VOL. 3. NO. 02 MARET 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

ANALISIS KLASIFIKASI TINGKAT KEMISKINAN MENGGUNAKAN ALGORITMA RANDOM FOREST PADA KABUPATEN KOTA DI INDONESIA

Anggy Setiawan[1], Nana Suarna[2], Agus Bahtiar[3], Faturrohman[4].

Program Studi Teknik Informatika[12]
Program Studi Sistem Informasi[3]
Program Studi Rekayasa Perangkat Lunak[4]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
setiawananggy32@gmail.com

(*) Corresponding Author : setiawananggy32@gmail.com
Published : 30 Maret 2026

Abstract—*Poverty is a multidimensional phenomenon involving deprivation in income, education, health, and access to basic services, and it varies substantially across Indonesian districts and municipalities. This study evaluates the effectiveness of the Random Forest algorithm in classifying multidimensional poverty levels at the district/city level using a supervised machine learning approach. A secondary dataset from Kaggle was employed, consisting of 514 records of socioeconomic indicators representing regions across Indonesia. The research pipeline includes data collection, cleaning and numeric standardization, feature selection, train–test splitting, Random Forest model training, and performance assessment using accuracy, precision, Recall, F1-score, confusion matrix, and Out-of-Bag (OOB) estimation. Results indicate an OOB score of 0.9854 and a test accuracy of 98%. Importantly, the model achieved perfect Recall for the minority (poor) class with zero false negatives, minimizing exclusion errors in poverty targeting. Feature importance analysis highlights sanitation access, average years of schooling, and per-capita expenditure as key predictors of poverty classification. Overall, the Random Forest model proves highly reliable, stable, and fair for multidimensional poverty classification in heterogeneous and imbalanced regional data. These findings support data-driven policymaking by enabling more precise identification of priority areas for poverty alleviation programs. Integrating machine learning models with official development indicators can further enhance the effectiveness and equity of regional development planning in Indonesia.*

Keywords ***:*** *Poverty level; Data classification; Random Forest algorithm; Socioeconomic indicators; Indonesian regencies/cities.*

Abstrak—Kemiskinan merupakan fenomena multidimensi yang mencakup keterbatasan pendapatan, pendidikan, kesehatan, dan akses layanan dasar, serta memperlihatkan disparitas kuat antar kabupaten/kota di Indonesia. Penelitian ini bertujuan mengevaluasi efektivitas algoritma *Random Forest* untuk mengklasifikasikan tingkat kemiskinan multidimensi pada level kabupaten/kota menggunakan pendekatan supervised learning. Dataset sekunder diperoleh dari Kaggle dan memuat 514 entri indikator sosial-ekonomi dari seluruh wilayah Indonesia. Tahapan penelitian meliputi pengumpulan data, pembersihan dan standarisasi nilai, pemilihan fitur, pembagian data latih-uji, pelatihan model *Random Forest*, serta evaluasi performa menggunakan metrik akurasi, precision, *Recall*, F1-score, confusion matrix, dan Out-of-Bag (OOB). Hasil menunjukkan skor OOB sebesar 0,9854 dengan akurasi pengujian 98%. Model mampu mendeteksi kelas minoritas (miskin) secara sempurna tanpa false negative, sehingga mengurangi risiko exclusion error dalam penargetan bantuan. Analisis feature importance mengindikasikan variabel seperti akses sanitasi layak, rata-rata lama sekolah, dan pengeluaran per kapita sebagai kontributor dominan terhadap klasifikasi. Kesimpulannya, *Random Forest* terbukti sangat andal, stabil, dan adil untuk klasifikasi kemiskinan multidimensi pada data kabupaten/kota yang heterogen dan cenderung tidak seimbang. Temuan ini mendukung pengembangan kebijakan publik berbasis data, terutama untuk pemetaan wilayah prioritas dan penajaman sasaran program pengentasan kemiskinan. Integrasi model pembelajaran mesin dengan indikator pembangunan resmi berpotensi memperkuat akurasi perencanaan pembangunan daerah.

Kata kunci: Tingkat kemiskinan; Klasifikasi data; Algoritma *Random Forest*; Indikator sosial ekonomi; Kabupaten/Kota Indonesia.

INTRODUCTION

Kemiskinan, ketimpangan, dan eksklusi sosial merupakan isu sosial ekonomi yang saling terkait dan menjadi tantangan utama dalam pembangunan berkelanjutan di banyak negara, termasuk Indonesia. Menurut [1], pengukuran kemiskinan tidak hanya terbatas pada aspek moneter seperti pendapatan, namun juga mencakup dimensi non-moneter seperti akses terhadap pendidikan, kesehatan, dan layanan dasar. Penilaian yang komprehensif terhadap kemiskinan memerlukan indikator multidimensional yang mampu menangkap kompleksitas realitas sosial.

Menurut [2] menambahkan bahwa eksklusi sosial mencerminkan ketidakmampuan individu atau kelompok dalam mengakses hak-hak dasar sosial-ekonomi, yang sering kali beririsan dengan kemiskinan. Indikator seperti status pekerjaan, tingkat pendidikan, dan kondisi rumah tangga menjadi tolok ukur penting dalam mengidentifikasi individu yang rentan terhadap eksklusi. Dalam konteks ini, pendekatan berbasis data diperlukan untuk menganalisis fenomena secara lebih akurat dan menyeluruh.

Penggunaan kecerdasan buatan, khususnya algoritma machine learning (ML), telah menjadi pendekatan yang menjanjikan untuk memetakan kemiskinan secara lebih efisien dan akurat. Menurut [3], algoritma seperti *Random Forest* mampu menangani data spasial dan sosial secara simultan, serta menghasilkan prediksi kemiskinan multidimensional dengan akurasi tinggi. Hal ini sejalan dengan temuan [4] dan [5]. yang menunjukkan bahwa model ML lebih unggul dibandingkan metode statistik tradisional dalam mengklasifikasikan status kemiskinan berdasarkan data rumah tangga.

Kelebihan *Random Forest* yang bersifat non-parametrik, kemampuannya menangani prediktor campuran (numerik dan kategorikal), serta kemampuannya memberikan estimasi error internal melalui out-of-bag (OOB) error menjadikannya sangat sesuai untuk pengklasifikasian sosial ekonomi yang kompleks [6]. Selain itu, algoritma ini juga memberikan insight penting mengenai fitur yang paling berpengaruh, yang sangat berguna dalam proses seleksi variabel pada studi kemiskinan.

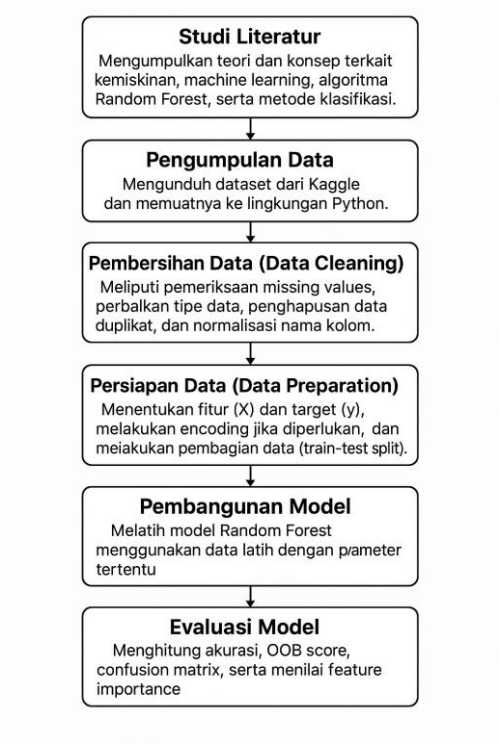
Penerapan pendekatan data-driven dalam perumusan kebijakan juga semakin menonjol. [7] dan [8]. Menunjukkan bagaimana pemanfaatan data spasial dan pencahayaan malam hari mampu memperbaiki ketepatan intervensi kebijakan di tingkat daerah. [9] menegaskan bahwa pendekatan machine learning dalam pemetaan kemiskinan sub-nasional memberikan presisi dan ketepatan waktu yang lebih tinggi dibandingkan metode konvensional, sehingga dapat meningkatkan efektivitas kebijakan pengentasan kemiskinan.

Dengan latar belakang tersebut, penting bagi penelitian ini untuk mengkaji lebih lanjut potensi dan tantangan dalam penggunaan machine learning, khususnya algoritma *Random Forest*, untuk mengklasifikasikan kemiskinan multidimensi di Indonesia. Pendekatan ini diharapkan dapat mendukung pengambilan keputusan berbasis data yang lebih adaptif dan responsif dalam merancang program pengentasan kemiskinan di tingkat lokal.

MATERIALS AND METHODS

Penelitian ini menggunakan pendekatan kuantitatif, yaitu pendekatan yang mengandalkan pada pengolahan data numerik dan analisis statistik untuk menjawab rumusan masalah. Pendekatan ini dipilih karena penelitian bertujuan menghasilkan model klasifikasi berdasarkan data yang bersifat terukur, seperti angka kemiskinan dan indikator sosial ekonomi lainnya.

Dari sisi desain, penelitian ini menggunakan desain eksperimen komputasional (computational experiment). Desain ini memanfaatkan perangkat lunak dan algoritma machine learning, khususnya *Random Forest*, untuk melakukan percobaan pemodelan, pelatihan, dan evaluasi secara sistematis. Seluruh tahapan analisis dilakukan dengan memproses dataset menggunakan teknik komputasi, sehingga hasilnya dapat direplikasi dengan mudah dan objektif. Desain Penelitian seperti pada gambar 1



Gambar 1 Desain Penelitian

Desain penelitian ini juga mencakup langkah-langkah eksplorasi data, pembersihan data, pembentukan model, evaluasi performa model, hingga interpretasi hasil. Pola ini memastikan bahwa setiap tahap menghasilkan keluaran yang valid dan mendukung tahap selanjutnya.

Prosedur penelitian dilaksanakan melalui beberapa tahapan yang terstruktur untuk memastikan alur kerja yang sistematis serta menghasilkan analisis yang valid. Adapun tahapan penelitian yang dilakukan adalah sebagai berikut:

Pengumpulan Data Pada tahap awal, peneliti mengunduh dataset dari platform Kaggle sebagai sumber data utama. Dataset tersebut kemudian dimuat ke dalam lingkungan kerja Python menggunakan pustaka pandas agar dapat diolah dan dianalisis lebih lanjut.

Pembersihan Data (*Data Cleaning*) Tahap ini mencakup identifikasi dan penanganan missing values, perbaikan tipe data yang tidak sesuai, serta penghapusan baris atau entri yang terdeteksi sebagai duplikat. Peneliti juga melakukan normalisasi dan standarisasi nama kolom untuk mencegah inkonsistensi serta memastikan seluruh variabel siap digunakan pada proses analisis berikutnya.

Persiapan Data (*Data Preparation*) Pada tahap ini dilakukan pemisahan variabel prediktor (X) dan variabel target (y). Jika terdapat fitur kategorikal, dilakukan encoding agar dapat diproses oleh algoritma pembelajaran mesin. Data kemudian dibagi menjadi data latih dan data uji (train-test split) untuk mengevaluasi performa model secara objektif.

Pembangunan Model Model klasifikasi dibangun menggunakan algoritma *Random Forest*. Pada tahap ini, peneliti menentukan parameter model yang relevan, seperti jumlah pohon keputusan (n_estimators) dan kedalaman maksimum pohon (max_depth). Model kemudian dilatih menggunakan data latih untuk mempelajari pola hubungan antara fitur dan target.

Evaluasi Model Setelah model terlatih, dilakukan evaluasi kinerja menggunakan berbagai metrik seperti akurasi, Out-of-Bag (OOB) score, confusion matrix, serta analisis feature importance untuk mengetahui kontribusi masing-masing fitur. Tahap ini bertujuan menilai sejauh mana model mampu melakukan prediksi secara akurat dan konsisten.

Interpretasi dan Pelaporan Hasil Hasil evaluasi kemudian dianalisis untuk memperoleh pemahaman mendalam mengenai performa model dan faktor-faktor yang mempengaruhi klasifikasi. Tahap akhir adalah penyusunan laporan penelitian secara komprehensif, yang mencakup penjelasan metodologi, hasil analisis, interpretasi temuan, serta implikasi dari penelitian yang dilakukan. Prosedur penelitian ini dirancang agar setiap tahapan saling mendukung dan membentuk alur kerja yang terarah, sehingga kualitas dan validitas hasil penelitian dapat terjamin.

RESULTS AND DISCUSSION

Setelah proses pelatihan model Random Forest selesai, langkah selanjutnya adalah melakukan evaluasi terhadap performa model menggunakan data uji sebanyak 103 baris. Hasil evaluasi dalam bentuk classification report, yang meliputi metrik evaluasi utama yaitu *precision*, *recall*, *f1-score*, dan *support* untuk masing-masing kelas target. Laporan klasifikasi seperti pada gambar 2

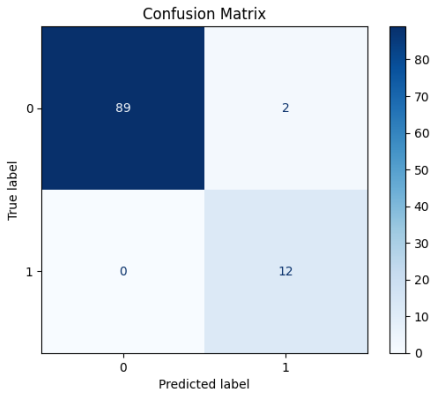
```
--- Laporan Klasifikasi (Classification Report) ---
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        91
           1       0.86      1.00      0.92        12

    accuracy                           0.98       103
   macro avg       0.93      0.99      0.96       103
weighted avg       0.98      0.98      0.98       103
```

Gambar 2 Laporan klasifikasi

Model menunjukkan performa yang sangat baik pada kelas mayoritas (kelas 0), dengan nilai precision sebesar 1.00, recall 0.98, dan f1-score 0.99. Untuk kelas minoritas (kelas 1), meskipun jumlah datanya hanya 12 (dibandingkan 91 untuk kelas 0), model masih mampu meraih recall sempurna (1.00) dan precision sebesar 0.86, menghasilkan f1-score sebesar 0.92.

Secara keseluruhan, model mencapai akurasi total sebesar 0.98, yang menandakan bahwa 98% dari data uji berhasil diklasifikasikan dengan benar. Nilai macro average f1-score sebesar 0.96 dan weighted average f1-score sebesar 0.98 juga menunjukkan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga memperlakukan kedua kelas dengan cukup seimbang, meskipun distribusi datanya tidak merata.

Sebagai bagian dari evaluasi akhir, confusion matrix untuk memperjelas distribusi prediksi model terhadap label aktual pada data uji. Matriks ini memberikan gambaran yang lebih detail tentang performa klasifikasi model Random Forest dalam membedakan dua kelas tingkat kemiskinan, yaitu kelas 0 (tidak miskin) dan kelas 1 (miskin). Confusion matrix seperti pada gambar 3



Gambar 4. 1 Confusion matrix

Dari hasil visualisasi, terlihat bahwa dari total 91 data aktual kelas 0, sebanyak 89 data berhasil diklasifikasikan dengan benar, dan hanya 2 data yang salah diklasifikasikan ke kelas 1. Sementara itu, untuk kelas 1 yang hanya memiliki 12 data, seluruhnya (12 data) berhasil diklasifikasikan dengan benar tanpa kesalahan (false negative = 0). Hal ini menunjukkan bahwa model memiliki sensitivitas tinggi terhadap data kelas minoritas, yang dalam banyak kasus sering diabaikan oleh model klasifikasi biasa.

Secara keseluruhan, hasil confusion matrix ini memperkuat temuan dari classification report sebelumnya, bahwa model mampu memberikan prediksi yang sangat akurat dan seimbang, meskipun terdapat ketidakseimbangan jumlah data antar kelas. Visualisasi ini juga menunjukkan bahwa kesalahan klasifikasi sangat minimal dan tidak mempengaruhi signifikan performa keseluruhan sistem.

Berdasarkan rangkaian eksperimen yang telah dilakukan, model *Random Forest* menunjukkan performa yang sangat kuat dalam melakukan klasifikasi. Proses pelatihan menghasilkan skor OOB (Out-of-Bag) sebesar 0.9854, yang mengindikasikan kemampuan model untuk melakukan generalisasi dengan baik bahkan terhadap data yang tidak terlibat dalam pelatihan.

Evaluasi lebih lanjut melalui laporan klasifikasi memperlihatkan bahwa model mampu mengenali kelas mayoritas maupun minoritas dengan tingkat ketepatan tinggi. Kelas 0 mencapai precision sempurna sebesar 1.00 dan *f1-score* 0.99, sementara kelas 1 yang jumlah datanya jauh lebih sedikit tetap dapat diprediksi dengan baik, terlihat dari nilai *Recall* 1.00 dan *f1-score* 0.92. Hal ini mengindikasikan bahwa model tidak hanya kuat pada kelas dominan, tetapi juga efektif dalam menangani data minoritas.

Confusion matrix menegaskan kekokohan kinerja model: dari 103 data uji, hanya terdapat dua kesalahan prediksi pada kelas 0, sementara seluruh data kelas 1 berhasil diklasifikasikan dengan benar. Ketidakmunculan false negative pada kelas 1 menjadi indikator positif, terutama jika kelas tersebut mewakili kondisi yang sensitif atau berisiko.

Secara keseluruhan, hasil ini menunjukkan bahwa model *Random Forest* layak digunakan sebagai alat prediksi dalam konteks data yang diuji. Kinerja tinggi pada seluruh metrik utama precision, *Recall*, f1-score, serta akurasi memberikan dasar kuat bahwa model dapat diandalkan untuk pengambilan keputusan, pelaporan analitis, maupun implementasi lanjutan di tahap operasional.

CONCLUSION

Penelitian ini menunjukkan bahwa algoritma *Random Forest* mampu mengklasifikasikan kondisi pembangunan antar daerah dengan tingkat akurasi yang sangat tinggi. Model menghasilkan skor Out-of-Bag sebesar 0,9854 serta akurasi pengujian 0,98, yang mengindikasikan kemampuan generalisasi yang kuat terhadap data baru. Selain itu, performa model pada kedua kelas baik kelas mayoritas maupun kelas minoritas sama-sama menunjukkan hasil yang stabil dan presisi, dengan tidak adanya

false negative pada kelas minoritas. Temuan ini menegaskan bahwa *Random Forest* bukan hanya efektif dalam mengenali pola dari variabel indikator pembangunan, tetapi juga mampu memberikan hasil prediksi yang konsisten dan dapat diandalkan.

Keberhasilan model tidak lepas dari kualitas data yang telah diolah melalui proses pembersihan, normalisasi, konversi tipe data, serta penanganan nilai kosong yang dilakukan secara komprehensif. Transformasi dataset yang awalnya penuh inkonsistensi menjadi data yang rapi dan seragam memungkinkan algoritma bekerja optimal dalam menangkap hubungan antar variabel. Proses standardisasi fitur juga berperan besar dalam menyamakan skala tiap variabel sehingga model dapat memproses seluruh informasi secara seimbang. Dengan demikian, penelitian ini menegaskan bahwa kualitas pemrosesan data merupakan fondasi utama yang mendorong tingginya kinerja model.

Secara keseluruhan, hasil penelitian ini menjawab pertanyaan penelitian dengan jelas: algoritma *Random Forest* terbukti efektif, stabil, dan akurat dalam melakukan klasifikasi pembangunan antar daerah ketika ditopang oleh data yang telah diproses dengan baik. Model ini tidak hanya mampu memberikan gambaran yang tepat mengenai kondisi pembangunan, tetapi juga memiliki potensi besar untuk diterapkan dalam proses pengambilan keputusan dan perencanaan kebijakan berbasis data. Dengan performa yang sangat kompetitif, pendekatan ini dapat menjadi alat analitis yang kuat untuk mendukung upaya peningkatan pemerataan pembangunan di tingkat kabupaten/kota.

REFERENCE

[1] G. E. Halkos and P.-S. C. Aslanidis, “Causes and Measures of Poverty, Inequality, and Social Exclusion: A Review,” *Economies*, vol. 11, no. 4, p. 110, 2023, doi: 10.3390/economies11040110.

[2] J. Cuesta, B. López Noval, and M. Niño Zarazúa, “Social exclusion concepts, measurement, and a global estimate,” *PLOS ONE*, vol. 19, no. 2, p. e0298085, 2024, doi: 10.1371/journal.pone.0298085.

[3] J. C. Muñetón Santa and J. A. Manrique Ruiz, “Random Forest approaches for spatial-socioeconomic poverty estimation,” *International Journal of Data Science*, 2023.

[4] Q. Li, S. Yu, D. Échevin, and M. Fan, “Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan,” *Socio Economic Planning Sciences*, vol. 81, p. 101195, 2022, doi: 10.1016/j.seps.2021.101195.

[5] S. M. Satapathy, S. Mohanty, and R. Dash, “Household-level poverty prediction using machine learning models,” *Expert Systems with Applications*, 2023.

[6] W. J. Browne, H. Goldstein, and J. Rasbash, “Random Forests for social and economic classification analysis,” *SAGE Publications*, 2021.

[7] Q. Zheng, H. Li, and Y. Sun, “Nighttime light data for poverty targeting and policy intervention accuracy,” *Remote Sensing of Environment*, 2024.

[8] B. Buttow, “Spatial data integration for poverty-targeting improvement: A policy-driven approach,” *Journal of Development Analytics*, 2025.

[9] P. Corral Rodas, “Sub-national poverty mapping using machine learning approaches,” *Development Policy Review*, 2024.