BAB II
TINJAUAN PUSTAKA

2.1 *Data Mining*

2.1.1 Pengertian *Data Mining*

Data Mining merupakan proses menemukan korelasi baru yang bermanfaat, pola dan tren dengan menambang sejumlah *repository* data dalam jumlah besar, menggunakan teknologi pengenalan pola seperti statistik dan teknik *matematika*.[1] *Data Mining* disebut juga dengan *knowledge discovery* in *database* (KDD) ataupun *pattern recognition*. [3] *Data Mining* dapat dibagi menjadi empat kelompok, yaitu model prediksi (*Prediction modelling*), analisis kelompok (*Cluster analysis*), analisis asosiasi (*association analysis*) dan deteksi anomali (*anomaly detection*). Menurut *Data Mining* merupakan proses semi otomatik yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan di dalam *database* besar.[4]

2.1.2 Fungsi *Data Mining*

a. Prediction

Prediction atau fungsi prediksi merupakan salah satu fungsi *Data Mining*. Maksudnya yaitu dari proses nanti akan menemukan pola tertentu dari suatu data. Pola tersebut dapat diketahui dari variabel-variabel yang ada pada data. Pola yang didapat bisa digunakan untuk memprediksi variabel lain yang belum diketahui nilai ataupun jenisnya. Karena itulah fungsi satu ini dikatakan sebagai fungsi prediksi. Nantinya bisa digunakan untuk memprediksi variabel tertentu yang tidak ada dalam suatu data. Hal ini tentunya memudahkan dan menguntungkan bagi mereka pemilik kepentingan yang memerlukan prediksi akurat untuk membuat hal penting tersebut menjadi lebih baik.[5]

b. *Description*

Deskripsi atau *Description* merupakan proses penting dalam analisis data yang bertujuan untuk mengidentifikasi ciri-ciri krusial yang terdapat dalam *database.* Melalui proses ini, peneliti dapat menemukan pola-pola yang relevan, yang dapat memberikan wawasan berharga tentang data yang dianalisis. Hasil dari fungsi deskripsi membantu dalam membuat prediksi yang lebih akurat dan mendukung pengambilan keputusan yang lebih baik.[6]

c. *Classification*

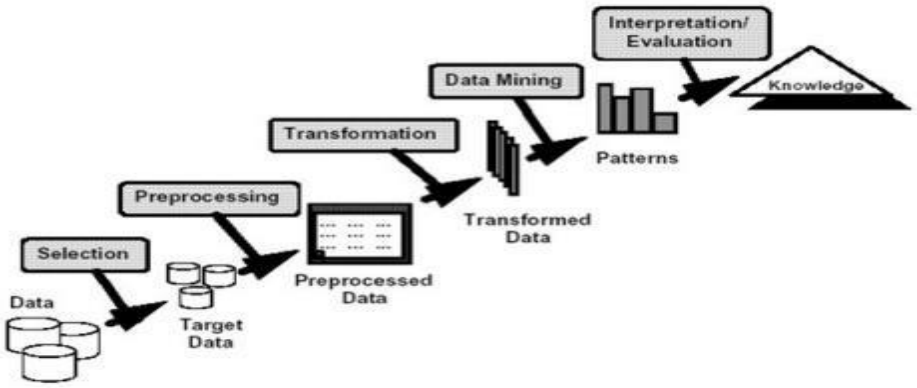
Klasifikasi atau *classification* adalah proses menemukan pola atau ciri untuk mendeskripsikan kelompok atau konsep dari data. Proses yang digunakan untuk mendeskripsikan data penting dan dapat memprediksi tren data di masa depan. Proses yang digunakan untuk menggambarkan data tersebut adalah hal yang terdapat pada masa mendatang. Contoh: Pelanggan suatu perusahaan telah berpindah ke pesaing perusahaan lainnya.[7]

d. *Association*

Asosiasi atau *association* adalah proses yang dipakai untuk menemukan suatu hubungan yang terdapat pada nilai atribut daripada sekumpulan data. Mengidentifikasikan hubungan antara kejadian-kejadian yang terjadi pada suatu waktu.[5]

2.1.3 Tahapan Dalam *Data Mining*

Tahapan yang dilakukan pada proses *Data Mining* diawali dari seleksi data dari data sumber ke data target, tahap *preprocessing* untuk memperbaiki kualitas data, transformasi, *Data Mining* serta tahap *interpretasi* dan evaluasi yang menghasilkan *output* berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik.[8] Secara detail dijelaskan sebagai berikut dan sumber gambar: [8]



Gambar 2. 1 Tahapan Data Mining

Tahap-tahap *Data Mining* adalah sebagai berikut:

a. **Data *selection***

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *KDD* dimulai. Data hasil seleksi yang digunakan untuk proses *Data Mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional. Data-data yang tidak relevan itu juga lebih baik dibuang karena keberadaannya bisa mengurangi mutu atau akurasi dari hasil *Data Mining* nantinya.[9]

b. ***Pre-processing /cleaning***

Sebelum melakukan proses *Data Mining*, perlu dilakukan proses pembersihan pada data yang difokuskan pada *KDD*. Proses pembersihan tersebut antara lain menghapus data duplikat, memeriksa konsistensi data, dan memperbaiki kesalahan data. Dari data yang diambil, pembersihan data dilakukan ketika data hilang, data duplikat, atau dicetak berlebih.[10]

c. ***Data Mining***

Data Mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik metode atau algoritma dalam *Data Mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *KDD* secara keseluruhan. Teknik, metode algoritma dalam *Data Mining* sangat bervariasi.[12]

d. ***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 *KDD* yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Pola informasi oleh proses *Data Minin*g wajib diperlihatkan ke bentuk yang mudah dipahami pihak yang mempunyai kepentingan. Fase ini menjadi bagian *KDD* yang dinamakan interpretasi. Tahapan ini meliputi pemeriksaan apakah siklus ataupun informasi yang diputuskan berlawanan dengan fakta ataupun asumsi yang sudah ada.[13]

2.2 Strategi Meningkatkan Mutu Sekolah

Strategi adalah suatu rencana yang menyeluruh, *komprehensif* dan terpadu serta berorientasi pada masa depan untuk mencapai tujuan perusahaan. Oleh karena itu, strategi merupakan alat yang penting untuk mencapai tujuan, sehingga strategi perlu diterapkan dengan baik. Menghadapi situasi penuh ketidakpastian strategi membantu perusahaan untuk mengatasi perubahan dan menyiapkan petunjuk serta upaya pengendalian yang tepat bagi perusahaan. Dengan strategi memungkinkan perusahaan membuat cara baru pada waktunya untuk mengambil keuntungan dari peluang baru dalam lingkungan dan mengurangi risikonya dengan mengimplementasikan sistem dan kebijakan yang tepat sehingga diharapkan pihak perusahaan mampu menjadikan ketidakpastian sebagai teman bukannya lawan.[14]

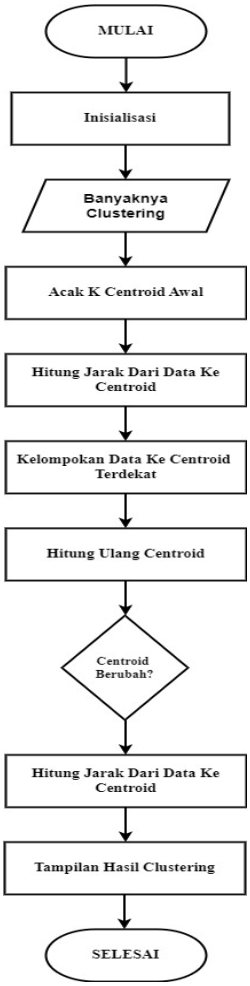
Mutu sekolah adalah kemampuan lembaga pendidikan dalam mendayagunakan sumber-sumber pendidikan untuk meningkatkan kemampuan belajar secara optimal. Dalam konteks pendidikan, pengertian kualitas atau mutu mengacu pada proses pendidikan dan hasil pendidikan. Kualitas sekolah merupakan kemampuan sekolah dalam mengembangkan ide dinamis yang meliputi *input*, *process*, *output* dan *outcame*. Beban kerja Kepala Sekolah untuk melaksanakan tugas pokok manajerial, pengembangan kewirausahaan, dan supervisi kepada guru dan tenaga kependidikan.

2.3 *Clustering*

Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode *Data Mining* yang bersifat tanpa arahan (*unsupervised*), maksudnya metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru serta tidak memerlukan target *output*. Dalam *Data Mining* ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering*.[16]

K-Means Clustering merupakan salah satu metode data *clustering* non-hierarki yang mengelompokkan data dalam bentuk satu atau lebih *cluster*/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster*/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokkan dengan *cluster*/kelompok yang lain sehingga data yang berada dalam satu *cluster*/kelompok memiliki tingkat variasi yang kecil.[8]

Langkah-langkah pada algoritma *K-Means* dapat dilihat pada gambar 2.2 di bawah:



Gambar 2. 2 *Flowchart* Proses Algoritma *K-Means*

Tahapan untuk perhitungan algoritma *K-Means* dijelaskan di bawah ini:

1. Tetapkan banyaknya jumlah *cluster (k)*
2. Pilih *random* titik pusat untuk *cluster (centroid)*
3. Pakai rumus *Euclidean Distance* untuk mendapatkan jarak tiap data terhadap *cendroid* rumusnya yakni :

   $(x,y) = \sqrt{\sum}\ ^{n}= 1(xi - yi)^2$

   Keterangan:
   $(x,)$: jarak data x dan y
   x: titik data objek

y: titik data *centroid*

i: banyaknya objek

4. Kelompokkan data yang sudah dikalkulasikan menurut jarak terkecil (minimum) antara data tersebut dengan pusat *cluster* atau data *centroid* dan mendapatkan *cluster* baru.
5. Lakukan perhitungan kembali berdasarkan data yang mengikuti *cluster* masing-masing untuk pusat *cluster (centroid*) baru. Nilai *centroid* baru didapatkan dari hasil perhitungan rata-rata data terhadap setiap *cluster*. Rumus sebagai Berikut:

$$C1 = \frac{X1+X2+X3+\cdots+Xn}{\sum x}$$

Keterangan:

CI : *centroid* baru

x1 : nilai *cluster* ke-1

xn : nilai *cluster* ke-n

x : jumlah data

6. Setelah dapat *centroid* baru, maka lakukan iterasi selanjutnya atau ulangi langkah c samapai e sampai tidak ditemukan data yang berpindah-pindah dari *cluster*.

2.4 *Rapidminer*

Rapidminer adalah *software* yang dapat diakses oleh siapa saja dan bersifat terbuka (*open source*). *RapidMiner* ini dijadikan sebuah solusi untuk menganalisa terhadap data *processing*. Pada *RapidMiner* ini digunakan berbagai teknik seperti teknik *deskriptif* dan *prediksi*. *RapidMiner* merupakan mesin pengolahan atau penambangan data yang dapat diintegrasikan ke dalam produknya sendiri dan tersedia sebagai perangkat lunak mandiri untuk analisis data.

2.5 Penelitian Terdahulu

Berikut adalah ringkasan dari beberapa penelitian sebelumnya yang terkait dengan Algoritma *K-Means*:

Tabel 2. 1 Penelitian Terdahulu

NO	Judul	Penulis/Tahun	Tujuan	Metode	Hasil
1.	Implementasi *K-Means* Untuk Identifikasi Penyakit Yang Disebabkan Oleh Nyamuk.	Sufiatul Maryana1), Agung Prajuhana Putra2), Penulis Faisal3) / 2019	Tujuan penelitian ini adalah mengidentifikasi penyakit yang disebabkan oleh nyamuk menggunakan metode *K-Means*.	Metode yang digunakan dalam penelitian ini adalah Metode *K-Means*.	Sistem telah melalui tahap uji coba yang mencakup uji coba struktural, fungsional, dan validasi guna memastikan kinerjanya. Proses uji coba ini bertujuan agar sistem dapat berfungsi dengan baik dan sesuai dengan kebutuhan. Hasil uji coba validasi menunjukkan bahwa pengelompokan data yang dilakukan oleh sistem cukup sesuai dengan data nyata yang ada. Hal ini membuktikan bahwa metode *K-Means* yang digunakan dalam penelitian ini cukup efektif untuk kasus yang diteliti. Ke depan, pengembangan aplikasi berbasis Android disarankan agar dapat mempermudah proses transfer *knowledge* serta meningkatkan aksesibilitas pengguna.
2.	Strategi *Marketi*	Melda Agarina1,	Penelitian ini memfokuska	Strategi promosi	

	ng Promosi Penerimaan Mahasiswa Baru Menggunakan *K-Means Clustering*.	Sutedi2, Arman Suryadi Karim3, Erlinda Ratna Sari4 / 2023	n pada strategi promosi melalui penerapan *Data Mining* untuk mahasiswa baru dengan metode *K-Means Clustering*.	penerimaan mahasiswa baru dirancang dengan *K-Means Clustering*, menganalisis data empat tahun terakhir menggunakan *Excel*, *RapidMiner*, dan *Tableau*.	Penelitian ini menunjukkan bahwa algoritma *K-Means Clustering* efektif dalam merancang strategi pemasaran penerimaan mahasiswa baru. Dengan menganalisis karakteristik calon mahasiswa, institusi dapat mengembangkan strategi yang lebih spesifik untuk meningkatkan jumlah pendaftar. Temuan ini diharapkan berkontribusi pada strategi pemasaran yang lebih efisien di perguruan tinggi.
3.	Implementasi Algoritma *K-Means Classifier* Sebagai Pendukung Keputusan Penerima Dana Bantuan Siswa Miskin (Studi Kasus : SMKN Sukoharjo).	Viv Fitria Yulia1 Handoyo Widi Nugroho2 / 2022		Metode yang digunakan dalam penelitian ini adalah Metode *K-Means*.	Penelitian ini menyimpulkan bahwa algoritma *K-Means* berhasil mengelompokkan data penerima bantuan siswa miskin menjadi tiga kategori: layak, dipertimbangkan, dan tidak layak. Hasil uji menunjukkan nilai *Davies-Bouldin Index* sebesar 0,262, menandakan kesamaan *cluster* yang baik. Hasil ini layak dijadikan

					rekomendasi penerima bantuan di SMKN Sukoharjo.
4.	Perbandingan Kinerja Algoritma *K-Medoids* Dan *K-Means* Untuk Klasifikasi Penyakit Kanker Serviks.	Dedi Arbain1a*, Sriyanto2b, Joko Triloka3c / 2023	Cluster Distance Penelitian ini membandingkan performa *K-Medoids* dan *K-Means Clustering* dalam pengelompokan *dataset* kanker serviks, yang terdiri dari 858 *record* dan 36 atribut. Penggunaan Performance yang tepat dapat meningkatkan efektivitas klastering.	Metode *K Medoids* dan *K-Means Clustering*	Hasil eksperimen menunjukkan bahwa *K-Means* lebih efektif dalam menangani *dataset* berukuran kecil dibandingkan dengan *K-Medoids*. Pada pemodelan menggunakan algoritma *K-Medoids*, terbentuk 361 pasien pada klaster positif dan 473 pasien pada klaster negatif. Sementara itu, pada algoritma *K-Means*, terbentuk 308 pasien pada klaster positif dan 526 pasien pada klaster negatif. Meskipun jumlah pasien berbeda, *K-Means* terbukti lebih efisien dalam pengelompokan data dengan ukuran kecil.

5.	Sistem Informasi Pemetaan Wilayah Rawan Kriminalitas Polresta Bandar Lampung Menggunakan *K-Means Clustering*.	Nurjoko1, Defi Dwirohayati2, Novi Herawadi Sudibyo3 / 2020	SIG yang dihasilkan dari penggabungan data spasial berdasarkan *clustering* dapat merekomendasikan wilayah dengan intensitas kejahatan tinggi untuk ditindaklanjuti oleh pihak berwenang. Sistem ini juga membantu dalam pembuatan laporan dan memberikan informasi kepada masyarakat Kota Bandar Lampung tentang lokasi rawan kejahatan serta daerah yang aman.	Salah satu metode dalam *clustering* adalah metode *K-Means*.	Penelitian ini menghasilkan sistem informasi pemetaan wilayah kriminalitas berbasis web dengan menggunakan metode *K-Means Clustering*. Sistem ini dirancang untuk membantu pihak POLRESTA dan POLSEK dalam menganalisis serta mengidentifikasi wilayah kriminal berdasarkan jenis dan tingkat kriminalitas yang terjadi. Dengan adanya sistem ini, diharapkan dapat memberikan informasi yang lebih akurat dan terstruktur sehingga dapat mendukung pengambilan keputusan yang lebih efektif dalam upaya pencegahan dan penanggulangan kejahatan.

6.	Implementasi *K-Means* Dalam Mengatasi Masalah *Cold Star* Pada *Collaborati*	Sylvia Sylvia 1a, Sri Lestari2 / 2022	Masalah *cold start* pada pengguna baru menghambat kinerja sistem rekomendasi karena	Metode penelitian ini menggunakan metode *K-Means*.	Mengelompokkan data dengan menggunakan algoritma *K-means* dilakukan dengan cara menentukan jumlah *cluster*, hitung jarak

	ve Filtering.		kurangnya data historis, sehingga sistem kesulitan menganalisis minat dan memberikan rekomendasi yang relevan.		terdekat dengan pusat *cluster*.
7.	Membandingkan Teknik *Data Mining* untuk Memprediksi Prestasi Akademik Mahasiswa.	Retno Dwi Handayani1) , Rini Nurlistiani2).	untuk mengidentifikasi faktor yang mempengaruhi tingkat keberhasilan mata kuliah dan tingkat keberhasilan siswa kemudian menggunakan faktor-faktor tersebut sebagai prediktor awal untuk tingkat keberhasilan yang diharapkan dan penanganan kelemahannya.	Metode yang digunakan penelitian ini adalah Teknik *Data Mining*.	Makalah ini mengulas berbagai penelitian tentang prediksi kinerja mahasiswa dengan menggunakan berbagai metode analitik. Sebagian besar penelitian menggunakan data seperti rata-rata nilai kumulatif (CGPA) dan penilaian internal sebagai *dataset*. Untuk teknik prediksi, metode klasifikasi, terutama *Neural Network* dan *Decision Tree*, sering digunakan dalam bidang *Data Mining* di pendidikan. Kesimpulannya, meta-analisis prediksi kinerja mahasiswa dapat mendorong penelitian lebih lanjut dan penerapan teknik-teknik ini di lingkungan pendidikan. Hal

					ini diharapkan dapat membantu sistem pendidikan untuk memantau kinerja mahasiswa secara lebih sistematis dan efisien.
8.	Komparasi Metode Apriori dan *FP-Growth Data Mining* Untuk Mengetahui Pola Penjualan.	Neni Purwati1*), Yogi Pedliyansah2 , Hendra Kurniawan3 , Sri Karnila4 , Riko Herwanto5 / 2023	Tujuan penelitian ini adalah untuk mengetahui pola penjualan produk terlaris dan untuk meningkatkan kuantitas penjualan produk parfum.	Metode penelitian yang digunakan pada penelitian ini adalah proses *KDD (Knowledge Discovery in Database)*.	Penelitian ini menghasilkan pola frekuensi tinggi untuk *itemsets* dengan *minimum support* 20%, menghasilkan produk teratas: Jo Malone (82,49%), Zarra (28,25%), dan Zwitsal (20,34%). Aturan asosiasi yang terbentuk dengan Min. Supp 20% dan Min. Conf 80% menghasilkan kombinasi 2 *itemsets* (Jo Malone dan Zarra) serta 3 itemsets (Jo Malone, Zarra, dan Baccarte). Kedua kombinasi ini valid dan kuat, terbukti dengan nilai *lift* lebih besar dari 1, sehingga aturan asosiasi ini dapat diterapkan.
9.	Penerapan Algoritma *K-Means Clustering* Untuk Pengelompokan Universitas	Faisal Dikarya, 2Sita Muharni / 2022	bertujuan untuk mengelompokkan universitas terbaik menjadi 3 *cluster* yaitu	metode algoritma *K-Means* untuk mengelompokkan universitas terbaik dari	Pengujian dari penelitian ini dilakukan iterasi *clustering* dengan data 2000 universitas teratas di dunia terjadi sebanyak 10 kali

	Terbaik Di Dunia.		*cluster* tinggi, sedang, dan rendah menggunakan 4 atribut antara lain *world rank*, *institution*, *country*, dan *score* dari 2000 data universitas teratas didunia.	banyaknya universitas.	terdapat tiga *cluster* yaitu *cluster* rendah, *cluster* sedang, dan *cluster* tinggi.
10.	Pemanfaatan *K Means Clustering* dalam Pengelompokan Judul Skripsi.	Nisar1), Wasilah2)Haris Kusumajaya3) / 2022	Penelitian ini bertujuan untuk mengelompokkan data skripsi mahasiswa program studi teknik informatika IIB Darmajaya. Pengelompokan dilakukan dengan menggunakan algoritma. *K-Means Clustering*.	Menggunakan metode *clustering K Means*.	Penelitian ini menunjukkan bahwa algoritma *K-Means* efektif dalam mengelompokkan skripsi dengan membagi data ke dalam k *cluster* yang ditentukan, menggunakan perhitungan jarak untuk mengukur kemiripan antar data. *K-Means* dalam Data Mining mampu mengelompokkan data besar dengan cepat, mempercepat proses pengelompokan.

Berdasarkan berbagai penelitian, metode *K-Means Clustering* banyak digunakan dalam berbagai bidang, termasuk kesehatan, pendidikan, pemasaran, dan keamanan. Algoritma ini terbukti efektif dalam mengelompokkan data, baik untuk identifikasi penyakit, strategi pemasaran, klasifikasi akademik, maupun pemetaan wilayah rawan kriminalitas. Beberapa penelitian juga membandingkan *K-Means* dengan metode lain, seperti *K-Medoids,* untuk mengevaluasi efektivitasnya. Secara keseluruhan, *K-Means* terbukti efisien dalam menangani *dataset* dengan jumlah besar dan menghasilkan pengelompokan yang akurat, meskipun efektivitasnya dapat bervariasi tergantung pada ukuran dan karakteristik *dataset* yang digunakan.