

Jenis artikel	Artikel Penelitian	ID artikel	TEK-2026-0001
Tautan permanen	https://doi.org/10.xxxx/teknovia.xxxx	Riwayat artikel	Masuk: 10-07-2026 Revisi: 31-07-2026
Keputusan & terbit	Diterima: 31-07-2026 Terbit: 31-01-2026	Lisensi	Creative Commons Attribution 4.0 International

Pengelompokan Pasien Diabetes dengan Metode K-NN pada Data yang Dibersihkan, Diseimbangkan, dan Dinormalisasi

Classification of diabetes patients using K-NN on cleaned, balanced, and normalized data

Dimas Maulana^{1*}, Muhammad Irsan Fadillah², Diky Jonathan³, Andrea Octavian⁴, Jordy Lasmana Putra⁵

^{1,2,3,4,5} Program Studi Informatika, Universitas Bina Sarana Informatika, Jakarta Pusat 10450, Indonesia

ORCID: [belum dicantumkan untuk seluruh penulis]

*Penulis korespondensi: Dimas Maulana (15220048@bsi.ac.id)

Abstract

Diabetes is a chronic metabolic disorder characterized by elevated blood glucose levels caused by insufficient insulin production and insulin resistance. In the medical field, decision support systems are needed to diagnose or predict diseases based on various factors that can be analyzed using data mining techniques. This study aims to classify diabetes patients using the K-Nearest Neighbor (K-NN) algorithm, with the objective of achieving higher classification accuracy by applying data preprocessing techniques, including data cleaning, data balancing, and data normalization, to minimize diagnostic errors in diabetes patients. The best results obtained in this study achieved an accuracy of 94.40%, precision of 90.36%, recall of 99.60%, and an F1-score of 94.71% with K = 3 on the preprocessed dataset. In contrast, the dataset without preprocessing produced slightly higher accuracy but was considered less reliable for evaluation because it contained numerous missing values and exhibited an imbalanced class distribution.

Keywords: *Data Mining; K-Nearest Neighbor (K-NN); Diabetes.*

Abstrak

Diabetes merupakan kelainan metabolisme kronis yang ditandai oleh kadar gula darah yang tinggi akibat kekurangan insulin dan resistensi insulin. Dalam dunia medis, dibutuhkan sistem atau alat bantu yang mampu melakukan diagnosis atau prediksi penyakit berdasarkan sejumlah faktor yang dapat dianalisis menggunakan teknik data mining. Penelitian ini bertujuan untuk mengklasifikasikan pasien diabetes dengan metode K-Nearest Neighbor (K-NN) dengan harapan memperoleh akurasi yang lebih tinggi dengan menggunakan pembersihan data, balancing dan normalisasi data untuk meminimalkan kesalahan diagnosis pada pasien diabetes. Hasil terbaik yang diperoleh dalam penelitian ini adalah akurasi 94,40%, presisi 90,36%, recall 99,60%, dan f1-score = 94,71% dengan nilai K = 3 pada data yang telah dipreprocessing, sedangkan data tanpa preprocessing menghasilkan akurasi lebih tinggi namun kurang valid secara evaluasi dikarenakan data masih banyak terdapat nilai kosong dan distribusi kelas tidak seimbang.

Kata kunci: *Data Mining; K-Nearest Neighbor (K-NN); Diabetes*

1. Pendahuluan

Kadar gula darah yang tinggi yang disebabkan oleh kekurangan insulin dan resistensi insulin adalah gejala metabolisme kronis yang dikenal sebagai diabetes melitus[1]. World Health Organization mengatakan bahwa pada tahun 2022, 14% orang berusia 18 tahun ke atas akan menderita diabetes, meningkat dari 7% pada tahun 1990. Selain itu, lebih dari 59% orang dewasa berusia 30 tahun ke atas

yang menderita diabetes tidak mendapatkan pengobatan yang cukup dan memadai. Negara-negara berpenghasilan rendah memiliki tingkat pengobatan diabetes terendah. Penyakit diabetes dapat menyebabkan komplikasi seperti stroke, gagal ginjal, penyakit jantung, nefropati, kebutaan, dan bahkan dapat menyebabkan penderita mengalami luka parah yang memerlukan amputasi. Selain komplikasi, diabetes juga dapat menimbulkan dampak sosial dan ekonomi yang signifikan, karena penyakit ini menyebabkan kerugian langsung maupun tidak langsung bagi penderitanya [2].

Saat di era modern ini, data dan informasi mengalir cukup cepat. Kemajuan teknologi memainkan peranan penting di berbagai sektor, salah satunya dalam bidang kesehatan. Dalam dunia medis, dibutuhkan sistem atau alat bantu yang mampu menganalisis dan klasifikasi penyakit berdasarkan sejumlah faktor menggunakan teknik *data mining* [3]. Penelitian ini bertujuan untuk mengklasifikasikan pasien diabetes dengan metode K-NN dengan harapan memperoleh akurasi yang lebih tinggi dengan menggunakan pembersihan data dan normalisasi data untuk meminimalkan kesalahan diagnosis pada pasien diabetes.

2. Tinjauan Pustaka

2.1 Diabetes

Diabetes merupakan kumpulan dari gangguan metabolik yang dicirikan dengan hiperglikemia yang disertai metabolisme karbohidrat, lemak dan protein yang abnormal. Pengetahuan yang harus dimiliki oleh pasien diabetes meliputi arti penyakit diabetes, penyebab diabetes mellitus, gejala yang sering menyertai dan pentingnya melakukan pengobatan yang teratur dan terus-menerus dalam jangka panjang serta mengetahui bahaya yang ditimbulkan jika tanpa adanya pengobatan[4]

2.2 Data mining

Data Mining adalah proses mendapatkan informasi yang berguna dari database besar, yang harus diekstraksi untuk menjadi informasi baru dan membantu dalam keputusan. *Data Mining* adalah kombinasi dari beberapa disiplin ilmu yang menggabungkan metode, pola, statistik, database, dan visualisasi untuk pembelajaran mesin, menyusun informasi dari database besar ke database besar [5]-[6].

2.3 Klasifikasi Pada Data mining

Klasifikasi adalah suatu proses pembelajaran yang memiliki fungsi untuk menentukan setiap himpunan atribut dari sebuah objek, atau dapat dikatakan juga klasifikasi merupakan pengelompokan[7].

2.4 SMOTE

Algoritma pra proses *Synthetic Minority Oversampling Technique* (SMOTE) dianggap sebagai standar “de facto” dalam kerangka pembelajaran dari data yang tidak seimbang. Hal ini disebabkan oleh kesederhanaannya dalam desain prosedur, serta kekokohannya ketika diterapkan pada berbagai jenis masalah [8].

2.5 Cross-Validation

Data Mining adalah proses mendapatkan informasi yang berguna dari database besar, yang harus diekstraksi untuk menjadi informasi baru dan membantu dalam keputusan. *Data Mining* adalah kombinasi dari beberapa disiplin ilmu yang menggabungkan metode, pola, statistik, database, dan visualisasi untuk pembelajaran mesin, menyusun informasi dari database besar ke database besar [5]-[6].

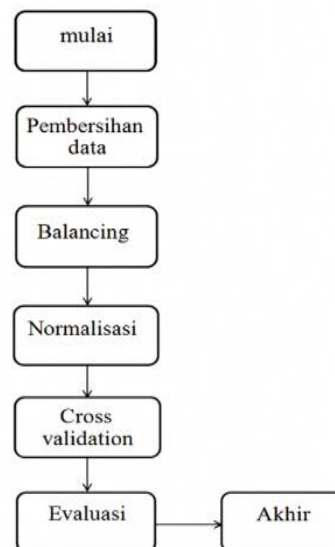
2.6 Resampling

Teknik resampling adalah teknik preprocessing untuk menyeimbangkan distribusi data untuk mengurangi efek dari kelas yang tidak seimbang, teknik resampling dilakukan dengan teknik menyeimbangkan data berdasarkan data asli yang diproses oleh algoritma sampling. Salah satu metode sampling adalah oversampling, yaitu meningkatkan data dari kelas minoritas agar seimbang dengan kelas mayoritas[10].

2.7 Algoritma K-NN

Algoritma K-Nearest Neighbor (KNN) merupakan salah satu metode klasifikasi dalam *data mining*. KNN mengklasifikasikan sekumpulan data berdasarkan data pembelajaran yang telah diberi *label*. KNN termasuk ke dalam *supervised learning* yang digunakan untuk mengklasifikasikan objek baru berdasarkan kedekatannya dengan objek lain. Hasil query instance yang baru akan diklasifikasikan berdasarkan kategori yang memiliki jumlah terbanyak atau mayoritas dalam KNN[7].

3. Metode Penelitian



Gambar 1. Tahap Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan algoritma *data mining* *K-Nearest Neighbor* (K-NN). Proses penelitian terdiri dari beberapa tahapan, yaitu pengumpulan data, pembersihan data, normalisasi data, penyeimbang kelas (balancing), pembentukan model, *Cross-validation*, evaluasi model dengan menggunakan *RapidMiner* sebagai alat bantu. Algoritma K-NN dipilih karena mampu memberikan performa klasifikasi yang baik pada data numerik.

3.1 Pengumpulan Data

Penelitian ini menggunakan *dataset* sekunder yang berjumlah 768 data pasien yang telah terdiagnosa dengan diabetes dan pasien yang tidak terdiagnosa dengan diabetes. Data ini memiliki beberapa atribut yang relevan dalam analisis penyakit diabetes, seperti umur, jumlah kehamilan, indeks massa tubuh (BMI), tekanan darah, Kadar gula, kadar insulin, riwayat keluarga penyakit diabetes dan ketebalan lipatan kulit.

3.2 Pembersihan Data

Proses pembersihan data dilakukan dengan menggunakan metode *Mean imputation* pada fitur-fitur yang terkait, yaitu mengganti nilai kosong menjadi nilai rata-rata pada fitur yang terkait. Tahap ini penting dilakukan agar mencegah model mengalami kesalahan perhitungan dikarenakan data yang bias/tidak valid dikarenakan nilai yang kosong pada *dataset*.

3.3 Normalisasi Data

Pada tahap ini dilakukan normalisasi data untuk menyamakan skala antar fitur agar seragam sehingga dapat meningkatkan performa model menggunakan metode z-transform.

3.4 Balancing

Dataset pada awalnya memiliki ketidakseimbangan kelas yang dapat dilihat pada *label* outcome 34,90% untuk kelas yes dan 65,10% untuk kelas no. Data yang tidak seimbangan ini dapat mempengaruhi performa model. Untuk mengatasi hal tersebut penelitian ini menggunakan teknik Synthetic Minority Over-sampling Technique (SMOTE), dan hanya diterapkan pada subset data *training* didalam *Cross-validation* agar tidak terjadi kebocoran data yang menyebabkan model menjadi bias.

3.5 Pembentukan Model K-NN

Setelah melakukan balancing, langkah berikutnya adalah pembentukan model menggunakan algoritma *K-Nearest Neighbor* (K-NN). Penelitian dilakukan dengan beberapa percobaan *hyperparameter*, yaitu nilai $K = 3, 5, 7$, dan 9 . Percobaan dilakukan dengan 2 kondisi, yang pertama pada data yang belum dibersihkan, balancing, dan dinormalisasi, dan yang kedua pada data yang sudah dibersihkan, balancing dan normalisasi untuk membandingkan dampak yang dihasilkan kepada model.

3.6 Cross-validation

Pada tahap ini dilakukan *Cross-validation* sebanyak 10 fold yang didalamnya terdapat normalisasi, balancing dan model pada subset *training* dan *performance validation* pada subset *testing*.

3.7 Evaluasi

Evaluasi terhadap performa model dilakukan dengan menggunakan matrix evaluasi seperti akurasi, presisi, recall dan f1-score untuk mengukur kualitas klasifikasi yang dihasilkan.

4. Hasil dan Pembahasan

Penelitian ini menggunakan algoritma K-NN yang bertujuan untuk menghasilkan klasifikasi pada pasien diabetes dengan akurasi yang tinggi.

4.1 Pengumpulan Data

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	yes
2	1	85	66	29	0	26.6	0.351	31	no
3	8	183	64	0	0	23.3	0.672	32	yes
4	1	89	66	23	94	28.1	0.167	21	no
5	0	137	40	35	168	43.1	2.288	33	yes
6	5	116	74	0	0	25.6	0.201	30	no
7	3	78	50	32	88	31	0.248	26	yes
8	10	115	0	0	0	35.3	0.134	29	no
9	2	197	70	45	543	30.5	0.158	53	yes
10	8	125	96	0	0	0	0.232	54	yes
11	4	110	92	0	0	37.6	0.191	30	no
12	10	168	74	0	0	38	0.537	34	yes
13	10	139	80	0	0	27.1	1.441	57	no
14	1	189	60	23	846	30.1	0.398	59	yes
15	5	166	72	19	175	25.8	0.587	51	yes
16	7	100	0	0	0	30	0.484	32	yes
17	0	118	84	47	230	45.8	0.551	31	yes
18	7	107	74	0	0	29.6	0.254	31	yes
19	1	103	30	38	83	43.3	0.183	33	no
20	1	115	70	30	96	34.6	0.529	32	yes
21	3	126	88	41	235	39.3	0.704	27	no
22	8	99	84	0	0	35.4	0.388	50	no
23	7	196	90	0	0	39.8	0.451	41	yes
24	9	119	80	35	0	29	0.263	29	yes
25									

Gambar 2. Dataset Diabetes

Pada gambar 2 menunjukan contoh data yang digunakan pada penelitian ini.

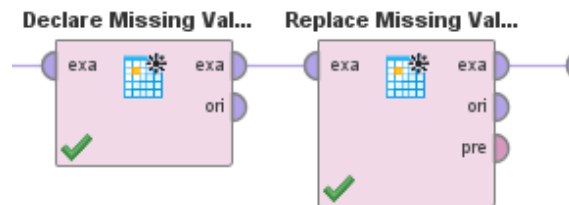
Tabel 1. Atribut Dataset

No	Atribut data
1	Pregnancies
2	Glucose
3	BloodPressure
4	SkinThickness
5	Insulin
6	BMI
7	DiabetesPedigreeFunction
8	Age
9	Outcome

Dari table 1 dapat dilihat data yang diperoleh diketahui pada dataset memiliki 8 fitur bertipe numerical dan 1 *label* dengan tipe nominal.

4.2 Pembersihan Data

Langkah selanjutnya adalah pembersihan data, penelitian ini menggunakan rapid maner dengan operator *Declare Missing Value* untuk menentukan nilai yang kosong dan *Replace missing value* untuk mengganti nilai yang kosong dengan nilai rata-rata.



Gambar 3. Operator pembersihan data

Pada gambar 3 menunjukkan penggunaan dua operator utama dalam proses pembersihan data pada *RapidMiner*. *Declare missing value* digunakan untuk mendeteksi nilai kosong dengan parameter tertentu, misalnya nilai 0 pada fitur-fitur tertentu. *Replace missing value* digunakan untuk mengubah dari data kosong menjadi data tertentu sesuai dengan parameter, metode yang digunakan untuk mengganti nilai yang kosong adalah *Mean imputation* atau mengganti nilai kosong menjadi rata-rata yang akan mempengaruhi performa model.

Row No.	Outcome	Pregnancies	Glucose	BloodPress...	SkinThicke...	Insulin	BMI	DiabetesPe...	Age
1	yes	6	148	72	35	0	33.600	0.627	50
2	no	1	85	66	29	0	26.600	0.351	31
3	yes	8	183	64	0	0	23.300	0.672	32
4	no	1	89	66	23	94	28.100	0.167	21
5	yes	0	137	40	35	168	43.100	2.288	33
6	no	5	116	74	0	0	25.600	0.201	30
7	yes	3	78	50	32	88	31	0.248	26
8	no	10	115	0	0	0	35.300	0.134	29
9	yes	2	197	70	45	543	30.500	0.158	53
10	yes	8	125	96	0	0	0	0.232	54
11	no	4	110	92	0	0	37.600	0.191	30
12	yes	10	168	74	0	0	38	0.537	34

Gambar 4. Data sebelum dibersihkan

Pada gambar 4 terlihat bahwa masih terdapat data kosong yang dapat mempengaruhi performa model.

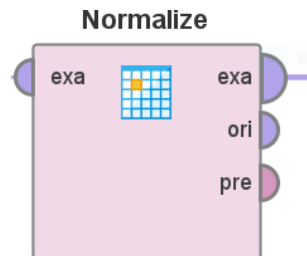
Row No.	Outcome	Pregnancies	Glucose	BloodPress...	SkinThicke...	Insulin	BMI	DiabetesPe...	Age
1	yes	6	148	72	35	156	33.600	0.627	50
2	no	1	85	66	29	156	26.600	0.351	31
3	yes	8	183	64	29	156	23.300	0.672	32
4	no	1	89	66	23	94	28.100	0.167	21
5	yes	4	137	40	35	168	43.100	2.288	33
6	no	5	116	74	29	156	25.600	0.201	30
7	yes	3	78	50	32	88	31	0.248	26
8	no	10	115	72	29	156	35.300	0.134	29
9	yes	2	197	70	45	543	30.500	0.158	53
10	yes	8	125	96	29	156	32.457	0.232	54
11	no	4	110	92	29	156	37.600	0.191	30
12	yes	10	168	74	29	156	38	0.537	34

Gambar 5. Data sesudah dibersihkan

Pada gambar 5 setelah dilakukan pembersihan data data kosong dirubah menjadi rata-rata pada fitur-fitur yang terkait.

4.3 Normalisasi Data

Tahap selanjutnya adalah normalisasi data , normalisasi dilakukan agar setiap fitur memiliki skala nilai yang seragam, sehingga algoritma K-Nearest Neighbor (K-NN) dapat menghitung jarak antar data dengan lebih akurat dan tidak terpengaruh oleh perbedaan skala antar atribut.



Gambar 6. Operator normalisasi data

Pada gambar 6 adalah operator normalisasi data pada *RapidMiner*, metode yang digunakan dalam penelitian ini adalah *z-transformation*.

Row No.	Outcome	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin	BMI	DiabetesPe...	Age
1	yes	6	148	72	35	156	33.600	0.627	50
2	no	1	85	66	29	156	26.600	0.351	31
3	yes	8	183	64	29	156	23.300	0.672	32
4	no	1	89	66	23	94	28.100	0.167	21
5	yes	4	137	40	35	168	43.100	2.288	33
6	no	5	116	74	29	156	25.600	0.201	30
7	yes	3	78	50	32	88	31	0.248	26
8	no	10	115	72	29	156	35.300	0.134	29
9	yes	2	197	70	45	543	30.500	0.158	53
10	yes	8	125	96	29	156	32.457	0.232	54
11	no	4	110	92	29	156	37.600	0.191	30
12	yes	10	168	74	29	156	38	0.537	34

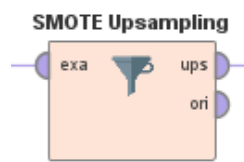
Gambar 7. Data sebelum normalisasi

Row No.	Outcome	Pregnancies	Glucose	BloodPress...	SkinThickne...	Insulin	BMI	DiabetesPe...	Age
1	yes	0.415	0.674	-0.081	0.645	-0.052	0.104	0.463	1.328
2	no	-1.213	-1.335	-0.587	-0.065	-0.052	-0.958	-0.391	-0.285
3	yes	1.066	1.790	-0.756	-0.065	-0.052	-1.459	0.603	-0.200
4	no	-1.213	-1.208	-0.587	-0.775	-0.773	-0.730	-0.961	-1.133
5	yes	-0.236	0.323	-2.780	0.645	0.088	1.546	5.606	-0.115
6	no	0.090	-0.347	0.087	-0.065	-0.052	-1.110	-0.855	-0.369
7	yes	-0.562	-1.559	-1.937	0.290	-0.843	-0.290	-0.710	-0.709
8	no	1.718	-0.378	-0.081	-0.065	-0.052	0.362	-1.063	-0.454
9	yes	-0.887	2.237	-0.250	1.828	4.452	-0.366	-0.989	1.583
10	yes	1.066	-0.060	1.943	-0.065	-0.052	-0.069	-0.759	1.668
11	no	-0.236	-0.538	1.605	-0.065	-0.052	0.712	-0.886	-0.369
12	yes	1.718	1.312	0.087	-0.065	-0.052	0.772	0.185	-0.030

Gambar 8. Data setelah normalisasi

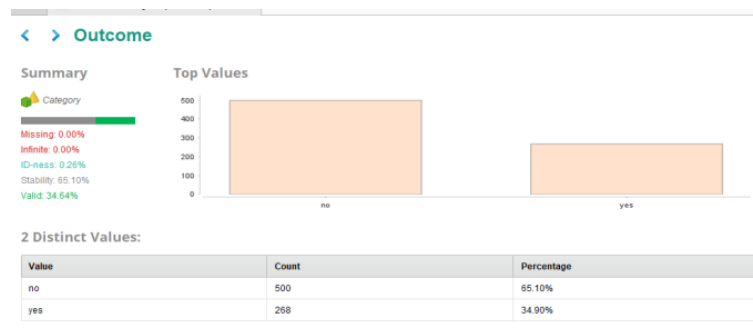
4.4 Penyeimbang Kelas(balancing)

Tahap berikutnya adalah balancing kelas menggunakan metode *Synthetic Minority Oversampling Technique (SMOTE)*. Pada penelitian ini digunakan operator SMOTE Upsampling pada *RapidMiner* untuk mengatasi ketidak seimbangan mayoritas dan minoritas kelas dengan parameter, *number of neighbor* = 5, *auto detect minority class*, *change rate* = 0.5.



Gambar 9. Operator smote upsampling

Pada gambar 9 adalah operator SMOTE Upsampling pada *RapidMiner* yang berfungsi menambahkan data kedalam kelas minoritas. Operator ini hanya digunakan pada subprocess *training* pada *Cross-validation* agar tidak terjadi kebocoran data dan model menjadi bias.



Gambar 10. Data sebelum balancing

Dapat dilihat dari gambar 10 bahwa penyebaran data pada kelas outcome belum seimbang yang akan mempengaruhi performa model

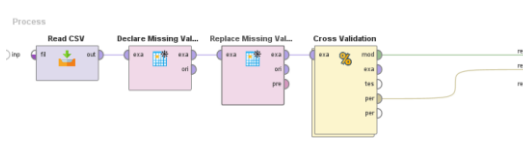


Gambar 11. Data setelah balancing

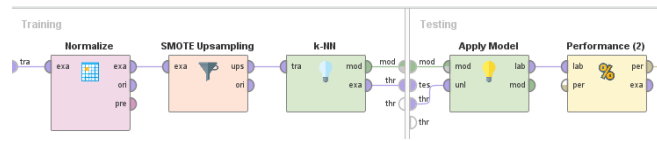
Pada gambar 11 adalah data yang sudah di balancing sehingga penyebaran data pada kelas outcome menjadi seimbang.

4.5 Pembentukan Model K-NN

Pada tahap ini dilakukan pemodelan klasifikasi menggunakan algoritma K-NN, proses pemodelan menggunakan tools *RapidMiner* yang dilakukan dengan beberapa percobaan *hyperparameter*, yaitu K=3, K=5, K=7, K=9. Percobaan dilakukan dengan 2 kondisi, yang pertama pada data yang belum dibersihkan, balancing, dan dinormalisasi, dan yang kedua pada data yang telah dibersihkan, balancing dan normalisasi sehingga dapat dibandingkan hasil klasifikasi terbaik yang didapatkan oleh model.



Gambar 12. Model k-nn

Gambar 13. Model sub *cross-validation*

Pada tahap *Cross-validation* digunakan sampling type automatic dan local random seed = 2001 sebagai standarisasi penelitian.

4.6 Evaluasi

Evaluasi model dilakukan dengan menggunakan matrix evaluasi yaitu, akurasi, presisi, recall dan f1-score, evaluasi dibagi menjadi 2 bagian, yaitu evaluasi pada data yang telah di *preprocessing*, balancing dan normalisasi dan yang ke dua pada data tanpa *preprocessing*.

Table 2. Hasil evaluasi pada data yang dipraprocessing, balancing dan normalisasi

K	Akurasi %	Presisi %	Recall %	F1%
3	94,40	90,36	99,60	94,71
5	86,21	81,91	92,96	87,08
7	84,19	80,30	90,62	85,15
9	82,40	78,70	88,87	83,47

Dari tabel 2 ditunjukan bahwa nilai K = 3 menghasilkan performa yang tinggi dalam segi akurasi, presisi, recall dan f1-score K = 9 menghasilkan performa yang rendah pada ujicoba yang dilakukan. Dari hasil tersebut dapat disimpulkan bahwa performa model lebih baik pada nilai K yang kecil pada data yang telah dipreprocessing, nomalisasi, dan balancing.

Table 3. Hasil evaluasi pada data tanpa *preprocessing*

K	Akurasi %	Presisi %	Recall %	F1 %
3	95,78	89,47	99,63	94,27
5	79,98	74,68	64,51	69,22
7	78,62	73,29	90,95	66,54
9	78,59	73,05	61,42	66,62

Dari tabel 3 ditunjukan bahwa nilai K = 3 menghasilkan performa yang tinggi dalam segi akurasi dengan data yang belum di *preprocessing*, dan K = 5,7,9 menunjukan penurunan performa yang signifikan. Namun hasil tersebut dianggap kurang valid karena data masih memiliki banyak nilai kosong dan kelas yang tidak seimbang yang dapat menyebabkan model menjadi bias. Artinya metode *preprocessing* data, normalisasi, dan balancing berperan besar bagi model untuk menghasilkan performa yang lebih optimal.

Penelitian ini menunjukan *preprocessing* data dengan pembersihan data, balancing data, dan normalisasi data memiliki pengaruh yang signifikan untuk optimalisasi model K-NN, yang dirujuk dari penelitian sebelumnya yang dilakukan oleh [11] dengan metode K-NN didapatkan hasil 74,6% untuk semua fitur dan 74,9% untuk empat fitur, jika dibandingkan dengan penelitian ini didapatkan hasil dengan akurasi paling tinggi adalah 94,40% pada data yang sudah di *preprocessing* dan 95,78% pada data tanpa *preprocessing*. Perbedaan hasil ini menunjukan penggunaan *preprocessing* pada data dapat meningkatkan keandalan pada model.

Selain itu, penelitian ini juga menggunakan metode *cross-validation* yang diintegrasikan ke dalam proses *preprocessing* pada data latih untuk menghindari kebocoran data, sehingga hasil evaluasi model dapat dikatakan valid dan tidak bias. Berbeda dengan penelitian yang dilakukan oleh [12], yang

meskipun telah menerapkan *preprocessing* secara lengkap, namun tidak menggunakan tahapan *cross-validation*, sehingga akurasi yang dihasilkan hanya sebesar 91%. Sementara itu, penelitian ini memperoleh hasil yang lebih tinggi dengan menerapkan validasi silang secara sistematis.

5. Kesimpulan

Dari hasil penelitian yang telah dilakukan diketahui bahwa pembersihan data, balancing dan normalisasai berdampak baik pada model dengan algoritma K-NN yang membuat hasil klasifikasi lebih optimal, hasil paling baik dihasilkan oleh $K = 3$ pada data yang telah di *preprocessing*. Walaupun hasil $K = 3$ pada data tanpa *preprocessing* menghasilkan matriks evaluasi yang lebih tinggi namun hasil tersebut dianggap kurang valid dikarenakan banyak data yang masih kosong dan distribusi data pada kelas belum seimbang yang mengakibatkan model menjadi bias. Jika dibandingkan dengan penelitian sebelumnya, hal ini menunjukkan adanya kenaikan yang signifikan jika menggunakan tahap *preprocessing* data pada algoritma K-NN yang ditunjukkan pada penelitian ini.

6. Referensi

- [1] C. Savan, D. Viroja, and A. Kyada, "An updated review on diabetes mellitus: Exploring its etiology, pathophysiology, complications and treatment approach," *IP Int. J. Compr. Adv. Pharmacol.*, vol. 9, no. 1, pp. 31–36, Mar. 2024, doi: 10.18231/ijcaap.2024.005.
- [2] F. Hans Tandra, FINASIM, "Diabetes Masih Bisa dikalahkan," vol. 3, no. 3, pp. 101–106, 2019.
- [3] G. Abdurrahman, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier," *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.)*, vol. 7, no. 1, pp. 59–66, 2022, doi: 10.32528/justindo.v7i1.4949.
- [4] H. R. Pramestutie, M. P. Sari, and R. K. Illahi, "PHARMACEUTICAL JOURNAL OF INDONESIA Tingkat Pengetahuan Pasien Diabetes mellitus tentang Penggunaan Obat di Puskesmas Kota Malang," *Fak. Kedokt.*, vol. 2, no. 1, pp. 7–11, 2016.
- [5] A. Kukuh Wahyudi, N. Azizah, and H. Saputro, "Data Mining Klasifikasi Kepribadian Siswa Smp Negeri 5 Jepara Menggunakan Metode Decision Tree Algoritma C4.5," *J. Inf. Syst. Comput.*, vol. 2, no. 2, pp. 8–13, 2022, doi: 10.34001/jister.v2i2.392.
- [6] S. N. Arofah and F. Marisa, "Penerapan Data Mining untuk Mengetahui Minat Siswa pada Pelajaran Matematika menggunakan Metode K-Means Clustering," *JOINTECS (Journal Inf. Technol. Comput. Sci.)*, vol. 3, no. 2, pp. 693–702, 2018, doi: 10.31328/jointecs.v3i2.787.
- [7] N. M. Putry, "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," *EVOLUSI J. Sains dan Manaj.*, vol. 10, no. 1, 2022, doi: 10.31294/evolusi.v10i1.12514.
- [8] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.
- [9] A. Nur Kirana, B. Nurhakim, S. Eka Permana, W. Prihartono, and G. Dwilestari, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Cuaca Menggunakan Rapidminer," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 2, pp. 1637–1642, 2024, doi: 10.36040/jati.v8i2.8967.
- [10] J. L. Putra and A. Subekti, "Prediksi Kinerja Siswa Pada E-Learning Moodle Platform Menggunakan Algoritma Adaptive Boosting," *J. Inform.*, vol. 10, no. 1, pp. 79–84, 2023, doi: 10.31294/inf.v10i1.15525.
- [11] D. S. F. Azzahrah and A. Alamsyah, "Comparison of Probabilistic Neural Network (PNN) and k-Nearest Neighbor (k-NN) Algorithms for Diabetes Classification," *Recursive J. Informatics*, vol. 1, no. 2, pp. 73–82, 2023, doi: 10.15294/rji.v1i2.66078.
- [12] A. H. Yudianto and E. R. Subhiyakti, "Perbandingan Kinerja Algoritma K-Nearest Neighbors dan Decision Tree untuk Klasifikasi Diabetes," vol. 6, no. 4, pp. 2601–2611, 2025, doi: 10.47065/bits.v6i4.6550.