

Prediksi Diabetes Berdasarkan Hipertensi dan Obesitas Menggunakan Algoritma *Logistic Regression* di Puskesmas Setiabudi

Intan Amelia Septiyani ^{1*}, Fuad Nur Hasan ², Lestari Yusuf ³

^{1*,2,3} Program Studi Informatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta, Indonesia.

Email: intanameliaaa22@gmail.com ^{1*}, fuad.fnu@bsi.ac.id ², lestari.lyf@bsi.ac.id ³

Histori Artikel:

Dikirim 31 Agustus 2026; Diterima dalam bentuk revisi 9 September 2026; Diterima 20 September 2026; Diterbitkan 30 September 2026. Semua hak dilindungi oleh Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMIK Indonesia Banda Aceh.

Abstrak

Penelitian ini bertujuan menerapkan algoritma Logistic Regression untuk memprediksi status diabetes pada pasien Puskesmas Setiabudi serta menganalisis hubungan variabel prediktor dengan probabilitas prediksi diabetes. Data sekunder yang digunakan mencakup 2.013 pasien, terdiri atas 1.962 pasien berstatus Normal dan 51 pasien berstatus Diabetes, dengan rasio kelas sekitar 38:1. Model menggunakan usia, jenis kelamin, tekanan darah, indeks massa tubuh (IMT), lingkar perut, dan Gula Darah Sewaktu sebagai prediktor, sedangkan status Gula Darah digunakan sebagai target. Evaluasi dilakukan melalui validasi silang 10-fold dengan prediksi *out-of-fold* (OOF), yaitu setiap pasien diprediksi ketika datanya berada pada fold validasi. Model memperoleh *accuracy* sebesar 99,95%, *precision* 98,08%, *recall* 100,00%, dan *F1-score* 99,03%. Hasil tersebut perlu ditafsirkan dengan mempertimbangkan ketidakseimbangan kelas. Selain itu, evaluasi OOF merupakan validasi internal pada dataset penelitian, bukan validasi menggunakan data uji independen atau eksternal. Oleh karena itu, hasilnya belum menunjukkan kemampuan generalisasi model pada populasi atau dataset lain.

Kata Kunci: Diabetes; Hipertensi; Obesitas; Regresi Logistik; Validasi Silang 10-Fold.

Abstract

This study aimed to apply logistic regression to predict diabetes status among patients at Setiabudi Public Health Center and to analyze the relationship between predictor variables and the predicted probability of diabetes. The secondary dataset comprised 2,013 patients: 1,962 with a Normal status and 51 with a Diabetes status, yielding a class ratio of approximately 38:1. The model used age, gender, blood pressure, body mass index (BMI), waist circumference, and random blood glucose as predictors, with blood glucose status as the target. Model performance was evaluated using 10-fold cross-validation with out-of-fold (OOF) predictions, in which each patient received a prediction while their record was in a validation fold. The model achieved an accuracy of 99.95%, precision of 98.08%, recall of 100.00%, and an F1-score of 99.03%. These results should be interpreted in light of the class imbalance. Moreover, OOF evaluation provides internal validation on the study dataset rather than validation on independent or external test data. The results therefore do not establish how well the model would generalize to other populations or datasets.

Keyword: Diabetes; Hypertension; Obesity; Logistic Regression; 10-Fold Cross-Validation.

1. Pendahuluan

Diabetes merupakan salah satu penyakit metabolik yang menjadi permasalahan kesehatan karena ditandai dengan tingginya kadar glukosa darah, yang dapat menimbulkan berbagai komplikasi apabila tidak ditangani dengan baik. Diabetes bukan penyakit menular, tetapi tetap berbahaya bagi kesehatan. Selain itu, jumlah penderitanya terus meningkat di berbagai wilayah (Mumuh *et al.*, 2026). Permasalahan utama bukan hanya tingginya jumlah penderita, melainkan juga besarnya proporsi kasus yang tidak terdiagnosis sejak dini (Astofa *et al.*, 2025). Peningkatan jumlah penderita diabetes berimplikasi pada bertambahnya beban biaya kesehatan dan menurunnya kualitas hidup masyarakat (Afandi *et al.*, 2026). Oleh karena itu, deteksi dini diperlukan untuk membantu mengetahui kondisi pasien lebih awal agar pencegahan dan penanganan dapat dilakukan dengan lebih tepat.

Tekanan darah, termasuk kondisi hipertensi, dapat digunakan sebagai salah satu variabel dalam prediksi diabetes. Diabetes tipe 1 dan tipe 2 memiliki karakteristik yang berbeda. Pasien diabetes tipe 1 memerlukan insulin melalui suntikan atau pompa insulin, sedangkan diabetes tipe 2 berkaitan dengan resistensi insulin, yaitu kondisi ketika sel tubuh tidak dapat menggunakan insulin dengan baik (Setyawan & Wakhidah, 2025). Gunawan *et al.* (2025) menggunakan algoritma *Logistic Regression* untuk memprediksi risiko diabetes melitus tipe 2 dengan variabel usia, jenis kelamin, indeks massa tubuh (IMT), tekanan darah sistolik dan diastolik, serta riwayat penyakit. Selain hipertensi, obesitas merupakan faktor yang berkaitan dengan diabetes, khususnya diabetes melitus tipe 2. Menurut D. Fitriyani *et al.* (2025), risiko diabetes tipe 2 meningkat seiring dengan peningkatan IMT, yang menunjukkan adanya hubungan antara obesitas dan diabetes melitus. Rohmawati *et al.* (2025) juga menyatakan bahwa obesitas berkaitan dengan peningkatan kadar gula darah pada penderita diabetes melitus tipe 2. Berdasarkan data pasien Puskesmas Setiabudi yang tersedia, penelitian ini menggunakan IMT dan lingkaran perut sebagai variabel yang merepresentasikan kondisi obesitas.

Pemanfaatan *machine learning* menjadi salah satu pendekatan untuk memprediksi diabetes berdasarkan data kesehatan pasien dan telah digunakan dalam penelitian prediksi dengan tingkat akurasi yang tinggi (Setiadi *et al.*, 2025). Salah satu algoritma untuk klasifikasi dengan variabel target biner adalah *Logistic Regression* (Mushtaq *et al.*, 2022). Variabel prediktornya dapat berupa data kontinu, data kategorikal, atau kombinasi keduanya (Tarimana *et al.*, 2024). Fadli Kurniawan dan Ayu Megawaty (2025) membandingkan *Logistic Regression*, *Random Forest*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN) untuk memprediksi diabetes menggunakan 100.000 data pasien. Hasil penelitian tersebut menunjukkan bahwa *Logistic Regression* merupakan salah satu algoritma yang dapat diterapkan untuk prediksi diabetes dan dibandingkan dengan algoritma lain menggunakan metrik evaluasi klasifikasi. Namun, penerapan *machine learning* di bidang kesehatan juga menghadapi tantangan dalam mengomunikasikan hasil prediksi secara efektif kepada tenaga medis dan pasien (Ibrahim *et al.*, 2025; Nasution & Indra, 2026).

Penggunaan *Logistic Regression* untuk prediksi diabetes telah dilakukan dalam penelitian terdahulu. Nugraha dan Syarif (2025), melalui penelitian berjudul “Optimalisasi Model *Logistic Regression* untuk Prediksi Diabetes Menggunakan Seleksi Fitur Berbasis Korelasi”, menemukan bahwa pemilihan fitur berdasarkan hubungan antarfitur dapat menghasilkan kinerja model yang lebih baik daripada penggunaan seluruh fitur pada data yang mereka teliti. Persamaan penelitian tersebut dengan penelitian ini terletak pada penggunaan algoritma *Logistic Regression* untuk memprediksi diabetes. Perbedaannya terletak pada variabel yang digunakan. Penelitian ini melibatkan enam variabel yang mencakup faktor hipertensi dan obesitas serta variabel pendukung lainnya, yaitu usia, jenis kelamin, tekanan darah, IMT, lingkaran perut, dan Gula Darah Sewaktu, untuk memprediksi status diabetes.

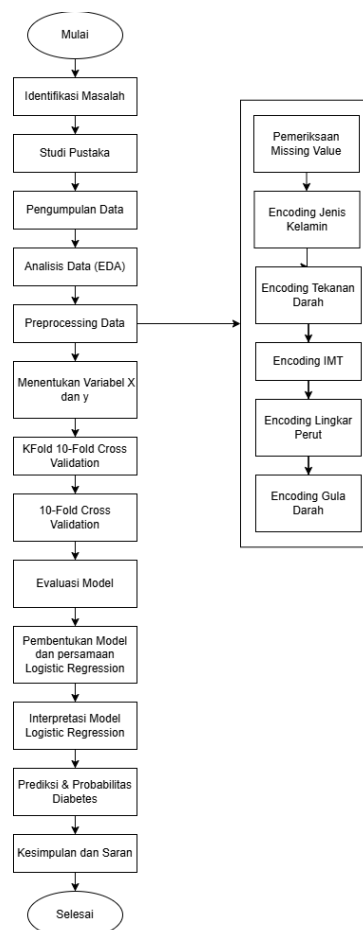
Penelitian lain yang berkaitan dengan penggunaan *Logistic Regression* dan analisis korelasi dalam pemrosesan data dilakukan oleh Kurniawati dan Brahma Arianto (2023) dalam artikel berjudul “Analisis Implementasi Seleksi Fitur Pada Klasifikasi Diabetes dengan Metode *Correlation Matrix* dan Algoritma *Logistic Regression*”. Penelitian tersebut menunjukkan bahwa penerapan *correlation matrix* dapat membantu mengidentifikasi atribut yang berkaitan dengan prediksi diabetes serta menelaah pengaruh seleksi fitur terhadap kinerja model klasifikasi. Berbeda dari penelitian tersebut, penelitian

ini tidak menjadikan seleksi fitur berbasis korelasi sebagai metode utama. Penelitian ini menggunakan enam variabel pada data pasien Puskesmas Setiabudi untuk memprediksi status diabetes berdasarkan faktor hipertensi dan obesitas serta variabel pendukung lainnya.

Berdasarkan latar belakang tersebut, penelitian ini diarahkan untuk mengimplementasikan algoritma *Logistic Regression* dalam membangun model prediksi status diabetes. Penelitian ini juga menganalisis hubungan usia, jenis kelamin, tekanan darah, IMT, lingkar perut, dan Gula Darah Sewaktu dengan probabilitas prediksi diabetes. Kinerja model dievaluasi melalui validasi silang *10-fold* berdasarkan prediksi *out-of-fold* (OOF) menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, dan *classification report* (Ashidiq *et al.*, 2025).

2. Metode Penelitian

Penelitian ini menggunakan metode kuantitatif dengan algoritma *Logistic Regression* untuk memprediksi status diabetes. Data yang digunakan merupakan data sekunder pasien Puskesmas Setiabudi tahun 2025 dalam format CSV yang diperoleh secara luring. Identitas pasien disamarkan melalui proses anonimisasi dengan mengacu pada Peraturan Menteri Kesehatan Nomor 269/MENKES/PER/III/2008 tentang Rekam Medis serta Undang-Undang Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi (Purnomo, 2024). *Dataset* terdiri atas 2.013 data pasien dan 12 variabel. Enam variabel prediktor yang digunakan ialah Usia, Jenis Kelamin, Tekanan Darah, IMT, Lingkar Perut, dan Gula Darah Sewaktu, sedangkan variabel target ialah Gula Darah dengan kategori Normal (0) dan Diabetes (1).



Gambar 1. Tahapan penelitian

Data yang tidak memiliki informasi yang diperlukan diperiksa pada tahap *preprocessing*. Tahapan penelitian meliputi analisis data eksploratif (EDA), pemeriksaan *missing value*, *encoding*, penentuan variabel prediktor X dan target y, pemodelan dengan *Logistic Regression*, validasi silang 10-fold dengan prediksi OOF, serta evaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, *confusion matrix*, dan *classification report*. Tahapan tersebut ditunjukkan pada Gambar 1. *Preprocessing* dilakukan sebelum pembagian data ke dalam *fold*, tanpa standarisasi atau normalisasi. Model menggunakan *LogisticRegression(max_iter=1000)*, sedangkan parameter lainnya mengikuti nilai bawaan Scikit-learn. Pembagian data menggunakan *KFold* dengan *n_splits=10*, *shuffle=True*, dan *random_state=42*. Hasil penelitian kemudian diinterpretasikan melalui koefisien, *logit*, probabilitas, dan *odds ratio*. Karena Gula Darah Sewaktu digunakan sebagai prediktor dan Gula Darah sebagai target, kedekatan informasi kedua variabel tersebut perlu diperhatikan karena berpotensi menyebabkan *target leakage*.

Penelitian ini memiliki beberapa keterbatasan yang perlu dipertimbangkan dalam interpretasi hasil. Distribusi kelas tidak seimbang, yaitu 1.962 data Normal dan 51 data Diabetes dengan rasio sekitar 38:1. Oleh karena itu, nilai *accuracy* perlu dipertimbangkan bersama *precision*, *recall*, *F1-score*, dan *confusion matrix*. Evaluasi dilakukan menggunakan validasi silang 10-fold dengan prediksi *out-of-fold* (OOF) pada *dataset* yang sama, sehingga hasilnya belum menggambarkan kinerja model pada data uji independen atau populasi eksternal. Selain itu, Gula Darah Sewaktu digunakan sebagai variabel prediktor, sedangkan Gula Darah digunakan sebagai variabel target. Kedekatan informasi antara keduanya berpotensi menyebabkan *target leakage* apabila Gula Darah Sewaktu berkaitan langsung dengan pembentukan label target. Oleh karena itu, penelitian selanjutnya disarankan menggunakan *dataset* independen dan validasi eksternal, serta mempertimbangkan kembali penggunaan Gula Darah Sewaktu sebagai prediktor untuk menguji kemampuan generalisasi model.

3. Hasil dan Pembahasan

3.1 Hasil

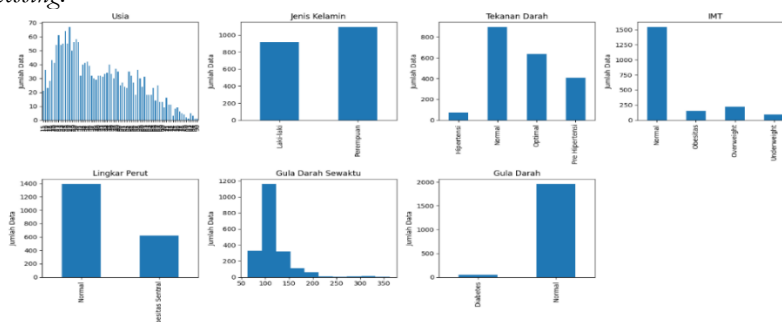
Analisis data eksploratif (*exploratory data analysis*/EDA) dilakukan untuk mengetahui karakteristik awal data sebelum tahap *preprocessing* dan pemodelan. Hasil pembacaan *dataset* ditunjukkan pada Gambar 2.

Jumlah data : 2013
Jumlah variabel: 12

Gambar 2. Hasil pembacaan *dataset* dan penampilan jumlah data.

Sumber: Penelitian (2026).

Berdasarkan hasil pembacaan, *dataset* terdiri atas 2.013 data pasien dan 12 variabel. Visualisasi masing-masing variabel disajikan pada Gambar 3 untuk memberikan gambaran mengenai distribusi data sebelum *preprocessing*.



Gambar 3. Hasil visualisasi variabel.

Sumber: Penelitian (2026).

Distribusi variabel target, yaitu Gula Darah, ditunjukkan pada Gambar 4. Sebanyak 1.962 data termasuk kelas Normal (0), sedangkan 51 data termasuk kelas Diabetes (1). Dengan demikian, jumlah data pada kelas Normal jauh lebih banyak daripada kelas Diabetes.

```
===== DISTRIBUSI GULA DARAH =====
Gula Darah
Normal      1962
Diabetes     51
Name: count, dtype: int64

Persentase:
Gula Darah
Normal      97.47
Diabetes     2.53
Name: proportion, dtype: float64
```

Gambar 4. Distribusi kelas Gula Darah.

Sumber: Penelitian (2026).

Tahap *preprocessing* mencakup pemeriksaan nilai kosong (*missing value*) dan pengkodean variabel kategorikal (*encoding*) menggunakan pemetaan yang telah ditentukan. Tahap ini dilakukan sebelum pembentukan variabel prediktor dan target serta sebelum pembagian data menggunakan *K-Fold*. Tidak dilakukan imputasi, normalisasi, standardisasi, *oversampling*, *undersampling*, maupun SMOTE. Pembersihan data dan penanganan nilai kosong merupakan bagian dari *preprocessing* dalam pembangunan model regresi logistik (Putra *et al.*, 2024). Hasil pemeriksaan nilai kosong sebelum pengkodean ditunjukkan pada Gambar 5. Seluruh variabel memiliki 0 nilai kosong, sehingga tidak diperlukan pengisian data dan proses dapat dilanjutkan ke tahap berikutnya.

```
===== PEMERIKSAAN MISSING VALUE =====
Usia          0
Jenis Kelamin 0
Tinggi Badan  0
Berat Badan   0
Lingkar Perut Dalam 0
Sistolik      0
Distolik      0
Gula Darah Sewaktu 0
Tekanan Darah 0
Gula Darah    0
IMT           0
Lingkar Perut 0
dtype: int64
Total missing value: 0
```

Gambar 5. Hasil pemeriksaan nilai kosong.

Sumber: Penelitian (2026).

Hasil pengkodean data ditunjukkan pada Gambar 6. Kategori pada variabel yang digunakan diubah menjadi nilai numerik sesuai dengan pemetaan pada Tabel 1.

```
===== HASIL ENCODING =====
Jenis Kelamin  Tekanan Darah  IMT  Lingkar Perut  Gula Darah
0              1              2    4              2          0
1              2              2    2              1          0
2              2              2    2              1          0
3              2              2    4              2          0
4              2              2    4              2          0
5              2              2    4              2          0
6              2              2    4              2          0
7              2              2    2              2          0
8              1              2    2              1          0
9              2              2    2              2          0
```

Gambar 6. Hasil pengkodean data.

Sumber: Penelitian (2026).

Tabel 1. Hasil pengkodean variabel

Variabel	Kategori	Kode
Jenis Kelamin	Laki-laki	1
	Perempuan	2
Tekanan Darah	Optimal	1
	Normal	2
	Prahipertensi	3
	Hipertensi	4
IMT	<i>Underweight</i>	1
	Normal	2
	<i>Overweight</i>	3
	Obesitas	4
Lingkar Perut	Normal	1
	Obesitas sentral	2
Gula Darah	Normal	0
	Diabetes	1

Sumber: Penelitian (2026).

Berdasarkan Tabel 1, seluruh variabel kategorikal yang digunakan telah dikonversi menjadi nilai numerik. Variabel Gula Darah dikodekan sebagai Normal (0) dan Diabetes (1) karena digunakan sebagai variabel target klasifikasi. Pemeriksaan setelah pengkodean, sebagaimana ditunjukkan pada Gambar 7, kembali menghasilkan 0 nilai kosong pada seluruh variabel. Jumlah data tetap sebanyak 2.013, sehingga proses pengkodean tidak menyebabkan kehilangan data.

```
===== MISSING VALUE SETELAH ENCODING =====
Usia          0
Jenis Kelamin 0
Tinggi Badan  0
Berat Badan   0
Lingkar Perut Dalam 0
Sistolik      0
Distolik      0
Gula Darah Sewaktu 0
Tekanan Darah 0
Gula Darah    0
IMT           0
Lingkar Perut 0
dtype: int64

Total missing value setelah encoding: 0
Jumlah data: 2013
```

Gambar 7. Hasil pemeriksaan nilai kosong setelah pengkodean.

Sumber: Penelitian (2026).

Selanjutnya, variabel prediktor, yaitu **X**, ditentukan dari Usia, Jenis Kelamin, Tekanan Darah, IMT, Lingkar Perut, dan Gula Darah Sewaktu. Variabel target, yaitu **y**, adalah status Gula Darah. Hasil penentuan variabel pada Gambar 8 menunjukkan bahwa **X** berukuran 2.013×6 , sedangkan **y** berisi 2.013 data. Distribusi kelas aktual pada **y** terdiri atas 1.962 data Normal (0) dan 51 data Diabetes (1).

```
Ukuran X : (2013, 6)
Ukuran y : (2013,)

Distribusi Data Aktual:
Gula Darah
0    1962
1     51
Name: count, dtype: int64
```

Gambar 8. Hasil penentuan variabel X dan y.

Sumber: Penelitian (2026).

Validasi model dilakukan menggunakan validasi silang 10-fold. Pada setiap *fold*, sembilan bagian data digunakan untuk melatih model dan satu bagian digunakan sebagai data validasi. Proses ini dilakukan bergantian hingga seluruh data memperoleh prediksi *out-of-fold* (OOF). Pembagian data menggunakan KFold dengan `n_splits=10`, `shuffle=True`, dan `random_state=42`. Kode pembagian data serta hasilnya ditunjukkan pada Gambar 9 dan Gambar 10.

```

K-FOLD 10-FOLD

kfold = KFold(
    n_splits=10,
    shuffle=True,
    random_state=42
)

folds = list(kfold.split(X, y))

print("Jumlah Fold :", len(folds))

for i, (train_idx, valid_idx) in enumerate(folds, 1):
    print(
        f"Fold {i:2d} : "
        f"Training = {len(train_idx)}, "
        f"Validasi = {len(valid_idx)}"
    )

```

Gambar 9. Kode pembagian data menggunakan K-Fold 10-fold.

Sumber: Penelitian (2026).

```

Jumlah Fold : 10
Fold 1 : Training = 1811, Validasi = 202
Fold 2 : Training = 1811, Validasi = 202
Fold 3 : Training = 1811, Validasi = 202
Fold 4 : Training = 1812, Validasi = 201
Fold 5 : Training = 1812, Validasi = 201
Fold 6 : Training = 1812, Validasi = 201
Fold 7 : Training = 1812, Validasi = 201
Fold 8 : Training = 1812, Validasi = 201
Fold 9 : Training = 1812, Validasi = 201
Fold 10 : Training = 1812, Validasi = 201

```

Gambar 10. Hasil pembagian data menjadi 10 fold.

Sumber: Penelitian (2026).

Hasil validasi pada Gambar 11 menunjukkan bahwa *fold* 1, 2, 3, 4, 6, 7, 8, 9, dan 10 masing-masing memperoleh *accuracy* sebesar 100,00%. *Fold* 5 memperoleh *accuracy* sebesar 99,50%. Rata-rata *accuracy* dari kesepuluh *fold* adalah 99,95%, dengan perhitungan sebagai berikut:

$$\bar{A} = \frac{(9 \times 10000) + 0,9950}{10} = 0,9995 = 99,95\%.$$

```

Fold 1 : 1.0000 (100.00%)
Fold 2 : 1.0000 (100.00%)
Fold 3 : 1.0000 (100.00%)
Fold 4 : 1.0000 (100.00%)
Fold 5 : 0.9950 (99.50%)
Fold 6 : 1.0000 (100.00%)
Fold 7 : 1.0000 (100.00%)
Fold 8 : 1.0000 (100.00%)
Fold 9 : 1.0000 (100.00%)
Fold 10 : 1.0000 (100.00%)

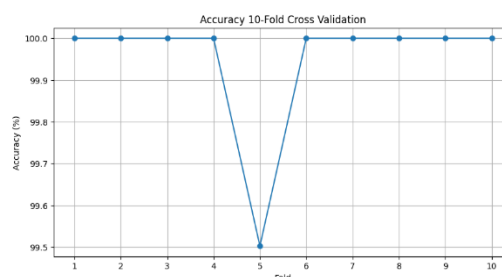
Mean Accuracy 10-Fold:
0.9995 (99.95%)

```

Gambar 11. Hasil validasi silang 10-fold.

Sumber: Penelitian (2026).

Pola *accuracy* pada setiap *fold* juga disajikan pada Gambar 12. Grafik tersebut menunjukkan sembilan *fold* dengan *accuracy* 100,00% dan satu *fold* dengan *accuracy* 99,50%.



Gambar 12. Grafik *accuracy* validasi silang 10-fold.

Sumber: Penelitian (2026).

Evaluasi berdasarkan seluruh prediksi OOF menghasilkan *accuracy* 99,95%, *precision* 98,08%, *recall* 100,00%, dan *F1-score* 99,03%. Hasil evaluasi tersebut ditunjukkan pada Gambar 13.

```
===== HASIL EVALUASI 10-FOLD =====
Accuracy : 99.95%
Precision : 98.08%
Recall    : 100.00%
F1-Score  : 99.03%
```

Gambar 13. Hasil evaluasi model berdasarkan validasi silang 10-fold.

Sumber: Penelitian (2026).

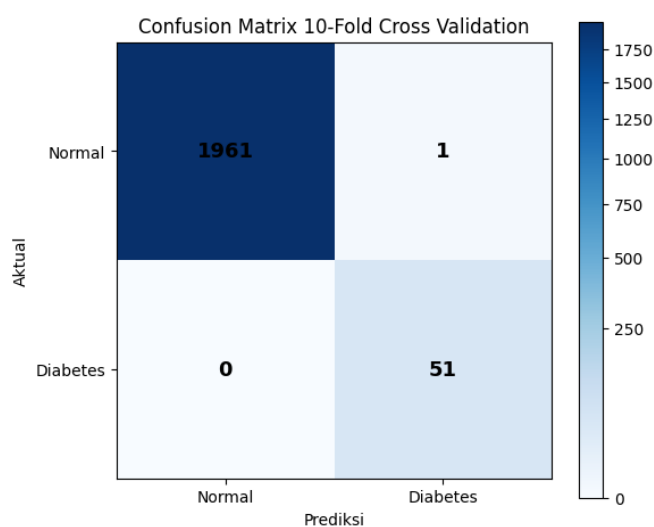
Confusion matrix pada Gambar 14 menunjukkan 1.961 *true negative* (TN), 1 *false positive* (FP), 0 *false negative* (FN), dan 51 *true positive* (TP). Artinya, sebanyak 1.961 data Normal dan 51 data Diabetes diklasifikasikan dengan benar. Satu data Normal diprediksi sebagai Diabetes, sedangkan tidak ada data Diabetes yang diprediksi sebagai Normal. Visualisasi hasil yang sama disajikan pada Gambar 15.

```
===== CONFUSION MATRIX 10-FOLD =====
[[1961  1]
 [  0  51]]

TN : 1961
FP : 1
FN : 0
TP : 51
```

Gambar 14. Hasil *confusion matrix* validasi silang 10-fold.

Sumber: Penelitian (2026).



Gambar 15. Visualisasi *confusion matrix* validasi silang 10-fold.

Sumber: Penelitian (2026).

Hasil *classification report* pada Gambar 16 menunjukkan bahwa kelas Normal memiliki *precision*, *recall*, dan *F1-score* yang masing-masing ditampilkan sebagai 1,00. Pada kelas Diabetes, *precision* ditampilkan sebesar 0,98, *recall* sebesar 1,00, dan *F1-score* sebesar 0,99. Nilai *support* masing-masing kelas adalah 1.962 untuk Normal dan 51 untuk Diabetes. Laporan tersebut menampilkan *accuracy* sebesar 1,00 setelah pembulatan, sedangkan nilai sebelum pembulatan berdasarkan seluruh prediksi OOF adalah 99,95%.

```

===== CLASSIFICATION REPORT 10-FOLD =====
              precision    recall  f1-score   support

   Normal         1.00        1.00        1.00       1962
   Diabetes        0.98        1.00        0.99         51

 accuracy          0.99        1.00        1.00       2013
 macro avg         0.99        1.00        1.00       2013
 weighted avg      1.00        1.00        1.00       2013

```

Gambar 16. Hasil *classification report* validasi silang 10-fold.

Sumber: Penelitian (2026).

Perhitungan metrik evaluasi berdasarkan *confusion matrix* adalah sebagai berikut:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% = \frac{51 + 1.961}{51 + 1.961 + 1 + 0} \times 100\% = \frac{2.012}{2.013} \times 100\% = 99,95\%.$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% = \frac{51}{51 + 1} \times 100\% = \frac{51}{52} \times 100\% = 98,08\%.$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% = \frac{51}{51 + 0} \times 100\% = \frac{51}{51} \times 100\% = 100,00\%.$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0,9808 \times 1,0000}{0,9808 + 1,0000} \approx 99,03\%.$$

Dengan demikian, dari 2.013 prediksi OOF, sebanyak 2.012 prediksi benar dan satu prediksi salah. Ringkasan hasil evaluasi disajikan pada Tabel 2.

Tabel 2. Hasil evaluasi model *Logistic Regression*

Metrik evaluasi	Nilai
<i>Accuracy</i> berdasarkan prediksi OOF	99,95%
<i>Precision</i> kelas Diabetes	98,08%
<i>Recall</i> kelas Diabetes	100,00%
<i>F1-score</i> kelas Diabetes	99,03%
Rata-rata <i>accuracy</i> validasi silang 10-fold	99,95%

Sumber: Penelitian (2026).

Setelah evaluasi silang dilakukan, model akhir *Logistic Regression* dibentuk menggunakan seluruh 2.013 data penelitian. Model menggunakan `max_iter=1000`, sedangkan parameter `solver`, `C`, dan `class_weight` tidak ditentukan secara eksplisit sehingga mengikuti nilai bawaan Scikit-learn. Kode pembentukan model dan hasilnya disajikan pada Gambar 17 dan Gambar 18.

```

# =====
# MODEL AKHIR LOGISTIC REGRESSION
# =====

model = LogisticRegression(
    max_iter=1000
)

# Model akhir menggunakan seluruh data
model.fit(
    X,
    y
)

print("Model Logistic Regression berhasil dibentuk.")
print("Jumlah data :", len(X))

```

Gambar 17. Kode pembentukan model akhir *Logistic Regression*.
Sumber: Penelitian (2026).

```

Model Logistic Regression berhasil dibentuk.
Jumlah data : 2013

```

Gambar 18. Hasil pembentukan model akhir *Logistic Regression*.
Sumber: Penelitian (2026).

Nilai *intercept* dan koefisien dari model akhir digunakan untuk menyusun persamaan regresi logistik berdasarkan enam variabel prediktor. Kode pengambilan nilai tersebut disajikan pada Gambar 19, sedangkan hasil persamaannya ditunjukkan pada Gambar 20. Nilai *intercept* yang diperoleh adalah -279,9988. Penelitian ini tidak melakukan standardisasi maupun normalisasi, sehingga koefisien dihitung berdasarkan skala variabel setelah *preprocessing*.

```

# =====
# PERSAMAAN LOGISTIC REGRESSION
# =====

intercept = model.intercept_[0]
coef = model.coef_[0]

print("==== PERSAMAAN LOGISTIC REGRESSION =====\n")

print(
    f"logit(p) = {intercept:.4f}"
)

for nama, nilai in zip(
    X.columns,
    coef
):
    tanda = "+" if nilai >= 0 else "-"

    print(
        f"
        {tanda}
        f"({abs(nilai):.4f} x {nama})"
    )

```

Gambar 19. Kode pembentukan persamaan *Logistic Regression*.
Sumber: Penelitian (2026).

```

==== PERSAMAAN LOGISTIC REGRESSION ====

logit(p) = -279.9988
- (0.0103 x Usia)
+ (0.2183 x Jenis Kelamin)
+ (0.0445 x Tekanan Darah)
+ (0.1172 x IMT)
- (0.1504 x Lingkar Perut)
+ (1.4112 x Gula Darah Sewaktu)

```

Gambar 20. Hasil persamaan *Logistic Regression*.
Sumber: Penelitian (2026).

Koefisien model akhir adalah $-0,0103$ untuk Usia, $0,2183$ untuk Jenis Kelamin, $0,0445$ untuk Tekanan Darah, $0,1172$ untuk IMT, $-0,1504$ untuk Lingkar Perut, dan $1,4112$ untuk Gula Darah Sewaktu. Hasil perhitungan *odds ratio* ditunjukkan pada Gambar 21.

```
===== ODDS RATIO =====
Usia                : 0.9898
Jenis Kelamin       : 1.2440
Tekanan Darah       : 1.0455
IMT                 : 1.1243
Lingkar Perut       : 0.8604
Gula Darah Sewaktu  : 4.1010
```

Gambar 21. Hasil perhitungan *odds ratio*.

Sumber: Penelitian (2026).

Berdasarkan hasil tersebut, *odds ratio* masing-masing variabel adalah $0,9898$ untuk Usia; $1,2440$ untuk Jenis Kelamin; $1,0455$ untuk Tekanan Darah; $1,1243$ untuk IMT; $0,8604$ untuk Lingkar Perut; dan $4,1010$ untuk Gula Darah Sewaktu. Model akhir selanjutnya digunakan untuk menghasilkan prediksi dan probabilitas pada seluruh data. Gambar 22 menunjukkan bahwa proses tersebut menghasilkan 2.013 label prediksi dan 2.013 nilai probabilitas.

```
Prediksi berhasil dibuat.
Jumlah prediksi : 2013
Jumlah probabilitas : 2013
```

Gambar 22. Hasil prediksi terhadap seluruh data.

Sumber: Penelitian (2026).

Hasil pengurutan probabilitas disajikan pada Gambar 23 dan Gambar 24. Sepuluh data dengan probabilitas Diabetes tertinggi ditampilkan dengan probabilitas 100% serta memiliki kelas aktual dan prediksi Diabetes (1). Sepuluh data dengan probabilitas Diabetes terendah ditampilkan dengan probabilitas 0% serta memiliki kelas aktual dan prediksi Normal (0).

```
===== PROBABILITAS TERTINGGI =====
Aktual  Prediksi  Probabilitas Diabetes (%)
65      1          1          100.000000%
62      1          1          100.000000%
1808    1          1          100.000000%
1828    1          1          100.000000%
60      1          1          100.000000%
1171    1          1          100.000000%
1200    1          1          100.000000%
108     1          1          100.000000%
1766    1          1          100.000000%
73      1          1          100.000000%
```

Gambar 23. Sepuluh data dengan probabilitas Diabetes tertinggi.

Sumber: Penelitian (2026).

===== PROBABILITAS TERENDAH =====			
Aktual	Prediksi	Probabilitas Diabetes (%)	
922	0	0	0.000000%
657	0	0	0.000000%
619	0	0	0.000000%
1654	0	0	0.000000%
899	0	0	0.000000%
592	0	0	0.000000%
650	0	0	0.000000%
606	0	0	0.000000%
656	0	0	0.000000%
1523	0	0	0.000000%

Gambar 24. Sepuluh data dengan probabilitas Diabetes terendah.

Sumber: Penelitian (2026).

Perbandingan distribusi kelas aktual dan kelas prediksi disajikan pada Gambar 25. Distribusi kelas aktual terdiri atas 1.962 data Normal dan 51 data Diabetes. Jumlah pada kelas prediksi perlu dicocokkan kembali dengan keluaran pada Gambar 25 karena hasil *confusion matrix* yang dilaporkan menunjukkan 1.961 prediksi Normal dan 52 prediksi Diabetes.

```

Distribusi Aktual:
Aktual
0    1962
1      51
Name: count, dtype: int64

Distribusi Prediksi:
Prediksi
0    1962
1      51
Name: count, dtype: int64

```

Gambar 25. Distribusi kelas aktual dan kelas prediksi.

Sumber: Penelitian (2026).

3.2 Pembahasan

Hasil evaluasi OOF menunjukkan bahwa model *Logistic Regression* mengklasifikasikan 2.012 dari 2.013 data dengan benar. *Accuracy* sebesar 99,95% dan rata-rata *accuracy* validasi silang 10-fold sebesar 99,95% menunjukkan kinerja yang tinggi pada data penelitian. Sembilan *fold* memperoleh *accuracy* 100,00%, sedangkan satu *fold* memperoleh 99,50%. Penggunaan validasi silang membantu memberikan gambaran kinerja model pada pembagian data yang berbeda (Qiu, 2024). Meskipun demikian, distribusi kelas dalam penelitian ini tidak seimbang: terdapat 1.962 data Normal dan hanya 51 data Diabetes. Oleh karena itu, *accuracy* tidak sebaiknya ditafsirkan sendiri. *Precision* kelas Diabetes sebesar 98,08% menunjukkan bahwa 51 dari 52 data yang diprediksi sebagai Diabetes merupakan data Diabetes. *Recall* sebesar 100,00% menunjukkan bahwa seluruh 51 data Diabetes dikenali pada prediksi OOF, sedangkan *F1-score* sebesar 99,03% merangkum keseimbangan antara *precision* dan *recall*. Nilai-nilai ini berlaku untuk evaluasi pada *dataset* penelitian; belum dapat dianggap sebagai bukti kinerja yang sama pada pasien dari populasi lain.

Confusion matrix memperjelas jenis kesalahan yang terjadi. Model menghasilkan satu *false positive*, yaitu data Normal yang diprediksi sebagai Diabetes, dan tidak menghasilkan *false negative* pada data OOF. Karena seluruh kasus Diabetes dalam data penelitian berhasil dikenali, hasil ini tampak baik untuk tujuan identifikasi awal. Namun, jumlah kasus Diabetes yang hanya 51 perlu diperhatikan ketika menafsirkan kestabilan hasil tersebut. Perbedaan antara *accuracy* 1,00 pada tampilan *classification report* dan 99,95% pada perhitungan bukanlah perbedaan hasil evaluasi, melainkan akibat pembulatan angka dalam laporan.

Koefisien model menunjukkan arah hubungan setiap prediktor dengan *logit* kelas Diabetes ketika prediktor lain dianggap tetap. Koefisien positif terdapat pada Jenis Kelamin, Tekanan Darah, IMT, dan Gula Darah Sewaktu, sedangkan koefisien negatif terdapat pada Usia dan Lingkar Perut. Untuk Gula Darah Sewaktu, *odds ratio* sebesar 4,1010 berarti peningkatan satu unit pada bentuk variabel yang dimasukkan ke model berkaitan dengan *odds* klasifikasi Diabetes sebesar 4,1010 kali, dengan prediktor lain tetap. Penafsiran “satu unit” harus disesuaikan dengan cara Gula Darah Sewaktu dicatat atau dikodekan dalam data. Usia memiliki koefisien $-0,0103$ dan *odds ratio* 0,9898. Dalam model ini, peningkatan satu unit Usia berkaitan dengan *odds* klasifikasi Diabetes sebesar 0,9898 kali, dengan variabel lain tetap. Lingkar Perut memiliki koefisien $-0,1504$ dan *odds ratio* 0,8604; perubahan satu unit pada kode Lingkar Perut berkaitan dengan *odds* sebesar 0,8604 kali. Arah negatif kedua variabel tersebut tidak membuktikan bahwa usia yang lebih tinggi atau obesitas sentral melindungi seseorang dari diabetes. Interpretasinya perlu mempertimbangkan pengkodean, distribusi kategori, karakteristik data, dan hubungan antarvariabel dalam model.

Koefisien Gula Darah Sewaktu sebesar 1,4112 merupakan yang terbesar secara numerik di antara koefisien yang dilaporkan. Akan tetapi, karena variabel tidak distandardisasi dan memiliki skala atau bentuk pengkodean yang berbeda, besar koefisien antarvariabel tidak dapat langsung dibandingkan untuk menyimpulkan prediktor yang paling berpengaruh. Nilai *intercept* $-279,9988$ juga perlu dibaca bersama skala dan nilai prediktor yang digunakan, bukan dibandingkan secara langsung dengan *intercept* model yang memakai fitur terstandarisasi. Hal penting lainnya adalah kedekatan informasi antara Gula Darah Sewaktu sebagai prediktor dan status Gula Darah sebagai target. Apabila kategori target dibentuk secara langsung dari Gula Darah Sewaktu, penggunaan variabel tersebut sebagai prediktor berpotensi menimbulkan *target leakage* dan membuat hasil evaluasi tampak sangat tinggi. Karena itu, cara pembentukan label Normal dan Diabetes harus dijelaskan dan diperiksa sebelum hasil model ditafsirkan sebagai kemampuan prediksi yang berlaku secara klinis. Hasil probabilitas yang ditampilkan sebagai 0% atau 100% juga perlu diperiksa nilai lengkapnya untuk memastikan apakah angka tersebut merupakan nilai persis atau hasil pembulatan tampilan.

Prediksi pada seluruh 2.013 data setelah model akhir dilatih menggunakan data yang sama merupakan keluaran model akhir, bukan pengujian pada data independen. Evaluasi kinerja dalam penelitian ini didasarkan pada prediksi OOF, sedangkan kemampuan generalisasi ke pasien lain masih memerlukan data uji independen atau validasi eksternal. Dengan demikian, hasil penelitian menunjukkan kinerja klasifikasi yang sangat tinggi pada data dan prosedur evaluasi yang digunakan, tetapi tetap perlu ditafsirkan dengan memperhatikan ketimpangan kelas, cara pengkodean, potensi *target leakage*, dan belum adanya validasi eksternal.

4. Kesimpulan

Berdasarkan penelitian mengenai prediksi diabetes berdasarkan hipertensi dan obesitas menggunakan algoritma *Logistic Regression* di Puskesmas Setiabudi, dapat disimpulkan bahwa algoritma tersebut berhasil diimplementasikan untuk membangun model prediksi status diabetes pada data pasien Puskesmas Setiabudi. Model menggunakan enam variabel prediktor, yaitu Usia, Jenis Kelamin, Tekanan Darah, IMT, Lingkar Perut, dan Gula Darah Sewaktu. Variabel targetnya adalah Gula Darah, yang terdiri atas kelas Normal (0) dan Diabetes (1).

Tekanan Darah, IMT, dan Lingkar Perut digunakan untuk merepresentasikan fokus penelitian terhadap faktor hipertensi dan obesitas, sedangkan Usia, Jenis Kelamin, dan Gula Darah Sewaktu digunakan sebagai variabel pendukung dalam proses prediksi. Interpretasi model dilakukan berdasarkan koefisien regresi dan *odds ratio* untuk melihat arah serta besaran hubungan masing-masing variabel dengan kecenderungan prediksi diabetes. Keenam variabel prediktor memiliki koefisien yang berbeda terhadap *log-odds* diabetes. Namun, koefisien dan *odds ratio* tersebut tidak digunakan sebagai dasar tunggal untuk menyatakan penerimaan atau penolakan hipotesis statistik. Pengujian signifikansi setiap koefisien memerlukan informasi inferensial, seperti *standard error*, nilai *z*, *p-value*, atau interval

kepercayaan. Karena pengujian tersebut belum dilakukan, penelitian ini tidak menyatakan penerimaan H1 secara statistik dan membatasi interpretasinya pada hubungan prediktor dengan probabilitas diabetes berdasarkan model *Logistic Regression*.

Proses validasi silang 10-fold menghasilkan prediksi *out-of-fold* (OOF) untuk seluruh 2.013 data. Rata-rata *accuracy* antarfold mencapai 99,95%, dengan sembilan *fold* memperoleh *accuracy* 100,00% dan satu *fold* memperoleh *accuracy* 99,50%. Evaluasi berdasarkan prediksi OOF menghasilkan *accuracy* 99,95%, *precision* 98,08%, *recall* 100,00%, dan *F1-score* 99,03%. Hasil *confusion matrix* menunjukkan TN = 1.961, FP = 1, FN = 0, dan TP = 51. Dengan demikian, terdapat 2.012 prediksi benar dari 2.013 data.

Hasil tersebut menunjukkan kinerja klasifikasi yang tinggi pada *dataset* penelitian, tetapi belum dapat digunakan untuk menyimpulkan kemampuan generalisasi model pada populasi lain atau *dataset* independen. Interpretasi kinerja model juga perlu mempertimbangkan penggunaan Gula Darah Sewaktu sebagai prediktor karena variabel tersebut berkaitan erat dengan status diabetes sebagai target. Oleh karena itu, hasil evaluasi perlu dibaca bersama dengan pemeriksaan proses pembentukan label. Apabila diperlukan, analisis perbandingan tanpa Gula Darah Sewaktu sebagai prediktor dapat dilakukan untuk menilai pengaruh penggunaan variabel tersebut terhadap kinerja model.

5. Daftar Pustaka

- Afandi, M., Riskianto, D. D., Ramadhan, M. R., & Sudriyanto. (2026). Analisis prediksi diabetes berbasis artificial neural network menggunakan data CDC diabetes health indicators. *Jurnal Riset Sistem dan Teknologi Informasi (RESTIA)*, 4(1), 22–29. <https://doi.org/10.30787/restia.v4i1.2096>
- Ashidiq, M. F., Muflikhah, L., & Setiawan, B. D. (2025). Deteksi nefropati diabetik pada pasien diabetes melitus menggunakan regresi logistik. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 9(2), 1–9. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/14478>
- Astofa, A., Rosyani, P., Rahmawati, & Apandi, S. (2025). Evaluasi komparatif algoritma machine learning untuk prediksi dini diabetes. *Bulletin of Computer Science Research*, 6(1), 558–565. <https://doi.org/10.47065/bulletincsr.v6i1.859>
- Badan Kebijakan Pembangunan Kesehatan. (2023). *Hasil utama SKI 2023*. Kemenkes BKPK. <https://www.badankebijakan.kemkes.go.id/daftar-frequently-asked-question-seputar-hasil-utama-ski-2023/hasil-utama-ski-2023/>
- Fitriyani, D., Dwi Ramadhani, R., Khabiza Syahdu, R., & Hajjah Purba, S. (2025). Pengaruh obesitas terhadap kejadian diabetes melitus: Studi literatur. *Jurnal Ilmu Kedokteran dan Kesehatan*, 12(2), 291–303. <https://doi.org/10.33024/jikk.v12i2.16198>
- Gunawan, S., Astuti, R., Prihartono, W., & Hamonangan, R. (2025). Prediksi diabetes mellitus tipe 2 dengan algoritma logistic regression untuk pendeteksian dini. *Jurnal Informatika dan Teknik Elektro Terapan*, 13(1), 848–894. <https://doi.org/10.23960/jitet.v13i1.5747>
- Ibrahim, M. C., Fachruddin, & Nurhadi. (2025). Comparison of diabetes prediction data using machine learning perbandingan data prediksi diabetes menggunakan machine learning. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5(October), 1423–1436. <https://doi.org/10.57152/malcom.v5i4.2301>

- Kurniawan, M. F., & Megawaty, D. A. (2025). Comparison of logistic regression, random forest, support vector machine (SVM) and K-nearest neighbor (KNN) algorithms in diabetes prediction. *Journal of Applied Informatics and Computing*, 9(5), 2154–2162. <https://doi.org/10.30871/jaic.v9i5.9815>
- Kurniawati, F., & Arianto, D. B. (2023). Analisis implementasi seleksi fitur pada klasifikasi diabetes dengan metode corellation matrix dan algoritma logistic regression. *Jurnal Informatik*, 19(3), 157–164. <https://doi.org/10.52958/iftk.v19i3.6019>
- Mumuh, U., Syafei, A., & Utami, R. P. (2026). Analisis keberhasilan atau kegagalan pasien diabetes melitus dalam mengadopsi gaya hidup sehat di Puskesmas Telagasari. *Masyarakat Sehat Indonesia*, 5(1), 61–69. <https://doi.org/10.70304/jmsi.v5i01.86>
- Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. *Mobile Information Systems*, 2022, 6521532. <https://doi.org/10.1155/2022/6521532>
- Nasution, A. N., & Indra, Z. (2026). Deteksi dini diabetes melitus tipe 2 berdasarkan data rekam medis pasien RSUD Pintu Padang Kabupaten Tapsel menggunakan algoritma random forest. *Jurnal Komputer Teknologi Informasi Sistem Komputer*, 5(2), 1633–1644. <https://doi.org/10.62712/juktisi.v5i2.1359>
- Nugraha, W., & Syarif, M. (2025). Optimalisasi model logistic regression untuk prediksi diabetes menggunakan seleksi fitur berbasis korelasi. *Jurnal ICT: Information Communication & Technology*, 25(2), 139–144. <https://doi.org/10.36054/jict-ikmi.v25i2.321>
- Purnomo, A. (2024). *Akibat hukum bagi rumah sakit terhadap penyebaran data rekam medis pasien* [Skripsi sarjana, Universitas Merdeka Pasuruan]. Repository Universitas Merdeka Pasuruan. <https://repository.unmerpas.ac.id/796/>
- Putra, F., Tahiyat, H. F., Ihsan, R. M., Rahmadden, & Efrizoni, L. (2024). Penerapan algoritma K-nearest neighbor menggunakan wrapper sebagai preprocessing untuk penentuan keterangan berat badan manusia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(1), 273–281. <https://doi.org/10.57152/malcom.v4i1.1085>
- Qiu, J. (2024). An analysis of model evaluation with cross-validation: Techniques, applications, and recent advances. *Proceedings of ICFITBA 2024 Workshop: Finance in the Age of Environmental Risks and Sustainability*, 99(1), 69–72. <https://doi.org/10.54254/2754-1169/99/2024OX0213>
- Rohmawati, R., Faizah, I., Nur Hasina, S., Yunita Sari, R., & Aditya Putri, R. (2025). Hubungan obesitas dengan peningkatan kadar gula darah pada penderita diabetes mellitus tipe 2. *Jurnal Penelitian Perawat Profesional*, 7(5), 33–40. <https://doi.org/10.37287/jppp.v7i5.434>
- Setiadi, W. T., Jollyta, D., & Setiawan, E. B. (2025). Prediksi risiko diabetes menggunakan model regresi logistik dan random forest. *Seminar Nasional Informatika (SENATIKA): Prosiding SENATIKA 2025*, 342–350. <https://ejournal.pelitaindonesia.ac.id/index.php/SENATIKA/article/view/5581>
- Setyawan, N. H., & Wakhidah, N. (2025). Analisis perbandingan metode logistic regression, random forest, gradient boosting untuk prediksi diabetes. *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 10(1), 150–162. <https://doi.org/10.29100/jupi.v10i1.5799>

Tarimana, A. A., Fajar, M. R. S., Saktiawan, M. A., & Saputra, R. A. (2024). Prediksi penyakit hipertensi menggunakan machine learning dengan algoritma regresi logistik. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(6), 12062–12068. <https://doi.org/10.36040/jati.v8i6.11793>.