KLASIFIKASI POTENSI PENYAKIT DIABETES MELLITUS TIPE II PADA PASIEN MENGGUNAKAN ALGORITME KNN (K-NEAREST NEIGHBOR)

Mohammad Sholikhul Fiqri, Henny Dwi Bhakti*
Teknik Informatika, Universitas Muhammadiyah Gresik
Jl. Sumatera No.101, Gn. Malang, Randuagung, Kec. Kebomas, Kabupaten Gresik, Jawa Timur 61121
hennydwi@umg.ac.id

ABSTRAK

Indonesia berada di peringkat kelima di dunia dalam jumlah penderita diabetes. Menurut laporan dari *International Diabetes Federation* (IDF), pada tahun 2021 terdapat 19,5 juta orang Indonesia berusia 20-79 tahun yang menderita diabetes. Selain itu, Indonesia juga menempati peringkat teratas di Asia Tenggara untuk jumlah penderita diabetes tipe satu. *Diabetes mellitus* adalah kondisi di mana tubuh tidak dapat memproduksi atau menggunakan insulin dengan baik. Insulin, hormon yang dihasilkan oleh pankreas, berperan penting dalam mengatur kadar gula darah dari makanan yang dikonsumsi agar dapat digunakan sebagai sumber energi oleh sel-sel tubuh. Penelitian ini bertujuan untuk mendeteksi dini penyakit diabetes mellitus dengan menggunakan algoritme klasifikasi *KNN (K-Nearest Neighbor)* dalam data mining. Metode yang digunakan dalam penelitian ini meliputi: Dataset, Preprocessing, Klasifikasi, Evaluasi, Prediksi, dan Penyimpanan Model. Dalam penelitian ini, telah dilakukan klasifikasi potensi penyakit diabetes mellitus tipe II pada pasien dengan menggunakan algoritme KNN K=1 dan K=3, serta menggunakan *Confusion Matrix* sebagai alat pengujinya. Hasilnya, akurasi sebesar 85% untuk KNN K=1 dan 75% untuk KNN K=3. Oleh karena itu, penelitian ini menunjukkan bahwa algoritme KNN dengan K=1 lebih efektif dibandingkan dengan KNN dengan K=3 dalam mengklasifikasi potensi penyakit diabetes mellitus tipe II berdasarkan dataset yang digunakan.

Kata kunci : *Diabetes Mellitus Tipe II, KNN (K-Nearest Neighbor), Klasifikasi, Machine Learning, Kesehatan.*

1. PENDAHULUAN

Indonesia berada di urutan kelima di dunia dalam hal jumlah penderita *diabetes*, dengan 19,5 juta penduduk berusia 20-79 tahun yang menderita *diabetes* pada tahun 2021 [1]. Selain itu, Indonesia memiliki jumlah pasien diabetes tipe satu tertinggi di Asia Tenggara.

Diabetes mellitus merupakan kondisi di mana tubuh mengalami masalah dalam memproduksi dan menggunakan insulin, yang penting untuk mengatur kadar glukosa darah. Penyakit ini dapat menyebabkan komplikasi jangka panjang yang berkembang perlahan dan semakin sulit dikendalikan seiring waktu, dengan risiko menyebabkan kecacatan dan bahkan mengancam jiwa [2].

Diabetes Mellitus dibagi menjadi dua tipe: (1) *Diabetes Mellitus* tipe I, yang disebabkan oleh produksi insulin yang sangat rendah atau tidak ada sama sekali oleh pankreas, biasanya terjadi pada anak-anak dan remaja yang tidak obesitas, dan (2) *Diabetes Mellitus* tipe II, yang terjadi karena pankreas tidak memproduksi cukup insulin atau karena sel-sel lemak dan otot tubuh menjadi resisten terhadap insulin. *Diabetes* tipe II umumnya dikaitkan dengan obesitas, tingkat aktivitas fisik, pola makan, dan faktor lainnya [3].

Menurut [4], Terdapat hubungan antara jenis kelamin dan *diabetes mellitus* tipe 2, di mana perempuan berisiko lebih tinggi dibandingkan laki-laki. Faktor-faktor penyebabnya mencakup tingginya kadar kolesterol serta perbedaan dalam aktivitas fisik dan gaya hidup. Perempuan memiliki persentase lemak tubuh sekitar 20-25%, sedangkan laki-laki sekitar 15-20%, sehingga risiko perempuan terkena *diabetes mellitus* 3-7 kali lebih besar. Menurut Damayanti, risiko ini juga dipengaruhi oleh peluang peningkatan indeks massa tubuh yang lebih tinggi pada perempuan [3].

Identifikasi potensi penyakit *diabetes mellitus* tipe II kini menjadi kebutuhan mendesak mengingat dampak serius dan fatal bagi kesehatan manusia. Untuk meningkatkan upaya pencegahan, penelitian harus difokuskan pada pengembangan sistem yang mampu mendeteksi dini penyakit ini, sehingga memungkinkan tindakan pencegahan lebih awal. Salah satu pendekatan yang mendukung tujuan ini adalah penerapan teknik penambangan data atau *data mining*.

Penambangan data atau *data mining* merupakan pendekatan yang efektif dalam mengatasi tantangan penanganan penyakit *diabetes mellitus* tipe II. Teknik-teknik penambangan data digunakan untuk mengidentifikasi pola dan informasi penting dari data pasien, seperti faktor risiko, gejala, dan catatan medis. Dengan demikian, dapat dilakukan klasifikasi untuk menentukan kemungkinan seseorang menderita *diabetes mellitus* tipe II, yang memungkinkan intervensi dini dan pengobatan yang tepat. Dalam penambangan data, salah satu teknik yang diterapkan adalah menganalisis data yang ada untuk membangun model, yang kemudian dapat digunakan untuk mengenali pola dalam data lain yang tidak terdapat dalam basis data yang tersedia.

Klasifikasi adalah salah satu teknik dalam *data mining* yang melibatkan evaluasi objek data untuk menempatkannya ke dalam salah satu dari beberapa

kategori yang tersedia. Dalam proses klasifikasi, terdapat dua tugas utama: pembangunan model sebagai prototipe yang disimpan dalam memori, dan penggunaan model tersebut untuk mengenali, mengklasifikasikan, atau memprediksi objek data lainnya guna menentukan kategorinya berdasarkan model yang telah disimpan [5].

Algoritme *K-Nearest Neighbor (KNN)* mampu menghasilkan hasil klasifikasi yang efektif ketika diterapkan pada dataset yang besar [6]. Sebuah penelitian yang menggunakan Algoritme *KNN* dilakukan oleh [7] dengan judul "Perbandingan Kinerja Algoritme *Gaussian Naïve Bayes* dan *K-Nearest Neighbor (KNN)* Untuk Mengklasifikasi Penyakit *Hepatitis C Virus (HCV)*". Penelitian ini menghasilkan nilai akurasi sebesar 91,80%.

Penelitian ini bertujuan untuk mendeteksi secara dini keberadaan penyakit *diabetes mellitus* dengan menerapkan Algoritme klasifikasi *KNN* dalam *data mining*. Pengujian dilakukan menggunakan dataset yang diperoleh dari Kaggle untuk mengetahui performa Algoritme *KNN* terhadap dataset tersebut. Pengujian dilakukan menggunakan *Confusion Matrix* dengan menggunakan bahasa pemrograman *Python*. Hasil implementasi akan diterapkan pada suatu sistem yang akan dibangun menggunakan bahasa *Python* dan *framework Streamlit*. Selanjutnya, sistem tersebut akan di-hosting menggunakan *Streamlit Share*.

2. TINJAUAN PUSTAKA.

2.1. Data Mining

Data mining adalah suatu proses setengah otomatis yang memanfaatkan teknik-teknik matematika, statistika, *machine learning*, dan kecerdasan buatan untuk mengenali dan mengekstraksi pengetahuan yang relevan dari beragam basis data yang besar [8]. Proses ini melibatkan pencarian dan pengidentifikasian pengetahuan baru, pola, serta tren dari volume data besar yang disimpan dalam repositori atau tempat penyimpanan, dengan menerapkan teknik statistika dan matematika untuk menyaring informasi yang penting [9]. *Data mining*, juga dikenal sebagai "*knowledge discovery*" membantu dalam pengambilan keputusan dengan menggali informasi baru dari volume data yang besar. Teknik-teknik dalam *data mining*, seperti membangun model dari data dan mengidentifikasi pola-pola yang signifikan, digunakan untuk tujuan prediksi. Pengelompokan data bertujuan untuk menemukan pola yang umum, sementara deteksi anomali dalam data transaksi penting untuk mengambil tindakan yang sesuai. Semua ini bertujuan untuk mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan dapat tercapai.

2.2. Klasifikasi

Klasifikasi adalah proses untuk mencari model atau fungsi yang membedakan atau menjelaskan konsep atau kelas data tertentu. Tujuannya adalah untuk memprediksi kelas dari objek yang tidak diketahui berdasarkan pola atau karakteristik yang telah dianalisis dalam data tersebut [10]. Proses ini melibatkan pembuatan pengklasifikasi menggunakan data latih yang sudah diberi label kelas, dan mengevaluasi kinerjanya dengan menggunakan metrik akurasi untuk menilai kemampuan pengklasifikasi dalam mengidentifikasi kelas objek yang tidak diketahui [11]. Proses klasifikasi melibatkan pelatihan pada fungsi target yang menghubungkan setiap kelompok atribut dengan label kelas yang tersedia. Model yang dihasilkan digunakan untuk memperkirakan label kelas pada data baru yang tidak diketahui. Saat mengembangkan model, digunakan berbagai Algoritme pembelajaran seperti *Artificial Neural Network (ANN), Support Vector Machine Learning (SVM), Decision Tree, KNN (K-Nearest Neighbor) dan Naïve Bayes* [5].

2.3. Confusion Matrix

Matriks konfusi atau *Confusion Matrix* adalah tabel yang mencatat hasil dari proses klasifikasi. Pada klasifikasi masalah biner, matriks konfusi terdiri dari dua kelas, yaitu kelas 0 dan 1. Setiap sel dalam matriks mencatat jumlah data dari kelas i yang diprediksi masuk ke kelas j. Sebagai contoh, sel f11 menunjukkan jumlah data dari kelas 1 yang benar diprediksi masuk ke kelas 1, sedangkan sel f10 menunjukkan jumlah data dari kelas 1 yang diprediksi masuk ke kelas 0 [5].

Tabel 1. Matrik Konfusi untuk Klasifikasi dua kelas

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 0
Kelas asli (i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Untuk menghitung akurasi digunakan formula

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi benar}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (1)$$

Untuk menghitung laju error (kesalahan prediksi) digunakan formula

$$\text{Laju error} = \frac{\text{Jumlah data yang diprediksi salah}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (2)$$

Tujuan Algoritme klasifikasi adalah mencapai tingkat akurasi yang tinggi, yang sering kali berhasil memprediksi dengan tepat ketika diterapkan pada data latih. Namun, penilaian yang sebenarnya terjadi saat menggunakan data uji untuk mengukur kinerja Algoritme tersebut [5].

2.4. KNN (K-Nearest Neighbor)

Algoritme *K-Nearest Neighbor (KNN)* adalah teknik klasifikasi yang memproyeksikan kelas dari data yang belum diketahui berdasarkan pada kelas dari tetangga terdekatnya (K) [6]. Jarak antara tetangga sering kali diukur dengan menggunakan *Euclidean Distance*, metode yang sesuai untuk menilai seberapa dekat dua data tersebut [12]. Perhitungan jarak dapat menggunakan metode *Euclidean, Manhattan, dan*

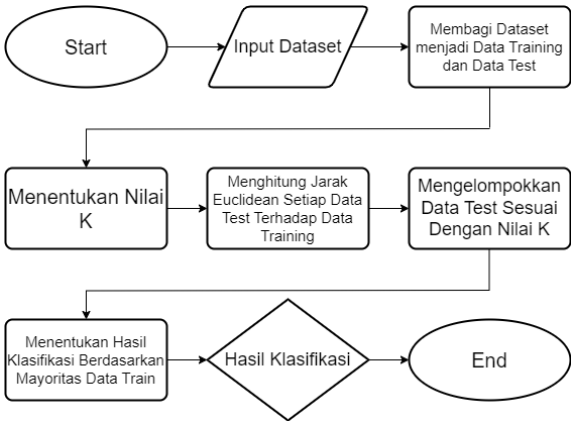
Minkowski untuk data dengan dimensi tinggi, sedangkan untuk data yang memiliki dimensi rendah lebih disarankan menggunakan metode *Chebyshev*. Dari ketiga metode tersebut, hanya *Manhattan* yang tidak melibatkan perhitungan kuadrat, sehingga memberikan hasil yang lebih baik pada K=1 dibandingkan dengan *Euclidean dan Minkowski*. [13]. Algoritme *KNN* dipilih karena kemampuannya untuk mengelompokkan data dengan memperkirakan jarak dari tetangga terdekat [14].

Adapun tahapan dari algoritme *KNN* dapat dijelaskan sebagai berikut :

a. Persiapan data training dan data testing
b. Melakukan normalisasi data menggunakan min – max, apabila ada nilai yang memiliki satuan ukuran yang berbeda. Berikut adalah rumusnya
c. $X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}}$ (3)
d. Menentukan nilai K
e. Menghitung jarak data testing ke setiap data training dengan rumsu *Euclidean Distance*. Berikut adalah rumusnya :

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

Berikut adalah alur *KNN*



Gambar 1. Alur KNN

Adapun keterangan dari gambar diatas adalah sebagai berikut :

a. Start : alur dimulai dengan memahami data yang digunakan
b. Input dataset : data input adalah dataset yang telah dikumpulkan dan dipersiapkan untuk analisis.
c. Pembagian dataset : data dibagi menjadi 2 yaitu data train dan data test.
d. Penentuan nilai K : pemilihan nilai K (jumlah tetangga terdekat) adalah Langkah penting dalam Algoritme *KNN*. Nilai K yang tepat akan mempengaruhi akurasi.
e. Perhitungan jarak *Euclidean* : setiap data test diukur menggunakan jarak *Euclidean* terhadap setiap data train.
f. Pengelompokan data test : berdasarkan nilai K yang telah ditentukan, data testing dikelompokkan dengan cara menentukan nilai K dari data training terdekat.
g. Penentuan hasil klasifikasi : dengan menggunakan mayoritas kelas dari tetangga terdekat, model *KNN* akan menentukan hasil klasifikasi untuk setiap data testing.
h. Hasil klasifikasi : setelah proses klasifikasi selesai, hasil klasifikasi akan muncul.
i. Selesai : dengan hasil klasifikasi yang diperoleh, alur *KNN* telah selesai.

2.5. Python

Python adalah sebuah bahasa pemrograman tingkat tinggi yang dibuat oleh *Guido Van Rossum* pada tahun 1991 dan telah mendapatkan popularitas yang signifikan dalam beberapa tahun terakhir. Bahasa ini terkenal karena memiliki sintaksis yang mudah dipahami dan berbagai pustaka yang lengkap. Salah satu kelebihan utamanya adalah dukungan yang kuat dari komunitas, karena bersifat *open source* [15]. *Python* menggabungkan kemampuan dan kapabilitas dengan sintaksis yang jelas, dengan pustaka standar yang komprehensif. Bahasa ini sering digunakan untuk mengembangkan aplikasi berbasis web [16].

2.6. Streamlit

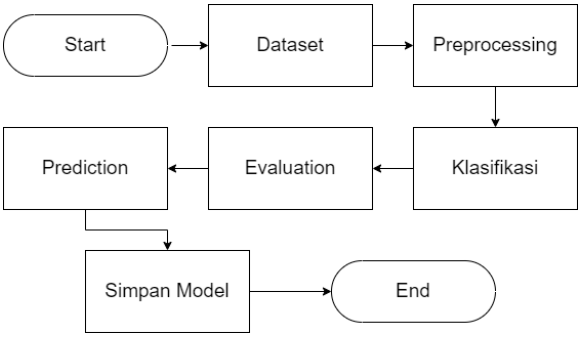
Streamlit adalah kerangka kerja sumber terbuka dari *Python* yang memungkinkan pembuatan aplikasi web dengan menggunakan bahasa pemrograman *Python*, terutama untuk menerapkan model dari bidang *machine learning* atau *data science* [17]. *Streamlit* digunakan oleh insinyur *machine learning* dan ilmuwan data untuk membuat alat-alat internal atau aplikasi web tanpa harus terjebak dalam bahasa pemrograman yang kompleks, sehingga memberikan kemudahan penggunaan [18]. Salah satu keunggulan utama *Streamlit* adalah pengembang tidak perlu memikirkan tata letak situs web menggunakan CSS, HTML, dan JavaScript, karena fitur-fitur tersebut telah tersedia melalui fungsi-fungsi terintegrasi, memungkinkan fokus pada logika aplikasi dan analisis data [19].

2.7. Penelitian Terdahulu

Terdapat penelitian yang pernah dilakukan sebelumnya oleh [7] dengan judul "Perbandingan Kinerja Algoritma *Gaussian Naïve Bayes* dan *K-Nearest Neighbor (KNN)* Untuk Mengklasifikasi Penyakit *Hepatitis C Virus (HCV)*" yang menghasilkan nilai akurasi sebesar 91,80% untuk algoritme *K-Nearest Neighbor (KNN)* dan 90,98% untuk algoritme *Gaussian Naïve Bayes*. Kemudian penelitian oleh [20] dengan judul "Perbandingan Metode *Naïve Bayes* dan *KNN (K-Nearest Neighbor)* dalam *klasifikasi* Penyakit *Diabetes*" menghasilkan nilai akurasi sebesar 90% untuk metode *klasifikasi KNN (K-Nearest Neighbor)* dan akurasi sebesar 78% untuk metode *klasifikasi Naïve Bayes.* Kemudian penelitian oleh [21] dengan judul "*Diabetes prediction using supervised machine learning*" yang membandingkan algoritme *KNN* dan *Naïve Bayes*, hasil yang didapat yaitu algoritme *Naïve Bayes* lebih

unggul dibandingkan dengan algoritme *KNN* dengan nilai akurasi dari algoritme *Naïve Bayes* sebesar 78,57% sedangkan nilai akurasi dari algoritme *KNN* sebesar 77,92%. Kemudian penelitian oleh [22] dengan judul "*Klasifikasi* Pengidap *Diabetes* Pada Perempuan Menggunakan Penggabungan Metode *Support Vector Machine (SVM)* dan *K-Nearest Neighbor (KNN)*" yang menghasilkan nilai akurasi sebesar 92%.

3. METODE PENELITIAN



Gambar 2. Alur Metode Penelitian

Metode yang digunakan untuk menyelesaikan penelitian ini adalah :

a. Start
   Memulai proses penelitian dengan merencanakan langkah-langkah yang akan diambil.
b. *Dataset*
   Pada penelitian ini, data didapatkan dari situs web Kaggle Diabetes_Dataset_With_18_Feature. Data set ini memiliki 17 variabel dan 1 label kelas dengan total 4.303 data yang terbagi menjadi dua yaitu 3000 data pasien tidak terkena diabetes dan 1303 data pasien terkena diabetes.
c. *Preprocessing*
   Dalam penelitian ini, tahapan preprocessing data menggunakan bahasa pemrograman *Python* sebagai alat analisis data. Pada tahap ini, variabel akan disortir menjadi 8 dan dilakukan pemisahan data antara atribut dan label. Selanjutnya, data akan dibagi menjadi data train dan data test, dengan total data train sebanyak 4.283 dan data test sebanyak 20.
d. *Klasifikasi*
   Dalam tahap ini, akan dilakukan pembangunan model *machine learning* menggunakan algoritme *klasifikasi Naïve Bayes*.
e. *Evaluation*
   Pada tahap ini, dataset akan diuji dan dievaluasi menggunakan *Confusion Matrix* serta akan dilakukan pengukuran tingkat akurasi.
f. *Prediction*
   Pada tahap ini akan dibuat model prediksi yang berfungsi untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpan.
g. Simpan model
   Pada tahap ini akan dilakukan penyimpanan model yang yang berguna sebagai mesin dalam website.
h. End
   Mengakhiri proses penelitian dengan memastikan semua langkah telah dijalankan dan hasilnya sesuai dengan tujuan penelitian.

4. HASIL DAN PEMBAHASAN

4.1. Perhitungan KNN K=1 dan K=3

a) Persiapan Data

Tabel 2. Data Train Perhitungan Manual KNN K=1 dan K=3

No	Age	Gender	BMI	SBP	DBP	FPG	Chol	FFPG	Diabetes
1	2	1	20,1	119	81	5,8	4,36	5,4	Negative
2	2	1	17,7	97	54	4,6	3,7	4,1	Negative
3	2	2	19,7	85	53	5,3	5,87	4,85	Negative
4	2	1	23,1	111	71	4,5	4,05	5,3	Negative
5	2	1	26,5	130	82	5,54	6,69	5,53	Negative
6	2	1	23,4	126	75	6,82	5	6,7	Positive
7	2	2	22,3	115	84	5,32	4,37	6,55	Positive
8	3	1	24,6	138	81	4,85	3,89	6,8	Positive
9	2	2	29	101	60	6,7	5,91	7	Positive
10	2	1	34,3	120	71	4,97	5,42	8,2	Positive
11	2	1	23,4	113	72	5,13	4,14	4,9	?

Tabel 3. Penjelasan Parame**ter**

	Nilai	Keterangan
Age	1	0 – 14 tahun
	2	15 – 65 tahun
	3	≥ 65 tahun
Gender	1	Female
	2	Male
BMI (Body Mass Index)	≤ 18,5	Berat Badan Kurang
	18,5 – 24,9	Berat Badan Normal

	Nilai	Keterangan
	25 – 29,9	Berat Badan Berlebihan
	≥ 30	Obesitas
SBP (Systolic Blood Pressure)	≤ 80	Tekanan Darah Rendah
	80 – 120	Normal
	120 – 139	prahipertensi
Tekanan yang dicatat selama kontraksi jantung	140 – 159	Tekanan Darah Tinggi (Hypertension Stadium 1)
	≥ 160	Tekanan Darah Tinggi (Hypertension Stadium 2)
	≥ 180	Krisis Tekanan Darah Tinggi
DBP (Diastolic Blood Pressure)	≤ 60	Tekanan Darah Rendah
	60 – 80	Normal
	80 – 89	prahipertensi
Tekanan yang dicatat selama relaksasi jantung	90 – 99	Tekanan Darah Tinggi (Hypertension Stadium 1)
	≥ 100	Tekanan Darah Tinggi (Hypertension Stadium 2)
	≥ 110	Krisis Tekanan Darah Tinggi
FPG (Fasting Plasma Glucose)	≤ 100 mg/dL (5.55 mmol/L)	Normal
	100 mg/dL – 125 mg/Dl (5.55 mmol/L – 6.9375 mmol/L)	Prediabetes
Kadar Glucose Puasa	≥ 126 mm/dL (7.003 mmol/L)	Diabetes
Cholesterol	≤ 200 mg/Dl (5.17 mmol/L)	Normal
	200 – 239 mg/dL (5.17 – 6.18 mmol/L)	Borderline Tinggi
	≥ 240 mg/Dl (6.21 mmol/L)	Tinggi
FFPG (Final Fasting Plasma Glucose)	≤ 100 mg/dL (5.56 mmol/L)	Normal
	100 mg/dL – 125 mg/dL (5.56 mmol/L – 6.94 mmol/L)	Prediabetes
Kadar Glucose Dua Jam Setelah Makan	≥ 126 mg/dL (7.00 mmol/L)	Diabetes

b) Normalisasi data
Langkah selanjutnya yaitu normalisasi dataset diatas menggunakan rumus min-max sebagai berikut :

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}}$$

Dalam contoh ini kita akan menghitung normalisasi data menggunakan min-max dari atribut SBP, maka formulanya adalah :

$$SBP = \frac{119 - 85}{138 - 85} = \frac{34}{53} = 0{,}641$$

Tabel 4. Hasil Normalisasi Data K=1 dan K=3

No	Age	Gender	BMI	SBP	DBP	FPG	Chol	FFPG	Diabetes
1	0,000	0	0,1446	0,6415	0,9032	0,5603	0,2207	0,3171	Negative
2	0,000	0	0,0000	0,2264	0,0323	0,0431	0,0000	0,0000	Negative
3	0,000	1	0,1205	0,0000	0,0000	0,3448	0,7258	0,1829	Negative
4	0,000	0	0,3253	0,4906	0,5806	0,0000	0,1171	0,2927	Negative
5	0,000	0	0,5301	0,8491	0,9355	0,4483	1,0000	0,3488	Negative
6	0,000	0	0,3434	0,7736	0,7097	1,0000	0,4348	0,6341	Positive
7	0,000	1	0,2771	0,5660	1,0000	0,3534	0,2241	0,5976	Positive
8	1,000	0	0,4157	1,0000	0,9032	0,1509	0,0635	0,6585	Positive
9	0,000	1	0,6807	0,3019	0,2258	0,9483	0,7391	0,7073	Positive
10	0,000	0	1,0000	0,6604	0,5806	0,2026	0,5753	1,0000	Positive
11	0,000	0	0,3434	0,5283	0,6129	0,2716	0,1472	0,1951	?

c) Tentukan Nilai K
Penentuan Nilai K ini tidak ada rumus pastinya. Namun apabila kelas berjumlah genap maka sebaiknya nilai K-nya ganjil, sebaliknya jika kelas berjumlah ganjil maka sebaiknya nilai K-nya genap. Seperti contoh diatas ada du akelas genap yaitu (Negative dan Positive), maka untuk nilai K nya akan ditentukan jadi K=1 dan K=3.

d) Hitung jarak antar datat test dan data train menggunakana rumus Euclidean sebagai berikut :

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$= \sqrt{((0–0)^2 + (0–0)^2 + (0{,}343 – 0{,}1446)^2 + (0{,}5283 – 0{,}6415)^2 + (0{,}6129 – 0{,}9032)^2 + (0{,}2716 – 0{,}5603)^2 + (0{,}1472 – 0{,}2207)^2 + (0{,}1951 – 0{,}3171)^2)}$

$= \sqrt{0{,}4902}$

Tabel 5. Hasil Perhitungan Jarak Euclidean K=1

No	Kelas asli	Jarak data test ke data train	Urutan Nilai Terkecil
1	Negative	0,4902	2
2	Negative	0,8112	3
3	Negative	1,4299	9
4	Negative	0,2949	1
5	Negative	1,0119	5
6	Positive	0,9358	4
7	Positive	1,1534	7
8	Positive	1,2443	8
9	Positive	1,5446	10
10	Positive	1,1338	6

Setelah itu pilih datu data dengan nilai urut terkecil, yaitu pada data negan nomor 4.

Tabel 6. Hasil Penentuan Data Menggunakan K=1

No	Kelas asli	Jarak data test ke data train	Urutan Nilai Terkecil
4	Negative	0,2949	1

Dari hasil diatas diketahui nilai K=1 berada pada nomor 4 dengan nilai jarak data test ke data train sebesar 0,2949 dan kelas asli "Negative". Sehingga pasien dengan Age = 2, Gender =1, BMI = 23,4 , SBP = 113, DBP = 72, FPG = 5,13 , Chol = 4,14 dan FFPG = 4,9 diklasifikasikan oleh algoritme KNN dan hasil kelasnya adalah **Negative**.

Tabel 7. Hasil 20 Data Testing KNN K=1

No	**Id**	**Age**	**Gender**	**BMI**	**SBP**	**DBP**	**FPG**	**Chol**	**FFPG**	**True Label**	**Predict Label**
1	2811	36	1	23.40	113	72	5.13	4.14	4.90	0	0
2	1818	36	2	22.30	113	72	4.23	4.72	5.23	0	0
3	1808	27	1	19.30	112	68	4.03	3.21	4.70	0	0
4	271	53	1	24.20	133	85	3.76	5.02	4.90	0	0
5	474	33	1	22.40	110	62	5.41	3.08	5.00	0	0
6	2972	53	2	24.24	155	77	4.57	5.01	5.58	0	0
7	1607	36	1	24.50	116	68	4.61	3.59	4.40	0	0
8	3390	71	1	25.20	143	91	5.00	4.00	7.10	1	1
9	979	29	2	17.80	113	72	4.01	4.52	4.00	0	0
10	1493	38	1	32.60	121	72	5.60	4.94	5.49	0	0
11	3957	66	1	25.90	100	61	4.92	4.10	8.00	1	1
12	3699	57	1	21.80	108	65	5.40	4.79	5.00	1	0
13	4102	47	1	35.00	125	69	5.04	4.97	6.70	1	0
14	552	31	1	24.20	125	81	4.80	4.58	6.10	0	0
15	1814	35	2	23.90	106	63	4.78	4.72	4.90	0	0
16	1428	80	1	24.00	177	85	5.27	5.67	5.90	0	0
17	3260	62	1	24.50	122	76	6.90	5.27	6.40	1	1
18	2441	36	2	20.70	112	61	4.70	4.10	4.63	0	0
19	3701	49	1	25.00	132	93	6.09	4.26	7.30	1	0
20	2522	31	1	16.50	123	69	4.30	3.78	4.87	0	0

Terdapat 3 error yaitu pada nomor 12, 13 dan 19, berikut adalah penjelasannya :

a. Nomor 12 True Label menunjukkan hasil Positive, sedangkan hasil Predict Label menunjukkan hasil Negative.
b. Nomor 13 True Label menunjukkan hasil Positive, sedangkan hasil Predict Label menunjukkan hasil Negative.
c. Nomor 19 True Label menunjukkan hasil Positive, sedangkan hasil Predict Label menunjukkan hasil Negative.

Tabel 8. Confusion Matrix Hasil Perhitungan KNN K=1

True Label	Negative	14	0
	Positive	3	3
		Negative	Positive
		Predict Label	

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3+14}{3+14+3+0} = \frac{17}{20} = 0{,}85$$

$$\text{Laju error} = \frac{FP+FN}{FP+FN+TP+TN} = \frac{3+0}{3+0+3+14} = \frac{3}{20} = 0{,}15$$

$$\text{Presisi} = \frac{TP}{TP+FP} = \frac{3}{3+3} = \frac{3}{6} = 0{,}5$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+0} = \frac{3}{3} = 1$$

Tabel 9. Hasil Perhitungan Jarak Eculidean K=3

No	Kelas asli	Jarak data test ke data train	Urutan Nilai Terkecil
1	Negative	0,4902	2
2	Negative	0,8112	3
3	Negative	1,4299	9
4	Negative	0,2949	1
5	Negative	1,0119	5
6	Positive	0,9358	4
7	Positive	1,1534	7
8	Positive	1,2443	8
9	Positive	1,5446	10
10	Positive	1,1338	6

Setelah itu pilih 3 data dengan nilai urutan terkecil, yaitu pada data dengan nomor 4, 1 dan 2.

Tabel 10. Hasil Penentuan Data Menggunakan K=3

No	Kelas asli	Jarak data test ke data train	Urutan Nilai Terkecil
4	Negative	0,2949	1
1	Negative	0,4902	2
2	Negative	0,8112	3

Dari hasil diatas diketahui kelas mayoritas dari 3 data diatas adalah "Negaive". Sehingga, Pasien dengan Age = 2, Gender = 1, BMI = 23,4 , SBP = 113, DBP = 72, FPG = 5,13 , Chol = 4,14 dan FFPG = 4,9 diklasifikasikan oleh algoritme KNN dan hasil kelasnya adalah **Negative.**

Tabel 11. Hasil 20 Data Testing KNN K=3

No	Id	Age	Gender	BMI	SBP	DBP	FPG	Chol	FFPG	True Label	Predict Label
1	2811	36	1	23.40	113	72	5.13	4.14	4.90	0	0
2	1818	36	2	22.30	113	72	4.23	4.72	5.23	0	0
3	1808	27	1	19.30	112	68	4.03	3.21	4.70	0	0
4	271	53	1	24.20	133	85	3.76	5.02	4.90	0	0
5	474	33	1	22.40	110	62	5.41	3.08	5.00	0	0
6	2972	53	2	24.24	155	77	4.57	5.01	5.58	0	0
7	1607	36	1	24.50	116	68	4.61	3.59	4.40	0	0
8	3390	71	1	25.20	143	91	5.00	4.00	7.10	1	0
9	979	29	2	17.80	113	72	4.01	4.52	4.00	0	0
10	1493	38	1	32.60	121	72	5.60	4.94	5.49	0	1
11	3957	66	1	25.90	100	61	4.92	4.10	8.00	1	1
12	3699	57	1	21.80	108	65	5.40	4.79	5.00	1	1
13	4102	47	1	35.00	125	69	5.04	4.97	6.70	1	0
14	552	31	1	24.20	125	81	4.80	4.58	6.10	0	0
15	1814	35	2	23.90	106	63	4.78	4.72	4.90	0	0
16	1428	80	1	24.00	177	85	5.27	5.67	5.90	0	1
17	3260	62	1	24.50	122	76	6.90	5.27	6.40	1	1
18	2441	36	2	20.70	112	61	4.70	4.10	4.63	0	0
19	3701	49	1	25.00	132	93	6.09	4.26	7.30	1	0
20	2522	31	1	16.50	123	69	4.30	3.78	4.87	0	0

Terdapat 5 error yaitu pada nomor 8, 10, 13, 16 dan 19, berikut adalah penjelasannya :

a. Nomor 8 True Label menunjukkan hasil Positive, sedangkan hasil Predict Label menunjukkan hasil Negative.
b. Nomor 10 True Label menunjukkan hasil Negative, sedangkan hasil Predict Label menunjukkan hasil Positive.
c. Nomor 13 True Label menunjukkan hasil Positive, sedangkan hasil Predict Label menunjukkan hasil Negative.
d. Nomor 16 True Label menunjukkan hasil Negative, sedangkan hasil Predict Label menunjukkan hasil Positive.
e. Nomor 19 True Label menunjukkan hasil Positive, sedangkan hasil Predict Label menunjukkan hasil Negative.

Tabel 12. Confusion Matrix Hasil Perhitungan KNN K=3

True Label	Negative	12	2
	Positive	3	3
		Negative	Positive
		Predict Label	

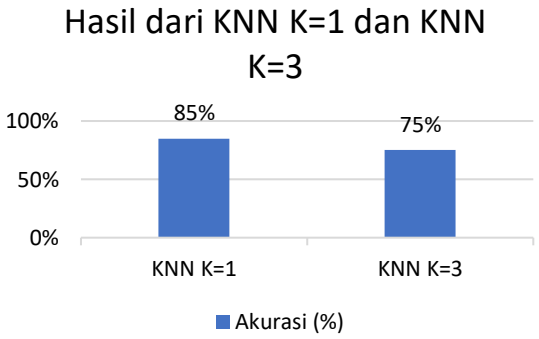
$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3+12}{3+12+3+2} = \frac{15}{20} = 0{,}75$$

$$\text{Laju error} = \frac{FP+FN}{FP+FN+TP+TN} = \frac{3+2}{3+2+3+14} = \frac{5}{20} = 0{,}25$$

$$\text{Presisi} = \frac{TP}{TP+FP} = \frac{3}{3+3} = \frac{3}{6} = 0{,}5$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+2} = \frac{3}{5} = 0{,}6$$

Berikut adalah hasil dari pengujian 20 data test *KNN K=1* dan *KNN K=3* pada sistem :



Gambar 3. Hasil Akurasi dari KNN K=1 dan K=3

Dibawah ini merupakan hasil dari pengujian data test pada sistem dan akan dihitung nilai akurasi, laju error, presisi dan recall menggunakan *Confusion Matrix*.

Tabel 13. Hasil dari pengujian data test

Metode	Akurasi	Laju Error	Presisi	Recall
KNN K=1	85%	15%	50%	100%
KNN K=3	75%	25%	50%	60%

Pada tabel di atas, algoritme *KNN* dengan *K=1* menghasilkan nilai akurasi sebesar 85%, yang menunjukkan bahwa model ini secara umum memberikan hasil klasifikasi yang baik. Nilai laju error

sebesar 15% menunjukkan bahwa ada ruang untuk perbaikan, meskipun angka ini dapat dianggap cukup baik tergantung pada konteks aplikasi.

4.2. Implementasi Sistem

Gambar dibawah ini merupakan tampilan awal sebelum melakukan pengklasifikasian potensi penyakit *diabetes mellitus* tipe II pada pasien menggunakan algoritme *KNN (K-Nearest Neighbor)*.

Gambar 4. Tampilan Awal Sebelum Klasifikasi

Gambar dibawah ini merupakan tampilan awal sesudah melakukan pengklasifikasian potensi penyakit *diabetes mellitus* tipe II pada pasien menggunakan algoritme *KNN (K-Nearest Neighbor)*. Yang mana setelah selesai proses input data pada parameter tersebut lalu tekan tombol "Klasifikasi Diabetes" maka akan muncul hasil klasifikasi, seperti pada contoh dibawah ini yang menunjukkan hasil bahwa "Pasien Tidak Terkena Diabetes". Dan juga data sudah otomatis terhubung ke database.

Gambar 5. Tampilan Awal Sesudah Klasifikasi

5. KESIMPULAN DAN SARAN

Pada penelitian ini telah dilakukan penelitian untuk mengklasifikasi potensi penyakit *diabetes mellitus* tipe II pada pasien menggunakan algoritme *KNN K=1 dan K=3* dengan menggunakan *Confusion Matrix* sebagai pengujiannya. Didapat akurasi sebesar 85% untuk *KNN K=1* dan 75% untuk *KNN K=3*. Maka pada penelitian ini menunjukkan bahwa algoritme *KNN K=1* efektif dalam mengklasifikasi potensi penyakit *diabetes mellitus* tipe II berdasarkan dataset yang digunakan. Adapun saran yang dapat diberikan kepada penelitian berikutnya agar menjadi lebih baik adalah dapat dilakukan perbandingan metode lain terhadap algoritme *KNN (K-Nearest Neighbor)*.

DAFTAR PUSTAKA

[1] M. Mahdi, "Penderita Diabetes Indonesia Terbesar Kelima di Dunia," DataIndonesia.id, 2 Februari 2022. [Online]. Available: https://dataindonesia.id/kesehatan/detail/penderita-diabetes-indonesia-terbesar-kelima-di-dunia. [Accessed 13 Juni 2024].

[2] D. N. Anisa and Jumanto, "KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITM NAIVE BAYES," *Jurnal Dinamika Informatika*, vol. 14, no. 1, pp. 33-42, 2022.

[3] S. Wahyuni and R. N. Alkaff, "DIABETES MELLITUS PADA PEREMPUAN USIA REPRODUKSI DI INDONESIA TAHUN 2007," *Jurnal Kesehatan Reproduksi*, vol. 3, no. 1, pp. 46-51, 2013

[4] Susilawati and R. Rahmawati, "Hubungan Usia, Jenis Kelamin dan Hipertensi dengan Kejadian Diabetes," *Jurnal ARKESMAS*, vol. 6, no. 1, p. 8, 2021

[5] E. Prasetyo, Data Mining - Konsep dan Aplikasi Menggunakan MATLAB, Yogyakarta: C.V ANDI OFFSET, 2012

[6] M. S. Mustafa and W. Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," *PROSIDING SEMINAR ILMIAH SISTEM INFORMASI DAN TEKNOLOGI INFORMASI*, vol. VIII, no. 1, pp. 1-10, 2019

[7] K. Y. Raharja, H. Oktavianto and R. Umilasari, "PERBANDINGAN KINERJA ALGORITMA GAUSSIAN NAIVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MENGKLASIFIKASI PENYAKIT HEPATITIS C VIRUS (HCV)," *Doctoral dissertation Universitas Muhammadiyah Jember*, 2021

[8] R. Rachman and R. N. Handayani, "Klasifikasi Algoritma Naive Bayes Dalam Memprediksi Tingkat Kelancaran Pembayaran Sewa Teras UMKM," *Jurnal Informatika*, vol. 8, no. 2, pp. 111-122, 2021

[9] M. H. Rifqo and A. Wijaya, "IMPLEMENTASI ALGORITMA NAIVE BAYES DALAM PENENTUAN PEMBERIAN KREDIT," *Jurnal Pseudocode*, vol. IV, no. 2, pp. 120-128, 2017.

[10] A. Setiawan, I. F. Astuti and A. H. Kridalaksana, "KLASIFIKASI DAN PENCARIAN BUKU REFERENSI AKADEMIK MENGGUNAKAN

METODE NAIVE BAYES CLASSIFIER (NBC) (STUDI KASUS: PERPUSTAKAAN DAERAH PROVINSI KALIMANTAN TIMUR)," *Jurnal Informatika Mulawarman*, vol. 10, no. 1, pp. 1-10, 2015.

[11] Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus : Data Mahasiswa Baru Perguruan Tinggi)," *Jurnal Ilmiah Ilmu Komputer*, vol. 2, no. 1, pp. 22-26, 2016.

[12] R. Amilia and E. Prasetyo, "KLASIFIKASI DIAGNOSA PENYAKIT DEMAM BERDARAH DENGUE PADA ANAK MENGGUNAKAN METODE K-NEAREST NEIGHBOR STUDI KASUS RUMAH SAKIT PKU MUHAMMADIYAH UJUNG PANGKAH GRESIK," *INDEXIA : Informatic and Computational Intelegent Journal*, vol. 2, no. 2, pp. 1 -10, 2020.

[13] F. Mandita and R. K. Pratama, "KLASIFIKASI PENERIMAAN TENAGA KERJA TERTUTUP MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (KNN)," *INDEXIA : Informatic and Computational Intelligent Journal*, vol. 6, no. 1, pp. 42-61, 2024

[14] K. Akmal, A. Faqih and F. Dikananda, "Perbandingan Metode Algoritma Naive Bayes dan K-Nearest Neighbors Untuk Klasifikasi Penyakit Stroke," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 1, pp. 470-477, 2023

[15] M. R. S. Alfarizi, M. Taufiqurrahman, M. Z. Al-Farish, G. Ardiansyah and M. Elgar, "PENGGUNAAN PYTHON SEBAGAI BAHASA PEMROGRAMAN UNTUK MACHINE LEARNING DAN DEEP LEARNING," *Jurnal Karimah Tauhid*, vol. 2, no. 1, pp. 1-6, 2023

[16] J. N. Semendawai, I. Febiola, B. Pamungkas and M. D. Rauliansyah, "Perancangan Aplikasi Otomatisasi Menggunakan Bahasa Pemrograman Python Pada Aktivitas Monitoring Pemakaian Data Harian Kartu Internet Of Things," *Jurnal Rekayasa Elektro Sriwijaya*, vol. 3, no. 1, pp. 193-198, 2021

[17] M. Ferdyandi, Y. Setiawan and F. A. Bachtiar, "Prediksi Potensi Penjualan Makanan Beku berdasarkan Ulasan Pengguna Shopee menggunakan Metode Decison Tree Algoritma C4.5 dan Random Forest (Studi Kasus Dapur Lilis)," *Jurnal Pengembangan Teknologi dan Ilmu Komputer*, vol. 6, no. 2, pp. 588-596, 2022

[18] M. Nofyantoro, D. K. Silalahi and N. Prihatiningrum, "Perancangan Sistem Prediksi Penggunaan Listrik Rumah Tangga Berbasis Website," *Jurnal e-Proceeding of Engineering*, vol. 9, no. 5, pp. 2253-2259, 2022

[19] A. Putranto, N. L. Azizah and I. R. I. Astutik, "Sistem Prediksi Penyakit Jantung Berbasis Web Menggunakan Metode SVM dan Framework Streamlit," *Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, vol. 4, no. 2, pp. 442-452, 2023

[20] I. L. F. Amien, W. Astuti and K. M. Lhaksamana, "Perbandingan Metode Naive Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes," *e-Proceeding of Engineering*, vol. 10, no. 2, pp. 1911-1920, 2023.

[21] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryaningrum and R. Yunanda, "Diabetes prediction using supervised learning," *Jurnal Procedia Computer Science*, vol. 216, pp. 21-30, 2023

[22] J. A. Putra and A. L. Akbar, "Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Pengabungan Metode Support Vectore Machine dan K-Nearest Neighbour," *Informatics Journal*, vol. 1, no. 2, pp. 47-52, 2016