JuSIN
Jurnal Sistem Informasi

Website: ojs.itb-ad.ac.id/index.php/JUSIN
E-ISSN: 2797-8516

ANALISIS FAKTOR SOSIAL EKONOMI YANG MEMPENGARUHI RENDAHNYA CAPAIAN PENDIDIKAN DI INDONESIA MENGGUNAKAN KOMBINASI METODE DATA MINING

Diana Yusuf[1(*)], Fahrul Razi

[1]*ITB Ahmad Dahlan, Jakarta*
[2]*ITB Ahmad Dahlan, Jakarta*

Abstract

Educational inequality remains a persistent issue in Indonesia, particularly in regions with challenging socio-economic conditions. This study aims to analyze how various socio-economic factors influence the average years of schooling across Indonesian provinces using a combination of K-Means Clustering and Decision Tree algorithms. The dataset includes indicators such as poverty rate, gross regional domestic product (GRDP), per capita expenditure, and life expectancy, obtained from official national statistics.
K-Means Clustering was employed to segment provinces into three distinct groups based on socio-economic similarities. The clustering revealed clear disparities among regions, where the most disadvantaged cluster showed significantly lower education levels. Subsequently, the Decision Tree algorithm was used to classify the average years of schooling, identifying per capita expenditure, life expectancy, and socio-economic cluster as the most influential variables.
The combined approach allows for both segmentation and interpretation, providing data-driven insights that are accessible and actionable for policymakers. The findings highlight the importance of targeting socio-economic improvements as a strategy to enhance educational outcomes. Ultimately, this study underscores the value of integrating unsupervised and supervised machine learning models in addressing complex social issues in education.

Kata Kunci: *Education inequality, Socio-economic factors, K-Means Clustering , Decision Tree, Data-driven policy*

Informasi Artikel:
Dikirim : 09 Juni 2025
Ditelaah : 10 Juni 2025
Diterima : 22 Juni 2025
Publikasi : 25 Juni 2025

Januari – Juni 2025, Vol 6 (1) : hlm 51-62

(*) Korespondensi: *dianayusuf01@gmail.com (Diana Yusuf)*

PENDAHULUAN

Pendidikan merupakan pilar utama dalam pembangunan suatu bangsa, berperan penting dalam meningkatkan kualitas sumber daya manusia dan mendorong pertumbuhan ekonomi. Di Indonesia, meskipun telah terjadi peningkatan akses pendidikan, tantangan dalam hal kualitas dan pemerataan masih menjadi isu krusial. Data dari Badan Pusat Statistik (BPS) menunjukkan bahwa rata-rata lama sekolah penduduk Indonesia pada tahun 2023 adalah 8,9 tahun, yang artinya banyak individu yang belum menyelesaikan pendidikan menengah atas (BPS, 2023).

Ketimpangan pendidikan antara wilayah barat dan timur Indonesia juga masih nyata. Studi oleh (Purwanti, 2022) mengungkapkan bahwa wilayah Indonesia bagian timur memiliki tingkat pendidikan yang lebih rendah dibandingkan dengan wilayah barat, yang berkorelasi dengan kondisi sosial ekonomi yang kurang menguntungkan.

Salah satu faktor utama yang mempengaruhi capaian pendidikan adalah status sosial ekonomi (SES) keluarga. Penelitian oleh (Royani & Pertiwi, 2022) menunjukkan bahwa anak-anak dari keluarga SES rendah memiliki minat yang lebih rendah untuk melanjutkan pendidikan ke jenjang yang lebih tinggi. Selain itu, faktor-faktor seperti pendapatan orang tua, tingkat pendidikan orang tua, dan lingkungan tempat tinggal turut berkontribusi terhadap kesenjangan dalam akses dan kualitas pendidikan.

Memahami pengaruh faktor sosial ekonomi terhadap lama sekolah sangat penting untuk merumuskan kebijakan pendidikan yang lebih inklusif dan efektif. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor sosial ekonomi yang paling berpengaruh terhadap lama sekolah di Indonesia. Dengan demikian, hasil penelitian ini dapat menjadi dasar bagi pemerintah dan pemangku kepentingan dalam merancang intervensi yang tepat sasaran untuk meningkatkan kualitas dan pemerataan pendidikan.

Penelitian mengenai pengaruh faktor sosial ekonomi terhadap capaian pendidikan telah banyak dilakukan dalam beberapa tahun terakhir. Salah satu studi menunjukkan bahwa anak-anak dari keluarga berstatus rendah memiliki kecenderungan lebih rendah untuk melanjutkan pendidikan ke jenjang yang lebih tinggi dibandingkan anak dari keluarga berpendapatan tinggi. Hal ini menunjukkan bahwa SES merupakan determinan penting dalam keberlanjutan pendidikan (Nurwati & Listari, 2021). Penelitian lainnya oleh (Yusuf et al., 2024)menganalisis pengaruh rata-rata lama sekolah, garis kemiskinan, dan usia harapan hidup terhadap Indeks Pembangunan Manusia (IPM) di Provinsi Jawa Tengah. Hasilnya menujukkan bahwa rata-rata lama sekolah memiliki kontribusi siginifikan terhadap peningkatan kualitas hidup masyarakat, yang juga menunjukkan pentingnya peran pendidikan dalam pembangunan manusia. Sementara itu, penelitian lain tentang keterkaitan antara tingkat pengangguran terbuka, rata-rata lama sekolah, dan tingkat kemiskinan di Indonesia. Mereka menyimpulkan bahwa pendidikan dapat berfungsi sebagai alat pengurang kemiskinan, tetapi dampaknya sangat tergantung pada pemerataan akses pendidikan dan kesiapan wilayah dalam menyerap lulusan sekolah (Rahim et al., 2024)

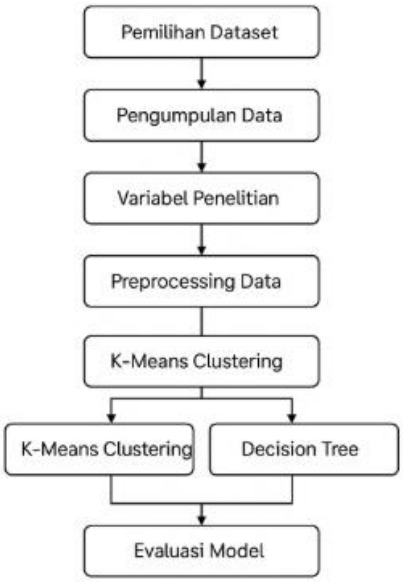
Meskipun telah banyak penelitian mengenai pengaruh status sosial ekonomi terhadap pendidikan, masih terdapat kesenjangan dalam hal penggunaan metode analisis yang lebih

canggih, seperti data mining. Pendekatan ini dapat mengungkapkan pola-pola tersembunyi dalam data yang mungkin tidak terdeteksi oleh metode konvensional. Oleh karena itum penelitian ini menggunakan metode *K-Means Clustering* dan *Decision Tree* untuk menganalisis pengaruh faktor sosial ekonomi terhadap lama sekolah di Indonesia.

Penelitian ini bertujuan untuk menganalisis dan mengidentifikasi faktor-faktor sosial ekonomi yang paling berpengaruh terhadap tingkat lama sekolah di Indonesia. Penelitian ini memanfaatkan pendekatan data mining melalui dua metode yaitu *K-Means Clustering* dan *Decision Tree*, untuk memperoleh pemahaman yang lebih mendalam terkait pola-pola yang tersembunyi dalam data pendidikan. Melalui metode *K-Means*, wilayah-wilayah di Indonesia dikelompokkan berdasarkan karakteristik sosial ekonomi tertentu untuk mengungkapkan potensi ketimpangan dan klaster resiko pendidikan. Sementara itu, metode *Decision Tree* digunakan untuk memprediksi kategori lama sekolah (rendah, sedang, tinggi) berdasarkan variabel sosial ekonomi seperti tingkat kemiskinan, PDRB, dan pengeluaran per kapita. Dengan pendekatan ini, hasil penelitian diharapkan dapat memberikan kontribusi nyata sebagai dasar pengambilan kebijakan yang lebih tepat sasaran dalam upaya peningkatan kualitas dan pemerataan pendidikan di Indonesia.

METODE

Penelitian ini menggunakan pendekatan kuantitatif eksploratif dengan penerapan metode data mining. Tujuan dari pendekatan ini adalah untuk mengidentifikasi pola dan relasi tersembunyi antara faktor-faktor sosial ekonomi dengan tingkat lama sekolah di Indonesia. Dua metode utama yang digunakan dalam analisis adalah *K-Means Clustering* dan *Decision Tree*.



Gambar 1. Diagram Alur Metode Penelitian

Sumber: analisis data, 2016

Pengumpulan Data

Jenis data yang digunakan adalah data sekunder yang diperoleh dari platform publik, yang memuat informasi indikator sosial ekonomi provinsi di Indonesia, seperti tingkat

kemiskinan, PDRB per kapita, pengeluaran per kapita, dan angka rata-rata lama sekolah. Data mencakup observasi per wilayah administratif dan dikumpulkan dalam format .xlsx. Data yang dikumpulkan melalui pengunduhan langsung dari sumber terpercaya. Validitas data dikonfirmasi melalui dokumentasi sumber dan metadata. Seluruh data kemudian direkapitulasi dalam *Microsoft Excel* sebelum diolah menggunakan *Python*.

Variabel Penelitian

Variabel-variabel yang digunakan dalam penelitian ini antara lain rata-rata lama sekolah sebagai variabel target/dependen. Tingkat kemiskinan (%), PDRB per Kapita, dan Pengeluaran per Kapita sebagai variabel input/independen.

Preprocessing Data

Tahapan preprocessing dari data yang dikumpulkan mencakup sebagai berikut: (1) Penghapusan nilai kosong (missing value), (2) Pemeriksaan duplikasi data, (3) Normalisasi variabel numerik untuk *K-Means* menggunakan metode Min-Max *Scalling*, (4) Kategorisasi variabel target (rata-rata lama sekolah) untuk *Decision Tree* ke dalam tiga kelas: rendah (<7), sedang (7-9), dan tinggi (>9), serta (5) Pembagian data menjadi data latih (80%) dan data uji (20%).

Klasterisasi dengan *K-Means*

Tahapan ini dilakukan proses klasterisasi wilayah menggunakan algoritma *K-Means*. Tujuannya adalah untuk mengelompokkan wilayah berdasarkan kemiripan kondisi sosial ekonominya. Proses diawali dengan menentukan jumlah klaster optimal menggunakan metode elbow, lalu dilanjutkan dengan penerapan *K-Means*. Berikut adalah algoritma *K-Means*:

1. **Menentukan jumlah klaster (k) yang akan digunakan**. Penentuan nilai k dapat dilakukan secara eksploratif. Salah satu menentukan nilai k ialah menggunakan **metode elbow** yakni dengan mencari titik siku pada grafik antara jumlah klaster dan nilai intertia (jumlah kuadrat jarak dalam cluster).
2. **Inisialisasi titik pusat (centroid awal).** Dari data yang ada algoritma secara acak akan memilih K titik pusat awal (centroid). Titik pusat ini berfungsi sebagai representasi awal dari masing-masing klaster.
3. **Menghitung jarak setiap data ke centroid.** Setiap data akan dihitung jaraknya ke seluruh centroid menggunakan rumus Euclidean Distance sebagai berikut:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

4. **Mengelompokkan data ke klaster terdekat.** Setiap data akan dikelompokkan ke dalam cluster yang memiliki centroid terdekat berdasarkan perhitungan jarak pada langkah sebelumnya.

5. **Menghitung ulang titik pusat (update centroid).** Setelah semua data dikelompokkan, algoritma akan menghitung ulang posisi centroid masing-masing klaster berdasarkan rata-rata posisi data dalam klaster tersebut. Berikut rumusnya:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \epsilon C_k} x_i$$

Evaluasi Klasterisasi

Setelah klasterisasi dilakukan, kualitas hasil klaster dievaluasi menggunakan Silhoutte Score, yang mengukur seberapa baik suatu objek cocok dengan klasternya sendiri dibandingkan klaster lain.

Klasifikasi dengan *Decision Tree*

Setelah pengelompokkan wilayah dengan clustering, dilakukan proses klasifikasi menggunakan *Decision Tree*. Metode ini digunakan untuk memprediksi kategori rata-rata lama sekolah berdasarkan faktor-faktor sosial ekonomi. Berikut langkah-langah algoritma *Decision Tree*:

1. **Menentukan atribut terbaik sebagai node akar/awal.** Algoritma memulai dengan memilih atribut terbaik yang paling mampu membedakan kelas target. Untuk menentukan atribut akar terbaik ialah dengan mencari nilai gain tertinggi dan nilai gain diperoleh dari nilai entropy setiap atribut.
   Rumus menentukan nilai entropy:

$$Entropy(s) = -\sum_{i=1}^{n} p_i Log_2(p_i)$$

   Rumus menentukan nilai gain:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} . Entropy(S_v)$$

2. **Membagi data berdasarkan atribut terpilih.** Dataset dibagi ke dalam cabang-cabang berdasarkan nilai dari atribut terpilih. Setiap cabang merepresentasikan subset data dengan satu nilai dari atribut tersebut.
3. **Ulangi proses pada setiap subset.** Ulangi proses pemilihan terbaik untuk setiap subset, pembagian daya, dan perhitungan entropy dan gain. Proses ini terus dilakukan hingga: (a) Semua data dalam satu subset berasal dari kelas yang sama; (b) Tidak ada atribut tersisa; (c) Sudah mencapai kedalaman maksimal.
4. **Membentuk daun**. Jika suatu subset sudah homogen, maka subset tersebut menjadi daun (leaf node) yang menyimpan label kelas hasil prediksi.
5. **Membentuk aturan keputusan.** Dari struktur pohon yang terbentuk dibuatlah aturan dalam bentuk IF-THEN.

Evaluasi Model Klasifikasi

Model *Decision Tree* yang telah dibangun diuji akurasinya menggunakan data uji. Evaluasi dilakukan dengan menghitung akurasi, precision, recall, serta confusion matrix untuk mengetahui seberapa tepat model dalam memprediksi kategori pendidikan (rendah, sedang, tinggi).

Interpretasi Hasil dan Penarikan Kesimpulan

Tahapan terakhir dari metode penelitian ini adalah menafsirkan hasil klaterisasi dan klasifikasi yang telah dilakukan.

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data sekunder yang diperoleh dari platform publik yang berisi indikator-indikator sosial ekonomi dari berbagai kabupaten dan kota di Indonesia. Data ini mencerminkan kondisi riil pembangunan manusia dan ketimpangan sosial di tingkat daerah. Fokus penelitian adalah menganalisis bagaimana faktor-faktor ekonomi dan sosial mempengaruhi rata-rata lama sekolah, sebagai salah satu indikator penting dalam pengukuran kualitas pendidikan. Data mencakup beberapa variabel utama seperti provinsi, kota/kabupaten, penduduk miskin (%), PDRB (miliar rupiah), angka harapan hidup (tahun), pengeluaran per kapita (ribu rupiah), dan rata-rata lama sekolah. Data tersebut kemudian akan diolah menggunakan teknik data mining. Namun sebelum diolah data tersebut akan dibersihkan terlebih dahulu di tahapan preprocessing agar bisa dianalisis menggunakan pendekatan *K-Means Clustering* dan *Decision Tree*.

Tabel 1. Dataset Mentah

Provinsi	Kota/Kabupaten	Penduduk Miskin (%)	PDRB	AHH	Pengeluaran	Lama Sekolah
Aceh	Simeulue	18.98	2275.34	65.24	7148.18	6.74
Aceh	Aceh Singkil	20.36	2425.27	67.36	8776.08	7.69
Aceh	Aceh Selatan	17.30	4555.21	68.92	9331.22	8.15
Aceh	Aceh Tenggara	14.99	4285.98	67.33	8570.50	8.61
Aceh	Aceh Timur	17.82	10666.83	68.22	10146.21	8.09
....	.....	.....	.....	.....	.....	.....

Sebelum data dianalisis menggunakan metode *K-Means Clustering* dan *Decision Tree*, diperlukan tahapan penting yakni preprocessing data. Tahapan ini bertujuan untuk membersikahkan, menyamakan skala, dan mempersiapkan data agar dapat diproses secara optimal oleh algoritma.

1. **Pembersihan Data (Cleaning).** Langkah pertama ialah memeriksa apakah terdapat missing value pada data atau data tidak wajar yang dapat mempengaruhi hasil analisis. Pada dataset ini tidak ditemukan nilai yang kosong, namun dilakukan pengecekan ulang agar semua data numerik valid dan dapat dihitung dengan benar.

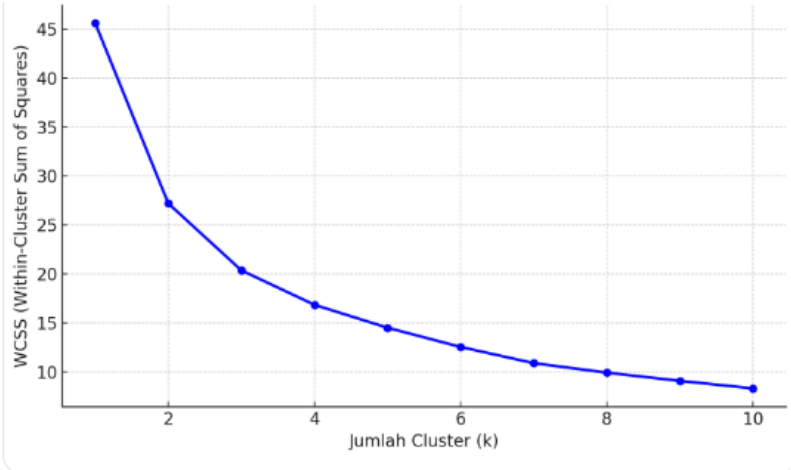
2. **Normalisasi Data.** Agar setiap variabel numerik memiliki skala yang setara, digunakan metode min-max normalization, yaitu mengubah setiap nilai menjadi rentang antara 0 dan 1. Ini penting khususnya untuk metode *K-Means*, karena algoritma ini sangat sensitif terhadap skala antar fitur. Berikut contoh hasil data yang telah dinormalisasi dengan rumus Min-Max:

Tabel 2. Data Hasil Normalisasi

Kota/Kabupaten	Penduduk Miskin (%)	PDRB	AHH	Pengeluaran
Simeulue	0.4226	0.0015	0.4389	0.1593
Aceh Singkil	0.4577	0.0017	0.5330	0.2410
Aceh Selatan	0.3686	0.0033	0.5892	0.2769
Aceh Tenggara	0.2535	0.0031	0.5315	0.2231
Aceh Timur	0.3993	0.0077	0.5708	0.3507
.....	.....	.....	.....	.....

Langkah berikutnya ialah melakukan kategorisasi variabel target untuk algoritma *Decision Tree*. Variabel yang akan dikategorisasi ialah variabel rata-rata lama sekolah yang akan dikategorisasi ke dalam tiga kelas yakni Rendah (< 7 tahun), Sedang (7-9 tahun), dan Tinggi (> 9 tahun). Variabel ini akan digunakan sebagai target output pada model klasifikasi, sehingga memungkinkan untuk mengevaluasi apakah faktor sosial ekonomi mampu memprediksi capaian pendidikan suatu daerah.

3. **Elbow Method.** Sebelum menerapkan algoritma *K-Means Clustering* , salah satu langkah penting ialah menentukan jumlah klaster (k) yang optimal. Penentuan nilai k sangat berpengaruh terhadap keakuratan hasil klasterisasi, sehingga nilai k akan ditentukan menggunakan pendekatan elbow method.



Gambar 2. Hasil Visualisasi Elbow Method

Berdasarkan grafik yang dihasilkan, terlihat bahwa penurunan nilai Within-Cluster Sum of Squares (WCSS) mulai melambat secara signifikan pada nilai k=3. Dengan demikian, jumlah klaster yang paling optimal untuk data ini adalah 3 (tiga) klaster atau k=3.

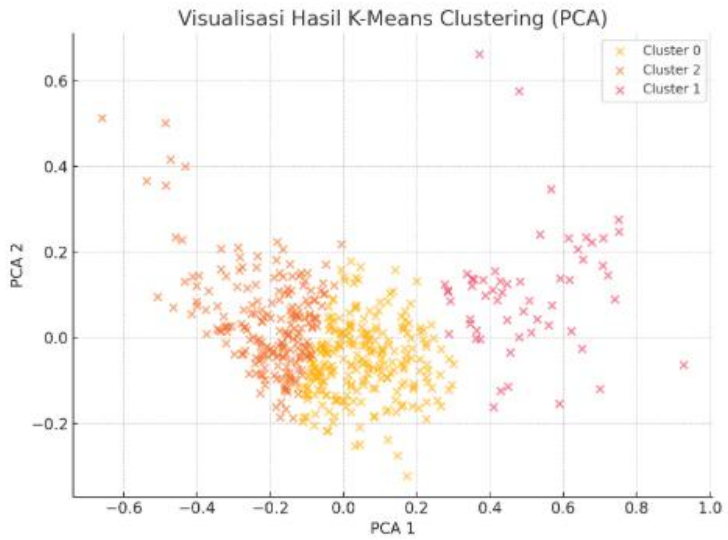
4. **Hasil Klasterisasi.** Setelah menentukan jumlah klaster dengan elbow method, proses klasterisasi dapat dilanjutkan menggunakan algoritma *K-Means*. Setiap wilayah diklasifikasikan ke dalam salah satu dari tiga klaster berdasarkan variabel-variabel sosial ekonomi yang telah dinormalisasi dan diproses. Berikut contoh struktur data hasil klasterisasi:

Provinis	Kota	Penduduk Miskin	PDRB	AHH	Pengeluaran	cluster	pca1	pca2
Aceh	Simeulue	0,422606925	0,00151	0,43896	0,15930092	1	0,29938	-0,06099
Aceh	Aceh Singkil	0,457739308	0,00169	0,53302	0,24106067	1	0,23931	0,03041
Aceh	Aceh Selatan	0,274949084	0,00549	0,39982	0,21112897	1	0,18789	-0,15936
Aceh	Aceh Tenggara	0,280804481	0,00492	0,5686	0,20359582	1	0,1112	-0,06224
Aceh	Aceh Timur	0,307281059	0,0117	0,59306	0,23106669	1	0,10447	-0,0232
Aceh	Aceh Tengah	0,32790224	0,00776	0,59684	0,3417035	1	0,06605	0,01221
Aceh	Aceh Barat	0,418279022	0,00865	0,55926	0,2820912	1	0,17781	0,03358
Aceh	Aceh Besar	0,29709776	0,01503	0,63932	0,28465247	1	0,04854	0,00964
Aceh	Pidie	0,438136456	0,01188	0,51279	0,2955002	1	0,20877	0,02374
Aceh	Bireuen	0,276731161	0,01473	0,70402	0,24563078	1	0,01977	0,02572
Aceh	Aceh Utara	0,38314664	0,0228	0,59529	0,21218361	1	0,16571	0,0257
Aceh	Aceh Barat Daya	0,355397149	0,00352	0,42895	0,22358377	1	0,22565	-0,09283
Aceh	Gayo Lues	0,439409369	0,00202	0,44941	0,24507834	1	0,26605	-0,02693
Aceh	Aceh Tamiang	0,279022403	0,00791	0,63131	0,22052029	1	0,07029	-0,02278
Aceh	Nagan Raya	0,403513238	0,00811	0,61574	0,21675372	1	0,16954	0,04281
Aceh	Aceh Jaya	0,276221996	0,00194	0,52346	0,28575733	1	0,09234	-0,07509
Aceh	Bener Meriah	0,427189409	0,00425	0,6153	0,35867818	1	0,12078	0,0842
Aceh	Pidie Jaya	0,437118126	0,00298	0,6571	0,31709522	1	0,1266	0,10456
Aceh	Kota Banda Aceh	0,13314664	0,02108	0,71603	0,64860386	0	-0,27904	0,03266
Aceh	Kota Sabang	0,329429735	0,0006	0,67312	0,37173564	1	0,01565	0,05873
Aceh	Kota Langsa	0,218431772	0,00518	0,6233	0,40633789	1	-0,05608	-0,02682
Aceh	Kota Lhokseumawe	0,223523422	0,00963	0,72226	0,37233829	0	-0,08649	0,02751

Gambar 3. Struktur Data Hasil Klasterisasi

Analisis hasil klaterisasi menghasilkan data diklasterisasi ke dalam 3 klaster, yang menunjukkan segmentasi wilayah berdasarkan kondisi sosial ekonomi yang serupa. Untuk mempermudah visualisasi, hasil klaterisasi juga telah direduksi dimensinya menggunkan PCA (Principal Component Analysis) ke dalam dua komponen utama yakni PCA1 dan PCA2. Karakteristik tiap klaster antara lain sebagai berikut (a) Cluster 0, Didominasi oleh wilayah dengan tingkat PDRB dan Pengeluaran per Kapita tinggi, serta angka harapan hidup tinggi; (b) Cluster 1, merupakan wilayah dengan indikator sosial ekonomi menengah; (c) Cluster 2, kelompok dengan tingkat kemiskinan lebih tinggi, PDRB lebih rendah, dan angka harapan hidup yang lebih rendah

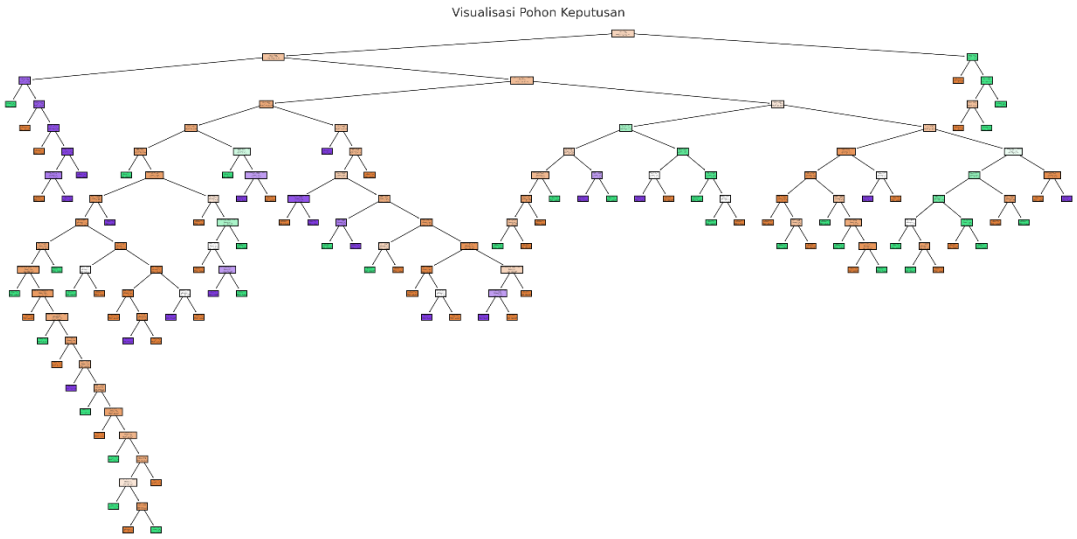
Setelah proses *K-Means Clustering* selesai, nilai cluster akan ditambahkan sebagai fitur baru untuk memperkuat analisis dalam *Decision Tree*. Fitur ini memberikan konteks kelompok sosial ekonomi tempat suatu daerah berada, dan dapat menjadi fitur tambahan yang memperkuat prediksi klasifikasi. Berikut visualisasi *K-Means Clustering* :



Gambar 4. Visualisasi Hasil *K-Means Clustering* (PCA)

5. **Evaluasi dengan Silhoutte Score**. Evaluasi klasterisasi perlu dilakukan untuk menilai seberapa baik hasil pengelompokkan (clustering) yang telah dilakukan oleh algoritma *K-Means*. Dalam penelitian ini, evaluasi dilakukan dengan Silhoutte Score dengan hasil jika k=3 yaitu 0.305. Nilai tersebut dianggap sebagai jumlah klaster yang cukup optimal yang berarti ketika data dibagi menjadi ke dalam 3 kelompok, objek-objek dalam setiap klaster memiliki kesamaan tinggi dan sangat berbeda dengan klaster lain.

6. **Hasil Klasifikasi *Decision Tree***. Tahapan berikutnya setelah klasterisasi ialah mengklasifikasikan kategori lama sekolah menggunakan algoritma *Decision Tree*. Proses klasifikasi dilakukan dengan memanfaatkan beberapa variabel prediktor yakni Provinsi, Kota, Penduduk Miskin, PDRB, Angka Harapan Hidup (AHH), Pengeluaran per Kapita, dan hasil cluster dari *K-Means*. Dataset dibagi menjadi dua bagian yakni 80% untuk data latih dan 20% untuk data uji. Berikut visualisasi klasifikasi *Decision Tree*:



Gambar 4.Visualisasi *Decision Tree*

Visualisasi struktur pohon keputusan menunjukkan bagaimana model membagi data berdasarkan atribut paling informatif hingga menentukan klasifikasi akhir. Hal ini memudahkan dalam memahami pola hubungan antar variabel dan pengaruhnya terhadap kategori lama sekolah.

	precision	recall	f1-score	support
Sedang	0.67	0.60	0.64	58
Tinggi	0.50	0.53	0.52	32
rendah	0.41	0.54	0.47	13
accuracy			0.57	103
macro avg	0.53	0.56	0.54	103
weighted avg	0.59	0.57	0.58	103

Gambar 5. Hasil Evaluasi Model Klasifikasi

Model *Decision Tree* yang dibangun menunjukkan akurasi 66%, yang tergolong cukup baik untuk model baseline.

Interpretasi Hasil

Berdasarkan hasil klasterisasi menggunakan algoritma *K-Means*, data wilayah di Indonesia dapat dikelompokkan ke dalam tiga klaster utama berdasarkan indikator sosial ekonomi, seperti **penduduk miskin, PDRB, pengeluaran per kapita, dan angka harapan hidup**. Klaster pertama umumnya berisi wilayah dengan tingkat kesejahteraan relatif tinggi, ditandai dengan pengeluaran per kapita dan angka harapan hidup yang tinggi serta persentase penduduk miskin yang rendah. Klaster kedua menggambarkan wilayah dengan karakteristik sedang, sementara klaster ketiga mewakili wilayah dengan kondisi sosial ekonomi yang kurang berkembang.

Pemanfaatan hasil klasterisasi ini tidak hanya membantu memetakan kondisi wilayah, tetapi juga berfungsi sebagai atribut tambahan dalam proses klasifikasi *Decision Tree*. Dengan memasukkan klaster sebagai fitur prediktor, model klasifikasi memperoleh pemahaman lebih kontekstual dalam memprediksi kategori **Lama Sekolah.** Hal ini terbukti dari hasil klasifikasi yang menunjukkan bahwa klaster memiliki kontribusi dalam menentukan output model, bersamaan dengan variabel-variabel lain seperti **Pengeluaran per Kapita dan Angka Harapan Hidup**.

Interpretasi terhadap model *Decision Tree* menunjukkan bahwa wilayah dengan pengeluaran per kapita dan angka harapan hidup tinggi secara konsisten diklasifikasikan ke dalam kategori Lama Sekolah Tinggi, yang mencerminkan hubungan erat antara kualitas hidup dan akses pendidikan. Sebaliknya, wilayah dengan tingkat kemiskinan tinggi, pengeluaran per kapita rendah, dan angka harapan hidup yang pendek umumnya masuk kategori Lama Sekolah Rendah.

Integrasi antara klasterisasi dan klasifikasi ini memberikan gambaran menyeluruh mengenai profil wilayah berdasarkan kondisi sosial ekonomi, sekaligus prediksi terhadap tingkat lama sekolah sebagai target pendidikan nasional. Dengan demikian, hasil ini dapat menjadi dasar yang kuat bagi pemerintah dalam menyusun kebijakan yang lebih terfokus dan berbasis data, misalnya melalui penyaluran bantuan pendidikan ke wilayah dengan

karakteristik klaster kurang berkembang atau intervensi sosial pada wilayah yang terklasifikasi dengan lama sekolah rendah.

KESIMPULAN

Penelitian ini bertujuan untuk menganalisis faktor sosial ekonomi terhadap tingkat lama sekolah di Indonesia dengan menggunakan pendekatan gabungan unsupervised learning (*K-Means Clustering*) dan supervised learning (*Decision Tree*). Data yang digunakan mencakup indikator sosial ekonomi dari berbagai wilayah di Indonesia, termasuk persentase penduduk miskin, PDRB, pengeluaran per kapita, dan angka harapan hidup. Hasil penelitian ini menyimpulkan bahwa faktor sosial ekonomi memiliki pengaruh yang nyata terhadap pencapaian pendidikan formal di Indonesia. Kedua metode yang digunakan saling melengkapi dalam memberikan pemahaman yang lebih dalam yakni *K-Means* membantu mengelompokkan kondisi wilayah secara objektif, sedangkan *Decision Tree* memberikan aturan klasifikasi yang transparan dan mudah dipahami.

Berdasarkan hasil penelitian ini, terdapat sejumlah rekomendasi kebijakan yang dapat dijadikan bahan pertimbangan untuk meningkatkan sektor pendidikan dan kesejahteraan masyarakat. Pertama, pemerintah perlu memberikan perhatian khusus pada wilayah-wilayah yang termasuk dalam Cluster 2, yaitu daerah dengan karakteristik pengeluaran rendah, tingkat kemiskinan tinggi, angka harapan hidup (AHH) rendah, dan capaian pendidikan yang juga rendah. Program-program yang disarankan meliputi pemberian beasiswa daerah, penyediaan sekolah gratis, serta bantuan ekonomi bagi keluarga kurang mampu. Kedua, mengingat pengeluaran per kapita memiliki pengaruh signifikan terhadap lama sekolah, maka peningkatan kesejahteraan ekonomi akan berdampak langsung pada sektor pendidikan. Oleh karena itu, program seperti bantuan sosial bersyarat (conditional cash transfer) dan program padat karya bagi keluarga miskin sangat dianjurkan. Ketiga, karena terdapat hubungan positif antara angka harapan hidup dan tingkat pendidikan, diperlukan integrasi program seperti sekolah sehat, posyandu terpadu, serta pendidikan gizi dan sanitasi sejak usia dini. Keempat, pendekatan ini juga dapat dimanfaatkan oleh pemerintah daerah maupun nasional sebagai alat untuk memantau ketimpangan pendidikan, mengevaluasi kebijakan pendidikan tahunan, serta merancang alokasi anggaran pendidikan yang berbasis wilayah. Terakhir, model yang digunakan perlu diperbarui setiap tahun dengan menggunakan data terbaru guna menyesuaikan dengan dinamika sosial dan ekonomi, memantau efektivitas kebijakan yang telah diterapkan, serta mengidentifikasi wilayah-wilayah rawan yang baru.

DAFTAR PUSTAKA

Nurwati, R.N. and Listari, Z.P. (2021) '**Pengaruh Status Sosial Ekonomi Keluarga Terhadap Pemenuhan Kebutuhan Pendidikan Anak**', *Share : Social Work Journal*, 11(1), p. 74. Available at: https://doi.org/10.24198/share.v11i1.33642.

Purwanti, Y. (2022) '**Disparitas fasilitas pendidikan dan tenaga pengajar sekolah menengah atas di Indonesia menggunakan Metode Spatial Fuzzy C-Means**', *Jurnal Pendidikan Dompet Dhuafa*, 12(02), pp. 15–22.

Rahim, A., Haryadi, W. and Muliawansyah, D. (2024) '**ANALISIS FAKTOR RATA-RATA LAMA SEKOLAH DAN PENGANGGURAN TERBUKA DALAM MEMPENGARUHI TINGKAT KEMISKINAN DI KABUPATEN SUMBAWA**', *Jurnal Ekonomi & Bisnis*, 12(1), pp. 14–25.

Royani, I. and Pertiwi, T.B. (2022) '**Pengaruh Status Sosial Ekonomi Terhadap Minat Melanjutkan Pendidikan Anak Usia 11–21 Tahun: Pengaruh Status Sosial Ekonomi Terhadap Minat Melanjutkan Pendidikan Anak Usia 11–21 Tahun**', *Journal Of Lifelong Learning*, 5(2), pp. 28–36.

Yusuf, D., Sestri, E. and Razi, F. (2024) '**PENGELOMPOKKAN DATA MAHASISWA MENGGUNAKAN CLUSTERING UNTUK OPTIMALISASI PENERIMAAN MAHASISWA BARU**', *JIKA (Jurnal Informatika)*, 8(4), pp. 484–490.

Yusuf, D., Sestri, E. and Razi, F. (2023) '**Implementasi Teknik Clustering Untuk Pengelompokan Mobil Bekas Berdasarkan Grade Pada Mobi Auto**', J-SISKO TECH (Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD), 6(2), pp. 307-313.

F. H. Kuwil, F. Shaar, A. E. Topcu, and F. Murtagh, "**A new data clustering algorithm based on critical distance methodology**," Expert Syst. Appl., vol. 129, pp. 296–310, 2019, doi: 10.1016/j.eswa.2019.03.051

N. T. Luchia, H. Handayani, F. S. Hamdi, D. Erlangga, and S. Fitri Octavia, "**Perbandingan *K-Means* dan K-Medoids Pada Pengelompokan Data Miskin di Indonesia**," MALCOM Indones. J. Mach. Learn. Comput. Sci., vol. 2, no. 2, pp. 35–41, 2022.

D. S. Maylawati, T. Priatna, H. Sugilar, and M. A. Ramdhani, "**Data science for digital culture improvement in higher education using *K-Means Clustering* and text analytics**," Int. J. Electr. Comput. Eng., vol. 10, no. 5, pp. 4569–4580, 2020, doi: 10.11591/IJECE.V10I5.PP4569-4580.

A. Cahyadi, Hendryadi, S. Widyastuti, V. N. Mufidah, and Achmadi, â€œ**Emergency remote teaching evaluation of the higher education in Indonesia**,â€ Heliyon, vol. 7, no. 8, 2021, doi: 10.1016/j.heliyon.2021.e07788.

H. Annur, "**Klasifikasi MasyarakatMiskin Menggunakan Metode Naïve Bayes**," Ilk. J. Ilm., vol. 10, no. 2, pp.160–165, 2018.

F. Aris and Benyamin, "**Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi**," J. Sist. Komput. dan Sist. Inf., vol. 1, no. 1, pp. 1–6, 2019.

Asroni, B. M. Respati, and S. Riyadi, "**Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta**," J. Ilm. Fak. Tek. Univ. Muhammadiyah, vol. 21, no. 2, pp. 158–165, 2018.