Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2764~2774
e-ISSN: 2548-6861

Comparison of Feature Selection Methods in Classifying Poverty Levels in Indonesia Using Comparative Machine Learning Methods

Nuniska Dwi Kamayanti [1]*, Ifnu Wisma Dwi Prastya [2]*, Sahri [3]*
* Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Nahdlatul Sunan Giri
nuniskakamayanti@gmail.com [1], ifnuprastya@unugiri.ac.id [2], sahriunugiri@gmail.com [3]

Article Info

Article history:

Received 2026-02-28
Revised 2026-05-04
Accepted 2026-05-25

Keyword:

Chi-Square,
Feature Selection,
Normalization,
Poverty Classification,
Random Forest

ABSTRACT

Poverty classification requires models capable of handling multidimensional data and imbalanced class distributions. This study aims to develop and compare several machine learning algorithms for classifying poverty levels in Indonesia, as well as to analyze the impact of feature selection and reduction methods on model performance. The study employs a comparative approach using a secondary dataset consisting of 514 districts/cities with socio-economic indicators and a binary target variable. The methodology includes data preprocessing, the application of Chi-Square, Pearson Correlation, and Principal Component Analysis (PCA), and the handling of imbalanced data using the Synthetic Minority Oversampling Technique (SMOTE). Modelling is conducted using Random Forest, Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Network (ANN), with evaluation performed using Stratified K-Fold Cross Validation and metrics including accuracy, precision, recall, and F1-score. The results indicate that Chi-Square and Pearson Correlation outperform PCA, with Random Forest achieving the best performance, attaining an accuracy of 0.9854 and an F1-score of 0.9507, while effectively detecting the minority class. Therefore, the combination of Chi-Square and Random Forest is identified as the most effective approach in this study, as it produces a model that is accurate, stable, and capable of handling imbalanced data.

This is an open access article under the CC–BY-SA license.

I. PENDAHULUAN

Kemiskinan merupakan hambatan yang kompleks dan mudah berubah terhadap pembangunan di tingkat nasional dan internasional. Meskipun Bank Dunia menyatakan bahwa prevalensi kemiskinan ekstrem secara global menurun, perkembangan ini rapuh dan sangat sensitif terhadap ketidakstabilan ekonomi global dan bencana [1]. Analisis Badan Pusat Statistik (BPS) menunjukkan bahwa, pada Maret 2024, 25,22 juta atau 9,03 % penduduk Indonesia hidup dalam kemiskinan [2]. Data ini menunjukkan adanya kesenjangan yang cukup besar dalam definisi, deskripsi, dan klasifikasi kemiskinan, dan pengembangan kerangka analitis yang canggih sangat penting untuk menciptakan kerangka kerja yang tepat bagi kebijakan intervensi yang terfokus dengan cara yang konsisten dengan SDGs [3].

Kemiskinan tidak selalu dapat dijelaskan hanya melalui indikator moneter, karena kondisi miskin juga tercermin dari deprivasi pada berbagai aspek kehidupan. Dalam perspektif multidimensi kemiskinan, kekurangan yang saling terkait dalam pendidikan, kesehatan, dan standar hidup memberikan gambaran yang lebih baik tentang kerentanan rumah tangga [4]. Dalam kasus Indonesia, terdapat segmen penduduk yang Indeks Kemiskinan Multidimensi (MPI) menunjukkan bahwa mereka terjebak atau berisiko berada dalam kemiskinan multidimensi, yang menekankan perlunya studi kemiskinan untuk melampaui pendekatan berbasis pendapatan [5].

Bersama dengan berbagai distribusi kemiskinan di berbagai wilayah dan strata sosial, sifat kemiskinan yang multifaset juga berasal dari kompleksitas variabel yang terkait dengannya, beberapa di antaranya mungkin tidak saling terkait dengan cara yang langsung. Metode statistik tradisional yang menghasilkan pengaturan linier dan berfokus pada satu variabel, tidak memadai untuk menangkap dinamika kemiskinan yang dihasilkan dari pertemuan pendidikan, kesehatan, dan standar hidup [6]. Dalam konteks

ini, metodologi pembelajaran mesin menyediakan kerangka analitis yang lebih adaptif dan berpusat pada data untuk menjelaskan hubungan non-linier dan interaksi yang rumit, dan karenanya dianggap sebagai sumber daya yang berharga untuk analisis kebijakan publik berbasis bukti [7].

Karena kebutuhan akan metode yang lebih mudah beradaptasi, beberapa penelitian sebelumnya telah menggunakan beragam algoritma machine learning untuk mengkategorikan kemiskinan atau keluarga miskin di Indonesia. Nuzula dkk. menggunakan *Support Vector Machine* (SVM), berbeda dengan CART, untuk mengkategorikan status kemiskinan rumah tangga di Kabupaten Wonosobo, dan mencapai akurasi yang tinggi. Namun, metode SVM secara inheren sensitif terhadap pemilihan kernel dan parameter dan memerlukan pra-pemrosesan, seperti penskalaan fitur, untuk memastikan kinerja yang konsisten di berbagai dataset [8]. Dalam kasus yang berbeda, Ramadhanti dkk. menggunakan *K-Nearest Neighbor* (KNN) untuk klasifikasi rumah tangga miskin dengan strategi penanganan data tidak seimbang (SMOTE/ADASYN) namun KNN cenderung bergantung pada pemilihan nilai k dan metrik jarak, serta dapat terpengaruh oleh distribusi data dan dimensi fitur sehingga hasilnya bisa berubah signifikan ketika karakter data bergeser [9]. Selain itu, Sholihah dkk. menerapkan *Random Forest* dengan bantuan SMOTE untuk klasifikasi status ekonomi rumah tangga walau kinerjanya dapat meningkat, penggunaan *oversampling* seperti SMOTE berpotensi menambah kompleksitas dan risiko *overfitting* pada pola sintetis, sehingga perlu evaluasi yang ketat agar model tidak sekadar bagus di data latih [10].

Seiring dengan meningkatnya kebutuhan akan metode klasifikasi yang adaptif dan akurat, berbagai penelitian telah memanfaatkan algoritma *machine learning* untuk menyelesaikan permasalahan klasifikasi pada berbagai domain. Metode *Logistic Regression* dikenal sebagai algoritma yang populer dan memiliki kemampuan prediksi yang baik serta interpretasi yang jelas terhadap hubungan antar variabel, sehingga banyak digunakan dalam berbagai bidang penelitian [11]. Selain itu, *Support Vector Machine* (SVM) memiliki keunggulan dalam membangun model dengan margin optimal serta mampu bekerja dengan baik pada data berdimensi tinggi dan menghasilkan generalisasi yang baik [12]. Di sisi lain, *Random Forest* sebagai metode berbasis *ensemble learning* memiliki kemampuan dalam meningkatkan akurasi prediksi dengan menggabungkan banyak pohon keputusan, serta mampu mengurangi risiko *overfitting* yang sering terjadi pada model tunggal [13]. Di bidang pendidikan, penggunaan ANN dengan teknik *backpropagation* dan mencapai akurasi 93% saat memprediksi tingkat kelulusan siswa [14]. Keunggulan-keunggulan tersebut menjadikan keempat metode ini banyak digunakan dalam berbagai penelitian klasifikasi, termasuk dalam analisis data dengan karakteristik kompleks dan tidak seimbang.

Dalam penelitian sebelumnya, evaluasi model klasifikasi banyak dilakukan menggunakan metode *K-Fold Cross Validation* karena mampu memberikan pengukuran performa yang lebih stabil dan representatif [15]. Metode ini bekerja dengan membagi data menjadi beberapa bagian, kemudian secara bergantian digunakan sebagai data latih dan data uji, sehingga seluruh data dapat dimanfaatkan secara optimal dan mengurangi bias evaluasi akibat pembagian data tunggal. Selain itu, permasalahan data tidak seimbang juga menjadi perhatian penting dalam proses klasifikasi. Salah satu pendekatan yang digunakan adalah *Synthetic Minority Oversampling Technique* (SMOTE), yaitu metode yang menghasilkan data sintetis pada kelas minoritas untuk menyeimbangkan distribusi data. Berdasarkan penelitian yang dilakukan, penerapan SMOTE terbukti mampu meningkatkan kinerja model klasifikasi, terutama dalam meningkatkan kemampuan model dalam mengenali kelas minoritas serta menghasilkan evaluasi yang lebih seimbang dibandingkan sebelum dilakukan penanganan data tidak seimbang [16]. Dengan demikian, penggunaan *K-Fold Cross Validation* dan SMOTE menjadi pendekatan yang efektif dalam meningkatkan keandalan dan kualitas hasil klasifikasi pada dataset dengan karakteristik tidak seimbang.

Namun dalam praktiknya, upaya klasifikasi kemiskinan mencoba untuk merangkum rentang faktor sosial-ekonomi yang saling terkait, menambah kompleksitas struktural data dan daya tarik model yang kurang canggih. Sementara penelitian sebelumnya telah mencatat berbagai faktor yang berkontribusi pada konstruksi model, klasifikasi yang dihasilkan masih jauh dari memuaskan. Misalnya, sebuah studi yang menganalisis teknik pembelajaran mesin untuk klasifikasi kemiskinan menggunakan data Susenas mencapai akurasi sekitar 73%, bahkan setelah mengendalikan sejumlah variabel penjelas dan menerapkan teknik untuk data tidak seimbang [17]. Analisis diskriminan induktif yang diterapkan pada data yang sama dan klasifikasi tingkat kemiskinan berikutnya menghasilkan rentang akurasi klasifikasi, yang salah satu prosidingnya mencatat sekitar 67,7% [18]. Ketidakseimbangan kelas positif, adanya multikolinearitas, dan ketidakhadiran struktur yang dapat dibedakan dalam populasi menciptakan situasi di mana hanya meningkatkan jumlah fitur tidak secara otomatis meningkatkan akurasi model. Agar model lebih baik mencapai tingkat kesalahan yang tidak dapat dikurangi ini, langkah-langkah seperti pengurangan dan pemilihan fitur menjadi sangat penting. Kompleksitas data tidak hanya bersumber dari banyaknya variabel, tetapi dari disparitas rentang nilai antarfitur yang dapat menyebabkan dominasi atribut tertentu dalam proses pembelajaran. Oleh karena itu, normalisasi *min–max* yang memetakan seluruh fitur ke dalam interval [0,1] diterapkan untuk menyeimbangkan kontribusi numerik setiap variabel, sehingga evaluasi terhadap respons model dapat dilakukan secara lebih terkontrol dan proporsional [19].

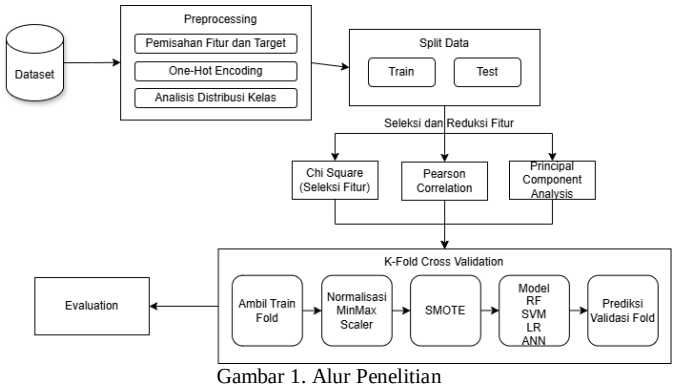
Salah satu cara untuk mencapai ini adalah melalui *Principal Component Analysis* (PCA), sebuah metode yang mengurangi kompleksitas data tanpa kehilangan pola penting

yang signifikan dengan mentransformasi sekelompok atribut yang berkorelasi menjadi sejumlah atribut yang tidak berkorelasi yang dikenal sebagai komponen utama yang menangkap sebagian besar variabilitas dalam distribusi data [20]. *Pearson Correlation* dapat diterapkan untuk mengidentifikasi dan menghapus salah satu dari sepasang fitur yang redundan agar mengurangi efek multikolinearitas dan meningkatkan pelatihan model [21]. Salah satu teknik yang dikenal untuk meningkatkan kinerja model klasifikasi serta efisiensi komputasi model adalah pemilihan fitur *Chi-Square* yang memeriksa derajat independensi masing-masing fitur dan variabel target [22]. Menggabungkan semua teknik yang dijelaskan kemungkinan besar akan meningkatkan model dalam hal efisiensi komputasi dan memiliki generalisasi yang akurat.

Sejalan dengan latar belakang tersebut, penelitian ini bertujuan untuk membangun dan membandingkan model klasifikasi tingkat kemiskinan di Indonesia menggunakan beberapa algoritma *machine learning*, yaitu *Random Forest*, *Support Vector Machine* (SVM), *Logistic Regression*, dan *Artificial Neural Network* (ANN). Penelitian ini juga bertujuan untuk menganalisis pengaruh metode seleksi dan reduksi fitur, yaitu *Chi-Square*, *Pearson Correlation*, dan *Principal Component Analysis* (PCA), dalam meningkatkan performa model klasifikasi. Selain itu, untuk mengatasi permasalahan data tidak seimbang, diterapkan metode *Synthetic Minority Oversampling Technique* (SMOTE), serta dilakukan evaluasi model menggunakan *Stratified K-Fold Cross Validation* dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Melalui pendekatan tersebut, penelitian ini diharapkan dapat menghasilkan model klasifikasi yang optimal serta mengidentifikasi fitur-fitur yang paling berpengaruh dalam menentukan tingkat kemiskinan.

II. METODE

Metode penelitian ini dirancang dalam alur kerja yang sistematis dan terstruktur untuk memastikan proses klasifikasi tingkat kemiskinan berlangsung secara konsisten dan dapat direplikasi pada setiap skenario eksperimen. Tahapan dimulai dari pengumpulan data, dilanjutkan dengan pra-pemrosesan yang mencakup pembersihan data, penanganan nilai hilang melalui imputasi, serta transformasi variabel kategorikal menggunakan *one-hot encoding*. Dataset kemudian dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20 menggunakan *stratified sampling* guna mempertahankan distribusi kelas, di mana ketidakseimbangan pada data pelatihan ditangani menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Selanjutnya, dilakukan seleksi dan reduksi fitur melalui metode *Chi-Square* untuk mengidentifikasi fitur yang paling relevan, *Pearson Correlation* untuk mengeliminasi redundansi antarfitur, serta *Principal Component Analysis* (PCA) untuk mereduksi dimensi data dengan mempertahankan variansi utama. Pada skenario tertentu, diterapkan normalisasi menggunakan *Min–Max Scaling* untuk menyamakan skala fitur numerik. Proses pemodelan dilakukan menggunakan beberapa algoritma klasifikasi, yaitu *Random Forest*, *Support Vector Machine* (SVM), *Logistic Regression*, dan *Artificial Neural Network* (ANN), yang dievaluasi menggunakan pendekatan *k-fold cross-validation* dengan variasi nilai k. Kinerja model diukur menggunakan metrik akurasi dan dianalisis melalui *confusion matrix* untuk menilai kemampuan model dalam mengklasifikasikan kelas secara tepat, khususnya pada kondisi data yang tidak seimbang.



Gambar 1. Alur Penelitian

A. Dataset

Data yang digunakan dalam penelitian ini bersumber dari dataset publik yang tersedia pada platform *Kaggle*, yang merepresentasikan kondisi sosial ekonomi wilayah di Indonesia pada tingkat kabupaten/kota. Dataset tersebut memuat seperangkat variabel prediktor yang mencerminkan dimensi kesejahteraan masyarakat, meliputi aspek pendidikan, kesehatan, ketenagakerjaan, dan ekonomi regional. Secara keseluruhan, data terdiri atas 10 fitur numerik, yaitu persentase penduduk miskin (*P0*), rata-rata lama sekolah, pengeluaran per kapita, Indeks Pembangunan Manusia (*IPM*), umur harapan hidup, akses sanitasi layak, akses air minum layak, tingkat pengangguran terbuka, tingkat partisipasi angkatan kerja, serta Produk Domestik Regional Bruto (*PDRB*), serta 1 fitur kategorikal berupa Provinsi. Variabel kategorikal tersebut selanjutnya ditransformasikan ke dalam representasi numerik menggunakan teknik *one-hot encoding* pada tahap pra-pemrosesan untuk mendukung kompatibilitas dengan model pembelajaran mesin.

Variabel target dalam penelitian ini berupa label biner yang merepresentasikan klasifikasi tingkat kemiskinan wilayah, yang terdiri atas dua kategori, yaitu tidak miskin (0) dan miskin (1). Penetapan label tersebut tidak dilakukan melalui konstruksi atau rekayasa dalam penelitian ini, melainkan telah tersedia secara inheren pada dataset asli yang diperoleh dari *Kaggle*, sehingga validitas label mengacu pada definisi yang digunakan oleh penyedia data. Dengan demikian, penelitian ini berfokus pada pemodelan hubungan antara variabel prediktor dan label target yang telah ditetapkan. Sebelum tahap pemodelan, dilakukan proses seleksi atribut untuk memastikan bahwa variabel yang digunakan bersifat relevan, konsisten, dan memiliki kontribusi prediktif, sementara atribut yang berfungsi sebagai identitas wilayah dieliminasi

guna menghindari terbentuknya pola semu (*spurious correlation*) dalam proses pembelajaran model.

B. Pre-processing Data

Tahap pra-pemrosesan dilakukan untuk memastikan kualitas dan konsistensi data sebelum digunakan dalam proses pemodelan. Proses ini diawali dengan pembersihan data, yang mencakup penghapusan duplikasi serta penyesuaian tipe data agar seluruh variabel numerik berada dalam format yang sesuai untuk analisis. Penanganan nilai hilang dilakukan menggunakan teknik imputasi, di mana nilai median digunakan untuk variabel numerik guna meminimalkan pengaruh pencilan, sedangkan nilai modus digunakan untuk variabel kategorikal. Selanjutnya, variabel kategorikal, seperti provinsi, ditransformasikan ke dalam bentuk numerik menggunakan metode *one-hot encoding* agar dapat diproses oleh algoritma pembelajaran mesin. Setelah itu, dilakukan pemisahan data menjadi data pelatihan dan pengujian menggunakan *stratified sampling* untuk menjaga proporsi distribusi kelas.

C. Split Data

Tahap pembagian data dilakukan untuk menguji kemampuan model dalam melakukan generalisasi terhadap data yang tidak terlibat dalam proses pelatihan. Dataset dibagi menjadi dua subset, yaitu data pelatihan sebesar 80% dan data pengujian sebesar 20% menggunakan metode *stratified sampling* untuk mempertahankan proporsi distribusi kelas pada kedua subset. Pendekatan ini digunakan untuk memastikan bahwa karakteristik distribusi kelas, baik mayoritas maupun minoritas, tetap representatif pada data pelatihan dan pengujian, sehingga evaluasi model dapat dilakukan secara lebih objektif. Data pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk mengukur kinerja model terhadap data baru. Dengan demikian, proses pemisahan data ini bertujuan untuk menghasilkan estimasi performa model yang lebih akurat dan tidak bias.

D. Seleksi dan Reduksi Fitur

Tahap seleksi dan reduksi fitur dilakukan untuk menyederhanakan struktur data sebelum proses pemodelan, sekaligus memastikan bahwa fitur yang digunakan benar-benar memberikan kontribusi terhadap proses klasifikasi. Pengurangan fitur yang tidak relevan atau bersifat redundan diperlukan agar model dapat belajar secara lebih efisien dan stabil. Dalam penelitian ini, digunakan tiga pendekatan yang berbeda, yaitu *Pearson Correlation*, *Chi-Square*, dan *Principal Component Analysis* (PCA), yang masing-masing diterapkan sesuai dengan tujuan dan karakteristik skenario pemodelan.

Pendekatan *Pearson Correlation* digunakan untuk mengevaluasi hubungan linier antar fitur numerik dengan tujuan mengidentifikasi potensi redundansi informasi akibat adanya fitur-fitur yang memiliki tingkat korelasi tinggi [23]. Koefisien *Pearson Correlation* dihitung menggunakan persamaan berikut :

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

di mana X_i dan Y_i merupakan nilai pengamatan ke-i, $\bar{X}$ dan $\bar{Y}$adalah nilai rata-rata masing-masing variabel, serta n menyatakan jumlah data.

Principal Component Analysis (PCA) merupakan metode statistik yang digunakan untuk menyederhanakan data berdimensi tinggi dengan mengekstraksi sejumlah komponen utama berdimensi lebih rendah yang tetap merepresentasikan variasi informasi paling dominan dalam data [24]. Transformasi PCA secara umum dapat dinyatakan sebagai berikut :

$$Z = XW \quad (2)$$

Dengan X sebagai matriks data awal, W sebagai matriks eigenvector dari matriks kovarians, dan Z sebagai matriks komponen utama. Penentuan jumlah komponen dilakukan berdasarkan nilai *cumulative explained variance*, yang dirumuskan sebagai:

$$\sum_{i=1}^{k} \lambda_i \quad (3)$$

di mana λ_i menunjukkan proporsi variansi yang dijelaskan oleh komponen ke-i.

Metode *Chi-Square* digunakan untuk menilai keterkaitan antara fitur dan variabel target, sehingga fitur yang memiliki hubungan statistik paling kuat dengan kelas target dapat diprioritaskan dalam proses klasifikasi [25]. Statistik *Chi-Square* dihitung menggunakan persamaan berikut :

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

dengan menyatakan frekuensi pengamatan, frekuensi harapan, dan jumlah kategori.

E. K-fold Cross Validation

Evaluasi model dalam penelitian ini dilakukan menggunakan metode *k-fold cross-validation* sebagai teknik validasi untuk mengukur kemampuan generalisasi model. Pada metode ini, data pelatihan dibagi menjadi *k* subset (*fold*) dengan ukuran yang relatif sama, di mana secara iteratif satu subset digunakan sebagai data validasi dan sisanya sebagai data pelatihan. Proses ini diulang sebanyak *k* kali sehingga setiap subset digunakan sebagai data validasi tepat satu kali [26]. Untuk menjaga konsistensi distribusi kelas pada setiap fold, digunakan pendekatan *stratified k-fold.* Dalam penelitian ini, nilai *k* divariasikan menjadi 3,5,7 dan 9 untuk memperoleh gambaran performa model yang lebih stabil terhadap berbagai skenario pembagian data. Hasil dari setiap iterasi kemudian dirata-ratakan untuk memperoleh estimasi kinerja model yang lebih representatif, yang secara matematis dirumuskan sebagai:

$$Metric_{avg} = \frac{1}{k}\sum_{i=1}^{k} Metric_i \quad (5)$$

di mana Metrici merupakan nilai metrik pada iterasi ke-*i*, dan *k* adalah jumlah fold yang digunakan. Pendekatan ini memungkinkan seluruh data pelatihan dimanfaatkan secara optimal dalam proses validasi tanpa mengganggu independensi data pengujian.

F. Normalisasi Fitur Numerik (Skenario Tertentu)

Normalisasi fitur numerik dilakukan pada skenario tertentu untuk menyamakan skala nilai antarvariabel sebelum proses pemodelan. Dalam penelitian ini, digunakan metode *Min–Max Scaling* yang mentransformasikan nilai setiap fitur ke dalam rentang [0,1] tanpa mengubah distribusi dasar data. Proses ini bertujuan untuk mengurangi perbedaan skala yang signifikan antarfitur, terutama pada variabel ekonomi yang memiliki rentang nilai jauh lebih besar dibandingkan variabel lainnya. Dengan skala yang seragam, kontribusi setiap fitur dalam proses pembelajaran model menjadi lebih seimbang, sehingga dapat meningkatkan stabilitas dan efisiensi proses pelatihan, khususnya pada algoritma yang sensitif terhadap perbedaan skala data [27].

G. SMOTE

Penanganan ketidakseimbangan kelas pada data pelatihan dilakukan menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Metode ini bekerja dengan menghasilkan data sintetis pada kelas minoritas melalui interpolasi antara sampel yang berdekatan dalam ruang fitur, sehingga distribusi kelas menjadi lebih seimbang tanpa sekadar menduplikasi data yang ada. Penerapan SMOTE bertujuan untuk meningkatkan kemampuan model dalam mengenali pola pada kelas minoritas yang cenderung kurang terwakili, sehingga mengurangi bias model terhadap kelas mayoritas. Dalam penelitian ini, SMOTE diterapkan hanya pada data pelatihan agar distribusi data pengujian tetap mencerminkan kondisi asli dataset, sehingga evaluasi kinerja model dapat dilakukan secara lebih objektif dan tidak bias.

H. Modelling

Pada tahap pemodelan, penelitian ini menggunakan beberapa algoritma klasifikasi untuk membandingkan kinerja dalam mengklasifikasikan tingkat kemiskinan, yaitu *Random Forest*, *Support Vector Machine (SVM)*, *Logistic Regression*, dan *Artificial Neural Network (ANN)*. Penggunaan beberapa algoritma ini bertujuan untuk memperoleh model terbaik yang mampu menangani karakteristik data, khususnya yang bersifat tidak seimbang dan memiliki hubungan non-linear antar variabel.

1) *Random Forest*

Random Forest merupakan metode berbasis *ensemble learning* yang membangun sejumlah *decision tree* dari subset data yang berbeda menggunakan teknik *bagging* [28]. Setiap pohon menghasilkan prediksi, kemudian hasil akhir ditentukan melalui mekanisme *majority voting*.

$$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_n(x)\} \quad (6)$$

dengan $h_i(x)$ merupakan prediksi dari pohon ke-*i*. Metode ini memiliki keunggulan dalam mengurangi variansi model serta mampu menangani data dengan hubungan non-linear secara efektif.

2) *Support Vector Machine (SVM)*

SVM merupakan metode klasifikasi yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan data antar kelas dengan margin maksimum [29]. Fungsi keputusan SVM dapat dinyatakan sebagai:

$$f(x) = w \cdot x + b \quad (7)$$

dengan tujuan optimasi:

$$\min \frac{1}{2} \|w\|^2 \text{ dengan kendala } y_i(w \cdot x_i + b) \geq 1. \quad (8)$$

3) *Logistic Regression*

Logistic Regression merupakan model klasifikasi berbasis probabilistik yang memetakan input ke dalam nilai probabilitas menggunakan fungsi logistik (*sigmoid*) [30]. Probabilitas suatu data termasuk ke dalam kelas tertentu dirumuskan sebagai:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

Klasifikasi dilakukan dengan menentukan ambang batas (*threshold*), umumnya sebesar 0,5.

4) *Artificial Neural Network (ANN)*

Artificial Neural Network (ANN) merupakan model yang meniru cara kerja jaringan saraf manusia, terdiri dari lapisan input, *hidden layer*, dan output [31]. Proses komputasi pada setiap neuron dapat dinyatakan sebagai:

$$z = w \cdot x + b, a = \sigma(z) \quad (9)$$

Dengan $sigma$ merupakan fungsi aktivasi. ANN mampu mempelajari pola yang kompleks dan non-linear, namun performanya dipengaruhi oleh jumlah data dan konfigurasi parameter yang digunakan.

I. Evaluasi

Evaluasi model dilakukan untuk mengukur kinerja algoritma klasifikasi dalam memprediksi tingkat kemiskinan. Mengingat dataset memiliki karakteristik tidak seimbang (imbalanced dataset), maka evaluasi tidak hanya menggunakan akurasi, tetapi juga metrik lain yang lebih representatif, yaitu precision, recall, dan F1-score.

Akurasi digunakan untuk mengukur proporsi hasil klasifikasi yang benar terhadap keseluruhan data, yang dirumuskan sebagai berikut:

$$\boldsymbol{Accuracy} = \frac{TP+FN}{TP+TN+FP+FN} \quad (10)$$

Presisi digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan data ke dalam kelas positif, yaitu perbandingan antara jumlah prediksi positif yang benar terhadap seluruh data yang diprediksi sebagai kelas positif. Rumus presisi dinyatakan sebagai:

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (11)$$

Recall digunakan untuk mengukur kemampuan model dalam mengidentifikasi seluruh data yang termasuk ke dalam kelas positif, yang merepresentasikan proporsi data kelas positif yang berhasil diklasifikasikan dengan benar. Rumus recall dituliskan sebagai:

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \quad (12)$$

F1-score digunakan untuk memberikan ukuran keseimbangan antara presisi dan recall, khususnya pada kondisi data dengan distribusi kelas yang tidak seimbang [32]. Metrik ini dinyatakan sebagai rata-rata harmonis dari presisi dan recall, sehingga mampu merepresentasikan kinerja model secara lebih adil terhadap kelas minoritas. Rumus F1-score dinyatakan sebagai:

$$\boldsymbol{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

Confusion matrix menyajikan ringkasan hasil prediksi model terhadap data aktual, sehingga kesalahan dan ketepatan klasifikasi pada setiap kelas dapat dianalisis secara sistematis [33]. Untuk melengkapi proses evaluasi, kinerja model juga dianalisis menggunakan kurva *Receiver Operating Characteristic* (ROC) dan nilai *Area Under the Curve* (AUC), yang menggambarkan kemampuan model dalam membedakan kelas positif dan kelas negatif secara keseluruhan [34]. Penggunaan kombinasi metrik evaluasi ini bertujuan untuk memperoleh penilaian kinerja model yang lebih komprehensif dan objektif, khususnya dalam menangani permasalahan klasifikasi dengan distribusi kelas yang tidak seimbang.

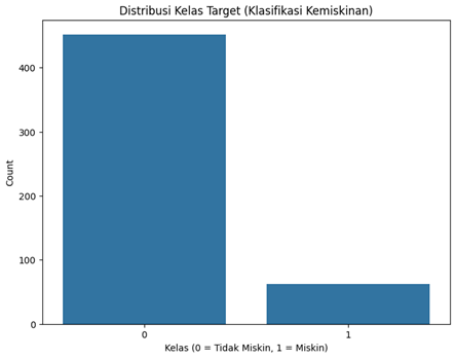
III. HASIL DAN PEMBAHASAN

A. Karakteristik Dataset

TABEL I
ACUAN UKURAN TEKS

Atribut	Data Type
Provinsi	Object
Kab/Kota	Object
Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)	Float64
Rata-rata Lama Sekolah Penduduk 15+ (Tahun)	Object
Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)	Int64
Indeks Pembangunan Manusia	Float64
Umur Harapan Hidup (Tahun)	Float64
Persentase rumah tangga yang memiliki akses terhadap sanitasi layak	Float64
Persentase rumah tangga yang memiliki akses terhadap air minum layak	Float64
Tingkat Pengangguran Terbuka	Float64
Tingkat Partisipasi Angkatan Kerja	Float64
PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)	Int64
Klasifikasi Kemiskinan	Int64

Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari platform Kaggle, yang berisi informasi tingkat kemiskinan pada kabupaten/kota di Indonesia dengan total 514 entri. Setiap entri merepresentasikan satu wilayah administratif dan terdiri dari sejumlah variabel sosio-ekonomi yang mencerminkan kondisi pendidikan, kesehatan, ketenagakerjaan, serta aspek ekonomi wilayah. Variabel target dalam dataset ini berupa klasifikasi tingkat kemiskinan yang dinyatakan dalam bentuk biner, yaitu kelas 0 (tidak miskin) dan kelas 1 (miskin).



Gambar 2. Distribusi Kelas Target

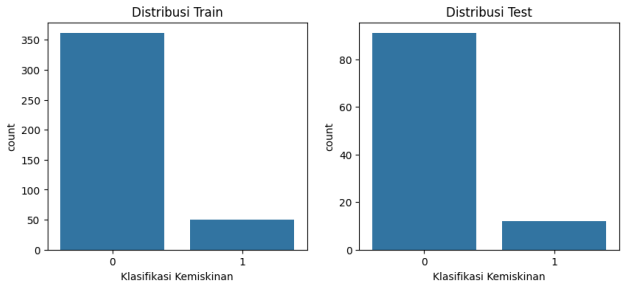
Distribusi kelas pada dataset menunjukkan ketidakseimbangan yang cukup signifikan antara kelas mayoritas dan kelas minoritas. Berdasarkan visualisasi distribusi kelas, jumlah data pada kelas 0 (*tidak miskin*) mencapai sekitar 450 entri, sedangkan kelas 1 (*miskin*) hanya sekitar 60 entri. Perbedaan ini menghasilkan rasio ketidakseimbangan sekitar 7:1, yang mengindikasikan bahwa dataset didominasi oleh kelas mayoritas. Ketimpangan distribusi ini berpotensi memengaruhi proses pembelajaran model, di mana algoritma cenderung lebih sering mempelajari pola dari kelas mayoritas dibandingkan kelas minoritas. Akibatnya, model dapat menghasilkan performa yang terlihat tinggi dari sisi akurasi, tetapi kurang optimal dalam mendeteksi kelas minoritas. Dalam konteks penelitian ini, kondisi tersebut menjadi penting karena kelas minoritas

(*miskin*) justru merupakan fokus utama yang perlu diidentifikasi secara akurat.

Selain itu, distribusi yang tidak seimbang juga mencerminkan karakteristik data sosial-ekonomi di mana jumlah wilayah dengan kategori tidak miskin lebih dominan dibandingkan wilayah dengan kategori miskin. Hal ini menunjukkan bahwa representasi data tidak merata dan dapat membatasi kemampuan model dalam mengenali pola yang lebih jarang muncul.

B. Split Data

Tahap pembagian data dilakukan menggunakan metode *stratified sampling* dengan rasio 80:20, sehingga diperoleh 411 data sebagai data pelatihan dan 103 data sebagai data pengujian, masing-masing dengan jumlah fitur sebanyak 931. Hasil pembagian ini menunjukkan bahwa proporsi data telah terbagi secara representatif antara kedua subset tanpa mengubah struktur fitur yang digunakan. Pendekatan ini bertujuan untuk memastikan bahwa data pengujian tetap independen, sehingga dapat digunakan untuk mengukur kemampuan generalisasi model secara objektif.



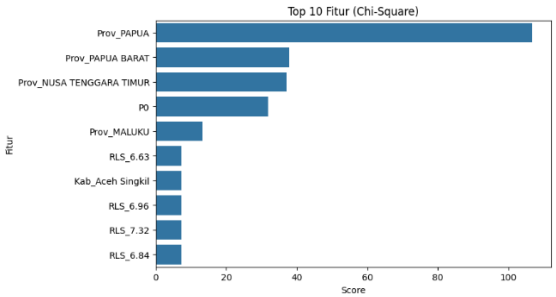
Gambar 3. Distribusi Data Train dan Data Test

Distribusi kelas pada data pelatihan dan data pengujian menunjukkan konsistensi yang sangat baik. Pada data pelatihan, kelas tidak miskin memiliki proporsi sebesar 0,878, sedangkan kelas miskin sebesar 0,122. Sementara itu, pada data pengujian, proporsi kelas tidak miskin sebesar 0,883 dan kelas miskin sebesar 0,117. Perbedaan proporsi antara kedua subset relatif kecil, yang mengindikasikan bahwa metode *stratified sampling* berhasil mempertahankan distribusi kelas sebagaimana pada dataset awal. Visualisasi distribusi kelas dan perbandingan proporsi juga memperkuat temuan ini, di mana pola distribusi antara data pelatihan dan pengujian terlihat serupa. Dengan demikian, proses pemisahan data ini menghasilkan pembagian yang seimbang dan representatif, sehingga mampu mendukung proses pelatihan dan evaluasi model secara lebih valid serta mengurangi potensi bias akibat ketidakseimbangan distribusi data.

C. Seleksi dan Reduksi Fitur

Pada tahap seleksi dan reduksi fitur dilakukan menggunakan tiga pendekatan, yaitu *Chi-Square*, *Pearson Correlation*, dan *Principal Component Analysis* (PCA). Tahap ini bertujuan untuk meningkatkan kualitas fitur yang digunakan dalam proses pemodelan dengan cara memilih fitur yang relevan, mengurangi redundansi, serta menyederhanakan representasi data.

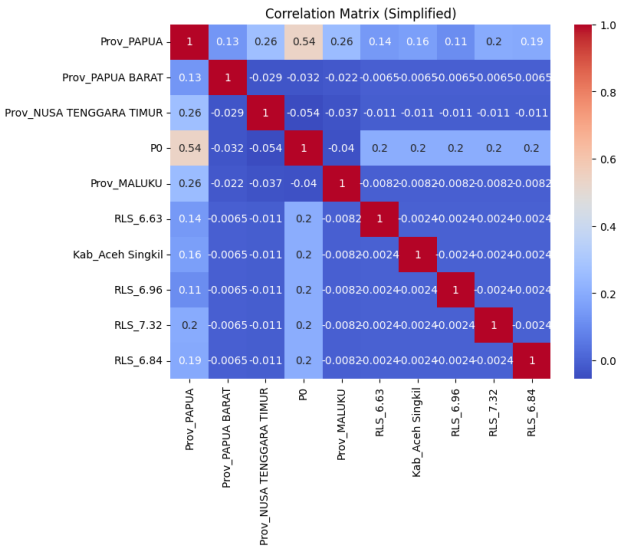
1) Chi-Square



Gambar 4. Top 10 Fitur (Chi-Square)

Berdasarkan visualisasi *Top 10 Fitur (Chi-Square)*, terlihat bahwa fitur dengan skor tertinggi didominasi oleh variabel berbasis wilayah, terutama Provinsi Papua, yang memiliki nilai paling tinggi dibandingkan fitur lainnya, diikuti oleh Provinsi Papua Barat dan Provinsi Nusa Tenggara Timur. Hal ini menunjukkan bahwa faktor geografis memiliki hubungan yang sangat kuat terhadap tingkat kemiskinan, yang mencerminkan adanya ketimpangan kondisi sosial-ekonomi antar wilayah di Indonesia. Selain itu, variabel Persentase Penduduk Miskin (P0) juga memiliki skor yang relatif tinggi, yang menegaskan bahwa indikator kemiskinan secara langsung menjadi faktor penting dalam proses klasifikasi. Sementara itu, fitur yang berkaitan dengan pendidikan seperti Rata-rata Lama Sekolah (RLS) turut berkontribusi, meskipun dengan pengaruh yang lebih kecil dibandingkan variabel wilayah. Secara keseluruhan, hasil ini menunjukkan bahwa metode *Chi-Square* efektif dalam mengidentifikasi fitur yang paling relevan, dengan dominasi faktor geografis dan indikator sosial-ekonomi utama sebagai penentu dalam klasifikasi tingkat kemiskinan.

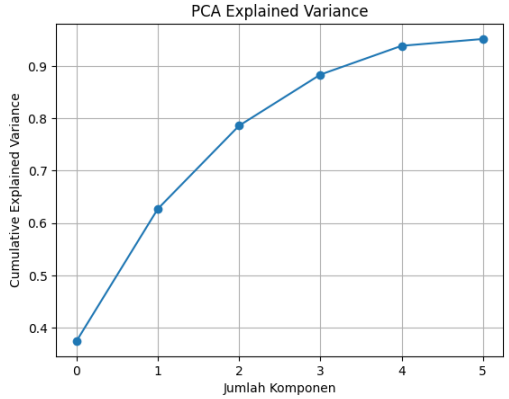
2) Pearson Correlation



Gambar 5. Correlation Matrix

Berdasarkan visualisasi *correlation matrix*, terlihat bahwa sebagian besar pasangan fitur memiliki nilai korelasi yang rendah, yang ditunjukkan oleh nilai koefisien yang mendekati nol. Hal ini mengindikasikan bahwa antar fitur tidak memiliki hubungan linier yang kuat, sehingga risiko terjadinya multikolinearitas relatif kecil. Beberapa fitur seperti Provinsi Papua dan P0 menunjukkan korelasi yang lebih tinggi dibandingkan fitur lainnya, namun masih dalam kategori sedang dan tidak menunjukkan hubungan yang terlalu kuat. Sementara itu, fitur-fitur berbasis pendidikan seperti Rata-rata Lama Sekolah (RLS) memiliki korelasi yang sangat rendah dengan fitur lainnya, yang menunjukkan bahwa variabel tersebut memberikan informasi yang berbeda dalam proses klasifikasi. Secara keseluruhan, hasil ini menunjukkan bahwa sebagian besar fitur bersifat independen satu sama lain, sehingga tetap layak dipertahankan dalam pemodelan, serta mengindikasikan bahwa proses reduksi fitur berbasis korelasi tidak menghilangkan terlalu banyak informasi penting dari dataset.

3) Principal Component Analysis



Gambar 6. Cumulative Explained Variance

Berdasarkan grafik *PCA Explained Variance*, terlihat bahwa penambahan jumlah komponen utama (*principal components*) secara bertahap meningkatkan nilai *cumulative explained variance*. Pada komponen pertama, variansi yang dapat dijelaskan masih relatif rendah, namun meningkat secara signifikan hingga mencapai sekitar 0,78 pada tiga komponen dan mendekati 0,90 pada empat komponen. Setelah itu, peningkatan variansi menjadi lebih kecil, yang menunjukkan adanya titik *diminishing return*, di mana penambahan komponen tidak lagi memberikan kontribusi yang signifikan terhadap informasi yang dipertahankan. Hal ini mengindikasikan bahwa sebagian besar informasi dalam dataset dapat direpresentasikan dengan jumlah komponen yang relatif sedikit, sehingga reduksi dimensi menggunakan PCA cukup efektif dalam menyederhanakan data. Namun demikian, meskipun variansi yang dijelaskan tinggi, hasil sebelumnya menunjukkan bahwa performa model tidak selalu meningkat, yang mengindikasikan bahwa PCA dapat menghilangkan informasi penting yang relevan terhadap variabel target dalam proses klasifikasi.

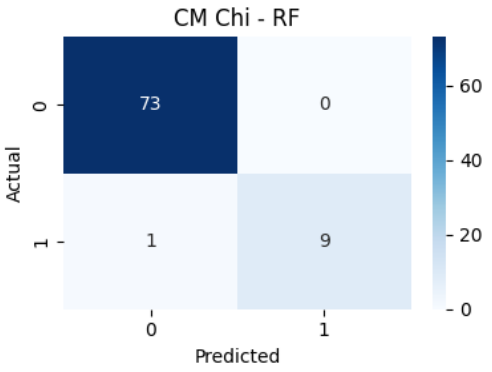
D. K-Fold Cross Validation

Pada tahap ini, dilakukan proses pemodelan menggunakan empat algoritma klasifikasi, yaitu *Random Forest*, *Support Vector Machine* (SVM), *Logistic Regression*, dan *Artificial Neural Network* (ANN). Setiap model diuji menggunakan tiga pendekatan pengolahan fitur, yaitu *Chi-Square*, *Pearson Correlation*, dan *Principal Component Analysis* (PCA), dengan variasi *k-fold cross-validation* (k = 3, 5, 7, dan 9).

Secara umum, hasil pemodelan menunjukkan bahwa seluruh model mampu menghasilkan performa yang cukup tinggi, terutama pada metode *Chi-Square* dan *Pearson Correlation*. Algoritma *Random Forest* secara konsisten memberikan nilai akurasi tertinggi pada hampir seluruh skenario pengujian.

1) Pemodelan dengan *Chi-Square*

Pada metode *Chi-Square*, model *Random Forest* menunjukkan performa terbaik dengan nilai akurasi sebesar 0,9854 dan *F1-score* sebesar 0,9507 pada k = 9. Selain itu, model ini juga menunjukkan kestabilan performa pada seluruh variasi nilai k, yang mengindikasikan kemampuan generalisasi yang baik. Model *Artificial Neural Network* (ANN) menunjukkan performa yang cukup kompetitif dengan nilai *F1-score* tertinggi sebesar 0,9125 pada k = 7. Sementara itu, *Support Vector Machine* (SVM) dan *Logistic Regression* menghasilkan performa yang relatif lebih rendah, khususnya pada nilai *F1-score*, yang menunjukkan keterbatasan dalam mendeteksi kelas minoritas.

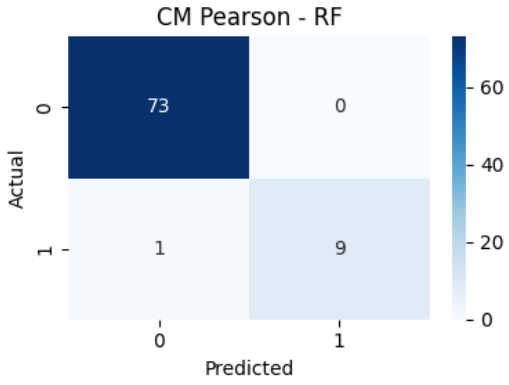


Gambar 7. Confusion Matrix Model Random Forest dengan Metode Chi-Square

Berdasarkan *confusion matrix*, model Random Forest dengan metode *Chi-Square* menunjukkan performa yang sangat baik dengan tingkat kesalahan yang sangat rendah. Model mampu mengklasifikasikan seluruh data kelas mayoritas secara sempurna tanpa kesalahan (FP = 0), serta memiliki kemampuan yang tinggi dalam mendeteksi kelas minoritas dengan nilai recall sebesar 90%. Hal ini menunjukkan bahwa model tidak hanya memiliki akurasi tinggi, tetapi juga mampu menangani ketidakseimbangan data secara efektif. Meskipun demikian, masih terdapat satu kasus *false negative*, yang menunjukkan bahwa terdapat data kemiskinan yang belum terdeteksi oleh model.

2) Pemodelan dengan *Pearson Correlation*

Pada metode *Pearson Correlation*, hasil pemodelan menunjukkan pola yang hampir serupa dengan metode *Chi-Square*. Model *Random Forest* kembali menghasilkan performa terbaik dengan nilai akurasi sebesar 0,9854 dan *F1-score* sebesar 0,9507 pada k = 9, serta menunjukkan kestabilan performa pada berbagai variasi nilai k. Model *Artificial Neural Network* (ANN) juga menunjukkan performa yang kompetitif, dengan nilai *F1-score* mencapai 0,9284 pada k = 7. Hal ini mengindikasikan bahwa pengurangan redundansi fitur melalui *Pearson Correlation* mampu memberikan kontribusi positif terhadap kinerja model.

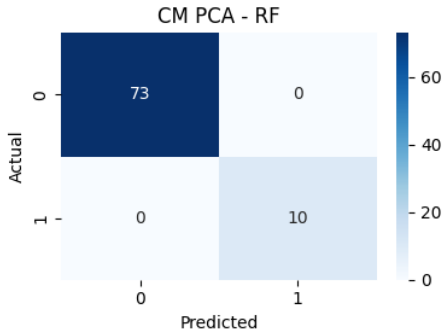


Gambar 8. Confusion Matrix model Random Forest dengan metode Pearson Correlation

Berdasarkan *confusion matrix*, model *Random Forest* dengan metode Pearson menunjukkan performa yang sangat baik dengan tingkat kesalahan yang sangat rendah. Model mampu mengklasifikasikan seluruh data kelas mayoritas secara sempurna tanpa kesalahan (*false positive* = 0), serta memiliki kemampuan yang tinggi dalam mendeteksi kelas minoritas dengan nilai *recall* sebesar 90%. Hal ini menunjukkan bahwa model tidak hanya memiliki akurasi yang tinggi, tetapi juga mampu menangani ketidakseimbangan data secara efektif. Namun demikian, masih terdapat satu kasus *false negative*, yang menunjukkan bahwa terdapat data kemiskinan yang belum berhasil terdeteksi oleh model.

3) Pemodelan dengan PCA

Pada metode *Principal Component Analysis* (PCA), performa model secara umum tetap tinggi, namun cenderung lebih rendah dibandingkan metode *Chi-Square* dan *Pearson Correlation*. Model *Random Forest* masih menjadi model terbaik dengan nilai *F1-score* sebesar 0,9444 pada k = 3 dan 0,9406 pada k = 9, serta menunjukkan kestabilan performa pada berbagai variasi nilai k. Model *Artificial Neural Network* (ANN) juga menunjukkan performa yang stabil dengan nilai akurasi sekitar 0,97, sedangkan *Support Vector Machine* (SVM) dan *Logistic Regression* menghasilkan nilai *F1-score* yang relatif lebih rendah dibandingkan model lainnya.



Gambar 9. Confusion Matrix Model Random Forest dengan Metode Principal Component Analysis

Hasil tersebut diperkuat oleh analisis *confusion matrix* pada model *Random Forest* dengan metode *PCA*, yang menunjukkan performa klasifikasi yang sangat tinggi. Model mampu mengklasifikasikan seluruh data dengan tepat tanpa kesalahan (*false positive* = 0 dan *false negative* = 0), sehingga menghasilkan nilai *precision*, *recall*, dan *F1-score* sebesar 1,00. Hal ini menunjukkan bahwa model mampu memisahkan kelas mayoritas dan minoritas secara sempurna pada data validasi.

Namun demikian, hasil tersebut perlu diinterpretasikan secara hati-hati. Performa yang sempurna ini kemungkinan dipengaruhi oleh distribusi data pada fold tertentu, mengingat secara keseluruhan metode PCA tidak selalu menunjukkan performa yang lebih unggul dibandingkan metode lainnya. Oleh karena itu, meskipun hasil pada confusion matrix menunjukkan kinerja optimal, evaluasi secara keseluruhan tetap perlu mempertimbangkan hasil rata-rata dari seluruh proses *cross-validation*.

Berdasarkan *confusion matrix*, model *Random Forest* dengan metode PCA menunjukkan performa yang sangat tinggi dengan tingkat klasifikasi yang sempurna, di mana seluruh data berhasil diprediksi dengan benar tanpa kesalahan (FP = 0 dan FN = 0). Model mampu mengklasifikasikan seluruh data kelas mayoritas maupun minoritas secara tepat, sehingga menghasilkan nilai *precision*, *recall*, dan F1-score sebesar 1,00. Meskipun demikian, hasil ini perlu diinterpretasikan secara hati-hati karena kemungkinan dipengaruhi oleh distribusi data pada *fold* tertentu, mengingat secara keseluruhan performa PCA tidak selalu lebih unggul dibandingkan metode lainnya.

E. Evaluation

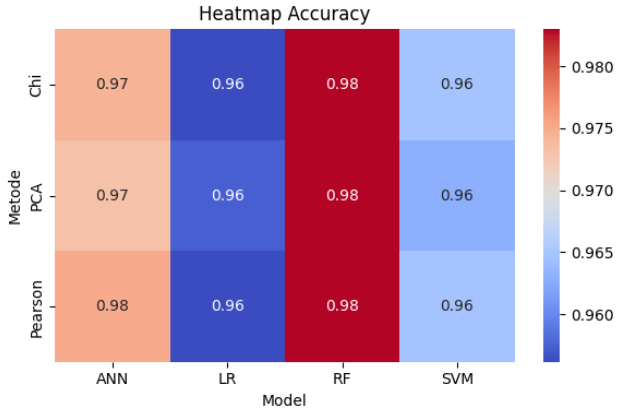
Evaluasi model dilakukan untuk menilai kinerja algoritma dalam mengklasifikasikan tingkat kemiskinan menggunakan metrik *accuracy* dan *F1-score*. Pengujian dilakukan dengan pendekatan *Stratified K-Fold Cross Validation* pada variasi nilai k = 3, 5, 7, dan 9. Hasil evaluasi menunjukkan bahwa metode seleksi dan reduksi fitur serta algoritma yang digunakan memberikan pengaruh signifikan terhadap performa model.

Secara umum, metode *Chi-Square* dan *Pearson Correlation* menunjukkan performa yang paling baik dan

relatif setara. Kedua metode ini secara konsisten menghasilkan nilai akurasi tertinggi sebesar 0,9854 dan *F1-score* sebesar 0,9507 pada model *Random Forest*. Sebaliknya, metode *Principal Component Analysis* (PCA) menghasilkan performa yang sedikit lebih rendah, khususnya pada nilai *F1-score*, yang mengindikasikan adanya penurunan kemampuan model dalam mendeteksi kelas minoritas akibat proses reduksi dimensi.

Dari sisi algoritma, *Random Forest* merupakan model dengan performa paling unggul pada seluruh metode pengolahan fitur. Model ini secara konsisten menghasilkan nilai akurasi di atas 0,98 serta *F1-score* tertinggi dibandingkan model lainnya, yang menunjukkan kemampuannya dalam menangani pola data yang kompleks serta kondisi data yang tidak seimbang. Model *Artificial Neural Network* (ANN) menunjukkan performa yang cukup kompetitif dan stabil, sedangkan *Support Vector Machine* (SVM) dan *Logistic Regression* menunjukkan performa yang relatif lebih rendah, terutama dalam nilai *F1-score*.

Performa model pada berbagai nilai k menunjukkan kestabilan yang baik. Perbedaan nilai akurasi dan *F1-score* antar variasi k tidak menunjukkan perubahan yang signifikan, yang mengindikasikan bahwa model memiliki kemampuan generalisasi yang baik. Namun demikian, pada beberapa skenario, nilai k yang lebih besar cenderung menghasilkan *F1-score* yang sedikit lebih tinggi.



Gambar 10. Heatmap Accuracy

Berdasarkan visualisasi tersebut, terlihat bahwa model *Random Forest* secara konsisten menghasilkan nilai akurasi tertinggi pada seluruh metode, diikuti oleh model ANN, SVM, dan Logistic Regression. Selain itu, metode *Chi-Square* dan *Pearson Correlation* menunjukkan performa yang lebih baik dibandingkan PCA, meskipun perbedaannya tidak terlalu signifikan. Visualisasi ini memperkuat hasil evaluasi numerik yang telah diperoleh sebelumnya.

IV. KESIMPULAN

Berdasarkan hasil penelitian, penerapan metode klasifikasi pada dataset tingkat kemiskinan di Indonesia yang terdiri dari 514 data wilayah menunjukkan bahwa model mampu menghasilkan performa yang sangat baik meskipun data bersifat tidak seimbang dengan rasio sekitar 7:1 antara kelas tidak miskin dan miskin. Dari sisi algoritma, *Random Forest* merupakan model dengan performa terbaik dibandingkan *Support Vector Machine* (SVM), *Logistic Regression*, dan *Artificial Neural Network* (ANN), dengan nilai akurasi tertinggi sebesar 0,9854 dan *F1-score* sebesar 0,9507. Hasil evaluasi menggunakan *Stratified K-Fold Cross Validation* menunjukkan bahwa model memiliki kestabilan performa yang tinggi serta kemampuan generalisasi yang baik. Selain itu, analisis *confusion matrix* menunjukkan bahwa model mampu mengklasifikasikan kelas mayoritas secara sempurna tanpa kesalahan, serta memiliki kemampuan yang tinggi dalam mendeteksi kelas minoritas dengan nilai *recall* mencapai 90%.

Dari sisi metode seleksi dan reduksi fitur, *Chi-Square* dan *Pearson Correlation* terbukti lebih efektif dibandingkan *Principal Component Analysis* (PCA), karena mampu mempertahankan fitur yang relevan terhadap variabel target sehingga menghasilkan performa model yang lebih optimal. Secara keseluruhan, kombinasi metode *Chi-Square* dan algoritma *Random Forest* merupakan pendekatan terbaik dalam penelitian ini karena menghasilkan performa yang tinggi, stabil, dan mampu menangani ketidakseimbangan data. Penerapan teknik *Synthetic Minority Oversampling Technique* (SMOTE) juga terbukti berkontribusi dalam meningkatkan kemampuan model dalam mendeteksi kelas minoritas, sehingga dapat menjadi solusi yang efektif dalam pengembangan sistem klasifikasi tingkat kemiskinan berbasis data.

DAFTAR PUSTAKA

[1] World Bank, “the World Bank Has Set a Clear Mission : Ending Extreme Poverty and Boosting Shared,” 2024.

[2] Badan Pusat Statistik (BPS), “Profil Kemiskinan di Indonesia Maret 2024,” *Present. Pendud. miskin Maret 2010-2024*, no. 50, 2024.

[3] Badan Pusat Statistik, *Indonesian Sustainable Development Goals Indicators 2024 - BPS-Statistics Indonesia*. 2024. Accessed: Dec. 19, 2025. [Online]. Available: https://www.bps.go.id/en/publication/2024/12/31/936a26d5d2b168b9971d3b02/indonesian-sustainable-development-goals-indicators-2024.html

[4] UNDP, “Unstacking Global Poverty: Data for high impact action,” *Glob. Multi-dimensional Poverty Index 2023*, pp. 1–2, 2023, [Online]. Available: https://hdr.undp.org/content/2023-global-multidimensional-poverty-index-mpi#/indicies/MPI

[5] UNDP & OPHI, “2023 Global Multidimensional Poverty Index (MPI) | Human Development Reports,” Jul. 2023. Accessed: Dec. 18, 2025. [Online]. Available: https://hdr.undp.org/content/2023-global-multidimensional-poverty-index-mpi

[6] S. Alkire *et al.*, “The global Multidimensional Poverty Index (MPI) 2023 disaggregation results and methodological note Sabina Alkire *, Usha Kanagaratnam **, Nicolai Suppa ***,” no. July, 2023.

[7] S. Alkire, U. Kanagaratnam, and N. Suppa, “The global Multidimensional Poverty Index,” no. October, pp. 2–17, 2023, doi: 10.18356/9789210028356c002.

[8] L. Nuzula, A. Prahutama, and A. R. Hakim, “Klasifikasi Status Kemiskinan Rumah Tangga Dengan Metode Support Vector Machines (Svm) Dan Classification And Regression Trees (Cart) Menggunakan Gui R (Studi Kasus di Kabupaten Wonosobo Tahun 2018),” *J. Gaussian*, vol. 9, no. 4, pp. 525–534, 2020, doi:

10.14710/j.gauss.v9i4.29449.
[9] D. V. Ramadhanti, R. Santoso, and T. Widiharih, "Perbandingan Smote Dan Adasyn Pada Data Imbalance Untuk Klasifikasi Rumah Tangga Miskin Di Kabupaten Temanggung Dengan Algoritma K-Nearest Neighbor," *J. Gaussian*, vol. 11, no. 4, pp. 499–505, 2023, doi: 10.14710/j.gauss.11.4.499-505.
[10] N. N. Sholihah and A. Hermawan, "Implementation of Random Forest and Smote Methods for Economic Status Classification in Cirebon City," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1387–1397, 2023, doi: 10.52436/1.jutif.2023.4.6.1135.
[11] R. Siringoringo, D. Arisandi, E. Kurniawan, and E. B. Nababan, "Model Klasifikasi Dengan Logistic Regression Dan Recursive Classification Model Using Logistic Regression And Recursive," vol. 11, no. 4, 2024, doi: 10.25126/jtiik.1148198.
[12] S. Rahayu and Y. Yamasari, "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)," vol. 05, pp. 440–446, 2024.
[13] V. J. S. Vimal, M. K. Mi, and Y. Lee, "AI - based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, 2021, doi: 10.1007/s11227-020-03481-x.
[14] Y. Muamar and A. Muhajirin, "Penerapan Jaringan Saraf Tiruan Dengan Metode Backpropagation Untuk Memprediksi Tingkat Kelulusan Mahasiswa Perguruan Tinggi," *Digit. Transform. Technol.*, vol. 4, no. 1, pp. 214–224, 2024, doi: 10.47709/digitech.v4i1.3810.
[15] L. Mardiana, D. Kusnandar, and N. Satyahadewi, "Analisis Diskriminan Dengan K Fold Cross Validation Untuk Klasifikasi Kualitas Air Di Kota Pontianak," vol. 11, no. 1, pp. 97–102, 2022.
[16] A. P. Sabiq Sofyan, "Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Tingkat Pendapatan Pekerja," vol. 2019, pp. 868–877, 2021.
[17] P. R. Sihombing and A. M. Arsani, "Comparison of Machine Learning Methods in Classifying Poverty in Indonesia in 2018," *J. Tek. Inform.*, vol. 2, no. 1, pp. 51–56, 2021, doi: 10.20884/1.jutif.2021.2.1.52.
[18] R. Ariefudin, M. Alfarizzi, N. Machmudah, E. Aminah, Y. Febiyanti, and Wasono, "Analisis Diskriminan untuk Klasifikasi Tingkat Kemiskinan di Perkotaan Menurut Provinsi Berdasarkan Bagian Wilayah di Indonesia Tahun 2022," *Pros. Semin. Nas. Mat. Stat. dan Apl.*, pp. 214–223, 2023.
[19] H. A. Ahmed, P. J. M. Ali, A. K. Faeq, and S. M. Abdullah, "An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method," pp. 29–37, 2022, doi: 10.14500/aro.10970.
[20] R. Firliana, R. Wulanningrum, and W. Sasongko, "Implementasi Principal Component Analysis (PCA) Untuk Pengenalan Wajah Manusia," *Nusant. Eng. ISSN 2355-6684*, vol. 2, no. 1, pp. 65–69, 2005.
[21] D. Leni, A. Dwiharzandis, R. Sumiati, and S. Afriyani, "Seleksi Fitur Berdasarkan Korelasi Pearson dalam Pemodelan Efisiensi Energi Bangunan Feature Selection Based on Pearson Correlation in Building Energy Efficiency Modeling," vol. 08, 2023.
[22] A. Ratna, R. Rakhmat, and I. Novita, "Perbandingan Metode Seleksi Fitur Chi-Square dan Information Gain untuk Peningkatan Interpretabilitas dan Optimasi Kinerja Model TabNet," vol. 03, pp. 253–262, 2025.
[23] P. Putu, N. Ardhaneswari, I. W. C. Suwitra, and J. J. I. S. Siwirabuda, "Analisis Korelasi Pearson Dalam Menentukan Hubungan Harga Dengan Volume Penjualan Wardah Matte Lip Cream Pada Platform E-Commerce Shopee," vol. 02, no. 02, pp. 151–156, 2024.
[24] R. A. Nugraha, E. W. Hidayat, N. I. Kurniati, and R. N. Shofa, "Klasifikasi Jenis Buah Jambu Biji Menggunakan Algoritma Principal Component Analysis dan K-Nearest Neighbor," vol. 7, no. 1, pp. 1–7, 2023.
[25] D. Herinanto, B. H. S. Utami, D. Arif, and M. Gumanti, "Analisis Chi Square Zona Wilayah Marketing Terhadap Penjualan Produk Ekonomi Kreatif," vol. 6, no. 09, pp. 1626–1637, 2024.
[26] J. R. J.M. Gorriz, R. Martin Clemente, F. Segovia, "I S K- Fold Cross Validation The Best Model Selection Method For M Achine L Earning ?," 2024.
[27] T. Wahyuningsih and E. Rahwanto, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," vol. 4, no. 1, pp. 13–20, 2021.
[28] W. Zhang, A. Bifet, X. Zhang, and L. G. Aug, "FARF: A Fair and Adaptive Random Forests Classifier," pp. 1–12.
[29] M. Arya and C. S. S. Bedi, "Survey on SVM and their application in image classification," 2018.
[30] S. Rahmah, H. Azis, D. Widyawati, and A. U. Tenripada, "Prediksi potensi donatur menggunakan model Logistic Regression," vol. 4, no. 1, pp. 31–37, 2023.
[31] R. Mawarni, "Analisis Determinan Financialdistress Dengan Menggunakan Metode Artificial Neural Network Dalam Perspektif Ekonomi Islam," 2022.
[32] E. Hasibuan and E. Allistair, "Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play," vol. 1, no. 3, pp. 13–24, 2022.
[33] T. Hidayati and B. F. Putra, "Scientia Sacra : Jurnal Sains , Teknologi dan Masyarakat Implementasi Deep Learning Untuk Image Classification menggunakan Convolutional Neural Network Pada Citra Wayang (Studi Kasus : SDN Leuwibatu 03)," vol. 4, no. 1, pp. 1–7, 2024.
[34] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and Rismayanti, "Analisis Algoritma Gradient Boosting , AdaBoost dan CatBoost dalam Klasifikasi Kualitas Air Analysis of Gradient Boosting , Adaboost , Catboost Algorithms in Water Quality Classification," vol. 8, pp. 392–402, 2022.