Jurnal ICT : Information Communication & Technology
Vol. 25 N0.2, Desember 2025, pp. 139 - 144
p-ISSN: 2302-0261, e-ISSN: 2303-3363, DOI: https://doi.org/10.36054/jict-ikmi.v25i2.321

Optimalisasi Model *Logistic Regression* untuk Prediksi Diabetes Menggunakan Seleksi Fitur Berbasis Korelasi

Wahyu Nugraha[1], Muhamad Syarif[2*]
[1]Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia
Email: [1]wahyu.whn@bsi.ac.id , [2]muhamad.mdx@bsi.ac.id

INFORMASI ARTIKEL

Histori artikel:

Naskah masuk, 5 November 2025
Direvisi, 24 November 2025
Diiterima, 18 Desember 2025

Kata Kunci:

Prediksi Diabetes
Logistic Regression
Machine Learning
Seleksi Fitur
Analisis Korelasi

ABSTRAK

Abstract- *Diabetes Mellitus is a pressing global health challenge, making early detection a key component of effective intervention. Machine learning has shown great potential in predicting diabetes risk. Among various models, Logistic Regression* (LR) *is often favored in a medical context due to its high interpretability, although its accuracy frequently lags behind more complex black-box models. LR performance is known to be highly sensitive to the quality and relevance of input features. This study aims to quantitatively evaluate the impact of a strict correlation-based feature selection strategy on the accuracy of the Logistic Regression model. Using the "Diabetes Health Indicators" dataset (N=100,000), this study compares two scenarios:* (1) *a baseline LR model using all features (All Input) and* (2) *an optimized LR model using only a subset of features (including engineered features) that have a high absolute correlation with diabetes diagnosis (Correlated Input). The results demonstrate a significant performance improvement. The All Input baseline model achieved an accuracy of 80.45%, while the Correlated Input model achieved an accuracy of 85.67%. The measurement using AUC on the Correlated Input model was 0.93, which is higher than the baseline All Input model at 0.88. Correlation-based feature selection increased the predictive power of the Logistic Regression (LR) model by up to +5.22% by eliminating noise from irrelevant features. This optimized Logistic Regression offers a strong balance between enhanced accuracy and interpretability, which is essential for clinical applications.*

Abstrak- Diabetes Mellitus merupakan tantangan kesehatan global yang mendesak, di mana deteksi dini menjadi kunci intervensi yang efektif. *Machine learning* telah menunjukkan potensi besar dalam prediksi risiko diabetes. Di antara berbagai model, Regresi Logistik (LR) sering disukai dalam konteks medis karena interpretasinya yang tinggi, meskipun akurasinya seringkali tertinggal dari model *black-box* yang lebih kompleks. Kinerja LR diketahui sangat sensitif terhadap kualitas dan relevansi fitur input. Penelitian ini bertujuan untuk mengevaluasi secara kuantitatif dampak dari strategi seleksi fitur berbasis korelasi yang ketat terhadap akurasi model Regresi Logistik. Menggunakan *dataset Diabetes Health Indicators* (N=100.000), penelitian ini membandingkan dua skenario: (1) model LR *baseline* yang menggunakan semua fitur (*All Input*) dan (2) model LR yang dioptimalkan, yang hanya menggunakan *subset* fitur (termasuk fitur hasil rekayasa) yang memiliki korelasi absolut tinggi dengan diagnosis diabetes (*Correlated Input*). Hasil penelitian menunjukkan peningkatan kinerja yang signifikan. Model *baseline All Input* mencapai akurasi 80.45%, sedangkan model *Correlated Input* mencapai akurasi 85.67%. Pengukuran menggunakan AUC

pada model *correlated input* sebesar 0.93 lebih tinggi dibandingkan dengan model *baseline all input* sebesar 0.88. Seleksi fitur berbasis korelasi meningkatkan kekuatan prediktif model Regresi Logistik (LR) hingga +5.22% dengan menghilangkan *noise* fitur yang tidak relevan. Regresi Logistik yang dioptimalkan ini memberikan keseimbangan yang kuat antara akurasi yang ditingkatkan dan interpretasi yang esensial untuk aplikasi klinis.

Copyright © 2025 LPPM - STMIK IKMI Cirebon
This is an open access article under the CC-BY license

***Penulis Korespondensi**:*

Muhamad Syarif
Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika
Universitas Bina Sarana Informatika
Jl. Abdul Rahman Saleh No.18, Kec. Pontianak Tenggara, Kota Pontianak, Kalimantan Barat - Indonesia
Email: muhamad.mdx@bsi.ac.id

1. Pendahuluan

Diabetes *Mellitus* telah menjadi salah satu tantangan kesehatan global terbesar di abad ke-21 [1][2]. Sebagai penyakit kronis yang ditandai dengan kadar glukosa darah tinggi, diabetes dapat menyebabkan komplikasi serius yang memengaruhi berbagai organ tubuh, termasuk penyakit kardiovaskular, gagal ginjal, dan neuropati [1]. Besarnya beban penyakit ini, baik dari segi morbiditas, mortalitas, maupun biaya ekonomi, telah mendorong urgensi untuk pengembangan strategi pencegahan dan intervensi dini [2][3]. Salah satu pilar utama dalam upaya ini adalah identifikasi individu yang berisiko tinggi terkena diabetes sebelum penyakit tersebut bermanifestasi secara klinis [4].

Sejalan dengan kemajuan teknologi komputasi, metode *machine learning* (ML) telah menunjukkan potensi besar sebagai alat bantu untuk prediksi risiko penyakit kronis, termasuk diabetes [5]. Algoritma ML mampu menganalisis pola yang kompleks dan *non-linear* dari multidimensi data pasien mencakup faktor demografis, klinis, gaya hidup, dan riwayat kesehatan untuk menghasilkan model prediktif [5][6]. Berbagai model, seperti Regresi Logistik (*Logistic Regression*), *Random Forest*, dan *Gradient Boosting*, telah banyak diterapkan untuk tugas klasifikasi ini [6]. Di antara model-model tersebut, Regresi Logistik sering menjadi pilihan populer dalam konteks medis karena kemampuannya yang relatif sederhana dan hasil koefisiennya yang dapat diinterpretasi, sehingga memberikan wawasan tentang faktor risiko mana yang paling berpengaruh [7].

Kinerja model *machine learning* bergantung pada kualitas dan relevansi fitur [8]. Pendekatan *all input* (menggunakan semua fitur) berisiko menimbulkan *noise*, redundansi, dan kompleksitas komputasi, yang dapat menurunkan akurasi dan generalisasi [9][10]. Oleh karena itu, seleksi fitur adalah langkah krusial [9][10]. Penelitian ini berfokus pada analisis korelasi sebagai metode seleksi fitur fundamental [11]. Kemudian menguji hipotesis model Regresi Logistik (LR) yang hanya menggunakan fitur yang memiliki korelasi statistik kuat (termasuk hasil rekayasa fitur) dengan diagnosis diabetes akan menghasilkan performa yang lebih unggul dibandingkan dengan model *baseline all input*.

Penelitian ini secara spesifik bertujuan untuk membandingkan kinerja akurasi model Regresi Logistik pada dua skenario: pertama, menggunakan seluruh *set* fitur yang telah dibersihkan; dan kedua, menggunakan *set* fitur yang telah disaring secara ketat, hanya menyisakan fitur-fitur (asli maupun hasil rekayasa) yang menunjukkan korelasi absolut tinggi misalnya, dengan diagnosis diabetes. Model ini menunjukkan peningkatan kinerja yang cukup signifikan. Model *baseline All Input* mencapai akurasi 80.45% dan model *Correlated* Input mencapai akurasi 85.67%. Sedangkan untuk nilai AUC pada model *Correlated Input* sebesar 0.93 lebih tinggi dibandingkan dengan model *baseline All Input* sebesar 0.88. Hasil perbandingan ini diharapkan dapat memberikan wawasan mengenai efektivitas seleksi fitur berbasis korelasi sebagai strategi untuk mengoptimalkan model prediksi diabetes.

2. Studi Literatur

2.1. Prediksi Penyakit Diabetes

Diabetes *Mellitus* adalah penyakit metabolik kronis yang prevalensinya terus meningkat secara global dan menjadi beban kesehatan masyarakat yang signifikan [1]. Komplikasi serius dari diabetes yang tidak terkelola seperti penyakit kardiovaskular, retinopati, dan nefropati menimbulkan kebutuhan

mendesak untuk deteksi dini dan intervensi preventif. Dalam konteks ini, model prediksi risiko telah menjadi alat bantu yang krusial [4].

Secara historis, model-model ini bersifat statistika konvensional. Namun, dengan digitalisasi data kesehatan (EHR) dan ketersediaan dataset yang besar, teknik *machine learning* telah menunjukkan keunggulan. Berbagai penelitian telah berhasil menerapkan algoritma seperti *Support Vector Machines* (SVM), *Naïve Bayes*, dan *Artificial Neural Networks* (ANN) untuk mengklasifikasikan pasien berisiko tinggi dengan akurasi yang menjanjikan [12].

2.2. Peran *Logistic Regression* dalam Prediksi Medis

Di antara berbagai algoritma ML, Regresi Logistik (*Logistic Regression* - LR) tetap menjadi salah satu model yang paling banyak digunakan dalam penelitian medis. Meskipun model yang lebih kompleks (seperti *deep learning* atau *ensemble trees*) seringkali menawarkan akurasi yang sedikit lebih tinggi, Regresi Logistik memiliki keunggulan utama yang tidak tertandingi yaitu kemampuan interpretasi (daya tafsir) [13].

Dalam bidang medis, alasan mengapa sebuah prediksi dibuat seringkali sama pentingnya dengan apakah prediksi itu akurat [13][14]. Regresi Logistik adalah model *linier* probabilistik yang menghasilkan koefisien (dan *odds ratio*) untuk setiap fitur input [15]. Koefisien ini memungkinkan praktisi klinis untuk memahami secara kuantitatif faktor risiko mana (misalnya, *hba1c* atau *bmi*) yang memiliki dampak terbesar terhadap probabilitas diagnosis [15]. Karena transparansinya ini, LR sering digunakan sebagai benchmark atau model dasar yang kuat dalam studi klinis.

2.3. Signifikansi Seleksi Fitur untuk Model Linier

Kelemahan utama dari Regresi Logistik adalah sensitivitasnya terhadap *input* data. Kinerjanya dapat menurun secara signifikan ketika dihadapkan pada:

a. Multikolinearitas: Korelasi tinggi antar fitur prediktor.
b. Fitur Tidak Relevan (*Noise*): Fitur yang tidak memiliki hubungan statistik dengan variabel target.

Ketika dataset memiliki dimensionalitas tinggi (banyak fitur), model LR akan mencoba memberikan "bobot" (koefisien) pada setiap fitur, termasuk fitur yang tidak relevan. Hal ini dapat mengaburkan sinyal dari prediktor yang benar-benar penting dan menurunkan akurasi model secara keseluruhan.

Fenomena ini kontras dengan model *ensemble* berbasis pohon (seperti *Random Forest* atau Gradient *Boosting*), yang memiliki mekanisme seleksi fitur internal. Algoritma tersebut secara inheren akan mengabaikan fitur yang tidak relevan saat membangun pohon keputusan. Oleh karena itu, model linier seperti Regresi Logistik secara khusus mendapat manfaat besar dari langkah *preprocessing* berupa seleksi fitur yang eksplisit.

2.4. Seleksi Fitur Berbasis Korelasi sebagai Metode *Filter*

Metode seleksi fitur secara umum dapat dibagi menjadi tiga kategori: *filter, wrapper*, dan *embedded*. Metode filter adalah yang paling sederhana dan paling cepat secara komputasi. Metode ini mengevaluasi relevansi fitur berdasarkan karakteristik statistik data itu sendiri, sebelum model dilatih.

Salah satu metode filter yang paling umum adalah analisis korelasi, seringkali menggunakan koefisien korelasi *Pearson*. Pendekatan ini mengukur kekuatan hubungan linier antara setiap fitur independen dan variabel target misalnya, *diagnosed_diabetes*. Dengan menetapkan ambang batas misalnya peneliti dapat dengan cepat menyaring sejumlah besar fitur dan hanya mempertahankan fitur-fitur yang memiliki hubungan statistik terkuat dengan hasil yang ingin diprediksi.

3. Celah Penelitian

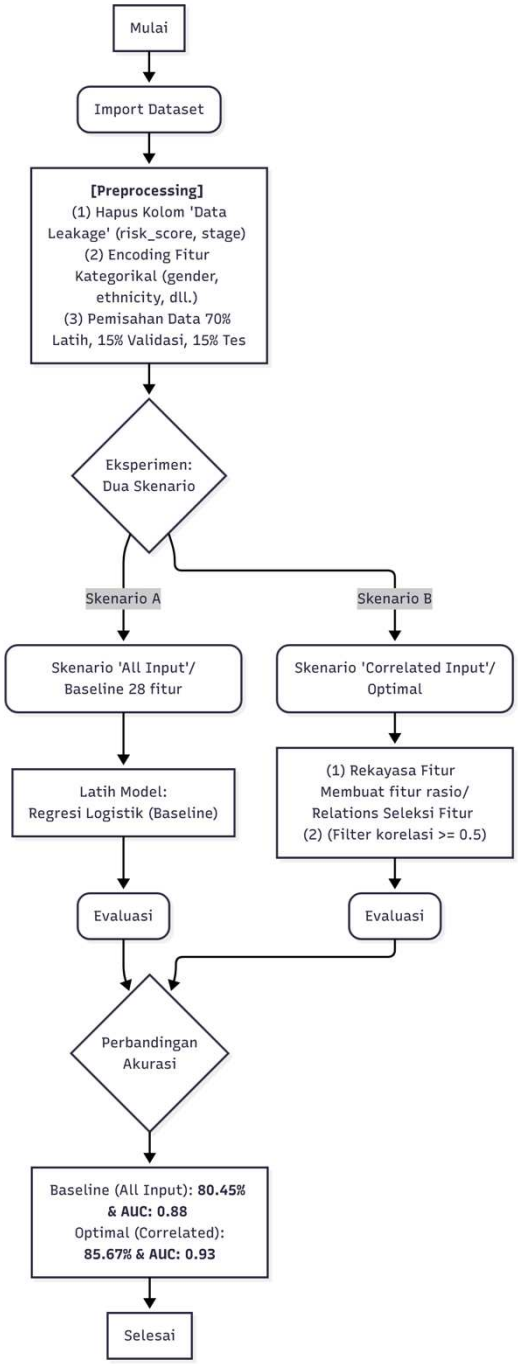
Meskipun banyak penelitian telah membandingkan Regresi Logistik melawan model *black-box* yang kompleks, masih terdapat fokus yang kurang pada optimalisasi model Regresi Logistik itu sendiri. Seringkali, model LR *baseline* (yang menggunakan semua fitur) disajikan dengan kinerja yang buruk, lalu diabaikan demi model yang lebih kompleks.

Celah penelitian yang ingin diisi oleh studi ini adalah untuk mengukur secara kuantitatif dampak langsung dari strategi seleksi fitur berbasis korelasi yang ketat terhadap kinerja Regresi Logistik. Penelitian ini berhipotesis bahwa model Regresi Logistik yang dilatih hanya dengan fitur-fitur yang sangat relevan (termasuk fitur hasil rekayasa) dapat mencapai peningkatan akurasi yang drastis, menjadikannya pilihan yang jauh lebih kompetitif yang tetap mempertahankan keunggulan interpretasinya.

4. Metode Penelitian

Metodologi penelitian ini dirancang untuk mengevaluasi efektivitas seleksi fitur berbasis korelasi dalam meningkatkan kinerja model Regresi Logistik untuk prediksi diabetes. Gambar 1 merupakan alur penelitian yang dibagi menjadi

beberapa tahapan utama yaitu akuisisi *dataset*, *preprocessing* data, desain eksperimen (termasuk rekayasa dan seleksi fitur), pemodelan, dan evaluasi.



Gambar 1. Tahapan Penelitian

4.1. Deskripsi *Dataset*

Penelitian ini menggunakan *dataset Diabetes Health Indicators Dataset* yang bersumber dari *platform* Kaggle. Gambar 2 merupakan *subset* dari *dataframe*. *Dataset* ini terdiri dari 100.000 sampel pasien dengan 31 atribut. Atribut-atribut ini mencakup data demografis (misalnya, *age*, *gender*, *ethnicity*), indikator gaya hidup (misalnya, *smoking_status*, *physical_activity_minutes_per_week*), riwayat kesehatan (misalnya, *family_history_diabetes*, *hypertension_history*), dan parameter klinis (misalnya, *bmi, glucose_fasting, hba1c*).

Variabel target adalah *diagnosed_diabetes*, sebuah variabel biner di mana "1" mengindikasikan pasien terdiagnosis diabetes dan "0" mengindikasikan tidak terdiagnosis.

Gambar 2. *Subset* dari *Dataframe*

4.2. *Preprocessing* Data

Tahap *preprocessing* data bertujuan untuk membersihkan dan mempersiapkan *dataset* agar siap digunakan untuk pemodelan.

a. Penanganan Kebocoran Data (*data leakage*): Kolom *diabetes_risk_score* dan *diabetes_stage* dihapus dari dataset. Kolom-kolom ini dianggap sebagai proksi atau hasil dari diagnosis diabetes itu sendiri, sehingga penggunaannya dalam prediksi akan menyebabkan kebocoran data dan hasil yang terlalu optimistis secara artifisial.
b. *Encoding* Fitur Kategorikal: Fitur non-numerik seperti *gender, ethnicity, education_level, income_level, employment_status*, dan *smoking_status* dikonversi menjadi representasi numerik menggunakan teknik label *encoding*.
c. Pemisahan Data (*Data Splitting*): *Dataset* yang telah bersih kemudian diacak (*shuffled*) dan dibagi menjadi tiga *subset* data yang berbeda:
   1) Data Latih (*Train Set*): 70% dari total data, digunakan untuk melatih model.
   2) Data Validasi (*Validation Set*): 15% dari total data, digunakan untuk mengevaluasi dan membandingkan kinerja model selama fase eksperimen.
   3) Data Uji (*Test Set*): 15% dari total data, digunakan untuk pengujian akhir pada model terbaik (meskipun dalam konteks perbandingan ini, fokus utama adalah pada set validasi).

4.3. Desain Eksperimen dan Seleksi Fitur

Eksperimen dirancang untuk membandingkan dua pendekatan dalam pemilihan fitur. Skenario 1: Model *All Input,* Model pertama yang berfungsi sebagai *baseline*, dilatih menggunakan seluruh fitur yang tersisa (28 fitur) setelah tahap *preprocessing data*. Skenario 2: Model *Correlated Input,* Model kedua dilatih menggunakan *subset* fitur yang telah melalui proses rekayasa dan seleksi fitur yang ketat:

a. Rekayasa Fitur (*Feature Engineering*): Fitur-fitur baru, yang disebut *Relations*, dibuat dengan menghitung rasio (pembagian) antara pasangan fitur numerik yang ada.
b. Analisis Korelasi: Koefisien korelasi *Pearson* dihitung antara setiap fitur (baik fitur asli maupun fitur *Relations* yang baru dibuat) dengan variabel target *diagnosed_diabetes*.
c. Seleksi Fitur: Hanya fitur-fitur yang menunjukkan korelasi absolut yang kuat (didefinisikan sebagai nilai) dengan target yang dipertahankan. Subset fitur yang telah disaring dan sangat berkorelasi ini kemudian disebut sebagai *set data Correlated Input*.

Berikut ini adalah *pseudocode* tahapan *Automated Feature Engineering*

```
INPUT:
    df_train, df_test, df_val   #dataset train,
    test, validation
    target = 'diagnosed_diabetes'
STEP 1: Ambil daftar kolom fitur
        colnames = semua kolom dari df_train
        hapus kolom target dari colnames
STEP 2: Buat fitur baru dari rasio antar kolom
    FOR setiap pasangan unik (col_i, col_j)
        dalam colnames:
        relation_col = df_train[col_i] /
        df_train[col_j]
        corr = | korelasi(relation_col,
        df_train[target]) |
        IF corr >= 0.5:
            relation_name = UUID unik
            simpan metadata pasangan (col_i,
            col_j) dengan nama relation_name
            tambahkan relation_col ke df_train
            dengan nama relation_name
            tambahkan fitur serupa ke df_test
            dan df_val
STEP 3: Seleksi fitur berdasarkan korelasi
        dengan target colnames_extended = semua
        kolom dari df_train_extended
    hapus kolom target dari colnames_extended
    FOR setiap col dalam colnames_extended:
        corr =|korelasi(df_train_extended[col],
        df_train_extended[target])|
        IF corr < 0.5:
            hapus col dari df_train_extended
            hapus col dari df_test_extended
            hapus col dari df_val_extended
STEP 4: Tampilkan bentuk akhir dataset
    print ukuran df_train_extended
    print ukuran df_test_extended
    print ukuran df_val_extended
OUTPUT:
    df_train_extended, df_test_extended,
    df_val_extended
    relations (metadata fitur baru)
```

4.4. Pemodelan Regresi Logistik

Algoritma *machine learning* yang menjadi fokus utama dalam perbandingan ini adalah Regresi Logistik (*Logistic Regression*). Model ini dipilih karena merupakan standar industri untuk masalah klasifikasi biner dan kemampuannya dalam memberikan koefisien yang dapat diinterpretasi. Dua model Regresi Logistik yang identik dilatih secara terpisah: satu pada *set* data *All Input* dan satu lagi pada *set* data *Correlated Input*.

4.5. Metrik Evaluasi

Kinerja dari kedua model Regresi Logistik (*All Input vs. Correlated Input*) dievaluasi dan dibandingkan menggunakan *set* data validasi. Metrik utama yang digunakan untuk perbandingan adalah Akurasi (*Accuracy*), yang didefinisikan sebagai:

$$Akurasi = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Sample}} \quad (1)$$

Di mana TP (*True Positive*) dan TN (*True Negative*) adalah jumlah prediksi yang benar untuk masing-masing kelas.

5. Hasil dan Pembahasan

Bagian ini menyajikan temuan kuantitatif dari eksperimen yang berfokus pada model Regresi Logistik. Tujuan utamanya adalah untuk mengevaluasi secara langsung dampak dari seleksi fitur berbasis korelasi terhadap kinerja model.

5.1. Hasil Kinerja Model Regresi Logistik

Model Regresi Logistik diuji dalam dua skenario fitur yang berbeda, seperti yang diuraikan dalam metodologi. Kinerja dievaluasi dengan metrik Akurasi pada *set* data validasi (15% dari total data). Hasil perbandingan disajikan pada Tabel 1 dan Tabel 2.

Tabel 1. Perbandingan Akurasi Model Regresi Logistik

Skenario Fitur	Jumlah Fitur (*Input*)	Akurasi (*Validation Set*)
All Input	28 Fitur	0.8045 (80.45%)
Correlated Input	7 Fitur*	**0.8567 (85.67%)**

Catatan: Jumlah fitur "Correlated Input" (7) diekstrak dari analisis notebook, yang menunjukkan dataset akhir memiliki 8 kolom (7 fitur + 1 target).

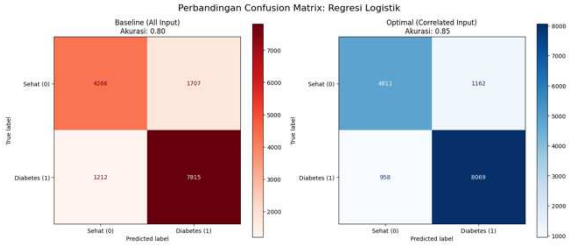
Tabel 2. Perbandingan Nilai AUC Model Regresi Logistik

Skenario Fitur	Jumlah Fitur (*Input*)	AUC
All Input	28 Fitur	0.8875
Correlated Input	7 Fitur*	**0.9316**

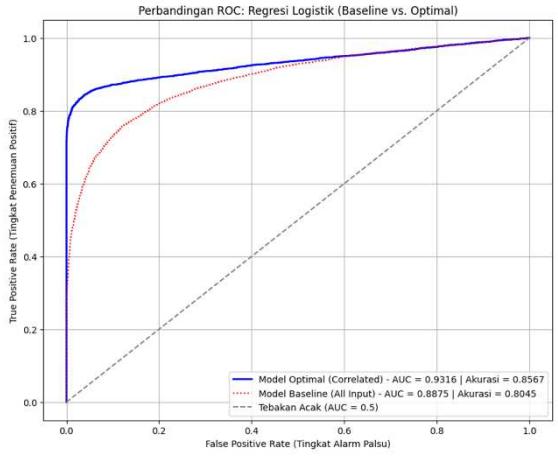
Confusion matrix digunakan untuk melihat di mana tepatnya model melakukan kesalahan deteksi atau "alarm palsu" atau justru "kebobolan" gagal mendeteksi orang sakit). Temuan ini diperkuat oleh *confusion matrix* pada Gambar 3, di mana model Optimal terbukti lebih aman secara klinis karena berhasil menekan angka *False Negative* (pasien diabetes yang tidak terdeteksi) dari 1.212 kasus menjadi hanya 958 kasus, sekaligus meningkatkan deteksi yang benar (*True Positive*) menjadi 8.069 pasien. Secara keseluruhan, seleksi fitur terbukti efektif mengurangi *noise*, meminimalkan kesalahan

diagnosis, dan meningkatkan akurasi prediksi secara substansial.

Model Regresi Logistik yang dilatih menggunakan *subset* fitur *Correlated Input* mencapai akurasi (85.67%), mengungguli model baseline *All Input* (80.45%) dengan selisih absolut sebesar +5.22%. Gambar 4 menunjukkan kurva ROC untuk perbandingan kedua skema penelitian. Visualisasi kurva ROC (*receiver operating characteristic*) di atas secara tegas mendemonstrasikan keunggulan model Regresi Logistik yang telah dioptimalkan melalui seleksi fitur. Garis biru, yang merepresentasikan model optimal (*correlated input*), secara konsisten melengkung lebih tinggi mendekati sudut kiri atas dibandingkan model *baseline* (garis merah), yang mengindikasikan kemampuan deteksi positif yang lebih baik dengan tingkat kesalahan (*false positive*) yang lebih rendah. Secara kuantitatif, superioritas ini dibuktikan dengan peningkatan nilai *area under the curve* (AUC) yang signifikan, dari 0.8875 pada model *baseline* menjadi 0.9316 pada model optimal. Peningkatan metrik AUC ini, yang berjalan lurus dengan kenaikan akurasi dari 80.45% menjadi 85.67%, mengonfirmasi bahwa eliminasi fitur *noise* tidak hanya mempertajam akurasi prediksi, tetapi juga secara drastis meningkatkan kemampuan diskriminatif model dalam membedakan pasien *diabetes* dan *non-diabetes* secara reliabel.



Gambar 3 Perbandingan Confusion Matrix



Gambar 4. Perbandingan ROC Regresi Logistik

5.2. Pembahasan

Temuan utama dari penelitian ini adalah bahwa seleksi fitur yang ketat dan berbasis korelasi secara drastis meningkatkan kinerja prediktif dari model Regresi Logistik. Pembahasan di bawah ini akan menguraikan implikasi dari hasil ini.

a. Dampak Kritis Seleksi Fitur pada Model *Linier*

Peningkatan akurasi sebesar 5.22% menyoroti kelemahan signifikan dari model Regresi Logistik (dan model linier lainnya) ketika dihadapkan pada data berdimensi tinggi yang mengandung *noise*. Model *All Input* (dengan 28 fitur) kemungkinan besar mengalami *overfitting* pada *noise* atau terdistorsi oleh fitur-fitur yang tidak memiliki hubungan prediktif (atau hanya hubungan yang sangat lemah) dengan diagnosis diabetes. Regresi Logistik bekerja paling baik ketika input-nya adalah prediktor yang kuat dan relevan. Dengan menyaring data dari 28 fitur menjadi hanya 7 fitur yang paling berkorelasi, kami secara efektif menghilangkan gangguan statistik. Hal ini memungkinkan algoritma untuk mengoptimalkan koefisiennya pada sinyal yang paling penting, menghasilkan batas keputusan yang jauh lebih akurat.

b. Efektivitas Rekayasa Fitur Rasio (*Relations*)

Penting untuk dicatat bahwa *set Correlated Input* tidak hanya terdiri dari fitur asli yang disaring, tetapi juga mencakup fitur-fitur hasil rekayasa (misalnya, *waist_to_hip_ratio* / *glucose_postprandial*). Fakta bahwa fitur-fitur rasio ini lolos dari saringan korelasi (memiliki korelasi) menunjukkan bahwa hubungan *non-linier* sederhana (dalam bentuk rasio) dapat menangkap interaksi biologis yang relevan. Misalnya, glucose_postprandial (glukosa setelah makan) mungkin berkorelasi sedang dengan diabetes. Namun, rasio antara *waist_to_hip_ratio* (indikator obesitas sentral) dan glukosa tersebut mungkin merupakan prediktor yang jauh lebih kuat, yang mencerminkan bagaimana tubuh mengelola beban glukosa dalam konteks obesitas. Keberhasilan ini menunjukkan rekayasa fitur yang digerakkan oleh hipotesis atau bahkan eksplorasi sangat berharga.

c. Implikasi: Keseimbangan antara Akurasi dan Interpretasi

Meskipun model *ensemble* yang kompleks seperti *CatBoost* atau *Random Forest* mungkin mencapai akurasi yang lebih tinggi, pendekatan yang divalidasi Regresi Logistik pada fitur yang sangat relevan memiliki keunggulan besar dalam hal interpretasi.

Model *Correlated Input* kami tidak hanya 85,67% akurat, tetapi juga dapat dijelaskan sepenuhnya. Model ini hanya didasarkan pada 7 faktor. Seorang dokter atau peneliti dapat dengan mudah memeriksa koefisien dari 7 fitur tersebut untuk memahami secara pasti faktor apa yang paling mendorong risiko diabetes menurut model (misalnya, *hba1c, glucose_postprandial*, dan rasio-rasio baru). Model *baseline All Input* dengan 28 fitur jauh lebih sulit untuk diinterpretasikan.

Dengan demikian, penelitian ini menunjukkan bahwa untuk aplikasi medis di mana *interpretability* (kemampuan untuk dijelaskan) sama pentingnya dengan akurasi, Regresi Logistik yang dikombinasikan dengan seleksi fitur berbasis korelasi yang ketat merupakan strategi metodologi yang sangat *valid* dan efektif.

6. Kesimpulan dan Saran

6.1. Kesimpulan

Penelitian ini mengevaluasi dampak dari strategi seleksi fitur berbasis korelasi terhadap model Regresi Logistik untuk prediksi diabetes. Hasilnya menunjukkan perbedaan kinerja yang sangat signifikan. Model *baseline* yang menggunakan semua fitur *All Input* hanya mencapai akurasi 80.45%. Sebaliknya, setelah menerapkan metode *Correlated Input* yang melibatkan rekayasa fitur dan filter korelasi yang ketat kinerja model melonjak drastis ke 85.67%, sebuah peningkatan absolut sebesar 5.22%.

Peningkatan substansial ini membuktikan bahwa Regresi Logistik adalah model yang sangat sensitif terhadap *noise* statistik, atau fitur-fitur yang tidak relevan dan berkorelasi lemah. Dengan menghilangkan gangguan ini, model dapat fokus pada sinyal prediktif yang paling kuat, sehingga akurasinya meningkat secara signifikan.

Pada akhirnya, model *Correlated Input* tidak hanya lebih akurat, tetapi juga menawarkan keunggulan signifikan dalam hal interpretasi. Model yang dihasilkan hanya dengan 7 fitur jauh lebih mudah untuk dijelaskan dan dianalisis oleh praktisi klinis dibandingkan dengan model *baseline* 28 fitur) Penelitian ini membuktikan bahwa strategi seleksi fitur yang ketat adalah metode yang sangat efektif untuk mengoptimalkan Regresi Logistik, menciptakan keseimbangan ideal antara akurasi dan kemampuan untuk dijelaskan dalam prediksi medis.

6.2. Saran

Berdasarkan temuan penelitian ini, terdapat implikasi praktis yang kuat bagi para peneliti yang menggunakan model linier dalam konteks medis. Sangat disarankan agar penggunaan pendekatan *brute-force All Input* dihindari. Sebaliknya, implementasi langkah seleksi fitur yang cermat harus dianggap sebagai bagian wajib dari alur kerja untuk mengoptimalkan model seperti Regresi Logistik. Selain itu, keberhasilan fitur rekayasa *Relations* menunjukkan potensi besar. Penelitian di masa depan disarankan untuk mengeksplorasi lebih lanjut interaksi antar variabel klinis, seperti rasio *BMI* terhadap glukosa atau tekanan darah terhadap *HBA1c*, yang mungkin dapat berfungsi sebagai prediktor tunggal yang lebih kuat.

Untuk langkah selanjutnya, validasi klinis terhadap model 7 fitur yang dioptimalkan (akurasi 85,67%) sangat penting. Model ini harus diuji menggunakan data prospektif dari populasi yang berbeda misalnya, rumah sakit lain untuk menguji kemampuannya di dunia nyata (generalisasi). Terakhir, penelitian ini berfokus pada metode *filter* korelasi *Pearson*. Studi di masa depan akan mendapat manfaat dari perbandingan efektivitas metode ini dengan teknik seleksi fitur lain, seperti metode *wrapper* (misalnya, RFE) atau *embedded* (regularisasi *L1/Lasso*), untuk melihat apakah akurasi Regresi Logistik dapat ditingkatkan lebih jauh lagi.

Daftar Pustaka

[1] World Health Organization, “Diabetes,” World Health Organization. Accessed: Nov. 03, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] K. L. Ong *et al.*, “Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021,” *The Lancet*, vol. 402, no. 10397, pp. 203–234, Jul. 2023, doi: 10.1016/S0140-6736(23)01301-6.

[3] D. J. Magliano, E. J. Boyko, and IDF Diabetes Atlas 10th edition scientific committee, Eds., *IDF DIABETES* ATLAS, 10th ed. Brussels: International Diabetes Federation, 2021. Accessed: Nov. 03, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK581934/

[4] A. D. A. P. P. Committee, “Prevention or Delay of Diabetes and Associated Comorbidities: Standards of Care in Diabetes-2024,” *Diabetes Care*, vol. 47, no. Supplement_1, pp. S43–S51, Dec. 2023, doi: 10.2337/dc24-S003.

[5] S. Shafi and G. Ansari, “Prediction of Diabetes Mellitus Using Machine Learning ,” in *Machine Learning and* Deep *Learning in Efficacy Improvement of Healthcare Systems*, 1st Edition., CRC Press Books, 2022. doi: 10.1201/9781003189053-5.

[6] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” *Procedia Comput Sci*, vol. 132, pp. 1578–1585, 2018, doi: https://doi.org/10.1016/j.procs.2018.05.122.

[7] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, “Interpretability of machine learning-based prediction models in healthcare,” *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 5, 2020, doi: https://doi.org/10.1002/widm.1379.

[8] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012, doi: 10.1145/2347736.2347755.
[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 1157–1182, Mar. 2003.
[10] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans Knowl Data Eng*, vol. 17, no. 4, pp. 491–502, 2005, doi: 10.1109/TKDE.2005.66.
[11] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.
[12] A. Al-Sideiri, Z. B. C. Cob, and S. B. M. Drus, "Machine Learning Algorithms for Diabetes Prediction: A Review Paper," in *Proceedings of the 2019 International* Conference *on Artificial Intelligence, Robotics and Control*, in AIRC 19. New York, NY, USA: Association for Computing Machinery, 2020, pp. 27–32. doi: 10.1145/3388218.3388231.
[13] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
[14] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD 15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 1721–1730. doi: 10.1145/2783258.2788613.
[15] S. Sperandei, "Understanding logistic regression analysis," *Biochem* Med *(Zagreb)*, vol. 24, pp. 12–18, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:7782682