Jurnal TICOM: Technology of Information and Communication
Volume 14, Nomor 3, Mei 2026 Hal: 129-136 E-ISSN: 2962-7982

Model Diagnosis Penyakit Diabetes Mellitus Menggunakan Smote dan Algoritme *Random Forest*

Fathimah Azzahra[1], Deni Mahdiana[2], Nidya Kusumawardhany[3]

[1,2,3]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta, Indonesia
Jl. Raya Ciledug, Petukangan Utara, Kebayoran Lama, Jakarta Selatan 12260
Email: [1]2112501016@student.budiluhur.ac.id, [2]deni.mahdiana@budiluhur.ac.id, [3]nidya.kusumawardhany@budiluhur.ac.id
(*: corresponding author)

Abstrak— **Diabetes Mellitus merupakan penyakit metabolik kronis yang prevalensinya terus meningkat secara global. Deteksi dini terhadap risiko diabetes sangat penting untuk mencegah munculnya komplikasi yang lebih serius di kemudian hari. Namun, proses ini menghadapi tantangan tersendiri, terutama dalam menentukan variabel risiko yang paling relevan dan memperoleh hasil klasifikasi yang akurat dalam mengidentifikasi individu yang berpotensi mengidap diabetes. Untuk menjawab tantangan tersebut, penelitian ini memanfaatkan algoritma Random Forest, yang dikenal memiliki performa baik dalam menangani masalah klasifikasi, termasuk dalam konteks medis seperti deteksi penyakit. Random Forest juga unggul dalam mengolah data berukuran besar serta mampu memberikan informasi mengenai tingkat kepentingan setiap fitur pada dataset. Penelitian ini juga menerapkan metode SMOTE (Synthetic Minority Oversampling Technique) guna mengatasi ketidakseimbangan distribusi data antar kelas, yang seringkali memengaruhi performa model secara keseluruhan. Dengan kombinasi algoritma Random Forest dan teknik SMOTE, diperoleh hasil evaluasi model dengan AUC sebesar 0.979, akurasi 91,36%, precision 98,46%, recall 72,32%, dan specificity 99,51%. Nilai-nilai ini menunjukkan bahwa model mampu memberikan performa yang tinggi dan andal. Pendekatan yang digunakan dalam penelitian ini tidak hanya meningkatkan akurasi dalam proses diagnosis dini diabetes, tetapi juga membantu mengungkap faktor-faktor risiko utama yang berperan, sehingga dapat menjadi dasar dalam strategi pencegahan dan pengendalian yang lebih tepat sasaran.**

Kata Kunci— **Diabetes, Random Forest, AUC, Klasifikasi, SMOTE.**

Abstract— ***Diabetes Mellitus is a chronic metabolic disease whose prevalence continues to increase globally. Early detection of diabetes risk is essential to prevent more serious complications in the future. However, this process faces several challenges, particularly in identifying the most relevant risk factors and achieving accurate classification results in predicting individuals at risk of developing diabetes.To address these challenges, this study utilizes the Random Forest algorithm, which is known for its strong performance in classification tasks, including in medical contexts such as disease detection. Random Forest is also effective in processing large datasets and provides valuable insights into the importance of each feature within the dataset. In addition, this study applies the SMOTE (Synthetic Minority Oversampling Technique) method to address class imbalance in the dataset, a common issue that can Negatively impact model performance. The combination of the Random Forest algorithm and SMOTE resulted in evaluation outcomes with an AUC of 0.979, accuracy of 91.36%, precision of 98.46%, recall of 72.32%, and specificity of 99.51%. These metrics indicate that the model demonstrates excellent and reliable performance.The approach used in this research not only improves the accuracy of early diabetes detection but also helps identify the key risk factors contributing to the disease, making it a valuable reference for more targeted prevention and control strategies.***

Keyword— ***Diabetes, Random Forest, AUC, Clasification, SMOTE.***

I. PENDAHULUAN

Diabetes mellitus (DM) merupakan penyakit metabolik kronis yang ditandai oleh gangguan fungsi insulin, sehingga tubuh tidak mampu mengatur kadar glukosa darah secara optimal dan menyebabkan kondisi hiperglikemia [1]. Prevalensi diabetes terus mengalami lonjakan dalam beberapa dekade terakhir, yang berdampak pada peningkatan beban tanggung jawab di tingkat personal maupun nasional secara global. Merujuk pada IDF Diabetes Atlas edisi ke-11, diabetes kini diderita oleh 11,1% populasi dewasa (usia 20-79 tahun), atau mencakup 1 dari 9 orang. Di balik angka tersebut, terdapat tantangan besar dalam deteksi dini, mengingat lebih dari 40% penderita belum teridentifikasi atau tidak menyadari bahwa mereka mengidap gangguan metabolik tersebut [2].

Survei Kesehatan Indonesia tahun 2023 memperlihatkan adanya lonjakan prevalensi diabetes yang cukup mencolok. Data menunjukkan bahwa jumlah penderita diabetes yang telah mendapat diagnosis dokter meningkat dari 1,5% pada 2018 menjadi 1,7% pada 2023. Pada kelompok usia 15 tahun ke atas, prevalensi juga mengalami kenaikan dari 2,0% menjadi 2,2%. Kondisi ini mencerminkan bahwa diabetes semakin menjangkau berbagai lapisan usia, sehingga menimbulkan dampak signifikan baik dari sisi ekonomi maupun social [3].

Melihat kerumitan dan besarnya dampak dari penyakit ini, sangat penting untuk mengambil tindakan deteksi dini untuk mengurangi risiko dan komplikasi yang mungkin terjadi. Pendekatan yang semakin banyak diterapkan dewasa ini adalah penggunaan teknologi analitik, khususnya data mining dan machine learning, dalam pengolahan dan analisis data. Pendekatan ini memungkinkan identifikasi risiko secara lebih cepat, sehingga penanganan dan pencegahan dapat dilakukan sebelum komplikasi serius berkembang.

Data Mining merupakan proses menggali informasi yang bernilai dari Kumpulan big data dengan memanfaatkan metode dari bidang matematika, statistika serta artificial intelligence. Hasil yang diperoleh melalu proses Data Mining dapat dimanfaatkan untuk mendukung proses pengambilan keputusan [4]. Data Mining adalah bagian penting dari proses

Knowledge Discovery in Database (KDD). KDD merujuk pada proses keseluruhan untuk mengubah data menjadi data bermanfaat melalui preprocessing, Data Mining, dan postprocessing [5]. Tujuan utama dari Data Mining adalah mengidentifikasi pola dan wawasan yang bermanfaat dari data. Hasil dari proses ini dapat digunakan dalam berbagai konteks, seperti membuat prediksi, mengoptimalkan proses, segmentasi konsumen, mendeteksi kecurangan, dan beragam aplikasi lainnya. Secara keseluruhan, Data Mining menyediakan landasan teoritis dan alat praktis untuk mengekstrak informasi bernilai dari data yang ada [6].

Machine learning adalah metode yang digunakan untuk memperoleh wawasan atau pengetahuan dari data. Bidang ini berada di titik temu antara statistik, kecerdasan buatan, dan ilmu komputer, serta sering disebut sebagai pembelajaran statistik atau analitik prediktif [7].

Terdapat metode pada penelitian - penelitian terdahulu yang telah melakukan deteksi dan prediksi penyakit diabetes menggunakan algoritma Extreme Gradient Boosting [8], Naive Bayes [9], Descision Tree [10], Logistic Regression [11], K-Nearest Neighbor [12], Support Vector Machine [13], dan Random Forest [14].

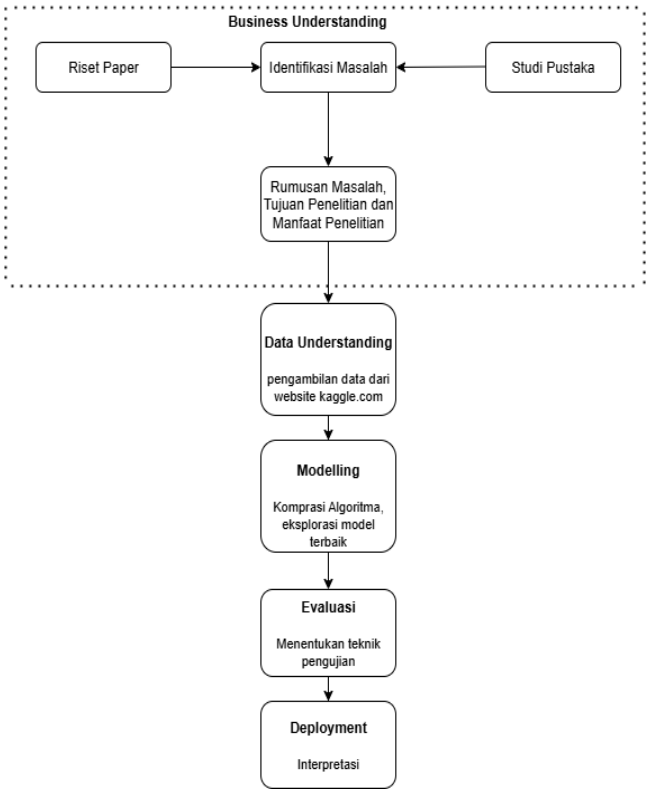
Setiap algoritma memiliki kelebihan dan keterbatasan. SVM unggul pada data berdimensi tinggi namun sulit dalam pemilihan kernel dan interpretasi [13]. Random Forest akurat serta mengurangi overfitting, tetapi membutuhkan komputasi besar [14], Regresi Logistik mudah dipahami untuk klasifikasi biner, meski tidak ideal untuk hubungan nonlinier [11]. KNN sederhana tanpa asumsi distribusi, namun kurang efektif pada dataset besar dan sensitif terhadap skala fitur [12]. Decision Tree intuitif tetapi rentan overfitting [10]. Naive Bayes efisien pada data besar, meski asumsi independensi sering tidak realistis [9]. XGBoost cepat dan akurat, tetapi memerlukan tuning parameter yang cermat [8].

Penelitian ini menggunakan algoritma Random Forest karena telah terbukti dapat dengan akurat mendeteksi risiko diabetes mellitus. Selain itu, SMOTE (*Synthetic Minority Oversampling Technique*) juga digunakan dalam penelitian ini untuk meningkatkan kinerja model. Dengan menggabungkan kedua algoritma tersebut, penelitian ini bertjuan untuk menyeimbangkan label class yang tidak seimbang menggunakan SMOTE dan algoritma Random Forest agar hasilnya lebih akurat.

II. METODE PENELITIAN

A. CRISP-DM

Proses penelitian ini mencakup serangkaian tahapan yang mengikuti metode CRISP-DM (*Cross Industry Standard for Data Mining*), sebagaimana disajikan pada Gambar 1:



Gambar 1. Tahapan Penelitian Menggunakan Metode CRISP-DM

Cross industry standart process for Data Mining (CRISPDM) merupakan sebuah model yang tidak terikat pada platform teknologi tertentu. Sehingga fleksibilitas untuk diimplementasikan di berbagai sektor industry. Keunggulan CRISP-DM terletak pada kemampuannya untuk membuat proyek Data Mining menjadi lebih mudah, lebih cepat, menjamin konsistensi hasil, serta memberikan keranka pengelolaan yang lebih terstruktur. Secara spesifik, CRIPSDM memberikan panduan operasional yang jelas mengenai tahapan dalam proses Data Mining dan batasan yang perlu diperhatikan [15].

B. Tahapan CRISP-DM

a. *Business Understanding*

Pada tahap business understanding akan dilakukan dengan cara riset pada jurnal-jurnal terdahulu. Pada tahap riset, jurnal dapat ditemukan secara online menggunakan google schoolar, elsivier, research gate, ataupun ieee. Jurnal yang diteliti hanya jurnal-jurnal yang berfokus pada prediksi ataupun deteksi pada penyakit diabetes menggunakan algoritma machine learning yang diterbitkan maksimal paling lama 5 tahun yang lalu. Tujuan dari riset jurnal terdahulu adalah untuk komparasi nilai akurasi yang didapatkan pada algortima yang digunakan masing-masing penelitian . setelah itu, bisa menentukan algoritma mana yang ingin digunakan dalam penelitian ini berdasarkan kecocokan dengan data serta akurasi yang paling tinggi

b. *Data Understanding*

Pada tahap data understanding akan dilakukan pengambilan data sekunder yang penulis dapatkan dari website www.kaggle.com. Data tersebut bersifat publik dengan total

100.000 *record* dan terdiri dari 9 atribut seperti pada Tabel 1 berikut.

TABEL I
ATRIBUT PADA DATASET

Nama Atribut	Tipe Atribut
Gender	*Categorical*
Age	*Numeric*
Hypertention	*Numeric*
Heart_Disease	*Numeric*
Smoking_History	*Categorical*
BMI	*Numeric*
HbA1c	*Numeric*
Blood_glucose_level	*Numeric*
Diabetes	*Numeric*

c. *Data Preparation*

Data sekunder yang telah peneliti temukan, sudah dalam keadaan bersih tanpa missing value, maka dari itu pada penelitian ini tidak diperlukan data cleansing.Berikut adalah tahap lainnya yang peneliti lakukan pada data preparation :

1) Merapihkan data dengan fungsi Delimited pada Microsoft Excel yang bertujuan untuk merapihkan data berdasarkan kolom yang sesuai.
2) Melakukan labelling pada atribut "diabetes" menggunakan operator Set Role pada RapidMiner.
3) Mengubah isi data pada atribut "*Hypertension*", "Heart_Disease", dan "Diabetes" dari yang tipe datanya integer menjadi binominal menggunakan operator Generate Attributes pada RapidMiner.
4) Melakukan metode SMOTE untuk menyeimbangkan label class.
5) Melakukan eksplorasi beberapa teknik validasi data, yaitu *Splitting data* dan *Cross Validation.*

d. *Modelling*

Tahap pemodelan data akan melibatkan berbagai metode pemrosesan data dengan menggunakan algoritma untuk mendapatkan hasil terbaik untuk penelitian ini. Proses pemodelan terdiri dari beberapa tahapan, termasuk membandingkan performa algoritma untuk menentukan model yang paling unggul. Penilaian dilakukan berdasarkan metrik evaluasi dari confusion matrix, bukan hanya dari akurasi semata. Setelah ditemukan model dengan kinerja terbaik, langkah selanjutnya adalah melakukan eksplorasi lebih lanjut terhadap model tersebut untuk menguji kinerja dan performanya secara mendalam.

e. *Evaluation*

Tahapan evaluasi kinerja model deteksi Diabetes Mellitus dilakukan dengan mengimplementasikan teknik confusion *matrix* sebagai basis pengukuran. Melalui pendekatan ini, peneliti dapat memvalidasi akurasi prediksi model baik pada kelompok penderita maupun kelompok sehat. Efektivitas model tersebut diuji menggunakan berbagai parameter kritis: *Accuracy* mengestimasi ketepatan prediksi secara general, sementara *Precision* menyoroti reliabilitas hasil pada klasifikasi positif. Di sisi lain, *Recall* atau *Sensitivity* mengevaluasi sejauh mana model mampu menjaring seluruh individu yang benar-benar sakit, didukung oleh *Specificity* untuk kelompok negatif. Sebagai indikator performa agregat, nilai AUC digunakan untuk menggambarkan kapasitas model dalam memisahkan kedua kelas diagnosis secara konsisten.

f. *Deployment*

Model klasifikasi yang telah dibangun dan dievaluasi diterapkan untuk diterapkan dalam situasi nyata pada tahap deployment. Model digunakan pada data medis dalam penelitian ini untuk mendeteksi risiko awal penyakit Diabetes Mellitus, khususnya pada orang yang menunjukkan gejala atau faktor risiko tertentu. Diharapkan model ini dapat membantu tenaga medis dan instansi kesehatan membuat keputusan yang lebih cepat dan tepat. Peneliti dan praktisi kesehatan dapat menggunakan hasil klasifikasi untuk membuat program pencegahan seperti skrining dini, instruksi gaya hidup sehat, dan kampanye kesadaran diabetes.

Tujuan dari penerapan model ini adalah untuk mendorong deteksi dini agar tindakan pencegahan dan intervensi dapat dilakukan sebelum komplikasi diabetes berkembang lebih lanjut.

III. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data yang telah dikumpulkum oleh peneliti bersifat sekunder yang didapatkan melalui situs www.kaggle.com. Dataset yang didapat terdiri dari 9 Attributes dengan total keseluruhan 100.000 *record*s yang sudah bersih tanpa adanya missing value. 8 Attributes akan digunakan sebagai regular Attribute dan 1 Attribute lainnya akan digunakan sebagai label class, seperti yang terlihat pada Tabel 1.

Atribut-atribut yang terdapat dalam dataset ini telah melalui proses validasi dan memiliki kejelasan yang baik. Setiap atribut memiliki hubungan atau korelasi yang erat dengan risiko terjadinya penyakit diabetes. Atribut gender menunjukkan jenis kelamin individu, yang terdiri dari kategori laki-laki, perempuan, dan lainnya. Jenis kelamin ini memiliki peranan dalam menentukan tingkat kerentanan seseorang terhadap diabetes karena faktor biologis yang berbeda pada masing-masing gender. Atribut *age* atau usia juga menjadi salah satu faktor penting karena semakin bertambah usia, risiko terkena diabetes cenderung meningkat. Dalam dataset ini, rentang usia berkisar antara 0 hingga 80 tahun.

Atribut hipertensi merujuk pada keadaan medis di mana tekanan vaskular berada dalam level tinggi secara berkelanjutan, yang secara klinis memperbesar probabilitas terjadinya diabetes. Selain tekanan darah, keberadaan penyakit jantung menjadi indikator penting dalam penilaian risiko ini; terdapat keterkaitan erat yang menunjukkan bahwa penurunan fungsi jantung sering kali beriringan dengan peningkatan kerentanan terhadap diabetes. Oleh karena itu, integrasi kedua atribut ini dalam model klasifikasi menjadi sangat krusial untuk memetakan komorbiditas yang dialami oleh pasien secara akurat.

Riwayat merokok seseorang direpresentasikan oleh atribut *smoking history*, yang terbagi menjadi beberapa kategori, yaitu *"not current", "former", "No Info", "current", "never",* dan *"ever"*. Riwayat merokok ini turut memengaruhi kondisi

kesehatan secara menyeluruh, termasuk dalam memperburuk komplikasi diabetes jika seseorang telah mengidapnya. Sementara itu, atribut BMI (*Body Mass Index* mengukur jumlah lemak tubuh berdasarkan berat dan tinggi badan. Dalam dataset ini, nilai BMI berkisar antara 10.16 hingga 71.55, dengan klasifikasi: Nilai IMT <18,5 menunjukkan status gizi kurang (*underweight*), IMT 18,5–24,9 mencerminkan status gizi normal, IMT 25–29,9 termasuk kelebihan berat badan (*overweight*), dan IMT ≥30 dikategorikan sebagai obesitas. Kelebihan berat badan berperan signifikan sebagai pemicu utama diabetes melitus tipe 2.

Atribut HbA1c level mengukur kadar gula darah rata-rata selama dua hingga tiga bulan terakhir, di mana nilai lebih dari 6.5% umumnya digunakan sebagai indikator bahwa seseorang mengidap diabetes. Selain itu, terdapat pula atribut blood glucose level atau kadar glukosa darah, yaitu jumlah gula dalam aliran darah pada satu waktu tertentu. Kadar glukosa darah yang tinggi merupakan ciri khas utama dari penderita diabetes. Terakhir, atribut diabetes merupakan variabel target dalam dataset ini, dengan nilai 1 menunjukkan individu positif mengidap diabetes dan nilai 0 menunjukkan tidak mengidap diabetes. Atribut ini dikaji berdasarkan semua variabel yang telah disebutkan sebelumnya.

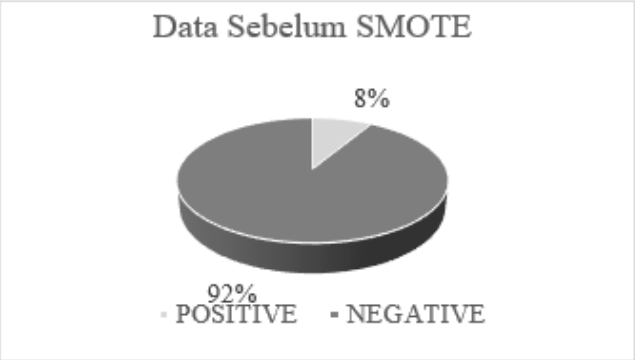
B. Data *Prepocessing*

Data yang berhasil didapatkan oleh peneliti sudah bersih tanpa adanya missing value maka, data preprocessing pada penelitian ini yang akan dilakukan yaitu data transformation dan SMOTE. Dalam penelitian ini akan diubah beberapa atribut yang memiliki *record* dengan nilai 0 atau 1, yang akan diubah menjadi *Positive* dan *Negative* pada label class, dan akan diubah menjadi yes dan no pada atribut heart_disease dan *hypertension*.

Penelitian ini menghadapi tantangan berupa distribusi kelas yang tidak merata, dengan kecenderungan data yang berpusat pada label negatif. Untuk mengoptimalkan kinerja algoritma, digunakan pendekatan SMOTE guna melakukan intervensi terhadap masalah *imbalanced class*. Teknik ini bekerja dengan meningkatkan jumlah observasi pada kelas minoritas melalui pembangkitan data sintetis, bukan sekadar melakukan duplikasi. Langkah ini diambil untuk memastikan bahwa model prediktif tidak hanya memiliki akurasi tinggi pada kelas mayoritas, tetapi juga memiliki sensitivitas yang baik dalam mendeteksi kategori positif [16].

Data memiliki distribusi yang sangat tidak seimbang, yaitu sekitar 92% untuk kelas negatif dan hanya 8% untuk kelas positif. Untuk meningkatkan proporsi kelas minoritas, dilakukan oversampling pada kelas positif menggunakan operator SMOTE dengan pengaturan *number of neighbors* sebanyak 5, yang berarti setiap data sintetis dibuat dengan mempertimbangkan 5 tetangga terdekat.

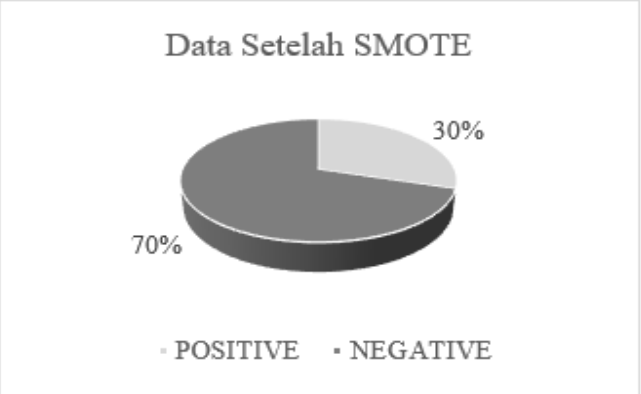
Parameter *upsampling size* ditentukan sebesar 30.714, sehingga jumlah data kelas positif meningkat dari semula 8.500 menjadi 39.214, dan komposisi data menjadi lebih seimbang, yaitu 70% kelas negatif (91.500 data) dan 30% kelas positif (39.214 data).



Gambar 2. Komposisi Data Sebelum SMOTE

Berikut adalah komposisi data awal 91.500 *record*s untuk kelas negatif dan hanya 8.500 *record*s untuk kelas positif dengan perbandingan 92:8 seperti pada Gambar 2 berikut.

Dan setelah dilakukan proses upsampling dengan perbandingan 70:30. Dengan jumlah kelas positif sebanyak 39.214 *record* dan kelas negatif sebanyak 91.500 *record*s seperti pada Gambar 3 berikut.



Gambar 3. Komposisi Data Setelah SMOTE

C. Hasil Data Latih dan Data Uji

Guna mencapai performa model yang optimal, pengujian dilakukan melalui pendekatan validasi ganda yang melibatkan pemisahan data (*data splitting*) serta *cross validation*. Peneliti membagi dataset menggunakan rasio komposisi 60:40, 70:30, dan 80:20 untuk memisahkan antara set pelatihan dan set pengujian. Set pelatihan digunakan secara spesifik untuk memfasilitasi proses pembelajaran algoritma, sedangkan set pengujian dimanfaatkan untuk memvalidasi ketepatan prediksi model. Selain pembagian secara manual, teknik *Cross Validation* turut diimplementasikan untuk melakukan iterasi pengujian, sehingga generalisasi model terhadap data baru dapat teruji dengan lebih akurat. Pada penelitian ini, *K-Fold Cross Validation* yang digunakan yaitu dengan *number of fold* sebanyak 10, yang artinya data akan dibagi ke dalam 10 partisi yang berbeda untuk proses pelatihan dan pengujian secara bergantian. *Cross Validation* ini diterapkan menggunakan algoritma *Decision Tree*, *Naïve Bayes*, *Random Forest*, dan *Logistic Regression*.

Berikut disajikan hasil komparasi nilai AUC yang diperoleh dari penerapan dua teknik validasi data, yaitu *data*

splitting dan *cross validation*, dengan perbandingan performa antara data yang diolah menggunakan SMOTE dan data tanpa SMOTE yang dpaat dilihat pada Tabel 2 berikut.

TABEL II
KOMPARASI AUC MENGGUNAKAN DATA SEBELUM SMOTE

Ratio Perbandingan	*Decision Tree*	*Random Forest*	*Naïve Bayes*	*Logistic Regression*
80:20	0.929	0.955	0.948	0.962
70:30	0.937	0.955	0.948	0.962
60:40	0.937	0.954	0.948	0.961
Cross Validation	0.923	0.974	0.949	0.962

Tabel 3 berikut menyajikan hasil komparasi nilai AUC pada data yang telah melalui proses SMOTE dengan menggunakan dua teknik validasi, yaitu *data splitting* dan *cross validation*.

TABEL III
KOMPARASI AUC MENGGUNAKAN DATA SETELAH SMOTE

Ratio Perbandingan	*Decision Tree*	*Random Forest*	*Naïve Bayes*	*Logistic Regression*
80:20	0.972	0.976	0.951	0.964
70:30	0.972	0.976	0.951	0.964
60:40	0.972	0.976	0.951	0.964
Cross Validation	0.973	**0.979**	0.951	0.963

Berdasarkan hasil komparasi nilai AUC berikut, dapat disimpulkan algoritma *Random Forest* dengan menggunakan *Cross Validation* dan SMOTE memiliki nilai AUC paling tinggi dibandingkan dengan algoritma lainnya. Meskipun model tanpa SMOTE memiliki akurasi lebih tinggi pada algortima *Random Forest*, SMOTE tetap digunakan karena bertujuan untuk mengatasi ketidakseimbangan kelas yang signifikan dalam dataset. Karena akurasi saja tidak menunjukkan kinerja yang sebenarnya dalam situasi di mana distribusi kelas tidak seimbang, dan model dengan akurasi tinggi dapat mengabaikan kelas minoritas yang sangat penting, seperti dalam hal deteksi penyakit diabetes mellitus ini.

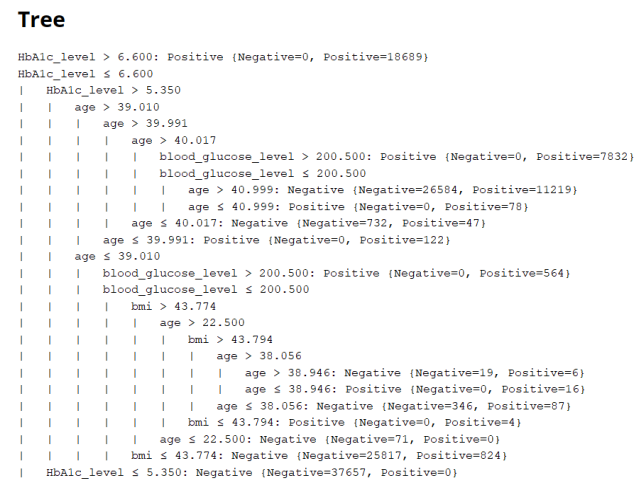
D. Pemodelan

1) Hasil Komparasi Model

Dalam rangka menentukan model terbaik, penelitian ini mengevaluasi algoritma *Random Forest*, *Logistic Regression*, *Naïve Bayes*, dan *Decision Tree* melalui serangkaian uji validasi. Setelah melalui tahap transformasi dan aplikasi SMOTE pada fase *preprocessing*, model diuji menggunakan teknik *data splitting* (rasio 8:2, 7:3, dan 6:4) serta *K-Fold Cross Validation* dengan 10 lipatan. Hasil eksperimen menunjukkan keunggulan metode *Cross Validation* dalam hal akurasi. Penilaian kualitas model secara menyeluruh kemudian dilakukan dengan mengekstraksi metrik dari *Confusion Matrix*, yang mencakup AUC, *precision*, *recall*, *specificity*, dan akurasi keseluruhan untuk menjamin validitas hasil deteksi.

2) Penyajian Model Terbaik

Berdasarkan hasil perbandingan kinerja model, algoritma *Random Forest* ditetapkan sebagai model terbaik. Oleh karena itu, pembahasan pada bagian ini difokuskan pada hasil pemrosesan menggunakan algoritma tersebut, mengingat kemampuannya dalam mengklasifikasikan penyakit diabetes dengan tingkat performa yang lebih tinggi dibandingkan *Naive Bayes*, *Decision Tree*, dan *Logistic Regression*. Algoritma hutan acak akan menghasilkan banyak pohon. ntuk mempersingkat trees yang ada maka digunakan number of trees 20 yang nantinya akan terbentuk 20 tree, dan menggunakan maximal depth 10 dengan akurasi sebesar 91.50% dan AUC 0,979, untuk membatasi pohon Keputusan yang terbentuk. Hasil pola dari data berupa peraturan If-Then, seperti pada Gambar berikut.



Gambar 4. *Rules If-Then Model Terbaik*

Rules:

1. Jika nilai HbA1c_level > 6.600, maka diklasifikasikan sebagai positif diabetes (*Positive* = 18.689, *Negative* = 0).
2. Jika nilai HbA1c_level ≤ 6.600, maka diperiksa nilai HbA1c_level lebih lanjut:
3. Jika HbA1c_level > 5.350, maka diperiksa usia (*age*):
4. Jika *age* > 39.010, maka diperiksa kembali nilai usia:
5. Jika *age* > 39.991, maka periksa kembali usia. Jika *age* > 40.017, maka periksa *blood_glucose_level*:
6. Jika *blood_glucose_level* > 200.500, maka diklasifikasikan sebagai positif diabetes (*Positive* = 7.832, *Negative* = 0).
7. Jika *blood_glucose_level* ≤ 200.500, maka periksa usia:
8. Jika *age* > 40.999, maka diklasifikasikan sebagai negatif diabetes (*Positive* = 11.219, *Negative* = 26.584).
9. Jika *age* ≤ 40.999, maka diklasifikasikan sebagai positif diabetes (*Positive* = 78, *Negative* = 0).
10. Jika *age* ≤ 40.017, maka diklasifikasikan sebagai negatif diabetes(*Positive* = 47, *Negative* = 732).
11. Jika *age* ≤ 39.991, maka diklasifikasikan sebagai positif diabetes (*Positive* = 122, *Negative* = 0).
12. Jika *age* ≤ 39.010, maka periksa *blood_glucose_level*:
13. Jika *blood_glucose_level* > 200.500, maka diklasifikasikan sebagai positif diabetes (*Positive* = 564, *Negative* = 0). 46
14. Jika *blood_glucose_level* ≤ 200.500, maka periksa nilai BMI:
15. Jika BMI > 43.774, maka periksa usia:
16. Jika *age* > 22.500, periksa kembali nilai BMI dan usia:

17. Jika BMI > 43.794, dan *age* > 38.056, maka: Jika *age* > 38.946, maka diklasifikasikan sebagai negatif diabetes (*Positive* = 6, *Negative* = 19).
18. Jika *age* ≤ 38.946, maka diklasifikasikan sebagai positif diabetes (*Positive* = 16, *Negative* = 0).
19. Jika *age* ≤ 38.056, maka diklasifikasikan sebagai negatif diabetes (*Positive* = 87, *Negative* = 346).
20. Jika BMI ≤ 43.794, maka diklasifikasikan sebagai positif diabetes (*Positive* = 4, *Negative* = 0).
21. Jika *age* ≤ 22.500, maka diklasifikasikan sebagai negatif diabetes (*Positive* = 0, *Negative* = 71).
22. Jika BMI ≤ 43.774, maka diklasifikasikan sebagai negatif diabetes (*Positive* = 824, *Negative* = 25.817).
23. Jika HbA1c_level ≤ 5.350, maka diklasifikasikan sebagai negatif diabetes (*Positive* = 0, *Negative* = 37.657)

3) Perhitungan dengan Algoritma Terpilih

Dalam penelitian ini digunakan sampel sebanyak 10 *record*, yang terdiri dari 5 data berlabel negatif dan 5 data berlabel positif, sebagai dasar perhitungan dengan algoritma terpilih sekaligus untuk tahap pengujian.

a) Perhitungan Entropy Dataset

Variabel diabetes pada memiliki total 10 *record* dengan total *Positive* diabetes 5 *record* dan total *Negative* diabetes 5 *record* sehingga didapatkan nilai entropy seperti pada Tabel 4 berikut.

TABEL IV
ENTROPY CLASS

Atribut	Jumlah	Positif	Negatif	Entropy
diabetes	10	5	5	1

$$Entropy(Y) = -\sum_i p(c|Y) log_2\, p(c|Y) \ldots\ldots \quad (1)$$
$$Entropy(Class) = -\left(\left(\frac{5}{10}\, log2\, \frac{5}{10}\right) + \left(\frac{5}{10}\, log2\, \frac{5}{10}\right)\right)$$
$$= -((0.5\,(-1) + 0.5\,(-1)))$$
$$= 1$$

b) Perhitungan Entropy Variabel *hypertension*

Variable hypertension memiliki total 10 *record* pada sample dataset dengan total *hypertension* "yes" yaitu 1 *record* dan total *hypertension* "No" 9 *record*, sehingga didapatkan hasilnya seperti pada Tabel 5 berikut

TABEL V
ENTROPY HYPERTENSION

Atribut		Positif	Negatif	Entropy
hypertension	Yes	1	0	0
	No	4	5	0.991

$$Entropy(Y) = -\sum_i p(c|Y) log_2\, p(c|Y)$$
$$Entropy(Yes) = -\left(\left(\frac{1}{1}\, log2\, \frac{1}{1}\right) + (0)\right)$$
$$= 0$$

$$Entropy(Y) = -\sum_i p(c|Y) log_2\, p(c|Y)$$
$$Entropy(No) = -\left(\left(\frac{4}{9}\, log2\, \frac{4}{9}\right) + \left(\frac{5}{9}\, log2\, \frac{5}{9}\right)\right)$$
$$= 0.991$$

c) Perhitungan *Information Gain*

Setelah perhitungan *entropy* untuk *variable class dan variable hypertension* selesai, perhitungan *Information Gain* akan dilakukan berdasarkan hasil *entropy* untuk *variable class* dan *variable hypertension.*

$$Gain(Y,a) = Entropy(Y) - \sum_{v \in value(a)} \frac{|Y_v|}{|Y_a|}\, Entropy(Y_v) \quad \ldots\ldots\ldots \quad (2)$$
$$Gain(Y,a) = Entropy(Class) - \sum_{v \in value(a)} \frac{|Y_v|}{|Y_a|} Entropy(hypertension)$$
$$= 1 - \left(\left(\frac{1}{10} \times 0\right) + \left(\frac{9}{10} \times 0.991\right)\right)$$
$$= 1 - 0.8919$$
$$= 0.1081$$

d) Perhitungan *Split Information*

Perhitungan *Split information* digunakan untuk mendapatkan hasil *gain ratio* dan akan dilakukan sebagai berikut:

$$Split\ Information = \sum_{i=1}^{c} \left(\frac{|Si|}{|S|}\right) log2 \left(\frac{|Si|}{|S|}\right) \ldots\ldots\ldots\ldots\ldots \quad (4)$$

$$Split\ Information = \left(\frac{1}{10} log2 \frac{1}{10}\right) + \left(\frac{9}{10} log2 \frac{9}{10}\right)$$

$$Split\ Information = 0.4689$$

e) Perhitungan *Gain Ratio*

Gain ratio didapatkan dengan cara membagi hasil *information gain* dengan *Split information* sebagai berikut:

$$Gain\ Ratio = \frac{Information\ Gain}{Split\ Information} \quad \ldots\ldots\ldots \quad (5)$$

$$Gain\ Ratio = \frac{0.1081}{0.4689}$$

$$= 0.2305$$

E. Pengujian

Pada tahap pengujian akan digunakan teknik pengujian dengan *Confusion Matrix* Pada tahap pemodelan dengan total 183.000 *record* dataset dengan total class *Positive* sebanyak 91.500 *record* dataset dan total class *Negative* sebanyak 91.500 *record* dataset. Berikut yang tertera pada table 4.16 merupakan hasil pengujian menggunakan *Confusion Matrix* dengan algoritma terbaik yaitu *Random Forest* menggunakan data latih uji dengan teknik 10 *Folds Cross Validation* dan *stratified random sampling* pada pada algoritma *Random Forest* dengan batasan pengujian menggunakan *number of tress* 5, *maximal depth* 5 serta dilakukannya *apply prunning.*

Tabel 6
Pengujian *Confusion Matrix*

	True Positive	***True Negative***	***Class Precision***
Pred. *Positive*	88.280	21.327	96.48%
Pred. *Negative*	3.220	70.173	95.61%
Class Recall	80.54%	76.69%	

Berikut adalah perhitungan *Accuracy* manual menggunakan rumus seperti di bawah ini

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad \text{.....(6)}$$

$$= \frac{158.453}{183.000} = 86.59\%$$

Berikut adalah perhitungan *Precision* manual menggunakan rumus seperti di bawah ini :

$$Precision = \frac{(TP)}{(TP+FP)} \quad \text{.....} \quad (7)$$

$$= \frac{(88.280)}{(88.280+3.220)} = \frac{(88.280)}{(91.500)}$$

$$= 96.48\%$$

Berikut adalah perhitungan *Recall* manual menggunakan rumus seperti di bawah ini :

$$Recall = \frac{(TP)}{(TP + FN)} \quad \text{.....} \quad (8)$$

$$= \frac{88.280}{(88.280 + 21.327)} = \frac{(88.280)}{(109.607)}$$

$$= 80.54\%$$

Berikut adalah perhitungan *Specificity* menggunakan rumus seperti dibawah ini

$$Specificity = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad \text{.....} \quad (9)$$

$$Specificity = \frac{70.173}{(70.173 + 21.327)} = 76.69\%$$

IV. PENUTUP

Berdasarkan rumusan masalah dan tujuan penelitian, dapat disimpulkan bahwa penerapan metode SMOTE terbukti mampu menyeimbangkan label kelas pada data penyakit diabetes, di mana jumlah label kelas Positive yang semula hanya 8.500 record berhasil meningkat menjadi 39.214 record, sementara label kelas Negative tetap sebanyak 91.500 record. Selanjutnya, penggunaan algoritma Random Forest terbukti memberikan hasil prediksi terbaik terhadap penyakit diabetes dengan capaian kinerja yang cukup tinggi, yakni akurasi sebesar 91,36%, precision 98,46%, recall 72,32%, specificity 99,51%, serta nilai AUC sebesar 0,979.

Adapun saran untuk penelitian berikutnya adalah agar dapat dikembangkan dengan mengombinasikan algoritma klasifikasi lainnya sehingga diperoleh peningkatan pada nilai akurasi, precision, recall, sensitivity, specificity, maupun AUC. Selain itu, peneliti maupun tenaga medis juga diharapkan mulai mempertimbangkan pemanfaatan algoritma machine learning seperti Random Forest, terutama jika dipadukan dengan teknik penyeimbangan data seperti SMOTE.

REFERENSI

[1] R. P. Febrinasari, T. A. Sholikah, D. N. Pakha, dan S. E. Putra, "Buku Saku Diabetes Melitus untuk Awam. Surakarta : UNS Press.," *Pnb. dan Pencetakan UNS (UNS Press.*, no. 1, hal. 79, 2020.
[2] *IDF Diabetes Atlas*. 2025.
[3] Kemenkes, "Survei Kesehatan Indonesia 2023," 2023. [Daring]. Tersedia pada: https://www.badankebijakan.kemkes.go.id/ski-2023-dalam-angka/
[4] Muttaqin *et al.*, *Pengenalan Data Mining*, no. July. 2023.
[5] S. Ningsih *et al.*, *pengantar DATA MINING & STATISTIK*. PT Penerbit Penamuda Media, 2024.
[6] F. SLN, "Basic Data Mining from A to Z," no. January, hal. 15–15, 2023.
[7] D. Chopra dan R. Khurana, *Introduction to Machine Learning with Python*. 2023. doi: 10.2174/9789815124422123010l.
[8] N. P. Miriyala, R. L. Kottapalli, G. P. Miriyala, G. Lorenzini, C. Ganteda, dan V. A. Bhogapurapu, "Diagnostic Analysis of Diabetes Mellitus Using Machine Learning Approach," *Rev. d'Intelligence Artif.*, vol. 36, no. 3, hal. 347–352, 2022, doi: 10.18280/ria.360301.
[9] C. D. K, I. P. C, R. Ameen, S. Kundur, dan S. G. M, "Diabetes Prediction using Machine Learning Algorithms," *Interantional J. Sci. Res. Eng. Manag.*, vol. 07, no. 02, hal. 1–9, 2023, doi: 10.55041/ijsrem17771.
[10] V. Vakil, S. Pachchigar, C. Chavda, dan S. Soni, "Explainable predictions of different machine learning algorithms used to predict Early Stage diabetes," no. Ml, 2021, [Daring]. Tersedia pada: http://arxiv.org/abs/2111.09939
[11] S. Yang dan H. Kong, "Diabetes Prediction Based on KNN , XGBoost , SVM and LR model," vol. 0, hal. 91–95, 2024, doi: 10.54254/2755-2721/104/20241142.
[12] M. Zhang, "Research on Diabetes Risk Prediction Using Multiple Machine Learning Models," hal. 0–8, 2024.
[13] S. Firdous, G. A. Wagai, dan K. Sharma, "A survey on diabetes risk prediction using machine learning approaches," *J. Fam. Med. Prim. Care*, vol. 6, no. 2, hal. 169–170, 2022, doi: 10.4103/jfmpc.jfmpc.
[14] L. P. Nguyen *et al.*, "The Utilization of Machine Learning Algorithms for Assisting Physicians in the Diagnosis of Diabetes," *Diagnostics*, vol. 13, no. 12, 2023, doi: 10.3390/diagnostics13122087.
[15] A. Massahiro, S. Instituto, D. P. Tecnológicas, D. S. Paulo, dan F. Univer-, "The evolution of CRISP-DM for Data Science : Methods , Processes and The evolution of CRISP-DM for Data Science : Methods ," no. October, 2024, doi: 10.5753/reviews.2024.3757.
[16] A. Nugroho dan E. Rilvani, "Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan." 2023.

Dikosongkan

Dikosongkan