

Komparasi Kinerja Algoritma XGBoost dengan Reduksi Dimensi PCA pada Klasifikasi Diabetes

Rivale Belano Nusa¹, Rina Dewi Indahsari^{2*}

^{1,2} Teknik Informatika, Fakultas Teknologi dan Desain, Institut Teknologi dan Bisnis Asia Malang, Indonesia

Info Artikel

Riwayat Artikel:

Diterima : **26 November 2025**

Direvisi : **19 Desember 2025**

Diterbitkan : **31 Desember 2025**

Kata Kunci:

PCA,
XGBoost,
Diabetes,
Machine Learning,
Klasifikasi

Keywords:

PCA,
XGBoost,
Diabetes,
Machine Learning,
Classification

ABSTRAK

Diabetes merupakan salah satu penyakit kronis dengan prevalensi tinggi di dunia dan memerlukan deteksi dini yang akurat untuk mencegah komplikasi jangka panjang. Dalam bidang analisis data medis, penerapan algoritma machine learning seperti XGBoost telah terbukti efektif dalam melakukan klasifikasi risiko penyakit. Penelitian ini bertujuan untuk membandingkan kinerja algoritma XGBoost sebelum dan sesudah penerapan Principal Component Analysis (PCA) dalam klasifikasi risiko diabetes menggunakan Early Stage Diabetes Risk Prediction Dataset. Tahapan penelitian meliputi pra-pemrosesan data pemeriksaan nilai hilang, label encoding, pembersihan outlier, normalisasi, serta penerapan PCA dengan retensi variansi sebesar 90%. Hasil eksperimen menunjukkan bahwa model XGBoost tanpa PCA memperoleh akurasi tertinggi sebesar 99,04%, sedangkan model dengan PCA menghasilkan akurasi 98,08%. Meskipun penerapan PCA sedikit menurunkan akurasi, teknik ini berhasil mengurangi jumlah fitur dari dan meningkatkan efisiensi komputasi tanpa kehilangan informasi penting. Dengan demikian, PCA terbukti efektif dalam menyederhanakan kompleksitas data dan mempertahankan kinerja model secara optimal.

ABSTRACT

Diabetes is one of the most prevalent chronic diseases worldwide and requires accurate early detection to prevent long-term complications. In the field of medical data analysis, the application of machine learning algorithms such as XGBoost has proven effective in classifying disease risk. This study aims to compare the performance of the XGBoost algorithm before and after applying Principal Component Analysis (PCA) in diabetes risk classification using the Early Stage Diabetes Risk Prediction Dataset. The research stages include data preprocessing involving missing value checking, label encoding, outlier removal, normalization, and followed by the application of PCA with a 90% variance retention threshold. The experimental results show that the XGBoost model without PCA achieved the highest accuracy of 99.04%, while the model with PCA achieved 98.08%. Although the application of PCA slightly reduced accuracy, this technique successfully decreased the number of features and improved computational efficiency without losing important information. Therefore, PCA is proven to be effective in simplifying data complexity while maintaining optimal model performance.

Penulis Korespondensi:

Rina Dewi Indahsari,
Program Studi Teknik Informatika,
Institut Teknologi dan Bisnis Asia Malang
Email: rinadewi@asia.ac.id

This is an open access article under the [CC BY-SA](#) license



1. PENDAHULUAN

Diabetes merupakan salah satu penyakit kronis dengan prevalensi tinggi di dunia dan memerlukan deteksi dini yang akurat untuk mencegah komplikasi jangka Panjang. Prevalensi diabetes mellitus terus meningkat dan

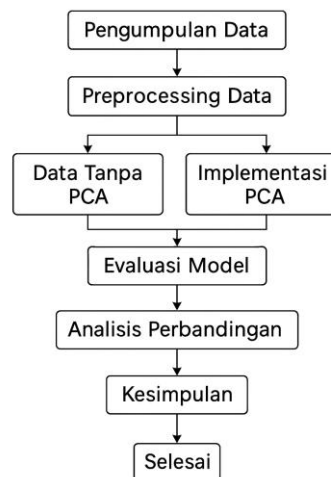
menjadi salah satu beban utama kesehatan global. Laporan terbaru IDF Diabetes Atlas edisi ke-11 memperkirakan sekitar 589 juta orang dewasa di dunia hidup dengan diabetes pada tahun 2024, dan jumlah ini diproyeksikan naik menjadi 853 juta pada 2050 [1]. Organisasi Kesehatan Dunia menegaskan bahwa peningkatan tersebut terjadi terutama di negara berpendapatan rendah dan menengah, dengan banyak penderita yang belum terdiagnosis maupun belum mendapat terapi yang memadai [2]. Kondisi ini menekankan pentingnya upaya deteksi dini dan prediksi risiko diabetes yang lebih akurat dan efisien untuk mencegah komplikasi jangka panjang.

Dalam beberapa tahun terakhir, pemanfaatan *machine learning* untuk prediksi diabetes berkembang pesat, termasuk penggunaan algoritma *ensemble* seperti *Extreme Gradient Boosting* (XGBoost). Berbagai penelitian melaporkan bahwa XGBoost mampu memberikan kinerja klasifikasi yang tinggi dan stabil, terutama ketika dikombinasikan dengan optimasi parameter atau *feature selection* berbasis algoritma evolusioner [3], [4]. Sejumlah penelitian mutakhir menunjukkan bahwa berbagai algoritma *machine learning* seperti logistic regression, decision tree, Naïve Bayes hingga SVM mampu memberikan performa yang kompetitif untuk deteksi dini diabetes tipe 2, dengan akurasi berada pada kisaran 77–80% pada dataset Pima Indians Diabetes [5]. Di sisi lain, dataset *Early Stage Diabetes Risk Prediction* dari UCI Machine Learning Repository berisi 520 baris dari data pasien dengan atribut gejala klinis dan label kelas risiko secara luas digunakan sebagai *benchmark* untuk pengembangan model prediksi diabetes tahap awal [6], [7]. Hal ini menjadikan dataset tersebut relevan untuk menguji berbagai rancangan model, termasuk skenario perbandingan kinerja XGBoost dengan dan tanpa teknik pengolahan fitur tertentu.

Selain pemilihan algoritma, kualitas dan dimensi fitur sangat berpengaruh terhadap performa model klasifikasi. Sejumlah studi terbaru menunjukkan bahwa *Principal Component Analysis* (PCA) dapat dimanfaatkan untuk reduksi dimensi dan penyederhanaan fitur pada data medis, sehingga meningkatkan efisiensi komputasi dan sering kali mempertahankan atau bahkan memperbaiki akurasi model [8]. Pada konteks prediksi diabetes, optimasi fitur juga terbukti berpengaruh signifikan terhadap kinerja model. Rahman dkk. (2025) menunjukkan bahwa kombinasi beberapa algoritma *feature selection* dengan *boosting classifier* seperti XGBoost dan LightGBM mampu mencapai akurasi hingga sekitar 85% sekaligus menurunkan jumlah fitur dan waktu pelatihan secara signifikan [9]. Kombinasi PCA dan XGBoost telah dikaji dalam konteks prediksi prediabetes dan deteksi dini diabetes, dengan hasil yang menunjukkan peningkatan kinerja dan penyederhanaan struktur fitur dibandingkan model tanpa reduksi dimensi [10], [11]. Di luar domain diabetes, integrasi PCA dengan XGBoost juga telah dimanfaatkan pada deteksi penyakit kronis lainnya. Manjula dkk. (2025) menggabungkan PCA untuk reduksi dimensi dengan XGBoost yang dioptimasi menggunakan Improved Brainstorm Optimization pada deteksi penyakit ginjal kronis, dan melaporkan akurasi hingga sekitar 97,5% sekaligus peningkatan efisiensi komputasi [12]. Berdasarkan hal tersebut, tujuan penelitian ini adalah membandingkan kinerja algoritma XGBoost sebelum dan sesudah penerapan PCA pada dataset Early Stage Diabetes Risk Prediction, ditinjau dari metrik akurasi, *precision*, *recall*, dan *F1-score*. Kontribusi yang diharapkan adalah memberikan bukti empiris mengenai efektivitas PCA sebagai teknik reduksi fitur pada model XGBoost untuk klasifikasi diabetes tahap awal, sekaligus menyajikan rekomendasi praktis bagi pengembangan sistem prediksi diabetes berbasis *machine learning* yang akurat dan efisien.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional. Tujuan utama penelitian adalah membandingkan performa model XGBoost sebelum dan sesudah penerapan *Principal Component Analysis* (PCA) dalam klasifikasi pada dataset Early Stage Diabetes Risk Prediction. Pendekatan ini mengikuti praktik umum penelitian komputasi kesehatan yang menguji efek teknik reduksi dimensi terhadap performa model *machine learning* [8], [10]. Untuk memberikan gambaran yang lebih jelas mengenai prosedur penelitian yang dilakukan, berikut disajikan diagram alur penelitian. Diagram ini menunjukkan tahapan utama mulai dari proses pengumpulan data, pra-pemrosesan, pemodelan dengan dan tanpa PCA, hingga tahap evaluasi serta analisis perbandingan untuk memperoleh kesimpulan akhir.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Dataset yang diuji adalah *Early Stage Diabetes Risk Prediction Dataset* dari UCI Machine Learning Repository, yang berisi 520 data dengan 16 fitur gejala klinis dan 1 label kelas [6]. Dataset ini telah banyak digunakan dalam penelitian deteksi dini diabetes karena strukturnya yang ringkas dan relevan [7]. Data terdiri atas atribut kategorikal biner yang merepresentasikan gejala umum diabetes seperti pada gambar 2.

Data columns (total 17 columns):			
#	Column	Non-Null Count	Dtype
0	Age	520 non-null	int64
1	Gender	520 non-null	object
2	Polyuria	520 non-null	object
3	Polydipsia	520 non-null	object
4	sudden weight loss	520 non-null	object
5	weakness	520 non-null	object
6	Polyphagia	520 non-null	object
7	Genital thrush	520 non-null	object
8	visual blurring	520 non-null	object
9	Itching	520 non-null	object
10	Irritability	520 non-null	object
11	delayed healing	520 non-null	object
12	partial paresis	520 non-null	object
13	muscle stiffness	520 non-null	object
14	Alopecia	520 non-null	object
15	Obesity	520 non-null	object
16	class	520 non-null	object
dtypes: int64(1), object(16)			

Gambar 2. Struktur Dataset

2.2 Pra-pemrosesan Data

Tahap pra-pemrosesan data diimplementasikan untuk memastikan bahwa dataset berada dalam kondisi siap digunakan oleh algoritma pembelajaran mesin dan mampu menghasilkan performa model yang optimal. Urutan Pra-pemrosesan data:

1. Memeriksa Nilai Hilang (Missing Values)

Proses dimulai dengan pemeriksaan nilai hilang (missing values) pada setiap variabel untuk memastikan tidak terdapat data kosong atau tidak valid yang berpotensi memengaruhi distribusi data maupun hasil pelatihan model. Setelah memastikan kualitas data, dilakukan pengecekan konsistensi label guna memastikan bahwa setiap kategori pada variabel biner seperti *Yes/No* atau *Male/Female* tidak memiliki variasi penulisan yang berbeda, misalnya penggunaan huruf kecil atau besar yang tidak seragam.

2. Label Encoding

Selanjutnya, seluruh variabel kategorikal diubah menjadi bentuk numerik melalui label encoding, yang mengonversi kategori menjadi nilai bilangan bulat. Proses ini penting karena sebagian besar algoritma pembelajaran mesin, termasuk PCA dan XGBoost, bekerja pada input numerik. Secara matematis, label encoding dapat direpresentasikan sebagai suatu fungsi pemetaan:

$$f : C \rightarrow \{0, 1\} \quad (1)$$

di mana C adalah himpunan kategori asli (*Yes, No, Male, Female*) dan outputnya berupa angka 0 atau 1 sesuai kategori yang direpresentasikan.

3. Pembersihan Outlier

Selanjutnya dilakukan tahap pembersihan outlier menggunakan metode Z-Score, yang bertujuan mengidentifikasi nilai fitur yang berada jauh dari distribusi normalnya. Outlier dapat menyebabkan distorsi pada proses pembelajaran model dan memengaruhi kestabilan PCA karena PCA sensitif terhadap deviasi ekstrem. Z-Score dihitung menggunakan rumus berikut:

$$Z = \frac{X - \mu}{\sigma} \quad (2)$$

dimana X merupakan nilai data, μ merupakan nilai rata-rata fitur, dan σ standar deviasi fitur. Nilai dianggap outlier jika $|Z| > 3$. Observasi yang memenuhi kondisi tersebut akan dibuang untuk menjaga kualitas distribusi data dan mengurangi noise yang tidak relevan. Pendekatan ini lazim digunakan pada dataset medis untuk meningkatkan stabilitas proses reduksi dimensi [13].

4. Normalisasi Data (Feature Scaling)

Tahap berikutnya adalah proses normalisasi menggunakan StandardScaler, yang bertujuan menyeragamkan skala antar fitur agar PCA tidak bias terhadap variabel dengan skala lebih besar. Normalisasi dilakukan menggunakan pendekatan *standardization*, yaitu mentransformasikan setiap fitur agar memiliki nilai mean = 0 dan standard deviation = 1. Secara matematis, proses standardisasi menggunakan rumus:

$$X - scaled = \frac{X - \mu}{\sigma} \quad (3)$$

dengan X adalah nilai asli, μ adalah nilai rata-rata fitur, dan σ adalah standar deviasi. Proses ini direkomendasikan dalam literatur PCA sebagai langkah krusial untuk mempertahankan stabilitas dan validitas komponen utama [13].

5. Pembagian Data (Split Data)

Setelah data dinormalisasi, dataset dibagi menjadi data latih (train set) dan data uji (test set) menggunakan rasio 80:20. Pembagian dilakukan dengan teknik stratified splitting, yang bertujuan menjaga proporsi kelas *Positive* dan *Negative* pada kedua subset tetap seimbang. Strategi ini penting untuk menghindari bias model akibat distribusi kelas yang tidak merata, sehingga evaluasi performa menjadi lebih representatif terhadap keseluruhan dataset.

Dengan demikian, seluruh proses pra-pemrosesan ini memastikan bahwa data yang digunakan siap dan sesuai untuk penerapan PCA dan pelatihan model XGBoost pada tahap berikutnya.

2.3 Implementasi PCA (Principal Component Analysis)

Penerapan Principal Component Analysis (PCA) dilakukan sebagai teknik reduksi dimensi untuk menyederhanakan struktur fitur pada dataset tanpa menghilangkan informasi penting yang berkaitan dengan variasi data. PCA bekerja dengan mentransformasikan sekumpulan fitur asli menjadi sekumpulan komponen utama (principal components) yang saling ortogonal dan mampu menjelaskan proporsi varians terbesar dalam data [14]. Pendekatan ini sangat efektif ketika dataset memiliki fitur yang saling berkorelasi, karena PCA dapat menghilangkan redundansi informasi di antara fitur-fitur tersebut.

Dalam penelitian ini, jumlah komponen ditentukan hingga mencapai nilai kumulatif $\geq 90\%$, sehingga mayoritas informasi dalam data tetap dipertahankan sesuai rekomendasi penelitian terdahulu [8]. Pemilihan

ambang 90% dianggap optimal karena memberikan keseimbangan antara penyederhanaan fitur dan retensi informasi data.

Setelah transformasi PCA selesai, dataset yang telah direduksi dimensi tersebut digunakan sebagai input untuk membangun model XGBoost, pada skenario pembandingan dengan model baseline (tanpa PCA). Dengan demikian, penelitian dapat mengevaluasi secara empiris sejauh mana PCA mempengaruhi perubahan performa model, baik dari aspek akurasi maupun efisiensi komputasi.

2.4 Membangun Model XGBoost

Model XGBoost dibangun menggunakan parameter dasar yang umum digunakan dalam penelitian, meliputi learning rate, n_estimators, dan max_depth, sebagaimana direkomendasikan pada literatur [15]. Pada penelitian ini, dua skenario pemodelan dikembangkan, yaitu:

1. model XGBoost tanpa reduksi dimensi sebagai baseline.
2. model XGBoost dengan fitur hasil transformasi PCA.

Kedua skenario menggunakan subset data yang sama agar perbandingan kinerja dapat dilakukan secara objektif.

Untuk memperoleh konfigurasi parameter terbaik, dilakukan proses hyperparameter tuning menggunakan metode Grid Search yang menguji beberapa kombinasi parameter inti. Pendekatan ini merujuk pada praktik penelitian modern dalam prediksi diabetes yang menunjukkan bahwa Grid Search mampu meningkatkan performa model boosting secara signifikan [3], [4]. Model terbaik dari masing-masing skenario kemudian dievaluasi dengan metrik akurasi, precision, recall, dan F1-score.

2.5 Evaluasi Model

Evaluasi performa model dilakukan menggunakan *confusion matrix* dan empat metrik utama, yaitu akurasi, precision, recall, dan F1-score, yang merupakan standar dalam penelitian klasifikasi biner karena mampu menggambarkan kemampuan model secara lebih komprehensif [10]. Keempat metrik tersebut dihitung berdasarkan nilai *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)* yang dihasilkan model.

1. Akurasi (Accuracy)

mengukur akurasi prediksi yang benar dari keseluruhan data uji:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

2. Precision

mengukur ketepatan dalam memprediksi kelas, yaitu seberapa besar prediksi positif yang benar:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

3. Recall (Sensitivity)

mengukur kemampuan model mendeteksi seluruh kasus positif dengan benar:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

4. F1-Score

mengukur rata-rata dari precision dan recall, digunakan ketika diperlukan keseimbangan antara kedua metrik tersebut:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Analisis perbandingan dilakukan dengan menilai perbedaan performa pada dua skenario: XGBoost tanpa PCA dan XGBoost dengan PCA, sehingga pengaruh reduksi dimensi terhadap akurasi maupun efisiensi dapat dinilai secara objektif.

2.6 Analisis Perbandingan

Tahap terakhir adalah melakukan analisis perbandingan terhadap performa model pada kedua skenario, yaitu XGBoost tanpa PCA dan XGBoost dengan PCA. Perbandingan dilakukan berdasarkan Confusion matrix dan hasil evaluasi (akurasi, precision, recall, dan F1-score). Hasil dari evaluasi dianalisis untuk menentukan apakah penerapan PCA memberikan peningkatan performa prediktif atau efisiensi komputasi yang signifikan.

Proses analisis dilakukan dengan menginterpretasikan selisih nilai metrik serta perubahan kompleksitas model setelah reduksi dimensi. Pendekatan ini sejalan dengan praktik penelitian komputasi kesehatan yang menilai dampak teknik reduksi fitur terhadap model klasifikasi, khususnya pada data medis dengan banyak atribut gejala [11], [13]. Dengan demikian, analisis perbandingan ini memberikan dasar empiris untuk menentukan efektivitas PCA dalam meningkatkan kualitas model XGBoost pada tugas klasifikasi diabetes tahap awal.

3. HASIL DAN ANALISIS

3.1 Memeriksa Nilai Hilang (Missing Values)

Pemeriksaan terhadap nilai hilang (*missing values*) dilakukan untuk memastikan integritas dataset sebelum tahap analisis lebih lanjut. Berdasarkan hasil pengecekan, seluruh fitur pada dataset *Early Stage Diabetes Risk Prediction* menunjukkan nilai 0 pada jumlah data yang hilang, sebagaimana ditunjukkan pada gambar di bawah ini.

Jumlah Missing Value tiap kolom:	
Age	0
Gender	0
Polyuria	0
Polydipsia	0
sudden weight loss	0
weakness	0
Polyphagia	0
Genital thrush	0
visual blurring	0
Itching	0
Irritability	0
delayed healing	0
partial paresis	0
muscle stiffness	0
Alopecia	0
Obesity	0
class	0
dtype: int64	

Gambar 3. Jumlah Missing Value

Dengan demikian, dataset ini dinyatakan bersih dari nilai kosong atau tidak valid, sehingga tidak diperlukan proses imputasi nilai.

3.2 Label Encoding

Berdasarkan hasil keluaran struktur data setelah dilakukan encoding (Tabel di bawah), setiap kolom telah bertipe data int64 dengan total 17 variabel dan 520 entri data tanpa nilai hilang. Perubahan tipe data ini menunjukkan bahwa proses konversi telah berhasil dilakukan dengan baik dan dataset kini sepenuhnya berada dalam format numerik sebagaimana ditunjukkan pada gambar 4.

Data columns (total 17 columns):

#	Column	Non-Null Count	Dtype
0	Age	520 non-null	int64
1	Gender	520 non-null	int64
2	Polyuria	520 non-null	int64
3	Polydipsia	520 non-null	int64
4	sudden weight loss	520 non-null	int64
5	weakness	520 non-null	int64
6	Polyphagia	520 non-null	int64
7	Genital thrush	520 non-null	int64
8	visual blurring	520 non-null	int64
9	Itching	520 non-null	int64
10	Irritability	520 non-null	int64
11	delayed healing	520 non-null	int64
12	partial paresis	520 non-null	int64
13	muscle stiffness	520 non-null	int64
14	Alopecia	520 non-null	int64
15	Obesity	520 non-null	int64
16	class	520 non-null	int64

dtypes: int64(17)

Gambar 4. Stuktur Dataset Setelah Encoding

Dengan demikian, data sudah siap digunakan untuk tahap pembersihan outlier dan normalisasi pada langkah selanjutnya.

3.3 Pembersihan Outlier

Berdasarkan hasil perhitungan, jumlah data sebelum pembersihan *outlier* adalah 520 baris, dan setelah proses pembersihan, jumlah data tersisa 516 baris. Artinya, terdapat 4 data yang terdeteksi sebagai *outlier* dan dihapus dari dataset sebagaimana ditampilkan pada tabel 1.

Tabel 1. Hasil Pembersihan Outlier

Jumlah Baris Sebelum Pembersihan Outlier	520
Jumlah Baris Setelah Pembersihan Outlier	516

Hasil ini menunjukkan bahwa proporsi data yang tereliminasi relatif kecil, sehingga proses pembersihan *outlier* tidak memengaruhi representativitas data secara keseluruhan. Dataset hasil pembersihan kemudian digunakan pada tahap normalisasi dan penerapan PCA untuk memastikan kestabilan komputasi serta mengurangi distorsi variansi akibat nilai ekstrem.

3.4 Implementasi PCA

Jumlah fitur pada data latih dan data uji setelah transformasi PCA masing-masing menjadi 12 fitur, sebagaimana ditunjukkan pada hasil eksekusi program. Hal ini menunjukkan bahwa proses reduksi dimensi berhasil menurunkan kompleksitas data tanpa kehilangan informasi signifikan, sesuai dengan tujuan PCA untuk meningkatkan efisiensi pemodelan.

Dengan demikian, PCA mampu mempertahankan mayoritas variasi informasi dalam dataset sekaligus mengurangi kemungkinan *redundansi antar fitur*. Dataset hasil reduksi ini kemudian digunakan sebagai input untuk pelatihan model XGBoost pada skenario kedua, guna dibandingkan performanya dengan model baseline tanpa PCA.

3.5 Pemodelan XGBoost

Proses penalaan parameter dilakukan menggunakan GridSearchCV dengan skema *5-fold cross-validation* untuk mencari kombinasi parameter yang memberikan performa terbaik. Hasil tuning menunjukkan bahwa setiap model menghasilkan konfigurasi parameter optimal yang berbeda, sebagaimana ditampilkan pada Tabel 2.

Tabel 2. Parameter Model Terbaik

Model	colsample_bytree	learning_rate	max_depth	n_estimators	subsample
XGBoost Tanpa PCA	0.5	0.05	5	200	0.9
XGBoost PCA	0.7	0.1	4	300	0.3

Hasil tersebut menunjukkan bahwa pada model tanpa PCA, nilai *learning_rate* yang optimal relatif lebih kecil (0.05), dengan kedalaman pohon (*max_depth*) lebih besar (5) dibandingkan model dengan PCA. Hal ini mengindikasikan bahwa model tanpa reduksi dimensi memerlukan kompleksitas lebih tinggi untuk mencapai performa optimal. Sebaliknya, pada model dengan PCA, nilai *learning_rate* yang lebih besar (0.1) dan *max_depth* yang lebih dangkal (4) menunjukkan bahwa PCA berhasil menyederhanakan representasi data, sehingga model dapat belajar lebih efisien dengan struktur yang lebih ringan.

3.6 Perbandingan Kinerja Model

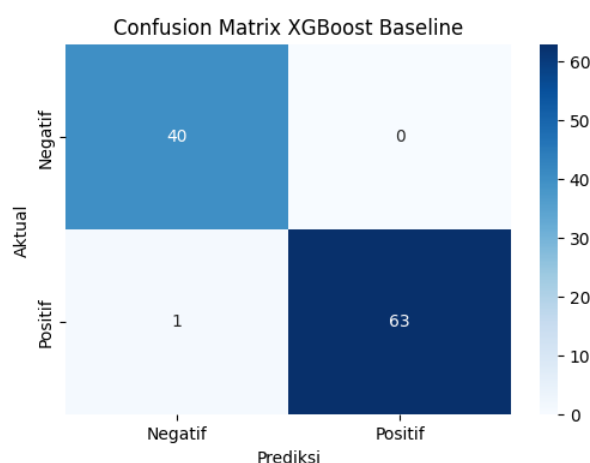
Hasil pengujian menunjukkan bahwa model XGBoost tanpa PCA memperoleh akurasi sebesar 99.04%, sedangkan model XGBoost dengan PCA mencapai akurasi 98.08%. Perbedaan ini menunjukkan bahwa penerapan PCA sedikit menurunkan akurasi model, meskipun secara umum kedua model sama-sama memiliki kinerja yang sangat baik. Ringkasan hasil *classification report* untuk kedua model sebagaimana ditampilkan pada Tabel 6.

Tabel 3. Perbandingan Kinerja Model

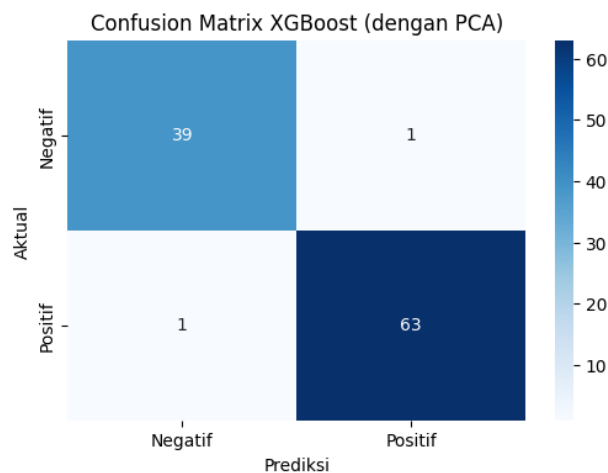
Model	Akurasi	Precision	Recall	F1-Score
XGBoost Tanpa PCA	99.04%	0.99	0.99	0.99
XGBoost PCA	98.08%	0.98	0.98	0.98

Pada model baseline menunjukkan bahwa model mampu mengidentifikasi kasus diabetes dengan tingkat kesalahan yang sangat rendah. Sedangkan pada model dengan PCA, precision dan recall sama-sama bernilai 0.98, menandakan bahwa proses reduksi dimensi sedikit memengaruhi sensitivitas model terhadap data positif.

Visualisasi *confusion matrix* pada Gambar berikut memperkuat hasil tersebut. Model XGBoost tanpa PCA mampu mengklasifikasikan semua data negatif dengan benar (40 dari 40) dan hanya salah satu pada kelas positif (63 dari 64). Sebaliknya, model dengan PCA menghasilkan dua kesalahan klasifikasi, masing-masing satu pada kelas negatif dan satu pada kelas positif. Meskipun perbedaan ini kecil, hasil tersebut mengindikasikan bahwa PCA sedikit mempengaruhi sensitivitas model terhadap variasi data.



Gambar 5. Visualisasi Confusion Matrix XGBoost Baseline



Gambar 6. Visualisasi Confusion Matrix XGBoost dengan PCA

Secara keseluruhan, kedua model menunjukkan performa tinggi dengan nilai F1-score di atas 0.98. Namun, model tanpa PCA memiliki sedikit keunggulan dalam akurasi dan ketepatan klasifikasi, sedangkan model dengan PCA unggul dari segi efisiensi komputasi karena jumlah fitur berkurang dari 16 menjadi 12. Dengan demikian, dapat disimpulkan bahwa penerapan PCA tidak secara signifikan meningkatkan performa akurasi model, tetapi berhasil mengoptimalkan efisiensi dan menyederhanakan kompleksitas model tanpa kehilangan banyak informasi penting.

4. KESIMPULAN

Berdasarkan hasil dan analisis, penelitian ini berhasil mencapai tujuan yang telah dinyatakan pada bagian Pendahuluan, yaitu membandingkan kinerja metode XGBoost sebelum dan sesudah penerapan Principal Component Analysis (PCA) dalam klasifikasi risiko diabetes tahap awal. Hasil penelitian menunjukkan bahwa model XGBoost tanpa PCA memperoleh akurasi tertinggi sebesar 99.04%, sedangkan model XGBoost dengan PCA mencapai akurasi 98.08%. Perbedaan yang relatif kecil ini membuktikan bahwa PCA mampu menyederhanakan kompleksitas data tanpa mengurangi performa model secara signifikan. Meskipun penerapan PCA sedikit menurunkan akurasi model, namun reduksi dimensi berhasil menurunkan kompleksitas fitur dari tanpa kehilangan informasi penting secara signifikan. Hal ini menandakan bahwa PCA efektif dalam meningkatkan efisiensi komputasi sekaligus mempertahankan performa model yang stabil.

Adapun prospek pengembangan selanjutnya dapat difokuskan pada penerapan teknik feature selection atau dimensionality reduction lain seperti Linear Discriminant Analysis (LDA) atau t-SNE, serta pengujian model pada dataset yang lebih besar dan beragam. Selain itu, hasil penelitian ini juga dapat diterapkan dalam implementasi teknik pembelajaran mesin untuk sistem pendukung keputusan di bidang kesehatan untuk deteksi dini penyakit kronis, sehingga berpotensi memberikan kontribusi nyata dalam bidang kesehatan preventif.

REFERENSI

- [1] I. D. Federation, "IDF Diabetes Atlas, 11th edition Global factsheet," International Diabetes Federation, 2025. [Daring]. Tersedia pada: <http://httpsdiabetesatlas.org>
- [2] W. H. Organization, "Urgent action needed as global diabetes cases increase four-fold over past decades," 2024, *World Health Organization*. [Daring]. Tersedia pada: <https://www.who.int/news/item/13-11-2024-urgent-action-needed-as-global-diabetes-cases-increase-four-fold-over-past-decades>
- [3] W. Li, Y. Peng, dan K. Peng, "Diabetes prediction model based on GA-XGBoost and stacking ensemble algorithm," *PLoS One*, vol. 19, no. 9, hlm. e0311222, Sep 2024, [Daring]. Tersedia pada: <https://doi.org/10.1371/journal.pone.0311222>
- [4] D. N. Jawza, M. I. Mazdadi, A. Farmadi, T. H. Saragih, D. Kartini, dan V. Abdullayev, "The Enhancing Diabetes Prediction Accuracy Using Random Forest and XGBoost with PSO and GA-Based Feature

- Selection," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2 SE-Electronics, Feb 2025, doi: 10.35882/jeeemi.v7i2.626.
- [5] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, dan M. Cabanillas-Carbonell, "Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes," 2023. doi: 10.3390/diagnostics13142383.
- [6] U. C. I. M. L. Repository, "Early Stage Diabetes Risk Prediction Dataset," 2020, *University of California, Irvine*. [Daring]. Tersedia pada: <https://archive.ics.uci.edu/dataset/529/early%2Bstage%2Bdiabetes%2Brisk%2Bprediction%2Bdataset>
- [7] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, dan W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study," 2022. doi: 10.3390/s22145247.
- [8] M. Tantowen, K. Putra, M. Isnain, dan B. Pardamean, "Principal Component Analysis Implementation on Machine Learning in Diabetes Classification," *Communications in Mathematical Biology and Neuroscience*, vol. 2024, hlm. 1–19, 2024, doi: 10.28919/cmbn/8492.
- [9] F. Rahman, S. Hossain, J.-J. Tiang, dan A.-A. Nahid, "Diabetes Prediction Using Feature Selection Algorithms and Boosting-Based Machine Learning Classifiers," 2025. doi: 10.3390/diagnostics15202622.
- [10] "Principal Component Analysis for Prediabetes Prediction using Extreme Gradient Boosting (XGBoost)," *Scientific Journal of Informatics*, vol. 11, no. 3 SE-Articles, hlm. 863–872, doi: 10.15294/sji.v11i3.13416.
- [11] R. Abdurrosyid, A. Teguh, dan W. Almais, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Deteksi Dini Diabetes menggunakan Machine Learning dengan Metode PCA dan XGBoost," vol. 11, no. 1, hlm. 51–56, 2025.
- [12] S. Manjula, N. H. Rajini, dan K. Chokkanathan, "Enhanced chronic kidney disease detection using XGBoost with improved brainstorm optimization for hyperparameter tuning," *Discover Applied Sciences*, vol. 7, no. 10, hlm. 1181, 2025, doi: 10.1007/s42452-025-07633-7.
- [13] J. Gupta, N. Sharma, dan S. Aggarwal, "Impact of Principal Component Analysis on the Performance of Machine Learning Models for the Prediction of Length of Stay of Patients," *EMITTER International Journal of Engineering Technology*, vol. 12, no. 2 SE-Articles, Des 2024, doi: 10.24003/emitter.v12i2.835.
- [14] I. T. Jolliffe dan J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, hlm. 20150202, 2016, doi: 10.1098/rsta.2015.0202.
- [15] T. Chen dan C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Mar 2016, doi: 10.48550/arXiv.1603.02754.