LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

Optimasi Prediksi Diabetes Dengan Algoritma XGBoost Dan Teknik Preprocessing Data

Andhika Brahmandjati[1], Abd Mizwar A. Rahim[2*], Firman Asharudin[3]

[1,2,3]Ilmu Komputer, Informatika, Universitas Amikom, Yogyakarta, Indonesia
Email: [1]andhikabrahmandjati@students.amikom.ac.id, [2*]abdulmizwar@amikom.ac.id,
[3]Firman_asharudin@amikom.ac.id
(* : coressponding author)

Abstrak − Diabetes adalah penyakit metabolik kronis yang menjadi tantangan kesehatan global dengan prevalensi yang terus meningkat. Dalam upaya meningkatkan deteksi dini, penelitian ini menggunakan algoritma XGBoost (Extreme Gradient Boosting) untuk klasifikasi data diabetes. Model dikembangkan dengan tahapan preprocessing data yang komprehensif dan diuji pada dataset yang seimbang. Evaluasi performa model dilakukan menggunakan confusion matrix dan classification report. Hasil penelitian menunjukkan bahwa model memiliki akurasi 98%, precision 98%, recall 97%-99%, dan F1-score 98% pada kedua kelas (positif dan negatif). Confusion matrix mencatat True Positive (TP) sebanyak 201, True Negative (TN) sebanyak 190, False Positive (FP) sebanyak 6, dan False Negative (FN) sebanyak 3. Analisis lebih lanjut menunjukkan bahwa model memiliki performa yang sangat konsisten tanpa bias terhadap salah satu kelas. Dibandingkan dengan penelitian sebelumnya, model ini menunjukkan peningkatan akurasi yang signifikan hingga lebih dari 15%, dengan metode terbaik sebelumnya, yaitu Naive Bayes, hanya mencapai akurasi 82,3%. Keunggulan ini menjadikan algoritma XGBoost yang diterapkan dalam penelitian ini sebagai pendekatan yang lebih andal dan efektif untuk deteksi diabetes. Hasil penelitian ini memberikan kontribusi baru dalam pengembangan sistem prediksi berbasis data untuk mendukung pengambilan keputusan klinis.

Kata Kunci: Diabetes; Machine Learning; XGBoost; Deteksi Dini; Confusion Matrix

Abstract *− Diabetes is a chronic metabolic disease that is a global health challenge with increasing prevalence. In an effort to improve early detection, this study uses the XGBoost (Extreme Gradient Boosting) algorithm for diabetes data classification. The model was developed with comprehensive data preprocessing stages and tested on a balanced dataset. Model performance evaluation was carried out using a confusion matrix and classification report. The results showed that the model had an accuracy of 98%, precision of 98%, recall of 97%-99%, and F1-score of 98% in both classes (positive and negative). The confusion matrix recorded 201 True Positive (TP), 190 True Negative (TN), 6 False Positive (FP), and 3 False Negative (FN). Further analysis showed that the model had very consistent performance without bias towards one class. Compared to previous studies, this model showed a significant increase in accuracy of more than 15%, with the previous best method, Naive Bayes, only achieving an accuracy of 82.3%. These advantages make the XGBoost algorithm applied in this study a more reliable and effective approach to diabetes detection. The results of this study provide new contributions to the development of data-based prediction systems to support clinical decision-making.*

Keywords*: Diabetes; Machine Learning; XGBoost; Early Detection; Confusion Matrix*

1. PENDAHULUAN

Diabetes adalah salah satu penyakit metabolik kronis yang telah menjadi masalah kesehatan global. Menurut World Health Organization (WHO), diabetes ditandai oleh peningkatan kadar glukosa darah yang dapat menyebabkan komplikasi serius, seperti kerusakan jantung, pembuluh darah, mata, ginjal, dan saraf(Wahyuni et al., 2019a). Diabetes tipe 2 adalah jenis diabetes yang paling umum, terjadi ketika tubuh menjadi resisten terhadap insulin atau tidak memproduksi insulin yang cukup(Sumah, 2018). Kondisi ini biasanya dialami oleh orang dewasa, dan prevalensinya telah meningkat secara signifikan di semua negara, baik berpenghasilan rendah, menengah, maupun tinggi, selama tiga dekade terakhir(Widyana & Afriansyah, 2022). Sementara itu, diabetes tipe 1, yang sebelumnya dikenal sebagai diabetes remaja, adalah kondisi kronis di mana pankreas hampir tidak menghasilkan insulin(J, 2019). Untuk penderita diabetes, akses terhadap pengobatan, termasuk insulin, sangat penting demi kelangsungan hidup mereka(L. A. R. Putri & Ellyani Abadi, 2021).

Di tengah meningkatnya jumlah penderita diabetes yang diperkirakan mencapai 830 juta orang secara global, lebih dari separuhnya tidak mendapatkan pengobatan yang memadai(D. P. Putri et al., 2021). Hal ini menekankan urgensi penelitian untuk menemukan solusi yang lebih efektif, termasuk dalam meningkatkan kemampuan deteksi dini diabetes. Salah satu pendekatan modern

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

yang mulai banyak digunakan adalah pemanfaatan metode Machine Learning (ML)(Putri Nurhayati et al., 2022). Dalam beberapa tahun terakhir, ML telah menjadi alat yang sangat efektif dalam berbagai bidang, termasuk kesehatan, agronomi, bisnis, dan ekonomi. Dalam konteks diabetes, ML telah digunakan untuk klasifikasi, prediksi, hingga pengambilan keputusan klinis berbasis data(Salasa et al., 2019).

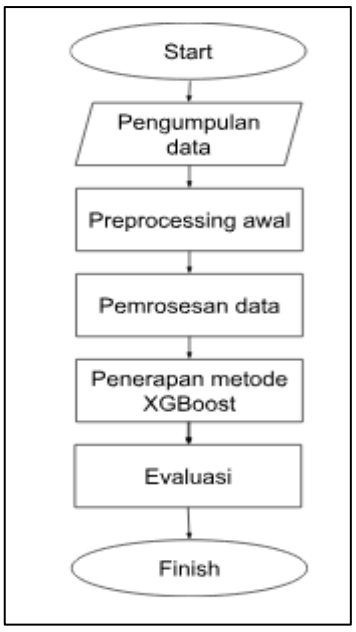
XGBoost (Extreme Gradient Boosting) adalah salah satu algoritma Machine Learning yang unggul untuk tugas-tugas klasifikasi. Diperkenalkan oleh Dr. Tianqi Chen pada tahun 2014, algoritma ini merupakan pengembangan dari gradient boosting dengan beberapa keunggulan, seperti pengurangan risiko overfitting melalui model yang lebih teratur dan peningkatan akurasi prediktor(Ramadona et al., 2021). XGBoost bekerja dengan membangun serangkaian pohon regresi yang secara iteratif mengurangi nilai error untuk menghasilkan prediksi akhir yang optimal(Nyayu Mevia Fiqi & Zulmansyah, 2021).

Penelitian sebelumnya menunjukkan bahwa berbagai algoritma telah digunakan untuk mendeteksi diabetes, seperti Random Forest yang mencapai akurasi 90,38%(Nunik Purnamasari et al., 2024), dan SVM dengan akurasi terbaik 77,73% (Anwari, 2021). Namun, banyak penelitian tersebut masih memiliki keterbatasan, terutama dalam tahapan preprocessing data, yang mencakup pengecekan missing value dan distribusi kelas untuk mengatasi masalah oversampling dan undersampling(Denggos, 2023).

Penelitian ini bertujuan untuk meningkatkan akurasi prediksi diabetes dengan menggunakan algoritma XGBoost. Pemilihan algoritma ini didasarkan pada kemampuannya dalam menangani data yang kompleks dan meningkatkan akurasi prediktor melalui pendekatan ensemble(Nazliansyah et al., 2022). Selain itu, penelitian ini akan mengintegrasikan tahapan preprocessing data yang lebih komprehensif untuk memastikan kualitas data yang optimal. Evaluasi kinerja model akan dilakukan menggunakan confusion matrix untuk menganalisis efektivitas prediksi dalam mendeteksi kasus diabetes(Nizar & Amelia, 2022).

2. METODE

Dalam Penelitian ini menerapkan metode eksperimen, dengan tahapan yang meliputi pemilihan bahan, prosedur yang dijalankan, serta analisis data hingga evaluasi metode klasifikasi yang diterapkan. Gambar 1 menggambarkan secara keseluruhan alur penelitian ini, mulai dari bahan yang digunakan hingga langkah-langkah yang dilakukan dalam proses penelitian.



Gambar 1. Tahapan Penelitian

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

Pada gambar 1 menggambarkan alur penelitian yang dilakukan. Tahapan dimulai dengan pengumpulan data, dilanjutkan dengan preprocessing data untuk menyiapkan data. Selanjutnya, data diproses menggunakan metode XGBoost, diikuti dengan penyetelan hiperparameter untuk meningkatkan kinerja, dan diakhiri dengan evaluasi guna memastikan hasil yang optimal.

2.1 Kumpulan Data

Terdapat dua dataset yang digunakan dalam penelitian ini diambil dari Kaggle sumber pertama : https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset, dan sumber kedua https://www.kaggle.com/datasets/mathchi/diabetes-data-set, lalu dilakukan penggabungan dan berjumlah 1536 entri. Dataset ini mencakup 9 variabel yang berkaitan dengan informasi mengenai diabetes, dapat dilihat pada tabel 1 Kumpulan Data.

Tabel 1. Jenis jenis database

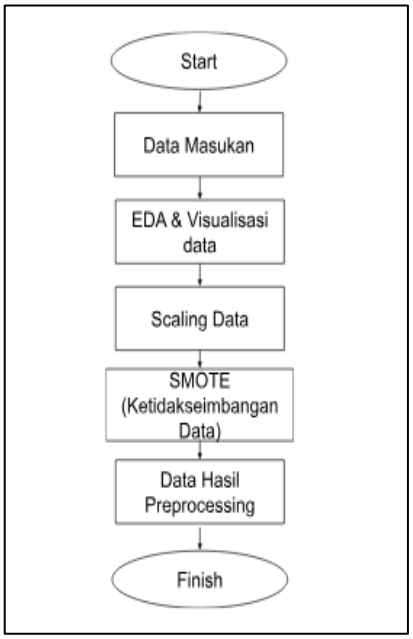
Atribut	Keterangan
Pregnancies	Jumlah Kehamilan
Glucose	Konsentrasi glukosa plasma 2 jam setelah uji toleransi glukosa oral
BloodPressure	Tekanan darah diastolik
SkinThickness	Ketebalan lipatan kulit trisep
Insulin	Insulin serum 2 jam
BMI	Indikator untuk menentukan kategori berat badan
DiabetesPedigreeFunction	Fungsi silsilah diabetes
Age	Umur
Outcome	kelas

Data ini sudah dalam kondisi optimal, bebas dari nilai yang hilang, dan setiap kolom telah disesuaikan dengan tipe data yang tepat, seperti integer dan float.

2.2. Preprocessing.

Data preprocessing adalah langkah esensial dalam analisis data yang bertujuan untuk membersihkan serta mengonversi data mentah menjadi format yang cocok untuk dianalisis(A. Rahim et al., 2023). Langkah ini berfokus pada peningkatan kualitas data dengan menyelesaikan masalah seperti kekosongan data, pencilan, dan ketidaksesuaian format(Rahim et al., 2022). Tahapan keseluruhan pre-prosessing data sebelum dilakukan pemodelan dengan XGBoost dapat dilihat pada gambar 2. Beberapa langkah kunci dalam tahap ini termasuk eksplorasi data dan visualisasi (EDA), proses scaling data, serta penanganan ketidakseimbangan data melalui metode SMOTE. Setelah tahapan ini selesai, data sudah siap untuk dimasukkan ke dalam model.

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125



Gambar 2. Tahap Preprocessing

1. Eksplorasi Data dan Visualisasi (EDA)

   Exploratory Data Analysis (EDA) adalah proses menganalisis dan menampilkan data bertujuan mendapatkan pemahaman yang lebih baik tentang wawasan dari data(KOWALSKI et al., 2017). Pada tahap awal, kami melakukan eksplorasi data untuk memperoleh pemahaman yang lebih mendalam mengenai karakteristik dataset(Cobre et al., 2021). Proses ini melibatkan analisis distribusi data dan identifikasi hubungan antara fitur dengan menggunakan berbagai jenis visualisasi, seperti histogram, diagram kotak (boxplot), dan diagram pencar (scatter plot)(Dai & Genton, 2018). Melalui visualisasi tersebut, kita dapat melihat bagaimana data tersebar, mengidentifikasi pola yang ada, serta mendeteksi kemungkinan adanya data yang tidak sesuai atau pencilan (outlier) (Hinterberger, 2018).

2. Proses Skalasi Data.

   Setelah melakukan eksplorasi data, langkah berikutnya adalah melakukan normalisasi pada data dengan skala yang berbeda antar fitur, hal ini dapat mengganggu proses pelatihan model, terutama pada algoritma yang sensitif terhadap skala(Wahyuni et al., 2019). Dalam penelitian ini, kami menggunakan Min-Max Scaler untuk mengubah nilai fitur ke dalam rentan 0 hingga 1, dengan cara ini, setiap fitur memiliki kontribusi yang setara terhadap model(Rahim et al., 2022).

3. Penanganan Ketidakseimbangan Data dengan SMOTE.

   Masalah ketidakseimbangan antara kelas dalam dataset sering muncul, terutama ketika satu kelas memiliki jumlah data yang sangat dominan(Krawczyk, 2016). Ketidakseimbangan ini dapat menyebabkan model cenderung memprediksi kelas mayoritas dengan lebih baik, namun mengabaikan kelas minoritas(Buda et al., 2018). Untuk mengatasi hal ini, kami menerapkan teknik SMOTE (Synthetic Minority Over-sampling Technique), yang menghasilkan sampel sintetis untuk kelas yang lebih sedikit(Fernández et al., 2018). Dengan demikian, distribusi kelas menjadi lebih seimbang, yang membantu model untuk mempelajari pola dari kedua kelas secara lebih adil dan meningkatkan akurasi prediksi(Haibo He & Garcia, 2009).

2.3. Split Data

Dataset pada penelitian ini dipisahkan menggunakan metode Train/Test Split dengan rasio 80% dan 20%. Proses ini menghasilkan 1600 data untuk pelatihan dan 400 data untuk pengujian,

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

dari total 2000 entri hasil resampling menggunakan metode SMOTE. Resampling dilakukan untuk memastikan jumlah data pada setiap kelas seimbang, yaitu masing-masing 1000 entri untuk kelas positif dan negatif.

Tahapan ini bertujuan untuk menguji kemampuan model dalam memprediksi data baru yang tidak digunakan selama pelatihan, sehingga performa model dapat dievaluasi secara objektif. Pembagian data pelatihan dan pengujian secara rinci dapat dilihat pada Tabel 2, yang menjelaskan proporsi data yang digunakan dalam penelitian ini.

Tabel 2. Split Data

Deskripsi	Data Pelatihan	Data Pengujian	Total
Proporsi	80%	20%	100%
Jumlah	1600	400	2000

2.4. Klasifikasi menggunakan algoritma Extreme Gradient Boosting

XGBoost adalah algoritma machine learning berbasis gradient boosting yang unggul dalam akurasi, efisiensi, dan pencegahan overfitting. Algoritma ini banyak digunakan di berbagai bidang, termasuk lingkungan, kesehatan, dan keuangan, untuk tugas-tugas prediksi kompleks. Dengan kemampuan komputasi paralel dan fitur regulasi, XGBoost sering menjadi pilihan utama dibanding algoritma lain seperti Random Forest dan LightGBM. Proses klasifikasi ini akan dilakukan dengan mengatur parameter metode juga, parameter tersebut dapat dilihat pada table 3.

Tabel 3. Parameter XGBoost

Parameter	Keterangan
max depth	Menentukan kedalaman maksimum setiap pohon.
eta (learning rate)	Mengontrol ukuran langkah pembaruan pada setiap iterasi.
min child weight	Menentukan jumlah minimum bobot sum dari sampel dalam leaf node.
n estimators	Jumlah total pohon (iterasi boosting) yang akan dibangun oleh model.
subsample	Persentase data yang digunakan untuk membangun setiap pohon.
gamma	Ambang batas minimal untuk split.
random state	Menetapkan seed untuk kontrol pengacakan, memastikan hasil eksperimen dapat direproduksi.

2.5. Evaluasi Metode

Penelitian ini menggunakan Confusion Matriks untuk mengevaluasi kinerja model dengan empat komponen utama yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Hasil akhir dari evaluasi berupa nilai akurasi, presisi, recall, dan F1-Score. Berikut ini beberapa persamaan dari nilai evaluasi.

1. Akurasi (ACC): Efektivitas keseluruhan dari klasifikasi, rumus akurasi dapat dilihat pada persamaan (1).

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

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

2. Presisi (PREC): Persentase label positif yang benar dari seluruh label yang diprediksi sebagai positif, rumus presisi dapat dilihat pada persamaan (2)..

$$\text{Presisi} = \frac{TP}{TP + FP}$$

(2)

3. Recall (REC), atau Sensitivitas: Efektivitas pengklasifikasi dalam mengidentifikasi label positif yang benar, rumus recall dapat dilihat pada persamaan (3)..

$$\text{Akurasi} = \frac{TP}{TP + FN}$$

(3)

4. F1-Score: Rata-rata harmonis antara presisi dan recall, yang memberikan gambaran lebih lengkap tentang kinerja model, terutama ketika data tidak seimbang, rumus F1-Score dapat dilihat pada persamaan (4).

$$\text{Akurasi} = 2\,x\frac{Presisi + Recall}{Presisi + Recall}$$

(4)

3. ANALISA DAN PEMBAHASAN

3.1 Dataset

Penelitian ini memanfaatkan dataset dari Kaggle, yang sering digunakan sebagai acuan dalam penelitian. Dataset ini memiliki 9 kolom, terdiri dari 8 kolom sebagai variabel independen, serta 1 kolom sebagai variabel target (dependen) untuk memprediksi kemungkinan terjadinya diabetes. Ilustrasi dataset dapat dilihat pada table 4.

Tabel 4. Dataset

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Tabel 4. dataset ini merupakan iluastrasi dari data pengolahan penelitian ini yang memiliki peran dalam menentukan apakah seseorang berpotensi terkena diabetes. Variabel-variabel yang digunakan sebagai fitur adalah: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, dan Age. Sementara itu, kolom Outcome bertindak sebagai label dengan dua kemungkinan nilai, yaitu 0 untuk pasien yang tidak memiliki diabetes, dan 1 untuk pasien yang didiagnosa mengidap diabetes.

3.2 Preprocessing

Setelah itu lakukan preprocessing dengan teknik normalisasi data menggunakan min-max scaler, dataset berisi 1536 entri dengan parameter kesehatan seperti jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, insulin, BMI, riwayat diabetes dalam keluarga, usia, dan hasil akhir. Hasil pre-prosessing ini dapat dilihat pada tabel 5.

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

Tabel 5. Hasil Preprocessing Dataset

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Tabel 5. Merupakan hasil pre-prosessing data dengan teknik normalisasi. Hasil tersebut diketahui bahwa Semua nilai telah distandarisasi dalam rentang 0 hingga 1. Kadar glukosa memiliki rata-rata 0.607510 dengan variasi 0.160614, sementara BMI rata-rata 0.476790 dengan variasi 0.117460. Dataset juga mencakup nilai kuartil bawah, median, dan kuartil atas, yang mempermudah analisis dan memastikan akurasi klasifikasi selanjutnya.

3.3. Split data.

Pembagianyang dilakukan penelitian ini yaitudata training dan data testing menjadi 80/20. Dengan jumlah pembagian tersebut mempunyai tujuan melihat model dalam memprediksi ketika mempunyai data test dengan jumlah 400, secara umum model machine learning mendapatkan hasil akurasi yang baik jika memiliki jumlah data testing yang sedikit dan data training yang banyak. Maka dalam penelitian ini meningkatkan data test dan menguji apakah model mendapatkan hasil yang baik atau tidak. Pada Tabel 5menggambarkan pembagian data yang di lakukan.

Tabel 6. Train/Test Split

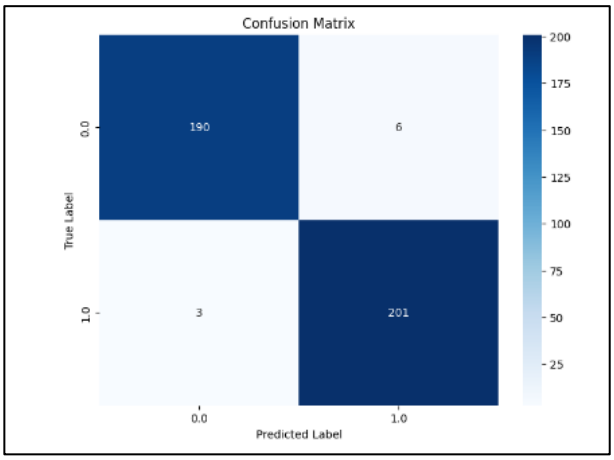
Keterangan	Data Training	Data Testing	Total
Proporsi	80%	20%	100%
Jumlah	1600	400	2000

Pada Tabel 5 di jelaskan bahwa pembagian data training dan data testing yang dilakukan menjadi 80/20, 80% untuk data training yang berjumlah 1600 data dan 20% untuk data testing yang berjumlah 400 data, dengan itu jumlah keseluruhan data dari dataset berjumlah 2000 data.

3.4. Evaluasi Metode.

Confusion Matrix menunjukkan perbandingan antara hasil prediksi model dan nilai sebenarnya dalam klasifikasi. Matriks ini terdiri dari dua kelas, yaitu 0 untuk data yang tidak terindikasi diabetes dan 1 untuk data yang terindikasi diabetes. Hasil yang diperoleh dari Confusion Matrix adalah: True Positive (TP) sebanyak 201, True Negative (TN) sebanyak 190, False Positive (FP) sebanyak 6, dan False Negative (FN) sebanyak 3. Berdasarkan hasil ini, dilakukan analisis lebih lanjut untuk mengevaluasi performa model. Model yang digunakan dalam penelitian ini adalah XGBoost (Extreme Gradient Boosting), dengan hasil terbaik menunjukkan akurasi sebesar 98%. Hasil dari confusion matriks ini dapat dilihat pada gambar 3 berikut.

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125



Gambar 3. Confusion Matrix.

Gambar 3 mengindikasikan bahwa model memiliki kinerja yang sangat baik dengan tingkat kesalahan yang rendah, berikut ini keseluruhan nilai evaluasi yang di tampilkan pada gambar 4 diantaranya akurasi, presisi, recall, dan f1-score.

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.98       196
         1.0       0.97      0.99      0.98       204

    accuracy                           0.98       400
   macro avg       0.98      0.98      0.98       400
weighted avg       0.98      0.98      0.98       400
```

Gambar 4. Classification Report.

Pada Gambar 4 Dari hasil di atas dapat dilihat bahwa model klasifikasi memiliki nilai accuracy, precision, recall, dan f1-score yang sangat baik. Accuracy sebesar 98% menunjukkan bahwa proporsi data yang diprediksi dengan benar oleh model adalah 98% dari total 400 data yang diuji. Precision menunjukkan sejauh mana prediksi kelas positif benar-benar sesuai. Untuk kelas 0, precision adalah 98%, sedangkan untuk kelas 1 adalah 97%. Nilai ini menandakan bahwa model sangat baik dalam meminimalkan kesalahan prediksi positif palsu (false positive). Recall mengukur seberapa baik model mengenali data positif sebenarnya. Untuk kelas 0, nilai recall adalah 97%, sementara untuk kelas 1 adalah 99%. Nilai recall tinggi pada kedua kelas ini menunjukkan kemampuan model untuk mendeteksi hampir semua data aktual di masing-masing kelas. F1-Score, yang merupakan kombinasi harmonik dari precision dan recall, bernilai 98% untuk kedua kelas. Ini menunjukkan bahwa model memiliki keseimbangan yang sangat baik antara precision dan recall. Dari hasil ini, dapat disimpulkan bahwa model bekerja dengan sangat efektif untuk kedua kelas (kelas 0 dan kelas 1) dalam dataset. Distribusi support sebesar 196 untuk kelas 0 dan 204 untuk kelas 1 menunjukkan bahwa dataset cukup seimbang, sehingga model dapat mempertahankan performa yang konsisten tanpa bias terhadap salah satu kelas.

Tabel 7. Perbandingan Penelitian

Pengarang	Metode	Ketepatan
Muhammad Salsabil, dkk (2024)	XGBoost	76%
N. Sneha, dkk (2019)	SVM	77%
	Naive Bayes	82.30%

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

Perbandingan penelitian menunjukkan hasil yang sangat unggul dibandingkan dengan penelitian sebelumnya yang menggunakan metode XGBoost, SVM, dan Naive Bayes. Dalam penelitian yang dilakukan oleh Muhammad Salsabil dkk. (2024), metode XGBoost menghasilkan akurasi sebesar 76%. Sementara itu, penelitian oleh N. Sneha dkk. (2019) menggunakan metode SVM dan Naive Bayes, dengan masing-masing akurasi sebesar 77% dan 82,3%. Sebaliknya, model yang dikembangkan berhasil mencapai akurasi yang jauh lebih tinggi, yaitu 98%.

Keunggulan ini tidak hanya terlihat dari akurasi, tetapi juga dari metrik lainnya seperti precision, recall, dan f1-score, yang semuanya menunjukkan nilai sebesar 98%. Hal ini menunjukkan bahwa model Anda memiliki performa yang sangat konsisten dalam mengklasifikasikan data pada kedua kelas, baik kelas positif maupun negatif. Dibandingkan dengan penelitian sebelumnya, model memberikan peningkatan performa yang signifikan, dengan selisih lebih dari 15% dibandingkan metode terbaik yang tercatat dalam penelitian sebelumnya, yaitu Naive Bayes. Peningkatan ini menunjukkan bahwa pendekatan tidak hanya lebih efektif tetapi juga lebih handal dalam mengatasi permasalahan klasifikasi yang serupa. Oleh karena itu, hasil penelitian ini memberikan kontribusi baru dan dapat menjadi solusi yang lebih baik untuk aplikasi serupa di masa depan.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma XGBoost dengan teknik preprocessing data yang komprehensif berhasil memberikan hasil yang sangat baik dalam klasifikasi diabetes. Model yang dikembangkan mencapai akurasi 98%, jauh lebih tinggi dibandingkan metode lain seperti Naive Bayes (82,3%) dan SVM (77%). Evaluasi menggunakan metrik seperti precision, recall, dan F1-score menunjukkan performa konsisten untuk kedua kelas, dengan semua metrik berada di angka 98%. Teknik preprocessing seperti SMOTE untuk menangani ketidakseimbangan data dan normalisasi dengan Min-Max Scaler turut berkontribusi terhadap keberhasilan ini.

Hasil ini menegaskan bahwa pendekatan yang digunakan lebih efektif dan dapat diandalkan untuk aplikasi serupa di masa depan. Penelitian lanjutan disarankan untuk mengintegrasikan lebih banyak data dari sumber berbeda dan mengeksplorasi teknik boosting lain untuk validasi lebih lanjut.

REFERENCES

A. Rahim, A. M., Inggrid Yanuar Risca Pratiwi, & Muhammad Ainul Fikri. (2023). Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier. *Indonesian Journal of Computer Science*, *12*(5). https://doi.org/10.33022/ijcs.v12i5.3413

Anwari, R. H. (2021). Dampak Konsumsi Kopi pada Penurunan Kadar Glukosa Darah Penderita Diabetes Mellitus Tipe 2. *Jurnal Penelitian Perawat Profesional*, *3*(3), 531–540. https://doi.org/10.37287/jppp.v3i3.543

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Cobre, A. de F., Stremel, D. P., Noleto, G. R., Fachi, M. M., Surek, M., Wiens, A., Tonin, F. S., & Pontarolo, R. (2021). Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *Computers in Biology and Medicine*, *134*, 104531. https://doi.org/10.1016/j.compbiomed.2021.104531

Dai, W., & Genton, M. G. (2018). Functional boxplots for multivariate curves. *Stat*, *7*(1). https://doi.org/10.1002/sta4.190

Denggos, Y. (2023). Penyakit Diabetes Mellitus Umur 40-60 Tahun di Desa Bara Batu Kecamatan Pangkep. *Healthcaring: Jurnal Ilmiah Kesehatan*, *2*(1), 55–61. https://doi.org/10.47709/healthcaring.v2i1.2177

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4

Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hinterberger, H. (2018). Exploratory Data Analysis. In *Encyclopedia of Database Systems* (pp. 1413–1414). Springer New York. https://doi.org/10.1007/978-1-4614-8265-9_1384

LOGIC : Jurnal Ilmu Komputer dan Pendidikan
Volume 3, No. 1, Desember Tahun 2024
ISSN 2985-4172 (media online)
Hal 116-125

J, A. (2019). GAMBARAN TINGKAT PENGETAHUAN PASIEN DIABETES MELITUS TIPE 2 TENTANG MANAJEMEN DIABETES MELITUS. *Media Keperawatan: Politeknik Kesehatan Makassar*, *10*(2), 19. https://doi.org/10.32382/jmk.v10i2.1334

KOWALSKI, P. A., LŁUKASIK, S., & KULCZYCKI, P. (2017). Methods of Collective Intelligence in Exploratory Data Analysis: A Research Survey. *Proceedings of the International Conference on Computer Networks and Communication Technology (CNCT 2016)*. https://doi.org/10.2991/cnct-16.2017.1

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Nazliansyah, Amiruddin, & Lubis, A. Y. S. (2022). Peningkatan Health Promotion Behavior pada Penderita Diabetes Mellitus Tipe 2 di Wilayah Kerja Puskesmas Tanjung Binga Kabupaten Belitung. *Jurnal Pengabdian Masyarakat Bestari*, *1*(7), 659–668. https://doi.org/10.55927/jpmb.v1i7.1596

Nizar, M., & Amelia, R. (2022). Hubungan Kadar Trigliserida Dengan Kadar Glukosa Pada Penderita Diabetes Melitus Tipe 2 di RS Krakatau Medika. *Journal of Medical Laboratory Research*, *1*(1), 7–12. https://doi.org/10.36743/jomlr.v1i1.432

Nunik Purnamasari, Nuzirwan Acang, & Harvi Puspa Wardani. (2024). Pengaruh Diabetes Melitus Tipe 2 Tidak Terkontrol terhadap Komplikasi Nefropati Diabetik. *Bandung Conference Series: Medical Science*, *4*(1), 495–501. https://doi.org/10.29313/bcsms.v4i1.11035

Nyayu Mevia Fiqi, & Zulmansyah. (2021). Gambaran Tingkat Pengetahuan Siswa SMA Negeri Kelas XII di Kota Bandung tentang Penyakit Diabetes Mellitus Tipe 2. *Jurnal Riset Kedokteran*, *1*(2), 66–70. https://doi.org/10.29313/jrk.v1i2.437

Putri, D. P., Prabowo, N. A., Myrtha, R., Apriningsih, H., & Hermawati, B. D. (2021). PENGELOLAAN PENYAKIT DIABETES MELLITUS TIPE 2 MELALUI PEMBERDAYAAN PENDERITA DIABETES MELLITUS DI RUMAH SAKIT UNS. *LOGISTA - Jurnal Ilmiah Pengabdian Kepada Masyarakat*, *5*(2), 224. https://doi.org/10.25077/logista.5.2.224-229.2021

Putri, L. A. R., & Ellyani Abadi. (2021). Pengaruh Pemberian Omega 3 (EPA+DHA) Terhadap Kadar Gula Darah Penderita Diabetes Melitus Tipe 2 di Kota Kendari. *Media Publikasi Promosi Kesehatan Indonesia (MPPKI)*, *5*(1), 91–94. https://doi.org/10.56338/mppki.v5i1.2106

Putri Nurhayati, Achmad Mujahidin Irham, I Dewa Bagus Ketut Widya Pramana, & Herpan Syafii Harahap. (2022). HbA1c Sebagai Kandidat Biomarker untuk Prediksi Progesivitas Gangguan Kognitif Terkait Diabetes Melitus Tipe 2. *Unram Medical Journal*, *11*(1), 732–738. https://doi.org/10.29303/jk.v11i1.4342

Rahim, A. M. A., Sunyoto, A., & Arief, M. R. (2022). Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *21*(3), 595–606. https://doi.org/10.30812/matrik.v21i3.1666

Ramadona, A., Rustam, E., & Syauqie, M. (2021). Hubungan Kepatuhan Minum Obat dengan Munculnya Gejala Neuropati Pada Pasien Diabetes Melitus Tipe 2 Di Puskesmas Andalas. *Jurnal Farmasi Higea*, *13*(1), 14. https://doi.org/10.52689/higea.v13i1.326

Salasa, R. A., Rahman, H., & Andiani, A. (2019). Faktor Risiko Diabetes Mellitus Tipe 2 Pada Populasi Asia: A systematic Review. *JURNAL BIOSAINSTEK*, *1*(01), 95–107. https://doi.org/10.52046/biosainstek.v1i01.306

Sumah, D. F. (2018). *HUBUNGAN AKTIVITAS FISIK DAN KUALITAS TIDUR DENGAN KADAR GULA DARAH SEWAKTU PADA PASIEN DIABETES MELITUS TIPE 2 DI POLIKLINIK PENYAKIT DALAM RSUD dr. M. HAULUSSY AMBON*. https://doi.org/10.31227/osf.io/9jxg8

Wahyuni, K. I., Prayitno, A. A., & Wibowo, Y. I. (2019a). Efektivitas Edukasi Pasien Diabetes Mellitus Tipe 2 Terhadap Pengetahuan dan Kontrol Glikemik Rawat Jalan di RS Anwar Medika. *Jurnal Pharmascience*, *6*(1), 1. https://doi.org/10.20527/jps.v6i1.6069

Wahyuni, K. I., Prayitno, A. A., & Wibowo, Y. I. (2019b). Efektivitas Edukasi Pasien Diabetes Mellitus Tipe 2 Terhadap Pengetahuan dan Kontrol Glikemik Rawat Jalan di RS Anwar Medika. *Jurnal Pharmascience*, *6*(1), 1. https://doi.org/10.20527/jps.v6i1.6069

Widyana, A. R., & Afriansyah, M. A. (2022). Penyuluhan dan Pemeriksaan Kadar HbA1c pada Pasien Diabetes Mellitus Tipe 2 di RSUD Suradadi. *JURNAL INOVASI DAN PENGABDIAN MASYARAKAT INDONESIA*, *1*(3), 6–9. https://doi.org/10.26714/jipmi.v1i3.23