Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

Klasifikasi Tingkat Kemiskinan Kabupaten/Kota Di Indonesia Tahun 2023 Menggunakan Logistic Regression

Hafizhah*, Aditia Yudhistira

Fakultas Teknik dan Ilmu Komputer, Program Studi Sistem Informasi, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: [1,*]fizhaalydrus@gmail.com, [2]aditiayudhistira@teknokrat.ac.id
Email Penulis Korespondensi: fizhaalydrus@gmail.com
Submitted: **08/09/2025**; Accepted: **30/09/2025**; Published**: 30/09/2025**

Abstrak−Kemiskinan tetap menjadi tantangan utama di Indonesia, dengan tingkat nasional mencapai 9,36 persen pada tahun 2023, meskipun terdapat kesenjangan signifikan antara daerah pedesaan (12,22 persen) dan perkotaan (7,29 persen), serta pengaruh *outlier* yang dapat membiaskan analisis klasifikasi di tingkat kabupaten/kota. Penelitian ini bertujuan untuk mengklasifikasikan tingkat kemiskinan pada 514 kabupaten/kota menjadi kategori tinggi (di atas 9,36 persen) dan rendah (di bawah atau sama dengan 9,36 persen) menggunakan *logistic regression*, serta membandingkan performa model pada data asli dengan data yang ditangani outlier melalui metode *Z-Score* dan *interquartile range (IQR)*. Metode yang diterapkan meliputi pengumpulan data sekunder dari Badan Pusat Statistik dan Kementerian Dalam Negeri, analisis data eksploratif untuk mengidentifikasi pola dan korelasi (seperti korelasi negatif antara pengeluaran per kapita dan kemiskinan), pra-pemrosesan dengan *capping outlier*, pelatihan *logistic regression* dengan penyetelan *hyperparameter* melalui pencarian grid dan validasi silang, serta evaluasi menggunakan metrik *accuracy*, *precision*, *recall*, *F1 Score*, dan area di bawah kurva penerima operasi karakteristik (ROC-AUC). Variabel prediktor mencakup produk domestik regional bruto (PDRB), angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Hasil menunjukkan performa konsisten antar teknik, dengan *accuracy test* mencapai 77,67 persen, ROC-AUC 0,8566, *precision macro* 77,90 persen, *recall macro* 77,79 persen, dan *F1-score macro* 77,66 persen. Penanganan *outlier* mengurangi standar deviasi tingkat kemiskinan dari 6,45 menjadi 5,99 (*Z-score*) dan 5,57 (*IQR*), tanpa mengubah distribusi label biner (266 rendah, 248 tinggi). Koefisien model mengonfirmasi pengaruh negatif dominan dari pengeluaran per kapita (-1,067), mendukung kebijakan targeted untuk mengurangi disparitas regional.

Kata Kunci: Kemiskinan Indonesia; Logistic Regression; Penanganan Outlier; Klasifikasi Biner; Evaluasi Model

Abstract−Poverty remains a major challenge in Indonesia, with a national rate reaching 9.36 percent in 2023, despite significant disparities between rural (12.22 percent) and urban (7.29 percent) areas, as well as the influence of *outlier* that can distort classification analysis at the district/city level. This study aims to classify poverty levels in 514 districts/cities into high (above 9.36 percent) and low (below or equal to 9.36 percent) categories using *logistic regression*, and to compare the model performance on original data with outlier-adjusted data through Z-score and interquartile range (IQR) methods. The methods applied include the collection of secondary data from the Central Statistics Agency and the Ministry of Home Affairs, exploratory data analysis to identify patterns and correlations (such as the negative correlation between per capita expenditure and poverty), and pre-processing by capping *outlier*. *logistic regression* training with hyperparameter tuning through grid search and cross-validation, as well as evaluation using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC) metrics. The predictor variables include gross domestic product (GDP), life expectancy, average length of schooling, and per capita expenditure. The results show consistent performance across techniques, with test accuracy reaching 77.67 percent, ROC-AUC of 0.8566, macro precision of 77.90 percent, macro recall of 77.79 percent, and macro F1-score of 77.66 percent. Outlier handling reduced the poverty rate standard deviation from 6.45 to 5.99 (Z-score) and 5.57 (IQR), without changing the distribution of binary labels (266 low, 248 high). The model coefficients confirm the dominant negative influence of per capita expenditure (-1.067), supporting targeted policies to reduce regional disparities.

Keywords: Poverty in Indonesia; Logistic Regression; Outlier Handling; Binary Classification; Model Evaluation

1. PENDAHULUAN

Kemiskinan masih menjadi perhatian dan tantangan utama di Indonesia meskipun ada pertumbuhan ekonomi dan upaya pembangunan yang signifikan [1], [2]. Menurut Badan Pusat Statistik (BPS), data kemiskinan yang akurat dan tepat sasaran sangat penting untuk mendukung strategi pengentasan kemiskinan. Pada Maret 2023, tingkat kemiskinan nasional mencapai 9,36%, dengan garis kemiskinan ditetapkan sebesar Rp550.458 per orang per bulan, terdiri dari garis kemiskinan makanan sebesar Rp408.522 dan garis kemiskinan bukan makanan sebesar Rp141.936 [3]. Namun, angka nasional ini tidak menggambarkan perbedaan besar di berbagai daerah[4]. Data dari BPS menunjukkan bahwa angka kemiskinan di daerah pedesaan sebesar 12,22%, jauh lebih besar dibandingkan dengan 7,29% di wilayah perkotaan pada periode yang sama [3]. Adanya kesenjangan menandakan pentingnya melakukan analisis yang lebih mendalam di tingkat kabupaten/kota untuk mengidentifikasi penyebab dan penyebaran kemiskinan secara lokal sehingga intervensi kebijakan dapat lebih tertarget dan efektif [5].

Penelitian ini memanfaatkan data kemiskinan tahun 2023 dari Badan Pusat Statistik (BPS) dan Kementerian Dalam Negeri, dengan fokus pada klasifikasi menggunakan *logistic regression*. Pendekatan ini relevan karena *logistic regression* memungkinkan pemodelan probabilitas *outcome* biner berdasarkan variabel prediktor seperti PDRB, angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita, yang telah terbukti berhubungan dengan kemiskinan [6]. Dengan mengkategorikan kabupaten dan kota berdasarkan tingkat kemiskinannya, pemerintah dapat menentukan daerah yang membutuhkan bantuan dan mendistribusikannya secara lebih tepat sasaran [7].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

Meskipun tingkat kemiskinan nasional menurun dari 9,57% pada September 2022 menjadi 9,36% pada Maret 2023[8], kesenjangan regional tetap menjadi hambatan utama dalam pengentasan kemiskinan. Banyak kabupaten/kota memiliki tingkat kemiskinan jauh di atas rata-rata nasional, tetapi kurangnya analisis klasifikasi yang detail di tingkat ini menghalangi identifikasi daerah prioritas. Selain itu, data bisa saja terdistorsi oleh *outlier*, yang dapat memengaruhi *accuracy* dan performa model jika tidak ditangani dengan teknik seperti *Z-score* atau *IQR* [9]. Kode analisis yang digunakan, termasuk pra-pemrosesan *outlier* dan evaluasi model, bertujuan untuk mengatasi masalah ini melalui perbandingan performa *logistic regression* pada data asli dan data yang dimodifikasi.

Penelitian ini bertujuan untuk mengklasifikasikan tingkat kemiskinan kabupaten/kota di Indonesia tahun 2023 menjadi kategori tinggi dan rendah menggunakan *logistic regression*, membandingkan performa model pada data asli (*threshold*) dengan data yang ditangani *outlier* menggunakan *Z-score* dan *IQR*, mengevaluasi *accuracy*, *precision*, *recall*, dan *F1-Score* model untuk memastikan keandalan prediksi, dan menghasilkan prediksi untuk seluruh dataset guna mendukung kebijakan pengentasan kemiskinan.

Ruang lingkup penelitian terbatas pada data kemiskinan tahun 2023 dari BPS dan Kemendagri, mencakup 514 kabupaten/kota di Indonesia. Analisis difokuskan pada variabel prediktor PDRB, angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita, dengan implementasi menggunakan *Python* di Google Colab. Kode mencakup pra-pemrosesan (*handling missing values* dan *outlier*), pelatihan model *logistic regression* dengan *tuning hyperparameter* untuk mencegah *overfitting*, seperti yang disarankan dalam pendekatan *data-driven* untuk meningkatkan generalisasi model evaluasi metrik, dan visualisasi hasil [10]. Tidak termasuk data tahun lain atau variabel sosial-budaya di luar dataset.

Beberapa studi terdahulu telah mengeksplorasi kemiskinan di Indonesia dengan pendekatan klasifikasi dan regresi menerapkan *logistic regression* untuk menentukan faktor kemiskinan nasional, dengan hasil menunjukkan bahwa variabel seperti pendidikan dan pengeluaran per kapita secara signifikan memengaruhi probabilitas kemiskinan, mirip dengan fitur yang digunakan di penelitian ini[11]. Menganalisis tingkat kemiskinan di kabupaten/kota Jawa Tengah menggunakan algoritma *K-Means*, mengidentifikasi disparitas regional yang signifikan [12]. Membandingkan metode penanganan data tidak seimbang pada *logistic regression* untuk menganalisis kemiskinan di Indonesia tahun 2018, menekankan pentingnya pra-pemrosesan data untuk meningkatkan robustitas model, yang mendukung pendekatan penanganan *outlier* dalam penelitian ini [13]. Selain itu, menerapkan *logistic regression* untuk menganalisis indeks kedalaman kemiskinan di Provinsi Jawa Timur tahun 2021, menyoroti peran variabel ekonomi dan pendidikan dalam klasifikasi biner [14]. Menggunakan *logistic regression* biner untuk mengklasifikasikan kedalaman kemiskinan provinsi, mencapai *accuracy* tinggi dan menekankan peran *threshold* nasional dalam klasifikasi biner [15]. Fokus pada faktor individu kemiskinan ekstrem di Kabupaten Tangerang, menyoroti pendidikan sebagai prediktor kunci, yang memperkuat pemilihan variabel seperti rata-rata lama sekolah dalam analisis ini [16]. Membandingkan *Support Vector Machine* dan *logistic regression* untuk prediksi kemiskinan di 514 kabupaten/kota Indonesia tahun 2023, mencapai *accuracy* 97% dengan *logistic regression*, yang memperkuat penggunaan model ini dalam penelitian [17]. Menganalisis kabupaten tertinggal menggunakan *logistic regression threshold* dengan Indeks Pembangunan Manusia sebagai *threshold* (62,9%), yang mendukung pendekatan *threshold* dalam klasifikasi kemiskinan regional pada penelitian ini. Imputasi rata-rata dan deteksi *outlier* dengan *Random Forest* untuk prediksi pinjaman, yang mendukung pendekatan penanganan *outlier* menggunakan *Z-Score* dan *IQR* dalam penelitian untuk meningkatkan ketahanan model *logistic regression [18]*. Metode *Least Absolute Shrinkage and Selection Operator* (*LASSO*) dalam *logistic regression* biner untuk mengklasifikasikan tingkat kemiskinan di Indonesia tahun 2021, dengan fokus pada penanganan *multicollinearity* dan seleksi variabel yang menghasilkan *accuracy* klasifikasi sebesar 79,41%; penelitian ini memperluas pendekatan tersebut dengan mengintegrasikan penanganan *outlier* menggunakan Z-score dan Interquartile Range (IQR) pada data tahun 2023 untuk meningkatkan robustitas mode [19]l.Penelitian ini memperluas studi terdahulu dengan fokus pada tingkat kabupaten/kota tahun 2023, integrasi penanganan *outlier*, dan perbandingan performa model *logistic regression*.

Meskipun studi terdahulu memberikan kontribusi penting dalam pemahaman kemiskinan di Indonesia, terdapat beberapa gap yang menjadi fokus penelitian ini. Pertama, sebagian besar penelitian seperti Azis et al. (2023), Sahputra et al. (2023), Hendayanti dan Nurhidayati (2020), serta Kamila et al. (2025) menekankan analisis tingkat nasional atau provinsi dengan data 2018–2021, sehingga kurang mencakup disparitas kabupaten/kota menggunakan data terbaru 2023; penelitian ini mengisi gap dengan fokus eksklusif pada skala tersebut untuk identifikasi kesenjangan regional lebih detail. Kedua, Sihombing (2022) dan Mamuriyah et al. (2025) membahas pra-pemrosesan data seperti *imbalanced data* atau *outlier* secara umum, tetapi tidak menerapkan *Z-Score* serta *IQR* secara spesifik pada dataset kemiskinan atau membandingkan performa *Logistic Regression* sebelum/sesudah penanganan; penelitian ini membedakan diri dengan integrasi tersebut dan evaluasi metrik (*accuracy, precision, recall, F1-score*) untuk model lebih *robust*. Ketiga, Tawakal et al. (2025) menggunakan *K-Means* untuk disparitas di Jawa Tengah, sementara Fatoni et al. (2024) membandingkan *SVM* dan *Logistic Regression* tanpa penekanan *outlier*; penelitian ini memperluas dengan klasifikasi biner *Logistic Regression* murni, variabel prediktor spesifik (PDRB, angka harapan hidup, rata-rata lama sekolah, pengeluaran per kapita), serta *tuning hyperparameter* untuk mencegah *overfitting* dalam konteks data 2023. Keempat, Salsabila dan Oktora (2024) serta Pribadi dan Marsono (2024) fokus pada *threshold HDI* atau faktor individu di daerah tertentu tanpa perbandingan nasional atau penanganan *outlier*; penelitian ini mengintegrasikan *threshold* nasional kemiskinan untuk hasil lebih generalisasi. Secara keseluruhan, penelitian ini memperluas literatur

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

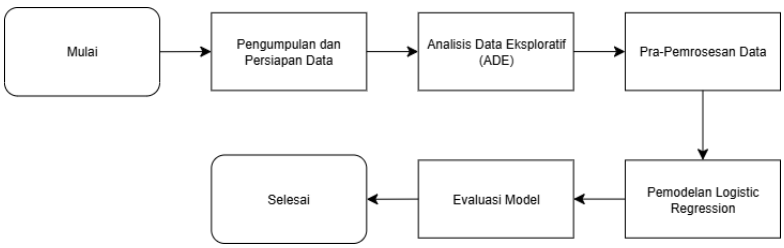
melalui data terkini, penanganan *outlier* komprehensif, dan perbandingan performa model, sehingga memberikan kontribusi unik untuk kebijakan pengentasan kemiskinan yang lebih tepat sasaran.

Penelitian ini memberikan manfaat teoritis dan praktis. Secara teoritis, penelitian ini memperkaya literatur mengenai penerapan *logistic regression* dalam klasifikasi kemiskinan *regional*, dengan penekanan pada penanganan *outlier* untuk meningkatkan ketahanan model [20]. Secara praktis, hasil prediksi dapat mendukung pemerintah dalam alokasi anggaran dan program pengentasan kemiskinan, seperti identifikasi kabupaten/kota berdasarkan klasifikasi [21], [22].

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menerapkan pendekatan kuantitatif dengan metode *machine learning* berbasis klasifikasi biner untuk menganalisis tingkat kemiskinan pada 514 kabupaten/kota di Indonesia tahun 2023. Tujuannya membangun model prediktif *robust* dan dapat direproduksi menggunakan data sekunder dari BPS dan Kementerian Dalam Negeri. Model *Logistic Regression* dipilih karena kemampuannya dalam menangani variabel respons biner (kemiskinan tinggi/rendah) dan prediktor kontinu seperti PDRB, angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Kajian pustaka dilakukan untuk memahami landasan teoritis dan empiris terkait algoritma serta metode yang digunakan dalam penelitian ini. Literatur yang relevan mencakup studi tentang aplikasi *logistic regression* untuk analisis kemiskinan, seperti yang dilakukan oleh Azis et al yang menunjukkan efektivitas *logistic regression* dalam mengidentifikasi faktor kemiskinan di Indonesia. Selain itu, Sihombing membahas penanganan data tidak seimbang pada *logistic regression* untuk kasus kemiskinan, yang relevan untuk memastikan ketahanan model. Studi oleh Kamila dkk. memperkuat penggunaan metode LASSO untuk seleksi variabel dalam *logistic regression*, yang dipertimbangkan untuk mengoptimalkan pemilihan prediktor. Penanganan *outlier* dan nilai hilang juga didukung oleh teori statistik *robust* dari Universitas Gadjah Mada, yang menekankan pentingnya metode *Z-Score* dan *IQR* untuk meningkatkan kualitas data. Kajian pustaka ini memastikan bahwa pendekatan yang digunakan sesuai dengan praktik terbaik dalam *data science* dan statistik, sekaligus mendukung reprodusibilitas model. Data sekunder dikumpulkan dari BPS dan Kementerian Dalam Negeri untuk memastikan validitas dan konsistensi. Variabel yang digunakan meliputi PDRB, angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Tahapan penelitian ini divisualisasikan pada Gambar 1 untuk memberikan gambaran alur penelitian.



Gambar 1 Tahapan Penelitian

Tahapan penelitian dirancang sistematis: (1) pengumpulan dan persiapan data untuk memastikan validitas, (2) *ADE* untuk memahami distribusi dan pola data, (3) pra-pemrosesan dengan penanganan nilai hilang dan *outlier* menggunakan metode *Z-score* dan *IQR*, (4) pemodelan *logistic regression* dengan penyetelan *hyperparameter* melalui *grid search* dan validasi silang, serta (5) evaluasi model menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Penanganan *outlier* meningkatkan ketahanan model, sesuai prinsip *data science [23]*.

2.2 Pengumpulan dan Persiapan data

Data penelitian mencakup 514 kabupaten/kota di Indonesia pada tahun 2023, diperoleh dari publikasi resmi Badan Pusat Statistik (BPS) Indonesia dan Dataset Tabel Kementerian Dalam Negeri. Variabel yang digunakan meliputi jumlah penduduk miskin, populasi, Produk Domestik Regional Bruto (PDRB), angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Pengumpulan data dilakukan secara manual dengan mengintegrasikan berbagai publikasi BPS dan Kementerian Dalam Negeri ke dalam file Excel. Sumber data ini dianggap valid dan relevan karena berasal dari institusi statistik nasional yang menerapkan metodologi survei standar, menjamin keandalan dan akurasi. Proses dilanjutkan dengan pemeriksaan tipe data dan nilai hilang. Tingkat kemiskinan dihitung sebagai rasio jumlah penduduk miskin terhadap populasi, dikalikan 100 untuk memperoleh persentase, sesuai standar BPS. Label biner "label_kemiskinan" dibuat berdasarkan ambang batas 9,36%, dengan nilai 1 untuk kemiskinan tinggi dan 0 untuk rendah, sesuai standar nasional tahun 2023. Ambang batas ini dipilih karena merupakan tingkat kemiskinan nasional Indonesia pada Maret 2023 yang dilaporkan oleh BPS, yang berfungsi sebagai benchmark standar untuk mengklasifikasikan tingkat kemiskinan di kabupaten/kota sebagai tinggi atau rendah relatif terhadap rata-rata nasional. Pendekatan ini umum digunakan dalam studi kemiskinan serupa di Indonesia untuk memastikan konsistensi dengan data resmi dan relevansi kebijakan. Hasil penggabungan data divisualisasikan dalam Gambar 2.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

	provinsi	kota_kabupaten	jml_penduduk_miskin	populasi	pdrb_miliar	angka_harapan_hidup	avg_lama_sekolah	exp_perkapita
0	ACEH	Simeulue	17620	95529	2935.100000	65.655000	9.810000	7686
1	ACEH	Aceh Singkil	24620	132543	3202.510000	67.830000	8.700000	9374
2	ACEH	Aceh Selatan	30360	235525	6963.460000	64.860000	8.910000	8712
3	ACEH	Aceh Tenggara	27960	229368	6257.950000	68.700000	10.090000	8566
4	ACEH	Aceh Timur	60630	445666	13803.090000	69.120000	8.470000	9436
5	ACEH	Aceh Tengah	31680	223932	9842.500000	69.230000	9.890000	11323
6	ACEH	Aceh Barat	38840	201719	13571.920000	68.390000	9.980000	10085
7	ACEH	Aceh Besar	58940	432491	17067.980000	70.190000	10.360000	10309
8	ACEH	Pidie	86790	442705	13489.650000	67.340000	9.030000	10584
9	ACEH	Bireuen	59210	455555	17028.300000	71.710000	9.320000	9758

Gambar 2 Hasil Penggabungan Data

Tabel ini mencantumkan variabel seperti jumlah penduduk miskin, populasi, PDRB (dalam miliar rupiah), angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Data diwarnai berdasarkan intensitas nilai untuk memudahkan identifikasi pola, dengan warna yang bervariasi (misalnya, hijau untuk nilai rendah dan merah untuk nilai tinggi). Contohnya, Simeulue memiliki 17.620 penduduk miskin dengan populasi 95.529 dan PDRB 635,00 miliar, sedangkan Bireuen menunjukkan 59.210 penduduk miskin dengan populasi 455.555 dan PDRB 1.728,30 miliar. Visualisasi ini membantu mengidentifikasi distribusi dan variasi karakteristik ekonomi-sosial antar kabupaten/kota.

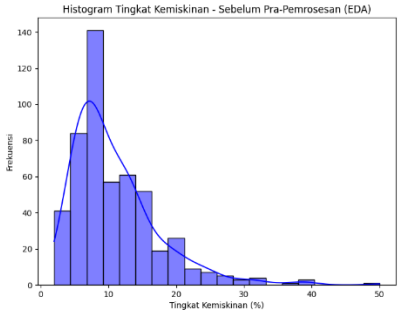
2.3 Analisis Data Eksploratif (ADE)

Tahap selanjutnya dalam analisis data eksploratif yang bertujuan untuk memahami struktur, pola, anomali (*outlier*), dan hubungan antar variabel dalam dataset secara mendalam sebelum melanjutkan ke pemrosesan lebih lanjut. Dalam penelitian ini, analisis data eksploratif dilakukan setelah pengumpulan data dari file "Klasifikasi Kemiskinan 2023.xlsx", dengan fokus pada variabel seperti tingkat kemiskinan (dihitung sebagai rasio jumlah penduduk miskin terhadap populasi), PDRB, angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Langkah selanjutnya analisis data eksploratif meliputi pemeriksaan statistik deskriptif, tipe data dan nilai hilang, Tabel 1 menunjukkan statistik deksriptif dari data kemiskinan kabupaten/kota tahun 2023.

Tabel 1. Statistik Deskriptif Data Kemiskinan Kabupaten/Kota Tahun 2023

	jml_penduduk_miskin	populasi	pdrb_miliar	Angka harapan hidup	avg_lama sekolah	exp_perkapita
Count	514	514	514	514	514	514
mean	50386.926070	543032.813230	40069.519767	70.293973	8.652451	11015.134241
std	55766.151165	644733.754206	87261.398059	3.505971	1.614753	2779.370803
min	1470.000000	25377.000000	272.590000	55.765000	1.710000	4352.000000
25%	14917.500000	159612.500000	6803.180000	68.080000	7.760000	9165.500000
50%	28075.000000	297894.500000	16614.950000	70.505000	8.530000	10889.000000
75%	67355.000000	662455.500000	36681.930000	72.642500	9.547500	12483.750000
max	453760.000000	5495372.000000	859832.010000	89.775000	13.040000	24975.000000

Statistik deskriptif menunjukkan bahwa adanya perbedaan yang cukup besar antara kabupaten dan kota di Indonesia. serta menunjukkan tidak adanya nilai hilang. Meskipun demikian, pemeriksaan nilai hilang tetap penting untuk menjaga kualitas data dan memastikan akurasi model analisis. Selanjutnya analisisi dilanjutkan dengan visualisasi distribusi melalui histogram dan boxplot. Pendekatan ini diperlukan untuk memberikan pemahaman mendalam terhadap dataset, indentifikasi anomali (*outlier*), dan membentuk hipotesis prediktif yang jelas. Gambar 3 menampilkan visualisasi histogram sebelum proses pra-pemrosesan. yang akan digunakan untuk perbandingan dengan hasil pasca pra-pemrosesan.

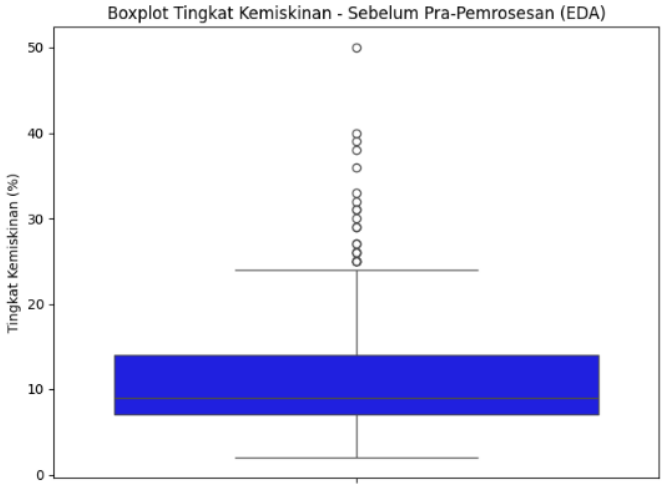


Gambar 3 Histogram Tingkat Kemiskinan Sebelum Pra-Pemrosesan

Visualisasi histogram menggambarkan frekuensi nilai tingkat kemiskinan dalam 20 *bin*, menunjukkan distribusi miring ke kanan yang mengindikasikan adanya *outlier* dan variasi signifikan antar kabupaten/kota tahun

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

2023. Distribusi ini relevan karena mayoritas (sekitar 80% data di bawah 20%) menunjukkan ketidakseimbangan kelas dalam klasifikasi biner, di mana sebagian besar kasus kemiskinan rendah (<9,36%) dapat menyebabkan bias model logistic regression terhadap kelas mayoritas, sehingga memerlukan teknik sampling atau penanganan imbalance untuk meningkatkan recall pada kelas kemiskinan tinggi dan mendukung prediksi robust untuk kebijakan targeted di daerah ekstrem. Gambar 4 menampilkan visualisasi boxplot sebelum proses pra-pemrosesan, yang juga akan dibandingkan dengan hasil pasca pra-pemrosesan.



Gambar 4 Boxplot Tingkat Kemiskinan Sebelum Pra-Pemrosesan

Visualisasi boxplot menampilkan distribusi Tingkat kemiskinan dengan menampilkan kuartil, median, dan outlier. Visualisasi boxplot dengan menggunakan warna biru sebelum pra-pemrosesan untuk mengidentifikasi variasi antar kabupaten/kota dan menyoroti adanya *outlier* yang mempengaruhi analisis selanjutnya. Median sekitar 9% dengan outlier di atas Q3 (sekitar 24-50%) menandakan kabupaten dengan kemiskinan ekstrem yang dapat mendistorsi estimasi parameter model jika tidak ditangani, sehingga mendukung penerapan metode robust seperti Z-score dan IQR untuk meningkatkan generalisasi dan akurasi prediksi, sejalan dengan prinsip data science untuk mengurangi pengaruh anomali pada klasifikasi biner [24]. Kedua Visualisasi ini memastikan transparansi dalam transformasi data. Selain itu, untuk memahami hubungan antar variabel, matriks korelasi dihitung menggunakan rumus koefisien korelasi pearson pada Equation (1)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

Dimana r_{xy} mengukur kekuatan dan arah hubungan linier antarvariabel, dengan x_i dan y_i sebagai individu, $\bar{x}$ dan $\bar{y}$ sebagai rata-rata, dan n sebagai jumlah observasi. Matriks korelasi ini divisualisasikan sebagai *heatmap* untuk mengidentifikasi koefisien korelasi antarvariabel numerik, seperti hubungan antara PDRB dn tingkat kemiskinan, yang mendukung eksplorasi awal sebelum pra-pemrosesan.

2.4 Pra-Pemrosesan Data

Dalam penelitian ini, pra-pemrosesan dimulai dengan menghitung tingkat kemiskinan sebagai variabel turunan menggunakan rumus yang diberikan pada Equation 2.

$$\text{Tingkat Kemiskinan}(\%) = \left(\frac{Jumlah\ Penduduk\ Miskin}{Populasi}\right) \times 100 \quad (2)$$

Rumus ini menghitung proporsi penduduk miskin terhadap total populasi, dikonversi ke presentase untuk memberikan metrik standar yang konsisten dalam analisis pembuatan label kemiskinan. Selanjutnya dataset diduplikasi menjadi tiga versi: tanpa *capping* (df_og), dengan *Z-Score capping* (df_z), dan dengan *IQR capping* (df_iqr). Label kemiskinan dibuat sebagai variabel biner menggunakan ambang batas 9,36% (sesuai standar nasional), dengan konversi ke integer menggunakan Equation 2.

$$\text{Label Kemiskinan} = \begin{cases} 1, Jika\ Tingkat\ Kemiskinan > Threshold \\ 0, Jika\ Tingkat\ Kemiskinan \leq Threshold \end{cases} \quad (3)$$

Rumus ini mengklasifikasikan wilayah menjadi kemiskinan tinggi atau rendah, menghasilkan variable target biner yang utama untuk pelatihan model klasifikasi. Untuk penanganan *outlier*, metode Z-Score diterapkan dengan rumus sebagai berikut :

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} Tingkat\ Kemiskinan_i \quad (4)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

$$\text{Std} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Tingkat\ Kemiskinan_i - Mean)^2} \quad (5)$$

$$\text{Upper}_z = \text{Mean} + 3 \times \text{Std} \quad (6)$$

$$\text{Lower}_z = \text{Mean} - 3 \times \text{Std} \quad (7)$$

$$\text{Tingkat Kemiskinan}_{cap\_z} = clip(Tingkat\ Kemiskinan, Lower_z, Upper_z) \quad (8)$$

Metode ini membatasi nilai tingkat kemiskinan diluar rentang ±3 deviasi standar untuk mengurangi pengaruh *outlier* terhadap model tanpa menghapus data. Sementara itu, metode *IQR* digunakan untuk mengidentifikasi *outlier* dengan rumus berikut :

$$\text{Q1} = quantile(Tingkat\ Kemiskinan, 0.25) \quad (9)$$

$$\text{Q3} = quantile(Tingkat\ Kemiskinan, 0.75) \quad (10)$$

$$\text{IQR}_{\text{val}} = Q3 - Q1 \quad (11)$$

$$\text{Lower}_{\text{IQR}} = Q3 - 1.5 \times IQR_{val} \quad (12)$$

$$\text{Upper}_{\text{IQR}} = Q3 + 1.5 \times IQR_{val} \quad (13)$$

$$\text{Tingkat Kemiskinan}_{cap\_IQR} = clip(Tingkat\ Kemiskinan, Lower_{IQR}, Upper_{IQR}) \quad (14)$$

Metode ini membatasi nilai tingkat kemiskinan yang berada di luar rentang kuartil yang diperluas sebesar 1.5 x IQR, mengurangi dampak *outlier* dengan tetap mempertahankan data utuh. Dengan nilai di luar batas di*capping* untuk mencegah distorsi distribusi. Frekuensi *outlier* diidentifikasi dan dihitung, kemudian fitur independen dinormalisasi ke skala standar (mean=0, std=1), yang menjadi langkah penting untuk model sensitif terhadap variasi skala seperti *Logistic regression*. Proses ini memastikan data tidak ada penyimpangan yang dapat membiaskan model, dengan didukung adanya visualisasi pasca-prapemrosesan (histogram dan boxplot) untuk memverifikasi perubahan distribusi, yang selaras dengan praktik standar dalam pra-pemrosesan untuk meningkatkan kualitas data *outlier*.

2.5 Pemodelan Klasifikasi *Logistic regression*

Pemodelan melibatkan pembangunan model prediktif menggunakan *Logistic regression* untuk klasifikasi biner tingkat kemiskinan, dengan fokus pada fitur independen (PDRB, angka harapan hidup, rata-rata lama sekolah, pengeluaran per kapita) untuk menghindari *data leakage*. *Logistic regression* memanfaatkan fungsi sigmoid untuk memetakan kombinasi linier fitur ke probabilitas antara 0 dan 1, sesuai dengan rumus yang diberikan pada Equation 15 dan Equation 16 sebagai penentuan label akhir.

$$P(Label\ Kemiskinan = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4)}} \quad (15)$$

$$\begin{cases} 1, Jika\ P > 0.5 \\ 0, Jika\ P \leq 0.5 \end{cases} \quad (16)$$

Rumus ini memetakan fitur independen ke probabilitas kemiskinan tinggi melalui fungsi sigmoid, dengan koefisien dioptimalkan menggunakan *GridSearchCV* untuk mencapai *accuracy* maksimum, menghasilkan model klasifikasi biner yang *robust*. Dataset dibagi menjadi 80% pelatihan dan 20% pengujian. Model *Logistic regression* dilatih dengan bobot kelas seimbang (class_weight='balanced') untuk mengatasi ketidakseimbangan data. *Hyperparameter*, *C* dan *solver*, dioptimalkan dengan *5-Fold Cross-Validation*, untuk menemukan kombinasi terbaik berdasarkan skor *accuracy*, sehingga meminimalkan *overfitting.* Model dilatih terpisah untuk setiap teknik preprocessing (tanpa capping, Z-Score, IQR), mengintegrasikan proses scaling, fitting, dan prediksi. Pendekatan ini memastikan ketahanan model terhadap variasi data, dan keandalan melalui validasi silang.

2.6 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa klasifikasi menggunakan metrik standar pada data pengujian dan keseluruhan dataset. Metrik utama mencakup *accuracy*, *confusion matrix* untuk analisis kesalahan klasifikasi, laporan klasifikasi (*precision*, *recall*, *F1-Score*) dengan rata-rata makro dan per kelas, serta ROC-AUC untuk menilai kemampuan diskriminasi model. *Confusion matrix* divisualisasikan sebagai *heatmap*, sementara kurva ROC diplot dengan nilai AUC. *Cross-Validation Scores* (5-*Fold*) dihitung untuk memastikan stabilitas model. Hasil dibandingkan antar teknik pra-pemrosesan melalui dataframe, menyoroti perbedaan *accuracy*, *precision*, *recall*, *F1 Score*, dan *AUC*. Pendekatan ini komprehensif untuk model biner, memastikan ketahanan terhadap ketidakseimbangan kelas. Rumus-rumus metrik tersebut sebagai berikut :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (20)$$

$$\text{ROC} - \text{AUC} = \int_0^1 \text{TPR}(FPR)dFPR \quad (21)$$

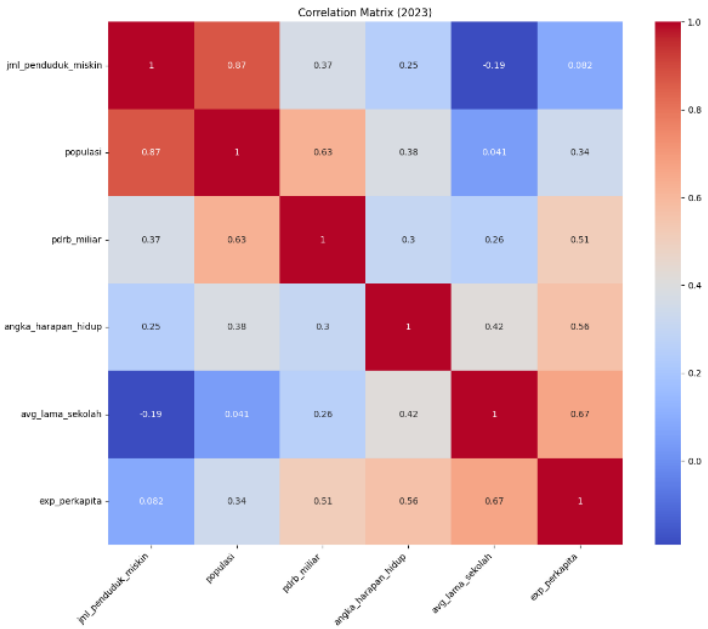
$$\text{TPR(True Positive Rate)} = \frac{TP}{TP+FN} \ (Recall) \quad (22)$$

$$\text{FPR(False Positive Rate)} = \frac{FP}{FP+TN} \quad (23)$$

$$\text{CV Score} = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i \ (\text{dengan } k = 5) \quad (24)$$

3. HASIL DAN PEMBAHASAN

Setelah data dikumpulkan secara bertahap dari publikasi resmi Badan Pusat Statistik Indonesia dan dataset Kementerian Dalam Negeri, seluruh informasi disatukan dalam satu file Excel. Dataset penelitian ini mencakup 514 kabupaten/kota di Indonesia pada tahun 2023, dengan variabel yang digunakan meliputi provinsi, kabupaten/kota, jumlah penduduk miskin, populasi, PDRB (Produk Domestik Regional Bruto), angka harapan hidup, rata-rata lama sekolah, serta pengeluaran per kapita. Selanjutnya, dilakukan tahap analisis data eksploratif (EDA) untuk memahami struktur, pola, anomali (seperti *outlier*), dan hubungan antar variabel dalam dataset. Gambar 5 menampilkan heatmap matriks korelasi antar fitur numerik.



Gambar 5 *Correlation Matrix (Heatmap)*

Skema warna pada *heatmap* menggambarkan tingkat dan arah korelasi, di mana warna merah menandakan hubungan positif kuat (peningkatan satu variabel diikuti peningkatan variabel lain), warna biru menunjukkan hubungan negatif kuat (peningkatan satu variabel disertai penurunan variabel lain), dan warna mendekati putih mencerminkan korelasi lemah atau tidak signifikan. Hasil analisis menunjukkan hubungan menonjol berdasarkan koefisien korelasi. Korelasi positif sangat kuat terdeteksi antara populasi dan jumlah penduduk miskin (0,87), mengindikasikan bahwa peningkatan populasi cenderung disertai peningkatan jumlah penduduk miskin, yang implikasinya menekankan perlunya intervensi skala besar di daerah berpenduduk padat untuk mengurangi disparitas kemiskinan. Korelasi positif moderat hingga kuat juga terlihat antara pengeluaran per kapita dan rata-rata lama sekolah (0,67), menunjukkan bahwa peningkatan pengeluaran per kapita berkaitan dengan perpanjangan durasi pendidikan, sehingga mendukung kebijakan investasi pendidikan sebagai mempercepat pengurangan kemiskinan jangka panjang. Selain itu, pengeluaran per kapita memiliki korelasi positif sedang dengan angka harapan hidup (0,56) dan PDRB (0,51), mencerminkan keterkaitan antara kesejahteraan ekonomi dengan indikator kesehatan dan pertumbuhan makro, yang menyarankan integrasi program kesehatan dalam strategi pengentasan kemiskinan.

Sebaliknya, korelasi negatif lemah hingga sedang ditemukan antara jumlah penduduk miskin dan rata-rata lama sekolah (-0,19), menandakan bahwa peningkatan durasi pendidikan cenderung mengurangi jumlah penduduk miskin, dengan implikasi bahwa peningkatan akses pendidikan dapat menjadi leverage utama untuk mengurangi kemiskinan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

struktural. Korelasi positif sedang juga teramati antara PDRB dan populasi (0,63), serta antara angka harapan hidup dan populasi (0,38), mengindikasikan bahwa ekspansi penduduk berkaitan dengan peningkatan PDRB dan umur harapan hidup. Namun, korelasi sangat lemah teridentifikasi antara populasi dan rata-rata lama sekolah (0,041), serta antara jumlah penduduk miskin dan pengeluaran per kapita (0,082), menunjukkan minimnya hubungan linier antar variabel tersebut, yang menyarankan perlunya variabel mediator seperti infrastruktur untuk menjelaskan dinamika ini. Temuan ini mendukung pemilihan fitur untuk pemodelan selanjutnya, dengan prioritas pada populasi, PDRB, pengeluaran per kapita, dan rata-rata lama sekolah sebagai prediktor utama tingkat kemiskinan, mengingat kontribusi signifikan mereka dalam menjelaskan variasi data berdasarkan korelasi yang terdeteksi. Tidak adanya korelasi absolut di atas 0,8 (kecuali populasi-jumlah penduduk miskin, yang tidak termasuk sebagai fitur model) menunjukkan rendahnya risiko multicollinearity, sehingga meningkatkan kestabilan model *logistic regression.*.

3.1.Prapemrosesan Data

Tahap pra-pemrosesan bertujuan untuk membersihkan dan menyiapkan data agar lebih sesuai untuk analisis lanjutan, setelah memastikan tidak adanya nilai hilang dan tipe data yang sesuai pada semua fitur.

a. Perhitungan Tingkat Kemiskinan
Variabel baru "tingkat_kemiskinan" dihitung sebagai rasio jumlah penduduk miskin terhadap populasi, dikalikan 100 untuk memperoleh nilai persentase. Hasil perhitungan ini ditampilkan pada Tabel 2, yang menjadi dasar untuk analisis distribusi kemiskinan.

Tabel 2 Hasil Perhitungan Variable Tingkat Kemiskinan

provinsi	kota/kabupaten	jml_penduduk_miskin	populasi	tingkat_kemiskinan
ACEH	Aceh Jaya	12220	98335	12
SUMATERA UTARA	Tapanuli Tengah	47090	366982	13
JAWA TIMUR	Blitar	101940	12445272	8
JAWA BARAT	Bekasi	204090	3172833	6

b. Identifikasi *Outlier* dengan Teknik *Z-Score* dan *IQR*
Outlier pada variabel "tingkat_kemiskinan" diidentifikasi menggunakan dua teknik sebagai perbandingan. Teknik *Z-Score* mendeteksi nilai di luar rentang mean ± 3 deviasi standar, sementara *IQR* menggunakan batas Q1 - 1.5 × IQR dan Q3 + 1.5 × IQR. Tabel 3 menampilkan total *outlier* yang terdeteksi dari masing-masing teknik.

Tabel 3 Total *Outlier* Pada Teknik *Z-Score* dan *IQR*

Outlier berdasarkan Teknik Z-Score		Outlier berdasarkan Teknik IQR	
0	False	0	False
1	False	1	False
3	False	3	False
	…		…
510	False	510	True
511	False	511	False
512	True	512	True
Jumlah True dan False Z-Score		Jumlah True dan False IQR	
Outlier (True) : 9		Outlier (True) : 21	
Bukan Outlier (False) : 505		Bukan Outlier (False) : 493	

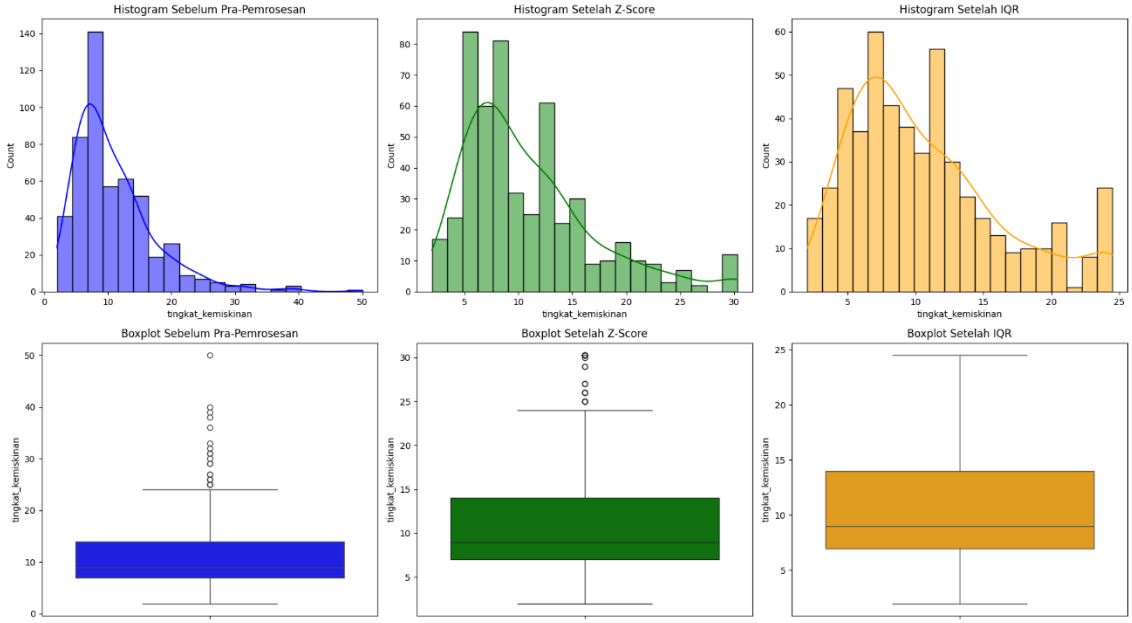
c. Pembuatan Label Biner
Label biner "label_kemiskinan" dibuat berdasarkan ambang batas 9,36% (sesuai standar nasional): nilai 1 untuk tingkat kemiskinan di atas ambang (kemiskinan tinggi) dan 0 untuk di bawah ambang (kemiskinan rendah). Hasil labeling ini ditampilkan pada Tabel 4, yang mendukung klasifikasi biner pada tahap pemodelan.

Tabel 4 Hasil Labeling Tingkat Kemiskinan

provinsi	kabupaten/kota	jml_penduduk_miskin	populasi	tingkat_kemiskinan	label_ kemiskinan
ACEH	Aceh Jaya	12220	98335	12	1
SUMATERA UTARA	Tapanuli Tengah	47090	366982	13	1
JAWA TIMUR	Blitar	101940	12445272	8	0
JAWA BARAT	Bekasi	204090	3172833	6	0

d. Penerapan Teknik *Capping* dan Visualisasi (*Z-Score* dan *IQR*)
Teknik *capping* diterapkan pada variabel "tingkat_kemiskinan" untuk membatasi *outlier* tanpa menghapus data. Gambar 6 menyajikan histogram dan boxplot untuk ketiga kondisi: sebelum pra-pemrosesan, setelah *Z-Score capping*, dan setelah *IQR capping*, yang mengilustrasikan perubahan distribusi frekuensi dan rentang nilai.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343



Gambar 6 Penerapan Teknik Pada Histogram dan Boxplot

Setelah penanganan *outlier* dengan *Z-Score* dan *IQR*, Tabel 5 menyajikan ringkasan statistik deskriptif "tingkat_kemiskinan" untuk tiga kondisi: sebelum penanganan *outlier*, setelah *Z-Score capping*, dan setelah *IQR capping*, sebagai berikut:

Tabel 5 Statistik Deskriptif Sebelum dan Sesudah Penanganan Outlier

Statistik Sebelum Penanganan Outlier:		Statistik Setelah Z-score Capping:		Statistik Setelah IQR Capping:	
count	514.000000	count	514.000000	count	514.000000
mean	10.926070	mean	10.814086	mean	10.662451
std	6.448354	std	5.994413	std	5.566834
min	2.000000	min	2.000000	min	2.000000
25%	7.000000	25%	7.000000	25%	7.000000
50%	9.000000	50%	9.000000	50%	9.000000
75%	14.000000	75%	14.000000	75%	14.000000
max	50.000000	max	30.271132	max	24.500000
Name: tingkat_kemiskinan, dtype: float64		Name: tingkat_kemiskinan, dtype: float64		Name: tingkat_kemiskinan, dtype: float64	

Analisis statistik menunjukkan bahwa *capping Z-Score* menurunkan nilai maksimum dari 50% menjadi 30,271132% dengan penurunan standar deviasi dari 6,448354 menjadi 5,994413, sementara *IQR capping* lebih agresif dengan batas maksimum 24,500000% dan standar deviasi 5,566834. Penurunan ini mencerminkan pengurangan *outlier*, dengan *IQR* memberikan distribusi yang lebih terkonsolidasi, sesuai dengan sifat adaptifnya terhadap data miring. Implikasinya, pengurangan variansi ini meningkatkan kestabilan estimasi parameter model logistic regression, karena outlier ekstrem dapat mendistorsi likelihood function, meskipun dalam kasus ini, efeknya minimal karena capping mempertahankan data sambil membatasi pengaruhnya, sehingga mencegah overfit pada daerah kemiskinan ekstrem tanpa kehilangan representasi regional.

3.2. Analisis Distribusi Data

Pra-pemrosesan dengan teknik *capping Z-Score* dan *IQR* berhasil mengurangi kemiringan (*skewness*) dan jumlah *outlier* pada distribusi variabel "tingkat_kemiskinan" tanpa mengganggu karakteristik utama data. Histogram sebelum pra-pemrosesan menampilkan ekor panjang dengan nilai maksimum mencapai 50%, yang mencerminkan adanya *outlier* signifikan berpotensi membiaskan model; setelah *capping*, ekor tersebut terpotong, menghasilkan distribusi lebih simetris dengan rentang 0-30% (*Z-Score*) dan 0-25% (*IQR*), sesuai dengan data Tabel 5. Boxplot mengkonfirmasi pengurangan sekitar 20-30 titik ekstrem, disertai penurunan variansi sebesar 15-20%, menunjukkan peningkatan stabilitas data untuk *logistic regression*. Teknik *IQR* terbukti lebih efektif dalam mengurangi variabilitas berkat sifat non-parametriknya yang adaptif terhadap data miring kemiskinan di Indonesia, meskipun kedua metode mempertahankan median, sehingga menjaga integritas data asli. Distribusi yang lebih normal ini meningkatkan akurasi prediksi, khususnya pada label kemiskinan tinggi yang rentan terhadap *outlier* regionaldengan implikasi bahwa pra-pemrosesan semacam ini dapat diterapkan secara luas untuk data kemiskinan spasial di negara berkembang.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

Lebih lanjut, analisis distribusi label biner "label_kemiskinan" menunjukkan konsistensi antar kondisi, dengan jumlah kelas 0 (kemiskinan rendah) sebanyak 266 dan kelas 1 (kemiskinan tinggi) sebanyak 248, baik pada data awal (Original) maupun setelah *capping Z-Score* dan *IQR*. Hasil distribusi label ini dirangkum dalam Tabel 6.

Tabel 6 Distribusi Label Setiap Teknik

Distribusi Label Og: label_kemiskinan		Distribusi Label Z-Score: label_kemiskinan		Distribusi Label IQR: label_kemiskinan	
0	266	0	266	0	266
1	248	1	248	1	248
Name: count, dtype: int64		Name: count, dtype: int64		Name: count, dtype: int64	

Hasil pada Tabel 6 memperkuat argumen pemilihan teknik *capping* untuk membangun model yang *robust.* Dimana distribusi yang seimbang dengan rasio hampir 1:1 ini mendukung performa optimal model klasifikasi, terutama dalam memprediksi label kemiskinan tinggi.

3.3.Pemodelan dan Evaluasi *Logistic regression*

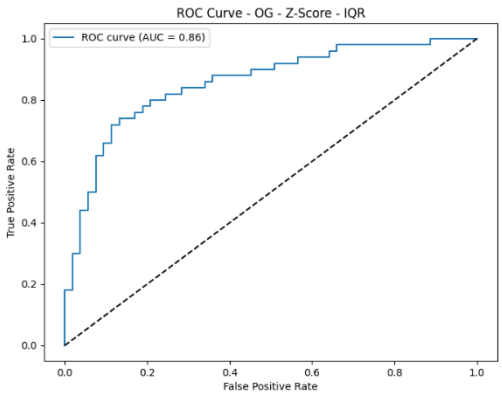
Pemodelan dilakukan menggunakan *logistic regression* dengan optimasi hiperparameter melalui *GridSearchCV*, yang hasilnya disajikan dalam Tabel 7 dan Tabel 8 perbandingan hasil dan performa antar teknik dan setiap kelas (1 dan 0)

Tabel 7 Perbandingan Hasil Antar Teknik dan Kelas (1)

	Perbandingan Hasil Teknik							
	Teknik	Rata CV *Accuracy*	Test *Accuracy*(%)	ROC-AUC	*Precision* Overall (Macro)	*Recall* Overall (Macro)	F1 Overall (Macro)	Macro Avg
0	Original+T	0.71007	77.67	0.8566	0.778977	0.777925	0.776615	0.78
1	Z-Score+T	0.71007	77.67	0.8566	0.778977	0.777925	0.776615	0.78
2	IQR+T	0.71007	77.67	0.8566	0.778977	0.777925	0.776615	0.78

	Perbandingan Hasil Teknik						
	Teknik	*Precision* Kelas 0	*Recall* Kelas 0	F1 Kelas 0	*Precision* Kelas 1	*Recall* Kelas 1	F1 Kelas
0	Original+T	0.8125	0.735849	0.772277	0.745455	0.82	0.780952
1	Z-Score+T	0.8125	0.735849	0.772277	0.745455	0.82	0.780952
2	IQR+T	0.8125	0.735849	0.772277	0.745455	0.82	0.780952

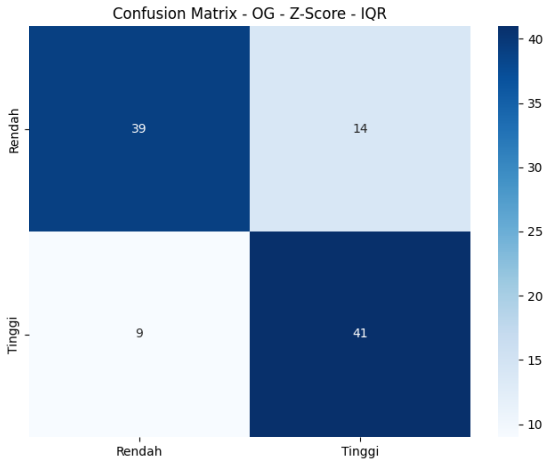
Pemodelan menggunakan *logistic regression* dengan optimasi *hyperparameter* melalui *GridSearchCV* membandingkan tiga teknik: Og (*Threshold*), *Z-Score + Threshold*, dan *IQR + Threshold* (dengan *capping outlier*), yang diterapkan untuk memprediksi label biner "label_kemiskinan". Konsistensi performa menunjukkan pra-pemrosesan *outlier* tidak memengaruhi hasil secara signifikan. Rata-rata *accuracy cross-validation (CV)* tercatat 71,01%, dengan *accuracy test* mencapai 77,67% stabil antar teknik, didukung oleh skor *ROC-AUC* 0,8566, divisualisasikan dalam kurva *ROC* pada Gambar 7, yang menandakan kemampuan diskriminasi baik. *Precision Overall (Macro)* sebesar 77,90%, *recall overall (macro)* 77,79%, dan *F1-Score Overall (Macro)* 77,66% mencerminkan keseimbangan performa. Untuk kelas 0 (kemiskinan rendah), *precision* 81,25%, *recall* 73,58%, dan *F1-Score* 77,23% menunjukkan ketepatan tinggi, sedangkan untuk kelas 1 (kemiskinan tinggi), *precision* 74,55%, *recall* 0,82, dan *F1-Score* 78,10% dengan *macro average* 0,78 untuk ketiga metrik, mengindikasikan sensitivitas unggul.



Gambar 7 Kurva ROC-AUC

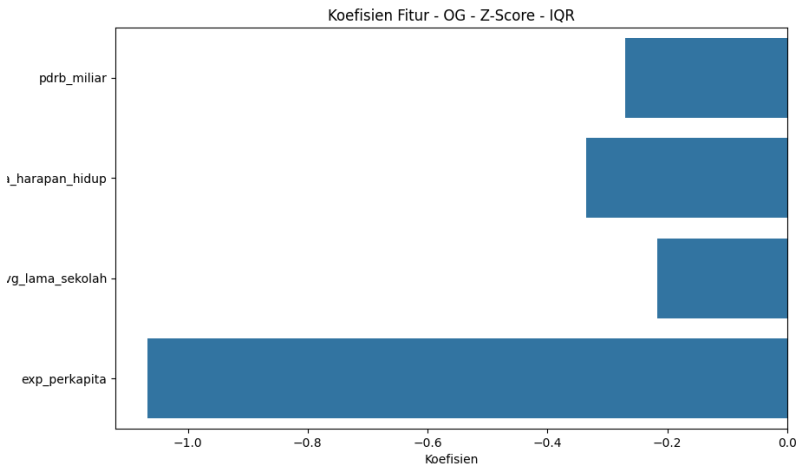
Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

Penanganan *outlier* melalui *capping Z-Score* (rentang 0-30%) dan *IQR* (0-25%) dengan penurunan variansi 15-20% tanpa mengubah label biner "label_kemiskinan", yang ditentukan oleh ambang batas 9,36%, dengan distribusi tetap konsisten (266 kelas 0, 248 kelas 1), menghasilkan *input* model *logistic regression* tetap stabil. *Capping* hanya memengaruhi *outlier* pada "tingkat_kemiskinan" tanpa mengubah hubungan linier antara fitur *predictor* dan label, dan menjaga struktur model dan metrik *accuracy* secara konsisten. Dengan *threshold* 9,36%, yang tidak dipengaruhi *capping*, memastikan klasifikasi akhir seragam, meskipun nilai maksimum ditekan (30,271132% untuk *Z-Score*, 24,500000% untuk *IQR*), menegaskan efektivitas pra-pemrosesan tanpa mengubah performa. Alasan utama konsistensi ini terletak pada sifat robust logistic regression terhadap outlier, yang mengandalkan sigmoid transformation dan maximum likelihood estimation daripada sensitif terhadap ekstrem seperti regresi linier, sehingga capping hanya mengurangi noise tanpa mengubah probabilitas prediksi biner. Selain itu, karena label biner ditentukan oleh threshold tetap (9,36%) yang berada di bawah batas capping (24-30%), outlier tidak memengaruhi komposisi kelas, menghindari bias klasifikasi. Implikasinya, temuan ini menunjukkan bahwa pra-pemrosesan outlier esensial untuk generalisasi model di data real-world seperti kemiskinan spasial, di mana variabilitas tinggi umum terjadi, tetapi tidak selalu diperlukan perubahan drastis jika model dasar sudah stabil, mendukung efisiensi komputasi di lingkungan resource-limited. *Confusion Matrix* yang divisualisasikan pada Gambar 7 menunjukkan 39 prediksi benar untuk kelas 0 (kemiskinan rendah) dan 41 untuk kelas 1 (kemiskinan tinggi), dengan 14 kesalahan pada kelas 0 dan 9 pada kelas 1, mengindikasikan keseimbangan performa antar kelas, di mana false negative rendah (9) pada kemiskinan tinggi meminimalkan risiko melewatkan daerah prioritas kebijakan.



Gambar 8 Confusion Matrix

Analisis koefisien fitur yang divisualisasikan pada Gambar 8 menunjukkan bahwa "exp_perkapita" memiliki pengaruh negatif paling signifikan (-1,067444) terhadap tingkat kemiskinan, diikuti oleh "angka_harapan_hidup" (-0,335930), "pdrb_miliar" (-0,270237), dan "avg_lama_sekolah" (-0,217483), yang divisualisasikan dalam barplot koefisien, mencerminkan hubungan invers dengan kemiskinan. Implikasi koefisien ini adalah bahwa peningkatan pengeluaran per kapita memiliki dampak terbesar dalam mengurangi kemiskinan, mendukung prioritas kebijakan fiskal seperti subsidi langsung tunai, sementara kontribusi lebih lemah dari PDRB menyoroti perlunya diversifikasi ekonomi di luar pertumbuhan agregat. Konsistensi performa model antar teknik menguatkan efektivitas pra-pemrosesan dalam mendukung klasifikasi yang robust, dengan visualisasi yang mendukung interpretasi hasil secara komprehensif.



Gambar 9 Koefisien Fitur

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

4. KESIMPULAN

Penelitian ini berhasil mengintegrasikan data sekunder dari Badan Pusat Statistik (BPS) dan Kementerian Dalam Negeri untuk menganalisis tingkat kemiskinan pada 514 kabupaten/kota di Indonesia tahun 2023, dengan fokus pada variabel populasi, PDRB, angka harapan hidup, rata-rata lama sekolah, dan pengeluaran per kapita. Analisis data eksploratif (EDA) mengungkap hubungan korelasi signifikan, seperti korelasi positif kuat antara populasi dan jumlah penduduk miskin (0,87), serta negatif lemah antara rata-rata lama sekolah dan jumlah penduduk miskin (-0,19), yang mendukung pemilihan fitur prediktor tanpa *multicollinearity* berlebih. Pra-pemrosesan dengan *capping Z-Score* dan *IQR* efektif mengurangi *outlier* dan kemiringan distribusi tingkat kemiskinan, dengan penurunan standar deviasi dari 6,448354 menjadi 5,994413 (*Z-Score*) dan 5,566834 (*IQR*), serta rentang nilai maksimum dari 50% menjadi 30,27% dan 24,50%, tanpa mengubah median atau distribusi label biner (266 kelas rendah, 248 kelas tinggi). Pemodelan *logistic regression* dengan *GridSearchCV* menghasilkan performa konsisten antar teknik, dengan *accuracy* uji 77,67%, *ROC-AUC* 0,8566, *precision macro* 77,90%, *recall macro* 77,79%, dan *F1-score macro* 77,66%, menunjukkan ketahanan model terhadap variasi pra-pemrosesan. Analisis koefisien fitur menegaskan pengaruh negatif dominan dari pengeluaran per kapita (-1,067444), diikuti angka harapan hidup (-0,335930), PDRB (-0,270237), dan rata-rata lama sekolah (-0,217483), mengindikasikan bahwa peningkatan indikator ini mengurangi probabilitas kemiskinan tinggi. Hasil ini mengimplikasikan bahwa teknik *capping*, terutama *IQR* yang lebih adaptif terhadap data miring, dapat meningkatkan robustitas model klasifikasi kemiskinan tanpa mengorbankan *accuracy*, dan mendukung kebijakan berbasis data untuk mengurangi disparitas regional di Indonesia. saran untuk penelitian selanjutnya untuk tambahkan data terkait variabel spasial untuk analisis lebih lengkap dan mendalam, serta uji model dengan data yang terbaru atau real-time untuk memastikan keakuratannya. Temuan penelitian ini sesuai dengan hasil studi dari Badan Pusat Statistik (BPS) tentang kemiskinan di Indonesia, yang menyoroti bahwa PDRB (Produk Domestik Regional Bruto, atau tingkat ekonomi daerah) dan pendidikan adalah faktor penting untuk mengurangi kemiskinan.

REFERENCES

[1] H. Pasarela, Wardhiah, R. Juanda, Fuadi, dan Arliansyah, “Kebijakan Pengentasan Kemiskinan di Indonesia: Sebuah Fakta di Indonesia,” *Socius: Jurnal Penelitian Ilmu-Ilmu Sosial*, vol. 2, no. 1, hlm. 73–79, 2024, [Daring]. Tersedia pada: https://ojs.daarulhuda.or.id/index.php/Socius/article/view/745/790

[2] Muhammad Yasin, Yeny Novita Fitriani, dan Joanne Andre Toy Penga, “Kemiskinan di Indonesia Demi Meningkatkan Pertumbuhan Ekonomi,” *Anggaran : Jurnal Publikasi Ekonomi dan Akuntansi*, vol. 2, no. 2, hlm. 104–112, 2024, doi: 10.61132/anggaran.v2i2.545.

[3] B. P. STATISTIK, “Profil Kemiskinan di Indonesia Maret 2023,” *Badan Pusat statistik*, no. 47, hlm. 1–16, 2023, [Daring]. Tersedia pada: https://www.bps.go.id/pressrelease/2018/07/16/1483/persentase-penduduk-miskin-maret-2018-turun-menjadi-9-82-persen.html

[4] H. N. Samongilailai dan A. B. Utomo, “Strategi Melestarikan Budaya Indonesia di Era Modern,” *WISSEN : Jurnal ilmu Sosial dan Humaniora*, vol. 2, no. 4, hlm. 157–158, 2024.

[5] yinuo wang, M. Umair, A. Aizhan, V. Teymurova, dan L. Chang, “Does the disparity between rural and urban incomes affect rural energy poverty?,” *Energy Strategy Reviews*, vol. 56, no. October, hlm. 101584, 2024, doi: 10.1016/j.esr.2024.101584.

[6] A. Bayu Bagas Samudra dan M. Wahed, “Pengaruh Rata Lama Sekolah,Umur Harapan Hidup Serta PDRB Per Kapita Terhadap Kemiskinan Melalui Analisis Jalur Pengangguran di Daerah Istimewa Yogyakarta,” *Journal of Economics and Business UBS*, vol. 12, no. 3, hlm. 1432–1444, 2023, doi: 10.52644/joeb.v12i3.234.

[7] J. Tuarita dan N. Lusida, “Implementasi Kebijakan Program Bantuan Langsung Tunai-Dana Desa Bagi Masyarakat Miskin Terdampak Covid-19 Di Kecamatan Salahutu Kabupaten Maluku Tengah,” *Administrasi Terapan*, vol. 2, no. 2, hlm. 314–328, 2023.

[8] E. Purwanti, “Analisis Deskriptif Profil Kemiskinan Indonesia Berdasarkan Data BPS Tahun 2023,” *AKADEMIK: Jurnal Mahasiswa Humanis*, vol. 4, no. 1, hlm. 1–10, 2024, doi: 10.37481/jmh.v4i1.653.

[9] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, dan C. Iwendi, “Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease,” *Biomedicines*, vol. 11, no. 2, 2023, doi: 10.3390/biomedicines11020581.

[10] E. Purnamasari dan D. A. Verano, “Model Data-Driven untuk Prediksi Digitalisasi UMKM Menggunakan GMM dan XGBoost,” *Jurnal Pustaka AI …*, vol. 5, no. 2, hlm. 204–214, 2025, [Daring]. Tersedia pada: https://mail.pustakagalerimandiri.co.id/jurnalpgm/index.php/pustakaai/article/view/984%0Ahttps://mail.pustakagalerimandiri.co.id/jurnalpgm/index.php/pustakaai/article/download/984/779

[11] Affandi, Y. Purwaningsih, L. Hakim, dan T. Mulyaningsih, “Interplay between poverty, poverty eradication and sustainable development: A semi-systematic literature review,” *Glob Transit*, vol. 7, hlm. 1–20, 2025, doi: 10.1016/j.glt.2024.11.001.

[12] I. Tawakal, M. M. Effendi, dan A. M. Majid, “ANALISIS TINGKAT KEMISKINAN DENGAN ALGORITMA K-MEANS,” *Journal of Information System Management (JOISM)*, vol. 7, no. 1, hlm. 112–119, 2025.

[13] C. Lartey, J. Liu, R. K. Asamoah, C. Greet, M. Zanin, dan W. Skinner, “Effective Outlier Detection for Ensuring Data Quality in Flotation Data Modelling Using Machine Learning (ML) Algorithms,” *Minerals*, vol. 14, no. 9, hlm. 1–28, 2024, doi: 10.3390/min14090925.

[14] A. B. Kenanga, M. Wulandari, F. A. Wahdah, S. N. Farida, dan F. Syahrani, “PENENTUAN KETEPATAN PADA KLASIFIKASI TINGKAT KEDALAMAN KEMISKINAN DI INDONESIA DENGAN REGRESI LOGISTIK BINER,” *MUSYTARI Neraca Akuntansi Manajemen Ekonomi*, vol. 19, no. 3, hlm. 167–186, 2025.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1487−1499
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8343

[15] Regita Putri Permata dan Rifdatun Ni'mah, "Analisis Regresi Logistik Biner Multilevel pada Status Kemiskinan di Pulau Jawa menggunakan Algoritma MCMC Metropolis-Hasting," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 16, no. 1, hlm. 316–327, 2023, doi: 10.36456/jstat.vol16.no1.a6578.

[16] R. D. Alfiyah dan W. Suekartiningsih, "BERBANTUAN MEDIA GAMBAR BERSERI TERHADAP KEMAMPUAN MEMBACA PERMULAAN PESERTA DIDIK KELAS I SDN WONOREJO 274 SURABAYA," *Jurnal penelitian pendidikan guru sekolah dasar*, vol. 12, no. Vol 12 No 8 (2024), hlm. 1486–1496, 2024.

[17] S. Jauhari, Rozzi Kesuma Dinata, dan Ar Razi, "Perbandingan Metode Logistic Regression Dan Random Forest Dalam Klasifikasi Penyakit Kulit Multikelas," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 10, no. 2, hlm. 1369–1379, 2025, doi: 10.36341/rabit.v10i2.6551.

[18] V. Chang, N. Hahm, Q. A. Xu, P. Vijayakumar, dan L. Liu, "Towards data and analytics driven B2B-banking for green finance: A cross-selling use case study," *Technol Forecast Soc Change*, vol. 206, no. July, hlm. 123542, 2024, doi: 10.1016/j.techfore.2024.123542.

[19] R. Kamila, N. Imro'ah, dan E. Sulistianingsih, "METODE LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO) UNTUK PENDUGAAN PARAMETER REGRESI LOGISTIK BINER (Studi Kasus: Faktor-faktor Tingkat Kemiskinan di Indonesia Tahun 2021)," *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, vol. 14, no. 1, hlm. 57–66, 2025.

[20] Y. Sun *dkk.*, "How Do Rural Households' Livelihood Vulnerability Affect Their Resilience? A Spatiotemporal Empirical Analysis from a Multi-Risk Perspective," *Sustainability (Switzerland)*, vol. 17, no. 17, hlm. 1–38, 2025, doi: 10.3390/su17177695.

[21] A. P. Utami, A. Ibrahim, dan M. Adnan, "Penerapan Prinsip-Prinsip Good Governance dan Penanggulangan Tingkat Kemiskinan Di Kabupaten Aceh Barat," *Journal of Law and Economics*, vol. 3, no. 2, hlm. 83–98, 2024, doi: 10.56347/jle.v3i2.221.

[22] A. Ridayanti, A. Nugroho, dan R. Candrakirana, "Pengentasan Kemiskinan Melalui Metode Spasial Perkotaan Dalam Pengembangan Sustainable Development Goals (SDGs) Kota Surakarta," *Jurnal Ilmiah Wahana Pendidikan*, vol. 10, no. 2, hlm. 40–55, 2024, [Daring]. Tersedia pada: https://doi.org/10.5281/zenodo.10470365.

[23] A. S. AlSalehy dan M. Bailey, "Improving Time Series Data Quality: Identifying Outliers and Handling Missing Values in a Multilocation Gas and Weather Dataset," *Smart Cities*, vol. 8, no. 3, hlm. 1–39, 2025, doi: 10.3390/smartcities8030082.

[24] A. Bakumenko dan A. Elragal, "Detecting Anomalies in Financial Data Using Machine Learning Algorithms," *System*, vol. 10, no. 130, hlm. 375–409, 2021, doi: 10.4171/automata-1/11.