Journal of Applied Informatics and Computing (JAIC)
Vol.x, No.x, Juli 201x, pp. xx~xx
e-ISSN: 2548-6861

Perbandingan Teknik Normalisasi Data terhadap Kinerja Klasifikasi KNN pada Dataset Diabetes

Yohanes Dimas Pratama [1]*, Abu Salam [2], Penulis Ketiga [3]***
* Teknik Informatika, Politeknik Negeri Batam
** Teknik Multimedia Jaringan, Politeknik Negeri Batam
mail1@polibatam.ac.id [1], mail2@polibatam.ac.id [2], mail3@polibatam.ac.id [3]

Article Info

Article history:

Received …
Revised …
Accepted …

Keyword:

Pilih maksimum lima kata kunci atau frase yang diurutkan menurut abjad, dan dipisahkan dengan koma. Keyword1, Keyword2, Keyword3.

ABSTRACT

Gunakan dokumen ini sebagai template untuk menyusunan artikel. Artikel dapat berupa full Bahasa Inggris (*-diutamakan-*) atau Bahasa Indonesia. Bagian abstrak memuat informasi terkait penelitian apa yang hendak akan dilakukan dan hasil yang diperoleh pada penelitian harus disampaikan. Setelah makalah diterima, dan perbaikan terakhir silahkan dikirimkan kepada kami, dokumen elektronik ini akan diformat lebih lanjut oleh redaksi JAIC. Dalam abstrak, anda seharusnya tidak merujuk publikasi lainnya. Buatlah abstrak dalam bahasa Inggris.

I. PENDAHULUAN

Diabetes merupakan salah satu penyakit tidak menular yang semakin meningkat prevalensinya di seluruh dunia, termasuk di Indonesia. Menurut data dari Organisasi Kesehatan Dunia (WHO), jumlah penderita diabetes meningkat dari 108 juta pada tahun 1980 menjadi 422 juta pada tahun 2014. Diperkirakan jumlah penderita diabetes akan mencapai 578 juta pada tahun 2030 dan 700 juta pada tahun 2045 [1]. Diabetes terjadi ketika tubuh tidak bisa mengatur kadar gula darah dengan baik, yang menyebabkan kadar gula darah terlalu tinggi [2][3][4]. Diabetes dapat menyebabkan berbagai komplikasi kesehatan serius, seperti kerusakan pada jantung, ginjal, mata, dan sistem saraf [4][5]. Oleh karena itu, deteksi dini dan prediksi risiko diabetes sangat penting untuk mencegah terjadinya komplikasi lebih lanjut. Dalam hal ini, teknologi informasi dan pembelajaran mesin dapat memainkan peran penting dalam mempermudah dan mempercepat diagnosis serta prediksi penyakit diabetes.

Salah satu algoritma pembelajaran mesin yang umum digunakan dalam klasifikasi adalah K-Nearest Neighbors (KNN). KNN merupakan metode klasifikasi yang bekerja berdasarkan kedekatan jarak antara titik data yang akan diklasifikasikan dengan data yang sudah terlabel [6]. Meskipun KNN sederhana dan mudah diimplementasikan, namun ada salah satu tantangan utama dalam menggunakan algoritma ini adalah sensitifitasnya terhadap data yang belum dinormalisasi. KNN mengandalkan perhitungan jarak, seperti Euclidean, untuk mengukur kedekatan antar data [7]. Jika data memiliki skala atau satuan yang berbeda-beda, hal ini dapat menyebabkan ketidakseimbangan dalam perhitungan jarak, yang pada gilirannya dapat menurunkan kinerja model [8]. Karena itu, normalisasi data menjadi langkah penting sebelum diterapkan pada algoritma KNN. Normalisasi bertujuan untuk mengubah nilai fitur ke dalam skala yang seragam, sehingga perhitungan jarak antar data dapat dilakukan secara akurat tanpa dipengaruhi oleh perbedaan skala [9]. Terdapat tiga teknik normalisasi yang relevan dengan perhitungan Euclidean distance, yaitu Min-Max Scaling, Z-Score, dan Decimal Scaling [10].

Beberapa penelitian sebelumnya telah mengkaji perbandingan teknik normalisasi terhadap kinerja berbagai algoritma klasifikasi, yang memiliki dampak signifikan pada akurasi prediksi dalam berbagai dataset. Sebagai contoh, penelitian Muasir Pagan et al. [11] mengkaji perbandingan teknik normalisasi terhadap kinerja algoritma K-Nearest Neighbor (K-NN) dengan menggunakan sepuluh dataset. Mereka mengevaluasi tiga teknik skala data (min-max normalization, Z-score, dan decimal scaling). Hasilnya menunjukkan bahwa Z-score dan decimal scaling

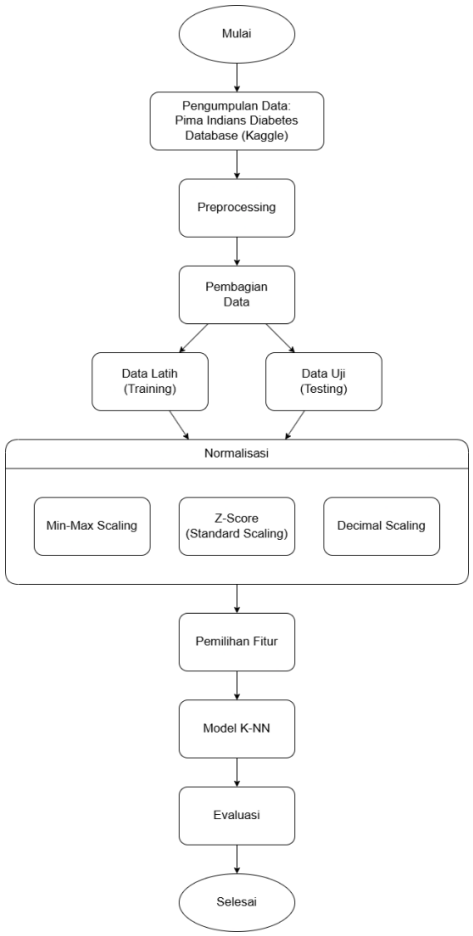
e-ISSN: 2548-6861

memberikan kinerja yang lebih baik dibandingkan min-max normalization, dengan Z-score secara konsisten menghasilkan akurasi, presisi, recall, dan F1-score yang lebih tinggi di sebagian besar dataset. Temuan ini menyoroti pentingnya pemilihan teknik normalisasi yang sesuai berdasarkan karakteristik dataset. Selanjutnya, penelitian Alshdaifat et al. [12] juga mengevaluasi dampak teknik normalisasi (Min-Max, Z-Score, dan Decimal Scaling) terhadap kinerja algoritma klasifikasi seperti SVM dan ANN, dengan temuan yang juga relevan untuk K-Nearest Neighbor (KNN). Hasil penelitian pada 18 dataset menunjukkan bahwa Z-Score Normalization sering kali memberikan hasil terbaik karena kemampuannya menangani outlier, sementara Decimal Scaling dianggap kurang efektif dalam meningkatkan akurasi model secara keseluruhan. Selain itu, penelitian Saichon Sinsomboonthong [13] membandingkan kinerja delapan teknik normalisasi dalam klasifikasi menggunakan ANN pada enam dataset. Hasil penelitian menunjukkan bahwa min-max normalization umumnya memberikan akurasi tertinggi dan MSE terendah. Namun, pada beberapa dataset seperti White Wine Quality dan Pima Indians Diabetes, Adjusted-2 min-max normalization memberikan hasil yang lebih baik. Teknik normalisasi lain seperti Statistical Column dan Decimal Scaling juga menunjukkan hasil kompetitif pada dataset tertentu, namun tidak mengungguli min-max normalization.

Penelitian ini bertujuan untuk menganalisis perbandingan berbagai teknik normalisasi terhadap kinerja model KNN dalam klasifikasi diabetes pada dataset Pima Indians Diabetes Database. Teknik normalisasi diterapkan sebagai bagian dari preprocessing data untuk mengevaluasi dampaknya terhadap kinerja model [14]. Selain itu, pemilihan fitur juga dilakukan untuk memastikan hanya fitur relevan yang digunakan, guna mengurangi overfitting dan meningkatkan interpretabilitas [15]. Dengan membandingkan teknik normalisasi yang berbeda, diharapkan dapat dibangun model yang lebih akurat dan efisien, mendukung deteksi dini diabetes, serta membantu tenaga medis dalam pengambilan keputusan yang lebih tepat.

II. METODE

Penelitian ini dimulai dengan pengumpulan data dari platform Kaggle, diikuti dengan tahapan preprocessing yang mencakup pemeriksaan tipe data, pemeriksaan nilai hilang (missing values), pemeriksaan duplikasi, pemeriksaan skala data numerik, dan pemeriksaan nilai unik pada data kategorikal. Dataset kemudian dibagi menjadi dua bagian, yaitu data latih (training) dan data uji (testing). Selanjutnya, dilakukan normalisasi data menggunakan tiga teknik, yaitu Min-Max Scaling, Z-Score Scaling (Standard Scaling), dan Decimal Scaling. Setelah normalisasi, dilakukan pemilihan fitur (feature selection), yang kemudian dilanjutkan dengan pelatihan model K-Nearest Neighbor (KNN) dan evaluasi kinerja model. Terakhir, dilakukan analisis terhadap hasil evaluasi model untuk setiap metode normalisasi guna mengetahui perbandingan masing-masing metode terhadap kinerja model. Diagram alur penelitian ini disajikan pada Gambar 1.



Gambar 1. Alur Penelitian

A. Pengumpulan Data

Pada tahap pengumpulan data, penelitian ini menggunakan dataset Pima Indians Diabetes Database yang diambil dari Kaggle. Dataset ini berasal dari National Institute of Diabetes and Digestive and Kidney Diseases dengan tujuan untuk memprediksi secara diagnostik apakah seorang pasien mengidap diabetes atau tidak berdasarkan berbagai fitur pengukur medis [16]. Dataset ini mencakup 768 baris data, di mana setiap baris mewakili satu pasien. Setiap baris terdiri dari 8 fitur independen yang digunakan untuk memprediksi kemungkinan diabetes serta satu fitur target yang menunjukkan hasil diagnosis diabetes. Fitur tersebut meliputi jumlah kehamilan (Pregnancies), konsentrasi glukosa plasma (Glucose), tekanan darah diastolik (BloodPressure), ketebalan lipatan kulit (SkinThickness), kadar insulin (Insulin), indeks massa tubuh (BMI), fungsi keturunan diabetes (DiabetesPedigreeFunction), usia (Age), serta fitur target yang menunjukkan hasil diagnosis diabetes (Outcome) [17]. Semua pasien dalam dataset ini adalah perempuan berusia

minimal 21 tahun dengan latar belakang keturunan Pima Indian [18]. Visualisasi dataset digambarkan pada Tabel 1.

TABEL 1
FITUR PIMA INDIANS DIABETES DATASET

Fitur	Deskripsi
Pregnancies	Jumlah kehamilan yang pernah dialami oleh pasien.
Glucose	Konsentrasi glukosa plasma saat berpuasa.
BloodPressure	Tekanan darah diastolik (tekanan darah saat relaksasi jantung).
SkinThickness	Ketebalan lipatan kulit yang diukur di lengan atas.
Insulin	Kadar insulin yang terdapat dalam darah pasien.
BMI	Indeks massa tubuh, yang mencerminkan berat badan pasien relatif terhadap tinggi badan.
Diabetes Pedigree Function	Fungsi keturunan diabetes, yang menggambarkan riwayat keluarga pasien terkait diabetes.
Age	Usia pasien.
Outcome	Variabel target yang menunjukkan apakah pasien didiagnosis mengidap diabetes (1) atau tidak (0).

B. Preprocessing

Preprocessing adalah tahapan penting dalam pengolahan dataset yang bertujuan untuk meningkatkan kualitas data sebelum digunakan dalam pemodelan. Proses ini dilakukan agar informasi yang diekstraksi lebih akurat, sehingga dapat berkontribusi pada peningkatan performa model [19]. Pada penelitian ini, preprocessing mencakup beberapa langkah utama yang dimulai dengan memeriksa tipe data agar setiap fitur memiliki format yang sesuai dan konsisten. Hal ini penting untuk mencegah kesalahan dalam pemrosesan data, terutama saat menerapkan algoritma pembelajaran mesin seperti KNN (K-Nearest Neighbors), yang hanya bisa menghitung jarak antar data jika data berformat numerik [20]. Selanjutnya, dilakukan identifikasi dan penanganan missing values pada setiap fitur. Dalam penelitian ini, data yang memiliki missing values akan dihapus untuk memastikan bahwa hanya data yang lengkap yang digunakan dalam analisis. Pendekatan ini dipilih karena jumlah data yang hilang relatif kecil, sehingga penghapusan tidak berdampak signifikan terhadap keseluruhan dataset [21]. Tahap berikutnya adalah deteksi dan penghapusan data duplikat, karena data yang terduplikasi dapat membuat model terlalu fokus pada pola tertentu, sehingga mengurangi kemampuan generalisasi dan menurunkan akurasi prediksi [22]. Setelah itu, fitur dalam dataset diklasifikasikan berdasarkan jenisnya menjadi fitur numerik dan kategorikal. Untuk fitur numerik, dilakukan analisis skala untuk memahami distribusi dan rentang nilainya. Sementara itu, untuk fitur kategorikal, dilakukan identifikasi terhadap nilai unik yang terdapat di dalamnya [23]. Khusus pada fitur Outcome, distribusi nilai unik dianalisis untuk memastikan keseimbangan kelas data.

C. Pembagian Data

Pada tahap pembagian data, dataset yang telah melalui proses preprocessing akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Pembagian ini dilakukan dengan proporsi 80% untuk data latih dan 20% untuk data uji. Data latih akan digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi performa model yang telah dilatih. Pembagian ini bertujuan untuk memungkinkan model mempelajari pola dari data latih dan menguji kemampuannya dalam memprediksi hasil pada data uji yang belum pernah dilihat sebelumnya, sehingga dapat dinilai kemampuan generalisasi model [24].

D. Normalisasi

Setelah pembagian dataset menjadi data pelatihan dan pengujian, langkah selanjutnya adalah normalisasi data. Normalisasi dilakukan untuk menyelaraskan skala fitur sehingga tidak ada fitur yang mendominasi perhitungan dalam algoritma berbasis jarak seperti K-Nearest Neighbors (KNN). Dalam algoritma ini, perhitungan jarak terutama Euclidean Distance sangat bergantung pada skala data, sehingga perbedaan rentang nilai antar fitur dapat menyebabkan distorsi dalam proses klasifikasi [25]. Oleh karena itu, normalisasi menjadi tahap krusial untuk memastikan setiap fitur memiliki bobot yang seimbang dalam analisis. Pada penelitian ini, normalisasi diterapkan menggunakan tiga metode, yaitu Min-Max Scaling, Z-Score Normalization (Standard Scaling), dan Decimal Scaling. Masing-masing metode memiliki karakteristik dan manfaat spesifik dalam mengubah distribusi data agar lebih optimal untuk analisis KNN. Penjelasan lebih lanjut mengenai masing-masing metode akan dijelaskan sebagai berikut:

1. Min-Max Scaling

   Min-Max Scaling diterapkan untuk menormalisasi data dengan mengubah rentang nilai fitur ke dalam skala 0 hingga 1. Proses ini dilakukan dengan merumuskan ulang setiap nilai berdasarkan nilai minimum dan maksimum dalam dataset. Dengan demikian, distribusi data tetap terjaga, tetapi dalam skala yang lebih seragam, sehingga model dapat mengolahnya tanpa bias akibat perbedaan skala antar fitur [26]. Rumus Min-Max Scaling adalah sebagai berikut:

   $$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

   Nilai x merupakan data asli sebelum dinormalisasi, sementara $\min(X)$ dan $\max(X)$ masing-masing mewakili nilai terkecil dan terbesar dalam dataset. Setelah dilakukan proses normalisasi, nilai x' menjadi data yang telah disesuaikan dalam rentang [0, 1].
2. Z-Score (Standard Scaling)

e-ISSN: 2548-6861

Z-Score atau Standard Scaling digunakan untuk menstandarisasi data dengan mereskalakan nilai fitur sehingga memiliki rata-rata 0 dan standar deviasi 1. Metode ini menghitung sejauh mana suatu nilai menyimpang dari rata-rata dalam satuan standar deviasi, sehingga memungkinkan perbandingan antar fitur yang memiliki skala berbeda [27]. Rumus Z-Score adalah sebagai berikut:

$$z = \frac{x - \mu}{\sigma}$$

Nilai x merupakan data asli sebelum distandarisasi, sedangkan μ mewakili rata-rata (mean) dari suatu fitur, dan σ adalah standar deviasi yang menggambarkan sebaran data. Setelah proses standardisasi, nilai z diperoleh untuk menunjukkan seberapa jauh x berada dari rata-rata dalam satuan deviasi standar.

3. Decimal Scaling

Decimal Scaling digunakan untuk menormalisasi data dengan membagi setiap nilai dengan pangkat sepuluh yang sesuai, sehingga semua nilai berada dalam rentang yang lebih kecil. Faktor pembagi ditentukan berdasarkan jumlah digit terbesar dalam dataset, sehingga skala data tetap proporsional tanpa mengubah distribusi relatif antar nilai [13]. Rumus Decimal Scaling adalah sebagai berikut:

$$x^* = \frac{x}{10^j}$$

Nilai x merupakan data asli sebelum dinormalisasi, sementara 10^j adalah faktor pembagi yang ditentukan berdasarkan jumlah digit desimal yang diperlukan agar nilai x berada dalam rentang yang lebih kecil. Setelah proses normalisasi menggunakan Decimal Scaling, nilai x^* diperoleh sebagai hasil pembagian x dengan faktor 10^j.

E. Pemilihan Fitur

Setelah tahap normalisasi, langkah selanjutnya adalah melakukan feature selection menggunakan metode Random Forest. Feature selection adalah proses penting untuk memilih fitur-fitur yang paling relevan dan signifikan dalam model, sehingga dapat meningkatkan akurasi serta efisiensi komputasi [28]. Random Forest, yang merupakan algoritma berbasis pohon keputusan, dapat digunakan untuk menentukan pentingnya setiap fitur dalam memprediksi target fitur [29]. Feature selection akan dilakukan pada dua versi data, yaitu data yang sudah ternormalisasi dan data yang tidak dilakukan normalisasi.

Pada tahap ini, Random Forest akan mengevaluasi kontribusi relatif dari setiap fitur dengan cara menghitung feature importance berdasarkan seberapa besar kontribusi masing-masing fitur dalam mengurangi ketidakpastian (impurity) dalam pohon keputusan. Fitur yang memiliki importance tinggi akan dipertahankan, sementara fitur dengan importance rendah bisa dihapus, sehingga model dapat fokus pada fitur yang lebih relevan [30]. Proses ini membantu mengurangi kompleksitas model dan mencegah overfitting, yang pada gilirannya dapat meningkatkan performa model dalam memprediksi data yang belum pernah dilihat sebelumnya.

F. Model KNN (K-Nearest Neighbors)

Setelah proses pemilihan fitur, tahap berikutnya adalah melatih model K-Nearest Neighbors (KNN) pada setiap metode normalisasi yang telah diterapkan sebelumnya. Pelatihan model akan dilakukan pada dua versi data, yaitu data yang sudah ternormalisasi dan data yang tidak dilakukan normalisasi. Model KNN akan diuji dengan berbagai nilai k, yaitu 1, 3, 5, 7, 9, dan 11, untuk menentukan nilai k yang memberikan performa terbaik. Dalam algoritma KNN, klasifikasi dilakukan berdasarkan kedekatan suatu data dengan sejumlah k tetangga terdekatnya dalam ruang fitur [6]. Oleh karena itu, pemilihan nilai k yang tepat menjadi faktor krusial dalam kinerja model. Nilai k yang terlalu kecil dapat menyebabkan model terlalu sensitif terhadap data training (overfitting), sedangkan nilai k yang terlalu besar dapat menyebabkan model menjadi terlalu umum (underfitting) [31][32]. Hasil dari tahap ini digunakan untuk menentukan nilai k terbaik berdasarkan akurasi pada data uji. Nilai k terbaik yang diperoleh kemudian digunakan sebagai acuan untuk membandingkan performa pada berbagai metode normalisasi dalam algoritma KNN.

G. Evaluasi

Pada tahap evaluasi, kinerja model K-Nearest Neighbors (KNN) yang telah dilatih akan diuji menggunakan berbagai metrik evaluasi untuk menilai efektivitas model dalam mengidentifikasi pasien dengan diabetes serta menganalisis perbandingan metode normalisasi terhadap performa klasifikasi. Setelah diperoleh nilai k terbaik berdasarkan hasil pengujian, analisis akan difokuskan pada bagaimana setiap metode normalisasi memengaruhi hasil klasifikasi.

Evaluasi model dilakukan dengan mempertimbangkan empat komponen utama dalam analisis klasifikasi, yaitu True Positive (TP), yang menunjukkan jumlah kasus di mana model benar dalam memprediksi pasien mengidap diabetes; True Negative (TN), yang mengindikasikan jumlah kasus ketika model dengan benar memprediksi pasien tidak mengidap diabetes; False Positive (FP), yang terjadi saat model salah memprediksi pasien mengidap diabetes padahal sebenarnya tidak; serta False Negative (FN), yang terjadi ketika model salah memprediksi pasien tidak mengidap diabetes padahal sebenarnya mengidap. Keempat komponen ini menjadi dasar dalam menghitung berbagai metrik evaluasi yang digunakan untuk mengukur kinerja model secara menyeluruh. Hasil evaluasi ini mencakup berbagai metrik yang digunakan untuk menilai performa model yang akan dijelaskan sebagai berikut:

1. Akurasi

Metrik ini mengukur sejauh mana model dapat melakukan prediksi yang benar dibandingkan dengan

total prediksi. Akurasi memberikan gambaran umum tentang performa model, tetapi kurang berguna pada data yang tidak seimbang. Akurasi dihitung dengan rumus:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision
Precision mengukur akurasi prediksi positif yang dilakukan oleh model. Metrik ini digunakan untuk mengetahui seberapa banyak prediksi positif yang akurat atau sesuai dengan kondisi yang sebenarnya. Precision dihitung dengan rumus:

$$Precision = \frac{TP}{TP + FP}$$

3. Recall
Recall mengukur kemampuan model dalam menemukan semua kasus positif yang sebenarnya. Metrik ini penting ketika sangat penting untuk mendeteksi sebanyak mungkin kasus positif. Recall dihitung dengan rumus:

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score
F1-Score adalah rata-rata harmonik antara precision dan recall. Metrik ini penting untuk memberikan gambaran keseimbangan antara kemampuan model dalam mendeteksi kelas positif dan negatif, terutama pada dataset yang tidak seimbang. F1-score dihitung dengan rumus:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. Specificity
Specificity mengukur kemampuan model dalam mengidentifikasi data negatif dengan benar. Metrik ini penting untuk memastikan model tidak salah mengklasifikasikan data negatif sebagai positif. Specificity dihitung dengan rumus:

$$Specificity = \frac{TN}{TN + FP}$$

6. ROC AUC
ROC-AUC mengukur kemampuan model dalam membedakan antara kelas positif dan negatif. Nilai AUC yang lebih tinggi menunjukkan model yang lebih baik dalam klasifikasi pada berbagai threshold.
7. Confusion Matrix
Confusion matrix memberikan informasi tentang jumlah prediksi yang benar dan salah dalam setiap kategori (positif atau negatif). Metrik ini memungkinkan analisis mendalam mengenai kesalahan klasifikasi dan membantu menghitung metrik lain seperti precision, recall, F1-score, dan specificity untuk evaluasi model.
8. P-value
P-Value mengukur seberapa signifikan hasil model dalam membedakan kelas positif dan negatif. Nilai p yang kecil (biasanya < 0,05) menunjukkan perbedaan yang signifikan secara statistik. P-value dihitung dengan uji statistik seperti uji t atau ANOVA untuk membandingkan kinerja model. Jika p-value rendah, perbedaan antara model atau metode normalisasi dianggap signifikan.

III. HASIL DAN PEMBAHASAN

Pada bab ini, akan dijelaskan hasil yang diperoleh dari setiap tahapan yang telah dijelaskan pada bagian metode penelitian. Berikut adalah penjelasan mengenai hasil yang didapat dari proses yang telah dilakukan:

A. Preprocessing Data

Pada tahap awal preprocessing, dilakukan pemeriksaan terhadap tipe data untuk memastikan bahwa setiap fitur dalam dataset memiliki format yang sesuai. Dalam KNN, perhitungan jarak, seperti menggunakan Euclidean distance, bergantung pada angka yang dapat dihitung secara matematis. Oleh karena itu, jika dataset berisi data non-numerik, algoritma ini tidak akan dapat melakukan perhitungan jarak yang benar, sehingga dapat mempengaruhi akurasi hasil prediksi.

TABEL 2
TIPE DATA PADA SETIAP FITUR DATASET

Fitur	Tipe Data
Pregnancies	int64
Glucose	int64
BloodPressure	int64
SkinThickness	int64
Insulin	int64
BMI	float64
DiabetesPedigreeFunction	float64
Age	int64
Outcome	int64

Dari Tabel 2, dapat dilihat bahwa sebagian besar fitur memiliki tipe data int64, kecuali fitur BMI dan DiabetesPedigreeFunction yang bertipe float64. Meskipun terdapat perbedaan antara int64 dan float64, kedua tipe data tersebut tetap termasuk dalam kategori numerik dan dapat diproses oleh algoritma KNN tanpa memerlukan konversi tipe data tambahan.

Pada tahap selanjutnya dalam preprocessing, dilakukan pemeriksaan terhadap missing values dan data duplikat. Jika ditemukan data yang memiliki missing values, data tersebut akan dihapus untuk memastikan hanya data yang lengkap yang digunakan dalam analisis. Begitu juga, jika ditemukan data duplikat, data tersebut akan dihapus untuk mencegah pengaruhnya terhadap hasil analisis, yang dapat menyebabkan model menjadi terlalu fokus pada pola tertentu atau memberikan hasil yang tidak akurat.

TABEL 3
MISSING VALUES DAN DATA DUPLIKAT PADA FITUR DATASET

Fitur	Missing Values	Data Duplikat
Pregnancies	0	0

e-ISSN: 2548-6861

Glucose	0	0
BloodPressure	0	0
SkinThickness	0	0
Insulin	0	0
BMI	0	0
DiabetesPedigreeFunction	0	0
Age	0	0
Outcome	0	0

Dari Tabel 3, dapat dilihat bahwa tidak ada fitur yang memiliki missing values, dan juga tidak ditemukan data duplikat pada dataset ini. Ini menunjukkan bahwa dataset sudah dalam kondisi yang baik, dengan data yang lengkap dan unik. Dengan demikian, tidak perlu ada penghapusan data atau penanganan lebih lanjut terkait missing values atau duplikat.

Langkah selanjutnya dalam preprocessing adalah mengklasifikasikan fitur dalam dataset berdasarkan jenisnya, yaitu fitur numerik dan kategorikal. Fitur numerik mencakup variabel yang memiliki nilai numerik yang dapat dihitung dan digunakan dalam perhitungan matematis. Sementara itu, fitur kategorikal berisi variabel yang mengelompokkan data ke dalam kategori atau kelas tertentu. Hasil klasifikasi fitur berdasarkan jenisnya pada Tabel 4.

TABEL 4
KLASIFIKASI FITUR BERDASARKAN JENISNYA

Jenis Data	Fitur
Numerikal	Pregnancies
	Glucose
	BloodPressure
	SkinThickness
	Insulin
	BMI
	DiabetesPedigreeFunction
	Age
Kategorikal	Outcome

Dari Tabel 4, kita dapat melihat bahwa seluruh fitur kecuali Outcome termasuk dalam kategori numerik. Outcome dianggap sebagai fitur kategorikal karena berisi informasi kelas atau hasil diagnosis diabetes (0 atau 1). Setelah pengklasifikasian ini, langkah selanjutnya adalah melakukan analisis lebih lanjut terhadap fitur numerik, seperti memahami distribusi dan rentang nilainya. Untuk fitur kategorikal, langkah berikutnya adalah memeriksa nilai unik yang terdapat dalam fitur tersebut. Analisis terhadap fitur numerik dan kategorikal telah dilakukan, seperti yang ditampilkan pada Tabel berikut:

1. Data Numerik
   Distribusi dan rentang nilai fitur numerik diperiksa menggunakan statistik deskriptif. Statistik ini memberikan informasi mengenai rata-rata (mean), standar deviasi (std), nilai minimum (min), kuartil (25%, 50%, 75%), dan nilai maksimum (max) untuk setiap fitur numerik.

TABEL 5
STATISTIK DESKRIPTIF PADA DATA NUMERIK

Fitur	Count	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	768	3.845	3.37	0	1	3	6	17
Glucose	768	120.895	31.973	0	99	117	140.25	199
BloodPressure	768	69.105	19.356	0	62	72	80	122
SkinThickness	768	20.536	15.952	0	0	23	40	99
Insulin	768	79.799	115.244	0	0	30.5	127.25	846
BMI	768	31.993	7.884	0	27.3	32	36.6	67.1
DiabetesPedigreeFunction	768	0.472	0.331	0.078	0.244	0.372	0.626	2.42
Age	768	33.241	11.76	21	24	29	41	81

Berdasarkan analisis statistik deskriptif pada fitur numerik yang dapat dilihat pada Tabel 5, terlihat bahwa terdapat variasi yang signifikan dalam rentang nilai beberapa fitur. Misalnya, fitur "Glucose" memiliki nilai antara 0 hingga 199, dengan rata-rata 120.90 dan deviasi standar 31.97. Sementara itu, fitur "Age" memiliki rentang nilai antara 21 hingga 81, dengan rata-rata 33.24 dan deviasi standar 11.76. Selain itu, fitur "Pregnancies" memiliki nilai maksimum 17 dan rata-rata 3.85, sementara fitur "Insulin" menunjukkan deviasi standar yang sangat tinggi, mencapai 115.24, yang menunjukkan adanya variasi yang besar dalam data. Fitur seperti "SkinThickness" memiliki nilai minimum 0, dan rata-rata yang cukup rendah (20.54), yang bisa mempengaruhi model karena perbedaan skala yang sangat besar antar fitur. Fitur-fitur seperti "BMI" juga menunjukkan variabilitas yang cukup besar, dengan rentang nilai antara 0 hingga 67.1, rata-rata 31.99, dan deviasi standar 7.88. Perbedaan skala yang sangat besar antar fitur ini dapat mempengaruhi kinerja model, karena fitur dengan rentang nilai yang lebih besar cenderung mendominasi pembelajaran model. Oleh karena itu, normalisasi diperlukan untuk menyelaraskan skala dan rentang nilai antar fitur, sehingga setiap fitur dapat berkontribusi secara seimbang dalam model yang dibangun.

2. Data Kategorikal
   Untuk fitur kategorikal, dilakukan pemeriksaan terhadap nilai unik yang ada. Dapat dilihat pada Tabel 6

menunjukkan bahwa fitur Outcome hanya memiliki dua nilai unik, yaitu [0, 1].

TABEL 6
NILAI UNIK PADA DATA KATEGORIKAL

Fitur	Nilai Unik
Outcome	0
	1

Fitur Outcome ini merupakan fitur target dalam dataset yang mengindikasikan apakah seorang pasien mengidap diabetes (1) atau tidak mengidap diabetes (0). Karena hanya memiliki dua nilai unik, fitur ini dapat diperlakukan sebagai fitur kategorikal biner. Dengan hanya dua kelas, fitur Outcome tidak memerlukan normalisasi.

B. Pembagian Data

Pada tahap ini, akan dilakukan pembagian dataset menjadi dua bagian, yaitu Data Latih dan Data Uji. Pembagian ini dilakukan dengan proporsi 80% untuk data latih dan 20% untuk data uji. Data latih digunakan untuk melatih model, sementara data uji digunakan untuk menguji performa model setelah dilatih. Dataset yang telah dibagi dapat dilihat pada Tabel 7. Data latih terdiri dari 614 baris dan 9 kolom, sementara data uji terdiri dari 154 baris dan 9 kolom. Dengan demikian, dataset sudah siap untuk digunakan dalam tahap pemodelan dan evaluasi selanjutnya.

TABEL 7
PEMBAGIAN DATASET LATIH DAN UJI

Data	Jumlah Baris	Jumlah Kolom
Data Training	614	9
Data Testing	154	9

C. Normalisasi

Setelah pembagian data menjadi data training dan testing, data akan dilakukan normalisasi. Data asli dapat dilihat pada Tabel 8 dan 9 yang memberikan gambaran tentang rentang dan variasi nilai pada setiap fitur dalam data training dan data testing sebelum proses normalisasi dilakukan. Dengan melihat data asli ini, dapat lebih jelas dipahami adanya perbedaan skala yang signifikan antar fitur.

TABEL 8
DATA TRAINING SEBELUM DILAKUKAN NORMALISASI

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	2	84	0	0	0	0	0.304	21	0
2	9	112	82	24	0	28.2	1.282	50	1
3	1	139	46	19	83	28.7	0.654	22	0
4	0	161	50	0	0	21.9	0.254	65	0
5	6	134	80	37	370	46.2	0.238	46	1
…	…	…	…	…	…	…	…	…	…
614	0	125	96	0	0	22.5	0.262	21	0

TABEL 9
DATA TESTING SEBELUM DILAKUKAN NORMALISASI

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	6	98	58	33	190	34	0.43	43	0
2	2	112	75	32	0	35.7	0.148	21	0
3	2	108	64	0	0	30.8	0.158	21	0
4	8	107	80	0	0	24.6	0.856	34	0
5	7	136	90	0	0	29.9	0.21	50	0
…	…	…	…	…	…	…	…	…	…
154	8	74	70	40	49	35.3	0.705	39	0

Berdasarkan analisis statistik deskriptif pada fitur numerik yang disajikan dalam Tabel 5, setiap fitur menunjukkan rentang nilai yang bervariasi. Beberapa fitur memiliki nilai minimum nol, sementara yang lain memiliki rentang yang jauh lebih besar. Perbedaan skala ini dapat memengaruhi proses pembelajaran model, karena fitur dengan nilai lebih besar cenderung mendominasi. Oleh karena itu, normalisasi diperlukan untuk menyamakan skala dan rentang nilai antar fitur. Selain itu, berdasarkan pemeriksaan nilai unik pada fitur kategorikal di Tabel 6, fitur Outcome hanya memiliki dua nilai, yaitu 0 dan 1. Karena hanya memiliki dua kelas, fitur ini termasuk dalam kategori biner sehingga tidak memerlukan normalisasi. Dengan demikian, normalisasi akan diterapkan pada semua fitur kecuali Outcome. Proses normalisasi akan diterapkan pada data training dan testing. Selanjutnya, akan dibahas lebih lanjut mengenai detail metode normalisasi yang diterapkan pada data ini:

1. Min-Max Scaling

e-ISSN: 2548-6861

Min-Max Scaling menormalisasi data dengan mengubah nilai fitur sehingga berada dalam rentang yang seragam, yaitu antara 0 hingga 1. Proses ini dilakukan dengan menghitung nilai minimum dan maksimum dari data training, kemudian menggunakan nilai tersebut untuk menyesuaikan skala fitur ke rentang tertentu, seperti 0 hingga 1. Dengan demikian, fitur yang memiliki rentang nilai besar akan diperkecil, sementara fitur dengan rentang nilai kecil akan diperbesar, tetapi tetap mempertahankan proporsi antar nilai. Pada data testing, transformasi dilakukan menggunakan nilai minimum dan maksimum yang telah dihitung dari data training tanpa melakukan perhitungan ulang. Hasil dari normalisasi ini dapat dilihat pada Tabel 10 untuk data training dan Tabel 11 untuk data testing.

TABEL 10
DATA TRAINING SETELAH DILAKUKAN MIN-MAX SCALING

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	0.117647059	0.422110553	0	0	0	0	0.096498719	0	0
2	0.529411765	0.56281407	0.672131148	0.380952381	0	0.420268256	0.514090521	0.483333333	1
3	0.058823529	0.698492462	0.37704918	0.301587302	0.098108747	0.427719821	0.245943638	0.016666667	0
4	0	0.809045226	0.409836066	0	0	0.326378539	0.075149445	0.733333333	0
5	0.352941176	0.673366834	0.655737705	0.587301587	0.437352246	0.68852459	0.068317677	0.416666667	1
…	…	…	…	…	…	…	…	…	…
614	0	0.628140704	0.786885246	0	0	0.335320417	0.078565329	0	0

TABEL 11
DATA TESTING SETELAH DILAKUKAN MIN-MAX SCALING

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	0.352941176	0.492462312	0.475409836	0.523809524	0.224586288	0.506706408	0.15029889	0.366666667	0
2	0.117647059	0.56281407	0.614754098	0.507936508	0	0.532041729	0.029888984	0	0
3	0.117647059	0.542713568	0.524590164	0	0	0.459016393	0.034158839	0	0
4	0.470588235	0.537688442	0.655737705	0	0	0.36661699	0.332194705	0.216666667	0
5	0.411764706	0.683417085	0.737704918	0	0	0.445603577	0.056362084	0.483333333	0
…	…	…	…	…	…	…	…	…	…
154	0.470588235	0.371859296	0.573770492	0.634920635	0.057919622	0.526080477	0.267719898	0.3	0

2. Z-Score (Standard Scaling)

Z-Score Scaling mengubah data dengan cara menstandarisasi setiap fitur sehingga memiliki rata-rata 0 dan deviasi standar 1. Proses Z-Score ini menggunakan nilai mean dan standar deviasi yang dihitung dari data training untuk menstandarkan fitur sehingga memiliki distribusi dengan mean nol dan standar deviasi satu. Parameter ini kemudian diterapkan pada data testing tanpa menghitung ulang statistik baru. Dengan cara ini, data pada data testing akan disesuaikan menggunakan parameter yang diperoleh dari data training, memastikan konsistensi dalam distribusi data. Hasil dari normalisasi ini dapat dilihat pada Tabel 12 untuk data training dan Tabel 13 untuk data testing.

TABEL 12
DATA TRAINING SETELAH DILAKUKAN Z-SCORE (STANDARD SCALING)

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	-0.526396861	-1.151397924	-3.75268255	-1.322773646	-0.701205532	-4.135255779	-0.49073479	-1.035940379	0
2	1.588045858	-0.276642826	0.680344855	0.233505192	-0.701205532	-0.489168806	2.41502991	1.487100846	1
3	-0.828460107	0.566871018	-1.265862298	-0.090719566	0.013448315	-0.424521874	0.549160552	-0.948938958	0
4	-1.130523353	1.254178595	-1.049617059	-1.322773646	-0.701205532	-1.303720151	-0.639291267	2.792122169	0

5	0.681856121	0.410664751	0.572222235	1.076489563	2.484600775	1.838120751	-0.68682934	1.139095159	1
…	…	…	…	…	…	…	…	…	…
614	-1.130523353	0.129493469	1.437203192	-1.322773646	-0.701205532	-1.226143832	-0.615522231	-1.035940379	0

TABEL 13
DATA TESTING SETELAH DILAKUKAN Z-SCORE (STANDARD SCALING)

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	0.681856121	-0.714020375	-0.617126581	0.817109756	0.934749058	0.260735607	-0.116372467	0.878090895	0
2	-0.526396861	-0.276642826	0.301915686	0.752264805	-0.701205532	0.480535176	-0.954231	-1.035940379	0
3	-0.526396861	-0.40160784	-0.292758722	-1.322773646	-0.701205532	-0.153004759	-0.924519704	-1.035940379	0
4	1.285982613	-0.432849094	0.572222235	-1.322773646	-0.701205532	-0.954626717	1.149328721	0.095078101	0
5	0.983919367	0.473147258	1.112835334	-1.322773646	-0.701205532	-0.269369236	-0.770020968	1.487100846	0
…	…	…	…	…	…	…	…	…	…
154	1.285982613	-1.463810459	0.031609137	1.271024418	-0.279301454	0.428817631	0.700688159	0.530085208	0

3. Decimal Scaling

Decimal Scaling menyesuaikan skala fitur dengan membagi nilai setiap fitur dengan pangkat sepuluh yang sesuai. Proses Decimal Scaling ini menyesuaikan skala fitur dengan membagi nilai setiap fitur dengan pangkat 10 berdasarkan jumlah digit terbesar dalam dataset. Karena skala ini ditentukan berdasarkan distribusi keseluruhan data, jika hanya dihitung dari data training, ada kemungkinan distribusi data testing berbeda sehingga skala menjadi tidak konsisten. Pembagian ini dilakukan agar nilai-nilai dalam dataset tidak terlalu besar atau kecil, tetapi tetap mempertahankan proporsi relatif antar data. Teknik ini sangat sederhana, karena hanya melibatkan pembagian dengan angka tetap, dan memastikan bahwa distribusi data tetap terjaga dalam rentang yang lebih kecil. Hasil dari normalisasi ini dapat dilihat pada Tabel 14 untuk data training dan Tabel 15 untuk data testing.

TABEL 14
DATA TRAINING SETELAH DILAKUKAN DECIMAL SCALING

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	0.02	0.084	0	0	0	0	0.0304	0.21	0
2	0.09	0.112	0.082	0.24	0	0.282	0.1282	0.5	1
3	0.01	0.139	0.046	0.19	0.083	0.287	0.0654	0.22	0
4	0	0.161	0.05	0	0	0.219	0.0254	0.65	0
5	0.06	0.134	0.08	0.37	0.37	0.462	0.0238	0.46	1
…	…	…	…	…	…	…	…	…	…
614	0	0.125	0.096	0	0	0.225	0.0262	0.21	0

TABEL 15
DATA TESTING SETELAH DILAKUKAN DECIMAL SCALING

No	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
1	0.06	0.098	0.058	0.33	0.19	0.34	0.043	0.43	0
2	0.02	0.112	0.075	0.32	0	0.357	0.0148	0.21	0
3	0.02	0.108	0.064	0	0	0.308	0.0158	0.21	0

e-ISSN: 2548-6861

4	0.08	0.107	0.08	0	0	0.246	0.0856	0.34	0
5	0.07	0.136	0.09	0	0	0.299	0.021	0.5	0
…	…	…	…	…	…	…	…	…	…
154	0.08	0.074	0.07	0.4	0.049	0.353	0.0705	0.39	0

D. Pemilihan Fitur

Judul makalah harus berukuran 24 pt, *Times New Roman*, rata tengah, seperti contoh. Nama penulis harus dalam ukuran 11 pt. Nama institusi penulis harus dalam ukuran 10 pt dan cetak miring (*italic*). Alamat email harus dalam ukuran 9 pt dengan teks *Courier*.

E. Model KNN (K-Nearest Neighbors)

Judul makalah harus berukuran 24 pt, *Times New Roman*, rata tengah, seperti contoh. Nama penulis harus dalam ukuran 11 pt. Nama institusi penulis harus dalam ukuran 10 pt dan cetak miring (*italic*). Alamat email harus dalam ukuran 9 pt dengan teks *Courier*.

F. Evaluasi

Judul makalah harus berukuran 24 pt, *Times New Roman*, rata tengah, seperti contoh. Nama penulis harus dalam ukuran 11 pt. Nama institusi penulis harus dalam ukuran 10 pt dan cetak miring (*italic*). Alamat email harus dalam ukuran 9 pt dengan teks *Courier*.

Setiap huruf pertama pada setiap kata pada judul diketik dalam huruf besar kecuali kata - kata penghubung seperti “di”, “dan”, “atau”, “dengan”, “ke”, “yang”, “untuk”, “dari”, “jika”, atau “dari”.

Gelar akademis (seperti Dr., Ir., atau ST.) maupun gelar profesional (seperti Direktur atau Manajer) tidak boleh dicantumkan dalam nama penulis.

Untuk menghindari kebingungan, nama belakang atau nama keluarga penulis harus dituliskan di akhir. Contohnya Widodo B. Wahyu.

Setiap penulis harus mencantumkan informasi afiliasi mereka, minimum nama institusi dan alamat di mana penulis bekerja. Apabila tidak bekerja di institusi perguruan tinggi, penulis juga tetap perlu mencantumkan nama dan alamat perusahaan tempat kerja.

Alamat email wajib dicantumkan sebagai informasi kontak pengarang.

G. Sub-Bab

Sub-bab tidak boleh lebih dari 3 tingkatan. Semua judul sub-bab harus diketik dalam teks 10 pt. Setiap huruf pertama pada setiap kata pada judul diketik dalam huruf besar kecuali kata-kata penghubung seperti “di”, “dan”, “atau”, “dengan”, “ke”, “yang”, “untuk”, “dari”, “jika”, atau “dari”.

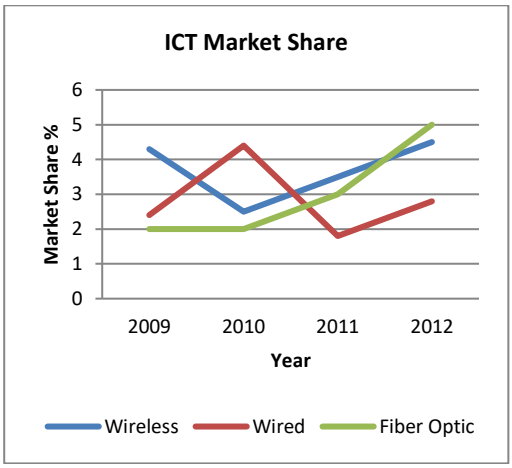
1) *Judul Bab tingkat 1*: Judul Bab tingkat 1 harus disusun dalam *Small Caps*, rata tengah dan dinomori dengan nomor romawi (I, II, III, dst). Contohnya lihat Bab I Pendahuluan.

2) *Judul Bab tingkat II:* Judul Bab tingkat 2 harus dibuat cetak miring (*italic*) dan diberi nomor A, B, C, D, dan seterusnya, diikuti titik. Untuk judul, setiap huruf pertama pada setiap kata pada judul diketik dalam huruf besar kecuali kata-kata penghubung seperti “di”, “dan”, “atau”, “dengan”, “ke”, “yang”, “untuk”, “dari”, “jika”, atau “dari”.

3) *Judul Bab tingkat III:* Judul Bab tingkat 3 harus diketik menjorok ke dalam. Kata-kata dalam cetak miring (*italic*) dan diberi nomor urut 1,2,3, dst. Judul Bab tingkat 3 diikuti dengan titik dua (:) seperti pada contoh di file ini. Isi dari Bab tingkat 3 harus langsung mengikuti tanda titik dua di paragraf yang sama. Contohnya, paragraf ini adalah isi untuk bab tingkat 3.

H. Gambar dan Tabel

Gambar-gambar dan Tabel-Tabel harus dibuat rata tengah dalam 1 kolom. Jika gambar dan Tabelnya sangat besar, dapat dibuat sepanjang lebar halaman menggunakan kedua kolom. Jika anda membuat Tabel dan gambar yang menggunakan lebih dari 1 kolom, maka Tabel atau gambar tersebut harus ditempatkan di paling atas atau paling bawah halaman yang bersangkutan.



Gambar 1.Contoh di atas adalah contoh grafik garis menggunakan warna yang berbeda yang dapat memberikan kontras yang baik di layar maupun di proceeding cetakan.

Gambar 1 menunjukkan contoh gambar yang memiliki resolusi rendah yang tidak akan diterima, sedangkan Gambar 3 menunjukkan gambar dengan resolusi yang cukup. Pastikan

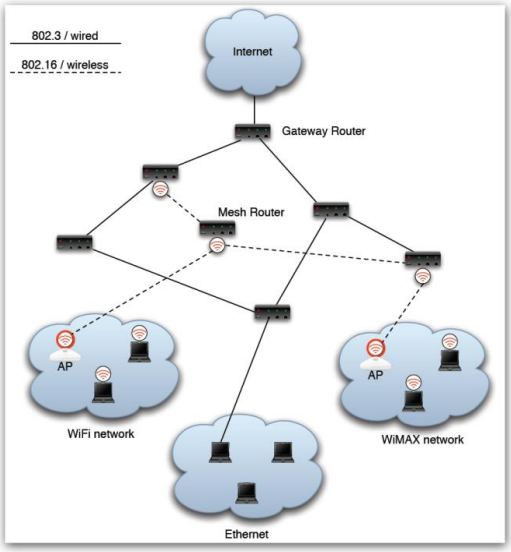
bahwa resolusi gambar sudah mencukupi untuk mendapatkan detil yang diperlukan.

Harap periksa semua gambar dalam makalah anda, baik di layar maupun dalam cetakan hitam putih. Pada saat anda memeriksa cetakan hitam putih, harap periksa:

- Semua warna yang digunakan memberikan kontras yang baik dalam bentuk warna maupun hitam putih.
- Gambar yang dipakai terlihat jelas,
- Semua tulisan yang ada pada gambar harus terbaca.

I. Judul Gambar

Gambar-gambar harus diberi nomor urut 1,2,3,....dst. font *Times new Roman*, 8 pt. Judul yang hanya 1 barus harus dibuat rata tengah (contoh Gambar 1), sementara judul yang lebih dari 1 baris (contoh Gambar 2) harus dibuat rata kiri kanan (*justified*). Penyebutan rujukan pada gambar harus segera ditempatkan setelah gambar yang bersangkutan contohnya pada Gambar 1.



Gambar 3. Contoh gambar dengan resolusi yang dapat diterima

J. Judul Tabel

Tabel-Tabel harus diberi nomor dengan angka Romawi I, II, III,..dst. Judul Tabel ditempatkan di tengah, dengan *Times New Roman*, 8 pt, *Small Caps*. Setiap kata pada Tabel harus dibuat dalam huruf besar, kecuali kata-kata sambung seperti "di", "dan", "atau", "dengan", "ke", "yang", "untuk", "dari", "jika", atau "dari". Paragraf yang merujuk pada Tabel yanbersangkutan harus ditempatkan sebelum Tabel tersebut seperti terlihat pada Tabel 1.

K. Nomor Halaman, Header, dan Footer

Tidak boleh ada penomoran halaman, header, maupun footer. Panitia akan menambahkan bagian ini sebelum prosiding dicetak.

L. Tautan dan Bookmarks

Semua tautan ke alamat internet dan *bookmarks* akan dihapuskan dari makalah pada saat pemrosesan makalah untuk publikasi. Jika anda perlu memberikan referensi pada alamat email tertentu di makalah anda, anda harus mengetikkan alamat email maupun tautan dengan menggunakan teks tanpa menggunakan format *links*.

M. Daftar Pustaka

Daftar pustaka menggunakan aplikasi/tool reference manager atau https://www.mendeley.com/ dengan format IEEE minimal **10 artikel jurnal**.

N. Jumlah Halaman Paper

Jumlah halaman antara 4-15 halaman untuk setiap artikel, sudah termasuk daftar pustaka.

IV. KESIMPULAN

Template JAIC 2020 adalah modifikasi dari template Distributed Framework & Applications 2010 (DFmA 2010) dari IEEE LaTeX yang disediakan oleh Causal Productions (www.causalproductions.com).

UCAPAN TERIMA KASIH

Judul untuk bagian ucapan terima kasih. Ucapan terima kasih diberikan pada institusi atau perusahaan yang mendanai riset.

DAFTAR PUSTAKA

[1] I. W. Suryasa, M. Rodríguez-Gámez, and T. Koldoris, "Health and Treatment of Diabetes Mellitus," *Int J Health Sci (Qassim)*, vol. 5, no. 1, pp. I–V, 2021, doi: 10.53730/IJHS.V5N1.2864.

[2] L. Ryden, G. Ferrannini, and E. Standl, "Risk prediction in patients with diabetes: is SCORE 2D the perfect solution?," Jul. 21, 2023, *Oxford University Press*. doi: 10.1093/eurheartj/ehad263.

[3] S. A. Antar *et al.*, "Diabetes mellitus: Classification, mediators, and complications; A gate to identify potential targets for the development of new effective treatments," Dec. 01, 2023, *Elsevier Masson s.r.l.* doi: 10.1016/j.biopha.2023.115734.

[4] S. Alam, M. K. Hasan, S. Neaz, N. Hussain, M. F. Hossain, and T. Rahman, "Diabetes Mellitus: Insights from Epidemiology, Biochemistry, Risk Factors, Diagnosis, Complications and Comprehensive Management," Jun. 01, 2021, *MDPI*. doi: 10.3390/diabetology2020004.

[5] S. Templer, S. Abdo, and T. Wong, "Preventing diabetes complications," *Intern Med J*, vol. 54, no. 8, pp. 1264–1274, Aug. 2024, doi: 10.1111/imj.16455.

[6] S. Zhang and J. Li, "KNN Classification With One-Step Computation," *IEEE Trans Knowl Data Eng*, vol. 35, no. 3, pp. 2711–2723, Mar. 2023, doi: 10.1109/TKDE.2021.3119140.

[7] N. Ukey, Z. Yang, B. Li, G. Zhang, Y. Hu, and W. Zhang, "Survey on Exact kNN Queries over High-Dimensional Data Space," Jan. 01, 2023, *MDPI*. doi: 10.3390/s23020629.

[8] S. Zhang, "Challenges in KNN Classification," *IEEE Trans Knowl Data Eng*, vol. 34, no. 10, pp. 4663–4675, Oct. 2022, doi: 10.1109/TKDE.2021.3049250.

[9] M. V. Polyakova and V. N. Krylov, "Data normalization methods to improve the quality of classification in the breast cancer diagnostic system," *Applied Aspects of Information Technology*, vol. 5, no. 1, pp. 55–63, Apr. 2022, doi: 10.15276/aait.05.2022.5.

[10] M. Zulkifilu and A. Yasir, "About Some Data Precaution Techniques For K-Means Clustering Algorithm," *UMYU Scientifica*, vol. 1, no. 1, pp. 12–19, 2022, doi: 10.47430/usci.1122.003.

[11] M. Pagan, M. Zarlis, and A. Candra, "Investigating the impact of data scaling on the k-nearest neighbor algorithm," *Computer Science and Information Technologies*, vol. 4, no. 2, pp. 135–142, Jul. 2023, doi: 10.11591/csit.v4i2.pp135-142.

[12] A. Alsarhan, F. Hussein, S. Moh, and F. S. El-Salhi, "The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance," *Data (Basel)*, vol. 6, no. 2, 2021, doi: 10.3390/data.

[13] S. Sinsomboonthong, "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification," *Int J Math Math Sci*, vol. 2022, 2022, doi: 10.1155/2022/3584406.

[14] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput Methods Programs Biomed*, vol. 220, Jun. 2022, doi: 10.1016/j.cmpb.2022.106773.

[15] A. M. Vommi and T. K. Battula, "A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study," *Expert Syst Appl*, vol. 218, May 2023, doi: 10.1016/j.eswa.2023.119612.

[16] Y. Zhao, "Comparative Analysis of Diabetes Prediction Models Using the Pima Indian Diabetes Database," *ITM Web of Conferences*, vol. 70, p. 02021, Jan. 2025, doi: 10.1051/itmconf/20257002021.

[17] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput Appl*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[18] V. Patil and D. R. Ingle, "Comparative Analysis of Different ML Classification Algorithms with Diabetes Prediction through Pima Indian Diabetics Dataset," in *2021 International Conference on Intelligent Technologies, CONIT 2021*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021. doi: 10.1109/CONIT51480.2021.9498361.

[19] H. Karamti *et al.*, "Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach," *Cancers (Basel)*, vol. 15, no. 17, Sep. 2023, doi: 10.3390/cancers15174412.

[20] M. N. Maskuri, K. Sukerti, and R. M. Herdian Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 4, no. 1, May 2022.

[21] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," Mar. 29, 2021, *Frontiers Media S.A.* doi: 10.3389/fenrg.2021.652801.

[22] O. Alotaibi, E. Pardede, and S. Tomy, "Cleaning Big Data Streams: A Systematic Literature Review," Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/technologies11040101.

[23] M. Arif, maruf Setiawan, A. Dwi Hartono, M. Arif Ma, and ruf Setiawan, "Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass," *Jurnal Informatika: Jurnal pengembangan IT*, vol. 10, no. 1, p. 2025, 2025, doi: 10.30591/jpit.v9ix.xxx.

[24] L. A. Demidova, "Two-stage hybrid data classifiers based on svm and knn algorithms," *Symmetry (Basel)*, vol. 13, no. 4, Apr. 2021, doi: 10.3390/sym13040615.

[25] P. J. Muhammad Ali, "Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements," *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, vol. 10, no. 1, pp. 85–91, Jun. 2022, doi: 10.14500/aro.10955.

[26] Henderi , T. Wahyuningsih , and E. Rahwanto, "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," *International Journal of Informatics and Information System*, vol. 4, no. 1, Mar. 2021, [Online]. Available: http://archive.ics.uci.edu/ml.

[27] M. R. Firmansyah and Y. P. Astuti, "Stroke Classification Comparison with KNN through Standardization and Normalization Techniques," *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, Jan. 2024, doi: 10.26877/asset.v6i1.17685.

[28] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," Jun. 27, 2022, *Frontiers Media SA*. doi: 10.3389/fbinf.2022.927312.

[29] M. Alduailij, Q. W. Khan, M. Tahir, M. Sardaraz, M. Alduailij, and F. Malik, "Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method," *Symmetry (Basel)*, vol. 14, no. 6, Jun. 2022, doi: 10.3390/sym14061095.

[30] R. A. Disha and S. Waheed, "Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique," *Cybersecurity*, vol. 5, no. 1, Dec. 2022, doi: 10.1186/s42400-021-00103-8.

[31] Emad Majeed Hameed and Hardik Joshi, "Improving Diabetes Prediction by Selecting Optimal K and Distance Measures in KNN Classifier," *Journal of Techniques*, vol. 6, no. 3, pp. 19–25, Aug. 2024, doi: 10.51173/jt.v6i3.2587.

[32] G. Fatima and S. Saeed, "A Novel Weighted Ensemble Method to Overcome the Impact of Under-fitting and Over-fitting on the Classification Accuracy of the Imbalanced Data Sets," *Pakistan Journal of Statistics and Operation Research*, vol. 17, no. 2, pp. 483–496, 2021, doi: 10.18187/pjsor.v17i2.3640.