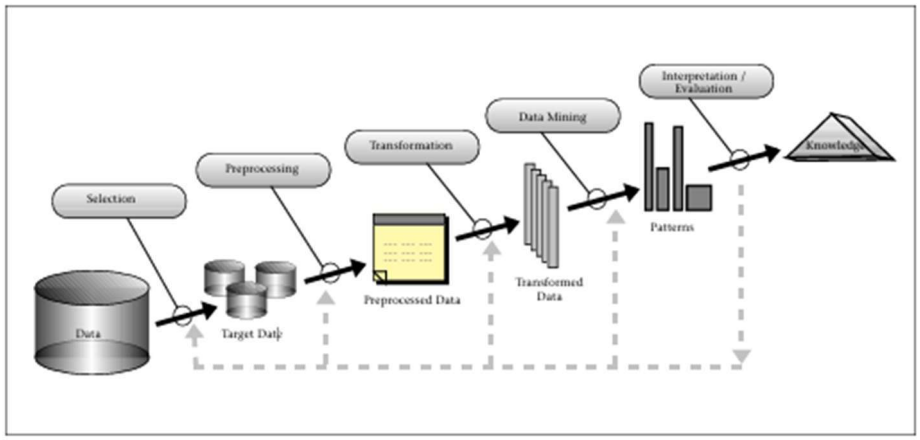
BAB II
TINJAUAN PUSTAKA

2.1. Data *Mining*

Data *mining* adalah sebuah teknik untuk mengidentifikasi suatu informasi secara menyeluruh terkait pada berbagai *dataset* yang besar untuk mendapat pengetahuan baru (R, Anggraeni, & Enri, 2022). Data *mining* dapat dijelaskan sebagai pengolahan untuk memperoleh informasi dengan menggunakan teknik matematika, statistik, *machine learning* dan *artificial intelligence* yang terikat dengan bermacam data *warehouse* atau *database* berukuran besar yang berguna untuk mengidentifikasi pengetahuan (Muttaqin & Defriani, 2020). Data *mining* bermanfaat untuk mengolah data mentah menjadi kumpulan informasi yang dapat dijadikan sebagai penunjang keputusan secara efektif (Sari, Sudewa, Lestari, & Jaya, 2020). Data *mining* adalah teknologi untuk menyaring data yang relevan dari sebuah *database* dengan menggunakan beberapa teknik dan algoritma seperti *Associations, Clustering* dan *Classification* (Pareek & Bhari, 2020).

Banyak orang yang menggunakan istilah *Knowledge Discovery in Database* atau KDD untuk data mining. Tahapan proses dalam penggunaan data mining merupakan salah satu dari rangkaian *Knowledge Discovery in Database* (KDD). KDD adalah sebuah kegiatan untuk mengekstrak suatu pengetahuan dari sekumpulan data, hasil data mining kemudian diubah menjadi informasi yang mudah dipahami (Afdal & Rosadi, 2019). Dibawah merupakan gambar untuk tahap-tahap data mining :



Gambar 2. 1 KDD

Sumber : *From Data Mining to Knowledge Discovery in Databases* (Fayyad, Piatetsky-Shapiro, & Smyth , 1996)

Terdapat beberapa langkah yang dilakukan dalam data *mining* yang merupakan salah satu rangkaian *Knowledge Discovery in Database* (KDD), yaitu :

1. Data *Selection* adalah menciptakan himpunan data target dengan memilih himpunan data dan hasilnya disimpan dalam suatu berkas yang terpisah dari basis data operasional.
2. *Pre-processing* dan *Cleaning* Data adalah proses penghapusan *noise* dengan menghilangkan duplikat data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data. Normalisasi juga dilakukan dalam tahap *pre-processing*. Normalisasi adalah proses mengubah skala atau rentang nilai dari data sehingga nilainya dapat dibandingkan. atau digunakan dengan lebih efisien. Normalisasi data terjadi pada data numerik untuk mengurangi perbedaan skala yang mungkin ada antar atribut.
3. *Transformation* adalah proses integrasi pada data yang telah dipilih sehingga data sesuai untuk proses data *mining*.
4. Data *Mining* adalah pemilihan teknik, metode atau algoritma yang tepat dan sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation* atau *Evaluation* adalah penerjemahan pola-pola yang dihasilkan dari data *mining* dalam bentuk yang mudah dimengerti. Tahap ini akan memeriksa apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang sudah ada sebelumnya.

Secara garis besar, data *mining* dapat dikelompokkan menjadi 2 kategori strategi utama (Arhami & Nasir, 2020), yaitu:

1. *Descriptive mining*, yaitu proses menemukan keteraturan data dan memperlihatkan pola yang dihasilkan atau dengan kata lain, analisis ini digunakan untuk menemukan subkelompok yang menarik pada data utama. *Descriptive mining* berfokus untuk meringkas dan mengkonversi data menjadi informasi yang bermanfaat dalam pelaporan dan pemantauan. Metode yang termasuk *descriptive mining* adalah *Clustering, Summarization, Asosiation Rules*, *Correlations, Characterization, Discrimination,* dan *Sequence Discovery*.
2. *Predictive mining*, yaitu proses menemukan pola dari data dengan menggunakan beberapa *variable* lain untuk memprediksi hasilnya di masa depan. Metode yang termasuk *predictive mining* adalah *Classification, Regression, Time Series Analysis, Prediction, Outlier Analysis,* dan *Evolution Analysis*.

Data *mining* memiliki beberapa tujuan yaitu *Explanatory* atau sebagai sarana menjelaskan, *Confirmatory* atau sebagai sarana konfirmasi, dan *Exploratory* sebagai saranan eksplorasi. *Explonatory* berarti data *mining* digunakan sebagai sarana untuk menjelaskan suatu kondisi penelitian. *Confirmatory* berarti data *mining* digunakan sebagai sarana untuk mengkonfirmasi suatu pernyataan atau

hipotesis yang telah ada. *Exploratory* berarti data *mining* digunakan sebagai sarana untuk menganalisis pola baru yang sebelumnya tidak diketahui (Dena, Rifqo, Reswan, & Mahfuzi, 2023).

Secara umum, beberapa metode yang digunakan dalam data *mining* (Dena, Rifqo, Reswan, & Mahfuzi, 2023), adalah:

1. *Association*, merupakan pendekatan yang berlandaskan aturan untuk mengidentifikasi hubungan dan keterkaitan antar variabel dalam suatu kumpulan data. Metode ini melibatkan pernyataan sederhana berbentuk "jika atau maka".
2. *Classification* merupakan proses untuk meramalkan kategori dari suatu objek.
3. *Regression* adalah metode yang menggambarkan variabel dependen melalui proses analisis variabel independen. Sebagai contoh, memperkirakan jumlah penjualan sebuah produk berdasarkan hubungan antara harga produk dan rata-rata pendapatan konsumen.
4. *Clustering* digunakan untuk mengelompokkan data ke dalam beberapa kategori berdasarkan kesamaan fitur yang dimilikinya. Salah satu contohnya adalah Segmentasi Pelanggan. Metode ini mengelompokkan pelanggan ke dalam beberapa kategori berdasar tingkat kesamaannya.

2.2. Metode Clustering

Clustering adalah pengelompokan satu set objek data ke dalam beberapa kelompok atau *cluster* sehingga objek dalam sebuah *cluster* memiliki jumlah kemiripan yang tinggi, tetapi sangat berbeda dengan objek di *cluster* lain. Ketidak miripan dan kesamaan dinilai berdasarkan nilai atribut yang menggambarkan objek

dan sering melihat prilaku jarak *clustering* sebagai alat data *mining* dapat diterapkan pada berbagai bidang (Azhara, Nugraha, & Piqri, 2021). Algoritma *clustering* terdiri dari dua bagian yaitu secara *hirarkis* dan secara *partitional*. Algoritma *hirarkis* menemukan *cluster* secara berurutan dimana *cluster* ditetapkan sebelumnya, sedangkan algoritma *partitional* menentukan semua kelompok pada waktu tertentu. *Clustering* juga bisa dikatakan suatu proses dimana membagi pola data menjadi beberapa jumlah data set sehingga akan membentuk pola yang serupa dan dikelompokkan pada *cluster* yang sama dan memisahkan diri dengan membentuk pola yang berbeda di *cluster* yang berbeda (Merliana, Ernawati, & Santoso, 2015). Pada data *mining* ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering* atau disebut dengan algoritma *K-Means* (Dwiarni & Setiyono, 2019).

2.3. Algoritma Metode *K-Means*

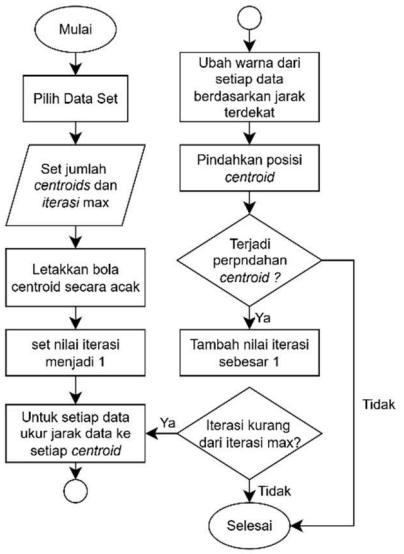
Algoritma *K-means* merupakan metode *clustering* yang paling sederhana dan umum. Hal ini dikarenakan *K-Means* mempunyai kemampuan mengelompokkan data dalam jumlah cukup besar dengan waktu komputasi yang cepat dan efisien (Dwiarni & Setiyono, 2019). Algoritma *K-Means* membagi data ke dalam k kelompok yang sudah ditentukan sebelumnya, di mana k adalah jumlah kelompok yang diinginkan. Tahapan awal dilakukan dengan memilih secara acak k titik pusat kelompok (*centroid*) di dalam ruang data, kemudian mengelompokkan setiap titik data ke dalam kelompok yang memiliki *centroid* terdekat. Kemudian, pusat dari setiap kelompok dihitung ulang dengan cara mengambil rata-rata dari titik data yang ada dalam kelompok itu, dan langkah ini dilakukan berulang kali

sampai tidak ada perubahan lagi dalam pemindahan titik data ke kelompok atau sampai batas iterasi yang telah ditetapkan tercapai. *K-Means Clustering* sering dimanfaatkan dalam pemisahan pelanggan, pengelompokan dokumen, analisis gambar, dan area lain karena kecepatan serta kemampuannya untuk skala, meskipun hasil yang diperoleh bisa terpengaruh oleh penempatan awal *centroid* (Hendrastuty, 2024).

Kelebihan dari metode *K-Means* adalah mampu mengelompokkan data besar dengan sangat cepat, sedangkan kekurangan dari metode *K-Means* adalah banyaknya gerombol harus ditentukan sebelumnya (Lathifaturrahmah, 2014). *K-Means* adalah dalam mengelompokkan data berdasarkan kesamaan atau *similarity*. *K-Means* memiliki kelebihan, diantaranya:

1. Sederhana dan konvergensi yang cepat,
2. Banyak digunakan dalam dataset besar karena efisien secara komputasi,
3. Mudah diimplementasikan dan hasilnya yang mudah diinterpretasi, dll.

K-Means juga memiliki kekurangan, salah satu diantaranya adalah peka terhadap inisialisasi yang berarti bahwa hasilnya sangat bergantung pada pusat cluster (*centroid*) awal yang dipilih secara acak (*random*) (Mutrofin, Wicaksono, & Murtadho, 2023). *Flowchart* dari algoritma *K-Means* seperti pada gambar 2.2.



Gambar 2. 2 *Flowchart* Algoritma *K-Means*
Sumber : *Visualisasi Algoritma sebagai Sarana Pembelajaran K-Means Clustering* (Suryadibrata & Young, 2020)

Langkah-langkah dalam menggunakan algoritma *K-Means* (Metisen & Sari, 2015), sebagai berikut:

1. Tentukan *k* sebagai jumlah *cluster* yang akan dibentuk.
2. Tentukan *k Centroid* (titik pusat *cluster*) awal secara random/acak.

$$v = \frac{\sum_{i=0}^{n} x_i}{n} \quad ; i = 1,2,3, \ldots . n \qquad (1)$$

Dimana;
v = *centroid* pada *cluster*
x_i = objek ke-i
n = banyaknya objek/jumlah objek yang menjadi anggota *cluster*

3. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Untuk menghitung jarak antara objek dengan *centroid* dapat menggunakan *Euclidian Distance*.

$$d(x,y) = |x-y| = \sqrt{\sum_{i=0}^{n}(x_i} - y_i)^2 \; ; i \qquad = 1,2,3, \ldots n \qquad (2)$$

Dimana;
x_i = objek x ke - i
y_i = daya y ke – i
n = banyaknya objek

4. Alokasikan masing-masing objek ke dalam *centroid* yang paling dekat.
5. Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan (1).
6. Ulangi langkah 3 jika posisi *centroid* baru tidak sama.

2.4. Metode *Elbow*

Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik (Merliana, Ernawati, & Santoso, 2015). Metode *Elbow* digunakan untuk menentukan jumlah *cluster* dari *dataset* sehingga total dalam variasi *cluster* dapat diminimalkan, penentuan jumlah *cluster* dimulai dari *k=2* dan terus bertambah di setiap langkah dengan ditambah 1 pada nilai *k* (Marisa, Ahmad, Yusof, Fachrudin, & Aziz, 2019). Berikut rumus untuk mendapatkan perbandingannya dengan menghitung SSE (*Sum Of Square Error*) dari setiap nilai *cluster*.

$$SSE = \sum_{K=1}^{K} \sum_{X_I \in S_K} \|X_i C_k\|_2^2 \qquad (3)$$

Keterangan:
k = jumlah *cluster*
X_i = nilai atribut data ke-i
C_k = nilai atribut titik pusat *cluster* ke-i

Langkah-langkah metode *Elbow* untuk menentukan nilai *k* sebagai berikut:

1. Inisialisasi awal nilai *k*.
2. Menaikkan nilai *k*.
3. Menghitung hasil *sum of square error* dari setiap nilai *k*.
4. Melihat hasil *sum of square error* dari nilai *k* yang turun secara drastis.
5. Menetapkan nilai *k* yang berbentuk siku.

2.5. Python

Python merupakan bahasa pemrograman yang bersifat dinamis dan berkelas tinggi yang mudah dipahami tetapi tetap tangguh dan sangat fleksibel, yang menjadikan bahasa pemrograman ini menarik untuk pengembangan aplikasi. Sintaks dan tipe data *Python* sangat adaptif, dan karena sifat interpretatifnya, bahasa ini sangat cocok untuk penulisan skrip dan pengembangan aplikasi secara cepat. *Python* mendukung berbagai pola pemrograman, termasuk pendekatan pemrograman berorientasi objek, imperatif, fungsional, serta prosedural. Keuntungan menggunakan bahasa pemrograman *Python* yaitu penggunaan fitur yang mudah dipelajari dan digunakan, memiliki bahasa yang ekspresif, dan memiliki kemampuan untuk menyelesaikan kode yang rumit hanya dengan sedikit baris kode. Cukup satu baris yang diperlukan untuk menjalankan program, yang membuat proses *debugging* menjadi lebih sederhana, fleksibel, dan mudah (Suharto, 2023).

2.6. Google *Collaboratory* atau Google *Colab*

Google *Collaboratory* atau yang lebih dikenal sebagai Google *Colab*, merupakan layanan berbasis *cloud* yang memungkinkan pengguna untuk menulis, menjalankan, dan berbagi kode *Python* melalui browser internet. Layanan ini

ditujukan untuk analis, pengembang, peneliti, dan pendidik yang berfokus pada *data science* dan pembelajaran mesin dengan menawarkan lingkungan komputasi yang dapat diakses dengan mudah dan fleksibel tanpa mengeluarkan biaya (Febrywinata, 2024). Keuntungan menggunakan Google *Colab* mencakup kemampuan untuk bekerja bersama secara *online* dengan *user* lain, menyediakan akses gratis ke GPU berkualitas tinggi (GPUTesla, RAM12GB, DISK300GB), serta memberi fleksibilitas dalam menjalankan program pembelajaran mendalam. Google *Colab* tidak memerlukan pengaturan tambahan karena memanfaatkan teknologi komputasi awan, dan mudah diintegrasikan dengan Google *Drive* dan GitHub (Hidayat, Muttaqin, & Irmayanti, 2024).

2.7. Penerapan *K-Means* untuk *Clustering* Wisata Terfavorit

Pada penelitian yang dilakukan oleh Sarah Astiti, Rika Harman, dan Darmansah yang berjudul "Pengelompokan Destinasi Wisata di Batam Berdasarkan Daya Tarik dan Fasilitas Menggunakan Metode *K-Means Clustering*" menghasilkan 3 *cluster* utama berdasarkan daya tarik dan fasilitasnya yaitu tinggi, menengah dan rendah. Adapun kelompok tinggi adalah *cluster_2* dengan jumlah tiga destinasi, kelompok menengah adalah *cluster_1* dengan satu destinasi dan rendah adalah *cluster_0* dengan enam destinasi wisata. Kelebihan pada penelitian ini yaitu penggunaan *K-Means clustering* yang efektif dalam mengelompokkan destinasi berdasarkan karakteristiknya. Kekurangan pada penelitian ini yaitu penggunaan data yang terbatas pada parameter tertentu, sehingga mungkin belum mencakup semua aspek penting (Astiti, Harman, & Darmansah, 2024).

Penelitian yang dilakukan oleh Anggun Helia, Mulyawan, Cep Lukman Rohmat, dan Fathurrohman yang berjudul "Analisis Pengelompokan Daya Tarik

Obyek Wisata Berdasarkan Jenisnya Menggunakan Metode *K-Means* pada Data Pemprov Jabar" berhasil mengelompokkan daya tarik objek wisata di Jawa Barat menggunakan algoritma *K-Means* menjadi 4 cluster dengan nilai DBI terbaik 0.146. *Cluster* terdiri dari objek wisata tingkat rendah (16 kabupaten), sedang (9 kabupaten), dan tinggi (2 wilayah: Bandung dan Bogor). Analisis ini membantu memetakan potensi wisata untuk strategi pengembangan. Kelebihan pada penelitian ini yaitu menggunakan data aktual dan metode *K-Means* yang efektif. Kekurangan pada penelitian ini yaitu atribut data yang terbatas, sehingga belum mencakup seluruh faktor penting (Helia, Mulyawan, Rohmat, & Fathurrohman, 2024).

Penelitian yang dilakukan oleh Andhika Utama, Wilda Imama Sabilla, Rokhimatul Wakhidah yang berjudul "Sistem Rekomendasi Tempat Wisata di Malang Raya Menggunakan Metode *K-Means Clustering*" menghasilkan pengembangkan sistem rekomendasi tempat wisata di Malang Raya menggunakan metode *K-Means Clustering*. Objek wisata dikelompokkan berdasarkan lokasi, jumlah wisatawan, *rating*, harga tiket, fasilitas, dan akses jalan. Hasilnya menunjukkan cluster optimal berbeda untuk setiap wilayah: 5 *cluster* di Kabupaten Malang, 5 *cluster* di Kota Malang, dan 4 *cluster* di Kota Batu. Sistem ini membantu wisatawan memilih destinasi yang sesuai. Kelebihan dari penelitian ini yaitu menggunakan metode *K-Means* yang efisien dan terukur serta mengintegrasikan *Silhouette Coefficient* untuk validasi *cluster*. Kekurangan dari penelitian ini yaitu data yang terbatas pada tahun tertentu (2019), sehingga kurang mencakup tren terkini (Utama, Sabilla, & Wakhidah, 2024).

Dengan studi pendahuluan yang dilakukan dan mengamati beberapa penelitian terkait yang relevan mengenai *clustering* wisata yang menggunakan algoritma *K-Means*, penelitian yang diambil memiliki kelebihan yaitu menggunakan data dengan atribut tujuan paket wisata, harga paket wisata, dan jumlah wisatawan yang di ambil dari data *Guestlist* dan sentimen komentar maupun *Direct Messenger* Instagram Malang Gateway. Selain itu, juga dilakukan *clustering* berdasarkan data tahun 2023 dan 2024. Selanjutnya penentuan nilai *k* dilakukan dengan menggunakan metode *Elbow* serta perhitungan kualitas *clustering* yang dihasilkan menggunakan metode *Silhouette Coefficient*.

2.8. Evaluasi *Clustering*

Dalam mengavaluasi hasil pengelompokan dapat dilakukan dengan validasi *cluster*. Validasi *cluster* berguna untuk mengukur tingkat kebaikan dari *cluster* yang terbentuk, salah satu internal *validation index* adalah *Silhouette Coefficient* (Widyadhan, Hastuti, Kharisudin, & Fauzi, 2021). *Silhouette Coefficient* digunakan pada tahap evaluasi ini dengan fungsi untuk menguji kualitas dari *cluster* yang telah dihasilkan. Selanjutnya dapat digunakan mengambil keputusan untuk menentukan digunakan atau tidaknya hasil dari data *mining* tersebut (R, Anggraeni, & Enri, 2022).

Silhouette Coefficient memiliki nilai dengan rentang -1 hingga 1, dengan ketentuan *cluster* yang semakin baik ketika nilai dari rata-rata yang dihasilkan dekat dengan rentang nilai 1 dan buruk apabila dekat dengan rentang -1. Rentang nilai ini dipergunakan untuk menunjukkan korelasi kemiripan dari data yang telah dikelompokkan ke dalam *cluster* (R, Anggraeni, & Enri, 2022).

Tahapan perhitungan *Silhouette Coefficient* adalah sebagai berikut (Handoyo, M, & Nasution, 2014):

1. Hitung rata-rata jarak dari suatu dokumen, misalkan I dengan semua dokumen lain yang berada dalam satu *cluster*.

$$a(i) = \frac{1}{|A| - 1} \sum j \ \in A, j \neq i \ d \ (i, j) \tag{4}$$

   Dengan j adalah dokumen lain dalam satu *cluster* A dan d(I,j) adalah jarak antara dokumen I dengan j.

2. Hitung rata-rata jarak dari dokumen i tersebut dengan semua dokumen di *cluster* lain, dan diambil nilai terkecilnya.

$$d(i, C) = \frac{1}{|A|} \sum j \ \in C \ d \ (i, j) \tag{5}$$

   Dengan d(i,C) adalah jarak rata-rata dokumen I dengan semua objek pada *cluster* lain C dimana $A \neq C$.

$$b(i) = min \ C \ \neq A \ d \ (i, C) \tag{6}$$

3. Nilai *Silhouette Coefficient* nya adalah:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{7}$$