Jurnal Integrasi Teknologi dan Inovasi
Volume 1, Nomor 2, 2026 pp. 77- 86
ISSN: 3110-8601
Open Access: https://journal.pintarologimedia.com/index.php/integrasi

Analisis Kelayakan Penerima Program Keluarga Harapan (PKH) di Kecamatan Pujud Kabupaten Rokan Hilir Menggunakan Logistic Regression

Seleksi[1] , Candra Adipradana[2], Imam Taufiq[3], Afifah Nurull Izzati[4]

[1,2,3,4] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Kahuripan Kediri, Kediri, Indonesia

ARTICLE INFO

Article history:
Received February 20, 2026
Revised February 23, 2026
Accepted Marh 30, 2026
Available online May 25, 2026

Kata Kunci:
PKH; Kelayakan Penerima; Bantuan Sosial; Logistic Regression; Klasifikasi

Keywords:
Please PKH, Recipient Eligibility, Social Assistance, Logistic Regression, Classification

ABSTRAK

Program Keluarga Harapan (PKH) merupakan bantuan sosial bersyarat untuk meningkatkan kesejahteraan keluarga prasejahtera, namun di lapangan masih ditemukan ketidaktepatan sasaran yang memicu inclusion error dan exclusion error. Penelitian ini menganalisis kelayakan penerima PKH di Kecamatan Pujud, Kabupaten Rokan Hilir menggunakan metode Logistic Regression berbasis data sosial-ekonomi. Data dikumpulkan dari 200 kepala keluarga dengan variabel pendapatan bulanan, jumlah tanggungan, pendidikan, pekerjaan, kondisi rumah, kepemilikan kendaraan, serta akses fasilitas dasar. Tahap praproses meliputi pembersihan atribut non-relevan, imputasi nilai hilang, one-hot encoding untuk fitur kategorikal, standardisasi fitur numerik, dan pembagian data 80:20. Hasil menunjukkan variabel pendapatan, jumlah tanggungan, dan akses fasilitas dasar berpengaruh dominan terhadap kelayakan. Model memberikan performa klasifikasi yang baik pada data uji dan mendukung seleksi penerima bantuan yang lebih objektif, transparan, dan tepat sasaran.

ABSTRACT

The Family Hope Program (PKH) is a conditional social assistance scheme intended to improve the welfare of underprivileged households; however, inaccurate targeting still occurs in practice, leading to inclusion and exclusion errors. This study analyzes PKH recipient eligibility in Pujud Subdistrict, Rokan Hilir Regency, using a Logistic Regression approach based on socioeconomic data. Data were collected from 200 household heads and included monthly income, number of dependents, education level, employment status, housing condition, vehicle ownership, and access to basic facilities. Data preprocessing involved removing non-relevant administrative attributes, imputing missing values, applying one-hot encoding for categorical variables, standardizing numerical features, and splitting the dataset into training and testing sets (80:20). The results indicate that income, number of dependents, and access to basic facilities are the most influential factors in determining eligibility. The Logistic Regression model demonstrates good classification performance on the test set and can support a more objective, transparent, and better-targeted mechanism for PKH recipient selection in the studied area.

1. PENDAHULUAN

Kemiskinan masih menjadi tantangan pembangunan di Indonesia sehingga kebijakan perlindungan sosial diperlukan untuk menurunkan kerentanan rumah tangga dan memperkuat akses layanan dasar. Salah satu intervensi yang dijalankan pemerintah adalah Program Keluarga Harapan (PKH) sebagai bantuan sosial bersyarat yang menekankan peningkatan akses kesehatan dan pendidikan bagi keluarga miskin/prasejahtera. Dalam perspektif kebijakan, PKH juga dapat dipahami sebagai instrumen pengurangan kemiskinan yang bekerja melalui relasi sosial dan mekanisme pendampingan agar perubahan perilaku dan akses layanan terjadi secara berkelanjutan (Anggraini, 2024). Secara operasional, pelaksanaan PKH diatur melalui pedoman teknis yang memuat ketentuan komponen, mekanisme, dan sasaran program

*Corresponding author
E-mail addresses: seleksi@students.kahuripan.ac.id (Seleksi)

(Kementerian Sosial RI, 2021). Meski demikian, pada level implementasi daerah, penentuan penerima masih berpotensi tidak tepat sasaran ketika proses seleksi bergantung pada verifikasi manual yang rentan subjektivitas, tidak konsisten, dan kurang efisien, sehingga berisiko memunculkan inclusion error (penerima tidak layak) maupun exclusion error (penerima layak tidak terakomodasi).

Kecamatan Pujud, Kabupaten Rokan Hilir, menjadi konteks penelitian ini karena memiliki variasi karakteristik sosial-ekonomi rumah tangga yang menuntut penguatan proses penetapan penerima bantuan berbasis data agar keputusan lebih objektif dan dapat dipertanggungjawabkan. Untuk itu, pendekatan analitik dengan model klasifikasi biner relevan digunakan: rumah tangga diklasifikasikan ke dalam kategori "layak" atau "tidak layak" menerima PKH. Dalam ranah machine learning terapan, Logistic Regression merupakan model klasifikasi yang luas digunakan karena sederhana, efisien, dan transparan untuk interpretasi keputusan (Ghosh, 2024). Secara konsep, Logistic Regression berakar pada pemodelan peluang kejadian biner melalui fungsi logit sehingga sesuai untuk kasus penentuan kelayakan (Binomial Logistic Regression, 2023). Model ini juga sering dijadikan baseline yang kuat dalam studi klasifikasi dan prediksi, termasuk pada perbandingan dengan Naïve Bayes pada berbagai domain (Ganesh & Kalaiarasi, 2022; Jeevan & Kanimozhi, 2022) serta pada problem prediksi berbasis data besar seperti churn pelanggan (Zhu, 2023). Selain itu, penelitian terapan berbasis machine learning menunjukkan bahwa pendekatan klasifikasi dapat digunakan pada data sosial/administratif untuk membangun sistem rekomendasi atau pengambilan keputusan berbasis data (Pratama et al., 2022; Prasad Lakurwar et al., 2022).

Berdasarkan latar belakang tersebut, penelitian ini bertujuan: (1) membangun model Logistic Regression untuk memprediksi kelayakan penerima PKH berdasarkan variabel sosial-ekonomi, (2) mengevaluasi performa model menggunakan metrik klasifikasi, dan (3) menginterpretasikan faktor yang paling berpengaruh sebagai masukan perbaikan pendataan dan seleksi penerima bantuan. Kontribusi utama penelitian adalah menyediakan kerangka seleksi berbasis data yang lebih objektif, interpretatif, dan dapat direplikasi pada wilayah dengan karakteristik serupa.

2. METODE

Data penelitian berasal dari pendataan rumah tangga di Kecamatan Pujud. Variabel yang digunakan merepresentasikan kondisi sosial-ekonomi, meliputi pendapatan bulanan, jumlah tanggungan, pendidikan, pekerjaan, kondisi rumah, akses fasilitas dasar, serta kepemilikan kendaraan. Target keluaran adalah label biner kelayakan penerima PKH.

Label kelayakan disusun berdasarkan indikator kerentanan yang digunakan pada skripsi, kemudian dilakukan praproses data (pembersihan atribut nonrelevan, penanganan nilai hilang, encoding fitur kategorikal, standardisasi fitur numerik) dan pembagian data latih serta data uji.

Model Logistic Regression dilatih pada data latih dan dievaluasi pada data uji menggunakan confusion matrix, precision, recall, F1-score, dan ROC-AUC. Berikut ini Persamaan Logistic Regression (logit dan fungsi sigmoid)

$$logit(p) = \beta 0 + \sum_{i} \beta i 3 i$$

Persamaan logit tersebut menunjukkan bahwa dalam Logistic Regression, peluang suatu kejadian p (misalnya peluang rumah tangga layak menerima PKH) tidak dihitung langsung, tetapi melalui nilai antara yang disebut logit (log-odds). Nilai logit ini merupakan kombinasi linear dari variabel-variabel penjelas x_i (contoh: pendapatan, jumlah tanggungan, kondisi rumah, akses fasilitas), masing-masing dikalikan koefisiennya βi , ditambah konstanta β0. Karena hasil kombinasi linear tersebut bisa bernilai sangat besar atau sangat kecil (tidak terbatas), ia kemudian diubah menjadi probabilitas dengan fungsi sigmoid, sehingga nilai p selalu berada pada rentang 0 sampai 1. Koefisien βi merepresentasikan arah dan kekuatan pengaruh variabel xi : jika βi positif maka peningkatan xi cenderung menaikkan probabilitas p, sedangkan jika βi negatif maka peningkatan xi cenderung menurunkan probabilitas p; probabilitas inilah yang kemudian digunakan untuk menentukan kelas (misalnya "layak" atau "tidak layak") berdasarkan ambang tertentu.

2.1. Data dan Variabel

Penelitian menggunakan data primer dari 200 kepala keluarga di wilayah Kecamatan Pujud, Kabupaten Rokan Hilir. Variabel prediktor mencakup: pendapatan bulanan, jumlah tanggungan, pendidikan tertinggi kepala keluarga, status pekerjaan, kondisi rumah, luas tempat tinggal, akses fasilitas dasar (listrik/air bersih/sanitasi), kepemilikan kendaraan, serta jumlah anak bersekolah.

Tabel 1. Contoh data setelah seleksi atribut

Pendapatan	Tanggungan	Pendidikan	Pekerjaan	Kondisi Rumah	Akses Fasilitas	Kendaraan
1.500.000	3	SMA	Buruh	Kurang Layak	Tidak Lengkap	Motor
3.000.000	1	Sarjana	Guru	Layak	Lengkap	Mobil

2.2. Pembentukan Label Target (Kelayakan)

Pada penelitian ini, variabel target (label) didefinisikan dalam bentuk klasifikasi biner, yaitu 1 = layak menerima PKH dan 0 = tidak layak. Penetapan label ini menjadi tahap krusial karena dataset awal yang digunakan belum menyediakan kolom kelayakan secara eksplisit (ground truth) dari instansi, sehingga diperlukan mekanisme pelabelan terstruktur agar proses pembelajaran model dapat dilakukan secara konsisten. Oleh karena itu, penelitian membangun aturan labeling berbasis indikator sosial-ekonomi yang merepresentasikan tingkat kerentanan rumah tangga, dengan tujuan mendekati logika seleksi bantuan sosial yang menitikberatkan pada kondisi kebutuhan dan keterbatasan akses.

Aturan pelabelan disusun menggunakan empat indikator utama, yaitu: (1) pendapatan bulanan < Rp2.000.000, yang diasumsikan menggambarkan keterbatasan daya beli; (2) jumlah tanggungan > 2, yang menunjukkan beban ekonomi rumah tangga lebih tinggi; (3) kondisi rumah tergolong kurang layak/tidak layak, sebagai indikator kualitas tempat tinggal yang rendah; serta (4) akses fasilitas dasar yang tidak lengkap, misalnya keterbatasan pada air bersih, sanitasi, atau listrik, yang mencerminkan deprivasi layanan dasar. Keempat indikator tersebut dipilih karena relatif mudah diukur dalam pendataan lapangan dan secara umum berkaitan langsung dengan kondisi kesejahteraan rumah tangga.

Berdasarkan kombinasi indikator tersebut, rumah tangga kemudian diberi label "layak" (1) apabila memenuhi minimal dua indikator (≥2 dari 4). Pendekatan ambang "dua indikator" dipakai untuk menghindari keputusan yang terlalu ketat atau terlalu longgar; artinya, sebuah rumah tangga tidak langsung dianggap layak hanya karena memenuhi satu kondisi, tetapi juga tidak harus memenuhi seluruh indikator sekaligus. Rumah tangga yang hanya memenuhi 0–1 indikator diberi label "tidak layak" (0). Dengan strategi ini, proses pembentukan label menjadi lebih objektif, dapat direplikasi pada data lain, dan menghasilkan target yang cukup representatif untuk melatih model klasifikasi dalam memprediksi kelayakan penerima PKH..

2.3. Praproses

Tahap praproses dilakukan untuk memastikan data siap digunakan dalam pemodelan serta meminimalkan bias dan kesalahan akibat kualitas data yang tidak seragam. Pada penelitian ini, praproses dirancang agar setiap variabel memiliki format yang konsisten, dapat diolah oleh algoritma Logistic Regression, dan menghasilkan performa model yang lebih stabil. Secara umum, tahapan praproses mencakup pembersihan atribut yang tidak relevan, penanganan nilai hilang, transformasi variabel kategorikal menjadi numerik, penyetaraan skala fitur numerik, serta pembagian data menjadi data latih dan data uji.

Tahapan praproses meliputi:

1. Pembersihan data
   Langkah pertama adalah pembersihan data (data cleaning), yaitu menghapus atribut administratif yang tidak berkontribusi terhadap penentuan kelayakan, seperti nama, alamat, RT/RW, dusun, atau informasi identitas lain yang berpotensi menambah noise dan memunculkan bias. Penghapusan atribut ini juga penting untuk menjaga fokus analisis pada variabel sosial-ekonomi yang benar-benar berpengaruh, sekaligus mendukung prinsip privasi data karena informasi sensitif tidak ikut diproses dalam model.
2. Imputasi nilai hilang
   Langkah kedua adalah imputasi nilai hilang (missing value imputation). Dalam data lapangan, nilai kosong dapat muncul akibat responden tidak mengisi, kesalahan pencatatan, atau perbedaan format pendataan. Agar dataset tetap dapat digunakan tanpa mengurangi jumlah sampel secara signifikan, penelitian menerapkan imputasi dengan pendekatan yang sederhana namun robust: nilai hilang pada variabel numerik diisi menggunakan median, karena median lebih tahan terhadap outlier dibanding mean; sedangkan nilai hilang pada variabel kategorikal diisi menggunakan modus, yaitu kategori yang paling sering muncul, sehingga mempertahankan distribusi kategori dominan pada data.
3. Encoding kategorikal
   Langkah ketiga adalah encoding variabel kategorikal, karena Logistic Regression membutuhkan input dalam bentuk numerik. Variabel seperti pendidikan, pekerjaan, kondisi rumah, kepemilikan kendaraan, dan akses fasilitas umumnya berbentuk kategori (misalnya "SD/SMP/SMA", "petani/buruh/wiraswasta", "layak/tidak layak"). Untuk menghindari asumsi urutan (ordinal) yang

keliru pada kategori nominal, penelitian menggunakan one-hot encoding, yaitu mengubah setiap kategori menjadi kolom biner (0/1). Dengan demikian, model dapat mempelajari kontribusi setiap kategori secara lebih tepat tanpa memaksakan hubungan numerik palsu antar kategori.

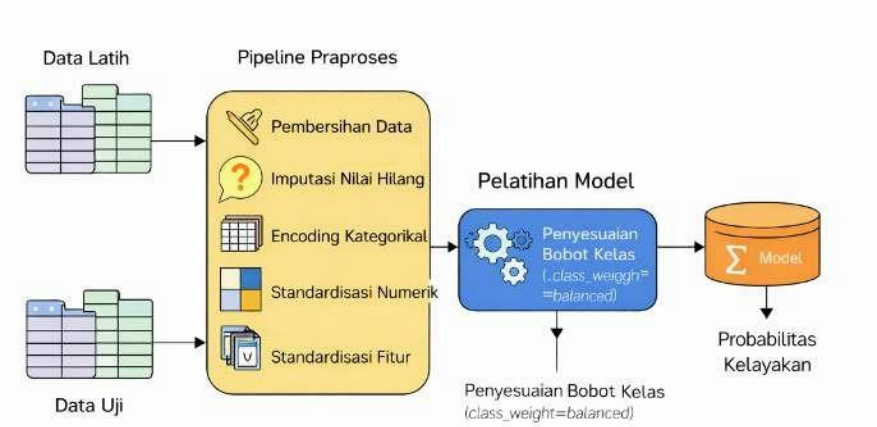
4. Standardisasi fitur numerik
   Langkah keempat adalah standardisasi fitur numerik menggunakan StandardScaler. Standardisasi diperlukan karena fitur numerik seperti pendapatan, jumlah tanggungan, luas rumah, dan jumlah anak sekolah memiliki skala yang berbeda-beda (misalnya pendapatan dalam jutaan rupiah, luas rumah dalam meter persegi). Jika tidak distandardisasi, fitur berskala besar dapat mendominasi proses optimasi dan memengaruhi koefisien model secara tidak proporsional. StandardScaler mengubah data agar memiliki rata-rata mendekati 0 dan deviasi standar 1, sehingga setiap fitur berada pada skala yang sebanding dan proses pelatihan menjadi lebih stabil.
5. Split data
   Langkah terakhir adalah pembagian data (data splitting) menjadi 80% data latih dan 20% data uji. Data latih digunakan untuk membangun dan menyesuaikan parameter model, sedangkan data uji dipakai untuk mengevaluasi performa model pada data yang tidak pernah dilihat sebelumnya, sehingga gambaran akurasi dan generalisasi model lebih objektif. Dengan rangkaian praproses ini, dataset menjadi lebih bersih, konsisten, dan siap digunakan untuk pelatihan Logistic Regression serta evaluasi yang lebih reliabel.

2.4. Pelatihan Model

Pada tahap ini, model klasifikasi dibangun menggunakan Logistic Regression untuk memprediksi kelayakan penerima PKH berdasarkan fitur sosial-ekonomi yang telah disiapkan. Agar proses pemodelan berjalan lebih terstruktur dan menghindari inkonsistensi transformasi data, penelitian menerapkan pipeline praproses. Pipeline ini menggabungkan seluruh langkah praproses (pembersihan atribut nonrelevan, imputasi nilai hilang, encoding variabel kategorikal, dan standardisasi fitur numerik) menjadi satu rangkaian proses yang diterapkan secara otomatis dan identik pada data latih maupun data uji. Dengan pendekatan ini, model selalu menerima bentuk data yang sama (fitur numerik hasil transformasi), serta mengurangi risiko kesalahan umum seperti perbedaan skala, kategori yang tidak sejajar, atau data leakage akibat transformasi yang dilakukan secara terpisah.

Pelatihan dilakukan pada data latih (training set), di mana parameter model Logistic Regression (koefisien β) dipelajari untuk memaksimalkan kemampuan membedakan kelas layak (1) dan tidak layak (0). Logistic Regression menghasilkan keluaran berupa probabilitas kelayakan untuk setiap sampel, kemudian probabilitas tersebut dapat dikonversi menjadi label kelas menggunakan ambang tertentu (misalnya 0,5) sesuai kebutuhan evaluasi. Selain itu, karena dalam data bantuan sosial sering terjadi ketidakseimbangan jumlah kelas (misalnya jumlah "tidak layak" lebih banyak daripada "layak" atau sebaliknya), penelitian menerapkan strategi penyesuaian bobot kelas (class_weight = balanced). Teknik ini bertujuan mengurangi bias model terhadap kelas mayoritas dengan memberikan bobot yang lebih besar pada kelas minoritas, sehingga kesalahan prediksi pada kelas minoritas "dihukum" lebih tinggi selama proses pelatihan. Dampaknya, model terdorong mempelajari pola dari kedua kelas secara lebih seimbang dan tidak hanya mengejar akurasi semu yang dominan pada kelas mayoritas.

Dengan konfigurasi tersebut, pelatihan model menjadi lebih robust untuk kasus penentuan kelayakan PKH yang sensitif terhadap kesalahan klasifikasi. Model yang terlatih diharapkan mampu memberikan prediksi yang lebih adil pada kedua kelas, sekaligus menghasilkan probabilitas yang dapat diinterpretasikan untuk mendukung proses verifikasi lapangan dan pengambilan keputusan penetapan penerima bantuan secara lebih objektif..



Gambar 1. Diagram Pipeline pelatihan menggunakan Logistic Regression dengan penyesuaian bobot kelas

2.5. Evaluasi

Tahap evaluasi dilakukan untuk menilai seberapa baik model Logistic Regression mampu memprediksi kelayakan penerima PKH pada data uji yang belum pernah dilihat saat pelatihan. Evaluasi tidak hanya berfokus pada satu metrik saja, karena pada permasalahan kelayakan bantuan sosial, kesalahan klasifikasi dapat berdampak signifikan—baik inclusion error (penerima tidak layak tetapi diprediksi layak) maupun exclusion error (penerima layak tetapi diprediksi tidak layak). Oleh karena itu, penelitian menggunakan beberapa alat evaluasi yang saling melengkapi, yaitu confusion matrix, classification report (precision, recall, F1-score), serta ROC Curve dan AUC.

Pertama, confusion matrix digunakan untuk melihat ringkasan performa model dalam bentuk jumlah prediksi benar dan salah pada masing-masing kelas. Confusion matrix membagi hasil prediksi menjadi empat kategori: True Positive (TP), yaitu sampel layak yang diprediksi layak; True Negative (TN), sampel tidak layak yang diprediksi tidak layak; False Positive (FP), sampel tidak layak namun diprediksi layak; dan False Negative (FN), sampel layak namun diprediksi tidak layak. Melalui matriks ini, peneliti dapat menilai pola kesalahan model secara lebih detail, termasuk tipe kesalahan mana yang lebih dominan dan berpotensi menimbulkan dampak kebijakan yang lebih besar.

Kedua, evaluasi diperkuat dengan classification report yang memuat metrik utama: precision, recall, dan F1-score untuk tiap kelas. Precision menggambarkan tingkat ketepatan prediksi kelas positif, yaitu seberapa banyak prediksi "layak" yang benar-benar layak; metrik ini penting untuk mengendalikan inclusion error. Recall (sensitivitas) mengukur kemampuan model menangkap semua sampel yang benar-benar layak; metrik ini penting untuk meminimalkan exclusion error. Sementara itu, F1-score adalah rata-rata harmonik antara precision dan recall yang memberikan ukuran keseimbangan ketika salah satu metrik cenderung lebih tinggi daripada yang lain. Dengan melihat ketiga metrik tersebut secara bersamaan, performa model dapat dinilai secara lebih adil, terutama ketika distribusi kelas tidak seimbang.

Ketiga, penelitian juga menggunakan ROC Curve (Receiver Operating Characteristic) untuk mengevaluasi kemampuan model dalam membedakan kelas pada berbagai ambang keputusan (threshold). ROC Curve menggambarkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) ketika threshold diubah-ubah, sehingga memberikan gambaran menyeluruh tentang trade-off antara menangkap kasus "layak" dan risiko salah mengklasifikasikan "tidak layak" sebagai "layak". Untuk merangkum performa ROC secara kuantitatif, digunakan nilai AUC (Area Under the Curve). Nilai AUC berada pada rentang 0–1; semakin mendekati 1, semakin baik kemampuan model dalam memisahkan kelas, sedangkan nilai sekitar 0,5 menunjukkan performa setara tebakan acak. Dengan kombinasi confusion matrix, classification report, serta ROC-AUC, evaluasi model menjadi lebih komprehensif dan dapat mendukung penentuan threshold yang paling sesuai dengan tujuan kebijakan, apakah lebih menekankan pencegahan inclusion error atau meminimalkan exclusion error.

3. HASIL DAN PEMBAHASAN

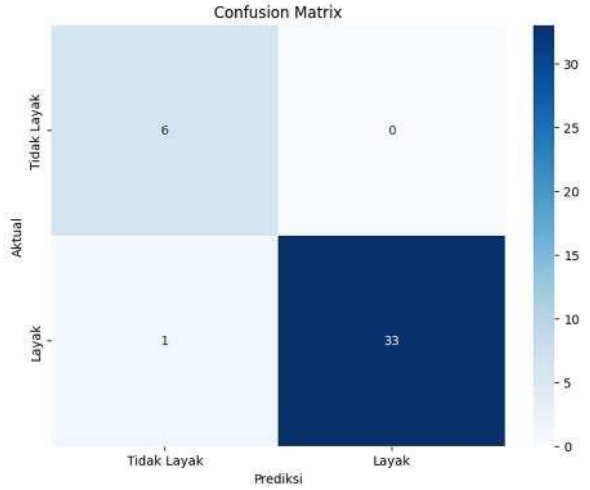
3.1. Deskripsi Data Responden

Data responden merepresentasikan variasi kondisi sosial-ekonomi: pendapatan bervariasi dari rendah hingga menengah, jumlah tanggungan 0–7 orang, serta perbedaan kondisi rumah, kepemilikan

kendaraan, dan akses fasilitas dasar. Gambaran ini penting karena kelayakan PKH bersifat multidimensi, tidak hanya bergantung pada satu variabel.

3.2. Hasil Pelatihan dan Performa Model

Model Logistic Regression yang dilatih pada data latih kemudian diuji pada data uji (20% dari 200; ±40 data). Hasil pada skripsi menunjukkan model dapat mengklasifikasikan sebagian besar sampel uji dengan benar, dengan performa kuat pada pemisahan kelas "layak" vs "tidak layak". Evaluasi berbasis ROC juga menunjukkan nilai AUC yang tinggi (sekitar 0,91) yang menandakan kemampuan diskriminasi model sangat baik.

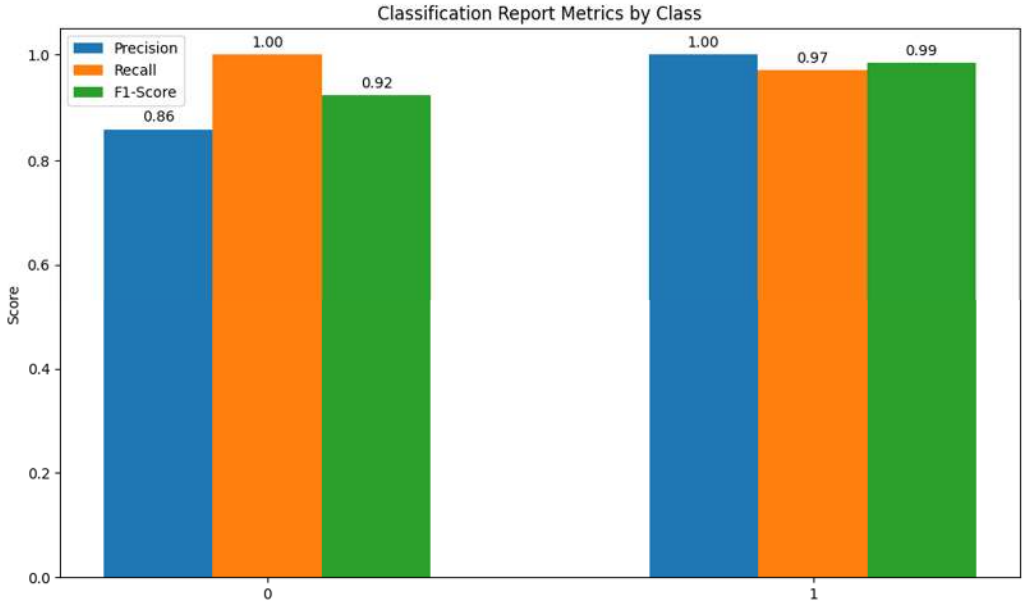


Gambar 2. Confusion Matrix

Perbandingan antara kondisi aktual dan prediksi model untuk dua kelas, yaitu Tidak Layak dan Layak. Dari total 40 data uji, terdapat 6 kasus Tidak Layak yang berhasil diprediksi Tidak Layak (true negative) dan 33 kasus Layak yang berhasil diprediksi Layak (true positive). Tidak ada kasus Tidak Layak yang keliru diprediksi Layak (false positive = 0), sehingga pada data uji ini model tidak menghasilkan inclusion error. Namun, masih terdapat 1 kasus Layak yang salah diprediksi Tidak Layak (false negative), yang berarti masih ada potensi exclusion error meskipun sangat kecil. Secara keseluruhan, hasil ini menggambarkan performa model yang sangat baik karena sebagian besar prediksi tepat, dengan akurasi sekitar 97,5%, precision untuk kelas Layak 100%, dan recall untuk kelas Layak sekitar 97,1%.

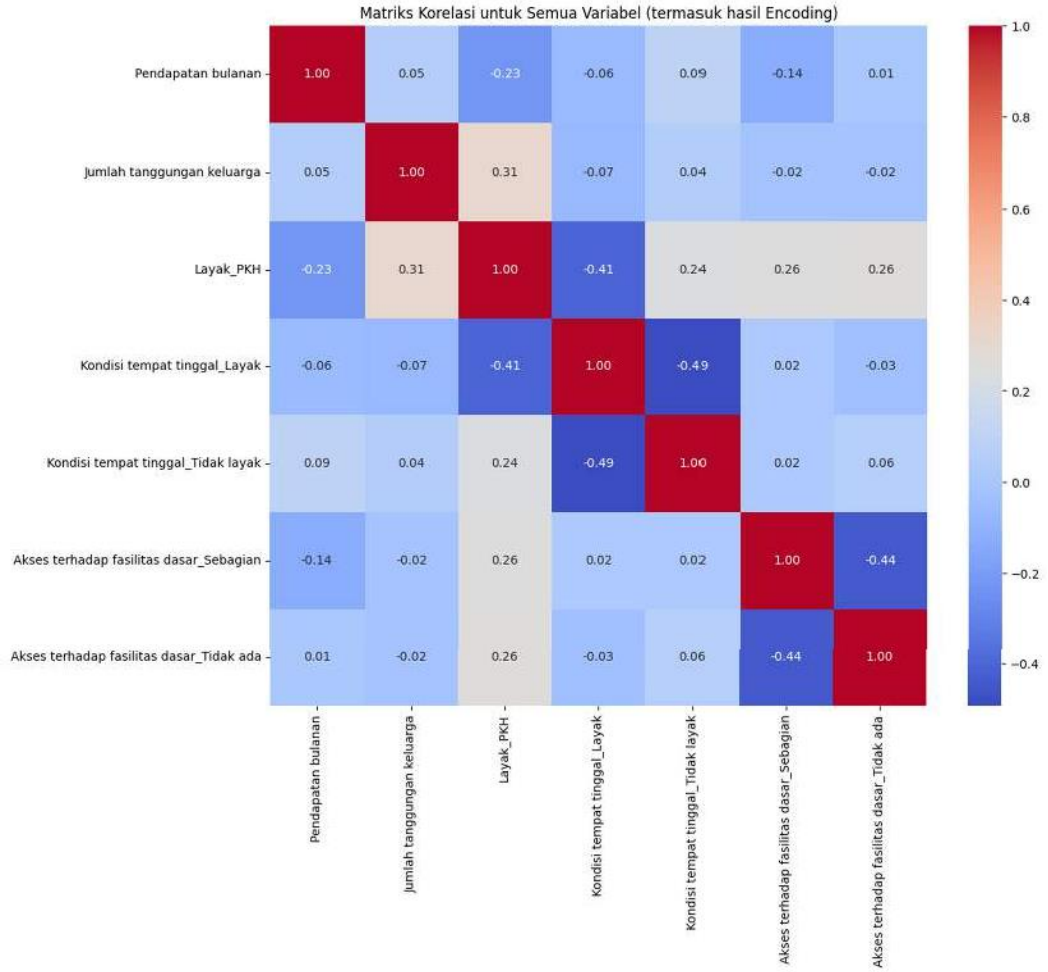
Tabel 2. Classification Report

Class	Precision	Recall	F1-score	Support
0 (tidak layak)	0.86	1.00	0.92	6
1 (layak)	1.00	0.97	0.99	34



Gambar 3. Classification ny Class

Grafik Classification Report Metrics by Class tersebut menampilkan tiga metrik evaluasi—precision, recall, dan F1-score—untuk masing-masing kelas 0 (Tidak Layak) dan 1 (Layak). Pada kelas 0 (Tidak Layak), nilai precision = 0,86 berarti dari seluruh prediksi model yang menyatakan "Tidak Layak", sekitar 86% benar (sisanya adalah kasus yang sebenarnya layak tetapi diprediksi tidak layak). Nilai recall = 1,00 menunjukkan model berhasil menangkap seluruh kasus "Tidak Layak" yang ada di data uji (tidak ada "Tidak Layak" yang lolos menjadi "Layak"), sehingga error tipe false positive untuk kelas "Layak" tidak terjadi. Kombinasi precision dan recall tersebut menghasilkan F1-score = 0,92, yang menandakan performa kelas "Tidak Layak" sangat baik meskipun masih ada sedikit kesalahan saat model memutuskan "Tidak Layak" untuk kasus yang sebenarnya "Layak". Sementara itu, pada kelas 1 (Layak), precision = 1,00 berarti setiap kali model memprediksi "Layak", prediksi tersebut selalu benar (tidak ada penerima yang tidak layak salah masuk sebagai layak), recall = 0,97 menunjukkan hampir semua kasus layak terdeteksi namun ada sebagian kecil yang terlewat (sekitar 3%), dan F1-score = 0,99 mengindikasikan kinerja model untuk mendeteksi "Layak" sangat tinggi dan stabil. Secara keseluruhan, grafik ini memperlihatkan model sangat kuat dalam menjaga agar prediksi "Layak" tidak keliru (menghindari inclusion error), dengan trade-off kecil berupa beberapa kasus layak yang masih bisa terlewat (exclusion error).



Gambar 4. Matriks Korelasi

Heatmap tersebut adalah matriks korelasi yang menunjukkan seberapa kuat hubungan linier antar variabel (termasuk variabel kategorikal yang sudah di-encoding menjadi dummy/one-hot). Nilai korelasi berada pada rentang -1 sampai +1: angka positif berarti dua variabel cenderung naik bersama, angka negatif berarti saat satu naik yang lain cenderung turun, sedangkan nilai mendekati 0 berarti hubungan liniernya lemah. Fokus utama biasanya pada korelasi terhadap variabel target Layak_PKH. Terlihat bahwa pendapatan bulanan berkorelasi negatif dengan Layak_PKH (-0,23), yang mengindikasikan semakin tinggi pendapatan, kecenderungan "layak PKH" cenderung menurun. Sebaliknya, jumlah tanggungan keluarga berkorelasi positif dengan Layak_PKH (0,31), yang berarti semakin banyak tanggungan, kecenderungan "layak" meningkat. Untuk variabel kondisi tempat tinggal yang di-encoding, dummy Kondisi tempat tinggal_Layak memiliki korelasi negatif yang cukup kuat terhadap Layak_PKH (-0,41), sedangkan dummy Kondisi tempat tinggal_Tidak layak berkorelasi positif (0,24); ini masuk akal karena rumah yang "layak" menurunkan peluang menerima bantuan, sementara kondisi "tidak layak" meningkatkan peluang. Pola serupa tampak pada akses fasilitas dasar: Akses_sebagian dan Akses_tidak ada berkorelasi positif terhadap Layak_PKH (keduanya sekitar 0,26), menandakan keterbatasan akses fasilitas dasar berkaitan dengan meningkatnya kelayakan. Selain itu, beberapa pasangan dummy variabel menunjukkan korelasi negatif antar sesamanya (misalnya Kondisi_Layak vs Kondisi_Tidak layak = -0,49 dan Akses_sebagian vs Akses_tidak ada = -0,44), yang wajar karena kategori-kategori tersebut saling eksklusif akibat one-hot encoding. Secara keseluruhan, heatmap ini menegaskan bahwa variabel yang paling "sejalan" dengan kelayakan PKH pada data Anda adalah tanggungan, kondisi rumah, dan akses fasilitas dasar, sementara pendapatan bergerak berlawanan arah dengan kelayakan.

3.3. Faktor Dominan dan Interpretasi

Dari interpretasi koefisien/odds ratio, penelitian menegaskan bahwa:

1. Pendapatan berpengaruh negatif terhadap peluang "layak" (semakin rendah pendapatan, semakin tinggi peluang layak).
2. Jumlah tanggungan berpengaruh positif (semakin banyak tanggungan, semakin tinggi peluang layak).
3. Akses fasilitas dasar yang tidak lengkap meningkatkan kerentanan sehingga meningkatkan peluang layak.

Temuan ini konsisten secara substantif dengan logika kebijakan bantuan sosial: kerentanan ekonomi (pendapatan rendah) dan beban keluarga (tanggungan tinggi), ditambah deprivasi fasilitas dasar, merupakan indikator kuat kebutuhan bantuan. Secara implikatif, model dapat dipakai sebagai alat bantu validasi untuk memprioritaskan verifikasi lapangan pada rumah tangga dengan probabilitas kelayakan tinggi, sehingga mengurangi beban kerja enumerator dan meningkatkan ketepatan sasaran.

3.4. Keterbatasan

Meskipun hasil penelitian menunjukkan performa model yang baik, terdapat beberapa keterbatasan yang perlu dicatat agar interpretasi temuan tetap proporsional dan menjadi landasan perbaikan pada penelitian lanjutan. Pertama, cakupan data penelitian hanya berasal dari Kecamatan Pujud, Kabupaten Rokan Hilir, sehingga karakteristik sosial-ekonomi yang tertangkap pada dataset sangat mungkin bersifat lokal, dipengaruhi oleh struktur mata pencaharian, pola pendapatan, serta kondisi infrastruktur wilayah tersebut. Akibatnya, generaliasi model ke kecamatan atau kabupaten lain perlu kehati-hatian, karena distribusi variabel, indikator kemiskinan, serta pola penerimaan bantuan bisa berbeda. Untuk meningkatkan daya generalisasi, penelitian berikutnya perlu memperluas sampel lintas desa/kecamatan atau bahkan lintas kabupaten sehingga model diuji pada variasi karakteristik yang lebih beragam.

Kedua, label target (kelayakan) pada penelitian ini dibentuk melalui aturan (rule-based labeling) yang ditetapkan peneliti karena label administratif resmi tidak tersedia langsung pada dataset awal. Meskipun aturan tersebut didasarkan pada indikator sosial-ekonomi yang logis, pendekatan ini tetap memiliki potensi bias, misalnya karena ambang batas (threshold) pendapatan atau ketentuan "memenuhi ≥2 indikator" bisa berbeda dengan kebijakan faktual di lapangan, serta dapat berubah mengikuti regulasi program. Selain itu, karena label belum melalui validasi administratif oleh instansi sosial/pengelola PKH, maka label yang digunakan belum dapat dipastikan sepenuhnya merepresentasikan keputusan resmi program. Penelitian lanjutan sebaiknya melakukan cross-check dengan data penerima PKH yang telah ditetapkan pemerintah atau melibatkan verifikasi bersama pendamping/instansi terkait agar label target lebih akurat dan memiliki legitimasi kebijakan.

Ketiga, penelitian ini hanya menggunakan Logistic Regression sebagai model utama tanpa melakukan pembandingan sistematis dengan algoritma klasifikasi lain seperti Decision Tree, Random Forest, Gradient Boosting, SVM, atau XGBoost. Logistic Regression memiliki keunggulan interpretabilitas dan kesederhanaan, namun pada dataset yang kompleks atau memiliki interaksi non-linear, algoritma lain mungkin menghasilkan performa yang lebih tinggi atau lebih stabil. Tanpa baseline pembanding, kesimpulan "model terbaik" belum dapat ditegaskan, karena belum diketahui apakah peningkatan performa bisa dicapai dengan metode lain, atau apakah Logistic Regression sudah optimal untuk karakteristik data yang digunakan.

Keempat, variabel yang digunakan masih berfokus pada indikator sosial-ekonomi dasar dan belum mencakup beberapa aspek penting yang dalam konteks bantuan sosial sering menjadi pembeda kerentanan rumah tangga, seperti kondisi kesehatan anggota keluarga, status disabilitas, beban utang, kepemilikan aset produktif, kepemilikan lahan, stabilitas pekerjaan musiman, atau komponen pengeluaran rutin (misalnya biaya pendidikan dan kesehatan). Tidak dimasukkannya variabel-variabel tersebut dapat membatasi kemampuan model dalam menangkap kondisi kerentanan secara lebih komprehensif, sehingga peluang misklasifikasi masih mungkin terjadi pada kasus-kasus yang secara ekonomi tampak "cukup" namun sebenarnya rentan karena beban kesehatan atau utang, ataupun sebaliknya.

Dengan memahami keterbatasan-keterbatasan ini, hasil penelitian tetap dapat dimanfaatkan sebagai alat bantu seleksi berbasis data pada konteks lokal, sekaligus menjadi pijakan untuk pengembangan model yang lebih kuat melalui perluasan wilayah, validasi label dengan instansi, benchmarking multi-algoritma, serta penambahan variabel kerentanan yang lebih lengkap..

.

4. SIMPULAN

Penelitian ini menunjukkan bahwa Logistic Regression dapat digunakan untuk menganalisis kelayakan penerima PKH berbasis variabel sosial-ekonomi rumah tangga. Faktor yang paling dominan dalam penentuan kelayakan adalah pendapatan bulanan, jumlah tanggungan, dan akses fasilitas dasar. Berdasarkan evaluasi pada data uji, model memiliki kemampuan klasifikasi yang baik dan AUC yang tinggi,

sehingga berpotensi mendukung proses seleksi penerima bantuan yang lebih objektif, transparan, dan tepat sasaran di tingkat lokal.

penelitian lanjutan disarankan menambah variabel (kesehatan, aset/lahan, beban utang), memperluas wilayah sampel, serta membandingkan beberapa algoritma klasifikasi untuk memperoleh model terbaik. Implementasi praktis juga memerlukan validasi bersama instansi terkait agar aturan labeling dan keputusan model selaras dengan kebijakan resmi.

5. UCAPAN TERIMA KASIH

Penulis menyampaikan ucapan terima kasih yang sebesar-besarnya kepada Program Studi Teknik Informatika Universitas Kahuripan Kediri atas dukungan akademik, fasilitas, serta lingkungan penelitian yang kondusif selama proses penyusunan dan pelaksanaan penelitian ini. Penulis juga mengucapkan terima kasih kepada dosen pembimbing yang telah memberikan arahan, masukan metodologis, koreksi, serta bimbingan secara berkelanjutan sehingga penelitian dapat diselesaikan dengan baik dan sesuai kaidah ilmiah. Selain itu, penulis menyampaikan apresiasi kepada pihak Kecamatan Pujud, Kabupaten Rokan Hilir, serta seluruh pihak yang terlibat dalam proses pendataan dan pengumpulan informasi lapangan, atas bantuan, akses data, dan kerja sama yang diberikan selama penelitian berlangsung. Ucapan terima kasih juga penulis sampaikan kepada rekan-rekan dan semua pihak yang tidak dapat disebutkan satu per satu, yang telah memberikan dukungan moril maupun teknis sehingga penelitian ini dapat diselesaikan hingga tahap penyusunan artikel ilmiah.

6. DAFTAR PUSTAKA

Anggraini, N. W. (2024). Program Keluarga Harapan (PKH): Examining the Social Relation Approach on Poverty Alleviation Policies in Indonesia. https://doi.org/10.2991/978-2-38476-228-6_22

Binomial Logistic Regression. (2023). https://doi.org/10.1017/9781108923071.012

Falah, M. H. (2023). Tantangan Koperasi Nelayan Sebagai Penyeimbang Rezim Pengelolaan Sumber Daya Kelautan Dan Perikanan Di Indonesia. https://doi.org/10.55981/brin.908.c765

Ganesh, R., & Kalaiarasi, S. (2022). Strange Approach of Movie Rating Prediction Using Logistic Regression Comparing to Gaussian Naive Bayes Algorithm. https://doi.org/10.3233/apc220041

Ghosh, D. K. (2024). Perspective Chapter: Linear Regression and Logistic Regression Models. https://doi.org/10.5772/intechopen.1003183

Harahap, N. R., Arma, N., Sipayung, N. A., & Syari, M. (2021). Faktor Yang Memengaruhi Ibu Terhadap Pemilihan Tempat Persalinan Di Desa Aek Badak Jae. Journal of Midwifery Senior, 5(1).

HS, N., & Yani, R. (2021). Faktor-Faktor Yang Memengaruhi Penentuan Jumlah Anak Dalam Keluarga Di Kecamatan Bukit Kabupaten Bener Meriah. Jurnal Ilmiah Cerebral Medika, 3(1). https://doi.org/10.53475/jicm.v3i1.79

Ing, L. Y., & Basri, M. C. (2022). COVID-19 in Indonesia. https://doi.org/10.4324/9781003243670

Jeevan, K., & Kanimozhi, K. (2022). Improved Accuracy for Fake News in Social Media Using Logistic Regression Comparing Naive Bayes Classifier. https://doi.org/10.3233/apc220068

Kementerian Sosial RI. (2021). Pedoman Pelaksanaan Program Keluarga Harapan 2021. In DIREKTUR JAMINAN SOSIAL KELUARGA DIREKTORAT JENDRAL PERLINDUNGAN DAN JAMINAN SOSIAL KEMENTRIAN SOSIAL ri (Vol. 5, Issue 2).

Mardiani, R., & Lhutfi, I. (2021). Faktor – Faktor Yang Mempengaruhi Minat Mahasiswa Dalam Pemilihan Jurusan Akuntansi (Studi Kasus Pada Mahasiswa Baru Di Jurusan Akuntansi Perguruan Tinggi Kota Cimahi). Jurnal Pendidikan Akuntansi & Keuangan, 9(1), 74–87. https://doi.org/10.17509/jpak.v9i1.30083

Muslihatun, W. N., & Santi, M. Y. (2022). Faktor yang Mempengaruhi Keterlibatan Ayah dalam Pengasuhan Anak Usia Dini. Window of Health : Jurnal Kesehatan, 404–418. https://doi.org/10.33096/woh.vi.131

Prasad Lakurwar, et.al. (2022). In Machine Learning Methods for Engineering Application Development. https://doi.org/10.2174/9789815079180122010 1

Pratama, A. R., Aryanto, R. R., & Pratama, A. T. M. (2022). Model Klasifikasi Calon Mahasiswa Baru Untuk Sistem Rekomendasi Program Studi Sarjana Berbasis Machine Learning. Jurnal Teknologi Informasi Dan Ilmu Komputer, 9(4), 725–734. https://doi.org/10.25126/jtiik.2022934311

Tangi, D. S., Karels, D. W., & Hangge, E. E. (2022). Pemilihan Moda Transportasi Angkutan Umum di Golewa Selatan Kabupaten Ngada. Jurnal Teknik Sipil, 11(1), 77–90. https://puslit2.petra.ac.id/index.php/jurnal-teknik-sipil/article/view/24659

Zhu, Z. H. (2023). Customer Churn Prediction Based on Big Data and Machine Learning Approaches. https://doi.org/10.2991/978-94-6463-142-5_