PREDIKSI PENYAKIT DIABETES DENGAN METODE K-NEAREST NEIGHBOR (KNN) DAN SELEKSI FITUR INFORMATION GAIN

Devian, Puspita Nurul Sabrina, Agus Komarudin
Teknik Informatika, Universitas Jenderal Achmad Yani
Jl. Terusan Jend. Sudirman, Cibeber, Kec. Cimahi Sel., Kota Cimahi, Jawa Barat
devian20@if.unjani.ac.id

ABSTRAK

Diabetes mellitus, yang sering disebut diabetes, merupakan masalah kesehatan global yang signifikan dan mempengaruhi sekitar 422 juta orang di seluruh dunia. Penyakit ini lebih umum terjadi di negara-negara dengan tingkat ekonomi rendah dan menengah, menjadikannya tantangan kesehatan yang mendesak. Pencegahan dan deteksi dini diabetes sangat penting untuk mengurangi dampaknya terhadap kesehatan individu dan masyarakat. Penelitian ini mengeksplorasi penerapan metode K-Nearest Neighbors (K-NN) untuk prediksi diabetes dengan penekanan pada penggunaan Information Gain untuk seleksi fitur. Information Gain digunakan untuk mengidentifikasi dan memilih subset atribut paling informatif dari dataset kesehatan, bertujuan untuk meningkatkan akurasi model K-NN dan mengurangi dimensi data yang tidak relevan. Penelitian dilakukan dengan menguji model K-NN dalam empat skenario rasio dataset yang berbeda (90%:10%, 80%:20%, 70%:30%, 60%:40%). Hasil eksperimen menunjukkan bahwa penerapan Information Gain secara konsisten meningkatkan akurasi model K-NN dibandingkan dengan model tanpa seleksi fitur. Pada rasio 90%:10%, model K-NN tanpa Information Gain mencapai akurasi 69,11%, sementara model dengan Information Gain mencapai akurasi 70,96%, dengan Pencarian nilai K terbaik menunjukkan bahwa model K-NN dengan Information Gain mencapai akurasi tertinggi sebesar 72,93% pada nilai K = 17.

Kata kunci : *KNN, data mining, diabetes, dataset, Information Gain.*

1. PENDAHULUAN

Ada peningkatan kadar gula darah (glukosa) di atas nilai normal yang merupakan gejala penyakit kronis yang dikenal sebagai diabetes melitus (DM). DM terjadi ketika tubuh tidak dapat menyerap glukosa ke dalam sel dan menggunakannya sebagai energi, yang mengakibatkan penumpukkan gula tambahan dalam aliran tubuh. Diabetes mellitus yang tidak diagnosis dan tidak dikontrol dengan baik dapat mengakibatkan kerusakan pada banyak jaringan dan organ tubuh, termasuk mata, saraf, jantung, ginjal, dan jantung[1]. Untuk mendiagnosis DM, ilmu kedokteran sangat penting. Namun, teknologi seperti pengajaran mesin (ML) dengan algoritma K-Nearest Neighbors (KNN) dapat membantu dalam mengetahui perkembangan dan prediksi penyakit diabetes karena ketidakseimbangan antara dokter dan jumlah pasien diabetes meningkat[1].

Data mining dapat digunakan sebagai acuan untuk memprediksi dan mendiagnosis berbagai jenis penyakit, termasuk diabetes. Data mining melibatkan pengenalan pola dalam data dengan tujuan tertentu dan dapat digunakan untuk klasifikasi, asosiasi, estimasi, prediksi, dan pengelompokan [2]. *K-Nearest Neighbors* (KNN) adalah algoritma pembelajaran terawasi yang mengklasifikasikan data berdasarkan kedekatan dengan data lain yang sudah dikenal dan perhitungan kinerja metode yang diuji juga pemilihin fitur yang mana yaitu pada tahapan ini menggunakan information gain, memiliki tujuan untuk memilih beberapa fitur dalam data yang memiliki kepentingan yang cenderung sama dan memilih fitur dengan informasi yang tidak terlalu terkait dan berlebihan. Dengan mengurangi indeks dari data, algoritme klasifikasi dapat bekerja dengan lebih efisien dan lebih cepat, sehingga menghasilkan akurasi yang lebih baik[3].

Dalam penelitian yang dilakukan Salim F. Marko menggunakan teknik data mining untuk mengidentifikasi karakteristik pasien diabetes mellitus, kecenderungan, dan jenis penyakit di RSUP Dr. Sardjito di Yogyakarta. Dengan rancangan cross-sectional ini, penelitian deskriptif observasional dilakukan. Selanjutnya, aplikasi Weka digunakan untuk menganalisis data yang dikumpulkan. Hasil: Antara tahun 2011 dan 2016, ada 1.554 pasien dengan diabetes mellitus dengan tren penurunan. Sebagian besar populasi (27,86 persen) berusia antara 56 dan 63 tahun. Diabetes mellitus tipe 2 merupakan bentuk yang paling prevalen, dan komplikasi utamanya melibatkan neuropati, nefropati, dan hipertensi. Dalam rangka menganalisis rekam medis pasien, diterapkan teknik data mining dengan penerapan algoritma decision tree J48, yang memberikan tingkat akurasi sebesar 88.42%. Proses analisis ini menghasilkan serangkaian peraturan atau aturan (rules) yang signifikan dari data medis pasien[3].

Penelitian ini bertujuan melakukan analisis faktor-faktor yang mempengaruhi kemampuan prediksi KNN yang didukung oleh informatin gain, yang mana Information Gain adalah algoritma seleksi fitur yang sangat terkenal dan efektif dalam memilih fitur terbaik, terutama ketika bekerja dengan data dengan dimensi tinggi. Dalam penelitian sebelumnya, tujuan dari seleksi fitur adalah untuk mengidentifikasi fitur-fitur yang paling relevan dalam kumpulan data dan menghilangkan fitur-fitur lainnya yang dianggap tidak relevan atau redundant. Proses seleksi fitur ini

mengurangi dimensi data, sehingga algoritma dapat bekerja lebih cepet[4].

Dalam penelitian ini, Information Gain digunakan untuk memilih fitur selama tahap preprocessing dan algoritma K-NN digunakan dalam proses pelatihan untuk memprediksi penyakit diabetes, dengan harapan dapat mencapai akurasi yang lebih tinggi. Dimana dengan penambahaan Information gain diharapkan dalam pemilihan fitur yang tepat dapat memberikan pengaruh terhadap nilai akurasinya, karena metode pemilihan fitur Informasi Gain dapat diterapkan pada metode klasifikasi dan mempengaruhi tingkat akurasi metode tersebut[4].

2. TINJAUAN PUSTAKA

2.1. Data Mining

Proses pencarian pengetahuan dalam database (KDD) adalah proses pencarian informasi-informasi baru dan bernilai yang terdapat di dalam kumpulan data atau database. Dalam penelitian ini, data mining digunakan dengan tahapan-tahapan yang terdapat dalam KDD. Proses KDD dimulai dari menetapkan tujuan hingga evaluasi Data mining sering kali melibatkan beberapa tahapan, termasuk pemrosesan data, transformasi, dan penyaringan untuk mempersiapkan data. data mining dapat digunakan untuk mengidentifikasi faktor risiko utama, mengklasifikasikan pola-pola klinis, atau bahkan memprediksi kemungkinan onset penyakit pada individu berdasarkan riwayat kesehatan mereka[2].

2.2. Diabetes

Diabetes Melitus (DM) merupakan penyakit kronis dengan gejala adanya peningkatan kadar gula darah (*glukosa*) diatas nilai normal. DM terjadi ketika tubuh tidak dapat menyerap glukosa ke dalam sel dan menjadikannya sebagai energi yang menyebabkan penumpukkan gula ekstra dalam aliran tubuh. DM yang tidak diagnosis dan dikontrol dengan baik dapat menyebabkan kerusakan pada berbagai organ dan jaringan tubuh seperti seperti ginjal, jantung, mata, dan saraf [1].

2.3. Faktor Risiko Diabetes

Interaksi dari berbagai faktor penyebab yang mempengaruhi perubahan gaya hidup masyarakat dapat menyebabkan diabetes mellitus (DM). Contohnya termasuk kurangnya aktivitas fisik dan pola makan tradisional yang mengandung banyak karbohidrat dan serat dari sayuran dibandingkan dengan pola makan Barat yang mengandung makanan yang terlalu banyak mengandung (protein, lemak, gula, garam, dan sedikit serat). Inilah yang menyebabkan sebagian besar orang baru mengetahui bahwa mereka menderita diabetes mellitus (DM) setelah mengalami sakit yang sangat parah[6].

2.4. K-Nearest Neighbor

Konsep K-Nearest Neighbor, yang didasarkan pada nilai k dalam ruang fitur, digunakan untuk mengklasifikasikan objek. Metode ini memerlukan ukuran jarak untuk mengetahui seberapa dekat dua objek satu sama lain. Metode ini mencari tetangga yang paling dekat dan memilih kelas mayoritas dari cluster tersebut, sehingga data uji akan diklasifikasikan berdasarkan tetangga terdekatnya

Langkah-langkah yang diambil untuk mengklasifikasikan data menggunakan algoritma K-Nearest Neighbor adalah sebagai berikut[5]:

a. Menentukan nilai K
b. Menghitung nilai jarak antara data latih dan data uji
c. Mengelompokkan data berdasarkan perhitungan jarak
d. Mengelompokkan data berdasarkan nilai tetangga terdekat
e. Memilih nilai tetangga terdekat yang paling sering muncul sebagai prediksi data selanjutnya

Tujuan utama dari algoritma KNN adalah untuk mengelompokkan objek baru dengan cara yang sesuai berdasarkan atributnya dan training sample yang telah ada. Cara kerja KNN didasarkan pada konsep kedekatan spasial antara data yang ada dalam set latihan. Ini berarti bahwa ketika kita memiliki data baru yang akan diklasifikasikan, kita mencari data pelatihan yang paling mirip dengannya berdasarkan sejumlah atribut. Kemiripan ini sering diukur dengan menggunakan metode jarak, seperti jarak Euclidean atau jarak Minkowski, yang menghitung seberapa dekat atau jauhnya data baru dengan data pelatihan yang sudah ada[7].

2.5. Euclidian Distance

Rumus persamaan (1) digunakan untuk menghitung jarak geometris antara titik data dan tetangga terdekat.

$$d(P,Q) = \sqrt{\sum_{i=1}^{n} (P_i - Q_i)^2} \quad (1)$$

Dimana n adalah jumlah data latihan, P adalah masukkan data ke-i dari data uji, dan Q adalah masukkan data ke-i dari data latihan[5].

2.6. Information Gain

Information Gain merupakan konsep yang umum digunakan dalam bidang pembelajaran mesin dan analisis data, terutama dalam konteks pohon keputusan dan pemilihan fitur. Dalam algoritma pohon keputusan, seperti ID3 (Iterative Dichotomiser 3) atau C4.5, information Gain digunakan untuk menentukan fitur yang paling cocok untuk memisahkan data di setiap node. Ide dasarnya adalah memilih atribut yang memberikan informasi paling banyak tentang variabel target, sehingga menghasilkan pengurangan ketidakpastian atau entropi yang paling besar[8].

Namun, ekstraksi fitur sering kali menghasilkan terlalu banyak fitur, yang dapat meningkatkan beban kerja dan menurunkan akurasi. Oleh karena itu, penting untuk mengeliminasi fitur-fitur yang tidak relevan atau berlebihan agar dapat memperoleh fitur yang relevan dan optimal tanpa mengurangi akurasi[8].

Pengujian Information Gain dilakukan dengan menggunakan 5 hingga 10 fitur karena tujuan seleksi fitur adalah untuk mengidentifikasi kombinasi fitur yang paling relevan dan efektif dalam meningkatkan akurasi klasifikasi. Jika hanya menggunakan 1 fitur, kemungkinan besar informasi yang diperoleh tidak cukup untuk membedakan antara kelas-kelas yang ada.

Adapun terdapat penelitian sebelumnya yang menggunakan information gain dalam penelitian tersebut Dengan menggunakan lebih dari satu fitur, ini memungkinkan model untuk memiliki pemahaman yang lebih baik tentang data yang sedang dianalisis yang mana Ini menunjukkan bahwa lebih banyak fitur yang relevan dapat memberikan informasi tambahan yang penting untuk klasifikasi yang lebih akurat.[8] Lalu pada penelitian [9] dilakukan seleksi fitur information gain menggunakan 5 fitur dan memberikan hasil yang cukup baik pada hasil klasifikasi, dengan fokus pada fitur-fitur yang paling informatif[9].

Penggunaan pemilihan fitur berguna untuk mengurangi jumlah fitur, mempercepat proses penambangan data algoritma, dan meningkatkan kinerja penambangan data dengan menghilangkan hal-hal yang tidak relevan, berlebihan[4].

2.7. Confusion Matrix

Tabel yang disebut confusion matrix digunakan untuk menilai kinerja model klasifikasi. True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN) adalah empat komponen utama tabel ini. Untuk menilai model tersebut, berbagai metrik evaluasi seperti akurasi, presisi, dan recall dapat dihitung dengan menggunakan matriks confusion[11].

a. True Positive (TP): Kasus positif yang salah diprediksi sebagai negatif oleh model.
b. True Negative (TN): Kasus negatif yang diprediksi dengan benar oleh model.
c. False Positive (FP): Kasus negatif yang salah diprediksi sebagai positif oleh model.
d. False Negative (FN): Kasus positif yang salah diprediksi sebagai negatif oleh model.

Dengan menggunakan nilai-nilai dari confusion matrix, kita dapat menghitung metrik-metrik sebagai berikut :

a. Persamaan yang dipakai untuk menilai akurasi dapat ditemukan pada persamaan (2).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

b. Persamaan yang diterapkan untuk menentukan nilai precision dijelaskan dalam persamaan (3).

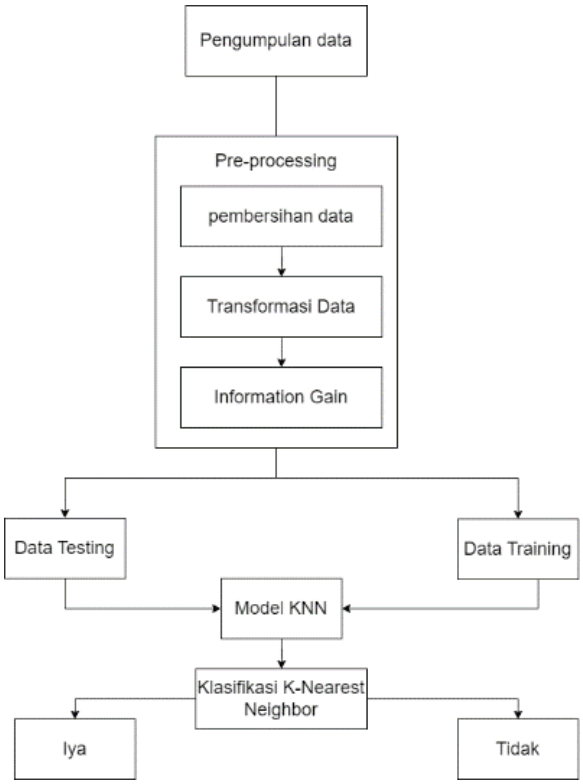
$$Precision = \frac{TP}{TP + FP} \quad (3)$$

c. Persamaan yang digunakan untuk menghitung nilai recall diuraikan dalam persamaan (4).

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

3. METODE PENELITIAN

Pada penelitian ini, terdapat beberapa tahap yang terdiri dari :



Gambar 1. Tahap Metode penelitian

3.1. Pengumpulan Data

Pada proses pengumpulan data ini bisa menggunakan beberapa metode, salah satunya yaitu studi literature,dan juga mengumpulkan dan mempelajari beberapa referensi dari berbagai macam sumber. Data yang menjadi subjek penelitian ini berasal dari dataset yang diunduh dari platform Kaggle dengan judul "Diabetes Dataset Prediction" dalam format Excel. Dataset ini terdiri dari 17 atribut yang mencakup Age, Sex, HighChol, CholCheck, BMI, Smoker, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, GenHlth, MentHlth, PhysHlth, DiffWalk, Stroke, HighBP, dan Diabetes. Setiap atribut ini memainkan peran penting dalam menyediakan informasi yang relevan untuk keperluan penelitian atau analisis yang dilakukan.

3.2. Pra-proses

Ini adalah tahap krusial dalam pembuatan model prediksi karena data dari basis data real-time sering kali tidak memadai, tidak seragam, dan tidak konsisten, yang dapat menghasilkan prediksi yang tidak akurat. Pada tahap ini, akan dilakukan proses preprocessing yang meliputi beberapa langkah, seperti pembersihan data, transformasi data, dan penggunaan Information Gain.

a. Pembersihan Data
   Proses ini melibatkan identifikasi dan penanganan nilai-nilai yang hilang, atau tidak sesuai.

b. Transformasi Data
Mengubah data ke dalam bentuk yang lebih sesuai untuk analisis dan model prediksi menggunakan KNN.[3] Dengan menggunakan Normalisasi atau Standarisasi yaitu menyesuaikan skala data untuk memastikan bahwa semua fitur memiliki dampak yang seimbang pada model. Lalu meggunakan Encoding Kategorikal Variables untuk mengubah variabel kategorikal menjadi bentuk numerik agar dapat digunakan oleh algoritma KNN[4].

c. Information Gain
Dalam metode pemilihan fitur berbasis Information Gain, berikut adalah tahapan yang dilakukan :

- Identifikasi Fitur
  Pada tahap ini, semua fitur atau atribut yang tersedia dalam dataset diidentifikasi. Ini melibatkan pengumpulan data terkait setiap fitur yang akan dipertimbangkan untuk pemilihan.
- Pemilihan Fitur
  Setelah fitur diidentifikasi, langkah selanjutnya adalah memilih fitur-fitur yang relevan untuk analisis. Fitur yang dianggap memiliki potensi untuk memberikan informasi penting terhadap target variabel dipilih untuk proses perhitungan lebih lanjut.
- Perhitungan Information Gain
  Pada tahap ini, Information Gain dihitung untuk setiap fitur. Information Gain mengukur seberapa banyak informasi yang diperoleh tentang target variabel dengan mengetahui nilai dari fitur tertentu. Perhitungan ini melibatkan analisis entropi sebelum dan setelah membagi data berdasarkan fitur tersebut.
- Pemilihan Fitur Terbaik
  Fitur-fitur Dipilih sebagai fitur terbaik karena memiliki nilai pendapatan informasi tertinggi. Fitur dengan Information Gain yang tinggi dianggap lebih informatif dan relevan untuk model klasifikasi atau prediksi yang akan dibangun.

3.3. Pemodelan KNN

Dalam metode KNN, pencarian nilai K melibatkan eksperimen untuk menemukan nilai optimal yang memaksimalkan akurasi prediksi berdasarkan data training. Sementara itu, dalam Information Gain, pencarian nilai K berkaitan dengan pemilihan fitur yang memberikan informasi paling signifikan terhadap variabel target, seperti prediksi penyakit diabetes. Model algoritma K-Nearest Neighbor dikembangkan dengan menggunakan ukuran jarak Euclidean Distance dan berbagai nilai K yang sering dipilih, yaitu K=1, 3, 5, 7, 11, 13, 15, 17, dan 19[12]. Selanjutnya, nilai K yang memberikan performa terbaik akan dipilih. Nilai K pada K-NN mengklasifikasikan data berdasarkan nilai K yang telah ditentukan sebelumnya. Idealnya, nilai K berupa bilangan ganjil selama proses klasifikasi, tetapi jika K-NN digunakan untuk melakukan prediksi, nilai K dapat berupa angka genap atau ganjil[13].

3.4. Evaluasi

Data yang telah diolah kemudian diukur nilai performancenya. Untuk mengukur nilai performance, dibutuhkan model untuk mengukur tingkat akurasi dari Nilai Performance. Dalam melakukan penilaian performance, semakin tinggi nilai yang didapat maka semakin bagus pula performance model tersebut untuk digunakan[14]. Model pengukuran yang akan digunakan adalah Accuracy. Ini sering diungkapkan sebagai persentase dan mewakili rasio instansi yang diprediksi dengan benar terhadap total jumlah instansi dalam dataset.

4. HASIL DAN PEMBAHASAN

Hasil dari penelitian "Prediksi Penyakit Diabetes Dengan Metode K-Nearest Neighbor (Knn) dan Seleksi Fitur Informasi Gain" akan dijelaskan secara rinci di bagian ini.

4.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah Diabetes, Hypertension, and Stroke Prediction yang diperoleh dari platform Kaggle. Dataset ini terdiri dari 70.693 record data dan 18 atribut. Berikut adalah deskripsi dari masing-masing atribut yang terdapat dalam dataset :

Tabel 1. Deskripsi atribut

No	Atribut	Deskripsi
1.	Age	Usia
2.	Sex	Jenis Kelamin
3.	Highchol	Tingkat Kolesterol
4.	Cholchek	Pemeriksaan Kolesterol
5.	BMI	Indeks Masa Tubuh
6.	Smoker	Perokok
7.	HeartDiseaseorAttack	Penyakit Jantung
8.	PhysActivity	Aktifitas Fisik
9.	Fruits	Konsumsi Buah
10.	Veggies	Konsumsi Sayuran
11.	HvyAlcoholConsump	Konsumsi Alkohol
12.	GenHlth	Kesehatan Umum
13.	MentHlth	Kesehatan Umum
14.	PhysHlth	Penyakit Fisik
15.	DiffWalk	Kesulitan Berjalan
16.	Stroke	Penyakit Stroke
17.	HighBP	Tekanan Darah Tinggi
18.	Diabetes	Penyakit Diabetes

Tabel 2 menunjukkan contoh data riwayat penyakit diabetes yang digunakan dalam penelitian ini, yang mencakup informasi tentang kondisi medis dan faktor risiko lainnya.

Tabel 2. Data penyakit diabetes

No	Age	Sex	Hidhchol	…	Diabetes
1	4.0	1.0	0.0	…	0.0
2	12.0	1.0	1.0	…	0.0
3	13.0	0.0	0.0	…	0.0
4	11.0	0.0	0.0	…	0.0
5	8.0	1.0	1.0	…	0.0

No	Age	Sex	Hidhchol	…	Diabetes
…	…	…	…	…	…
70689	6.0	0.0	1.0	…	1.0
70690	10.0	1.0	1.0	…	1.0
70691	13.0	0.0	1.0	…	1.0
70692	11.0	0.0	1.0	…	1.0
70693	9.0	0.0	1.0	…	1.0

4.2. Pra-proses

Pada tahap ini bertujuan untuk mempersiapkan data sebelum dilakukan proses klasifikasi.

4.3. Pembersihan Data

Pada tahap preprocessing, data cleaning dilakukan untuk menangani masalah seperti missing value. Namun, dalam dataset ini tidak ditemukan missing value, seperti yang ditunjukkan pada Gambar 2 di bawah ini.

```
Missing Value :
Age                    0
Sex                    0
HighChol               0
CholCheck              0
BMI                    0
Smoker                 0
HeartDiseaseorAttack   0
PhysActivity           0
Fruits                 0
Veggies                0
HvyAlcoholConsump      0
GenHlth                0
MentHlth               0
PhysHlth               0
DiffWalk               0
Stroke                 0
HighBP                 0
Diabetes               0
dtype: int64
```

Gambar 2. Cek data yang hilang

4.4. Transformasi Data

Data perlu dimodifikasi atau disusun ulang ke dalam format yang sesuai agar dapat diproses oleh teknik data mining tertentu, karena beberapa teknik memerlukan format spesifik sebelum data dapat digunakan[10]. Dalam penelitian ini, data yang digunakan sudah berada dalam format variabel numerik, sehingga tidak memerlukan transformasi tambahan sebelum penerapan metode KNN dan Information Gain.

4.5. Seleksi Fitur dengan Information Gain

Untuk mengidentifikasi dan memilih fitur yang paling informatif guna meningkatkan akurasi dan efisiensi model dengan mengurangi dimensi data dan fokus pada fitur yang paling berpengaruh terhadap hasil prediksi, penelitian ini menggunakan metode Information Gain untuk membantu dalam pemilihan fitur. Tahapan preprocessing dengan menggunakan Information Gain biasanya dilakukan sebelum pelatihan model. Langkah-langkah umumnya adalah sebagai berikut :

a. Hitung Information Gain untuk setiap fitur dalam dataset.
b. Urutkan fitur-fitur berdasarkan nilai Information Gain-nya.
c. Pilih sejumlah fitur dengan Information Gain tertinggi untuk digunakan dalam pelatihan model. Dalam skenario ini, fitur yang dipilih akan mencakup 5 hingga 10 fitur dengan Information Gain tertinggi untuk memastikan bahwa model menggunakan informasi yang paling relevan.

Dengan demikian, Information Gain merupakan alat yang penting dalam tahap preprocessing untuk memastikan bahwa hanya fitur-fitur yang paling relevan yang digunakan dalam model prediksi, seperti pada penelitian prediksi penyakit diabetes menggunakan metode KNN.

4.6. Pemodelan KNN

Setelah kita melakukan pengujian fitur dengan information gain maka di dapatkan lah 5 fitur terpilih oleh information gain yaitu (GenHlth, HighBp, BMI, Age, dan HighChol) yang mana di sini dengan menggunakan perbandingan data atau rasio 90:10 di dapatkan hasil akurasi dengan menggunakan 5 fitur sebesar 70.96 %. Lalu kita melakukan pencarian nilai K terbaik di mulai dari k = 3 sampai K = 19 dengan menggunakan rumus Eucludian Distance untuk melakukan pencarian nilai k tetangga terdekat, yang mana bertujuan agar mendapatkan nilai K tetangga terdekat dengan nilai optimal.

Tabel 3. Hasil pencarian nilai K

Nilai K	Akurasi
K=3	0.7096
K=4	0.6943
K=5	0.7171
K=6	0.6992
K=7	0.7168
K=8	0.7045
K=9	0.7158
K=10	0.7175
K=11	0.7248
K=12	0.7232
K=13	0.7158
K=14	0.7264
K=15	0.7249
K=16	0.7233
K=17	0.7293
K=18	0.7235
K=19	0.7243

Berdasarkan Hasil pencarian nilai k terbaik dapat di lihat pada Tabel 2 yang mana nilai Akurasi terbaik di dapatkan pada nilai k17 dengan hasil akurasi sebesar 0.7293.

4.7. Confusion Matrix

Tujuan utama dari confusion matrix adalah untuk memberikan gambaran komprehensif tentang kinerja model klasifikasi dengan membandingkan hasil prediksi model dengan nilai sebenarnya dari dataset.

Tabel 4.Confusion Matrix

	Prediksi Positif	Prediksi Negatisf
Aktual Positif	*TP*	*FN*
Aktual Negatif	*FP*	*TN*

Rumus untuk berbagai metrik kinerja yang diturunkan dari matriks kebingungan adalah sebagai berikut

a. Accuracy

Nilai Accuracy mengukur rasio antara jumlah data yang terklasifikasi dengan benar dan total data yang ada. Akurasi dapat dihitung menggunakan persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} * 100\% \quad (5)$$

b. Precision

Nilai Precision menunjukkan proporsi data kategori positif yang diklasifikasikan dengan benar dibandingkan dengan seluruh data yang dikategorikan sebagai positif. Presisi dapat dihitung menggunakan persamaan berikut:

$$Precision = \frac{TP}{TP + FP} * 100\% \quad (6)$$

c. Recall

Recall mengukur persentase data kategori positif yang berhasil diklasifikasikan dengan benar oleh sistem. Nilai recall dapat dihitung menggunakan persamaan berikut:

$$Recall = \frac{TP}{TP + FN} * 100\% \quad (7)$$

4.8. Evaluasi

Pada evaluasi dan pengujian, digunakan rasio 90:10, di mana 90% data digunakan untuk training dan 10% untuk testing. Model klasifikasi yang diterapkan adalah K-Nearest Neighbors (KNN) serta K-Nearest Neighbors (KNN) dengan pendekatan Information Gain. Dalam pendekatan ini, nilai K yang diambil adalah K=17, dan 5 fitur teratas dipilih berdasarkan Information Gain. Tujuan penggunaan Information Gain adalah untuk memilih fitur yang tepat, sehingga dapat memberikan pengaruh positif terhadap akurasi model. Evaluasi model diperoleh sebagai berikut :

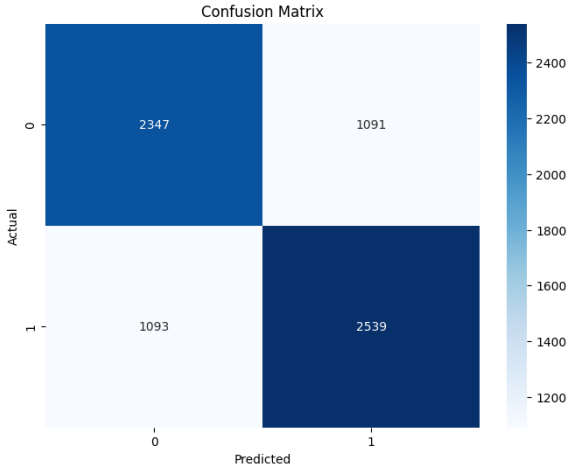
a. Perbandingan Model K-NN dengan Information Gain dan tanpa Information Gain

Tabel 5. Akurasi perbandingan

Model	Akurasi
K-Nearest Neighbors (KNN)	69.11%
K-Nearest Neighbors (KNN) & Information Gain	72.93%

b. Confusion Matrix K-NN tanpa Information Gain

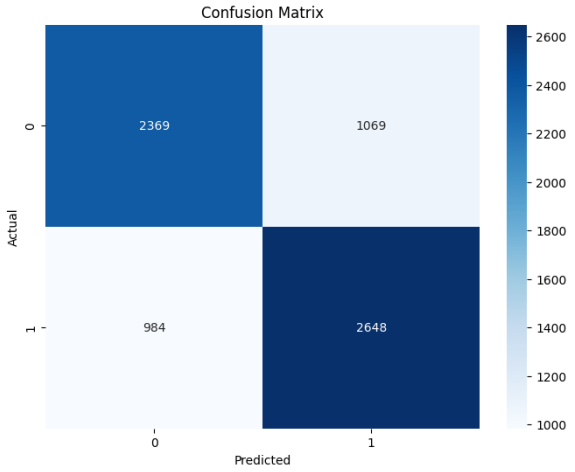
Dapat dilihat Pada Gambar 3 hasil Confusion matrix menunjukan untuk Confusion Matrix K-NN tanpa Information Gain, nilai TP=2539, lalu TN=2347, FP=1091, dan FN=1093.



Gambar 3. Confusion matrix KNN tanpa IG

c. Confusion Matrix K-NN dengan Information Gain

Dapat dilihat Pada Gambar 3 hasil Confusion matrix menunjukan untuk Confusion Matrix K-NN dengan Information Gain, nilai TP=2648, lalu TN=2369, FP=1069, dan FN=984.



Gambar 4. Confusion matrix KNN dengan IG

5. KESIMPULAN DAN SARAN

Penelitian ini menggunakan dataset Diabetes yang mencakup 17 atribut, termasuk Age, Sex, HighChol, CholCheck, BMI, Smoker, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, GenHlth, MentHlth, PhysHlth, DiffWalk, Stroke, HighBP, dan 70693 record, yang diambil dari platform Kaggle. Penelitian ini bertujuan untuk mengembangkan model prediksi diabetes menggunakan metode K-NN dan seleksi fitur Information Gain. Hasil penelitian menunjukkan bahwa model K-NN yang dilatih dengan Information Gain menunjukkan akurasi yang akurat daripada dengan model K-NN yang tidak menggunakan Information Gain. Model tanpa Information Gain mencapai akurasi sebesar 0,6922, sementara model yang menerapkan Information Gain memperoleh akurasi yang lebih tinggi, yaitu 0,7096, dengan perbandingan data 90% untuk pelatihan dan 10%

untuk pengujian. Selanjutnya, dilakukan pencarian nilai K terbaik dengan menggunakan data 90% untuk data latih dan 10% untuk data uji dan 5 fiutr terpilih oleh information gain. Hasil pencarian menunjukkan bahwa nilai K terbaik adalah K = 17, dengan akurasi tertinggi mencapai 0,7293. Ini menunjukkan bahwa model K-NN dengan nilai K = 17 memberikan performa yang optimal dalam prediksi diabetes pada dataset ini. Secara keseluruhan, penelitian ini mengindikasikan bahwa seleksi fitur dengan Information Gain dapat meningkatkan akurasi model K-NN, dan pemilihan nilai K yang tepat juga berpengaruh signifikan terhadap kinerja model dalam klasifikasi diabetes. Berdasarkan temuan penelitian yang telah dilakukan, rekomendasi untuk penelitian lanjutan adalah melakukan eksplorasi tambahan terhadap parameter KNN seperti jumlah tetangga (k), metrik jarak yang digunakan, dan teknik bobot jika diperlukan. Melakukan analisis lebih lanjut terhadap kualitas dan relevansi atribut dalam dataset. Penambahan atribut baru atau penggabungan fitur dapat meningkatkan keakuratan prediksi. Selain Information Gain, Anda bisa mempertimbangkan penggunaan metode seleksi fitur lain seperti Gain Ratio, Chi-square, atau ReliefF. Membandingkan kinerja model dengan berbagai metode seleksi fitur dapat memberikan wawasan yang lebih dalam tentang atribut mana yang paling relevan untuk prediksi diabetes.

DAFTAR PUSTAKA

[1] N. Azizah, M. R. Firdaus, R. Suyaningsih, and F. Indrayatna, “Penerapan Algoritma Klasifikasi K-Nearest Neighbor pada Penyakit Diabetes,” 2023, [Online]. Available: http://prosiding.snsa.statistics.unpad.ac.id.

[2] Zai, Charles. "Implementasi Data Mining Sebagai Pengolahan Data." Jurnal Portal Data 2.3 (2022).”

[3] Hasibuan, M. R., & Marji, M. (2020). Pemilihan Fitur dengan Information Gain untuk Klasifikasi Penyakit Gagal Ginjal menggunakan Metode Modified K-Nearest Neighbor (MKNN). Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 3(11), 10435–10443. Diambil dari https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/6691.

[4] Arifin, M. (2015). IG-KNN Untuk Prediksi Customer Churn Telekomunikasi. Jurnal Simetris, 6, 1-10.

[5] D. Cahyanti, A. Rahmayani, and S. Ainy Husniar, “Indonesian Journal of Data and Science Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara,” vol. 1, no. 2, pp. 39–43, 2020.

[6] Azis, Wa Ode Azfari, and Rifandi Saputra. "Faktor resiko kejadian diabetes mellitus pada lansia." Poltekita: Jurnal Ilmu Kesehatan 15.4 (2022): 346-354

[7] P. B. Utomo, E. Utami, and S. Raharjo, “Implementasi Metode K-Nearest Neighbor dan Regresi Linear dalam Prediksi Harga Emas,” J. Inf. Interaktif, vol. 4, no. 3, pp. 155–159, 2019.

[8] F. Y. Nabella, Y. A. Sari, and R. C. Wihandika, “Seleksi Fitur Information Gain Pada Klasifikasi Citra Makanan Menggunakan Hue Saturation Value dan Gray Level Co-Occurrence Matrix,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 3, no. 2, 2019.

[9] Sari, Betha Nurina. "Implementasi teknik seleksi fitur information gain pada algoritma klasifikasi machine learning untuk prediksi performa akademik siswa." Semnasteknomedia Online 4.1 (2016): 2-9

[10] S. Widaningsih and S. Yusuf, “Penerapan Data Mining Untuk Memprediksi Siswa Berprestasi Dengan Menggunakan Algoritma K Nearest Neighbor,” Jurnal Teknik Informatika dan Sistem Informasi, vol. 9, no. 3, 2022, [Online]. Available: http://jurnal.mdp.ac.id.

[11] B. P. Pratiwi, “Pengukuran Kinerja Sistem Kualitas Udara,” J. Inform. UPGRIS, vol. 6, no. 2, pp. 66–75, 2020.

[12] Riza Adrianti Supono, & Muhammad Azis Suprayogi. (2021). Perbandingan Metode TF-ABS dan TF-IDF Pada Klasifikasi Teks Helpdesk Menggunakan K-Nearest Neighbor . Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(5), 911 - 918. https://doi.org/10.29207/resti.v5i5.3403.

[13] Khairi, Ahmad, et al. "Implementasi K-Nearest Neighbor (KNN) untuk Klasifikasi Masyarakat Pra Sejahtera Desa Sapikerep Kecamatan Sukapura." TRILOGI: Jurnal Ilmu Teknologi, Kesehatan, dan Humaniora 2.3 (2021): 319-323.

[14] B. A. Wijaya and M. Nuryatno, “Pengaruh Environmental Performance Dan Environmental Disclosure Terhadap Economic Performance,” J. Informasi, Perpajakan, Akuntansi, Dan Keuang. Publik, vol. 9, no. 2, pp. 141–152, 2019, doi: 10.25105/jipak.v9i2.4530.