JURNAL TEKNOLOGI INFORMASI DIGITAL (JTID)
INTEGRASI PUBLIKASI DIGITAL
Vol. 2 No. 2, Juli 2026 E-ISSN (xxx-xxx)
Website : https://jurnal.ipdig.id/index.php/jtid

Peningkatan Kinerja Prediksi Diabetes Menggunakan *Random Forest* melalui kombinasi *Threshold Tuning* dan *Class Weight Balance*

Cindy Setyowati[1*], Aditya Pratama Werdana[2], Ariska Nur Anggraini[3]

[1,2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa, Indonesia

Dikirimkan: 28-05-2026
Diterbitkan: 08-06-2026

Keywords:
Diabetes Prediction;
Feature Importance;
Machine Learning;
Random Forest;
Threshold Tuning.

E-mail Penulis korespondensi:
cindyssetyo@gmail.com

Abstrak. Diabetes merupakan penyakit metabolik kronis yang ditandai oleh peningkatan kadar *glukosa* darah dan berpotensi menyebabkan komplikasi serius apabila tidak terdeteksi secara dini. Penelitian ini bertujuan membangun model prediksi diabetes menggunakan algoritma *Random Forest* pada *dataset diabetes prediction* yang terdiri dari 100.000 data dan setelah proses penghapusan duplikasi menjadi 96.146 data. Tahapan penelitian meliputi pembersihan data, analisis distribusi kelas, *preprocessing* menggunakan *StandardScaler* untuk fitur numerik dan *OneHotEncoder* untuk fitur kategorik, pelatihan model *Random Forest* dengan *class_weight balanced*, serta evaluasi performa secara komprehensif. Evaluasi dilakukan menggunakan *accuracy*, *precision*, *recall*, *F1-score*, *ROC-AUC, confusion matrix, ROC curve, precision-recall curve, calibration curve, learning curve,* visualisasi *PCA*, dan analisis *feature importance*. Hasil pengujian menunjukkan bahwa model *Random Forest* memperoleh *accuracy* sebesar 96,91%, *precision* 94,15%, *recall* 69,28%, *F1-score* 79,82%, dan *ROC-AUC* 96,39% pada *threshold* standar 0,5. Proses *threshold tuning* menunjukkan bahwa *threshold* optimal berada pada nilai 0,75 dengan peningkatan *F1-score* menjadi 80,80% serta *accuracy* menjadi 97,16%. Analisis *feature importance* menunjukkan bahwa *HbA1c_level, blood_glucose_level, age,* dan *BMI* merupakan faktor paling berpengaruh dalam prediksi diabetes. Hasil penelitian menunjukkan bahwa *Random Forest* dengan optimasi *threshold* mampu memberikan performa prediksi yang tinggi dan berpotensi digunakan sebagai pendekatan pendukung deteksi dini diabetes berbasis *machine learning.*

Abstract. *Diabetes is a chronic metabolic disease characterized by elevated blood glucose levels and may lead to serious complications if not detected early. This study aims to develop a diabetes prediction model using the Random Forest algorithm on a diabetes prediction dataset consisting of 100,000 records, which became 96,146 records after duplicate removal. The research stages included data cleaning, class distribution analysis, preprocessing using StandardScaler for numerical features and OneHotEncoder for categorical features, Random Forest model training with balanced class weighting, and comprehensive performance evaluation. Model evaluation was conducted using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, ROC curve, precision-recall curve, calibration curve, learning curve, PCA visualization, and feature importance analysis. Experimental results showed that the Random Forest model achieved an accuracy of 96.91%, precision of 94.15%, recall of 69.28%, F1-score of 79.82%, and ROC-AUC of 96.39% at the default threshold of 0.5. Threshold tuning indicated that the optimal threshold was 0.75, improving the F1-score*

to 80.80% and accuracy to 97.16%. Feature importance analysis revealed that HbA1c_level, blood_glucose_level, age, and BMI were the most influential factors in diabetes prediction. The findings indicate that Random Forest combined with threshold optimization provides high predictive performance and has strong potential as a machine learning-based approach for early diabetes detection.

1. Pendahuluan

Diabetes merupakan salah satu penyakit metabolik kronis yang ditandai oleh tingginya kadar *glukosa* dalam darah akibat gangguan produksi insulin, kerja insulin, atau keduanya [1]. Penyakit ini menjadi perhatian global karena dapat memicu berbagai komplikasi serius seperti gangguan *kardiovaskular*, kerusakan ginjal, *neuropati*, hingga gangguan penglihatan apabila tidak ditangani secara tepat [2]. Selain memengaruhi kualitas hidup penderita, diabetes juga menimbulkan beban ekonomi yang besar bagi individu maupun sistem pelayanan kesehatan [3]. Kondisi tersebut menjadikan diabetes sebagai salah satu masalah kesehatan yang membutuhkan perhatian khusus, terutama dalam upaya pencegahan dan deteksi dini.

Peningkatan jumlah penderita diabetes terjadi secara signifikan dalam beberapa dekade terakhir seiring perubahan gaya hidup masyarakat, pola konsumsi yang kurang sehat, rendahnya aktivitas fisik, serta meningkatnya prevalensi obesitas [4]. Faktor usia, hipertensi, penyakit jantung, indeks massa tubuh (*BMI*), kadar *HbA1c*, dan kadar glukosa darah juga diketahui memiliki hubungan erat dengan risiko diabetes [5]. Banyak penderita tidak menyadari kondisi yang dialami karena gejala awal sering kali tidak spesifik atau bahkan tidak muncul secara jelas. Akibatnya, diagnosis sering terlambat dilakukan ketika komplikasi telah berkembang, sehingga diperlukan pendekatan yang mampu membantu identifikasi risiko secara lebih cepat dan akurat.

Proses diagnosis diabetes secara konvensional umumnya dilakukan melalui pemeriksaan klinis dan pengujian laboratorium seperti kadar glukosa darah serta HbA1c [6]. Meskipun metode tersebut memiliki tingkat akurasi yang baik, penerapannya memerlukan waktu, biaya, serta akses terhadap fasilitas kesehatan yang memadai. Selain itu, tingginya jumlah data kesehatan yang tersedia saat ini sering kali belum dimanfaatkan secara optimal untuk mendukung pengambilan keputusan medis. Kondisi ini membuka peluang bagi pengembangan sistem prediksi berbasis data yang mampu membantu proses skrining awal dan memberikan dukungan terhadap diagnosis klinis [7].

Perkembangan teknologi informasi dan data mining telah mendorong pemanfaatan *machine learning* dalam berbagai bidang, termasuk sektor kesehatan. *Machine learning* memungkinkan komputer mempelajari pola dari data historis untuk menghasilkan prediksi atau klasifikasi secara otomatis [8]. Dalam konteks kesehatan, pendekatan ini telah banyak diterapkan untuk mendeteksi penyakit, memprediksi risiko pasien, serta membantu proses pengambilan keputusan medis berdasarkan data yang tersedia. Kemampuan *machine learning* dalam mengolah data berskala besar dan menemukan hubungan kompleks antarvariabel menjadikannya sebagai pendekatan yang menjanjikan dalam pengembangan sistem prediksi diabetes [9].

Salah satu tantangan dalam pengembangan model prediksi diabetes adalah ketidakseimbangan *kelas (imbalanced data),* yaitu kondisi ketika jumlah data pada satu kelas jauh lebih dominan dibandingkan kelas lainnya [10]. Pada dataset diabetes yang digunakan, proporsi data tidak diabetes mencapai sekitar 91%, sedangkan data diabetes hanya sekitar 9%. Ketidakseimbangan ini dapat menyebabkan model *machine learning* cenderung lebih fokus mempelajari pola pada kelas mayoritas sehingga kemampuan mendeteksi kelas minoritas menjadi menurun [11]. Akibatnya, model dapat menghasilkan nilai accuracy yang tinggi tetapi memiliki recall atau sensitivitas yang rendah dalam mengidentifikasi pasien diabetes [12]. Oleh karena itu, penanganan imbalanced data menjadi aspek penting dalam penelitian prediksi diabetes.

Salah satu algoritma *machine learning* yang banyak digunakan dalam permasalahan klasifikasi adalah *Random Forest*. Algoritma ini merupakan metode *ensemble* yang membangun sejumlah pohon keputusan dan menggabungkan hasil prediksi untuk meningkatkan stabilitas serta akurasi model [13]. *Random Forest* memiliki beberapa keunggulan, antara lain mampu menangani data berdimensi tinggi, mengurangi risiko *overfitting*, serta bekerja dengan baik pada data yang memiliki hubungan nonlinier antarfitur [14]. Selain itu, algoritma ini juga menyediakan informasi mengenai tingkat pengaruh setiap fitur melalui *feature importance* [15], sehingga tidak hanya menghasilkan prediksi tetapi juga membantu memahami faktor-faktor yang berkontribusi terhadap risiko diabetes.

Beberapa penelitian terdahulu menunjukkan bahwa penerapan *machine learning* mampu memberikan performa yang baik dalam klasifikasi diabetes. Penelitian menggunakan algoritma *Multi-Layer Perceptron (MLP)* pada dataset *diabetes* sebanyak 96.146 data setelah pembersihan berhasil memperoleh *accuracy* sebesar 97,15%, *precision* 99,39%, *recall* 68,16%, *F1-score* 80,87%, dan *ROC-AUC* 97,49%, yang menunjukkan kemampuan model dalam membedakan kelas diabetes dan non-diabetes dengan baik [16]. Penelitian lain yang mengevaluasi metode *ensemble learning* berbasis *boosting*, yaitu *Gradient Boosting, XGBoost,* dan *CatBoost*, juga menghasilkan tingkat akurasi sekitar 97% dengan *F1-score* sebesar 0,81, meskipun nilai *recall* pada kelas diabetes masih relatif rendah sekitar 0,69 sehingga menunjukkan adanya tantangan dalam mendeteksi seluruh kasus positif [17]. Selain itu, penelitian mengenai penanganan *class imbalance* menggunakan metode *oversampling SMOTE* dan *ADASYN* pada algoritma *CatBoost* menunjukkan bahwa pendekatan *SMOTE* mampu meningkatkan performa model dengan *accuracy* sebesar 97,27%, *precision* 97,37%, *recall* 69,71%, *F1-score* 81,25%, dan *ROC-AUC* 97,96% [18].

Berbagai penelitian telah menunjukkan bahwa penerapan *machine learning* mampu meningkatkan performa prediksi, meskipun masih menghadapi tantangan pada penanganan data tidak seimbang dan sensitivitas terhadap kelas minoritas [19]. Penelitian yang membandingkan algoritma *K-Nearest Neighbor (KNN)* dan *Naïve Bayes* menunjukkan bahwa *Naïve Bayes* memperoleh akurasi tertinggi sebesar 80%, sedangkan *KNN* menghasilkan *recall* tertinggi sebesar 0,92 sehingga lebih baik dalam mendeteksi kasus positif diabetes. Penelitian lain menggunakan *Support Vector Machine (SVM)* pada dataset Pima Indians Diabetes juga menunjukkan kemampuan yang baik dalam membangun model prediktif melalui *preprocessing* dan *exploratory data analysis* [20]. Pendekatan deep learning melalui *Deep Neural Network (DNN)* dengan optimasi hiperparameter berbasis *Bayesian Optimization* bahkan menghasilkan performa tinggi dengan *accuracy* sebesar 97,14%, *recall* 91,76%, dan *AUC* 0,9764 pada *threshold* optimal [20]. Selain itu, penelitian berbasis ensemble learning menggunakan *Softvoting* yang dikombinasikan dengan *SMOTE-ENN* dan optimasi *Bayesian* berhasil mencapai akurasi hingga 99,80% pada *Diabetes Prediction Dataset*, menunjukkan efektivitas penanganan data tidak seimbang dan optimasi model [21]. Studi lain mengenai *comparative analysis* pada *data imbalanced* dengan pendekatan *SMOTE* juga menunjukkan bahwa *Random Forest* memberikan *accuracy* sebesar 95,70% dan *F1-score* terbaik dibanding beberapa algoritma lain, sementara *Logistic Regression* memiliki *recall* tertinggi dalam mendeteksi kelas positif [22]. Temuan serupa pada domain lain menunjukkan bahwa penerapan *SMOTE* mampu meningkatkan *recall* dan *F1-score*, meskipun terkadang diikuti penurunan akurasi secara keseluruhan [23]. Hasil-hasil penelitian tersebut menunjukkan bahwa pemilihan algoritma, strategi penanganan class imbalance, dan optimasi model menjadi faktor penting dalam menghasilkan sistem prediksi diabetes yang akurat sekaligus sensitif terhadap deteksi kasus positif.

Meskipun berbagai penelitian terdahulu menunjukkan performa yang baik dalam klasifikasi diabetes, masih terdapat beberapa keterbatasan yang perlu diperhatikan. Sebagian penelitian lebih berfokus pada pencapaian nilai accuracy yang tinggi tanpa mengevaluasi secara mendalam kemampuan model dalam mendeteksi kelas diabetes sebagai kelas minoritas. Pada kondisi data yang tidak seimbang, *accuracy* yang tinggi belum tentu menunjukkan performa model yang optimal karena model dapat cenderung bias terhadap kelas mayoritas. Selain itu, beberapa penelitian masih terbatas pada penggunaan evaluasi dasar seperti *accuracy* dan *confusion matrix* tanpa disertai analisis probabilitas, kalibrasi model, maupun interpretasi faktor-faktor yang memengaruhi prediksi. Kondisi tersebut menunjukkan adanya kebutuhan untuk mengembangkan pendekatan prediksi diabetes yang tidak hanya akurat, tetapi juga memiliki sensitivitas yang baik serta mampu memberikan interpretasi terhadap hasil prediksi.

Penelitian ini menawarkan pendekatan yang berbeda melalui penerapan algoritma *Random Forest* dengan penanganan *class imbalance* menggunakan *class_weight balanced*, disertai proses *threshold tuning* untuk memperoleh titik keputusan klasifikasi yang lebih optimal. Tidak hanya berfokus pada nilai *accuracy*, penelitian ini juga melakukan evaluasi model secara komprehensif melalui *precision, recall, F1-score, ROC-AUC, precision-recall curve, calibration curve, learning curve,* visualisasi *PCA,* serta analisis *feature importance*. Pendekatan ini memberikan *novelty* berupa kombinasi optimasi *threshold* dan evaluasi multi-metrik yang bertujuan menghasilkan model prediksi yang lebih seimbang antara kemampuan mendeteksi pasien diabetes dan kestabilan performa secara keseluruhan. Selain itu, analisis *feature importance* pada *Random Forest* memungkinkan identifikasi variabel kesehatan yang paling berpengaruh terhadap prediksi diabetes sehingga model tidak hanya bersifat prediktif tetapi juga lebih interpretatif.

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung deteksi dini diabetes berbasis *machine learning* yang memiliki tingkat akurasi tinggi sekaligus sensitivitas yang baik terhadap identifikasi pasien berisiko. Selain memberikan model prediksi yang andal, penelitian ini juga diharapkan mampu memperkaya kajian mengenai penerapan *Random Forest* pada data kesehatan yang tidak seimbang serta menjadi referensi bagi penelitian selanjutnya dalam mengembangkan metode klasifikasi diabetes yang lebih adaptif dan interpretatif. Dengan demikian, hasil penelitian tidak hanya memiliki nilai akademik dalam bidang *data mining*

dan kecerdasan buatan, tetapi juga berpotensi memberikan manfaat praktis bagi dunia kesehatan dalam mendukung pengambilan keputusan secara lebih cepat dan berbasis data.

2. Metode Penelitian

2.1. Metode Pendekatan

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *machine learning* untuk membangun model prediksi diabetes menggunakan algoritma *Random Forest*. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengolahan data numerik, pengukuran performa model, serta analisis statistik terhadap hasil klasifikasi [23]. Algoritma *Random Forest* digunakan karena memiliki kemampuan yang baik dalam menangani data klasifikasi, mengurangi risiko *overfitting* melalui metode *ensemble*, serta mampu memberikan interpretasi berupa tingkat kepentingan fitur *(feature importance)* [24]. Penelitian dilakukan melalui beberapa tahapan yang meliputi pengumpulan data, *preprocessing*, pelatihan model, evaluasi, serta optimasi *threshold* untuk memperoleh performa klasifikasi yang optimal.

2.2. *Data Collection*

Dataset yang digunakan dalam penelitian ini adalah *Diabetes Prediction Dataset* yang dapat diakses melalui link *https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset*. yang terdiri dari 100.000 data dengan 9 atribut, yaitu *gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level,* dan diabetes sebagai variabel target. Data dikumpulkan dari sumber dataset publik yang telah banyak digunakan pada penelitian prediksi diabetes.

2.3. Exploratory Data Analysis (EDA)

Tahap *Exploratory Data Analysis* dilakukan untuk memahami karakteristik data dan hubungan antarvariabel sebelum dilakukan pemodelan [25]. Analisis dilakukan melalui visualisasi distribusi kelas diabetes menggunakan *countplot*, Selain itu, digunakan *heatmap correlation* untuk melihat hubungan antarfitur numerik terhadap variabel diabetes [26]. Tahap ini penting untuk memahami kondisi dataset, mengidentifikasi pola awal, serta mendeteksi adanya ketidakseimbangan kelas *(class imbalance)* pada data [27].

2.4. Data Preprocessing

Tahap *preprocessing* bertujuan mempersiapkan data agar dapat diproses secara optimal oleh algoritma machine learning [28]. Fitur kategorikal seperti *gender* serta *smoking_history* ditransformasikan menggunakan *OneHotEncoder* sehingga dapat direpresentasikan dalam bentuk numerik dan sementara itu, fitur numerik seperti *age, hypertension, heart_disease, bmi, HbA1c_level,* dan *blood_glucose_level* dilakukan standardisasi menggunakan *StandardScaler* untuk menyeragamkan rentang nilai fitur [29]. Seluruh proses *preprocessing* diintegrasikan menggunakan *ColumnTransformer* dan *Pipeline* agar proses transformasi data dilakukan secara konsisten dan sistematis [27].

2.5. Pembagian Data

Setelah *preprocessing*, dataset dibagi menjadi data *training* dan data *testing* menggunakan metode *train-test split* dengan rasio 80:20 [30]. Pembagian dilakukan menggunakan teknik *stratified sampling* untuk mempertahankan proporsi kelas diabetes dan non-diabetes pada kedua kelompok data [31]. Hasil pembagian menghasilkan *data training* sebanyak 76.916 data dan *data testing* sebanyak 19.230 data. Pendekatan ini digunakan agar model dapat dilatih menggunakan sebagian besar data dan tetap memiliki data independen untuk proses evaluasi.

2.6. Pemodelan *Random Forest*

Model klasifikasi dibangun menggunakan algoritma *Random Forest* yang merupakan metode *ensemble* berbasis kumpulan pohon keputusan *(decision tree).* Parameter yang digunakan pada penelitian ini meliputi *n_estimators* sebanyak 300, *criterion gini, max_features sqrt, bootstrap true,* dan *class_weight balanced.* Penggunaan *class_weight balanced* bertujuan membantu model menangani kondisi imbalanced data dengan memberikan bobot yang lebih proporsional pada kelas minoritas diabetes [32]. Model dibangun melalui *Pipeline* sehingga proses preprocessing dan klasifikasi berjalan dalam satu alur terintegrasi.

2.7. Evaluasi Model

Evaluasi model dilakukan secara komprehensif menggunakan beberapa metrik performa klasifikasi, yaitu *accuracy, precision, recall, F1-score, dan ROC-AUC.* Selain itu, digunakan *classification report* dan *confusion matrix* untuk menganalisis kemampuan model dalam mengidentifikasi masing-masing kelas [33]. Visualisasi *ROC*

curve dan *precision-recall curve* juga digunakan untuk mengevaluasi kemampuan diskriminatif model terhadap kelas diabetes dan non-diabetes [34].

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

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

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

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

2.8. *Threshold Tuning* dan Analisis Model

Penelitian ini menerapkan proses threshold tuning terhadap probabilitas prediksi model pada rentang *threshold* 0,10 hingga 0,90 untuk memperoleh *threshold* terbaik berdasarkan nilai *F1-score*. Pendekatan ini bertujuan menghasilkan keseimbangan yang lebih baik antara *precision* dan recall, khususnya pada kelas diabetes sebagai kelas minoritas [35]. Selain *threshold tuning,* dilakukan analisis tambahan melalui *calibration curve* untuk mengevaluasi reliabilitas probabilitas prediksi, *learning curve* untuk melihat kestabilan model terhadap jumlah data pelatihan, visualisasi *PCA* dua dimensi untuk memetakan hasil prediksi, serta *feature importance* untuk mengidentifikasi variabel yang paling berpengaruh dalam prediksi diabetes.

3. Hasil dan Pembahasan

Dataset yang digunakan dalam penelitian ini adalah *Diabetes Prediction Dataset* yang dapat diakses melalui link *https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset*. yang terdiri dari 100.000 data dengan 9 atribut, yaitu *gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level,* dan diabetes sebagai variabel target seperti terlihat pada Tabel 1.

Tabel 1. Dataset Penelitian

Age	Gender	BMI	HbA1c_level	…	Diabetes
80	Female	25.19	6.6	…	0
54	Male	27.32	6.6	…	0
28	Female	27.32	5.7	…	0
36	Male	23.45	5.0	…	0
76	Female	20.14	4.8	…	0
44	Male	19.31	6.5	…	1
67	Female	32.45	7.2	…	1
29	Male	24.89	5.4	…	0
58	Female	30.11	6.8	…	1
….	…	…	…	…	…
62	Male	28.76	6.9	…	1

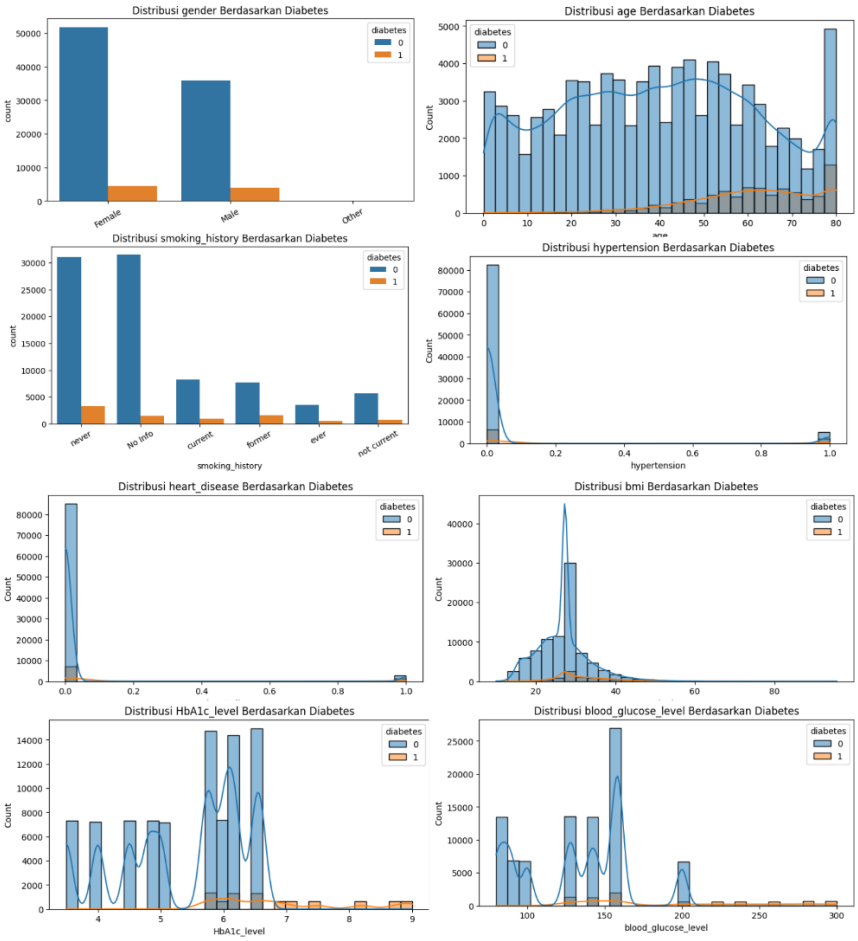
Pada tahap awal dilakukan pemeriksaan kualitas data melalui identifikasi *missing value* dan data duplikat, seperti terlihat pada Tabel 2.

Tabel 2. Tabel Missing Value

Atribut	Missing Value
gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

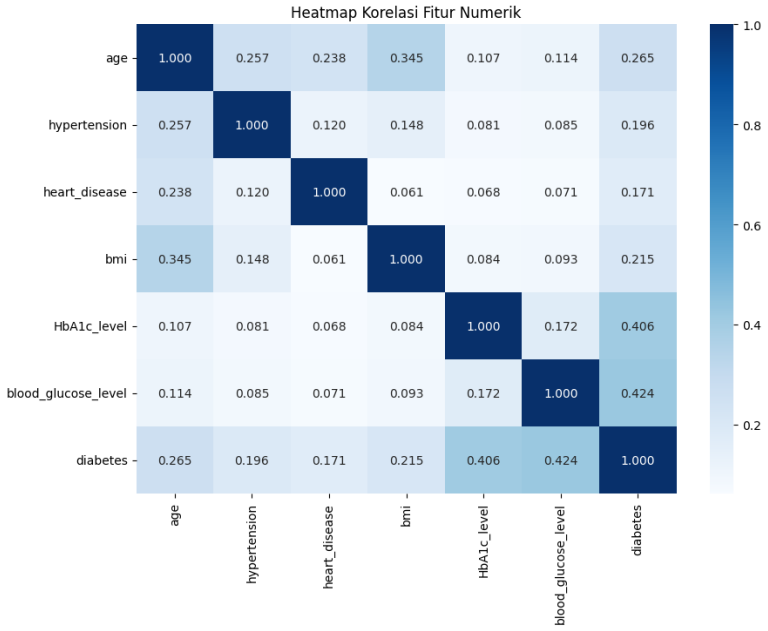
Berdasarkan Gambar 1 terlihat visualisasi distribusi fitur terhadap status diabetes, terlihat bahwa beberapa variabel memiliki hubungan yang cukup kuat dengan kejadian diabetes. Distribusi usia (*age*) menunjukkan bahwa kasus diabetes lebih banyak ditemukan pada kelompok usia lanjut, terutama di atas 50 tahun. Variabel *hypertension* dan *heart disease* juga memperlihatkan proporsi penderita diabetes yang lebih tinggi pada individu dengan riwayat hipertensi dan penyakit jantung dibandingkan yang tidak memiliki kondisi tersebut. Pada variabel *BMI*, penderita diabetes cenderung berada pada rentang indeks massa tubuh yang lebih tinggi, menunjukkan adanya hubungan

antara obesitas dan risiko diabetes. Selain itu, fitur *HbA1c level* dan *blood glucose level* memperlihatkan pemisahan distribusi yang paling jelas antara kelas diabetes dan non-diabetes, di mana penderita diabetes cenderung memiliki nilai yang lebih tinggi. Sementara itu, distribusi *gender* dan *smoking history* menunjukkan perbedaan yang tidak terlalu signifikan antar kelas, meskipun kategori tertentu tetap memperlihatkan variasi risiko. Temuan ini mengindikasikan bahwa faktor metabolik seperti HbA1c dan kadar glukosa darah menjadi indikator paling dominan dalam prediksi diabetes, yang selaras dengan hasil *feature importance* pada model *Random Forest*.



Gambar 1. Hubungan Faktor dan Keterkaitan dengan Penyebab Diabetes

Berdasarkan *heatmap* korelasi fitur numerik pada Gambar 2, terlihat bahwa hubungan antar variabel terhadap status diabetes umumnya berada pada tingkat rendah hingga sedang. Fitur *blood_glucose_level* memiliki korelasi tertinggi terhadap diabetes sebesar 0,424, diikuti oleh *HbA1c_level* sebesar 0,406, yang menunjukkan bahwa kadar glukosa darah dan HbA1c merupakan indikator paling kuat dalam membedakan pasien diabetes dan non-diabetes. Variabel *age* juga menunjukkan korelasi positif terhadap diabetes sebesar 0,265, diikuti *BMI* sebesar 0,215, *hypertension* sebesar 0,196, dan *heart_disease* sebesar 0,171, yang mengindikasikan bahwa peningkatan usia, indeks massa tubuh, serta adanya hipertensi dan penyakit jantung berkontribusi terhadap peningkatan risiko diabetes. Selain itu, terdapat korelasi sedang antara *age* dan *BMI* (0,345) serta antara *HbA1c_level* dan *blood_glucose_level* (0,172). Tidak ditemukan korelasi yang sangat tinggi antar fitur independen, sehingga risiko *multicollinearity* relatif rendah. Hasil ini memperkuat temuan sebelumnya bahwa faktor metabolik, khususnya *HbA1c* dan kadar glukosa darah, merupakan variabel yang paling dominan dalam proses prediksi diabetes menggunakan model *Random Forest.*



Gambar 2. *Heatmap Correlation*

Hasil *classification report* menunjukkan bahwa model *Random Forest* memiliki performa klasifikasi yang sangat baik pada *data testing* dengan *accuracy* sebesar 96,91% (Tabel 3). Pada kelas *non-diabetes* (kelas 0), model memperoleh *precision* 97,10%, *recall* 99,58%, dan *F1-score* 98,33%, yang menunjukkan kemampuan sangat tinggi dalam mengenali data non-diabetes secara tepat. Sementara itu, pada kelas diabetes (kelas 1), model menghasilkan *precision* sebesar 94,15%, *recall* 69,28%, dan *F1-scor*e 79,82%. Nilai *precision* yang tinggi menunjukkan bahwa sebagian besar prediksi positif diabetes yang dihasilkan model adalah benar, namun nilai *recall* yang lebih rendah mengindikasikan bahwa masih terdapat beberapa kasus diabetes yang belum berhasil terdeteksi *(false negative).* Perbedaan performa antara kedua kelas dipengaruhi oleh kondisi *imbalanced data* dengan dominasi kelas *non-diabetes,* sehingga meskipun model menunjukkan tingkat akurasi tinggi, evaluasi menggunakan *precision, recall,* dan *F1-score* menjadi penting untuk memberikan gambaran performa model yang lebih komprehensif dalam mendeteksi diabetes.

Tabel 3. *Classification Report* dari *Random Forest* tanpa *Improvment*

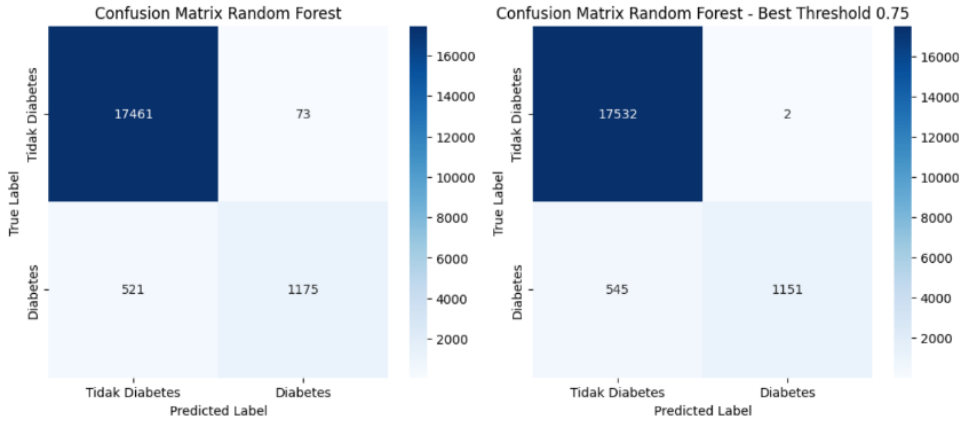
Kelas	Precision	Recall	F1-Score	Support
Non-Diabetes (0)	0.9710265821	0.9958366602	0.9832751436	17534
Diabetes (1)	0.9415064103	0.6928066038	0.7982336957	1696
Accuracy			0.9691107644	19230
Macro Average	0.9562664962	0.8443216320	0.8907544196	19230
Weighted Average	0.9684230350	0.9691107644	0.9669553154	19230

Hasil *classification report* setelah penerapan *best threshold* sebesar 0,75 menunjukkan adanya peningkatan performa model Random Forest dibandingkan threshold standar 0,5 (Tabel 4). Model memperoleh *accuracy* sebesar 97,16% dengan *weighted F1-score* sebesar 96,91%. Pada kelas diabetes (kelas 1), *precision* meningkat menjadi 99,83% dengan *F1-score* sebesar 80,80%, menunjukkan bahwa hampir seluruh prediksi positif diabetes yang dihasilkan model merupakan prediksi yang benar. Namun demikian, *recall* kelas diabetes sedikit menurun menjadi 67,87%, yang menunjukkan bahwa masih terdapat sebagian kasus diabetes yang belum berhasil terdeteksi. Sementara itu, kelas non-diabetes tetap menunjukkan performa sangat tinggi dengan *recall* mencapai 99,99% dan *F1-score* sebesar 98,46%. Hasil ini menunjukkan bahwa optimasi *threshold* berhasil meningkatkan ketepatan prediksi (*precision*) dan keseimbangan performa model melalui peningkatan *F1-score*, sehingga model menjadi lebih selektif dan andal dalam mengidentifikasi kasus diabetes.

Tabel 4. *Classification Report Random Forest* dengan *Best Threshold (0,75)*

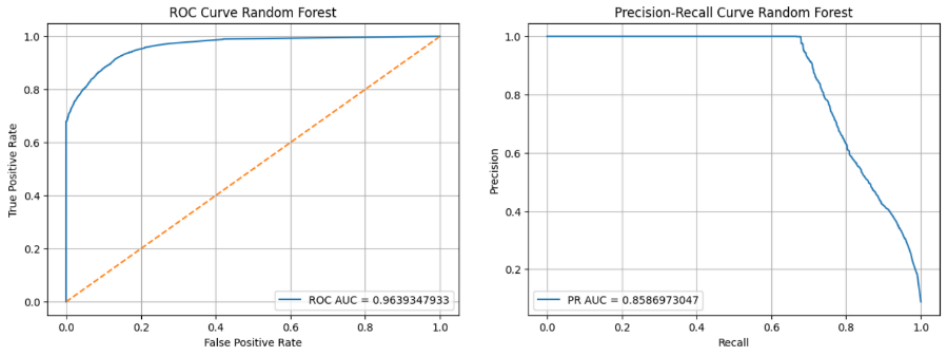
Kelas	Precision	Recall	F1-Score	Support
Non-Diabetes (0)	0.9698511921	0.9998859359	0.9846395777	17534
Diabetes (1)	0.9982653946	0.6786556604	0.8080028080	1696
Accuracy			0.9715548622	19230
Macro Average	0.9840582934	0.8392707981	0.8963211928	19230
Weighted Average	0.9723571977	0.9715548622	0.9690610045	19230

Berdasarkan *confusion matrix* sebelum dan sesudah *threshold tuning* pada Gambar 3, terlihat adanya perubahan pola prediksi model Random Forest. Pada threshold standar 0,5, model menghasilkan *true negative* sebanyak 17.461, *true positive* 1.175, *false positive* 73, dan *false negative* 521. Setelah dilakukan optimasi dengan *best threshold* 0,75, jumlah *true negative* meningkat menjadi 17.532 dan *false positive* menurun drastis menjadi hanya 2 kasus, menunjukkan peningkatan kemampuan model dalam menghindari kesalahan prediksi positif pada kelas non-diabetes. Namun, *true positive* sedikit menurun menjadi 1.151 dan *false negative* meningkat menjadi 545, yang menunjukkan bahwa model menjadi lebih selektif dalam memprediksi diabetes sehingga beberapa kasus positif tidak terdeteksi. Kondisi ini menunjukkan adanya *trade-off* antara *precision* dan *recall*, di mana threshold 0,75 meningkatkan ketepatan prediksi diabetes dengan mengurangi *false positive*, tetapi pada saat yang sama menurunkan sensitivitas model dalam mendeteksi seluruh kasus diabetes. Oleh karena itu, pemilihan threshold perlu disesuaikan dengan tujuan implementasi, apakah lebih menekankan minimisasi kesalahan positif atau peningkatan deteksi dini diabetes.



Gambar 3. Perbandingan *Confusion Matrix*

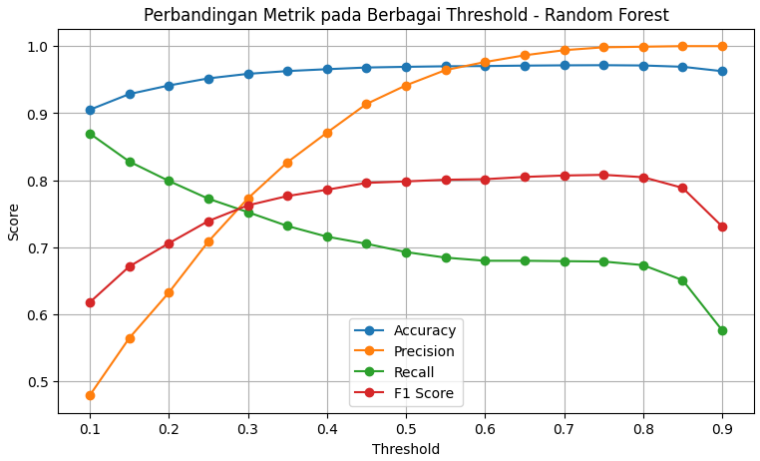
Berdasarkan kurva *ROC* dan *Precision-Recall*, model *Random Forest* pada Gambar 4 menunjukkan kemampuan klasifikasi yang sangat baik dalam memprediksi diabetes. Nilai *ROC-AUC* sebesar 0,9639 mengindikasikan bahwa model memiliki kemampuan yang sangat tinggi dalam membedakan kelas diabetes dan non-diabetes, terlihat dari kurva *ROC* yang berada jauh di atas garis diagonal sebagai representasi prediksi acak. Selain itu, nilai *PR-AUC* sebesar 0,8587 menunjukkan performa yang baik dalam menjaga keseimbangan antara *precision* dan *recall*, terutama pada kondisi *imbalanced data* di mana jumlah kelas non-diabetes jauh lebih dominan. Kurva *Precision-Recall* memperlihatkan bahwa model mampu mempertahankan nilai *precision* tinggi pada berbagai tingkat *recall*, meskipun terjadi penurunan *precision* ketika *recall* mendekati nilai maksimum. Hasil ini menunjukkan bahwa Random Forest tidak hanya memiliki kemampuan diskriminatif yang tinggi berdasarkan *ROC-AUC*, tetapi juga efektif dalam mendeteksi kelas minoritas diabetes secara lebih representatif melalui evaluasi *PR-AUC*.



Gambar 4. *ROC Curve* dan *Precision-Recall Curve*

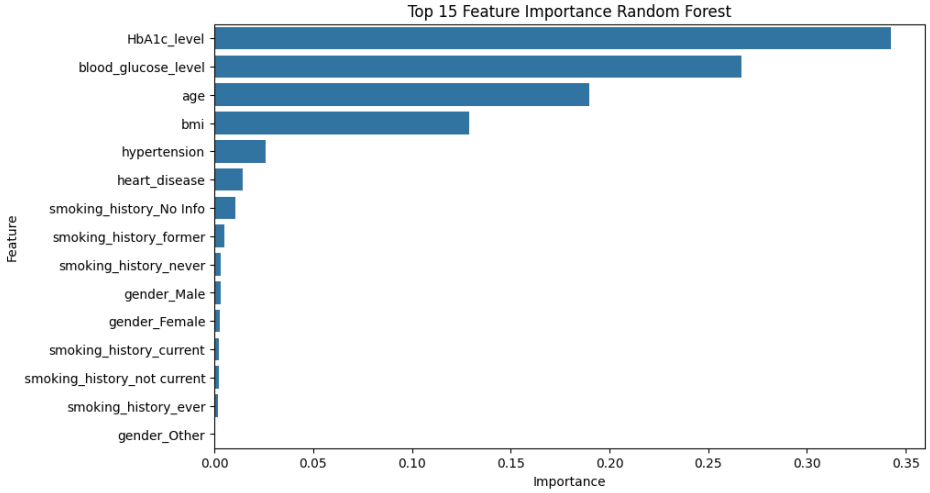
Berdasarkan grafik perbandingan metrik pada berbagai *threshold* pada Gambar 5, terlihat bahwa perubahan nilai *threshold* memberikan pengaruh signifikan terhadap performa model Random Forest. Seiring meningkatnya *threshold* dari 0,1 hingga 0,9, nilai *precision* meningkat secara konsisten dari sekitar 0,48 hingga mendekati 1,00, menunjukkan bahwa model menjadi semakin selektif dalam memprediksi kelas diabetes dan menghasilkan lebih sedikit *false positive*. Sebaliknya, nilai *recall* mengalami penurunan dari sekitar 0,87 menjadi 0,58, yang menandakan semakin banyak kasus diabetes yang tidak terdeteksi (*false negative*) pada *threshold* tinggi. Nilai

accuracy relatif stabil dan tinggi pada kisaran 0,95–0,97, sedangkan *F1-score* meningkat hingga mencapai titik optimal pada *threshold* sekitar 0,75 dengan nilai sekitar 0,81 sebelum kembali menurun. Hasil ini menunjukkan adanya *trade-off* antara *precision* dan *recall*, sehingga pemilihan *best threshold* sebesar 0,75 dilakukan karena memberikan keseimbangan performa terbaik berdasarkan *F1-score*, sekaligus meningkatkan ketepatan prediksi tanpa menurunkan akurasi secara signifikan.



Gambar 5. Perbandingan Metrik pada Berbagai *Threshold*

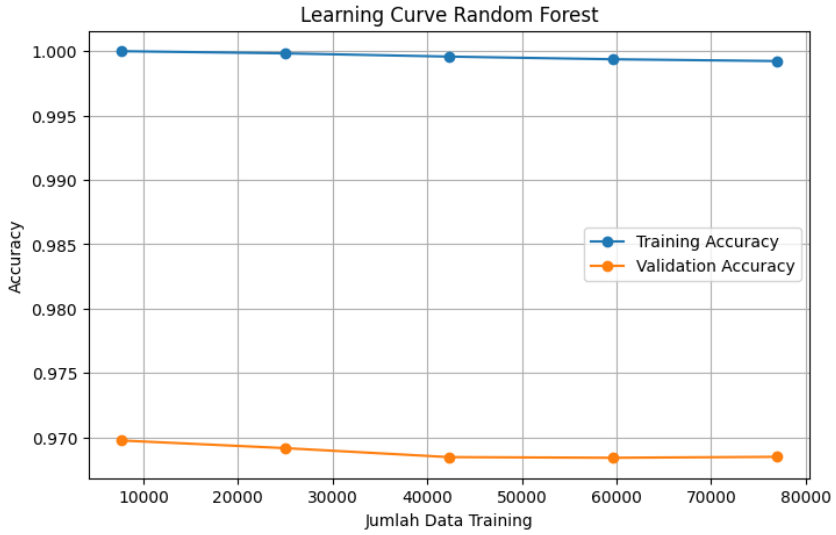
Berdasarkan grafik *Top 15 Feature Importance Random Forest* pada Gambar 6, dapat dilihat bahwa fitur yang paling berpengaruh dalam prediksi diabetes adalah *HbA1c_level* dengan nilai *importance* tertinggi sekitar 0,34, diikuti oleh *blood_glucose_level* sebesar 0,27. Hal ini menunjukkan bahwa kadar *HbA1c* dan kadar glukosa darah memiliki kontribusi paling besar dalam menentukan hasil prediksi model. Selain itu, fitur *age* dan *bmi* juga memberikan pengaruh yang cukup signifikan dibandingkan fitur lainnya. Sementara itu, variabel seperti *hypertension*, *heart_disease,* dan beberapa kategori smoking_history memiliki pengaruh yang relatif kecil. Fitur terkait *gender* menunjukkan nilai *importance* paling rendah, sehingga kontribusinya terhadap prediksi diabetes dalam model *Random Forest* ini tidak terlalu signifikan. Secara keseluruhan, model lebih banyak mengandalkan indikator medis utama dibandingkan faktor demografis atau riwayat merokok.



Gambar 6. Top 15 Feature Importance Random Forest

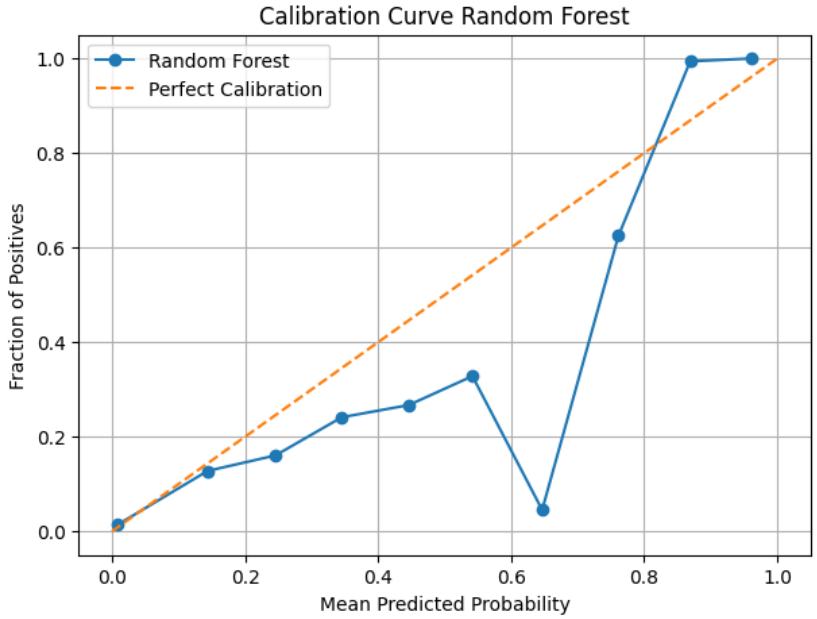
Berdasarkan grafik *Learning Curve Random Forest* pada Gambar 7, terlihat bahwa akurasi *data training* berada pada nilai yang sangat tinggi, mendekati 100% pada seluruh jumlah *data training*, sedangkan akurasi validasi berada di sekitar 96,8%–97,0%. Seiring bertambahnya jumlah *data training,* akurasi *training* sedikit menurun namun tetap stabil, sementara akurasi validasi cenderung konstan. Perbedaan antara *kurva training* dan *validation* menunjukkan adanya sedikit indikasi *overfitting*, karena model sangat baik dalam mempelajari *data training* tetapi performanya sedikit lebih rendah pada data validasi. Namun, selisih tersebut tidak terlalu besar sehingga model

Random Forest masih dapat dikatakan memiliki kemampuan generalisasi yang baik dan stabil dalam melakukan prediksi.



Gambar 7. *Learning Curve*

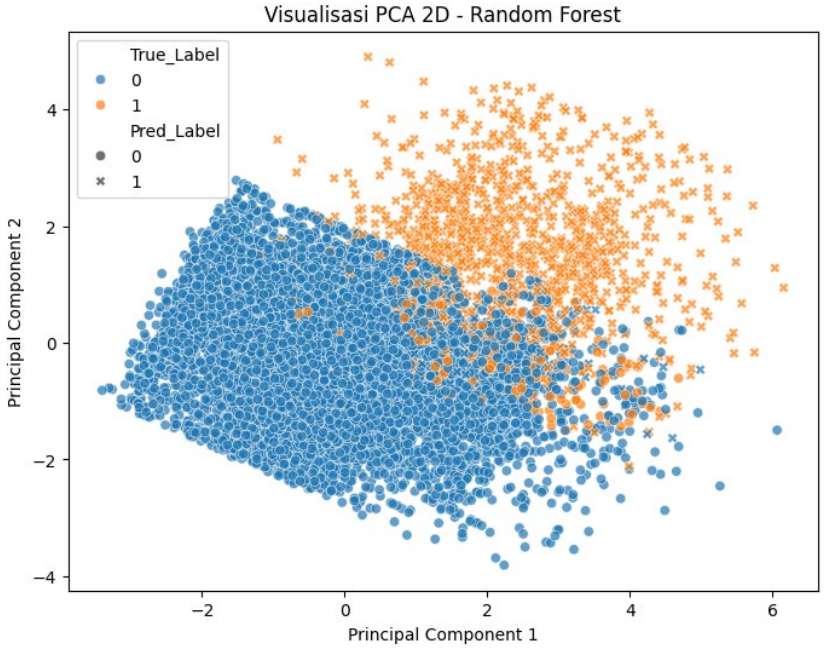
Berdasarkan grafik *Calibration Curve Random Forest* pada Gambar 8, terlihat bahwa probabilitas prediksi model belum sepenuhnya terkalibrasi dengan baik terhadap kondisi ideal (*perfect calibration*). Hal ini ditunjukkan oleh kurva *Random Forest* yang beberapa kali berada cukup jauh dari garis diagonal putus-putus sebagai representasi kalibrasi sempurna. Pada probabilitas prediksi rendah hingga menengah, model cenderung *underestimate* karena fraksi positif aktual lebih rendah dibanding probabilitas prediksi. Namun, pada probabilitas tinggi, model menunjukkan peningkatan yang signifikan dan mendekati nilai kalibrasi sempurna. Secara keseluruhan, model *Random Forest* sudah mampu memberikan prediksi probabilitas yang cukup baik, tetapi masih terdapat beberapa penyimpangan pada interval probabilitas tertentu sehingga kalibrasi model dapat ditingkatkan lebih lanjut agar hasil probabilitas prediksi menjadi lebih akurat dan konsisten.



Gambar 8. *Calibration Curve*

Berdasarkan visualisasi *PCA 2D Random Forest* pada Gambar 9, terlihat bahwa data berhasil direduksi menjadi dua komponen utama (*Principal Component 1* dan *Principal Component 2*) untuk mempermudah analisis pola distribusi kelas. Titik berwarna biru merepresentasikan kelas 0, sedangkan titik oranye merepresentasikan kelas 1. Hasil visualisasi menunjukkan bahwa sebagian besar data kelas 0 terkonsentrasi pada area tertentu dengan

penyebaran yang lebih padat, sementara kelas 1 cenderung berada pada area dengan nilai komponen utama yang lebih tinggi. Meskipun masih terdapat beberapa area tumpang tindih (*overlap*) antara kedua kelas, pola pemisahan antar kelas sudah cukup terlihat, yang menandakan bahwa model *Random Forest* mampu membedakan karakteristik masing-masing kelas dengan cukup baik. Selain itu, simbol prediksi menunjukkan bahwa sebagian besar hasil prediksi model sesuai dengan label sebenarnya, sehingga mengindikasikan performa klasifikasi yang cukup baik pada data yang telah direduksi menggunakan *PCA*.



Gambar 9. PCA 2D Visualization

4. Kesimpulan

Berdasarkan hasil penelitian, algoritma *Random Forest* mampu memberikan performa yang sangat baik dalam prediksi penyakit diabetes menggunakan *Diabetes Prediction Dataset* yang terdiri dari 96.146 data setelah proses penghapusan duplikasi. Model menghasilkan *accuracy* sebesar 96,91%, *precision* 94,15%, *recall* 69,28%, *F1-score* 79,82%, dan *ROC-AUC* 96,39% pada *threshold* standar 0,5, yang menunjukkan kemampuan tinggi dalam membedakan pasien diabetes dan non-diabetes. Analisis *feature importance* menunjukkan bahwa *HbA1c level* dan *blood glucose level* merupakan fitur paling dominan dalam proses prediksi diabetes, diikuti oleh *age* dan *BMI*. Selain itu, penerapan *threshold tuning* berhasil meningkatkan performa model dengan *best threshold* sebesar 0,75, yang menghasilkan peningkatan *accuracy* menjadi 97,16% dan *F1-score* menjadi 80,80%, serta meningkatkan *precision* kelas diabetes hingga 99,83%. Hasil evaluasi komprehensif melalui *confusion matrix*, kurva *ROC*, *Precision-Recall Curve*, serta analisis berbagai *threshold* menunjukkan bahwa *Random Forest* merupakan metode yang efektif dan andal untuk mendukung deteksi dini diabetes, meskipun tantangan dalam meningkatkan *recall* pada kelas minoritas masih menjadi aspek yang perlu dikembangkan pada penelitian selanjutnya.

5. Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat dipertimbangkan untuk pengembangan penelitian selanjutnya. Pertama, peningkatan kemampuan model dalam mendeteksi kelas diabetes perlu menjadi perhatian utama karena nilai recall pada kelas diabetes masih berada pada kisaran 67–69%, yang menunjukkan masih adanya kasus false negative. Oleh karena itu, penelitian berikutnya dapat menerapkan teknik penanganan imbalanced data seperti SMOTE, ADASYN, atau SMOTE-ENN untuk meningkatkan sensitivitas model terhadap kelas minoritas. Kedua, optimasi model dapat dikembangkan melalui hyperparameter tuning menggunakan metode seperti Grid Search, Random Search, atau Bayesian Optimization agar diperoleh kombinasi parameter Random Forest yang lebih optimal. Ketiga, penelitian selanjutnya disarankan melakukan perbandingan Random Forest dengan algoritma lain seperti XGBoost, CatBoost, LightGBM, atau Deep Neural Network guna memperoleh model dengan performa yang lebih baik, khususnya pada metrik recall dan F1-score. Keempat, penggunaan data yang lebih beragam dan bersumber dari populasi atau institusi kesehatan yang berbeda perlu

dilakukan agar model memiliki kemampuan generalisasi yang lebih tinggi. Terakhir, model yang dihasilkan dapat dikembangkan ke dalam bentuk sistem pendukung keputusan atau aplikasi berbasis web maupun mobile untuk membantu tenaga medis dalam melakukan deteksi dini diabetes secara lebih cepat dan efisien.

Daftar Rujukan

[1] H. Husain, S. Ramadani, and N. Magfirah Ilyas, "Literature Review: Analisis Faktor Penyebab Penyakit Degeneratif (Diabetes Mellitus) pada Metabolisme Karbohidrat," *Indones. J. Sci. Public Heal.*, vol. 2, no. 2 SE-Articles, pp. 258–270, Sep. 2025, [Online]. Available: https://yici-journal.id/ijsph/article/view/29

[2] N. N. Rosyidah and E. A. Cahyono, "DIABETES MELITUS TIPE 2 ; ARTIKEL REVIEW," *Enferm. Cienc.*, vol. 3, no. 1, pp. 44–63, Feb. 2025, doi: 10.56586/ec.v3i1.74.

[3] D. Rahmawati, "Kualitas Hidup Pasien Diabetes Melitus dan Hipertensi dalam Program Penyakit Kronis (Prolanis) di Indonesia: Narative Review," *J. Mandala Pharmacon Indones.*, vol. 10, no. 1 SE-Review Article, pp. 116–122, Jun. 2024, doi: 10.35311/jmpi.v10i1.531.

[4] L. Muhaziroh *et al.*, "Edukasi Pola Makan Sehat dan Aktivitas Fisik Sebagai Upaya Pencegahan Diabetes pada Transisi Epidemiologi," *BERNAS J. Pengabdi. Kpd. Masy.*, vol. 7, no. 2 SE-Articles, pp. 1248–1257, Apr. 2026, doi: 10.31949/jb.v7i2.17680.

[5] I. Restika BN, S. Suarnianti, and S. Syamsuriah, "Trend Diabetes Melitus Tipe 2 pada Remaja: Literatur Review," *J. Penelit. Sains dan Kesehat. Avicenna*, vol. 4, no. 3 SE-Artikel, pp. 249–252, Sep. 2025, doi: 10.69677/avicenna.v4i3.192.

[6] F. Sartika and N. Hestiani, "Kadar HbA1c pada Pasien Wanita Penderita Diabetes Mellitus Tipe 2 di Rsud Dr. Doris Sylvanus Palangka Raya: HbA1c Levels in Patients Female with Type 2 Diabetes Mellitus in RSUD Dr. Doris Sylvanus Palangka Raya," *Borneo J. Med. Lab. Technol.*, vol. 2, no. 1 SE-Articles, pp. 97–100, Oct. 2019, doi: 10.33084/bjmlt.v2i1.1086.

[7] L. Najma Rachmawati, C. Rievania Khairunisa Fitri, and M. Exsanni Araf Octaviana, "PELUANG DAN TANTANGAN ARTIFICIAL INTELLIGENCE TERHADAP OPTIMALISASI LAYANAN KESEHATAN," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 882–890, Dec. 2024, doi: 10.36040/jati.v9i1.12514.

[8] N. Rokhman, S. Sumaryanto, F. N. Hakim, and P. A. Maulan, "Integrasi Machine Learning dalam Homebase Sistem Informasi untuk Analisis Produktivitas Akademik," *Go Infotech J. Ilm. STMIK AUB; Vol 31, No 2 December*, 2025, doi: 10.36309/goi.v31i2.424.

[9] J. J. Hidayat, F. F. Azhari, T. M. Husna, A. N. Fahmayani, N. N. Pradana, and C. Setyowati, "Perbandingan Kinerja Algoritma Machine Learning Dalam Prediksi Kesehatan Mental Dan Burnout Mahasiswa," *J. Surya Inform.*, vol. 16, no. 1 SE-Articles, pp. 32–42, May 2026, doi: 10.48144/suryainformatika.v16i1.2420.

[10] A. Salam, L. Azhari, R. S. Septarini, and N. Heriyani, "Pendekatan Hybrid K-Means SMOTE dan Logistic Regression Untuk Deteksi Dini Diabetes Mellitus Pada Imbalanced Data," *Bull. Comput. Sci. Res.*, vol. 5, no. 3, pp. 219–227, Apr. 2025, doi: 10.47065/bulletincsr.v5i3.502.

[11] M. Samodro, "Analisis Pengaruh Ketidakseimbangan Data terhadap Kinerja Model Klasifikasi Penyakit Jantung," *J. Softw. Eng. Inf. Syst.*, vol. 6, no. 1 SE-Articles, pp. 56–62, Feb. 2026, [Online]. Available: https://ejurnal.umri.ac.id/index.php/SEIS/article/view/11050

[12] K. A. Saputro, E. M. Atsir, and H. Hasanah, "Perbandingan Tingkat Akurasi Penyakit Diabetes Menggunakan Metode Regresi Logistik dan Random Forest," *TAMIKA J. Tugas Akhir Manaj. Inform. Komputerisasi Akunt.*, vol. 4, no. 2, pp. 159–166, Dec. 2024, doi: 10.46880/tamika.Vol4No2.pp159-166.

[13] R. Harahap, M. Irpan, M. A. Dinata, L. Efrizoni, and R. Rahmaddeni, "Perbandingan Algoritma Random Forest dan XGBoost untuk Klasifikasi Penyakit Paru-Paru Berdasarkan Data Demografi Pasien," *J. Ilm. BETRIK Besemah Teknol. Inf. dan Komput.*, vol. 15, no. 2, pp. 130–141, 2024, [Online]. Available: https://ejournal.pppmitpa.or.id/index.php/betrik/article/view/231

[14] Gullam Almuzadid and Egia Rosi Subhiyakto, "Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and Random Forest," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 828–837, Jun. 2025, doi: 10.30871/jaic.v9i3.9528.

[15] M. Haris Khoirul Anam, D. Kurnianingtyas, and A. Andy Soebroto, "Implementasi Algoritma Random Forest Untuk Prediksi Churn Pada Pelanggan Retail Online," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 4 SE-Artikel, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/16262

[16] J. J. Hidayat, D. E. Sujianto, M. R. Saputra, E. A. Ramdhani, M. Jihansyah, and Y. Nandya, "Klasifikasi Penyakit Diabetes Melitus Berbasis Jaringan Syaraf Tiruan Menggunakan Algoritma Multi-Layer Perceptron," *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 5, no. 1, pp. 401–411, May 2026, doi: 10.62712/juktisi.v5i1.1042.

[17] J. J. Hidayat, M. R. Saputra, A. R. Sigand, A. L. N. Fadilah, M. D. I. Amin, and A. R. Ramadhan, "Evaluasi Kinerja Algoritma Ensemble Learning Pada Klasifikasi Penyakit Diabetes Berbasis Boosting Method," *J. Surya Inform.*, vol. 16, no. 1 SE-Articles, pp. 71–80, May 2026, doi: 10.48144/suryainformatika.v16i1.2424.

[18] Z. Rozikin and J. J. Hidayat, "Perbandingan Metode Oversampling SMOTE dan ADASYN pada Klasifikasi Diabetes Menggunakan Algoritma CatBoost," *J. Manaj. Inform. Teknol.*, vol. 6, no. 1, pp. 151–164, 2026, doi: 10.51903/mifortekh.v6i1.1157.

[19] N. Nanda Pradana, A. Agung Subekti, and E. Rilvani, "DETEKSI TRANSAKSI MENCURIGAKAN MENGGUNAKAN DECISION TREE DAN LOGISTIC REGRESSION DENGAN MITIGASI KETIDAKSEIMBANGAN KELAS," *J. Media Akad.*, vol. 3, no. 8 SE-Articles, 2025, doi: 10.62281/v3i8.2680.

[20] P. R. P. Rosalya Putri and R. Alit, "Klasifikasi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine (SVM)," *J. Informatics Comput. Sci.*, vol. 6, no. 03, pp. 740–746, Jan. 2025, doi: 10.26740/jinacs.v6n03.p740-746.

[21] S. Ernawati and I. Maulana, "Meningkatkan Klasifikasi Penyakit Diabetes Menggunakan Metode Ensemble Softvoting Dengan SMOTE-ENN dan Optimasi Bayesian," *Evolusi J. Sains dan Manaj.*, vol. 13, no. 1, pp. 71–86, Mar. 2025, doi: 10.31294/evolusi.v13i1.8267.

[22] A. Nugroho, Wiyanto, and D. Maulana, "COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS IN HANDLING IMBALANCED DATA WITH SMOTE OVERSAMPLING APPROACH," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 11, no. 2, pp. 487–495, Nov. 2025, doi: 10.33480/jitk.v11i2.6956.

[23] N. Surojudin, S. Butsianto, and A. Firmansyah, "Perbandingan Kinerja Naïve Bayes dengan dan Tanpa SMOTE untuk Klasifikasi Gangguan Kecemasan Mahasiswa pada Data Tidak Seimbang," *Bull. Comput. Sci. Res.*, vol. 6, no. 2, pp. 804–812, Feb. 2026, doi: 10.47065/bulletincsr.v6i2.1021.

[24] R. Amin and A. S. F. Utami, "Prediksi Nilai Ujian Berdasarkan Kebiasaan Siswa Menggunakan Algoritma Random Forest Regressor," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 10, no. 2, p. 149, Dec. 2025, doi: 10.51211/isbi.v10i2.3722.

[25] A. Syaifudin, R. Risqiati, and Hermanus Wim Hapsoro, "IMPLEMENTASI EXPLORATORY DATA ANALYSIS UNTUK ANALISIS DATA LEMAK TUBUH," *IC Tech Maj. Ilm.*, vol. 20, no. 1, pp. 1–10, Apr. 2025, doi: 10.47775/ictech.v20i1.328.

[26] A. Astofa, P. Rosyani, R. Rahmawati, and S. Apandi, "Evaluasi Komparatif Algoritma Machine Learning untuk Prediksi Dini Diabetes,"

Bull. Comput. Sci. Res., vol. 6, no. 1 SE-, pp. 558–565, Dec. 2025, doi: 10.47065/bulletincsr.v6i1.859.
[27] A. Setiawan, Adelina, D. M. Hutabalian, R. Irnanda, H. Fredi, and Iswanto, "EVALUASI PERBANDINGAN KINERJA MODEL MACHINE LEARNING UNTUK PREDIKSI DIABETES: STUDI KASUS XGBOOST, RANDOM FOREST, DAN SVM," *INFOKOM (Informatika & Komputer)*, vol. 12, no. 2 SE-Articles, Dec. 2024, doi: 10.56689/infokom.v12i2.2350.
[28] J. J. Hidayat, A. H. Anshor, and M. S. Anwar, "Pemodelan Deteksi dan Klasifikasi Fraktur Tulang pada Radiografi X-Ray Menggunakan YOLOv8 dan Preprocessing CLAHE," *J. FASILKOM*, vol. 16, no. 1, pp. 31–45, 2026, doi: 10.37859/jf.v16i1.11241.
[29] A. I. K. Akbar and Y. P. Astuti, "Lung Cancer Classification using the Naïve Bayes Method with SMOTE," *SISTEMASI*, vol. 14, no. 6, p. 2954, Nov. 2025, doi: 10.32520/stmsi.v14i6.5607.
[30] A. Z. Kamalia, Choiriyatun Nisa Latansa, and Zaenur Rozikin, "Klasifikasi Kondisi Pasar Harga Emas ANTAM Indonesia Menggunakan Algoritma Decision Tree," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 3, pp. 2087–2098, Jan. 2026, doi: 10.62712/juktisi.v4i3.800.
[31] E. Tri Armawan, R. Safitri, and L. Riyandari, "Evaluasi dan Interpretabilitas Model Machine learning untuk Prediksi Diabetes dengan Nested cross-validation dan SHAP," *J. Pustaka AI (Pusat Akses Kaji. Teknol. Artif. Intell.*, vol. 6, no. 1 SE-Artikel, pp. 12–24, Mar. 2026, doi: 10.55382/jurnalpustakaai.v6i1.1751.
[32] S. Ijayanti and D. W. Utomo, "Implementasi Stacking Ensemble Berbasis Cross Domain untuk Klasifikasi Diabetes," *J. INFOTEKMESIN*, vol. 17, no. 01, pp. 48–56, 2026, doi: 10.35970/infotekmesin.v17i1.3000.
[33] J. J. Hidayat, C. Setyowati, and A. P. Werdana, "Perancangan Sistem Prediksi Penyakit pada Tanaman Padi Berbasis Image Processing Menggunakan Algoritma VGG-16 Transfer Learning dan K-Means Segmentation," *J. Pract. Comput. Sci.*, vol. 5, no. 1, pp. 1–15, May 2025, doi: 10.37366/jpcs.v5i1.5759.
[34] A. Ichwani, R. I. Kesuma, A. Setiawan, I. E. Wicaksono, and R. Hanifah, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning," *J. Tek. Inform.*, vol. 7, no. 1, pp. 391–410, Feb. 2026, doi: 10.52436/1.jutif.2026.7.1.5490.
[35] D. O. E. Wanto, C. Harmon, and J. Jupron, "Penerapan Metode Algoritma C.48 untuk Klasisifikasi Penyakit Diabetes," *J. Janitra Inform. dan Sist. Inf.*, vol. 5, no. 2, pp. 180–188, Oct. 2025, doi: 10.59395/qyzpt451.