BAB II
LANDASAN TEORI

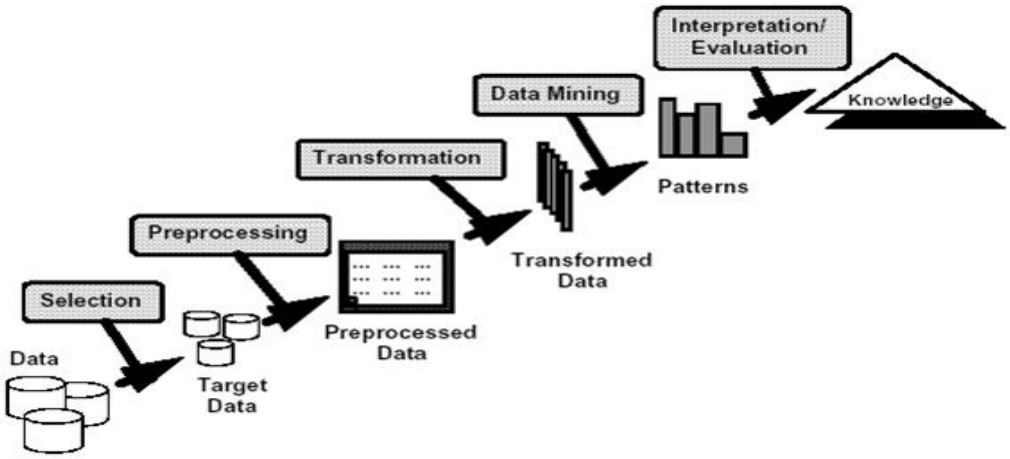
2.1 *Data Mining* dalam Pendidikan

Data mining atau penambangan data merupakan proses analisis untuk menemukan pola-pola tersembunyi, korelasi, serta informasi bermakna dari kumpulan data besar. Menurut [1], *Data mining* merupakan proses mencari pola atau informasi yaitu menarik data terpilih dengan menggunakan teknik atau metode tertentu. Teknik ini melibatkan kombinasi antara metode statistik, komputasi, dan algoritma pembelajaran mesin (*machine learning*) untuk mendukung pengambilan keputusan berbasis data. Dalam ranah pendidikan, pendekatan ini dikenal sebagai Educational *Data Mining* (EDM), yang berfokus pada pemanfaatan data pendidikan untuk meningkatkan kualitas proses belajar mengajar dan pengelolaan akademik. Menurut [2], EDM memungkinkan sekolah untuk memahami kecenderungan akademik siswa melalui analisis data nilai, kehadiran, serta aktivitas belajar. EDM juga berperan dalam memberikan saran kebijakan akademik kepada pemangku kepentingan pendidikan melalui pemrosesan dan interpretasi data pendidikan secara sistematis. [3] menyatakan bahwa penerapan EDM dapat digunakan sebagai instrumen pendukung dalam merumuskan kebijakan sekolah berdasarkan data aktual. Terdapat berbagai teknik dalam *data mining* yang digunakan dalam konteks analisis pendidikan, salah satunya ialah teknik *clustering*. *Clustering* merupakan teknik eksploratif yang digunakan ketika data tidak memiliki label kelas. Teknik ini

berguna untuk menemukan struktur alami dalam data, misalnya untuk mengelompokkan siswa berdasarkan pola nilai Matematika atau gaya belajar. *clustering* efektif dalam membantu guru memahami karakteristik kelompok siswa. Penelitian oleh [4] juga menegaskan pentingnya metode ini dalam menemukan segmentasi alami di lingkungan belajar.

Penerapan *data mining* dalam konteks Sekolah Menengah Atas (SMA) menunjukkan efektivitas tinggi dalam mendukung proses pembelajaran yang lebih adaptif dan berbasis data. [5] menyatakan bahwa metode *clustering* bermanfaat dalam memetakan tingkat pemahaman siswa dan mendeteksi kelompok siswa yang membutuhkan perhatian khusus. [6] menambahkan bahwa data nilai Matematika dapat digunakan untuk mengelompokkan siswa dalam kategori prestasi tertentu. Dengan memahami segmentasi ini, guru dapat merancang pendekatan pembelajaran yang disesuaikan dengan kebutuhan masing-masing kelompok siswa, sehingga meningkatkan motivasi dan hasil belajar. Evaluasi pembelajaran berbasis data juga menjadi landasan dalam peningkatan kualitas pengajaran. [7] mengungkapkan bahwa *data mining* membantu guru dalam mengevaluasi efektivitas metode mengajar dan menyusun perbaikan berbasis hasil analisis data.

Penggunaan teknologi digital dalam pembelajaran juga mendukung pemanfaatan *data mining* secara lebih luas. *Digital learning*, jika digabungkan dengan analitik data, dapat memberikan gambaran yang lebih jelas tentang kemajuan belajar siswa. [8] menambahkan bahwa penggunaan media berbasis teknologi, seperti video animasi, dapat meningkatkan pemahaman siswa jika diterapkan dengan strategi pengajaran yang berbasis data.



Gambar 2. 1 Proses Umum *Data Mining*

2.2 *Clustering*

Clustering adalah proses pengelompokan data ke dalam kelompok-kelompok berdasarkan kemiripan karakteristik, sehingga data yang berada dalam satu kelompok (klaster) akan memiliki kemiripan tinggi, dan berbeda signifikan dengan data dari kelompok lain. Teknik ini termasuk dalam metode *unsupervised learning*, karena tidak memerlukan informasi atau label kelas sebelumnya. Teknik *clustering* sangat relevan digunakan dalam dunia pendidikan, terutama untuk mengelompokkan siswa berdasarkan tingkat pemahaman, minat belajar, atau hasil akademik mereka. Dengan *clustering*, pendidik dapat memahami lebih dalam tentang segmentasi siswa dan menyesuaikan strategi pembelajaran sesuai dengan kebutuhan masing-masing kelompok. Tujuannya adalah untuk menemukan pola tersembunyi dalam data akademik yang tidak mudah diidentifikasi menggunakan pendekatan konvensional

Secara umum, terdapat dua pendekatan dalam teknik *clustering*:

1. *Hierarchical Clustering*, membentuk struktur pohon (dendrogram) yang menunjukkan proses penggabungan atau pemisahan klaster secara bertahap.
2. *Partitional Clustering*, mempartisi data ke dalam sejumlah klaster tetap, seperti *algoritma K-Means*, berdasarkan iterasi terhadap jarak dari pusat klaster.

Tabel 2. 1 Perbandingan *Hierarchical* dan *Partitional Clustering*

Kriteria	***Hierarchical Clustering***	***Partitional Clustering (K-Means)***
Struktur output	Dendrogram	Klaster tetap
Penentuan jumlah klaster	Tidak diperlukan sebelumnya	Diperlukan di awal
Kompleksitas komputasi	Tinggi	Rendah
Skalabilitas	Rendah	Tinggi

Penelitian ini menggunakan pendekatan *partitional clustering* dengan *algoritma K-Means* karena memiliki efisiensi komputasi tinggi dan lebih cocok untuk data numerik seperti nilai akademik siswa

[9] menjelaskan bahwa *algoritma K-Means* bekerja dengan menentukan sejumlah klaster (k) terlebih dahulu, lalu mengelompokkan data berdasarkan jarak *Euclidean* terhadap pusat klaster (*centroid*) yang dihitung ulang secara iteratif hingga hasilnya konvergen. Dalam bidang pendidikan, aplikasi *clustering* menjadi sangat penting dalam mendukung strategi pembelajaran yang adaptif. Sebagai contoh, [10] menerapkan *K-Means* untuk merekomendasikan pemilihan jurusan bagi siswa SMK berdasarkan nilai akademik, sehingga siswa dapat menentukan jalur pendidikan lanjutan yang sesuai dengan kemampuan mereka. Penelitian yang

dilakukan oleh [11] membuktikan bahwa *K-Means* sangat efektif digunakan dalam mengelompokkan siswa berprestasi berdasarkan nilai akademik, sehingga hasil *clustering* tersebut dapat digunakan oleh pihak sekolah dalam memberikan intervensi pembelajaran yang lebih tepat sasaran.

Clustering, khususnya dengan *K-Means*, tidak hanya mendukung personalisasi dalam pendidikan, tetapi juga meningkatkan efektivitas pengambilan keputusan berbasis data. Oleh karena itu, penerapan metode ini dapat memberikan kontribusi besar dalam perencanaan intervensi yang tepat sasaran dalam lingkungan pembelajaran.

2.3 *Algoritma K-Means*

Algoritma K-Means merupakan metode partisi yang paling banyak digunakan karena kemudahannya dalam implementasi dan efisiensi waktu komputasi. Menurut [12] *Algoritma K-Means* adalah salah satu teknik klasterisasi paling populer yang digunakan dalam analisis data yang keunggulannya terletak pada kemampuannya untuk mengelompokkan data ke dalam kumpulan-kumpulan yang homogen berdasarkan fitur-fitur tertentu. Menurut [11], *K-Means* bekerja dengan membagi data ke dalam sejumlah klaster yang telah ditentukan sebelumnya (k). Menurut [13] Pengelompokan *K-Means* digunakan untuk mengklasifikasikan data ke dalam kelompok-kelompok metodenya adalah dengan melihat angka-angka dalam variabel *K-Means*. Algoritma ini bekerja secara iteratif dari setiap titik hingga terbentuk kelompok *K-Means*. Berdasarkan fitur praktisnya, data umum akan dipertahankan. Probabilitas untuk tergabung dalam kelompok yang sama meningkat seiring dengan kesamaan data.

Algoritma ini bekerja melalui lima tahap utama, yaitu:

1. Inisialisasi dan Pemilihan *Centroid* Awal: Proses dimulai dengan memilih k pusat klaster awal. Pemilihan ini dapat dilakukan secara acak atau menggunakan metode seperti *K-Means++* yang memberikan hasil lebih stabil

2. Penugasan Data ke Klaster: Setiap data akan dihitung jaraknya ke setiap *centroid* menggunakan metrik seperti jarak *Euclidean*. Data kemudian ditugaskan ke klaster dengan *centroid* terdekat. Rumus *Euclidean Distance* :

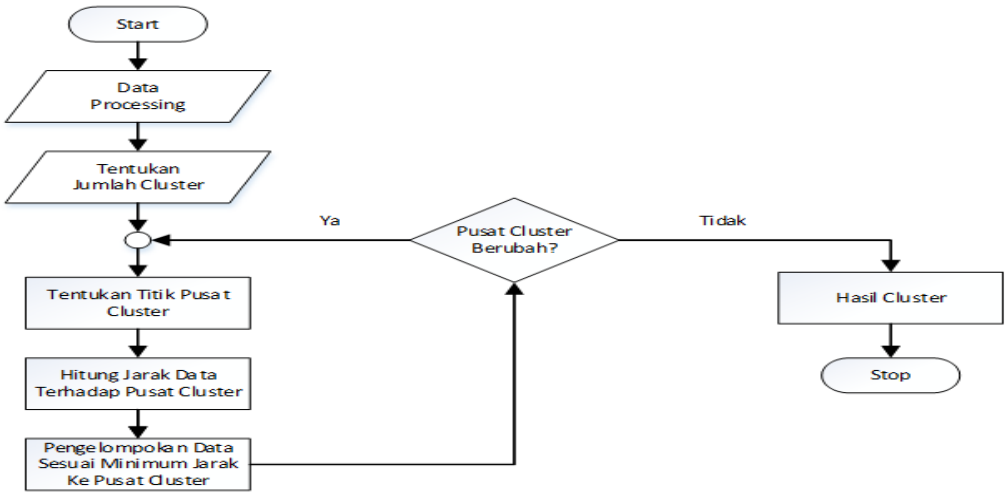
$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$$

3. Penghitungan Ulang *Centroid*: Setelah semua data ditugaskan ke klaster, *centroid* setiap klaster dihitung ulang sebagai rata-rata dari seluruh data dalam klaster tersebut [14]

$$v = \frac{\sum_{i=1}^{n} xi}{n} \qquad ; i = 1,2,3, \dots n$$

4. Iterasi Proses: Langkah penugasan dan penghitungan *centroid* diulang hingga tidak ada perubahan signifikan dalam komposisi klaster atau mencapai iterasi maksimum.

5. Evaluasi dan Validasi Hasil: Evaluasi kualitas klaster dapat dilakukan menggunakan indeks seperti *Silhouette Score* atau *Davies-Bouldin Index* untuk mengukur konsistensi dan efektivitas pengelompokan



Gambar 2. 2 Diagram Alir Metode *K-Means*

Algoritma K-Means memiliki berbagai keunggulan, antara lain:

1. Kesederhanaan dan Implementasi Cepat: Algoritma ini mudah dipahami dan diimplementasikan, sehingga menjadikannya populer dalam penelitian akademik.

2. Efisiensi Komputasi: *K-Means* mampu memproses dataset besar dengan waktu yang relatif singkat.

3. Kemampuan Menangani Data Numerik: Algoritma ini sangat cocok digunakan dalam konteks data numerik, seperti nilai ujian siswa [15]

4. Fleksibel untuk Berbagai Aplikasi Pendidikan: Digunakan dalam berbagai studi seperti segmentasi siswa, rekomendasi bidang studi, dan klasifikasi prestasi akademik .

Namun, algoritma ini juga memiliki beberapa keterbatasan:

1. Sensitivitas terhadap Inisialisasi *Centroid*: sensitivitas terhadap pemilihan jumlah klaster (K) yang harus ditentukan sebelumnya dan ketergantungan pada inisialisasi awal *centroid* yang dapat memengaruhi hasil akhir pengelompokan. [13]
2. Asumsi Bentuk Klaster: *K-Means* mengasumsikan bahwa klaster berbentuk bola dengan ukuran seragam, yang tidak selalu sesuai dengan kondisi nyata dalam pendidikan.
3. Rentan terhadap *Outlier* dan *Noise*: Adanya data ekstrem dapat mempengaruhi posisi *centroid* dan hasil pengelompokan secara signifikan.
4. Menentukan Jumlah Klaster (k): Tidak adanya metode otomatis untuk menentukan nilai k yang optimal dapat menyebabkan bias dalam analisis

2.4 Evaluasi Model *Clustering*

Evaluasi diperlukan untuk mengetahui seberapa baik hasil klaster mewakili struktur data yang sesungguhnya. Penelitian ini menggunakan dua metode evaluasi:

1. *Davies-Bouldin Index* (DBI): Mengukur rasio antara rata-rata dispersi intra-klaster dengan jarak antar klaster. DBI dihitung dengan rumus:

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} \frac{max}{j \neq i} \begin{pmatrix} S_i + S_j \\ M_{ij} \end{pmatrix}$$

Keterangan:

k = jumlah klaster

Si = rata-rata jarak dalam klaster

Sj = jarak antara *centroid* klaster dan

Nilai DBI yang lebih kecil menandakan kualitas pengelompokan yang lebih baik, dengan nilai ideal di bawah 1.0. DBI digunakan untuk evaluasi global antar klaster dan sensitif terhadap penyebaran serta pemisahan klaster ([16]; [17]; [18]).

Tabel 2. 2 Interpretasi Nilai DBI

Nilai DBI	Interpretasi
0.0 – 0.5	Klaster sangat baik (kompak & terpisah)
0.5 – 1.0	Klaster cukup baik
> 1.0	Klaster buruk (tumpang tindih tinggi)

2. *Silhouette Coefficient*: Mengukur kesesuaian data terhadap klasternya dibandingkan klaster lain. Rumus *silhouette* adalah:

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

Keterangan:

s(i)= nilai *silhouette* untuk data titik iii

a(i)= rata-rata jarak antara titik iii dan semua titik lain dalam klaster yang sama (disebut *intra-cluster distance*)

b(i)= rata-rata jarak titik iii ke semua titik dalam klaster terdekat lainnya (disebut *nearest-cluster distance*)

max{a(i),b(i)}= nilai maksimum dari dua jarak di atas

Nilai *Silhouette* berkisar dari -1 hingga 1. Nilai mendekati 1 menunjukkan pengelompokan sangat baik, nilai antara 0.5 hingga 1 menunjukkan pemisahan

yang baik, dan nilai negatif menunjukkan kemungkinan kesalahan pengelompokan ([19]; [20]).

Tabel 2. 3 Interpretasi Nilai *Silhouette* Score

Nilai s(i)s(i)s(i)	Interpretasi
Mendekati **+1**	Titik sangat cocok dengan klasternya sendiri dan sangat berbeda dari klaster lain
Sekitar **0**	Titik berada di batas dua klaster
Mendekati **-1**	Titik kemungkinan salah ditempatkan dalam klaster

Kombinasi antara DBI dan *Silhouette Coefficient* disebut sebagai evaluasi ganda (*multi-metric evaluation*) dan memberikan hasil yang lebih holistik. Hal ini disampaikan oleh [21] dalam penelitiannya yang menekankan pentingnya penggunaan lebih dari satu metrik untuk mengevaluasi hasil klaster agar tidak bias pada satu aspek saja. Misalnya, DBI lebih sensitif terhadap sebaran data dalam klaster, sedangkan *Silhouette* lebih berfokus pada posisi relatif setiap data terhadap klaster lainnya. [22] menyebutkan bahwa dengan bantuan visualisasi DBI dan grafik *Silhouette* di *RapidMiner*, pengguna dapat secara interaktif menentukan klaster yang paling menggambarkan struktur data yang sebenarnya.

2.5 *RapidMiner*

RapidMiner merupakan salah satu perangkat lunak berbasis GUI yang menyediakan lingkungan pengembangan terpadu (IDE) untuk kebutuhan data *science*, khususnya dalam domain *data mining*, *machine learning*, dan analisis statistik. Menurut [23], *RapidMiner* memiliki keunggulan utama berupa antarmuka yang intuitif dan sistem *drag-and-drop*, sehingga sangat cocok digunakan oleh peneliti pendidikan yang tidak memiliki latar belakang teknis mendalam di bidang

data *science*. Perangkat ini juga memungkinkan pengguna untuk merancang proses analisis data secara modular, dimulai dari tahap pemrosesan awal (*preprocessing*), pemodelan, hingga evaluasi model.

1. **Tahap *Preprocessing***

Pada tahap ini, *RapidMiner* menyediakan berbagai operator yang memungkinkan pembersihan data dari nilai kosong, duplikasi, hingga *noise*. *RapidMiner* menawarkan teknik imputasi dan penghapusan data yang hilang, serta transformasi skala seperti normalisasi dan standardisasi. Proses feature engineering juga didukung dengan fitur ekstraksi atribut dan seleksi fitur yang berguna untuk meningkatkan akurasi model dan efisiensi pemrosesan data.

2. **Tahap Modeling**

RapidMiner memungkinkan integrasi langsung *algoritma K-Means* dan algoritma lainnya ke dalam workflow analitik. [22] menambahkan bahwa fleksibilitas ini memudahkan peneliti untuk melakukan eksplorasi terhadap berbagai konfigurasi model dan membandingkan hasilnya secara langsung. [24] juga mencatat bahwa *RapidMiner* menyediakan antarmuka intuitif dalam konfigurasi dan eksekusi *K-Means*. Hal ini memungkinkan pemilihan model terbaik berdasarkan kinerja metrik evaluasi yang diinginkan.

3. **Tahap Evaluasi**

Evaluasi hasil klasterisasi menjadi bagian penting dalam proses analisis data. *RapidMiner* menyediakan metrik seperti *Davies-Bouldin Index dan Silhouette Coefficient*. [24] membuktikan bahwa evaluasi berbasis metrik ini memberikan

wawasan mendalam terkait kohesi dan pemisahan antar klaster. Visualisasi seperti scatter plot atau diagram radial dapat membantu dalam menilai sebaran data, keterkaitan antar klaster, dan pengambilan keputusan berdasarkan pola yang teridentifikasi.

Penggunaan *RapidMiner* dalam pengolahan *data mining* pendidikan memiliki sejumlah kelebihan dibandingkan dengan alat lain yang serupa. Berikut adalah beberapa aspek yang menonjol dari *RapidMiner* dan bagaimana keunggulan ini berkontribusi dalam pengolahan data akademik:

1. Antarmuka Pengguna yang Intuitif, *RapidMiner* menawarkan antarmuka grafis yang memungkinkan pengguna, bahkan mereka yang tidak memiliki latar belakang teknis yang kuat, untuk melakukan analisis data dengan relatif mudah. Kemampuan *drag-and-drop* memfasilitasi pemodelan dan *preprocessing* data tanpa memerlukan keterampilan pemrograman yang mendalam [25] Ketersediaan Berbagai Algoritma dan Metode, *RapidMiner* memiliki banyak algoritma *machine learning* dan teknik pengolahan data yang siap digunakan, termasuk *K-Means Clustering* dan banyak algoritma lainnya. Hal ini memberi fleksibilitas kepada peneliti dan praktisi pendidikan untuk mengeksplorasi berbagai pendekatan analitis dalam satu platform, mempercepat proses percobaan dan pengujian model [26]

2. Kemudahan dalam Proses *Preprocessing Data*, *RapidMiner* dilengkapi dengan berbagai alat *preprocessing* yang kuat dan efisien. Pengguna dapat melakukan manipulasi data seperti normalisasi, transformasi, pemilihan

fitur, dan penanganan data yang hilang dengan relatif mudah. Alat ini sangat penting dalam pengolahan data akademik karena memastikan bahwa data siap untuk dianalisis dan menghasilkan hasil yang akurat [27].

3. Integrasi Data dan Kompatibilitas dengan Banyak Format, *RapidMiner* mampu mengimpor data dari berbagai sumber, seperti database SQL, file CSV, dan aplikasi spreadsheet. Ini memungkinkan pengumpulan data yang lebih luas dari sistem informasi akademik yang berbeda, sehingga memudahkan analisis data yang lebih komprehensif [25]; [28].

4. Opsi Pemodelan dan Evaluasi yang Mendalam**,** *RapidMiner* tidak hanya mendukung pemodelan saja, tetapi juga menyediakan alat evaluasi yang kuat untuk menilai performa model. Dengan metrik seperti *Silhouette Coefficient* dan *Davies-Bouldin Index*, pengguna dapat melakukan evaluasi yang komprehensif terhadap hasil *clustering* dan memahami kualitas model yang dihasilkan ([29]; [27]).

5. Kemampuan Visualisasi Data, Visualisasi data yang baik sangat penting untuk memahami pola dalam data akademik. *RapidMiner* menawarkan berbagai alat visualisasi yang memungkinkan pengguna untuk mengeksplorasi hasil analisis melalui grafik dan visual lainnya, membantu dalam mengambil keputusan berbasis data [30]; [31].

6. Komunitas dan Sumber Daya yang Luas, *RapidMiner* didukung oleh komunitas besar yang aktif, menyediakan dokumentasi lengkap, tutorial,

dan forum komunitas. Ini memudahkan pengguna untuk mendapatkan bantuan dan berbagi pengalaman serta teknik analisis.

2.6 Prestasi Belajar Matematika

Prestasi belajar Matematika adalah kemampuan siswa dalam memahami dan menerapkan konsep, menyelesaikan soal, serta menunjukkan kemajuan akademik melalui evaluasi yang terstandar. Menurut [32] , prestasi belajar dalam Matematika sangat dipengaruhi oleh interaksi antara faktor internal siswa dan faktor eksternal dari lingkungan belajar. Faktor internal mencakup motivasi, minat, disiplin belajar, efikasi diri, serta gaya belajar. Sedangkan faktor eksternal meliputi metode pengajaran, dukungan orang tua, kondisi sekolah, dan ketersediaan fasilitas belajar.

1. Keterlibatan Orang Tua, [33] yang menyimpulkan bahwa siswa yang mendapatkan bimbingan belajar dari orang tua cenderung memiliki prestasi yang lebih stabil dibandingkan mereka yang belajar tanpa arahan. Dukungan dan interaksi aktif orang tua dalam pendidikan memberikan kontribusi signifikan terhadap prestasi belajar siswa.

2. Motivasi Belajar, Motivasi intrinsik siswa berperan penting dalam pencapaian prestasi akademik [3] menyatakan bahwa siswa yang memiliki motivasi intrinsik yang tinggi cenderung lebih aktif dalam proses pembelajaran, lebih tekun dalam menyelesaikan soal, dan menunjukkan hasil akademik yang lebih baik dibandingkan siswa yang kurang termotivasi.

3. Kedisiplinan Belajar, Kedisiplinan dalam mengatur waktu belajar dan menyelesaikan tugas berkontribusi langsung terhadap peningkatan hasil belajar. [34] menunjukkan bahwa siswa dengan kedisiplinan tinggi memiliki rata-rata nilai Matematika yang lebih baik secara signifikan dibandingkan siswa yang sering absen atau lalai mengerjakan tugas.

4. Gaya Belajar, Preferensi gaya belajar, seperti visual, auditori, dan kinestetik, memengaruhi cara siswa menyerap informasi. Ketidak sesuaian antara gaya belajar siswa dan metode pengajaran dapat menurunkan efektivitas pembelajaran matematika.

5. Efikasi Diri, Tingkat kepercayaan siswa terhadap kemampuannya sendiri dalam menyelesaikan tugas matematika sangat berkaitan dengan hasil belajar. Menurut [35] efikasi diri yang tinggi mendorong siswa untuk tidak mudah menyerah ketika menghadapi kesulitan, serta meningkatkan kepercayaan diri dalam menjawab soal-soal dengan tingkat kesulitan tinggi.

6. Lingkungan Belajar, Faktor eksternal seperti suasana kelas, fasilitas sekolah, dan hubungan sosial turut menentukan kualitas pembelajaran.

7. Kemandirian Belajar, Siswa yang mampu mengatur sendiri proses belajarnya menunjukkan hasil akademik yang lebih baik. [36] menyatakan bahwa kemandirian belajar berkorelasi dengan kemampuan manajemen waktu dan tanggung jawab akademik.

8. Kesehatan Mental dan Emosional, Kondisi psikologis seperti kecemasan atau stres akademik berdampak negatif pada prestasi belajar matematika.

kecemasan matematis menjadi salah satu faktor penghambat utama dalam pencapaian akademik.

Prestasi belajar dapat diukur melalui pendekatan berbasis nilai akademik dan psikometrik. Masing-masing pendekatan memiliki peran dalam mengevaluasi keberhasilan siswa secara objektif dan subjektif.