

PERBANDINGAN KINERJA ALGORITMA C4.5 DAN RANDOM FOREST DALAM KLASIFIKASI KATEGORI GARIS KEMISKINAN DI INDONESIA

Rendi Wijayakusuma¹, Zaehol Fatah², Syarif Aminul Khoiri³

Universitas Ibrahimy

rendiwk39@gmail.com

ABSTRAK

Perkembangan teknologi data mining dan machine learning semakin banyak diterapkan dalam analisis data sosial-ekonomi, termasuk klasifikasi data kemiskinan di Indonesia. Permasalahan dalam penelitian ini adalah belum banyaknya penelitian yang membandingkan performa algoritma C4.5 dan Random Forest pada klasifikasi kategori garis kemiskinan di Indonesia. Penelitian ini bertujuan untuk membandingkan kinerja kedua algoritma dalam mengklasifikasikan kategori garis kemiskinan berdasarkan data Badan Pusat Statistik (BPS) periode 2013–2022. Metode penelitian menggunakan pendekatan data mining dengan tahapan CRISP-DM yang meliputi business understanding, data understanding, data preparation, modeling, dan evaluation. Proses klasifikasi dilakukan menggunakan algoritma C4.5 dan Random Forest dengan pembagian data training sebesar 70% dan testing sebesar 30%. Hasil penelitian menunjukkan bahwa algoritma C4.5 dan Random Forest memperoleh nilai accuracy, precision, recall, dan F1-score sebesar 100%. Selain itu, penelitian ini juga menghasilkan aplikasi berbasis web menggunakan Streamlit yang mampu melakukan komparasi algoritma dan simulasi klasifikasi kategori garis kemiskinan secara interaktif.

Kata kunci: : Data Mining, C4.5, Random Forest, Klasifikasi, Garis Kemiskinan.

ABSTRACT

The development of data mining and machine learning technologies is increasingly being applied to socio-economic data analysis, including poverty data classification in Indonesia. The problem addressed in this research is the limited number of studies comparing the performance of C4.5 and Random Forest algorithms in classifying poverty line categories in Indonesia. This study aims to compare the performance of both algorithms in classifying poverty line categories based on Statistics Indonesia (BPS) data for the 2013–2022 period. The research method employs a data mining approach using the CRISP-DM framework, which includes business understanding, data understanding, data preparation, modeling, and evaluation. The classification process was conducted using C4.5 and Random Forest algorithms with a data split of 70% for training and 30% for testing. The results show that both the C4.5 and Random Forest algorithms achieved accuracy, precision, recall, and F1-score values of 100%. Furthermore, this research produced a web-based application using Streamlit capable of performing algorithmic comparisons and interactive poverty line category classification simulations.

Keywords: Data Mining, C4.5, Random Forest, Classification, Poverty Line.

Pendahuluan

Kemiskinan merupakan salah satu permasalahan sosial dan ekonomi yang masih menjadi tantangan utama dalam pembangunan di Indonesia. Tingginya tingkat kemiskinan berdampak terhadap rendahnya kualitas hidup masyarakat, terbatasnya akses pendidikan dan kesehatan, serta menurunnya produktivitas sumber daya manusia [1]. Kondisi tersebut menyebabkan pemerintah terus melakukan berbagai upaya penanggulangan kemiskinan melalui program bantuan sosial dan pembangunan ekonomi masyarakat [2]. Salah satu indikator yang digunakan untuk mengukur tingkat kesejahteraan masyarakat adalah garis kemiskinan yang ditetapkan oleh Badan Pusat Statistik (BPS) [3]. Garis kemiskinan digunakan sebagai acuan untuk menentukan kategori masyarakat miskin berdasarkan tingkat pengeluaran minimum untuk memenuhi kebutuhan dasar makanan dan non-makanan [4]. Oleh karena itu, klasifikasi kategori garis kemiskinan menjadi hal penting dalam mendukung penyaluran bantuan sosial yang tepat sasaran serta pengambilan kebijakan berbasis data.

Perkembangan teknologi informasi, khususnya pada bidang data mining dan machine learning, memberikan peluang besar dalam proses analisis data kemiskinan [5]. Pemanfaatan algoritma klasifikasi mampu membantu proses identifikasi pola dan pengelompokan data secara lebih cepat, objektif, dan akurat dibandingkan metode konvensional [6]. Salah satu metode klasifikasi yang banyak digunakan adalah algoritma C4.5 yang bekerja dengan membangun pohon keputusan berdasarkan nilai entropy dan information gain. Selain itu, algoritma Random Forest juga banyak digunakan karena mampu meningkatkan stabilitas dan akurasi klasifikasi melalui pendekatan ensemble learning [6]. Beberapa penelitian sebelumnya menunjukkan bahwa kedua algoritma tersebut memiliki performa yang baik dalam proses klasifikasi data sosial dan penerima bantuan sosial [7]. Namun, penelitian yang secara khusus membandingkan performa algoritma

C4.5 dan Random Forest dalam klasifikasi kategori garis kemiskinan di Indonesia masih terbatas.

Selain itu, meningkatnya jumlah data sosial-ekonomi yang tersedia dari berbagai instansi pemerintah menuntut adanya metode pengolahan data yang efektif dan efisien. Data yang besar dan kompleks akan sulit dianalisis secara manual sehingga diperlukan pendekatan berbasis komputasi untuk menghasilkan informasi yang lebih akurat dan cepat. Penggunaan algoritma klasifikasi pada data kemiskinan dinilai mampu membantu proses pengambilan keputusan karena dapat mengidentifikasi pola tertentu berdasarkan karakteristik data yang dimiliki. Dengan demikian, penerapan machine learning dalam klasifikasi garis kemiskinan dapat menjadi solusi dalam meningkatkan kualitas analisis data sosial-ekonomi di Indonesia.

Penerapan algoritma klasifikasi dalam penelitian kemiskinan juga memiliki kelebihan dalam menghasilkan model prediksi yang dapat digunakan secara berkelanjutan. Algoritma C4.5 memiliki kemampuan menghasilkan aturan keputusan yang mudah dipahami sehingga dapat membantu proses interpretasi hasil klasifikasi. Sementara itu, Random Forest memiliki keunggulan dalam meningkatkan akurasi klasifikasi dan mengurangi risiko overfitting melalui pembentukan banyak pohon keputusan secara acak. Perbedaan karakteristik kedua algoritma tersebut menjadi alasan penting dilakukannya penelitian komparasi untuk mengetahui algoritma yang lebih optimal dalam klasifikasi kategori garis kemiskinan berdasarkan data sosial-ekonomi yang digunakan.

Berdasarkan permasalahan tersebut, penelitian ini dilakukan untuk membandingkan kinerja algoritma C4.5 dan Random Forest dalam mengklasifikasikan kategori garis kemiskinan di Indonesia menggunakan dataset garis kemiskinan periode 2013–2022. Penelitian ini menerapkan metode CRISP-DM yang meliputi tahapan business understanding, data understanding, data preparation,

modeling, evaluation, dan deployment [8]. Selain melakukan proses komparasi algoritma, penelitian ini juga menghasilkan aplikasi berbasis web menggunakan framework Streamlit yang mampu melakukan simulasi klasifikasi kategori garis kemiskinan secara interaktif. Hasil penelitian diharapkan dapat memberikan kontribusi dalam pengembangan penerapan data mining pada bidang sosial-ekonomi serta menjadi referensi bagi pemerintah dalam mendukung kebijakan penanggulangan kemiskinan yang lebih efektif dan tepat sasaran.

Materi dan Metode

Penelitian ini menggunakan metode kuantitatif dengan pendekatan data mining untuk membandingkan kinerja algoritma C4.5 dan Random Forest dalam klasifikasi kategori garis kemiskinan di Indonesia. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengolahan data numerik dan pengukuran performa algoritma berdasarkan nilai evaluasi tertentu. Proses penelitian dilakukan secara sistematis menggunakan metodologi CRISP-DM (*Cross-Industry Standard Process for Data Mining*) yang terdiri dari tahapan business understanding, data understanding, data preparation, modeling, evaluation, dan deployment [8].

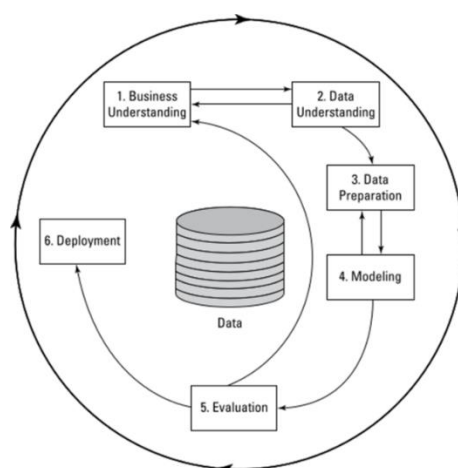
Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Badan Pusat Statistik (BPS) melalui platform Kaggle. Dataset berisi data garis kemiskinan per provinsi di Indonesia periode 2013–2022 dengan pengambilan data pada bulan Maret dan September setiap tahunnya. Variabel yang digunakan meliputi provinsi, tahun, periode, daerah, jenis pengeluaran, dan nilai garis kemiskinan sebagaimana disajikan pada Tabel 1.

Tabel 1. Variabel dataset garis kemiskinan

Variabel	Tipe Data	Keterangan
Provinsi	Kategorikal	Nama provinsi di Indonesia
Tahun	Numerik	Tahun pengamatan
Periode	Kategorikal	Maret atau September

Daerah	Kategorikal	Perkotaan/Perdesaan
Jenis Pengeluaran	Kategorikal	Makanan, Non-makanan, Total
Nilai GK	Numerik	Nilai garis kemiskinan
Kategori Kemiskinan	Kategorikal	Rendah, Menengah, Tinggi

Metode penelitian ini mengadopsi tahapan CRISP-DM sebagai berikut [9], sebagaimana digambarkan pada Gambar 1.



Gambar 1. Tahapan metodologi CRISP-DM

Business Understanding

Tahap ini dilakukan untuk memahami tujuan penelitian dan permasalahan yang akan diselesaikan. Fokus utama penelitian adalah membandingkan performa algoritma C4.5 dan Random Forest dalam mengklasifikasikan kategori garis kemiskinan di Indonesia. Hasil penelitian diharapkan dapat membantu proses analisis data kemiskinan secara lebih akurat dan mendukung pengambilan kebijakan berbasis data.

Data Understanding

Pada tahap ini dilakukan proses pengumpulan dan pemahaman dataset yang digunakan. Dataset dianalisis untuk mengetahui struktur data, tipe atribut, jumlah data, distribusi data, serta mendeteksi adanya missing value atau data tidak konsisten. Analisis awal dilakukan menggunakan Python dengan pustaka Pandas dan NumPy untuk

memperoleh informasi statistik deskriptif dari dataset.

Data Preparation

Tahap persiapan data dilakukan sebelum proses pemodelan. Proses yang dilakukan meliputi:

- a. Data Cleaning — membersihkan data yang memiliki nilai kosong (missing value) dan memperbaiki inkonsistensi data.
- b. Transformasi Data — mengubah data kategorikal menjadi bentuk numerik agar dapat diproses oleh algoritma klasifikasi.
- c. Feature Selection — memilih atribut yang relevan untuk proses klasifikasi.
- d. Pembagian Data — dataset dibagi menjadi data latih (training data) sebesar 70% dan data uji (testing data) sebesar 30%.

Tahapan ini bertujuan menghasilkan dataset yang siap digunakan dalam proses klasifikasi.

Algoritma C4.5

Algoritma C4.5 digunakan untuk membentuk model klasifikasi berbasis pohon keputusan (decision tree). Proses pembentukan pohon dilakukan dengan menghitung nilai entropy dan information gain pada setiap atribut. Atribut dengan nilai gain tertinggi akan dipilih sebagai akar pohon keputusan. Pada penelitian ini, algoritma C4.5 digunakan untuk mengklasifikasikan kategori garis kemiskinan berdasarkan atribut yang terdapat pada dataset.

Algoritma Random Forest

Random Forest merupakan metode ensemble learning yang bekerja dengan membangun banyak pohon keputusan secara acak. Setiap pohon dibentuk menggunakan teknik bootstrap sampling dan pemilihan atribut secara acak. Hasil klasifikasi akhir diperoleh melalui proses majority voting dari seluruh pohon keputusan. Algoritma ini

digunakan untuk meningkatkan stabilitas dan akurasi hasil klasifikasi.

Evaluasi Model

Evaluasi model dilakukan untuk mengetahui performa masing-masing algoritma dalam proses klasifikasi kategori garis kemiskinan. Pengukuran performa dilakukan menggunakan confusion matrix dengan beberapa metrik evaluasi, yaitu:

- a. Accuracy — mengukur tingkat ketepatan prediksi model terhadap seluruh data.
- b. Precision — mengukur tingkat ketepatan model dalam memprediksi kelas positif.
- c. Recall — mengukur kemampuan model dalam menemukan seluruh data positif.
- d. F1-Score — mengukur keseimbangan antara precision dan recall.

Hasil evaluasi dari algoritma C4.5 dan Random Forest kemudian dibandingkan untuk menentukan algoritma yang memiliki performa terbaik dalam klasifikasi kategori garis kemiskinan di Indonesia.

Hasil dan Pembahasan

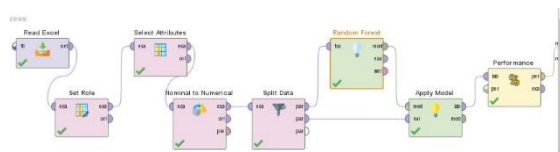
Deskripsi Dataset

Pada penelitian ini digunakan dataset garis kemiskinan Indonesia periode 2013–2022 yang diperoleh dari Badan Pusat Statistik (BPS) melalui Kaggle. Dataset berisi informasi mengenai garis kemiskinan berdasarkan provinsi, tahun, periode pengamatan, daerah, jenis pengeluaran, dan nilai garis kemiskinan. Dataset kemudian diklasifikasikan ke dalam tiga kategori, yaitu kategori rendah, sedang, dan tinggi. Tahap awal dilakukan dengan proses pembacaan dataset menggunakan RapidMiner dan Python. Dataset yang telah diinput kemudian diperiksa struktur datanya untuk memastikan tidak terdapat missing value maupun inkonsistensi data. Selanjutnya dilakukan proses transformasi data kategorikal menjadi bentuk numerik agar dapat diproses

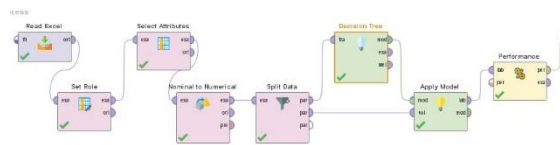
oleh algoritma klasifikasi. Berdasarkan hasil pengamatan, dataset memiliki distribusi kelas yang cukup seimbang sehingga dapat digunakan dengan baik dalam proses klasifikasi menggunakan algoritma C4.5 dan Random Forest.

Hasil Pengujian

Berdasarkan pengujian melalui proses RapidMiner dengan menggunakan algoritma Random Forest, diperoleh hasil seperti pada Gambar 2, Gambar 3, dan Gambar 4 berikut.



Gambar 2. Hasil pengujian algoritma Random Forest pada RapidMiner



Gambar 3. Hasil pengujian algoritma Random Forest pada RapidMiner

PerformanceVector

```
PerformanceVector:
accuracy: 100.00%
ConfusionMatrix:
True:  Sedang  Rendah  Tinggi
Sedang: 479    0      0
Rendah: 0     463    0
Tinggi: 0     0      533
classification_error: 0.00%
ConfusionMatrix:
True:  Sedang  Rendah  Tinggi
Sedang: 479    0      0
Rendah: 0     463    0
Tinggi: 0     0      533
weighted_mean_recall: 100.00%, weights: 1, 1, 1
ConfusionMatrix:
True:  Sedang  Rendah  Tinggi
Sedang: 479    0      0
Rendah: 0     463    0
Tinggi: 0     0      533
weighted_mean_precision: 100.00%, weights: 1, 1, 1
ConfusionMatrix:
True:  Sedang  Rendah  Tinggi
Sedang: 479    0      0
Rendah: 0     463    0
Tinggi: 0     0      533
```

Gambar 4. Hasil pengujian algoritma Random Forest pada RapidMiner

Berdasarkan hasil pengujian, algoritma C4.5 dan Random Forest sama-sama menghasilkan performa klasifikasi yang sama-sama sangat tinggi. Perbandingan performa kedua algoritma dapat dilihat pada Tabel 2 berikut.

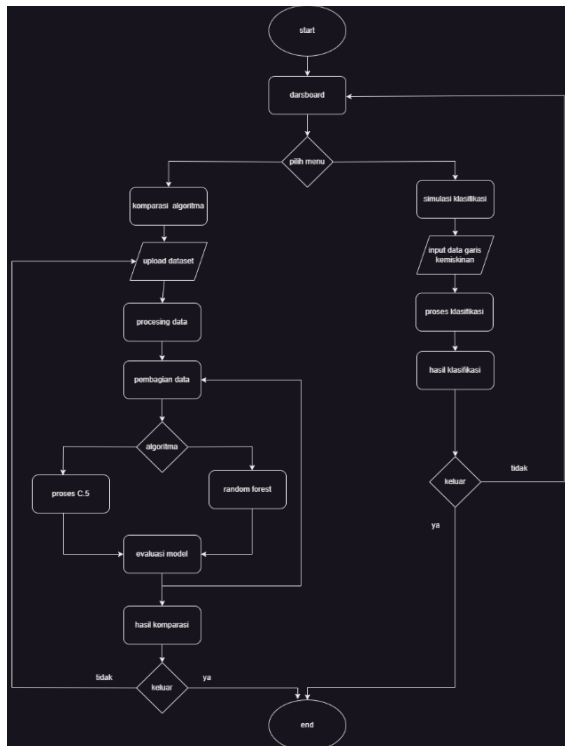
Tabel 2. Perbandingan performa algoritma C4.5 dan Random Forest

Algoritma	Accuracy	Precision	Recall	Error
C4.5	100%	100%	100%	0%
Random Forest	100%	100%	100%	0%

Berdasarkan hasil pengujian yang telah dilakukan, algoritma C4.5 dan Random Forest sama-sama memiliki kemampuan klasifikasi yang sangat baik pada dataset garis kemiskinan di Indonesia. Kedua algoritma berhasil memperoleh nilai accuracy, precision, dan recall sebesar 100%, sehingga menunjukkan bahwa seluruh data dapat diklasifikasikan dengan benar. Meskipun memiliki tingkat akurasi yang sama, kedua algoritma mempunyai karakteristik yang berbeda dalam proses klasifikasi data.

Analisis Perancangan Sistem

Analisis sistem pada penelitian ini dilakukan berdasarkan perancangan sistem yang telah dibuat, sebagaimana ditampilkan pada Gambar 5.



Gambar 5. Rancangan alur sistem aplikasi

Sistem dimulai dari halaman dashboard sebagai tampilan utama aplikasi. Pada tahap ini pengguna dapat memilih menu yang tersedia, yaitu menu komparasi algoritma dan simulasi klasifikasi. Pada menu komparasi algoritma, pengguna melakukan upload dataset garis kemiskinan yang kemudian diproses melalui tahapan preprocessing data dan pembagian data menjadi data training dan testing. Setelah itu sistem melakukan proses klasifikasi menggunakan algoritma C4.5 dan Random Forest untuk memperoleh hasil evaluasi model. Hasil evaluasi kemudian ditampilkan dalam bentuk perbandingan nilai accuracy, precision, recall, dan F1-score guna mengetahui algoritma dengan performa terbaik.

Pembuatan Aplikasi

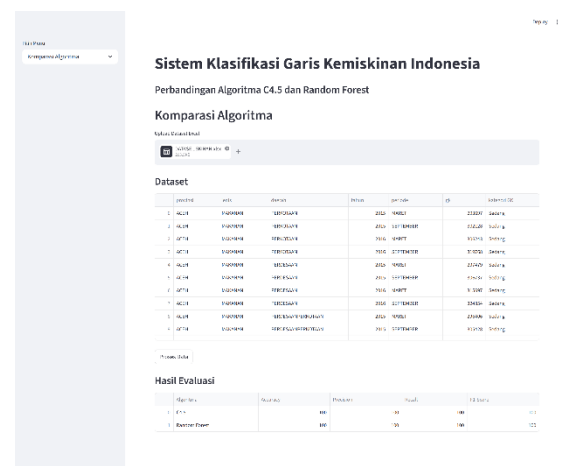
Pembuatan aplikasi dilakukan menggunakan bahasa pemrograman Python dengan framework Streamlit. Aplikasi dibangun berbasis web sehingga dapat

digunakan secara interaktif untuk melakukan proses klasifikasi kategori garis kemiskinan di Indonesia. Gambar 6 menunjukkan halaman dashboard aplikasi.



Gambar 6. Halaman dashboard aplikasi

Halaman tersebut merupakan tampilan utama aplikasi yang berisi informasi sistem klasifikasi garis kemiskinan Indonesia. Pada halaman ini tersedia menu navigasi yang digunakan untuk mengakses fitur komparasi algoritma dan simulasi klasifikasi. Tampilan sidebar aplikasi ditunjukkan pada Gambar 7.



Sistem Klasifikasi Garis Kemiskinan Indonesia
Perbandingan Algoritma C4.5 dan Random Forest

Komparasi Algoritma

Upload Dataset

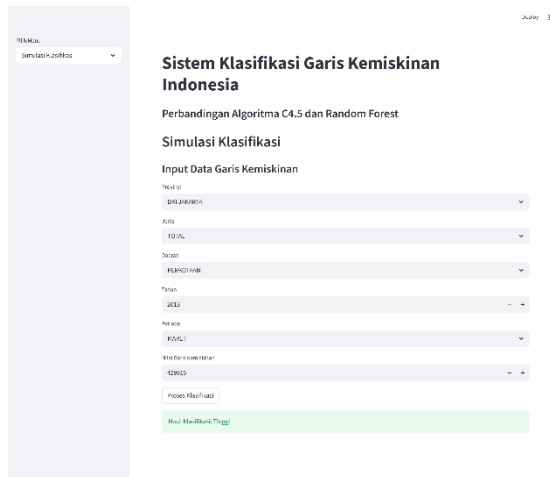
id	nama	alamat	umur	pendidikan	pekerjaan	status	jenis kelamin	pendapatan
1	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
2	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
3	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
4	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
5	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
6	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
7	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
8	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
9	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
10	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000

Hasil Evaluasi

id	nama	alamat	umur	pendidikan	pekerjaan	status	jenis kelamin	pendapatan
1	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000
2	ALDI	INDONESIA	18	SD	PEKERJA	LAJAK	LELAH	1000000

Gambar 7. Tampilan sidebar aplikasi

Halaman komparasi algoritma digunakan untuk melakukan pengujian performa algoritma C4.5 dan Random Forest. Pengguna dapat melakukan upload dataset kemudian sistem akan menampilkan dataset serta hasil evaluasi model berupa accuracy, precision, recall, dan F1-score. Berdasarkan hasil pengujian, kedua algoritma memperoleh nilai evaluasi sebesar 100%. Halaman simulasi klasifikasi ditunjukkan pada Gambar 8.



Gambar 8. Halaman simulasi klasifikasi

Halaman ini digunakan untuk melakukan klasifikasi kategori garis kemiskinan secara langsung. Pengguna memasukkan data seperti provinsi, jenis pengeluaran, daerah, tahun, periode, dan nilai garis kemiskinan. Sistem kemudian menampilkan hasil klasifikasi kategori garis kemiskinan berupa rendah, sedang, atau tinggi sesuai nilai input yang diberikan.

Pengujian Sistem

Pengujian sistem dilakukan menggunakan metode black box testing, yaitu pengujian yang dilakukan dengan menguji fungsi-fungsi utama aplikasi tanpa melihat kode program di dalam sistem. Pengujian dilakukan untuk memastikan seluruh fitur aplikasi berjalan sesuai kebutuhan pengguna, sebagaimana dirangkum pada Tabel 3.

Tabel 3. Hasil pengujian black box

No	Pengujian	Keluaran yang Diharapkan	Keterangan
1	Memilih menu Dashboard	menampilkan halaman utama aplikasi	Berhasil
2	Memilih menu Komparasi Algoritma	menampilkan halaman komparasi algoritma	Berhasil

3	Upload dataset Excel	Sistem menerima dataset garis kemiskinan	Berhasil
4	Klik tombol Proses Data	Sistem melakukan preprocessing dan klasifikasi data	Berhasil
5	Menampilkan dataset	Sistem menampilkan isi dataset yang diupload	Berhasil
6	Menampilkan hasil evaluasi	Sistem menampilkan accuracy, precision, recall, dan F1-score	Berhasil
7	Memilih menu Simulasi Klasifikasi	Sistem menampilkan halaman simulasi klasifikasi	Berhasil
8	Input provinsi	Sistem menerima input provinsi	Berhasil
9	Input jenis pengeluaran	Sistem menerima input jenis pengeluaran	Berhasil
10	Input daerah	Sistem menerima input daerah	Berhasil
11	Input tahun	Sistem menerima input tahun	Berhasil
12	Input periode	Sistem menerima input periode	Berhasil
13	Input nilai garis kemiskinan	Sistem menerima nilai garis kemiskinan	Berhasil
14	Klik tombol Proses Klasifikasi	Sistem menampilkan hasil klasifikasi kategori garis kemiskinan	Berhasil

Berdasarkan hasil pengujian pada tabel di atas, seluruh fitur aplikasi dapat berjalan

dengan baik sesuai fungsi yang diharapkan. Sistem mampu melakukan proses komparasi algoritma C4.5 dan Random Forest, menampilkan hasil evaluasi model, serta melakukan simulasi klasifikasi kategori garis kemiskinan secara interaktif. Dengan demikian, aplikasi klasifikasi garis kemiskinan berbasis web ini dapat digunakan sebagai media analisis dan klasifikasi data kemiskinan di Indonesia secara efektif.

Penelitian Terdahulu

Penelitian mengenai penerapan algoritma klasifikasi dalam bidang sosial-ekonomi telah banyak dilakukan, khususnya dalam penentuan penerima bantuan sosial dan klasifikasi tingkat kemiskinan. Salah satu penelitian dilakukan oleh Kurniawan dkk. mengenai implementasi algoritma Random Forest untuk menentukan penerima bantuan Raskin [10]. Penelitian tersebut bertujuan menentukan penerima bantuan Raskin menggunakan algoritma klasifikasi berbasis machine learning dengan dataset dari Kelurahan Gunungparang, Kota Sukabumi, dan variabel seperti jenis lantai rumah, jenis dinding rumah, pekerjaan, sumber penerangan, dan kepemilikan aset. Hasil penelitian menunjukkan bahwa algoritma Random Forest memperoleh nilai akurasi sebesar 97,26%, precision 97%, recall 97%, dan ROC sebesar 0,970 sehingga dinilai lebih baik dibandingkan algoritma SVM dan Naive Bayes.

Penelitian selanjutnya dilakukan oleh Suseno dkk. mengenai klasifikasi kemiskinan kabupaten dan kota di Indonesia berdasarkan indikator sosial-ekonomi menggunakan algoritma C4.5 [11]. Penelitian ini menggunakan indikator seperti pendidikan, pengangguran, pengeluaran per kapita, dan indeks pembangunan manusia dengan dataset mencakup 514 data kabupaten/kota di Indonesia. Hasil penelitian menunjukkan bahwa algoritma C4.5 mampu menghasilkan pola pohon keputusan yang jelas dan mudah dipahami, dengan variabel pengeluaran per kapita menjadi faktor dominan dalam menentukan status kemiskinan suatu wilayah.

Penelitian lain dilakukan oleh Ernawati dkk. mengenai penerapan data mining untuk klasifikasi penduduk miskin di Kabupaten Labuhanbatu menggunakan Random Forest dan K-Nearest Neighbors [12]. Dataset yang digunakan terdiri dari 21.137 data penduduk dengan atribut pekerjaan, pendapatan, pendidikan, dan tempat tinggal. Berdasarkan hasil pengujian menggunakan metrik akurasi, precision, recall, F1-score, dan AUC, algoritma Random Forest memperoleh performa yang lebih baik dibandingkan KNN dengan nilai akurasi sebesar 0,6023 dan AUC sebesar 0,5681.

Penelitian berikutnya dilakukan oleh Ramdani dkk. mengenai algoritma klasifikasi data mining untuk memprediksi masyarakat dalam menerima bantuan sosial [13]. Penelitian tersebut menerapkan algoritma C4.5 di Kecamatan Binong, Kabupaten Subang, dengan variabel penghasilan, pekerjaan, kepemilikan rumah, dan status kelayakan penerima bantuan sosial. Hasil pengujian menggunakan metode Cross Validation dan Confusion Matrix menunjukkan bahwa algoritma C4.5 memperoleh nilai accuracy, precision, dan recall sebesar 100%, membuktikan bahwa algoritma ini sangat efektif digunakan dalam proses klasifikasi bantuan sosial.

Data Mining

Data mining merupakan proses penggalian informasi, pola, dan pengetahuan dari sekumpulan data berukuran besar dengan menggunakan teknik statistik, matematika, dan kecerdasan buatan. Proses ini bertujuan menemukan hubungan atau pola tersembunyi yang dapat digunakan sebagai dasar pengambilan keputusan. Dalam perkembangannya, data mining banyak diterapkan pada berbagai bidang, termasuk ekonomi, kesehatan, pendidikan, dan sosial [14]. Dalam penelitian ini, data mining digunakan untuk menganalisis data garis kemiskinan di Indonesia berdasarkan beberapa variabel sosial-ekonomi. Melalui teknik data mining, data yang sebelumnya hanya berupa kumpulan angka dapat diolah menjadi

informasi yang berguna untuk mengklasifikasikan kategori garis kemiskinan. Proses data mining dalam penelitian ini meliputi tahapan pengumpulan data, pembersihan data, transformasi data, pemodelan, serta evaluasi hasil klasifikasi.

Klasifikasi

Klasifikasi merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam kelas atau kategori tertentu berdasarkan karakteristik yang dimiliki. Teknik klasifikasi bekerja dengan membangun model dari data pelatihan (training data) yang telah memiliki label kelas, kemudian model tersebut digunakan untuk memprediksi kelas pada data baru [15]. Klasifikasi digunakan untuk mengelompokkan data garis kemiskinan ke dalam kategori tertentu, seperti kategori rendah, menengah, dan tinggi. Proses klasifikasi dilakukan menggunakan algoritma C4.5 dan Random Forest untuk mengetahui algoritma mana yang memiliki performa terbaik dalam mengolah data kemiskinan di Indonesia.

Garis Kemiskinan

Garis kemiskinan merupakan ukuran yang digunakan untuk menentukan batas minimum pengeluaran seseorang dalam memenuhi kebutuhan dasar hidup. Badan Pusat Statistik (BPS) mendefinisikan garis kemiskinan sebagai jumlah pengeluaran minimum per kapita per bulan untuk memenuhi kebutuhan makanan dan non-makanan [16]. Garis kemiskinan menjadi indikator penting dalam mengukur tingkat kesejahteraan masyarakat. Semakin tinggi nilai garis kemiskinan suatu wilayah, maka semakin besar pula kebutuhan minimum yang harus dipenuhi masyarakat di wilayah tersebut. Dalam penelitian ini, data garis kemiskinan digunakan sebagai objek utama analisis untuk diklasifikasikan berdasarkan kategori tertentu menggunakan metode machine learning [17].

Algoritma C4.5

Algoritma C4.5 merupakan metode klasifikasi berbasis decision tree yang

dikembangkan oleh J. Ross Quinlan. Algoritma ini membentuk pohon keputusan berdasarkan nilai entropy dan information gain untuk menentukan atribut terbaik sebagai node [18]. Nilai entropy digunakan untuk mengukur tingkat ketidakpastian data, sebagaimana ditunjukkan pada Gambar 9.

$$Entropy(S) = - \sum_{i=1}^n p_i \log_2 p_i$$

Gambar 9. Rumus perhitungan entropy

Sedangkan nilai information gain dihitung menggunakan rumus pada Gambar 10.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i)$$

Gambar 10. Rumus perhitungan information gain

Atribut dengan nilai gain tertinggi akan dipilih sebagai akar pohon keputusan. Algoritma C4.5 memiliki keunggulan dalam menangani data numerik dan kategorikal serta menghasilkan aturan klasifikasi yang mudah dipahami. Pada penelitian ini, algoritma C4.5 digunakan untuk mengklasifikasikan kategori garis kemiskinan berdasarkan variabel sosial-ekonomi [19].

Random Forest

Random Forest merupakan algoritma ensemble learning yang membangun banyak decision tree secara acak untuk meningkatkan akurasi klasifikasi dan mengurangi overfitting. Metode ini menggunakan teknik bootstrap sampling dan pemilihan atribut secara acak pada setiap pohon [20]. Hasil klasifikasi Random Forest diperoleh melalui metode majority voting dengan rumus pada Gambar 11.

$$H(x) = \text{mode}\{h_1(x), h_2(x), \dots, h_k(x)\}$$

Gambar 11. Rumus majority voting Random Forest

Keterangan: $H(x)$ adalah hasil klasifikasi akhir, $h_i(x)$ adalah prediksi pohon ke- i , dan k

adalah jumlah pohon keputusan. Random Forest memiliki keunggulan dalam menangani data berukuran besar dan menghasilkan performa klasifikasi yang stabil. Dalam penelitian ini, algoritma Random Forest digunakan untuk mengklasifikasikan kategori garis kemiskinan dan dibandingkan dengan algoritma C4.5 berdasarkan nilai akurasi, precision, recall, dan F1-score.

CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) merupakan metodologi standar yang digunakan dalam proses data mining. Metodologi ini terdiri dari enam tahapan utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment [21]. Pada penelitian ini, CRISP-DM digunakan sebagai kerangka kerja dalam proses analisis data kemiskinan. Tahapan tersebut dimulai dari memahami tujuan penelitian, mempersiapkan dataset garis kemiskinan, membangun model klasifikasi menggunakan algoritma C4.5 dan Random Forest, hingga melakukan evaluasi hasil klasifikasi untuk menentukan algoritma terbaik.

Python

Python merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan dalam bidang data science, machine learning, dan pengembangan aplikasi. Python memiliki sintaks yang sederhana serta didukung oleh berbagai pustaka seperti Pandas, NumPy, Scikit-learn, dan Matplotlib yang memudahkan proses analisis data [22]. Dalam penelitian ini, Python digunakan untuk proses pengolahan data, implementasi algoritma C4.5 dan Random Forest, serta evaluasi hasil klasifikasi. Selain itu, Python juga digunakan untuk membangun aplikasi berbasis web menggunakan framework Streamlit.

Streamlit

Streamlit merupakan framework berbasis Python yang digunakan untuk membangun aplikasi web interaktif secara sederhana dan

cepat. Streamlit banyak digunakan dalam pengembangan aplikasi machine learning dan data science karena mampu menampilkan visualisasi data dan hasil analisis secara interaktif [23]. Pada penelitian ini, Streamlit digunakan untuk membangun aplikasi klasifikasi garis kemiskinan. Aplikasi tersebut berfungsi untuk menampilkan hasil pengolahan data, visualisasi performa algoritma, serta hasil perbandingan klasifikasi antara algoritma C4.5 dan Random Forest.

Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, algoritma C4.5 dan Random Forest mampu digunakan dengan baik dalam proses klasifikasi kategori garis kemiskinan di Indonesia menggunakan dataset garis kemiskinan periode 2013–2022. Proses penelitian dilakukan menggunakan metode CRISP-DM yang meliputi tahapan pemahaman data, persiapan data, pemodelan, dan evaluasi model. Berdasarkan hasil pengujian, kedua algoritma memperoleh nilai accuracy, precision, recall, dan F1-score sebesar 100%, sehingga menunjukkan bahwa seluruh data berhasil diklasifikasikan dengan baik. Selain itu, aplikasi berbasis web yang dibangun menggunakan Python dan Streamlit mampu menjalankan proses komparasi algoritma serta simulasi klasifikasi kategori garis kemiskinan secara interaktif dan sesuai kebutuhan pengguna. Untuk penelitian selanjutnya, disarankan menggunakan dataset yang lebih besar dan kompleks serta menambahkan algoritma klasifikasi lain agar hasil perbandingan performa model menjadi lebih optimal dan bervariasi.

Ucapan terima kasih

Penulis mengucapkan terima kasih kepada Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Ibrahimy atas dukungan yang diberikan selama proses penelitian ini. Dan juga kami ucapkan Terima kasih kepada pihak penyelenggara seminar nasional yaitu Universitas Muhammadiyah Muara Bungo.

Daftar Pustaka

- [1] Akib, F. H. Y., Novriansyah, A., dan Wolok, E. 2025. "Dinamika Kemiskinan di Provinsi Gorontalo: Tren, Tantangan, dan Strategi Penanggulangan 2012–2022." *Jurnal Ilmiah Manajemen dan Bisnis* 7(3):1203–8.
- [2] Salsabila, N., Muna, N., Pradana, V. H., dan Nurcahya, W. F. 2024. "Analisis Efektivitas Bantuan Sosial (Bansos) dalam Mengatasi Kemiskinan di Indonesia." *Journal of Macroeconomics and Social Development* 1(4):1–13. doi: 10.47134/jmsd.v1i4.317.
- [3] Samsul Hadi, Taufiki, A., Ahmad, M. Z., dan Siswanto, A. 2024. "Analisis Tingkat Kemiskinan Tahun 2023 di Badan Pusat Statistik (BPS) Kabupaten Jember." *Journal of Indonesian Social Society (JISS)* 2(1):32–39. doi: 10.59435/jiss.v2i1.224.
- [4] Nugraha Rahmansyah dan Lusinia, S. A. 2022. "Analisa Sistem Pendukung Keputusan Kemiskinan Menurut Kabupaten Kota Provinsi Sumatera Barat." *Jurnal Karya Ilmiah Multidisiplin (JURKIM)* 2(1):76–82. doi: 10.31849/jurkim.v2i1.9195.
- [5] Hardianto, H., Kunang, Y. N., Negara, E. S., dan Sutabri, T. 2025. "Model Data Mining dalam Menganalisis Faktor-Faktor yang Mempengaruhi Kemiskinan di Sumatera Selatan." *JATI (Jurnal Mahasiswa Teknik Informatika)* 9(2):2195–2202. doi: 10.36040/jati.v9i2.12989.
- [6] Wijaya, A. P., Penulis, dan Diajukan, K. 2025. "Perbandingan Algoritma Klasifikasi Random Forest dengan Naïve Bayes Classifier pada Studi Penyakit Berdasarkan Pola Nutrisi." *REMIK: Riset dan E-Jurnal Manajemen Informatika Komputer* 9(1):429–38.
- [7] Susanto, B. E., Putra, A., Ridwan, Ginting, S., dan Amin, M. 2026. "Penerapan Algoritma Decision Tree (C4.5) dalam Menentukan Kelayakan Penerima Bantuan Sosial." *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)* 4(3):1955–61. doi: 10.62712/juktisi.v4i3.793.
- [8] Sastya, N. C., dan Nugraha, I. 2023. "Penerapan Metode CRISP-DM dalam Menganalisis Data untuk Menentukan Customer Behavior di MeatSolution." *Unistek* 10(2):103–15. doi: 10.33592/unistek.v10i2.3079.
- [9] Faisal Halim, M., Maulana, M. R., Darmawan, A. S., Arochman, A., dan Ilyas, A. 2023. "Sistem Rekomendasi Pemilihan Kerja Siswa SMK dengan Menggunakan Algoritma Decision Tree." *Smart Comp: Jurnalnya Orang Pintar Komputer* 12(3):825–33. doi: 10.30591/smartcomp.v12i3.4957.
- [10] Kurniawan, I., Buani, D. C. P., Apriliah, W., dan Saputra, R. A. 2023. "Implementasi Algoritma Random Forest untuk Menentukan Penerima Bantuan Raskin." *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)* 10(2):421–28. doi: 10.25126/jtiik.202396225.
- [11] Suseno, L. A., Hutapea, T. B., dan Kurniawan, A. 2025. "Classification of Poverty Levels in Indonesian Regencies and Cities Based on Socio Economic Indicators Using the C4.5 Algorithm." *Journal of Big Data Analytic and Artificial Intelligence* 8(2):48–53.
- [12] Ernawati, A., Sitorus, Z., Iqbal, M., dan Nasution, D. 2025. "Penerapan Data Mining untuk Klasifikasi Penduduk Miskin di Kabupaten Labuhanbatu Menggunakan Random Forest dan K-Nearest Neighbors." *Bulletin of Information Technology (BIT)* 6(2):23–35. doi: 10.47065/bit.v5i2.1783.
- [13] Ramdani, A., Sofyan, C. D., Ramdani, F., Tama, M. F. A., dan Rachmatsyah, M. A. 2022. "Algoritma Klasifikasi Data Mining untuk Memprediksi Masyarakat dalam Menerima Bantuan Sosial." *Jurnal Ilmiah Sistem Informasi (JUSI)* 1(2):39–47.
- [14] Agus Nur K. 2020. "Teknik Data Mining: Algoritma K-Means Clustering." *Ilmu Komputer*, 1–12.
- [15] Pebdika, A., Herdiana, R., dan Solihudin, D. 2023. "Klasifikasi Menggunakan Metode Naive Bayes untuk Menentukan Calon Penerima PIP." *JATI (Jurnal Mahasiswa Teknik Informatika)* 7(1):452–58. doi: 10.36040/jati.v7i1.6303.
- [16] Marbun, D., dan Suryahadi, A. 2025. "Kriteria Kemiskinan Konsumsi: Praktik di Indonesia dan Beberapa Catatan." *Jurnal Analisis Sosial*, 13–30.
- [17] Mulia, R. A., dan Putri, R. P. 2022. "Pengaruh Tingkat Kemiskinan dan Produk Domestik Regional Bruto Terhadap Kesejahteraan Masyarakat." *Jiee: Jurnal Ilmiah Ekotrans & Erudisi* 2(1):22–33.
- [18] Avianto, D., dan Wibowo, A. P. 2024. "Pembentukan Pohon Keputusan untuk

-
- Penerima Bantuan Beras Miskin Menggunakan Algoritma Decision Tree C4.5." *Networking Engineering Research Operation* 9(1):59–68. doi: 10.21107/nero.v9i1.28020.
- [19] Muhammad Hamdan Sukri, dan Handrianto, Y. 2024. "Penerapan Algoritma C4.5 dalam Menentukan Prediksi Prestasi Siswa pada SMPN 51 Jakarta." *Informatics and Computer Engineering Journal* 4(1):11–24. doi: 10.31294/icej.v4i1.2582.
- [20] Putra Argadinata, A., Abdul Fatah, D., dan Sukri, H. 2025. "Klasifikasi Kualitas Buah Apel Menggunakan Metode Random Forest." *JATI (Jurnal Mahasiswa Teknik Informatika)* 9(2):2016–22. doi: 10.36040/jati.v9i2.12854.
- [21] Ruswanti, D., Susilo, D., dan Riani, R. 2024. "Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C4.5." *Go Infotech: Jurnal Ilmiah STMIK AUB* 30(1):111–21. doi: 10.36309/goi.v30i1.266.
- [22] Maulana, dan Sulthan Thaha Saifuddin. 2024. "Python Bahasa Pemrograman yang Ramah bagi Pemula." *JISCO: Journal of Information System and Computing* 2(2):73–78.
- [23] Th, A. T. 2023. Implementasi Algoritma Machine Learning Berbasis Web dengan Framework Streamlit. Probolinggo: Universitas Ibrahimy.