Jurnal Sistem Informasi dan Teknologi Peradaban (JSITP)
Vol. 6, No. 2, 2025
e-ISSN 2797-1570

Klasifikasi resiko Diabetes Mengunakan Algoritma Decision Tree

Stefanus Charles Selvianto[1], Alexander Novaldi, Andri Wijaya[3*]

[1,2,3] Universitas katolik Musi Charitas
Email: [1]stefanusc33@gmail.com, [2]novaldinovaldi77@gmail.com, [3]andri_wijaya@ukmc.ac.id

Abstrak

Peningkatan kasus Diabetes *Mellitus* menuntut adanya metode deteksi dini yang efektif untuk mencegah komplikasi serius pada penderita. Penelitian ini bertujuan mengklasifikasikan risiko diabetes menggunakan algoritma *Decision Tree* yang mampu menghasilkan aturan keputusan yang mudah diinterpretasikan oleh tenaga medis. Penelitian memanfaatkan dataset *Pima Indians Diabetes* dari repositori UCI *Machine Learning* yang diolah menggunakan perangkat lunak *RapidMiner*. Melalui tahapan *preprocessing* dan pembagian data latih serta uji dengan rasio 80:20, model dievaluasi menggunakan *Confusion Matrix* dan kurva ROC. Hasil pengujian menunjukkan model mencapai akurasi 70.13%, presisi 70.00%, *recall* 25.93%, dan nilai AUC sebesar 0.736 (*fair performance*). Meskipun nilai *recall* rendah mengindikasikan keterbatasan sensitivitas, tingginya nilai presisi menunjukkan model sangat andal dalam meminimalkan kesalahan diagnosis positif palsu. Secara spesifik, model menemukan aturan klinis bahwa kadar glukosa di atas 127.5 mg/dL merupakan indikator risiko tinggi, diikuti oleh *Body Mass Index* (BMI) dan usia sebagai faktor determinan sekunder pada pasien dengan gula darah normal. Penelitian ini menyimpulkan bahwa metode *Decision Tree* efektif digunakan sebagai sistem pendukung keputusan medis berbasis aturan (*rule-based decision support*) untuk identifikasi profil risiko pasien.

Keyword: *Diabetes, Data Mining, Decision Tree, Klasifikasi, RapidMiner.*

I. PENDAHULUAN

Diabetes Mellitus merupakan salah satu penyakit kronis yang jumlah penderitanya terus meningkat setiap tahun, baik di dunia maupun di Indonesia. Penyakit ini terjadi ketika kadar gula dalam darah terlalu tinggi dalam waktu yang lama, sehingga dapat merusak organ tubuh seperti jantung, ginjal, mata, dan saraf. Menurut laporan International Diabetes *Federation (IDF)*, diabetes telah menjadi masalah kesehatan global yang membutuhkan perhatian serius karena dampaknya yang besar terhadap kualitas hidup masyarakat [1].

Di Indonesia sendiri, kasus diabetes juga terus meningkat. Banyak penderita yang tidak menyadari bahwa dirinya berisiko mengidap diabetes karena gejala awalnya sering tidak terasa. Oleh karena itu, deteksi dini risiko diabetes menjadi sangat penting agar pencegahan dan pengendalian bisa dilakukan lebih cepat [2].

Seiring berkembangnya teknologi, bidang data mining dan *machine learning* semakin banyak dimanfaatkan dalam dunia kesehatan. Teknologi ini dapat membantu menganalisis data dalam jumlah besar untuk menemukan pola yang tidak mudah dilihat secara manual. Beberapa penelitian terbaru menunjukkan bahwa teknik *machine learning* efektif digunakan untuk memprediksi risiko berbagai penyakit, termasuk diabetes [3].

Salah satu algoritma yang sering digunakan dalam klasifikasi adalah *Decision Tree*. Algoritma ini memiliki kelebihan karena mampu menghasilkan aturan-aturan keputusan yang mudah dipahami oleh manusia. Dengan *Decision Tree*, hasil prediksi tidak hanya berupa angka, tetapi juga dapat dijelaskan dalam bentuk aturan seperti: "jika kadar gula darah tinggi dan BMI besar, maka risiko diabetes tinggi". Hal ini membuat *Decision Tree* sangat cocok digunakan dalam bidang kesehatan yang membutuhkan hasil yang mudah diinterpretasikan oleh tenaga medis dan masyarakat umum [4].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengklasifikasikan risiko diabetes menggunakan algoritma *Decision Tree* dengan memanfaatkan dataset diabetes publik yang bersumber dari UCI *Machine Learning Repository*. Penelitian ini diharapkan dapat membantu memberikan gambaran faktor-faktor yang paling berpengaruh terhadap risiko diabetes, sekaligus menjadi referensi dalam pengembangan sistem deteksi risiko diabetes yang lebih dini dan akurat.

II. TINJAUAN PUSTAKA

A. Klasifikasi

Klasifikasi adalah proses mengelompokkan data ke dalam kategori tertentu berdasarkan pola yang dipelajari dari data sebelumnya. Dalam dunia data mining dan machine learning, klasifikasi sangat sering digunakan untuk memprediksi suatu kondisi, misalnya apakah seseorang

berisiko terkena suatu penyakit atau tidak. Teknik ini membantu peneliti dan praktisi untuk mengambil keputusan dengan lebih cepat karena sistem dapat mempelajari pola dari data yang sudah ada dan menerapkannya pada data baru [5].

B. Diabetes

Diabetes Mellitus merupakan penyakit kronis yang terjadi ketika tubuh tidak mampu mengatur kadar gula darah dengan baik, baik karena produksi insulin yang berkurang maupun karena tubuh tidak dapat menggunakan insulin secara efektif. Penyakit ini berkembang secara perlahan dan sering kali tidak menimbulkan gejala yang jelas pada tahap awal, sehingga banyak orang tidak menyadari bahwa dirinya berisiko. Berbagai penelitian menunjukkan bahwa jumlah penderita diabetes terus meningkat setiap tahun, baik di tingkat global maupun di Indonesia, sehingga diperlukan metode deteksi dini yang efektif [6].

C. Algoritma

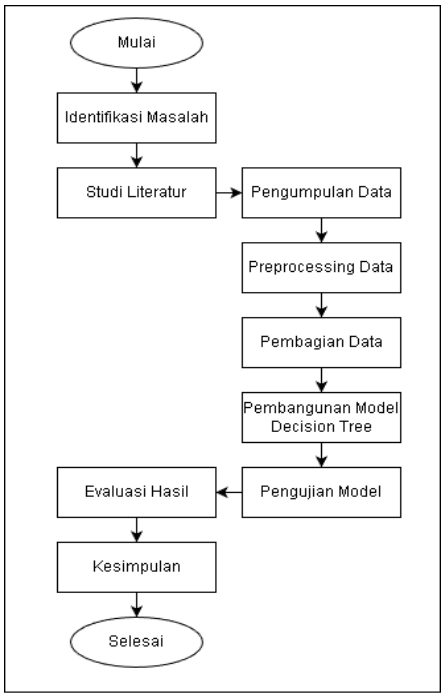
Algoritma dalam konteks *machine learning* dapat dipahami sebagai langkah-langkah logis yang digunakan komputer untuk mempelajari pola dari data dan membuat prediksi. Berbeda dengan pemrograman biasa yang menggunakan aturan tetap, *machine learning* memungkinkan sistem belajar dari data sehingga kinerjanya dapat semakin baik seiring waktu. Banyak penelitian menunjukkan bahwa pemilihan algoritma yang tepat sangat berpengaruh terhadap akurasi hasil prediksi, khususnya dalam bidang kesehatan [7].

D. *Decision Tree*

Decision Tree adalah salah satu algoritma yang bekerja dengan membentuk struktur seperti pohon yang berisi aturan-aturan keputusan yang mudah dipahami. Cara kerjanya menyerupai cara manusia mengambil keputusan, yaitu dengan menanyakan kondisi tertentu secara bertahap hingga diperoleh suatu kesimpulan. Karena tampilannya yang visual dan mudah dipahami, algoritma *Decision Tree* sangat cocok digunakan dalam penelitian di bidang kesehatan, termasuk untuk memprediksi risiko diabetes. Penelitian internasional dan nasional sudah banyak membuktikan bahwa *Decision Tree* mampu memberikan hasil yang cukup akurat sekaligus mudah dijelaskan [8].

III. METODE PENELITIAN

Pada penelitian ini sebuah metode penelitian dalam bentuk *flowchart* untuk menggambarkan alur kerja penelitian secara jelas dan terstruktur, mulai dari tahap awal hingga akhir. Penelitian ini bertujuan untuk mengklasifikasikan risiko penyakit diabetes menggunakan algoritma *Decision Tree*, yaitu salah satu metode dalam *machine learning* yang bekerja dengan membentuk struktur pohon keputusan berdasarkan data yang dianalisis. Algoritma ini banyak digunakan karena hasilnya mudah dipahami dan mampu membantu dalam memprediksi tingkat risiko diabetes berdasarkan faktor-faktor yang dimiliki oleh seseorang [9]. Dengan adanya kerangka penelitian ini, diharapkan alur penelitian dapat dipahami dengan lebih mudah dan sistematis.



Gambar 1. Metode Penelitian

A. Indentifikasi Masalah

Pada tahap ini, peneliti mengamati permasalahan yang ada di masyarakat, khususnya terkait meningkatnya jumlah penderita diabetes dan kurangnya metode yang efektif untuk memprediksi risiko penyakit ini sejak dini. Tahap ini penting karena menjadi dasar utama dalam menentukan arah dan fokus penelitian

B. Studi Literatur

Setelah masalah ditentukan, peneliti mempelajari berbagai jurnal, buku, dan penelitian terdahulu yang berkaitan dengan diabetes, data mining, dan algoritma Decision Tree. Tujuannya adalah agar penelitian memiliki dasar teori yang kuat dan tidak mengulang penelitian yang sudah ada, melainkan mengembangkan atau melengkapinya [10].

C. Pengumpulan data

Pada tahap ini, data dikumpulkan dari sumber sekunder yang terpercaya, yaitu dataset *Pima Indians Diabetes* yang berasal dari UCI *Machine Learning Repositor***y** (dapat diakses melalui Kaggle). Pemilihan sumber ini didasarkan pada validitas data yang telah banyak digunakan sebagai standar pengujian dalam penelitian *data mining* kesehatan. Data ini berisi informasi klinis seperti kadar glukosa, tekanan darah, berat badan (BMI), dan usia pasien yang sangat dibutuhkan dalam proses analisis risiko diabetes.

D. *Preprocessing Data*

Data yang sudah dikumpulkan belum tentu langsung bisa digunakan. Oleh karena itu, dilakukan proses pembersihan data, seperti menangani data yang kosong, menghilangkan data yang duplikat, dan memperbaiki data yang tidak

konsisten. Tahap ini sangat penting karena kualitas data yang baik akan menghasilkan model yang lebih akurat.

E. Pembagian Data

Data yang sudah bersih kemudian dibagi menjadi dua bagian, yaitu data latih *(training)* dan data uji *(testing).* Pembagian ini dilakukan agar model dapat belajar dari sebagian data dan diuji menggunakan data yang belum pernah dilihat sebelumnya, sehingga hasil pengujian menjadi lebih objektif.

F. Pembanguan Model Decision Tree

Pada tahap ini, algoritma Decision Tree digunakan untuk membuat model yang dapat mempelajari pola hubungan antara variabel input dan status risiko diabetes. Model ini menghasilkan struktur pohon keputusan yang mudah dipahami melalui aturan-aturan *if–then* [11].

G. Pengujian Model

Model yang telah dibuat kemudian diuji menggunakan data uji untuk melihat kesesuaian hasil prediksi dengan kondisi sebenarnya. Tahap ini bertujuan untuk mengetahui seberapa baik performa model dalam memprediksi risiko diabetes [12].

H. Evaluasi Model

Pada tahap ini, dilakukan analisis hasil pengujian menggunakan metrik evaluasi seperti akurasi, precision, recall, dan confusion matrix. Dari tahap ini, peneliti dapat mengetahui kelebihan dan kekurangan model yang telah dibuat [13] .

I. Kesimpulan

Tahap akhir penelitian adalah menyusun kesimpulan berdasarkan hasil analisis dan memberikan saran untuk pengembangan penelitian selanjutnya. Tahap ini penting untuk menegaskan kontribusi ilmiah dari penelitian yang telah dilakukan.

IV. HASIL DAN PEMBAHASAN

A. Sumber Data

Penelitian ini menggunakan dataset *Pima Indians Diabetes* yang diperoleh dari repositori UCI *Machine Learning* (yang diakses melalui Kaggle). Dataset ini terdiri dari 768 data rekam medis pasien dengan 8 atribut fitur klinis dan 1 atribut label kelas. Sebelum dilakukan pemrosesan lebih lanjut, dilakukan observasi terhadap sampel data untuk memastikan validitasnya. Berikut adalah cuplikan sampel data yang digunakan:

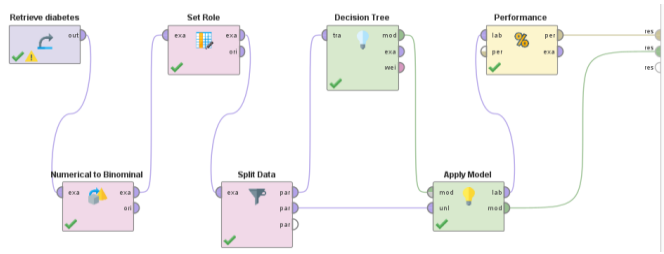
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

Gambar 2. Sampel Data

Cuplikan data ini memperlihatkan karakteristik atribut numerik dan variasi nilai pada setiap fitur medis. Kolom terakhir, yaitu *Outcome*, merepresentasikan label kelas target yang bernilai 1 (Positif) atau 0 (Negatif) yang akan menjadi acuan prediksi model.

B. Implementasi Sistem

Implementasi sistem klasifikasi risiko diabetes dibangun menggunakan perangkat lunak *RapidMiner*. Desain proses dirancang secara sistematis yang terdiri dari rangkaian operator *data mining* yang saling terhubung. Alur kerja dimulai dari operator *Retrieve* untuk memuat dataset, dilanjutkan dengan tahap *preprocessing* yang mencakup konversi tipe data label dan penetapan atribut target. Selanjutnya, aliran data masuk ke operator *Split Data* untuk membagi data latih dan uji, sebelum akhirnya diproses oleh algoritma inti *Decision Tree*. Visualisasi desain proses lengkap yang telah diimplementasikan dapat dilihat pada Gambar 1 berikut:



Gambar 3. Desain Alur Proses Klasifikasi

Berdasarkan Gambar 1, alur kerja sistem klasifikasi terdiri dari beberapa tahapan operator yang memiliki fungsi spesifik sebagai berikut:

1. *Retrieve*: Memasukan dataset *Pima Indians Diabetes* ke dalam lembar kerja RapidMiner agar siap diproses.
2. *Numerical to Binominal*: Operator ini berfungsi mengubah atribut target 'Outcome' yang semula bertipe numerik (0 dan 1) menjadi tipe nominal (kategori). Langkah ini krusial untuk menegaskan bahwa tugas yang dilakukan adalah klasifikasi (membedakan kelas), bukan regresi (memprediksi nilai angka).
3. *Set Role*: Operator ini berfungsi untuk menetapkan peran (*role*) variabel dalam dataset. Atribut *Outcome* didefinisikan secara eksplisit sebagai *Label* (target prediksi), sedangkan atribut lainnya ditetapkan sebagai fitur input (*regular attributes*).
4. *Split Data*: Operator ini berfungsi untuk mempartisi dataset menjadi dua bagian terpisah guna validasi model. Pembagian dilakukan dengan rasio 0.8 (80%) sebagai data latih (*training data*) untuk pembentukan model, dan 0.2 (20%) sebagai data uji (*testing data*) untuk evaluasi.

5. *Decision Tree*: Ini adalah operator inti pemodelan yang berfungsi menjalankan algoritma pohon keputusan. Algoritma mempelajari pola dari data latih menggunakan kriteria *Information Gain* untuk menyusun aturan klasifikasi.
6. *Apply Model*: Operator ini berfungsi untuk menerapkan model yang telah terbentuk ke data uji (*testing data*). Hasilnya adalah label prediksi untuk setiap data baru yang belum pernah dilihat model sebelumnya.
7. *Performance (Binominal Classification)*: Operator evaluasi ini berfungsi untuk mengukur kinerja model dengan cara membandingkan hasil prediksi sistem dengan label data aktual. Output dari operator ini berupa matriks kebingungan (*confusion matrix*) serta nilai akurasi, presisi, *recall*, dan AUC.

C. Hasil Evaluasi Kinerja

Untuk mengukur keandalan model yang telah dibangun, dilakukan pengujian (*testing*) terhadap data uji (20% dari populasi). Evaluasi dilakukan menggunakan metode *Confusion Matrix* untuk memetakan distribusi prediksi benar dan salah.

Evaluasi kinerja model dilakukan untuk memvalidasi kemampuan prediksi sistem. Distribusi hasil klasifikasi, yang memetakan perbandingan antara prediksi model dengan data aktual (termasuk *True Positive* dan *False Negative*), disajikan dalam matriks kebingungan (*Confusion Matrix*) pada Gambar 4 di bawah ini:

```
PerformanceVector:
accuracy: 70.13%
ConfusionMatrix:
True:   false   true
false:  94      40
true:   6       14
AUC: 0.736 (positive class: true)
precision: 70.00% (positive class: true)
ConfusionMatrix:
True:   false   true
false:  94      40
true:   6       14
recall: 25.93% (positive class: true)
ConfusionMatrix:
True:   false   true
false:  94      40
true:   6       14
f_measure: 37.84% (positive class: true)
ConfusionMatrix:
True:   false   true
false:  94      40
true:   6       14
```

Gambar 4. Hasil Uji Coba *Confusion Matrix*

Berdasarkan Gambar 4, hasil pengukuran kinerja model dapat dianalisis sebagai berikut:

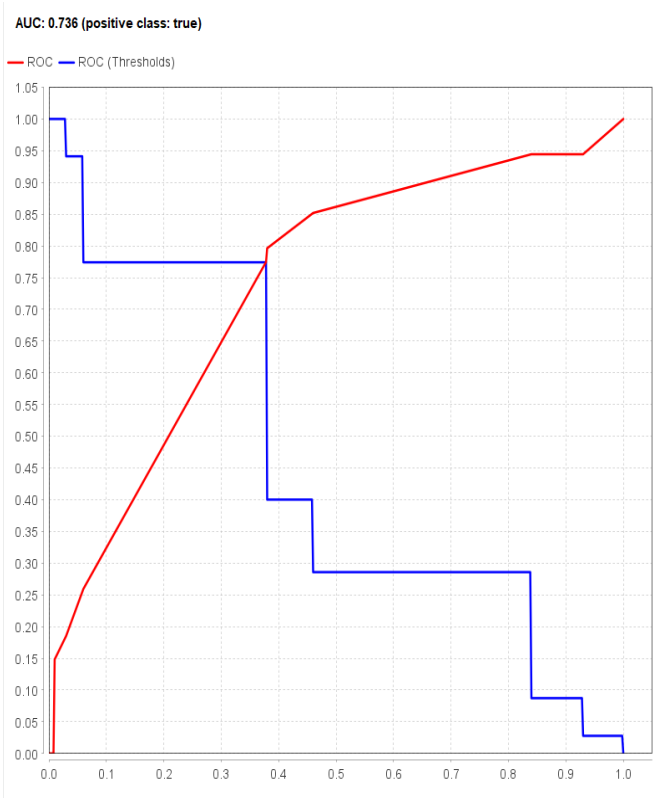
1. *Accuracy*: Secara global, model *Decision Tree* berhasil memprediksi status kesehatan pasien dengan benar sebanyak 70.13% dari total data uji. Angka ini menunjukkan bahwa model memiliki kemampuan yang cukup moderat (sedang) dalam membedakan antara pasien diabetes dan non-diabetes secara umum.

2. *Precision*: Nilai presisi sebesar 70.00% menunjukkan tingkat "kepercayaan" sistem terhadap prediksi positif. Artinya, ketika model memvonis seorang pasien "Berisiko Diabetes", probabilitas diagnosa tersebut benar adalah sebesar 70%. Nilai ini tergolong baik, yang mengindikasikan bahwa model cukup handal dalam meminimalisir alarm palsu (*False Positive*). Model tidak sembarangan memvonis orang sehat sebagai sakit.

3. *Recall*: Berbeda dengan presisi, nilai *recall* tercatat rendah di angka 25.93%. Nilai ini menunjukkan bahwa dari seluruh pasien yang secara fakta medis menderita diabetes, model baru mampu mendeteksi sekitar 26% di antaranya, sedangkan 74% sisanya terklasifikasi sebagai *False Negative*. Rendahnya nilai *recall* ini mengindikasikan bahwa model, dalam konfigurasi saat ini, lebih condong bersifat 'konservatif' sangat berhati-hati dalam memberikan label positif. Rendahnya nilai recall ini mengindikasikan bahwa model cenderung konservatif dalam memberikan label positif. Hal ini kemungkinan besar disebabkan oleh ketidakseimbangan kelas (*imbalanced class*) pada dataset, di mana jumlah sampel data negatif (sehat) jauh lebih mendominasi, sehingga algoritma kesulitan mengenali pola minoritas pada kelas positif (sakit).

4. *F-Measure*: F-Measure menggambarkan keseimbangan antara Presisi dan Recall dalam satu angka tunggal. Nilai yang rendah (37.84%) menegaskan adanya ketimpangan performa model, di mana model sangat presisi namun kurang sensitif. Nilai yang relatif kecil ini disebabkan oleh ketimpangan yang signifikan antara Presisi (tinggi) dan Recall (rendah). Hal ini mengonfirmasi bahwa kinerja model saat ini lebih menitikberatkan pada ketepatan prediksi positif (*Precision*) dibandingkan kemampuan menjaring seluruh kasus positif (*Recall*).

Selain metrik di atas, kualitas klasifikasi juga divalidasi menggunakan kurva ROC (*Receiver Operating Characteristic*) untuk melihat nilai AUC (*Area Under Curve*) sebagaimana ditampilkan pada Gambar 5.



Gambar 5. Grafik Kurva ROC

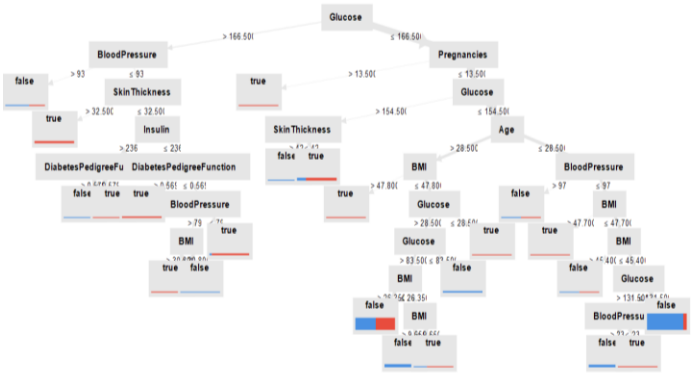
Hasil eksperimen menunjukkan nilai AUC tercatat sebesar 0.736. Secara statistik, nilai ini menunjukkan bahwa model memiliki kemampuan diskriminasi yang cukup memadai (*fair performance*) dalam mengidentifikasi risiko diabetes. Meskipun terdapat irisan antara kelas yang menyebabkan beberapa kesalahan prediksi, nilai AUC di atas 0.7 menegaskan bahwa atribut klinis yang digunakan terutama Glukosa dan BMI memiliki pola yang cukup kuat untuk dikenali oleh algoritma, sehingga model ini layak dipertimbangkan sebagai alat bantu skrining awal.

D. Analisis Pola Keputusan dan Faktor Determinan

Visualisasi pohon keputusan di bawah ini menggambarkan bagaimana sistem memecah kompleksitas data pasien menjadi logika 'JIKA-MAKA' yang sederhana. Melalui grafik ini, dapat diamati variabel mana saja yang menjadi penentu utama dalam klasifikasi risiko diabetes dan di titik ambang batas (*threshold*) mana seorang pasien dikategorikan berisiko tinggi. Berikut adalah visualisasi struktur pohon keputusan yang dihasilkan:

1. Analisis Struktur Pohon Keputusan (*Tree Graph*)

Selain metrik evaluasi statistik, penelitian ini menghasilkan tiga luaran (*output*) utama yang merepresentasikan nilai fungsional dari algoritma *Decision Tree*. Ketiga luaran tersebut meliputi visualisasi struktur keputusan, kemampuan prediksi status pasien, dan ekstraksi pengetahuan mengenai faktor determinan penyakit.



Gambar 6. Visualisasi Tree (Tight)

Berdasarkan visualisasi yang dihasilkan pada Gambar 6, struktur pohon keputusan membagi populasi pasien ke dalam tiga jalur logika utama (*Decision Paths*) sebagai berikut:

1) Identifikasi Risiko Tinggi (*High Risk Path*)
Logika**:** JIKA Glucose > 127.5 MAKA Class = Positif (1). Algoritma mengidentifikasi bahwa pasien dengan kadar glukosa plasma di atas 127.5 mg/dL memiliki probabilitas sangat tinggi untuk menderita diabetes. Pada visualisasi grafik, hal ini ditunjukkan oleh ketebalan batang (*bar*) warna kelas positif yang dominan pada cabang sebelah kiri. Jalur ini menunjukkan bahwa hiperglikemia (gula darah tinggi) adalah indikator tunggal yang cukup kuat untuk memvonis risiko tanpa perlu melihat atribut lain.
2) dentifikasi Kelompok Sehat (*Low Risk Path*)
Logika: JIKA Glucose ≤ 127.5 & BMI ≤ 29.950 MAKA Class = Negatif (0). Pada pasien dengan kadar glukosa yang terkontrol (≤ 127.5), algoritma melakukan pemeriksaan lanjutan terhadap berat badan. Jika pasien memiliki Indeks Massa Tubuh (BMI) di bawah 29.950 (kategori berat badan normal hingga *overweight* ringan), sistem memprediksi pasien tersebut Sehat/Negatif. Ini mengonfirmasi bahwa gula darah normal yang dikombinasikan dengan berat badan terjaga adalah indikator kesehatan metabolik yang kuat.
3) Identifikasi Risiko Bersyarat (*Conditional Risk Path*)
Logika: JIKA Glucose ≤ 127.5 & BMI > 29.950 & Age > 28.5 MAKA Class = Positif (1). Algoritma menemukan pola bahwa pasien dengan gula darah normal tetap berisiko terkena diabetes apabila memenuhi dua syarat: mengalami obesitas (BMI > 29.950) dan berusia di atas 28.5 tahun. Jalur ini menangkap kelompok pasien "tersembunyi" di mana risiko diabetes bukan disebabkan oleh lonjakan gula

darah semata, melainkan oleh kombinasi resistensi insulin akibat obesitas dan faktor degeneratif usia.

2. Implikasi Status Pasien
   Dalam penerapannya sebagai sistem pendukung keputusan medis (*Clinical Decision Support System*), model ini memberikan status prediksi yang dapat digunakan untuk triase awal pasien:

   1) Peringatan Dini (*Early Warning*): Model sangat efektif mendeteksi pasien yang memiliki profil risiko gabungan (Gula Darah Normal + Obesitas + Usia Dewasa). Kelompok ini sering kali luput dari pemeriksaan diabetes konvensional karena kadar gulanya belum mencapai ambang batas diabetik, namun algoritma sudah menandainya sebagai "Berisiko" (Prediksi 1).
   2) Objektivitas Diagnosa: Dengan mengikuti aturan *threshold* yang kaku (misal: batas BMI 29.950), sistem menghilangkan subjektivitas dalam penilaian risiko, memberikan standar diagnosa yang konsisten bagi setiap pasien.

3. Faktor Determinan (*Feature Importance*)
   Analisis *Decision Tree* mengungkap hierarki kepentingan fitur (*Feature Importance*) berdasarkan posisi atribut dalam struktur pohon. Atribut yang berada di level atas memiliki nilai *Information Gain* terbesar. Berdasarkan struktur pohon yang terbentuk, ditemukan urutan determinan risiko sebagai berikut:

   1) Determinan Utama (Ranking 1): Kadar Glukosa (*Glucose*) Konsisten menempati posisi *Root Node* (Akar). Temuan ini memvalidasi fakta medis bahwa kadar glukosa adalah variabel paling diskriminatif. Perubahan nilai pada atribut ini memberikan dampak terbesar terhadap perubahan status prediksi.
   2) Determinan Sekunder (Ranking 2): Obesitas (*BMI*) Muncul sebagai simpul keputusan kedua. Hal ini menunjukkan bahwa BMI adalah "penyaring" utama bagi pasien dengan gula darah normal. Obesitas teridentifikasi sebagai faktor komorbiditas terbesar yang memicu risiko diabetes pada dataset ini.
   3) Determinan Tersier (Ranking 3): Usia (*Age*) Muncul pada level keputusan ketiga. Faktor usia menjadi penentu akhir (*tie-breaker*) pada pasien yang memiliki gula darah normal namun mengalami obesitas.

V. KESIMPULAN

Penelitian ini menyimpulkan bahwa algoritma *Decision Tree* memiliki kinerja yang memadai (*fair performance*) dalam mengklasifikasikan risiko diabetes, dibuktikan dengan capaian nilai AUC sebesar 0.736 dan akurasi 70.13%. Analisis ekstraksi aturan keputusan (*rule extraction*) secara spesifik mengidentifikasi bahwa kadar Glukosa di atas 127.5 mg/dL merupakan determinan risiko paling dominan, sementara kombinasi BMI (>29.950) dan Usia (>28.5 tahun) menjadi indikator krusial pada pasien dengan kadar gula darah normal. Meskipun model menunjukkan keandalan tinggi dalam memprediksi kasus positif (Presisi 70.00%), rendahnya nilai *recall* (25.93%) mengindikasikan adanya keterbatasan sensitivitas model dalam mendeteksi keseluruhan populasi penderita, sehingga penggunaannya saat ini lebih disarankan sebagai instrumen pendukung keputusan medis (*decision support system*) dengan validasi lanjutan.

VI. SARAN

Berdasarkan temuan dan keterbatasan penelitian, saran untuk pengembangan selanjutnya adalah:

1. Menerapkan teknik penanganan ketidakseimbangan data (*imbalanced data handling*) seperti SMOTE (*Synthetic Minority Over-sampling Technique*) untuk meningkatkan nilai *Recall* (sensitivitas) agar model lebih peka dalam menjaring pasien positif.
2. Menguji algoritma *ensemble* seperti *Random Forest* atau *Gradient Boosting* untuk membandingkan apakah performa klasifikasi dapat ditingkatkan secara signifikan dibandingkan *Decision Tree* tunggal.

DAFTAR PUSTAKA

[1] R. Daud, S. Rahma, and S. F. M. Arsad, “Hubungan Dukungan Keluarga dengan Kepatuhan Minum Obat pada Pasien Diabetes Melitus Disertai Hipertensi Di Wilayah Kerja Puskesmas Kabila,” *J. Kolaboratif Sains*, vol. 8, no. 7, pp. 4869–4879, 2025, doi: 10.56338/jks.v8i7.8322.

[2] R. Fadila, M. P. Via, A. A. I. C. Dewiyani, and A. Ardhiasti, “Analisis Pencapaian Indikator Kapitasi Berbasis Kinerja Pada Masa Pandemi Covid 19,” *J. Kesehat. Qamarul Huda*, vol. 11, no. 1, pp. 241–249, 2023, doi: 10.37824/jkqh.v11i1.2023.446.

[3] A. Hennebelle, L. Ismail, H. Materwala, J. Al Kaabi, and R. Janardhanan, “Secure and Privacy-Preserving Automated Machine Learning Operations into End-to-End Integrated IoT-Edge-Artificial Intelligence-Blockchain Monitoring System for Diabetes Mellitus Prediction,” *arXiv*, pp. 1–41, 2023.

[4] E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, and S. Karnila, “Prediksi Penyakit Diabetes Melitus Menggunakan Algoritma Machine Learning,” *J. Sist. dan Teknol. Inf.*, vol. 12, no. 4, pp. 760–766, 2024,

doi: 10.26418/justin.v12i4.84620.

[5] A. Wijaya and W. Bismi, "Penerapan Algoritma Machine Learning Dalam Mengklasifikasi Data Masa Studi di Indonesia Berdasarkan Jenis Kelamin," *JIEET (Journal Inf. Eng. Educ. Technol.*, vol. 08, no. 02, pp. 62–70, 2024.

[6] M. Wahidin *et al.*, "Projection of diabetes morbidity and mortality till 2045 in Indonesia based on risk factors and NCD prevention and control programs," *Sci. Rep.*, pp. 1–17, 2024, doi: 10.1038/s41598-024-54563-2.

[7] P. Data, I. Kementerian, K. Republik, and K. Kunci, "PENERAPAN MACHINE LEARNING DALAM PREDIKSI TINGKAT KASUS PENYAKIT DI INDONESIA," *J. Inf. Syst. Manag.*, vol. 5, no. 1, pp. 40–45, 2023.

[8] N. Wayan, E. Rosiana, I. M. S. Putra, E. Simanungkalit, T. Informasi, and U. Udayana, "Comparison of decision tree and naive bayes methods in glioma classification based on clinical and molecular factors," *J. Mandiri IT*, vol. 13, no. 4, pp. 408–418, 2025.

[9] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. TRENDS*, vol. 02, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[10] M. Ridwan, B. Ulum, F. Muhammad, and U. I. Indragiri, "Pentingnya Penerapan Literature Review pada Penelitian Ilmiah," *J. Masohi*, vol. 02, no. 1, pp. 42–51, 2021.

[11] M. Faisal, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree," *J. Inform.*, vol. 10, no. 2, pp. 143–149, 2023.

[12] D.M.W., POWERS "EVALUATION : FROM PRECISION , RECALL AND F-MEASURE TO ROC , INFORMEDNESS , MARKEDNESS & CORRELATION," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[13] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.