Pengembangan Model *Machine Learning* untuk Deteksi Penyakit Diabetes Menggunakan Analisis *Gini Importance*

Muhammad Aulia Alfarisi[1], Budi Sumanto[1,*], I Putu Fadya Rachmawan[1]
[1]Departemen Teknik Elektro dan Informatika, Sekolah Vokasi, Universitas Gadjah Mada;
malfarisi80@mail.ugm.ac.id
putu.f@mail.ugm.ac.id
[*]Korespondensi: budi.sumanto@ugm.ac.id;

Abstract – Diabetes mellitus is a disease that causes blood sugar levels to rise and has the potential to cause more serious complications. Diabetes screening is very important to prevent this problem. This issue can be overcome by developing a machine learning system to predict diabetes. This research focuses on developing a machine learning classification model for diabetes detection. Five classification models such as Random Forest, Support Vector Machine, Logistic Regression, Linear Discriminant Analysis, and Artificial Neuron Network were evaluated to identify the most accurate model for predicting diabetes. The dataset used was taken from Kaggle under the name "Diabetes Prediction Dataset" with 100,000 data points. The results of feature contribution analysis using the Gini Importance method showed that the blood_glucose, bmi, and age features contributed the most to predicting diabetes. The model was trained using the hybrid resampling technique and tested using 10-fold cross validation and data division (80% training data and 20% testing data). Results showed that the Random Forest model had the highest accuracy at 87%, followed by Support Vector Machine (86%), Artificial Neural Network (83%), Logistic Regression (82%), and Linear Discriminant Analysis (81%). Based on this study, the Random Forest model with the Gini Importance method provides the best performance for early diabetes detection.

Keywords – diabetes detection, machine learning, feature importance, classification

Intisari – Diabetes mellitus adalah penyakit yang mengakibatkan kadar gula darah naik dan berpotensi menyebabkan komplikasi yang lebih serius. Skrining diabetes menjadi sangat penting agar tercegahnya terjadi masalah ini. Masalah tersebut dapat diatasi dengan perancangan sistem *machine learning* untuk memprediksi penyakit diabetes. Penelitian ini bertujuan untuk mengembangkan model klasifikasi machine learning untuk deteksi penyakit diabetes. Lima model klasifikasi seperti *Random Forest, Support Vector Machine, Logistic Regression, Linear Discriminant Analysis,* dan *Artificial Neuron Network* dibandingkan untuk menentukan model terbaik dalam memprediksi diabetes. Dataset yang digunakan diambil dari Kaggle dengan nama "*Diabetes Prediction Dataset*" sebanyak 100.000 data. Hasil Analisis kontribusi fitur menggunakan metode *Gini Importance* didapatkan hasil yaitu fitur *blood_glucose*, bmi, dan *age* merupakan fitur yang paling berkontribusi dalam memprediksi diabetes. Model dilatih menggunakan teknik *hybrid resampling* dan diuji menggunakan 10-*fold cross validation* serta pembagian data *training* (80%) dan data *testing* (20%). Hasil menunjukkan model *Random Forest* memiliki akurasi tertinggi sebesar 87%, diikuti oleh *Support Vector Machine* (86%), *Artificial Neuron Network* (83%), *Logistic Regression* (82%), dan *Linear Discriminant Analysis* (81%). Berdasarkan penelitian ini, model *Random Forest* dengan metode *Gini Importance* memberikan performa terbaik untuk deteksi dini diabetes.

Kata kunci – deteksi penyakit diabetes, machine learning, feature importance, klasifikasi

I. PENDAHULUAN

Menurut data *International Diabetes Federation* (IDF) diperkirakan 382 juta orang mengidap penyakit diabetes per tahun 2013 dan diestimasi naik sampai sekitar 592 juta orang pada tahun 2035. Berdasarkan jumlah tersebut, terdapat 175 juta pengidap diabetes yang belum terdiagnosis dan berisiko berkembang dan berlanjut menjadi komplikasi apabila tidak ada tindakan pencegahan [1].

Indonesia sendiri menganggap diabetes mellitus menjadi isu kesehatan yang membutuhkan atensi serius. Penyakit diabetes mellitus menjadikan Indonesia bertengger di peringkat ke-empat dengan prevalensi diabetes berada di urutan setelah India, China, dan Amerika Serikat. *World Health Organization* (WHO) memprediksi, pengidap Diabetes mellitus tipe 2 di Indonesia akan terus melonjak hingga 21,3 juta orang per tahun 2030 [2].

Diabetes mellitus yang semakin meningkat prevalensinya setiap tahun menuntut pendekatan yang lebih canggih dan efektif dalam pendeteksian. Pendekatan menggunakan *machine learning* dapat digunakan untuk deteksi awal diabetes untuk menurunkan kemungkinan terjadinya komplikasi serius yang berhubungan dengan penyakit diabetes [3]. Namun, sebagian besar penelitian sebelumnya masih berfokus pada penggunaan dimensi fitur yang luas, yang seringkali sulit diimplementasikan pada layanan kesehatan primer dengan fasilitas terbatas. Oleh karena itu, penelitian ini menawarkan kebaruan melalui pengembangan model klasifikasi yang parsimonious (sederhana namun kuat) dengan mengintegrasikan seleksi fitur berbasis *Gini Importance* dan teknik *hybrid resampling*. Kontribusi utama penelitian ini adalah membuktikan bahwa penggunaan hanya tiga fitur klinis fundamental (*blood_glucose*, bmi, dan *age*) mampu memberikan akurasi diagnosa yang optimal (87%), sehingga model ini lebih efisien dan aplikabel untuk diintegrasikan ke dalam sistem skrining kesehatan digital.

II. DASAR TEORI

Studi yang dilakukan oleh [4], merupakan salah satu contoh pendekatan yang berhasil dalam penggunaan *machine*

learning untuk mendeteksi penyakit diabetes. Mereka menerapkan berbagai metode *machine learning* seperti *Support Vector Machine* (SVM), *Naive Bayes*, dan *Light Gradient Boosting Machine* (*LightGBM*) untuk menganalisis data. Dataset yang digunakan berasal dari situs UCI dan diperoleh dari Rumah Sakit Diabetes Sylhet di Bangladesh. Hasil penelitian mereka menunjukkan bahwa model yang mereka usulkan mampu mengungguli berbagai teknik lainnya. Akurasi keseluruhan yang dicapai menunjukkan SVM mengungguli model lain dengan nilai akurasi sebesar 96,54% sedangkan model lainnya seperti *Naive Bayes classifier* dan *LightGBM* mencapai 93,27% dan 88,46% [4].

Terdapat kekurangan pada metodologi penelitian ini yaitu kurangnya proses pembersihan data dan *preprocessing*. Data *preprocessing* merupakan langkah yang sangat penting untuk menangani masalah yang terjadi data yang berjumlah besar seperti menghapus *noise*, redundansi dan data yang tidak relevan. Oleh karena itu diperlukan proses data *preprocessing* untuk membantu memastikan bahwa data yang digunakan data yang bersih, konsisten dan siap digunakan [5].

Penelitian lain karya [6] juga memberikan informasi penting dalam pengembangan model *machine learning* untuk mendeteksi diabetes. Penelitian dilakukan menggunakan dua model *machine learning*, yaitu *Logistic Regression* dan *Random Forest*, dan dataset yang diambil dari *Kaggle* dengan nama "*Early Stage Diabetes Risk Prediction Dataset*." yang mencakup 16 fitur numerik dan satu variabel target yang menunjukkan apakah pasien yang diuji negatif atau positif terhadap diabetes. Hasilnya menunjukkan bahwa model *Random Forest* unggul dengan akurasi mencapai 99,03%. Meskipun demikian, klasifikasi menggunakan *logistic regression* juga memberikan hasil yang layak dengan akurasi sebesar 94,23%. Secara garis besar, perbandingan 2 model *machine learning* yang digunakan berjalan dengan baik ditandai dengan akurasi pada masing-masing model yang tinggi [6].

Terkait dengan penelitian sebelumnya yang telah dibahas, penelitian lain yang dilakukan oleh [7], juga memiliki dampak signifikan dalam meningkatkan akurasi dari model *machine learning* untuk mendeteksi penyakit diabetes. Pada penelitian ini, mereka menerapkan 6 model yang berbeda seperti KNN, *Logistic Regression*, SVM, *Random Forest*, *LightGBM* dan *XGBoost* dengan pembagian data untuk *training* dan *test* sebesar 60%-40%, 70%-30%, 80%-20% untuk mendeteksi diabetes mellitus tipe 2. Dataset yang digunakan didapat dari arsip UCI yaitu *Pima Indian Diabetes Dataset*. Pada penelitian ini, dilakukan sebuah proses *Feature Engineering* yaitu *Feature Importance* dan *Recursive Feature Elimination*. Proses ini dilakukan untuk mengukur kontribusi masing-masing fitur pada akurasi dari model *machine learning* yang digunakan. Hal ini bertujuan untuk menyeleksi fitur dan mencari fitur yang paling penting untuk digunakan. Hasilnya menunjukkan bahwa *LightGBM* unggul dari model lain dengan akurasi sebesar 91,47% dengan pembagian data *training* dan tes sebesar 80% - 20 %.

Berdasarkan hasil penelitian sebelumnya, penelitian ini akan mengembangkan model *machine learning* untuk mendeteksi dini diabetes dengan menggunakan lima algoritma yaitu *Random Forest, Support Vector Machine, Logistic Regression, Linear Discriminant Analysis* dan *Multilayer Perceptron* atau *Artificial Neuron Network* seperti yang sudah diterapkan sebelumnya pada penelitian oleh [4, 6]. Selain itu, untuk meningkatkan akurasinya akan digunakan metode *Feature Importance* yang telah dijelaskan pada penelitian oleh [7] sehingga dataset yang telah didapat akan diukur kontribusinya pada masing-masing fitur atau atribut untuk mengetahui fitur yang paling penting untuk digunakan.

III. METODOLOGI

A. Deskripsi Data

Data yang digunakan dalam penelitian ini diperoleh dari Kaggle melalui *Diabetes Prediction Dataset* dengan total 100.000 sampel. Dataset ini terdiri dari 9 variabel, di mana variabel target memiliki distribusi kelas yang tidak seimbang yaitu 91.500 sampel negatif (91,5%) dan 8.500 sampel positif (8,5%). Delapan variabel sisanya merupakan variabel prediktor, namun penelitian ini melakukan eliminasi terhadap satu fitur yang dianggap tidak relevan pada tahap *data preprocessing* guna meningkatkan efisiensi komputasi dan menghindari *noise* pada model.

B. Data Preprocessing

Tahap pra-pemrosesan data diawali dengan penghapusan fitur '*smoking_history*' karena memiliki tingkat ambiguitas data yang tinggi dan korelasi yang rendah terhadap target prediksi, sehingga hanya 7 fitur prediktor yang dipertahankan. Selanjutnya, dilakukan pembersihan terhadap 3.854 data duplikat dan penghapusan entri '*Other*' pada variabel gender guna menjamin kualitas data. Untuk menangani ketidakseimbangan kelas, diterapkan teknik *hybrid resampling* (kombinasi *undersampling* dan SMOTE) yang diikuti dengan *feature scaling* menggunakan StandardScaler untuk menyeragamkan rentang nilai fitur. Tahap akhir dari pra-pemrosesan adalah pembagian data (*data splitting*) menjadi data latih sebesar 80% dan data uji sebesar 20% untuk kebutuhan evaluasi model.

C. Feature Importance

Feature Importance adalah sebuah metode untuk mengukur tingkat kontribusi dari variabel pada sebuah model *machine learning*. *Feature Importance* digunakan pada model untuk mengetahui fitur apa yang penting bagi performa model [8].

Breiman pada tahun 2001 menyebutkan bahwa salah satu metode *Feature Importance* yang paling umum digunakan adalah *Mean Decrease of Impurity*. *Impurity importance* pada variabel x_i dihitung dengan menjumlahkan semua *impurity decrease* (penurunan ketidakmurnian) di semua *node* model *Random Forest* pada tempat pemisahan pada x_i dilakukan, kemudian dinormalisasi dengan jumlah pohon. *Mean Decrease of Impurity* atau MDI terkadang disebut *Gini Importance* karena parameter untuk mengukur seberapa

penting suatu fitur menggunakan nilai *Gini Impurity* dari suatu *node* [9,10].

Gini impurity dapat dihitung berdasarkan Persamaan 1 :

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \quad (1)$$

D merupakan dataset yang berisi sampel kelas k p_i merupakan probabilitas data pada kelas i.

Selanjutnya apabila dataset D terbagi pada atribut A menjadi dua *subset* D_1 dan D_2 dengan ukuran n_1 dan n_2, masing-masing, maka *Gini Impurity* dapat didefinisikan pada Persamaan 2:

$$Gini_A(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2) \quad (2)$$

Kemudian untuk menghitung pengurangan dari ketidakmurnian data atau *decrease of impurity*, dapat didefinisikan *Gini Gain* pada Persamaan 3 :

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \quad (3)$$

D. *Algoritma Machine Learning*

Penelitian ini mengevaluasi lima algoritma klasifikasi *Random Forest*, *Logistic Regression*, *Support Vector Machine* (SVM), *Linear Discriminant Analysis* (LDA), dan *Artificial Neural Network* (ANN), karena efektivitasnya yang teruji dalam menangani data medis yang kompleks. Penggunaan variasi model mulai dari pendekatan linear hingga berbasis jaringan saraf bertujuan untuk membandingkan stabilitas prediksi terhadap dataset diabetes, di mana *Random Forest* dipilih sebagai fokus utama karena kemampuannya dalam melakukan perangkingan fitur secara internal yang mendukung analisis *feature importance*.

1. *Random Forest*

Random Forest merupakan metode klasifikasi berbasis komputasi yang dikembangkan oleh Leo Breiman pada tahun 2001. Metode *Classification and Regression Trees (CART)* yang diperkenalkan oleh Breiman pada tahun 1984 dikembangkan menjadi *Random Forest*. Breiman menyebutkan di dalam jurnalnya pada tahun 2001, *Random Forest* terdiri dari gabungan pohon klasifikasi yang masing-masing melakukan pelatihan pada sampel data atau fitur yang dipilih secara *random* (acak) di awal dan hasil prediksi dari model *Random Forest* ditentukan berdasarkan *majority voting* [9].

Random Forest adalah model klasifikasi yang secara otomatis memilih fitur-fitur yang paling penting dari dataset untuk membangun model [11]. Dengan kata lain, proses pembuatan model *Random Forest* secara otomatis juga melakukan proses *Feature Importance* sehingga metode *Gini Importance* dapat digunakan.

2. *Logistic Regression*

Logistic Regression merupakan salah satu metode analisis data yang digunakan untuk menemukan hubungan antara variabel target biner dengan satu atau lebih variabel prediktor, yang bisa berupa data berskala ordinal atau rasio.

Menurut Hosmer & Lemeshow [12], dalam regresi linier, variabel target diasumsikan berdistribusi normal, sementara pada regresi logistik biner, variabel target mengikuti distribusi Bernoulli dengan fungsi probabilitas yang ditunjukkan oleh Persamaan 4 :

$$f(y) = \pi^y (1-\pi)^{1-y}; y = 0 \; atau \; 1 \quad (4)$$

Ketika y = 0 maka$f(y) = 1 - \pi$ kemudian ketika y = 1 maka $f(y) = \pi$ di mana π adalah probabilitas dari nilai positif biner (misalnya 1 atau ya).

Model regresi logistik dapat dituliskan seperti pada Persamaan 5 :

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1, + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1, + \cdots + \beta_p x_p}} \quad (5)$$

Dimana :
p = jumlah variabel prediktor x_i
β_0 = konstanta nilai dasar dari fungsi logit ketika semua variabel prediktor bernilai 0

$\beta_1, \beta_2, \ldots, \beta_p$, = koefisien regresi untuk masing-masing variabel $x_1, x_2, \ldots, x_p$ [13].

3. *Support Vector Machine*

Support Vector Machine atau (SVM) merupakan algoritma pembelajaran mesin yang diperkenalkan oleh Vapnik pada tahun 1992. Prinsip dasar SVM adalah memisahkan dua kelas data dengan garis linier. Model ini melakukan pemetaan vektor *input* ke dalam ruang dimensi yang lebih tinggi, dimana *hyperplane* pemisah dapat dibangun [14].

SVM menggunakan *hyperplane* untuk memisahkan data dari dua kelas yang berbeda. *Hyperplane* dengan jarak yang paling jauh akan membuat model SVM memberi hasil prediksi sampel yang lebih optimal. Model ini bekerja efektif pada himpunan data berdimensi tinggi, karena *noise* dan *outlier* lebih mudah dihilangkan di ruang dimensi yang lebih tinggi [15].

4. *Linear Discriminant Analysis*

Linear Discriminant Analysis atau (LDA) adalah metode klasifikasi yang dapat digunakan untuk memisahkan data dari dua kelas atau lebih. LDA menggunakan *hyperplane* untuk memisahkan data dari kelas yang berbeda. LDA berasumsi bahwa data dari kelas yang berbeda dapat dipisahkan secara linier. Artinya, data dari kelas yang berbeda dapat dipisahkan dengan garis lurus. LDA memproyeksikan data ke *hyperplane*

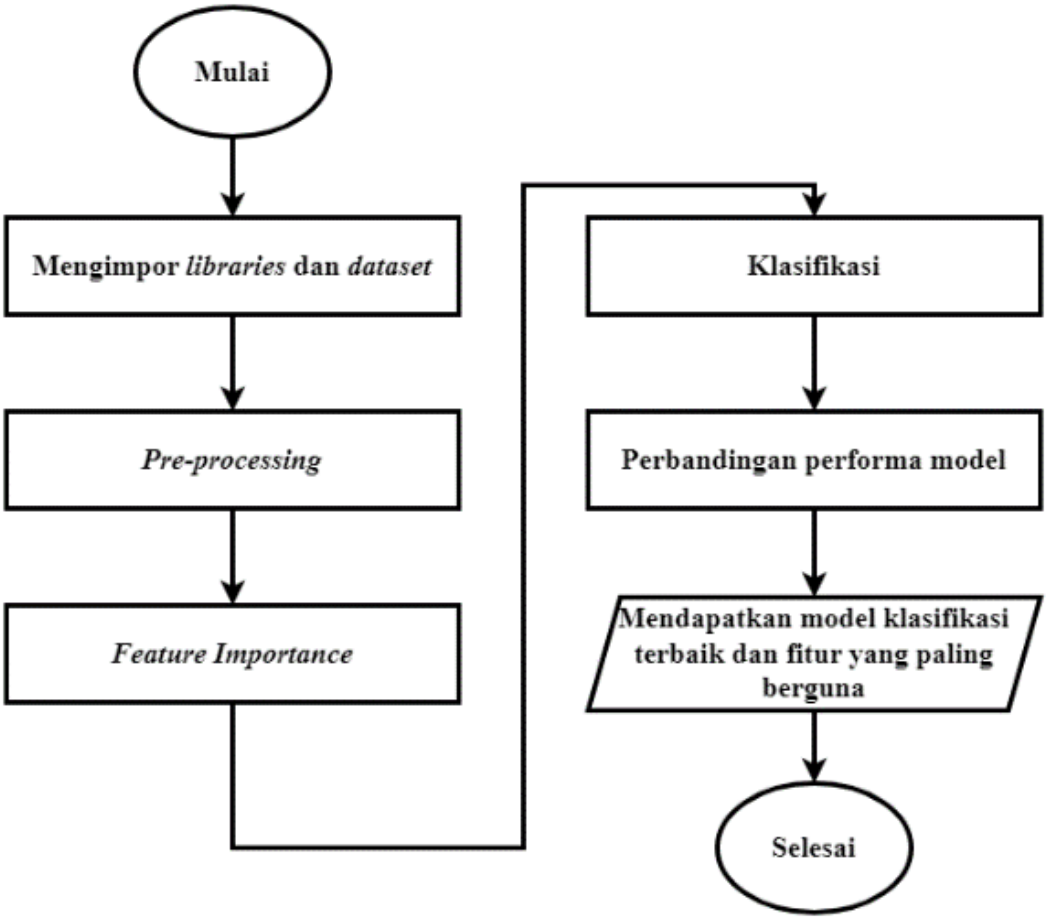
untuk memaksimalkan pemisahan antara kelas yang berbeda [16].

5. *Artificial Neuron Network*

Artificial Neuron Network (ANN) adalah sistem cerdas yang digunakan untuk mengolah data. ANN terpengaruh dari cara kerja sistem saraf manusia. Para ilmuwan menciptakan algoritma matematis yang meniru pola kerja saraf manusia. struktur ANN yang terdiri dari tiga lapisan: *input layer*, *hidden layer*, dan *output layer*. *Input layer* menerima informasi melalui bobot yang telah ditentukan. Bobot-bobot ini akan dikumpulkan dan diakumulasikan oleh lapisan tersembunyi. Kemudian hasilnya dibandingkan dengan nilai ambang (*threshold*) yang telah ditetapkan sebagai nilai aktivasi. Informasi yang melewati ambang batas akan diteruskan ke *output layer*.

E. Tahapan Penelitian

Tahapan penelitian ini terdiri dari *import libraries* dan dataset, *preprocessing*, *feature importance*, klasifikasi data, perbandingan performa model sehingga didapatkan hasil yang diharapkan berupa model klasifikasi dan fitur yang paling berguna seperti yang ditunjukkan Gambar 1.



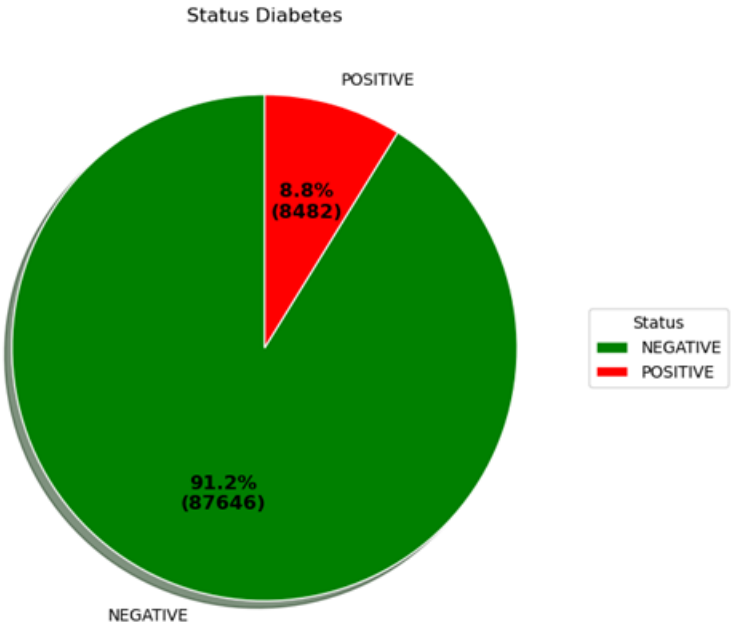
Gambar 1. Diagram alir penelitian

IV. HASIL DAN PEMBAHASAN

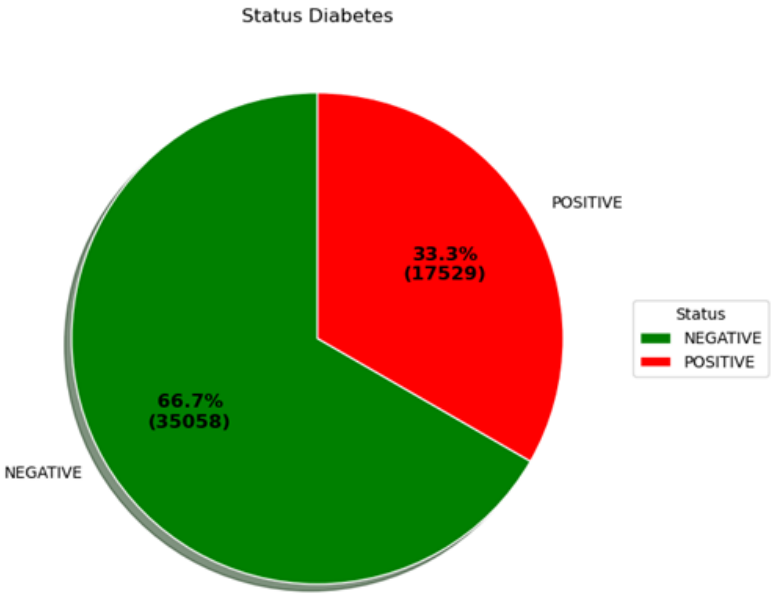
A. Hasil Preprocessing

Tahap pra-pemrosesan dilakukan untuk menjamin kualitas dan integritas data sebelum proses pemodelan. Langkah awal meliputi deduplikasi dengan menghapus 3.854 data duplikat serta membersihkan data pada variabel gender yang bernilai '*Other*' (0,195%) guna menghindari bias. Selanjutnya, dilakukan *Label Encoding* pada fitur kategorikal dan *Feature Scaling* menggunakan *StandardScaler* untuk menyeragamkan rentang nilai fitur.

Karakteristik dataset awal menunjukkan ketidakseimbangan kelas yang signifikan, sebagaimana terlihat pada Gambar 2, di mana terdapat 87.646 sampel negatif (91,2%) dan hanya 8.482 sampel positif (8,8%). Untuk mengatasi hal ini dan meningkatkan sensitivitas model, diterapkan strategi *Hybrid Resampling*. Langkah pertama adalah melakukan *Random Undersampling* pada kelas mayoritas guna mereduksi jumlah sampel negatif secara proporsional. Langkah kedua adalah menerapkan SMOTE (*Synthetic Minority Over-sampling Technique*) pada kelas minoritas untuk mensintesis sampel baru melalui interpolasi antar data positif yang ada.



Gambar 2. Status Diabetes sebelum *resampling* data



Gambar 3. Status Diabetes sesudah *resampling* data

Hasil dari proses resampling ini ditunjukkan pada Gambar 3, di mana komposisi data menjadi lebih seimbang dengan total 52.587 sampel (35.058 negatif dan 17.529 positif). Setelah dataset mencapai keseimbangan yang lebih baik, data dipartisi menjadi data training (80%) dan data testing (20%). Penting untuk dicatat bahwa proses *resampling* dan *scaling* dilakukan dengan pengawasan ketat untuk memastikan tidak terjadi *data leakage* antara set pelatihan dan set pengujian.

B. Hasil *Feature Importance*

Pengukuran kontribusi masing-masing fitur dalam klasifikasi data diabetes dilakukan menggunakan metode *Gini Importance*. Langkah pertama dalam proses ini adalah menghitung nilai *Gini Impurity* awal dari keseluruhan dataset (Gini(D)). Berdasarkan komposisi data setelah proses *resampling*, terdapat 17.529 sampel positif diabetes dan

35.058 sampel negatif. Menggunakan Persamaan (2), didapatkan nilai Gini(D) sebesar 0,44.

Selanjutnya, dilakukan perhitungan nilai *Gini Impurity* untuk setiap fitur guna menentukan seberapa efektif fitur tersebut dalam membagi data. Hasil perhitungan nilai *Gini Impurity* untuk masing-masing fitur disajikan pada Tabel 1. Berdasarkan nilai tersebut, langkah berikutnya adalah menghitung *Gini Gain* menggunakan Persamaan (2.3), yang menunjukkan tingkat pengurangan ketidakmurnian (*impurity*) yang dihasilkan oleh setiap fitur. Nilai Gini Gain untuk setiap fitur dirangkum dalam Tabel 2.

Untuk memberikan gambaran yang lebih jelas mengenai kontribusi fitur, nilai Gini Gain tersebut divisualisasikan dalam bentuk grafik batang yang disusun dari nilai terbesar hingga terkecil sebagaimana ditunjukkan pada Gambar 4.

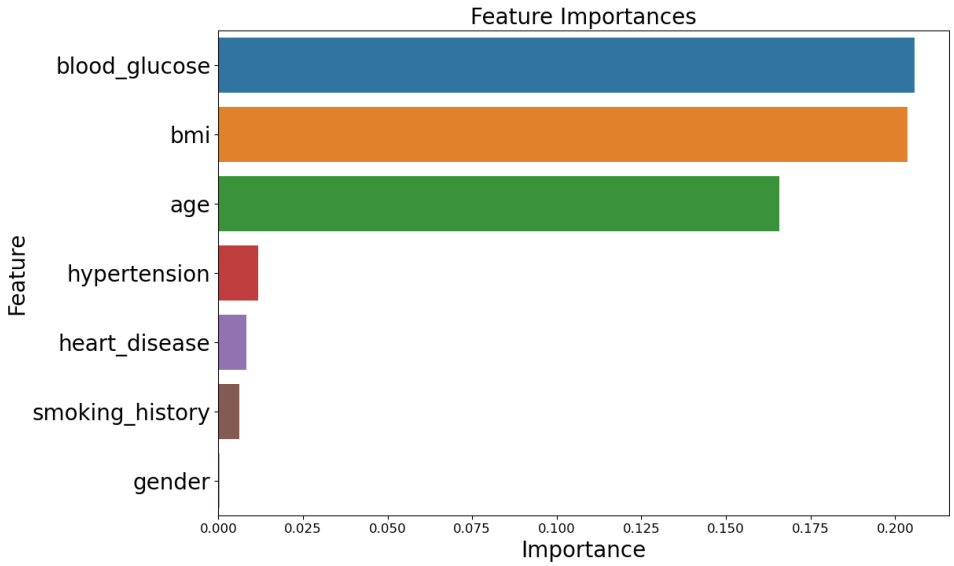
Berdasarkan Gambar 4, fitur blood_glucose memiliki nilai Gini Gain tertinggi (0,21), yang mengindikasikan bahwa fitur ini merupakan prediktor yang paling berkontribusi dan signifikan dalam klasifikasi diabetes. Selain itu, fitur bmi (0,20) dan age (0,17) juga menunjukkan nilai kepentingan yang tinggi. Sebaliknya, fitur lainnya seperti hypertension, heart_disease, smoking_history, dan gender memiliki nilai Gini Gain yang sangat rendah (rentang 0,004 hingga 0,01).

Tabel 1. Nilai *Gini Impurity* masing-masing fitur

Fitur	*Gini Impurity*
Gender	0,444
Age	0,277
Hypertension	0,433
heart_disease	0,436
smoking_history	0,438
Bmi	0,240
blood_glucose	0,238

Tabel 2. Nilai *Gini Gain* masing-masing fitur

Fitur	*Gini Gain*
gender	0,004
age	0,17
hypertension	0,01
heart_disease	0,01
smoking_history	0,01
bmi	0,20
blood_glucose	0,21



Gambar 4. Hasil nilai *Gini Gain*

Rendahnya nilai kontribusi pada fitur-fitur tersebut menunjukkan bahwa variabel tersebut tidak memberikan dampak signifikan terhadap peningkatan performa pemisahan kelas pada model. Oleh karena itu, dalam penelitian ini diputuskan untuk hanya menggunakan tiga fitur utama, yaitu *blood_glucose*, bmi, dan *age*, sebagai input dalam pemodelan *machine learning*. Strategi seleksi fitur ini bertujuan untuk menghasilkan model yang lebih efisien secara komputasi (*parsimonious*) dan mengurangi risiko *overfitting* yang disebabkan oleh fitur-fitur yang kurang relevan (*noise*), tanpa mengorbankan nilai diagnostik utama secara klinis.

Meskipun *Gini Importance* efektif dalam mengidentifikasi prediktor utama, penting untuk mendiskusikan potensi bias yang melekat pada metode ini. Secara teoretis, *Gini Importance* cenderung memberikan bobot lebih pada fitur kontinu atau fitur dengan kardinalitas tinggi dibandingkan fitur kategorikal. Dalam penelitian ini, fitur *blood_glucose*, bmi, dan *age* yang terpilih semuanya merupakan variabel kontinu. Hal ini memungkinkan

tingginya skor pada fitur tersebut dipengaruhi oleh banyaknya jumlah titik potong (*split points*) yang tersedia bagi algoritma untuk optimasi *impurity*. Namun, mengingat signifikansi klinis yang kuat dari ketiga variabel tersebut dalam literatur medis diabetes, hasil seleksi ini tetap dianggap valid dan representatif sebagai dasar klasifikasi.

C. Hasil Klasifikasi

Pada tahap ini, dilakukan proses pelatihan dan pengujian terhadap lima model *machine learning* (*Random Forest*, LDA, *Logistic Regression*, SVM, dan ANN) untuk mendeteksi penyakit diabetes. Berdasarkan analisis *Gini Importance*, dipilih tiga fitur paling signifikan sebagai prediktor utama, yaitu *blood_glucose*, bmi, dan *age*.

Guna memastikan stabilitas dan generalisasi model, penelitian ini menerapkan metode *10-fold cross-validation* dalam proses evaluasi, di samping pembagian data awal menjadi *training set* (80%) dan *testing set* (20%). Optimasi *hyperparameter* pada setiap model dilakukan menggunakan metode *grid search* untuk mencapai performa yang maksimal. Selanjutnya, model dilatih dengan parameter optimal tersebut dan dievaluasi secara berulang melalui mekanisme *cross-validation* untuk meminimalisir risiko *overfitting* serta bias pada pembagian data. Hasil evaluasi komprehensif dari masing-masing model dapat dilihat pada tabel berikut ini.

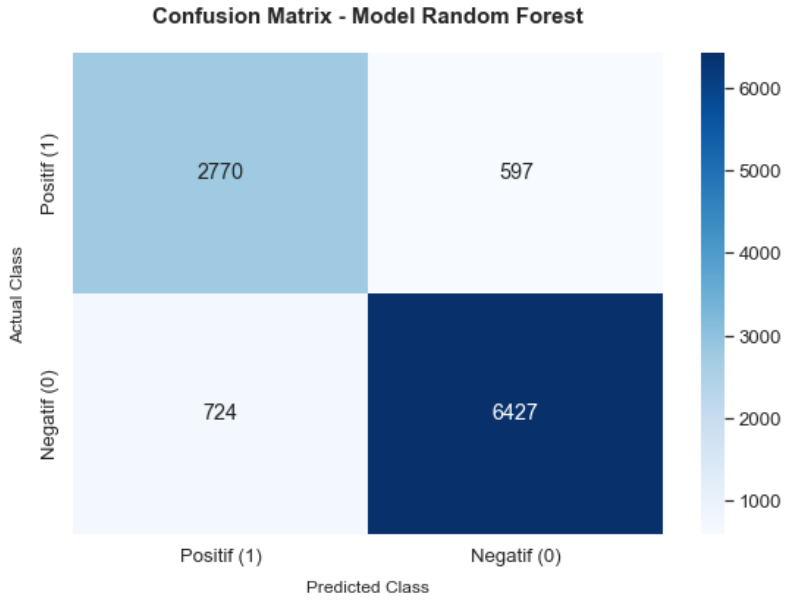
Berdasarkan hasil evaluasi yang disajikan pada Tabel 3, model Random Forest menunjukkan performa terbaik dibandingkan empat model lainnya dalam deteksi penyakit diabetes. Model ini meraih tingkat akurasi tertinggi sebesar 87%, didukung oleh nilai AUC sebesar 0,85. Nilai AUC tersebut mengindikasikan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan antara pasien diabetes dan non-diabetes.

Tabel 3.Perbandingan Akurasi 5 Model

Model	*Akurasi*	*Recall*	*Presisi*	*F1-Score*	*AUC*
Random Forest	87%	79%	82%	81%	0,85
Logistic Regression	82%	65%	76%	70%	0,78
SVM	86%	72%	78%	75%	0,81
LDA	81%	60%	79%	68%	0,76
ANN	83%	73%	75%	74%	0,80

Selain akurasi, performa *Random Forest* juga unggul secara konsisten pada metrik lainnya dengan nilai *Recall* 79%, Presisi 82%, dan *F1-Score* 81%. Dalam konteks medis, tingginya nilai *Recall* sangat krusial karena menunjukkan kemampuan model dalam meminimalisir risiko penderita yang tidak terdeteksi (*false negatives*). Keunggulan menyeluruh ini membuktikan bahwa penggunaan analisis *Gini Importance* efektif dalam menyeleksi fitur-fitur yang paling relevan, sehingga mampu mengoptimalkan performa model klasifikasi dibandingkan model linear seperti *Logistic Regression* atau LDA.

Untuk memahami lebih lanjut bagaimana model terbaik bekerja dalam mengklasifikasikan data, dilakukan analisis mendalam terhadap hasil prediksi model *Random Forest* seperti yang ditampilkan pada Gambar 5.



Gambar 5. *Heatmap Confusion Matrix* Model *Random Forest*

Detail performa klasifikasi model *Random Forest* disajikan pada Gambar 5. Dari total data uji, model berhasil mengklasifikasikan secara akurat 2.770 sampel positif diabetes dan 6.427 sampel negatif. Meskipun terdapat 597 kasus *false negative*, nilai ini relatif rendah dibandingkan model lainnya, menunjukkan bahwa integrasi *Gini Importance* mampu menjaga sensitivitas model dalam mendeteksi gejala diabetes secara efektif.

Guna menunjang implementasi klinis, model Random Forest ini dirancang dalam arsitektur berbasis API yang memungkinkan integrasi langsung dengan Sistem Informasi Rumah Sakit (HIS). Melalui pipa ML (ML *pipeline*) yang terotomatisasi, data pasien dari HIS akan melewati tahap pra-pemrosesan secara *real-time* sebelum diprediksi oleh model yang telah di-*deploy* menggunakan teknologi kontainer (*Docker*) untuk menjamin stabilitas sistem. Selain itu, aspek penjaminan kualitas (*Quality Assurance*) dilakukan melalui pemantauan performa berkala guna mendeteksi data *drift*, sehingga model tetap akurat dan andal saat digunakan sebagai alat bantu pengambilan keputusan medis di lingkungan pelayanan kesehatan digital.

V. SIMPULAN

Penelitian ini menyimpulkan bahwa tahapan *preprocessing* data yang komprehensif, khususnya penanganan data duplikat dan penerapan *hybrid resampling* (SMOTE dan *undersampling*), sangat krusial dalam menghasilkan dataset yang seimbang untuk klasifikasi diabetes. Analisis kontribusi fitur menggunakan metode *Gini Importance* menunjukkan bahwa *blood_glucose* merupakan prediktor paling signifikan, diikuti oleh bmi dan *age*. Hasil evaluasi terhadap lima model *machine learning* menunjukkan bahwa *Random Forest* memberikan performa terbaik dengan akurasi sebesar 87%.

Meskipun memberikan hasil yang optimal, penelitian ini memiliki keterbatasan pada penggunaan dataset sekunder yang mungkin belum mencakup diversitas demografis yang lebih luas. Pekerjaan di masa depan diharapkan dapat memvalidasi model ini menggunakan data klinis *real-time* serta mengeksplorasi integrasi sistem informasi kesehatan berbasis cloud untuk meningkatkan aksesibilitas deteksi dini diabetes di fasilitas kesehatan primer.

REFERENSI

[1] M. F. Salim dan S. Sugeng, "Analisis Rekam Medis Pasien Diabetes Mellitus Melalui Implementasi Teknik Data Mining di RSUP Dr. Sardjito Yogyakarta," *Jurnal Kesehatan Vokasional*, vol. 2, no. 2, hlm. 167, Mei 2018, doi: 10.22146/jkesvo.30331.

[2] M. A. Nurjana dan N. N. Veridiana, "Hubungan Perilaku Konsumsi dan Aktivitas Fisik dengan Diabetes Mellitus di Indonesia," *Buletin Penelitian Kesehatan*, vol. 47, no. 2, hlm. 97–106, Agu 2019, doi: 10.22435/bpk.v47i2.667.

[3] Y. Jian, M. Pasquier, A. Sagahyroon, dan F. Aloul, "A machine learning approach to predicting diabetes complications," *Healthcare (Switzerland)*, vol. 9, no. 12, Des 2021, doi: 10.3390/healthcare9121712.

[4] J. Xue, F. Min, dan F. Ma, "Research on Diabetes Prediction Method Based on Machine Learning," *J Phys Conf Ser*, vol. 1684, no. 1, hlm. 012062, Nov 2020, doi: 10.1088/1742 6596/1684/1/012062.

[5] V. Gulati dan N. Raheja, "Efficiency Enhancement of Machine Learning Approaches through the Impact of Preprocessing Techniques," *Proceedings of IEEE International Conference on Signal Processing,Computing and Control*, vol. 2021-October, hlm. 191–196, 2021, doi: 10.1109/ISPCC53510.2021.9609474.

[6] S. Mahajan, P. K. Sarangi, A. K. Sahoo, dan M. Rohra, "Diabetes Mellitus Prediction using Supervised Machine Learning Techniques," dalam *2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, hlm. 587–592. doi: 10.1109/InCACCT57535.2023.10141734.

[7] C. Charitha, A. D. Chaitrasree, P. C. Varma, dan C. Lakshmi, "Type-II Diabetes Prediction Using Machine Learning Algorithms," dalam *2022 International Conference on Computer Communication and Informatics (ICCCI)*, Institute of Electrical and Electronics Engineers Inc., 2022, hlm. 1–5. doi: 10.1109/ICCCI54379.2022.9740844.

[8] A. Baughman, K. Yogaraj, R. Hebbar, S. Ghosh, R. Haq, dan Y. Chhabra, *Study of Feature Importance for Quantum Machine Learning Models*. 2022.

[9] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, hlm. 5–32, 2001, doi: 10.1023/A:1010933404324.

[10] S. Nembrini, I. R. König, dan M. N. Wright, "The revival of the Gini importance?," *Bioinformatics*, vol. 34, no. 21, hlm. 3711–3718, Nov 2018, doi: 10.1093/bioinformatics/bty373.

[11] B. H. Menze *dkk.*, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, no. 1, hlm. 213, Des 2009, doi: 10.1186/1471-2105-10-213.

[12] D. W. Hosmer dan S. Lemeshow, *Applied Logistic Regression*. Wiley, 2000. doi: 10.1002/0471722146.

[13] I. Zain dan A. Yanti, "Pemodelan Regresi Logistik Biner terhadap Peminat ITS di Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN) 2014," *Jurnal Sains dan Seni ITS*, vol. 4, no. 1, 2015, doi: 10.12962/j23373520.v4i1.9402.

[14] J. P. Maulana dan I. Irhamah, "Klasifikasi Kabupaten di Provinsi Jawa Timur Berdasarkan Indikator Daerah Tertinggal dengan metode Support Vector Machine (SVM) dan Entropy Based Fuzzy Support Vector Machine (EFSVM)," *Inferensi*, vol. 1, no. 1, hlm. 9, Sep 2018, doi: 10.12962/j27213862.v1i1.6715.

[15] Y. S. Pamuji, D. Safitri, dan A. Prahutama, "KLASIFIKASI PENERIMA PROGRAM BERAS MISKIN (RASKIN) DI KABUPATEN WONOSOBO DENGAN METODE SUPPORT VECTOR MACHINE MENGGUNAKAN LibSVM," *Jurnal Gaussian*, vol. 4, no. 4, hlm. 1087–1096, 2015, [Daring]. Tersedia pada: http://ejournal-s1.undip.ac.id/index.php/gaussian

[16] Vaibhaw, J. Sarraf, dan P. K. Pattnaik, "Brain–computer interfaces and their applications," dalam *An Industrial IoT Approach for Pharmaceutical Industry Growth*, V. E. Balas, V. K. Solanki, dan R. Kumar, Ed., Elsevier, 2020, hlm. 31–54. doi: 10.1016/B978-0-12-821326-1.00002-4.