Prediksi Penyakit Diabetes Menggunakan Algoritma KNN Dengan Mengkomparasikan Nilai K dan K-Fold Cross-Validation

Dosen Pengampu:

Abu Salam, M.Kom

Disusun Oleh:

R Bagus Ario Arlianda Dwiputra (A11.2020.12796)

Program Studi Teknik Informatika

Fakultas Ilmu Komputer Universitas Dian Nuswantoro

I. DAFTAR ISI

II. PENDAHULUAN

Diabetes merupakan penyakit tidak menular yang cukup serius dimana insulin tidak dapat diproduksi secara maksimal oleh pancreas (Safitri & Nurhayati, 2019). Kelompok penyakit metabolik dengan karakteristik hiperglikemia karena kelainan sekresi insulin, kerja insulin atau kedua duanya. Insulin yang tidak bekerja dengan kuat membuat kadar glukosa darah dalam darah akan meninggi. Penyakit ini banyak dialami oleh masyarakat yang dimana menjadi prioritas dalam memecahakan masalah kesehatan. Penyakit Diabetes Mellitus merupakan ranking keenam penyebab kematian di Dunia, hal ini diungkapkan oleh dunia World Health Organization (WHO) (Wicaksono, 2015). Data yang didapatkan bahwa kematian yang disebabkan karena diabetes ada sekitar 1,3 juta dan yang meninggal sebelum usia 70 tahun sebanyak 4 persen dengan mayoritas kematian diabetes pada usia 45-54 tahun (Kistianita, Yunus, & Gayatri, 2018). Kebanyakan pasien tersebut meninggal akibat komplikasi dari penyakit diabetes mellitus. Komplikasi ini timbul tergantung dari lamanya penyakit ini diderita atau dari keparahan penyakit itu sendiri. Komplikasi yang dimaksud disini adalah komplikasi makrovaskuler dan mikrovaskuler.

Penyakit diabetes mellitus atau yang dikenal dengan sebutan kencing manis adalah suatu penyakit gangguan metabolisme kronis yang ditandai dengan peningkatan kadar gula disertai dengan gangguan metabolisme karbohidrat, lipid dan protein, sebagai akibat oleh defisiensi produksi insulin oleh pankreas, atau sel-sel tubuh kurang responsif terhadap insulin, atau bisa kedua-duanya. Penelitian Kusnadi yang menyatakan seseorang dengan riwayat keluaraga DM akan berisiko 6 kali lebih besar dibandingkan dengan seseorang tanpa ada riwayat keluarga DM (Kusnadi, Murbawani, & Fitranti, 2017). Faktor resiko lain yang dapat dimodifikasi adalah faktor pola makan, kebiasaan merokok, obesitas, hipertensi, stress, kurangnya aktifitas fisik, alcohol dan lain sebaginya. Adanya kaitan obesitas dengan kadar glokosa darah

dimana IMT > 23 dapat menyebabkan peningkatan glukosa darah (Tandra, 2017). Adapun Klasifikasi diabetes mellitus berdasarkan etiologinya adalah sebagai berikut :

1. Diabetes Mellitus Tipe 1 : defisiensi insulin absolut akibat destruksi sel beta
2. Diabetes Mellitus Tipe 2 : bervariasi mulai dari resistensi insulin predominan disertai defisiensi insulin relatif, sampai defek sekresi insulin predominan disertai resistensi insulin
3. Diabetes Tipe Lain : akibat defek genetik fungsi sel beta, defek genetik kerja insulin, pankreas, penyakit endokrinopati, eksokrin obat/bahan kimia, infeksi, imunologi, dan sindroma genetik lain. 4. Diabetes Gestasional : pada kehamilan
4. Pra-diabates : IFG (Impaired Fasting Glucose) dan IGT (Impaired Glucose Tolerance)
5. Komplikasi Penyakit Diabetes Mellitus
6. Adapun penyakit kompilkasi yang diakibatkan dari penyakit diabetes mellitus yaitu:

I. Hipoglikemia Serangan hipoglikemia ditandai dengan perasaan pusing, lemas, gemetar, mata berkunang-kunang, keringat dingin, detak jantung meningkat, sampai hilang kesadaran. Hipoglikemnia biasanya timbul bila kadar glukosa darah < 50 mg/dl, dan ini terjadi apabila dosis obat anti diabetes atau insulin terlalu tinggi, makan terlalu sedikit, olahraga terlalu berat, minum alkohol atau depresi.

II. Hiperglikemia Hiperglikemia yang dimaksud disini adalah suatu keadaan dimana kadar gula darah tiba-tiba melonjak. Hal ini disebabkan antara lain oleh stress, infeksi, dan konsumsi obat-obatan tertentu. Hiperglikemia ditandai dengan poliuria, polidipsia, polifagia, kelelahan yang parah, dan Hipergikemia pandangan dapat kabur. memperburuk gangguan-gangguan kesehatan seperti gastroparesis, disfungsi ereksi, dan infeksi jamur pada vagina. Hiperglikemia yang berlangsung berkembang lama menjadi dapat keadaan metabolisme yang berbahaya antara lain ketoasidosis diabetik (Diabetic Ketoacidosis), yang

dapat berakibat fatal dan membawa kematian. Hiperglikemia dapat dicegah dengan kontrol kadar gula darah yang ketat.

III. Komplikasi Makrovaskuler Komplikasi makrovaskular yang umum berkembang pada penderita diabetes adalah penyakit jantung koroner, penyakit pembuluh darah otak, dan penyakit pembuluh darah perifer 1,2. Komplikasi makrovaskular lebih sering timbul pada DM tipe 2, yang umumnya menderita hipertensi, dislipidemia dan atau kegemukan, walaupun komplikasi makrovaskular dapat juga terjadi pada DM tipe 1.

IV. Komplikasi Mikrovaskuler Komplikasi ini terutama terjadi pada penderita diabetes tipe 1. Komplikasi mikrovaskuler yang timbul antara lain retinopati, nefropati, dan neuropati. Disamping karena kondisi hiperglikemia, ketiga komplikasi ini juga dipengaruhi oleh faktor genetik. Untuk berkembang kearah komplikasi mikrovaskular, tergantung lamanya (durasi) sakit dan tingkat keparahan diabetes. Satu-satunya cara untuk mencegah atau memperlambat jalan perkembangan mikrovaskular adalah komplikasi dengan pengendalian kadar gula darah yang ketat.

V. Dari permasalahan tersebut perlu adanya deteksi sejak dini agar dapat dilakukan penanganan dengan cepat terhadap pengidap diabetes (Gunawan et al., 2020). Model prediksi diabetes perlu dikembangkan agar dapat memberikan dampak yang signifikan dalam mendeteksi penyakit diabetes sejak dini. Dengan kemajuan teknologi yang begitu pesat terutama di bidang kecerdasan buatan model prediksi dapat dikembangkan dengan bantuan pembelajaran komputer tentang machine learning (Manongga et al., 2022). . Dari berbagai percobaan yang dilakukan dengan menggunakan algoritma tunggal dalam memprediksi penyakit diabetes belum mendapatkan hasil yang terbaik. Hal ini disebabkan setiap algoritma bergantung pada karakteristik data. Upaya dalam mengatasi permasalahan tersebut dengan menggabungkan algoritma tunggal menjadi satu model yang baik dalam menangani perbedaan karakteristik data. seperti pada penelitian yang dilakukan oleh (Kibria et al., 2022) yang mencoba melakukan ensemble learning atau menggabungkan algoritma pohon keputusan dalam memprediksi penyakit diabetes, dari

percobaan tersebut teknik ensemble learning mendapatkan nilai akurasi tinggi daripada menggunakan algoritma tunggal. Dengan menggabungkan beberapa algoritma, diharapkan dapat membantu dengan kelebihan masing-masing algoritma dan mengurangi kelemahan algoritma yang digabungkan(Masacgi & Rohman, 2023)

VI. Penelitian ini bertujuan melakukan peningkatan performa model Hard Voting Classifier dalam memprediksi penyakit diabetes dengan mengatasi permasalahan yang ada pada dataset dari penelitian ini dengan menggunakan teknik oversampling ADASYN. Dengan menerapkan penanganan ketidakseimbangan data terdapat peningkatan hasil akurasi, presisi, recall, dan f1 score model prediksi.

III. SUMBER DATASET PENELITIAN

Dataset ini mencakup informasi demografis dan klinis dari 1000 pasien dengan berbagai kondisi kesehatan. Kolom yang terdapat dalam dataset adalah sebagai berikut.

	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
0	0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
1	1	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23.0	0
2	0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
3	0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24.0	0
4	1	33	7.1	46	4.9	4.9	1.0	0.8	2.0	0.4	21.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
995	1	71	11.0	97	7.0	7.5	1.7	1.2	1.8	0.6	30.0	2
996	1	31	3.0	60	12.3	4.1	2.2	0.7	2.4	15.4	37.2	2
997	1	30	7.1	81	6.7	4.1	1.1	1.2	2.4	8.1	27.4	2
998	1	38	5.8	59	6.7	5.3	2.0	1.6	2.9	14.0	40.5	2
999	1	54	5.0	67	6.9	3.8	1.7	1.1	3.0	0.7	33.0	2

1000 rows × 12 columns

Dataset tersebut tampaknya berfokus pada profil metabolik dan faktor risiko kardiovaskular. Ini sering ditemukan dalam studi terkait diabetes, penyakit jantung, dan sindrom metabolik.

- Gender: Jenis kelamin pasien
- AGE: Usia pasien
- Urea: Kadar urea dalam darah
- Cr: Kadar kreatinin dalam darah
- HbA1c: Hemoglobin A1c, indikator kontrol gula darah
- Chol: Kolesterol total
- TG: Trigliserida
- HDL: High-Density Lipoprotein (Kolesterol baik)
- LDL: Low-Density Lipoprotein (Kolesterol jahat)
- VLDL: Very Low-Density Lipoprotein
- BMI: Indeks massa tubuh
- Class penyakit

Pada bagian penentuan data objek dilakukan beberapa proses untuk mempersiapkan data tersebut agar dapat digunakan untuk proses pemodelan. Setelah dilihat grafik visualisasi sebaran kelas seluruh fitur, terlihat kolom 'ID' dan 'No_Pation' mempunyai sebaran kelas yang bervariasi dan tentunya atribut ini juga mencerminkan nomor id dan jumlah pasien, sehingga atribut ini tidak tidak perlu digunakan dalam pemodelan sehingga pada proses ini kolom dihilangkan.

IV. DETERMINE DATA OBJECT

A. Identifikasi Atribut

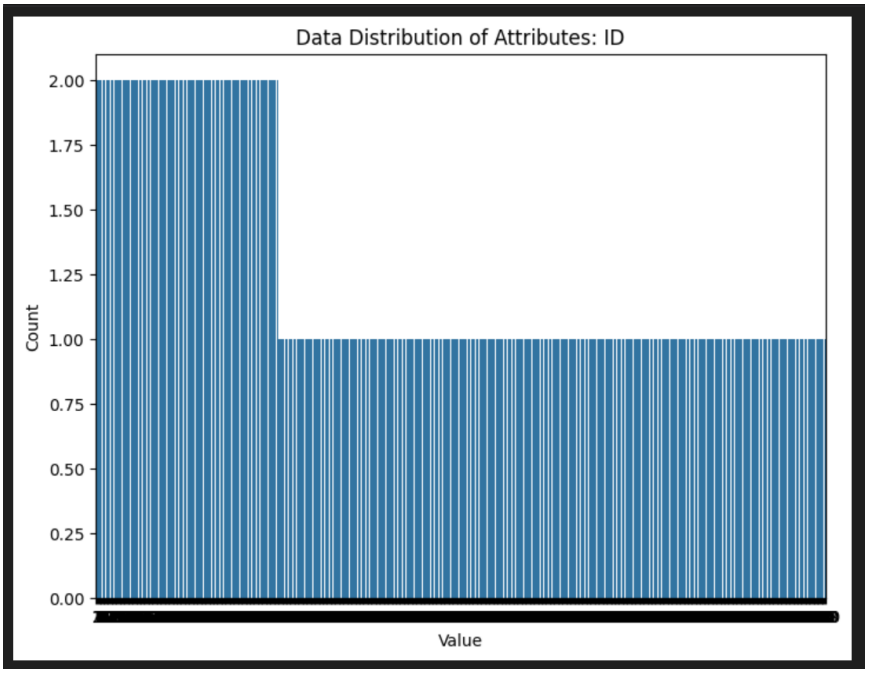
Atribut "ID" adalah atribut unik yang biasanya digunakan untuk mengidentifikasi setiap entitas dalam dataset. Ini bisa berupa nomor urut, kode unik,

atau pengenal lain yang memastikan bahwa setiap baris data dapat dibedakan dari baris lainnya.

B. Distribusi Data

Grafik menunjukkan jumlah entitas yang memiliki nilai "ID" tertentu. Distribusi yang Anda berikan menunjukkan dua hal:

- **Banyaknya Baris Data dengan Nilai Unik ID**: Setiap nilai "ID" muncul tepat sekali dalam dataset, yang berarti tidak ada duplikasi pada kolom ini. Ini adalah karakteristik umum dari pengenal unik.
- **Tidak Ada Nilai yang Hilang**: Setiap baris dalam dataset memiliki nilai "ID", yang berarti tidak ada data yang hilang untuk atribut ini.



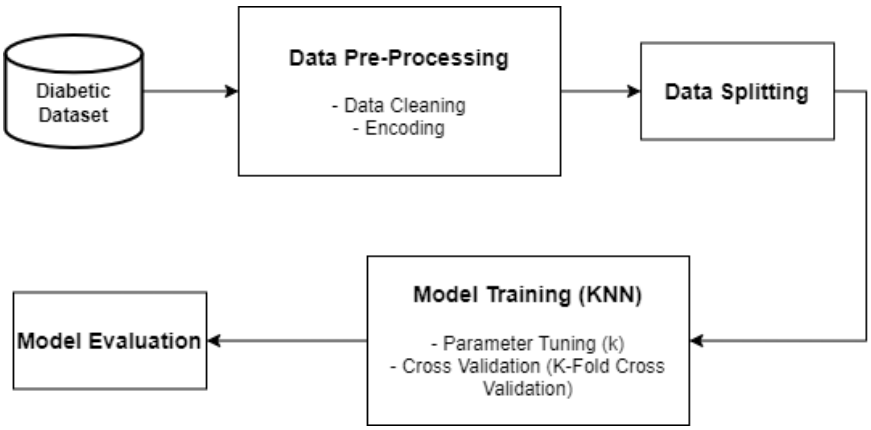
Gambar 1. *Gambar Data Distribusi*

Determine data object untuk atribut "ID" dalam dataset ini adalah sebagai pengenal unik untuk setiap entitas dalam dataset. Grafik distribusi menunjukkan bahwa setiap nilai "ID" adalah unik dan tidak ada yang hilang, yang berarti data ini bersih dalam hal pengenal unik. Atribut ini penting untuk tujuan identifikasi dan referensi, tetapi tidak berguna untuk analisis statistik atau pembelajaran mesin secara langsung.

V. METODOLOGI PENELITIAN

Pada penelitian ini menggunakan Teknik K-Fold Cross Validation. Tetapi sebelum melewati proses tersebut ada beberapa tahap yang harus dilewati. Dalam prediksi penyakit diabetes melewati proses data cleaning dan encoding, fungsi dari kedua tahap itu merupakan pembersihan data yaitu mencari missing value atau duplicate value kemudian encoding adalah untuk mengubah data kategorikal menjadi numerik. Selanjutnya data splitting yaitu untuk membagi data antara data sample dan juga data uji. Kemudian masuk ke model training (KNN) yaitu parameter tuning, mempunyai fungsi untuk meningkatkan performa uji model sehingga memberikan prediksi lebih akurat dan lebih baik. Dan yang terakhir yaitu tahap Cross Validation (K-Fold Cross Validation)

A.



Gambar 1. Contoh Diagram Alir/Flowchart Penelitian

Pembelajaran mesin ini adalah pembuatan prediktif berdasarkan data. Proses pembelajaran mesin yang ada pada gambar tersebut mencakup beberapa langkah yang perlu di perhatikan untuk membuat model klasifikasi yang akurat dan efisien Pertama, data disiapkan melalui pra pemrosesan, Sebelum memasuki tahap yang

lebih mendalam, memastikan bahwa data telah bersih merupakan hal yang penting dalam pemrosesan data (Setiawan et al., 2024), setelah melakukan pembersihan data kemudian encoding, encoding merupakan salah satu tahap preprocessing. Cara ini untuk mengubah value dari fitur yang berupa teks menjadi angka (Kristiawan & Widjaja, 2021) dan dinormalisasi untuk memastikan kualitas data. Data tersebut kemudian dibagi menjadi set pelatihan dan set pengujian untuk pelatihan dan evaluasi model. Ketidakseimbangan kelas dalam data selanjutnya diatasi dengan menggunakan teknik random

VI. HASIL DAN PEMBAHASAN

K-Fold Cross Validation adalah metode evaluasi model yang membagi dataset menjadi kkk bagian (folds) yang sama besar.

Tabel 1. Hasi Eksperimen

K	K-FOLD	ACCURACY	PRECISION	RECALL	F1 SCORE
2	5	87%	66,22%	71,35%	65,36%
3	5	87.75%	65,56%	66,25%	65,45%
5	5	87.37%	64.94%	66.30%	64.97%
2	10	87.37%	61.52%	69.85%	64.21%
3	10	87.75%	63.49%	65.72%	63.51%
5	10	87.62%	63.12%	64.77%	63.09%
2	15	88%	63.98%	72.05%	66.63%
3	15	87.63%	64.84%	65.63%	64.23%
5	15	87.25%	61.54%	63.96%	61.77%
2	20	87.75%	62%	67.54%	63.07%
3	20	87.37%	64.02%	65.71%	63.02%
5	20	87.12%	58.48%	61.77%	58.46%
2	25	88%	64.05%	70.72%	64.35%
3	25	87.25%	61.88%	65.83%	60.71%
5	25	87.5%	60.51%	63.76%	59.82%

parameter k (jumlah tetangga) dan cv_fold (jumlah lipatan validasi silang)

VII. Analisis Kinerja Model

1) **Parameter K dan CV_Fold** :

- k mewakili jumlah tetangga yang dipertimbangkan dalam algoritma K-Nearest Neighbors (KNN).
- cv_fold mewakili jumlah lipatan dalam validasi silang (cross-validation), yang menentukan berapa kali data akan dipecah menjadi pelatihan dan pengujian.

2) **Nilai K Optimal** :

- k=2 dan k=3k = 3k=3 cenderung memberikan kinerja yang lebih baik pada validasi silang dan data uji.
- Validasi silang dengan 5 atau 10 lipatan memberikan hasil yang konsisten dan baik. Jumlah lipatan yang terlalu besar (20 atau 25) dapat menyebabkan penurunan kinerja model.

VIII. KESIMPULAN

Penelitian ini bertujuan untuk memprediksi penyakit diabetes menggunakan algoritma K-Nearest Neighbors (KNN) dengan membandingkan nilai K dan teknik K-Fold Cross-Validation. Hasil penelitian menunjukkan bahwa nilai K optimal adalah K=2 dan K=3, dengan performa terbaik diperoleh melalui validasi silang 5 atau 10 lipatan. Penggunaan validasi silang dengan 5 dan 10 lipatan menghasilkan akurasi dan kestabilan model yang baik, sementara jumlah lipatan yang terlalu besar (20 atau 25) cenderung menurunkan kinerja. Model prediksi ini diharapkan dapat membantu dalam diagnosa dini diabetes, memungkinkan tindakan pencegahan dan penanganan yang lebih efektif. Penelitian ini menyoroti pentingnya pemilihan parameter yang tepat dan teknik validasi silang dalam algoritma KNN untuk meningkatkan akurasi prediksi, serta memberikan kontribusi signifikan dalam aplikasi pembelajaran mesin untuk prediksi penyakit diabetes.

IX. DAFTAR PUSTAKA

Setiawan, D., Nugraha, A., & Luthfiarta, A. (2024). Komparasi Teknik Feature Selection Dalam Klasifikasi Serangan IoT Menggunakan Algoritma Decision Tree. Jurnal Media Informatika Budidarma, 8(1), 83-93.

Sidik, A. P., Amin, M., & Wilana, A. (2023). Implementasi Perancangan Klasifikasi Kualitas Buah Jeruk Berdasarkan Warna. JOURNAL ZETROEM, 5(1), 72-76.

Rovy, N. W. (2018). Hubungan beberapa faktor yang dapat di modifikasi denga kejadian diabetes melitus tipe 2 pada calon Jemaah haji di kabupaten Magetan DISS, Stikes Bhakti Husada