LAPORAN TUGAS BESAR DATA MINING

"Penerapan Algoritma K-Means Clustering Untuk Mengidentifikasi Keterkaitan Antara Putus Sekolah Pada Tingkat SMA dan Jumlah Kemiskinan Tiap Provinsi di Indonesia"

Dosen Pengampu :

Dwi Welly Sukma Nirad, M.T.

Aina Hubby Aziira, M.Eng

Oleh :

Kelompok 2

Fahri Zamzami	2311521014
Nayla Nurul Afifah	2311522002
Ahmad Rasha Radya Aufa Lubis	2311522008
Zakky Aulia Aldrin	2311522018
Muhammad Rafi Asytar	2311522030

DEPARTEMEN SISTEM INFORMASI
FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS ANDALAS

2025

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga kami, Kelompok 2, dapat menyelesaikan laporan tugas besar mata kuliah Data Mining yang berjudul *"Penerapan Algoritma K-Means Clustering untuk Mengidentifikasi Keterkaitan Antara Putus Sekolah pada Tingkat SMA dan Jumlah Kemiskinan Tiap Provinsi di Indonesia"*.

Laporan ini disusun sebagai bentuk penerapan pemahaman kami terhadap konsep dan teknik data mining, khususnya penggunaan algoritma *K-Means Clustering* dalam menganalisis data sosial. Dengan topik ini, kami berusaha mengidentifikasi pola keterkaitan antara angka putus sekolah di tingkat SMA dan jumlah kemiskinan di setiap provinsi di Indonesia. Harapannya, hasil analisis ini dapat memberikan gambaran awal yang bermanfaat bagi pihak-pihak yang berkepentingan dalam perumusan kebijakan sosial.

Kami mengucapkan terima kasih yang sebesar-besarnya kepada Ibu Dwi Welly Sukma Nirad, M.T. dan Ibu Aina Hubby Aziira, M.Eng. selaku dosen pengampu mata kuliah Data Mining, yang telah memberikan ilmu, arahan, serta bimbingan selama proses perkuliahan hingga tersusunnya laporan ini. Ucapan terima kasih juga kami sampaikan kepada seluruh anggota Kelompok 2 yang telah bekerja sama dengan penuh semangat, serta kepada semua pihak yang telah mendukung baik secara langsung maupun tidak langsung.

Kami menyadari bahwa laporan ini masih memiliki kekurangan, baik dari segi isi maupun penyajiannya. Oleh karena itu, kami sangat terbuka terhadap kritik dan saran yang bersifat membangun demi penyempurnaan di masa yang akan datang. Semoga laporan ini dapat memberikan manfaat serta menjadi tambahan referensi dalam pengembangan analisis data pada permasalahan sosial di Indonesia.

Padang, 15 Juni 2025

Penyusun,
Kelompok 2

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendidikan merupakan fondasi utama dalam pembangunan suatu negara. Kualitas sumber daya manusia sangat ditentukan oleh sejauh mana masyarakat dapat mengakses dan menyelesaikan pendidikan secara merata dan berkelanjutan. Namun, di berbagai wilayah di Indonesia, ketimpangan dalam pendidikan masih menjadi permasalahan yang belum terselesaikan secara optimal. Salah satu indikator yang mencerminkan ketimpangan tersebut adalah angka putus sekolah, khususnya pada jenjang Sekolah Menengah Atas (SMA).

Putus sekolah adalah kondisi di mana peserta didik tidak menyelesaikan jenjang pendidikan yang sedang dijalani, baik karena alasan ekonomi, sosial, maupun budaya. Permasalahan ini berdampak jangka panjang terhadap kehidupan individu, seperti terbatasnya akses pekerjaan, rendahnya kualitas hidup, serta potensi memperbesar lingkaran kemiskinan. Fenomena ini banyak dijumpai pada daerah-daerah dengan kondisi ekonomi masyarakat yang masih tergolong rendah.

Kemiskinan menjadi salah satu faktor utama penyebab siswa putus sekolah. Keluarga dengan penghasilan terbatas sering kali harus memprioritaskan kebutuhan dasar sehari-hari dibandingkan pendidikan. Dalam banyak kasus, anak-anak dari keluarga miskin terpaksa berhenti sekolah dan ikut bekerja untuk membantu ekonomi keluarga. Hal ini menunjukkan adanya kemungkinan korelasi yang kuat antara tingkat kemiskinan dengan angka putus sekolah.

Untuk melihat hubungan tersebut secara objektif dan berbasis data, diperlukan analisis yang tepat. Salah satu pendekatan yang dapat digunakan adalah *data mining*, yaitu proses penggalian informasi atau pola dari kumpulan data besar. Metode *K-Means Clustering* dalam *data mining* sangat berguna untuk mengelompokkan objek berdasarkan kemiripan data yang dimiliki.

Dalam tugas besar ini, kami menerapkan algoritma *K-Means Clustering* untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan data angka putus sekolah tingkat SMA dan jumlah penduduk miskin. Dataset yang digunakan merupakan data tahun 2024, yang mencakup 39 provinsi di Indonesia. Melalui analisis ini, diharapkan dapat ditemukan pola keterkaitan antara kemiskinan dan pendidikan, serta gambaran klaster provinsi yang memiliki karakteristik serupa.

1.2 Rumusan Masalah

Berikut adalah beberapa rumusan masalah yang dapat dijelaskan dalam laporan penelitian ini:

1. Bagaimana penerapan algoritma *K-Means Clustering* dalam mengelompokkan provinsi di Indonesia berdasarkan angka putus sekolah tingkat SMA dan jumlah penduduk miskin?
2. Apa saja kelompok atau klaster yang terbentuk berdasarkan data tersebut?
3. Apakah terdapat pola keterkaitan yang signifikan antara tingkat kemiskinan dan angka putus sekolah berdasarkan hasil klastering?

1.3 Batasan Masalah

Agar analisis yang dilakukan lebih terarah dan fokus, maka ruang lingkup atau batasan dari tugas besar ini ditetapkan sebagai berikut:

1. Data yang digunakan terbatas pada angka putus sekolah jenjang SMA/sederajat dan jumlah penduduk miskin per provinsi di Indonesia tahun 2024.
2. Teknik analisis yang digunakan adalah algoritma *K-Means Clustering* tanpa membandingkan dengan metode *clustering* lainnya.
3. Data yang dianalisis bersifat statis (bukan time series).
4. Hasil pengelompokan tidak dimaksudkan untuk membuat generalisasi menyeluruh, melainkan untuk eksplorasi dan visualisasi pola yang muncul.
5. Analisis ini tidak mengkaji hubungan kausal secara langsung, melainkan bersifat eksploratif melalui pendekatan *unsupervised learning*.

1.4 Tujuan

Adapun tujuan dari penyusunan tugas besar ini adalah sebagai berikut:

1. Menerapkan algoritma *K-Means Clustering* pada data sosial untuk mengelompokkan provinsi berdasarkan angka putus sekolah dan tingkat kemiskinan tahun 2024.
2. Menemukan klaster atau kelompok provinsi yang memiliki karakteristik serupa dalam hal angka putus sekolah dan jumlah penduduk miskin.
3. Memberikan gambaran awal mengenai keterkaitan antara pendidikan dan kemiskinan sebagai bahan evaluasi dan pertimbangan dalam penyusunan kebijakan.

BAB II
LITERATUR

2.1 Data Mining

Data mining adalah proses untuk mengekstraksi pengetahuan yang berguna dan tersembunyi dari kumpulan data yang besar dan kompleks. Proses ini melibatkan analisis otomatis atau semi-otomatis terhadap data untuk menemukan pola, korelasi, tren, atau hubungan yang sebelumnya tidak diketahui.

Data mining merupakan inti dari proses *Knowledge Discovery in Databases (KDD)*, yang mencakup serangkaian tahapan mulai dari pembersihan data, integrasi data, seleksi, transformasi, hingga penambangan informasi dan interpretasi hasil. Tujuannya adalah mengubah data mentah menjadi pengetahuan yang bermakna dan dapat digunakan untuk pengambilan keputusan atau pemecahan masalah.

Dalam praktiknya, data mining sering diterapkan dalam berbagai bidang seperti bisnis, pendidikan, kesehatan, dan pemerintahan, karena kemampuannya untuk mengolah data dalam jumlah besar dan mengungkap informasi yang tidak tampak secara langsung melalui analisis manual biasa.

2.2 K-Means

K-Means adalah algoritma clustering yang digunakan untuk membagi sekumpulan data ke dalam beberapa kelompok atau klaster berdasarkan kemiripan karakteristik. Algoritma ini bekerja secara iteratif dengan cara menentukan sejumlah klaster awal (*k*), memilih titik pusat klaster (*centroid*) secara acak, lalu mengelompokkan setiap data ke centroid terdekat menggunakan ukuran jarak, seperti Euclidean Distance. Setelah semua data dikelompokkan, centroid diperbarui dengan menghitung rata-rata posisi data dalam setiap klaster. Proses ini terus diulang hingga posisi centroid tidak berubah lagi secara signifikan atau mencapai jumlah iterasi maksimum.

K-Means banyak diaplikasikan dalam berbagai bidang seperti pengenalan pola, pengolahan citra, segmentasi pasar, dan analisis sosial karena efisiensinya dalam menangani dataset berukuran besar dan kemampuannya menghasilkan hasil pengelompokan yang cepat dan cukup akurat untuk data yang memiliki distribusi yang jelas.

2.3 Pendidikan

Pendidikan memiliki peran sentral dalam membentuk sumber daya manusia yang berkualitas, baik dalam aspek pengetahuan, keterampilan, maupun sikap. Melalui pendidikan, individu dapat mengembangkan kemampuan berpikir kritis, menyelesaikan masalah, serta beradaptasi dengan dinamika sosial dan perubahan

ekonomi. Pendidikan juga membuka akses yang lebih luas terhadap peluang kerja dan sumber penghidupan yang layak. Semakin tinggi jenjang dan mutu pendidikan yang diperoleh, semakin besar pula peluang individu untuk memperoleh penghasilan yang lebih baik dan keluar dari kondisi ekonomi yang terbatas.

Dampak positif pendidikan tidak hanya dirasakan secara individual, tetapi juga menimbulkan efek berkelanjutan dalam lingkungan keluarga dan masyarakat, karena generasi berikutnya cenderung memiliki akses pendidikan yang lebih baik dan kesempatan hidup yang lebih sejahtera.

2.4 Putus Sekolah

Putus sekolah adalah kondisi ketika peserta didik menghentikan proses belajar sebelum menyelesaikan jenjang pendidikan yang seharusnya. Fenomena ini menjadi indikator rendahnya keberhasilan sistem pendidikan, serta mencerminkan masih adanya hambatan dalam mempertahankan partisipasi siswa. Putus sekolah juga menunjukkan belum optimalnya upaya dalam menjamin kelangsungan pendidikan bagi seluruh warga negara. Selain berdampak pada perkembangan individu, masalah ini turut memengaruhi kualitas sumber daya manusia secara umum karena mengurangi jumlah lulusan yang dapat melanjutkan ke tingkat pendidikan yang lebih tinggi.

2.5 Kemiskinan

Kemiskinan merupakan keadaan di mana seseorang atau suatu keluarga tidak mampu memenuhi kebutuhan dasar yang diperlukan untuk hidup layak, seperti makanan, tempat tinggal, pendidikan, dan layanan kesehatan. Keadaan ini tidak hanya mencerminkan kekurangan secara ekonomi, tetapi juga menunjukkan keterbatasan dalam akses terhadap berbagai layanan publik.Kemiskinan bersifat multidimensi dan memiliki dampak luas terhadap kehidupan sosial masyarakat, termasuk keterbatasan dalam pengembangan diri, partisipasi sosial, dan perlindungan terhadap risiko kehidupan sehari-hari.

2.6 Penelitian Terdahulu

BAB III
METODOLOGI

3.1 Objek Penelitian

Objek penelitian dalam tugas besar ini adalah provinsi-provinsi di Indonesia yang dianalisis berdasarkan dua indikator sosial utama, yaitu angka putus sekolah pada jenjang Sekolah Menengah Atas (SMA) dan jumlah penduduk miskin. Fokus penelitian ini adalah mengidentifikasi pola keterkaitan antara kedua variabel tersebut dan mengelompokkan provinsi ke dalam klaster-klaster tertentu berdasarkan kemiripan karakteristik sosial-ekonominya.

Secara lebih spesifik, objek penelitian ini mencakup:

1. Angka Putus Sekolah SMA/sederajat: Menunjukkan jumlah peserta didik yang tidak melanjutkan pendidikan hingga lulus pada tingkat SMA. Faktor ini mencerminkan sejauh mana akses dan kesinambungan pendidikan tersedia di suatu daerah.
2. Jumlah Penduduk Miskin (dalam ribu jiwa): Merupakan indikator tingkat kesejahteraan ekonomi masyarakat di masing-masing provinsi. Kemiskinan sering kali menjadi penyebab utama siswa terpaksa keluar dari sekolah untuk membantu ekonomi keluarga.

Penelitian ini memusatkan perhatian pada data tahun 2024 untuk seluruh provinsi di Indonesia. Dengan menggunakan pendekatan unsupervised learning melalui algoritma K-Means, penelitian ini bertujuan mengelompokkan provinsi ke dalam beberapa klaster berdasarkan kesamaan kondisi sosial yang ditunjukkan oleh kedua variabel tersebut.

3.2 Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari dua sumber resmi yang dapat diakses secara publik. Data angka putus sekolah pada jenjang SMA/sederajat diambil dari portal data milik Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia melalui situs data.dikdasmen.kemdikbud.go.id. Sementara itu, data jumlah penduduk miskin per provinsi pada tahun 2024 diperoleh dari situs resmi Badan Pusat Statistik Indonesia (bps.go.id). Kedua data ini kemudian digabungkan menjadi satu dataset yang utuh untuk keperluan analisis klaster. Sebelum dilakukan proses pengelompokan, data yang diperoleh melalui kedua situs tersebut terlebih dahulu melalui tahap pembersihan (cleaning) dan normalisasi agar dapat dianalisis secara seimbang menggunakan algoritma data mining. Penggunaan data dari sumber resmi pemerintah memastikan bahwa informasi yang digunakan bersifat akurat, mutakhir, dan relevan dengan tujuan penelitian ini.

3.3 Metode

Penelitian ini menggunakan algoritma *K-Means Clustering*, yaitu metode pembelajaran tanpa pengawasan (*unsupervised learning*) yang digunakan untuk mengelompokkan data berdasarkan kemiripan nilai antar objek. Algoritma ini cocok untuk data numerik dan mampu mengidentifikasi pola tersembunyi dalam

data dengan membagi data ke dalam beberapa klaster yang memiliki karakteristik serupa.

Dalam konteks penelitian ini, K-Means digunakan untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan angka putus sekolah dan jumlah penduduk miskin. Metode ini dipilih karena kesederhanaannya serta kemampuannya dalam menghasilkan pengelompokan yang dapat memberikan gambaran visual dan interpretatif terhadap kondisi sosial ekonomi antar wilayah. Hasil klastering diharapkan dapat menjadi dasar dalam memahami hubungan antara kemiskinan dan pendidikan secara lebih sistematis.

3.4 Alur Kerja

3.4.1 Perhitungan Manual K-Means

a. Siapkan Data dan Tentukan berapa k (Kluster) yang akan digunakan.
   Pada tahap ini, persiapan data dilakukan untuk memastikan perbandingan yang adil antara dua fitur yang digunakan: 'Siswa/i Putus Sekolah' dan 'Penduduk Miskin'. Terlihat bahwa skala kedua data ini sangat berbeda data putus sekolah berada di ratusan, sementara data penduduk miskin mencapai jutaan. Tanpa normalisasi, fitur 'Penduduk Miskin' akan mendominasi perhitungan jarak, dan fitur 'Siswa/i Putus Sekolah' hampir tidak akan berpengaruh. Oleh karena itu, metode Normalisasi Min-Max diterapkan untuk mengubah skala kedua fitur ke rentang yang sama, yaitu antara 0 dan 1, sehingga keduanya memiliki bobot yang setara dalam analisis.

Data Asli:

Provinsi	Putus Sekolah	Penduduk Miskin
Jawa Barat	286	3,668,350
Jawa Tengah	226	3,396,340
Jawa Timur	737	3,893,820
Sumatera Utara	980	1,110,920
Sumatera Barat	208	315,43
Sumatera Selatan	281	948,84
Kalimantan Barat	211	333,99
Sulawesi Selatan	409	711,77
Maluku	186	293,99
Nusa Tenggara Barat	385	658,6

Data Setelah Normalisasi:

Provinsi	X (Siswa Putus Sekolah)	Y (Penduduk Miskin)
Jawa Barat	0,126	0,937

Jawa Tengah	0,05	0,862
Jawa Timur	0,694	1
Sumatera Utara	1	0,227
Sumatera Barat	0,028	0,006
Sumatera Selatan	0,12	0,182
Kalimantan Barat	0,031	0,011
Sulawesi Selatan	0,281	0,116
Maluku	0	0
Nusa Tenggara Barat	0,251	0,101

b. Inisialisasi Centroid

Setelah data dinormalisasi, langkah selanjutnya adalah menentukan titik awal untuk pusat dari tiga klaster (K=3) yang akan dibentuk. Pada perhitungan ini, tiga data provinsi—Jawa Barat, Sumatera Utara, dan Sumatera Barat—dipilih untuk menjadi centroid awal. Titik-titik ini berfungsi sebagai "jangkar" atau posisi awal dari pusat masing-masing klaster sebelum proses pengelompokan iteratif dimulai.

Inisiasi Centroid Awal:

Centroid Awal	X	Y
C1 (Jawa Barat)	0,126	0,937
C2 (Sumatera Utara)	1	0,227
C3 (Sumatera Barat)	0,028	0,006

c. Inisiasi 1(Penugasan dan Pembaruan)

Ini adalah siklus pertama dari proses pengelompokan. Pada tahap ini, jarak Euclidean dari setiap data provinsi dihitung ke ketiga centroid awal, kemudian setiap provinsi dikelompokkan ke dalam klaster dengan centroid terdekat. Sebagai contoh, Jawa Timur ditempatkan ke dalam Klaster 1 karena jaraknya lebih dekat ke Centroid 1 (Jawa Barat) dibandingkan ke dua centroid lainnya. Setelah pengelompokan awal ini selesai, posisi setiap centroid diperbarui dengan cara menghitung nilai rata-rata dari seluruh anggota yang ada di dalam klaster tersebut, sehingga pusat klaster bergeser ke lokasi yang lebih representatif. Centroid baru untuk Klaster 1, misalnya, adalah hasil rata-rata dari data Provinsi Jawa Timur, Jawa Barat, dan Jawa Tengah.

Iterasi 1:

Provinsi	X	Y	Jarak ke C1	Jarak ke C2	Jarak ke C3	Klaster Terdekat
Jawa Barat	0,126	0,937	0	1,126	0,9361	1

Jawa Tengah	0,05	0,862	0,1068	1,1427	0,8563	1
Jawa Timur	0,694	1	0,5715	0,8314	1,1965	1
Sumatera Utara	1	0,227	1,126	0	0,9968	2
Sumatera Barat	0,028	0,006	0,9361	0,9968	0	3
Sumatera Selatan	0,12	0,182	0,755	0,8811	0,1986	3
Kalimantan Barat	0,031	0,011	0,9309	0,9928	0,0058	3
Sulawesi Selatan	0,281	0,116	0,8355	0,7275	0,2759	3
Maluku	0	0	0,9454	1,0254	0,0286	3
Nusa Tenggara Barat	0,251	0,101	0,8453	0,7595	0,2424	3

d. Iterasi 2(Pengecekan Konvergensi)

Proses penugasan dan pembaruan diulang kembali, namun kali ini menggunakan posisi centroid baru yang didapat dari Iterasi 1. Jarak dari setiap provinsi dihitung ulang ke centroid-centroid baru ini. Hasil dari Iterasi 2 menunjukkan bahwa keanggotaan klaster untuk setiap provinsi tidak berubah sama sekali dibandingkan dengan hasil Iterasi 1. Fenomena di mana tidak ada lagi perpindahan anggota antar klaster ini disebut konvergensi. Ini menandakan bahwa proses pengelompokan telah mencapai titik stabil dan optimal.

Centroid Baru:

Centroid Baru	X	Y
C1_new	0,29	0,933
C2_new	1	0,227
C3_new	0,1185	0,0693

Iterasi 2:

Provinsi	X	Y	Jarak ke C1_new	Jarak ke C2_new	Jarak ke C3_new	Klaster Terdekat
Jawa Timur	0,694	1	0,4095	0,8314	1,0942	1
Jawa Barat	0,126	0,937	0,164	1,126	0,8677	1
Jawa Tengah	0,05	0,862	0,2503	1,1427	0,7956	1
Sumatera Utara	1	0,227	1,0013	0	0,8955	2

Sumatera Selatan	0,12	0,182	0,77	0,8811	0,1127	3
Sulawesi Selatan	0,281	0,116	0,817	0,7275	0,1691	3
Nusa Tenggara Barat	0,251	0,101	0,8329	0,7595	0,1362	3
Sumatera Barat	0,028	0,006	0,9633	0,9968	0,1105	3
Maluku	0	0	0,977	1,0254	0,1373	3
Kalimantan Barat	0,031	0,011	0,9577	0,9928	0,1052	3

e. Hasil Akhir

Karena proses telah konvergen pada iterasi kedua, maka pengelompokan final telah didapatkan. Hasil analisis ini menunjukkan terbentuknya tiga kelompok yang berbeda. Klaster pertama, yang dapat dikategorikan sebagai klaster 'Tinggi', secara eksklusif berisi provinsi-provinsi di Pulau Jawa (Jawa Timur, Jawa Barat, dan Jawa Tengah) yang memiliki karakteristik nilai penduduk miskin yang sangat dominan. Berbeda dengan itu, klaster kedua merupakan sebuah *outlier* yang hanya beranggotakan Provinsi Sumatera Utara, yang membentuk klaster sendiri karena memiliki angka siswa putus sekolah yang ekstrem tinggi. Adapun klaster ketiga adalah kelompok 'Rendah-Menengah' yang merupakan klaster terbesar, terdiri dari enam provinsi sisanya (Sumatera Selatan, Sulawesi Selatan, NTB, Sumatera Barat, Maluku, dan Kalimantan Barat) yang secara umum memiliki angka putus sekolah dan kemiskinan yang relatif lebih rendah dibandingkan dua klaster lainnya.

Klaster 1 (Tinggi)	Klaster 2 (Outlier)	Klaster 3 (Rendah)
Jawa Timur	Sumatera Utara	Sumatera Selatan
Jawa Barat		Sulawesi Selatan
Jawa Tengah		Nusa Tenggara Barat
		Sumatera Barat
		Maluku
		Kalimantan Barat

3.4.2 Penerapan K-Means dengan Jupyter Notebook

a. Pengumpulan dan persiapan data

- Mengimport library yang diperlukan

  Pertama, import library yang akan digunakan. Library pandas digunakan untuk membaca dan mengelola data dalam bentuk tabel (dataframe), sedangkan numpy (diimpor sebagai nm) digunakan untuk melakukan operasi numerik dan manipulasi

array. Visualisasi data dilakukan dengan bantuan matplotlib.pyplot (diimpor sebagai mtp) dan seaborn untuk membuat grafik yang lebih informatif dan estetis. Selain itu, plotly.express (diimpor sebagai px) digunakan untuk membuat visualisasi data interaktif yang dapat membantu dalam eksplorasi data secara lebih dinamis. Untuk memastikan semua fitur memiliki skala yang sama, digunakan MinMaxScaler dari sklearn.preprocessing yang mengubah nilai atribut ke dalam rentang 0 hingga 1. Terakhir, algoritma K-Means dari sklearn.cluster digunakan untuk membentuk kelompok (cluster) berdasarkan kemiripan data.

```
import pandas as pd
import numpy as nm
import matplotlib.pyplot as mtp
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
import plotly.express as px
```

- Membaca dataset
  Selanjutnya baca dataset dan simpan sebagai data frame dalam variabel df kemudian tampilkan data pada data frame tersebut untuk melihat data yang ada.

```
df = pd.read_csv('datasettb.csv')
df.head(10)
```

	Provinsi	Siswa_X	Siswa_XI	Siswa_XII	jumlah_siswa	Siswi_X	Siswi_XI	Siswi_XII	jumlah_siswi	Siswa_putussekolah	Siswi_putussekolah	Siswa/i_putussekolah	Penc
0	Prov. D.K.I. Jakarta	31177	30596	30863	92636	32097	32307	33103	97507	8	5	13	
1	Prov. Jawa Barat	127402	122130	120881	370413	161401	155902	155044	472347	161	125	286	
2	Prov. Jawa Tengah	67167	64258	59790	191215	91534	90326	90473	272333	112	114	226	
3	Prov. D.I. Yogyakarta	8912	8660	8757	26329	12343	12104	11963	36410	0	0	0	
4	Prov. Jawa Timur	83600	82317	80277	246194	105812	103703	103605	313120	383	354	737	
5	Prov. Aceh	22554	22396	20683	65633	25875	25812	24370	76057	105	88	193	
6	Prov. Sumatera Utara	61632	59524	57601	178757	72398	72242	70796	215436	563	417	980	
7	Prov. Sumatera Barat	24490	22094	21807	68391	30690	29151	29041	88882	111	97	208	
8	Prov. Riau	28940	27300	25520	81760	33913	32976	31052	97941	74	51	125	
9	Prov. Jambi	13597	12870	12188	38655	15449	15296	14494	45239	50	31	81	

b. Data preprocessing
   - Memilih atribut yang akan digunakan
     Kami menggunakan 2 atribut yaitu “Siswa/i putussekolah” dan “Penduduk_Miskin” yang ada pada index 11 dan 12 dengan

menggunakan iloc dan simpan ke kembali ke dalam variabel df sebagai data frame untuk mengganti isi dari variabel df tersebut. Setelah itu tampilkan data dari data frame yang baru dibuat tersebut untuk melihat apakah data sudah sesuai.

```
df = df.iloc[:,[11,12]]
df.head(10)
```

	Siswa/i_putussekolah	Penduduk_Miskin
0	13	449070.0
1	286	3668350.0
2	226	3396340.0
3	0	430470.0
4	737	3893820.0
5	193	718960.0
6	980	1110920.0
7	208	315430.0
8	125	473040.0
9	81	272700.0

- Melakukan pengecekan nilai null dan duplikasi data lalu membersihkan data
  Selanjutnya lakukan pengecekan nilai null dan duplikasi data yang ada pada data frame.

```
print("Data Null")
print(df.isnull().sum())

print("\nDuplikasi Data")
print(df.duplicated().sum())
```

```
Data Null
Siswa/i_putussekolah    0
Penduduk_Miskin         1
dtype: int64

Duplikasi Data
0
```

Dari hasil pengecekan terdapat 1 nilai null pada kolom “Penduduk_Miskin” dan tidak ada duplikasi data pada data frame tersebut. Sehingga hanya perlu melakukan drop data null tersebut dengan menggunakan dropna() untuk menghapus semua nilai null.

```
df = df.dropna()

print("Data Null")
print(df.isnull().sum())

print("\nDuplikasi Data")
print(df.duplicated().sum())
```

Setelah dilakukan penghapusan nilai null, dapat dilihat pada output nilai null sekarang sudah tidak ada.

```
Data Null
Siswa/i_putussekolah    0
Penduduk_Miskin         0
dtype: int64

Duplikasi Data
0
```

- Konversi data frame menjadi array
  Selanjutnya data frame akan dirubah menjadi array. Fungsi nm.set_printoptions(suppress=True) digunakan untuk menonaktifkan notasi ilmiah (scientific notation) saat menampilkan array NumPy, sehingga angka ditampilkan dalam format desimal biasa. Selanjutnya, df.values digunakan untuk mengonversi data dari bentuk dataframe (pandas) menjadi array NumPy (x_array) agar dapat diproses lebih lanjut.

```
nm.set_printoptions(suppress=True)
x_array = df.values
x_array
```

```
array([[    13.,  449070.],
       [   286., 3668350.],
       [   226., 3396340.],
       [     0.,  430470.],
       [   737., 3893820.],
       [   193.,  718960.],
       [   980., 1110920.],
       [   208.,  315430.],
       [   125.,  473040.],
       [    81.,  272700.],
       [   281.,  948840.],
       [   118.,  939300.],
       [   211.,  333990.],
       [    98.,  149240.],
       [    34.,  180200.],
       [    57.,  211880.],
       [    24.,  173300.],
       [   159.,  358330.],
       [   409.,  711770.],
       [   154.,  305270.],
       [   186.,  293990.],
       [    11.,  176210.],
       [   385.,  658600.],
       [   548., 1107940.],
       [    33.,  161070.],
       [    52.,  261150.],
       [   141.,   79690.],
       [   177.,  777490.],
       [    12.,   78580.],
       [   147.,  170030.],
       [    23.,  124960.],
       [    51.,  108280.],
       [    90.,  155910.],
       [    39.,   41110.],
       [   151.,  287540.],
       [    23.,  103020.],
       [   158.,  331120.],
       [    95.,   96810.]])
```

- Melakukan Scaling untuk menyamakan skala semua atribut
  Setelah data frame diubah menjadi array maka selanjutnya akan dilakukan proses normalisasi data menggunakan MinMaxScaler dari library sklearn.preprocessing. Objek scaler diinisialisasi sebagai instance dari MinMaxScaler(), yang berfungsi untuk mengubah nilai-nilai dalam array x_array ke dalam skala antara 0 hingga 1. Proses transformasi ini dilakukan dengan metode fit_transform(), yang pertama-tama menghitung nilai minimum dan maksimum dari setiap fitur, lalu menerapkan transformasi ke data. Hasilnya disimpan dalam variabel x_scaled, yang berisi data yang telah dinormalisasi dan siap digunakan untuk algoritma seperti K-Means Clustering, agar semua fitur memiliki kontribusi yang seimbang.

```
scaler = MinMaxScaler()
x_scaled = scaler.fit_transform(x_array)
x_scaled
```

```
array([[0.01326531, 0.1058891 ],
       [0.29183673, 0.94147756],
       [0.23061224, 0.87087531],
       [0.        , 0.10106133],
       [0.75204082, 1.        ],
       [0.19693878, 0.17594109],
       [1.        , 0.27767727],
       [0.2122449 , 0.07120183],
       [0.12755102, 0.1121107 ],
       [0.08265306, 0.06011093],
       [0.28673469, 0.23560818],
       [0.12040816, 0.233132  ],
       [0.21530612, 0.07601922],
       [0.1       , 0.02806596],
       [0.03469388, 0.03610186],
       [0.05816327, 0.04432464],
       [0.0244898 , 0.03431091],
       [0.1622449 , 0.08233685],
       [0.41734694, 0.17407487],
       [0.15714286, 0.06856472],
       [0.18979592, 0.06563692],
       [0.01122449, 0.03506623],
       [0.39285714, 0.1602742 ],
       [0.55918367, 0.27690379],
       [0.03367347, 0.03113652],
       [0.05306122, 0.05711305],
       [0.14387755, 0.01001373],
       [0.18061224, 0.19113299],
       [0.0122449 , 0.00972562],
       [0.15      , 0.03346216],
       [0.02346939, 0.0217639 ],
       [0.05204082, 0.01743448],
       [0.09183673, 0.02979721],
       [0.03979592, 0.        ],
       [0.15408163, 0.06396277],
       [0.02346939, 0.01606921],
       [0.16122449, 0.07527429],
       [0.09693878, 0.01445736]])
```

c. Penentuan jumlah cluster

- Menggunakan metode elbow untuk mencari jumlah cluster yang optimal

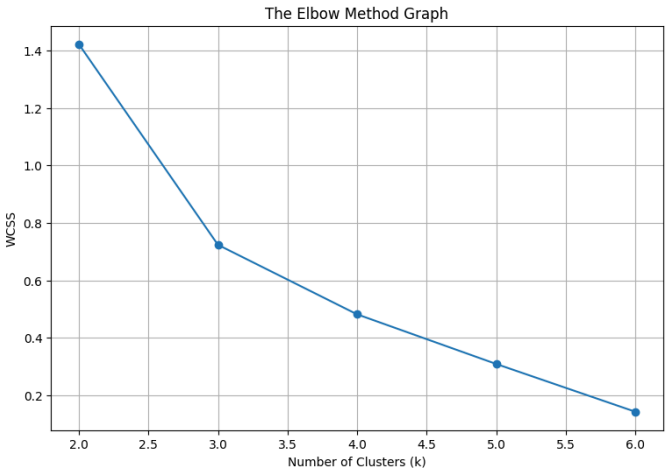
  Metode Elbow digunakan pada algoritma K-Means Clustering. Variabel wcss_list digunakan untuk menyimpan nilai WCSS (*Within-Cluster Sum of Squares*) dari berbagai jumlah klaster. Melalui perulangan for, model K-Means dilatih dengan jumlah klaster dari 2 hingga 6 (n_clusters=i), dan setiap nilai WCSS disimpan ke dalam wcss_list setelah model dilatih dengan data yang telah dinormalisasi (x_scaled). Selanjutnya, data WCSS divisualisasikan menggunakan matplotlib, di mana sumbu X menunjukkan jumlah klaster dan sumbu Y menunjukkan nilai WCSS. Grafik ini membantu mengidentifikasi *"elbow point"*, yaitu titik di mana penurunan WCSS mulai melambat, yang biasanya dianggap sebagai jumlah klaster optimal.

```
wcss_list = []
for i in range(2, 7):
         kmeans  =  KMeans(n_clusters=i,  init='k-means++',
random_state=42)
    kmeans.fit(x_scaled)
    wcss_list.append(kmeans.inertia_)

mtp.figure(figsize=(9, 6))
mtp.plot(range(2, 7), wcss_list, marker='o')
```

```
mtp.title('The Elbow Method Graph')
mtp.xlabel('Number of Clusters (k)')
mtp.ylabel('WCSS')
mtp.grid(True)
mtp.show()
```

Kode tersebut akan memberikan output berupa visualisasi dengan line chart dan dari visualisasi ini dapat dilihat bahwa jumlah cluster optimal yang dapat digunakan adalah 3.



d. Pembuatan dan pelatihan model

- Melakukan pelatihan model K-Means

  Setelah data dipersiapkan dan dinormalisasi menggunakan MinMaxScaler, langkah selanjutnya adalah membuat model K-Means dan melatihnya menggunakan data tersebut

```
kmeans = KMeans(n_clusters=3, init='k-means++',
random_state=42)
y_kmeans = kmeans.fit_predict(x_scaled)
kmeans
```

```
KMeans
KMeans(n_clusters=3, random_state=42)
```

Kode di atas akan membentuk 3 cluster menggunakan metode inisialisasi 'k-means++' untuk mempercepat konvergensi. Parameter random_state=42 digunakan agar hasil clustering bersifat reproducible.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Clustering

Setelah model K-Means dilatih dan label cluster diperoleh melalui fit_predict, label tersebut ditambahkan ke dalam salinan dataframe asli untuk analisis lebih lanjut.

```
df_clustered = df.copy()
df_clustered['cluster'] = y_kmeans
df_clustered.head(10)
```

Pada bagian ini akan menghasilkan dataframe baru bernama df_clustered yang memuat data asli beserta kolom baru bernama 'cluster' yang menunjukkan keanggotaan cluster masing-masing baris data seperti dibawah ini

	Siswa/i_putussekolah	Penduduk_Miskin	cluster
0	13	449070.0	0
1	286	3668350.0	2
2	226	3396340.0	2
3	0	430470.0	0
4	737	3893820.0	2
5	193	718960.0	0
6	980	1110920.0	1
7	208	315430.0	0
8	125	473040.0	0
9	81	272700.0	0

4.2 Visualisasi dan Interpretasi

Visualisasi Cluster Dengan Scatter Plot

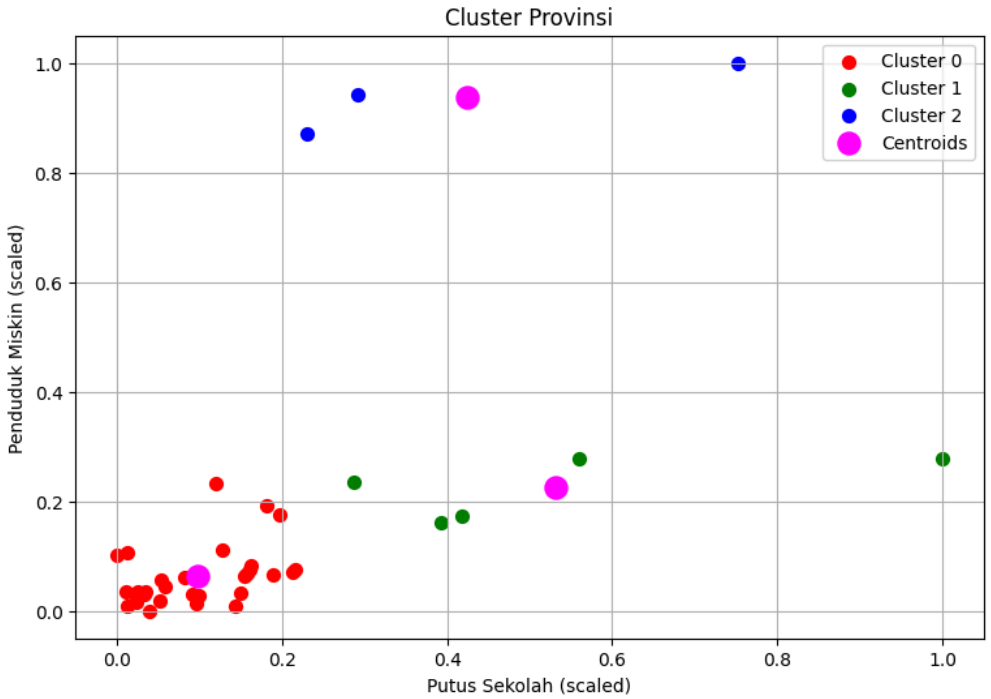
```
kmeans = KMeans(n_clusters=3, init='k-means++', random_state=42)
y_kmeans = kmeans.fit_predict(x_scaled)

mtp.figure(figsize=(9, 6))
colors = ['red', 'green', 'blue']

for i in range(3):
    mtp.scatter(x_scaled[y_kmeans == i, 0], x_scaled[y_kmeans == i, 1],
                s=50, c=colors[i], label=f'Cluster {i}')

mtp.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
            s=150, c='magenta', marker='o', label='Centroids')
```

```
mtp.title('Clusters of Provinces')
mtp.xlabel('Putus Sekolah (scaled)')
mtp.ylabel('Penduduk Miskin (scaled)')
mtp.legend()
mtp.grid(True)
mtp.show()
```



Visualisasi ini menunjukkan hasil pengelompokan beberapa provinsi berdasarkan dua hal: tingkat putus sekolah dan jumlah penduduk miskin. Sebelum dikelompokkan, kedua data tersebut telah diskalakan agar bisa dibandingkan secara adil.

Setiap titik pada grafik mewakili satu provinsi. Warna yang berbeda—merah, hijau, dan biru—menunjukkan kelompok atau klaster yang terbentuk dari proses K-Means. Artinya, provinsi yang warnanya sama cenderung punya karakteristik yang mirip dalam hal tingkat putus sekolah dan jumlah penduduk miskin.

Titik besar berwarna magenta di tengah-tengah masing-masing klaster adalah pusat kelompok atau yang disebut centroid. Titik ini menunjukkan posisi rata-rata dari provinsi-provinsi dalam kelompok tersebut. Jadi, dari posisi titik-titik itu kita bisa tahu seperti apa ciri umum tiap kelompok.

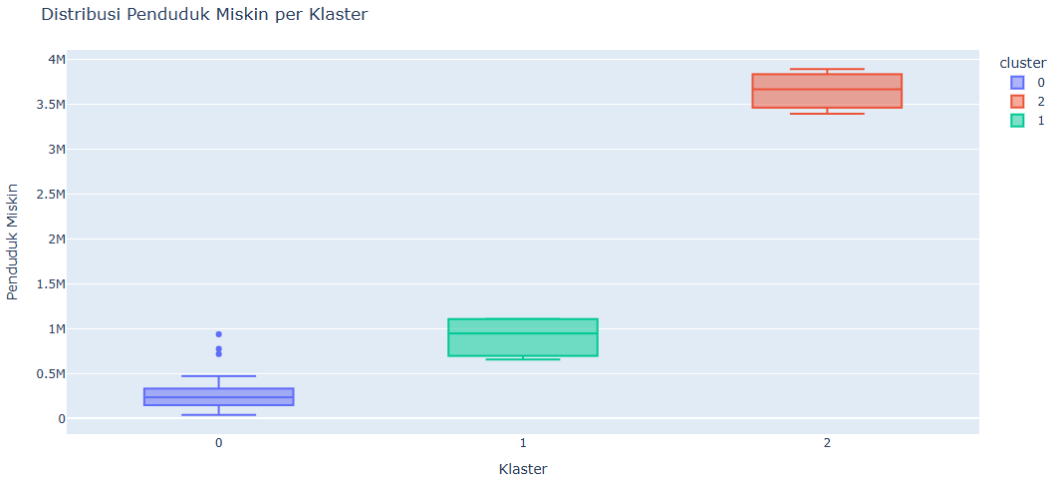
Dengan melihat hasil ini, kita bisa mengetahui bahwa ada pola atau kemiripan tertentu antara provinsi-provinsi, dan pola itu bisa berguna untuk membantu perumusan kebijakan, misalnya dalam penanganan kemiskinan atau pendidikan.

Distribusi Penduduk Miskin per Klaster dengan box plot

```
df_clustered = df.copy()
df_clustered['cluster'] = y_kmeans

fig = px.box(df_clustered, x='cluster', y='Penduduk_Miskin',
             hover_data=['Penduduk_Miskin'], color='cluster')

fig.update_layout(
    title="Distribusi Penduduk Miskin per Klaster",
    xaxis_title="Klaster",
    yaxis_title="Penduduk Miskin"
)
fig.show()
```



Visualisasi ini menggunakan boxplot (diagram kotak) untuk menampilkan distribusi jumlah penduduk miskin di setiap klaster yang terbentuk dari algoritma K-Means. Data yang digunakan telah ditambahkan kolom baru bernama "cluster" yang menunjukkan hasil pengelompokan dari proses sebelumnya.

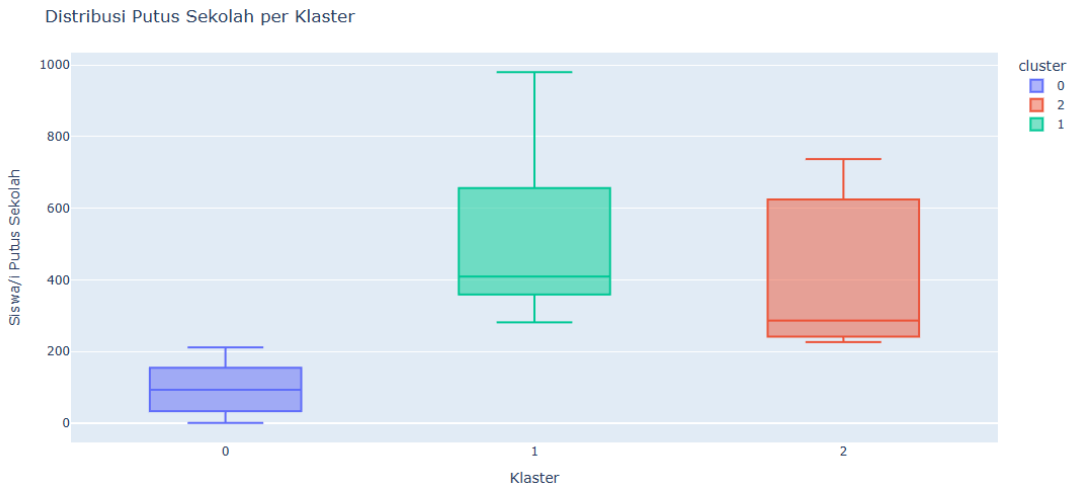
Setiap kotak pada grafik mewakili satu klaster, dengan nilai pada sumbu Y menunjukkan jumlah penduduk miskin (dalam skala asli, bukan yang sudah diskalakan). Kotak ini menunjukkan bagaimana data tersebar di dalam klaster tersebut—mulai dari nilai minimum, kuartil pertama, median, kuartil ketiga, hingga nilai maksimum. Selain itu, outlier atau data yang menyimpang dari pola umum juga bisa terlihat di luar batas kotak.

Dengan melihat perbedaan tinggi dan posisi kotak antar klaster, kita bisa membandingkan secara visual seberapa besar perbedaan tingkat kemiskinan antar kelompok. Misalnya, jika salah satu klaster memiliki median yang lebih tinggi dibanding klaster lain, berarti provinsi dalam kelompok tersebut cenderung memiliki jumlah penduduk miskin yang lebih banyak. Sebaliknya, klaster dengan median rendah cenderung berisi provinsi dengan jumlah penduduk miskin yang lebih sedikit.

Visualisasi ini sangat membantu untuk memahami karakteristik sosial ekonomi dari masing-masing klaster, terutama dalam hal distribusi dan variasi jumlah penduduk miskin. Ini juga bisa menjadi dasar untuk menganalisis kebutuhan atau pendekatan kebijakan yang berbeda-beda untuk setiap kelompok provinsi.

Distribusi Putus Sekolah per Klaster

```
fig2 = px.box(df_clustered, x='cluster', y='Siswa/i_putussekolah',
              hover_data=['Siswa/i_putussekolah'], color='cluster')
fig2.update_layout(
    title="Distribusi Putus Sekolah per Klaster",
    xaxis_title="Klaster",
    yaxis_title="Siswa/i Putus Sekolah"
)
fig2.show()
```



Visualisasi ini menampilkan distribusi jumlah siswa atau siswi yang putus sekolah di setiap klaster yang terbentuk dari proses K-Means. Setiap titik data dalam boxplot berasal dari provinsi yang telah dikelompokkan ke dalam klaster berdasarkan kemiripan karakteristik sosial, termasuk data putus sekolah.

Setiap kotak dalam grafik mewakili sebaran data jumlah siswa putus sekolah dalam satu klaster. Sumbu X menunjukkan nomor klaster, sedangkan sumbu Y menunjukkan jumlah siswa putus sekolah. Bentuk dan posisi kotak memberikan informasi penting: bagian tengah kotak menunjukkan median atau nilai tengah dari data dalam klaster tersebut, sedangkan ujung atas dan bawah kotak menunjukkan rentang nilai antara kuartil pertama dan kuartil ketiga. Garis vertikal di luar kotak (whiskers) memperlihatkan jangkauan data yang masih dianggap wajar, dan titik-titik di luar whiskers merepresentasikan outlier atau nilai yang jauh berbeda dari mayoritas data di dalam klaster.

Melalui visualisasi ini, kita dapat dengan cepat membandingkan seberapa besar masalah putus sekolah di antara masing-masing klaster. Misalnya, jika satu klaster

memiliki median dan sebaran nilai yang jauh lebih tinggi dari klaster lain, maka klaster tersebut mencakup provinsi-provinsi dengan tingkat putus sekolah yang lebih parah. Sebaliknya, klaster dengan nilai lebih rendah cenderung berisi provinsi dengan jumlah siswa putus sekolah yang lebih sedikit.

Visualisasi ini berguna untuk memahami perbedaan kondisi pendidikan di berbagai kelompok provinsi, yang bisa dijadikan dasar dalam menyusun strategi intervensi yang lebih tepat sasaran untuk mengatasi masalah putus sekolah.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan penerapan algoritma K-Means Clustering terhadap data angka putus sekolah tingkat SMA dan jumlah penduduk miskin per provinsi di Indonesia tahun 2024, terbentuk tiga klaster utama yang merepresentasikan kondisi sosial ekonomi yang berbeda-beda, yaitu Klaster 0, Klaster 1 dan Klaster 2. Klaster 0 mencakup provinsi-provinsi dengan tingkat kemiskinan dan angka putus sekolah yang rendah, yang menunjukkan kondisi sosial yang relatif stabil dan lebih baik. Klaster 1 terdiri dari provinsi dengan tingkat sedang pada kedua variabel tersebut, yang dapat dikategorikan sebagai wilayah dengan kondisi menengah. Sementara itu, Klaster 2 dihuni oleh provinsi-provinsi dengan tingkat kemiskinan dan angka putus sekolah yang sama-sama tinggi, yang menjadikannya sebagai kelompok dengan permasalahan paling kompleks dan perlu mendapat prioritas dalam penanganan.

Hasil pengelompokan ini memperlihatkan adanya kecenderungan keterkaitan antara tingginya angka kemiskinan dengan tingginya angka putus sekolah, meskipun tidak dapat disimpulkan sebagai hubungan kausal. Klastering ini memberikan gambaran visual dan interpretatif yang kuat mengenai bagaimana karakteristik sosial tiap provinsi dapat dikelompokkan secara objektif. Oleh karena itu, penggunaan K-Means dalam konteks ini terbukti efektif untuk mengeksplorasi pola-pola sosial yang tersembunyi dalam data, dan hasilnya dapat dijadikan sebagai dasar awal dalam penyusunan strategi kebijakan di bidang pendidikan dan pengentasan kemiskinan yang lebih terarah dan berbasis data.

5.2 Saran

Penggunaan metode clustering seperti K-Means dapat dikembangkan lebih lanjut dengan memasukkan variabel-variabel lain yang relevan, seperti tingkat pengangguran, indeks pembangunan manusia (IPM), atau akses terhadap fasilitas pendidikan, guna menghasilkan pengelompokan yang lebih komprehensif. Hasil dari proses clustering ini dapat dimanfaatkan sebagai salah satu bahan pertimbangan oleh pemerintah atau pihak terkait dalam merancang kebijakan dan program intervensi yang lebih tepat sasaran, khususnya dalam bidang pendidikan dan upaya pengentasan kemiskinan di berbagai provinsi di Indonesia.

DAFTAR PUSTAKA

Han, J., Kamber, M., & Pei, J. (2012). Data Mining : Concepts and Techniques : Concepts and Techniques (3rd Edition). In *Data Mining*. books.google.com. http://linkinghub.elsevier.com/retrieve/pii/B9780123814791000010

Ali, M., Raza, M., Ali, A., Khan, S. A., & Zahid, S. (2022). A comprehensive survey on K-means clustering algorithm: Variants, applications, and challenges. *Information Sciences, 607*, 441–478. https://doi.org/10.1016/j.ins.2022.05.105

Zulyanto, A. (2022). Pendidikan dan pengentasan kemiskinan dalam pembangunan berkelanjutan (SDGs). In *Convergence: The Journal of Economic Development*. https://pdfs.semanticscholar.org/9be4/8c9e029fc4734bbe67b0b72f33e516213f70.pdf