PREDIKSI PENYAKIT STROKE MENGGUNAKAN METODE K-NEAREST NEIGBORS DAN INFORMATION GAIN

Fabian Nabiel Syahreza, Puspita Nurul Sabrina, Edvin Ramadhan
Teknik Informatika, Universitas Jendral Achmad Yani
Jl. Terusan Jend. Sudirman, Cimahi, Jawa Barat, Kota Cimahi, Jawa Barat 40525
fabiannabiel929@gmail.com

ABSTRAK

Stroke adalah penyebab utama kecacatan global, dengan lebih dari 12 juta kasus baru setiap tahunnya menurut WHO. Kondisi ini ditandai oleh defisit neurologis yang berkembang cepat dan dapat berlangsung lama atau menyebabkan kematian. Untuk memprediksi risiko stroke secara akurat, penelitian ini mengembangkan model prediksi menggunakan metode K-Nearest Neighbors (KNN) yang digabungkan dengan Information Gain (IG). KNN, yang mengklasifikasikan data berdasarkan kesamaan fitur, dioptimalkan melalui seleksi fitur menggunakan IG untuk memilih fitur yang paling relevan. Hasil evaluasi menunjukkan bahwa model KNN yang mengintegrasikan IG mencapai akurasi tertinggi 99,85% dengan 5 fitur terpilih (hypertension, avg_glucose_level_category, heart_disease, ever_married, work_type) dan nilai k=25, dibandingkan dengan akurasi 88,46% dari model KNN tanpa IG. Akurasi tinggi ini didukung oleh precision 99,71%, recall 100%, dan F1-Score 99,86%, yang menegaskan efektivitas IG dalam meningkatkan performa model KNN. Penelitian ini menunjukkan bahwa integrasi IG secara signifikan meningkatkan akurasi prediksi risiko stroke dengan memilih fitur-fitur yang paling relevan.

Kata kunci *: Stroke, K-Nearest Neighbors (KNN), Information Gain, pemilihan fitur, prediksi penyakit, data mining*

1. PENDAHULUAN

Stroke menjadi penyebab utama kecacatan di seluruh dunia dan setiap tahunnya lebih dari 12 juta orang menderita stroke. Menurut WHO *(World Health Organization)*, stroke merupakan suatu keadaan di mana ditemukan tanda-tanda klinis yang berkembang cepat berupa deficit neurologik fokal dan global yang dapat memberat dan berlangsung lama selama 24 jam atau lebih dan atau dapat menyebabkan kematian [1]. Dalam konteks ini, fokus pada penelitian sebelumnya tertuju pada penerapan data mining dengan mengadopsi metode K-Nearest Neighbors (KNN) untuk melakukan prediksi risiko penyakit stroke pada tingkat individu. Metode KNN, yang didasarkan pada prinsip kesamaan antar data, diharapkan dapat memberikan kontribusi dalam mengenali faktor-faktor risiko yang relevan dan mendukung inisiatif pencegahan penyakit stroke [2].

Penelitian ini mengembangkan model prediksi risiko stroke pada tingkat individu dengan menggabungkan metode KNN dan Information Gain (IG), menyoroti potensi peningkatan akurasi prediksi melalui seleksi fitur yang lebih efektif. Penelitian terkait menunjukkan bahwa penerapan KNN dengan parameter k yang tepat dapat memberikan hasil yang baik dalam mendeteksi penyakit, seperti pada kasus deteksi penyakit jantung dengan akurasi 90% [3]. Selain itu, penelitian tentang seleksi fitur menggunakan Information Gain menunjukkan keefektifan dalam memilih fitur yang relevan dan mengatasi tantangan data yang kompleks. Dalam studi sebelumnya, penggunaan KNN dengan IG telah terbukti memberikan akurasi yang lebih tinggi dibandingkan dengan KNN saja, memperkuat argumen bahwa pendekatan ini dapat meningkatkan kemampuan prediksi risiko stroke [6].

Pada penelitian sebelumnya, dijelaskan bahwa atribut yang paling berpengaruh untuk memprediksi penyakit stroke menggunakan metode KNN adalah jenis kelamin, umur, hipertensi, riwayat penyakit jantung, status menikah, tipe pekerjaan, tipe tempat tinggal, rerata kadar glukosa, BMI, dan status merokok [2]. Sedangkan menurut *World Stroke Organization* (WSO), menyebutkan bahwa faktor utama dari penyatkit stroke yaitu Hipertensi (tekanan darah tinggi) [4]. Sehingga dengan banyak nya atribut yang dapat menyebakan resiko stroke, maka dapat menyebabkan juga kurangnya efisiensi dalam performa untuk memprediksi penyakit stroke. Maka penelitian ini akan menggunakan pendekatan Information gain untuk mengeliminasi atribut yang di intregasikan dengan Metode KNN sehingga akan meningkatkan akurasi prediksi penyakit stroke.

Penelitian ini mengembangkan pemodelan prediksi risiko penyakit stroke pada tingkat individu dengan menggabungkan metode K-Nearest Neighbors (KNN) dan Information Gain (IG). Penelitian ini menyoroti potensi peningkatan seleksi fitur dengan menggabungkannya dengan metode Information Gain pada dataset penyakit stroke. Melalui pendekatan ini, dapat ditemukan faktor-faktor risiko yang lebih relevan dan meningkatkan akurasi prediksi.

2. TINJAUAN PUSTAKA

2.1. Stroke

Stroke adalah kondisi medis yang terjadi ketika aliran darah ke area tertentu di otak terhenti atau berkurang, menyebabkan kerusakan pada sel-sel otak

dan fungsi tubuh yang terkait dengan area tersebut. Penyebab umum stroke melibatkan pembekuan atau pecahnya pembuluh darah di dalam otak. Terdapat dua jenis utama stroke, yaitu stroke *iskemik*, yang disebabkan oleh penyumbatan pembuluh darah otak, dan stroke *hemoragik*, yang disebabkan oleh perdarahan di dalam otak. Stroke merupakan penyebab ketiga kecacatan di dunia dan memiliki tingkat ketergantungan yang signifikan pada populasi usia di atas 60 tahun. Faktor risiko yang berhubungan dengan kematian akibat stroke meliputi jenis kelamin, usia, tekanan darah tinggi, indeks massa tubuh (*BMI*), fungsi kognitif, dan nutrisi. Pencegahan stroke melibatkan edukasi kesehatan yang mencakup definisi, penyebab, tanda dan gejala, pencegahan, serta penanganan dan perawatan stroke di rumah. Diharapkan, dengan pengetahuan yang memadai, lansia dapat menerapkan gaya hidup sehat untuk mencegah terjadinya stroke [5].

2.2. Data Mining

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 [6]. Data mining juga didefinisikan sebagai proses menemukan pola dalam data secara otomatis atau semi-otomatis, di mana pola yang ditemukan harus memiliki arti dan memberikan keuntungan, biasanya secara ekonomi [7]

2.3. K-Nearest Neigbors

KNN termasuk dalam kategori algoritma supervised learning, di mana hasil klasifikasi untuk instance baru yang diujikan ditentukan berdasarkan mayoritas kategori dari tetangga terdekat dalam KNN. Prinsip dasar operasional K-NN melibatkan perhitungan jarak antara titik data baru dengan titik yang paling dekat dengannya dalam ruang fitur yang relevan[8].Perhitungan jarak dengan menggunakan fungsi Euclidean dijelaskan oleh persamaan 1 berikut ini

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

Dengan label :

- d (x,y) : Jarak Euclidean antara dua vector x dan y
- n : Jumlah atribut atau dimensi pada setiap vector
- xi dan yi : nilai atribut ke-i pada vector x dan y

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 [9]

2.4. Information Gain

Gain informasi dihitung melalui pengurangan nilai *entropy* sebelum proses pemisahan dari nilai entropy setelah proses pemisahan. Pengukuran ini berfungsi sebagai langkah awal dalam menetapkan atribut yang akan dipertahankan atau dieliminasi. Atribut yang memenuhi kriteria pembobotan akan dipertimbangkan untuk dimasukkan dalam tahap klasifikasi suatu algoritma [10].

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 [11].

Rumus perhitungan Information Gain (IG) adalah sebagai berikut:

$$entropy(S) = \sum_{i=1}^{N} - pi^* \, log_2 \, pi$$

Dengan label sebagai berikut :

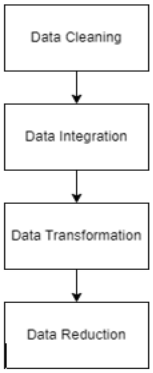
- IG(A) adalah Information Gain untuk atribut A.
- Entropi(Sebelum) adalah Entropi sebelum pemilihan atribut

2.5. KNN Sebagai Alat Prediksi

KNN, sebuah algoritma berbasis klasifikasi yang mengandalkan kemiripan antara data, telah diterapkan dalam prediksi berbagai penyakit dengan tingkat akurasi yang menjanjikan [12]. Sementara itu, Information Gain, yang digunakan untuk pemilihan atribut, telah berhasil dalam meningkatkan keakuratan model [13].

2.6. Pre-Processing

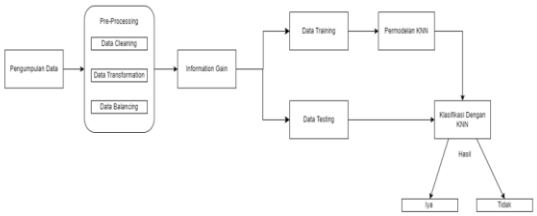
Preprocessing Data merupakan langkah yang terkait dengan persiapan dan transformasi data dalam konteks Data Mining. Langkah ini dilakukan secara simultan untuk meningkatkan efisiensi dalam menemukan pengetahuan dari kumpulan data.



Gambar 1. Tahapan Pre-processing

3. METODE PENELITIAN

Pada bagian ini memuat metode yang digunakan pada penelitian ini



Gambar 2. Diagram Penelitian

3.1. Pengumpulan Data

Dalam proses akuisisi data, berbagai metode dapat diterapkan, seperti studi literatur dan ekstraksi informasi dari berbagai referensi. Data yang digunakan dalam penelitian ini diperoleh dari dataset yang diunduh dari platform Kaggle dengan judul "Prediction of Stroke Dataset" dalam format Excel. Dataset tersebut terdiri dari sepuluh atribut, yaitu *Age, Sex, Hypertension, HeartDisease, EverMarried, WorkType, ResidenceType, AvgGlucoseLevel, BMI, dan SmokingStatus.* Data ini akan diolah untuk mendukung tujuan penelitian.

3.2. Pengolahan Data

Pada pengolahan data ini akan melakukan proses Preprocesing ,yang dimana di dalam Preprocesing itu ada beberapa tahap lagi ,diantaranya :

- Pembersihan Data (Data Cleaning):
- Transformasi Data:
- Pengelompokan (Aggregation):
- Pemfilteran (Filtering):
- Integrasi Data:

3.3. Permodelan K-Nearest Neighbors

Setelah penyeleksian fitur oleh Information Gain sehingga menghasilkan beberapa atribut yang paling berpengaruh, selanjutnya yaitu menerapkan algoritma *K-Nearest Neighbor* untuk melakukan prediksi. Algoritma ini bekerja dengan mencari K tetangga terdekat dalam ruang fitur dan menggunakan informasi dari tetangga-tetangga ini untuk menentukan kelas atau nilai dari data baru.

Dengan nilai K yang telah ditentukan, model KNN kemudian dilatih menggunakan set pelatihan, kemudian model KNN menghitung jarak antara data baru dan semua data dalam set pelatihan, kemudian memilih K tetangga terdekat berdasarkan metrik jarak Euclidean atau metrik jarak lainnya yang sesuai.

3.4. Evaluasi Model

Evaluasi model akan dilakukan menggunakan Confusion Matrix sebagai alat utama untuk mengukur kinerja prediktif dari model yang dikembangkan. Confusion Matrix akan memberikan gambaran yang lebih mendalam mengenai sejauh mana model dapat mengklasifikasikan instance dengan benar ke dalam kelas positif dan negatif. Parameter seperti *True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN)* akan dihitung untuk membentuk Confusion Matrix, yang nantinya akan digunakan untuk mengukur precision, recall, F1-score, dan akurasi model.

4. HASIL DAN PEMBAHASAN

Memuat hasil, Pengujian dan pembahasan tentang skripsi yang telah dilakukan

4.1 Pengumpulan Data

Dalam tahap ini, dijelaskna setiap atribut dalam dataset. Tujuannya untuk memahami variabel-variabel yang tersedia dan bagaimana hubungannya dengan kondisi penyakit stroke dan juga dapat ditentukan atribut mana yang paling penting dalam memprediksi penyakit stroke dan merencanakan analisis data yang sesuai.

Tabel 1. Deskripsi Atribut

No	Atribut	Deskripsi
1.	sex	Jenis kelamin dari subjek
2.	age	Umur Subjek
3.	hypertension	Riwayat hipertensi
4.	heart_disease	Riwayat penyakit jantung
5.	ever_maried	Status pernikahan
6.	work_type	Tipe pekerjaan
7.	residence_type	Tipe tempat tinggal
8.	avg_glucose	Rata-rata glukosa
9.	bmi	Indeks massa tubuh
10.	smoking_status	Status Merokok

Berikut ini adalah beberapa contoh data Riwayat penyakit stroke yang dimana mencangkup informasi tentang penyakit stroke ,dan data di bawah digunakan dalam penelitian ini, yang dapat di temukan di Tabel 2

Tabel 2. Contoh Dataset

No	Sex	Age	hypertensi	…	Stroke
1	1	63.0	0	…	1
2	1	42.0	0	…	1
3	0	61.0	0	…	1
4	1	41.0	1	…	1
5	1	85.0	0	…	1
6	1	55.0	1	…	1
…	…	…	…	…	…
40907	1	38.0	0	…	0
40908	0	53.0	0	…	0
40909	1	32.0	0	…	0
40910	1	42.0	0	…	0

4.2 Pra-Proces

Pada tahap ini bertujuan untuk mempersiapkan data sebelum dilakukan proses klasifikasi. Pra-proses melibatkan tiga tahapan utama, yaitu pembersihan data, transformasi data, dan penyeimbangan data.

4.3 Data Cleaning

Di dalam proses Data Cleaning ini ,masalah utamanya yaitu keberadaan nilai yang hilang (Missing Value). Di dalam penelitian ini yang dimana menggunakan metode KNN dan pendekatan

Information gain ,terdapat 3 nilai yang hilang. Dapat di lihat pada gambar di bawah ini

```
Missing Value :
sex                3
age                0
hypertension       0
heart_disease      0
ever_married       0
work_type          0
Residence_type     0
avg_glucose_level  0
bmi                0
smoking_status     0
stroke             0
dtype: int64
```

Gambar 3. Dataset Sebelum di Cleaning

Dapat dilihat di gambar diatas ,terdapat nilai yang hilang pada data di bagian atribut sex sebanyak 3 buah nilai. Setelah melakukan pengecekan Missing Value , dilakukan proses imputasi dengan mangisi nilai yang hilang dengan nilai yang sering muncul pada atribut tersebut tersebut sehingga menghasilkan keluaran seperti gambar di bawah ini.

```
Missing Value :
sex                0
age                0
hypertension       0
heart_disease      0
ever_married       0
work_type          0
Residence_type     0
avg_glucose_level  0
bmi                0
smoking_status     0
stroke             0
dtype: int64
```

Gambar 4. Dataset sesudah dicleaning

dilakukan proses imputasi dengan mangisi nilai yang hilang dengan nilai yang sering muncul pada atribut tersebut tersebut sehingga menghasilkan keluaran seperti gambar di bawah ini.

4.4 Data Tranformasi

Pada tahap transformation data ini, bertujuan untuk mengubah data kategorikal menjadi numerikal agar dapat di proses dengan lebih efisien oleh machine learning. Dalam penelitian ini, melakukan 2 kali transformasi data yaitu dari data numerikal ke kategorikal dan sebaliknya.

	sex	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1.0	63.0	0	1	1	4	1	228.69	36.6	1	1
1	1.0	42.0	0	1	1	4	0	105.92	32.5	0	1
2	0.0	61.0	0	0	1	4	1	171.23	34.4	1	1
3	1.0	41.0	1	0	1	3	0	174.12	24.0	0	1
4	1.0	85.0	0	0	1	4	1	186.21	29.0	1	1

Gambar 5. Data sebelum transformasi

Pada gambar 5 menunjukan data sebelum diubah ke dalam data kategorik, yang dimana ketika data sudah di ubah ke dalam kategorik akan dilakukan perhitungan untuk mengentahui nilai information dan nilai entropy dari setiap atributnya. Data yang sudah di ubah ke dalam kategorik di tunjukan pada gambar 6.

	sex	hypertension	heart_disease	ever_married	work_type	Residence_type	smoking_status	stroke	age_category	avg_glucose_level_category	bmi_category
0	male	no	yes	yes	Private	Urban	yes	yes	adult	High	Obesity
1	male	no	yes	yes	Private	Rural	no	yes	adult	Medium	Obesity
2	female	no	no	yes	Private	Urban	yes	yes	adult	High	Obesity
3	male	yes	no	yes	Self-employed	Rural	no	yes	adult	High	Ideal
4	male	no	no	yes	Private	Urban	yes	yes	adult	High	Overweight

Gambar 6. Data sesudah transformasi

4.5 Seleksi Fitur Information Gain

Dalam penelitian ini, pengujian dilakukan untuk membandingkan akurasi model *KNN* dalam prediksi penyakit stroke dengan menggunakan Information Gain dan tanpa Information Gain. Eksperimen ini mencakup pengujian akurasi untuk berbagai rasio pembagian data, evaluasi pemilihan fitur menggunakan Information Gain, dan pengujian nilai K. Semua eksperimen ini bertujuan untuk menemukan hasil akurasi prediksi penyakit stroke terbaik untuk model yang dikembangkan. Nilai gain dari setiap atribut ditunjukkan pada Tabel

Tabel 3. Nilai Gain setiap atribut

No	Atribut	Nilai Gain
1.	sex	0.00887
2.	age	0.00364
3.	hypertension	0.04947
4.	heart_disease	0.03832
5.	ever_maried	0.02440
6.	work_type	0.02118
7.	residence_type	0.00009
8.	avg_glucose	0.04937
9.	bmi	
10.	smoking_status	0.00340

4.6 Permodelan KNN

Tahapan permodelan K-Nearest Neighbors (KNN) dengan metode Information Gain dalam penelitian ini dilakukan melalui beberapa langkah penting. Pertama, dilakukan seleksi fitur menggunakan Information Gain (IG) untuk menentukan fitur-fitur yang memiliki kontribusi terbesar terhadap variabel target (stroke). Lima fitur utama yang terpilih adalah Hypertension, Avg_glucose_level, Heart_disease, Ever_married, dan Work_type. Fitur-fitur ini dianggap memberikan informasi yang relevan dalam membedakan kelas stroke dan non-stroke. Selanjutnya, dilakukan pemilihan nilai k dalam algoritma KNN, di mana nilai k diuji dalam rentang 10 hingga 50. Hasil menunjukkan bahwa nilai k = 25 memberikan akurasi tertinggi sebesar 99,85%, menjadikannya nilai k yang optimal untuk model ini. Setelah itu, dilakukan uji coba dengan menambah jumlah fitur menjadi 6 hingga 9, namun penambahan fitur ini justru menurunkan akurasi model, menunjukkan bahwa fitur tambahan tidak memberikan informasi yang relevan dan menambah kompleksitas model. Pembagian data latih dan uji dengan rasio 80:20 dipilih untuk memberikan keseimbangan optimal, dengan 80% data untuk pelatihan dan 20% untuk pengujian. Rasio ini

memberikan hasil yang optimal, memastikan model mampu memberikan generalisasi yang baik terhadap data baru. Berdasarkan tahapan yang dijalankan, model KNN dengan nilai k = 25 dan 5 fitur utama memberikan performa terbaik dengan akurasi 99,85%. Tahapan ini menunjukkan pentingnya seleksi fitur yang tepat dan pemilihan parameter k yang optimal dalam menghasilkan model KNN yang akurat untuk prediksi stroke.

Tabel 4. Pencarian Nilai K dan perbandingan jumlah atribut

No	Total Attributes	K-Value	Accuracy
1	4 (hypertension, avg_glucose_level, heart_disease, ever_married)	10	99.76%
		15	99.76%
		20	99.76%
		25	99.76%
		30	99.76%
		35	99.76%
		40	99.76%
		....	....
		45	99.76%
		50	99.76%
2	5 (hypertension, avg_glucose_level, heart_disease, ever_married, work_type)	10	99.85%
		15	99.85%
		20	99.85%
		25	99.85%
		30	99.85%
		35	99.85%
		40	99.85%
		....	....
		45	99.85%
		50	99.85%
3	6 (hypertension, avg_glucose_level, heart_disease, ever_married, work_type, sex)	10	99.79%
		15	99.79%
		20	99.79%
		25	99.79%
		30	99.76%
		35	99.76%
		40	99.68%
		....	....
		45	99.68%
		50	99.68%
4	7 (hypertension, avg_glucose_level, heart_disease, ever_married, work_type, sex, bmi)	10	99.84%
		15	99.84%
		20	99.84%
		25	99.84%
		30	99.80%
		35	99.80%
		40	99.77%
		....	....
		45	99.77%
		50	99.73%
5	8 (hypertension, avg_glucose_level, heart_disease, ever_married, work_type, sex, bmi, age)	10	84.25%
		15	83.21%
		20	82.62%
		25	81.65%
		30	81.28%
		35	80.79%
		40	80.47%
		...	....
		45	80.19%
		50	79.83%
6	9 (hypertension, avg_glucose_level, heart_disease, ever_married, work_type, sex, bmi, age, smoking_status)	10	86.19%
		15	85.02%
		20	84.12%
		25	83.57%
		30	83.24%
		35	82.58%
		40	82.09%
		....	....
		45	81.96%
		50	81.59%
Akurasi terbesar di 5 fitur dengan nilai K - 25 : 99.85%			

4.7 Evaluasi

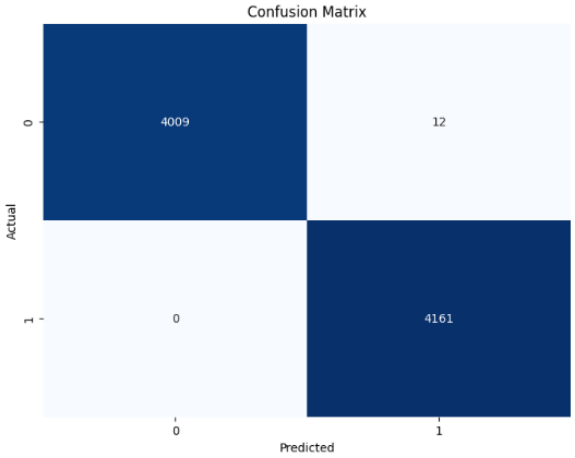
Hasil evalusi pengujian menunjukkan bahwa akurasi tertinggi sebesar 99,85% dicapai ketika menggunakan 5 fitur utama, yaitu hypertension, avg_glucose_level, heart_disease, ever_married, dan work_type, dengan k=25. Penambahan fitur setelah 5 fitur utama justru mengurangi kinerja model, terutama ketika jumlah fitur mencapai 8 atau 9. Digunakan dua model klasifikasi, yaitu dengan Information Gain dan tanpa Information Gain. Model KNN dengan Information Gain mendapatkan performa yang lebih tinggi dibandingkan dengan model KNN tanpa Information Gain. Evalusi model diperoleh sebagai berikut :

4.8 Akurasi perbandingan Model KNN dengan IG dan tanpa IG

Tabel 5. Perbandingan akurasi Model

Nilai K	KNN tanpa Information Gain	KNN dengan Information Gain (5 fitur)
10	88.30%	99.85%
15	87.18%	99.85%
20	86.43%	99.85%
25	85.86%	99.85%
30	85.14%	99.85%
35	84.94%	99.85%
40	84.33%	99.85%
45	84.21%	99.85%
50	83.68%	99.85%

4.9 Confusion Matrix KNN dengan IG

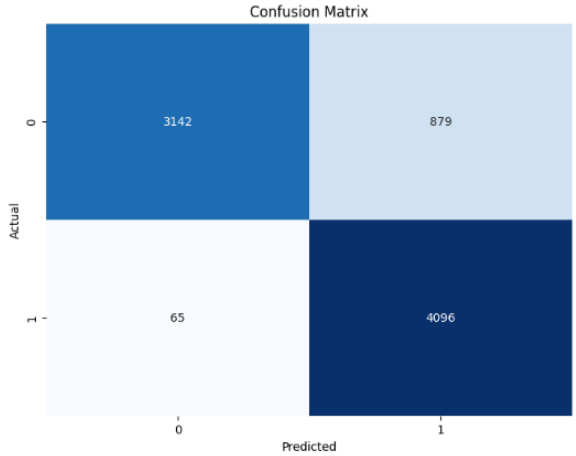


Gambar 7. Confusion Matrix KNN+IG

Dalam matriks ini, terdapat 4009 prediksi benar untuk kelas 0 (True Negative), 4161 prediksi benar untuk kelas 1 (True Positive), 12 prediksi salah di mana model memprediksi 1 padahal sebenarnya 0 (False Positive), dan tidak ada prediksi salah di mana model memprediksi 0 padahal sebenarnya 1 (False Negative).

4.10 Confusion Matrix KNN tanpa IG

Dalam matriks ini, terdapat 4096 prediksi benar untuk kelas 1 (True Positive), 3142 prediksi benar untuk kelas 0 (True Negative), 879 prediksi salah di mana model memprediksi 1 padahal sebenarnya 0 (False Positive), dan 65 prediksi salah di mana model memprediksi 0 padahal sebenarnya 1 (False Negative).



Gambar 8. Confusion Matrix KNN tanpa IG

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa penerapan Information Gain dalam metode K-Nearest Neighbors (KNN) secara signifikan meningkatkan performa model dalam memprediksi risiko stroke. Dengan menggunakan data sebanyak 40.910 record, model KNN yang mengintegrasikan Information Gain mencapai akurasi tertinggi 99,85% menggunakan 5 fitur terpilih (hypertension, avg_glucose_level_category,heart_disease,ever_married, work_type) dan nilai k=25. Akurasi tinggi ini didukung oleh precision 99,71%, recall 100%, dan F1-Score 99,86%, menandakan bahwa model berhasil mengklasifikasikan sebagian besar data dengan benar dan seimbang dalam identifikasi kasus positif. Sebaliknya, model KNN tanpa Information Gain hanya mencapai akurasi 88,46%. Hasil ini menegaskan bahwa integrasi Information Gain secara efektif meningkatkan akurasi prediksi stroke. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi metode pemilihan fitur lainnya seperti Chi-Square dan Recursive Feature Elimination (RFE), guna menemukan kombinasi fitur yang lebih optimal dan mungkin memberikan peningkatan performa yang lebih signifikan.

DAFTAR PUSTAKA

[1] E. J. Benjamin *et al.*, *Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association*, vol. 139, no. 10. 2019. doi: 10.1161/CIR.0000000000000659.

[2] M. N. Maskuri, Harliana, K. Sukerti, and R. M. H. Bhakti, "Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 4, no. 1, pp. 130–140, 2022.

[3] S. Novianti, N. Arif, A. M. Siregar, S. Faisal, and A. R. Juwita, "Klasifikasi Penyakit Serangan Jantung Menggunakan Metode Machine Learning K-Nearest Neighbors (KNN) dan Support Vector Machine (SVM)," vol. 8, pp. 1617–1626, 2024, doi: 10.30865/mib.v8i3.7844.

[4] M. P. Lindsay *et al.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2019," *Int. J. Stroke*, vol. 14, no. 8, pp. 806–817, 2019, doi: 10.1177/1747493019881353.

[5] Y. Oktarina, N. Nurhusna, K. Kamariyah, and S. Mulyani, "Edukasi Kesehatan Penyakit Stroke Pada Lansia," *Med. Dedication J. Pengabdi. Kpd. Masy. FKIK UNJA*, vol. 3, no. 2, pp. 106–109, 2021, doi: 10.22437/medicaldedication.v3i2.11220.

[6] A. A. Nababan, M. Khairi, and B. S. Harahap, "Implementation of K-Nearest Neighbors (KNN) algorithm in classification of data water quality," *J. Mantik*, vol. 6, no. 1, pp. 30–35, 2022.

[7] D. Papakyriakou and I. S. Barbounakis, "Data Mining Methods: A Review," *Int. J. Comput. Appl.*, vol. 183, no. 48, pp. 5–19, 2022, doi: 10.5120/ijca2022921884.

[8] M. S. Mustafa and I. W. Simpen, "Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba," vol. VIII, no. 1, pp. 1–10.

[9] 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.

[10] T. Setiyorini, "Penerapan Information Gain pada K-Nearest Neighbor untuk Klasifikasi Tingkat Kognitif Soal pada Taksonomi Bloom," no. 1, pp. 57–62, 2021.

[11] U. I. Gain, "JURNAL RESTI Using Information Gain and KNN Methods," vol. 5, no. 158, pp. 185–192, 2023.

[12] A. Noviriandini, P. Handayani, and Syahriani, "Prediksi Penyakit Liver Dengan Menggunakan Metode," *Pros. TAU SNAR-TEK Semin. Nas. Rekayasa dan Teknol.*, no. November, pp. 75–80, 2019.

[13] U. Mutmainnah, B. D. Setiawan, and C. Dewi, "Pengaruh Seleksi Fitur Information Gain pada K-Nearest Neighbor untuk Klasifikasi Tingkat Kelancaran Pembayaran Kredit Kendaraan," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 9, pp. 8882–8888, 2019.