BAB II

LANDASAN TEORI

2.1 Teori dan Konsep yang Mendasari Penelitian

Pendidikan modern semakin bergantung pada pengambilan keputusan berbasis data (*data-driven decision making*) untuk meningkatkan prestasi belajar siswa dan kualitas pembelajaran. Integrasi antara kerangka teoretis dengan analisis data memberikan landasan ilmiah dalam mengidentifikasi pola pembelajaran dan memperbaiki strategi pengajaran. Sintesis ini mengkaji empat pertanyaan utama: (1) kerangka teoretis apa yang menjelaskan hubungan antara prestasi belajar dan model pembelajaran berbasis data; (2) bagaimana *data mining* didefinisikan dan diterapkan dalam konteks pendidikan; (3) faktor-faktor utama apa yang memengaruhi kinerja akademik di pendidikan dasar berdasarkan penelitian terkini; dan (4) bagaimana analisis data pendidikan dapat mendukung pembelajaran yang dipersonalisasi dan adaptif. Bagian ini menjelaskan landasan ilmiah yang menghubungkan prestasi belajar dengan proses pengukuran berbasis data, sekaligus memberikan dasar konseptual bagi penerapanalgoritma *K-Means* dalam pengelompokan nilai siswa.

2.1.1. Kerangka Teoretis Prestasi Belajar dan Pembelajaran Berbasis Data

Hubungan antara prestasi belajar dan model pembelajaran berbasis data dapat dipahami melalui sejumlah teori dasar. Teori psikometrik menyediakan dasar pengukuran untuk menganalisis nilai prestasi dengan menganggapnya sebagai indikator dari konstruk belajar laten. Menurut (Meyer and Reynolds 2022),

Pendekatan sosial-kognitif menyoroti aspek motivasi dan efikasi diri yang memengaruhi nilai siswa, sehingga variasi prestasi yang muncul dapat dipertimbangkan sebagai dasar pengelompokan dalam analisis K-Means.untuk memahami hasil belajar dengan lebih baik(Martín et al. 2024). Perspektif sosial-ekologis menempatkan prestasi siswa dalam konteks system pendidikan, yang mendukung interpretasi klaster dengan mempertimbangkan lingkungan belajar.dengan menekankan pengaruh perilaku guru, hubungan teman sebaya, dan faktor sistemik terhadap variasi prestasi(Hernández-Ortiz, Precht, and Cudina 2021). Semua kerangka ini menunjukkan bahwa analisis data pendidikan harus mempertimbangkan dimensi individual dan kontekstual untuk mencapai interpretasi yang adil dan dapat dipertanggung jawabkan.

Dari sisi komputasional, teori pembelajaran adaptif menjelaskan bagaimana algoritma menyesuaikan instruksi secara waktu nyata berdasarkan data siswa. (Lin et al. 2024) mengembangkan sistem adaptif berbasis *knowledge graph* yang memetakan hubungan kognitif untuk mempersonalisasi lintasan belajar.(Taufik Gusman and Muhammad Umar Huzein 2022) memperkenalkan *Adaptive Cognitive Enhancement Model (ACEM)* yang menghubungkan profil kognitif dengan intervensi pembelajaran yang dipersonalisasi. Pendekatan ini menegaskan bahwa sistem berbasis data yang efektif harus memadukan ketepatan psikometrik, konteks motivasional, dan adaptivitas algoritmik untuk mengoptimalkan hasil belajar.

Pemahaman terhadap landasan teoretis ini memastikan bahwa penggunaan *data mining* dalam pendidikan menghormati prinsip validitas pengukuran serta kompleksitas proses belajar manusia. Hal ini memberikan dasar bagi integrasi teori

psikometrik dan sosial-kognitif dengan kerangka algoritmik untuk menafsirkan serta meningkatkan prestasi siswa.

2.1.2 Definisi dan Penerapan Data Mining Dalam Dunia Pendidikan

Data mining dalam pendidikan atau *Educational Data Mining (EDM)* merupakan proses identifikasi pola pembelajaran dari data akademik, termasuk nilai siswa, untuk mendukung analisis dan pengambilan keputusan berbasis data.(Lin et al. 2024) menjelaskan bahwa EDM mencakup analisis prediktif, deskriptif, dan preskriptif yang digunakan untuk mengidentifikasi tren, memprediksi risiko, serta merekomendasikan intervensi. Dalam praktiknya,Metode klaster seperti K-Means digunakan untuk mengelompokkan siswa berdasarkan kesamaan pola nilai, sehingga setiap klaster dapat mewakili tingkat prestasi belajar tertentu.(Amalia et al. 2021). Hasil pengelompokan ini membantu pendidik merancang program dukungan yang lebih tepat sasaran.

Penerapan EDM meliputi sistem peringatan dini bagi siswa berisiko, optimalisasi desain kurikulum, serta analisis perilaku belajar untuk memperbaiki strategi pengajaran (Perveen et al. 2025). Alat visualisasi dan dasbor interaktif memudahkan guru memahami hasil analisis, sementara umpan balik berbasis data membantu pengambilan keputusan instruksional(Li et al. 2025). Selain itu, aspek etika seperti kualitas data, interpretabilitas, dan keterlibatan pemangku kepentingan sangat penting untuk menjaga transparansi (Talib, Majid, and Sahran 2023)

Definisi dan penerapan EDM dalam pendidikan memastikan adanya pendekatan sistematis dalam interpretasi data belajar, menjadi dasar pengembangan sistem pengelompokan, prediksi, dan rekomendasi yang memperkuat pengambilan keputusan berbasis bukti.

2.1.3 Faktor Utama yang Mempengaruhi Prestasi Akademik di Pendidikan Dasar

Penelitian empiris menunjukkan bahwa berbagai faktor memengaruhi prestasi akademik di tingkat dasar. Variabel kelas dan guru, seperti kualitas pengajaran, kecukupan sarana, dan keterlibatan guru, memiliki korelasi positif dengan hasil belajar (Hernández-Ortiz, Precht, and Cudina 2021). Latar belakang sosial ekonomi serta dukungan orang tua juga sangat berperan, di mana ketimpangan ekonomi berpengaruh langsung terhadap akses pembelajara(Huang, Wang, and Lubis 2023). Selain itu, faktor psikososial seperti resiliensi, efikasi diri, dan motivasi terbukti memberikan kontribusi signifikan terhadap variasi prestasi antar siswa (Giménez et al. 2024)

Integrasi teknologi pembelajaran menghasilkan data akademik yang lebih lengkap sehingga mendukung analisis prestasi melalui pendekatan klasterisasi.(Liu 2021). Faktor-faktor ini menegaskan bahwa keberhasilan belajar bergantung pada interaksi kompleks antara dimensi pedagogis, sosial, teknologi, dan pengukuran.

Justifikasi Pemahaman terhadap determinan prestasi belajar membantu perancang kebijakan dan analis pendidikan membangun model komprehensif yang mempertimbangkan berbagai dimensi kontekstual (Nasyuha, Zulham, and Rusydi 2022)

2.1.4. Analisis Data Pendidikan untuk Pembelajaran Adaptif dan Personal

Analisis data pendidikan memungkinkan pembelajaran yang dipersonalisasi melalui sistem adaptif yang menyesuaikan instruksi secara dinamis. menjelaskan bahwa algoritma adaptif berbasis *knowledge graph* menyesuaikan konten dengan struktur kognitif siswa, sedangkan (Susanti, Triyana, and Nurwiyeni 2023) menekankan model berbasis AI yang mempersonalisasi lintasan belajar melalui evaluasi berkelanjutan. Pendekatan ini meningkatkan keterlibatan dan efisiensi belajar dengan menyesuaikan strategi pengajaran terhadap kebutuhan individu.

(Du 2022)menambahkan bahwa sistem pengambilan keputusan berbasis data dapat mengoptimalkan desain kurikulum serta distribusi sumber daya pendidikan. Kombinasi otomatisasi dan peran guru sebagai pengambil keputusan akhir menciptakan keseimbangan antara presisi algoritma dan konteks pedagogis (Bahtiar et al. 2021). Selain itu, (Obaid and Alabbas 2024) menyoroti pentingnya Visualisasi hasil klaster *K-Means* memungkinkan guru menilai distribusi prestasi siswa dan menetapkan strategi pembelajaran yang lebih personal.

2.2 *Knowledge Discovery In Database* (KDD)

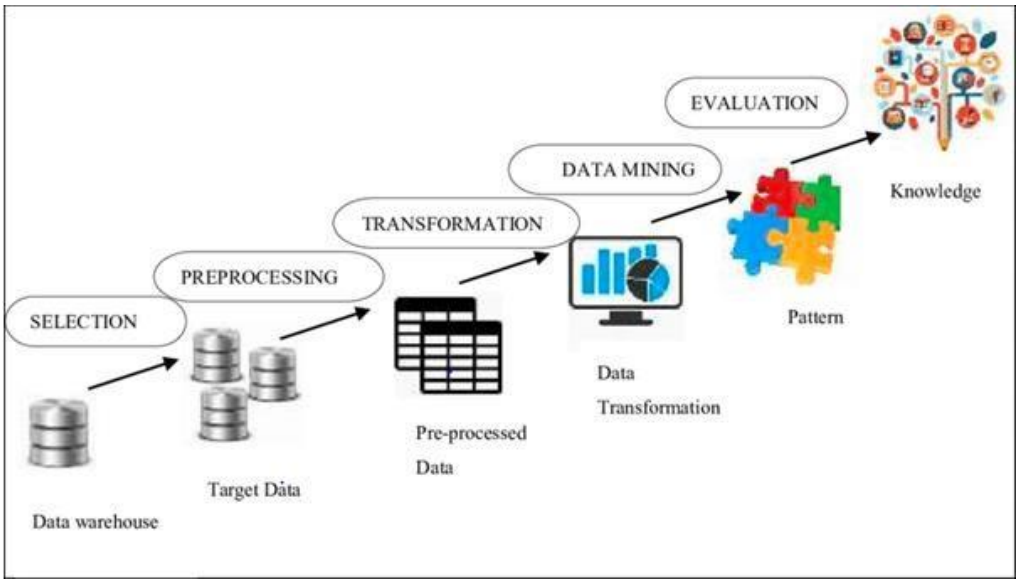
Knowledge Discovery in Database (KDD) merupakan suatu proses yang melibatkan pengumpulan dan pemanfaatan data historis untuk mengidentifikasi pola, keteraturan, serta hubungan dalam kumpulan data yang berukuran besar. KDD memiliki hubungan yang erat dengan data mining karena berfokus pada penggalian informasi tersembunyi dalam basis data. Secara umum, KDD dapat diartikan sebagai metode untuk mengekstraksi pengetahuan dari data dalam skala besar guna

menemukan pola-pola baru yang dapat menghasilkan informasi dan wawasan yang bermanfaat.(Ngueajio, Washington, and Rawat n.d.)

Lebih lanjut, KDD merupakan suatu bidang keilmuan yang bersifat multidisipliner karena mengintegrasikan berbagai disiplin seperti statistik, basis data, kecerdasan buatan (artificial intelligence), visualisasi data, serta komputasi paralel. Melalui proses ini, dapat diidentifikasi pola maupun kecenderungan yang relevan dari data, yang kemudian diolah menjadi informasi yang akurat, tepat, dan mudah dipahami oleh pengguna.(Alam, Misba, and Atulasimha n.d.)

Knowledge Discovery in Database (KDD) merupakan suatu proses yang bertujuan untuk menemukan serta mengidentifikasi pola (pattern) yang terdapat dalam suatu basis data. Dalam pelaksanaannya, KDD terdiri dari beberapa tahapan yang saling Knowledge Discovery in Database (KDD) merupakan suatu proses yang bertujuan untuk menemukan serta mengidentifikasi pola (pattern) yang terdapat dalam suatu basis data. Dalam pelaksanaannya, KDD terdiri dari beberapa tahapan yang saling berkaitan untuk menghasilkan pengetahuan yang bermakna

dari data yang dianalisis.



2.3 Algoritma yang Digunakan: K-Means Clustering

2.3.1 Prinsip Matematis K-Means Clustering

K-Means adalah algoritma klaster yang bekerja dengan meminimalkan nilai withinclustersum of squares (WCSS) untuk mendapatkan pemisahan kelompok yang optimal pada data nilai siswa. Misalkan terdapat himpunan data $X = \{x_1, x_2, \dots, x_n\} \subset R^d$dan jumlah klaster yang diinginkan adalah k Tujuan algoritma ini adalah mencari partisi $C = \{C_1, C_2, \dots, C_k\}$dan centroid $\mu_1, \dots, \mu_k$fungsi objektif berikut diminimalkan: $J(C, \mu) = \sum_{j=1}^{k} \sum_{x \in C_j} |x - \mu_j|^2$Persamaan ini dikenal sebagai *Within Cluster Sum of Squares* (WCSS) atau *inertia*, yang merepresentasikan total variasi data dalam setiap klaster Proses optimasi dilakukan melalui pendekatan *alternating minimization* atau *Lloyd's algorithm*, yang terdiri dari dua langkah utama: (1) *assignment step*, di mana setiap titik data dikaitkan dengan centroid terdekat menggunakan jarak Euclidean, dan (2) *update step*, di mana setiap centroid dihitung ulang sebagai rata-rata aritmetika titik-titik yang

menjadi anggotanya (Latifah, Surono, and Suparman 2022). Proses ini diulang hingga konvergen, yaitu ketika tidak ada lagi perubahan signifikan pada posisi centroid.

Dari sudut pandang geometris, *K-Means* mengasumsikan bahwa setiap klaster berbentuk sferis dan memiliki varians yang seragam. Karena *K-Means* menghitung jarak *Euclidean*, maka fitur berskala besar dapat mendominasi perhitungan sehingga normalisasi diperlukan untuk menjaga keseimbangan kontribusi setiap variabel.(Ginting, Efendi, and Suwilo 2022) Oleh karena itu, dalam penerapan pada data pendidikan, penting untuk melakukan normalisasi fitur agar hasil pengelompokan tidak dipengaruhi oleh skala atribut yang berbeda.

2.3.2 Mekanisme Pengelompokan Berdasarkan Ukuran Kemiripan

Proses pengelompokan dalam *K-Means* berlandaskan pada konsep kemiripan *(similarity)*, di mana ukuran yang digunakan adalah jarak Euclidean. Setiap titik data akan ditempatkan pada klaster dengan centroid yang memiliki jarak paling kecil. Proses ini membentuk *Voronoi partition* dalam ruang fitur, di mana setiap wilayah menunjukkan area pengaruh satu centroid tertentu (Dai, Wong, and Wong 2024) Penetapan klaster ditentukan dengan menghitung jarak Euclidean setiap data kesemua centroid, kemudian menetapkannya pada centroid dengan jarak paling kecil.(Daoudi et al. 2021)

Namun, hasil akhir algoritma sangat dipengaruhi oleh nilai awal centroid dan jumlah klaster kkk yang ditentukan. Untuk mengurangi bias akibat inisialisasi yang tidak tepat, berbagai strategi seperti *K-Means++* dan metode berbasis metaheuristik telah dikembangkan. Strategi *K-Means++* memilih centroid awal

secara probabilistik berdasarkan jarak antar titik, sehingga dapat mempercepat konvergensi dan meningkatkan kualitas hasil pengelompokan (Obaid and Alabbas 2024)

2.3.3 Keunggulan dan Keterbatasan dalam Analisis Data Pendidikan

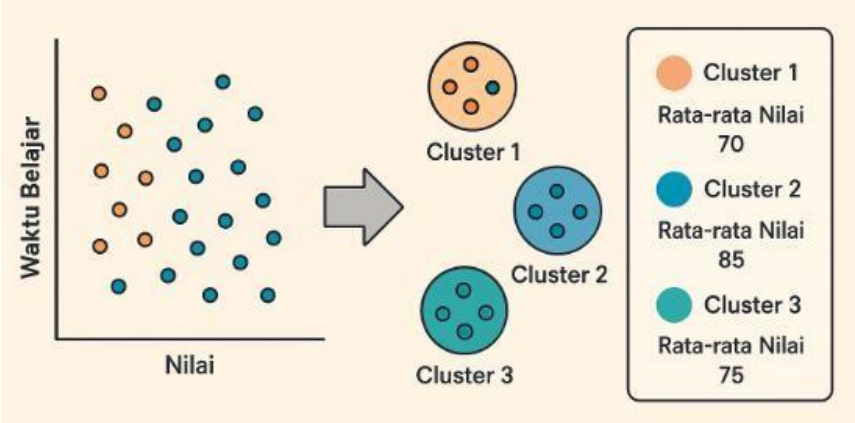
Dalam konteks pendidikan, *K-Means* banyak digunakan karena sifatnya yang sederhana, efisien, dan mudah diinterpretasikan. Algoritma ini memungkinkan institusi pendidikan untuk mengelompokkan siswa berdasarkan kesamaan nilai, tingkat kehadiran, atau perilaku belajar. *K-Means* unggul untuk analisis nilai siswa karena cepat, mudah diinterpretasikan, serta mampu memetakan variasi prestasi dengan efisien. serta kemudahan interpretasi hasil karena setiap klaster dapat dijelaskan melalui nilai rata-rata (centroid) fitur-fitur utama siswa (Amalia et al. 2021)

Meski demikian, algoritma ini memiliki beberapa keterbatasan. Pertama, *K-Means* mensyaratkan jumlah klaster k ditentukan di awal, yang bisa memengaruhi akurasi jika nilai kkk tidak sesuai dengan struktur alami data. Kedua, algoritma sensitif terhadap outlier dan skala atribut, sehingga data perlu dinormalisasi agar hasil tidak bias. Ketiga, karena bergantung pada jarak Euclidean, algoritma ini kurang efektif untuk data kategorikal atau atribut non-linier yang sering ditemukan pada catatan pendidikan (Zainuddin and Risal 2024) Dalam menghadapi tantangan ini, sejumlah penelitian mengusulkan varian seperti *K-Medoids* dan *Constrained K-Means* untuk memperbaiki stabilitas dan keadilan hasil pengelompokan.

2.3.4 Implementasi *K-Means* pada Studi Prestasi Belajar

Berbagai penelitian membuktikan bahwa K-Means efektif dalam memisahkan siswa ke dalam kelompok prestasi, sehingga hasil klaster dapat digunakan sebagai dasar intervensi pembelajaran.. menggunakan algoritma ini untuk menganalisis nilai akademik dan mengembangkan *dashboard* visualisasi performa siswa.menerapkan *K-Means* untuk membentuk kelas unggulan dengan tujuan meningkatkan (Syahputra and Hutagalung 2022) pencapaian akademik. memanfaatkan algoritma ini dalam pengelompokan penerima bantuan Program Indonesia Pintar (PIP) untuk memastikan pemerataan manfaat pendidikan dasar. Studi lain oleh (Amalia et al. 2021) menggunakan *K-Means* untuk pengelompokan mahasiswa penerima beasiswa, menghasilkan model prioritas berdasarkan profil nilai dan keaktifan akademik.

Implementasi algoritma ini juga dikombinasikan dengan metode *Support Vector Machine* (SVM) untuk klasifikasi lanjutan (Talib, Majid, and Sahran 2023) Pendekatan hibrida tersebut memperlihatkan bahwa hasil klaster *K-Means* dapat dijadikan masukan (input features) bagi algoritma pembelajaran terawasi dalam mengidentifikasi pola performa siswa secara prediktif. Dalam konteks ini, *K-Means* berperan sebagai tahap eksploratif yang memperkaya analisis deskriptif maupun inferensial di bidang *educational data mining.*



Gambar 2. 1 Implementasi K-Means pada Studi Prestasi Belajar

2.3.6 Justifikasi Pembahasan

Pembahasan mengenai algoritma *K-Means* menjadi inti penelitian ini karena metode tersebut merupakan pendekatan utama dalam menentukan kelompok prestasi belajar siswa. Pemahaman terhadap teori matematis, mekanisme kerja, serta penerapan algoritma ini memberikan dasar ilmiah yang kuat bagi analisis data akademik. Integrasi metode ini dalam sistem evaluasi pendidikan memungkinkan pengambilan keputusan yang lebih objektif dan berbasis data, khususnya dalam mengidentifikasi kebutuhan intervensi bagi kelompok siswa tertentu. Dengan demikian, landasan teoritis ini berfungsi tidak hanya sebagai acuan metodologis, tetapi juga sebagai pijakan konseptual dalam mengembangkan model pengelompokan yang adaptif dan berkeadilan(Mcgregor et al. n.d.).

2.4 Langkah-Langkah dalam Machine Learning

2.4.1 Tahapan Umum *Machine Learning* untuk Clustering

Tahapan analisis mencakup pengumpulan nilai siswa, preprocessing data numerik, penerapan algoritma K-Means, evaluasi klaster, dan interpretasi hasil untuk kebutuhan pembelajaran. Pada tahap pertama, data dikumpulkan dari

berbagai sumber seperti catatan akademik, kehadiran, atau hasil ujian siswa. (Ariska 2021) keberhasilan analisis data pendidikan sangat bergantung pada kualitas data mentah dan keberagaman atribut yang digunakan untuk menggambarkan perilaku belajar siswa. Data yang baik tidak hanya mencakup nilai akademik, tetapi juga indikator non-kognitif seperti motivasi dan partisipasi kelas(Shutaywi 2021).

Tahap berikutnya adalah *data preprocessing*, di mana data disiapkan agar dapat digunakan secara efektif dalam model *machine learning*. Proses ini meliputi pembersihan data dari nilai yang hilang (*missing values*), penghapusan duplikasi, serta transformasi atribut menjadi format numerik yang sesuai. menekankan bahwa tahapan *preprocessing* yang sistematis dapat meningkatkan efisiensi dan stabilitas algoritma *K-Means*, terutama karena algoritma ini sensitif terhadap skala dan rentang nilai variabel. Setelah data siap, tahap pemilihan model dilakukan dengan mempertimbangkan karakteristik data dan tujuan penelitian. Dalam konteks pengelompokan nilai siswa**,** *K-Means clustering* menjadi pilihan populer karena kemampuannya dalam mengidentifikasi pola dan kelompok dengan kesamaan perilaku belajar (Nasyuha, Zulham, and Rusydi 2022)

Tahap pelatihan dan evaluasi dilakukan dengan mengoptimalkan parameter algoritma untuk memperoleh pembagian kelompok yang paling representatif. Evaluasi hasil *clustering* sering menggunakan metrik seperti *Silhouette Score* dan *Elbow Method* untuk menentukan jumlah klaster terbaik yang merepresentasikan variasi dalam data. Seperti yang diungkapkan oleh (Lavandaia Dharma Bali PELATIHAN JARINGAN DAN TROUBLESHOOTING KOMPUTER et al.

2022) pendekatan berbasis metrik evaluasi sangat penting dalam konteks *unsupervised learning*, karena tidak adanya label kebenaran (ground truth) yang bisa digunakan untuk validasi langsung. Tahap terakhir adalah interpretasi hasil, di mana tiap kelompok atau klaster diidentifikasi berdasarkan ciri-ciri khasnya untuk menghasilkan informasi yang dapat digunakan oleh pendidik atau pembuat kebijakan.

2.4.2 Perbedaan *Unsupervised Learning* dan *Supervised Learning*

Perbedaan mendasar antara *unsupervised learning* dan *supervised learning* terletak pada keberadaan label dalam data latih. *Supervised learning* menggunakan data yang telah diberi label untuk memprediksi hasil tertentu, seperti klasifikasi nilai atau kelulusan siswa. Sebaliknya, *unsupervised learning* berfokus pada menemukan struktur atau pola tersembunyi tanpa menggunakan label (Chong 2021) Dalam konteks pendidikan, *unsupervised learning* digunakan untuk mengelompokkan siswa berdasarkan kesamaan karakteristik belajar tanpa mengetahui kategori sebelumnya, seperti kelompok siswa berprestasi tinggi, sedang, dan rendah.(Pitafi, Anwar, and Sharif 2023)

Menurut(Wang 2023), keunggulan utama *unsupervised learning* adalah kemampuannya untuk mengekstraksi pengetahuan baru dari data besar yang belum terlabel, memungkinkan analisis eksploratif terhadap perilaku dan pola pembelajaran siswa. Namun, pendekatan ini juga memiliki tantangan, seperti kesulitan dalam mengevaluasi hasil *clustering* dan risiko interpretasi yang subjektif. Dalam praktiknya, banyak penelitian menggabungkan *unsupervised* dan *supervised learning* dalam pendekatan hibrida, Hasil klaster nilai siswa dapat dijadikan

variabel tambahan untuk analisis prediktif seperti perkiraan risiko penurunan prestasi.(Obaid and Alabbas 2024)Pendekatan ini membantu meningkatkan akurasi analisis dan menghasilkan wawasan yang lebih aplikatif di bidang pendidikan.

2.4.3 Teknik *Preprocessing* untuk Dataset Pendidikan

Tahap *preprocessing* merupakan aspek kritis dalam setiap proyek *machine learning* karena memengaruhi stabilitas dan kinerja algoritma yang digunakan. Dalam konteks data pendidikan, *Preprocessing* diperlukan untuk memastikan nilai antar mata pelajaran memiliki skala yang seragam sehingga klasterisasi menghasilkan pola yang representatif. (Lin et al. 2024)menyatakan bahwa data pendidikan sering kali bersifat heterogen, mencakup nilai numerik, teks, dan data waktu. Oleh karena itu, teknik seperti *feature scaling* dan *one-hot encoding* sangat diperlukan untuk mengonversi data ke format numerik yang seragam(Remondino et al. 2017).

Selain itu, seleksi fitur yang relevan menjadi langkah penting untuk menghindari overfitting dan meningkatkan interpretabilitas hasil menegaskan bahwa fitur seperti nilai rata-rata per mata pelajaran, tingkat kehadiran, dan waktu belajar efektif merupakan prediktor kuat dalam menentukan prestasi belajar siswa.(Rouillard et al. 2016) Dengan demikian, peneliti harus melakukan *feature engineering* secara hati-hati agar variabel yang dipilih benar-benar mewakili karakteristik akademik dan perilaku belajar siswa. (Ginting, Efendi, and Suwilo 2022) juga menambahkan bahwa penghapusan outlier atau data ekstrem sangat penting, karena nilai-nilai ekstrem dapat menggeser pusat klaster (*centroid*) dan menghasilkan hasil pengelompokan yang bias.

Lebih lanjut, (Daoudi et al. 2021)menguraikan bahwa pemilihan teknik *preprocessing* yang tepat tidak hanya meningkatkan efisiensi algoritma tetapi juga membantu mengatasi keterbatasan algoritma *K-Means* yang sensitif terhadap skala dan distribusi data. Sebagai contoh, penggunaan *Principal Component Analysis (PCA)* dapat membantu mengurangi dimensi data dan mengeliminasi korelasi antar variabel, sehingga meningkatkan akurasi pengelompokan. Kombinasi antara *feature selection* dan *dimensionality reduction* menjadi strategi umum dalam penelitian pendidikan berbasis *machine learning*(Pastorello 2020).

2.4.4 Dampak Normalisasi Data terhadap Proses Clustering

Normalisasi data merupakan langkah penting dalam proses *clustering* menggunakan algoritma *K-Means* Normalisasi sangat penting karena K-Means hanya mengandalkan jarak; tanpa normalisasi, mata pelajaran dengan rentang nilai lebih besar akan mendominasi proses klasterisasi.. Menurut(Kurniawan et al. 2024), tanpa normalisasi, fitur dengan nilai numerik besar akan mendominasi perhitungan jarak Euclidean, sehingga mengarah pada hasil klaster yang tidak representatif. Dalam penelitian yang melibatkan nilai siswa, perbedaan skala antara nilai ujian, jumlah kehadiran, dan waktu belajar dapat menyebabkan bias terhadap atribut dengan rentang nilai lebih besar.

(Rezki, Ihsan, and Ambarsari 2021)menekankan bahwa penggunaan teknik normalisasi *seperti Min-Max Scaling* dan *Z-score Standardization* sangat direkomendasikan untuk meningkatkan performa *K-Means*. Dengan normalisasi, setiap variabel diperlakukan secara proporsional dalam proses perhitungan jarak, menghasilkan pembagian klaster yang lebih akurat. (Feng, Fan, and Ao 2022) juga

menemukan bahwa penerapan *data normalization* sebelum pengelompokan dapat mempercepat konvergensi algoritma dan mengurangi risiko jebakan pada solusi lokal. Selain itu, normalisasi membantu meningkatkan konsistensi hasil ketika algoritma dijalankan dengan inisialisasi yang berbeda.(Tes et al. n.d.)

Dalam konteks pendidikan, normalisasi data berperan penting dalam menjaga keadilan analisis, terutama ketika hasil *clustering* digunakan untuk pengambilan keputusan administratif seperti penentuan kelompok remedial atau pemberian beasiswa. Hasil studi (Fang 2023)menunjukkan bahwa kesalahan dalam tahap normalisasi dapat menimbulkan bias terhadap kelompok tertentu, sehingga penting untuk memastikan bahwa proses ini dilakukan secara sistematis dan transparan. Dengan demikian, normalisasi tidak hanya berfungsi sebagai langkah teknis tetapi juga memiliki implikasi etis dan kebijakan dalam penerapan *machine learning* di bidang Pendidikan(Ridho et al. 2023)

2.5 Teknik Evaluasi Algoritma di Machine Learning

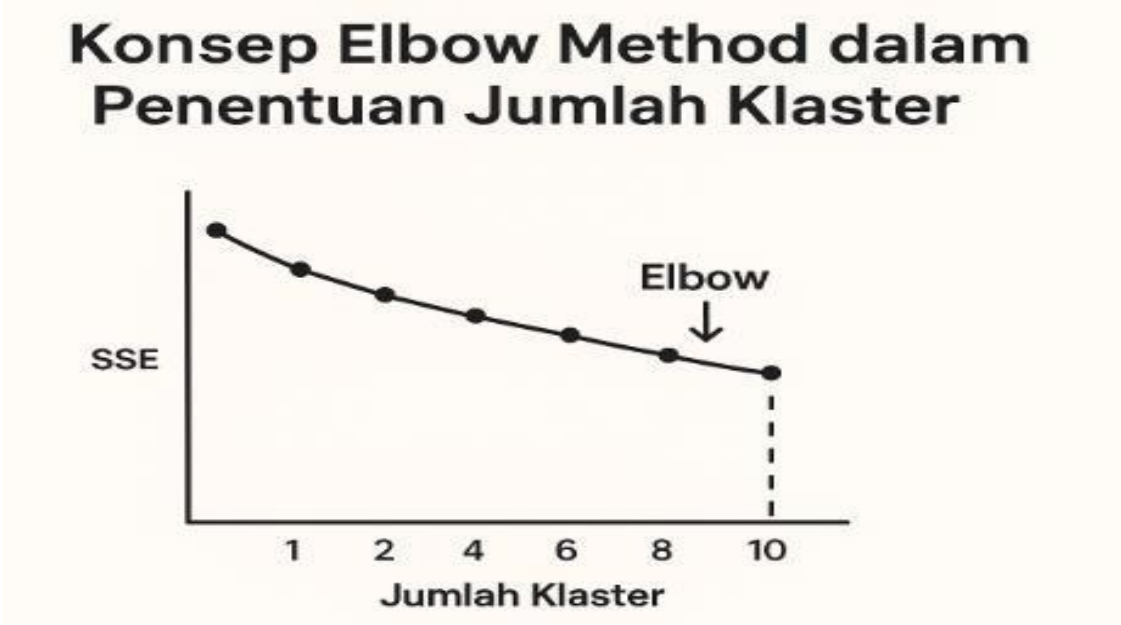
2.5.1 Konsep *Elbow Method* dalam Penentuan Jumlah Klaster

Elbow Method digunakan untuk mengidentifikasi titik ketika penambahan klister tidak lagi memberikan peningkatan kualitas yang signifikan**.**Secara konseptual, metode ini menggambarkan grafik jumlah klaster terhadap nilai WCSS, di mana titik tekuk (*elbow point*) mengindikasikan nilai *k* terbaik. Sebelum titik tersebut, penambahan jumlah klaster secara signifikan mengurangi WCSS; setelahnya, penurunan menjadi tidak signifikan. Dengan demikian, *Elbow Method* membantu mencegah *overfitting* dengan memilih jumlah klaster yang merepresentasikan keseimbangan antara kompleksitas model dan akurasi

pemisahan data (Saputra, Saputra, and Oswari 2020)Secara matematis, rumus WCSS dituliskan sebagai:

$$WCSS = \sum_{j=1}^{k} \sum_{x \in C_j} |x - \mu_j|^2$$

Nilai WCSS yang terlalu kecil bisa menandakan bahwa model terlalu kompleks, sedangkan nilai yang terlalu besar mengindikasikan bahwa data belum terkelompok dengan baik. Dalam konteks pendidikan, *Elbow Method* sering digunakan untuk menentukan jumlah kelompok siswa berdasarkan performa akademik atau tingkat kehadiran (Aulia and Nurahman 2023)



Gambar 2.3 Konsep **Elbow Method**

2.5.2 Konsep *Silhouette Score* dan Interpretasi Kualitas Klaster

Selain *Elbow Method*, *Silhouette Score* digunakan untuk mengevaluasi kualitas pemisahan antar klaster berdasarkan jarak antar titik data. Nilai *silhouette coefficient* dihitung menggunakan rumus:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Di mana *a(i)a(i)a(i)* adalah rata-rata jarak antara satu titik dengan semua titik lain dalam klaster yang sama, dan *b(i)b(i)b(i)* adalah jarak rata-rata ke titik-titik di klaster terdekat(Ridho et al. 2023).

Silhouette Score mengukur seberapa dekat data dalam satu klaster dan seberapa jauh dari klaster lain, sehingga nilai tinggi menunjukkan klaster yang terpisah dengan baik. Nilai *Silhouette Score* berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa data sangat cocok dengan klasternya, sedangkan nilai negatif menandakan bahwa data mungkin berada pada klaster yang salah (Latifah, Surono, and Suparman 2022)Nilai di sekitar 0 menunjukkan bahwa data berada di batas dua klaster. Menurut (Nasyuha, Zulham, and Rusydi 2022)kombinasi *Silhouette Score* dan *Elbow Method* memungkinkan validasi hasil *clustering* yang lebih objektif, terutama dalam analisis *Educational Data Mining*.

Gambar 2. 4.Konsep Silhouette Score

2.5.3 Elbow Method **dan** Silhouette Score **dalam Validasi Klaster**

Kedua metode evaluasi tersebut saling melengkapi. *Elbow Method* fokus pada penurunan *distortion error*, sementara *Silhouette Score* menilai jarak relatif antar klaster dan homogenitas internal(Juanita, Cahyono, and Luhur 2024). Kombinasi keduanya dapat memastikan bahwa model tidak hanya menghasilkan klaster dengan jarak yang optimal tetapi juga dengan pemisahan yang baik. Dalam penelitian pendidikan, kombinasi metode ini dapat digunakan untuk mengidentifikasi kelompok siswa berdasarkan gaya belajar, tingkat keterlibatan, atau performa akademik secara objektif (Zainuddin and Risal 2024) . Evaluasi menggunakan Elbow Method dan Silhouette Score dipilih untuk memastikan jumlah klaster optimal sekaligus menilai pemisahan klaster secara objektif.(Carr et al. 2021) Dengan demikian, keduanya menjadi alat validasi ganda yang memperkuat kredibilitas analisis *K-Means.*

2.5.4 Tantangan dalam Evaluasi Algoritma *Unsupervised Learning*

Salah satu tantangan utama dalam mengevaluasi algoritma *unsupervised learning* adalah ketiadaan *ground truth* atau label kebenaran yang diketahui. Karena tidak ada kategori yang ditetapkan sebelumnya, hasil pengelompokan harus dievaluasi berdasarkan metrik internal seperti *Silhouette Score* atau eksternal seperti *Davies-Bouldin Index*. Tantangan lainnya termasuk sensitivitas hasil terhadap *scaling*, inisialisasi centroid, serta distribusi data yang tidak seragam.(Ginting, Efendi, and Suwilo 2022), ketika data pendidikan memiliki karakteristik yang bervariasi (misalnya, nilai akademik, absensi, partisipasi siswa), evaluasi algoritma harus mempertimbangkan normalisasi dan penyeimbangan bobot agar tidak bias terhadap atribut tertentu. Selain itu (Khan et al. 2024) menegaskan bahwa kualitas klaster juga sangat dipengaruhi oleh pemilihan parameter awal dan metode *seeding*, yang dapat menyebabkan variasi hasil.

2.5.5 Pentingnya Evaluasi dalam Analisis Data Pendidikan

Evaluasi hasil *clustering* memiliki dampak langsung terhadap interpretasi kebijakan dan keputusan pendidikan. Misalnya, jika klaster siswa dibentuk berdasarkan performa akademik tanpa evaluasi yang tepat, kelompok yang terbentuk dapat bersifat bias atau tidak representatif. Penggunaan *Elbow Method* dan *Silhouette Score* membantu memastikan bahwa hasil pengelompokan mencerminkan struktur data yang sebenarnya dan dapat digunakan untuk pengambilan keputusan berbasis data. Dalam konteks *Educational Data Mining*, hasil evaluasi ini digunakan untuk memantau dinamika kelas, mengidentifikasi siswa berisiko rendah, menengah, dan tinggi, serta merancang intervensi pembelajaran yang lebih adaptif Selain itu, kombinasi evaluasi numerik dan

interpretasi pedagogis menjadi penting untuk memastikan keberlanjutan dan keadilan dalam penerapan model analisis data (Junaedi and Arifin 2019)

2.5.6 Rumusan Evaluasi dalam Penelitian ini

Dalam penelitian ini, evaluasi algoritma dilakukan melalui dua tahap utama: (1) penerapan *Elbow Method* untuk menentukan jumlah klaster optimal berdasarkan nilai WCSS, dan (2) penghitungan *Silhouette Score* untuk menilai seberapa baik setiap data cocok dengan klasternya. Tahapan evaluasi ini mengikuti praktik umum dalam literatur terbuka seperti yang dijelaskan Secara konseptual, pendekatan ini menjamin bahwa hasil analisis bersifat *replicable, interpretable,* dan sesuai dengan struktur alami data pendidikan yang digunakan. Dalam konteks penerapan, metode ini juga mendukung prinsip *fair clustering*, yakni pengelompokan yang tidak bias terhadap atribut demografis atau sosial siswa (Zainuddin and Risal 2024)

2.6 Alat Bantu Pemrograman dan Tools Pendukung

2.6.1. Alat Bantu terhadap Keakuratan dan Efisiensi Analisis RapidMiner

RapidMiner merupakan perangkat lunak berbasis *open source* yang memberikan fleksibilitas kepada pengguna dalam menyesuaikan serta memodifikasi fitur sesuai dengan kebutuhan analisis sesuai kebutuhan analisis yang spesifik. Hal ini memberikan fleksibilitas tinggi dalam proses pengolahan data dan memungkinkan pengguna untuk menyesuaikan proses analisis agar sesuai dengan tujuan tertentu. Perangkat ini menawarkan lingkungan kerja yang terpadu dan mendukung berbagai metode analisis, seperti seperti *data mining*, *text mining*, serta analisis prediktif dan deskriptif, sehingga menjadikannya mampu mengolah data

dalam berbagai kondisi yang kompleks secara lebih efisien. (Sudarsono and Leo 2021).

Salah satu keunggulan utama RapidMiner terletak pada desain antarmukanya yang user-friendly, sehingga dapat digunakan oleh pengguna tanpa latar belakang teknis maupun kemampuan pemrograman yang mendalam. Antarmuka yang intuitif memungkinkan pengguna menyusun alur kerja (workflow) analisis melalui fitur drag-and-drop, sehingga proses seperti preprocessing data, penerapan algoritma K-Means, hingga evaluasi hasil clustering dapat dilakukan dengan lebih mudah dan terstruktur (Fatmawati1 and Agus Perdana Windarto2 2018).

Selain itu, RapidMiner mampu menangani berbagai format data, seperti CSV, Excel, maupun data yang berasal dari basis data relasional, sehingga proses impor dan pengolahan data menjadi lebih fleksibel. Perangkat ini juga dilengkapi dengan fitur visualisasi yang beragam dan interaktif, seperti scatter plot, heatmap, serta grafik dinamis lainnya, yang dapat membantu pengguna dalam memahami serta mengevaluasi hasil clustering secara visual dengan lebih jelas(Sari, Wanto, and Windarto 2018)

Gambar 2. 5.Rapidminer

2.6.2. Tools Pendukung

Pemilihan perangkat pemrograman dan pustaka pendukung yang tepat memiliki dampak langsung terhadap keakuratan dan efisiensi hasil analisis. Pustaka seperti *scikit-learn* menyediakan fungsi evaluasi otomatis seperti *silhouette score* yang membantu menilai kualitas klasterisasi secara objektif(Purnomo et al. 2025). Sementara itu, *NumPy* memastikan perhitungan matematis dilakukan dengan presisi tinggi, dan *Matplotlib* memungkinkan pengguna untuk memverifikasi hasil analisis secara visual.(Nasution et al. 2025)

Justifikasi pemilihan pustaka ini juga berkaitan dengan konteks pendidikan. Dengan data yang cenderung heterogen, seperti nilai ujian, kehadiran, dan keterampilan sosial, efisiensi perhitungan dan kemampuan visualisasi menjadi sangat penting. Penggunaan pipeline Python memungkinkan peneliti untuk mengeksekusi model, melakukan *cross-validation*, dan menghasilkan grafik evaluasi tanpa perlu berpindah antar platform.(Arvi et al. 2024)

2.7 Penelitian Terdahulu dan Kelebihan Penelitian

2. 1 Tabel Penelitian Terdahulu

Peneliti	**Judul**	**Tahun**	**Data & Algoritma**	**Hasil**
Hendrastuty	Analisis Pengelompokan Siswa Berdasarkan Hasil Belajar Menggunakan Metode Clustering	2024	Nilai hasil belajar siswa; Algoritma: K-Means	Mendapatkan Silhouette Score 0,9168 yang menunjukkan kualitas klaster sangat baik.
Flomina	Academic Performance Mapping Using K-Means	2025	Data nilai akademik mahasiswa; Algoritma: K-Means	Menghasilkan pemetaan performa akademik untuk mendukung layanan akademik.
Santosa	K-Means untuk Prediksi GPA	2021	Dataset akademik; K-Means	K-Means mampu mengelompokkan mahasiswa berdasarkan GPA
Nurdiansyah	Pengelompokan Sekolah Dasar	2023	Data fasilitas sekolah; K-Means	Menghasilkan 3 cluster berdasarkan kapasitas dan fasilitas
Meng	Integrasi Konsep Hijau dalam Pendidikan	2024	Data kurikulum; K-Means	Menghasilkan cluster terkait integrasi konsep rendah karbon
Muhammad Rizki	Analisis Pengelompokan Nilai Siswa Kelas VI SDN 05 Bilah Barat	2025	Data nilai siswa; K-Means + Elbow + Silhouette	Sedang di Proses

Penelitian ini memiliki beberapa kelebihan yang membedakannya dari penelitian-penelitian terdahulu yang membahas pengelompokan nilai siswa

maupun penerapan algoritma *clustering* dalam konteks pendidikan. Sebagian besar penelitian sebelumnya hanya berfokus pada penggunaan algoritma *K-Means* untuk mengelompokkan data pendidikan tanpa mengintegrasikan metode evaluasi yang komprehensif seperti *Elbow Method* dan *Silhouette Coefficient*. Pada penelitian ini, kedua metode tersebut digunakan secara simultan untuk menentukan jumlah klaster optimal sekaligus mengevaluasi kualitas pemisahan klaster, sehingga hasil pengelompokan menjadi lebih valid dan dapat dipertanggungjawabkan. Selain itu, penelitian ini memanfaatkan data riil siswa kelas VI di SD Negeri 05 Bilah Barat, sehingga temuan yang dihasilkan lebih kontekstual dan relevan untuk kebutuhan sekolah dasar, berbeda dengan kebanyakan penelitian sebelumnya yang menggunakan data simulasi atau data dari pendidikan tinggi.

2.8 Flowchart Kerangka Penelitian