

ANALISIS FAKTOR RISIKO STATUS HIV DENGAN MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION (STUDI KASUS: PUSKESMAS PAMULANG)

Farhan Yorah Pane^{1*}, Ari Mulyoto²,

^{1,2}Universitas Pamulang; Jl. Raya Puspiptek No. 46, Tangerang Selatan, telp/fax. (021) 7412566

Keywords:

HIV,
Logistic Regression,
Faktor Risiko,
Machine Learning,
Puskesmas Pamulang.

Correspondent Email:

farhanyorahpane@gmail.com

Abstrak. Kasus HIV di Puskesmas Pamulang tersebar pada berbagai Kelompok Populasi Kunci dengan proporsi kasus positif yang sangat kecil dibandingkan populasi umum, sehingga diperlukan pendekatan berbasis data untuk mengidentifikasi faktor risiko yang paling berpengaruh secara objektif. Penelitian ini bertujuan menganalisis faktor risiko status HIV serta menerapkan algoritma Logistic Regression untuk memprediksi potensi status HIV pasien. Penelitian menggunakan metode kuantitatif dengan pendekatan prediktif, memanfaatkan data sekunder pasien VCT/PITC periode 2024–2025 sebanyak 3.410 observasi. Data diproses melalui pembersihan, encoding, seleksi fitur, dan penanganan ketidakseimbangan kelas menggunakan SMOTE, sebelum dilatih dengan Logistic Regression yang dioptimasi melalui GridSearchCV. Model gabungan mencapai AUC-ROC 0,937, sedangkan model produksi dengan tiga variabel stabil (Jenis Kelamin, Kelompok Umur, dan Kelompok Risiko Tinggi) mencapai AUC-ROC 0,803 dengan recall 75%. Kelompok Risiko Tinggi terbukti sebagai faktor paling dominan ($OR = 17,29$; 95% CI 13,96–21,41; $p < 0,001$). Sistem deteksi risiko HIV berbasis web berhasil diimplementasikan dan seluruh skenario pengujian black-box dinyatakan berhasil, sehingga dapat mendukung skrining awal di Puskesmas Pamulang.



Copyright © 2026 The Author(s),
Published by Jurnal Informatika
dan Teknik Elektro Terapan. This
is an open-access article
distributed under the terms of the
Creative Commons Attribution-
NonCommercial 4.0 International
License (CC BY-NC 4.0).

Abstract. HIV cases at Puskesmas Pamulang are distributed across various Key Population Groups, with a markedly small proportion of positive cases compared to the general population, necessitating a data-driven approach to objectively identify the most influential risk factors. This study aims to analyze HIV status risk factors and apply the Logistic Regression algorithm to predict a patient's potential HIV status. The research employs a quantitative method with a predictive approach, using secondary data from VCT/PITC patients for the 2024–2025 period, comprising 3,410 observations. The data were processed through cleaning, encoding, feature selection, and class-imbalance handling using SMOTE, before being trained with Logistic Regression optimized through GridSearchCV. The combined model achieved an AUC-ROC of 0.937, while the production model with three stable variables (Gender, Age Group, and High-Risk Group) achieved an AUC-ROC of 0.803 with 75% recall. The High-Risk Group proved to be the most dominant factor ($OR = 17.29$; 95% CI 13.96–21.41; $p < 0.001$). A web-based HIV risk detection system was successfully implemented and all black-box testing scenarios were declared successful, so that it can support

1. PENDAHULUAN

Human Immunodeficiency Virus (HIV) masih menjadi salah satu masalah kesehatan masyarakat yang belum terselesaikan hingga saat ini. Virus ini menginfeksi dan menghancurkan limfosit yang memiliki antigen permukaan CD4 sehingga sistem kekebalan tubuh melemah secara progresif dan penderitanya menjadi rentan terhadap berbagai infeksi oportunistik [1]. Sifat virus yang mudah bermutasi serta memiliki banyak sub tipe menyebabkan infeksi baru terus bermunculan setiap tahun di berbagai wilayah dunia, termasuk Indonesia [1], [2].

Di Indonesia, penularan HIV sebagian besar terjadi pada kelompok usia produktif yang aktif secara sosial dan seksual. Hubungan seksual tanpa kondom, berganti-ganti pasangan, penggunaan jarum suntik secara bergantian, serta transmisi vertikal dari ibu ke anak merupakan jalur penularan utama yang telah banyak dilaporkan [1], [2]. Faktor sosial-ekonomi seperti tingkat pendidikan, akses informasi, dan pengetahuan mengenai kesehatan reproduksi turut memengaruhi besarnya paparan risiko pada masyarakat [3]. Di sisi lain, minimnya edukasi dalam lingkungan keluarga serta stigma yang masih melekat pada isu HIV membuat banyak individu tidak menyadari tingkat risiko yang mereka hadapi dan cenderung menunda pemeriksaan [4].

Secara epidemiologis, risiko infeksi HIV tidak tersebar merata pada seluruh populasi, melainkan terkonsentrasi pada kelompok populasi kunci tertentu [5]. Karakteristik ini membuka peluang bagi pendekatan berbasis data untuk memperkirakan risiko individu secara objektif. Sejumlah penelitian internasional telah mengembangkan risk-score HIV berbasis karakteristik sosiodemografi dan perilaku untuk menentukan sasaran pemeriksaan, antara lain di Kenya [6] dan Afrika Selatan [7], serta tinjauan sistematis atas model serupa di Amerika Serikat [8]. Model-model tersebut umumnya memprioritaskan sensitivitas, karena kegagalan mendeteksi kasus positif dinilai jauh lebih merugikan dibandingkan kelebihan rujukan pemeriksaan lanjutan.

Di Indonesia, pemanfaatan metode statistik dan machine learning untuk menganalisis faktor risiko HIV masih terbatas. Fathonah dan Mawarti [9] menerapkan regresi logistik untuk menganalisis hubungan faktor risiko dengan status HIV, sementara Aziz dan Sulistiani [10] membandingkan Random Forest, Support Vector Machine, dan regresi logistik pada data agregat populasi. Penerapan algoritma klasifikasi pada domain lain juga telah banyak dilakukan, mulai dari analisis sentimen berbasis Naive Bayes Classifier [11], klasifikasi risiko penyakit jantung [12], analisis kadar gula darah pasien diabetes [13], hingga identifikasi faktor prestasi akademik [14], [15]. Meskipun demikian, sebagian besar penelitian tersebut berhenti pada tahap analisis statistik dan belum diwujudkan menjadi perangkat bantu yang dapat digunakan langsung oleh petugas di fasilitas kesehatan.

Kesenjangan tersebut semakin nyata pada tingkat layanan kesehatan primer. Puskesmas secara rutin mencatat data pasien melalui Sistem Informasi HIV/AIDS (SIHA), namun data tersebut umumnya hanya dimanfaatkan untuk keperluan pelaporan administratif dan belum dijadikan dasar analisis faktor risiko. Padahal, data rekam medis HIV pada layanan primer memiliki karakteristik yang menantang secara metodologis, yaitu proporsi kasus positif yang sangat kecil sehingga menimbulkan ketidakseimbangan kelas yang ekstrem, serta struktur variabel kelompok populasi kunci yang bersifat multi-label dan saling tumpang tindih. Kebaruan penelitian ini terletak pada tiga hal: pertama, penggunaan data riil pasien VCT/PITC pada satu puskesmas periode 2024–2025 yang belum pernah dianalisis sebelumnya; kedua, penanganan eksplisit terhadap multikolinearitas dan quasi-complete separation melalui pemisahan antara model analitik dan model produksi; dan ketiga, integrasi model ke dalam sistem berbasis web yang dapat langsung digunakan sebagai alat bantu skrining awal.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menerapkan algoritma Logistic Regression dalam menganalisis faktor risiko yang memengaruhi status HIV pasien di

Puskesmas Pamulang, mengidentifikasi faktor yang paling dominan berdasarkan koefisien dan odds ratio model, serta mengimplementasikan model tersebut ke dalam sistem deteksi risiko HIV berbasis web.

2. TINJAUAN PUSTAKA

HIV merupakan retrovirus yang menyerang sistem kekebalan tubuh melalui kerusakan sel limfosit CD4. Tanpa penanganan yang tepat, infeksi dapat berkembang menjadi Acquired Immunodeficiency Syndrome (AIDS) ketika sistem imun telah kehilangan kemampuannya melawan infeksi [1]. Program pengendalian HIV nasional menempatkan kelompok populasi kunci—lelaki seks lelaki (LSL), wanita pekerja seks (WPS), waria, pengguna napza suntik (penasun), pasangan ODHIV, dan pelanggan pekerja seks—sebagai sasaran prioritas pemeriksaan karena prevalensi pada kelompok tersebut jauh di atas populasi umum [2], [5].

Logistic Regression merupakan metode analisis yang memodelkan hubungan antara satu variabel dependen kategorik dengan satu atau lebih variabel independen. Berbeda dengan regresi linear yang menghasilkan nilai kontinu, regresi logistik memetakan kombinasi linear prediktor ke dalam rentang probabilitas 0 hingga 1 melalui fungsi sigmoid, sehingga sesuai digunakan ketika variabel respons bersifat biner [9], [12]. Keunggulan utama metode ini adalah interpretabilitasnya: koefisien model dapat dikonversi menjadi odds ratio yang menyatakan besarnya perubahan kecenderungan suatu kejadian untuk setiap kenaikan satu satuan variabel prediktor, sehingga cocok digunakan pada penelitian faktor risiko penyakit [13].

Metode ini juga memiliki keterbatasan. Regresi logistik tidak menangkap hubungan non-linear yang kompleks tanpa transformasi variabel, dan estimasi koefisiennya menjadi tidak stabil ketika terdapat kategori dengan prevalensi sangat rendah atau prediktor yang saling kolinear. Kondisi ekstrem dari persoalan ini dikenal sebagai quasi-complete separation, yang ditandai dengan koefisien bermagnitudo sangat besar dan standard error yang melebar [14]. Penurunan performa juga dapat terjadi ketika distribusi kelas pada data sangat tidak seimbang.

Ketidakseimbangan kelas ditangani melalui Synthetic Minority Over-sampling Technique (SMOTE), yang membangkitkan sampel sintetis pada kelas minoritas berdasarkan interpolasi antar tetangga terdekat. Zhou et al. [16] menunjukkan bahwa integrasi SMOTE pada model regresi logistik secara konsisten meningkatkan nilai AUC dibandingkan pemodelan tanpa penanganan ketidakseimbangan kelas, sementara Aziz dan Sulistiani [10] menekankan penerapan SMOTE hanya pada data latih agar data uji tetap merepresentasikan kondisi lapangan. Pada tahap seleksi fitur, uji Chi-Square umum digunakan untuk menguji signifikansi hubungan antara fitur kategorik dan variabel target sebelum pemodelan klasifikasi risiko penyakit [17].

Aspek terakhir yang relevan adalah penentuan ambang klasifikasi. Pada konteks skrining kesehatan, ambang tidak selalu ditetapkan pada nilai default 0,5, melainkan diturunkan untuk mengejar sensitivitas yang lebih tinggi, sebagaimana dipraktikkan pada berbagai model risk-score HIV [6], [7]. Agar hasil analisis dapat dimanfaatkan secara langsung oleh pengguna akhir, model perlu diintegrasikan ke dalam antarmuka berbasis web yang mudah diakses [18].

3. METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan prediktif dan desain cross-sectional, yaitu data dikumpulkan dan dianalisis pada satu periode waktu tertentu tanpa pengamatan longitudinal. Lokasi penelitian adalah Puskesmas Pamulang, Kota Tangerang Selatan, dengan pelaksanaan analisis pada periode September 2025 hingga Juni 2026.

Data yang digunakan adalah data sekunder rekam medis pasien yang telah menjalani pemeriksaan HIV melalui layanan Voluntary Counseling and Testing (VCT) dan Provider-initiated Testing and Counseling (PITC) periode 2024–2025. Pengambilan sampel menggunakan teknik total sampling terhadap 3.424 rekam kunjungan. Sebanyak 14 observasi dieksklusi karena status HIV tercatat sebagai “Belum Tahu” sehingga tidak dapat digunakan sebagai label target yang definitif, menyisakan 3.410 observasi sebagai sampel

akhir. Pengumpulan data dilakukan melalui prosedur administrasi resmi dengan izin dari Puskesmas Pamulang, Dinas Kesehatan Kota Tangerang Selatan, dan Badan Kesatuan Bangsa dan Politik, serta seluruh data bersifat anonim sesuai etika penelitian kesehatan.

Variabel dependen adalah status HIV pasien yang dikategorikan secara biner, yaitu reaktif (bernilai 1) dan non-reaktif (bernilai 0). Variabel independen terdiri atas Kelompok Umur (kategorik ordinal: 0–14, 15–24, 25–49, dan ≥ 50 tahun), Jenis Kelamin (kategorik nominal), dan Kelompok Populasi Kunci yang tercatat pada SIHA. Variabel Gejala Klinis, Stadium Klinis WHO, dan Status ART tidak dilibatkan karena tidak tersedia pada rekam medis dan berpotensi menimbulkan data leakage, mengingat nilainya hanya tercatat pada pasien yang telah terkonfirmasi positif. Variabel Status Kawin dieksklusi karena tingkat data hilang mencapai 99,3%, sedangkan Jenis Layanan dieksklusi karena sebarannya sangat timpang (87,8% pada satu kategori) dengan satu kategori hanya memuat dua observasi.

Tahap preprocessing meliputi pembersihan data, encoding, dan seleksi fitur. Kolom dengan proporsi data hilang di atas 95% serta kolom identitas personal dihapus untuk kepentingan deidentifikasi. Encoding dilakukan secara berbeda sesuai karakteristik variabel: label encoding berurutan untuk Kelompok Umur, binary encoding untuk Jenis Kelamin, dan multi-label binarization untuk Kelompok Populasi Kunci. Pendekatan terakhir dipilih karena seorang pasien dapat tercatat pada lebih dari satu kelompok sekaligus, sebagaimana ditemukan pada 127 dari 3.410 observasi, sehingga tidak dapat direpresentasikan dengan one-hot encoding konvensional. Sebanyak 17 label mentah pada dokumentasi SIHA disederhanakan menjadi 15 kategori efektif setelah penggabungan dua label yang bermakna sama. Normalisasi tidak diterapkan karena seluruh variabel bersifat kategorik ordinal atau biner sehingga tidak terdapat perbedaan skala yang perlu dikoreksi.

Seleksi fitur dilakukan melalui dua pendekatan yang saling melengkapi. Variance Threshold dengan ambang 0,002 digunakan untuk menggabungkan kategori berprevalensi sangat rendah ke dalam satu kategori KP_Lainnya, sedangkan uji Chi-Square

digunakan untuk menguji signifikansi hubungan setiap fitur kategorik terhadap variabel target pada taraf $\alpha = 0,05$, mengikuti metodologi seleksi fitur pada penelitian klasifikasi risiko penyakit [17]. Selain itu, dibentuk variabel turunan Kelompok_Risiko_Tinggi sebagai nilai maksimum dari enam kategori populasi kunci berisiko tinggi (LSL, WPS, waria, penasun, pasangan ODHIV, dan pelanggan PS).

Data dibagi menjadi data latih dan data uji dengan proporsi 80:20 menggunakan fungsi `train_test_split` dengan parameter `stratify` dan `random_state = 42` agar proporsi kelas tetap terjaga. Penanganan ketidakseimbangan kelas dilakukan menggunakan SMOTE yang diterapkan hanya pada data latih setelah pembagian data, guna mencegah kebocoran informasi dari data uji [16], [10].

Logistic Regression memetakan kombinasi linear variabel prediktor ke dalam probabilitas melalui fungsi sigmoid. Kombinasi linear prediktor dirumuskan pada Persamaan (1), transformasi sigmoid pada Persamaan (2), dan konversi koefisien menjadi odds ratio pada Persamaan (3).

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

$$P(y = 1 | x) = 1 / (1 + e^{-z}) \quad (2)$$

$$OR_j = \exp(\beta_j) \quad (3)$$

$$P(y = 1 | x)$$

menyatakan probabilitas pasien berstatus HIV positif, β_0 adalah intercept, β_1 hingga β_n adalah koefisien masing-masing variabel prediktor, dan x_1 hingga x_n adalah nilai variabel prediktor. Optimasi hyperparameter dilakukan menggunakan GridSearchCV dengan skema 5-fold Stratified Cross Validation pada data latih hasil SMOTE, menguji kombinasi parameter kekuatan regularisasi $C \in \{0,01; 0,1; 1; 10; 100\}$ dan jenis penalti l1 serta l2 dengan solver liblinear, menggunakan AUC-ROC sebagai fungsi objektif pencarian.

Ambang klasifikasi tidak menggunakan nilai default 0,5, melainkan ditetapkan sebesar 0,35. Penetapan ini merupakan keputusan sadar untuk memprioritaskan sensitivitas, sesuai konteks skrining HIV yang menempatkan risiko mengabaikan kasus positif sebagai konsekuensi yang jauh lebih besar dibandingkan kelebihan rujukan pemeriksaan lanjutan [6], [7]. Evaluasi model menggunakan accuracy, precision, recall, F1-

score, dan AUC-ROC, dengan recall dan AUC-ROC sebagai metrik prioritas. Validitas statistik diperkuat melalui pemeriksaan Variance Inflation Factor (VIF), uji Wald beserta selang kepercayaan 95% odds ratio, serta estimasi performa menggunakan repeated stratified 5-fold cross-validation yang diulang sepuluh kali.

Model hasil analisis selanjutnya diimplementasikan ke dalam sistem berbasis web yang dibangun menggunakan PHP dan MySQL. Koefisien, intercept, dan ambang klasifikasi disimpan pada tabel basis data tersendiri sehingga model dapat diperbarui tanpa mengubah kode program. Pengujian sistem dilakukan menggunakan metode black-box testing terhadap seluruh fungsi yang dirancang.

:

4. HASIL DAN PEMBAHASAN

1. Seleksi Fitur Uji Chi-Square

Penerapan Variance Threshold menggabungkan tiga kategori berprevalensi sangat rendah, yaitu Penyakit Hepatitis (0,059%), Waria (0,059%), dan WBP (0,029%), ke dalam kategori KP_Lainnya. Selanjutnya, uji Chi-Square dilakukan terhadap seluruh fitur kandidat untuk menguji signifikansi hubungannya dengan status HIV. Hasil pengujian disajikan pada Tabel 1.

Tabel 1. Hasil Uji Chi-Square Fitur terhadap Status HIV ($\alpha = 0,05$)

Fitur	Chi2	p-value	Signifikan
Kelompok_Risiko_Tinggi	313,911	0,0000	Ya
KP_LSL	232,724	0,0000	Ya
KP_Pelanggan_PS	91,728	0,0000	Ya
KP_Pasangan_ODHIV	80,022	0,0000	Ya
KP_Ibu_Hamil	23,856	0,0000	Ya
KP_Populasi_Umum	15,998	0,0001	Ya
KP_Calon_Pengantin	7,662	0,0056	Ya
Jenis_Kelamin_Enc	6,686	0,0097	Ya
KP_WPS	6,226	0,0126	Ya
KP_Penyakit_T	3,786	0,0517	Tidak

B			
Kelompok_Umur_Enc	1,333	0,2483	Tidak
KP_Lainnya	0,059	0,8075	Tidak

Variabel Kelompok_Risiko_Tinggi memperoleh nilai Chi-Square tertinggi (313,911), diikuti KP_LSL (232,724) dan KP_Pelanggan_PS (91,728). Kelompok_Umur_Enc tidak signifikan secara statistik ($p = 0,2483$), namun tetap dipertahankan pada model akhir mengingat relevansi klinisnya yang kuat dalam literatur epidemiologi HIV [1], sedangkan KP_Lainnya dan kategori dengan p-value tinggi lainnya dikeluarkan dari model.

2. Penanganan Ketidak Seimbangan Kelas

Distribusi kelas target pada data latih sangat timpang, dengan kasus positif hanya sebesar 1,2%. Penerapan SMOTE menyamakan jumlah kelas minoritas dengan kelas mayoritas secara sintesis, sebagaimana disajikan pada Tabel 2.

Tabel 2. Distribusi Kelas Sebelum dan Sesudah SMOTE pada Data Latih

Kelas	Sebelum SMOTE	Sesudah SMOTE
Negatif (0)	2.696	2.696
Positif (1)	32	2.696

Penyeimbangan ini mencegah model bias terhadap kelas mayoritas pada tahap pelatihan. Data uji sengaja tidak diberi perlakuan SMOTE agar tetap merepresentasikan prevalensi sesungguhnya di lapangan, sehingga metrik evaluasi yang dihasilkan mencerminkan kondisi operasional nyata.

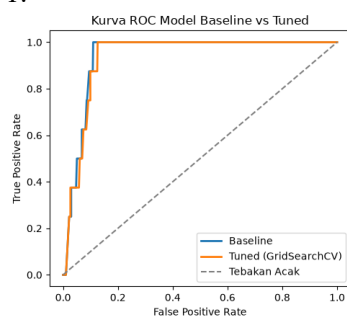
3. Pelatihan dan Perbandingan Model

Model baseline dilatih menggunakan parameter default, kemudian dioptimasi melalui GridSearchCV. Kombinasi parameter terbaik yang diperoleh adalah $C = 100$ dengan penalti l2 dan solver liblinear, menghasilkan AUC rata-rata 5-fold cross-validation sebesar 0,985 pada data latih. Perbandingan performa kedua model pada data uji disajikan pada Tabel 3.

Tabel 3. Perbandingan Model Baseline dan Model Tuned pada Data Uji

Model	Threshold	Accuracy	Precision	Recall	F1-Score	AUC
Baseline	0,50	94,43%	10,53%	50,00%	17,39%	0,942
Tuned (Grid Search CV)	0,35	93,11%	8,51%	50,00%	14,55%	0,937

Kedua model menghasilkan recall yang sama sebesar 50%. Nilai AUC model tuned (0,937) sedikit lebih rendah di bandingkan baseline (0,942) pada data uji, meskipun AUC cross-validation pada data latih jauh lebih tinggi (0,985). Kesenjangan ini mengindikasikan bahwa model tuned lebih menyesuaikan diri terhadap pola data latih, sementara performanya pada data uji tetap stabil dan sebanding dengan baseline. Mengingat data uji hanya memuat delapan kasus positif, selisih satu atau dua kasus berdampak besar terhadap nilai metrik sehingga perbedaan antar model perlu diinterpretasikan secara hati-hati. Model tuned dengan ambang 0,35 dipilih sebagai model akhir atas pertimbangan sensitivitas klinis. Kurva ROC kedua model ditampilkan pada Gambar 1.



Gambar 1. Kurva ROC Model Baseline dan Model Tuned

Kedua kurva berada jauh di atas garis diagonal yang merepresentasikan tebakan acak, menandakan bahwa variabel Kelompok Umur, Jenis Kelamin, dan Kelompok Populasi Kunci secara kolektif memiliki daya diskriminasi yang kuat terhadap status HIV pasien.

4. Evaluasi Model Akhir

Rincian prediksi model akhir pada data uji disajikan dalam bentuk confusion matrix pada

Tabel 4. Confusion Matrix Model Akhir (threshold 0,35)

	Prediksi Negatif (0)	Prediksi Positif (1)
Aktual Negatif (0)	631 (TN)	43 (FP)
Aktual Positif (1)	4 (FN)	4 (TP)

Model memperoleh accuracy sebesar 93,11%. Angka ini tampak tinggi, namun didominasi oleh keberhasilan mendeteksi 631 dari 674 kasus negatif; pada kelas positif model hanya berhasil mendeteksi empat dari delapan kasus. Temuan ini menegaskan bahwa accuracy tidak memadai sebagai indikator tunggal pada klasifikasi dengan distribusi kelas yang timpang. Precision sebesar 8,51% berarti dari sekitar dua belas pasien yang diprediksi berpotensi, hanya satu yang benar-benar positif. Nilai ini merupakan konsekuensi langsung dari penurunan ambang klasifikasi dan prevalensi kasus positif yang sangat rendah pada populasi. Dalam konteks skrining awal, precision rendah masih dapat diterima selama setiap hasil yang diklasifikasikan berpotensi ditindaklanjuti dengan pemeriksaan laboratorium, bukan diperlakukan sebagai diagnosis final — prinsip yang sama dianut model risk-score HIV pada penelitian terdahulu [6], [7].

5. Identifikasi Faktor Dominan

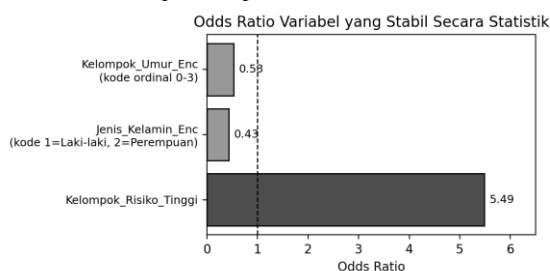
Faktor dominan diidentifikasi dari koefisien regresi yang telah dikonversi menjadi odds ratio. Ringkasan koefisien dan odds ratio pada model gabungan disajikan pada Tabel 5.

Tabel 5. Koefisien dan Odds Ratio Variabel pada Model Gabungan

Variabel	Koefisien	Odds Ratio	Keterangan
Kelompok_Risiko_Tinggi	1,703	5,488	Stabil, dapat diinterpretasikan
Jenis_Kelamin_Enc	-0,840	0,432	Stabil, dapat diinterpretasikan
Kelompok_Umur_Enc	-0,639	0,528	Stabil, tidak signifikan
KP_LSL	-3,159	0,0425	Kolinear dengan Kelompok_Risiko_Tinggi

			gi
KP_Pelanggan_PS	-3,937	0,0195	Kolinear dengan Kelompok_Risiko_Tinggi
KP_Pasangan_ODHIV	-5,276	0,0051	Kolinear dengan Kelompok_Risiko_Tinggi
KP_Ibu_Hamil	-10,728	0,00002	Terindikasi separation

Sebagian besar kategori Kelompok Populasi Kunci memiliki koefisien bermagnitudo sangat besar dengan odds ratio mendekati nol. Nilai ekstrem ini bukan mencerminkan efek protektif yang sesungguhnya, melainkan artefak statistik akibat multikolinearitas. Variabel Kelompok_Risiko_Tinggi dibentuk sebagai nilai maksimum dari enam kategori yang lima di antaranya turut dimasukkan sebagai variabel terpisah dalam model yang sama; pengecekan lanjutan menunjukkan nilai keduanya identik pada 99,94% observasi, sehingga model mengalami kondisi mendekati quasi-complete separation. Odds ratio kategori populasi kunci individual karenanya tidak dapat diinterpretasikan secara langsung sebagai ukuran asosiasi yang valid. Ketiga variabel yang tidak terlibat dalam relasi kolinear tersebut ditampilkan pada Gambar 2



Gambar 2. Odds Ratio Variabel yang Stabil Secara Statistik

Kelompok_Risiko_Tinggi merupakan faktor dominan dengan odds ratio 5,49, yang berarti pasien yang tergabung dalam kelompok populasi berisiko tinggi memiliki kecenderungan berstatus HIV positif sekitar 5,5 kali lebih besar dibandingkan pasien di luar kelompok tersebut setelah mengendalikan variabel lain. Jenis_Kelamin_Enc menghasilkan odds ratio 0,432 untuk perempuan relatif terhadap laki-laki, setara

dengan odds pasien laki-laki sekitar 2,3 kali lebih tinggi. Kelompok_Umur_Enc menghasilkan odds ratio 0,528 yang mengindikasikan kecenderungan menurun pada kelompok usia lebih tua, namun hubungan ini tidak signifikan secara statistik sehingga arah maupun besaran efeknya hanya dapat diperlakukan sebagai kecenderungan awal.

6. Model Produksi dan Validasi Statistik

Karena model gabungan tidak menghasilkan odds ratio yang valid untuk kategori individual, model dilatih ulang menggunakan hanya tiga variabel yang terbukti stabil, yaitu Jenis Kelamin, Kelompok Umur, dan Kelompok Risiko Tinggi, tanpa menyertakan kategori penyusunnya sebagai variabel terpisah. Hasil estimasi beserta uji Wald dan selang kepercayaan 95% disajikan pada Tabel 6.

Tabel 6. Koefisien, Odds Ratio, dan Uji Wald Model Produksi

Variabel	Koefisien	Odds Ratio	p-value	CI 95% bawah	CI 95% atas
Jenis_Kelamin_Enc	-2,1489	0,117	0,0000	0,097	0,141
Kelompok_Umur_Enc	0,0472	1,048	0,5660	0,892	1,232
Kelompok_Risiko_Tinggi	2,8500	17,287	0,0000	13,957	21,412

Kelompok_Risiko_Tinggi (OR = 17,29; 95% CI 13,96–21,41; $p < 0,001$) dan Jenis_Kelamin_Enc (OR = 0,117; 95% CI 0,097–0,141; $p < 0,001$) signifikan secara statistik dengan selang kepercayaan yang sempit dan tidak melewati angka satu, sedangkan Kelompok_Umur_Enc (OR = 1,048; $p = 0,566$) tidak signifikan dan selang kepercayaannya melewati satu — konsisten dengan hasil uji Chi-Square pada Tabel 1. Odds ratio Kelompok_Risiko_Tinggi pada model ini jauh lebih besar dibandingkan pada model gabungan karena variabel tersebut kini menyerap seluruh sinyal dari kategori penyusunnya tanpa terbagi. Sebagai perbandingan, estimasi ulang model gabungan

menggunakan Maximum Likelihood gagal mencapai konvergensi dengan standard error yang melebar hingga tak berhingga pada beberapa kategori berprevalensi rendah, yang merupakan bukti statistik langsung dari kondisi quasi-complete separation.

Pemeriksaan Variance Inflation Factor mengonfirmasi diagnosis tersebut secara kuantitatif, sebagaimana disajikan pada Tabel 7.

Tabel 7. Lima Nilai VIF Tertinggi pada Model Gabungan

Variabel	VIF
Kelompok_Risiko_Tinggi	37,79
KP_LSL	19,73
KP_Ibu_Hamil	8,81
KP_WPS	6,76
KP_Calon_Pengantin	6,10

Nilai VIF Kelompok_Risiko_Tinggi sebesar 37,79 dan KP_LSL sebesar 19,73 berada jauh di atas ambang batas umum, yaitu VIF lebih dari 10 sebagai indikasi multikolinearitas serius. Sebaliknya, pemeriksaan yang sama pada model produksi menghasilkan VIF mendekati satu untuk seluruh variabel (Jenis_Kelamin_Enc = 1,175; Kelompok_Umur_Enc = 1,003; Kelompok_Risiko_Tinggi = 1,179), menegaskan bahwa model produksi bebas dari masalah multikolinearitas.

Mengingat data uji pada satu kali pembagian hanya memuat delapan kasus positif, performa model produksi divalidasi ulang menggunakan repeated stratified 5-fold cross-validation sebanyak sepuluh pengulangan dengan SMOTE diterapkan ulang pada setiap fold. Perbandingan hasil kedua skema evaluasi disajikan pada Tabel 8.

Metrik	Data uji (1 split)	Rata-rata 10×5-fold	Std. Deviasi	Rentan g 95%
AUC-ROC	0,803	0,878	0,081	0,700–0,987
Recall	75,00 %	82,25 %	15,02 %	52,81–100,00 %
Precision	4,29%	5,97%	2,23 %	3,74–11,00%
F1-Score	8,11%	10,97 %	3,60 %	7,12–19,09%

Accurac y	80,06 %	82,72 %	5,62 %	77,27–93,24%
-----------	---------	---------	--------	--------------

Hasil cross-validation berulang menunjukkan AUC-ROC rata-rata 0,878 dan recall rata-rata 82,25%, sedikit lebih tinggi namun konsisten dengan hasil pada satu kali pembagian data yang berada di dalam rentang tersebut. Precision tetap rendah pada seluruh fold, yang menegaskan bahwa nilai tersebut merupakan konsekuensi konsisten dari prevalensi kasus positif yang sangat kecil, bukan artefak dari satu pembagian data tertentu. Standar deviasi yang cukup besar, terutama pada recall dan precision, menunjukkan pentingnya menafsirkan angka performa sebagai estimasi titik dari rentang yang lebih luas.

7. Pembahasan

Temuan bahwa keanggotaan pada kelompok populasi berisiko tinggi merupakan faktor paling dominan sejalan dengan penelitian Fathonah dan Mawarti [9] yang menemukan perilaku seksual berisiko dan riwayat infeksi menular seksual sebagai variabel paling signifikan terhadap status HIV, serta dengan temuan Slaymaker et al. [5] mengenai konsentrasi risiko pada kelompok populasi kunci. Perbedaanannya terletak pada tingkat granularitas variabel: penelitian terdahulu menggunakan variabel perilaku yang dikumpulkan melalui kuesioner, sedangkan penelitian ini bekerja pada variabel kategori populasi yang tersedia dalam pencatatan rutin layanan. Hal ini menunjukkan bahwa data administratif yang sudah tersedia di puskesmas pun sudah cukup untuk menghasilkan model dengan daya diskriminasi yang baik, tanpa memerlukan instrumen pengumpulan data tambahan.

Nilai AUC-ROC yang diperoleh, yaitu 0,937 pada model gabungan dan 0,878 pada rata-rata cross-validation model produksi, sebanding dengan performa model risk-score HIV pada penelitian internasional [6], [7] yang umumnya berada pada rentang 0,70 hingga 0,85. Namun demikian, perbandingan ini perlu diperlakukan dengan hati-hati karena penelitian tersebut menggunakan populasi dan variabel perilaku yang jauh lebih kaya. Precision yang rendah pada penelitian ini merupakan konsekuensi struktural dari

prevalensi kasus positif sebesar 1,2%, bukan kelemahan algoritma; pola serupa juga dilaporkan pada penelitian klasifikasi risiko HIV berbasis data agregat [10].

Kontribusi metodologis penelitian ini terletak pada penanganan eksplisit terhadap quasi-complete separation yang timbul ketika variabel turunan dimasukkan bersama kategori penyusunnya ke dalam satu model. Persoalan ini jarang dibahas secara terbuka pada penelitian regresi logistik di bidang kesehatan, padahal berdampak langsung pada validitas interpretasi odds ratio. Strategi pemisahan antara model analitik, yang digunakan untuk memetakan hubungan bivariat seluruh kategori, dan model produksi, yang digunakan untuk prediksi per pasien, terbukti menghasilkan estimasi yang stabil dengan VIF mendekati satu dan selang kepercayaan yang sempit. Pendekatan ini dapat diadopsi pada penelitian sejenis yang menggunakan variabel komposit.

Secara praktis, hasil penelitian ini berimplikasi pada penentuan prioritas skrining di layanan kesehatan primer. Dengan odds ratio 17,29, keanggotaan pada kelompok berisiko tinggi dapat dijadikan penanda utama dalam penjarangan sasaran pemeriksaan VCT/PITC, dan model tiga variabel yang stabil ini berpotensi diintegrasikan ke dalam sistem informasi puskesmas sebagai alat bantu skrining awal, bukan sebagai pengganti diagnosis laboratorium. Keterbatasan penelitian ini terletak pada jumlah kasus positif yang sangat sedikit (40 dari 3.410 observasi), cakupan data yang terbatas pada satu fasilitas kesehatan, serta ketiadaan variabel perilaku dan klinis yang pada penelitian lain terbukti berpengaruh kuat. Generalisasi hasil ke wilayah lain karenanya memerlukan validasi eksternal dengan data yang lebih luas.

5. KESIMPULAN

1. Algoritma Logistic Regression berhasil diterapkan untuk menganalisis faktor risiko status HIV pasien VCT/PITC di Puskesmas Pamulang menggunakan 3.410 observasi periode 2024–2025. Melalui tahapan preprocessing, multi-label encoding, penanganan ketidakseimbangan kelas dengan SMOTE, dan optimasi hyperparameter menggunakan GridSearchCV, model akhir

dengan ambang klasifikasi 0,35 memperoleh AUC-ROC sebesar 0,937 pada data uji.

2. Faktor paling dominan dan stabil secara statistik adalah keanggotaan pada Kelompok Risiko Tinggi, dengan odds ratio 5,49 pada model gabungan dan 17,29 pada model produksi tiga variabel yang bebas multikolinearitas (95% CI 13,96–21,41; $p < 0,001$). Jenis kelamin laki-laki juga berhubungan dengan kecenderungan status HIV positif yang lebih tinggi dibandingkan perempuan, sedangkan kelompok umur tidak signifikan secara statistik.

3. Odds ratio kategori Kelompok Populasi Kunci individual pada model gabungan tidak dapat diinterpretasikan secara langsung akibat multikolinearitas dengan variabel turunan Kelompok Risiko Tinggi. Untuk penelitian sejenis, variabel komposit sebaiknya tidak digunakan bersamaan dengan kategori penyusunnya dalam satu model yang sama.

4. Model produksi tiga variabel yang dihasilkan bersifat stabil, ringkas, dan layak dijadikan dasar bagi pengembangan alat bantu skrining awal di layanan kesehatan primer, meskipun precision-nya masih rendah sebagai konsekuensi struktural dari rendahnya prevalensi kasus positif, sehingga setiap hasil deteksi tetap perlu ditindaklanjuti dengan pemeriksaan laboratorium.

5. Pengembangan selanjutnya disarankan mengumpulkan data pada rentang waktu yang lebih panjang agar jumlah kasus positif bertambah, melakukan validasi eksternal pada puskesmas lain, mengeksplorasi algoritma alternatif atau penambahan variabel perilaku sepanjang tidak menimbulkan data leakage, serta mengintegrasikan model ke dalam sistem informasi puskesmas sebagai tindak lanjut penelitian.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak terkait yang telah memberi dukungan terhadap penelitian ini.

DAFTAR PUSTAKA

- [1] D. Rohmatullailah and D. Fikriyah, "Faktor risiko kejadian HIV pada kelompok usia produktif di Indonesia," *Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan*,

- vol. 2, no. 1, pp. 45–59, 2021, doi: 10.7454/bikfokes.v2i1.1022.
- [2] K. K. Munawar and A. I. Purnamasari, "Implementasi algoritma K-Means clustering pada klasterisasi kasus HIV di Jawa Barat," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 7, no. 2, 2023.
- [3] I. A. Ismail, A. Febriyanti, D. Alif, A. Namira, S. Wicaksono, and R. S. Nadeak, "Hubungan pengetahuan dan sikap terhadap pencegahan HIV/AIDS pada remaja," Jurnal Ilmiah Kesehatan Masyarakat, 2022.
- [4] L. Khairani, W. Masitah, S. Hajar, and M. Ibrahim, "Edukasi pencegahan dan pengendalian HIV/AIDS melalui peran perempuan dalam keluarga," Jurnal Pengabdian Masyarakat, pp. 82–90, 2022.
- [5] E. Slaymaker, C. Calvert, M. Marston, K. Risher, and J. W. Imai-Eaton, "Individual and population-level risk factors for new HIV infections among adults in Eastern and Southern Africa," Nature Communications, vol. 17, art. 1195, 2026, doi: 10.1038/s41467-025-67966-0.
- [6] H. Muttai et al., "Development and validation of a sociodemographic and behavioral characteristics-based risk-score algorithm for targeting HIV testing among adults in Kenya," AIDS and Behavior, vol. 25, no. 2, pp. 297–310, 2021, doi: 10.1007/s10461-020-02962-7.
- [7] R. Moyo, D. Govindasamy, P. S. Nyasulu, and S. Manda, "A prediction risk score for HIV among adolescent girls and young women in South Africa: identifying those in need of HIV pre-exposure prophylaxis," African Journal of AIDS Research, 2023, doi: 10.1080/25787489.2023.2221377.
- [8] M. Patel et al., "HIV risk score and prediction model in the United States: a scoping review protocol," medRxiv, 2023, doi: 10.1101/2023.12.18.23300171.
- [9] I. Fathonah and L. Mawarti, "Analisis faktor risiko terhadap status HIV dengan menggunakan pendekatan regresi logistik," Jurnal Sains dan Teknologi Kesehatan, pp. 175–181, 2022.
- [10] A. Aziz and H. Sulistiani, "Klasifikasi risiko HIV berbasis data agregat populasi menggunakan Random Forest, SVM, dan Logistic Regression," Jurnal Algoritma, vol. 23, no. 1, pp. 1732–1743, 2026, doi: 10.33364/algoritma/v.23-1.3492.
- [11] D. Duei Putri, G. F. Nama, and W. E. Sulistiono, "Analisis sentimen kinerja Dewan Perwakilan Rakyat (DPR) pada Twitter menggunakan metode Naive Bayes Classifier," Jurnal Informatika dan Teknik Elektro Terapan (JITET), vol. 10, no. 1, Jan. 2022.
- [12] T. Sianturi, L. N. Siahaan, and S. D. Aritonang, "Analisis aspek-aspek yang memengaruhi penyakit jantung menggunakan model regresi logistik biner menggunakan software RStudio," Jurnal Statistika dan Komputasi, vol. 24, no. 1, pp. 73–83, 2025.
- [13] M. Mayawi, N. Nurhayati, T. Talib, A. W. Bustan, and N. S. Laamena, "Ordinal logistic regression analysis of factors affecting the blood sugar levels of diabetes mellitus patients," Pattimura International Journal of Mathematics (PIJMath), vol. 2, no. 1, pp. 33–42, 2023, doi: 10.30598/pijmathvol2iss1pp33-42.
- [14] Nursalam, Q. Hamzah, Baharuddin, A. Kusumayanti, and A. Sriyanti, "Identify factors affecting student achievement using the binary logistic regression analysis," MaPan: Jurnal Matematika dan Pembelajaran, vol. 11, no. 2, pp. 368–385, 2023.
- [15] F. Aprilia, R. A. Anggraini, and Y. D. Putri, "Prediksi kelulusan siswa dengan algoritma pembelajaran mesin: aplikasi regresi linear dan logistik pada faktor-faktor pendidikan," Jurnal Riset dan Teknologi, vol. 3, no. 1, pp. 55–64, 2025, doi: 10.25181/rt.v3i1.3897.
- [16] R. Zhou et al., "Prediction model for infectious disease health literacy based on synthetic minority oversampling technique algorithm," Computational and Mathematical Methods in Medicine, vol. 2022, art. 8498159, 2022, doi: 10.1155/2022/8498159.
- [17] D. Kuswandani and R. Umar, "Comparison of filter and wrapper feature selection methods for heart disease risk classification using k-Nearest Neighbors (k-NN)," Sistemasi: Jurnal Sistem Informasi, vol. 15, pp. 871–885, 2026.
- [18] G. M. Rafsyjanatni and A. Fauzi, "Pengembangan media promosi berbasis website dengan aplikasi CMS WordPress untuk menarik minat beli konsumen pada UMKM Aneka Batik Pamekasan," Jurnal Teknologi Informasi, vol. 8, no. 1, pp. 1–6, 2022.