TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi
Vol. 4 No. 2 (Desember 2024)
https://doi.org/10.46880/tamika.Vol4No2.pp159-166 e-ISSN: 2828-1276

PERBANDINGAN TINGKAT AKURASI PENYAKIT DIABETES MENGGUNAKAN METODE REGRESI LOGISTIK DAN RANDOM FOREST

Khoirul Adi Saputro✉, Ega Muhammad Atsir , Herliyani Hasanah
Fakultas Ilmu Komputer, Universitas Duta Bangsa, Surakarta, Indonesia
Email: khoiruladis06@gmail.com

ABSTRACT

This study analyzes and compares the performance of Logistic Regression and Random Forest methods in diabetes prediction. Diabetes is an increasingly common chronic condition. If left untreated, diabetes can cause major problems. Predicting and detecting diabetes early is very important. This study uses a dataset from Kaggle consisting of 9 demographic and clinical variables. The research method includes data collection and pre-processing, model selection, model training, and model evaluation. The results showed that the Random Forest model had 97.0% accuracy and 95.6% precision, which was higher than the Logistic Regression which had 95.9% accuracy and 86.4% precision. Although both models faced challenges in predicting positive classes, Random Forest performed better with a recall of 68.4% and F1-Score of 79.7%, compared to Logistic Regression which had a recall of 61.3% and F1-Score of 71.8%. This study advances our knowledge of how machine learning techniques can be applied to diabetes prediction and helps in the creation of more accurate and productive diagnostic tools for use in clinical settings.

Keyword: Diabetes Prediction, Logistic Regression, Random Forest, Machine Learning, Diagnostic Methods.

ABSTRAK

Penelitian ini menganalisis dan membandingkan performa metode Regresi Logistik dan Random Forest dalam prediksi diabetes. Diabetes adalah kondisi kronis yang semakin umum terjadi. Jika tidak diobati, diabetes dapat menyebabkan masalah besar. Memprediksi dan mendeteksi diabetes sejak dini sangatlah penting. Penelitian ini menggunakan dataset dari Kaggle yang terdiri dari 9 variabel demografi dan klinis. Metode penelitian mencakup pengumpulan dan pra-proses data, pemilihan model, pelatihan model, dan evaluasi model. Hasil penelitian menunjukkan bahwa model Random Forest memiliki akurasi 97,0% dan presisi 95,6%, yang lebih tinggi dibandingkan dengan Regresi Logistik yang memiliki akurasi 95,9% dan presisi 86,4%. Meskipun kedua model menghadapi tantangan dalam memprediksi kelas positif, Random Forest menunjukkan performa yang lebih baik dengan recall 68,4% dan F1-Score 79,7%, dibandingkan dengan Regresi Logistik yang memiliki recall 61,3% dan F1-Score 71,8%. Studi ini memajukan pengetahuan kita tentang bagaimana teknik pembelajaran mesin dapat diterapkan pada prediksi diabetes dan membantu dalam pembuatan alat diagnostik yang lebih akurat dan produktif untuk digunakan dalam pengaturan klinis.

Kata Kunci: Prediksi Diabetes, Regresi Logistik, Random Forest, Pembelajaran Mesin, Metode Diagnostik.

PENDAHULUAN

Diabetes merupakan salah satu penyakit kronis yang prevalensinya semakin meningkat secara global dan dapat menyebabkan berbagai komplikasi serius seperti penyakit jantung, kerusakan saraf, dan gagal ginjal jika tidak ditangani dengan baik (Yusnita et al., 2021). Menurut data dari Kementerian Kesehatan Republik Indonesia, prevalensi diabetes di Indonesia meningkat dari 8,5% pada tahun 2018 menjadi 11,2% pada tahun 2023. Prediksi dan deteksi dini diabetes mellitus sangat penting untuk mengurangi risiko komplikasi dan meningkatkan kualitas hidup pasien.

Permasalahan utama yang dihadapi dalam dunia medis adalah bagaimana mengidentifikasi individu yang berisiko tinggi terkena diabetes dengan akurasi tinggi menggunakan data klinis yang tersedia. Teknik prediksi yang akurat dapat membantu mengurangi biaya perawatan kesehatan dan memungkinkan implementasi langkah-langkah pencegahan yang tepat waktu. Tantangan utama dalam prediksi diabetes meliputi pemilihan fitur yang relevan dari data kesehatan yang kompleks, mengatasi masalah *overfitting* dan *underfitting* dalam model prediksi, serta menentukan metode prediksi yang paling efektif dan efisien dalam berbagai konteks klinis (Hovi et al., 2022).

Berbagai teknik pembelajaran mesin telah diterapkan untuk prediksi diabetes, dengan hasil yang beragam dalam hal akurasi. Misalnya, sebuah studi yang dilakukan oleh Gunawan menunjukkan akurasi

sebesar 83,33% (Gunawan et al., 2020) menggunakan metode Regresi Logistik. Penelitian lain oleh Mulyo Widodo juga menggunakan Regresi Logistik dan mencatat akurasi sebesar 92,3% (Widodo et al., 2021). Selain itu, Muhammad Salsabil dalam penelitiannya tentang prediksi diabetes menggunakan metode Random Forest mencapai akurasi sebesar 74% (Salsabil et al., 2024), sementara Apriliah dalam studi serupa mencatat akurasi yang lebih tinggi, yaitu sebesar 97,88%(Apriliah et al., 2021). Meskipun banyak penelitian yang menunjukkan hasil positif, masih terdapat kesenjangan pengetahuan mengenai performa komparatif dari metode yang berbeda dalam berbagai set data dan kondisi klinis. Penelitian lebih lanjut diperlukan untuk mengevaluasi efektivitas metode-metode ini dalam berbagai konteks klinis guna meningkatkan akurasi prediksi dan aplikasi di lapangan.

Penelitian ini bertujuan untuk menganalisis dan membandingkan performa metode Regresi Logistik dan Random Forest dalam prediksi diabetes. Penelitian ini akan mengevaluasi akurasi, sensitivitas, spesifisitas, dan nilai prediktif positif dari kedua model serta mengidentifikasi fitur-fitur kunci yang berkontribusi signifikan terhadap prediksi diabetes. Dengan demikian, penelitian ini tidak hanya bertujuan untuk meningkatkan pemahaman tentang aplikasi teknik-teknik pembelajaran mesin dalam prediksi diabetes, tetapi juga untuk memberikan kontribusi terhadap pengembangan metode diagnostik yang lebih efektif dan efisien dalam praktik medis.

METODE PENELITIAN

Model penelitian studi ini menggunakan teknik random forest dan regresi logistik untuk mempermudah analisis prediksi diabetes.

Sumber Data

Penelitian ini menggunakan dataset Diabetes Prediction Dataset yang diperoleh dari website Kaggle https://www.kaggle.com/datasets/amirmohammadparvizi/diabetes-prediction-dataset . Dataset ini terdiri dari 9 variabel yang mencakup informasi demografi dan klinis dari pasien.

Variable Data

Dataset ini berisi data yang dikumpulkan dari berbagai sumber medis yang digabungkan, dengan 9 atribut, di mana 8 atribut merupakan atribut independen, dan 1 atribut sebagai class yang menunjukkan apakah pasien menderita diabetes.

a. Atribut Independen:

Atribut independen adalah variabel yang digunakan untuk memprediksi atau menjelaskan variabel target (atribut dependen). Dalam konteks dataset untuk prediksi diabetes, atribut independen adalah faktor-faktor yang dapat mempengaruhi apakah seseorang menderita diabetes atau tidak (Manalu et al., 2015). Berikut penjelasan untuk masing-masing atribut independen dalam dataset ini:

Table 1. Atribut Independen

No.	Atribut	Penjelasan
1.	Gender	Jenis kelamin pasien
2.	Age	Umur pasien
3.	Hypertension	Apakah pasien menderita hipertensi (1: Ya, 0: Tidak)
4.	Heart disease	Apakah pasien menderita penyakit jantung (1: Ya, 0: Tidak)
5.	Smoking history	Riwayat merokok pasien
6.	Bmi	Indeks massa tubuh pasien
7.	Hba1c level	Level HbA1c dalam darah pasien
8.	Blood glucose level	Level glukosa dalam darah pasien

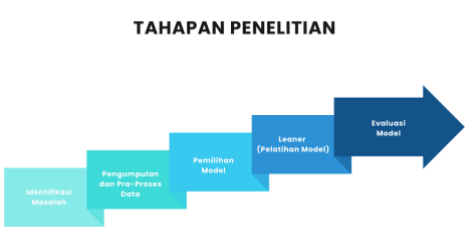
b. Atribut Dependen:

Atribut dependen adalah variabel target yang diprediksi atau dijelaskan oleh atribut independent (Setyorini et al., 2022). Dalam dataset ini, atribut dependen adalah:

Table 2. Atribut Dependen

No.	Atribut	Penjelasan
1.	Diabetes	Apakah pasien menderita diabetes (1: Ya, 0: Tidak)

Tahapan Penelitian



Gambar 1. Tahapan Penelitian

Berikut ini langkah-langkah yang dilakukan dalam penelitian:

a. Identifikasi Masalah:
   Penelitian ini mengidentifikasi bagaimana memprediksi apakah seorang pasien menderita diabetes berdasarkan data klinis yang tersedia. Penelitian ini bertujuan untuk mengembangkan model prediksi yang akurat dan andal menggunakan data klinis.
b. Pengumpulan dan Pra-Proses Data:
   Data untuk penelitian ini dikumpulkan dari sumber publik yang terpercaya, memastikan data yang digunakan memiliki kualitas dan relevansi yang tinggi. Proses pengumpulan data ini melibatkan beberapa tahap penting untuk menjamin integritas dan kesesuaian data dengan tujuan penelitian. Setelah data dikumpulkan, dilakukan langkah-langkah pra-proses sebagai berikut:
   a. Penanganan Missing Values:
      Tahap ini melibatkan identifikasi dan penanganan data yang hilang (missing values) dalam dataset. Missing values dapat menyebabkan bias dan mengurangi keakuratan model prediksi. Oleh karena itu, teknik seperti imputasi (mengisi missing values dengan nilai rata-rata, median, atau modus) atau penghapusan baris/kolom yang memiliki terlalu banyak missing values digunakan untuk memastikan kelengkapan dataset.
   b. Normalisasi Data:
      Untuk memastikan performa model yang optimal, data dinormalisasi. Normalisasi adalah proses mengubah data ke dalam skala yang konsisten, biasanya dalam rentang 0 hingga 1 atau -1 hingga 1. Hal ini penting karena berbagai fitur dalam dataset mungkin memiliki skala yang berbeda, dan normalisasi membantu dalam mempercepat konvergensi algoritma pembelajaran mesin dan meningkatkan keakuratan model.
   c. Pembagian Data:
      Setelah mengatasi nilai yang hilang dan dinormalisasi, data dibagi menjadi dua set: set pelatihan dan set pengujian. Untuk menjamin validitas model, pembagian ini dilakukan dengan rasio yang tepat, misalnya, 80% untuk pelatihan dan 20% untuk pengujian. Set pengujian digunakan untuk menilai kinerja model pada data yang belum pernah dilihat sebelumnya, memberikan prediksi yang tepat dari perilaku model dalam skenario dunia nyata. Set pelatihan digunakan untuk melatih model.
c. Pemilihan Model
   Penelitian ini memilih dua model pembelajaran mesin, yaitu Regresi Logistik dan Random Forest, karena kemampuan mereka dalam menangani data klinis dan memberikan hasil yang akurat. Implementasi model dilakukan menggunakan bahasa pemrograman Python dan pustaka scikit-learn, yang menyediakan berbagai alat dan fungsi untuk pengembangan model machine learning.
d. *Leaner*(Pelatihan Model):
   Pada tahap ini, model dilatih menggunakan dataset yang telah diproses. Proses ini mencakup pemilihan parameter optimal melalui teknik cross-validation untuk mengurangi overfitting dan memastikan generalisasi yang baik. Algoritma regresi logistik dan random forest digunakan untuk mempelajari pola dalam data, di mana regresi logistik memberikan probabilitas prediksi untuk kelas positif dan random forest meningkatkan akurasi dengan menggabungkan beberapa pohon keputusan. Model divalidasi secara internal menggunakan k-fold cross-validation untuk memastikan performa yang baik pada data baru. Dengan pelatihan yang teliti, model diharapkan memiliki prediksi yang akurat dan andal untuk diagnosis diabetes.
e. Evaluasi Model:
   Model dievaluasi menggunakan metrik akurasi, sensitivitas, spesifisitas, dan nilai prediktif positif. Selain itu, dilakukan analisis lebih lanjut terhadap fitur-fitur yang berkontribusi signifikan dalam prediksi diabetes.
   a. *AUC (Area Under the Curve)*: Mengevaluasi kapasitas model untuk membedakan antara kelas-kelas yang positif dan negatif. Nilai AUC berkisar antara 0 hingga 1, dengan nilai mendekati 1 yang menunjukkan kinerja yang unggul.
   b. *CA (Classification Accuracy)*: menghitung proporsi prediksi akurat yang dihasilkan relatif terhadap semua prediksi yang dibuat. Penilaian akurasi prediksi keseluruhan model diberikan oleh CA.
   c. *F1 Score*: Merupakan harmonisasi dari precision dan recall, memberikan gambaran seimbang tentang performa model, terutama pada dataset yang tidak seimbang.

d. *Precision*: Mengukur proporsi prediksi positif yang benar. Precision tinggi menunjukkan bahwa model menghasilkan sedikit false positives.
e. *Recall*: Mengukur kemampuan model dalam mengidentifikasi kasus positif dengan benar. Recall tinggi menunjukkan bahwa model menghasilkan sedikit false negatives.
f. *Confusion Matrix*: Menyediakan tabel yang menggambarkan kinerja model dengan menunjukkan jumlah *true positives*, *true negatives*, *false positives*, dan *false negatives*. Confusion matrix membantu dalam memahami di mana model membuat kesalahan.
g. ROC (*Receiver Operating Characteristic)* Analysis: Menggunakan kurva ROC untuk mengevaluasi kemampuan model dalam memisahkan kelas positif dan negatif pada berbagai threshold. Analisis ROC membantu dalam memilih threshold yang optimal untuk memaksimalkan performa model.

Metode Regresi Logistik

Metode statistik untuk mensimulasikan hubungan antara variabel dependen biner dan satu atau lebih variabel independen disebut regresi logistik. Metode ini sering diterapkan dalam analisis data untuk menunjukkan bagaimana variabel respons terkait dengan variabel penjelasnya. Regresi logistik menjadi semakin populer karena kemampuannya dalam menangani variabel respons yang biner atau dikotomik. Dalam konteks prediksi diabetes, regresi logistik digunakan untuk memperkirakan probabilitas seseorang menderita diabetes berdasarkan variabel klinis seperti kadar glukosa darah, tekanan darah, dan indeks massa tubuh. Persamaan regresi logistik dinyatakan sebagai berikut(Britanthia et al., 2020):

$$Y = \beta 0 + \beta 1\, X1 + \cdots + \beta kXk + \varepsilon$$

Keterangan :
Y adalah variabel respon
X adalah variabel penjelas
β adalah konstanta atau intersep
βi adalah koefisien X, untuk i=1,2…,k
ε adalah residu atau galat

Dalam regresi logistik, untuk menguji hipotesis dan membuat interval kepercayaan parameter, diasumsikan bahwa:

a. Rata-rata kesalahan adalah nol untuk menguji hipotesis dan menghasilkan interval kepercayaan untuk parameter.
b. Varians kesalahan (homoskedastik) adalah konstan.
c. Kesalahan tidak menunjukkan tanda-tanda autokorelasi.
d. Ada distribusi normal dari kesalahan.

Akibatnya, parameter kesalahan dalam model regresi logistik sering kali dikosongkan. Karena nilai sebenarnya dari parameter tidak dapat ditentukan, maka nilai galat tidak berdampak pada model jika tidak bernilai nol.

Metode Random Forest

Berdasarkan jumlah pohon yang terbentuk, Random Forest merupakan teknik yang umum digunakan untuk klasifikasi data berskala besar karena akurasi prediksinya yang tinggi (Reinardus et al., 2022). Pohon-pohon dalam algoritma ini dibentuk secara acak dan kemudian digabungkan. Prosedur untuk menjalankan random forest adalah sebagai berikut:

1. Pilih sampel acak ukuran-n dengan pengembalian. Tahap bootstrap adalah apa yang kita sebut sebagai tahap ini.
2. Tanpa pemangkasan, kembangkan pohon ke ukuran maksimum menggunakan sampel bootstrap. Pada setiap tingkat proses seleksi, seleksi fitur acak digunakan untuk membentuk pohon; yaitu, k variabel penjelas dipilih secara acak.
3. Lanjutkan langkah 1 dan 2 hingga terbentuk hutan dengan sejumlah pohon (Reka, 2022).

Random forest hanya dapat menunjukkan relevansi dari sebuah variabel; ia tidak dapat menentukan nilai yang tepat dari setiap variabel karena bergantung pada ensemble pohon keputusan. Rumus berikut ini digunakan untuk menghitung tingkat kepentingan variabel. Dengan mengasumsikan satu set q variabel penjelas dengan h = 1, 2, ..., q, signifikansi dari setiap variabel penjelas Xh ditentukan oleh Mean Decrease in Gini (MDG) (Sandri & Zuccolotto, 2008) :

$$\mathrm{DGh} = 1\ \mathrm{k}\ \sum [d(h,t)I(h,t)]\ t$$

Keterangan
k adalah banyaknya pohon dalam random forest
$d(h,t)$ adalah besar penurunan indeks Gini untuk variabel penjelas Xh pada simpul t
$I(h,t)$ adalah { 1; Xh memilah simpul t 0; selanjutnya

HASIL DAN PEMBAHASAN

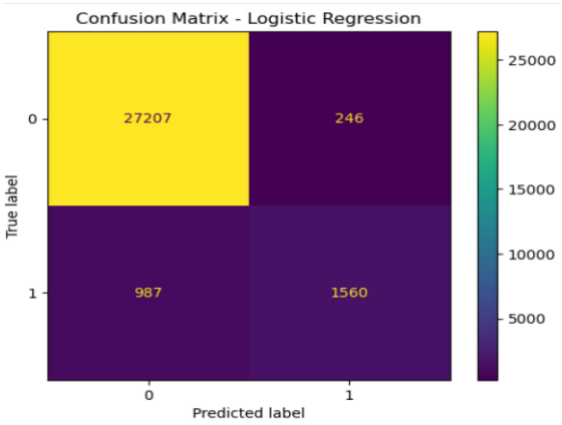
Olah Data

Dataset yang digunakan dalam penelitian ini adalah Diabetes Prediction Dataset dari Kaggle, yang dirancang untuk memprediksi kemungkinan seseorang mengidap diabetes. Dataset ini mencakup 100.000 instance dengan 9 atribut, termasuk kondisi klinis dan gaya hidup pasien, serta satu atribut target yang menunjukkan ada tidaknya diabetes. Untuk analisis, kolom gender diubah menjadi kode numerik (Male: 0, Female: 1) dan riwayat merokok dikategorikan sebagai berikut: 0 (never), 1 (No Info), 2 (former), 3 (current), 4 (not current), dan 5 (ever). Transformasi ini memudahkan analisis lebih lanjut dan memberikan wawasan mendalam tentang faktor risiko diabetes.

Table 3. Olah Data Prediksi Diabetes

Atribut	0	1	...	99998	99999
Gender	0	0		0	0
Age	80.0	54.0		24.0	57.0
Hypertension	0	0		0	0
Heart_disease	1	0		0	0
Smoking_history	0	1		0	3
Bmi	25.19	27.32		35.42	22.43
HbA1c_level	6.6	6.6		4.0	6.6
Blood_glucose_level	140	80		100	90

Evaluasi



Gambar 2. Confusion Matrix Regresi Logistik

Gambar 2 menunjukkan *Confusion Matrix* untuk model *Logistic Regression*. Berikut adalah pembacaan dan penjelasan dari nilai-nilai dalam matriks tersebut:

a. *True Negative (TN)*: 27,207 (Nilai pada baris 0, kolom 0) - Jumlah sampel yang diprediksi oleh model akan menjadi negatif dan yang benar-benar negatif.
b. *False Positive (FP)*: 246 (Nilai pada baris 0, kolom 1) - Jumlah sampel yang diprediksi oleh model sebagai positif, tetapi pada kenyataannya negatif.
c. *False Negative (FN)*: 987 (Nilai pada baris 1, kolom 0) - Jumlah sampel yang diprediksi oleh model sebagai negatif tetapi sebenarnya positif.
d. *True Positive (TP)*: 1,560 (Nilai pada baris 1, kolom 1) - Jumlah sampel yang diprediksi oleh model sebagai positif dan yang sebenarnya positif.

Dengan nilai-nilai ini, kita dapat menghitung beberapa metrik evaluasi kinerja model seperti berikut:

a. Akurasi *(Accuracy)*:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1560 + 27207}{1560 + 27207 + 246 + 987} = 95{,}9\%$$

b. Presisi *(Precision)*:

$$Presisi = \frac{TP}{TP + FP} = \frac{1560}{1560 + 246} = 86{,}4\%$$

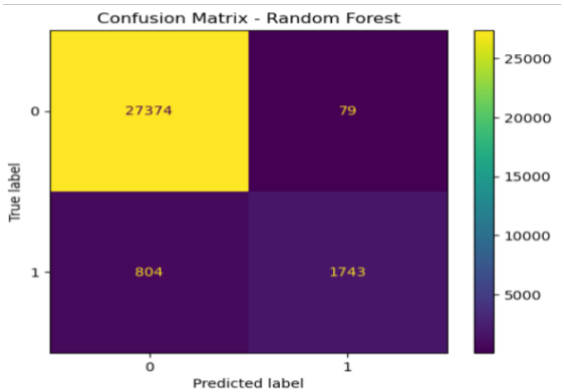
c. *Recall*:

$$Recall = \frac{TP}{TP + FN} = \frac{1560}{1560 + 987} = 61{,}3\%$$

d. *F1-Score*:

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} = 2 \times \frac{0{,}864 \times 0{,}613}{0{,}864 + 0{,}613} = 0{,}718$$

Confusion matrix ini menunjukkan bahwa model *Logistic Regression* cukup baik dalam memprediksi kelas negatif dengan jumlah yang tinggi (TN = 27,207), namun masih terdapat kesalahan dalam memprediksi kelas positif (FN = 987).



Gambar 3. Confusion Matrix Random Forest

Gambar 3 tersebut adalah Confusion Matrix untuk model Random Forest. Berikut adalah pembacaan dan penjelasan dari nilai-nilai dalam matriks tersebut:

a. *True Negative (TN)*: 27,374 (Nilai pada baris 0, kolom 0) - Jumlah sampel yang diprediksi oleh model akan menjadi negatif dan yang benar-benar negatif.
b. *False Positive (FP)*: 79 (Nilai pada baris 0, kolom 1) - Jumlah sampel yang diprediksi oleh model sebagai positif, tetapi pada kenyataannya negatif.
c. *False Negative (FN)*: 804 (Nilai pada baris 1, kolom 0) - Jumlah sampel yang diprediksi oleh model sebagai negatif tetapi sebenarnya positif.
d. *True Positive (TP)*: 1,743 (Nilai pada baris 1, kolom 1) - Jumlah sampel yang diprediksi oleh model sebagai positif dan yang sebenarnya positif.

Dengan nilai-nilai ini, kita dapat menghitung beberapa metrik evaluasi kinerja model seperti berikut:

1. Akurasi *(Accuracy)*:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1743 + 27374}{1743 + 27374 + 79 + 804} = 97{,}0\%$$

2. Presisi *(Precision)*:

$$Presisi = \frac{TP}{TP + FP} = \frac{1743}{1743 + 79} = 95{,}6\%$$

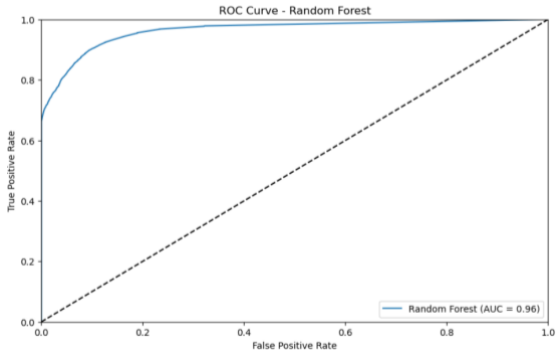
3. *Recall*:

$$Recall = \frac{TP}{TP + FN} = \frac{1743}{1743 + 804} = 68{,}4\%$$

4. *F1-Score*

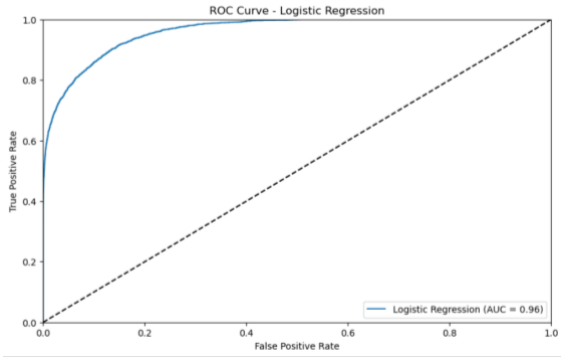
$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} = 2 \times \frac{0{,}956 \times 0{,}684}{0{,}956 + 0{,}684} = 0{,}797$$

Confusion matrix ini menunjukkan bahwa model Random Forest cukup baik dalam memprediksi kelas negatif dengan jumlah yang tinggi (TN = 27,374), namun masih terdapat kesalahan dalam memprediksi kelas positif (FN = 804).



Gambar 4. ROC Regresi Logistik

Model regresi logistik melakukan pekerjaan yang sangat baik dalam membedakan antara kelas positif dan negatif, yang dibuktikan dengan nilai AUC sebesar 0,96. *Area Under the Curve*, atau AUC, adalah statistik yang mengindikasikan seberapa baik sebuah model dapat membedakan antara dua kelas. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan yang tinggi untuk membedakan antara kejadian positif dan negatif. Dalam grafik ROC, kurva yang mendekati sumbu kiri atas menandakan bahwa model memiliki tingkat *false positive* yang rendah dan *true positive* rate yang tinggi, yang berarti model ini jarang membuat kesalahan dengan mengklasifikasikan kelas negatif sebagai positif dan sebaliknya. Hal ini menunjukkan bahwa model regresi logistik sangat andal dalam mengidentifikasi kelas yang benar.



Gambar 5. ROC Random Forest

Selain itu, nilai AUC sebesar 0,96 untuk model Random Forest menunjukkan bahwa model ini memiliki kinerja yang sangat baik dalam membedakan antara klasifikasi positif dan negatif. Mirip dengan regresi logistik, nilai AUC yang tinggi dari model Random Forest menunjukkan bahwa model ini dapat secara efektif membedakan kedua kelompok. Kurva ROC untuk model Random Forest juga mendekati sumbu kiri atas, yang berarti model ini memiliki tingkat

false positive yang rendah dan *true positive rate* yang tinggi. Dengan kata lain, model ini efektif dalam memprediksi kelas positif dengan benar tanpa banyak kesalahan dalam mengklasifikasikan kelas negatif sebagai positif. Performanya yang konsisten dengan regresi logistik menunjukkan bahwa Random Forest juga merupakan pilihan yang sangat andal untuk tugas klasifikasi ini.

Table 4. Kinerja Algoritma

Model	AUC	CA	F1	Precision	Recall
Regresi Logistik	0,96%	95,9%	0,718	86,4%	61,3%
Random Forest	0,96%	97%	0,797	95,6%	68,4%

Secara keseluruhan, kedua model ini memiliki performa yang mengagumkan dalam membedakan antara kelas positif dan negatif(ditunjukkan oleh nilai AUC yang sama), Random Forest menunjukkan performa yang lebih baik dalam hampir semua metrik lainnya. Random Forest memiliki akurasi, *F1 score*, *Precision*, dan *Recall* yang lebih tinggi, menjadikannya pilihan yang lebih unggul untuk tugas klasifikasi ini dibandingkan dengan Regresi Logistik.

KESIMPULAN

1. Akurasi Model: Model Random Forest menunjukkan akurasi yang lebih tinggi (97.0%) dibandingkan dengan model Regresi Logistik (95.9%). Hal ini menunjukkan bahwa Random Forest lebih unggul dalam memprediksi diabetes secara keseluruhan.
2. Presisi dan *Recall*:
   a. Presisi model Random Forest (95.6%) juga lebih tinggi dibandingkan dengan Regresi Logistik (86.4%). Ini menunjukkan bahwa Random Forest lebih baik dalam meminimalisir *false positives*.
   b. *Recall* untuk Random Forest (68.4%) lebih tinggi dibandingkan dengan Regresi Logistik (61.3%), yang berarti Random Forest lebih efektif dalam mendeteksi kasus diabetes yang sebenarnya *(true positives)*.
3. *F1-Score*: *F1-Score* untuk Random Forest (79.7%) lebih tinggi dibandingkan dengan Regresi Logistik (71.8%). Hasilnya menunjukkan bahwa Random Forest memiliki keseimbangan yang lebih baik dalam memprediksi diabetes daripada *F1-Score*, yang memberikan pemahaman tentang pertukaran antara presisi dan recall.
4. ROC-AUC: Model Regresi Logistik menunjukkan tingkat kinerja yang tinggi dalam membedakan antara kelas positif dan negatif, yang dibuktikan dengan nilai AUC sebesar 0,96. Meski demikian, evaluasi lebih lanjut menunjukkan bahwa Random Forest tetap lebih unggul dalam performa keseluruhan.

DAFTAR PUSTAKA

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). *SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest* (Vol. 10, Issue 1). http://sistemasi.ftik.unisi.ac.id

Britanthia, L., Tanujaya, C., Susanto, B., & Saragih, A. (2020). Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify. *Indonesian Journal of Data and Science (IJODAS)*, *1*(3), 68–78.

Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(3), 280. https://doi.org/10.26418/jp.v6i3.40718

Hovi, H. S. W., Id Hadiana, A., & Rakhmat Umbara, F. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). *Informatics and Digital Expert (INDEX)*, *4*(1), 40–45. https://doi.org/10.36423/index.v4i1.895

Salsabil, M., Azizah, N. L., & Eviyanti, A. (2024). Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost. *Jurnal Ilmiah Komputasi*, *23*(1). https://doi.org/10.32409/jikstik.23.1.3507

Manalu, E., Sianturi, F. A., & Manalu, M. R. (2015). (). *Jl. Iskandar Muda No.1 Medan*, *1*(2), 1.

Reinardus, Haristu, A., & Rosa, P. H. P. (2022.). *Penerapan Metode Random Forest untuk Prediksi Win Ratio Pemain Player Unknown Battleground*. *4*(2). http://ejournal.ust.ac.id/index.php/Jurnal_Means

Reka, C. (2022). Bulletin of Data Science Penerapan Data Mining Analisa Data Penjualan Obat Menggunakan Metode Random Forest. *Media Online)*, *1*(3), 117. https://ejurnal.seminar-id.com/index.php/bulletinds

Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611–628. https://doi.org/10.1198/106186008X344522

Setyorini, W., Jayusman, H., FarihFR, M., Manajemen, P., & Ekonomi, F. (2022). *Pengaruh Atribut Produk Terhadap Keputusan Pembelian (Studi Kasus : Rumah Makan Soto Lamongan Imam Jl. Iskandar Kel. Madurejo)* (Vol. 11, Issue 1).

Widodo, A. M., Anggraeni, Y. S., Anwar, N., Ichwani, A., & Sekti, B. A. (2021). Performansi K-NN, J48, Naive Bayes dan Regresi Logistik sebagai Algoritma Pengklasifikasi Diabetes. *Prosiding SISFOTEK*, 27–33.

Yusnita, Y., Djafar, M. H. A., & Tuharea, R. (2021). Risiko Gejala Komplikasi Diabetes Mellitus Tipe II di UPTD Diabetes Center Kota Ternate. *Media Publikasi Promosi Kesehatan Indonesia (MPPKI)*, *4*(1), 60–73.