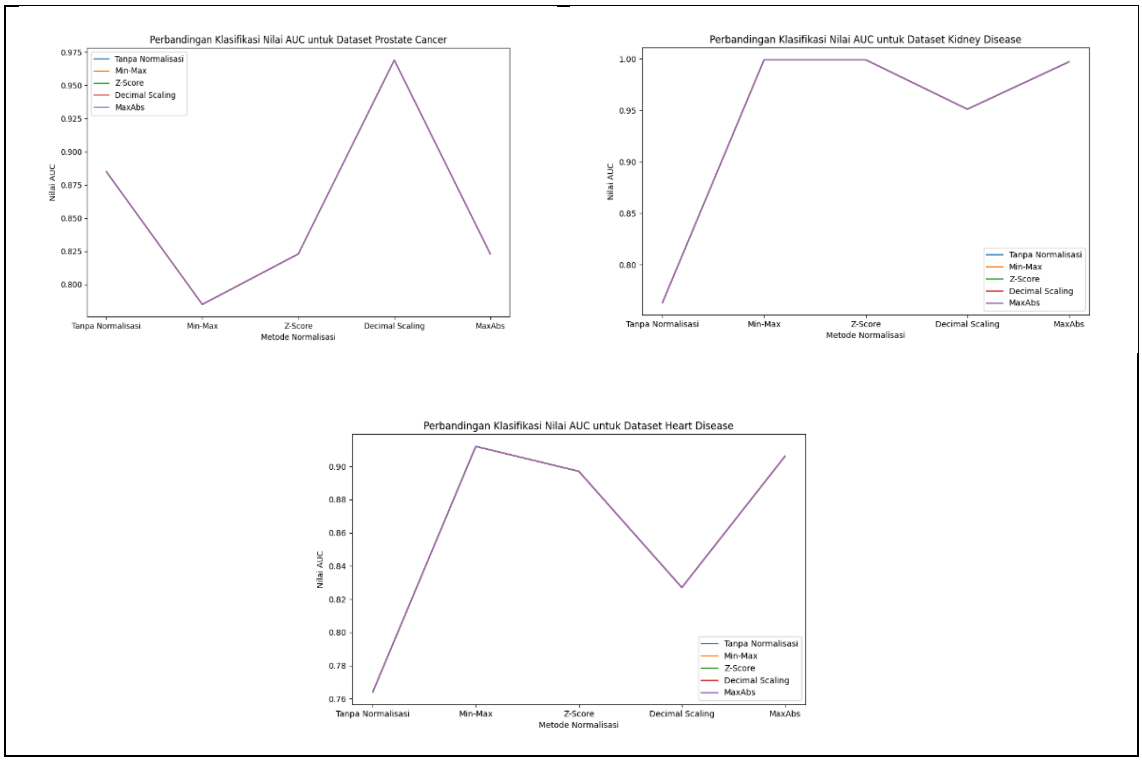
3.4 Hasil Analisis

Berdasarkan proses analisis diperoleh hasil perbandingan nilai akurasi dan nilai AUC untuk masing-masing *dataset* yang ditunjukkan melalui grafik Gambar 8 dan Gambar 9. Gambar 8 dan Gambar 9 menunjukkan apabila tidak menggunakan normalisasi data pada ketiga *dataset,* maka akan menghasilkan akurasi yang rendah yaitu ≤ 65%. Sedangkan jika menggunakan normalisasi data, akurasi algoritma K-NN dapat lebih tinggi.

Pada *dataset* penyakit kanker prostat terlihat metode *Decimal Scaling* menghasilkan akurasi tertinggi sebesar 90,00% dan nilai AUC sebesar 0.969. Pada *dataset* penyakit ginjal, metode *Min-Max* dan *Z-Score* menghasilkan akurasi tertinggi sebesar 97,50% dan nilai AUC sebesar 0,999. Metode *MaxAbs* juga menghasilkan akurasi yang tinggi pada *dataset* penyakit ginjal sebesar 96,25% dan nilai AUC sebesar 0,997. Pada *dataset* penyakit jantung, metode *Min-Max* dan *MaxAbs* menghasilkan akurasi yang cukup baik dibanding kedua metode lainnya sebesar 82,93% dan 81,95% serta nilai AUC sebesar 0,912 dan 0,906. Perbedaan penemuan metode normalisasi data terbaik ini disebabkan oleh karakteristik data yaitu, jumlah data, jumlah fitur, dan distribusi data (Pagan et al., 2023). Karakteristik dari ketiga *dataset* tersebut ditunjukkan pada Tabel 7.



Gambar 8 Perbandingan Nilai Akurasi



Gambar 9 Perbandingan Nilai AUC

Tabel 7 Karakteristik *Dataset*

Dataset	Total Data	Jumlah Fitur	Distribusi Data
Penyakit Kanker Prostat	100	9	Normal
Penyakit Ginjal	400	14	Tidak Normal
Penyakit Jantung	1025	14	Tidak Normal

Berdasarkan nilai akurasi pada Gambar 8, nilai AUC pada Gambar 9, dan karakteristik data pada Tabel 7, maka diperoleh beberapa pengetahuan antara lain: 1) *Decimal Scaling* cenderung cocok dengan karakteristik *dataset* yang memiliki jumlah data yang kecil, jumlah fitur yang sedikit dengan rentang nilai yang kecil, dan data yang berdistribusi normal. 2) *Min-Max* cenderung cocok dengan karakteristik *dataset* yang memiliki jumlah data yang besar, jumlah fitur yang banyak, dan data yang tidak mengikuti distribusi normal. Metode *Min-Max* mereskal nilai-nilai fitur ke dalam rentang yang telah ditentukan (biasanya antara 0 dan 1) dengan mempertahankan bentuk distribusi data. Ini berarti *Min-Max* tidak memerlukan asumsi tentang distribusi data. Kecocokan antara *Min-Max* dengan karakteristik tersebut didukung oleh penelitian sebelumnya yaitu Henderi et al. (2021) dan Sholeh et al. (2022) yang juga menghasilkan *Min-Max* sebagai metode terbaik dengan karakteristik tersebut. 3) Sama seperti *Min-Max*, *MaxAbs* juga cenderung cocok dengan karakteristik data yang memiliki jumlah data yang besar, jumlah fitur yang banyak, dan data yang tidak mengikuti distribusi normal. Namun, jika dibandingkan dengan metode *Min-Max* yang mampu menghasilkan akurasi 97,50% pada *dataset* penyakit ginjal dan 82,93% pada *dataset* penyakit jantung, metode *MaxAbs* justru hanya mampu menghasilkan akurasi 96,25% pada *dataset* penyakit ginjal dan 81,95% pada *dataset* penyakit jantung. Hal ini menunjukkan bahwa performa *MaxAbs* tidak sebaik *Min-Max*. 4) *Z-Score* biasanya cenderung cocok untuk *dataset* yang memiliki jumlah fitur yang relatif besar atau kecil, dan cocok untuk data yang berdistribusi normal atau mendekati distribusi normal (McLeod, 2023). Kecocokan metode *Z-Score* dengan karakteristik tersebut didukung oleh penelitian sebelumnya (Badugu, 2020) yang juga menghasilkan *Z-Score* sebagai metode terbaik. Jika meninjau lebih jauh distribusi datanya, *Z-Score* juga cenderung cocok pada distribusi yang tidak terlalu jauh dari normalitas, di mana pada penelitian ini ditemukan bahwa *dataset* penyakit ginjal memiliki nilai statistik uji yang tidak terlalu

jauh dari nilai tabel Kolmogorov-Smirnov. Hal tersebut membuktikan bahwa *Z-Score* juga dapat diterapkan pada *dataset* yang tidak berdistribusi normal, terutama jika *dataset* tersebut tidak memiliki *outlier* yang signifikan dan distribusinya tidak terlalu jauh dari normalitas (Indeed Editorial Team, 2024).

4. KESIMPULAN

Penelitian ini membandingkan empat metode normalisasi data (*Min-Max, Z-Score, Decimal Scaling,* dan *MaxAbs*) untuk meningkatkan akurasi klasifikasi penyakit menggunakan algoritma K-NN. Hasil penelitian mengindikasikan bahwa kinerja setiap metode dipengaruhi oleh karakteristik unik dari setiap *dataset*, seperti jumlah data, jumlah fitur, dan distribusi data. Metode *Decimal Scaling* memberikan akurasi tertinggi sebesar 90,00% pada *dataset* kanker prostat, sementara *Min-Max* dan *Z-Score* masing-masing mencapai akurasi 97,50% pada *dataset* penyakit ginjal. Metode *MaxAbs* juga menunjukkan performa yang baik dengan akurasi 96,25% pada *dataset* yang sama, sedangkan pada *dataset* penyakit jantung, *Min-Max* dan *MaxAbs* mencapai akurasi masing-masing sebesar 82,93% dan 81,95%.

Berdasarkan hasil penelitian ini, disarankan agar pemilihan metode normalisasi dilakukan secara cermat dengan mempertimbangkan karakteristik spesifik dari *dataset* yang akan dianalisis. Untuk *dataset* yang lebih kecil dan sederhana, *Decimal Scaling* dapat menjadi pilihan yang baik. Namun, untuk *dataset* yang lebih besar dan kompleks, *Min-Max* atau *MaxAbs* umumnya lebih disarankan. Metode *Z-Score* dapat menjadi opsi yang fleksibel dan dapat diterapkan pada berbagai jenis *dataset*, terutama jika data tersebut tidak memiliki *outlier* yang signifikan.

Penelitian ini memberikan kontribusi penting dalam pemahaman tentang pengaruh normalisasi data terhadap kinerja algoritma K-NN dalam konteks klasifikasi penyakit. Hasil penelitian ini dapat menjadi acuan bagi peneliti dan praktisi *data mining* untuk meninjau karakteristik *dataset* secara mendalam sebelum memilih metode normalisasi data. Pendekatan yang tepat dapat meningkatkan akurasi prediksi algoritma klasifikasi secara signifikan, yang pada akhirnya dapat menghasilkan keputusan yang lebih baik dalam analisis data kesehatan. Selain itu, penelitian lebih lanjut diperlukan untuk menguji efektivitas metode normalisasi lainnya atau mengombinasikan beberapa metode guna mencapai hasil yang lebih optimal.

DAFTAR PUSTAKA

Ambarwari, A., Jafar Adrian, Q., & Herdiyeni, Y. (2020). Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(1), 117–122. https://doi.org/10.29207/resti.v4i1.1517

Badugu, S. (2020). Prediction of Heart Problems for Diabetic Patients using Classification Algorithms. *Journal of Advanced Research in Dynamic and Control Systems*, *Volume 12*(02-Special Issue), 904–913. https://doi.org/10.5373/JARDCS/V12SP2/SP20201148

Barus, F. M., & Sutarman, S. (2023). Mendeteksi Outlier pada Data Multivariat dengan Metode Jarak Mahalanobis-Minimum Covariance Determinant (MMCD). *IJM: Indonesian Journal of Multidisciplinary*, *1*(3), 1164–1172. https://journal.csspublishing.com/index.php/ijm/article/view/287

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/10.33096/ijodas.v1i2.13

Chandra, R., Chaudhary, K., & Kumar, A. (2022). Comparison of Data Normalization for Wine Classification Using K-NN Algorithm. *IJIIS: International Journal of Informatics and Information Systems*, *5*(4), 175–180. https://doi.org/10.47738/ijiis.v5i4.145

Henderi, H., Wahyuningsih, T., & Rahwanto, E. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

HS, H., Azmi, N., Hazriani, H., & Yuyun, Y. (2023). Klasifikasi Status Gizi Balita Menggunakan Algoritma K-Nearest Neighbor (KNN) | Prosiding SISFOTEK. *Prosiding SISFOTEK*, *7*(1), 313–318. https://seminar.iaii.or.id/index.php/SISFOTEK/article/view/396

Indeed Editorial Team. (2024, August 16). *Normalization Formula: How To Use It on a Data Set | Indeed.com*. Indeed. https://www.indeed.com/career-advice/career-development/normalization-formula

Jain, A. K., Duin, P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. https://doi.org/10.1109/34.824819

Kusnaidi, M. R., Gulo, T., & Aripin, S. (2022). Penerapan Normalisasi Data Dalam Mengelompokkan Data Mahasiswa Dengan Menggunakan Metode K-Means Untuk Menentukan Prioritas Bantuan Uang Kuliah Tunggal. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 330–338. https://doi.org/10.47065/josyc.v3i4.2112

Marlina, D., & Bakri, M. (2021). Penerapan Data Mining untuk Memprediksi Transaksi Nasabah dengan Algoritma C4.5. *Jurnal Teknologi Dan Sistem Informasi*, *2*(1), 23–28. https://doi.org/10.33365/JTSI.V2I1.627

McLeod, S. (2023, October 6). *Z-Score: Definition, Formula, Calculation & Interpretation*. Simply Psychology. https://www.simplypsychology.org/z-score.html

Pagan, M., Zarlis, M., & Candra, A. (2023). Investigating the impact of data scaling on the k-nearest neighbor algorithm. *Computer Science and Information Technologies*, *4*(2), 135–142. https://doi.org/10.11591/csit.v4i2.p135-142

Permana, I., & Salisah, F. N. S. (2022). Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation. *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, *2*(1), 67–72. https://doi.org/10.57152/ijirse.v2i1.311

Riaz, M., Bashir, M., & Younas, I. (2022). Metaheuristics based COVID-19 detection using medical images: A review. *Computers in Biology and Medicine*, *144*, 105344. https://doi.org/10.1016/j.compbiomed.2022.105344

Sholeh, M., Andayati, D., & Rachmawati, Rr. Y. (2022). Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor dengan Normalisasi untuk Prediksi Penyakit Diabetes. *TeIKa*, *12*(02), 77–87. https://doi.org/10.36342/teika.v12i02.2911

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Whendasmoro, R. G., & Joseph, J. (2022). Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN. *JURIKOM (Jurnal Riset Komputer)*, *9*(4), 872. https://doi.org/10.30865/jurikom.v9i4.4526