ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

Analisis Faktor Dominan Yang Mempengaruhi Kemiskinan Di Sumatera Selatan Menggunakan *Machine Learning*

Annisa Maulina[1)], **Terttiaavini**[2)], **Agustina Heryati**[3)]

[1), 3)] *Sistem Informasi,* [2)] *Ilmu Komputer,* [1), 2), 3)] *Universitas Indo Global Mandiri*
Jl. Jend. Sudirman No. 629 Km. 4 Palembang, 30129
Email : annisamaulina14@gmail.com[1)], avini.saputra@uigm.ac.id[2)], agustina.heryati@uigm.ac.id[3)]

ABSTRACT

Poverty is a complex social issue in South Sumatra Province. Although the poverty rate has declined, it remains above the national average, indicating the need for a deeper analysis of the factors influencing it. In the digital era, poverty analysis can no longer rely solely on traditional methods but must incorporate information technology–based approaches and data analytics to uncover underlying socio-economic patterns. This study aims to analyze the dominant factors affecting poverty. The data used were obtained from the official publications of the Central Bureau of Statistics (BPS) of South Sumatra Province for the period 2020–2024, consisting of 85 observations from 17 regencies/cities, covering indicators of education, health, economy, and demography. The analysis was conducted through data preprocessing, correlation analysis, and predictive modeling using Machine Learning algorithms. The XGBoost algorithm was initially employed; however, the evaluation results indicated suboptimal performance. Therefore, a comparison was conducted with the Random Forest algorithm, which demonstrated better performance, achieving an R^2 value of 0.654, and was thus selected as the final model. Interpretation using the SHAP method revealed that increases in Life Expectancy, Gross Regional Domestic Product (GRDP) per capita, Mean Years of Schooling, Expected Years of Schooling, and the Human Development Index (HDI) contribute to reducing poverty predictions. In contrast, Population Size and Expenditure per capita contribute to increasing poverty predictions. These findings are expected to serve as a reference for local governments in designing more targeted policies and programs.

Keywords : *Poverty, Machine Learning, Random Forest, SHAP.*

ABSTRAK

Kemiskinan merupakan permasalahan sosial yang kompleks di Provinsi Sumatera Selatan. Meskipun mengalami penurunan, persentase kemiskinan masih berada di atas rata-rata nasional, sehingga diperlukan analisis yang lebih mendalam terhadap faktor-faktor yang memengaruhinya. Dalam era digital, analisis kemiskinan tidak cukup hanya dengan mengandalkan metode tradisional, tetapi perlu melibatkan pendekatan berbasis teknologi informasi dan analisis data untuk mengungkap pola ekonomi sosial yang mendasarinya. Penelitian ini bertujuan untuk menganalisis faktor dominan yang memengaruhi kemiskinan. Data yang digunakan bersumber dari Badan Pusat Statistik (BPS) Provinsi Sumatera Selatan periode 2020 hingga 2024 dengan total 85 data dari 17 kabupaten/kota, mencakup indikator pendidikan, kesehatan, ekonomi, dan demografi. Analisis dilakukan melalui tahapan pra-pemrosesan data, analisis korelasi, serta pemodelan prediktif menggunakan algoritma *Machine Learning*. Algoritma XGBoost digunakan pada tahap awal pemodelan, namun hasil evaluasi menunjukkan kinerja yang kurang optimal. Oleh karena itu, dilakukan perbandingan dengan algoritma *Random Forest* yang menghasilkan performa lebih baik dibandingkan XGBoost dengan nilai R^2 sebesar 0.654, sehingga ditetapkan sebagai model akhir. Interpretasi menggunakan metode SHAP mengidentifikasi bahwa peningkatan Umur Harapan Hidup (UHH), PDRB per kapita, Rata-rata Lama Sekolah, Harapan Lama Sekolah, dan Indeks Pembangunan Manusia (IPM) berkontribusi menurunkan prediksi kemiskinan, sedangkan variabel Jumlah Penduduk dan Pengeluaran per kapita berkontribusi meningkatkan prediksi kemiskinan. Temuan ini diharapkan dapat menjadi bahan pertimbangan bagi pemerintah daerah dalam merancang kebijakan serta program yang lebih tepat sasaran.

Kata Kunci : Kemiskinan, *Machine Learning, Random Forest*, SHAP.

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

1. Pendahuluan

Kemiskinan merupakan kondisi ketika individu atau kelompok tidak mampu memenuhi kebutuhan dasar seperti makanan, tempat tinggal, pendidikan, pakaian, dan pekerjaan untuk hidup layak (Pratiwi et al., 2023). Di era digital, upaya pengentasan kemiskinan tidak hanya mengandalkan metode tradisional, tetapi juga memerlukan pemanfaatan teknologi informasi dan analisis data untuk mengungkap pola sosial ekonomi yang memengaruhinya.

Provinsi Sumatera Selatan masih menghadapi tingkat kemiskinan yang cukup tinggi meskipun terus mengalami penurunan. Pada Maret 2024, tingkat kemiskinan tercatat sebesar 10,97% atau sekitar 984,24 ribu jiwa, lebih tinggi dibanding rata-rata nasional sebesar 9,03% (BPS Indonesia, 2024). Tingkat kemiskinan di perkotaan sebesar 10,04%, sedangkan di perdesaan mencapai 11,53% (Badan Pusat Statistik Provinsi Sumatera Selatan, 2024). BPS juga menetapkan garis kemiskinan sebesar Rp554.197 per kapita per bulan, yang sebagian besar dipengaruhi oleh kebutuhan makanan (Badan Pusat Statistik Provinsi Sumatera Selatan, 2024). Kondisi ini mengindikasikan bahwa kemiskinan masih menjadi masalah besar di wilayah tersebut.

Kemiskinan dipengaruhi oleh berbagai faktor, seperti pendidikan, kesehatan, tingkat pengangguran, tempat tinggal, dan pendapatan masyarakat. Melihat kompleksitas kemiskinan, penting untuk mengetahui secara tepat faktor-faktor dominan yang memengaruhinya. Pemahaman ini akan menjadi dasar dalam merancang kebijakan pengentasan kemiskinan yang lebih efektif. Untuk itu, penelitian ini bertujuan untuk mengidentifikasi faktor-faktor dominan yang berperan dalam memengaruhi tingkat kemiskinan di Provinsi Sumatera Selatan melalui pendekatan berbasis data.

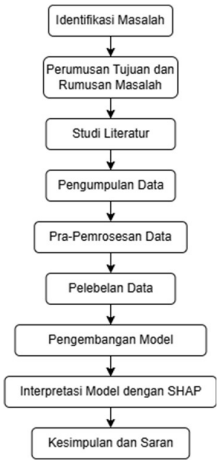
Beberapa penelitian sebelumnya telah menggunakan pendekatan statistik dan *Machine Learning* untuk menganalisis kemiskinan. Penelitian-penelitian tersebut menunjukkan bahwa *Machine Learning* mampu memberikan performa prediksi yang baik. Namun, penelitian terkait kemiskinan yang mampu memberikan interpretasi terhadap kontribusi masing-masing variabel masih terbatas. Selain itu, faktor-faktor yang digunakan dalam penelitian ini diperoleh dari beberapa penelitian terdahulu dan disesuaikan dengan ketersediaan data Badan Pusat Statistik (BPS).

Penelitian ini menerapkan algoritma XGBoost untuk membangun model prediksi dan mengidentifikasi faktor dominan yang paling berpengaruh. XGBoost dipilih karena memiliki performa yang baik dalam berbagai penelitian prediksi (Wijaya et al., 2024). Selain itu, metode Shapley Additive Explanations (SHAP) digunakan untuk menginterpretasikan hasil model dan mengetahui kontribusi masing-masing variabel terhadap prediksi kemiskinan (Mumpuni, 2024).

Hasil penelitian ini diharapkan dapat berkontribusi pada perkembangan ilmu, terutama di bidang data science dan ekonomi pembangunan, dengan penerapan interpretasi *Machine Learning* dalam menganalisis masalah sosial. Selain itu, hasil penelitian juga diharapkan menjadi bahan pertimbangan bagi Badan Perencanaan Pembangunan Daerah (Bappeda) dan Dinas Sosial dalam menyusun kebijakan dan program pengentasan kemiskinan yang lebih tepat sasaran.

2. Metodologi Penelitian

Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis faktor-faktor yang memengaruhi tingkat kemiskinan di Provinsi Sumatera Selatan. Penelitian ini memanfaatkan algoritma *Machine Learning*, yaitu Extreme Gradient Boosting (XGBoost), serta metode SHAP (Shapley Additive Explanations) untuk menginterpretasikan hasil model.



Gambar 1. *Tahapan Penelitian*

A. Pengumpulan Data

Penelitian ini memperoleh data dari sumber sekunder yaitu melalui publikasi resmi dari Badan Pusat Statistik (BPS) Provinsi Sumatera Selatan melalui situs https://sumsel.bps.go.id/id. Data ini mencakup seluruh 17 kabupaten dan kota Provinsi Sumatera Selatan selama periode 2020-2024, sehingga diperoleh total 85 observasi.

Variabel yang digunakan dalam penelitian ini meliputi Persentase Penduduk Miskin, Harapan Lama Sekolah, Rata-rata Lama Sekolah, Pengeluaran per Kapita, Jumlah Fasilitas Kesehatan, Indeks Pembangunan Manusia (IPM), Umur Harapan Hidup, Jumlah Penduduk, dan Produk Domestik Regional Bruto (PDRB) per Kapita.

Penelitian ini menerapkan metode sampel jenuh, di mana setiap individu dalam populasi dijadikan sebagai sampel (Fitra Anisa et al., 2024).

Penelitian dilakukan analisis statistik deskriptif sebagai langkah awal sebelum memasuki tahapan pengolahan dan analisis data lebih lanjut. Statistik Deskriptif merupakan statistik yang digunakan untuk menyajikan dan menggambarkan karakteristik suatu kumpulan data, dapat dikatakan bahwa statistik deskriptif hanya berhubungan dengan menjelaskan atau memberikan informasi mengenai data, kondisi, atau fenomena tertentu (Andrianingsih et al., 2024).

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

B. Pra-pemprosesan data

Penelitian ini melakukan tahap Pra-pemrosesan data guna memastikan bahwa data yang digunakan dalam penelitian berada dalam siap digunakan dalam proses analisis lebih lanjut. Pada tahap ini, data yang diperoleh akan mengalami pembersihan, termasuk penanganan nilai yang hilang, duplikat, ketidaksesuaian, serta kesalahan dalam penulisan atau format. Selain itu, juga dilakukan pemilihan dan pengubahan data, yaitu mengubah data agar sesuai dengan kebutuhan proses *data mining* (Laila Hasanah et al., 2025).

1. Penanganan Data Hilang

Tahapan pra-pemrosesan diawali dengan melakukan pemeriksaan terhadap keberadaan nilai kosong (*missing values)* pada seluruh atribut. Berdasarkan hasil pemeriksaan tidak ditemukan adanya nilai yang hilang, sehingga tidak diperlukan tindakan imputasi maupun penghapusan data.

2. Penanganan *Outlier*

Peneliti melakukan identifikasi dan penanganan *Outlier* sebagai bagian dari tahap pra-pemrosesan data. *Outlier* adalah nilai yang secara mencolok berbeda dari pola umum yang terdapat dalam kumpulan data. Keberadaan *Outlier* dapat mempengaruhi hasil analisis statistik serta prediksi yang dihasilkan oleh model (Jailani & Erna, 2025).

Proses deteksi awal dilakukan melalui visualisasi *boxplot* untuk memperoleh gambaran distribusi dan nilai ekstrem dari masing-masing variabel. Untuk mengidentifikasi *Outlier* secara lebih sistematis, digunakan metode *Interquartile Range (IQR)*.

$$IQR = Q_3 - Q_1 \quad (1)$$

Hitung Q1, Q3, dan IQR. Setelah itu, tentukan batas bawah dan batas atas dari batas bawah.

$$Lower\ Bound = Q_1 - 1.5 \times IQR \quad (2)$$
$$Upper\ Bound = Q_3 + 1.5 \times IQR \quad (3)$$

Keterangan:

Q_1 (Kuartil 1): Nilai yang memisahkan 25% data terbawah.

Q_3 (Kuartil 3): Nilai yang memisahkan 75% data terbawah.

Hasil dari *Outlier* yang telah teridentifikasi oleh *boxplot* dan metode IQR, sejumlah variabel diketahui mengandung *Outlier* dalam jumlah yang signifikan, maka dilakukan penanganan *Outlier* menggunakan metode *Capping* atau Winsorisasi, *Winsorizing* adalah teknik yang mengurangi dampak dari *Outlier*s dengan memangkas data di kedua ujung distribusi dan menggantikannya dengan nilai batas tertentu (Naufal Hibatulloh et al., 2025). Metode ini dipilih karena dianggap efektif dalam mengurangi pengaruh nilai-nilai ekstrem (*Outlier*) tanpa harus menghapus data dari *dataset*, sehingga ukuran sampel tetap terjaga.

3. Normalisasi data

Normalisasi data merupakan tahapan lanjutan setelah dilakukan penanganan *Outlier*. Metode normalisasi data yaitu langkah untuk memastikan bahwa berbagai variabel mencapai rentang nilai yang serupa, sehingga tidak ada yang terlalu tinggi atau rendah, sehingga mempermudah analisis statistik (Kusnaidi et al., 2022). Dalam penelitian ini, normalisasi dilakukan menggunakan teknik *Min-Max Scaling*. MinMaxScaler merupakan salah satu metode yang mengubah nilai asli suatu fitur menjadi rentang antara 0 hingga 1 (Jasril et al., 2025).

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (4)$$

Keterangan:
X adalah nilai asli dari fitur,
Xmin : nilai minimum pada fitur tersebut
Xmax : nilai maksimum pada fitur tersebut
Xnorm : nilai setelah dinormalisasi

4. Korelasi Antar Variabel

Dalam penelitian ini, analisis korelasi dilakukan untuk menemukan hubungan atau korelasi antara dua atau lebih variabel untuk memahami sejauh mana tingkat hubungan itu (Nurhaswinda et al., 2025)

Gambar yang dihasilkan adalah *Heatmap* korelasi antar variabel, yang menggambarkan hubungan antar variabel. Hasil visualisasi *Heatmap* ini menunjukkan hubungan antara variabel ditunjukkan oleh nilai korelasi berkisar -1 hingga 1, dimana:

a. Nilai yang mendekati angka 1 menunjukkan hubungan positif yang kuat.
b. Nilai yang mendekati angka -1 menunjukkan hubungan negatif yang kuat.
c. Nilai yang mendekati angka 0 menunjukkan hubungan lemah atau tidak signifikan (Desmarita et al., 2023)

C. Penentuan Label *Clustering*

Dalam penelitian ini, data tidak memiliki label kategori kemiskinan, sehingga diperlukan proses penentuan label. Melakukan *Clustering* terhadap data numerik (kemiskinan), menggunakan empat algoritma berbeda, lalu membandingkan performanya dengan tiga metrik evaluasi kualitas *Cluster*.

1. Cluster

Dalam penelitian ini digunakan empat algoritma *Clustering* yang berbeda untuk mengetahui pengelompokan data indikator kemiskinan. Pemilihan beberapa algoritma dilakukan untuk memperoleh hasil pengelompokan yang lebih komprehensif, serta untuk

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

membandingkan karakteristik dan performa masing-masing metode dalam konteks data yang dianalisis.

a. Kmeans
Algoritma *K-Means* merupakan teknik pengelompokan data yang bertujuan mengelompokkan informasi berdasarkan persamaan karakteristik antar data data (Srirahmawati et al., 2025).

b. *Agglomerative Clustering*
Agglomerative Clustering merupakan metode hierarki yang mengelompokkan data melalui proses penggabungan kelompok kecil menjadi kelompok yang lebih besar sampai seluruh data terkumpul dalam satu kelompok (Zahra Nur Fadhilah et al., 2025).

c. Gaussian Mixture Model (GMM)
Salah satu cara yang paling ampuh untuk mendapatkan pemahaman yang tepat mengenai objek dalam rata-rata adalah dengan memanfaatkan *Gaussian Mixture Model* (GMM). Dalam kerangka GMM, terdapat dua jenis nilai yang diuraikan, yaitu bobot dari komponen campuran serta nilai rata-rata, dan juga varians/kovarians dari komponen-komponen yang berperan sebagai parameter penting (Fitri Syalsabilla et al., 2024).

d. HDBSCAN
HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) adalah teknik pengelompokan untuk mengatasi data yang mengandung *Outlier* (Ghaly Ghiffary et al., 2024).

2. Evaluasi Hasil *Clustering*

Untuk menentukan algoritma mana yang paling sesuai digunakan dalam konteks data kemiskinan, digunakan tiga metrik yaitu Skor Silhouette, IndeksDavies-Bouldin, dan Indeks Calinski-Harabasz. Skor Silhouette menilai kekompakan dan terpisahnya klaster, Indeks Davies-Bouldin menilai sejauh mana klaster saling mendekati, dan Indeks Calinski-Harabasz menilai perbedaan varians antar klaster dan dalam klaster.

a. *Silhouette Score* mengukur seberapa mirip suatu data dengan klaster tempatnya berada dibandingkan dengan klaster lain. Nilai berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa data berada dalam klaster yang tepat dan terpisah secara jelas dari klaster lainnya (Ishak, 2024).
b. *Davies-Bouldin Index* (DBI) mengevaluasi rata-rata kesamaan antar klaster berdasarkan rasio varians dalam klaster dan jarak antar klaster. Semakin kecil nilai DBI, semakin baik kualitas klaster (Ishak, 2024).
c. *Calinski-Harabasz Index* (CHI) mengukur rasio antara variansi antar klaster dan variansi dalam klaster. Nilai CHI yang lebih tinggi menunjukkan klaster lebih jelas dan kompak (Putri Azizah et al., 2025)

Hasil evaluasi menunjukkan bahwa algoritma HDBSCAN memberikan performa *Clustering* terbaik untuk data kemiskinan yang digunakan dalam penelitian ini.

D. Pengembangan Model

Dalam penelitian ini, model XGBoost dan Random Forest diterapkan untuk memprediksi tingkat kemiskinan menggunakan kumpulan data yang telah diolah. Evaluasi dilakukan untuk mengetahui seberapa baik model dalam melakukan prediksi terhadap data yang tersedia. Tiga metrik evaluasi yang digunakan dalam penelitian ini adalah:

1. Mean Absolute Error (MAE)

MAE digunakan untuk menghitung rata-rata selisih absolut antara nilai prediksi dan nilai aktual. Semakin rendah nilai MAE yang dihasilkan, maka semakin akurat model dalam melakukan prediksi (Yulis et al., 2025).

$$MAE = \frac{1}{n}\sum_{t}^{n} = 1|X_t - F_t| \quad (5)$$

Keterangan:
Ft: Nilai ramalan
Xt: Nilai actual
n: Jumlah data error (Bufa Al badi et al., 2025).

2. *Mean Squared Error* (MSE)

Mean Squared Error (MSE) adalah metrik evaluasi yang umum digunakan karena perhitungannya sederhana dan mudah diterapkan. Semakin kecil nilai MSE, maka semakin baik kinerja model karena menunjukkan kesalahan prediksi yang lebih kecil (Togatorop, 2024).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (6)$$

Keterangan:
n: Jumlah data
yi: Aktual dari data ke-i
ŷ: Prediksi dari data ke-i (Togatorop, 2024).

3. *R^2 Score*

R^2 Score atau koefisien determinasi merupakan ukuran statistik dalam regresi yang digunakan untuk menilai seberapa baik sebuah model dapat memprediksi nilai dari suatu data (Tri et al., 2025).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(yi - \bar{y})^2} \quad (7)$$

Keterangan:
R^2: Koefisien determinasi
yi: Aktual dari data ke-i
ŷ: Prediksi dari data ke-i
ȳ: Rata-rata dari selulruh nilai aktual
n: Jumlah total data (Tri et al., 2025).

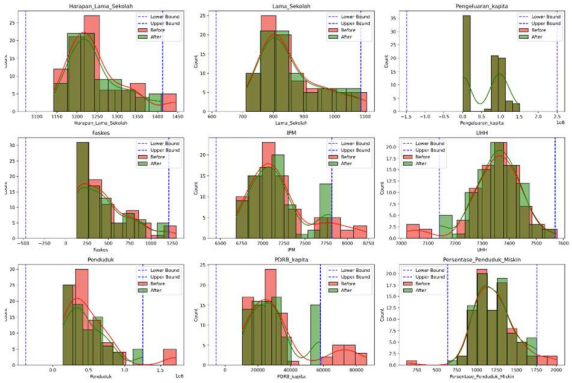
ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

3. Hasil dan Pembahasan

Tahap pra-pemrosesan bertujuan untuk memastikan bahwa data yang digunakan menjadi lebih terstruktur dan konsisten, serta memenuhi standar kelayakan untuk digunakan dalam proses pemodelan dengan algoritma *Machine Learning*.

A. Hasil penanganan *Outlier*

Proses deteksi awal dilakukan menggunakan visualisasi *boxplot* yang mampu menunjukkan sebaran data dan nilai ekstrem setiap variabel. Berdasarkan hasil identifikasi tersebut, beberapa variabel diketahui mengandung *Outlier* dalam jumlah yang signifikan, maka dilakukan penanganan *Outlier* menggunakan metode *Capping* atau Winsorisasi, yaitu membatasi nilai-nilai ekstrem agar tidak melampaui batas bawah dan batas atas yang ditentukan berdasarkan IQR.



Gambar 2 Penanganan *Outlier*

Gambar 2 memperlihatkan visualisasi menggunakan histogram dari sembilan variabel sebelum (warna merah) dan sesudah (warna hijau) dilakukan penyesuaian terhadap *Outlier* menggunakan metode Winsorization. Garis biru putus-putus menunjukkan batas bawah dan batas atas yang dihitung menggunakan metode Interquartil Range (IQR).

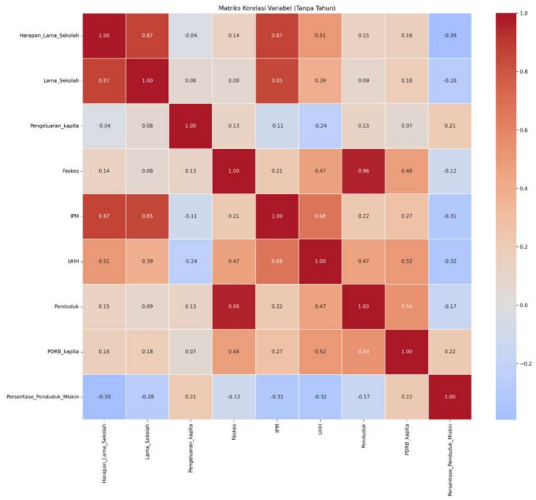
B. Hasil Normalisasi Data

Setelah dilakukan penanganan terhadap *Outlier*, tahap selanjutnya adalah normalisasi data untuk memastikan setiap variabel memiliki skala yang sama. Pada penelitian ini, normalisasi dilakukan menggunakan teknik *Min-Max Scaler*, yang mengubah nilai asli suatu variabel ke dalam rentang nilai yang serupa antara 0 hingga 1.

Hasil normalisasi menunjukkan bahwa seluruh nilai telah berhasil dipetakan dalam rentang 0 hingga 1 sesuai karakterisktik metode *Min-Max Scaler*. Dengan demikian, data yang telah dinormalisasi siap digunakan untuk proses analisis dan pemodelan selanjutnya.

C. Hasil Korelasi Antar Variabel

Analisis korelasi dilakukan untuk mengetahui hubungan antar variabel. Divisualisasikan dalam bentuk *Heatmap*, yang merepresentasikan tingkat hubungan antara satu variabel dengan variabel lainnya. Nilai korelasi yang ditampilkan berada dalam rentang -1 hingga 1, yang menunjukkan seberapa kuat hubungan antar variabel tersebut.



Gambar 3 Hasil Variabel Korelasi

Berdasarkan visualisasi matriks korelasi, diperoleh beberapa temuan sebagai berikut:

1. Variabel Harapan Lama Sekolah dan Lama Sekolah menunjukkan korelasi positif yang sangat kuat, yaitu sebesar 0,87. Hubungan ini mengindikasikan bahwa semakin tinggi harapan lama sekolah, maka secara umum lama sekolah aktual juga meningkat.
2. Variabel antara IPM dan Harapan Lama Sekolah juga sangat kuat dengan nilai 0,87, menunjukkan bahwa peningkatan pembangunan manusia berkaitan erat dengan indikator pendidikan.
3. Demikian pula, IPM dan Lama Sekolah memiliki korelasi kuat sebesar 0,85, yang semakin menegaskan peran pendidikan dalam pembangunan manusia.
4. Hubungan antara Jumlah Penduduk dan Fasilitas Kesehatan memiliki korelasi yang sangat kuat, yaitu 0,96. Hal ini mencerminkan bahwa daerah dengan jumlah penduduk yang lebih besar cenderung memiliki lebih banyak fasilitas kesehatan.

Variabel Persentase Penduduk Miskin memiliki korelasi negatif terhadap beberapa variabel diantaranya:

1. Harapan Lama Sekolah sebesar -0,39,
2. IPM sebesar -0,31,
3. Umur Harapan Hidup (UHH) sebesar -0,32, dan
4. Lama Sekolah sebesar -0,28.

Nilai-nilai ini menunjukkan bahwa peningkatan dalam kualitas pendidikan dan kesehatan cenderung berkorelasi dengan penurunan tingkat kemiskinan.

Peneliti menemukan bahwa nilai korelasi antara Pengeluaran per Kapita dan Persentase Penduduk Miskin adalah 0,21, yang termasuk dalam kategori korelasi

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

positif lemah. Hal ini bisa menunjukkan bahwa meskipun pengeluaran meningkat, belum tentu menurunkan kemiskinan, mungkin karena distribusinya tidak merata atau karena faktor lain yang mempengaruhi kesejahteraan.

D. Hasil Evaluasi *Clustering*

Untuk menentukan klasterisasi yang paling optimal digunakan dala konteks data kemiskinan ini, penelitian ini membandingkan empat algoritma *Clustering*, yaitu *K-Means*, *Agglomerative Clustering*, *Gaussian Mixture Model* (GMM), dan HDBSCAN. Pemilihan algoritma terbaik didasarkan pada hasil evaluasi menggunakan tiga metrik pengukuran, yaitu *Silhouette Score*, Davies-Bouldin Index, dan *Calinski-Harabasz Index*.

Tabel 1 Hasil *Clustering*

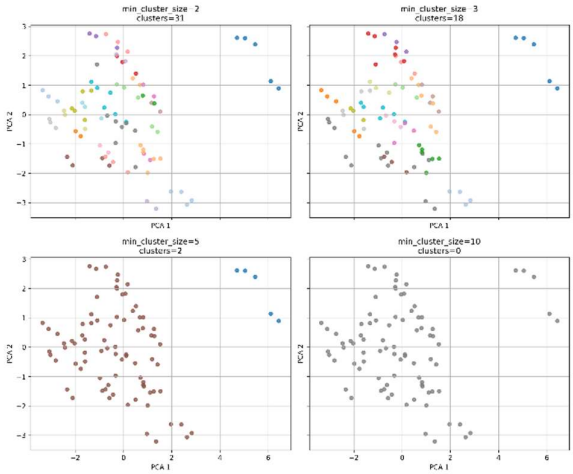
Algoritma *Clustering*	*Silhouette Score*	*Davies-Bouldin Index*	Calinski-Harabasz Index
K-Means	0.2703	1.3306	24.2564
Agglomerative Clustering	0.2741	1.2808	29.6973
Gaussian Mixture Model (GMM)	0.2718	1.3331	24.3315
HDBSCAN	0.3892	0.9736	26.7532

1. HDBSCAN menunjukkan performa terbaik secara keseluruhan, dengan *Silhouette Score* tertinggi (0,3892), *Davies-Bouldin Index* terendah (0,9736), dan nilai *Calinski-Harabasz Index* yang cukup kompetitif (26,7532). Ini menunjukkan bahwa HDBSCAN menghasilkan klaster yang paling terpisah dan paling kompak.
2. *Agglomerative Clustering* memiliki nilai Calinski-Harabasz tertinggi (29,6973), menandakan struktur klaster yang jelas, namun *Silhouette Score* dan Davies-Bouldin-nya masih kalah dari HDBSCAN.
3. KMeans dan GMM memiliki performa yang lebih rendah secara keseluruhan, dengan *Silhouette Score* dan *Davies-Bouldin Index* yang tidak lebih baik dibanding metode lainnya.

Hasil evaluasi menunjukkan bahwa algoritma HDBSCAN memberikan performa *Clustering* terbaik untuk data kemiskinan yang digunakan dalam penelitian ini.

E. Hasil *Clustering* HDBSCAN

Hasil pengelompokkan wilayah ditunjukkan pada Gambar 4, yang menggambarkan persebaran wilayah berdasarkan dua dimensi menggunakan PCA. Masing-masing parameter min *Cluster* size menunjukkan perbedaan signifikan dalam jumlah klaster yang terbentuk.



Gambar 4 Visualisasi Klaster

Berdasarkan hasil visualisasi pada gambar 4 menunjukkan dua dimensi menggunakan PCA, terlihat bahwa data terbagi ke dalam beberapa klaster yang berbeda, dengan karakteristik sebagai berikut:

1. Min *Cluster* size =2 menghasilkan 31 klaster, klaster-klaster terlihat sangat kecil dan tidak stabil. Hampir setiap beberapa titik yang berdekatan dikategorikan sebagai klaster terpisah.
2. Min *Cluster* size =3 menghasilkan 18 klaster, masih menunjukkan pembentukan klaster yang cenderung kecil-kecil.
3. Min *Cluster* size =5 hanya menghasilkan 2 klaster, terlihat dua kelompok yang besar dominan. Menunjukkan bahwa hanya kelompok besar yang berhasildikenali sebagai klaster, sementara kelompok kecil dianggap sebagai *noise*.
4. Min *Cluster* size =10 tidak terbentuk klaster sama sekali, yang menunjukkan bahwa tidak ada kelompok yang memenuhi batas minimum ukuran klaster, sehingga seluruh data diperlakukan sebagai *noise*.

Berdasarkan proses klasterisasi menggunakan HDBSCAN dengan parameter min *Cluster* size, hasil menunjukkan bahwa tidak terdapat klaster yang kuat dan stabil. Hasil tersebut mengindikasikan bahwa untuk nilai min *Cluster* size 2 dan 3 algoritma cenderung menghasilkan terlalu banyak klaster. Sedangkan untuk nilai min *Cluster* size 5 dan 10, jumlah turun drastis bahkan menjadi nol yang menunjukkan tidak adanya kelompok wilayah yang cukup padat untuk dibentuk sebagai klaster oleh HBDSCAN.

F. Hasil Evaluasi Model

Dalam penelitian ini, model XGBoost digunakan tahap awal, namun menghasilkan performa yang kurang optimal. Sebagai alternatif, algoritma Random Forest diterapkan untuk memprediksi tingkat kemiskinan menggunakan kumpulan data yang telah diolah. Evaluasi

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

dilakukan untuk mengetahui seberapa baik model dalam melakukan prediksi terhadap data yang tersedia. Tiga metrik evaluasi yang digunakan dalam penelitian ini adalah *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *R² Score*.

Tabel 2 Pemodelan

Model	MAE	MSE	R^2Score
Random Forest	0.412941	0.294388	0.654154
XGBoost	0.397932	0.441010	0.481902

Berdasarkan hasil evaluasi, Random Forest menunjukkan performa yang lebih baik dibandingkan model XGBoost. Hal ini terlihat dari nilai *R² Score* sebesar 0.654154, yang lebih tinggi dibandingkan XGBoost 0.481902. Artinya, model Random Forest mampu menjelaskan sekitar 65,42% variasi data, sedangkan XGBoost hanya sekitar 48,19%.

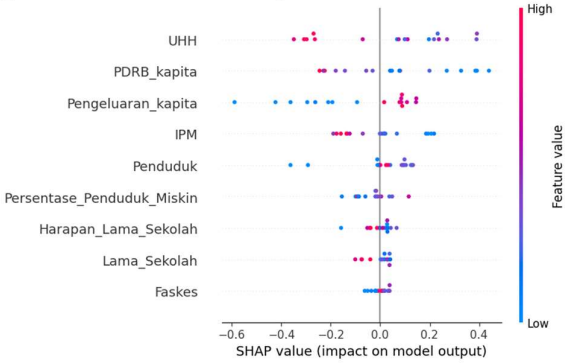
Meskipun nilai MAE XGBoost sedikit lebih rendah (0.3979) dibandingkan Random Forest (0.4129), namun secara keseluruhan, model Random Forest memiliki MSE yang lebih rendah dan R² yang lebih tinggi, yang menunjukkan kestabilan dan akurasi prediksi yang lebih baik. Dengan performa tersebut, Random Forest dinilai cukup layak sebagai model yang paling optimal dalam memprediksi tingkat kemiskinan pada data ini, dan hasilnya akan digunakan dalam proses interpretasi lebih lanjut menggunakan metode SHAP.

G. Evaluasi SHAP

Penelitian ini memanfaatkan *SHAP Summary Plot* untuk mengidentifikasi fitur-fitur yang memberikan kontribusi paling signifikan terhadap kinerja model, serta menggunakan *SHAP Dependence Plot* untuk melihat hubungan antara masing-masing fitur dengan hasil prediksi secara lebih transparan dan interpretatif.

1. *SHAP Summary Plot*

Gambar 5 merupakan *SHAP Summary Plot* yang menampilkan kontribusi masing-masing fitur terhadap prediksi model dalam menganalisis kemiskinan.



Gambar 5 *SHAP Summary Plot*

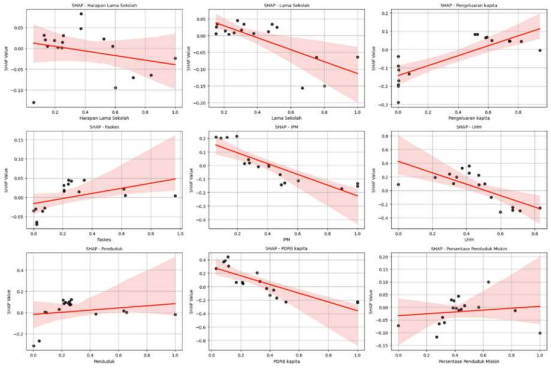
Berdasarkan hasil *SHAP Summary Plot*, faktor-faktor yang paling berpengaruh dalam menurunkan kemiskinan adalah Umur Harapan Hidup (UHH), PDRB per Kapita, Indeks Pembangunan Manusia (IPM), Lama Sekolah, dan Harapan Lama Sekolah. Nilai tinggi pada variabel-variabel tersebut menunjukkan kualitas kesehatan, pendidikan, dan aktivitas ekonomi yang lebih baik sehingga berkontribusi terhadap penurunan tingkat kemiskinan.

Sebaliknya, Jumlah Penduduk, Pengeluaran per Kapita, dan Persentase Penduduk Miskin memiliki kontribusi positif terhadap peningkatan kemiskinan. Tingginya jumlah penduduk dapat menambah beban ekonomi dan sosial, sedangkan pengeluaran per kapita yang tinggi dapat mencerminkan tingginya biaya hidup atau pengeluaran yang kurang efisien. Sementara itu, variabel Fasilitas Kesehatan (Faskes) menunjukkan pengaruh yang relatif kecil dan tidak konsisten dalam model.

Secara keseluruhan, hasil ini menunjukkan bahwa peningkatan kualitas pendidikan, kesehatan, dan ekonomi menjadi faktor utama dalam upaya pengentasan kemiskinan di Provinsi Sumatera Selatan.

2. *SHAP Dependence Plot*

SHAP Dependence Plot digunakan untuk menunjukkan visualisasi hubungan antara nilai fitur dan kontribusinya terhadap prediksi tingkat kemiskinan. Metode ini memungkinkan interpretasi yang lebih mendalam, dapat diperoleh gambaran yang lebih jelas mengenai faktor-faktor dominan yang berkontribusi terhadap tingkat kemiskinan di Sumatera Selatan.



Gambar 6 *Dependence Plot*

Berdasarkan *SHAP Dependence Plot*, sebagian besar variabel menunjukkan hubungan yang konsisten terhadap prediksi kemiskinan. Harapan Lama Sekolah, Lama Sekolah, IPM, Umur Harapan Hidup (UHH), dan PDRB per Kapita memiliki pola menurun, di mana peningkatan nilai variabel tersebut menghasilkan nilai SHAP yang semakin negatif. Hal ini menunjukkan bahwa peningkatan kualitas pendidikan, kesehatan, dan ekonomi berkontribusi dalam menurunkan tingkat kemiskinan.

Sebaliknya, Pengeluaran per Kapita, Jumlah Penduduk, dan Persentase Penduduk Miskin menunjukkan pola positif, yang berarti peningkatan nilai

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

variabel tersebut cenderung meningkatkan prediksi kemiskinan. Sementara itu, variabel Fasilitas Kesehatan (Faskes) memiliki pola yang relatif datar dan tidak konsisten, sehingga pengaruhnya terhadap model tergolong kecil.

Secara keseluruhan, hasil SHAP *Dependence Plot* menegaskan bahwa faktor pendidikan, kesehatan, dan pertumbuhan ekonomi memiliki peran penting dalam pengurangan kemiskinan, sedangkan tingginya jumlah penduduk dan pengeluaran per kapita cenderung meningkatkan risiko kemiskinan.

4. Kesimpulan & Saran

Penelitian ini bertujuan menganalisis faktor-faktor dominan yang mempengaruhi kemiskinan di Provinsi Sumatera Selatan menggunakan pendekatan *Machine Learning*, yaitu HDBSCAN untuk *Clustering* dan XGBoost yang diinterpretasikan dengan SHAP. Hasil penelitian menunjukkan bahwa algoritma HDBSCAN kurang efektif dalam membentuk klaster wilayah karena struktur data tidak menghasilkan pemisahan yang jelas berdasarkan kepadatan. Pada tahap pemodelan, XGBoost menghasilkan performa yang kurang optimal sehingga Random Forest dipilih sebagai model akhir dengan nilai R^2 sebesar 0,654.

Analisis SHAP menunjukkan bahwa faktor-faktor yang berperan dalam menurunkan kemiskinan adalah Umur Harapan Hidup (UHH), PDRB per Kapita, Lama Sekolah, Harapan Lama Sekolah, dan Indeks Pembangunan Manusia (IPM). Sebaliknya, Jumlah Penduduk dan Pengeluaran per Kapita berkontribusi meningkatkan risiko kemiskinan. Hasil ini menegaskan bahwa peningkatan kualitas hidup, pendidikan, kesehatan, dan pertumbuhan ekonomi menjadi dasar utama dalam pengentasan kemiskinan. Temuan lain menunjukkan bahwa tingginya pengeluaran per kapita dapat berkorelasi dengan meningkatnya kemiskinan, yang kemungkinan dipengaruhi oleh tingginya biaya hidup, inflasi, atau pengeluaran yang kurang produktif.

Berdasarkan hasil penelitian, peneliti selanjutnya disarankan menggunakan periode data yang lebih panjang agar hasil analisis lebih stabil dan relevan untuk kebijakan jangka panjang. Selain itu, metode *Clustering* lain seperti *K-Means* dapat digunakan untuk menghasilkan pengelompokan wilayah yang lebih terstruktur, serta dilakukan perbandingan Random Forest dengan model *Machine Learning* lain yang lebih kompleks untuk memperoleh performa yang lebih optimal.

Bagi Bappeda Provinsi Sumatera Selatan, program pembangunan dan alokasi anggaran perlu difokuskan pada peningkatan IPM, UHH, pendidikan, dan PDRB per Kapita sebagai investasi jangka panjang dalam mengurangi kemiskinan. Selain itu, diperlukan strategi pengelolaan jumlah penduduk yang sejalan dengan kapasitas ekonomi dan lingkungan daerah agar tidak menambah beban kemiskinan.

Daftar Pustaka

A'Fena, Elisa Tri Ammah, Aidina Ristyawan, and Arie Nugroho. "Pengaruh Normalisasi Fitur Pada Perbandingan Algoritma Random Forest Regressor Dan KNN Regressor Dalam Menentukan Harga Smartphone." *KRESNA: Jurnal Riset dan Pengabdian Masyarakat* 5, no. 1 (2025): 74-83. https://doi.org/10.36080/kresna.v5i1.226

Andrianingsih, V., Sahrul, S., Asmoro, E. I., & Samsudin, S. (2024). Analisis Kesulitan Mahasiswa dalam Menyelesaikan Soal Statistik pada Prodi Teknik Informasi di Universitas X Nusa Tenggara Barat. Jurnal Pendidikan Matematika, 1(4), 6. https://doi.org/10.47134/ppm.v1i4.852

Azizah, Fathin Putri, Shofa Shofiah Hilabi, Tukino Tukino, and Agustia Hananto. "Perbandingan Algoritma *K-Means* dan Hierarchical Untuk Klasterisasi Data Kehadiran Karyawan." *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi* 14, no. 1 (2025).

Badan Pusat Statistik (2024). Persentase Penduduk Miskin Maret 2024 Turun Menjadi 9,03 Persen.

Badan Pusat Statistik (2024). Persentase Penduduk Miskin Maret 2024 Turun Menjadi 9,03 Persen.

Badan Pusat Statistik Provinsi Sumatera Selatan. (2024). Persentase Penduduk Miskin Provinsi Sumatera Selatan Maret 2024 Sebesar 10,97 Persen. Diakses 1 Juli 2024.

Bufa Al badi, Mufti Ari Bianto, Rifki Noor Rohman, & Muhammad Syaifuddin Alamsyah. (2025). SISTEM PREDIKSI HARGA KOMODITAS CABAI DIWILAYAH JAWA TIMUR MENGGUNAKAN SIMPLE MOVING AVERAGE. Jurnal Informatika Dan Teknik Elektro Terapan, 13(3). https://doi.org/10.23960/jitet.v13i3.7345

Desmarita, L., Muchlisinalahuddin, Maimuzar, Haris, & HEndra. (2023). Analisis *Heatmap* Korelasi dan Scatterplot untuk Mengidentifikasi Faktor-Faktor yang Mempengaruhi Pelabelan AC efisiensi Energi. Jurnal Rekayasa Material, Manufaktur Dan Energi, 6(1). https://doi.org/10.30596/rmme.v6i1.13133

Ermila, Ermila, Ghefira Zahra Nur Fadhilah, Ni Wayan Erdiani, and Rizal Adi Saputra. "Klasterisasi Pola Gejala Depresi Menggunakan Agglomerative Hierarchical *Cluster* Analysis Berdasarkan Skor Depresi PHQ-9." *Jurnal ilmiah Sistem Informasi dan Ilmu Komputer* 5, no. 2 (2025): 01-17. https://doi.org/10.55606/juisik.v5i2.993

Fitra Anisa, M., Saleh, Y. T., & Pratiwi, A. S. (2024). Pengaruh metode jarimatika terhadap kecepatan berhitung dan hasil belajar siswa pada materi perkalian kelas 68 IV di SD Muhammadiyah. Maret 2024 Griya Journal of Mathematics Education and Application, 4(1).

Ghiffary, Ghardapaty Ghaly, Kevin Alifviansyah, Anwar Fitrianto, Erfiani Erfiani, and LM Risman Dwi Jumansyah. Perbandingan Algoritma Hdbscan Dan

ISSN PRINT : 2302-500X
ISSN ONLINE : 2477-3786

Agglomerative Hierarchical *Clustering* Dalam Klasterisasi Pada Data Yang Mengandung Pencilan. *Jurnal Riset dan Aplikasi Matematika (JRAM)* 8, no. 2 (2024): 122-135.

Hendra Wijaya, Dandy Pramana Hostiadi, and Evi Triandini. "Meningkatkan Prediksi Penjualan Retail Xyz Dengan Teknik Optimasi Random Search Pada Model Xgboost." In *Seminar Hasil Penelitian Informatika dan Komputer (SPINTER)| Institut Teknologi dan Bisnis STIKOM Bali*, pp. 829-833. 2024.

Hibatulloh, Muh Naufal, Gilang Danu Prakoso, Adiestiana Dwi Putri Yunus, and Tommy Dwi Putra. "Prediksi Harga Rumah di Bandung 2024 Menggunakan Ensemble Learning: Analisis Komparatif dan Interpretabilitas." *Jurnal Informatika: Jurnal Pengembangan IT* 10, no. 2 (2025): 484-493. https://doi.org/10.30591/jpit.v10i2.8200

Ishak, Rezqiwati, Nurmawanti Nurmawanti, and Amiruddin Bengnga. "Optimization of *K-Means* in Disease *Clustering* of Pregnant Women Using Random Forest." *Jambura Journal of Electrical and Electronics Engineering* 7, no. 1 (2025): 41-47.

J Yulis, N., Alif Anhar, M., & Rombe, A. S. (2025). Analisis Perbandingan Peramalan Penggunaan Bahan Baku Menggunakan Metode Weighted Moving Average (WMA) dan Evaluasi dengan *Mean Absolute Error* (MAE)

Jailani, A. K., & Erna, A. (2025). Deteksi Anomali pada Rasio Jam Belajar dan Aktivitas Sosial terhadapPerforma Akademis Mahasiswa menggunakan Metode Local *Outlier* Factor (LOF): Vol. XIV (Issue 1).

Jasril, J., Al Fiqri, M. F., Sanjaya, S., Handayani, L., & Insani, F. (2025). Pengelompokan Data Kondisi Mesin Screw Press Menggunakan Algoritma Fuzzy C-Means. Information System Journal, 8(01), 60–70. https://doi.org/10.24076/infosjournal.2025v8i01.2133

Kusnaidi, Muhammad Rafli, Timotius Gulo, and Soeb Aripin. "Penerapan normalisasi data dalam mengelompokkan data mahasiswa dengan menggunakan metode *K-Means* untuk menentukan prioritas bantuan uang kuliah tunggal." *Journal of Computer System and Informatics (JoSYC)* 3, no. 4 (2022): 330-338.

Laila Hasanah, N., Ahadi Ningrum, A., & Kamarudin. (2025). Penerapan *K-Means Clustering* Menentukan Lokasi Promosi Penerimaan Mahasiswa Baru. Universitas Muhammadiyah Banjarmasin.

Mumpuni Enggar and Harison. “Implementasi Shap Pada Catboost Untuk Meningkatkan Akurasi Prediksi Temperatur Udara Di Kota Pekanbaru.” Journal of Multidisciplinary Studies 8 No 6 (2024).

Nurhaswinda, Pratasya, M., Laila Qurrotun, N., Tri Nanda, R., Neftihanan, Harnida, S., Pratiwi, U., Mauluddin, A., Taskia, S., & Alpenita, V. (2025). Penelitian Korelasi

Pandi, Eunike Cantika Kusuma, and Iqbal Kharisudin. "Pemodelan CEEMDAN LSTM, CEEMDAN GRU dan Metode Dasarnya untuk Peramalan Emisi Karbon Dioksida di Indonesia." *MATHunesa: Jurnal Ilmiah Matematika* 13, no. 1 (2025): 25-33. https://doi.org/10.26740/mathunesa.v13n1.p25-33

Pratiwi, Devani, Sisiliana Sisiliana, Edy Suprayetno, Ulfah Setia Iswara, and Dewi Mahrani Rangkuty. "Studi kajian tingkat kemiskinan di Kota Medan." *Student Research Journal* 1, no. 4 (2023): 142-150. https://repository.stiesia.ac.id/id/eprint/6506/

Srirahmawati, Eneng Okta, Ade Irma Purnamasari, Agus Bahtiar, and Edi Tohidi. "PENGELOMPOKAN PRESTASI AKADEMIK SISWA SD MENGGUNAKAN ALGORITMA *K-MEANS*." *Jurnal Informatika dan Rekayasa Elektronik* 8, no. 1 (2025): 81-86.

Syalsabilla, Alya Fitri, and Agus Fachrur Rozy. "Klasifikasi Tingkat Sengketa Pemilu 2024 di Indonesia Menggunakan Metode *Gaussian Mixture Model*." *Prosiding PITNAS Widyaiswara* 1 (2024): 404-416.

Togatorop, Tita Larose. "Analisis Pengaruh Learning Rate dalam Menentukan Mean Square Error (Mse) pada Algoritma Long Short Term Memory (LSTM)." (2024).