JURNAL TEKNIK INFORMATIKA

Vol. 6 No. 1 Tahun 2026
ISSN 2809-1353
DOI: 10.58794/jekin.v6i2.2131

Analisis Data Pendidikan Menggunakan Metode Data Mining dengan Algoritma *Decision Tree* dan *K-Means Clustering*

Hisyam Akmal Maulana[1], Nilatul Fitriyah[2], Dika El Baradei[3], Farid Ardiyanto[4], Muhammad Arifin[5]
[1,2,3,4,5)] Program Studi Sistem Informasi, Universitas Muria Kudus
e-mail: *[1]202353086@std.umk.ac.id , [2]202353082@std.umk.ac.id, [3]202353072@std.umk.ac.id, [4]202353079@std.umk.ac.id, [5]arifin.m@umk.ac.id

Abstract* - This study aims to analyze educational data patterns to identify various factors that influence the quality of learning in elementary schools using a data mining approach. The methods applied in this study include classification techniques using the Decision Tree algorithm (C4.5) and clustering techniques with the K-Means algorithm applied in the Knowledge Discovery in Database (KDD) stage. The data used is secondary data in Excel format containing information on the number of students who repeat classes based on region and level of education. The research process was carried out through several stages, namely data preprocessing, data transformation, data mining application, and model evaluation. In the evaluation stage, the classification model was tested using the hold-out method with a data division of 80% for training data and 20% for testing data. The results showed that the Decision Tree algorithm obtained an accuracy level of 62.62%, although there were indications of potential overfitting caused by high correlation between variables. Meanwhile, the application of the K-Means algorithm produced three main groups with fairly good clustering quality based on the silhouette coefficient value. This research contributes in the form of integration of classification and clustering methods to produce more comprehensive educational data analysis and support data-based decision-making processes in an effort to improve the quality of learning.

Keywords: data mining, learning quality, Decision Tree, K-Means clustering, elementary schools

Abstrak* - Penelitian ini bertujuan untuk menganalisis pola data pendidikan guna mengidentifikasi berbagai faktor yang berpengaruh terhadap kualitas pembelajaran di sekolah dasar dengan memanfaatkan pendekatan data mining. Metode yang diterapkan dalam penelitian ini mencakup teknik klasifikasi menggunakan algoritma *Decision Tree* (C4.5) serta teknik *clustering* dengan algoritma *K-Means* yang diterapkan dalam tahapan *Knowledge Discovery in Database* (KDD). Data yang digunakan berupa data sekunder dalam format Excel yang memuat informasi mengenai jumlah peserta didik yang mengulang berdasarkan wilayah dan jenjang pendidikan. Proses penelitian dilakukan melalui beberapa tahapan, yaitu *preprocessing* data, transformasi data, penerapan data mining, hingga evaluasi model. Pada tahap evaluasi, model klasifikasi diuji menggunakan metode *hold-out* dengan pembagian data sebesar 80% untuk data pelatihan dan 20% untuk data pengujian. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* memperoleh tingkat akurasi sebesar 62,62%, meskipun ditemukan indikasi adanya potensi *overfitting* yang disebabkan oleh tingginya korelasi antar variabel. Sementara itu, penerapan algoritma *K-Means* menghasilkan tiga kelompok utama dengan kualitas pengelompokan yang tergolong cukup baik berdasarkan nilai *silhouette coefficient*. Penelitian ini memberikan kontribusi dalam bentuk integrasi metode klasifikasi dan *clustering* untuk menghasilkan analisis data pendidikan yang lebih komprehensif serta mendukung proses pengambilan keputusan berbasis data dalam upaya meningkatkan kualitas pembelajaran.

Kata Kunci : data mining, kualitas pembelajaran, *Decision Tree, K-Means clustering,* sekolah dasar

I. PENDAHULUAN

Pendidikan dasar merupakan jenjang pendidikan yang memiliki posisi strategis dalam membangun kemampuan akademik, pembentukan karakter, serta pengembangan keterampilan sosial peserta didik sebagai bekal untuk melanjutkan pendidikan pada tingkat berikutnya. Dengan demikian, mutu proses pembelajaran di sekolah dasar menjadi salah satu aspek penting dalam menentukan keberhasilan sistem pendidikan secara keseluruhan. Pembelajaran yang bermutu tidak hanya tercermin dari tingginya hasil belajar siswa, tetapi juga dari kemampuan peserta didik dalam mengembangkan pola pikir kritis, kreatif, serta memiliki motivasi belajar yang konsisten dan berkelanjutan [1].

Meskipun demikian, pelaksanaan pembelajaran di sekolah dasar hingga saat ini masih menghadapi berbagai tantangan yang cukup kompleks. Salah satu permasalahan yang sering ditemukan ialah dominannya penggunaan pendekatan pembelajaran yang berorientasi pada guru (*teacher-centered*), sehingga siswa cenderung kurang aktif dalam mengikuti proses pembelajaran [2]. Kondisi tersebut berdampak pada rendahnya tingkat partisipasi dan keterlibatan peserta didik selama kegiatan belajar berlangsung. Di samping itu, penerapan strategi pembelajaran yang inovatif, interaktif, dan relevan dengan konteks kehidupan siswa juga masih belum optimal. Akibatnya, pemahaman konsep yang dimiliki siswa menjadi kurang mendalam dan keterlibatan mereka dalam proses pembelajaran belum berkembang secara maksimal.

Berbagai penelitian menunjukkan bahwa kualitas pembelajaran dipengaruhi oleh faktor internal dan eksternal yang saling berinteraksi. Faktor internal, seperti motivasi belajar, kesiapan siswa, dan kondisi psikologis, memiliki pengaruh signifikan terhadap keberhasilan pembelajaran [3]. Sementara itu, faktor eksternal, seperti metode pembelajaran, lingkungan belajar, dukungan keluarga, serta ketersediaan sarana dan prasarana pendidikan dan kompetensi guru, juga berperan penting dalam meningkatkan efektivitas proses pembelajaran [4][5][6].

Meskipun berbagai penelitian telah mengkaji faktor-faktor tersebut, sebagian besar masih dilakukan secara parsial dan menggunakan pendekatan deskriptif atau korelasional sederhana. Pendekatan tersebut umumnya hanya mampu menunjukkan hubungan antar variabel secara terbatas dan belum mampu mengungkap pola yang lebih kompleks dan tersembunyi (*hidden patterns*) dalam data pendidikan [7]. Kondisi tersebut menunjukkan bahwa masih terdapat kesenjangan penelitian (*research gap*) dalam pengembangan analisis data pendidikan yang dilakukan secara lebih menyeluruh, sistematis, dan berbasis pada pemanfaatan data yang akurat.

Untuk mengatasi berbagai permasalahan tersebut, diperlukan suatu pendekatan analisis yang lebih terstruktur dan berbasis pada pengolahan data secara sistematis, salah satunya melalui penerapan data mining. Data mining merupakan metode yang digunakan untuk mengeksplorasi data dalam jumlah besar guna menemukan pola, hubungan, maupun struktur tertentu secara otomatis dan akurat. Dalam penelitian ini, algoritma Decision Tree digunakan untuk melakukan proses klasifikasi sekaligus mengidentifikasi faktor-faktor yang paling dominan mempengaruhi data penelitian. Selain itu, algoritma *K-Means Clustering* dimanfaatkan untuk mengelompokkan data berdasarkan tingkat kesamaan karakteristik tertentu, sehingga mampu menghasilkan kelompok data yang lebih terorganisasi dan mudah dianalisis [8].

Berbeda dengan penelitian sebelumnya yang umumnya menggunakan satu metode secara terpisah, penelitian ini mengintegrasikan metode klasifikasi dan *clustering* dalam satu kerangka analisis. Pendekatan tersebut diharapkan dapat menghasilkan analisis yang lebih menyeluruh dalam memahami pola data pendidikan, terutama dalam mengkaji jumlah peserta didik yang mengulang sebagai salah satu indikator untuk menggambarkan kondisi dan kualitas pembelajaran pada suatu wilayah tertentu.

Berdasarkan uraian latar belakang tersebut, penelitian ini bertujuan untuk mengidentifikasi pola data pendidikan dengan menerapkan pendekatan data mining melalui kombinasi metode *Decision Tree* dan *K-Means Clustering*. Integrasi kedua metode ini diharapkan dapat memberikan analisis yang lebih mendalam dalam mendukung proses pengambilan keputusan yang berbasis data, khususnya dalam upaya peningkatan kualitas pembelajaran di lingkungan pendidikan [7].

II. PENELITIAN YANG TERKAIT

Penelitian terdahulu merupakan komponen penting dalam menunjukkan posisi dan kontribusi suatu penelitian terhadap kajian yang telah ada. Dalam konteks ini, penelitian difokuskan pada penerapan metode data mining di bidang pendidikan, khususnya melalui penggunaan algoritma klasifikasi dan *clustering*.

Berbagai studi sebelumnya menunjukkan bahwa penerapan *data mining* dalam analisis data pendidikan dapat menghasilkan informasi yang efektif serta bersifat informatif dalam mendukung pengambilan keputusan. Salah satu metode yang paling sering digunakan adalah algoritma *Decision Tree* C4.5, yang dikenal memiliki kemampuan membangun model klasifikasi dengan tingkat akurasi yang relatif baik. Selain itu, metode ini juga memiliki keunggulan dari sisi *interpretabilita*s, karena hasil model yang dihasilkan mudah dipahami oleh pengguna sehingga memudahkan proses analisis [3][5]. Secara teknis, algoritma ini bekerja dengan membentuk struktur pohon keputusan berdasarkan perhitungan *entropy* dan *information gain* pada setiap atribut data. Proses tersebut memungkinkan sistem untuk menentukan atribut yang paling berpengaruh terhadap hasil klasifikasi secara terukur dan sistematis. Dengan demikian, pola yang terdapat dalam data dapat diidentifikasi dan dipahami dengan lebih jelas serta terstruktur.

Selain itu, metode *K- Means Clustering* juga banyak diterapkan dalam proses pengelompokan data, khususnya pada bidang pendidikan. Hasil berbagai penelitian sebelumnya menunjukkan bahwa algoritma ini mampu mengelompokkan data berdasarkan tingkat kesamaan karakteristik secara efektif, sehingga mempermudah proses analisis distribusi data secara lebih terstruktur [4][6]. Melalui pendekatan ini, data dapat dikelompokkan ke dalam beberapa kategori tertentu, seperti kelompok rendah, sedang, dan tinggi berdasarkan nilai atau atribut yang dimiliki.

Dalam bidang pendidikan, penerapan data mining telah dimanfaatkan untuk berbagai keperluan, seperti analisis hasil belajar, pengelompokan siswa, serta evaluasi kondisi pembelajaran. Penelitian sebelumnya menunjukkan bahwa data pendidikan dapat diolah untuk menemukan pola tersembunyi yang tidak tampak secara langsung, sehingga dapat dijadikan dasar dalam pengambilan keputusan [1][2].

Selain itu, beberapa penelitian juga menekankan pentingnya faktor-faktor seperti kualitas pembelajaran, lingkungan belajar, dan metode pengajaran dalam mempengaruhi hasil pendidikan [8]. Namun, sebagian besar penelitian tersebut masih menggunakan pendekatan statistik maupun kualitatif, sehingga masih terbuka peluang untuk pengembangan pendekatan berbasis data mining [9].

Berbeda dengan penelitian sebelumnya, penelitian ini tidak secara langsung menganalisis faktor-faktor seperti motivasi atau metode pengajaran, melainkan menggunakan jumlah peserta didik yang mengulang sebagai indikator dalam melihat pola kondisi pembelajaran. Pendekatan ini memberikan perspektif yang berbeda dalam analisis data pendidikan, khususnya dalam mengidentifikasi karakteristik wilayah berdasarkan data yang tersedia.

Dengan mengkombinasikan metode klasifikasi menggunakan *Decision Tree* dan metode *clustering* melalui *K-Means*, penelitian ini diharapkan mampu menghasilkan analisis yang lebih komprehensif, baik dalam hal pengelompokan data

maupun dalam memahami pola yang terbentuk [10]. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan pemanfaatan data mining pada bidang pendidikan, khususnya dalam analisis data yang berbasis wilayah. Kontribusi tersebut mencakup upaya untuk menghasilkan informasi yang lebih terstruktur dan berbasis data dalam memahami kondisi pendidikan di berbagai daerah. Selain itu, hasil penelitian ini juga diharapkan dapat menjadi dasar dalam perumusan strategi peningkatan kualitas pendidikan yang lebih tepat sasaran. Dengan demikian, penerapan data mining tidak hanya bersifat analitis, tetapi juga memiliki nilai praktis dalam mendukung kebijakan pendidikan berbasis data [11].

III. METODE PENELITIAN

Penelitian ini menggunakan pendekatan data mining dengan tujuan untuk menemukan pola-pola serta hubungan antar variabel yang terdapat dalam data pendidikan yang sedang dianalisis [12]. Pemilihan pendekatan tersebut didasarkan pada kemampuannya dalam mengekstraksi pola-pola tersembunyi (*hidden patterns*) serta menghasilkan model analisis yang bersifat objektif dan berbasis data, sehingga hasil yang diperoleh dapat dijadikan dasar yang kuat dalam proses pengambilan keputusan. Dalam penelitian ini, teknik *data mining* dimanfaatkan untuk mengkaji jumlah peserta didik yang mengulang sebagai salah satu indikator dalam menilai kualitas proses pembelajaran [9]. Melalui pendekatan tersebut, diharapkan dapat diperoleh pemahaman yang lebih komprehensif mengenai kondisi proses pembelajaran berdasarkan data yang telah tersedia, sehingga hasil analisis dapat memberikan kontribusi yang signifikan dalam pengambilan kebijakan pendidikan.

A. Tahapan Penelitian

Penelitian ini menggunakan tahapan *Knowledge Discovery in Database* (KDD) yang mencakup proses pemilihan data (*data selection)*, *pra-pemrosesan* (*preprocessing*), *transformasi* data (*data transformation*), proses data mining, hingga tahap evaluasi hasil [13]. Pemilihan model KDD didasarkan pada kemampuannya dalam menyediakan alur kerja yang terstruktur dan sistematis dalam proses pengolahan data. Setiap tahapan dalam KDD saling terintegrasi, dimulai dari proses pemilihan data hingga menghasilkan informasi yang memiliki nilai guna. Dengan penerapan kerangka tersebut, proses analisis data dapat berjalan lebih sistematis dan terstruktur. Di samping itu, pendekatan KDD juga membantu peneliti dalam memastikan bahwa setiap tahapan pengolahan data dilakukan secara tepat sehingga menghasilkan hasil yang lebih optimal. Oleh karena itu, model ini dinilai tepat untuk mendukung pencapaian tujuan penelitian dalam menghasilkan informasi yang akurat dan bermakna.

1. Data *Selection*

Dataset yang digunakan dalam penelitian ini merupakan data sekunder dalam bentuk file Microsoft Excel yang memuat informasi jumlah peserta didik yang mengulang berdasarkan wilayah dan jenjang pendidikan. Dataset terdiri dari 1.027 data (*record*) dengan 8 variabel (*atribut*) utama yang relevan dengan tujuan penelitian. Pemilihan atribut dilakukan secara selektif dengan mempertimbangkan relevansinya terhadap indikator kualitas pembelajaran. Oleh karena itu, hanya variabel-variabel yang memiliki hubungan langsung dengan tujuan penelitian yang diikutsertakan dalam proses analisis. Pendekatan ini bertujuan untuk memastikan bahwa hasil analisis yang diperoleh lebih terarah dan sesuai dengan fokus penelitian yang telah ditetapkan [14].

2. Data *Preprocessing*

Tahap *preprocessing* dilakukan untuk memperbaiki dan meningkatkan kualitas data sebelum proses analisis dilakukan. Proses ini meliputi pembersihan data (*data cleaning*), penghapusan data duplikat, serta penanganan *missing values*. Selain itu, dilakukan juga reduksi data dengan menghilangkan atribut yang tidak relevan. Proses ini bertujuan untuk memastikan bahwa data yang digunakan memiliki kualitas yang layak, sehingga dapat mendukung peningkatan akurasi model yang dihasilkan [15].

3. Data *Transformation*

Pada Proses ini, data yang telah dibersihkan kemudian diolah kembali agar sesuai dengan kebutuhan analisis. Proses transformasi dilakukan dengan mengubah data kategorik menjadi bentuk numerik menggunakan teknik encoding, serta melakukan normalisasi untuk menyamakan skala pada setiap variabel. Selain itu, dilakukan pula analisis korelasi guna mengidentifikasi hubungan antar variabel dan mendeteksi kemungkinan terjadinya multikolinearitas. Variabel-variabel yang menunjukkan tingkat korelasi tinggi kemudian diseleksi dan dieliminasi melalui proses feature selection. Langkah ini bertujuan untuk mengurangi redudansi data serta meningkatkan kualitas dan performa model dalam tahap analisis selanjutnya.

4. Data Mining

Tahap data mining menjadi bagian utama dalam penelitian ini dengan menggunakan dua metode, yaitu klasifikasi dan *clustering.*

1. Klasifikasi (*Decision Tree C4.5*)

Metode klasifikasi yang digunakan pada penelitian ini digunakan untuk membangun model prediktif yang bertujuan mengidentifikasi faktor-faktor yang mempengaruhi kualitas pembelajaran. Algoritma yang diterapkan adalah *Decision Tree* dengan menggunakan kriteria entropy sebagai dasar dalam menentukan atribut terbaik pada setiap proses pembentukan pohon keputusan [16]. Dataset yang digunakan kemudianmdipisahkan menjadi dua bagian, yaitu data pelatihan (training) dan data pengujian (*testing*) dengan rasio 80:20 menggunakan metode *hold-out*. Pembagian ini bertujuan agar model yang dikembangkan dapat diuji secara objektif menggunakan data yang tidak terlibat dalam proses pelatihan. Dengan cara ini, kinerja model dapat dievaluasi secara lebih tepat dalam menggambarkan kemampuan prediksi terhadap data yang belum pernah diproses sebelumnya [17].

2. Clustering (K-Means)

Metode *clustering* dalam penelitian ini dimanfaatkan untuk mengelompokkan data berdasarkan kesamaan atau kemiripan karakteristik yang dimiliki oleh setiap data yang dimiliki. Algoritma yang digunakan dalam penelitian ini adalah *K-Means*, di mana jumlah *cluster* (k) ditentukan menggunakan metode *Elbow* untuk mendapatkan jumlah cluster yang paling optimal.

Pendekatan ini memungkinkan identifikasi titik keseimbangan antara jumlah *cluster* dan variasi dalam data. Hasil dari proses *clustering* kemudian menghasilkan pembagian data menjadi beberapa kelompok yang mempunyai karakteristik serupa. Dengan demikian, pola-pola dalam dataset dapat terlihat lebih jelas dan terstruktur, sehingga memudahkan dalam proses interpretasi dan analisis lebih lanjut [18].

5. Evaluasi Model

Tahap evaluasi dalam penelitian ini bertujuan untuk melakukan pengukuran serta penilaian terhadap kinerja yang diamati yang telah dibangun. Pada metode klasifikasi, proses evaluasi dilakukan dengan menggunakan metrik akurasi serta confusion matrix guna mengetahui tingkat ketepatan hasil prediksi yang dihasilkan oleh model. Penggunaan kedua metrik tersebut memungkinkan penilaian yang lebih komprehensif terhadap performa klasifikasi. Sementara itu, pada metode *clustering*, evaluasi dilakukan dengan memanfaatkan silhouette coefficient untuk menilai sejauh mana kualitas pengelompokan data yang terbentuk. Nilai ini memberikan gambaran mengenai tingkat keterpisahan antar *cluster* serta kedekatan data dalam satu kelompok. Secara keseluruhan, tahap evaluasi ini berperan penting dalam memastikan bahwa model yang dihasilkan memiliki tingkat keandalan dan validitas yang baik sehingga dapat mendukung hasil analisis secara optimal.

B. Tools dan Perangkat Analisis

Proses analisis data dalam penelitian ini dilakukan menggunakan bahasa pemrograman Python yang dijalankan melalui platform Google Colab. Pada tahap pengolahannya, penelitian memanfaatkan sejumlah library pendukung untuk menunjang kebutuhan analisis data secara optimal. Library Pandas digunakan dalam proses pengelolaan, pembersihan, dan manipulasi dataset, sedangkan *NumPy* dimanfaatkan untuk mendukung perhitungan numerik dan operasi matematis. Selain itu, *Scikit-learn* digunakan dalam penerapan metode *machine learning*, sementara *Matplotlib* berperan dalam menyajikan visualisasi data agar hasil analisis lebih mudah dipahami. Pemanfaatan Python memberikan kemudahan dalam mengelola data secara lebih terstruktur karena didukung oleh sintaks yang sederhana dan fleksibel. Dukungan berbagai pustaka yang lengkap juga membantu meningkatkan efisiensi proses analisis sehingga tahapan penelitian dapat dilakukan secara sistematis dan terorganisasi dengan baik. Oleh karena itu, penggunaan Python dinilai mampu mendukung proses analisis data secara lebih efektif dan akurat dalam penelitian ini.

IV. HASIL DAN PEMBAHASAN

A. Data Selection

Pada tahap awal penelitian, dilakukan kegiatan seleksi data yang bertujuan untuk menetapkan dataset yang akan digunakan dalam proses analisis. Dataset yang digunakan dalam penelitian ini berupa data sekunder yang diperoleh dalam format file Excel, yang berisi informasi mengenai jumlah peserta didik yang mengulang berdasarkan wilayah serta jenjang pendidikan. Proses pemilihan data tersebut didasarkan pada tingkat kesesuaiannya dengan tujuan penelitian, yakni untuk mengkaji kualitas pembelajaran melalui indikator pengulangan siswa. Data yang telah melalui tahap seleksi ini selanjutnya digunakan sebagai landasan utama dalam pelaksanaan analisis pada tahap berikutnya. Berikut dibawah ialah *resume* data awal sebelum diubah.

	JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
1	1	38	38	514	514	1	142	102	87	98	80	13	255
2	0	0	0	4	0	0	56	2	2	81	48	4	192
3	0	0	0	4	0	2	0	0	0	0	0	0	0
4	0	0	0	8	1	0	10	45	82	12	69	1	177
5	0	0	0	8	1	2	0	0	0	1	0	0	1

Gambar 1. Data Awal

Proses pra-pemrosesan tersebut dilakukan dengan memanfaatkan bahasa pemrograman python yang dioperasikan melalui platform Google Colab, dengan serangkaian tahapan yang dirancang secara sistematis dan terstruktur.

B. Data Preprocessing

Tahap preprocessing merupakan langkah penting yang dilakukan untuk memperbaiki dan meningkatkan mutu data sebelum dilakukan proses analisis lebih lanjut. Pada tahap ini dilakukan berbagai proses seperti pembersihan data (*data cleaning*), penghapusan data duplikat, serta penanganan nilai yang hilang *(missing values*). Data yang tidak lengkap atau tidak konsisten dapat menimbulkan bias dan mengurangi validitas hasil analisis, sehingga perlu ditangani secara terstruktur. Melalui tahap *preprocessing* ini, data yang awalnya belum tersusun dengan baik kemudian diolah menjadi lebih rapi, konsisten, dan siap digunakan pada proses data mining.

Selanjutnya, data dibersihkan untuk memastikan kualitas dataset sebelum dilakukan analisis lebih lanjut. Proses ini meliputi penghapusan karakter yang tidak diperlukan seperti tanda koma, titik, dan simbol lainnya yang dapat mengganggu proses komputasi. Setelah itu, seluruh data diubah ke dalam format numerik dengan menggunakan fungsi konversi numerik pada Python agar dapat diproses pada tahap analisis berikutnya.

	JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Periode	Wilayah	Kode Wilayah	Kota/Kab	Kode Kota/ Kab	Status Sekolah	Tingkat - I	Tingkat - II	Tingkat - III	Tingkat - IV	Tingkat - V	Tingkat - VI	Jumlah
2	2024	PROV. ACEH	11	KAB. ACEH SELATAN	1101	Negeri	21	10	10	7	4	2	54
3	2024	PROV. ACEH	11	KAB. ACEH SELATAN	1101	Swasta	0	0	0	0	0	0	0
4	2024	PROV. ACEH	11	KAB. ACEH TENGGARA	1102	Negeri	11	3	9	13	7	1	44

Gambar 2. Hasil Insert Dataset

C. Data Cleaning

Tahap *data cleaning* dilakukan untuk mengidentifikasi sekaligus memperbaiki berbagai kesalahan maupun ketidakkonsistenan yang terdapat dalam dataset. Data yang diperoleh dari hasil kajian artikel memiliki kemungkinan mengandung duplikasi, perbedaan dalam penulisan, serta informasi yang tidak lengkap. Oleh karena itu, proses pembersihan data mencakup beberapa langkah penting, seperti menghapus data duplikat yang memiliki isi serupa atau berulang agar tidak mempengaruhi hasil analisis, menangani data yang hilang (*missing value*) baik dengan menghapusnya maupun menggantinya dengan nilai tertentu yang relevan, menyeragamkan format data terutama pada kategori yang memiliki makna sama tetapi ditulis secara berbeda, serta memperbaiki inkonsistensi data yang disebabkan oleh kesalahan input atau penggunaan istilah yang tidak konsisten. Melalui tahapan ini, diharapkan data yang digunakan dalam proses analisis telah bersih, akurat, dan memiliki kualitas yang baik sehingga mampu menghasilkan output yang valid dan reliabel.

	JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
1	Periode	Wilayah	Kode Wilayah	Kota/Kab	Kode Kota/ Kab	Status Sekolah	Tingkat - I	Tingkat - II	Tingkat - III	Tingkat - IV	Tingkat - V	Tingkat - VI	Jumlah
2	2024	PROV. ACEH	11	KAB. ACEH SELATAN	1101	Negeri	21	10	10	7	4	2	54
3	2024	PROV. ACEH	11	KAB. ACEH SELATAN	1101	Swasta	0	0	0	0	0	0	0
4	2024	PROV. ACEH	11	KAB. ACEH TENGGARA	1102	Negeri	11	3	9	13	7	1	44
5	2024	PROV. ACEH	11	KAB. ACEH TENGGARA	1102	Swasta	0	0	0	1	0	0	1

Gambar 3. Data Setelah *Cleaning*

D. Encoding dan Normalisasi Data

Tahapan transformasi data dilaksanakan setelah seluruh proses preprocessing selesai dilakukan. Pada fase ini, data yang sebelumnya telah melalui proses pembersihan dan normalisasi kemudian disesuaikan kembali agar dapat digunakan dalam proses analisis menggunakan algoritma *Decision Tree* serta *K-Means Clustering*. Proses transformasi dilakukan untuk memastikan bahwa bentuk dan struktur data telah memenuhi kebutuhan algoritma machine learning sehingga tahapan klasifikasi maupun clustering dapat dijalankan secara maksimal. Berdasarkan hasil transformasi, seluruh data berhasil diubah ke dalam format numerik yang terstruktur tanpa ditemukan *missing values* ataupun data duplikat yang berpotensi mempengaruhi hasil analisis. Kondisi data yang sudah konsisten dan tersusun dengan baik tersebut memungkinkan proses data mining berlangsung lebih efisien serta mendukung pembentukan model analisis dengan tingkat akurasi yang lebih optimal.

E. Data Reduction

Tahap data *reduction* dilakukan dengan tujuan menyederhanakan dataset melalui pemilihan atribut atau variabel yang paling relevan dengan kebutuhan analisis. Langkah ini penting untuk menekan tingkat kompleksitas data sekaligus meningkatkan efisiensi proses serta akurasi model yang digunakan. Dalam tahap ini, terdapat beberapa teknik yang diterapkan, di antaranya seleksi fitur (*feature selection*), yaitu memilih variabel yang memiliki pengaruh signifikan terhadap variabel target. Selain itu, dilakukan penghapusan atribut yang tidak relevan, yakni variabel yang tidak memberikan kontribusi terhadap hasil analisis. Proses reduksi juga mencakup pengurangan dimensi data agar proses komputasi menjadi lebih cepat dan efisien, serta menghindari redudansi dengan menghapus variabel yang memiliki informasi serupa atau berulang. Melalui tahapan ini, data yang digunakan menjadi lebih ringkas, terfokus, serta mampu menghasilkan analisis yang lebih akurat dan mudah untuk diinterpretasikan.

```
=== SETELAH DATA REDUCTION ===
   JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI  Unnamed: 1  Unnamed: 3  Unnamed: 5
1                                                                          0          38         514           1
2                                                                          1           0           4           0
3                                                                          1           0           4           2
4                                                                          1           0           8           0
5                                                                          1           0           8           2
Jumlah fitur sebelum: 13
Jumlah fitur sesudah: 4
```

Gambar 4. Data Setelah *Reduction*

F. Data Transformation

Pada tahap transformasi, data yang sebelumnya telah melalui proses pembersihan selanjutnya dikonversi ke dalam bentuk yang lebih sesuai untuk kebutuhan analisis. Proses ini mencakup pengubahan data kategorikal menjadi representasi numerik (*encoding*), normalisasi nilai data, serta penyesuaian struktur dataset agar selaras dengan kebutuhan algoritma yang akan digunakan. Selain itu, tahap ini juga melibatkan analisis korelasi antar variabel guna memahami keterkaitan antara atribut dalam dataset. Hasil pengujian menunjukkan bahwa terdapat beberapa variabel dengan tingkat korelasi yang sangat tinggi, yang mengindikasikan adanya gejala multikolinearitas. Kondisi tersebut berpotensi menurunkan kualitas model, sehingga dilakukan proses seleksi fitur (*feature selection*) dengan mengeliminasi variabel-variabel yang memiliki korelasi tinggi.

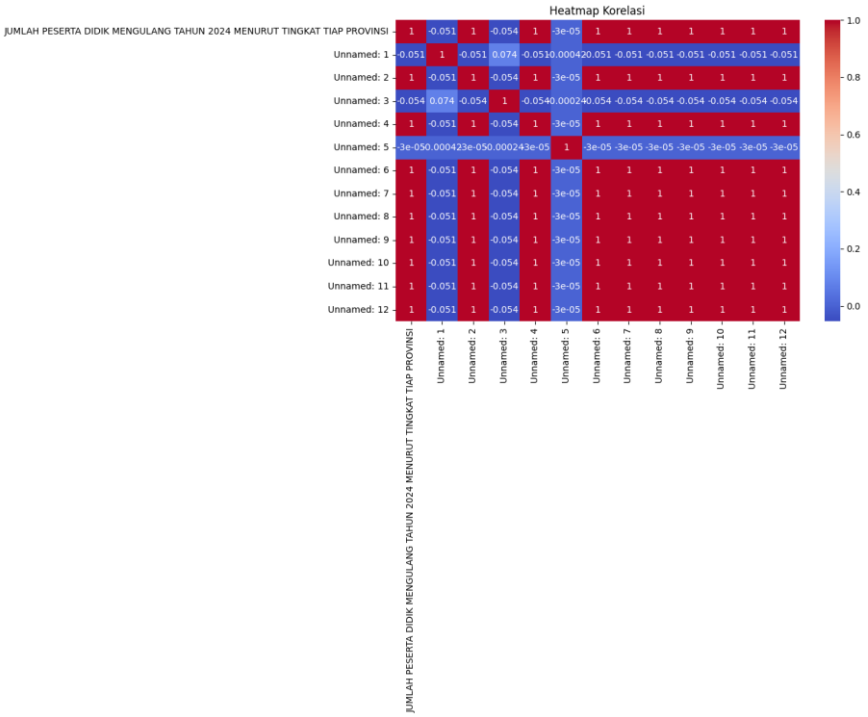
Langkah ini bertujuan untuk meminimalkan redundansi data sekaligus meningkatkan kinerja dan stabilitas model dalam proses analisis selanjutnya.

Data yang telah melewati tahap preprocessing selanjutnya diubah menggunakan teknik normalisasi dengan metode *StandardScaler*. Proses ini dilakukan untuk menyetarakan rentang nilai pada setiap variabel agar tidak ada atribut yang memiliki pengaruh lebih dominan selama proses *clustering* berlangsung. Normalisasi menjadi tahapan yang penting, terutama pada algoritma *K-Means*, karena algoritma tersebut sangat dipengaruhi oleh perbedaan skala antar data. Dengan dilakukannya normalisasi, setiap variabel dapat memberikan kontribusi yang lebih proporsional dalam proses pengelompokan sehingga hasil clustering menjadi lebih optimal dan akurat.

=== SETELAH TRANSFORMASI ===

	JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
1	0	38	0	514	0	1	0	0	0	0	0	0	0
2	1	0	1	4	1	0	1	1	1	1	1	1	1
3	1	0	1	4	1	2	1	1	1	1	1	1	1
4	1	0	1	8	1	0	1	1	1	1	1	1	1
5	1	0	1	8	1	2	1	1	1	1	1	1	1

Gambar 6. Data Setelah *Transformation*



Gambar 5. Grafik *Heatmap* Korelasi

G. Analisis Heatmap Korelasi

Analisis korelasi pada penelitian ini digunakan untuk mengukur tingkat hubungan antara variabel yang terdapat dalam dataset. Proses tersebut dilakukan untuk mengetahui seberapa besar keterkaitan antar atribut sekaligus mendeteksi kemungkinan terjadinya multikolinearitas pada data penelitian. Hasil perhitungan korelasi kemudian divisualisasikan menggunakan heatmap sehingga pola hubungan antar variabel dapat diamati secara lebih jelas dan informatif. Dalam analisis korelasi, nilai koefisien berada pada rentang −1 hingga 1. Nilai yang mendekati angka 1 menunjukkan hubungan yang sangat kuat dan memiliki arah yang sama, sedangkan nilai yang mendekati 0 mengindikasikan hubungan yang rendah atau lemah. Sementara itu, nilai yang mendekati −1 menandakan adanya hubungan yang kuat namun memiliki arah yang berlawanan antar variabel.

Berdasarkan visualisasi heatmap yang diperoleh, sebagian besar variabel memiliki tingkat korelasi yang tinggi bahkan mendekati angka 1. Hal tersebut menunjukkan bahwa beberapa atribut mempunyai pola data yang hampir sama sehingga berpotensi menyebabkan redundansi informasi dalam dataset. Di sisi lain, terdapat pula beberapa variabel dengan nilai korelasi rendah atau mendekati 0 yang mengindikasikan bahwa hubungan antar atribut tersebut relatif lemah dan tidak terlalu signifikan. Tingginya nilai korelasi pada sejumlah variabel juga dapat menjadi tanda adanya multikolinearitas, yaitu kondisi ketika antar atribut memiliki hubungan yang terlalu kuat dan dapat mempengaruhi performa model analisis.

Dalam penerapan data mining, multikolinearitas dapat memberikan pengaruh terhadap hasil analisis yang dilakukan. Pada algoritma *Decision Tree*, dampak dari multikolinearitas umumnya tidak terlalu dominan, tetapi tetap dapat menyebabkan struktur pohon keputusan menjadi kurang efektif. Sementara itu, pada algoritma *K-Means Clustering*, korelasi yang terlalu tinggi antar variabel dapat mempengaruhi proses pembentukan cluster karena distribusi data menjadi kurang mampu merepresentasikan perbedaan karakteristik setiap kelompok. Oleh sebab itu, diperlukan tahapan *preprocessing* tambahan, seperti menghapus atribut yang kurang relevan, mengurangi variabel dengan korelasi sangat tinggi, serta menerapkan feature

selection untuk meningkatkan kualitas dataset sebelum proses klasifikasi dan *clustering* dijalankan.

Secara umum, hasil analisis heatmap menunjukkan bahwa dataset memiliki hubungan antar variabel yang cukup kuat. Akan tetapi, tingginya korelasi pada beberapa atribut menandakan bahwa dataset masih memerlukan proses optimasi lebih lanjut agar model yang dihasilkan mampu memberikan tingkat akurasi dan kualitas analisis yang lebih baik.

H. Data Mining (Klasifikasi – Decision Tree)

Pada tahap data mining, proses klasifikasi dilakukan menggunakan algoritma *Decision Tree* C4.5 dengan metode *entropy* sebagai dasar dalam membentuk struktur pohon keputusan. Algoritma ini dipilih karena mampu menentukan atribut yang paling berpengaruh dalam proses klasifikasi melalui perhitungan information gain. Selain itu, *Decision Tree* memiliki keunggulan dalam menghasilkan model yang mudah dipahami karena dapat divisualisasikan dalam bentuk pohon keputusan, sehingga interpretasi hasil analisis menjadi lebih jelas dan terstruktur.

Sebelum proses klasifikasi dijalankan, dataset terlebih dahulu dibagi menjadi dua bagian, yaitu data training dan data testing dengan rasio 80:20 menggunakan metode *hold-out validation*. Data training digunakan untuk membangun model klasifikasi, sedangkan data testing dimanfaatkan untuk menguji kemampuan model dalam memprediksi data baru yang sebelumnya belum pernah diproses. Pembagian dataset tersebut dilakukan untuk meminimalkan potensi bias pada model sekaligus memastikan bahwa proses evaluasi dapat dilakukan secara lebih objektif.

I. Hasil Akurasi Model

Hasil pengujian menunjukkan bahwa model *Decision Tree* memperoleh tingkat akurasi sebesar 62,62%. Nilai tersebut menandakan bahwa model memiliki kemampuan yang sangat baik dalam mengklasifikasikan data berdasarkan atribut yang digunakan. Tingginya nilai akurasi menunjukkan bahwa algoritma mampu mengenali pola hubungan antar variabel dalam dataset dengan cukup baik. Akan tetapi, hasil akurasi yang sangat tinggi juga perlu dikaji lebih lanjut karena dapat mengindikasikan adanya overfitting pada model, terutama karena hasil analisis korelasi sebelumnya memperlihatkan adanya hubungan yang sangat kuat antar beberapa variabel dalam dataset.

Selain memanfaatkan nilai akurasi, evaluasi terhadap model juga dilakukan menggunakan *confusion matrix* untuk memperoleh gambaran hasil klasifikasi secara lebih mendetail. Metode ini digunakan untuk menilai seberapa tepat model dalam mengelompokkan data ke dalam kelas yang sesuai serta mengetahui kemungkinan adanya kesalahan prediksi pada proses klasifikasi. Melalui *confusion matrix*, performa model dapat dianalisis lebih menyeluruh karena mampu menunjukkan distribusi prediksi benar maupun prediksi yang keliru. Oleh sebab itu, evaluasi tidak hanya berfokus pada nilai akurasi semata, tetapi juga mempertimbangkan kemampuan model dalam mengenali pola dan melakukan generalisasi terhadap data baru dengan lebih baik.

Hasil Output :

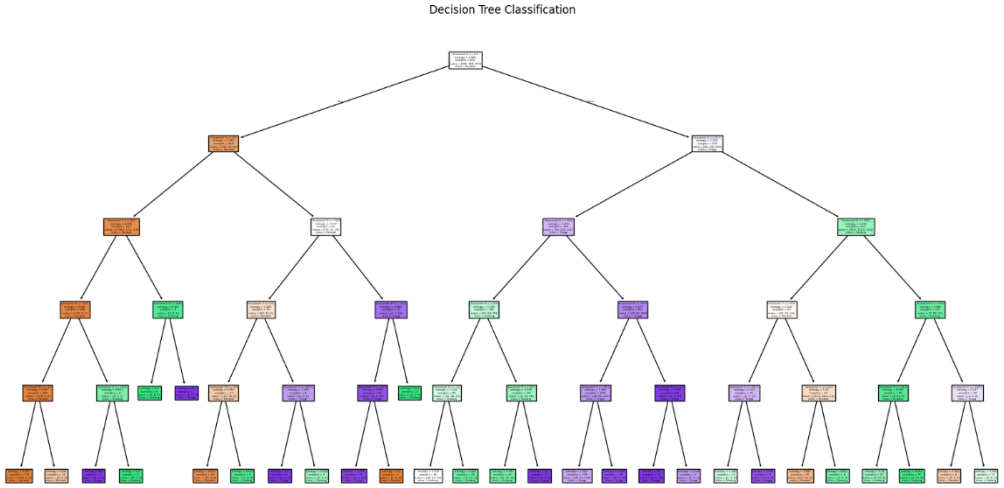
```
==================================
MENENTUKAN TARGET
==================================
Kolom target yang digunakan:
Unnamed: 12

==================================
HASIL KATEGORI TARGET
==================================
   Unnamed: 12 Kategori
1          255   Tinggi
2          192   Tinggi
3            0   Rendah
4          177   Tinggi
5            1   Rendah

==================================
DATA TRAINING & TESTING
==================================
Jumlah Data Training : 822
Jumlah Data Testing  : 206

==================================
AKURASI MODEL
==================================
Accuracy : 0.6262135922330098
```

J. *Confusion Matrix* dan Evaluasi Klasifikasi

Visualisasi pohon keputusan yang dihasilkan menunjukkan bahwa proses klasifikasi terbentuk melalui beberapa percabangan berdasarkan atribut yang memiliki nilai information gain tertinggi. Setiap node pada pohon keputusan merepresentasikan proses pengambilan keputusan berdasarkan kondisi tertentu hingga menghasilkan kelas akhir pada *leaf*

node. Struktur pohon tersebut memperlihatkan bahwa terdapat beberapa atribut yang memiliki pengaruh lebih dominan dibandingkan atribut lainnya dalam menentukan hasil klasifikasi. Dengan demikian, Decision Tree tidak hanya berfungsi sebagai model prediksi, tetapi juga dapat digunakan untuk memahami pola hubungan antar variabel dalam dataset secara lebih interpretatif.

Walaupun model menunjukkan performa yang sangat baik, penelitian ini masih memiliki beberapa keterbatasan, terutama terkait kemungkinan *overfitting* akibat tingginya korelasi antar variabel dan kurangnya variasi data. Oleh karena itu, penelitian selanjutnya disarankan untuk menerapkan metode validasi yang lebih kuat, seperti cross-validation, serta melakukan perbandingan dengan algoritma klasifikasi lainnya agar diperoleh model yang lebih stabil dan memiliki kemampuan generalisasi yang lebih optimal.



Gambar 6. Visualisasi Model *Decision Tree* C4.5 Menggunakan Kriteria *Entropy*

K. Data Mining (Clustering – K-Means)

Selain melakukan klasifikasi, penelitian ini juga menerapkan teknik *clustering* dengan menggunakan algoritma *K-Means* untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik yang dimiliki. Penentuan jumlah cluster dilakukan melalui metode *Elbow*, yang menunjukkan bahwa jumlah cluster yang paling optimal adalah tiga. Berdasarkan hasil proses *clustering*, data terbagi ke dalam tiga kelompok utama yang mencerminkan kategori rendah, sedang, dan tinggi dilihat dari tingkat pengulangan siswa. Penerapan *clustering* ini memberikan pemahaman yang lebih mendalam terkait distribusi data berdasarkan karakteristik tertentu. Dengan demikian, pola-pola yang terbentuk dalam dataset dapat diidentifikasi dengan lebih jelas dan sistematis, sehingga mendukung proses interpretasi hasil analisis secara keseluruhan.

Setelah tahapan *preprocessing* dan transformasi data ke dalam format numerik selesai dilakukan, proses analisis dilanjutkan dengan penerapan metode *K-Means Clustering*. Metode ini digunakan untuk mengelompokkan data jumlah peserta didik yang mengulang berdasarkan tingkat kemiripan karakteristik pada masing-masing wilayah. Dengan pendekatan ini, data dapat dibagi ke dalam beberapa kelompok yang memiliki pola serupa sehingga mempermudah proses analisis.

```
=== DATA SETELAH CLEANING ===
```

	JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI	Unnamed: 2	Unnamed: 4	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
2	2024.0	11.0	1101.0	21.0	10.0	10.0	7.0	4.0	2.0	54.0
3	2024.0	11.0	1101.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	2024.0	11.0	1102.0	11.0	3.0	9.0	13.0	7.0	1.0	44.0
5	2024.0	11.0	1102.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
6	2024.0	11.0	1103.0	49.0	27.0	27.0	35.0	12.0	0.0	150.0

```
Jumlah data: (1027, 10)
```

Gambar 7. Data Setelah *Cleaning*

Berdasarkan hasil data setelah proses *cleaning*, dapat dilihat bahwa dataset yang digunakan dalam penelitian telah melalui tahap pembersihan sehingga siap untuk dianalisis lebih lanjut. Dataset akhir terdiri atas 1.027 baris data dengan 10 atribut, yang memuat informasi jumlah peserta didik mengulang berdasarkan jenjang pendidikan di setiap provinsi.

Tahap cleaning dilakukan untuk memastikan bahwa dataset yang digunakan telah terbebas dari data kosong, duplikasi, maupun atribut yang tidak memiliki keterkaitan dengan kebutuhan penelitian. Proses ini bertujuan untuk meningkatkan kualitas data sehingga hasil analisis yang dihasilkan dapat menjadi lebih tepat dan terpercaya. Setelah tahapan pembersihan selesai dilakukan, data berada dalam kondisi yang lebih rapi, terstruktur, dan siap digunakan pada proses pengelompokan menggunakan algoritma *K-Means*. Dengan kualitas data yang lebih baik, proses *clustering* dapat berjalan secara lebih optimal serta menghasilkan pengelompokan yang lebih sesuai dengan karakteristik data yang dianalisis.

=== HASIL CLUSTERING ===

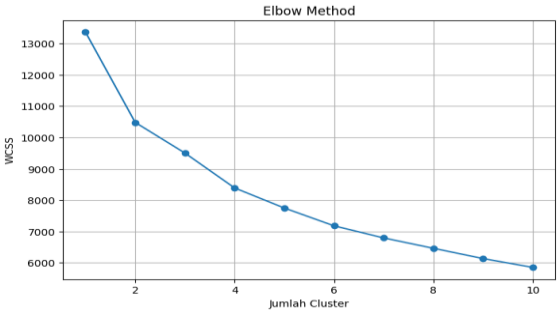
	JUMLAH PESERTA DIDIK MENGULANG TAHUN 2024 MENURUT TINGKAT TIAP PROVINSI	Unnamed: 2	Unnamed: 4	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	Cluster
2	2024.0	11.0	1101.0	21.0	10.0	10.0	7.0	4.0	2.0	54.0	2
3	2024.0	11.0	1101.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2
4	2024.0	11.0	1102.0	11.0	3.0	9.0	13.0	7.0	1.0	44.0	2
5	2024.0	11.0	1102.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	2
6	2024.0	11.0	1103.0	49.0	27.0	27.0	35.0	12.0	0.0	150.0	2

Gambar 8. Data Hasil *Clustering*

Pada hasil proses *clustering*, terlihat bahwa algoritma *K-Means Clustering* berhasil mengelompokkan data ke dalam beberapa kelompok berdasarkan kemiripan karakteristiknya. Hasil dari proses ini ditandai dengan penambahan atribut baru berupa *cluster*, yang menunjukkan kategori kelompok dari setiap data.

Setiap data dikelompokkan ke dalam salah satu dari tiga *cluster*, yaitu *cluster* 0, *cluster* 1, dan *cluster* 2. Pembagian ini didasarkan pada pola nilai atribut yang digunakan dalam analisis, sehingga data yang memiliki karakteristik serupa akan berada pada kelompok yang sama. Dengan adanya label cluster tersebut, proses identifikasi wilayah berdasarkan jumlah peserta didik mengulang menjadi lebih mudah, khususnya dalam membedakan kategori rendah, sedang, dan tinggi.

L. Penentuan Jumlah Cluster dengan Elbow Method



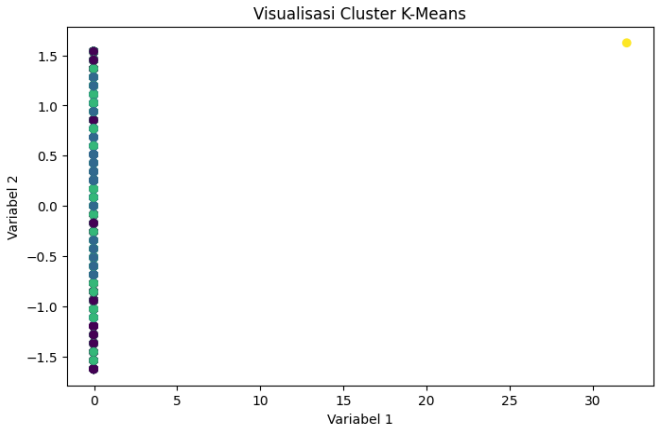
Gambar 9. Grafik Elbow

Penentuan jumlah cluster optimal pada penelitian ini dilakukan menggunakan metode Elbow dengan menghitung nilai *Within Cluster Sum of Squares* (WCSS) pada beberapa jumlah cluster yang berbeda. Metode ini digunakan untuk mengetahui jumlah cluster yang paling sesuai berdasarkan pola penurunan nilai WCSS. Semakin kecil nilai WCSS, maka jarak antar data dengan centroid dalam satu cluster semakin dekat, sehingga kualitas pengelompokan menjadi lebih baik.

Berdasarkan grafik Elbow Method yang ditampilkan, terlihat bahwa penurunan nilai WCSS terjadi cukup signifikan dari jumlah cluster 1 hingga 4. Namun, setelah melewati nilai K=4, penurunan grafik mulai melandai dan tidak menunjukkan perubahan yang terlalu signifikan. Kondisi tersebut menunjukkan adanya titik siku (*elbow point*) pada nilai K=4, sehingga jumlah cluster optimal dalam penelitian ini ditetapkan sebanyak empat cluster.

Pemilihan empat cluster dianggap mampu memberikan keseimbangan antara kualitas pengelompokan dan kompleksitas model. Dengan menggunakan jumlah cluster yang terlalu sedikit, variasi karakteristik data menjadi kurang tergambarkan secara optimal. Sebaliknya, penggunaan cluster yang terlalu banyak dapat menyebabkan proses pengelompokan menjadi terlalu kompleks dan sulit diinterpretasikan. Oleh karena itu, penggunaan K=4 dinilai paling sesuai untuk merepresentasikan pola distribusi data dalam penelitian ini.

M. Hasil Clustering K-Means



Gambar 10. Visualisasi *Clustering K-Means*

Berdasarkan hasil visualisasi *clustering* menggunakan algoritma *K-Means*, data terlihat terbagi ke dalam beberapa kelompok cluster yang dibedakan melalui variasi warna pada grafik. Visualisasi tersebut digunakan untuk menunjukkan pola persebaran data berdasarkan tingkat kemiripan karakteristik antar data setelah melewati proses normalisasi dan pengelompokan. Setiap titik pada grafik merepresentasikan satu data yang telah ditempatkan pada cluster tertentu sesuai

dengan jarak terdekat terhadap *centroid* masing-masing *cluster*.

Berdasarkan tampilan grafik, sebagian besar titik data tampak terkonsentrasi pada area yang saling berdekatan, terutama pada nilai variabel pertama yang relatif rendah. Kondisi ini menunjukkan bahwa mayoritas data memiliki karakteristik yang hampir sama sehingga pemisahan antar cluster belum dapat dilakukan secara optimal. Selain itu, terlihat pula adanya satu titik data yang berada cukup jauh dari kumpulan data lainnya. Hal tersebut mengindikasikan adanya data yang memiliki karakteristik berbeda secara signifikan dibandingkan data mayoritas, sehingga berpotensi dikategorikan sebagai outlier dalam proses clustering.

Hasil visualisasi tersebut sejalan dengan nilai *silhouette coefficient* yang sebelumnya diperoleh sebesar 0,203. Nilai tersebut menunjukkan bahwa kualitas pemisahan antar cluster masih berada pada kategori rendah sehingga batas antar kelompok data belum terbentuk secara jelas. Rendahnya kualitas *cluster* dapat disebabkan oleh beberapa faktor, seperti tingginya korelasi antar variabel, kurangnya keberagaman data, serta distribusi data yang tidak merata. Kondisi tersebut menyebabkan beberapa cluster masih memiliki karakteristik yang saling mirip sehingga proses pengelompokan belum sepenuhnya optimal.

Meskipun demikian, proses clustering yang dilakukan tetap mampu memberikan gambaran awal mengenai pola pengelompokan data berdasarkan karakteristik tertentu. Hasil ini menunjukkan bahwa metode *K-Means* masih dapat digunakan sebagai pendekatan eksploratif untuk mengidentifikasi pola persebaran data pendidikan. Namun, agar hasil *cluster* menjadi lebih representatif dan memiliki kualitas yang lebih baik, diperlukan tahapan preprocessing yang lebih optimal serta pemilihan atribut yang lebih relevan terhadap karakteristik data yang dianalisis.

N. Evaluation

Tahap evaluasi dilakukan untuk mengetahui serta menilai tingkat kualitas model yang dihasilkan pada proses klasifikasi maupun clustering. Pada model klasifikasi, pengukuran performa dilakukan menggunakan metrik akurasi dan menghasilkan nilai sebesar 62,62%. Hasil tersebut menunjukkan bahwa algoritma *Decision Tree* memiliki kemampuan yang sangat baik dalam melakukan klasifikasi data dengan tingkat ketepatan yang tinggi. Meskipun demikian, nilai akurasi yang terlalu tinggi juga perlu diperhatikan karena dapat mengindikasikan kemungkinan terjadinya overfitting pada model. Oleh sebab itu, diperlukan pengujian tambahan seperti confusion matrix dan cross-validation untuk memastikan bahwa model memiliki kemampuan generalisasi yang baik dan tidak hanya optimal pada dataset tertentu.

Di sisi lain, evaluasi terhadap proses clustering dilakukan menggunakan silhouette coefficient untuk mengukur kualitas hasil pengelompokan data. Berdasarkan pengujian yang telah dilakukan, diperoleh nilai silhouette coefficient sebesar 0,203. Nilai tersebut menunjukkan bahwa kualitas cluster yang terbentuk masih tergolong rendah hingga cukup lemah sehingga pemisahan antar *cluster* belum berjalan secara optimal. Kondisi ini menandakan bahwa masih terdapat kemiripan karakteristik data pada beberapa *cluster*, sehingga batas antar kelompok belum dapat dibedakan dengan jelas. Hasil tersebut dapat dipengaruhi oleh beberapa faktor, seperti tingginya hubungan antar variabel, minimnya variasi data, maupun adanya atribut yang memiliki pola distribusi hampir sama.

Walaupun demikian, proses clustering yang dilakukan tetap mampu memberikan gambaran awal mengenai pola pengelompokan data berdasarkan karakteristik tertentu. Hasil evaluasi tersebut menunjukkan bahwa metode *K-Means* masih cukup relevan digunakan dalam tahap eksplorasi data, meskipun kualitas cluster yang dihasilkan masih perlu ditingkatkan. Upaya peningkatan tersebut dapat dilakukan melalui *preprocessing* yang lebih optimal, pemilihan atribut yang lebih representatif, serta pengujian menggunakan metode *clustering* lain yang lebih sesuai dengan karakteristik dataset. Dengan demikian, hasil evaluasi dapat disimpulkan bahwa model yang dibangun telah mampu menjalankan proses analisis data dengan baik, tetapi masih memerlukan pengembangan lanjutan agar performa yang dihasilkan menjadi lebih optimal dan stabil.

O. Pembahasan

Hasil penelitian ini menunjukkan bahwa penerapan pendekatan data mining mampu menghasilkan analisis yang lebih menyeluruh dibandingkan dengan metode konvensional. Pemanfaatan algoritma *Decision Tree* memungkinkan peneliti untuk mengidentifikasi faktor-faktor yang berpengaruh dalam proses klasifikasi data secara lebih terstruktur. Di sisi lain, penggunaan *K-Means Clustering* memberikan gambaran yang lebih jelas terkait distribusi data berdasarkan kesamaan karakteristik tertentu. Kombinasi kedua metode ini menjadi keunggulan utama dalam penelitian, karena mampu mengintegrasikan pendekatan prediktif dan eksploratif dalam satu kerangka analisis yang terpadu. Dengan demikian, hasil yang diperoleh tidak hanya bersifat deskriptif, tetapi juga mampu memberikan insight yang lebih mendalam terhadap pola data yang dianalisis.

Meskipun demikian, terdapat beberapa aspek yang perlu diperhatikan dalam interpretasi hasil penelitian ini. Salah satunya adalah adanya indikasi *overfitting* pada model klasifikasi, yang tercermin dari nilai akurasi yang sangat tinggi. Selain itu, keterbatasan jumlah variabel yang digunakan serta metode validasi yang belum optimal juga berpotensi mempengaruhi kualitas hasil yang diperoleh. Oleh karena itu, penelitian selanjutnya disarankan untuk menggunakan dataset dengan cakupan yang lebih luas, menerapkan teknik validasi yang lebih kuat, serta melakukan komparasi dengan algoritma lain guna meningkatkan performa dan generalisasi model. Langkah-langkah tersebut diharapkan dapat menghasilkan model yang lebih *robust* dan mampu memberikan hasil analisis yang lebih reliabel. Sebagai tahap lanjutan, berikut adalah hasil pembahasan mengenai penelitian ini.

Interpretasi Hasil Klasifikasi

Berdasarkan hasil klasifikasi dengan menggunakan algoritma *Decision Tree*, diketahui bahwa variabel jumlah peserta didik memiliki kontribusi yang signifikan dalam pembentukan model klasifikasi. Temuan ini mengindikasikan bahwa jumlah siswa yang mengulang kelas dapat dijadikan sebagai salah satu indikator dalam menilai kondisi pembelajaran.

Semakin besar jumlah siswa yang mengulang, maka hal tersebut dapat menjadi tanda adanya permasalahan dalam proses

pembelajaran di suatu wilayah. Sebaliknya, jika jumlah siswa yang mengulang relatif rendah, hal ini menunjukkan bahwa proses pembelajaran berlangsung dengan lebih efektif dan baik.
Tingginya kontribusi variabel jumlah peserta didik terhadap hasil klasifikasi menunjukkan bahwa indikator tersebut memiliki keterkaitan yang cukup kuat terhadap kondisi pembelajaran. Dalam konteks pendidikan, jumlah siswa yang mengulang kelas sering dikaitkan dengan efektivitas proses pembelajaran, kualitas pengajaran, serta kesiapan peserta didik dalam memahami materi pembelajaran. Oleh karena itu, hasil klasifikasi ini dapat menjadi dasar awal dalam mengidentifikasi wilayah yang memerlukan evaluasi lebih lanjut terhadap sistem pembelajaran yang diterapkan.

Interpretasi Hasil *Clustering*

Hasil pengelompokan menggunakan metode *K-Means* menunjukkan bahwa data dapat dibagi ke dalam beberapa cluster berdasarkan tingkat kesamaan karakteristik yang dimiliki. Pengelompokan ini secara tidak langsung memberikan gambaran mengenai sebaran kualitas pembelajaran di berbagai wilayah.
Cluster dengan nilai tinggi mengindikasikan wilayah yang memiliki jumlah siswa mengulang relatif besar, sedangkan cluster dengan nilai rendah mencerminkan kondisi pembelajaran yang lebih baik. Informasi ini dapat dimanfaatkan sebagai dasar dalam pengambilan kebijakan di bidang pendidikan, khususnya untuk mengidentifikasi wilayah yang memerlukan perhatian dan penanganan lebih lanjut.

Keterkaitan dengan Konteks Pendidikan

Meskipun penelitian ini menerapkan pendekatan data mining, temuan yang dihasilkan tetap memiliki keterkaitan yang kuat dengan konteks pendidikan. Jumlah siswa yang mengulang kelas dapat dipengaruhi oleh berbagai faktor, seperti metode pembelajaran yang diterapkan, kompetensi guru, serta kondisi lingkungan belajar siswa.
Namun demikian, penelitian ini tidak melakukan pengukuran secara langsung terhadap faktor-faktor tersebut. Oleh karena itu, hasil yang diperoleh lebih bersifat indikatif, yaitu memberikan gambaran atau petunjuk awal, bukan menunjukkan hubungan sebab-akibat (kausal) secara pasti.

Implikasi Penelitian

Hasil penelitian ini memberikan sejumlah implikasi penting. Pertama, penelitian ini mampu menyajikan gambaran kondisi pendidikan berdasarkan data yang aktual dan terukur. Kedua, temuan yang dihasilkan dapat dimanfaatkan sebagai dasar dalam pengambilan keputusan yang lebih objektif dan berbasis
data. Ketiga, hasil penelitian ini juga berpotensi menjadi landasan bagi pengembangan penelitian selanjutnya yang lebih mendalam dan komprehensif.

Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Pertama, data yang digunakan bersifat privat sehingga akses dan kelengkapan informasi menjadi terbatas. Kedua, tidak seluruh variabel yang berkaitan dengan pendidikan dapat dimasukkan ke dalam analisis, sehingga cakupan penelitian belum sepenuhnya komprehensif. Ketiga, hasil yang diperoleh masih bersifat umum, sehingga interpretasinya belum mampu menggambarkan kondisi secara mendalam dan spesifik.

V. KESIMPULAN

Penerapan metode data mining dengan memanfaatkan algoritma *Decision Tree* (C4.5) dan *K-Means Clustering* terhadap data jumlah peserta didik yang mengulang menunjukkan bahwa pendekatan ini mampu mengolah data pendidikan secara sistematis dan menghasilkan informasi yang lebih bermakna. Setiap tahapan yang dilalui, mulai dari *preprocessing*, transformasi data, hingga proses pemodelan, memiliki kontribusi penting dalam menghasilkan model yang tidak hanya memiliki tingkat akurasi yang baik, tetapi juga mudah untuk dipahami dan diinterpretasikan.

Pada proses analisis klasifikasi, algoritma *Decision Tree* menunjukkan performa baik dalam mengelompokkan data sesuai karakteristik. Variabel dalam dataset memiliki pola untuk pengambilan keputusan. Namun, terdapat potensi *overfitting* yang perlu diperhatikan, sehingga perlu pengujian lebih lanjut untuk memastikan model tetap akurat pada data baru.

Sementara itu, hasil pengelompokan menggunakan metode *K-Means* menunjukkan bahwa data dapat terbagi ke dalam beberapa cluster berdasarkan tingkat kesamaan karakteristik. Pembagian ini memberikan gambaran yang lebih jelas mengenai distribusi jumlah peserta didik yang mengulang di berbagai wilayah, sehingga dapat membantu dalam mengidentifikasi kondisi wilayah secara lebih spesifik.

Di sisi lain, analisis korelasi memperlihatkan adanya hubungan yang sangat kuat antar variabel, yang berpotensi menimbulkan redundansi dalam data. Temuan ini menegaskan pentingnya proses preprocessing, khususnya dalam melakukan seleksi fitur, agar dapat menghindari masalah multikolinearitas yang dapat mempengaruhi kinerja model.

Secara umum, jumlah peserta didik yang mengulang dapat digunakan sebagai indikator tidak langsung untuk menggambarkan kondisi pembelajaran di suatu wilayah. Wilayah dengan tingkat pengulangan yang tinggi cenderung memerlukan perhatian lebih dalam upaya peningkatan kualitas pembelajaran. Selain itu, pendekatan berbasis data mining juga menunjukkan potensi yang besar sebagai alat bantu dalam pengambilan keputusan di bidang pendidikan, terutama dalam mengolah data dalam skala besar serta mengungkap pola-pola yang tidak terlihat secara langsung.

DAFTAR PUSTAKA

[1] B. Arya, A. Shidik, Y. T. Apriliyanto, A. Salsabilla, and E. S. Aulia, "Peningkatan literasi siswa sdn 2 kaligelang melalui pojok baca dan bimbingan belajar," vol. 6, pp. 54–65, 2025. https://doi.org/10.38048/jailcb.v6i1.4806

[2] A. M. Salsabilla, I. N. Cahaya, T. Latipah, and O. Farhurohman, "Meningkatkan Kualitas Pembelajaran Di Sekolah Dasar Melalui Metode Aktif Dan Inovatif Improving," vol. 5,pp. 1603–1618, 2025. https://doi.org/10.58578/arzusin.v5i4.6028

[3] U. Hasanah, S. Masitoh, Z. K. Dealova, M. Yunus, G. R. Frimananda, and M. P. Interaktif, "Faktor Penunjang Keberhasilan Dalam Proses Pembelajaran Siswa Sekolah Dasar," vol. 8,pp. 1184–1188, 2025. https://doi.org/10.31004/jrpp.v8i1.41516

[4] Muh. Asharif Suleman and Zulfi Idayanti, "Faktor-Faktor Yang Mempengaruhi Keberhasilan Proses Pembelajaran di SD/MI Muh. Asharif Suleman 1 , Zulfi Idayanti 2 1,2 Universitas Islam Negeri Sunan Kalijaga Yogyakarta, Indonesia," vol. 2, no. 3, pp. 235–244, 2024. https://doi.org/10.59689/ment.v2i3.1511

[5] A. P. Riani, A. Voutama, and T. Ridwan, "Penerapan K-Means Clustering Dalam Pengelompokan Hasil Belajar Peserta Didik Dengan Metode Elbow," vol. 6, pp. 164–172, 2023. https://doi.org/10.53513/jsk.v6i1.7351

[6] Y. M. Gultom, F. Syahputra, and S. Syahrial, "Pengaruh Evaluasi Pembelajaran terhadap Pembelajaran Guru di Sekolah Dasar Kualitas," no. 3, pp. 1–8, 2024. https://doi.org/10.47134/pgsd.v1i3.543

[7] B. Ardiansyah, I. Daulay, and R. Hutagaol, "K-Means and Decision Tree Algorithm for Prediction of Postgraduate Students Admission in University of Indonesia Algoritma K-Means dan Decision Tree untuk Prediksi Penerimaan Calon Mahasiswa Pascasarjana pada Universitas Indonesia," pp. 154–161, 2022.

[8] Idrus and D. W. Sari, "Penerapan Data Mining Menggunakan Algoritma Decision Tree C4.5 Untuk Memprediksi Mahasiswa Drop Out Di Universitas Wiraraja," vol. 1, no. Juni,pp. 1–7, 2023. https://doi.org/10.24929/jars.v1i02.2684

[9] K. Istiqomah and V. Sofica, "Penerapan Data Mining Menggunakan Algoritma Decision Tree Untuk Menganalisis Penggunaan Media Sosial Dengan Konsentrasi Belajar Mahasiswa," vol. 4, no. 4, pp. 53–67, 2025. https://doi.org/10.31004/riggs.v4i4.3228

[10] Asmana, Y. A. Wijaya, and Martanto, "K-Means di Sekolah Menengah Kejuruan Wahidin Kota Cirebon," vol. 6, no. 2, pp. 552–559, 2022. https://doi.org/10.36040/jati.v6i2.5236

[11] R. Hidayat, "Pemanfaatan Data Mining untuk Melihat Minat Siswa Setelah Menyelesaikan Pendidikan Sekolah Menengah Atas (SMA) dengan Algoritma K-Means Clustering," vol. 1, no. 2, 2022. https://doi.org/10.32639/tiij.v1i2.220

[12] Farokha and S. Pradikto, "Analisis Peran Lingkungan Keluarga dan Teman Pergaulan dalam Membangun Motivasi Belajar Siswa SMA," vol. 2, no. 1, pp. 1–7, 2025. https://doi.org/10.59923/jiim.v2i1.331

[13] Y. B. Utomo, I. Kurniasari, and I. Yanuartanti, "Penerapan Knowledge Discovery In Database," vol. 7, no. 1, 2023. https://doi.org/10.59697/jtik.v7i1.61

[14] N. Afiasari, N. Suarna, and N. Rahaningsih, "Implementasi Data Mining Transaksi Penjualan Menggunakan Algoritma Clustering dengan Metode K-Means E-commerce K-Means melakukan analisis penerapan Data Mining dalam mengelompokkan jumlah," vol. 9, pp. 100–110, 2023. https://doi.org/10.33020/saintekom.v13i1.402

[15] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning Data Text Pre-Processing Sentiment Analysis In Data Mining Using Machine LearninG School of Computer Science and Technology, Harbin Institute of Technology," vol. 4, no. 2, pp. 16–22, 2021. https://doi.org/10.30813/jbase.v4i2.3000

[16] S. V. Natasya and R. M. Awangga, "Profiling Mahasiswa Dan Alumni Menggunakan Metode Decision Tree Systematic Literature Review," vol. 7, no. 2, pp. 1359–1363, 2023.

[17] U. Suriani, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma," vol. 3, no. 2, pp. 55–66, 2023. https://doi.org/10.36040/jati.v7i2.6824

[18] E. A. Firdaus, S. Maulani, and A. B. Dharmawan, "Pengukuran Minat Baca Mahasiswa Dengan Metode Clustering Di Perpustakaan Akademi Keperawatan Rs.Dustira Cimahi Menggunakan Data Mining," vol. 15, pp. 32–40, 2021. https://doi.org/10.25134/nuansa.v15i1.3856