

BAB II

LANDASAN TEORI

2.1 Knowledge Discovery in Database (KDD)

Konsep data mining umumnya dikenal sebagai salah satu alat bantu yang sangat penting dalam manajemen pengolahan informasi. Hal ini disebabkan data mining dapat mengolah data yang sangat besar jumlahnya dan yang seperti yang kita ketahui data akan selalu bertambah. Data mining merupakan suatu langkah dalam knowledge discovery in database. Knowledge discovery in database (KDD) merupakan salah satu tahapan atau kegiatan yang meliputi kegiatan pengumpulan data dan diikuti pemakaian data historis untuk menemukan sebuah pola yang ada dalam hubungan antar data dalam sebuah database yang berukuran besar (K. Handoko, 2016). Knowledge discovery in database sering dikaitkan dengan data mining dikarenakan sering terjadinya kesalahpahaman tentang kedua istilah tersebut. Kedua istilah tersebut memiliki konsep yang berbeda, namun sangat berikatkan satu sama lain karena data mining merupakan salah satu tahapan dalam proses knowledge discovery in database. Dalam melakukan KDD, terdapat beberapa ciri penting dalam sebuah pencarian pengetahuan, yaitu (Aranski & Handoko, 2019):

1. Data yang dikerjakan dalam jumlah yang besar.
2. Volume data yang dikerjakan harus diefisiensikan.
3. Akurasi menjadi prioritas utama.
4. Menggunakan bahasa tingkat tinggi.
5. Memanfaatkan beberapa bentuk pembelajaran otomatis.
6. Hasil yang dihasilkan berupa pengetahuan yang bermanfaat

Secara keseluruhan, proses KDD dapat dibagi menjadi beberapa tahapan yaitu sebagai berikut (Mardi Yuli, 2023)

1. Data Selection / Pemilihan Data

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam Knowledge Discovery in Database (KDD) dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data

operasional.

2. Pre-processing / Cleaning

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus Knowledge Discovery in Database (KDD). Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Juga dilakukan proses enrichment, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk Knowledge Discovery in Database (KDD), seperti data atau informasi eksternal lainnya yang diperlukan.

3. Transformation

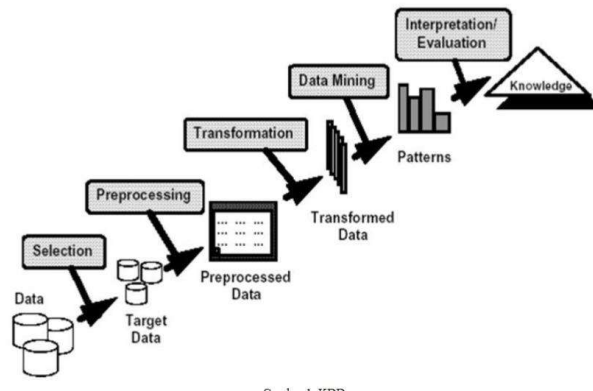
Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam Knowledge Discovery in Database (KDD) merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data Mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses Knowledge Discovery in Database (KDD) secara keseluruhan.

5. Interpretation / Evaluation

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses Knowledge Discovery in Database (KDD) yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.



Gambar 2. 1 KKD

2.2 Data Mining

Data mining dikenal sejak tahun 1990-an, ketika adanya suatu pekerjaan yang memanfaatkan data menjadi suatu hal yang lebih penting dalam berbagai bidang, seperti marketing dan bisnis, sains dan teknik, serta seni dan hiburan. Sebagian ahli menyatakan bahwa data mining merupakan suatu langkah untuk menganalisis pengetahuan dalam basis data atau biasa disebut Knowledge Discovery in Database (KDD). (Hasugian & Raphita Sagala, 2022)

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Menurut Kusri Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan machine learning untuk menginteraksi mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terikat dari berbagai database besar. (Kurniawan, 2019)

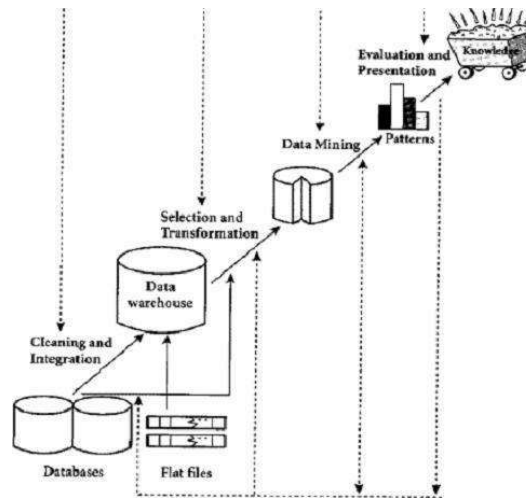
Data Mining merupakan suatu proses yang memproses satu atau lebih teknik pembelajaran komputer (machine learning) untuk menganalisis dan mengekstraksi pengetahuan (knowledge) secara otomatis. (Zulkarnaen Saputra dkk, 2023)

Adapun Menurut Charu Aggarwal (2020) Data mining adalah proses penemuan pengetahuan yang bermanfaat dari jumlah data yang sangat besar (Charu Aggarwal 2020).

- a. Beberapa teknik dan sifat data mining adalah sebagai berikut:

1. *Classification*

2. *Clustering*
3. *Association Rule*
4. *Regression*
5. *Deviation Detection*



Gambar 2. 2 Tahapan data mining

b. Tahap-Tahap Data Mining

Berikut tahapan data mining ditunjukkan pada gambar 1

Tahapan data mining dibagi menjadi bagian-bagian yaitu:

1. Pembersihan data (*Data cleaning*)

Tahapan pertama yang dilakukan adalah *Data Cleaning* Han, Kamber & Pei menjelaskan *Data Cleaning* merupakan proses untuk dapat mengatasi nilai yang hilang, noise dan data yang tidak konsisten. (Jasmine Putri dkk, 2022).

2. Integrasi data (*Data integration*)

Data integration merupakan proses menggabungkan data dari banyak sumber data. Proses ini dapat membantu mengurangi redundansi data dan data yang tidak konsisten yang disebabkan pengambilan data dari banyak sumber data. Hal ini tentu saja akan berpengaruh terhadap kecepatan dan akurasi saat melakukan data mining. (Farrel Putro dkk, 2021).

3. Seleksi data (*Data Selection*)

Data Selection merupakan proses meminimalkan jumlah data yang digunakan untuk proses mining dengan tetap merepresentasikan data

aslinya. Mengurangi jumlah data yang digunakan untuk proses mining akan lebih efisien mengingat hasil yang didapatkan sama (atau hampir sama) secara analitikal. (Gunawan Sudarsono dkk, 2021) Untuk data selection pada data nilai rapor semua data hasil dari cleaning dan integration digunakan seluruhnya agar data yang diolah mencakup keseluruhan data.

4. Transformasi data (*Data Transformation*)

Data transformation adalah tahap kerja selanjutnya dalam proses data preprocessing. dilakukan untuk mengubah bentuk dan format data hal ini tentunya akan sangat membantu memudahkan pengguna dalam proses mining ataupun memahami hasil yang didapat (Farrel Putro dkk, 2021).

5. Proses mining

Proses inti yang menggunakan sebuah metode tertentu untuk memperoleh sebuah pola dari data.

6. Evaluasi pola (pattern evaluation)

Proses mengidentifikasi pola. (Farrel Putro dkk, 2021)

7. Presentasi pengetahuan (knowledge presentation)

Presentasi Pengetahuan adalah Menampilkan pola informasi yang dihasilkan dari proses data mining, visualisasi ini membantu mengkomunikasikan hasil data mining dalam bentuk yang mudah dimengerti. Proses yang dapat mempresentasikan informasi yang dibutuhkan dimana informasi yang telah didapatkan kemudian digunakan oleh pemilik data (Farrel Putro dkk, 2021).

Berdasarkan beberapa pendapat para ahli diatas, dapat ditarik kesimpulan bahwa data mining adalah suatu proses pengolahan data yang bertujuan untuk mendapatkan, menggali dan menemukan pengetahuan yang tersembunyi dari sebuah dataset atau sekumpulan data yang berukuran sangat besar.

2.3 Clustering

Clustering merupakan salah satu teknik dari salah satu fungsionalitas data mining, algoritma clustering merupakan algoritma pengelompokan sejumlah data

menjadi kelompok-kelompok data tertentu (cluster). Pada proses clustering, tahapan menentukan atau mendeskripsikan nilai kuantitatif dari tingkat kemiripan atau ketidakmiripan data (proximity measure) memiliki peranan sangat penting, sehingga perlu dilakukannya perbandingan beberapa metode yang sering digunakan, yaitu jarak euclidean, manhattan, dan minkowski. (Kurniawan, 2019)

Clustering adalah suatu alat untuk analisa data, yang memecahkan permasalahan pengelompokan. Obyeknya ialah untuk kasus pendistribusian (orang-orang, objek, peristiwa dan lainnya) ke dalam kelompok, sedemikian hingga derajat tingkat keterhubungan antar anggota cluster yang sama adalah kuat dan lemah antar anggota dari cluster yang berbeda. Dengan cara ini masing-masing cluster menguraikan, dalam kaitan dengan kumpulan atau koleksi data, class dimana milik anggota-anggotanya. Cluster disebut juga data item dikelompokkan menurut pilihan konsumen atau hubungan logis. (Zulkarnaen Saputra dkk, 2023)

Clustering adalah proses mengelompokkan suatu objek ke dalam kelas objek yang memiliki karakteristik yang sama. Salah satu metode clustering yang dapat digunakan dalam permasalahan tersebut adalah K-Means. (Imantika dkk, 2019)

Dalam *clustering* dikenal empat tipe data. Keempat tipe data pada tersebut ialah:

1. Variabel berskala interval
2. Variabel biner
3. Variabel nominal, ordinal dan rasio
4. Variable dengan tipe lainnya

Metode *clustering* juga harus dapat mengukur kemampuannya sendiri dalam usaha untuk menemukan suatu pola tersembunyi pada data yang sedang diteliti. Terdapat berbagai metode yang dapat digunakan untuk mengukur nilai kesamaan antar objek-objek yang dibandingkan. Salah satunya ialah dengan *weighted Euclidean Distance*, *Euclidean distance* menghitung jarak dua buah point dengan mengetahui nilai dari masing- masing atribut pada kedua poin tersebut. Berikut formula yang digunakan untuk menghitung jarak dengan *Euclidean distance* (C. I. Loho dkk, 2022);

$$Distance(p,q) = (\sum_{k=1}^N \mu_k |P_k - q_k|^r)^{1/r} \quad \dots\dots\dots \text{Persamaan (1)}$$

Keterangan :

$$N = \text{Jumlah record data} \quad K = \text{Urutan feld data} \quad r = 2$$

μ_k = Bobot *field* yang diberikan *user*

Jarak adalah pendekatan yang umum dipakai untuk menentukan kesamaan atau ketidaksamaan dua vektor fitur yang dinyatakan dengan ranking. Apabila nilai ranking yang dihasilkan semakin kecil nilainya maka semakin dekat/tinggi kesamaan antara kedua vektor tersebut. Teknik pengukuran jarak dengan metode *Euclidean* menjadi salah satu metode yang paling umum digunakan (Safitri Sihombing, 2020) Pengukuran jarak dengan metode *Euclidean* dapat dituliskan dengan persamaan berikut :

$$J(v_1, v_2) = \sqrt{\sum_{k=1}^N (v_1(k) - v_2(k))^2} \quad \dots\dots\dots \text{Persamaan (2)}$$

Dimana v_1 dan v_2 adalah dua vektor yang jaraknya akan dihitung dan N menyatakan panjang vector

Berdasarkan beberapa pendapat para ahli diatas, dapat ditarik kesimpulan bahwa clustering adalah alat analisis data yang efektif untuk mengidentifikasi struktur tersembunyi dalam data dengan mengelompokkan data berdasarkan kesamaan atau karakteristik tertentu. Pemilihan metode pengukuran kemiripan dan algoritma yang tepat, seperti K-Means, sangat penting untuk mencapai hasil clustering yang optimal.

2.4 Algoritma K-Means

K-Means pertama kali dipublikasikan oleh Stuart Lloyd pada tahun 1984 dan merupakan algoritma clustering yang banyak digunakan. K-Means bekerja dengan mensegmentasi objek yang ada ke dalam kelompok atau yang disebut dengan segmen sehingga objek yang berada dalam masing-masing kelompok lebih serupa satu sama lain dibandingkan dengan objek dalam kelompok yang berbeda.(Aditya dkk, 2020)

Algoritma K-Means adalah algoritma clustering yang paling sederhana dibandingkan dengan algoritma yang lain. Algoritma ini mempunyai kelebihan yaitu mudah diterapkan dan dijalankan, relative cepat, mudah untuk diadaptasi, dan paling banyak dipraktekkan dalam tugas data mining. Algoritma ini termasuk salah satu algoritma paling penting dalam data mining.(Zulkarnaen Saputra dkk, 2023)

K-Means Clustering merupakan metode pengelompokan non-hirarki yang terkenal karena kecepatan dan kemudahannya. Metode K-Means ini mengelompokkan data ke dalam satu atau lebih cluster. Data dengan karakteristik serupa dikelompokkan bersama dalam satu cluster, sedangkan data dengan perbedaan karakteristik ditempatkan dalam cluster yang berbeda. (Amoy Parapat & Balqis, 2024) Dengan demikian, hasilnya adalah data dalam satu cluster memiliki tingkat variasi yang minimal.

Berikut adalah langkah-langkah Algoritma K-Means:

- a. Tetapkan nilai K sebagai jumlah cluster yang akan dibentuk
- b. Tetapkan pusat (Centroid) cluster awal untuk masing-masing cluster.

Proses ini melibatkan penggunaan rumus berikut sebagai dasar:

$$C_i = \frac{1}{M} \sum_{j=1}^M X_j \quad \dots\dots\dots \text{Persamaan (3)}$$

Keterangan:

C_i : centroid pada cluster

x_j : Objek ke-j

M: Banyaknya objek/jumlah objek yang menjadi anggota cluster

- c. Menghitung jarak antara data dan pusat klaster.
- d. Setelah jumlah populasi menentukan kedekatan dengan salah satu centroid yang ada, secara otomatis data populasi tersebut diklasifikasikan ke dalam kelas yang memiliki centroid yang bersesuaian.
- e. Melakukan iterasi, kemudian menentukan posisi centroid baru menggunakan persamaan.
- f. Mengulangi proses pada langkah ketiga jika terdapat data yang berpindah kelompok atau terjadi perubahan nilai centroid. (Putu dkk., 2022.)



Gambar 2. 3 Flowchart algoritma k-means

Berdasarkan beberapa pendapat para ahli diatas, dapat ditarik kesimpulan bahwa Algoritma K-Means adalah metode yang efisien dan sederhana untuk clustering data, dengan proses yang melibatkan penetapan jumlah cluster, penghitungan jarak ke centroid, dan iterasi untuk memastikan data dikelompokkan secara optimal. Keberhasilannya dalam praktek data mining terletak pada kemudahannya dan efektivitasnya dalam mengelompokkan data dengan karakteristik yang serupa ke dalam cluster yang sama.

2.5 RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik

deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari UNaiversity of Dortmund. *RapidMiner* didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan menggunakan *RapidMiner* di lebih dari 40 negara.

RapidMiner sebagai *software open source* untuk *data mining* tidak perlu diragukan lagi karena *software* ini sudah terkemuka di dunia. *RapidMiner* menempati peringkat pertama sebagai *Software data mining* pada polling oleh KDnuggets, sebuah portal *data mining* pada 2010-2011.

2.6 Software Pendukung

Software pendukung yang akan digunakan oleh peneliti pada penelitian ini adalah software RapidMiner. RapidMiner sebelumnya dikenal sebagai YALE (Yet Another Learning Environment). YALE dikembangkan pada tahun 2001 oleh sekelompok orang yang berasal dari Unit Kecerdasan Buatan Universitas Teknik Dortmund (A.I Unit in TU Dortmund). Kelompok tersebut terdiri dari Ralf Klinkenberg, Ingo Mierswa serta Simon Fisher. Dilanjutkan pada tahun 2006, Ralf dan Ingo mendirikan sebuah perusahaan yang bernama Rapid-I. Ralf dan Ingo melanjutkan untuk membangun serta mengembangkan YALE sehingga pada setahun kemudian pada tahun 2007, YALE mengalami perubahan nama menjadi RapidMiner. Pada tahun 2013, perusahaan dirian Ralf dan Ingo mengalami perubahan nama dari Rapid-I menjadi RapidMiner sesuai dengan nama aplikasi yang diciptakan oleh mereka. RapidMiner adalah platform perangkat lunak ilmu data yang dikembangkan oleh perusahaan yang memiliki nama yang sama. RapidMiner menyediakan lingkungan terintegrasi untuk melakukan persiapan data, data mining, AI, word mining guna mendukung penelitian, pendidikan, pelatihan,

rapid prototyping serta pengembangan aplikasi. Aplikasi RapidMiner juga mendukung dalam melakukan proses pembelajaran mesin termasuk persiapan data, hasil visualisasi, validasi model dan optimasi data. Walaupun aplikasi RapidMiner sangat membantu dalam tugasnya, aplikasi ini masih bersifat opensource yang berarti model inti terbuka dan dapat didapatkan secara gratis ¹⁷ pada website yang disediakan dengan syarat penggunaan RapidMiner harus bertujuan dalam membantu pendidikan. Jika digunakan secara komersil, maka akan dikenakan biaya. RapidMiner adalah salah satu software untuk pengolahan data mining. Pekerjaan yang dilakukan oleh RapidMiner text mining adalah berkisar dengan analisis teks, mengekstrak pola-pola dari data set yang besar dan mengkombinasikannya dengan metode statistika, kecerdasan buatan, dan database. Tujuan dari analisis teks ini adalah untuk mendapatkan informasi bermutu tertinggi dari teks yang diolah. RapidMiner menyediakan prosedur data mining dan machine learning, di dalamnya termasuk: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI. Penyajiannya dituliskan dalam bahasa pemrograman Java. RapidMiner merupakan software / perangkat lunak untuk pengolahan data. Aplikasi RapidMiner sangat membantu dalam melakukan data mining karena di dalam aplikasi RapidMiner sudah terdapat prinsip serta algoritma yang sering digunakan dalam proses data mining, yang dimana salah satunya adalah algoritma clustering K-means. Dengan memanfaatkan algoritma yang telah dimilikinya, aplikasi RapidMiner dapat mengekstrak pola-pola dari database besar yang telah dimasukkan ke dalam aplikasi tersebut dengan mengkombinasikan metode statistika dan kecerdasan buatan yang dimilikinya. RapidMiner sangat membantu dalam memudahkan pengguna melakukan perhitungan data yang memiliki jumlah yang sangat banyak serta yang memiliki unsur operator yang banyak. Dalam penampilan hasil, aplikasi RapidMiner sangat ¹⁸ user-friendly dikarenakan hasil yang ditampilkan oleh aplikasi RapidMiner berupa visual dengan grafik yang ditemani oleh deskripsi penjelasan dari visual tersebut sehingga pengguna dapat mengerti dari visual yang ditampilkan.

2.7 Penelitian Terdahulu

Penelitian terdahulu adalah sebagai acuan dalam penelitian yang bertujuan mempelajari studi literatur yang telah dilakukan oleh peneliti sebelumnya dan seorang penulis atau peneliti mengumpulkan data dan informasi dengan mencari, membaca, mencatat intisari dan mempelajari karya ilmiah, jurnal, skripsi serta tinjauan literatur yang berkenaan dengan topik yang dapat dijadikan acuan pembahasan dalam penyusunan penelitian ini

1. Penelitian oleh (Aditya dkk, 2020) Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. Klastering merupakan kegiatan yang bertujuan untuk mengelompokan suatu data yang memiliki sebuah kemiripan antara satu data dengan data lainnya. K-Means clustering merupakan salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster/kelompok. Pada penelitian ini dilakukan clustering dengan menggunakan algoritma K-Means menggunakan data capaian Ujian Nasional Sekolah Menengah Pertama pada tahun 2018/2019 yang diperoleh dari website resmi Pusat Penilaian Pendidikan dan Kebudayaan Kementerian Pendidikan dan Kebudayaan Republik Indonesia. Hasil cluster dengan algoritma K-Means didapatkan untuk cluster 1 terdapat 14 provinsi, cluster 2 terdapat 5 provinsi, dan cluster 3 terdapat 15 provinsi. Tingkatan cluster 1 adalah cluster dengan nilai ujian nasional Tinggi, cluster 2 adalah cluster dengan nilai ujian nasional Rendah dan cluster 3 adalah cluster dengan nilai ujian nasional Sedang. Sedangkan hasil evaluasi dari algoritma K-Means dengan jumlah kluster 3 menghasilkan nilai evaluasi Connectivity 11.916, Dunn 0.246 dan Silhouette 0.464.
2. Penelitian oleh (Zulkarnaen Saputra dkk, 2023) Klasterisasi Nilai Ujian Sekolah Menggunakan Metode Algoritma K-Means. penelitian ini adanya beberapa masalah yaitu bagaimana menerapkan Algoritma Clustering K-Means untuk mengelompokan Nilai ujian nasional berdasarkan Kabupaten Kota di Indonesia berdasarkan Nilai UN tingkat Sekolah Menengah Pertama dan bagaimana hasil pengelompokan Kabupaten dan Kota di Indonesia berdasarkan Nilai UN tingkat Sekolah Menengah Atas dengan Algoritma

Clustering K-Means. Tujuan penelitian ini yaitu untuk mengelompokkan nilai ujian tingkat Sekolah Menengah Atas. Metode yang digunakan yaitu Algoritma Clustering K-Means. Hasil dari penelitian ini yaitu untuk mengetahui pengelompokan nilai ujian sekolah siswa terbaik dengan tingkat akurasi Jumlah anggota atribut Kelas pada Cluster 0: IPS:943, Bahasa: 909, IPA: 549, Jemaah angora attribute Kolas pada Cluster 1: IPS: 422, Bahasa: 456, IPA: 276.

3. Penelitian oleh (Amos Saputra & Nataliani, 2021) Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means. Di dunia pendidikan seringkali terjadi permasalahan bagaimana cara menentukan tingkat prestasi murid dengan kategori rendah, cukup, dan tinggi serta menemukan top rank murid unggulan dalam kelas. Dalam hal ini data mining dengan pendekatan K-Means Clustering dapat digunakan untuk mengelompokkan data menjadi kumpulan data. Dalam sistem analisis, pendekatan data mining berdasarkan algoritma K-Means dapat digunakan untuk pengelompokan prestasi murid. Dalam penelitian ini data nilai siswa kelas X-XII Bahasa SMAN 1 Tenganan tahun 2014-2017, dari semester satu sampai lima dikelompokkan berdasar nilai rapor. Clustering digunakan dalam pembangunan program analitik ini untuk menilai dampak data murid terhadap kecenderungan keberhasilan murid di setiap kelompok yang dapat dibuktikan dengan kelulusan murid yang menduduki top rank serta dari hasil wawancara guru pengajar maupun wali kelas serta data nilai yang diperoleh dari Dapodik. Hasil dari penelitian ini membuktikan bahwa teknik clustering K-Means dapat dimanfaatkan oleh pengajar untuk mengkategorikan murid berdasarkan nilai mata pelajaran dan absensi, serta menggunakannya untuk menganalisis prestasi murid dengan mengelompokkan dari kategori prestasi rendah, rata-rata, dan tinggi.