



E-ISSN 3032-601X & P-ISSN 3032-7105

Vol. 3, No. 1, 2026

MISTER

**Journal of Multidisciplinary Inquiry in Science,
Technology and Educational Research**

**Jurnal Penelitian Multidisiplin dalam Ilmu
Pengetahuan, Teknologi dan Pendidikan**

**UNIVERSITAS SERAMBI MEKKAH
KOTA BANDA ACEH**

mister@serambimekkah.ac.id

Journal of Multidisciplinary Inquiry in Science Technology
and Educational Research

Journal of MISTER

Vol. 3, No. 1, 2026

Pages: 1091–1100

Penerapan *Logistic Regression* Untuk Klarifikasi Pendapatan Berdasarkan Variabel Sosial Ekonomi

Mahfudin Adnan, Hasbi Firmansyah, Wahyu Asriyani, Ria Indah Fitria

Universitas Pancasakti Tegal

Article in Journal of MISTER

Available at : <https://jurnal.serambimekkah.ac.id/index.php/mister/index>

DOI : <https://doi.org/10.32672/mister.v3i1.4035>

How to Cite this Article

APA : Adnan, M., Hasbi Firmansyah, Wahyu Asriyani, & Ria Indah Fitria. (2025). Penerapan Logistic Regression Untuk Klarifikasi Pendapatan Berdasarkan Variabel Sosial Ekonomi. *Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research*, 3(1), 1091 – 1100. <https://doi.org/10.32672/mister.v3i1.4035>

Others Visit : <https://jurnal.serambimekkah.ac.id/index.php/mister/index>

MISTER: *Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research* is a scholarly journal dedicated to the exploration and dissemination of innovative ideas, trends and research on the various topics include, but not limited to functional areas of Science, Technology, Education, Humanities, Economy, Art, Health and Medicine, Environment and Sustainability or Law and Ethics.

MISTER: *Journal of Multidisciplinary Inquiry in Science, Technology and Educational Research* is an open-access journal, and users are permitted to read, download, copy, search, or link to the full text of articles or use them for other lawful purposes. Articles on Journal of MISTER have been previewed and authenticated by the Authors before sending for publication. The Journal, Chief Editor, and the editorial board are not entitled or liable to either justify or responsible for inaccurate and misleading data if any. It is the sole responsibility of the Author concerned.



ISSN 3032-7105

ISSN 3032-601X



Penerapan *Logistic Regression* Untuk Klarifikasi Pendapatan Berdasarkan Variabel Sosial Ekonomi

Mahfudin Adnan^{1*}, Hasbi Firmansyah², Wahyu Asriyani³, Ria Indah Fitria⁴

Program Studi Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Pancasakti Tegal, Tegal, Indonesia^{1,2,4}

Program Studi Pendidikan dan Sastra Indonesia, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Pancasakti Tegal, Tegal, Indonesia³

Email: mahfudinadnan@gmail.com, hasbifirmansyah@upstegal.ac.id, asriyani1409@gmail.com, ria_indah@upstegal.ac.id

Diterima: 12-12-2025

| Disetujui: 22-12-2025

| Diterbitkan: 24-12-2025

ABSTRACT

This study aims to analyze the effectiveness of the Logistic Regression algorithm in classifying individual income levels based on socio-economic variables. The research employs a quantitative experimental approach using the Adult Census Income dataset, which consists of demographic and employment-related attributes. Data preprocessing stages include handling missing values, encoding categorical attributes, and splitting the dataset into training and testing subsets with an 80:20 ratio. Model development and evaluation were conducted using RapidMiner Studio to ensure a structured and reproducible workflow. The performance of the proposed model was evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. Experimental results show that the Logistic Regression model achieved an accuracy of 80.91%, indicating a reliable capability in distinguishing income categories $\leq 50K$ and $> 50K$. The model demonstrates strong performance in identifying low-income individuals, while challenges remain in improving precision for the high-income class due to class imbalance. Overall, the findings confirm that Logistic Regression remains a relevant and interpretable baseline model for income classification tasks, particularly in socio-economic analysis and policy-oriented studies where transparency and interpretability are essential.

Keywords: Logistic Regression, Income Classification, Machine Learning, Socio-Economic Data, RapidMiner

ABSTRAK

Penelitian ini bertujuan untuk menganalisis efektivitas algoritma Logistic Regression dalam mengklasifikasikan tingkat pendapatan individu berdasarkan variabel sosial-ekonomi. Penelitian menggunakan pendekatan kuantitatif dengan metode eksperimen pada dataset Adult Census Income yang mencakup atribut demografis dan pekerjaan. Tahapan penelitian meliputi pra-pemrosesan data, pengkodean atribut kategorikal, serta pembagian data menjadi data latih dan data uji dengan rasio 80:20. Proses pembangunan dan evaluasi model dilakukan menggunakan RapidMiner Studio untuk memastikan alur eksperimen yang terstruktur dan dapat direplikasi. Evaluasi performa model dilakukan menggunakan metrik akurasi, precision, recall, F1-score, serta analisis confusion matrix. Hasil pengujian menunjukkan bahwa model Logistic Regression menghasilkan akurasi sebesar 80,91%, yang menandakan kemampuan yang cukup baik dalam membedakan kategori pendapatan $\leq 50K$ dan $> 50K$. Model menunjukkan kinerja yang sangat baik dalam mengidentifikasi individu berpendapatan rendah, namun masih menghadapi tantangan pada precision kelas pendapatan tinggi akibat ketidakseimbangan distribusi data. Secara keseluruhan, penelitian ini membuktikan bahwa Logistic Regression tetap relevan sebagai model baseline yang seimbang antara performa prediksi dan interpretabilitas, khususnya untuk analisis sosial dan perumusan kebijakan publik.

Kata Kunci: Logistic Regression, Klarifikasi Pendapatan, Machine Learning, Data Sosial-Ekonomi, RapidMiner

PENDAHULUAN

Pendapatan individu merupakan salah satu indikator utama kesejahteraan ekonomi dan stabilitas sosial dalam masyarakat modern. Di tengah ketimpangan ekonomi dan pertumbuhan populasi yang cepat, kemampuan untuk memprediksi status pendapatan berdasarkan variabel demografis, pendidikan dan pekerjaan menjadi semakin relevan baik untuk keperluan riset sosial, kebijakan publik, maupun evaluasi program kesejahteraan. Metode tradisional analisis statistik seringkali terbatas dalam menangani volume data besar, variabel kategorikal kompleks, serta interaksi non-linear antar variabel. Oleh karena itu, pendekatan machine learning muncul sebagai solusi yang menjanjikan dalam klarifikasi pendapatan.

Dalam literatur terkini, banyak penelitian yang mengaplikasikan algoritma klarifikasi untuk memprediksi penghasilan individu atau rumah tangga. Sebagai contoh, studi terbaru menggunakan dataset publik dan membandingkan berbagai algoritma termasuk Logistic Regression (LR) untuk menentukan apakah pendapatan seseorang berada di atas atau di bawah ambang tertentu (*binary income classification*). Hasilnya menunjukkan bahwa metode berbasis machine learning dapat mencapai tingkat akurasi dan realibilitas yang cukup baik dalam klarifikasi pendapatan (Wang, 2022)(Shuvo et al., 2024).

Namun demikian, penggunaan model seperti Random Forest, ensemble learning, atau model yang lebih kompleks sering menawarkan performa yang lebih tinggi terutama dalam menangani data dengan banyak fitur dan ketidakseimbangan kelas (Wang, 2022)(Alexander & Supriyanto, 2023)(Siringoringo et al., 2024). Meski demikian, “kotak hitam” (*black-box*) pada model kompleks ini dapat menyulitkan interpretasi hasil, terutama ketika hasil prediksi dipakai dalam konteks kebijakan publik atau analisis sosial, di mana transparansi dan penjelasan kontribusi tiap variabel sangat penting.

Dalam konteks tersebut, Logistic Regression muncul sebagai opsi menarik meskipun sederhana, metode ini tetap populer karena kemampuannya menghasilkan model yang interpretable, memungkinkan untuk mengukur kontribusi relatif tiap variabel terhadap probabilitas outcome, serta lebih hemat komputasi dibanding metode kompleks. Hal ini membuat Logistic Regression relevan bila tujuan penelitian bukan hanya akurasi, tetapi juga pemahaman mendalam terhadap determinan pendapatan (Shuvo et al., 2024)(Siringoringo et al., 2024).

Akan tetapi, penerapan model klasifikasi pendapatan dengan dataset sosial-ekonomi nyata sering menghadapi tantangan signifikan: imbalance antara kategori (misalnya jumlah orang berpendapatan rendah jauh lebih banyak), fitur kategorikal banyak dan beragam, serta kebutuhan pra-pemrosesan data (encoding, normalisasi, feature selection) agar model bisa belajar dengan baik tanpa bias (Wang, 2022)(Siringoringo et al., 2024)(Ardana et al., 2023). Oleh sebab itu, proses pra-pemrosesan data serta pemilihan fitur menjadi aspek krusial sebelum model dibangun, agar hasil prediksi tidak hanya akurat tapi juga valid dan dapat diinterpretasikan dengan benar.

Beberapa penelitian terkini menunjukkan bahwa kombinasi Logistic Regression dengan teknik balancing kelas serta seleksi fitur mampu menghasilkan performa klasifikasi yang kompetitif, sekaligus mempertahankan kejelasan dalam interpretasi model (Siringoringo et al., 2024)(Susetyoko et al., 2022)(Wang, 2022). Dengan demikian, meskipun algoritma kompleks menawarkan akurasi tinggi, penggunaan Logistic Regression tetap dibenarkan ketika fokus penelitian diarahkan pada analisis faktor-faktor penentu pendapatan dan efek variabel independen terhadap probabilitas pendapatan tinggi atau rendah.

Berdasarkan uraian di atas, penelitian ini mengusulkan penerapan Logistic Regression untuk mengklasifikasikan status pendapatan ($\leq 50K$ vs $> 50K$) berdasarkan variabel sosial ekonomi dan

demografis. Penelitian akan menggunakan dataset publik, menjalani pra-pemrosesan menyeluruh, dan mengevaluasi kinerja model melalui metrik seperti akurasi, precision, recall, serta analisis kontribusi variabel prediktor. Dengan pendekatan ini, diharapkan penelitian dapat memberikan dua kontribusi: secara akademik dengan validasi metodologis terhadap model prediksi pendapatan terkini, dan secara praktis dengan pemahaman faktor determinan pendapatan yang transparan dan dapat dijadikan dasar kebijakan sosial.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen untuk menganalisis kemampuan algoritma *Logistic Regression* dalam melakukan klasifikasi pendapatan berdasarkan variabel sosial-ekonomi. Tahapan penelitian meliputi proses pengumpulan dataset, pra-pemrosesan data, pembagian dataset ke data pelatihan dan data pengujian, pembangunan model *Logistic Regression*, serta evaluasi performa model. Seluruh tahapan dilakukan menggunakan perangkat lunak RapidMiner Studio karena aplikasi ini menyediakan *workflow environment* yang memudahkan proses eksperimen tanpa perlu melakukan pemrograman langsung, serta mendukung operator lengkap untuk *preprocessing*, *training*, dan *evaluation* model *machine learning* sebagaimana dibuktikan dalam riset terapan terbaru (Wang, 2022)(Alexander & Supriyanto, 2023)(Dey & Roy, 2023).

Rancangan Metode

1. Pengumpulan Dataset

Table 1. Adult Census Income

Age	JobType	MaritalStatus	Occupation	Relationship	Gender	Capital Gain	Capital Loss	HoursPer Week	SalStat
39	State-gov	Never-married	Adm-clerical	Not-in-family	Male	2174	0	40	<=50K
50	Self-emp-not-inc	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	13	<=50K
38	Private	Divorced	Handlers-cleaners	Not-in-family	Male	0	0	40	<=50K
53	Private	Married-civ-spouse	Handlers-cleaners	Husband	Male	0	0	40	<=50K
28	Private	Married-civ-spouse	Prof-specialty	Wife	Female	0	0	40	<=50K
37	Private	Married-civ-spouse	Exec-managerial	Wife	Female	0	0	40	<=50K
49	Private	Married-spouse-absent	Other-service	Not-in-family	Female	0	0	16	<=50K
52	Self-emp-not-inc	Married-civ-spouse	Exec-managerial	Husband	Male	0	0	45	>50K
31	Private	Never-married	Prof-specialty	Not-in-family	Female	14084	0	50	>50K

42	Private	Married-civ-spouse	Exec-managerial	Husband	Male	5178	0	40	>50K
----	---------	--------------------	-----------------	---------	------	------	---	----	------

Dataset penelitian ini berupa kumpulan data sosial-ekonomi dengan atribut numerik dan kategorikal yang mencakup usia, jenis pekerjaan, status pernikahan, pekerjaan, hubungan, jenis kelamin, keuntungan modal, kerugian modal, jam kerja perminggu, dan pendapatan. Dataset dipilih karena struktur datanya sesuai untuk membangun model klasifikasi pendapatan dengan permasalahan biner, yaitu pendapatan $\leq 50K$ dan $>50K$ (Chu, 2025a). Sejalan dengan penelitian terdahulu mengenai income classification, tahap awal pengolahan data dilakukan untuk mengevaluasi kualitas dataset, termasuk mendeteksi keberadaan *missing values*, nilai duplikat, ketidakseimbangan distribusi kelas, serta kesesuaian tipe data dengan format numerik atau kategorikal. Tahap ini penting karena studi terkini menunjukkan bahwa kualitas *preprocessing* sangat menentukan akurasi klasifikasi pada domain sosial-ekonomi (Pradesh, 2025).

2. Preprosesing

Setelah kualitas data diverifikasi, tahap *preprocessing* dilanjutkan dengan imputasi *missing values* menggunakan *strategy* yang relevan sesuai tipe data, normalisasi variabel numerik, serta *encoding* fitur kategorikal menggunakan *One-Hot Encoding* agar dapat diolah oleh algoritma *Logistic Regression*. Selanjutnya, fitur yang berperan sebagai penentu pendapatan dianalisis melalui proses *feature selection* untuk mengurangi redundansi dan *noise* pada model. Tahap ini mengikuti rekomendasi model pipeline penelitian *income prediction* berbasis ML yang menekankan bahwa *encoding* dan *feature selection* berkontribusi langsung terhadap stabilitas dan generalisasi model (Wan, 2023).

3. Pengolahan Data



Gambar 1. Alur Proses Menggunakan Logistic Regression pada Rapidminer

Proses pengolahan data dilakukan menggunakan perangkat lunak RapidMiner sesuai alur pada Gambar 1. Tahapan diawali dengan operator *Retrieve* untuk memanggil dataset *Adult Census Income* dari repository. Selanjutnya operator *Select Attributes* digunakan untuk memilih atribut yang relevan, kemudian *Set Role* menentukan label pendapatan sebagai variabel target ($\leq 50K$ atau $>50K$). Tahap berikutnya menerapkan operator *Numerical to Binominal* agar atribut pendapatan memiliki format biner yang sesuai untuk proses klasifikasi. Operator *Replace Missing Values* digunakan untuk memastikan tidak ada nilai kosong yang dapat mengganggu proses pelatihan model (Islam et al., 2024).

Setelah data siap, operator *Split Data* membagi dataset menjadi bagian pelatihan dan pengujian agar performa model dapat dievaluasi secara objektif. Model kemudian dibangun menggunakan operator *Logistic Regression*, yang mempelajari hubungan antara variabel sosial ekonomi terhadap kategori pendapatan. Model yang dihasilkan diterapkan ke data pengujian melalui operator *Apply Model*, dan tahap akhir menggunakan operator *Performance* untuk menghitung metrik evaluasi seperti akurasi, precision, dan recall. Dengan alur proses ini, seluruh tahapan pemrosesan data mulai dari pemanggilan dataset hingga evaluasi performa model dapat dilakukan secara terstruktur dan konsisten (Wan, 2023).

4. Pembagian Data



Gambar 2. Pembagian Data Pada Operator Split Data

Dataset dibagi menjadi dua bagian menggunakan *Split Data*, yakni 80% data untuk pelatihan dan 20% data untuk pengujian. Proporsi ini dipilih untuk menghindari *overfitting* dan memastikan jumlah data yang memadai pada tahap evaluasi. Pembagian data dilakukan secara *stratified* agar distribusi kelas pendapatan sama pada kedua subset sehingga model tetap stabil dalam mendeteksi kelas minoritas. Studi terkini yang berfokus pada income prediction menegaskan bahwa evaluasi model akan bias jika pembagian data tidak mempertahankan proporsi kelas (Thapa, 2023).

5. Pembangunan Model

Model dibangun menggunakan algoritma *Logistic Regression* dengan parameter standar RapidMiner meliputi optimisasi iteratif, *regularization*, batas iterasi maksimum, serta aktivasi *L2 penalty* guna mengurangi risiko multikolinearitas antar atribut. Logistic Regression dipilih karena interpretabilitasnya, kemampuannya memberikan kontribusi koefisien setiap variabel dalam bentuk *log-odds*, dan tingkat keakuratannya yang kompetitif pada klasifikasi pendapatan dibandingkan model yang lebih kompleks pada domain serupa. Berbagai penelitian menunjukkan bahwa meskipun model berbasis *ensemble* seperti *Random Forest* dan *XGBoost* dapat menghasilkan akurasi yang lebih tinggi, Logistic Regression tetap unggul dalam aspek interpretasi sebab-akibat antar variabel, transparansi model, dan kemudahan evaluasi untuk konteks sosial (Islam et al., 2024).

6. Evaluasi Performa

Pengujian performa dilakukan menggunakan operator *Performance (Classification)* pada RapidMiner. Metrik utama yang dihitung adalah akurasi, *precision*, *recall*, dan *F1-score* untuk mengevaluasi kinerja klasifikasi pendapatan biner secara menyeluruh. Selain itu, *confusion matrix* digunakan untuk mengidentifikasi tingkat kesalahan prediksi pada masing-masing kelas, terutama pada kelas pendapatan >50K yang biasanya berjumlah lebih sedikit dibanding kelas ≤50K. Studi tahun 2021–2025 pada topik income prediction menekankan bahwa penggunaan evaluasi tunggal berbasis akurasi saja tidak memadai karena model dapat bias terhadap kelas dominan, sehingga penggunaan *precision*, *recall*, dan *F1-score* menjadi wajib dalam penelitian ini (Chu, 2025b).

Persamaan Matematika

1. Model Logistic Regression

Logistic Regression memodelkan hubungan antara variabel input (x_i) dan probabilitas suatu data termasuk ke kelas tertentu ($>50K$ atau $\leq 50K$) dengan fungsi logit sebagai berikut:

$$P(Y = 1|X) = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad 1$$

Keterangan:

- $P(Y = 1|X)$: Probabilitas Pendapatan $>50K$
 β_0 : Intercept
 $\beta_1 \dots \beta_n$: Koefisien variabel prediktor
 $X_1 \dots X_n$: Variabel sosial-ekonomi dalam dataset
 e : Konstanta Euler

Model ini dipilih karena mampu mengukur pengaruh variabel sosial-ekonomi terhadap probabilitas pendapatan melalui log-odds yang mudah diinterpretasikan (Wang, 2022)(Islam et al., 2024).

2. Log-Likelihood Function

Parameter Logistic Regression (β) diestimasi dengan memaksimalkan fungsi log-likelihood berikut:

$$LL(\beta) = \sum_{i=1}^m [y_i \ln(P_i) + (1 - y_i) \ln(1 - P_i)] \quad 2$$

Keterangan:

- m : Jumlah Sampel
 y_i : Label data ke- i ($1 = >50K$, $0 = \leq 50K$)
 P_i : Probabilitas hasil prediksi model

Pendekatan ini menjadi standar pada penelitian prediksi pendapatan karena mampu mengoptimalkan estimasi probabilitas secara efisien (Wan, 2023).

3. Regularization (L2 Penalty)

Untuk mencegah *overfitting* dan multikolinearitas, Logistic Regression dioptimalkan menggunakan regularisasi L2 sebagai berikut:

$$J(\beta) = -LL(\beta) + \lambda \sum_{j=1}^n \beta_j^2 \quad 3$$

Keterangan:

- $J(\beta)$: Fungsi tujuan setelah regulasi
 λ : Parameter regulasi
 β_j : Koefisien model

Regularisasi terbukti meningkatkan stabilitas klasifikasi pendapatan pada dataset sosial-ekonomi yang memiliki banyak fitur kategorikal (Alexander & Supriyanto, 2023)(Sharifzadeh et al., 2023).

4. Confusion Matrix

Evaluasi performa klasifikasi dilakukan menggunakan *confusion matrix* sebagai berikut:

Table 2. Confusion Matrix

	PREDIKSI POSITIF	PREDIKSI NEGATIF
AKTUAL POSITIF	True Positive (TP)	False Negative (FN)
AKTUAL NEGATIF	False Positive (FP)	True Negative (TN)

Dari *confusion matrix* dihitung metrik berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

4

Penggunaan kombinasi *accuracy*, *precision*, *recall*, dan *F1-score* direkomendasikan karena dataset pendapatan umumnya mengalami ketidakseimbangan kelas (Siringoringo et al., 2024)(Alghamdi, 2024)(Li et al., 2021).

HASIL DAN PEMBAHASAN

Hasil Pengujian Model Logistic Regression

accuracy: 80.91%			
	true <=50K	true >50K	class precision
pred. <=50K	4091	390	81.30%
pred. >50K	853	1178	88.00%
class recall	82.75%	75.13%	

Gambar 3. Confusion Matrix Hasil Klasifikasi Pendapatan Menggunakan Logistic Regression

Hasil pengujian model Logistic Regression diperoleh melalui operator *Performance (Classification)* pada RapidMiner. Berdasarkan eksperimen yang dilakukan, model menghasilkan nilai akurasi sebesar 80,91%, yang menunjukkan bahwa mayoritas data uji berhasil diklasifikasikan dengan benar ke dalam kategori pendapatan $\leq 50K$ dan $> 50K$.

Hasil confusion matrix menunjukkan bahwa model mampu mengklasifikasikan 4091 data berpendapatan $\leq 50K$ secara benar, sementara 1178 data berpendapatan $> 50K$ juga berhasil diprediksi dengan tepat. Namun demikian, masih terdapat 853 data $\leq 50K$ yang salah diprediksi sebagai $> 50K$, serta 390 data $> 50K$ yang keliru diklasifikasikan sebagai $\leq 50K$. Temuan ini menunjukkan bahwa meskipun performa model secara umum cukup baik, tantangan utama masih terdapat pada pemisahan kelas

pendapatan tinggi yang relatif lebih sedikit jumlahnya dibandingkan kelas pendapatan rendah.

Nilai precision untuk kelas $\leq 50K$ mencapai 91,30%, yang mengindikasikan bahwa model sangat akurat dalam mengidentifikasi individu dengan pendapatan rendah. Sebaliknya, precision pada kelas $> 50K$ sebesar 58,00%, menunjukkan bahwa prediksi pendapatan tinggi masih mengandung tingkat kesalahan yang cukup signifikan. Sementara itu, nilai recall kelas $> 50K$ mencapai 75,13%, yang menandakan bahwa sebagian besar individu berpendapatan tinggi berhasil terdeteksi oleh model meskipun dengan risiko false positive yang relatif tinggi.

Pembahasan

Hasil eksperimen menunjukkan bahwa *Logistic Regression* mampu memberikan kinerja yang cukup stabil dalam permasalahan klasifikasi pendapatan berbasis data sosial-ekonomi. Akurasi sebesar 80,91% menegaskan bahwa model linear ini masih relevan untuk memodelkan hubungan antara variabel demografis dan status pendapatan, khususnya ketika tujuan penelitian tidak hanya berfokus pada akurasi, tetapi juga pada interpretabilitas model.

Nilai precision yang tinggi pada kelas $\leq 50K$ menunjukkan bahwa *Logistic Regression* cenderung konservatif dalam mengklasifikasikan pendapatan tinggi, sehingga lebih berhati-hati dalam memberikan prediksi $> 50K$. Hal ini berdampak pada rendahnya precision kelas $> 50K$, tetapi sekaligus menghasilkan nilai *recall* yang cukup baik. Fenomena ini mencerminkan *trade-off* antara *precision* dan *recall* yang umum terjadi pada permasalahan klasifikasi dengan distribusi kelas yang tidak seimbang.

Dari sisi interpretasi, *Logistic Regression* memiliki keunggulan dibandingkan model berbasis ensemble karena setiap variabel independen berkontribusi secara eksplisit terhadap probabilitas prediksi melalui koefisien regresi. Hal ini menjadikan model lebih mudah dianalisis untuk memahami faktor-faktor sosial-ekonomi yang memengaruhi peluang individu berada pada kategori pendapatan tertentu. Dalam konteks kebijakan publik dan analisis sosial, transparansi ini menjadi nilai tambah yang signifikan.

Meskipun model yang lebih kompleks seperti *Random Forest* atau *Gradient Boosting* berpotensi menghasilkan akurasi yang lebih tinggi, hasil penelitian ini menunjukkan bahwa *Logistic Regression* tetap menjadi pilihan yang rasional ketika dibutuhkan keseimbangan antara performa klasifikasi, stabilitas model, dan kemudahan interpretasi. Dengan akurasi di atas 80% dan performa recall yang memadai pada kelas minoritas, model ini dapat dijadikan baseline yang kuat untuk pengembangan sistem prediksi pendapatan berbasis *machine learning*.

KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Logistic Regression* mampu digunakan secara efektif untuk mengklasifikasikan status pendapatan individu berdasarkan variabel sosial-ekonomi dengan pendekatan pembelajaran terawasi. Hasil pengujian menggunakan dataset *Adult Census Income* menghasilkan tingkat akurasi sebesar 80,91%, yang mengindikasikan bahwa model memiliki kemampuan prediksi yang cukup baik dalam membedakan kelas pendapatan $\leq 50K$ dan $> 50K$.

Evaluasi performa menggunakan *confusion matrix*, *precision*, *recall*, dan *F1-score* memperlihatkan bahwa *Logistic Regression* sangat andal dalam mengidentifikasi individu dengan pendapatan $\leq 50K$, ditunjukkan oleh nilai precision yang tinggi pada kelas tersebut. Meskipun demikian, performa pada kelas pendapatan $> 50K$ masih menghadapi tantangan, khususnya pada aspek *precision*, yang dipengaruhi oleh

ketidakseimbangan distribusi kelas dalam dataset. Temuan ini menegaskan pentingnya penggunaan metrik evaluasi yang komprehensif agar kinerja model dapat dinilai secara objektif dan tidak bias terhadap kelas mayoritas.

Dari sisi metodologis, penggunaan RapidMiner Studio terbukti mempermudah proses eksperimen mulai dari pra-pemrosesan data, pembangunan model, hingga evaluasi performa secara terstruktur dan konsisten. Pendekatan ini mendukung replikasi penelitian serta meminimalkan kesalahan implementasi algoritma, sehingga hasil yang diperoleh dapat dipertanggungjawabkan secara ilmiah.

Secara keseluruhan, Logistic Regression tetap relevan sebagai model baseline dalam klasifikasi pendapatan karena keseimbangan antara performa prediksi dan interpretabilitas model. Kemampuan model dalam menjelaskan kontribusi variabel sosial-ekonomi terhadap probabilitas pendapatan menjadikannya sesuai untuk konteks analisis sosial dan kebijakan publik yang menuntut transparansi. Penelitian selanjutnya disarankan untuk mengombinasikan Logistic Regression dengan teknik penanganan ketidakseimbangan kelas atau membandingkannya dengan model non-linear guna meningkatkan performa pada kelas pendapatan tinggi tanpa mengorbankan interpretabilitas.

DAFTAR PUSTAKA

- Alexander, D., & Supriyanto, R. (2023). *Comparative Analysis of Machine Learning Models for Predicting Default in Home Credit Companies*. 671–679.
- Alghamdi, H. M. (2024). *Unveiling Sentiments : A Comprehensive Analysis of Arabic Hajj-Related Tweets from 2017 – 2022 Utilizing Advanced AI Models*.
- Ardana, N. K. K., Amany, N., Kevin, T., Karunia, R., & Fauzia, S. (2023). *Perbandingan Metode KNN , Naive Bayes , dan Regresi Logistik Binomial dalam Pengklasifikasian Status Ekonomi Negara*. 5(2), 404–418.
- Chu, H. (2025a). *Adult Income Prediction and Key Factors Analysis Based on Machine Learning Algorithm*. 0, 143–147. <https://doi.org/10.54254/2754-1169/2025.BL29341>
- Chu, H. (2025b). *Adult Income Prediction and Key Factors Analysis Based on Machine Learning Algorithm*. 0, 142–146. <https://doi.org/10.54254/2754-1169/2025.BL29341>
- Dey, A. R., & Roy, N. (2023). An Investigation into the Prediction of Annual Income Levels through the Utilization of Demographic Features Employing the Modified UCI Adult Dataset. *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 1080–1086. <https://doi.org/10.1109/ICCCIS60361.2023.10425394>
- Islam, M. A., Nag, A., Roy, N., Dey, A., Fahim, S., & Ghosh, A. (2024). *An Investigation into the Prediction of Annual Income Levels Through the Utilization of Demographic Features Employing the Modified UCI Adult Dataset*. <https://doi.org/10.1109/ICCCIS60361.2023.10425394>
- Li, L., Zhao, K., Gan, J., Cai, S., Liu, T., Mu, H., & Sun, R. (2021). Robust Adaptive Semi-supervised Classification Method based on Dynamic Graph and Self-paced Learning. *Information Processing & Management*, 58(1), 102433. <https://doi.org/https://doi.org/10.1016/j.ipm.2020.102433>
- Pradesh, A. (2025). *ADULT INCOME CLASSIFICATION USING MACHINE LEARNING TECHNIQUES*. 16(05), 1195–1209.
- Sharifzadeh, A., Azad, S., & Ameli, M. T. (2023). Modern Smart Multi-Dimensional Infrastructure Energy Systems – State of the Arts. In *Coordinated Operation and Planning of Modern Heat and Electricity Incorporated Networks* (pp. 57–78). IEEE. <https://doi.org/10.1002/9781119862161.ch4>
- Shuvo, S., Mohanty, J., & Patel, D. (2024). *Predicting Annual Income of Individuals using Classification Techniques*. <https://doi.org/10.13140/RG.2.2.30102.15680>

- Siringoringo, R., Arisandi, D., Kurniawan, E., & Nababan, E. B. (2024). *MODEL KLASIFIKASI DENGAN LOGISTIC REGRESSION DAN RECURSIVE CLASSIFICATION MODEL USING LOGISTIC REGRESSION AND RECURSIVE*. 11(4). <https://doi.org/10.25126/jtiik.1148198>
- Susetyoko, R., Yuwono, W., & Purwantini, E. (2022). Model Klasifikasi Pada Seleksi Mahasiswa Baru Penerima KIP Kuliah Menggunakan Regresi Logistik Biner. *Jurnal Informatika Polinema*, 8, 31–40. <https://doi.org/10.33795/jip.v8i4.914>
- Thapa, S. (2023). *Adult Income Prediction Using various ML Algorithms*.
- Wan, Z. (2023). *Performances evaluation of machine learning models on income forecasting*. 0, 24–29. <https://doi.org/10.54254/2755-2721/27/20230111>
- Wang, J. (2022). *Research on Income Forecasting based on Machine Learning Methods and the Importance of Features*. <https://doi.org/10.4108/eai.17-6-2022.2322745>