JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

Analisis Kinerja Algoritma K-Nearest Neighbor dengan Variasi Validasi Untuk Prediksi Diabetes Mellitus

Performance Analysis of K-Nearest Neighbor Algorithm with Validation Variations for Diabetes Mellitus Prediction

Amelia Laura Ardianti[1*], Hasbi Firmansyah [2], Wahyu Asriyani[3], Ria Indah Fitria[4]
Universitas Pancasakti Tegal
Email: ameliardianti08@gmail.com[1], hasbifirmansyah@upstegal.ac.id[2], asriyani1409@gmail.com[3], ria_indah@upstegal.ac.id[4]*

Article Info

Article history :
Received : 12-12-2025
Revised : 14-12-2025
Accepted : 16-12-2025
Pulished : 18-12-2025

Abstract

Diabetes Mellitus is a non-communicable disease with an increasing prevalence, highlighting the need for accurate prediction methods to support early detection. This study aimed to analyze the performance of the K-Nearest Neighbor (KNN) algorithm in classifying diabetes status using the PIMA Indian Diabetes dataset. The model evaluation was conducted using various k values (k = 3, 5, 7, 9, and 11) and two validation techniques, namely a 70:30 split data approach and 10-fold cross-validation. All modeling processes were performed using RapidMiner Studio with Min–Max normalization. Model performance was evaluated based on accuracy, precision, recall, and confusion matrix metrics. The results indicated that 10-fold cross-validation produced more stable and representative performance compared to the split data technique, achieving the highest accuracy of 73.84% at k = 9 with a relatively low standard deviation. The variation of k values significantly affected the performance of the KNN algorithm, where moderate k values provided the best balance between accuracy and diabetes detection capability. However, the recall value for the positive class indicated the presence of false negative cases. Therefore, the KNN model in this study is more suitable as a decision support system for early Diabetes Mellitus screening rather than a primary diagnostic tool.

Keywords* : *K-Nearest Neighbor, Diabetes Mellitus, RapidMiner

Abstrak

Diabetes Mellitus ialah penyakit tidak menular dengan tingkat prevalensi yang terus meningkat, sehingga diperlukan metode prediksi yang akurat untuk mendukung upaya deteksi dini. Studi ini bertujuan menganalisis kinerja algoritma K-Nearest Neighbor (KNN) dalam mengklasifikasikan status diabetes menggunakan dataset PIMA Indian Diabetes. Evaluasi model dilaksanakan dengan variasi nilai k, yaitu k = 3, 5, 7, 9, dan 11, serta dua teknik validasi, yaitu split data dengan proporsi 70% data latih serta 30% data uji serta 10-fold cross validation. Seluruh proses pemodelan dilakukan menggunakan RapidMiner Studio dengan penerapan normalisasi Min–Max. Kinerja model dievaluasi berlandaskan metrik akurasi, precision, recall, serta confusion matrix. Temuan studi mengindikasikan bahwasanya teknik 10-fold cross validation menghasilkan performa yang lebih stabil dan representatif dibandingkan teknik split data, dengan nilai akurasi tertinggi sebesar 73,84% pada k = 9 serta simpangan baku yang relatif kecil. Variasi nilai k terbukti memengaruhi kinerja algoritma KNN, di mana nilai k pada rentang menengah memberikan keseimbangan terbaik antara akurasi dan kemampuan deteksi kelas diabetes. Meskipun demikian, nilai recall pada kelas

positif masih menunjukkan adanya potensi kesalahan false negative, sehingga model KNN dalam penelitian ini lebih tepat digunakan sebagai sistem pendukung keputusan untuk skrining awal Diabetes Mellitus.

Kata Kunci : K-Nearest Neighbor, Diabetes Mellitus, RapidMiner

PENDAHULUAN

Diabetes Mellitus ialah penyakit tidak menular yang prevalensinya terus bertambah dari tahun ke tahun. Jumlah penderita diabetes di seluruh dunia telah mencapai lebih dari 422 juta jiwa dan terus bertambah dari waktu ke waktu, seiring dengan perubahan gaya hidup, pola makan, serta faktor genetik yang berperan menjadi penyebab utama berkembangnya penyakit ini (World Health Organization, 2021). Kondisi tersebut menjadikan deteksi dini sebagai langkah penting guna mencegah komplikasi jangka panjang yakni penyakit jantung, gagal ginjal, serta neuropati. Pemanfaatan teknologi informasi dalam analisis data kesehatan kemudian menjadi sangat relevan guna mendukung pengambilan keputusan klinis yang lebih cepat, akurat, serta terukur.

Dalam bidang data mining, algoritma K-Nearest Neighbor (KNN) banyak digunakan di studi medis dikarenakan sederhana, mampu bekerja dengan baik pada data numerik, dan tidak memerlukan asumsi distribusi tertentu. Berbagai studi menunjukkan bahwa KNN dapat memberikan performa prediksi yang baik pada kasus medis, termasuk prediksi Diabetes Mellitus, terutama ketika nilai k dan teknik validasi ditentukan secara tepat (Prasetyo & Laksana, 2022). Temuan lain juga mengungkapkan bahwa pemilihan nilai k yang tepat berpengaruh besar terhadap kualitas klasifikasi, khususnya pada dataset PIMA Indian Diabetes yang terdiri dari variabel-variabel numerik dengan keterkaitan satu sama lain (Noori & Yassin, 2021).

Teknik validasi model turut berperan penting dalam menjaga konsistensi hasil prediksi. Metode split data merupakan pendekatan sederhana yang sering digunakan, namun hasil akurasinya dapat berbeda-beda tergantung pada proporsi pembagian data. Penelitian sebelumnya menunjukkan bahwa split data dapat menghasilkan performa yang baik tetapi memiliki sensitivitas tinggi terhadap skema pembagiannya (Wijayanti et al., 2024). Sebaliknya, teknik k-fold cross-validation dinilai lebih stabil karena seluruh data akan bergantian menjadi data latih serta data uji dalam beberapa iterasi. Beberapa penelitian membuktikan bahwa cross-validation menghasilkan akurasi yang lebih konsisten serta mampu memberikan gambaran performa model yang lebih reliabel (Pratiwi, 2021; Tembusai et al., 2021).

Dataset PIMA Indian Diabetes yang dipergunakan di studi ini terdiri atas beragam indikator medis penting, seperti tekanan darah, kadar glukosa, kadar insulin, indeks massa tubuh, usia, serta jumlah kehamilan. Dataset ini banyak digunakan dalam penelitian diagnosis diabetes karena struktur datanya komprehensif dan relevan secara klinis, sehingga memungkinkan analisis prediksi yang akurat serta dapat dibandingkan dengan penelitian lain (Kaggle, 2021).

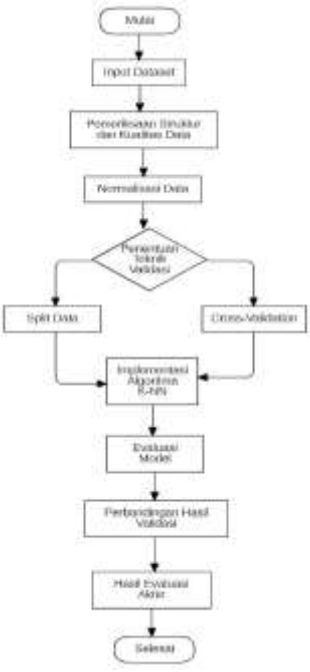
Walaupun penelitian mengenai prediksi diabetes dengan algoritma KNN telah banyak dilakukan, masih terdapat celah penelitian terkait konsistensi performa model berdasarkan variasi nilai k dan perbedaan teknik validasi. Sebagian besar penelitian hanya menggunakan satu metode

validasi, sehingga hasil evaluasi belum menggambarkan stabilitas model secara menyeluruh. Selain itu, penggunaan perangkat RapidMiner Studio dalam konteks penelitian medis belum banyak dijelaskan secara detail dalam literatur terbaru, padahal perangkat ini banyak digunakan dalam kegiatan akademik dan pembelajaran data mining.

Berlandaskan latar belakang tersebut, penelitian ini bertujuan menganalisis performa algoritma K-Nearest Neighbor pada dataset PIMA Indian Diabetes menggunakan 2 teknik validasi, yakni split data serta k-fold cross-validation, dengan beberapa variasi nilai k. Temuan penelitian diharapkan dapat memberikan gambaran menyeluruh mengenai efektivitas dan kestabilan KNN dalam memprediksi Diabetes Mellitus serta menjadi landasan bagi penelitian lanjutan dalam pengembangan model prediksi penyakit.

METODE PENELITIAN

Studi ini disusun mempergunakan pendekatan kuantitatif dengan desain eksperimen komputasional guna mengevaluasi performa algoritma KNN pada dataset PIMA Indian Diabetes. Seluruh proses analisis dilakukan menggunakan RapidMiner Studio agar alur pemodelan dapat divisualisasikan dengan jelas, ditelusuri secara sistematis, dan mudah direplikasi. Fokus utama penelitian adalah membandingkan pengaruh variasi nilai k serta 2 teknik validasi yakni split data dan k-fold cross-validation terhadap performa klasifikasi, sehingga setiap tahapan dirancang secara terstruktur, ringkas, namun tetap komprehensif untuk menjawab tujuan penelitian. Metodologi ini mencakup rangkaian proses dimulai dari persiapan dataset, preprocessing, pemilahan teknik validasi, pemodelan, hingga evaluasi model, sehingga menghasilkan gambaran utuh mengenai efektivitas K-Nearest Neighbor dalam memprediksi Diabetes Mellitus. Untuk memperjelas tahapan penelitian, flowchart dibuat secara sistematis dan menggambarkan seluruh langkah penting penelitian.



Gambar 1. Tahapan Penelitian

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

Dataset Penelitian

Dataset yang digunakan adalah PIMA Indian Diabetes yang terdiri dari 768 baris data pasien perempuan keturunan Indian-Pima. Dataset ini dipilih karena struktur datanya bersifat numerik dan telah banyak digunakan dalam penelitian prediksi diabetes berbasis KNN, sehingga memudahkan pembandingan hasil penelitian. Dalam studi lain, dataset ini dimanfaatkan untuk menguji normalisasi pada algoritma KNN (Sholeh et al., n.d.) dan digunakan untuk mengevaluasi performa KNN sebagai metode klasifikasi penyakit diabetes (Oktaviana et al., 2024). Hal tersebut menunjukkan bahwa dataset ini relevan dan telah divalidasi secara akademik. Dataset terdiri dari delapan atribut prediktor dan satu atribut kelas. Atribut-atribut yang digunakan dijelaskan secara ringkas di tabel 1 sebagai berikut:

Tabel 1. Atribut PIMA Indian Dataset

Atribut	Tipe Data	Deskripsi
Pregnancies	Numerik	Jumlah kehamilan yang pernah di alami
Glucose	Numerik	Kadar glukosa plasma 2 jam setelah tes toleransi glukosa
Blood Pressure	Numerik	Tekanan darah diastolik
Skin Thickness	Numerik	Ketebalan lipatan kulit trisep
Insulin	Numerik	Kadar insulin serum
BMI (Body Mass Index)	Numerik	Indeks massa tubuh berdasarkan berat dan tinggi
Diabetes Pedigree	Numerik	Risiko diabetes berdasarkan riwayat keluarga
Age	Numerik	Usia pasien
Outcome	Nominal (0/1)	Status diabetes (0 = tidak, 1 = diabetes)

Persiapan Data dan Normalisasi

Sebelum data dianalisis menggunakan KNN, seluruh atribut harus berada pada skala yang seragam karena algoritma ini berbasis pada perhitungan jarak. Jika salah satu atribut memiliki rentang nilai yang lebih besar dari atribut lainnya, atribut tersebut akan memiliki pengaruh dominan pada proses klasifikasi. Maka itu, studi ini menggunakan teknik normalisasi Min–Max yang umum diterapkan dalam penelitian KNN untuk dataset medis. Normalisasi Min–Max membantu mempertahankan pola distribusi data sekaligus meratakan skala antar atribut agar tidak terjadi bias jarak (Allorerung et al., 2024). Normalisasi Min–Max dihitung mempergunakan rumus ini:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Pada rumus tersebut, X merupakan nilai asli dari suatu atribut, sementara X′ ialah nilai baru yang telah dinormalisasi. Nilai Xmin dan Xmax ialah nilai minimum dan maksimum atribut yang sama. Proses ini memastikan bahwa setiap atribut berada pada rentang 0 - 1 sehingga seluruh fitur memberikan kontribusi yang seimbang. Karena dataset PIMA Indian Diabetes tidak memiliki data kosong (missing value), proses normalisasi dapat dilakukan langsung tanpa tahap imputasi.

Agar rentang awal setiap atribut dapat terlihat dengan jelas sebelum proses normalisasi dilakukan, penelitian ini juga menyajikan tabel berisi nilai minimum dan maksimum dari masing-masing fitur. Nilai-nilai inilah yang menjadi acuan dalam perhitungan Min–Max.

Tabel 2. Nilai Minimum dan Maksimum Atribut PIMA Indian Dataset

Atribut	Min	Max
Pregnancies	0	17
Glucose	0	199
Blood Pressure	0	122
Skin Thickness	0	99
Insulin	0	846
BMI (Body Mass Index)	0	67,1
Diabetes Pedigree	0,078	2,420
Age	21	81
Outcome	0	1

Penentuan Nilai k pada K-Nearest Neighbor

Nilai k merupakan parameter paling penting dalam algoritma KNN karena menetapkan jumlah tetangga yang dipergunakan pada proses klasifikasi. Variasi nilai k yang diuji pada penelitian ini adalah 3, 5, 7, 9, dan 11. Pemilihan rentang tersebut mempertimbangkan beberapa aspek, terutama ukuran dataset yang berada pada kategori menengah sehingga ideal untuk pengujian nilai k kecil hingga menengah. Nilai k pada rentang 3 hingga 11 sering memberikan performa terbaik pada dataset diabetes karena mampu menangkap pola lokal sekaligus tetap menjaga stabilitas prediksi (Oktaviana et al., 2024; Sholeh et al., n.d.).

Pemilihan nilai ganjil dilakukan agar tidak terjadi hasil voting yang imbang dalam proses klasifikasi. Nilai k yang lebih kecil seperti 3 atau 5 cenderung lebih sensitif terhadap pola lokal, sementara nilai k yang lebih besar seperti 9 dan 11 memberikan efek perataan (smoothing) yang dapat meningkatkan stabilitas prediksi pada data yang memiliki variabilitas tinggi. Dengan menguji lima variasi berbeda, penelitian ini dapat mengamati bagaimana perubahan parameter memengaruhi sensitivitas dan akurasi model.

Teknik Validasi Model

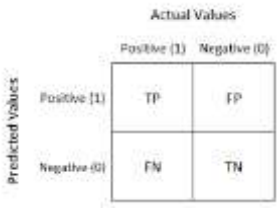
Penelitian ini menggunakan dua teknik validasi untuk memperoleh gambaran menyeluruh mengenai performa model. Teknik pertama adalah split data dengan proporsi 70% data latih serta 30% data uji. Skema ini banyak digunakan dalam penelitian klasifikasi karena mencerminkan kondisi nyata ketika model diterapkan pada data baru. Di studi yang dilaksanakan Hutagalung et al. menggunakan pembagian yang sama dan menemukan bahwa rasio tersebut memberikan keseimbangan terbaik antara jumlah data latih serta data uji. Teknik split data menghasilkan satu kali pengujian sehingga memberikan gambaran awal mengenai performa model pada konfigurasi data tertentu (Sholeh et al., n.d.).

Teknik kedua adalah k-fold cross-validation dengan jumlah lipatan sejumlah 10 fold. Dalam metode ini, dataset dibagi atas 10 bagian dan setiap bagian bergantian menjadi data uji, sementara 9 bagian lainnya dipergunakan menjadi data latih. Teknik ini menghasilkan 10 nilai

akurasi yang kemudian dirata-ratakan guna mendapatkan estimasi performa model lebih stabil. Pada data medis 10-fold cross-validation direkomendasikan karena memberikan hasil yang konsisten dan tidak mudah dipengaruhi oleh variasi pembagian data (Oktaviana et al., 2024).

Metrik Evaluasi Model

Evaluasi model dilaksanakan dengan empat metrik yang penting dalam konteks klasifikasi medis, yaitu akurasi, presisi, recall, serta confusion matrix. Akurasi memberikan gambaran umum mengenai persentase prediksi yang benar. Namun karena penelitian ini berkaitan dengan diagnosis penyakit, akurasi saja tidak cukup. Presisi digunakan untuk mengukur kemampuan model menghindari false positive, yaitu kondisi ketika seseorang diprediksi menderita diabetes padahal tidak. Recall sangat krusial di konteks medis karena mengukur kemampuan model mendeteksi pasien yang benar-benar menderita diabetes. Nilai recall yang rendah berarti model sering melewatkan kasus positif, yang secara klinis sangat berisiko. Sementara itu, confusion matrix memberi gambaran lengkap terkait distribusi prediksi benar dan salah sehingga pola kesalahan model dapat dianalisis lebih rinci.



Gambar 2. Format Confusion Matrix yang Digunakan dalam Evaluasi Model

Ringkasan Metodologi

Seluruh tahapan metode di studi ini disusun guna memberi gambaran yang jelas, runtut, dan mudah ditelusuri mengenai proses analisis performa algoritma K-Nearest Neighbor pada dataset PIMA Indian Diabetes. Normalisasi Min–Max digunakan untuk menyeragamkan skala data sebelum pemodelan, sementara variasi nilai k dipilih untuk melihat bagaimana sensitivitas parameter tersebut memengaruhi hasil klasifikasi. Dua teknik validasi yakni split data serta k-fold cross-validation diterapkan agar performa model dapat dievaluasi secara lebih menyeluruh, baik dari sisi pengujian satu kali maupun pengujian berulang. Penerapan metrik akurasi, presisi, recall, dan confusion matrix memastikan bahwa evaluasi model tetap relevan dengan kebutuhan analisis medis. Seluruh proses dirancang dan dijalankan melalui RapidMiner Studio sehingga alur pemodelan dapat divisualisasikan dengan jelas dan mudah direplikasi dalam penelitian lanjutan.

HASIL DAN PEMBAHASAN

Hasil penelitian ini menyajikan temuan pengujian sekaligus pembahasan kinerja algoritma KNN dalam mengklasifikasikan status diabetes pada dataset PIMA Indian Diabetes. Evaluasi dilakukan mempergunakan 2 teknik validasi, yakni teknik split data serta teknik 10-fold cross validation, dengan variasi nilai parameter k sebagaimana telah dijelaskan pada Bab Metodologi. Analisis difokuskan pada perbandingan performa model berdasarkan metrik evaluasi klasifikasi,

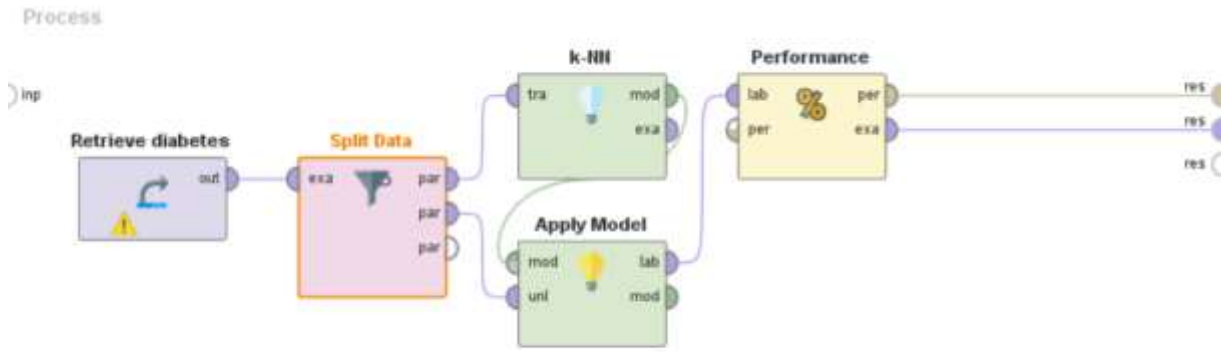
JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824

meliputi akurasi, recall, precision, dan confusion matrix serta pengaruh variasi nilai k terhadap kestabilan dan efektivitas model pada konteks data medis.

Alur Pengujian KNN pada RapidMiner Studio

Sebelum membahas hasil pengujian secara kuantitatif, perlu dijelaskan terlebih dahulu alur pengujian algoritma KNN yang dipergunakan di studi ini. Seluruh proses pemodelan dan evaluasi dilakukan menggunakan perangkat lunak RapidMiner Studio dengan pendekatan workflow visual, sehingga setiap tahapan pengolahan data dapat ditelusuri secara sistematis dan transparan.

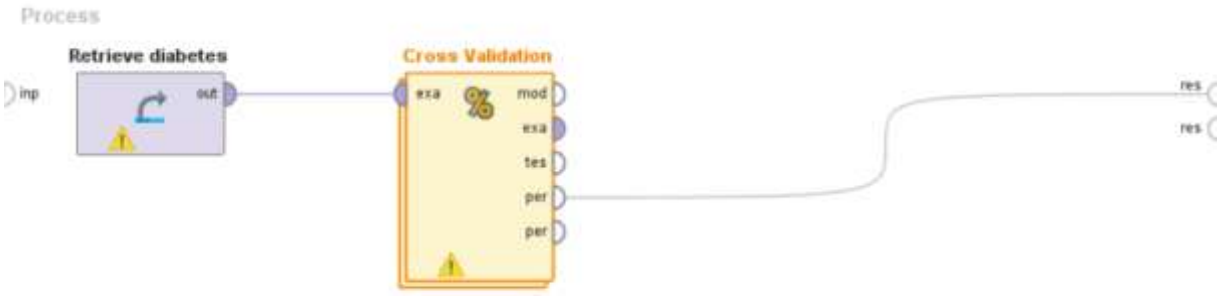
Pada teknik validasi split data, alur pengujian dimulai dengan operator Retrieve untuk memanggil dataset PIMA Indian Diabetes. Dataset kemudian diproses menggunakan operator Normalize dengan metode Min–Max untuk menyamakan skala seluruh atribut numerik. Setelah itu, operator Split Data dipergunakan guna membagi dataset menjadi data latih serta data uji dengan proporsi 70:30. Model KNN dibangun pada data latih menggunakan operator k-NN, kemudian diterapkan pada data uji melalui operator Apply Model. Evaluasi performa klasifikasi dilakukan menggunakan operator Performance, yang menghasilkan metrik akurasi, precision, recall, serta confusion matrix. Alur pengujian ini ditampilkan di Gambar 3.



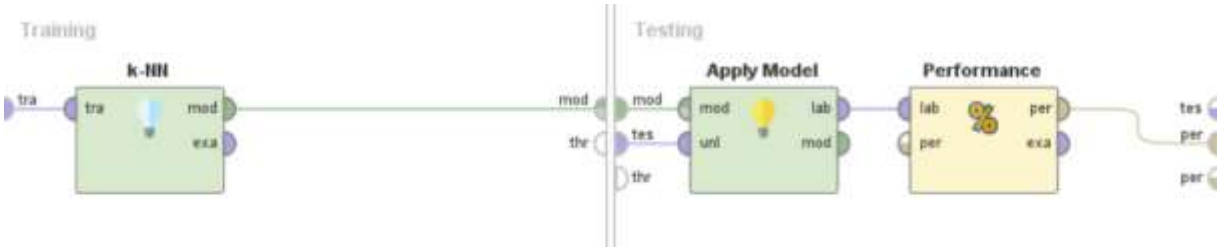
Gambar 3. Alur Proses Pengujian KNN dengan Validasi Split Data pada RapidMiner Studio

Pada teknik 10-fold cross validation, alur pengujian menggunakan operator Cross Validation yang secara internal membagi dataset menjadi sepuluh subset. Pada bagian training, dataset digunakan untuk membangun model KNN, sedangkan pada bagian testing, model diterapkan pada data uji menggunakan Apply Model dan dievaluasi dengan Performance. Proses ini diulang hingga seluruh subset berperan menjadi data uji, sehingga diperoleh hasil evaluasi yang lebih stabil serta representatif. Struktur alur ini ditampilkan di Gambar 4 serta 5.

JIIC: JURNAL INTELEK INSAN CENDIKIA
https://jicnusantara.com/index.php/jiic
Vol : 2 No: 12, Desember 2025
E-ISSN : 3047-7824



Gambar 4. Alur Proses Pengujian KNN dengan Validasi 10-Fold Cross-Validation pada RapidMiner Studio



Gambar 5. Alur Proses Pengujian KNN dengan Validasi 10-Fold Cross-Validation pada RapidMiner Studio

Penggunaan alur operator RapidMiner ini memastikan bahwa seluruh proses pengujian dilakukan secara konsisten, terstruktur, dan mudah direplikasi pada penelitian selanjutnya.

Hasil Pengujian Algoritma KNN Menggunakan Teknik Validasi Split Data

Pengujian awal pada penelitian ini dilaksanakan mempergunakan teknik validasi split data dengan proporsi 70% data latih serta 30% data uji. Teknik ini digunakan untuk memperoleh gambaran awal mengenai performa algoritma KNN pada dataset PIMA Indian Diabetes. Pengujian dilaksanakan dengan lima variasi nilai k, yaitu k = 3, 5, 7, 9, dan 11, guna melihat pengaruh perubahan parameter terhadap kinerja model klasifikasi.

Temuan pengujian mengindikasikan bahwasanya nilai akurasi cenderung meningkat seiring dengan bertambahnya nilai k hingga mencapai titik tertentu. Nilai akurasi terendah didapat pada k = 3 sebesar 66,52%, sedangkan nilai akurasi tertinggi diperoleh pada k = 7 dan k = 11 dengan nilai yang sama, yaitu 72,17%. Pola ini mengindikasikan bahwa penggunaan nilai k yang terlalu kecil mengakibatkan model menjadi lebih sensitif pada variasi lokal serta noise pada data, sehingga performa klasifikasi belum optimal.

Ringkasan hasil evaluasi algoritma KNN menggunakan teknik validasi split data ditampilkan di Tabel 3.

Tabel 3. Hasil Evaluasi KNN Menggunakan Teknik Validasi Split Data

Nilai k	Akurasi %	Recall (1)	Recall (0)	Precision (1)	Precision (0)
3	66.52%	45.00%	78.00%	52.17%	72.67%
5	71.74%	48.75%	84.00%	61.90%	75.45%
7	72.17%	50.00%	84.00%	62.50%	75.90%
9	71.30%	50.00%	82.67%	60.61%	75.61%
11	72.17%	48.75%	84.67%	62.90%	75.60%

Berlandaskan Tabel 3, terlihat bahwasanya performa klasifikasi pada kelas non-diabetes (kelas 0) secara konsisten lebih baik dibandingkan kelas diabetes (kelas 1). Nilai recall kelas 0 berada di atas 78% pada seluruh variasi nilai k, sedangkan recall kelas 1 masih berada di bawah 50%. Kondisi ini memperlihatkan bahwasanya model cenderung lebih akurat dalam mengidentifikasi pasien non-diabetes dibandingkan pasien diabetes, yang merupakan karakteristik umum pada dataset medis dengan distribusi kelas yang tidak seimbang.

Untuk memperjelas pola kesalahan klasifikasi yang terjadi, rincian confusion matrix untuk setiap variasi nilai k pada teknik validasi split data disajikan di Tabel 4.

TabeL 4. Rincian Confusion Matrix KNN Menggunakan Teknik Validasi Split Data

Nilai k	TP	FP	FN	TN
3	36	33	44	117
5	39	24	41	126
7	40	24	40	126
9	40	26	40	124
11	39	23	41	127

Berdasarkan Tabel 4, terlihat bahwa nilai false negative (FN) masih relatif tinggi pada seluruh variasi nilai k. Hal ini menunjukkan bahwa masih terdapat sejumlah kasus diabetes yang salah diklasifikasikan sebagai non-diabetes. Temuan ini menjadi perhatian penting karena kesalahan jenis ini berpotensi menimbulkan dampak serius apabila model digunakan pada konteks klinis.

Hasil Pengujian Algoritma KNN Menggunakan Teknik Validasi 10-Fold Cross Validation

Pengujian selanjutnya dilakukan mempergunakan teknik 10-fold cross validation untuk mendapat estimasi performa model lebih stabil serta representatif. Pada teknik ini, dataset dibagi atas 10 subset yang digunakan secara bergantian sebagai data latih dan data uji, sehingga seluruh data berkontribusi dalam proses evaluasi.

Hasil pengujian mengindikasikan bahwasanya teknik cross validation memberikan performa lebih baik dibanding teknik split data. Nilai akurasi tertinggi diperoleh pada k = 9 sebesar 73,84% dengan simpangan baku ± 3,82% dan micro average sebesar 73,83%.

Simpangan baku (Standard Deviation/SD) menunjukkan tingkat variasi nilai akurasi antar fold. Nilai SD yang relatif kecil mengindikasikan bahwa model memiliki kestabilan performa yang baik pada berbagai subset data. Sementara itu, micro average merepresentasikan performa

keseluruhan model yang dihitung berdasarkan akumulasi hasil confusion matrix dari seluruh fold, sehingga lebih representatif untuk dataset dengan distribusi kelas tidak seimbang.

Ringkasan hasil evaluasi algoritma KNN menggunakan teknik 10-fold cross validation ditampilkan di Tabel 5.

Tabel 5. Hasil Evaluasi KNN Menggunakan Teknik Validasi 10-Fold Cross Validation

Nilai k	Akurasi % +/- SD	Recall (1)	Recall (0)	Precision (1)	Precision (0)
3	69.94% +/- 7.54%	52.99%	79.00%	57.49%	75.82%
5	71.75% +/- 3.13%	51.87%	82.40%	61.23%	76.16%
7	72.80% +/- 4.25%	52.24%	83.80%	63.35%	76.60%
9	73.84% +/- 3.82%	53.36%	84.80%	65.30%	77.23%
11	73.70% +/- 3.67%	54.10%	84.20%	64.73%	77.39%

Rincian confusion matrix untuk teknik cross validation disajikan di Tabel 6.

TabeL 6. Rincian Confusion Matrix KNN Menggunakan Teknik Validasi 10-Fold Cross Validation

Nilai k	TP	FP	FN	TN
3	142	105	126	395
5	139	88	129	412
7	140	81	128	419
9	143	76	125	424
11	145	79	123	421

Terlihat bahwa jumlah true positive meningkat dan false negative relatif menurun dibandingkan teknik split data, yang menunjukkan kemampuan deteksi kelas diabetes yang lebih baik.

Perbandingan Teknik Split Data dan Cross Validation

Perbandingan hasil pengujian menunjukkan bahwa teknik 10-fold cross validation menghasilkan nilai akurasi yang lebih tinggi serta performa yang lebih stabil dibanding teknik split data. Ini disebabkan oleh mekanisme evaluasi cross validation yang tidak bergantung pada satu skema pembagian data tertentu, sehingga mampu mengurangi bias evaluasi. Ringkasan perbandingan performa terbaik dari kedua teknik validasi disajikan pada Tabel 7.

Tabel 7. Perbandingan Performa Terbaik KNN

Teknik Validasi	Nilai k Terbaik	Akurasi	Kestabilan
Split Data (70:30)	7/11	72.17%	Rendah
CV 10-fold	9	73.84%	Tinggi

Analisis Pengaruh Variasi Nilai k terhadap Kinerja Algoritma KNN

Variasi nilai k merupakan parameter utama yang secara langsung memengaruhi kinerja algoritma K-Nearest Neighbor. Nilai k menetapkan jumlah tetangga terdekat yang dipergunakan pada proses pengambilan keputusan kelas, sehingga pemilihan nilai k yang tidak tepat dapat menurunkan performa klasifikasi.

Berdasarkan hasil pengujian pada kedua teknik validasi, terlihat bahwa nilai k yang terlalu kecil, seperti k = 3, cenderung menghasilkan performa lebih rendah. Ini diakibatkan meningkatnya sensitivitas model terhadap noise dan variasi lokal data, yang berakibat pada ketidakstabilan prediksi. Sebaliknya, nilai k yang terlalu besar berpotensi mengurangi kemampuan model dalam menangkap pola lokal yang relevan, karena keputusan klasifikasi menjadi terlalu dipengaruhi oleh mayoritas data di sekitarnya.

Hasil penelitian mengindikasikan bahwasanya nilai k pada rentang 7 hingga 9 memberikan keseimbangan terbaik antara akurasi, precision, dan recall, baik pada teknik split data serta cross validation. Di rentang ini, model mampu mengurangi pengaruh noise sekaligus mempertahankan kemampuan diskriminatif terhadap kelas diabetes. Temuan ini memperlihatkan bahwa pemilihan nilai k yang moderat lebih sesuai untuk dataset medis yang memiliki karakteristik distribusi kelas tidak seimbang dan kompleksitas data yang tinggi.

Implikasi Hasil Penelitian

Berdasarkan keseluruhan hasil pengujian dan analisis yang dilaksanakan, algoritma KNN memperlihatkan performa yang cukup baik dalam mengklasifikasikan status diabetes pada dataset PIMA Indian Diabetes. Penerapan teknik validasi 10-fold cross validation terbukti menghasilkan evaluasi performa yang lebih stabil dan representatif dibanding teknik split data.

Meskipun demikian, nilai recall pada kelas diabetes yang belum optimal menunjukkan bahwa model masih berpotensi menghasilkan kesalahan false negative, yaitu kondisi ketika pasien diabetes diklasifikasikan sebagai non-diabetes. Dalam konteks medis, kesalahan ini mempunyai implikasi yang signifikan sebab dapat berdampak pada keterlambatan penanganan pasien.

Oleh karena itu, model KNN yang dihasilkan dalam penelitian ini lebih tepat digunakan menjadi sistem pendukung keputusan (decision support system) guna membantu proses skrining awal, dan bukan sebagai alat diagnosis utama. Temuan penelitian ini juga membuka peluang untuk pengembangan lanjutan, seperti penerapan teknik penyeimbangan data atau kombinasi algoritma lain guna meningkatkan kemampuan deteksi kelas positif.

KESIMPULAN

Berlandaskan hasil pengujian dan pembahasan yang telah dilaksanakan, disimpulkan bahwasanya algoritma K-Nearest Neighbor (KNN) mampu memberi performa klasifikasi yang cukup baik dalam memprediksi status Diabetes Mellitus pada dataset PIMA Indian Diabetes. Penerapan normalisasi Min–Max terbukti penting dalam menjaga keseimbangan kontribusi setiap atribut numerik sehingga proses perhitungan jarak pada KNN dapat berjalan secara optimal.

Hasil penelitian menunjukkan bahwa teknik validasi 10-fold cross validation menghasilkan performa yang lebih stabil dan representatif dibanding teknik split data. Ini ditunjukkan oleh nilai akurasi yang lebih tinggi serta simpangan baku (standard deviation) yang relatif kecil, yang mengindikasikan konsistensi performa model pada berbagai subset data. Nilai micro average yang mendekati nilai akurasi rata-rata juga menegaskan bahwa evaluasi model tidak dipengaruhi secara signifikan oleh ketidakseimbangan kelas pada dataset.

Variasi nilai k berpengaruh langsung terhadap kinerja algoritma KNN. Nilai k yang terlalu kecil cenderung meningkatkan sensitivitas terhadap noise, sedangkan nilai k yang terlalu besar berpotensi mengurangi kemampuan model dalam menangkap pola lokal data. Berlandaskan hasil pengujian, nilai k pada rentang 7 hingga 9 memberikan keseimbangan terbaik antara akurasi, precision, dan recall, terutama pada teknik 10-fold cross validation.

Meskipun demikian, nilai recall pada kelas diabetes yang masih relatif lebih rendah dibandingkan kelas non-diabetes menunjukkan adanya potensi kesalahan false negative. Oleh karena itu, model KNN yang dihasilkan dalam penelitian ini lebih tepat digunakan menjadi sistem pendukung keputusan (decision support system) untuk skrining awal Diabetes Mellitus, bukan sebagai alat diagnosis utama. Penelitian selanjutnya disarankan untuk mengombinasikan KNN dengan teknik penyeimbangan data atau algoritma klasifikasi lain guna meningkatkan kemampuan deteksi kasus positif diabetes.

DAFTAR PUSTAKA

Allorerung, P. P., Erna, A., & Bagussahrir, M. (2024). *Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit*. *9*(3), 178–191. https://doi.org/10.14421/jiska.2024.9.3.178-191

Kaggle. (2021). *Pima Indians Diabetes Database*. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Noori, N. A., & Yassin, A. A. (2021). *A Comparative Analysis for Diabetic Prediction Based on Machine Learning Techniques*. *1*(1), 180–190.

Oktaviana, A., Wijaya, D. P., Pramuntadi, A., & Heksaputra, D. (2024). *Prediction of Type 2 Diabetes Mellitus Using The K-Nearest Neighbor (K-NN) Algorithm Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)*. *4*(July), 812–818. https://doi.org/10.57152/malcom.v4i3.1268

Prasetyo, A. B., & Laksana, T. G. (2022). *Optimasi Algoritma K-Nearest Neighbors dengan Teknik Cross Validation Dengan Streamlit (Studi Data : Penyakit Diabetes)*. *6*(2), 194–204. https://doi.org/10.30871/jaic.v6i2.4182

Pratiwi, I. (2021). *Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular*. *2*(1), 21–28. https://doi.org/10.33096/ijodas.v2i1.25

Sholeh, M., Andayanti, D., & Rachmawati, Y. (n.d.). *DATA MINING MODEL KLASIFIKASI MENGGUNAKAN K-NEAREST NEIGHBOR WITH NORMALIZATION FOR DIABETES PREDICTION*. 77–87. https://doi.org/10.36342/teika.v12i02.2911

Tembusai, Z. R., Mawengkang, H., Zarlis, M., Info, A., & Process, A. H. (2021). *K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification*. *2*(1), 1–8. https://doi.org/10.25008/ijadis.v2i1.1204

Wijayanti, D., Hermawan, A., Avianto, D., Informasi, M. T., Yogyakarta, U. T., Sakit, R., & Daerah, U. (2024). *PENERAPAN ALGORITMA K-NEAREST NEIGHBOR UNTUK DETEKSI DINI STATUS GIZI PASIEN DEWASA*. *10*(2). https://doi.org/10.31961/positif.v10i2.2255

World Health Organization. (2021). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes%0A