Optimalisasi *Feature Selection* Untuk Mendeteksi Penyakit Diabetes Mellitus Menggunakan Metode *Decision Tree*

Aplea Pameka[1], Rudi Heriansyah[2], Lastri Widya Astuti*[3]
[1,2,3] Fakultas Ilmu Komputer dan Sains, Universitas Indo Global Mandiri, Palembang
e-mail: [1]2020110038@students.uigm.ac.id, [2]rudi@uigm.ac.id, ***[3]lastriwidya@uigm.ac.id**

Abstrak

Penyakit diabetes mellitus tipe 2 merupakan permasalahan kesehatan dengan tingkat prevalensi tinggi di seluruh dunia. International Diabetes Federation (IDF) kawasan Asia Pasifik Barat terdiri dari 20 negara dimana Indonesia menjadi salah satu anggota. Di dunia terdapat 536,6 juta orang menderita penyakit diabetes dan 206 juta orang di kawasan Asia Pasifik Barat. Hingga tahun 2045 jumlah ini akan terus meningkat menjadi 260 juta di Kawasan Asia Pasifik Barat dan sebanyak 783,7 juta penderita diabetes di dunia. Pola hidup yang kurang sehat menyebabkan penyakit diabetes, sehingga didapatkan temuan penderita diabetes bukan lagi berasal dari orang yang berusia lanjut. Pemanfaatan machine learning telah banyak digunakan untuk mengenali beberapa pola penyakit sebagai upaya awal deteksi. Matrik akurasi machine learning dapat ditingkatkan dengan menggunakan algoritma decision tree dengan menambahkan perbaikan pada proses pemilihan fitur menggunakan information gain. Penelitian ini menggunakan beberapa atribut yang diduga memiliki informasi dalam mendeteksi penyakit diabetes mellitus. Dengan menggunakan metode Information Gain dalam proses pemilihan subset fitur didapatkan 5 fitur dengan nilai tertinggi. Selanjutnya, algoritma klasifikasi Decision Tree diterapkan pada subset fitur terpilih dan penerapan Algoritma decision tree dengan penggunaan information gain meningkatkan akurasi sebanyak 96.25%. Diharapkan dengan adanya hasil penelitian ini dapat mengurangi jumlah penderita penyakit diabetes.

Kata kunci*— Deteksi, Diabetes Mellitus, Feature Selection, Information Gain, Decision Tree*

Abstract

Diabetes mellitus type 2 is a health problem with a high prevalence rate throughout the world. The International Diabetes Federation (IDF) in the West Asia Pacific region consists of 20 countries, of which Indonesia is a member. In the world, 536.6 million people have diabetes and 206 million in the West Asia Pacific region. Until 2045, this number will continue to increase to 260 million in the West Asia Pacific Region and as many as 783.7 million diabetes sufferers worldwide. An unhealthy lifestyle causes diabetes, so it is found that people with diabetes no longer come from older people. Machine learning has been widely used to recognize several disease patterns as an initial detection effort. The machine learning accuracy matrix can be improved using a decision tree algorithm by adding improvements to the feature selection process using information gain. This research uses several attributes that are thought to have information on detecting diabetes mellitus. Five features with the highest scores were obtained using the Information Gain method in the feature subset selection process. Next, the Decision Tree classification algorithm is applied to a subset of selected features, and applying the Decision Tree algorithm using information gain increases accuracy by 96.25%. It is hoped that the results of this research can reduce the number of people with diabetes.

Keywords*— Detection, Diabetes Mellitus, Feature Selection, Information Gain, Decision Tree*

1. PENDAHULUAN

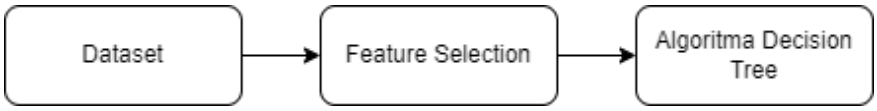
Penyakit diabetes mellitus, juga dikenal sebagai "kencing manis", adalah keadaan di mana gula dalam darah meningkat karena tubuh tidak bisa menciptakan atau mengeluarkan cukup insulin [1] Pankreas menghasilkan hormon insulin, yang bertanggung jawab untuk memungkinkan glukosa yang terkandung dalam makanan mengalir ke sel-sel tubuh dan kemudian diproses untuk menghasilkan energi[2]. Indonesia menduduki peringkat kelima di dunia pada total penderita diabetes, dengan total 19,47 juta orang pada tahun 2021 dan diperkirakan akan terus meningkat hingga mencapai angka 47% tahun selanjutnya, dengan total 28,57 juta orang[3]. Pasien diabetes memiliki gejala antara lain: berat badan turun secara signifikan, frekuensi buang air kecil meningkat, rasa lelah dan mengantuk yang cepat, rasa haus yang berlebihan, kulit menjadi bermasalah, penyembuhan yang lambat terhadap penyakit lain, infeksi jamur, pandangan yang kabur, kesemutan atau mati rasa, dan selalu merasa lapar[4].

Pada era teknologi informasi saat ini penggunaan *machine learning* dapat membantu dalam proses deteksi penyakit secara akurat dan juga dapat menghemat waktu. Metode pohon keputusan mengubah beragam fakta yang merepresentasikan pengetahuan dalam bentuk pohon keputusan yang terstruktur. Data dipohon keputusan disajikan dalam bentuk tabel atribut dan skor, yang bermanfaat untuk mendeskripsikan hubungan antara variabel masukan potensial dan variabel target. Salah satu atributnya adalah atribut yang memuat data hasil untuk setiap data item, yang dikenal sebagai atribut output [5]. Akurasi metode klasifikasi pada proses pengenalan pola penyakit sangat dipengaruhi oleh jumlah atribut yang digunakan dalam proses komputasi. Dimensi data yang tinggi dapat direduksi melalui proses ekstrasi fitur atau seleksi fitur untuk meningkatkan kinerja pengklasifikasian.

Pada beberapa penelitian terdahulu menyatakan bahwa *decision tree* yang digabung dengan metode seleksi fitur menghasilkan berbagai tingkat akurasi. Pada penelitian yang dilakukan oleh Hanif dan Setiaji menyatakan bahwa model *decision tree* memperoleh akurasi sebesar 96.16% dalam melakukan prediksi penyakit diabtes[6]. Sedangkan penelitian yang dilakukann oleh Astuti dkk memperoleh hasil bahwa model *decision tree* hanya memperoleh 71% tingkat akurasi dalam prediksi penyakit diabetes[7]. Penelitian ini menerapkan metode *information gain* pada *feature selection* dan metode *decision tree* untuk melihat seberapa tinggi tingkat akurasi yang didapatkan dan seberapa optimal *feature selection* dalam meningkatkan tingkat akurasi deteksi penyakit. Dengan pendekatan ini diharapkan dapat mengoptimalisasi dalam meminimalkan jumlah *features* untuk meningkatkan hasil akurasi deteksi penyakit diabetes mellitus.

2. METODE PENELITIAN

Penelitian mempunyai tahapan dalam melakukan proses perhitungan. Berikut tahapan pelaksanaan penelitian ini.



Gambar 1 Bagan Alir Penelitian

2.1. Dataset

Dataset pada studi ini merupakan data primer yang bersumber dari puskesmas Betung Semendawai Barat dengan menggunakan 80 dataset penderita penyakit diabetes dengan tiga belas atribut yaitu *riwayat penyakit, umur, jenis kelamin, kehamilan, riwayat penyakit keluarga, insulin, sistol, diastol, lingkar pinggang, imt, glukosa urin, glukosa puasa, dan diagnosa*. Metode *k-fold cross validation* akan membagi dataset menjadi *data testing* serta *data training*, kemudian data tersebut akan di hitung tingkat akurasinya menggunakan *confusion matrix*.

2.2. Feature Selection

Feature selection adalah proses untuk memilih subset dari fitur-fitur asli yang paling berharga. Tujuannya adalah untuk menghasilkan subset optimal dari fitur yang relevan tanpa menambah kompleksitas model atau algoritma. Manfaat dari feature selection meliputi mengurangi dimensi data, menghilangkan fitur yang tidak relevan, meningkatkan pemahaman terhadap data, serta potensial untuk menumbuhkan akurasi algoritma dengan fokus pada fitur yang paling penting.[8] *Information gain* adalah salah satu metoda pemilihan *feature* yang popular yang ada pada *feature selection*. *Feature* ini sering digunakan untuk menentukan peringkat atribut yang dianggap memiliki pengaruh terbesar pada kelas. *Information gain* digunakan untuk membantu menciptakan atribut yang relevan untuk digunakan oleh kelas target, karena setiap atribut mempunyai nilai serta bisa dipilih berdasarkan nilai terbaik[9].

2.3. Decision Tree

Decision tree adalah salah satu algoritma yang populer untuk prediksi dan klasifikasi data, yang mengadopsi struktur berbentuk pohon. Algoritma ini mampu menyederhanakan data yang kompleks menjadi struktur hierarkis yang lebih mudah dimengerti. Salah satu keunggulan utama dari algoritma decision tree adalah kemampuannya dalam menginterpretasikan masalah kompleks dan menyajikan solusi secara intuitif dalam proses pengambilan keputusan. Hal ini membuatnya menjadi pilihan yang sering digunakan dalam penelitian dan aplikasi di berbagai bidang[10]. Pohon keputusan yakni suatu taknik data *mining* yang menyajikan struktur mirip pohon bertujuan guna menetapkan peraturan pada proses klasifikasi. Ada dua jenis pohon keputusan: pohon klasifikasi dan pohon regresi. Pohon klasifikasi dipakau guna klasifikasi, yang prosesnya melibatkan pelabelan dan penempatan rekaman ke dalam kelas yang telah ditentukan. Sedangkan pohon regresi digunakan untuk klasifikasi guna membuat estimasi nilai variabel target numerik[11].

Gambar 2 Konsep *Decision Tree*

2.4.Confusion Matrix

Evaluasi matrik pada model dalam penelitian ini menggunkan *confusion matrix. Confusion* matrik yakni Teknik yang dipakai guna mengetahui tingkat akurasi sebuah model[12]. Berikut yakni representasi perolehan tahap klasifikasi dan rumus dari *confusion matrix*[13]:

Tabel 1 *Confusion Matrix*

Confusion Matrix	*Predicted Class*	
	Positive	*Negative*
True Class	*True Positive (TP)*	*False Positive (FP)*
	False Negative (FN)	*True Negative (TN)*

Keterangan:

- *TP : Nilai data sample positif yang diklasifikasikan sebagai True*
- *FP : Nilai data sample negatif yang dilasifikasikan sebagai True*

- *TN : Nilai data sample negatif yang diklasifikasikan sebagai False*
- *FN : Nilai data sample positif yang diklasifikasikan sebagai False*

Berdasarkan nilai diatas pada tabel 1 maka bisa didapatkan nilai akurasi, presisi serta *recall*. Nilai akurasi menunjukkan keakuratan model dalam mengklasifikasi data dengan benar, yang ditunjukan pada persamaan 1.

Nilai presisi ialah total data positif yang dikelompokan dengan benar kemudian dibagi dengan total data positif yang ditunjukkan pada persamaan 2. Recall adalah jumlah data positif yang diklasifikasikan dengan benar dibagi total data positif diakumulasi dengan data salah positif ditunjukkan oleh persamaan 3.

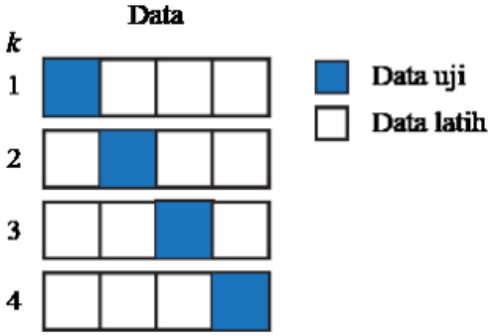
$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

2.5. Cross Validation

Cross Validation merupakan langkah dimana keakuratan prediksi dapat divalidasi[14]. Bentuk dari *cross validation* adalah *K-Fold Cross Validation* merupakan himpunan data yang disampaikan dibagi menjadi sejumlah K bagian atau fold, dengan setiap *fold* dipakai sebagai set uji di beberapa titik[15]. Metode ini bertujuan untuk bertujuan untuk mengukur sejauh mana model yang dikembangkan dapat memberikan hasil yang konsisten dan handal pada berbagai subset data.



Gambar 3 Ilustrasi K-*Fold Cross Validation*

3. HASIL DAN PEMBAHASAN

Analisis, perolehan serta uji pada penelitian yang telah dilaksanakan akan dijelaskan pada bagian ini. Tahap analisis hasil serta uji dilaksanakan dengan memakai Bahasa pemrograman *python* dan memanfaatkan *tool Google Collab.* Berikut penjelasan mengenai skenario pengujian yang akan diuraikan di tabel 2.

Tabel 2 Skenario Pengujian

Skenario	Informasi
Pembagian Data	Komposisi data training dan data testing, 80:20
Atribut	Riwayat Penyakit, Umur, Jenis Kelamin, Kehamilan, Riwayat Penyakit Keluarga, Insulin, Sistol, Diastol, Lingkar Pinggang, IMT, Glukosa Urin, Glukosa Puasa, dan Diagnosa.
Jumlah Data	80 data: 60 data penderita penyakit diabetes, 20 data normal

3.1. Data Cleansing

Tahap pertama yang dilakukan adalah mengimpor dataset lalu dilanjutkan dengan *data cleansing* yang dimulai dengan pemeriksaan apakah terdapat data yang kosong (*null*) dan data duplikat pada *dataset.* Hasil Perkodingan ditunjukkan pada gambar berikut.

```
Tidak terdapat data duplikat dalam dataset.
Jumlah data yang kosong untuk setiap atribut:
UMUR                    0
JK                      0
DIAGNOSA                0
RIW PENYAKIT            0
RIW PENYAKIT KLRGA      0
GLUKOSA URIN            0
GLUKOSA PUASA (mg/dL)   0
IMT                     0
KEHAMILAN               0
LINGKAR PINGGANG        0
SISTOL                  0
DIASTOL                 0
INSULIN                 0
dtype: int64
```

Gambar 3 Tampilan Data Kosong (*null*)

Setelah dipastikan tidak ada data kosong dan data duplikat maka akan dilanjutkan dengan tahap selanjutnya.

3.2. K-Fold Cross Validation

Setelah melalui tahap data *cleansing*, langkah selanjutnya adalah menerapkan metode validasi silang k-*fold*. Pada penelitian ini penulis membagi metode *K-Fold Cross Validation* yang membagi *dataset* menjadi 5 bagian (*fold*), di mana setiap *fold* berperan sebagai set pengujian (*testing*) satu kali sementara bagian lainnya digunakan sebagai set pelatihan (*training*). Proses ini diulangi sebanyak 5 kali, sehingga setiap *fold* pernah menjadi set pengujian.

Gambar 4 Ilustrasi Implementasi *K-fold Cross Validation* 5 *folds*

3.3. Implementasi Feature Selection

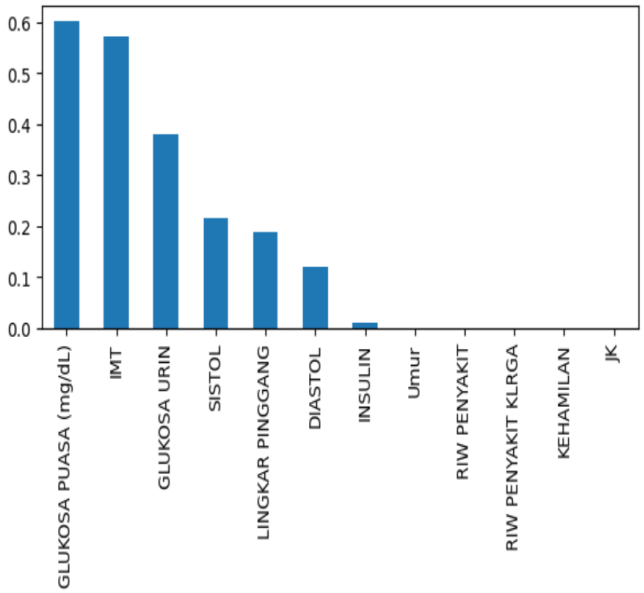
Proses selanjutnya adalah melakukan perhitungan terhadap atribut menggunakan *feature selection*. Tabel 3 menyajikan hasil perhitungan dari setiap atribut menggunakan *information gain* setelah melalui tahap perkodingan.

Tabel 3 Nilai Atribut dengan perhitungan *feature selection*

Atribut	Nilai
Glukosa Puasa	0,602023
IMT	0,572856
Glukosa Urin	0,379886
Sistol	0,217550
Lingkar Pinggang	0,188531
Diastol	0,121656

Atribut	Nilai
Insulin	0,009708
Umur	0,000000
Riwayat Penyakit	0,000000
Riwayat Penyakit Keluarga	0,000000
Kehamilan	0,000000
Jenis Kelamin	0,000000

Pada tabel diatas telah menunjukkan nilai hasil dari setiap atribut dan didapatkan nilai yang berbeda-beda untuk setiap atribut dan didapat juga atiribut yan tidak memiliki nilai atau mempunyai hasil nol. Sehingga dipilih lima atribut dengan nilai tertinggi yaitu Glukosa puasa, IMT (Indeks Massa Tubuh), Glukosa Urin, Sistol, dan Lingkar Pinggang. Hasil tabel tersebut akan disajikan pada diagram yang bisa diamati pada gambar ini:



Gambar 5 Hasil *Feature Selection* penderita penyakit diabetes mellitus

3.4. Implementasi Algoritma

Setelah melalui tahap perhitungan *feature selection* didapatkan 5 atribut dengan nilai *gain* tertinggi, yaitu Glukosa puasa, IMT, glukosa urin, sistol, dan lingkar pinggang. Selanjutnya adalah implementasi algoritma *decision tree* yang dipakai guna membangun model klasifikasi berdasarkan pola-pola yang teridentifikasi dalam data *training*. Model *decision tree* akan dibentuk dari *dataset* ini, memungkinkan algoritma untuk mempelajari pola dan hubungan antara fitur-fitur terpilih dengan variabel target.

Dengan model *decision tree* yang telah terbentuk, program selanjutnya akan menggunakannya untuk melakukan klasifikasi terhadap data *testing*. Proses ini memungkinkan model untuk membuat prediksi terhadap kelas target dari data *testing*, berdasarkan pembelajaran yang diperoleh dari data *training.*

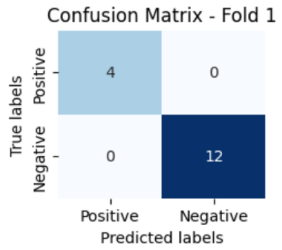
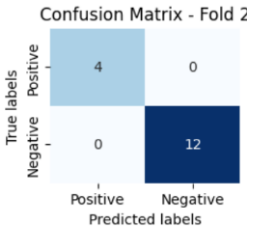
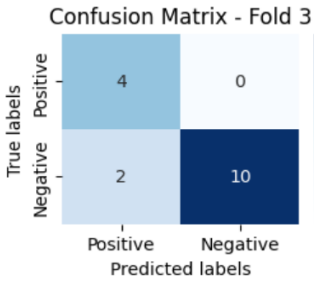
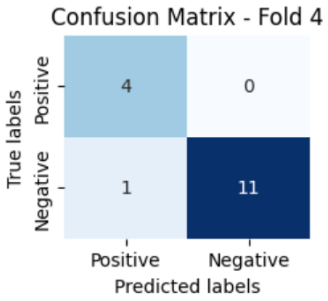
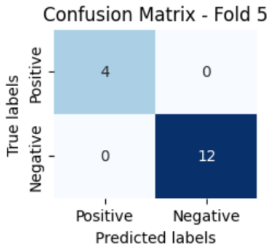
Decision Tree menggunakan parameter berikut:

Criterion : "*entropy*"
Random State : 12
Max Depth : 4
Min Sample Leaf : 13

3.5.Pengujian

Tahapan pengujian merupakan kegiatan yang dilakukan untuk membuktikan tingkat keberhasilan dalam penelitian pada saat mendesain sebuah model. Penelitian ini menggunakan matrik konfusi sebagai alat pengukuran untuk menentukan akurasi, presisi, dan recall. Pengujian menggunakan dataset sejumlah 80 data yang terdiri dari 60 dataset pasien penderita penyakit dan 20 dataset pasien dengan kondisi normal. Komposisi data pengujian disajikan pada tabel 4

Tabel 4 *Confussion matrix* pada model *decision tree* yang telah dibuat

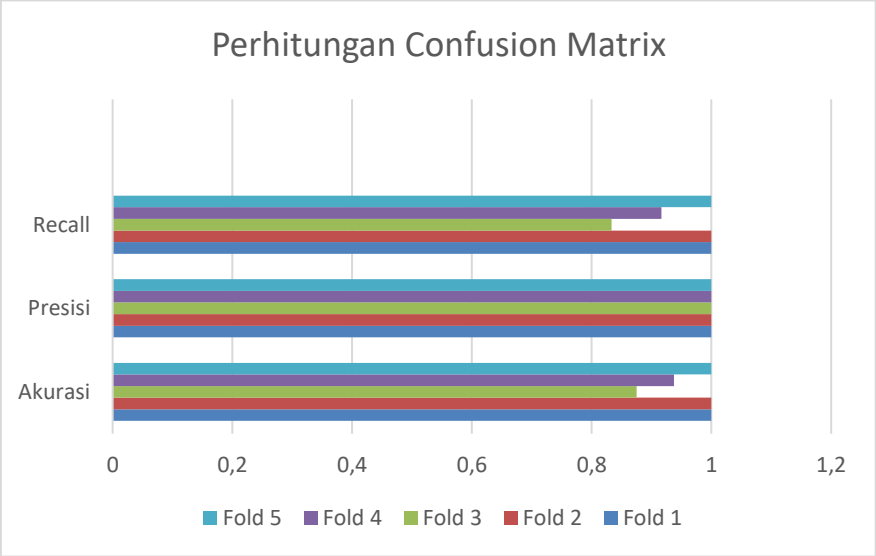
Model *Decision Tree - Fold* 1	**Model *Decision Tree - Fold* 2**
	
Model *Decision Tree - Fold* 3	**Model *Decision Tree - Fold* 4**
	
Model *Decision Tree - Fold* 5	
	

Dari tabel 4 diketahui akurasi, presisi, dan recall untuk setiap *fold* berbeda. Hasil dari perhitungan *confusion matrix* setiap *fold* dijelaskan ditabel 5.

Tabel 5 Perhitungan hasil *confusion matrix* setiap *fold*

Fold	**Algoritma *Decision Tree***		
	Akurasi	**Presisi**	**Recall**
1	1,0000	1,0000	1,0000
2	1,0000	1,0000	1,0000
3	0,8750	1,0000	0,8333
4	0,9375	1,0000	0,9167
5	1,0000	1,0000	1,0000
Average	**0,9625**	**1,0000**	**0,9500**

Deskripsi nilai perhitungan *confusion matrix* diatas diperjelas dengan diagram batang pada gambar 6, untuk menunjukkan hasil perhitungan masing masing *fold* dengan menambahkan *feature selection*.



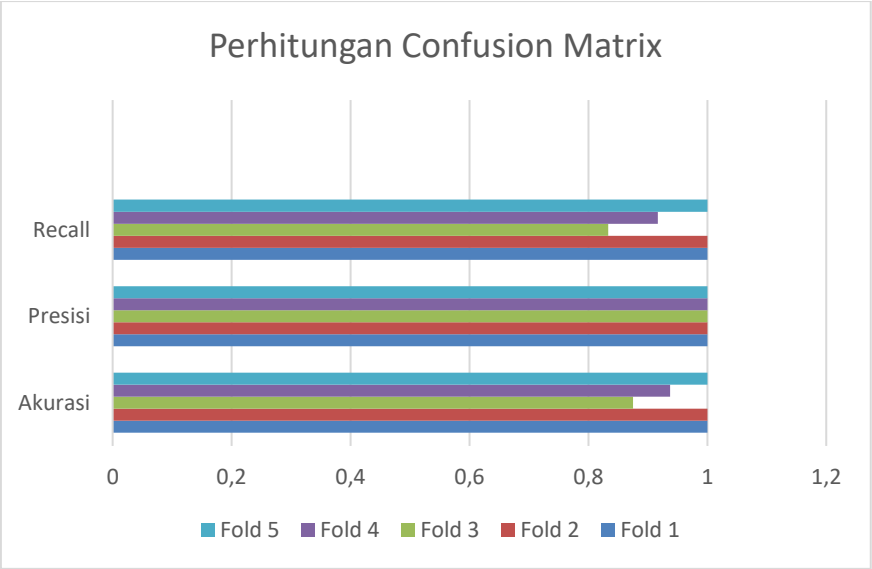
Gambar 6 Diagram Batang Perhitungan setiap *Fold* menggunakan *Feature Selection*

Dari tabel diatas didapatkan hasil akurasi keseluruhan yaitu sebesar 0,9625 atau sebesar 96%, presisi didaparkan nilai 1,0000 atau sebesar 100%, dan recall sebesar 0,9500 atau sebesar 95%. Pada penelitian ini juga memuat hasil dari metode *decision tree* tanpa melibatkan f*eature selection* sebagai pembanding seberapa berpengaruh *feature selection* metode *information gain* dalam mendeteksi penyakit diabetes mellitus. Berikut hasil perhitungan *confusion matrix* menggunakan metode *decision tree* tanpa melibatkan *feature selection*.

Tabel 6 hasil *confusion matrix* model *decision tree* pada masing masing *fold* tanpa *feature selection*

Fold	**Algoritma *Decision Tree***		
	Akurasi	**Presisi**	**Recall**
1	1,0000	1,0000	1,0000
2	1,0000	1,0000	1,0000
3	0,8750	1,0000	0,8333
4	0,9375	1,0000	0,9167
5	1,0000	1,0000	1,0000
Average	**0,9625**	**1,0000**	**0,9500**

Hasil perhitungan *fold* tanpa *feature selelection* pada tabel 6 akan diperjelas melalui diagram batang yang ditunjukkan pada gambar 7.



Gambar 7 Diagram Batang Perhitungan setiap *Fold* tanpa *Feature Selection*

Dari tabel diatas didapatkan hasil akurasi keseluruhan yaitu sebesar 0,9625 atau sebesar 96%, oresisi didaparkan nilai 1,0000 atau sebesar 100%, dan recall sebesar 0,9500 atau sebesar 95%. Penerapan *feature selection* membantu mengidentifikasi subset fitur yang paling relevan, mengurangi redundansi, dan meminimalkan risiko *overfitting*. Didapatkan bahwa penggunaan metode *feature selection* dan tidak melibatkan *feature selection* pada model *Decision Tree* menunjukkan perbedaan proses dalam pemodelan, namun secara keseluruhan tidak memberikan perbedaan dalam performa akhir model. Meskipun terdapat perbedaan dalam kompleksitas dan jumlah fitur yang digunakan, hasil akhir untuk tingkat akurasi dan recall pada kedua pendekatan cenderung serupa.

4. KESIMPULAN

Berdasarkan analisis dan hasil uji coba sebelumnya, maka didapatkan kesimpulan sebagai berikut:

1. Implementasi metode *feature selection* menggunakan *Information Gain* pada model *Decision Tree* menghasilkan tingkat akurasi sebesar 96,25% menunjukkan bahwa hasil ini meningkat dibandingkan dengan beberapa penelitian terdahulu. Melalui hasil tersebut, dapat ditarik kesimpulan bahwa penerapan *feature selection* dengan menggunakan metode *Information Gain* dan penggunaan model *Decision Tree* dalam konteks deteksi penyakit diabetes mellitus meningkatkan tingkat akurasi.
2. Penelitian ini berhasil menemukan lima fitur utama setelah menggunakan seleksi fitur dengan menggunakan dataset primer yang terdiri dari tigabelas atribut. Kelima fitur tersebut meliputi Glukosa Urin, Glukosa Puasa, Indeks Massa Tubuh (IMT), Sistol, dan Lingkar Pinggang.

5. SARAN

Berdasarkan analisis proses dan hasil pelaksanaan penelitian ini, terdapat ruang untuk pengembangan model penelitian untuk meningkatkan akurasi dan pengembangan model sehingga disarankan untuk penelitian ke depan sebagai berikut:

1. Jumlah data yang digunakan sangat terbatas untuk pengujian sehingga untuk penelitian kedepannya disarankan mengembangkan jumlah dataset penderita penyakit diabetes agar pada proses penelitian lanjutan dapat meningkatkan hasil akurasi pengujian.
2. Hasil dari analisis ini model dapat dikembangkan lagi dengan menggunakan metode *machine learning dan feature selection* lainnya yang dikomparasi dengan beberapa metode *machine learning* atau melakukan perbaikan data dengan menambahkan proses atau metode *preprocessing*.

UCAPAN TERIMAKASIH

Ungkapan pujian dan syukur peneliti panjatkan hanya untuk Allah SWT atas anugrah serta kemudahan hingga pada akhirnya penelitian ini dapat diselesaikan. Rasa terima kasih untuk kedua orangtua yang mendukung semua kinerja dan tahapan penelitian ini. Juga kepada teman-teman yang memberikan dukungan yang tak ternilai selama proses penelitian. Terima kasih juga kepada Universitas Indo Global Mandiri atas dukungan material selama periode pelaksanaan penelitian ini.

DAFTAR PUSTAKA

[1] F. Elfaladonna and A. Rahmadani, "Analisa Metode Classification-Decission Tree dan Algoritma C. 45 untuk Memprediksi Penyakit Diabetes dengan Menggunakan Aplikasi Rapid Miner," *SINTECH (Science and Information Technology) Journal*, vol. 2, no. 1, pp. 10–17, 2019.

[2] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020.

[3] Diabetes Indonesia, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," https://diabetes-indonesia.net/2022/02/idf-diabetes-atlas-global-regional-and-country-level-diabetes-prevalence-estimates-for-2021-and-projections-for-2045/.

[4] Kementerian Kesehatan, "Tanda dan Gejala Diabetes," https://p2ptm.kemkes.go.id/artikel-sehat/tanda-dan-gejala-diabetes.

[5] N. T. Romadloni, I. Santoso, and S. Budilaksono, "Perbandingan Metode Naive Bayes, Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl Commuter Line," *IKRA-ITH Informatika: Jurnal Komputer dan Informatika*, vol. 3, no. 2, pp. 1–9, 2019.

[6] M. B. Hanif and G. G. Setiaji, "Meningkatkan Kinerja Decision Tree C4. 5 Dengan Seleksi Fitur Korelasi Pearson Pada Deteksi Penyakit Diabetes," *Indonesian Journal of Computer Science*, vol. 11, no. 2, 2022.

[7] L. W. Astuti, I. Saluza, E. Yulianti, and D. Dhamayanti, "Feature Selection Menggunakan Binary Wheal Optimizaton Algorithm (Bwoa) Pada Klasifikasi Penyakit Diabetes," *Jurnal Ilmiah Informatika Global*, vol. 13, no. 1, 2022.

[8] F. K. Fikriah, N. Hayati, and J. K. No, "Feature Selection dengan Komparasi Algoritma untuk Prediksi Telemarketing Bank".

[9] M. Ahsan and W. Harianto, "KOMPARASI TINGKAT AKURASI INFORMATION GAIN DAN GAIN RATIO PADA METODE K-NEAREST NEIGHBOR," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 1, pp. 386–391, 2022.

[10] Mellina & Ayu Dian Fitri, "Prediksi deteksi penyakit kanker payudara dengan menggunakan algoritma decision tree," 2023.

[11] M. Defriani and I. Jaelani, "Algoritma J48 Dan Logistic Model Tree Untuk Memprediksi Predikat Kelulusan Mahasiswa: Studi Kasus STT XYZ," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 3, no. 2, pp. 129–140, 2020.

[12] I. Düntsch and G. Gediga, "Confusion matrices and rough set data analysis," in *Journal of Physics: Conference Series*, IOP Publishing, 2019, p. 012055.

[13] L. B. C. Tanujaya, B. Susanto, and A. Saragih, "The comparison of logistic regression methods and random forest for spotify audio mode featurre classification," *indonesian journal of data and science*, vol. 1, no. 3, pp. 68–78, 2020.

[14] A. Prasatya, R. R. A. Siregar, and R. Arianto, "Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes," *PETIR*, vol. 13, no. 1, pp. 86–100, Mar. 2020, doi: 10.33322/petir.v13i1.925.

[15] A. Peryanto, A. Yudhana, and R. Umar, "Klasifikasi Citra Menggunakan Convolutional Neural Network dan K Fold Cross Validation," *Journal of Applied Informatics and Computing*, vol. 4, no. 1, pp. 45–51, 2020.