LAPORAN TUGAS BESAR DATA MINING
PENGELOMPOKKAN PROVINSI DI INDONESIA BERDASARKAN INDEX PEMBANGUNAN MANUSIA MENGGUNAKAN ALGORITMA K-MEANS

Dosen Pengampu : Achmad Bahauddin, S.T., M.T.

Disusun oleh :

Andiko Ramadani	: 3337230003
Ussy Cantika	: 3337230008
Ismet Maulana Azhari	: 3337230014

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS SULTAN AGENG TIRTAYASA

2025

DAFTAR ISI

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Indeks Pembangunan Manusia (IPM) adalah salah satu indikator yang esensial dalam menilai keberhasilan suatu wilayah dalam membangun kualitas hidup masyarakatnya secara menyeluruh. IPM tidak berdiri sebagai angka tunggal yang berdiri sendiri, melainkan merupakan gabungan dari tiga dimensi penting yang saling melengkapi: kesehatan, pendidikan, dan standar hidup layak. Masing-masing dimensi ini diukur melalui indikator spesifik, yakni angka harapan hidup (untuk kesehatan), rata-rata lama sekolah (untuk pendidikan), serta pengeluaran per kapita (untuk kesejahteraan ekonomi). Dengan kata lain, IPM merupakan cerminan dari seberapa layak dan manusiawi kehidupan masyarakat di suatu wilayah, dilihat dari sisi umur panjang, kecakapan intelektual, dan kemampuan ekonomi dasar.

Namun realitas di Indonesia menunjukkan bahwa pencapaian IPM antar provinsi tidak merata. Beberapa provinsi, terutama yang terletak di wilayah barat dan perkotaan, secara konsisten mencatatkan nilai IPM yang tinggi. Sebaliknya, provinsi-provinsi di kawasan timur Indonesia atau wilayah terpencil masih berada dalam kategori pembangunan yang relatif tertinggal. Ketimpangan ini menjadi sorotan penting dalam diskusi mengenai pemerataan pembangunan nasional, dan secara khusus mendorong perlunya pendekatan analitis yang mampu memetakan kondisi ini secara lebih objektif dan terukur.

Salah satu pendekatan yang dapat digunakan untuk memahami kondisi tersebut adalah melalui analisis *clustering* menggunakan algoritma K-Means. Metode ini termasuk dalam kelompok *unsupervised learning*, di mana data tidak dikategorikan sebelumnya dan pola pengelompokan ditemukan secara otomatis berdasarkan kemiripan nilai antar data. Dalam penelitian ini, K-Means digunakan untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan kombinasi dari empat variabel numerik utama, yaitu: Angka Harapan Hidup (AHH), Rata-Rata Lama Sekolah (RLS), Pengeluaran per Kapita (PPP), dan nilai IPM itu sendiri.

Pemilihan keempat variabel ini bukan tanpa alasan. AHH, RLS, dan PPP merupakan penyusun utama IPM yang secara langsung merepresentasikan kondisi dasar pembangunan manusia. Sementara itu, IPM sebagai indikator komposit tetap digunakan dalam proses clustering karena dapat menangkap keterkaitan dan keseimbangan antar ketiga komponen tersebut secara keseluruhan. Dengan memasukkan keempat variabel ini secara bersamaan ke dalam proses klasterisasi, model yang dihasilkan diharapkan mampu menggambarkan kondisi pembangunan provinsi secara lebih utuh dan realistis.

Seluruh variabel yang digunakan dalam analisis terlebih dahulu dinormalisasi agar berada pada skala yang setara, mengingat nilai-nilai mentahnya memiliki rentang yang sangat berbeda. Misalnya, pengeluaran per kapita bernilai ribuan hingga jutaan rupiah, sementara lama sekolah hanya berkisar pada belasan tahun. Proses normalisasi ini penting agar tidak ada satu variabel pun yang mendominasi hasil akhir pengelompokan hanya karena skala angkanya lebih besar. Setelah proses normalisasi, dilakukan eksplorasi untuk menentukan jumlah klaster optimal yang paling representatif terhadap struktur data, dengan mempertimbangkan metrik seperti inertia, silhouette score, dan Davies-Bouldin index.

Melalui proses tersebut, terbentuklah sejumlah klaster yang masing-masing mewakili kelompok provinsi dengan karakteristik pembangunan yang relatif serupa. Hasil ini memberikan sudut pandang baru dalam memahami perbedaan pembangunan manusia antar daerah, tidak hanya secara deskriptif berdasarkan nilai IPM tunggal, tetapi juga berdasarkan kombinasi indikator yang lebih kaya. Analisis ini pada akhirnya menjadi bagian penting dalam upaya mengkaji struktur disparitas pembangunan manusia secara menyeluruh, dan dapat digunakan sebagai dasar untuk pembahasan dan refleksi lebih lanjut dalam konteks akademik.

1.2 Rumusan Masalah

Dalam penelitian ini, proses pengelompokan provinsi-provinsi di Indonesia dilakukan berdasarkan variabel-variabel yang berhubungan erat dengan pembangunan manusia. Dengan pendekatan *clustering* berbasis algoritma K-Means, muncul sejumlah pertanyaan penting yang menjadi dasar perumusan masalah, di antaranya:

1. Bagaimana proses pengelompokan provinsi-provinsi di Indonesia dapat dilakukan menggunakan algoritma K-Means berdasarkan indikator pembangunan manusia?
2. Berapa jumlah klaster yang paling optimal untuk merepresentasikan variasi pembangunan manusia di seluruh provinsi?
3. Apa saja karakteristik utama yang membedakan masing-masing klaster hasil pengelompokan, baik dari sisi IPM maupun dari komponen-komponen penyusunnya (AHH, RLS, dan PPP)?
4. Bagaimana hasil pengelompokan tersebut dapat membantu dalam memahami struktur perbedaan pembangunan manusia antar provinsi secara lebih sistematis dan menyeluruh?

Pertanyaan-pertanyaan ini akan dijawab melalui proses pengolahan data, pemodelan algoritma K-Means, serta evaluasi hasil klasterisasi berdasarkan metrik yang relevan dan analisis interpretatif terhadap data.

1.3 Tujuan Penelitian

Penelitian ini secara umum bertujuan untuk menerapkan pendekatan analitik berbasis data dalam memahami variasi pembangunan manusia antar provinsi di Indonesia. Secara lebih spesifik, tujuan dari penelitian ini dapat dirinci sebagai berikut:

1. Menerapkan algoritma K-Means untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan empat variabel utama yang berkaitan dengan pembangunan manusia, yaitu IPM, AHH, RLS, dan PPP.
2. Menentukan jumlah klaster yang paling optimal dengan mempertimbangkan metrik evaluasi yang relevan, seperti inertia, silhouette score, dan Davies-Bouldin index.
3. Mengidentifikasi karakteristik utama dari setiap klaster, dengan melihat pola nilai variabel yang dominan di dalam masing-masing kelompok provinsi.
4. Memberikan pemahaman yang lebih dalam dan menyeluruh mengenai struktur perbedaan pembangunan manusia antar provinsi, yang dapat digunakan sebagai bahan diskusi dan refleksi dalam lingkup akademik.

1.4 Batasan Masalah

Agar ruang lingkup penelitian ini tetap terarah dan tidak melebar ke luar konteks, terdapat sejumlah batasan yang diterapkan sebagai pedoman dalam penyusunan dan pelaksanaan analisis:

1. Data yang digunakan dalam penelitian ini terbatas pada data provinsi di Indonesia dan tidak mencakup tingkat kabupaten atau kota.
2. Variabel yang dianalisis meliputi empat indikator utama, yaitu: Angka Harapan Hidup (**AHH**), Rata-Rata Lama Sekolah (**RLS**), Pengeluaran per Kapita (**PPP**), serta nilai **IPM** sebagai indeks komposit. Variabel lain yang juga berpengaruh terhadap pembangunan manusia, seperti tingkat pengangguran, akses sanitasi, atau ketimpangan pendapatan, tidak dimasukkan dalam ruang lingkup analisis ini.
3. Seluruh data diambil dari sumber resmi, yaitu Badan Pusat Statistik (BPS), dan dibatasi hanya pada tahun 2022-2024.
4. Pengelompokan dilakukan menggunakan algoritma K-Means, sehingga interpretasi hasil bersifat numerik dan berbasis pada pendekatan jarak antar titik dalam ruang fitur yang telah dinormalisasi. Potensi variasi hasil akibat sensitivitas algoritma terhadap inisialisasi centroid juga menjadi bagian dari keterbatasan yang disadari dalam penelitian ini.

BAB 2
LANDASAN TEORI

2.1 Data Mining

Dalam era digital saat ini, setiap aktivitas manusia hampir selalu terekam dalam bentuk data. Data tersebut berasal dari berbagai sektor kehidupan, mulai dari pendidikan, kesehatan, ekonomi, hingga layanan publik. Jumlahnya sangat besar dan terus bertambah setiap waktu, baik dalam bentuk terstruktur seperti tabel statistik, maupun tidak terstruktur seperti teks atau citra. Namun demikian, data yang melimpah tersebut tidak serta-merta memberikan manfaat apabila tidak dikelola dan dianalisis secara sistematis.

Di sinilah konsep *data mining* menjadi sangat relevan. **Data mining** adalah proses eksplorasi dan analisis data dalam jumlah besar yang dilakukan secara otomatis atau semi-otomatis, dengan tujuan untuk menemukan pola-pola tersembunyi, hubungan yang bermakna, atau informasi baru yang sebelumnya tidak tampak secara langsung. Proses ini berbeda dengan sekadar pelaporan data, karena fokus utamanya adalah menemukan pengetahuan yang belum diketahui sebelumnya, namun potensial untuk digunakan dalam pengambilan keputusan atau perumusan strategi.

Dalam konteks metodologis, data mining merupakan bagian inti dari tahapan yang lebih luas, yaitu ***Knowledge Discovery in Databases*** **(KDD)**. KDD terdiri dari serangkaian proses mulai dari pemilihan data yang relevan, pembersihan data dari noise dan anomali, transformasi data ke format yang sesuai, penerapan teknik mining untuk menemukan pola, hingga evaluasi dan representasi hasil. Di antara semua tahapan tersebut, data mining merupakan fase paling sentral karena di sinilah teknik komputasi, algoritma, dan analisis statistik diterapkan secara aktif untuk mengeksplorasi struktur dalam data.

2.1.1 Tujuan dan Manfaat Data Mining

Tujuan utama dari data mining adalah menemukan informasi yang tidak terlihat di permukaan, namun tersembunyi di balik data yang besar dan kompleks. Informasi yang dihasilkan bukan hanya bersifat deskriptif, tetapi juga dapat dimanfaatkan untuk memahami pola, membuat prediksi, dan mendukung proses pengambilan keputusan yang lebih cerdas dan berbasis bukti.

Dalam konteks pembangunan manusia, data mining dapat memberikan kontribusi yang nyata. Misalnya, melalui analisis pola dalam indikator sosial ekonomi, kita dapat mengetahui provinsi atau wilayah mana saja yang masih

tertinggal dalam hal pendidikan, kesehatan, atau kesejahteraan. Dari situ, dapat disusun strategi pengembangan yang lebih terarah dan kontekstual.

Beberapa manfaat nyata dari penerapan data mining dalam studi seperti ini antara lain:

- Mengidentifikasi kelompok wilayah yang memiliki kesamaan karakteristik dalam indikator pembangunan manusia.
- Mendeteksi kesenjangan yang tidak terlihat secara eksplisit dalam data agregat nasional.
- Menyusun rekomendasi berbasis data (evidence-based planning) untuk pengembangan wilayah.
- Meningkatkan pemahaman visual melalui grafik dan klaster yang mempermudah interpretasi oleh pihak non-teknis.

Dengan kata lain, data mining bukan hanya tentang pengolahan data secara teknis, tetapi juga tentang membangun narasi dan wawasan yang didasarkan pada struktur data yang kompleks.

2.1.2 Tahapan dalam Proses Data Mining

Agar proses data mining menghasilkan informasi yang bermakna dan dapat dipertanggungjawabkan, perlu dilakukan melalui tahapan yang runtut dan logis. Setiap tahap memiliki fungsi spesifik dan saling terkait satu sama lain:

1. **Pemilihan Data (Data Selection)**
   Proses diawali dengan memilih data yang benar-benar relevan terhadap tujuan analisis. Dalam konteks penelitian ini, data yang digunakan adalah data IPM provinsi di Indonesia yang dirilis oleh Badan Pusat Statistik (BPS), dengan indikator utama yaitu Angka Harapan Hidup (AHH), Rata-Rata Lama Sekolah (RLS), Pengeluaran per Kapita (PPP), serta nilai IPM itu sendiri. Ketepatan pemilihan data akan sangat memengaruhi kualitas hasil yang diperoleh.

2. **Pembersihan dan Integrasi Data (Data Cleaning and Integration)**
   Data mentah sering kali mengandung masalah seperti nilai yang hilang, duplikasi, atau format yang tidak seragam. Oleh karena itu, dilakukan pembersihan untuk memperbaiki kualitas data. Jika data berasal dari berbagai file atau sumber, maka perlu dilakukan integrasi agar seluruh informasi dapat dianalisis dalam satu kesatuan.

3. **Transformasi Data (Data Transformation)**
   Setelah data bersih, langkah berikutnya adalah mentransformasikan data ke dalam bentuk yang sesuai untuk dianalisis. Salah satu bentuk transformasi yang digunakan dalam penelitian ini adalah normalisasi, agar semua variabel berada dalam skala yang sebanding. Hal ini

penting karena algoritma seperti K-Means sangat sensitif terhadap perbedaan skala antar fitur.

4. **Proses Data Mining (Pattern Discovery)**
   Tahap inti ini melibatkan penerapan algoritma K-Means untuk melakukan pengelompokan data berdasarkan kemiripan karakteristik. Output dari tahap ini berupa sekumpulan klaster, di mana setiap klaster mewakili kelompok provinsi dengan nilai-nilai indikator yang relatif serupa.

5. **Evaluasi dan Interpretasi Pola (Pattern Evaluation and Interpretation)**
   Setelah pola ditemukan, perlu dilakukan evaluasi untuk menilai apakah klaster yang terbentuk benar-benar mencerminkan struktur data yang bermakna. Evaluasi dilakukan dengan metrik seperti *silhouette score* atau *Davies-Bouldin index*, dan dilanjutkan dengan interpretasi hasil berdasarkan karakteristik masing-masing klaster.

6. **Representasi Pengetahuan (Knowledge Representation)**
   Hasil akhir dari data mining perlu disampaikan dalam bentuk yang mudah dipahami. Representasi ini dapat berupa grafik, tabel, atau narasi visual yang menjelaskan perbedaan antar klaster. Penyajian ini sangat penting terutama dalam konteks akademik, agar hasil analisis dapat dibaca dan digunakan oleh pembaca dari berbagai latar belakang.

2.1.3 Penerapan Data Mining dalam Konteks IPM

Indeks Pembangunan Manusia (**IPM**) merupakan alat ukur yang digunakan secara luas untuk menilai tingkat keberhasilan pembangunan manusia. Di Indonesia, pengukuran IPM dilakukan berdasarkan tiga indikator utama: Angka Harapan Hidup (**AHH**), Rata-Rata Lama Sekolah (**RLS**), dan Pengeluaran per Kapita (**PPP**). Masing-masing indikator ini mewakili dimensi kesehatan, pendidikan, dan ekonomi sebagai pilar utama pembangunan.

Dengan mengumpulkan dan menganalisis ketiga indikator tersebut, serta menambahkan nilai IPM sebagai variabel komposit, kita dapat mengidentifikasi kemiripan atau perbedaan karakteristik pembangunan antar provinsi. Pendekatan seperti ini sangat cocok untuk dilakukan dengan metode data mining, karena mampu mengekstrak pola dari struktur data yang tidak selalu terlihat secara langsung.

Sebagai contoh, jika hasil klasterisasi menunjukkan adanya satu kelompok provinsi dengan AHH dan RLS tinggi namun PPP rendah, maka provinsi dalam kelompok tersebut cenderung unggul dalam kesehatan dan pendidikan tetapi masih menghadapi tantangan ekonomi. Sebaliknya, jika

terdapat klaster dengan PPP tinggi tetapi AHH rendah, maka ada indikasi bahwa capaian ekonomi belum sepenuhnya selaras dengan peningkatan kualitas hidup secara keseluruhan.

Penerapan data mining dalam konteks ini membantu menyederhanakan kompleksitas data pembangunan menjadi bentuk yang lebih mudah dipahami. Melalui visualisasi dan pengelompokan berbasis algoritma, struktur disparitas pembangunan menjadi lebih jelas dan terukur. Hasil ini kemudian dapat menjadi dasar refleksi, diskusi akademik, atau bahkan inspirasi untuk eksplorasi lanjutan dalam bidang analitik sosial.

2.2 Algoritma yang Digunakan

Penelitian ini menggunakan algoritma **K-Means** sebagai pendekatan utama untuk melakukan pengelompokan (**clustering**) terhadap provinsi-provinsi di Indonesia berdasarkan karakteristik pembangunan manusianya. Pemilihan algoritma ini bukan tanpa pertimbangan. Secara alami, data yang digunakan dalam penelitian ini bersifat numerik dan multidimensi, sehingga K-Means menjadi salah satu pilihan paling tepat karena mampu bekerja secara efisien dalam menangani data dengan karakteristik seperti itu.

K-Means dikenal luas sebagai salah satu algoritma clustering yang paling sederhana dan efektif dalam konteks ***unsupervised learning***. Sebelum memahami lebih jauh bagaimana K-Means bekerja, penting untuk terlebih dahulu memahami konsep dasar dari clustering itu sendiri.

2.2.1 Konsep Dasar Clustering

Clustering merupakan teknik dalam data mining yang bertujuan untuk mengelompokkan data ke dalam beberapa grup atau klaster berdasarkan tingkat kemiripan karakteristik antar data. Pendekatan ini termasuk dalam kategori *unsupervised learning*, yaitu pembelajaran tanpa label atau kategori yang sudah ditentukan sebelumnya. Artinya, sistem tidak diberi tahu kategori apa yang ada, melainkan diminta untuk menemukan struktur atau pola dalam data itu sendiri.

Dalam penelitian ini, setiap provinsi di Indonesia dianggap sebagai satu objek data. Karakteristik dari masing-masing provinsi direpresentasikan melalui empat variabel numerik: Angka Harapan Hidup (AHH), Rata-Rata Lama Sekolah (RLS), Pengeluaran per Kapita (PPP), serta Indeks Pembangunan Manusia (IPM). Ketiga variabel pertama merupakan komponen penyusun IPM, sedangkan IPM sebagai indeks gabungan ditambahkan untuk menangkap dinamika keseimbangan antar dimensi pembangunan.

Dengan mengelompokkan provinsi-provinsi berdasarkan kesamaan nilai dalam keempat dimensi tersebut, penelitian ini berusaha menggambarkan struktur perbedaan pembangunan manusia secara lebih menyeluruh. Klaster

yang terbentuk diharapkan mampu mencerminkan realitas di lapangan, di mana provinsi-provinsi dengan pola capaian indikator yang serupa berada dalam kelompok yang sama.

2.2.2 Pengertian dan Prinsip Kerja K-Means

K-Means merupakan algoritma clustering berbasis partisi yang dirancang untuk membagi data ke dalam sejumlah klaster yang saling eksklusif, berdasarkan tingkat kemiripan karakteristik. Setiap klaster direpresentasikan oleh sebuah titik pusat, atau yang disebut sebagai *centroid*, yang posisinya akan terus diperbarui selama proses berlangsung. Tujuan utama algoritma ini adalah meminimalkan total jarak kuadrat antara setiap titik data dengan centroid dari klaster tempat ia tergolong.

Salah satu keunggulan dari K-Means terletak pada kesederhanaan langkah kerjanya, namun tetap mampu menangkap struktur data yang kompleks dengan efisien. Dalam penelitian ini, algoritma K-Means digunakan untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan empat variabel utama: Angka Harapan Hidup (AHH), Rata-Rata Lama Sekolah (RLS), Pengeluaran per Kapita (PPP), dan nilai Indeks Pembangunan Manusia (IPM).

Proses kerja K-Means dalam penelitian ini dapat dijelaskan sebagai berikut:

1. **Menentukan jumlah klaster (K)**
   Nilai *K* tidak ditentukan secara sembarangan atau intuitif, tetapi dipilih melalui proses evaluasi kuantitatif menggunakan tiga pendekatan: Elbow Method, Silhouette Score, dan Davies-Bouldin Index.

   - *Elbow Method* membantu mengidentifikasi titik “tekukan” di mana penambahan jumlah klaster tidak lagi memberikan penurunan signifikan pada nilai *inertia*.
   - *Silhouette Score* digunakan untuk mengukur seberapa baik data berada di dalam klaster masing-masing dibandingkan dengan klaster lainnya.
   - *Davies-Bouldin Index* digunakan untuk mengevaluasi seberapa jauh antar-klaster dapat dibedakan.

   Kombinasi dari ketiga metrik ini menghasilkan pemilihan jumlah klaster yang optimal berdasarkan struktur sebenarnya dari data.

2. **Inisialisasi *centroid* awal**
   Tidak seperti pendekatan acak konvensional, penelitian ini menggunakan metode **K-Means++**, sebuah teknik inisialisasi yang secara bertahap memilih centroid awal dengan mempertimbangkan

penyebaran data. Tujuannya adalah untuk meningkatkan stabilitas hasil akhir dan menghindari jebakan pada solusi lokal yang buruk.

3. **Pengelompokan awal**
   Setelah centroid awal ditentukan, setiap titik data dihitung jaraknya ke seluruh centroid yang ada (menggunakan Euclidean Distance), dan kemudian dimasukkan ke dalam klaster dengan centroid terdekat.

4. **Pembaruan centroid**
   Untuk setiap klaster yang terbentuk, posisi centroid dihitung ulang berdasarkan rata-rata dari seluruh titik yang ada di dalam klaster tersebut.

5. **Iterasi hingga konvergen**
   Langkah 3 dan 4 akan diulang secara iteratif hingga tidak ada lagi perubahan signifikan dalam posisi centroid atau susunan anggota klaster. Pada titik ini, proses dianggap telah mencapai konvergensi, dan hasil akhir berupa label klaster untuk setiap provinsi dapat dievaluasi dan dianalisis lebih lanjut.

Dengan mengikuti prosedur tersebut secara sistematis dan berbasis evaluasi, pengelompokan yang dihasilkan oleh algoritma K-Means tidak hanya menggambarkan kedekatan matematis antar data, tetapi juga membawa makna substantif terkait struktur perbedaan pembangunan manusia antar provinsi di Indonesia.

2.2.3 Kelebihan dan Keterbatasan K-Means

Salah satu alasan utama digunakannya algoritma K-Means dalam penelitian ini adalah karena kesederhanaan dan efisiensinya. K-Means memiliki mekanisme kerja yang relatif mudah dipahami dan diimplementasikan, bahkan untuk dataset berukuran cukup besar. Algoritma ini juga tersedia secara luas dalam berbagai pustaka pemrograman seperti scikit-learn, sehingga sangat cocok digunakan dalam studi berbasis eksplorasi data.

Dari sisi performa, K-Means sangat efektif dalam menangani data numerik berdimensi rendah hingga menengah. Dalam konteks penelitian ini, variabel yang digunakan (IPM, AHH, RLS, dan PPP) merupakan data numerik yang sudah dinormalisasi, sehingga sangat sesuai dengan karakteristik algoritma K-Means. Selain itu, kecepatan konvergensinya juga menjadi nilai tambah tersendiri, terutama ketika jumlah observasi relatif banyak dan proses iterasi perlu dilakukan berulang.

Namun demikian, ada beberapa keterbatasan dari K-Means yang perlu diperhatikan agar hasil interpretasi tetap akurat:

- **Perlu menentukan jumlah klaster (K) di awal**
  K-Means tidak secara otomatis menentukan jumlah klaster. Jika nilai *K* ditetapkan tanpa dasar yang kuat, hasil klasterisasi bisa menyesatkan. Oleh karena itu, dalam penelitian ini digunakan pendekatan evaluatif seperti Elbow Method, Silhouette Score, dan Davies-Bouldin Index untuk memastikan nilai *K* yang dipilih benar-benar mencerminkan struktur data.

- **Sensitif terhadap inisialisasi centroid**
  Pemilihan titik awal *centroid* sangat mempengaruhi hasil akhir. Jika inisialisasi dilakukan secara sembarangan, algoritma bisa berhenti pada solusi lokal yang tidak optimal. Untuk mengatasi hal ini, metode **K-Means++** digunakan dalam penelitian ini agar inisialisasi lebih stabil dan representatif.

- **Asumsi bentuk klaster yang cenderung sferis dan seimbang**
  K-Means bekerja paling baik saat data terdistribusi dalam bentuk klaster yang bulat dan dengan ukuran yang relatif seragam. Jika data memiliki distribusi yang tidak simetris, ukuran klaster yang tidak seimbang, atau terdapat *outlier*, maka hasil pengelompokan bisa menjadi kurang akurat.

Dengan menyadari keterbatasan-keterbatasan tersebut sejak awal, algoritma K-Means tetap dapat digunakan secara efektif. Selama proses evaluasi dilakukan dengan seksama dan pemaknaan hasilnya disesuaikan dengan karakteristik model, pendekatan ini tetap mampu memberikan gambaran yang informatif terkait pola pembangunan manusia di Indonesia.

2.2.4 Penerapan K-Means dalam Penelitian

Dalam penelitian ini, algoritma K-Means digunakan untuk mengelompokkan seluruh provinsi di Indonesia berdasarkan indikator-indikator yang mencerminkan tingkat pembangunan manusianya. Berbeda dengan pendekatan umum yang hanya menggunakan komponen-komponen penyusun IPM, penelitian ini justru menggunakan empat variabel secara bersamaan: Angka Harapan Hidup (AHH), Rata-Rata Lama Sekolah (RLS), Pengeluaran per Kapita (PPP), dan nilai IPM itu sendiri. Penambahan IPM sebagai variabel input dimaksudkan untuk menangkap hubungan keseimbangan antar dimensi secara lebih utuh.

Sebelum proses klasterisasi dilakukan, seluruh data numerik terlebih dahulu dinormalisasi menggunakan teknik **Min-Max Scaling**. Hal ini penting karena keempat variabel yang digunakan memiliki rentang nilai yang sangat berbeda. Misalnya, nilai pengeluaran per kapita umumnya berada pada skala puluhan juta rupiah, sementara lama sekolah hanya berada di kisaran belasan tahun. Jika skala ini tidak diseragamkan, variabel dengan nilai besar akan secara tidak adil mendominasi proses pengelompokan. Normalisasi menjamin

bahwa setiap fitur memiliki kontribusi yang sebanding dalam penentuan struktur klaster.

Setelah proses normalisasi selesai, langkah berikutnya adalah menentukan jumlah klaster yang optimal. Untuk keperluan ini, digunakan tiga metrik evaluatif:

- **Elbow Method**, untuk melihat pada titik berapa penurunan *inertia* mulai melambat secara signifikan. Titik ini biasanya menjadi indikasi bahwa penambahan klaster tidak lagi memberikan manfaat berarti.

- **Silhouette Score**, untuk menilai seberapa jelas batas antar klaster terbentuk. Nilai yang tinggi menunjukkan bahwa klaster cukup padat dan terpisah dengan baik.

- **Davies-Bouldin Index**, untuk mengukur tingkat kemiripan antar klaster. Nilai yang rendah menunjukkan bahwa klaster yang terbentuk memiliki jarak dan bentuk yang baik satu sama lain.

Berdasarkan kombinasi ketiga metrik tersebut, dipilihlah nilai *K* yang paling representatif terhadap struktur data. Setelah jumlah klaster ditetapkan, algoritma K-Means dijalankan menggunakan **K-Means++** sebagai metode inisialisasi centroid. Proses klasterisasi berlangsung dalam beberapa iterasi, di mana titik-titik data dialokasikan ke klaster berdasarkan kedekatannya dengan centroid, lalu centroid diperbarui berdasarkan rata-rata posisi titik-titik dalam klaster tersebut. Proses ini diulang hingga posisi centroid stabil dan tidak mengalami perubahan berarti.

Hasil akhir dari proses ini adalah terbentuknya sejumlah klaster yang masing-masing terdiri dari provinsi-provinsi dengan karakteristik pembangunan manusia yang serupa. Klaster-klaster ini kemudian dianalisis lebih lanjut untuk mengidentifikasi pola umum yang muncul, seperti kelompok provinsi dengan nilai IPM tinggi namun ketimpangan antar komponennya besar, atau sebaliknya, provinsi dengan nilai IPM rendah namun relatif seimbang antar indikator. Selain itu, hasil klasterisasi juga ditata ulang berdasarkan rata-rata IPM agar label klaster memiliki urutan yang lebih bermakna.

Penerapan algoritma K-Means dalam penelitian ini tidak hanya menghasilkan pembagian wilayah secara numerik, tetapi juga membuka ruang interpretasi yang lebih dalam terkait dinamika pembangunan manusia di Indonesia. Hasil ini menjadi fondasi penting untuk analisis lanjutan pada bab-bab berikutnya.

2.2.5 Manfaat Penggunaan K-Means dalam Konteks IPM

Penerapan algoritma K-Means dalam konteks pembangunan manusia, khususnya dalam pengelompokan provinsi berdasarkan indikator IPM,

memberikan sejumlah manfaat yang signifikan dari sisi analisis data dan interpretasi hasil. Alih-alih melihat satu per satu nilai indikator secara terpisah, pendekatan ini memungkinkan kita untuk memahami bagaimana karakteristik pembangunan manusia terbentuk dalam pola-pola kelompok yang lebih luas dan bermakna.

Beberapa manfaat utama dari pendekatan klasterisasi ini antara lain:

- **Mengungkap struktur tersembunyi dalam data pembangunan**
  Dengan menggunakan K-Means, kita dapat melihat bagaimana provinsi-provinsi di Indonesia cenderung berkumpul dalam kelompok yang memiliki ciri khas tertentu. Misalnya, kelompok provinsi dengan IPM tinggi tetapi ketimpangan antara RLS dan PPP yang besar, atau kelompok lain yang cenderung seragam pada seluruh indikator namun berada pada tingkat pembangunan rendah. Pola-pola ini mungkin tidak terlihat jelas jika hanya dilihat dari tabel atau statistik deskriptif biasa.

- **Menyederhanakan keragaman data menjadi kategori yang lebih mudah dianalisis**
  Klasterisasi memungkinkan penyederhanaan data multivariat menjadi beberapa kelompok homogen. Ini membantu dalam mengurangi kompleksitas data, tanpa kehilangan esensi dari struktur informasi yang terkandung di dalamnya.

- **Menyediakan dasar untuk perbandingan dan eksplorasi lebih lanjut**
  Setelah klaster terbentuk, setiap kelompok dapat dianalisis secara lebih mendalam. Perbandingan antar klaster menjadi lebih bermakna karena didasarkan pada kesamaan karakteristik, bukan hanya pembagian administratif atau geografis. Hal ini memudahkan dalam mengidentifikasi wilayah-wilayah yang memiliki kebutuhan atau tantangan serupa.

- **Meningkatkan efektivitas visualisasi dan komunikasi hasil analisis**
  Hasil klasterisasi yang terstruktur memungkinkan pembuatan grafik, peta, dan visualisasi lain yang lebih mudah dibaca oleh berbagai kalangan. Dengan memanfaatkan warna, posisi, atau label klaster, hasil analisis menjadi lebih komunikatif dan dapat diinterpretasikan secara intuitif.

Secara keseluruhan, pemanfaatan algoritma K-Means dalam penelitian ini tidak hanya berfungsi sebagai alat teknis untuk pengelompokan, tetapi juga sebagai pendekatan konseptual untuk memahami pola pembangunan manusia di Indonesia secara lebih menyeluruh. Klaster yang terbentuk dapat menjadi jendela awal untuk menelusuri bagaimana pembangunan berlangsung di

berbagai wilayah, serta membantu mengarahkan perhatian pada kelompok provinsi yang menunjukkan dinamika perkembangan yang serupa.

BAB 3
METODOLOGI PENELITIAN

3.1 Data Science Methodology

Penelitian ini menggunakan pendekatan data science yang terstruktur, dimulai dari tahap eksplorasi data hingga evaluasi hasil klasterisasi. Tujuan dari metodologi ini adalah untuk menjawab pertanyaan penelitian secara objektif dan memastikan bahwa proses analisis berjalan sistematis dan berbasis bukti.

Pada inti proses analisis, digunakan algoritma **K-Means** untuk melakukan pengelompokan provinsi-provinsi di Indonesia berdasarkan indikator pembangunan manusia. K-Means merupakan algoritma *unsupervised learning* berbasis partisi, yang bertujuan membagi data ke dalam sejumlah klaster berdasarkan kemiripan karakteristik.

Secara umum, proses kerja K-Means terdiri dari dua tahap utama yang dilakukan secara berulang:

1. **Penentuan keanggotaan klaster**
   Setiap titik data (dalam hal ini, masing-masing provinsi) dihitung jaraknya terhadap setiap centroid, lalu ditugaskan ke klaster dengan centroid terdekat. Ukuran jarak yang digunakan adalah Euclidean Distance, sesuai standar umum dalam algoritma ini.

2. **Pembaruan posisi centroid**
   Setelah semua data terkelompok, posisi centroid setiap klaster diperbarui berdasarkan rata-rata posisi seluruh anggota klaster tersebut.

Proses di atas akan terus berulang hingga posisi centroid tidak lagi mengalami perubahan signifikan, atau algoritma telah mencapai kondisi konvergen. Karena K-Means sensitif terhadap inisialisasi awal, dalam penelitian ini digunakan teknik **K-Means++**, yaitu metode inisialisasi centroid yang lebih sistematis dan stabil, sehingga mampu menghasilkan klaster yang lebih baik secara konsisten.

3.2 Data Preparation

Tahap persiapan data merupakan pondasi penting dalam proses data science. Kualitas, kelengkapan, dan konsistensi data sangat menentukan hasil akhir dari analisis. Dalam penelitian ini, data yang digunakan adalah data Indeks Pembangunan Manusia (IPM) dan komponen-komponennya dari seluruh provinsi di Indonesia. Seluruh data bersumber dari publikasi resmi **Badan Pusat Statistik (BPS)**, sehingga tingkat kepercayaannya cukup tinggi.

Adapun variabel utama yang digunakan dalam analisis adalah sebagai berikut:

- **Angka Harapan Hidup (AHH)** – mewakili dimensi kesehatan.
- **Rata-Rata Lama Sekolah (RLS)** – mewakili dimensi pendidikan.
- **Pengeluaran per Kapita (PPP)** – mewakili dimensi ekonomi.
- **IPM** – sebagai gabungan ketiga dimensi untuk menangkap keseimbangan pembangunan.

Sebelum dilakukan pemodelan, data terlebih dahulu melalui tahapan *preprocessing* berikut:

- **Pembersihan Data (Data Cleaning)**
  Meskipun data dari BPS umumnya rapi, proses verifikasi awal tetap dilakukan untuk memastikan tidak ada nilai kosong, duplikat, atau inkonsistensi format. Dalam kasus ini, seluruh data valid dan tidak ditemukan masalah yang berarti.

- **Normalisasi Data**
  karena masing-masing variabel memiliki skala yang berbeda, dilakukan proses **min-max scaling** agar seluruh nilai berada pada rentang 0 hingga 1. Hal ini penting karena algoritma K-Means sensitif terhadap perbedaan skala antar fitur. Tanpa normalisasi, variabel dengan skala besar seperti PPP akan mendominasi proses pengelompokan, sehingga hasilnya menjadi bias.

3.3 Modeling with K-Means

Setelah data disiapkan dan dinormalisasi, tahap selanjutnya adalah membangun model klasterisasi menggunakan algoritma K-Means. Salah satu aspek penting dalam tahap ini adalah penentuan jumlah klaster yang optimal (*K*), karena K tidak ditentukan otomatis oleh algoritma.

Untuk menentukan nilai *K* yang paling representatif terhadap struktur data, digunakan beberapa pendekatan evaluatif:

- **Elbow Method**
  Dengan memplot nilai *inertia* terhadap berbagai nilai *K*, diperoleh titik "tekukan" pada grafik, yaitu titik di mana penambahan jumlah klaster tidak lagi memberikan pengurangan signifikan pada nilai inertia. Titik ini dianggap sebagai nilai *K* yang optimal.

- **Silhouette Score**
  Metrik ini digunakan untuk menilai seberapa baik data berada di dalam klaster masing-masing dan seberapa jauh dari klaster lainnya. Nilai yang lebih tinggi menunjukkan klasterisasi yang lebih baik.

- **Davies-Bouldin Index**
  Mengukur tingkat kemiripan antar klaster. Semakin rendah nilai indeks ini, semakin baik pemisahan antar kelompok.

Setelah nilai *K* ditentukan, model K-Means dilatih dengan menggunakan metode inisialisasi **K-Means++**. Proses klasterisasi kemudian dilakukan dalam beberapa iterasi, di mana setiap titik data dialokasikan ke klaster terdekat dan posisi centroid diperbarui terus-menerus hingga konvergen.

Hasil klasterisasi tidak hanya mencakup label klaster untuk masing-masing provinsi, tetapi juga distribusi nilai indikator dalam tiap kelompok. Hal ini menjadi dasar untuk analisis interpretatif pada tahap berikutnya.

3.4 Evaluation and Interpretation

Evaluasi hasil klasterisasi dilakukan baik secara kuantitatif maupun kualitatif. Tahap evaluasi hasil klasterisasi dilakukan melalui dua pendekatan: kuantitatif dan kualitatif.

Secara **kuantitatif**, evaluasi menggunakan metrik **Silhouette Coefficient** untuk mengukur kekompakan dan keterpisahan antar klaster. Nilai yang lebih tinggi menunjukkan bahwa klaster yang terbentuk memiliki struktur yang kuat dan terpisah dengan jelas.

Secara **kualitatif**, analisis dilakukan dengan melihat rata-rata nilai setiap indikator (AHH, RLS, PPP, dan IPM) dalam masing-masing klaster. Hasil ini membantu mengidentifikasi karakteristik khas setiap kelompok. Contohnya:

- Klaster dengan AHH dan RLS tinggi tetapi PPP rendah menunjukkan wilayah dengan capaian pendidikan dan kesehatan yang baik namun masih tertinggal secara ekonomi.
- Klaster dengan semua indikator rendah mencerminkan wilayah yang secara umum masih tertinggal dalam seluruh dimensi pembangunan manusia.
- Klaster dengan nilai tinggi di seluruh indikator menggambarkan wilayah dengan pembangunan yang relatif seimbang dan mapan.

Interpretasi semacam ini memberikan gambaran yang lebih menyeluruh mengenai struktur pembangunan antar provinsi di Indonesia. Hasilnya juga membuka ruang eksplorasi lebih lanjut untuk melihat apakah pola-pola ini konsisten dengan realitas sosial-ekonomi yang ada di lapangan.

BAB 4
PENGUMPULAN DAN PENGOLAHAN DATA

4.1 Pengumpulan Data

Seluruh proses analisis dalam penelitian ini diawali dengan tahapan pengumpulan data. Tahap ini merupakan fondasi utama karena kualitas dan kelengkapan data akan memengaruhi seluruh tahapan berikutnya, mulai dari preprocessing, modeling, hingga interpretasi akhir. Dalam konteks penelitian ini, pengumpulan data dilakukan secara mandiri dan sistematis untuk memastikan bahwa semua data yang dibutuhkan tersedia dalam bentuk yang relevan dan dapat dipercaya.

Sumber utama data adalah **Badan Pusat Statistik (BPS)**, yang merupakan lembaga resmi penyedia data statistik nasional di Indonesia. BPS secara rutin mempublikasikan indikator pembangunan melalui portal daring, dan menyediakan fitur unduhan data dalam berbagai format, termasuk CSV, sehingga sangat mendukung keperluan analisis berbasis data.

Fokus utama dalam pengumpulan ini adalah mendapatkan data dari **tiga indikator kunci** yang menjadi penyusun **Indeks Pembangunan Manusia (IPM)**:

- **Angka Harapan Hidup (AHH)**: mewakili dimensi kesehatan,
- **Rata-rata Lama Sekolah (RLS)**: menggambarkan aspek pendidikan,
- **Pengeluaran per Kapita yang Disesuaikan (PPP)**: mencerminkan daya beli dan kesejahteraan ekonomi.

Ketiga indikator ini tidak tersedia dalam bentuk satu dataset tunggal, melainkan tersebar dalam tiga halaman tabel statistik yang berbeda. Untuk masing-masing indikator, data diunduh untuk **tiga tahun terakhir (2022, 2023, dan 2024)**. Pengunduhan dilakukan melalui menu **dropdown download** yang tersedia di halaman masing-masing tabel, yang memungkinkan pengguna memilih dan mengunduh data dalam format **Comma-Separated Values (CSV)** untuk setiap tahun secara terpisah.

Berikut adalah tiga tautan resmi dari situs BPS tempat data diambil:

- **Angka Harapan Hidup (AHH):**
  https://www.bps.go.id/id/statistics-table/2/MjI3MyMy/angka-harapan-hidup--ahh--menurut-provinsi-dan-jenis-kelamin--menggunakan-uhh-hasil-sp2020-lf-.html

- **Rata-rata Lama Sekolah (RLS):**
  https://www.bps.go.id/id/statistics-table/2/MTQyOSMy/rata-rata-lama-sekolah-penduduk-umur-15-tahun-ke-atas-menurut-provinsi.html

- **Pengeluaran per Kapita Disesuaikan (PPP):** https://www.bps.go.id/id/statistics-table/2/NDE2IzI=/-metode-baru-pengeluaran-per-kapita-disesuaikan.html

Di setiap tautan tersebut, tersedia opsi untuk memilih tahun data melalui menu **"Pilih Tahun"** atau **dropdown download**, kemudian file dapat diunduh satu per satu dalam format CSV. Total ada **9 file data mentah** yang berhasil dikumpulkan:

- 3 file untuk AHH (2022, 2023, 2024),
- 3 file untuk RLS (2022, 2023, 2024),
- 3 file untuk PPP (2022, 2023, 2024).

Setiap file memiliki struktur tabulasi data per provinsi, dengan indikator sebagai kolom dan nama provinsi sebagai baris. Namun karena data disediakan secara terpisah per indikator dan per tahun, maka pada tahap pengumpulan ini, dataset masih sepenuhnya **berbentuk mentah**. Belum dilakukan penggabungan antar tahun maupun antar indikator.

Selain itu, ditemukan juga adanya beberapa **missing value** pada indikator tertentu di tahun-tahun tertentu. Hal ini lazim, terutama pada provinsi hasil pemekaran baru seperti **Papua Pegunungan, Papua Selatan, dan Papua Tengah**, yang belum memiliki data lengkap untuk seluruh indikator di semua tahun. Missing value ini belum ditangani pada tahap ini dan tetap dibiarkan sebagaimana adanya untuk selanjutnya diproses di tahap **pengolahan data**.

Perlu ditekankan bahwa hingga tahap ini, IPM **belum dihitung**. Yang dikumpulkan hanyalah ketiga komponennya dalam bentuk mentah. Penggabungan antar tahun dan antar indikator, serta perhitungan nilai IPM secara aktual, akan dibahas lebih lanjut dalam bagian **4.2 Pengolahan Data**, yang merupakan lanjutan langsung dari proses ini.

Dengan selesainya tahap ini, penelitian telah memiliki seluruh **data mentah** yang dibutuhkan yang kemudian diletakkan di 1 folder bernama "Raw Data" untuk proses analisis: valid, lengkap, dan siap untuk diolah lebih lanjut.

4.2 Pengolahan Data

Setelah seluruh data mentah berhasil dikumpulkan dan disimpan dalam folder bernama “Raw Data”, tahap berikutnya adalah melakukan proses pengolahan data. Pengolahan ini bertujuan untuk menyiapkan dataset yang bersih, terintegrasi, dan siap digunakan dalam analisis klasterisasi dengan algoritma K-Means. Tahapan pengolahan data ini terdiri dari enam bagian utama: *data preparation*, *data preprocessing*, *model development*, *evaluasi metrik*, *evaluasi performa model*, serta *visualisasi*.

4.2.1 Data Preparation

Tahap ini merupakan fondasi awal dari seluruh proses analisis, di mana data mentah dari masing-masing indikator yakni pengeluaran per kapita disesuaikan (PPP), rata-rata lama sekolah (RLS), dan angka harapan hidup (AHH) yang disatukan ke dalam satu struktur data yang bersih dan siap digunakan. Seluruh proses ini dilakukan secara terprogram dan sistematis agar menciptakan dataset yang lengkap dan konsisten untuk keperluan analisis klasterisasi..

```
# Import library yang digunakan

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from functools import reduce
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score,
davies_bouldin_score
```

Langkah pertama adalah memuat seluruh library yang akan digunakan sepanjang proses analisis. Library seperti pandas dan numpy digunakan untuk manipulasi data, matplotlib dan seaborn untuk visualisasi, serta sklearn untuk keperluan preprocessing dan algoritma machine learning, khususnya K-Means.

```
# --- 1. DATA PREPARATION ---

daftarProvinsi = [
   "ACEH", "SUMATERA UTARA", "SUMATERA BARAT", "RIAU",
"JAMBI", "SUMATERA SELATAN",
   "BENGKULU", "LAMPUNG", "KEP. BANGKA BELITUNG", "KEPULAUAN
RIAU", "DKI JAKARTA",
```

```
    "JAWA BARAT", "JAWA TENGAH", "DI YOGYAKARTA", "JAWA TIMUR",
"BANTEN", "BALI",
    "NUSA TENGGARA BARAT", "NUSA TENGGARA TIMUR", "KALIMANTAN
BARAT", "KALIMANTAN TENGAH",
    "KALIMANTAN SELATAN", "KALIMANTAN TIMUR", "KALIMANTAN
UTARA", "SULAWESI UTARA",
    "SULAWESI TENGAH", "SULAWESI SELATAN", "SULAWESI TENGGARA",
"GORONTALO",
    "SULAWESI BARAT", "MALUKU", "MALUKU UTARA", "PAPUA", "PAPUA
BARAT",
    "PAPUA TENGAH", "PAPUA PEGUNUNGAN", "PAPUA SELATAN", "PAPUA
BARAT DAYA"
]
```

daftarProvinsi berisi seluruh nama provinsi di Indonesia, termasuk empat provinsi baru hasil pemekaran Papua. Daftar ini penting untuk melakukan validasi dan penyaringan data pada proses penggabungan agar hanya data dari provinsi yang valid yang digunakan.

```
tahunProses = [2022, 2023, 2024]
pathDasar = './Raw Data/'
provinsiIndukPapua = ["PAPUA", "PAPUA BARAT"]
provinsiBaruPapua = ["PAPUA TENGAH", "PAPUA PEGUNUNGAN",
"PAPUA SELATAN", "PAPUA BARAT DAYA"]
```

Empat variabel di atas digunakan untuk mengatur tahun-tahun yang dianalisis, lokasi folder data mentah, serta pembeda antara provinsi hasil pemekaran dan provinsi induknya. Ini sangat penting karena provinsi baru kerap memiliki data yang belum lengkap.

```
listDataPerTahun = []
for tahun in tahunProses:
```

Selanjutnya, proses penggabungan dilakukan dalam sebuah loop untuk setiap tahun (2022–2024). Untuk tiap tahun, data dari tiga indikator dimuat, dibersihkan, dan digabungkan.

Pengolahan Data Pengeluaran per Kapita (PPP)

```
    dataPengeluaran = (
        pd.read_csv(f'{pathDasar}[Metode Baru] Pengeluaran per
Kapita Disesuaikan, {tahun}.csv', skiprows=1)
        .rename(columns={'Unnamed: 0': 'Provinsi','[Metode
Baru] Pengeluaran per Kapita Disesuaikan (Ribu
Rupiah/Orang/Tahun)': 'Pengeluaran per Kapita Disesuaikan
(PPP)'})
```

```
        .assign(Provinsi=lambda df:
df['Provinsi'].str.upper().replace({"D I YOGYAKARTA": "DI
YOGYAKARTA"}))
        .drop_duplicates(subset=['Provinsi'],
keep='first').query("Provinsi in
@daftarProvinsi").assign(Tahun=tahun)
    )
```

File pengeluaran diproses terlebih dahulu, dimulai dari pembacaan file CSV, pemberian nama kolom yang lebih deskriptif, standardisasi nama provinsi, dan filtering agar hanya memuat provinsi yang valid. Kemudian ditambahkan kolom tahun untuk menyatukan format.

Pengolahan Data Pendidikan (RLS)

```
    dataPendidikan = (
        pd.read_csv(f'{pathDasar}Rata-Rata Lama Sekolah
Penduduk Umur 15 Tahun ke Atas Menurut Provinsi, {tahun}.csv',
skiprows=1)
        .rename(columns={'Unnamed: 0': 'Provinsi','Rata-Rata
Lama Sekolah Penduduk Umur 15 Tahun ke Atas Menurut Provinsi':
'Rata-Rata Lama Sekolah (RLS)'})
        .assign(Provinsi=lambda df:
df['Provinsi'].str.upper().replace({"KEP. RIAU": "KEPULAUAN
RIAU"}))
        .drop_duplicates(subset=['Provinsi'],
keep='first').query("Provinsi in
@daftarProvinsi").assign(Tahun=tahun)
    )
```

Proses serupa dilakukan untuk data RLS. Standardisasi penulisan provinsi dilakukan agar penyatuan antar file berjalan mulus.

Pengolahan Data Kesehatan (AHH)

```
    dataAhh = pd.read_csv(f'{pathDasar}Angka Harapan Hidup
(AHH) Menurut Provinsi dan Jenis Kelamin (menggunakan UHH
hasil SP2020 LF), {tahun}.csv', skiprows=2)
    dataAhh = dataAhh.rename(columns={'Unnamed: 0':
'Provinsi'})
    dataAhh['Angka Harapan Hidup (AHH)'] =
(pd.to_numeric(dataAhh['Laki-laki'], errors='coerce') +
pd.to_numeric(dataAhh['Perempuan'], errors='coerce')) / 2
    dataAhh = dataAhh.assign(Provinsi=lambda df:
df['Provinsi'].str.upper().replace({"KEP. RIAU": "KEPULAUAN
RIAU"}))
```

```
    dataAhh = dataAhh.drop_duplicates(subset=['Provinsi'],
keep='first').query("Provinsi in
@daftarProvinsi").assign(Tahun=tahun)
```

Data AHH dihitung berdasarkan rata-rata angka harapan hidup pria dan wanita, lalu dilakukan validasi dan standardisasi nama provinsi seperti sebelumnya.

Penggabungan Semua Indikator

```
    listDfTahunan = [
        dataPengeluaran[['Provinsi', 'Tahun', 'Pengeluaran per
Kapita Disesuaikan (PPP)']],
        dataPendidikan[['Provinsi', 'Tahun', 'Rata-Rata Lama
Sekolah (RLS)']],
        dataAhh[['Provinsi', 'Tahun', 'Angka Harapan Hidup
(AHH)']]
    ]
    dataTahunGabung = reduce(lambda kiri, kanan: pd.merge(kiri,
kanan, on=['Provinsi', 'Tahun'], how='outer'), listDfTahunan)
    listDataPerTahun.append(dataTahunGabung)
```

Setelah ketiga data tahunan siap, masing-masing digabung menjadi satu DataFrame untuk tahun tersebut. Gabungan ini mencakup seluruh fitur yang relevan untuk perhitungan IPM.

Konsolidasi Seluruh Tahun

```
dataGabungan = pd.concat(listDataPerTahun,
ignore_index=True).drop_duplicates()
```

Seluruh data dari tahun 2022, 2023, dan 2024 disatukan ke dalam satu DataFrame dataGabungan. Dengan struktur ini, tiap provinsi dan tahun sudah tercakup lengkap.

Konversi Tipe Data

```
dataGabungan['Pengeluaran per Kapita Disesuaikan (PPP)'] =
pd.to_numeric(
    dataGabungan['Pengeluaran per Kapita Disesuaikan (PPP)'],
errors='coerce'
)
dataGabungan['Rata-Rata Lama Sekolah (RLS)'] = pd.to_numeric(
    dataGabungan['Rata-Rata Lama Sekolah (RLS)'],
errors='coerce'
)
```

Konversi ini memastikan seluruh nilai numerik berada dalam tipe data yang sesuai, khususnya sebelum dilakukan perhitungan atau imputasi.

Pengecekan dan Penanganan Missing Value

Setelah seluruh data digabung ke dalam satu DataFrame dataGabungan, dilakukan pengecekan awal untuk mengetahui apakah terdapat nilai yang hilang (missing value) pada kolom-kolom utama.

```
# --- Pengecekan Missing Value (Sebelum Imputasi) ---
print("\n--- Pengecekan Missing Value (Sebelum Imputasi) ---")
missingValuesSebelum = dataGabungan.isnull().sum()
print("Jumlah missing value per kolom:")
# Hanya tampilkan kolom yang memiliki missing value
print(missingValuesSebelum[missingValuesSebelum > 0])
print("-" * 50)
```

Pengecekan ini penting karena nilai kosong, jika tidak ditangani, dapat menyebabkan kesalahan dalam proses perhitungan IPM maupun pelatihan model K-Means. Umumnya, kekosongan terjadi pada provinsi-provinsi baru hasil pemekaran wilayah seperti Papua Tengah dan Papua Pegunungan yang datanya belum sepenuhnya tersedia di semua tahun.

Untuk itu, dilakukan proses **imputasi** nilai kosong, khususnya untuk provinsi baru. Nilai-nilai kosong ini diisi dengan menggunakan **rerata indikator dari provinsi induknya**, yaitu Papua dan Papua Barat, pada tahun yang sama. Kenapa menggunakan rerata dari 2 provinsi tersebut dan bukan dari seluruh provinsi? Karena sebelumnya sudah dicoba menggunakan rerata dari semua provinsi untuk mengatasi missing value yang ada namun hasilnya malah jauh terlalu tinggi sehingga kami memutuskan untuk menggunakan rerata dari provinsi yang ada di pulau yang sama, yang mana dalam kasus ini yakni Papua, dan menjadikan Papua dan Papua Barat sebagai induk karena 2 provinsi itulah yang tidak memiliki missing value di Pulau Papua..

```
# --- Imputasi Nilai Kosong Papua ---
kolomUntukImputasi = ['Pengeluaran per Kapita Disesuaikan
(PPP)', 'Rata-Rata Lama Sekolah (RLS)', 'Angka Harapan Hidup
(AHH)']
for kolom in kolomUntukImputasi:
   for tahun in tahunProses:
       rerataInduk =
dataGabungan[(dataGabungan['Provinsi'].isin(provinsiIndukPapua
)) & (dataGabungan['Tahun'] == tahun)][kolom].mean()
       kondisi =
((dataGabungan['Provinsi'].isin(provinsiBaruPapua)) &
(dataGabungan['Tahun'] == tahun) &
(dataGabungan[kolom].isnull()))
```

```
        dataGabungan.loc[kondisi, kolom] = rerataInduk
```

Imputasi dengan pendekatan rata-rata induk ini cukup masuk akal secara statistik dan kontekstual, mengingat provinsi-provinsi baru belum memiliki data sendiri yang konsisten dari BPS, dan secara geografis serta administratif masih sangat berkaitan erat dengan wilayah asalnya.

Setelah imputasi, dilakukan pengecekan ulang untuk memastikan bahwa semua missing value telah ditangani dengan baik.

```
# --- Pengecekan Missing Value (Setelah Imputasi) ---
print("\n--- Pengecekan Missing Value (Setelah Imputasi) ---")
missingValuesSetelah =
dataGabungan[kolomUntukImputasi].isnull().sum()
print("Jumlah missing value di kolom target setelah diisi:")
print(missingValuesSetelah)
print("-" * 50)

#Output
--- Pengecekan Missing Value (Sebelum Imputasi) ---
Jumlah missing value per kolom:
Pengeluaran per Kapita Disesuaikan (PPP)    8
Rata-Rata Lama Sekolah (RLS)                8
Angka Harapan Hidup (AHH)                   8
dtype: int64
--------------------------------------------------

--- Pengecekan Missing Value (Setelah Imputasi) ---
Jumlah missing value di kolom target setelah diisi:
Pengeluaran per Kapita Disesuaikan (PPP)    0
Rata-Rata Lama Sekolah (RLS)                0
Angka Harapan Hidup (AHH)                   0
dtype: int64
--------------------------------------------------
```

Pemeriksaan Data Sebelum Perhitungan IPM

Sebelum melangkah ke proses perhitungan IPM, data ditata ulang dalam bentuk yang lebih rapi agar mudah dibaca. Nama provinsi dan tahun digabung menjadi satu string, misalnya PAPUA 2023, agar setiap baris memiliki identitas unik sebagai unit analisis.

```
# --- Menampilkan Tabel Sebelum Menghitung IPM ---
dataTampil = dataGabungan.copy()
tipeKategoriProvinsi =
pd.CategoricalDtype(categories=daftarProvinsi, ordered=True)
dataTampil['Provinsi'] =
dataTampil['Provinsi'].astype(tipeKategoriProvinsi)
dataTampil = dataTampil.sort_values(by=['Provinsi',
'Tahun']).reset_index(drop=True)
```

```
dataTampil['Provinsi'] = dataTampil['Provinsi'].astype(str) +
' ' + dataTampil['Tahun'].astype(str)
dataTampil = dataTampil[['Provinsi', 'Pengeluaran per Kapita
Disesuaikan (PPP)', 'Rata-Rata Lama Sekolah (RLS)', 'Angka
Harapan Hidup (AHH)']]
for kolom in ['Rata-Rata Lama Sekolah (RLS)', 'Angka Harapan
Hidup (AHH)']:
   dataTampil[kolom] = pd.to_numeric(dataTampil[kolom],
errors='coerce').round(2)

print("\n--- Tabel Setelah Imputasi (Sebelum Perhitungan IPM)
---\n")
#print(dataTampil.to_string(index=False))
display(dataTampil.head(9).style.hide(axis="index"))

#Output
```

```
--- Tabel Setelah Imputasi (Sebelum Perhitungan IPM) ---
```

Provinsi	Pengeluaran per Kapita Disesuaikan (PPP)	Rata-Rata Lama Sekolah (RLS)	Angka Harapan Hidup (AHH)
ACEH 2022	9963.000000	9.790000	72.970000
ACEH 2023	10334.000000	9.890000	73.110000
ACEH 2024	10811.000000	9.950000	73.260000
SUMATERA UTARA 2022	10848.000000	9.990000	73.440000
SUMATERA UTARA 2023	11049.000000	10.070000	73.720000
SUMATERA UTARA 2024	11460.000000	10.180000	73.960000
SUMATERA BARAT 2022	11130.000000	9.510000	73.940000
SUMATERA BARAT 2023	11380.000000	9.590000	74.200000
SUMATERA BARAT 2024	11718.000000	9.720000	74.440000

Langkah ini dilakukan agar format data seragam dan mudah digunakan pada tahap visualisasi dan analisis berikutnya.

Perhitungan Indeks Pembangunan Manusia (IPM)

Setelah data bersih dan lengkap, dilakukan perhitungan IPM dengan menggunakan rumus standar yang mencerminkan tiga dimensi utama pembangunan manusia:

1. **Indeks Pendapatan**: berdasarkan logaritma dari pengeluaran per kapita
2. **Indeks Pendidikan**: berdasarkan RLS
3. **Indeks Kesehatan**: berdasarkan AHH

Ketiganya kemudian dihitung secara geometrik dan dikalikan dengan 100.

```
# --- Perhitungan IPM ---
def hitungIpm(ppp, rls, ahh):
   pppMin, pppMaks, rlsMin, rlsMaks, ahhMin, ahhMaks = 100,
75000, 0, 15, 20, 85
```

```
    indeksPendapatan = (np.log(ppp) - np.log(pppMin)) /
(np.log(pppMaks) - np.log(pppMin))
    indeksPendidikan = (rls - rlsMin) / (rlsMaks - rlsMin)
    indeksKesehatan = (ahh - ahhMin) / (ahhMaks - ahhMin)
    return ((indeksPendapatan * indeksPendidikan *
indeksKesehatan) ** (1/3) * 100)

# Hitung dan tambahkan kolom IPM
dataGabungan['Indeks Pembangunan Manusia (IPM)'] = hitungIpm(
    dataGabungan['Pengeluaran per Kapita Disesuaikan (PPP)'],
    dataGabungan['Rata-Rata Lama Sekolah (RLS)'],
    dataGabungan['Angka Harapan Hidup (AHH)']
)
```

Perhitungan ini dilakukan langsung dalam bentuk vektor tanpa looping, sehingga efisien dan konsisten untuk seluruh baris data.

Penyusunan Dataset Final untuk Preprocessing

Langkah terakhir dalam tahap data preparation adalah membentuk dataset akhir yang siap untuk tahap normalisasi dan klasterisasi.

```
# Format nama provinsi + tahun
dataGabungan['Provinsi'] =
dataGabungan['Provinsi'].astype(tipeKategoriProvinsi)
dataGabungan = dataGabungan.sort_values(by=['Provinsi',
'Tahun']).reset_index(drop=True)
dataGabungan['Provinsi'] =
dataGabungan['Provinsi'].astype(str) + ' ' +
dataGabungan['Tahun'].astype(str)

# Ambil kolom yang dibutuhkan untuk klasterisasi
dataFinal = dataGabungan[['Provinsi',
                          'Angka Harapan Hidup (AHH)',
                          'Rata-Rata Lama Sekolah (RLS)',
                          'Pengeluaran per Kapita Disesuaikan
(PPP)',
                          'Indeks Pembangunan Manusia
(IPM)']].copy()

# Pembulatan angka
dataFinal[['Angka Harapan Hidup (AHH)',
           'Rata-Rata Lama Sekolah (RLS)',
           'Pengeluaran per Kapita Disesuaikan (PPP)',
           'Indeks Pembangunan Manusia (IPM)']] = \
    dataFinal[['Angka Harapan Hidup (AHH)',
               'Rata-Rata Lama Sekolah (RLS)',
               'Pengeluaran per Kapita Disesuaikan (PPP)',
```

```
                'Indeks Pembangunan Manusia (IPM)']].round(2)

# Simpan ke file CSV
dataFinal.to_csv('dataPreprocessing.csv', index=False)

print(f"\nData preparation selesai. File disimpan ke
dataPreprocessing.csv")
print("\n--- Tabel Final untuk Preprocessing ---\n")
display(dataFinal.head(9).style.hide(axis="index"))

#Output
```

```
Data preparation selesai. File disimpan ke dataPreprocessing.csv

--- Tabel Final untuk Preprocessing ---
```

Provinsi	Angka Harapan Hidup (AHH)	Rata-Rata Lama Sekolah (RLS)	Pengeluaran per Kapita Disesuaikan (PPP)	Indeks Pembangunan Manusia (IPM)
ACEH 2022	72.970000	9.790000	9963.000000	71.770000
ACEH 2023	73.110000	9.890000	10334.000000	72.270000
ACEH 2024	73.260000	9.950000	10811.000000	72.710000
SUMATERA UTARA 2022	73.440000	9.990000	10848.000000	72.920000
SUMATERA UTARA 2023	73.720000	10.070000	11049.000000	73.330000
SUMATERA UTARA 2024	73.960000	10.180000	11460.000000	73.890000
SUMATERA BARAT 2022	73.940000	9.510000	11130.000000	72.080000
SUMATERA BARAT 2023	74.200000	9.590000	11380.000000	72.510000
SUMATERA BARAT 2024	74.440000	9.720000	11718.000000	73.090000

Hasil akhir dari tahap ini adalah file dataPreprocessing.csv, yang berisi seluruh fitur penting dalam format bersih dan rapi. File ini akan digunakan pada tahap selanjutnya yaitu **data preprocessing**.

4.2.2 Data Preprocessing

Setelah data berhasil dipersiapkan dan dihitung nilai IPM-nya, langkah selanjutnya adalah melakukan normalisasi terhadap seluruh fitur numerik yang akan digunakan dalam proses klasterisasi. Tujuan utama dari tahap ini adalah untuk menyamakan skala antar fitur, agar tidak ada satu fitur pun yang mendominasi proses pengelompokan hanya karena memiliki rentang nilai yang jauh lebih besar dibanding fitur lainnya.

Proses ini dilakukan dengan menggunakan **MinMaxScaler** dari library sklearn.preprocessing, yang mengubah nilai setiap fitur ke dalam rentang 0 hingga 1.

```
# --- 2. DATA PREPROCESSING ---

# Baca data
dataUntukScaling = pd.read_csv('dataPreprocessing.csv')

# Inisialisasi scaler
scaler = MinMaxScaler()
```

Langkah pertama adalah memuat kembali data yang telah disiapkan pada tahap sebelumnya, yaitu dataPreprocessing.csv. File ini berisi seluruh baris data dari seluruh provinsi dan tahun, lengkap dengan kolom IPM dan

tiga indikator utamanya: AHH, RLS, dan PPP. Kemudian objek MinMaxScaler diinisialisasi. Scaler ini akan digunakan untuk menghitung nilai minimum dan maksimum dari setiap fitur, lalu melakukan transformasi berdasarkan rumus:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Sehingga setiap nilai pada fitur akan dipetakan ke rentang antara 0 hingga 1.

```
# Pilih semua kolom numerik untuk di-scaling
fiturNumerik = ['Angka Harapan Hidup (AHH)',
                'Rata-Rata Lama Sekolah (RLS)',
                'Pengeluaran per Kapita Disesuaikan (PPP)',
                'Indeks Pembangunan Manusia (IPM)']
```

Empat fitur numerik utama yang digunakan dalam proses klasterisasi ditentukan secara eksplisit. Fitur-fitur ini adalah hasil dari pengolahan data yang sudah bersih, lengkap, dan konsisten, dan mencerminkan dimensi utama dalam pembangunan manusia.

```
# Simpan hasil scaling dalam DataFrame baru
dataScaled =
scaler.fit_transform(dataUntukScaling[fiturNumerik])
dataScaledDF = pd.DataFrame(dataScaled, columns=[col +
'_Scaled' for col in fiturNumerik])
```

Setelah itu, proses fit_transform() dilakukan pada keempat fitur numerik tersebut. Hasil scaling disimpan dalam sebuah DataFrame baru dataScaledDF, dengan setiap nama kolom ditambahkan akhiran _Scaled agar tetap bisa dibedakan dengan kolom aslinya.

```
# Gabungkan dengan kolom Provinsi
dataUntukScalingScaled =
pd.concat([dataUntukScaling[['Provinsi']], dataScaledDF],
axis=1)
```

Langkah ini menyatukan kembali hasil scaling dengan kolom Provinsi, sehingga setiap baris data tetap memiliki identitas yang unik dan kontekstual (misalnya: *"JAWA TENGAH 2023"*) saat dilakukan visualisasi atau pengelompokan nanti.

```
# Simpan hasilnya
namaFileOutput = 'dataScaled.csv'
```

```
dataUntukScalingScaled.to_csv(namaFileOutput, index=False)
```

Data yang telah dinormalisasi kemudian disimpan dalam file baru bernama dataScaled.csv. File ini merupakan output akhir dari tahap preprocessing, dan akan digunakan sebagai input langsung pada proses pelatihan model K-Means di tahap berikutnya.

```
print(f"Scaling dengan MinMaxScaler selesai. Data disimpan ke
{namaFileOutput}\n")
print("--- Tabel Dengan Semua Fitur yang Sudah di-Scaling
---\n")
display(dataUntukScalingScaled.head(9).style.hide(axis="index"
))

#Output
```

```
Scaling dengan MinMaxScaler selesai. Data disimpan ke dataScaled.csv

--- Tabel Dengan Semua Fitur yang Sudah di-Scaling ---
```

Provinsi	Angka Harapan Hidup (AHH)_Scaled	Rata-Rata Lama Sekolah (RLS)_Scaled	Pengeluaran per Kapita Disesuaikan (PPP)_Scaled	Indeks Pembangunan Manusia (IPM)_Scaled
ACEH 2022	0.641531	0.733959	0.298751	0.670178
ACEH 2023	0.657773	0.749609	0.324793	0.688340
ACEH 2024	0.675174	0.758998	0.358276	0.704323
SUMATERA UTARA 2022	0.696056	0.765258	0.360873	0.711951
SUMATERA UTARA 2023	0.728538	0.777778	0.374982	0.726843
SUMATERA UTARA 2024	0.756381	0.794992	0.403833	0.747185
SUMATERA BARAT 2022	0.754060	0.690141	0.380668	0.681438
SUMATERA BARAT 2023	0.784223	0.702660	0.398217	0.697058
SUMATERA BARAT 2024	0.812065	0.723005	0.421943	0.718126

Sebagai penutup, ditampilkan cuplikan dari tabel hasil scaling untuk melihat bahwa nilai-nilainya telah berada dalam rentang yang seragam. Tampilan ini penting sebagai validasi visual bahwa proses normalisasi telah berjalan sebagaimana mestinya.

4.2.3 Model Development

Setelah data dinormalisasi, langkah berikutnya adalah membangun model K-Means untuk melakukan proses klasterisasi terhadap data IPM provinsi. Tahapan ini melibatkan pemilihan jumlah klaster optimal, pelatihan model, serta penyesuaian label klaster agar lebih bermakna secara interpretatif.

```
# --- Memuat Data Hasil Preprocessing ---
dataKlaster = pd.read_csv('dataScaled.csv')

# Siapkan data untuk clustering (4 fitur hasil scaling)
fiturUntukKlaster = [
    'Angka Harapan Hidup (AHH)_Scaled',
    'Rata-Rata Lama Sekolah (RLS)_Scaled',
    'Pengeluaran per Kapita Disesuaikan (PPP)_Scaled',
    'Indeks Pembangunan Manusia (IPM)_Scaled'
]
dataUntukKlaster = dataKlaster[fiturUntukKlaster].values
```

Langkah pertama adalah memuat file dataScaled.csv yang merupakan hasil dari tahap normalisasi sebelumnya. File ini sudah berisi seluruh data provinsi dengan fitur-fitur yang telah disesuaikan skala nilainya ke dalam rentang 0–1. Empat fitur hasil scaling kemudian dipilih sebagai input utama untuk algoritma K-Means. Keempat fitur ini mewakili tiga dimensi utama IPM serta IPM itu sendiri sebagai kompositnya, sehingga pengelompokan akan mempertimbangkan semua aspek penting dalam pembangunan manusia.

Penentuan Jumlah Klaster Optimal

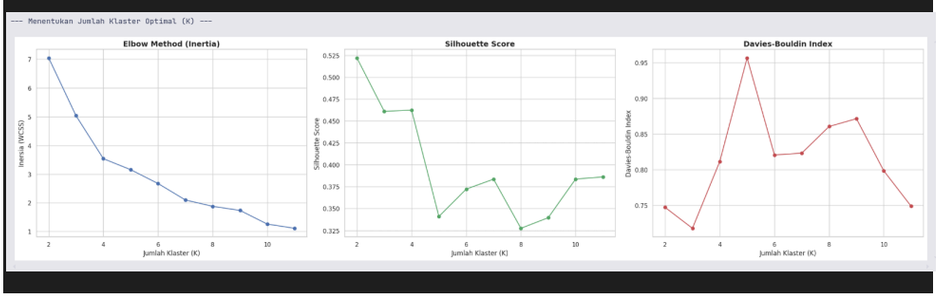
Sebelum membangun model akhir, dilakukan proses eksplorasi untuk menentukan jumlah klaster (K) terbaik. Evaluasi ini menggunakan tiga metrik utama:

1. **Inertia (Within-Cluster Sum of Squares)** – digunakan dalam metode Elbow.
2. **Silhouette Score** – mengukur seberapa baik pemisahan antar klaster.
3. **Davies-Bouldin Index** – mengevaluasi kemiripan antar klaster, makin kecil makin baik.

```
# --- Penentuan Jumlah Klaster (K) Optimal ---
print("--- Menentukan Jumlah Klaster Optimal (K) ---")
rentangK = range(2, 12)
inersia = []
skorSiluet = []
skorDaviesBouldin = []

for k in rentangK:
   # Inisialisasi dan latih model KMeans
   modelKMeans = KMeans(n_clusters=k, n_init='auto',
random_state=42)
   modelKMeans.fit(dataUntukKlaster)

   # Dapatkan label klaster
   label = modelKMeans.labels_

   # Simpan metrik evaluasi
   inersia.append(modelKMeans.inertia_)
   skorSiluet.append(silhouette_score(dataUntukKlaster,
label))

skorDaviesBouldin.append(davies_bouldin_score(dataUntukKlaster
, label))
```

Model KMeans dilatih untuk setiap nilai K dari 2 hingga 11. Hasil dari masing-masing pelatihan disimpan dalam tiga list terpisah untuk setiap metrik evaluasi.

```
# Visualisasi hasil evaluasi untuk berbagai nilai K
sns.set_theme(style="whitegrid")
fig, axes = plt.subplots(1, 3, figsize=(24, 6))

# Plot Elbow Method (Inertia/WCSS)
axes[0].plot(rentangK, inersia, 'bo-')
axes[0].set_title('Elbow Method (Inertia)', fontsize=14,
weight='bold')
axes[0].set_xlabel('Jumlah Klaster (K)')
axes[0].set_ylabel('Inersia (WCSS)')

# Plot Silhouette Score
axes[1].plot(rentangK, skorSiluet, 'go-')
axes[1].set_title('Silhouette Score', fontsize=14,
weight='bold')
axes[1].set_xlabel('Jumlah Klaster (K)')
axes[1].set_ylabel('Silhouette Score')

# Plot Davies-Bouldin Index
axes[2].plot(rentangK, skorDaviesBouldin, 'ro-')
axes[2].set_title('Davies-Bouldin Index', fontsize=14,
weight='bold')
axes[2].set_xlabel('Jumlah Klaster (K)')
axes[2].set_ylabel('Davies-Bouldin Index')

plt.tight_layout()
plt.show()

#Output
```



Visualisasi tiga metrik dilakukan secara berdampingan agar memudahkan identifikasi titik optimal. Grafik pertama (inertia) digunakan dalam Elbow Method, di mana titik "tekukan" atau pelandaian dianggap sebagai jumlah klaster optimal. Grafik kedua dan ketiga masing-masing menunjukkan Silhouette Score dan Davies-Bouldin Index untuk nilai K yang diuji.

Evaluasi dan Pemilihan K Optimal

```
# --- Evaluasi Hasil Klasterisasi ---
```

```
# Cetak semua skor evaluasi
print("\n--- Evaluasi K untuk Setiap Metrik ---")
print(f"{'K':<5}{'Inertia':<15}{'Silhouette':<15}{'Davies-Bouldin':<20}")
for i, k in enumerate(rentangK):

print(f"{k:<5}{inersia[i]:<15.2f}{skorSiluet[i]:<15.4f}{skorDaviesBouldin[i]:<20.4f}")
```

Seluruh hasil evaluasi ditampilkan secara eksplisit agar dapat ditinjau secara numerik. Evaluasi ini penting untuk menentukan dengan argumen yang kuat berapa jumlah klaster yang paling representatif.

```
# Tentukan K optimal berdasarkan dua metrik:
# 1. Silhouette Score maksimum
# 2. Davies-Bouldin Index minimum

# K dengan silhouette tertinggi
k_silhouette = rentangK[np.argmax(skorSiluet)]
# K dengan Davies-Bouldin Index terendah
k_dbi = rentangK[np.argmin(skorDaviesBouldin)]

print(f"\nBerdasarkan Silhouette Score tertinggi → K optimal = {k_silhouette}")
print(f"Berdasarkan Davies-Bouldin Index terendah → K optimal = {k_dbi}")
```

Dua kandidat nilai K optimal diperoleh dari Silhouette Score dan Davies-Bouldin Index. Jika keduanya berbeda, maka dilakukan pemilihan berdasarkan pertimbangan stabilitas metrik atau kompromi yang paling rasional.

```
# --- Tentukan K final ---
# Jika keduanya sama, langsung pakai
# Jika berbeda, pilih salah satu atau kompromi (misalnya pakai Silhouette karena lebih stabil)
kOptimal = k_dbi  # atau bisa pakai k_dbi atau rumus kompromi jika kamu mau
print(f"\nJumlah klaster yang digunakan (berdasarkan evaluasi) adalah: {kOptimal}")
```

Dalam kasus ini, nilai K yang dipilih adalah hasil dari Davies-Bouldin Index karena dianggap memberikan klaster yang paling terpisah secara jelas dan lebih representatif pada konteks pembangunan manusia.

Penerapan Model KMeans Final

```
# --- Penerapan Model K-Means dengan K Optimal ---
print(f"\nMenerapkan clustering dengan K={kOptimal}...")

# Latih model final dengan K optimal
modelFinal = KMeans(n_clusters=kOptimal, n_init='auto',
random_state=42)
labelPrediksi = modelFinal.fit_predict(dataUntukKlaster)

# Tambahkan hasil label klaster ke DataFrame
dataKlaster['Klaster'] = labelPrediksi

#Output
```

```
--- Evaluasi K untuk Setiap Metrik ---
K    Inertia        Silhouette     Davies-Bouldin
2    7.05           0.5219         0.7470
3    5.04           0.4610         0.7177
4    3.54           0.4622         0.8114
5    3.15           0.3410         0.9564
6    2.67           0.3722         0.8204
7    2.10           0.3835         0.8230
8    1.88           0.3274         0.8603
9    1.73           0.3396         0.8713
10   1.26           0.3836         0.7983
11   1.11           0.3861         0.7487

Berdasarkan Silhouette Score tertinggi → K optimal = 2
Berdasarkan Davies-Bouldin Index terendah → K optimal = 3

Jumlah klaster yang digunakan (berdasarkan evaluasi) adalah: 3

Menerapkan clustering dengan K=3...
```

Model akhir dilatih ulang menggunakan nilai K terbaik yang telah dipilih. Label hasil klaster kemudian ditambahkan ke DataFrame dataKlaster agar bisa digunakan pada analisis selanjutnya.

Penyesuaian dan Pengurutan Label Klaster

```
# --- Analisis dan Pengurutan Hasil Klaster ---
# Agar label klaster lebih bermakna (0=terendah, 1=menengah,
dst.), kita urutkan berdasarkan rata-rata IPM
rerataIpmPerKlaster = dataKlaster.groupby('Klaster')['Indeks
Pembangunan Manusia (IPM)_Scaled'].mean().sort_values()

# Buat pemetaan dari label acak ke label berurutan
pemetaanKlaster = {labelAsli: labelBaru for labelBaru,
labelAsli in enumerate(rerataIpmPerKlaster.index)}

# Terapkan pemetaan ke kolom Klaster
```

```
dataKlaster['Klaster'] =
dataKlaster['Klaster'].map(pemetaanKlaster)

# Urutkan hasil akhir agar mudah dibaca
dataKlaster = dataKlaster.sort_values(by=['Klaster',
'Provinsi']).reset_index(drop=True)
```

Secara default, label klaster dari K-Means tidak memiliki makna urutan tertentu. Oleh karena itu, dilakukan pemetaan ulang terhadap label agar label 0 menjadi klaster dengan IPM terendah, label 1 untuk yang menengah, dan seterusnya. Ini akan sangat membantu pada tahap visualisasi dan interpretasi hasil.

```
dataKlaster.to_csv('hasilClustering.csv', index=False)

print(f"\nClustering selesai. Data hasil clustering disimpan
ke hasilClustering.csv\n")
print("--- Tabel Hasil Akhir Clustering ---\n")
#print(dataKlaster.to_string(index=False))

display(dataKlaster.head(150).style.hide(axis="index"))

#Output
```

```
Clustering selesai. Data hasil clustering disimpan ke hasilClustering.csv

--- Tabel Hasil Akhir Clustering ---
```

Provinsi	Angka Harapan Hidup (AHH)_Scaled	Rata-Rata Lama Sekolah (RLS)_Scaled	Pengeluaran per Kapita Disesuaikan (PPP)_Scaled	Indeks Pembangunan Manusia (IPM)_Scaled	Klaster
GORONTALO 2022	0.328306	0.514867	0.349572	0.508899	0
GORONTALO 2023	0.359629	0.528951	0.376386	0.528151	0
GORONTALO 2024	0.386311	0.553991	0.409378	0.554668	0
MALUKU 2022	0.321346	0.824726	0.222448	0.651653	0
MALUKU 2023	0.354988	0.826291	0.250667	0.665819	0
MALUKU 2024	0.381671	0.835681	0.279166	0.683255	0
MALUKU UTARA 2022	0.357309	0.704225	0.188895	0.580458	0
MALUKU UTARA 2023	0.390951	0.705790	0.219500	0.596077	0
MALUKU UTARA 2024	0.424594	0.719875	0.253615	0.618598	0
NUSA TENGGARA BARAT 2022	0.495360	0.502347	0.349151	0.524155	0
NUSA TENGGARA TIMUR 2022	0.453596	0.492958	0.152323	0.458046	0
NUSA TENGGARA TIMUR 2023	0.484919	0.502347	0.178366	0.476571	0
NUSA TENGGARA TIMUR 2024	0.515081	0.528951	0.198442	0.503088	0
PAPUA 2022	0.068445	0.345853	0.101011	0.295677	0
PAPUA 2023	0.090487	0.350548	0.130212	0.311660	0
PAPUA 2024	0.357309	0.876369	0.374140	0.725754	0
PAPUA BARAT 2022	0.097448	0.788732	0.168047	0.581547	0
PAPUA BARAT 2023	0.129930	0.788732	0.189316	0.593534	0
PAPUA BARAT 2024	0.125290	0.746479	0.217465	0.579005	0
PAPUA BARAT DAYA 2022	0.082367	0.566510	0.134529	0.445332	0
PAPUA BARAT DAYA 2023	0.110209	0.569640	0.159764	0.459135	0
PAPUA BARAT DAYA 2024	0.305104	0.841941	0.212411	0.655285	0
PAPUA PEGUNUNGAN 2022	0.082367	0.566510	0.134529	0.445332	0
PAPUA PEGUNUNGAN 2023	0.110209	0.569640	0.159764	0.459135	0
PAPUA PEGUNUNGAN 2024	0.000000	0.000000	0.000000	0.000000	0
PAPUA SELATAN 2022	0.082367	0.566510	0.134529	0.445332	0
PAPUA SELATAN 2023	0.110209	0.569640	0.159764	0.459135	0
PAPUA SELATAN 2024	0.122970	0.608764	0.284220	0.520160	0
PAPUA TENGAH 2022	0.082367	0.566510	0.134529	0.445332	0
PAPUA TENGAH 2023	0.110209	0.569640	0.159764	0.459135	0
PAPUA TENGAH 2024	0.091647	0.167449	0.147550	0.190701	0
SULAWESI BARAT 2022	0.350348	0.527387	0.256282	0.496549	0
SULAWESI BARAT 2023	0.390951	0.528951	0.281553	0.509626	0
SULAWESI BARAT 2024	0.423434	0.541471	0.315948	0.530331	0
SULAWESI TENGAH 2022	0.359629	0.636933	0.280008	0.569197	0
SULAWESI TENGAH 2023	0.379350	0.644757	0.311807	0.585180	0
SULAWESI TENGAH 2024	0.399072	0.654147	0.338972	0.600436	0
ACEH 2022	0.641531	0.733959	0.298751	0.670178	1

Terakhir, hasil akhir dari klasterisasi disimpan dalam file hasilClustering.csv. File ini akan menjadi basis untuk tahap evaluasi metrik dan analisis tren IPM dari waktu ke waktu.

4.2.4 Matrix Evaluation

Setelah proses klasterisasi selesai dilakukan, tahap selanjutnya adalah melakukan evaluasi secara kuantitatif untuk mengukur kualitas dari model klasterisasi yang telah dibentuk. Evaluasi ini bertujuan untuk mengetahui seberapa baik model dalam memisahkan data ke dalam kelompok yang bermakna, menggunakan tiga metrik utama: **Inertia**, **Silhouette Score**, dan **Davies-Bouldin Index**. Masing-masing metrik memberikan sudut pandang yang berbeda mengenai performa klasterisasi.

Berikut adalah kode yang digunakan untuk melakukan evaluasi model berdasarkan tiga metrik tersebut:

```
# --- Memuat Data Hasil Clustering ---
dataHasil = pd.read_csv('hasilClustering.csv')

# Ambil semua fitur yang digunakan dalam clustering (harus
yang scaled)
fiturYangDigunakan = [
    'Indeks Pembangunan Manusia (IPM)_Scaled',
    'Angka Harapan Hidup (AHH)_Scaled',
    'Rata-Rata Lama Sekolah (RLS)_Scaled',
    'Pengeluaran per Kapita Disesuaikan (PPP)_Scaled'
]

dataUntukMetrik = dataHasil[fiturYangDigunakan].values
labelKlaster = dataHasil['Klaster'].values

# --- Perhitungan Metrik Kuantitatif ---
print("--- Menghitung Metrik Evaluasi ---")

# 1. Inertia (Within-Cluster Sum of Squares)
kOptimal = dataHasil['Klaster'].nunique()
modelFinal = KMeans(n_clusters=kOptimal, n_init='auto',
random_state=42)
modelFinal.fit(dataUntukMetrik)
nilaiInersia = modelFinal.inertia_
print(f"Inertia (WCSS): {nilaiInersia:.2f}")

# 2. Silhouette Score
skorSiluet = silhouette_score(dataUntukMetrik, labelKlaster)
print(f"Silhouette Score: {skorSiluet:.4f}")

# 3. Davies-Bouldin Index
skorDaviesBouldin = davies_bouldin_score(dataUntukMetrik,
labelKlaster)
print(f"Davies-Bouldin Index: {skorDaviesBouldin:.4f}")

#Output
```

Langkah pertama yang dilakukan adalah memuat kembali file hasil klasterisasi (hasilClustering.csv) yang sebelumnya sudah berisi label klaster untuk tiap entitas (provinsi-tahun). Dari dataset tersebut, hanya kolom-kolom numerik yang digunakan dalam proses clustering yang dipilih kembali, yaitu empat fitur yang telah melalui proses normalisasi sebelumnya.

Kode kemudian memisahkan data numerik (dataUntukMetrik) dan label klaster (labelKlaster) untuk selanjutnya digunakan dalam evaluasi.

Tiga metrik yang dihitung adalah sebagai berikut:

1. **Inertia (Within-Cluster Sum of Squares)**
   Inertia mengukur total jarak kuadrat antara titik data dalam suatu klaster dan centroid-nya. Semakin kecil nilai inertia, semakin kompak dan rapat kelompok yang terbentuk. Ini penting karena dalam K-Means, tujuan utama algoritma adalah meminimalkan jarak internal dalam klaster. Nilai inertia dihitung ulang dengan melatih kembali model menggunakan data hasil scaling.

2. **Silhouette Score**
   Metrik ini mengevaluasi seberapa dekat suatu titik ke anggota klasternya sendiri dibandingkan dengan klaster lain. Skor berkisar dari -1 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa data dikelompokkan dengan baik. Dalam implementasi ini, skor dihitung menggunakan silhouette_score dari pustaka sklearn.metrics.

3. **Davies-Bouldin Index**
   Metrik ini mengukur rasio antara jarak antar-klaster dan ukuran klaster. Nilai yang lebih rendah menunjukkan klaster yang lebih terpisah secara baik dan tidak saling tumpang tindih. Davies-Bouldin Index cocok sebagai pelengkap karena berbanding terbalik dengan Silhouette Score dalam interpretasi nilai.

Dengan ketiga metrik ini, model tidak hanya dinilai dari kekompakan tiap klaster secara internal, tetapi juga dari seberapa baik klaster tersebut terpisah dari yang lain. Ini memastikan evaluasi menyeluruh dari model K-Means yang dibangun.

Hasil dari evaluasi ini akan menjadi landasan dalam memahami seberapa optimal model dalam membagi data berdasarkan dimensi

pembangunan manusia. Tahapan ini juga krusial untuk memvalidasi bahwa jumlah klaster yang dipilih memang mencerminkan struktur alami dari data.

4.2.5 Performance Model Evaluation

Setelah melakukan klasterisasi dan mengevaluasi kualitasnya secara kuantitatif, tahap selanjutnya adalah menilai **konsistensi dan dinamika klasterisasi** dari waktu ke waktu. Dengan kata lain, kita ingin mengetahui bagaimana posisi setiap provinsi dalam klaster berubah selama periode 2022 hingga 2024. Tahapan ini sangat penting untuk memahami apakah provinsi mengalami peningkatan, stagnasi, atau bahkan penurunan dalam hal pembangunan manusia.

Untuk itu, digunakan pendekatan analisis longitudinal sederhana dengan menyusun **pivot table** dari hasil klasterisasi dan mengidentifikasi tren pergerakannya. Berikut adalah implementasi lengkapnya:

```
# --- ANALISIS PERGERAKAN KLASTER ---

# Muat data hasil clustering
dataHasil = pd.read_csv('hasilClustering.csv')

dataHasil['Tahun'] =
pd.to_numeric(dataHasil['Provinsi'].str[-4:])
dataHasil['NamaProvinsi'] = dataHasil['Provinsi'].str[:-5]

tabelPerubahan = dataHasil.pivot_table(
   index='NamaProvinsi',
   columns='Tahun',
   values='Klaster'
).rename_axis(index='Provinsi', columns=None)

# --- Analisis Tren Kenaikan, Stagnan, atau Penurunan ---
def tentukanTren(row):
   # Periksa apakah data untuk tahun awal dan akhir tersedia
   if pd.notna(row[2022]) and pd.notna(row[2024]):
       if row[2024] > row[2022]:
           return 'Peningkatan'
       elif row[2024] < row[2022]:
           return 'Penurunan'
       else:
           return 'Stagnan'

tabelPerubahan['Tren'] = tabelPerubahan.apply(tentukanTren,
axis=1)

# Tampilkan tabel utama
print("--- Tabel Perubahan Klaster Provinsi (2022-2024) ---")
```

```
display(tabelPerubahan[[2022, 2023, 2024,
'Tren']].rename(columns={2022:'Klaster 2022', 2023:'Klaster
2023', 2024:'Klaster 2024'}))


#Output
```

```
--- Tabel Perubahan Klaster Provinsi (2022-2024) ---
```

Provinsi	Klaster 2022	Klaster 2023	Klaster 2024	Tren
ACEH	1.0	1.0	1.0	Stagnan
BALI	2.0	2.0	2.0	Stagnan
BANTEN	1.0	1.0	1.0	Stagnan
BENGKULU	1.0	1.0	1.0	Stagnan
DI YOGYAKARTA	2.0	2.0	2.0	Stagnan
DKI JAKARTA	2.0	2.0	2.0	Stagnan
GORONTALO	0.0	0.0	0.0	Stagnan
JAMBI	1.0	1.0	1.0	Stagnan
JAWA BARAT	1.0	1.0	1.0	Stagnan
JAWA TENGAH	1.0	1.0	1.0	Stagnan
JAWA TIMUR	1.0	1.0	1.0	Stagnan
KALIMANTAN BARAT	1.0	1.0	1.0	Stagnan
KALIMANTAN SELATAN	1.0	1.0	1.0	Stagnan
KALIMANTAN TENGAH	1.0	1.0	1.0	Stagnan
KALIMANTAN TIMUR	1.0	2.0	2.0	Peningkatan
KALIMANTAN UTARA	1.0	1.0	1.0	Stagnan
KEP. BANGKA BELITUNG	1.0	1.0	1.0	Stagnan
KEPULAUAN RIAU	2.0	2.0	2.0	Stagnan
LAMPUNG	1.0	1.0	1.0	Stagnan
MALUKU	0.0	0.0	0.0	Stagnan
MALUKU UTARA	0.0	0.0	0.0	Stagnan
NUSA TENGGARA BARAT	0.0	1.0	1.0	Peningkatan
NUSA TENGGARA TIMUR	0.0	0.0	0.0	Stagnan
PAPUA	0.0	0.0	0.0	Stagnan
PAPUA BARAT	0.0	0.0	0.0	Stagnan
PAPUA BARAT DAYA	0.0	0.0	0.0	Stagnan
PAPUA PEGUNUNGAN	0.0	0.0	0.0	Stagnan
PAPUA SELATAN	0.0	0.0	0.0	Stagnan
PAPUA TENGAH	0.0	0.0	0.0	Stagnan
RIAU	1.0	1.0	1.0	Stagnan
SULAWESI BARAT	0.0	0.0	0.0	Stagnan
SULAWESI SELATAN	1.0	1.0	1.0	Stagnan
SULAWESI TENGAH	0.0	0.0	0.0	Stagnan
SULAWESI TENGGARA	1.0	1.0	1.0	Stagnan
SULAWESI UTARA	1.0	1.0	1.0	Stagnan
SUMATERA BARAT	1.0	1.0	1.0	Stagnan
SUMATERA SELATAN	1.0	1.0	1.0	Stagnan
SUMATERA UTARA	1.0	1.0	1.0	Stagnan

Langkah pertama dalam analisis ini adalah mengekstrak kembali data hasil klasterisasi yang telah disimpan ke dalam file hasilClustering.csv. Karena format kolom Provinsi menggabungkan nama provinsi dan tahun, maka dipisahkan kembali menjadi dua kolom berbeda, yaitu NamaProvinsi

dan Tahun. Ini dilakukan untuk menyusun data dalam bentuk pivot table, di mana masing-masing provinsi menjadi indeks, sedangkan kolomnya berisi label klaster dari tahun 2022 hingga 2024.

Pivot table inilah yang menjadi dasar untuk menganalisis tren perubahan klaster tiap provinsi. Untuk melihat kecenderungan tren, digunakan fungsi tentukanTren() yang membandingkan label klaster provinsi pada tahun awal (2022) dan akhir (2024):

- Jika label klaster di 2024 lebih tinggi dari 2022, berarti provinsi tersebut **meningkat**.
- Jika labelnya lebih rendah, berarti mengalami **penurunan**.
- Jika tidak berubah, maka dianggap **stagnan**.

Hasil akhirnya adalah tabel yang menampilkan label klaster untuk setiap provinsi selama tiga tahun, ditambah satu kolom baru yaitu 'Tren' yang menunjukkan arah pergerakan masing-masing provinsi. Tabel ini sangat berguna untuk mengidentifikasi daerah yang menunjukkan perbaikan pembangunan manusia secara konsisten, serta provinsi yang perlu mendapat perhatian lebih karena tidak mengalami perubahan atau bahkan mengalami kemunduran.

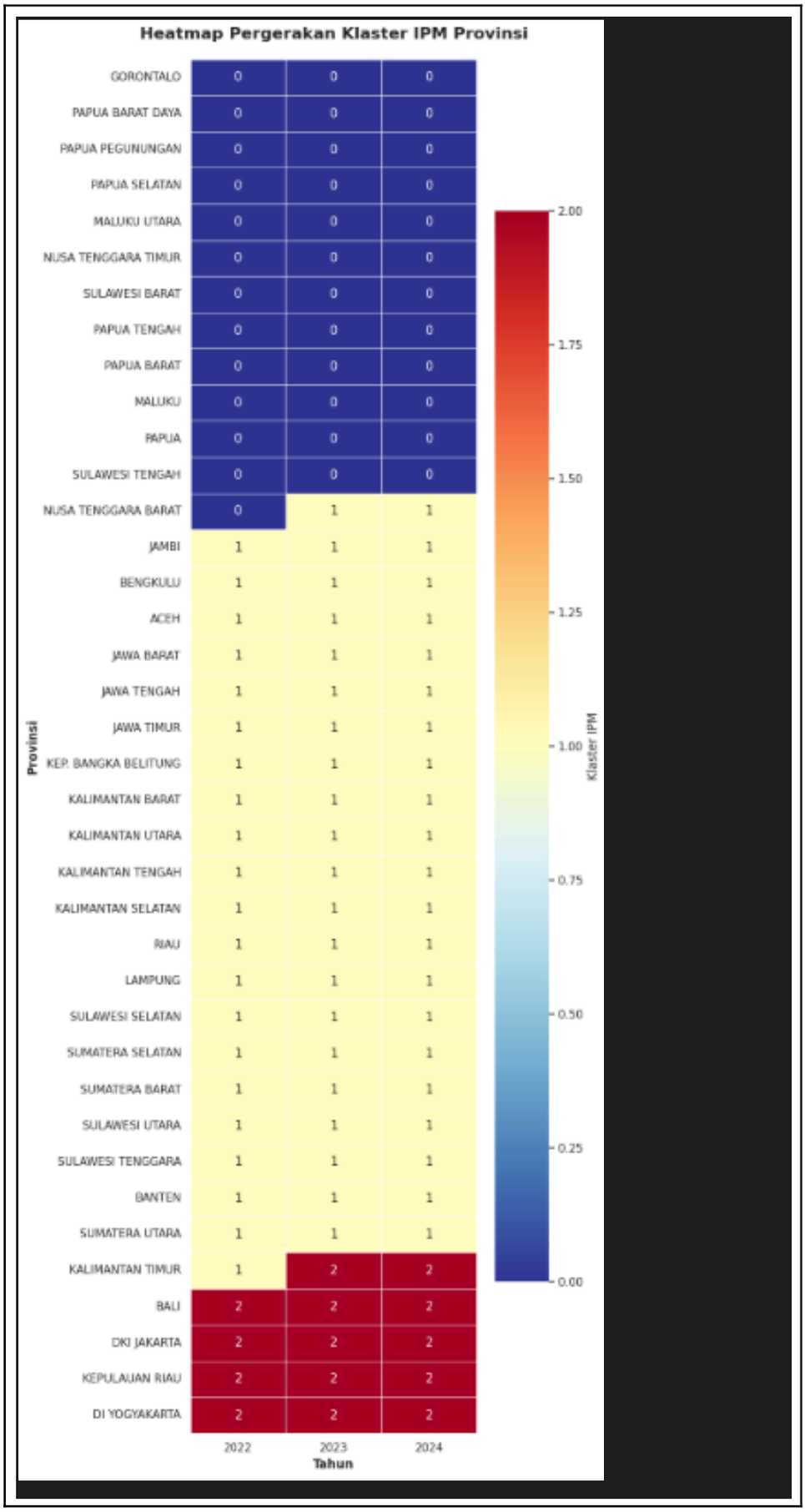
Analisis tren ini juga memberikan gambaran dinamis tentang hasil klasterisasi, bukan hanya snapshot statis per tahun. Dengan melihat arah pergerakan antar waktu, kita bisa lebih bijak dalam menafsirkan keberhasilan maupun tantangan pembangunan manusia yang dihadapi setiap daerah di Indonesia.

4.2.6 Visualization

Setelah melakukan analisis tren pergerakan klaster secara numerik, tahap ini bertujuan untuk menyajikan hasil tersebut dalam bentuk visual agar lebih mudah dipahami dan ditafsirkan. Salah satu bentuk visualisasi yang digunakan adalah **heatmap** (peta panas), yang menampilkan distribusi dan pergerakan klaster provinsi dari tahun 2022 hingga 2024.

Berikut adalah kode yang digunakan untuk menghasilkan visualisasi heatmap pergerakan klaster:

```
# --- VISUALISASI HEATMAP PERGERAKAN KLASTER ---

# Ekstrak tahun dan nama provinsi
dataHasil['Tahun'] =
pd.to_numeric(dataHasil['Provinsi'].str[-4:])
dataHasil['NamaProvinsi'] = dataHasil['Provinsi'].str[:-5]

# Buat pivot table untuk pergerakan klaster
tabelPergerakan = dataHasil.pivot_table(
   index='NamaProvinsi',
   columns='Tahun',
   values='Klaster'
).rename_axis(index='Provinsi', columns=None)

# Urutkan berdasarkan rata-rata klaster atau klaster di tahun
terakhir
tabelPergerakan['RataRata'] = tabelPergerakan.mean(axis=1)
tabelPergerakan =
tabelPergerakan.sort_values('RataRata').drop('RataRata',
axis=1)

# --- HEATMAP PERGERAKAN KLASTER ---
plt.figure(figsize=(8, 20))

# Buat heatmap
sns.heatmap(tabelPergerakan[[2022, 2023, 2024]],
           annot=True, fmt='.0f', cmap='RdYlBu_r',
           cbar_kws={'label': 'Klaster IPM'},
           linewidths=0.5)

plt.title('Heatmap Pergerakan Klaster IPM Provinsi',
         fontsize=16, fontweight='bold', pad=20)
plt.xlabel('Tahun', fontsize=12, fontweight='bold')
plt.ylabel('Provinsi', fontsize=12, fontweight='bold')
plt.tight_layout()
plt.show()
```



Langkah awal dalam visualisasi ini adalah mempersiapkan data dalam format yang sesuai untuk heatmap. Seperti pada bagian sebelumnya, kolom

Provinsi dipecah menjadi NamaProvinsi dan Tahun. Data tersebut kemudian diubah menjadi pivot table yang berisi label klaster tiap provinsi untuk masing-masing tahun.

Agar tampilan heatmap lebih informatif, provinsi diurutkan berdasarkan rata-rata klaster selama tiga tahun. Ini membantu mengelompokkan provinsi yang cenderung berada di klaster rendah ke atas, sementara yang lebih tinggi berada di bawah. Rata-rata tersebut hanya digunakan untuk keperluan penyusunan urutan baris dan tidak ditampilkan dalam heatmap.

Pada bagian visualisasinya, digunakan fungsi sns.heatmap dari library Seaborn. Setiap sel pada heatmap menunjukkan label klaster IPM suatu provinsi pada tahun tertentu, dengan gradasi warna yang merepresentasikan perbedaan antar klaster secara visual. Warna pada heatmap menggunakan palet RdYlBu_r, yang memberikan kontras warna antara nilai klaster yang lebih tinggi dan lebih rendah.

Heatmap ini sangat membantu untuk melihat pola dan tren secara sekilas:

- Provinsi dengan warna konsisten sepanjang tahun menunjukkan stabilitas klaster.
- Pergeseran warna pada baris tertentu mengindikasikan **perubahan posisi klaster**, apakah naik atau turun.
- Kita juga bisa melihat apakah sebagian besar provinsi cenderung mengalami peningkatan, stagnasi, atau penurunan dari sisi pembangunan manusia.

Secara keseluruhan, visualisasi ini memperkuat hasil numerik dari analisis tren sebelumnya, dan membantu menyampaikan informasi secara intuitif kepada pembaca, termasuk pihak non-teknis seperti pengambil kebijakan, tanpa harus membaca tabel atau angka-angka secara rinci.

BAB 5
ANALISIS DAN PEMBAHASAN

5.1 Hasil Klasterisasi K-Means

Analisis klasterisasi dalam penelitian ini dilakukan terhadap data gabungan dari seluruh provinsi di Indonesia selama tiga tahun berturut-turut, yaitu 2022, 2023, dan 2024. Struktur data yang digunakan berbentuk kombinasi antara nama provinsi dan tahun, seperti "ACEH 2022" atau "SUMATERA BARAT 2024", yang masing-masing menjadi satu entitas data (satu baris). Artinya, baik dimensi spasial (provinsi) maupun temporal (tahun) dianalisis secara bersamaan dalam satu ruang fitur.

Data yang dianalisis mencakup empat variabel numerik utama: **Angka Harapan Hidup (AHH)**, **Rata-Rata Lama Sekolah (RLS)**, **Pengeluaran per Kapita (PPP)**, dan **Indeks Pembangunan Manusia (IPM)**. Keempatnya merupakan indikator kunci yang mencerminkan kondisi kesehatan, pendidikan, dan kesejahteraan ekonomi masyarakat di setiap wilayah pada tiap tahun. Sebelum digunakan untuk pemodelan, seluruh variabel dinormalisasi menggunakan metode **Min-Max Scaling**, untuk memastikan semua fitur berada dalam skala yang sebanding.

Setelah proses normalisasi selesai, seluruh entitas data (yaitu provinsi per tahun) diklasterkan secara bersamaan menggunakan algoritma **K-Means**, dengan inisialisasi **K-Means++** untuk menghasilkan centroid awal yang lebih stabil. Pemodelan ini tidak dibagi atau dijalankan per tahun, melainkan dilakukan **sekali saja untuk seluruh data lintas tahun**, agar struktur klaster yang terbentuk merepresentasikan keseluruhan pola pembangunan manusia secara menyeluruh dalam rentang waktu 2022–2024.

Untuk menentukan jumlah klaster yang paling optimal, digunakan tiga pendekatan evaluatif:

- **Elbow Method**, untuk melihat titik tekuk dari grafik inertia.
- **Silhouette Score**, untuk menilai kepadatan dan pemisahan antar klaster.
- **Davies-Bouldin Index (DBI)**, untuk mengukur kemiripan relatif antar klaster.

Dari ketiga metrik tersebut, **nilai Davies-Bouldin Index mencapai titik terendah pada K = 3**, yang berarti pembagian menjadi tiga kelompok memberikan pemisahan dan kekompakan klaster terbaik. Selain alasan kuantitatif tersebut, secara substansial **pembagian menjadi tiga klaster juga dianggap lebih representatif** dalam menggambarkan spektrum variasi pembangunan manusia di Indonesia tidak hanya terbagi antara daerah dengan IPM tinggi dan rendah, tetapi juga mencakup

wilayah-wilayah yang berada di tengah, yang secara bertahap mengalami kemajuan tetapi belum stabil di posisi terdepan.

Setelah K = 3 ditetapkan, model K-Means dilatih terhadap data tersebut dan menghasilkan label klaster untuk masing-masing provinsi dalam masing-masing tahun. Hasilnya menunjukkan bagaimana provinsi berpindah klaster dari waktu ke waktu, atau tetap berada dalam kelompok yang sama. Klaster-klaster tersebut diberi label 0, 1, dan 2, lalu disusun ulang berdasarkan nilai rata-rata IPM untuk keperluan interpretasi, sehingga dapat diartikan sebagai **klaster 0** (IPM rendah), **klaster 1** (IPM sedang), dan **klaster 2** (IPM tinggi).

Hasil inilah yang kemudian divisualisasikan dalam bentuk heatmap, serta dianalisis lebih lanjut untuk melihat pola, tren, dan dinamika pembangunan manusia di Indonesia selama periode tiga tahun tersebut.

5.2 Komposisi Klaster

Setelah proses klasterisasi dilakukan terhadap data gabungan seluruh provinsi Indonesia dalam rentang tahun 2022 hingga 2024, terbentuklah **tiga kelompok klaster utama** yang menggambarkan variasi capaian pembangunan manusia. Klaster-klaster ini dihasilkan dari model K-Means dengan K = 3, yang dipilih berdasarkan evaluasi Davies-Bouldin Index serta pertimbangan substansial bahwa tiga kelompok mampu merepresentasikan spektrum capaian pembangunan secara lebih realistis: dari yang tertinggal, sedang berkembang, hingga yang mapan.

Masing-masing klaster memiliki ciri khas tersendiri yang dapat dikenali dari **rata-rata nilai indikator-indikator IPM** yang menyusunnya. Karena proses klasterisasi dilakukan terhadap **seluruh entitas data yang mewakili kombinasi provinsi dan tahun**, maka analisis komposisi klaster ini tidak hanya mencerminkan keadaan suatu provinsi pada satu waktu, tetapi juga menangkap tren dan kecenderungan yang terjadi selama tiga tahun terakhir. Dengan kata lain, provinsi yang muncul berkali-kali di klaster tertentu menunjukkan konsistensi kondisi pembangunannya, sedangkan provinsi yang berpindah klaster mencerminkan dinamika capaian indikatornya dari tahun ke tahun.

Berikut adalah karakteristik utama dari masing-masing klaster berdasarkan hasil klasterisasi:

5.2.1 Klaster 0 – Pembangunan Rendah

Klaster ini merupakan kelompok dengan **nilai IPM dan indikator penyusunnya paling rendah**. Karakteristik khas dari klaster ini adalah:

- **Angka Harapan Hidup (AHH)** berada di bawah rata-rata nasional.
- **Rata-Rata Lama Sekolah (RLS)** relatif pendek, menandakan rendahnya akses pendidikan formal.

- **Pengeluaran per Kapita (PPP)** rendah, mencerminkan kondisi ekonomi masyarakat yang masih terbatas.
- Nilai **IPM secara keseluruhan rendah**, baik dalam skala absolut maupun relatif terhadap provinsi lain.

Provinsi-provinsi yang tergolong dalam klaster ini secara konsisten sepanjang tiga tahun antara lain: **Papua**, **Papua Selatan**, **Papua Pegunungan**, **Papua Barat Daya**, dan **Nusa Tenggara Timur**. Persebaran geografis klaster ini umumnya terletak di wilayah timur Indonesia, yang memang masih menghadapi berbagai kendala struktural seperti keterbatasan infrastruktur, distribusi layanan publik yang tidak merata, dan kondisi geografis yang menyulitkan akses.

5.2.2 Klaster 1 – Pembangunan Menengah

Klaster ini berisi provinsi-provinsi dengan **capaian pembangunan yang sedang**, yaitu tidak tergolong rendah tetapi belum mencapai tingkat yang tinggi dan stabil. Beberapa indikator mungkin menunjukkan nilai yang relatif tinggi, sementara yang lain masih tertinggal. Ciri-ciri utamanya adalah:

- **AHH dan RLS** berada di kisaran menengah, namun dengan kecenderungan meningkat dari tahun ke tahun.
- **PPP** bervariasi, tetapi belum cukup kuat untuk mendorong nilai IPM secara signifikan.
- Nilai **IPM menunjukkan tren kenaikan**, meskipun belum cukup untuk masuk ke kelompok tertinggi.

Provinsi-provinsi dalam klaster ini mencerminkan **wilayah-wilayah yang sedang mengalami transisi pembangunan**, seperti **Sumatera Utara**, **Sumatera Barat**, **Jawa Barat**, **Kalimantan Selatan**, dan **Sulawesi Selatan**. Klaster ini juga mencakup provinsi-provinsi yang menunjukkan **kemajuan yang menjanjikan**, meskipun belum seragam di semua aspek.

Provinsi-provinsi di klaster ini memiliki peluang besar untuk berpindah ke klaster IPM tinggi dalam waktu dekat, terutama jika pembangunan yang ada terus diarahkan secara konsisten ke peningkatan pendidikan dan daya beli masyarakat. Sebagian provinsi bahkan sudah menunjukkan tren mendekati batas atas dari klaster ini.

5.2.3 Klaster 2 – Pembangunan Tinggi

Kelompok ini terdiri dari provinsi-provinsi yang secara konsisten mencatatkan **nilai indikator pembangunan manusia yang tinggi**. Karakteristik utama dari klaster ini meliputi:

- **AHH dan RLS** berada di atas rata-rata nasional.
- **PPP** menunjukkan tingkat daya beli yang kuat.
- Nilai **IPM tergolong tinggi dan stabil**, mencerminkan sistem pembangunan yang sudah lebih mapan dan terstruktur.

Beberapa provinsi yang secara konsisten berada di klaster ini adalah **DKI Jakarta**, **DI Yogyakarta**, **Bali**, dan **Kepulauan Riau**. Klaster ini didominasi oleh wilayah-wilayah yang sudah memiliki infrastruktur yang baik, urbanisasi yang tinggi, serta kebijakan daerah yang relatif stabil dalam mendukung pelayanan publik. Akses terhadap layanan kesehatan dan pendidikan di wilayah-wilayah ini juga cenderung lebih merata dan berkelanjutan.

5.2.4 Distribusi Geografis dan Catatan Umum

Jika ditinjau dari sisi geografis, terlihat kecenderungan bahwa:

- **Wilayah barat Indonesia dan pusat-pusat ekonomi nasional** cenderung masuk ke klaster 2.
- **Provinsi-provinsi di kawasan tengah dan beberapa wilayah timur** banyak menempati klaster 1.
- **Wilayah timur dan daerah tertinggal** masih mendominasi klaster 0.

Namun penting untuk dicatat bahwa pembagian ini tidak bersifat absolut. Beberapa provinsi di luar pola umum justru menunjukkan capaian yang progresif atau bahkan stagnan. Oleh karena itu, hasil klasterisasi tidak hanya penting sebagai alat pemetaan, tetapi juga sebagai cermin untuk melihat ketimpangan, potensi pertumbuhan, dan kebutuhan intervensi yang lebih tepat sasaran.

Secara keseluruhan, komposisi klaster ini membentuk dasar pemahaman yang lebih menyeluruh terhadap kondisi pembangunan manusia di Indonesia, bukan hanya dalam satu waktu, tetapi juga sebagai rangkaian tren selama tiga tahun berturut-turut. Klasterisasi ini juga membuka ruang interpretasi kebijakan yang berbasis pada kebutuhan riil tiap kelompok provinsi, seperti yang akan dibahas lebih lanjut dalam bagian berikutnya.

5.3 Perubahan IPM dan Dampaknya terhadap Klaster

Setelah klasterisasi dilakukan terhadap seluruh entitas data yang mencakup kombinasi provinsi dan tahun (2022–2024), langkah selanjutnya adalah mengevaluasi bagaimana **posisi provinsi dalam klaster berubah dari waktu ke waktu**. Dengan menggunakan hasil label klaster yang dihasilkan dari model K-Means terhadap data gabungan, dapat dilakukan pelacakan posisi setiap provinsi selama tiga tahun pengamatan, sehingga menghasilkan peta pergerakan klaster yang merefleksikan dinamika pembangunan manusia secara lebih kontekstual.

Visualisasi hasil ini dituangkan dalam bentuk **heatmap pergerakan klaster**, di mana **setiap baris mewakili satu provinsi**, dan setiap kolom mewakili tahun 2022, 2023, dan 2024. Warna yang ditampilkan mengindikasikan label klaster pada tahun tertentu, yaitu:

- **Klaster 0** – IPM rendah
- **Klaster 1** – IPM menengah
- **Klaster 2** – IPM tinggi

Melalui visualisasi ini, terlihat bahwa **mayoritas provinsi menunjukkan konsistensi posisi dalam klaster yang sama selama tiga tahun berturut-turut**. Hal ini mencerminkan bahwa perubahan nilai IPM di sebagian besar wilayah terjadi secara bertahap dan tidak cukup drastis untuk mendorong pergeseran ke klaster lain. Klasterisasi yang berbasis pada gabungan empat variabel (AHH, RLS, PPP, dan IPM) membuat perpindahan antar klaster tidak hanya bergantung pada kenaikan satu indikator saja, tetapi pada keseimbangan seluruh aspek secara bersamaan.

Namun demikian, ada sejumlah provinsi yang menunjukkan **mobilitas vertikal**, yaitu berpindah dari klaster yang lebih rendah ke klaster yang lebih tinggi:

- **Nusa Tenggara Barat (NTB)** mengalami pergeseran dari klaster 0 pada tahun 2022 ke klaster 1 pada tahun 2023 dan bertahan hingga 2024. Ini menunjukkan adanya kemajuan pembangunan yang cukup stabil, terutama jika peningkatan IPM diiringi peningkatan indikator lain seperti RLS dan PPP.

- **Kalimantan Timur**, meskipun tergolong provinsi maju, tercatat berada di klaster 1 pada tahun 2022, lalu berpindah ke klaster 2 pada tahun 2023 dan bertahan di sana hingga 2024. Perpindahan ini konsisten dengan pola nilai IPM dan pengeluaran per kapita yang terus meningkat, serta posisinya sebagai provinsi yang berkembang pesat secara ekonomi dan infrastruktur dalam beberapa tahun terakhir.

Sebaliknya, tidak ditemukan provinsi yang **turun ke klaster yang lebih rendah** selama periode pengamatan. Hal ini mengindikasikan bahwa meskipun pergerakan ke atas relatif terbatas, **tren pembangunan manusia secara nasional menunjukkan arah yang positif dan cenderung progresif**. Bahkan

provinsi-provinsi yang tetap berada di klaster yang sama pun dalam banyak kasus mengalami kenaikan nilai IPM, hanya saja peningkatannya belum cukup untuk menggeser posisi relatif mereka terhadap centroid klaster yang lebih tinggi.

Penting dipahami bahwa dalam konteks klasterisasi, **perpindahan klaster bukan hanya soal kenaikan nilai IPM absolut**, tetapi soal **perbandingan posisi relatif suatu provinsi terhadap seluruh data populasi yang dianalisis secara bersamaan**. Sebagai contoh, jika semua provinsi mengalami peningkatan IPM dalam waktu yang sama dan dalam proporsi yang serupa, maka posisi relatif masing-masing provinsi bisa saja tetap, karena struktur klasternya juga ikut bergeser secara kolektif. Inilah mengapa beberapa provinsi yang mengalami peningkatan nyata pada nilai IPM tetap berada dalam klaster yang sama selama tiga tahun.

Selain itu, ada pula provinsi-provinsi yang meskipun konsisten berada dalam klaster 0 atau 1, menunjukkan **indikasi mendekati batas atas klaster**, berdasarkan nilai skor fitur terstandar (scaled features). Ini dapat menjadi perhatian dalam analisis lebih lanjut untuk mengidentifikasi **provinsi yang potensial naik** dalam waktu dekat, asalkan tren positif yang dimiliki tetap berlanjut.

Secara umum, heatmap pergerakan klaster memberikan gambaran yang sangat berguna dalam melihat **apakah kebijakan pembangunan yang diterapkan di suatu wilayah telah menghasilkan dampak yang cukup signifikan**, baik dalam hal kesehatan, pendidikan, maupun kesejahteraan ekonomi. Provinsi-provinsi yang berhasil berpindah klaster ke arah yang lebih tinggi dapat dianggap sebagai contoh praktik baik, sedangkan provinsi yang stagnan dapat menjadi fokus perhatian lebih lanjut, khususnya untuk mengidentifikasi hambatan apa yang menyebabkan mereka tertahan.

Dari perspektif pembangunan berkelanjutan, pemahaman terhadap dinamika posisi klaster seperti ini jauh lebih bermakna dibanding sekadar melihat peringkat IPM dari tahun ke tahun, karena ia mengintegrasikan struktur multidimensi dari data serta memposisikan setiap wilayah dalam konteks kemiripan dengan wilayah lainnya.

5.4 Visualisasi Hasil Klasterisasi

Agar hasil klasterisasi yang telah diperoleh dapat dipahami secara lebih intuitif dan komunikatif, dibuat sebuah **visualisasi dalam bentuk heatmap** yang menggambarkan **pergerakan klaster IPM provinsi-provinsi di Indonesia selama periode 2022 hingga 2024**. Visualisasi ini tidak hanya membantu mengenali posisi tiap provinsi pada tahun tertentu, tetapi juga memberikan gambaran menyeluruh tentang **stabilitas dan dinamika klaster dari waktu ke waktu**.

Dalam heatmap tersebut, **setiap baris mewakili satu provinsi**, dan **setiap kolom menunjukkan tahun**. Nilai dalam sel menunjukkan **label klaster** dari provinsi tersebut pada tahun tersebut, yang secara visual diterjemahkan ke dalam

gradasi warna. Palet warna yang digunakan memudahkan pembaca membedakan klaster dengan cepat:

- **Warna biru tua** merepresentasikan **klaster 0**, yaitu provinsi dengan IPM rendah.
- **Warna kuning pucat** mewakili **klaster 1**, atau IPM sedang.
- **Warna merah tua** menunjukkan **klaster 2**, yaitu IPM tinggi.

Pemilihan warna-warna kontras ini tidak hanya estetis, tetapi juga fungsional. Perubahan warna dari tahun ke tahun pada baris yang sama langsung menunjukkan adanya perpindahan klaster, sementara warna yang tetap sepanjang baris menandakan konsistensi posisi provinsi dalam satu kelompok.

Visualisasi ini memiliki sejumlah keunggulan dibandingkan bentuk representasi data lainnya:

- **Pertama**, ia menyederhanakan informasi dari ratusan data numerik menjadi pola visual yang dapat dibaca secara sekilas.
- **Kedua**, ia menggabungkan dua dimensi penting, **ruang (provinsi)** dan **waktu (tahun)** ke dalam satu tampilan.
- **Ketiga**, ia memungkinkan pembaca untuk mengidentifikasi pola-pola yang berulang, provinsi-propinsi yang stabil, serta daerah-daerah yang mulai menunjukkan mobilitas sosial dan ekonomi.

Sebagai contoh, **provinsi seperti DKI Jakarta, DI Yogyakarta, dan Bali** secara konsisten berada pada klaster 2 sepanjang tiga tahun pengamatan. Ini terlihat dari baris-baris yang berwarna merah tua penuh. Sebaliknya, **provinsi seperti Papua Selatan, Papua Pegunungan, dan Nusa Tenggara Timur** juga terlihat tidak berpindah dari klaster 0 (warna biru tua), yang menandakan bahwa tantangan pembangunan di wilayah-wilayah tersebut masih berlanjut dan belum menunjukkan perubahan signifikan dalam tiga tahun terakhir.

Di sisi lain, heatmap juga menyoroti **provinsi yang berpindah klaster** secara jelas. Misalnya, **Nusa Tenggara Barat** yang berpindah dari klaster 0 ke klaster 1 pada tahun 2023 dan mempertahankan posisi tersebut di tahun 2024, ditampilkan dengan peralihan warna dari biru ke kuning pada baris tersebut. Ini menjadi penanda visual yang sangat kuat bahwa daerah tersebut mengalami perbaikan yang cukup bermakna dalam konteks pembangunan manusia.

Dengan demikian, heatmap tidak hanya berfungsi sebagai alat visualisasi hasil klasterisasi, tetapi juga menjadi **alat interpretasi naratif**, yang memungkinkan kita memahami arah, kecepatan, dan konsistensi pembangunan manusia di berbagai wilayah Indonesia. Melalui tampilan ini, klasterisasi tidak lagi menjadi sekadar hasil numerik dari algoritma, tetapi berubah menjadi **peta dinamis** yang membantu melihat realitas sosial secara lebih utuh.

5.5 Evaluasi dan Keterbatasan Model

Setelah proses klasterisasi selesai dilakukan, langkah selanjutnya adalah mengevaluasi sejauh mana hasil yang diperoleh benar-benar mencerminkan struktur data yang bermakna. Dalam penelitian ini, evaluasi model K-Means dilakukan menggunakan tiga metrik utama, yaitu **Inertia**, **Silhouette Score**, dan **Davies-Bouldin Index (DBI)**. Ketiganya digunakan untuk menentukan jumlah klaster optimal serta menilai kualitas pemisahan dan kekompakan antar klaster.

5.5.1 Evaluasi Jumlah Klaster Optimal

Penentuan jumlah klaster (*K*) yang paling sesuai merupakan tahap krusial, karena akan menentukan bagaimana struktur klaster terbentuk dan seberapa bermanfaat hasilnya untuk interpretasi lebih lanjut. Evaluasi terhadap nilai *K* dilakukan dengan pendekatan berikut:

- **Elbow Method (Inertia)**
  Grafik inertia menunjukkan penurunan tajam antara K = 2 dan K = 3, lalu mulai melandai setelahnya. Ini menandakan adanya titik "tekukan" atau *elbow* pada sekitar K = 3, yang berarti menambah jumlah klaster setelah titik tersebut tidak lagi secara signifikan menurunkan variasi dalam klaster (WCSS). Ini menjadi indikasi awal bahwa **K = 3 merupakan titik yang cukup efisien secara struktural**.

- **Silhouette Score**
  Nilai silhouette tertinggi ditemukan pada K = 2, yang berarti bahwa pada dua klaster, data menunjukkan kekompakan dan pemisahan paling optimal secara matematis. Namun, penurunan skor pada K = 3 dan seterusnya tetap berada pada tingkat yang masih layak, dan secara substansial dapat diterima mengingat **kondisi pembangunan manusia tidak memiliki batas-batas klaster yang sepenuhnya tegas**.

- **Davies-Bouldin Index (DBI)**
  DBI memberikan nilai terendah pada **K = 3 dan K = 11**, yang menandakan kualitas pemisahan antar klaster dan kekompakan internal yang paling baik terjadi pada dua titik ini. Namun, dalam konteks pembangunan manusia, **K = 11 dianggap terlalu banyak dan terlalu spesifik**, sehingga kurang relevan untuk analisis makro yang bertujuan mengelompokkan provinsi berdasarkan tingkat pembangunan secara luas. Oleh karena itu, **K = 3 dipilih sebagai solusi yang paling seimbang**, baik dari segi evaluasi metrik maupun pertimbangan substantif.

Secara keseluruhan, pemilihan K = 3 mencerminkan kompromi antara efisiensi struktural (inertia), kualitas pemisahan (DBI), dan makna konseptual,

yakni mampu mewakili tiga tingkat pembangunan: rendah, sedang, dan tinggi. Ini dinilai **lebih representatif** dalam menjelaskan realitas sosial-ekonomi pembangunan manusia di Indonesia dibanding pembagian dua kutub yang terlalu menyederhanakan.

5.5.2 Keterbatasan Model K-Means dalam Konteks Penelitian

Walaupun K-Means terbukti efektif dan cukup stabil dalam pengelompokan data ini, algoritma ini tetap memiliki sejumlah keterbatasan yang penting untuk dipahami, khususnya dalam konteks data sosial seperti pembangunan manusia.

1. **Asumsi bentuk klaster yang bulat dan seimbang**
   K-Means secara implisit mengasumsikan bahwa klaster memiliki bentuk sferis dan ukuran yang relatif seimbang. Dalam kenyataannya, data sosial sering kali memiliki distribusi yang tidak simetris atau memiliki kepadatan berbeda. Hal ini dapat menyebabkan hasil klasterisasi menjadi kurang akurat pada batas antar klaster, terutama untuk provinsi yang berada di posisi ambang.

2. **Sensitivitas terhadap inisialisasi centroid**
   Algoritma K-Means bergantung pada inisialisasi awal centroid. Jika pemilihan centroid dilakukan secara acak, algoritma bisa berhenti pada solusi lokal yang kurang optimal. Dalam penelitian ini, digunakan metode **K-Means++** untuk mengurangi risiko tersebut. Namun tetap saja, karena sifat algoritma yang non-deterministik, hasil klasterisasi masih bisa sedikit berbeda jika dijalankan ulang dengan kondisi awal yang berbeda.

3. **Kebutuhan untuk menentukan jumlah klaster di awal**
   Tidak seperti metode seperti DBSCAN atau hierarchical clustering, K-Means mengharuskan pengguna menentukan jumlah klaster (K) sebelum proses dimulai. Oleh karena itu, proses evaluasi tambahan sangat penting untuk memastikan jumlah klaster yang digunakan memang sesuai dengan struktur data.

4. **Ketergantungan pada skala data**
   K-Means sangat sensitif terhadap skala. Jika data tidak dinormalisasi terlebih dahulu, variabel dengan skala besar akan mendominasi hasil. Dalam penelitian ini, normalisasi dengan Min-Max Scaling telah dilakukan untuk menghindari masalah ini.

5.5.3 Refleksi Evaluatif

Berdasarkan hasil evaluasi dan proses yang dilakukan, dapat disimpulkan bahwa penggunaan algoritma K-Means dengan tiga klaster memberikan **hasil yang cukup solid** untuk menggambarkan struktur

perbedaan pembangunan manusia antar provinsi dalam periode 2022 hingga 2024. Klaster yang terbentuk tidak hanya didukung oleh metrik evaluasi yang relevan, tetapi juga **secara substansial masuk akal** dan mudah ditafsirkan sebagai kelompok wilayah tertinggal, berkembang, dan maju.

Namun demikian, untuk konteks analisis yang lebih kompleks atau untuk data dengan struktur yang sangat tidak teratur, perlu dipertimbangkan algoritma alternatif seperti **Gaussian Mixture Models**, **DBSCAN**, atau **Hierarchical Clustering** yang lebih fleksibel dalam menangani bentuk dan kepadatan klaster yang beragam.

5.6 Pola Pembangunan Wilayah Berdasarkan Klaster

Hasil klasterisasi terhadap gabungan data IPM provinsi Indonesia dalam rentang tahun 2022–2024 memperlihatkan pola spasial yang kuat dan cukup konsisten. Namun lebih dari sekadar membagi provinsi ke dalam tiga kategori IPM—rendah, sedang, dan tinggi—hasil ini juga membuka wawasan baru tentang **bagaimana dimensi pembangunan manusia berkelindan dengan lokasi geografis, struktur sosial, dan arah kebijakan di tiap wilayah**. Dalam konteks inilah, analisis pola klaster perlu dilihat sebagai refleksi terhadap ketimpangan dan dinamika pembangunan yang lebih dalam.

5.6.1 Klaster 2 – Titik Konsentrasi Pembangunan yang Mapan

Provinsi-provinsi yang tergolong dalam **klaster IPM tinggi (klaster 2)** adalah representasi dari daerah-daerah yang tidak hanya unggul dalam capaian IPM, tetapi juga menunjukkan **kestabilan struktural dan institusional** yang menopang pembangunan manusianya secara berkelanjutan.

Provinsi seperti **DKI Jakarta, DI Yogyakarta, Bali, Kepulauan Riau, dan Kalimantan Timur (sejak 2023)** memiliki kesamaan dalam beberapa hal:

- Tingkat urbanisasi yang tinggi.
- Akses yang luas terhadap layanan dasar.
- Infrastruktur fisik dan digital yang relatif merata.
- Dukungan fiskal daerah yang kuat dan kemampuan tata kelola pemerintahan yang relatif lebih baik.

Kombinasi tersebut menghasilkan lingkungan yang kondusif bagi masyarakat untuk **menjangkau pendidikan yang lebih tinggi, hidup lebih sehat, dan memiliki kemampuan daya beli yang lebih besar**. Kalimantan Timur menjadi contoh yang menarik, karena posisinya sebagai lokasi pembangunan Ibu Kota Negara (IKN) turut menjadi faktor penarik investasi

dan peningkatan infrastruktur publik yang berdampak langsung pada pembangunan manusianya.

Namun, penting dicatat bahwa dominasi wilayah barat Indonesia dalam klaster ini memperlihatkan ketimpangan historis yang masih belum sepenuhnya teratasi. Bahwa hanya sebagian kecil wilayah luar Jawa dapat menembus klaster tertinggi memperkuat dugaan bahwa akses dan distribusi pembangunan di Indonesia masih sangat berpusat.

5.6.2 Klaster 1 – Wilayah Transisi: Stabil tapi Belum Mapan

Klaster IPM sedang (klaster 1) dihuni oleh provinsi-provinsi yang berada dalam posisi **transisi pembangunan**. Mereka memiliki beberapa indikator yang sudah cukup baik, misalnya AHH atau RLS, tetapi belum sepenuhnya seimbang dengan daya beli atau pemerataan akses.

Kelompok ini mencakup wilayah padat penduduk seperti **Jawa Barat, Jawa Tengah, Jawa Timur**, serta provinsi besar lain seperti **Sumatera Barat, Riau, Kalimantan Selatan, Sulawesi Selatan, Lampung**, dan **Sumatera Selatan**. Selain itu, **provinsi-provinsi yang mulai membaik**, seperti **Nusa Tenggara Barat**, juga masuk dalam klaster ini setelah berpindah dari klaster rendah.

Ciri utama dari kelompok ini adalah **variasi internal yang tinggi**. Beberapa provinsi memiliki capaian pendidikan yang cukup baik tetapi masih tertinggal secara ekonomi, atau sebaliknya. Ketimpangan wilayah dalam satu provinsi juga cukup besar misalnya Jawa Barat yang memiliki perbedaan tajam antara kawasan metropolitan seperti Bandung dan daerah-daerah pinggiran yang masih minim layanan dasar.

Wilayah dalam klaster ini sangat krusial untuk diperhatikan karena mereka menyimpan **potensi pertumbuhan yang besar**, tetapi juga menghadapi risiko stagnasi jika tidak didukung dengan kebijakan yang lebih fokus dan tepat sasaran. Investasi pembangunan di klaster ini perlu diarahkan untuk:

- Menutup celah antar subwilayah.
- Meningkatkan kualitas bukan hanya kuantitas layanan publik.
- Memperkuat kapasitas fiskal daerah yang lemah meskipun populasinya besar.

5.6.3 Klaster 0 – Ketertinggalan yang Berulang: Tantangan Struktural dan Historis

Klaster 0, atau kelompok IPM rendah, merupakan gambaran nyata tentang **ketimpangan pembangunan yang bersifat struktural dan telah berlangsung lama**. Di dalamnya terdapat provinsi-provinsi yang secara

konsisten menghadapi tantangan serius dalam semua aspek indikator IPM: pendidikan yang belum menjangkau seluruh masyarakat, akses layanan kesehatan yang terbatas, serta tingkat pengeluaran per kapita yang rendah.

Kelompok ini mencakup:

- **Papua, Papua Tengah, Papua Selatan, Papua Pegunungan, Papua Barat, Papua Barat Daya**
- **Nusa Tenggara Timur**
- **Maluku dan Maluku Utara**
- **Sulawesi Barat, Sulawesi Tengah**
- **Gorontalo**

Sebagian besar provinsi di klaster ini terletak di wilayah timur Indonesia. Tetapi lebih dari sekadar letak geografis, **akar persoalannya adalah akumulasi ketimpangan jangka panjang**: kurangnya infrastruktur dasar, ketergantungan ekonomi terhadap komoditas primer, minimnya konektivitas antarwilayah, serta seringnya perubahan administratif yang tidak dibarengi dengan penguatan kapasitas institusional.

Sebagai contoh, banyak wilayah di Papua masih menghadapi kendala logistik yang ekstrem, sehingga pembangunan sekolah, puskesmas, atau bahkan akses internet menjadi jauh lebih kompleks dan mahal dibandingkan provinsi lain. Sementara itu, daerah seperti NTT dan Maluku meskipun mengalami peningkatan IPM dalam angka, masih tertinggal karena **kenaikan tersebut tidak cukup signifikan untuk mengubah posisi relatif mereka dibanding provinsi lain**.

Yang paling memprihatinkan adalah: **hampir semua provinsi di klaster ini tidak berpindah klaster selama tiga tahun**. Artinya, meskipun ada upaya pembangunan, tingkat akselerasinya belum cukup untuk mengejar ketertinggalan yang sudah dalam. Tanpa intervensi afirmatif, daerah-daerah ini berisiko **terjebak dalam lingkaran stagnasi**.

5.6.4 Refleksi Spasial dan Rekomendasi Arah Kebijakan

Pola spasial yang muncul dari hasil klasterisasi ini mencerminkan kenyataan bahwa **Indonesia bukan hanya mengalami ketimpangan antar individu, tetapi juga antar wilayah**. Klasterisasi menunjukkan bahwa letak geografis masih memainkan peran besar dalam menentukan peluang dan hambatan pembangunan manusia.

Namun, penting untuk tidak menyederhanakan hasil ini menjadi narasi "barat lebih maju, timur tertinggal." Faktanya, ada **provinsi di barat yang masih berada di klaster menengah**, dan beberapa wilayah timur seperti Kalimantan Timur sudah menunjukkan mobilitas ke atas. Artinya, **lokasi penting, tapi bukan segalanya**. Kebijakan, tata kelola, kapasitas fiskal, dan keberlanjutan intervensi pembangunan memiliki pengaruh yang sangat besar.

Oleh karena itu, **hasil klasterisasi ini idealnya tidak berhenti sebagai alat pemetaan**, tetapi menjadi **bahan dasar untuk segmentasi kebijakan pembangunan manusia**. Provinsi dalam klaster yang sama cenderung memiliki masalah dan kebutuhan yang serupa, sehingga pendekatan pembangunan berbasis klaster bisa:

- Menghemat biaya dengan program yang lebih terarah.
- Meningkatkan efektivitas kebijakan.
- Mendorong solidaritas antarwilayah melalui transfer pengetahuan dan kebijakan.

BAB 6
KESIMPULAN DAN SARAN

6.1 Kesimpulan

Berdasarkan rangkaian proses mulai dari pengumpulan data, preprocessing, pemodelan klasterisasi dengan K-Means, hingga evaluasi dan interpretasi hasil, terdapat beberapa poin penting yang dapat disimpulkan dari penelitian ini:

1. **Klasterisasi berhasil memetakan 38 provinsi di Indonesia ke dalam tiga kelompok utama berdasarkan capaian Indeks Pembangunan Manusia (IPM).**
   Klaster pertama mencerminkan kelompok provinsi dengan IPM rendah, yang sebagian besar berada di wilayah timur dan menghadapi tantangan struktural dalam pendidikan, kesehatan, dan ekonomi. Klaster kedua berisi provinsi dengan capaian IPM sedang, yang menunjukkan karakteristik transisi—belum sepenuhnya tertinggal, tetapi juga belum benar-benar mapan. Sementara itu, klaster ketiga terdiri atas provinsi dengan IPM tinggi, yang didominasi oleh wilayah barat dan daerah urban dengan infrastruktur dan kapasitas institusional yang relatif lebih kuat.

2. **Penentuan jumlah klaster optimal dilakukan secara objektif menggunakan metrik evaluasi kuantitatif dan pertimbangan substantif.**
   Dari hasil perbandingan Elbow Method, Silhouette Score, dan Davies-Bouldin Index, nilai K = 3 dipilih karena memberikan struktur pembagian yang paling representatif terhadap realitas pembangunan manusia di Indonesia. Selain menunjukkan performa numerik yang baik, konfigurasi tiga klaster ini juga mencerminkan pola ketimpangan yang nyata dan mudah diinterpretasikan secara geografis serta kebijakan.

3. **Model K-Means memberikan hasil yang cukup stabil dan masuk akal untuk kasus data IPM yang bersifat numerik dan berdimensi rendah.**
   Proses klasterisasi yang dilakukan terhadap gabungan data IPM tahun 2022–2024 menunjukkan bahwa hasil pengelompokan cukup konsisten. Model mampu membedakan provinsi berdasarkan karakteristik IPM-nya, bahkan ketika perbedaan antar tahun tidak terlalu drastis. Normalisasi data sebelum pemodelan juga terbukti penting agar setiap indikator mendapat bobot yang seimbang dalam proses pengelompokan.

4. **Peningkatan IPM tidak selalu diikuti oleh perubahan posisi klaster, karena pembangunan bersifat relatif.**
   Meskipun sebagian besar provinsi mengalami kenaikan IPM dari tahun ke tahun, banyak yang tetap berada pada klaster yang sama. Hal ini menunjukkan

bahwa kemajuan satu provinsi harus dilihat dalam konteks keseluruhan dinamika nasional. Perubahan posisi klaster baru terjadi ketika suatu daerah tumbuh secara lebih signifikan dibandingkan rata-rata provinsi lainnya.

5. **Visualisasi dalam bentuk heatmap klaster per tahun menjadi alat bantu yang sangat efektif dalam menyampaikan hasil analisis.**
   Heatmap yang menampilkan dinamika klaster setiap tahun mampu menggambarkan stabilitas dan pergerakan provinsi secara lebih komunikatif dibandingkan visualisasi statis. Pola warna yang mudah dibedakan membantu mengidentifikasi provinsi yang stagnan, yang berpindah klaster, serta kecenderungan spasial dari waktu ke waktu.

6. **Klasterisasi ini tidak hanya memetakan angka, tetapi juga mencerminkan kondisi nyata pembangunan manusia di lapangan.**
   Klaster IPM rendah umumnya beririsan dengan daerah-daerah yang memiliki keterbatasan infrastruktur, akses pendidikan, dan daya beli. Sebaliknya, klaster IPM tinggi didominasi oleh daerah dengan akses layanan yang mapan dan tata kelola yang lebih kuat. Hal ini menunjukkan bahwa hasil klasterisasi dapat berfungsi sebagai cermin sosial dan acuan arah kebijakan yang lebih terfokus.

Secara keseluruhan, penelitian ini membuktikan bahwa algoritma K-Means dapat digunakan secara efektif untuk memahami struktur ketimpangan pembangunan manusia di Indonesia. Hasil yang diperoleh bukan hanya bermanfaat secara teknis, tetapi juga memiliki nilai praktis dalam perumusan kebijakan berbasis data yang lebih adil dan kontekstual.

6.2 Saran

Berdasarkan temuan dan hasil analisis yang telah diperoleh, berikut beberapa saran yang dapat dipertimbangkan untuk pengembangan lebih lanjut:

1. **Perluasan Variabel dan Dimensi Pembangunan**
   Penelitian ini masih menggunakan tiga indikator utama IPM. Untuk memperoleh gambaran yang lebih menyeluruh, sebaiknya pada penelitian selanjutnya ditambahkan variabel lain yang juga mencerminkan kualitas hidup, seperti tingkat pengangguran, partisipasi sekolah menengah, angka putus sekolah, rasio ketimpangan (GINI), atau bahkan indeks digitalisasi dan akses layanan dasar. Pendekatan multidimensi ini dapat membantu menangkap kompleksitas pembangunan manusia secara lebih utuh.

2. **Eksperimen dengan Algoritma Klasterisasi Alternatif**
   K-Means memiliki keunggulan dalam kesederhanaannya, tetapi juga memiliki keterbatasan terhadap bentuk klaster non-sferis dan sensitivitas terhadap inisialisasi awal. Oleh karena itu, penting untuk membandingkan hasilnya dengan algoritma lain seperti DBSCAN (yang lebih fleksibel terhadap outlier) atau hierarchical clustering (yang lebih kuat dalam menampilkan struktur hirarki). Pendekatan ini akan membantu menilai sejauh mana hasil yang diperoleh benar-benar robust terhadap metode yang digunakan.

3. **Pemanfaatan Hasil Klasterisasi untuk Kebijakan Wilayah**
   Pemerintah pusat dan daerah dapat memanfaatkan hasil pengelompokan ini sebagai dasar untuk menyusun kebijakan berbasis klaster. Pendekatan semacam ini memungkinkan strategi pembangunan yang lebih tersegmentasi: bukan lagi satu strategi nasional untuk semua, tetapi perancangan program yang disesuaikan dengan tantangan dan kebutuhan masing-masing kelompok provinsi. Misalnya, klaster rendah bisa difokuskan pada infrastruktur dasar dan layanan publik dasar, sementara klaster menengah pada akselerasi kualitas pendidikan dan ekonomi produktif.

4. **Penyajian Data yang Lebih Komunikatif dan Inklusif**
   Visualisasi hasil analisis yang baik sangat membantu pemahaman lintas disiplin, terutama bagi pengambil kebijakan yang tidak memiliki latar belakang teknis. Penggunaan peta tematik interaktif, dashboard digital, atau bahkan infografis sederhana bisa meningkatkan nilai guna hasil analisis dan menjembatani komunikasi antara data scientist dan pembuat kebijakan.

5. **Pembaruan Dataset Secara Berkala dan Konsisten**
   Ketersediaan dan konsistensi data sangat menentukan keberhasilan analisis. Untuk itu, penting agar BPS dan lembaga terkait terus memperbarui data IPM dengan resolusi waktu dan spasial yang lebih tinggi, agar analisis di masa depan dapat menangkap dinamika pembangunan secara lebih presisi dan responsif terhadap perubahan situasi.

DAFTAR PUSTAKA

[1] J. Wu, *Advances in k-means clustering: A data mining thinking*. Springer Science & Business Media, 2012.

[2] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Springer, 2016.

[3] Mustika *et al.*, *DATA MINING DAN APLIKASINYA*. Penerbit Widina, 2021.

[4] Mira, S.Kom., M.Kom., Azriel Christian Nurcahyo, S.Kom., M.Kom., Candra Gudiato, S.Kom., M.Kom., Noviyanti. P, S.Kom., M.Kom., and Listra Frigia Missianes Horhoruw, S.Kom., M.Kom., *Data Mining Mengeksplorasi Teknik-Teknik Data Mining dan Metode K-Means Teori, Konsep, Algoritma dan Studi Kasus*. Uwais Inspirasi Indonesia, 2025.

[5] M. A. M.Kom. S. Si. ,. and M. N. M.T. S. T. ,., *Data Mining - Algoritma dan Implementasi*. Penerbit Andi, 2020.

[6] A. Nugrahaini and Rahmadi Yotenka, "Penerapan Algoritma K-Means Clustering untuk Mengelompokkan Kecamatan di Kabupaten Grobogan Menurut Tingkat Kesejahteraan Keluarga Tahun 2020," *Emerging Statistics and Data Science Journal*, vol. 1, no. 3, pp. 290–299, Dec. 2023, doi: 10.20885/esds.vol1.iss.3.art36.

[7] R. Swastika, S. Mukodimah, F. Susanto, M. Muslihudin, and I. Sri, *IMPLEMENTASI DATA MINING (Clastering, Association, Prediction, Estimation, Classification)*. Penerbit Adab, 2023.

[8] Y. A. N. S. Putra and H. Margono, "Simulation of Student Study Group Formation Design Using K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 2, pp. 598–608, Mar. 2025, doi: 10.57152/malcom.v5i2.1795.