JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 84~92
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Pendekatan *Data-Driven* untuk Pengembangan Model Prediksi Tingkat Kemiskinan di Provinsi Indonesia

[1]Evi Purnamasari, [2]Dwi Asa Verano
[1,2]Universitas Indo Global Mandiri, Indonesia
[1]evi.ps@uigm.ac.id; [2]dwiasa@uigma.c.id;

Article Info

Article history:

Received, 2024-12-31
Revised, 2025-01-02
Accepted, 2025-01-07

Kata Kunci:

Tingkat Kemiskinan
Data-Driven
K-Means
Naive Bayes
Cross-Validation
Akurasi

ABSTRAK

Tingkat kemiskinan di Indonesia masih menjadi masalah utama yang memerlukan perhatian serius, terutama di tingkat provinsi. Berbagai faktor, seperti akses terhadap pendidikan, kesehatan, dan lapangan pekerjaan, mempengaruhi tingkat kemiskinan tersebut. Penelitian ini bertujuan untuk mengembangkan model prediksi tingkat kemiskinan dengan pendekatan data-driven menggunakan analisis klaster dan klasifikasi. Metode yang digunakan dalam klasterisasi adalah K-Means, *Hierarchical Clustering*, dan DBSCAN, sedangkan untuk klasifikasi diterapkan algoritma *Random Forest*, *Naive Bayes*, dan *Support Vector Machine* (SVM). Hasil analisis klasterisasi menunjukkan bahwa K-Means menghasilkan pembagian klaster yang lebih jelas dengan *Calinski-Harabasz Index* tertinggi (nilai 179,45). Pada pengujian model klasifikasi, *Naive Bayes* memberikan hasil terbaik dengan akurasi 99,42%, yang lebih tinggi dibandingkan model lainnya. Penelitian ini berhasil mengidentifikasi faktor-faktor yang mempengaruhi tingkat kemiskinan di provinsi-provinsi Indonesia, yang dapat digunakan sebagai dasar bagi kebijakan pemerintah dalam upaya pengentasan kemiskinan Untuk mengatasi masalah *overfitting*, dilakukan pengujian menggunakan *cross-validation* dengan nilai *Mean Accuracy* sebesar 99,32% dan *Standard Deviation* 0,23%. Hasil yang dicapai memberikan kontribusi signifikan terhadap pengembangan model prediksi yang lebih akurat dan efektif dalam menangani masalah kemiskinan di Indonesia.

Keywords:

Poverty Level
Data-Driven
K-Means
Naive Bayes
Cross-Validation
Accuracy

ABSTRACT

Poverty in Indonesia remains a major issue that requires serious attention, particularly at the provincial level. Various factors, such as access to education, healthcare, and employment opportunities, affect the poverty rate. This study aims to develop a poverty prediction model using a data-driven approach through cluster analysis and classification. The methods used in clustering are K-Means, Hierarchical Clustering, and DBSCAN, while for classification, the algorithms applied are Random Forest, Naive Bayes, and Support Vector Machine (SVM). The clustering analysis results show that K-Means provides clearer cluster divisions with the highest Calinski-Harabasz Index value (179.45). In classification model testing, Naive Bayes provides the best results with an accuracy of 99.42%, which is higher than the other models. To address overfitting, cross-validation testing was conducted, yielding a Mean Accuracy of 99.32% and a Standard Deviatin of 0.23%. This study successfully identifies the factors influencing poverty levels in Indonesia's provinces, which can be used as a basis for government policies in poverty alleviation efforts. The results achieved contribute significantly to the development of a more accurate and effective predictive model for addressing poverty issues in Indonesia.

Penulis Korespondensi:

Evi Purnamasari,
Program Teknik Informatika,
Universitas Indo Global Mandiri
Email: evi.ps@uigm.ac.id

1. PENDAHULUAN

Tingkat kemiskinan merupakan salah satu indikator utama dalam mengukur kesejahteraan sosial-ekonomi suatu wilayah. Di Indonesia, kemiskinan masih menjadi masalah yang memerlukan perhatian serius, terutama di tingkat provinsi. Berbagai faktor, seperti akses terhadap pendidikan, kesehatan, lapangan pekerjaan, dan infrastruktur, mempengaruhi tingkat kemiskinan di suatu daerah. Oleh karena itu, penting untuk mengembangkan model yang dapat memprediksi tingkat kemiskinan secara akurat, sehingga kebijakan yang lebih efektif dapat diterapkan untuk mengatasi permasalahan ini. Pendekatan data-driven, yang memanfaatkan analisis data besar dan teknik machine learning, menawarkan potensi besar dalam menciptakan model prediksi tingkat kemiskinan yang lebih tepat.

Salah satu tantangan utama dalam memprediksi tingkat kemiskinan adalah kompleksitas data yang digunakan. Setiap provinsi di Indonesia memiliki karakteristik sosial, ekonomi, dan geografis yang berbeda, yang menjadikan analisis data kemiskinan cukup menantang. Selain itu, data yang tidak lengkap, seperti nilai yang hilang (*missing values*) dan adanya *outlier*, dapat mempengaruhi hasil prediksi secara signifikan. Untuk itu, tahap praproses data menjadi penting, di mana penanganan missing value dan outlier perlu dilakukan untuk meningkatkan kualitas data yang akan digunakan dalam model.

Penelitian ini menggunakan pendekatan data-driven yang menggabungkan teknik klasifikasi dan klasterisasi untuk memprediksi tingkat kemiskinan di provinsi-provinsi Indonesia [1]. Model klasterisasi digunakan untuk mengelompokkan provinsi berdasarkan kesamaan faktor-faktor yang mempengaruhi tingkat kemiskinan, sementara model klasifikasi digunakan untuk memprediksi tingkat kemiskinan di masing-masing provinsi. Tujuan utama dari penelitian ini adalah untuk mengembangkan model prediksi yang dapat memberikan gambaran yang akurat tentang tingkat kemiskinan di setiap provinsi, serta menjadi dasar bagi kebijakan pemerintah dalam upaya pengentasan kemiskinan. Selain itu, penelitian ini juga bertujuan untuk mengidentifikasi faktor-faktor utama yang mempengaruhi tingkat kemiskinan di Indonesia melalui analisis data.

Penelitian terdahulu yang telah dilakukan untuk memprediksi tingkat kemiskinan dengan berbagai metode, yaitu J48 *Decision Tree* [2], Metode Arima [3], algoritma *Back Propagation* [4] , *Support Vector Machine* (SVM) dengan *Regresi Linear* [5], *Single Moving Average* dan *Double Moving Average* [6], *Moving Average* [7]. Beberapa Penelitan sebelumnya lebih berfokus pada penerapan satu metode untuk analisis prediksi kemiskinan. Namun, pada penelitian ini mengisi gap dengan menggabungkan beberapa pendekatan, yaitu *clusterisasi* untuk mengelompokkan provinsi berdasarkan karakteristik kemiskinan, dan klasifikasi untuk meningkatkan akurasi prediksi. Selain itu, penelitian ini menerapkan teknik data *preprocessing* dan *exploratory data analysis* (EDA) yang lebih mendalam untuk menangani variabilitas data, serta menggunakan model yang lebih terintegrasi untuk menghasilkan prediksi yang lebih akurat dan holistik dibandingkan dengan metode-metode yang digunakan dalam penelitian sebelumnya.

Metode yang digunakan dalam penelitian ini mencakup analisis klasterisasi dan klasifikasi [8]. Pada analisis klasterisasi, tiga metode diuji, yaitu *K-Means*, *Hierarchical Clustering*, dan DBSCAN, untuk mengelompokkan provinsi berdasarkan karakteristik sosial-ekonomi dan tingkat kemiskinan [9]. Metrik evaluasi yang digunakan untuk mengukur kinerja klasterisasi meliputi *Silhouette Coefficient*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index* [10]. Sementara itu, untuk model klasifikasi, beberapa algoritma diuji, termasuk *Random Forest*, *Naive Bayes*, dan *Support Vector Machine* (SVM) [11]. Semua model klasifikasi dilatih dengan data yang telah diproses sebelumnya, menggunakan teknik imputasi untuk menangani nilai yang hilang dan teknik *Winsorizing* untuk menangani *outlier*. Pengujian model dilakukan dengan teknik *cross-validation* untuk mengatasi masalah *overfitting*, dan evaluasi performa dilakukan berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score*.

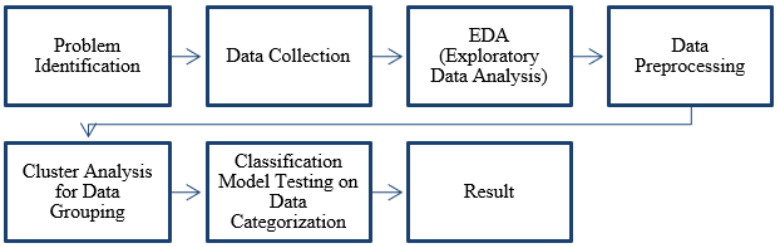
Penelitian ini bertujuan menghasilkan model klasterisasi yang mampu mengelompokkan provinsi-provinsi di Indonesia secara akurat berdasarkan karakteristik kemiskinan yang dimiliki serta memprediksi tingkat kemiskinan di setiap provinsi dengan tingkat akurasi yang tinggi. Dengan pemahaman yang lebih mendalam mengenai faktor-faktor yang memengaruhi kemiskinan, penelitian ini dapat mendukung penerapan kebijakan yang lebih tepat sasaran untuk mengurangi kemiskinan di berbagai wilayah.

Hasil penelitian ini diharapkan memberikan kontribusi signifikan dalam upaya pengentasan kemiskinan di Indonesia melalui pendekatan berbasis data yang lebih akurat dan efektif. Selain itu, penelitian ini dapat membantu pemerintah dalam pengalokasian sumber daya serta perencanaan program-program yang lebih terarah untuk mengatasi permasalahan kemiskinan secara menyeluruh.

2. METODE PENELITIAN

Pada sub bab ini akan dijelaskan metode yang digunakan dalam penelitian, yang mencakup beberapa tahapan penting. Dimulai dengan *Problem Identification* untuk mengidentifikasi masalah dan tujuan penelitian, dilanjutkan dengan *Data Collection* untuk mengumpulkan data yang relevan. Selanjutnya, dilakukan *Exploratory Data Analysis* (EDA) untuk memahami pola dan hubungan dalam data, diikuti dengan *Data*

Preprocessing untuk mempersiapkan data agar siap dianalisis. Setelah itu, *Cluster Analysis for Data Grouping* dilakukan untuk mengelompokkan data, kemudian diikuti dengan *Classification Model Testing on Data Categorization* untuk menguji akurasi model klasifikasi [12]. Akhirnya, hasil penelitian disajikan dalam tahap Result, yang merangkum temuan dan interpretasi dari analisis yang dilakukan**.** Gambar 1 merupakan diagram yang menggambarkan tahapan-tahapan dalam penelitian ini.



Gambar 1. Tahapan penelitian

Pengambilan data (*Data Collection*)

Tahap ini merupakan langkah penting untuk memperoleh informasi yang valid dan relevan guna mendukung analisis yang dilakukan. Data yang digunakan dalam penelitian ini diperoleh Badan Pusat Statistik (BPS) kota Palembang. Dataset tersebut berasal dari data kemiskinan di provinsi-provinsi di Indonesia pada tahun 2023, dengan total 514. Data tersebut terdiri dari nama provinsi, nama kota, Persentase Penduduk Miskin, Rata-rata Lama Sekolah, Pengeluaran per Kapita, Indeks Pembangunan Manusia, Umur Harapan Hidup, Persentas Rumah tangga memiliki sanitasi layak, Persentase Rumah tangga memiliki akses terhadap air minum layak, Tingkat Pengangguran, Tingkat Partisipasi Angkatan Kerja, Produk Domestik Regional Bruto (PDRB).

EDA (*Exploratory Data Analysis*)

Pada tahapan *Exploratory Data Analysis* (EDA), dilakukan peninjauan mendalam terhadap data untuk memahami karakteristiknya, serta mengidentifikasi adanya nilai yang hilang (*missing value*) dan pencilan (*outlier*) [13]. Tahapan ini mencakup analisis statistik deskriptif, visualisasi data menggunakan *boxplot* dan *scatterplot* untuk melihat distribusi data, serta analisis korelasi untuk mengukur hubungan antar fitur. Hasil dari EDA memberikan wawasan awal yang penting untuk mendukung proses pemodelan selanjutnya, termasuk identifikasi fitur-fitur yang paling berpengaruh terhadap tingkat kemiskinan di daerah tertentu [14].
Dataset yang digunakan diberi inisial untuk memudahkan dalam proses pengolahan data, yaitu Persentase Penduduk Miskin (X1), Rata-rata Lama Sekolah (X2), Pengeluaran per Kapita (X3), Indeks Pembangunan Manusia (X4), Umur Harapan Hidup (X5), Persentase Rumah Tangga Memiliki Sanitasi Layak (X6), Persentase Rumah Tangga Memiliki Akses terhadap Air Minum Layak (X7), Tingkat Pengangguran (X8), Tingkat Partisipasi Angkatan Kerja (X9), dan PDRB (X10).
Untuk memberikan wawasan yang lebih komprehensif mengenai karakteristik dataset, dilakukan analisis statistik deskriptif. Analisis ini bertujuan untuk memberikan gambaran mendalam tentang karakteristik data yang telah dikumpulkan, tanpa melakukan generalisasi terhadap populasi yang lebih luas. Tabel 1 menunjukkan hasil statistik deskriptif dari data kemiskinan di provinsi-provinsi di Indonesia pada tahun 2023.

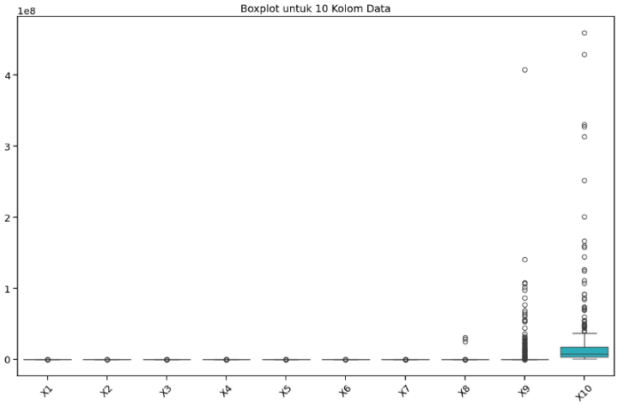
Tabel 1. Hasil Statistik Deskriptif Data Kemiskinan Di Provinsi Indonesia tahun 2023

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
count	514	514	514	514	514	513	512	512	510	425
mean	12.32	8.99	8500.61	58.50	71.54	78.60	71.03	164897.8	4900576	20472050
std	7.47	5.66	4456.08	26.16	8.81	19.41	32.53	2159458	23317910	47476470
min	2.38	1.42	63.00	0.12	12.11	0.26	1.24	0.41	56.39	147485
25%	7.15	7.56	7153.00	63.93	67.61	71.34	64.21	3.465	65.5625	3474011
50%	10.50	8.36	9434.50	68.55	70.34	83.39	85.19	5.15	69.84	7709700
75%	14.88	9.47	11296.00	72.06	73.40	92.25	94.24	9.215	76.57	17500610
max	41.66	72.49	23888.00	87.18	99.97	100.00	100.00	31176540	407726800	460081000

Tabel 1 menunjukkan hasil statistik deskriptif untuk berbagai indikator terkait kemiskinan di provinsi-provinsi Indonesia pada tahun 2023. Secara umum, data mencakup berbagai fitur yang menggambarkan kondisi sosial dan ekonomi.
Hasil statistik deskriptif menunjukkan adanya variasi yang cukup besar antar provinsi. Data juga menunjukkan adanya beberapa nilai yang hilang (*missing value*) pada beberapa fitur, yaitu kolom Persentase RT Memiliki Sanitasi Layak, Persentase RT Memiliki Akses terhadap Air Minum Layak, Tingkat Pengangguran, Tingkat Partisipasi Angkatan Kerja, dan PDRB. Penanganan *missing value* yang tepat sangat penting untuk menjaga kualitas data dan keakuratan model analisis yang dihasilkan. Namun, untuk mengetahui adanya *outlier*, maka analisis lebih lanjut diperlukan dengan menggunakan metode visualisasi menggunakan boxplot. Gambar 2 menunjukkan distribusi dari masing-masing fitur, di mana outlier dapat diidentifikasi

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 84~92
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

dengan jelas sebagai titik-titik yang terletak di luar batas jangkauan interkuartil (IQR).



Gambar 2. Visualisasi Outlier dari data kemiskinan di Indonesia

Secara umum, analisis boxplot menunjukkan bahwa sebagian besar fitur memiliki variasi yang signifikan, terutama pada X8 (Tingkat Pengangguran), X9 (Tingkat Partisipasi Angkatan Kerja), dan X10 (PDRB), yang menunjukkan adanya nilai ekstrem dan outlier yang besar. Distribusi fitur-fitur ini cenderung tidak simetris, dengan kecenderungan nilai yang sangat tinggi. Di sisi lain, fitur seperti X1 (Rata-rata Lama Sekolah) menunjukkan variasi rendah dan tidak memiliki outlier signifikan. Beberapa fitur lain seperti X2 (Pengeluaran per Kapita) juga menunjukkan variasi yang cukup tinggi dengan beberapa outlier pada nilai tinggi. Fitur X3 (Indeks Pembangunan Manusia), X4 (Umur Harapan Hidup), X5 (Persentase Rumah Tangga Memiliki Sanitasi Layak), dan X6 (Persentase Rumah Tangga Memiliki Akses terhadap Air Minum Layak) menunjukkan variasi rendah dengan beberapa outlier pada nilai rendah. Keseluruhan, data ini menunjukkan adanya ketimpangan yang cukup besar antar wilayah terkait tingkat pengangguran, partisipasi angkatan kerja, dan PDRB.

3. HASIL DAN PEMBAHASAN

Data Preprocessing

Setelah melakukan tahap *Exploratory Data Analysis* (EDA), langkah selanjutnya adalah tahap data processing. Pada tahap ini, data yang telah dianalisis akan diproses lebih lanjut. Proses ini mencakup penanganan nilai yang hilang dan penyesuaian terhadap outlier yang terdeteksi. Dengan demikian, tahap *data processing* memastikan bahwa data yang digunakan memiliki kualitas yang tinggi dan siap untuk diolah lebih lanjut dalam penelitian atau pemodelan.

Penanganan *Missing Value*

Fitur-fitur yang mengandung *missing value* adalah Persentase RT Memiliki Sanitasi Layak (1 *missing value*), Persentase RT Memiliki Akses terhadap Air Minum Layak (2 *missing values*), Tingkat Pengangguran (2 *missing values*), dan Tingkat Partisipasi Angkatan Kerja (4 *missing values*). Sementara itu, PDRB memiliki jumlah missing value yang cukup besar, yaitu 89. Penanganan *missing value* untuk kolom Persentase RT Memiliki Sanitasi Layak, Persentase RT Memiliki Akses terhadap Air Minum Layak, Tingkat Pengangguran, dan Tingkat Partisipasi Angkatan Kerja dilakukan dengan mengisi nilai yang hilang menggunakan nilai minimum dari masing-masing kolom. Sedangkan untuk kolom PDRB, penanganan *missing value* dilakukan dengan menggunakan metode *KNN Imputation*. Berikut adalah kode Python yang digunakan:

```
# ------------------------------------------------
# Handle missing values for minimum values
# ------------------------------------------------
columns_to_fill_min = [
    'Persentas RT Memiliki sanitasi layak',
    'Persentase RT memiliki akses terhadap air minum layak',
    'Tingkat Partisipasi Angkatan Kerja']
for col in columns_to_fill_min:
    min_value = data_selected[col].min()
    data_selected[col] = data_selected[col].fillna(min_value)
# ------------------------------------------------------------------------
# Handling missing values in the PDRB column using KNN Imputation
# ------------------------------------------------------------------------
knn_imputer = KNNImputer(n_neighbors=5)
data_selected['PDRB'] = knn_imputer.fit_transform(data_selected[['PDRB']])
```

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 84~92
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

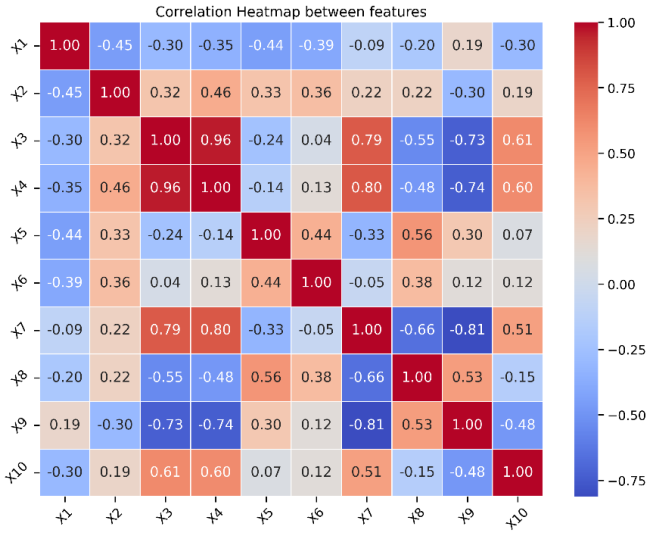
Penangan *Outlier*

Data tingkat kemiskinan di provinsi-provinsi Indonesia menunjukkan adanya outlier pada Sebagian besar fitur. Penanganan terhadap *outlier* dan distribusi data yang tidak simetris menjadi langkah penting dalam proses analisis data lebih lanjut [15]. Teknik *Winsorizing* digunakan untuk menangani *outlier* dengan cara menyesuaikan nilai-nilai ekstrem ke dalam persentil tertentu, tanpa menghilangkan data asli. Pendekatan ini bertujuan untuk mempertahankan sebanyak mungkin data asli, termasuk nilai ekstrem yang masih berada dalam rentang normal, sehingga variasi dalam dataset tetap tercermin. Selain itu, penggunaan teknik *Winsorizing* juga mengurangi potensi bias akibat penghapusan data, sehingga menghasilkan kesimpulan penelitian yang lebih akurat dan reliabel. Kode pyhton yang di gunakan adalah sebagai berikut

```
# ---------------------------------------------
# Winsorization to handle outliers
# ---------------------------------------------
# Winsorizing with iteration and percentile adjustment
for _ in range(2):  # Repeat the Winsorizing process 2 times
for col in data_winsorized.columns:
lower_limit = np.percentile(data_winsorized[col], 5)
upper_limit = np.percentile(data_winsorized[col], 95)
data_winsorized[col] = np.where(data_winsorized[col] < lower_limit,lower_limit,
data_winsorized[col])
data_winsorized[col] = np.where(data_winsorized[col] > upper_limit,upper_limit,
data_winsorized[col])
```

Korelasi antar Fitur

Setelah data bersih dari *missing value* dan outlier, langkah selanjutnya adalah melakukan visualisasi data dalam bentuk heatmap untuk melihat hubungan antar fitur. Heatmap dapat memberikan gambaran yang jelas mengenai korelasi antara fitur dalam dataset, sehingga memudahkan dalam mengidentifikasi pola atau hubungan yang signifikan. Gambar 3 merupakan Heatmap Korelasi antar fitur dalam dataset.



Gambar 3. *Heatmap* Korelasi antar fitur

Heatmap korelasi digunakan untuk menggambarkan kekuatan dan arah hubungan antara variabel-variabel dalam dataset. Warna pada heatmap menunjukkan tingkat korelasi, di mana warna merah menunjukkan korelasi positif yang kuat. Dengan demikian, peningkatan nilai suatu variabel biasanya diikuti oleh peningkatan nilai variabel lain. Sebaliknya, warna biru menunjukkan korelasi negatif yang kuat, yang berarti peningkatan nilai pada satu variabel cenderung disertai penurunan nilai pada variabel lainnya. Sementara itu, warna putih atau mendekati putih mencerminkan korelasi yang sangat lemah atau tidak ada korelasi sama sekali.
Berdasarkan analisis *heatmap* yang ditampilkan, korelasi positif sangat kuat ditemukan antara variabel Pengeluaran per Kapita (X3) dan Indeks Pembangunan Manusia (X4), dengan nilai korelasi sebesar 0,96, yang menunjukkan bahwa peningkatan pada Pengeluaran per Kapita (X3) cenderung diikuti oleh peningkatan pada Indeks Pembangunan Manusia (X4). Korelasi positif cukup kuat juga terlihat antara Rata-rata Lama Sekolah (X2) dan Indeks Pembangunan Manusia (X4) dengan nilai korelasi sebesar 0,46, sedangkan korelasi positif moderat terjadi antara Pengeluaran per Kapita (X3) dan Persentase Rumah Tangga Memiliki Akses terhadap Air Minum Layak (X7) dengan nilai korelasi sebesar 0,79.

Selain itu, korelasi negatif yang sangat kuat ditemukan antara Pengeluaran per Kapita (X3) dan Tingkat Pengangguran (X8) dengan nilai korelasi sebesar -0,73, menunjukkan bahwa peningkatan pada Pengeluaran per Kapita (X3) cenderung diikuti oleh penurunan pada Tingkat Pengangguran (X8). Hubungan negatif yang serupa juga terlihat antara Indeks Pembangunan Manusia (X4) dan Tingkat Pengangguran (X8) dengan nilai korelasi sebesar -0,74. Namun, beberapa pasangan variabel, seperti Persentase Penduduk Miskin (X1) dengan Umur Harapan Hidup (X5) dan Rata-rata Lama Sekolah (X2) dengan Umur Harapan Hidup (X5), menunjukkan korelasi yang sangat lemah atau hampir tidak ada hubungan sama sekali.
Oleh karena itu, fitur-fitur yang akan digunakan untuk tahap selanjutnya adalah Rata-rata Lama Sekolah (X2), Pengeluaran per Kapita (X3), Indeks Pembangunan Manusia (X4), Persentase Rumah Tangga Memiliki Akses terhadap Air Minum Layak (X7), Tingkat Pengangguran (X8)

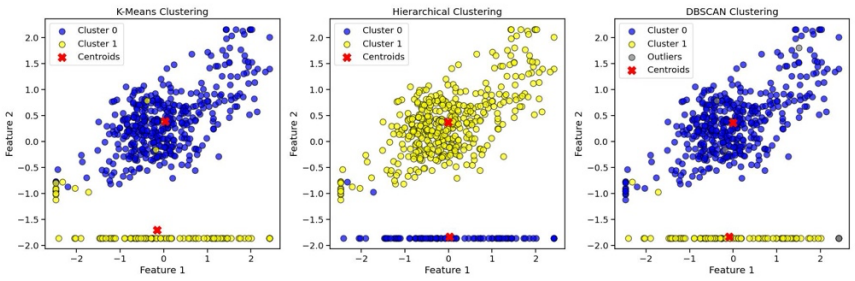
Analisis Klaster untuk Pengelompokan Data (*Cluster Analysis for Data Grouping*)

Pada tahap ini, dilakukan analisis klaster untuk mengelompokkan data berdasarkan kemiripan atau kesamaan karakteristik antar unit data. Klasterisasi merupakan teknik yang digunakan untuk memisahkan data menjadi beberapa grup atau klaster, di mana setiap grup memiliki data dengan sifat yang serupa satu sama lain [16], [17]. Tujuan utama dari analisis klaster adalah untuk mengidentifikasi pola atau struktur tersembunyi dalam data [12]. Melalui klasterisasi, dapat ditemukan segmen-segmen atau kelompok-kelompok dalam data yang memiliki kesamaan yang signifikan. Proses klasterisasi ini dilakukan dengan menggunakan berbagai algoritma, seperti *K-Means, Hierarchical Clustering*, atau DBSCAN, dengan menggunakan 5 fitur yang telah terpilih [15], [18].
Pada tahap ini, tiga metode klasterisasi digunakan untuk mengelompokkan data yang telah distandarisasi, yaitu *K-Means, Hierarchical Clustering*, dan DBSCAN. Pengukuran kinerja model dilakukan dengan menggunakan metrik evaluasi yang menghitung nilai *Silhouette Coefficient, Davies-Bouldin Index, Calinski-Harabasz Index* [19]. Hasil perhitungan metrik evaluasi model klasterisasi ditampilkan pada tabel 2.

Tabel 2. Metrik evaluasi penerapan model Klasterisasi

Model	*K-Means*	*Hierarchical*	*DBSCAN*
Silhouette Coefficient	0.6086	0.612	0.557
Davies-Bouldin Index	0.6223	0.541	1.360
Calinski-Harabasz Index	609.18	592.8	314.9

Hasil dari evaluasi model klasterisasi untuk tiga metode yang digunakan, Nilai yang lebih tinggi menunjukkan klasterisasi yang lebih baik. Berdasarkan nilai *Silhouette*, *model Hierarchical*, memperolah hasil terbaik dengan skor 0.612, sedikit lebih tinggi dibandingkan K-Means yang memperoleh skor 0.6086, dan DBSCAN yang memperoleh skor 0.557. Selanjutnya, *Davies-Bouldin Index* mengukur kualitas klaster berdasarkan jarak antar klaster dan kedekatan dalam klaster tersebut. Nilai yang lebih rendah menunjukkan kualitas klaster yang lebih baik. Model *Hierarchical* menghasilkan nilai terendah (0.541), menunjukkan bahwa klaster-klaster yang dihasilkan lebih terpisah dengan struktur yang lebih baik, dibandingkan dengan K-Means yang mendapatkan nilai 0.6223 dan DBSCAN yang memiliki nilai tertinggi 1.360.
Terakhir, *Calinski-Harabasz Index* mengukur rasio antara variasi antar klaster dan variasi dalam klaster. Nilai yang lebih tinggi menunjukkan pembagian klaster yang lebih baik. Dalam hal ini, K-Means memperoleh nilai tertinggi (609.18), menunjukkan bahwa klaster yang dihasilkan memiliki variasi antar klaster yang lebih tinggi dan lebih terpisah dibandingkan dengan Hierarchical (592.8) dan DBSCAN (314.9).
Berdasarkan kombinasi hasil dari ketiga metrik evaluasi, meskipun *Hierarchical* memberikan nilai terbaik pada dua metrik pertama, K-Means dianggap sebagai model terbaik untuk dataset ini karena menghasilkan nilai *Calinski-Harabasz Index* tertinggi, yang menunjukkan pembagian klaster yang lebih baik dan terpisah dengan jelas. Untuk melihat perbandingan antara hasil klasterisasi dengan ketiga metode tersebut, dibuatlah visualisasi untuk masing-masing model yang ditampilkan pada Gambar 4.



Gambar 4. Visualisasi perbandingan hasil klasterisasi menggunakan metode K-Means, *Hierarchical Clustering*, dan DBSCAN

Pengujian Model Klasifikasi pada Kategorisasi Data (*Classification Model Testing on Data Categorization*

Pengujian algoritma klasifikasi pada kategorisasi data bertujuan untuk mengevaluasi kemampuan algoritma dalam mengklasifikasikan data ke dalam kategori yang tepat. Dalam pengujian ini, digunakan empat algoritma klasifikasi, yaitu *Random Forest, Support Vector Machine (*SVM*), Naïve Bayes,* dan *Neural Network*, yang masing-masing memiliki karakteristik dan keunggulan dalam menangani berbagai jenis data serta masalah klasifikasi [20].
Pada pengujian ini, menggunaka matriks evaluasi yang meliputi akurasi, presisi, *recall*, dan *F1-score* untuk setiap algoritma klasifikasi. Metrik-metrik ini digunakan untuk menilai seberapa baik algoritma dalam mengklasifikasikan data ke dalam kategori yang sesuai. Akurasi mengukur persentase prediksi yang benar, presisi menunjukkan sejauh mana hasil yang diprediksi positif benar-benar positif, *recall* mengukur sejauh mana algoritma dapat mendeteksi semua data positif, dan *F1-score* memberikan gambaran keseimbangan antara presisi dan recall. Hasil dari pengujian menggunakan matriks evaluasi tersebut, diuraikan pada tabel 3 berikut.

Tabel 3. Metrik evaluasi penerapan model Klasifikasi

	Random Forest	*SVM*	*Naïve Bayes*	*Neural Network*
Accuracy	0.9903	0.9709	0.9806	0.9806
Precision	0.9907	0.9719	0.9806	0.9811
Recall	0.9903	0.9709	0.9806	0.9806
F1 Score	0.9904	0.9709	0.9806	0.9803

Berdasarkan hasil klasifikasi, model *RandomForest* menunjukkan kinerja terbaik dengan akurasi 99,03%, *precision* 99,07%, *recall* 99,03%, dan *F1 score* 99,04%. Model ini sangat efektif dalam mengklasifikasikan data dan memiliki keseimbangan yang baik antara *precision* dan *recall*. Sementara itu, *Support Vector Machine* (SVM) memberikan hasil yang sedikit lebih rendah dengan akurasi 97,09%, *precision* 97,19%, *recall* 97,09%, dan *F1 score* 97,02%. Meskipun demikian, SVM tetap menunjukkan performa yang sangat baik, meskipun tidak sebaik *RandomForest*. *Naive Bayes* menunjukkan hasil yang cukup konsisten dengan semua metrik mencapai 98,06%. *Precision*, *recall*, dan *F1 score* yang serupa menunjukkan bahwa model ini bekerja dengan baik dalam mengklasifikasikan data, meskipun sedikit lebih rendah dibandingkan RandomForest. Begitu pula dengan *Neural Network*, yang menghasilkan akurasi 98,06%, *precision* 98,11%, *recall* 98,06%, dan *F1 score* 98,03%, dengan hasil yang hampir identik dengan *Naive Bayes*.

Secara keseluruhan, *RandomForest* adalah model yang paling unggul dengan performa terbaik di semua metrik, diikuti oleh SVM, *Naive Bayes*, dan *Neural Network*, yang semuanya memberikan hasil yang solid meskipun sedikit lebih rendah. RandomForest terbukti sebagai model yang sangat baik dalam hal akurasi dan keseimbangan metrik, sementara model-model lain juga menunjukkan kinerja yang baik meskipun berada di bawah *RandomForest*.

Hasil perhitungan klasifikasi menunjukkan nilai yang hampir sempurna, yang dapat mengindikasikan adanya *overfitting*. Untuk mengantisipasi hal tersebut, pengujian dilakukan menggunakan *cross-validation* [21]. Teknik ini membagi dataset menjadi beberapa subset (*folds*), di mana model dilatih pada sebagian data dan diuji pada sisa data. Proses ini diulang untuk setiap fold, sehingga estimasi yang lebih stabil dan akurat mengenai kinerja model pada data yang tidak terlihat sebelumnya dapat diperoleh, serta mengurangi bias akibat pembagian data yang tidak representatif.

Cross-validation menghitung nilai Mean Accuracy dan Standard Deviation (Std Deviation) untuk mengukur konsistensi dan keakuratan performa model. Penerapan *cross-validation* pada seluruh model memberikan gambaran yang lebih baik mengenai performa model, mengurangi risiko *overfitting*, dan memastikan model lebih tahan terhadap variabilitas data. Dengan menggunakan *cross-validation*, model yang *optimal* dan konsisten saat menghadapi data baru dapat dipilih. Hasil dari perhitungan *Cross-validation* diuraikan pada tabel 4 berikut.

Tabel 4. Metrik evaluasi penerapan model Klasifikasi

Model	*Mean Accuracy*	*Std Deviation*
RandomForest	0.9883	0.0143
SVM	0.9922	0.0155
NaiveBayes	0.9942	0.0078
NeuralNetwork	0.9883	0.0188

Berdasarkan hasil *cross-validation* terhadap 4 model, maka dapat disimpulkan bahwa :

1. RandomForest memiliki mean accuracy sebesar 98.83% dan standard deviation 0.0143. Ini menunjukkan bahwa model ini memiliki performa yang cukup stabil di berbagai *fold cross-validation*. Meskipun akurasi relatif tinggi, variasi performa masih ada, yang menunjukkan adanya sedikit fluktuasi pada hasil yang diperoleh.

2. SVM memperoleh mean accuracy sebesar 99.22% dengan standard deviation 0.0155. Nilai akurasi ini menunjukkan performa yang sangat baik, namun variasinya sedikit lebih tinggi dibandingkan RandomForest. Ini dapat mengindikasikan bahwa meskipun model SVM sangat akurat, ada sedikit perbedaan pada hasil prediksi dari satu fold ke fold lainnya.
3. NaiveBayes mencatat mean accuracy 99.42% dengan standard deviation yang sangat rendah, yaitu 0.0078. Nilai ini menunjukkan bahwa model *Naive Bayes* tidak hanya sangat akurat, tetapi juga sangat stabil dalam klasifikasi pada setiap fold. Variasi yang rendah menunjukkan bahwa model ini memiliki performa yang konsisten.
4. NeuralNetwork memiliki mean accuracy 98.83% dengan standard deviation 0.0188. Meskipun akurasi ini sama dengan RandomForest, standard deviation yang lebih tinggi menunjukkan bahwa Neural Network lebih sensitif terhadap variasi pada dataset yang digunakan, yang bisa mengindikasikan sedikit ketidakstabilan pada model ini.

Berdasarkan hasil yang ada, tidak ditemukan indikasi jelas bahwa model-model ini mengalami *overfitting*, karena *cross-validation* memberikan gambaran tentang kemampuan model dalam menggeneralisasi ke data yang berbeda. Meskipun terdapat variansi yang lebih tinggi pada beberapa model, seperti NeuralNetwork dan SVM, tidak ada perbedaan signifikan yang mengarah pada indikasi *overfitting*.
Model Terbaik : berdasarkan *mean accuracy* dan *standard deviation*, *NaiveBayes* menunjukkan kombinasi performa terbaik dengan akurasi yang sangat tinggi (99.42%) dan variasi yang sangat rendah (0.0078). Ini menunjukkan bahwa model ini tidak hanya akurat tetapi juga sangat stabil, memberikan prediksi yang konsisten di seluruh *fold cross-validation*. Oleh karena itu, berdasarkan hasil ini, *NaiveBayes* bisa dianggap sebagai model yang paling optimal untuk digunakan pada dataset ini.

4. KESIMPULAN

Analisis klaster dan klasifikasi dilakukan untuk memprediksi tingkat kemiskinan di provinsi-provinsi Indonesia. Masalah pada data, seperti nilai yang hilang dan *outlier*, diatasi dengan mengisi nilai minimum, menggunakan KNN *Imputation*, dan menerapkan teknik *Winsorizing* untuk menjaga kualitas data. Dalam pengembangan model klasterisasi, tiga metode yang diuji adalah K-Means, *Hierarchical Clustering*, dan DBSCAN. Meskipun *Hierarchical Clustering* menunjukkan kinerja yang baik pada dua metrik evaluasi (*Silhouette Coefficient* dan *Davies-Bouldin Index*), K-Means dinyatakan sebagai model terbaik berdasarkan nilai *Calinski-Harabasz Index* tertinggi. Hal ini menunjukkan bahwa K-Means mampu memberikan pembagian klaster yang lebih jelas dan terpisah. Untuk model klasifikasi, algoritma *Random Forest* menunjukkan akurasi tinggi sebesar 99,03%. Namun, untuk mengatasi potensi *overfitting* pada model yang dihasilkan, dilakukan pengujian dengan *cross-validation* dan menghitung nilai *Mean Accuracy* serta *Standard Deviation* (*Std Deviation*). Hasilnya menunjukkan bahwa *Naive Bayes* memberikan kombinasi performa terbaik dengan akurasi sangat tinggi (99,42%). Secara keseluruhan, penelitian ini berhasil mengidentifikasi masalah dalam data dan mengembangkan model klaster serta klasifikasi yang efektif untuk memprediksi tingkat kemiskinan di Indonesia. Hasil ini dapat menjadi dasar dalam merumuskan kebijakan pengentasan kemiskinan di tingkat provinsi.

REFERENSI

[1] M. Sihag *et al.*, "A Data-Driven Approach for Finding Requirements Relevant Feedback from TikTok and YouTube," *Proceedings of the IEEE International Conference on Requirements Engineering*, vol. 2023-September, pp. 111–122, 2023, doi: 10.1109/RE57278.2023.00020.

[2] F. Joanda Kaunang, "Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia," *Cogito Smart Journal*, vol. 4, no. 2, pp. 348–358, Dec. 2018, [Online]. Available: https://cogito.unklab.ac.id/index.php/cogito/article/view/141

[3] R. I. Prasetyono and D. Anggraini, "Analisis Peramalan Tingkat Kemiskinan Di Indonesia Dengan Model Arima," *Jurnal Ilmiah Informatika Komputer*, vol. 26, no. 2, pp. 95–110, 2021, doi: 10.35760/ik.2021.v26i2.3699.

[4] E. Pujiana, I. Purnama Sari, V. Melia Mardika, and M. Putri, "Analisis Algoritma Back Propagation Dalam Prediksi Angka Kemiskinan Di Indonesia," *Pendekar : Jurnal Pendidikan Berkarakter*, vol. 3, no. 1, pp. 11–17, 2020, doi: 10.31764.

[5] A. Karim, "Perbandingan Prediksi Kemiskinan di Indonesia Menggunakan Support Vector Machine (SVM) dengan Regresi Linear," *Jurnal Sains Matematika dan Statistika*, vol. 6, no. 1, pp. 107–113, Jan. 2020.

[6] F. Kusuma *et al.*, "Prediksi Jumlah Penduduk Miskin Indonesia menggunakan Metode Single Moving Average dan Double Moving Average," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 3, no. 2, pp. 105–109, Sep. 2021.

[7] A. D. Mukmin, R. Irsyada, and H. Audytra, "Penerapan Metode Moving Average Pada Sistem Informasi Prediksi Angka Kemiskinan," *Jurnal Multidisciplinary Applications of Quantum Information Science (al-mantiq*, vol. 1, no. 1, pp. 43–50, Sep. 2021, doi: 10.35314/isi.v2i1.112.

[8] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," pp. 3–21, 2020, doi: 10.1007/978-3-030-22475-2_1.

[9] A. Avram, O. Matei, C.-M. Pintea, P. C. Pop, and C. A. Anton, "Comparative Analysis of Clustering Techniques for a Hybrid Model Implementation," in *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020)*, Á. Herrero, C. Cambra, D. Urda, J. JSedano, H. Quintián, and E. Corchado, Eds., Springer, Cham, 2021, pp. 22–32. doi: 10.1007/978-3-030-57802-2_3.

[10] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pp. 747–748, Oct. 2020, doi: 10.1109/DSAA49011.2020.00096.

[11] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[12] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "Big Data Preprocessing: Enabling Smart Data," *Big Data Preprocessing: Enabling Smart Data*, pp. 1–186, Jan. 2020, doi: 10.1007/978-3-030-39105-8/COVER.

[13] T. Milo and A. Somech, "Automating Exploratory Data Analysis via Machine Learning: An Overview," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA: ACM, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[14] Hartatik *et al.*, *Data Science For Business*. 2023. [Online]. Available: www.sonpedia.com

[15] B. J. J. Kremers, A. Ho, J. Citrin, and K. L. van de Plassche, "Two step clustering for data reduction combining DBSCAN and k-means clustering," *Journal Metrics: Contributions to Plasma Physics*, Nov. 2021, doi: 10.1002/ctpp.202200177.

[16] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[17] . T. *et al.*, "Clustering Analysis of Premier Research Fields," *International Journal of Engineering & Technology*, vol. 7, no. 4.44, 2018, doi: 10.14419/ijet.v7i4.44.26860.

[18] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 293–301, Nov. 2023, doi: 10.57152/malcom.v3i2.952.

[19] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jan. 2020. doi: 10.1088/1757-899X/725/1/012128.

[20] M. A. A. Hakim and T. Terttiaavini, "Predictive Buyer Behavior Model as Customer Retention Optimization Strategy in E-commerce," *Journal of Intelligent System and Computation*, vol. 6, no. 1, pp. 32–38, Apr. 2024, doi: 10.52985/insyst.v6i1.379.

[21] K. Furmańczyk, K. Paczutkowski, M. Dudziński, and D. Dziewa-Dawidczyk, "Classification Methods Based on Fitting Logistic Regression to Positive and Unlabeled Data," in *Computational Science – ICCS 2022: 22nd International Conference, London, UK,* D. Groen, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds., London: Springer Cham, Jun. 2022, pp. 31–45. doi: 10.1007/978-3-031-08751-6_3.