ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

Klasifikasi Resiko Diabetes terhadap Gaya Hidup dengan Algoritma *K-Nearest Neighbor* (k-NN) dan *Naive Bayes*

Thaufik Darma Jonhar[1)], Maulana Fiananta[2)], Dicky Purwanto[3)], Hendri Mahmud Nawawi[4)]

[1), 2),3), 4)] *Informatika, Universitas Nusa Mandiri*
[Jl. Raya Jatiwaringin No.2, RT.8/RW.13, Cipinang Melayu, Kec. Makasar, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta 13620]
Email :12250068@nusamandiri.ac.id[1)], 12250065@nusamandiri.ac.id[2)], 12250082@nusamandiri.ac.id[3)], hendri.hiw@nusamandiri.ac.id[4)]

ABSTRACT

Diabetes mellitus is a metabolic disorder that occurs when the pancreas is unable to produce sufficient insulin or when the body cannot use insulin effectively, resulting in uncontrolled blood glucose levels and potential long-term complications. This study applies a data mining approach to classify diabetes risk levels using two supervised learning algorithms, namely Naive Bayes and k-Nearest Neighbor (k-NN), with the objective of comparing their predictive performance and identifying patterns within the dataset. The research process follows the Knowledge Discovery in Databases (KDD) stages, including data selection, preprocessing, transformation, modeling, and evaluation. All experiments were conducted using RapidMiner with an 80:20 data split scheme through stratified sampling and a fixed random seed to ensure result consistency. Model evaluation was performed using a confusion matrix, accuracy, precision, recall, and F1-score. The results indicate that k-NN achieved an accuracy of 81.72%, higher than Naive Bayes at 75.42%. In terms of precision and recall, Naive Bayes obtained a precision of 86.69%, recall of 93.12%, and F1-score of 89.79%, while k-NN achieved a precision of 90.68%, recall of 80.33%, and F1-score of 85.19%. Although k-NN outperformed in overall accuracy and precision, Naive Bayes demonstrated a more balanced performance through higher recall and F1-score values, particularly in detecting positive cases. These findings emphasize that model selection should consider not only overall accuracy but also class-level performance, especially in imbalanced healthcare datasets.

Keywords : *Diabetes, RapidMiner, K-Nearest neighbor, Naive Bayes, Classification*

ABSTRAK

Diabetes melitus merupakan gangguan metabolik yang terjadi ketika pankreas tidak mampu memproduksi insulin dalam jumlah yang cukup atau ketika tubuh tidak dapat menggunakan insulin secara efektif, sehingga menyebabkan kadar gula darah tidak terkontrol dan berpotensi menimbulkan komplikasi jangka panjang. Penelitian ini menerapkan pendekatan data mining untuk mengklasifikasikan tingkat risiko diabetes menggunakan dua algoritma supervised learning, yaitu Naive Bayes dan k-Nearest Neighbor (k-NN), dengan tujuan membandingkan kinerja prediksi serta mengidentifikasi pola dari dataset. Proses penelitian mengikuti tahapan Knowledge Discovery in Database (KDD), meliputi seleksi data, pra-pemrosesan, transformasi, pemodelan, dan evaluasi. Seluruh eksperimen dilakukan menggunakan RapidMiner dengan skema pembagian data 80:20 melalui stratified split dan random seed tetap untuk menjaga konsistensi hasil. Evaluasi model menggunakan confusion matrix, akurasi, precision, recall, dan F1-score. Hasil menunjukkan bahwa k-NN memperoleh akurasi sebesar 81,72%, lebih tinggi dibandingkan Naive Bayes sebesar 75,42%. Dari sisi precision dan recall, Naive Bayes menghasilkan precision 86,69%, recall 93,12%, dan F1-score 89,79%, sedangkan k-NN memperoleh precision 90,68%, recall 80,33%, dan F1-score 85,19%. Meskipun k-NN unggul dalam akurasi dan precision, Naive Bayes menunjukkan keseimbangan performa yang lebih baik melalui nilai recall dan F1-score yang lebih tinggi, khususnya dalam mendeteksi kelas positif. Temuan ini menegaskan bahwa pemilihan model tidak hanya mempertimbangkan akurasi keseluruhan, tetapi juga performa pada tiap kelas, terutama pada dataset kesehatan yang cenderung tidak seimbang.

Kata Kunci : *Diabetes, RapidMiner, K-nearest Neighbor, Naive Bayes, Klasifikasi*

1. Pendahuluan

Diabetes Mellitus (DM) merupakan penyakit metabolik yang bersifat kronis, dimana penderita mengalami gangguan dalam produksi insulin atau ketidakmampuan tubuh dalam menggunakan insulin secara optimal. Kondisi ini mengakibatkan peningkatan kadar gula darah dalam tubuh yang sering kali tidak disadari pada tahap awal, sehingga gejalanya baru dirasakan setelah terjadi komplikasi pada organ-organ tertentu (Kudus et al., 2020).

Salah satu faktor yang berkontribusi terhadap

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

meningkatnya jumlah penderita diabetes adalah keterlambatan dalam melakukan diagnosis. Kondisi ini menyebabkan penyakit tidak terdeteksi sejak dini sehingga sulit dikendalikan, yang pada akhirnya dapat menimbulkan berbagai komplikasi serius dan berpotensi membahayakan nyawa penderita. Berbagai upaya pencegahan untuk menurunkan angka penderita Diabetes Mellitus (DM) telah dilakukan, namun hingga saat ini hasil yang dicapai masih belum optimal. Oleh karena itu, diperlukan penelitian lanjutan guna mengembangkan sistem pendeteksian diabetes yang lebih efektif sebagai langkah pencegahan dan penanganan penyakit tersebut (Levi Sabili et al., 2024).

Berdasarkan hasil survei, sekitar 30% penderita diabetes tidak menyadari bahwa dirinya telah mengidap penyakit tersebut dan baru mengetahuinya setelah menjalani pemeriksaan laboratorium. Bahkan, sekitar 25% penderita telah mengalami komplikasi mikrovaskular yang berdampak pada kerusakan organ seperti ginjal, indera perasa, dan mata akibat gangguan pada pembuluh darah kecil. Diabetes merupakan masalah kesehatan global yang terjadi di berbagai negara, dan diperkirakan jumlah penderitanya akan mencapai sekitar 439 juta orang pada tahun 2030 (Prasatya et al., 2020).

Seiring dengan meningkatnya jumlah penderita Diabetes Mellitus (DM), diperlukan metode pengolahan data yang mampu melakukan klasifikasi secara efektif, salah satunya melalui pendekatan data mining (Li, Y.; Li, H.; Yao, 2018). Data mining digunakan untuk memprediksi kemungkinan terjadinya penyakit diabetes berdasarkan gejala dan indikator yang relevan. Proses data mining bertujuan untuk menemukan pola atau hubungan antar data yang sebelumnya tidak diketahui, sehingga dapat membantu dalam pengambilan keputusan. Secara umum, data mining memiliki beberapa fungsi utama, antara lain estimation, prediction, classification, clustering, dan association (A'yuniyah et al., 2022).

Dalam mendukung proses diagnosis penyakit diabetes, diperlukan teknik klasifikasi berbasis komputer yang mampu menggali informasi dari dataset penyakit Diabetes Mellitus (DM). Klasifikasi merupakan proses pembentukan model yang dapat membedakan dan mendeskripsikan kelas-kelas data tertentu, sehingga model tersebut dapat digunakan untuk memprediksi kelas dari data atau objek yang belum diketahui sebelumnya (Nurdiana & Algifari, 2020).

Pada penelitian ini digunakan dua algoritma klasifikasi, yaitu *K-Nearest Neighbo*r (k-NN) dan *Naïve Bayes*, untuk membandingkan kinerja dalam mengklasifikasikan risiko penyakit diabetes. Pemilihan kedua algoritma tersebut didasarkan pada beberapa penelitian sebelumnya, diantaranya penelitian yang dilakukan oleh (Ridwan, 2020) menggunakan algoritma *Naïve Bayes* dengan objek penelitian berupa data penderita Diabetes Mellitus, dan memperoleh tingkat akurasi sebesar 90,20% dalam proses klasifikasinya. Penelitian lain oleh (Sugara, B., Adidarma, D., & Budilaksono, 2018) melakukan perbandingan algoritma *Naïve Bayes* dan C4.5, namun objek penelitian yang digunakan bukan penyakit diabetes melainkan deteksi dini gangguan autisme pada anak. Hasil penelitian tersebut menunjukkan bahwa algoritma Naïve Bayes memiliki tingkat akurasi yang lebih tinggi dibandingkan C4.5, yaitu sebesar 73,33%. Sementara itu, penelitian yang dilakukan oleh (Lestari et al., 2021) secara khusus menggunakan dataset penyakit Diabetes Mellitus dengan menerapkan algoritma *K-Nearest Neighbor*, dan menghasilkan tingkat akurasi sebesar 96% pada nilai k = 23.

Meskipun demikian, sebagian besar penelitian sebelumnya masih berfokus pada klasifikasi penyakit diabetes berdasarkan data klinis atau hasil pemeriksaan medis, sementara aspek gaya hidup sebagai faktor risiko utama belum banyak dikaji secara mendalam. Padahal, gaya hidup seperti rendahnya aktivitas fisik, pola makan yang tidak sehat, dan kebiasaan harian lainnya terbukti memiliki hubungan yang signifikan dengan peningkatan risiko terjadinya Diabetes Mellitus (DM) (Ismail et al., 2021) Oleh karena itu, penelitian ini menghadirkan kebaruan dengan melakukan klasifikasi risiko diabetes berdasarkan faktor gaya hidup menggunakan algoritma *K-Nearest Neighbor* (k-NN) dan *Naïve Bayes*, sekaligus membandingkan kinerja kedua algoritma tersebut. Diharapkan hasil penelitian ini dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan yang lebih preventif, khususnya dalam mengidentifikasi risiko diabetes sejak dini berdasarkan gaya hidup individu (Hunafa & Hermawan, 2023).

2. Pembahasan

2.1. Diabetes Melitus dan Konsep Risiko

Diabetes Mellitus (DM) merupakan penyakit kronis yang ditandai oleh peningkatan kadar glukosa darah (hiperglikemia) akibat gangguan produksi insulin, kerja insulin, atau keduanya. Kondisi hiperglikemia yang berlangsung lama dapat memicu komplikasi pada berbagai organ seperti pembuluh darah dan jantung, mata, ginjal, saraf, serta meningkatkan risiko infeksi, sehingga pencegahan dan deteksi dini berbasis faktor risiko menjadi penting dalam konteks kesehatan masyarakat (Kandhasamy & Balamurali, 2015).

Pada penelitian berbasis data kesehatan populasi, "risiko diabetes" umumnya dimaknai sebagai peluang seseorang berada pada kategori tidak diabetes / prediabetes / diabetes berdasarkan indikator perilaku dan kondisi kesehatan. Dataset Diabetes Health Indicators memfasilitasi pendekatan ini dengan label kelas 0 (no diabetes), 1 (prediabetes), 2 (diabetes) sehingga keluaran model dapat ditafsirkan sebagai tingkat risiko/kelas status diabetes yang berkaitan dengan gaya hidup dan indikator kesehatan (Mohamed et al., 2024).

Gaya hidup berperan kuat pada peningkatan risiko DM, misalnya pola makan, aktivitas fisik, status berat badan (BMI), serta kebiasaan merokok/alkohol yang berkontribusi terhadap kondisi metabolik. Karena faktor-faktor tersebut bersifat observasional dan dapat

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

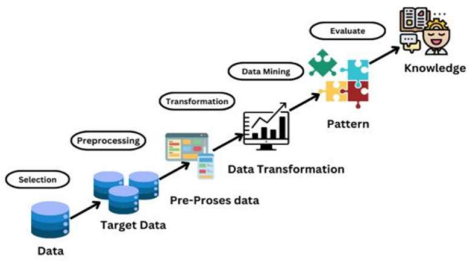
dikumpulkan dalam skala besar, pendekatan klasifikasi berbasis data mining relevan untuk memetakan kelompok risiko sebagai dasar edukasi dan pencegahan (Lestari et al., 2021).

2.2. Data Mining pada Bidang Kesehatan

Data mining dipahami sebagai proses penggalian pengetahuan dari data berukuran besar untuk menemukan pola yang bermanfaat dalam pengambilan keputusan. Dalam konteks kesehatan, data mining sering digunakan untuk membangun model prediksi/klasifikasi sehingga membantu identifikasi dini suatu kondisi dari indikator yang tersedia (Kandhasamy & Balamurali, 2015).

Klasifikasi merupakan salah satu teknik utama dalam data mining yang memetakan data ke dalam kelas tertentu berdasarkan atribut masukan. Pada penelitian risiko diabetes, klasifikasi digunakan untuk memprediksi kelas status diabetes berdasarkan indikator kesehatan dan perilaku, sehingga hasilnya dapat dipakai untuk penapisan (screening) atau prioritas intervensi (Sisodia & Sisodia, 2018).

Kerangka kerja yang sering dipakai adalah KDD (*Knowledge Discovery in Databases*) yang meliputi *data selection*, *preprocessing*, *data cleaning*, *transformation*, proses *data mining*, serta evaluasi/interpretasi hasil. Struktur tahapan ini membantu penelitian berjalan sistematis dari data mentah hingga keluaran model yang siap dianalisis (Gamadarenda et al., 2020).



Sumber: (Binus, 2021)

Gambar 1. Proses *Knowledge Discovery in Database Process* (KDD)

A. Data Original

Merupakan data awal yang diperoleh dari sumber data terpilih, sebelum dilakukan pembersihan maupun transformasi. Pada tahap ini, data masih merepresentasikan kondisi apa adanya dan menjadi dasar untuk tahap pra-pemrosesan berikutnya (Aprianti, B., Aulia, A., & Purnamasari, 2022).

B. Data Selection

Merupakan tahap memilih atribut dan record yang relevan dengan tujuan penelitian. Pada tahap ini dilakukan identifikasi variabel yang dibutuhkan dan penyisihan data yang tidak berkaitan, sehingga proses pemodelan menjadi lebih fokus dan hasil model dapat lebih optimal (Arta, J. K. I., Indarawan, G., & Dantes, 2019).

C. Data Cleaning

Merupakan bagian dari pra-pemrosesan yang berfokus pada pembersihan data dari noise dan ketidakkonsistenan, misalnya nilai kosong, nilai tidak valid, duplikasi, atau kesalahan input. Tujuannya adalah meningkatkan kualitas data agar proses pemodelan berjalan lebih baik dan hasil yang diperoleh lebih dapat dipercaya (Arta, J. K. I., Indarawan, G., & Dantes, 2019).

D. Data Transformation

Transformasi data adalah proses mengubah bentuk data agar sesuai untuk tahap analisis atau pemodelan, misalnya pengkodean atribut, normalisasi, maupun penyusunan kembali format data. Transformasi dilakukan agar data menjadi lebih seragam dan siap digunakan pada langkah berikutnya (Arta, J. K. I., Indarawan, G., & Dantes, 2019).

E. Data Mining

adalah proses yang mengelola data menjadi informasi dengan terdapat beragam metode dan teknik. Data mining adalah proses pengelolaan sebuah data yang besar sehingga dari data-data tersebut dapat memberikan informasi yang akurat dan dapat mempermudah dalam pemecahan masalah dan pengambilan keputusan (Asroni, A., Respati, B. M., & Riyadi, 2018)

F. Evaluation

Tahapan evaluasi bertujuan untuk melakukan evaluasi terhadap model atau pola yang telah terbentuk dari proses sebelumnya yaitu proses data mining sehingga hasil dari pola/model tersebut memiliki hasil akurasi yang dapat dipercaya (Putry et al., 2022).

2.3. *K-Nearest Neighbor* (k-NN)

Merupakan algoritma klasifikasi yang menyimpan data latih dan mengklasifikasikan data baru berdasarkan ukuran kemiripan (umumnya fungsi jarak). Kelas sebuah instance ditentukan melalui voting mayoritas dari k-NN, sehingga perilaku model sangat bergantung pada definisi jarak dan pilihan nilai k**.** Pemilihan nilai k merupakan komponen penting; peningkatan nilai k dapat memperhalus keputusan, tetapi berpotensi mengurangi sensitivitas pada pola lokal. Studi komparatif menunjukkan k-NN dapat memberikan akurasi yang lebih baik pada rentang k antara 3 sampai 10, dan pemilihan k juga dapat dipandu dengan validasi silang (Kandhasamy & Balamurali, 2015)**.**

Jarak Euclidean sering digunakan pada k-NN untuk mengukur kedekatan antar instance. Dalam implementasi praktis, langkah k-NN dapat diringkas sebagai: menentukan nilai k, menghitung jarak data uji ke seluruh data latih, memilih k jarak terkecil, lalu menetapkan

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

kelas berdasarkan mayoritas kelas tetangga tersebut (Lestari et al., 2021).

Kelemahan yang perlu diperhatikan adalah sensitivitas k-NN terhadap jumlah/jenis fitur dan keberadaan *missing value*, sehingga kualitas *pre-processing* menjadi faktor penentu performa (Gamadarenda et al., 2020).

$$d(x, y = \sqrt{\sum_{i=1}^{n} \quad (xi - yi)^2} \qquad (1)$$

Keterangan:
xi = data latih
yi = data uji
i = data variabel
n = dimensi data

2.4. *Naive Bayes*

Naive Bayes adalah metode klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes untuk menghitung peluang suatu kelas diberikan fitur-fitur tertentu. Asumsi "*naive*" menyatakan bahwa antar fitur dianggap saling independen bersyarat terhadap kelas, sehingga perhitungan peluang dapat disederhanakan namun tetap efektif pada banyak kasus klasifikasi. Dalam praktik, *Naive Bayes* kerap disukai karena prosesnya relatif cepat, sederhana, dan dapat diterapkan pada data berukuran besar. Beberapa kajian juga menekankan performanya yang cukup baik ketika data memiliki ketidakseimbangan kelas atau terdapat nilai hilang (dengan strategi penanganan yang sesuai) (Sisodia & Sisodia, 2018).

Untuk fitur numerik, varian yang umum digunakan adalah *Gaussian Naive Bayes* (mengasumsikan distribusi normal pada fitur kontinu). Dalam penelitian berbasis RapidMiner, pemodelan *Naive Bayes* sering disandingkan dengan k-NN untuk membandingkan karakteristik metode probabilistik vs berbasis jarak pada dataset kesehatan (Lubis et al., 2024).

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \qquad (2)$$

Keterangan:
y = data dengan kelas yang belum diketahui
x = hipotesis data y merupakan suatu kelas spesifik
$P(x|y)$ = probabilitas hipotesis x berdasar kondisi y (*posteriori probabilit*y)
P(x) = probabilitas hipotesis x (*prior probability*)
P($y|x$) = probabilitas y berdasarkan kondisi pada hipotesis x
P(y) = probabilitas dari y

2.5. *Result and Discussion*

Dataset yang digunakan pada penelitian ini berasal dari *website* Kaggle: Diabetes Health Indicators Dataset (sumber BRFSS—Behavioral Risk Factor Surveillance System). Dataset ini memuat lebih sekitar 200,000 baris data dengan variabel target diabetes_012 (0/1/2) dengan label kelas 0 (no diabetes), 1 (prediabetes), 2 (diabetes) dan sejumlah fitur indikator kesehatan serta kebiasaan, serta dicatat memiliki kondisi kelas yang tidak seimbang (imbalanced).

Dataset ini relevan karena mencakup indikator perilaku dan kebiasaan yang berhubungan dengan risiko diabetes, seperti aktivitas fisik, konsumsi buah/sayur, perilaku merokok, konsumsi alkohol berat, dan indikator terkait layanan kesehatan (misalnya akses pemeriksaan atau asuransi), selain indikator klinis/riwayat seperti tekanan darah tinggi dan kolesterol. Dengan demikian, model klasifikasi dapat mempelajari keterkaitan pola perilaku dengan kelas risiko/status diabetes.

Karena data berskala besar dan variabelnya beragam (biner dan numerik), dibutuhkan *preprocessing* yang konsisten agar algoritma k-NN dan *Naive Bayes* dapat bekerja stabil serta menghasilkan performa evaluasi yang dapat dibandingkan secara adil.

Tabel 1. Atribut dalam Dataset

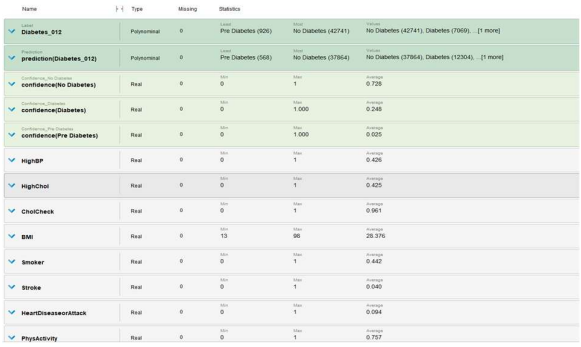
No	Attribute Name	Data Type	Description
1	Diabetes_012	Numerical (Interval Scale)	Disease status 0 = no diabetes 1 = prediabetes 2 = diabetes
2	HighBP	Boolean	Blood pressure 0 = no high BP 1 = high BP
3	HighChol	Boolean	Cholesterol 0 = no high cholesterol 1 = high cholesterol
4	CholCheck	Boolean	Cholesterol check 0 = no cholesterol check in 5 years 1 = yes, having cholesterol check in 5 years
5	BMI	Numeric (Continuous)	Value of body Mass Index based on patient weight and height
6	Smoker	Boolean	Have ever smoked more than 100 cigarettes (5 packs) in a lifetime 0 = no 1 = yes
7	Stroke	Boolean	Have ever experienced a stroke 0 = no 1 = yes
8	HeartDiseaseorAttack	Boolean	Have ever experiencing coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
9	PhysActivity	Boolean	Physical activity within the past 30 days 0 = no 1 = yes
10	Fruits	Boolean	Consume fruit at least one or more times per day 0 = no 1 = yes
11	Veggies	Boolean	Consume vegetables at least one or more times per day

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

			0 = no 1 = yes
12	HvyAlcoholConsump	Boolean	Heavy drinkers · Adult men – having more than 14 drinks per week · Adult women – having more than 7 drinks per week 0 = no 1 = yes
13	AnyHealthcare	Boolean	Have any type of health care coverage (health insurance, prepaid plans like HMO) 0 = no 1 = yes
14	NoDocbcCost	Boolean	In the past 12 months, have you had a need to see a doctor but couldn't because of the cost? 0 = no 1 = yes
15	GenHlth	Numerical (Interval Scale)	Would you say that your overall health is: Scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
16	MentHlth	Numerical (Interval Scale)	In the last 30 days, how many days did you have poor mental health, including stress, depression, emotional issues? Scale 1-30 days
17	PhysHlth	Numerical (Interval Scale)	In the last 30 days, how many days did you have poor physical health, including physical illness and injury? Scale 1-30 days
18	DiffWalk	Boolean	Have trouble walking or climbing stairwells? 0 = no 1 = yes
19	Sex	Nominal Scale	Gender 0 = female 1 = male
20	Age	Numerical (Ratio Scale)	Level age category 1 = Age 18 to 24 2 = Age 25 to 29 3 = Age 30 to 34 4 = Age 35 to 39 5 = Age 40 to 44 6 = Age 45 to 49 7 = Age 50 to 54 8 = Age 55 to 59 9 = Age 60 to 64 10 = Age 65 to 69 11 = Age 70 to 74 12 = Age 75 to 79 13 = Age 80 or older
21	Education	Numerical (Ratio Scale)	Education level Scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
22	Income	Numerical (Ratio Scale)	Income Scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

2.6. *Data Preprocessing*

Tahap berikutnya adalah *data cleaning*, yaitu menata kualitas data sebelum pemodelan dengan cara meninjau dan menangani data yang berpotensi mengganggu analisis, seperti nilai yang hilang, data tidak valid, ketidak konsistenan, duplikasi, maupun kesalahan input. *Data cleaning* berpengaruh terhadap *data mining* akibat pengurangan jumlah dan kompleksitas data. Tahap preprocessing berawal di kisaran 200,000 baris data. Setelah proses pembacaan, data kemudian dibagi menggunakan Split Data dengan rasio 80:20 (80% untuk data latih dan 20% untuk data uji). Dari hasil split tersebut, didapatkan data uji berjumlah sekitar 50,000 baris data mewakili 20% dari keseluruhan data.



Gambar 2. Hasil *Missing Value*

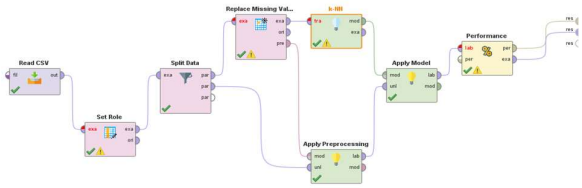
Setelah melalui proses missing value menggunakan RapidMiner pada gambar diatas dapat dikatakan bahwa hasilnya adalah nol, yang berarti bahwa data yang digunakan telah menjadi bersih, tidak ada lagi nilai yang hilang atau kosong dalam dataset tersebut, sehingga dapat digunakan untuk proses selanjutnya.

2.7. *Implementation Algorithms*

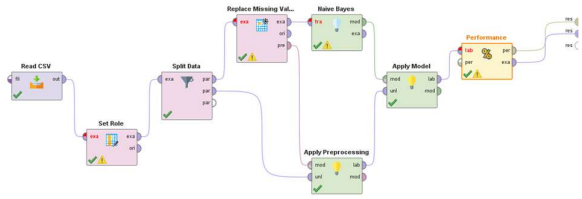
Perangkat lunak RapidMiner digunakan dalam proses pengolahan data pada tingkat klasifikasi ini. Dalam penelitian ini diterapkan dua algoritma klasifikasi, yaitu *Naive Bayes*, dan k-NN, untuk mengolah data yang tersedia. Tujuan penelitian ini adalah membandingkan tingkat akurasi yang dihasilkan

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

serta menarik kesimpulan berdasarkan hasil klasifikasi individu yang menderita diabetes melitus.



Gambar 3. Penerapan model k-NN



Gambar 4. Penerapan model *Naive Bayes*

A. *K-Nearest Neighbors* (k-NN)

Berikut adalah hasil pengolahan algoritma K-Nearest Neighbor menggunakan K-Fold dengan jumlah k=3, dimana penentuan kelas berdasarkan hitungan jarak terdekat data baru ke masing-masing data point di data latih.

Tabel 2. *Confusion matrix of k-NN*

***Accuracy*: 81.72%**

	true No Diabetes	true Diabetes	true Pre Diabetes	class precision
pred. No Diabetes	39799	5351	761	86.69%
pred. Diabetes	2782	1652	157	35.98%
pred. Pre Diabetes	160	66	8	3.42%
class recall	93.12%	23.37%	0.86%	

Berdasarkan hasil confusion matrix pada model k-NN, akurasi yang diperoleh sebesar 81,72%. Jika dilihat per kelas, k-NN sangat dominan memprediksi kelas No Diabetes dari 42.741 data No Diabetes, sebanyak 39.799 berhasil diprediksi benar dengan recall 93,12% dan precision 86,69%, sehingga model cukup kuat untuk mengenali kelas mayoritas. Sebaliknya, performa untuk kelas Diabetes justru rendah karena dari 7.069 data Diabetes yang sebenarnya, hanya 1.652 yang terdeteksi benar (recall 23,37%), sedangkan sebagian besar kasus Diabetes malah diprediksi sebagai No Diabetes (5.351), yang menunjukkan banyak false negative. Kondisi yang lebih berat terjadi pada kelas Pre Diabetes, karena dari 926 data Pre Diabetes hanya 8 yang diprediksi benar (recall 0,86%) dan precision-nya juga rendah (3,42%), sehingga hampir seluruh data Pre Diabetes menjadi No Diabetes (761) atau Diabetes (157). Jadi, meskipun akurasinya terlihat tinggi, hasil ini menunjukkan k-NN dalam konfigurasi saat ini cenderung mengikuti kelas mayoritas dan belum efektif untuk membedakan kelas minoritas, terutama Pre Diabetes.

B. *Naive Bayes Classifier*

Tabel 3 menampilkan hasil evaluasi penerapan algortima Naive Bayes dengan confusion matrix

Tabel 3. *Confusion matrix of Naive Bayes*

***Accuracy*: 75.42%**

	true No Diabetes	*true Diabetes*	*true Pre Diabetes*	*class precision*
pred. No Diabetes	34335	3012	517	90.68%
pred. Diabetes	7998	3914	392	31.81%
pred. Pre Diabetes	408	143	17	2.99%
class recall	80.33%	55.37%	1.84%	

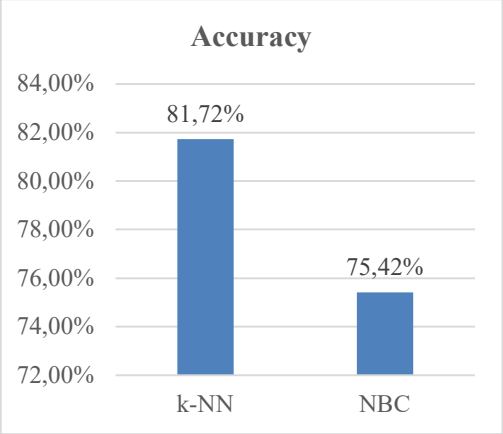
Berdasarkan hasil pengujian menggunakan metode *Naive Bayes,* model ini terlihat paling baik pada kelas No Diabetes dengan precision 90,68% dan recall 80,33%, sehingga sebagian besar data pada kelas mayoritas dapat dikenali dengan cukup konsisten. Pada kelas Diabetes, model menunjukkan recall 55,37% yang berarti lebih dari setengah kasus diabetes berhasil terdeteksi, namun precision masih rendah yaitu 31,81%. Ini mengindikasikan bahwa ketika model memprediksi Diabetes, cukup banyak prediksi tersebut yang sebenarnya berasal dari kelas lain (false positive), terutama dari kelas No Diabetes. Sementara itu, kelas Pre Diabetes memiliki nilai precision 2,99% dan recall 1,84%, yang menunjukkan bahwa model hampir tidak mampu membedakan kelas ini secara akurat. Kondisi tersebut wajar terjadi karena jumlah data Pre Diabetes jauh lebih sedikit dibanding kelas lain, serta karakteristiknya cenderung berada di area transisi sehingga mudah tertukar dengan No Diabetes maupun Diabetes. Dengan demikian, meskipun akurasi total model mencapai 75,42%, evaluasi per kelas memperlihatkan bahwa model masih kurang optimal untuk mendeteksi kelas minoritas, khususnya Pre Diabetes.

2.8. *Accuracy Comparison*

Setelah data diolah menggunakan kedua algoritma, maka langkah selanjutnya adalah membandingkan

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

akurasi antara Naive Bayes dan k-NN untuk menemukan hasil yang terbaik. Berikut hasil perbandingannya, hasil tersebut dapat dilihat pada gambar dibawah.

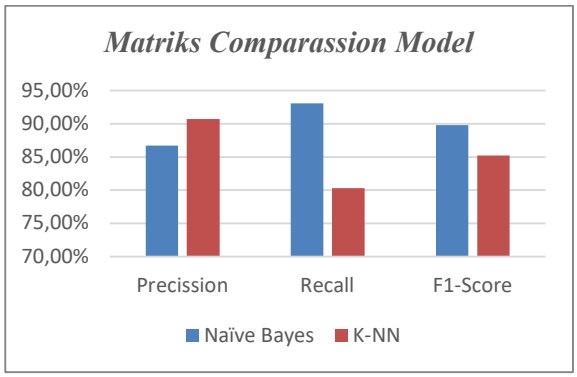


Gambar 5. Hasil Perbandingan Akurasi k-NN dan NBC

Berdasarkan hasil pengujian pada gambar 5, model k-NN memperoleh akurasi sebesar 81,72%, sedangkan Naive Bayes menghasilkan akurasi sebesar 75,42%. Selisih sebesar 6,30 poin persentase menunjukkan bahwa k-NN memiliki kemampuan klasifikasi yang lebih baik pada dataset yang digunakan. Peningkatan ini mengindikasikan bahwa pendekatan berbasis kedekatan jarak (instance-based learning) pada k-NN lebih efektif dalam menangkap pola distribusi data dibandingkan pendekatan probabilistik pada Naive Bayes.

Secara karakteristik algoritma, Naive Bayes bekerja dengan asumsi independensi antar fitur. Pada dataset kesehatan seperti diabetes, asumsi ini sering kali tidak sepenuhnya terpenuhi karena variabel seperti kadar gula darah, BMI, tekanan darah, dan usia cenderung saling berkorelasi. Ketidaksesuaian terhadap asumsi independensi tersebut dapat menyebabkan penurunan performa model Naive Bayes. Sebaliknya, k-NN tidak bergantung pada asumsi distribusi tertentu dan menentukan kelas berdasarkan kemiripan langsung terhadap data tetangga terdekat, sehingga lebih fleksibel dalam menangani pola non-linear.

Namun demikian, distribusi kelas yang tidak seimbang pada dataset juga perlu diperhatikan. Dominasi kelas *No Diabetes* berpotensi meningkatkan nilai akurasi secara keseluruhan tanpa menjamin peningkatan kemampuan deteksi pada kelas minoritas seperti *Diabetes* dan *Pre Diabetes*. Oleh karena itu, meskipun k-NN menunjukkan performa akurasi yang lebih tinggi, evaluasi kinerja model sebaiknya tidak hanya bergantung pada metrik akurasi, tetapi juga mempertimbangkan *precision, recall*, dan *F1-score* untuk menilai kemampuan model dalam mendeteksi kasus yang menjadi fokus penelitian. Matriks lainnya yang menunjukan bahwa model naive bayes lebih unggul dapat di lihat pada gambar 6 berikut



Gambar 6. Perbandingan Kinerja Model

Berdasarkan hasil evaluasi kinerja model pada gambar 6, diketahui bahwa Naïve Bayes menghasilkan nilai precision sebesar 86,69%, recall sebesar 93,12%, dan F1-Score sebesar 89,79%, sedangkan *K-Nearest Neighbor* (K-NN) memperoleh precision sebesar 90,68%, recall sebesar 80,33%, dan F1-Score sebesar 85,19%. Meskipun K-NN memiliki nilai precision yang lebih tinggi, Naïve Bayes menunjukkan kemampuan yang lebih baik dalam menangkap data positif yang relevan, hal ini dapat dilihat dari nilai recall dan F1-Score yang lebih tinggi. Hal ini menunjukkan bahwa secara keseluruhan Naïve Bayes memiliki keseimbangan performa yang lebih optimal dibandingkan K-NN pada dataset yang digunakan, sehingga model tersebut lebih direkomendasikan untuk kasus klasifikasi diabetes dalam penelitian ini. Namun, interpretasi kinerja model tetap harus mempertimbangkan keseimbangan data dan tujuan utama deteksi.

3. Kesimpulan

Setelah menerapakan percobaan dengan alur yang sama seperti pra-pemrosesan, split data 80:20, stratified sampling, dapat disimpulkan bahwa kedua model punya karakter yang berbeda. Naive Bayes memberikan akurasi 75,42% dan performanya relatif lebih seimbang dibanding k-NN karena masih mampu mendeteksi kelas Diabetes dengan recall yang lebih baik, meskipun konsekuensinya prediksi Diabetes masih cukup banyak yang salah dimana false positive cukup tinggi. Di sisi lain, k-NN menghasilkan akurasi yang lebih tinggi yaitu 81,72%, tetapi model ini cenderung sangat mengikuti kelas mayoritas (No Diabetes) sehingga prediksi untuk kelas minoritas, terutama Pre Diabetes, masih sangat lemah dan banyak kasus Diabetes/Pre Diabetes yang terlewat karena diprediksi sebagai No Diabetes. Dengan dilakukan penelitian ini diharapkan dapat memberikan informasi bermanfaat mengenai penyakit diabetes melitus dan algoritma yang lebih unggul diantara algoritma KNN dan Naïve Bayes sehingga dapat menjadi acuan referensi serta pengembangan ilmu pengetahuan untuk penelitian selanjutnya.

Daftar Pustaka

A'yuniyah, Q., Tasia, E., Nazira, N., Pratama, P. F., Anugrah, M. R., Adhiva, J., & Mustakim, M. (2022). Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(1), 72. https://doi.org/10.30865/json.v4i1.4781

Aprianti, B., Aulia, A., & Purnamasari, I. (2022). Algoritma Linear Regression dalam Memprediksi Pertumbuhan Jumlah Penduduk Menurut Provinsi dan Jenis Kelamin. *Jurnal Ilmiah Wahana Pendidikan*, 89–92. https://doi.org/https://doi.org/10.5281/zenodo.6408866

Arta, J. K. I., Indarawan, G., & Dantes, R. G. (2019). Data Mining Rekomendasi Calon Mahasiswa Berprestasi Di STMIK Denpasar Menggunakan Metode Technique For Others Reference By Similarity To Ideal Solution. *Jurnal Ilmu Komputer Indonesia (JIKI)*. https://doi.org/https://doi.org/10.23887/jstundiksha.v5i2.8549

Asroni, A., Respati, B. M., & Riyadi, S. (2018). *Penerapan Algoritma C4 . 5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta*. *21*(2), 158–165. https://doi.org/10.18196/st.212222

Binus. (2021). *No Title*. Proses Data Mining KDD. https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/

Gamadarenda, I. W., Waspada, I., Diponegoro, U., Korespondensi, P., Atribut, S., & Backward, A. (2020). IMPLEMENTASI DATA MINING UNTUK DETEKSI PENYAKIT GINJAL KRONIS (PGK) MENGGUNAKAN K-NEAREST NEIGHBOR (K-NN) DENGAN BACKWARD DATA MINING IMPLEMENTATION FOR DETECTION OF CHRONIC KIDNEY (CKD) USING K-NEAREST NEIGHBOR (K-NN) WITH BACKWARD ELIMINATION. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(2), 417–426. https://doi.org/10.25126/jtiik.202071896

Hunafa, M. R., & Hermawan, A. (2023). KLIK: Kajian Ilmiah Informatika dan Komputer Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes. *Media Online*, *4*(3), 1551–1561. https://doi.org/10.30865/klik.v4i3.1486

Ismail, L., Materwala, H., & Al, J. (2021). Association of risk factors with type 2 diabetes : A systematic review. *Computational and Structural Biotechnology Journal*, *19*, 1759–1785. https://doi.org/10.1016/j.csbj.2021.03.003

Kandhasamy, J. P., & Balamurali, S. (2015). Performance Analysis of Classifier Models to Predict Diabetes Mellitus. *Procedia - Procedia Computer Science*, *47*, 45–51. https://doi.org/10.1016/j.procs.2015.03.182

Kudus, U. M., Ganesha, J., & Kudus, P. (2020). Fida Maisa Hana. *Jurnal Sistem Komputer Dan Kecerdasan Buatan (Siskom-KB)*, *4*(1), 32–39. https://doi.org/https://jurnal.tau.ac.id/index.php/siskom-kb/article/view/173

Lestari, U. I., Nadhiroh, A. Y., & Novia, C. (2021). Penerapan Metode K-Nearest Neighbor Untuk Sistem Pendukung Keputusan Identifikasi Penyakit Diabetes Melitus. *Teknik Informatika Dan Sistem Informasi*, *8*(4), 2071–2082. https://doi.org/https://doi.org/10.35957/jatisi.v8i4.1235

Levi Sabili, N., Rakhmat Umbara, F., & Melina, M. (2024). KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA CATEGORICAL BOOSTING DENGAN FAKTOR RISIKO DIABETES. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(6), 11391–11398. https://doi.org/10.36040/jati.v8i6.11447

Li, Y.; Li, H.; Yao, H. (2018). Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China, 2016-2017. Computational and Mathematical Methods in Medicine. *Computational and Mathematical Methods in Medicine*. https://doi.org/10.1155/2018/7207151

Lubis, A. F., Haq, H. Z., Lestari, I., & Iltizam, M. (2024). Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors , Naïve Bayes and Decission Tree. *Classification of Diabetes Mellitus*, *2*(July), 44–51.

Mohamed, M. H., Khafagy, M. H., Mohamed, N., Kamel, M., & Said, W. (2024). DIABETIC MELLITUS PREDICTION WITH BRFSS DATA SETS. *DIABETIC MELLITUS PREDICTION*, *102*(3).

Nurdiana, N., & Algifari, A. (2020). Studi Komparasi Algoritma Id3 Dan Algoritma *Naive Bayes* Untuk Klasifikasi Penyakit Diabetes Mellitus. *INFOTECH Journal*, *6*(2), 18–23. https://ejournal.unma.ac.id/index.php/infotech/article/view/816

Prasatya, A., Siregar, R. R. A., & Arianto, R. (2020). Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes. *PETIR*, *13*(1), 86–100. https://doi.org/10.33322/petir.v13i1.925

Putri, A. I., Husna, N. A., Cia, N. M., & Arba, M. A. (2024). *Implementation of K-Nearest Neighbors , Naïve Bayes Classifier , Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction*. *2*(July), 26–33.

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

Putry, N. M., Sari, B. N., Kom, M., Informatika, T., & Karawang, U. S. (2022). *KOMPARASI ALGORITMA K-NN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELITUS*. *10*(1). https://www.researchgate.net/publication/362418111_KOMPARASI_ALGORITMA_K-NN_DAN_NAIVE_BAYES_UNTUK_KLASIFIKASI_DIAGNOSIS_PENYAKIT_DIABETES_MELLITUS

Ridwan, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *Siskom-Kb*, *IV*(September), 15–21. https://doi.org/https://doi.org/10.47970/siskom-kb.v4i1.169

Sisodia, D., & Sisodia, D. S. (2018). ScienceDirect Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*(Iccids), 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Sugara, B., Adidarma, D., & Budilaksono, S. (2018). Perbandingan akurasi algoritma c4.5 dan naïve bayes untuk deteksi dini gangguan autisme pada anak. *Komputer Dan Informatika*, *3*(18), 119–128. https://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/view/308