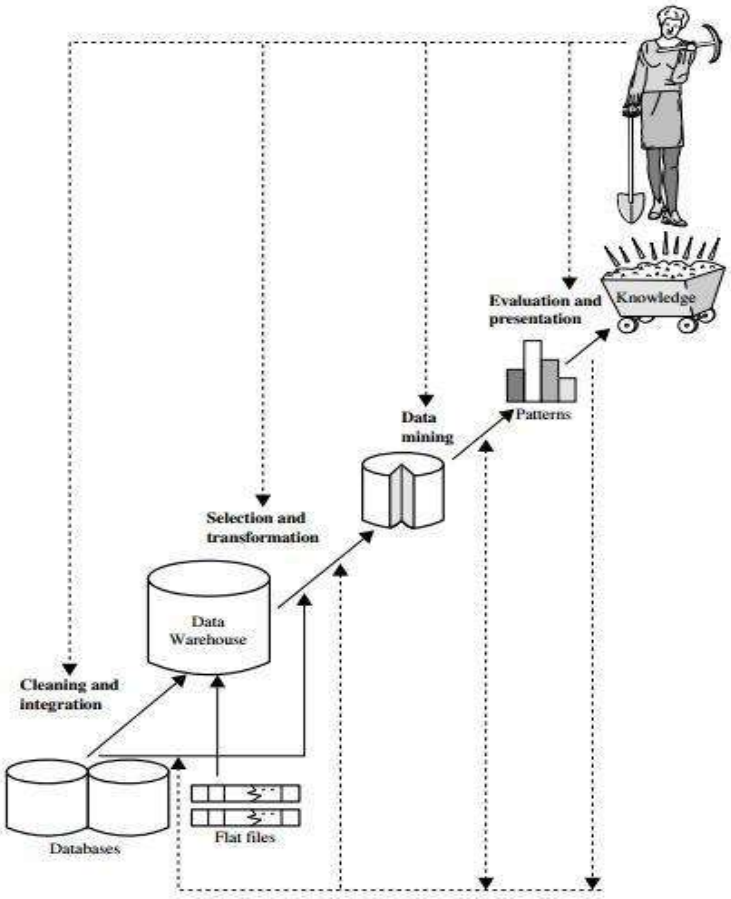
BAB II

TINJAUAN PUSTAKA

2.1 *Knowledge Discovery of Database (KDD)*

Knowledge Discovery of Database (KDD) yaitu proses mengolah informasi yang bermanfaat serta belum diketahui sebelumnya dari kumpulan data. Proses pada KDD ditampilkan pada gambar berikut:



Gambar 2. 1 Tahapan *Knowledge Discovery Database (KDD)*
Sumber: (Gustientiedina et al., 2019)

Berikut urutan proses KDD (Gustientiedina et al., 2019):

1. Pembersihan Data (*Data Cleaning*)

   Data yang tidak diperlukan akan dilakukan pembersihan atau dihilangkan. Pada tahapan *Cleaning,* data yang tersisa adalah atribut yang diperlukan untuk proses pengolahan selanjutnya.

2. Integrasi Data (*Data Integration*)

   Kombinasi beberapa sumber data disebut Integrase Data. Tahap ini bertugas melakukan penggabungan data untuk dibentuk penyimpanan data yang koheren.

3. Seleksi Data (*Data Selection*)

   Seleksi data adalah proses pemilihan data yang memiliki kaitan dalam penelitian. Dilakukan pengurangan representasi dari data untuk menghilangkan informasi yang tidak perlu. Ini meliputi proses mengurangi atributt dan mengkompresi data.

4. Transformasi Data (*Data Transformation*)

   Dengan meringkas atau menggabungkan operasi, data dikonsolidasikan menjadi bentuk yang cocok untuk dilakukan penambangan data.

5. Penambangan Data (*Data Mining*)

   Komponen kunci dari proses KDD adalah penambangan data. Dengan menggunakan metode tertentu, data mining merupakan proses untuk mencari pola atau informasi menarik pada data yang dipilih.

6. Evaluasi Pola (*Pattern Evaluation*)

Pengetahuan dasar tentang langkah-langkah adalah fokus dari tahap ini, yaitu menentukan apakah polanya sudah benar.

7. Representasi Pengetahuan (*Knowledge Presentation*)

   Pengguna diperlihatkan tentang pengetahuan visual dari penemuan. pada tahapan ini berguna membantu mereka memahami dari hasil penambangan data dan mempelajari lebih lanjut tentang metode tersebut.

2.2 Data *Mining*

Penambangan data merupakan proses penggunaan teknologi pengenalan pola statistik dan matematis untuk menemukan korelasi, pola, dan tren baru yang berarti disimpannya data dalam jumlah besar (Annur, 2019). Proses penggalian pengetahuan dan pola dari sejumlah besar data dikenal sebagai data mining. Elemen dasar yang diperlukan untuk proses penambangan data adalah sumber data. Data dalam penelitian ini disimpan dalam database relasional dengan tabel yang berisi bebrapa kolom dan baris yang menampilkan atribut tertentu.

Ada tiga tehnik yang terkenal pada *data mining,* yaitu:

1. Aturan Asosiatif (*Association rule mining*)

   Teknik penambangan yang disebut aturan asosiatif digunakan untuk menemukan aturan asosiatif antara satu set item dengan item yang lain. Untuk menunjukkan bagaimana objek data terkait, aturan asosiatif digunakan. Dua langkah berbeda membuat aturan asosiatif: Temukan dukungan minimum yang digunakan untuk menentukan keseluruhan pengulangan itemset dalam database terlebih dahulu.

Kedua, mengulangi kumpulan item dan menetapkan batas kepercayaan minimum aturan.

2. Klasifikasi (*Classification*)

   prosedur menemukan model yang berfungsi mendeskripsikan atau membedakan konsep *class* data untuk memperkirakan kelas dari objek yang tidak memiliki label. Jaringan saraf (*neural network*), pohon keputusan *(decision tree),* atau aturan jika-maka *(if-then)* bisa menjadi model itu sendiri.

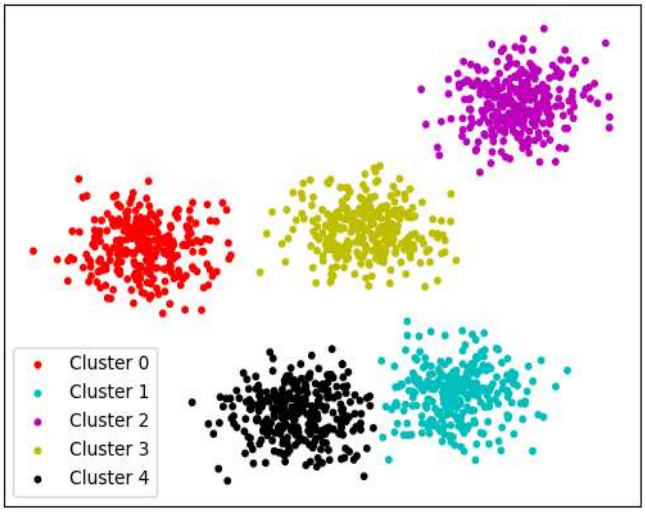
3. Pengelompokan (*Clustering*)

   Tanpa didasarkan pada kelas data tertentu, clustering mengelompokkan data. Kelas data yang tidak diketahui dapat diberi label dengan bantuan pengelompokan. Tujuan clustering adalah untuk meminimalkan kesamaan antar kelas sekaligus memaksimalkan kesamaan antara anggota satu kelas(Wahyudi et al., 2020).

2.3 Metode Data *Mining (Clustering)*

Metode *clustering* digunakan dalam penelitian ini. Proses pengelompokan sekumpulan objek data menjadi beberapa kelompok atau *cluster* sehingga objek dalam satu cluster memiliki banyak kesamaan dengan yang ada di *cluster* lain tetapi sangat berbeda. Evaluasi kesamaan dan ketidaksamaan bergantung pada jumlah atribut yang menggambarkan objek dan seringkali termasuk perlakuan jarak. Bidang seperti biologi, keamanan, intelijen bisnis, dan pencarian web, dapat memperoleh manfaat dari penggunaan pengelompokan sebagai alat penambangan data (Priyadi et al., 2019).

Tujuan *clustering* yaitu untuk meminimalkan kesamaan tiap cluster sambil memaksimalkan kesamaan antara anggota satu kelas. Data dengan banyak atribut yang dipetakan sebagai ruang multidimensi dapat digunakan untuk pengelompokan. Pada Gambar 2.2, sebuah bidang dua dimensi yang menggambarkan lokasi pelanggan toko dapat digunakan sebagai ilustrasi pengelompokan, dengan pusat setiap *cluster* ditandai dengan tanda positif (+). Jika datanya numerik, algoritma-algoritma dalam tehnik *clustering* memakai perhitungan jarak minimum untuk menentukan seberapa mirip dua set data (Nasir, 2021).



Gambar 2. 2 Visualisasi *Cluster*
Sumber : (Nasir, 2021).

Secara umum *clustering*bisa dikelompokkan menjadi metode-metode sebagai berikut:

1. Metode partisi *(Partitioning Method)*

   Metode ini bekerja dengan membagi database dari n objek atau data Tupelo menjadi k partisi, yang mana setiap partisi merepresentasikan suatu

cluster dan k lebih besar atau sama dengan n. Prasyaratnya adalah sebagai berikut: (1) Setidaknya satu objek harus ada di setiap grup, dan (2) setiap objek hanya boleh memiliki satu grup. Awalnya basis data dipartisi menjadi k partisi.

Setelah itu, menggunakan metode relokasi iteratif dan berusaha memperbaiki perpecahan dengan berpindah kelompok. Kriteria umum untuk partisi yang baik adalah objek dalam satu cluster sangat mirip satu sama lain.

Daftar lengkap semua kemungkinan partisi akan diperlukan untuk mencapai optimalitas global dalam pengelompokan berbasis partisi. Algoritma k-means, di mana setiap segmen diwakili oleh nilai rata-rata objek dalam cluster, dan algoritma k-medoids, di mana setiap segmen diwakili oleh salah satu objek yang terletak dekat dengan pusat massa, adalah dua di antaranya. metode heuristik yang paling umum. Untuk database berukuran sedang, teknik pengelompokan heuristik ini bekerja dengan baik untuk menemukan kelompok bola kecil (Wahyudi et al., 2020).

2. Metode Hirarki (*Hierarchi Method*)

Sebuah metode hirarkis menghasilkan satu set objek data menjadi dekomposisi hirarkis. Metode hierarki bersifat agglomeratif atau divisif, tergantung pada bagaimana hierarki dipecah. *Bottom-up* dan *top-down* adalah dua pendekatan yang membentuk pendekatan agglomerative. Metode bottom-up dimulai dengan setiap objek membentuk kelompoknya sendiri. secara

progresif menggabungkan objek atau grup yang dekat satu sama lain, baik hingga terjadi kondisi penghentian atau hingga semua grup digabungkan menjadi satu (tingkat hierarki tertinggi). Di sisi lain, *metode top-down* dimulai dengan kelompok objek yang lebih kecil yang termasuk dalam cluster yang sama dan berjalan ke bawah ke semua objek atau hingga kondisi terminasi terjadi.

3. Metode Berbasis Kerapatan (*Density Based Method*)

Sebagian besar metode *cluster* mempartisi objek berdasarkan jarak antara objek. Metode semacam itu hanya dapat menemukan *cluster* berbentuk bola dan mengalami kesulitan dalam menemukan *cluster* berbentuk sembarang. Metode pengelompokan lain telah dikembangkan berdasarkan gagasan kerapatan. Ide secara umumnya adalah terus tumbuhnya *cluster* yang diberikan selama densitas (jumlah objek atau pusat massa) di "*neighborhood*(lingkungan)" melebihi ambang batas tertentu, yaitu untuk setiap titik data dalam *cluster* tertentu, lingkungan radius tertentu setidaknya harus memuat minimal jumlah titik. Metode tersebut dapat digunakan untuk menyaring *outlier* dan menemukan bentuk kelompok sembarang.

Beberapa algoritma yang termasuk kedalam metode berbasis kerapatan, yaitu: DBSCAN, OPTICS, DENCLUE. DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) merupakan algoritma yang memperluas wilayah dengan kepadatan yang tinggi ke dalam cluster dan menempatkan *cluster* irregular pada basis data spasial dengan *noise*. Algoritma ini

mendefiniskan *cluster* sebagai kumpulan maksimal dari titik-titik kepadatan yang terkoneksi. OPTICS (Ordering Points To Identify the Clustering Structure) merupakan algoritma pada metode hirarki yang diusulkan untuk mengatasi kesulitan *user* dalam menentukan parameter yang digunakan untukmenemukan *cluster* yang bisa diterima. DENCLUE (*Density Based Clustering*) merupakan algoritma *clustering* yang berdasarkan suatu set fungsidistribusi kerapatan.

4. Metode berbasis *Grid*

   Ruang objek dikuantisasi menggunakan metode berbasis *grid* ke dalam struktur jaringan dari sejumlah sel yang terbatas. Ruang terkuantisasi, atau struktur jaringan, adalah dasar untuk semua operasi pengelompokan. Waktu pemrosesan tercepat adalah manfaat utama dari strategi ini, yang hanya bergantung pada jumlah sel di setiap dimensi ruang terkuantisasi daripada jumlah objek data. Metode berbasis grid menggunakan algoritma berikut:: *STING, Wave Cluster*, dan lainnya. *STING* (*Statistical Information Grid*) merupakan algoritma *clustering* yang bekerja dengan membagi daerah spatial menjadi sel-sel *rectanguleri*. Wave Cluster adalah algoritma pengelompokan yang meringkas data dengan mengidentifikasi struktur grid multidimensi ruang data..

5. Metode Berbasis Model

   Pendekatan berbasis model membangun model untuk setiap kelompok dan memilih salah satu yang paling sesuai dengan data. Dengan membangun fungsi kerapatan yang menggambarkan distribusi spasial titik

data, *algoritme* berbasis model dapat menemukan *cluster.* Hal ini m berdasarkan standar statistik, membawa "*noise*" atau *outlier* ke dalam perhitungan, menghasilkan strategi pengelompokan yang andal. Pendekatan berbasis model mencakup algoritma yaitu COBWEB dan SOM. COBWEB adalah algoritma pembelajaran konseptual yang menggunakan konsep sebagai model dan melakukan analisis probabilitas. untuk *cluster*. SOM (mengorganisir diri berfitur peta) adalah algoritma pengelompokan berdasarkan jaringan saraf yang mengubah data dimensi tinggi menjadi peta fitur 2D atau 3D untuk digunakan dalam visualisasi data.

Secara umum, tujuan dari teknik clustering adalah untuk menemukan kelompok data yang memiliki kesamaan yang sangat tinggi dalam satu kelompok tetapi sangat sedikit kesamaannya dengan kelompok lain.

Algoritma *k-means* adalah algoritma klasik untuk menyelesaikan masalah *clustering*, yang relatif sederhana dan cepat Algoritma *k-means clustering* lebih sering dikenal karena kemampuannya dalam mengelompokkan data dalam jumlah besar dengan cepat dan efisien. Algoritma *k- mean* sangat rawan dipusat-pusat *cluster* awal, karena pusat *cluster* awal diproduksi secara acak. Algoritma *k-means* tidak menjanjikan hasil pengelompokan yang khas. Efisiensi keaslian algoritma *k-mean* sangat bergantung pada titik pusat *cluster* (*centroid*) awal (Sani, 2018).

Langkah kerja dari algoritma *k-means* adalah sebagai berikut :

1. Menanyakan kepada pengguna berapa banyak *k cluster dataset* yang akan

dipartisi.

2. Menetapkan secara acak *k record* yang menjadi lokasi pusat *cluster* awal.
3. Setiap *record* dicari *centroid cluster* terdekatnya. Artinya setiap *centroid cluster* "memiliki" subset dari *record*, sehingga merepresentasikan sebuah partisi dari *dataset*. Didapatkan *k* cluster, $C_1, C_2, \ldots, C_k$.
4. Setiap *k cluster* dicari *centroid*nya dan memperbarui lokasi setiap pusat *cluster* untuk nilai *centroid* baru.
5. Ulangi langkah 3 sampai 5, sampai terjadi konvergensi atau terjadi penghentian.

Algoritma berakhir ketika titik pusat *cluster* tidak lagi berubah. Dengan kata lain, algoritma berakhir ketika dari seluruh *cluster* $C_1, C_2, \ldots, C_k$, semua *record* yang dimiliki oleh masing-masing pusat *cluster* tetap dalam *cluster* itu. Atau, algoritma dapat berhenti ketika beberapa kriteria konvergensi terpenuhi, seperti ada penyusutan yang tidak signifikan dalam jumlah kuadrat *error* (*sum of squared errors*):

Rumus 2. 1 Rumus *SSE (Sum of Squared Errors)*

$$SSE = \sum_{i=1}^{k} \sum_{p \in C} d(p, m_i)^2$$

dimana $p \in C_i$ melambangkan setiap titik dalam *cluster i* dan m_i merupakan pusat*cluster i* .

Berikut contoh permasalahan yang menerapkan algoritma *k-means* untuk menyelesaikannya. Terdapat 4 jenis obat (A, B, C, D) dan setiap jenis obat memiliki dua atribut, seperti yang ditunjukkan pada Tabel 2.1. Tujuan dari

kasusini adalah untuk mengelompokkan jenis obat-obat tersebut menjadi *k*=2 kelompok berdasarkan dua fitur (pH dan indeks berat badan).

Tabel 2. 1 Contoh Kasus *Clustering*

Objek	Atribut 1 (X): indeks berat badan	Atribut 2 (Y): pH
Obat A	1	1
Obat B	2	1
Obat C	4	3
Obat D	5	4

Sumber: (Peneliti, 2023)

Dengan ini, diketahui bahwa objek memiliki dua kelompok obat (*cluster* 1 dan *cluster* 2). Masalahnya sekarang adalah setiap obat berada pada *cluster* yang mana, apakah *cluster* 1 atau *cluster* lainnya. Setiap obat merepresentasikan satu titik dengan dua komponen koordinat. Dasar algoritma *k-means clustering* tergolong sederhana seperti berikut:

Algoritma k-means. merupakan *Algoritma* untuk pemisah, dimana setiap pusat klasterdiwakili dari nilai rata-rata objek didalam klaster.

Input:

- k: jumlah klaster,
- D: satu setdata yang berisi n objek.

Output: satu set k klasterMetode:

(1). Ambil secara *random* k objek dari D sebagai pusat klaster awal;

(2). Ulangi lagi hingga

(3). Kembali menetapkan objek kedalam kelompok terhadap objek yang paling mirip, berdasarkan nilai rata-rata objek didalam klauster;

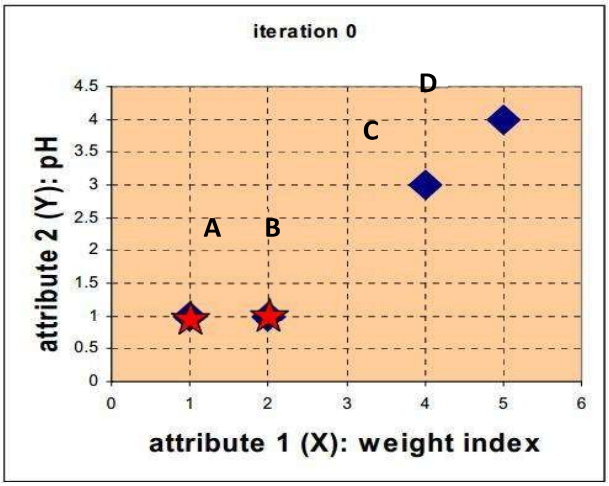
(4). Memperbaharui nilai rata-rata klaster, yaitu dengan menghitung nilai rata-rata objek dalam klaster;

(5). Lakukan proses sampai tidak lagi terjadi perubahan data;

Algoritma *k-means* yang dijelaskan sebagai berikut:

1. Untuk menemukan titik *centroid* awal maka tentukan jumlah *cluster* k pada D.
2. Melakukan penghitungan jarak objek ke *centroid* .
3. Objek dikelompokkan berdasarkan jarak minimum ke *centroid.*
4. Jika terjadi perpindahan atau data tidak stabil, maka dilakukan pencarian *centroid* baru dengan cara menghitung nilai rata-rata dari anggota *cluster*.
5. Setelah melakukan langkah 1-4, jika sudah tidak terjadi perpindahan atau data sudah stabil, maka algoritma selesai.

Dari kasus diatas, jenis obat direpresentasikan oleh satu titik, satu titik tersebut memiliki dua fitur, yang bisa dikatakan koordinat dalam ruang fitur, berikut lebih jelasnya ditunjukkan berikut.



Gambar 2. 3 Ruang Fitur pada Iterasi Awal
Sumber: (Larose, 2005:153)

Contoh penyelesaian akan diuraikan sebagai berikut:

- Iterasi Awal (iterasi 0)

1. Menentukan titik awal *centroid*. Obat A dan obat B disebut *centroid* awal dan dilambangkan dengan c_1 dan c_2 sebagai koordinat *centroid,* maka $c_1 = (1,1)$ dan$c_2 = (2,1)$
2. Jarak *centroid* ke objek mulai ditentukan, yang artinya menghitung jarak antara *centroid cluster* dengan setiap objek. Dapat dilakukan menggunakan rumus jarak *Euclidean* yang selanjutnya akan didapatkan matriks jarak pada iterasi 0 (d^0) berikut:

   Bentuk matrik dari data objek,

$$\begin{matrix} A & B & C & D & \\ \begin{bmatrix} 1 & 2 & 4 & 5 & X \\ 1 & 1 & 3 & 4 & Y \end{bmatrix} \end{matrix}$$

   Bentuk matriks dari jarak yang didapat pada iterasi 0,

$$d^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix}$$

Dengan $c_1 = (1,1)$ kelompok 1
$c_2 = (2,1)$ kelompok 2

Kolom pada setiap matrik jarak melambangkan objek. Baris pertama pada matriks jarak sesuai dengan jarak masing-masing objek menuju *centroid* awal dan baris kedua merupakan jarak setiap objek menuju *centroid* kedua. Misalnya, jarak dari obat *C = (4,3)* ke *centroid* pertama c_1 = *(1,1)* adalah

$\sqrt{(4-1)^2 + (3-1)^2} = 3.61$ dan jarak ke *centroid* kedua c_2 adalah

$\sqrt{(4-2)^2 + (3-1)^2} = 2.83$, dan seterusnya.

3. Langkah berikutnya dilanjutkan dengan pengelompokan objek. Pertama menetapkan keberadaan setiap objek berdasarkan jarak minimum. Dengan demikian, obat A ditetapkan menjadi kelompok 1, obat B ditetapkan kelompok 2, obat C ditetapkan kelompok 2 dan obat-obatan D ditetapkan kelompok 2. Unsur dari kelompok matriks berikut bernilai 1 jika dan hanya jika objek ditugaskan ke kelompok itu.

$$G^0 = \begin{matrix} A & B & C & D \\ \end{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{kelompok 1} \\ \text{kelompok 2} \end{matrix}$$

- Iterasi ke 1

1. Untuk menyelesaikan langkah ini, pertama cari centroidnya. Setelah mengidentifikasi anggota dari masing-masing kelompok, centroid baru dari masing-masing kelompok dihitung dengan menggunakan anggota baru. Karena hanya ada satu anggota di Grup 1, pusat massa ditetapkan., yaitu $c_1 = (1,1)$. Kelompok 2 sekarang terdiri dari tiga anggota, jadi pusat massanya adalah rata-rata dari koordinat mereka.
2. Iterasi 1 pada jarak *centroid* objek. Menghitung jarak setiap objek adalah langkah selanjutnya ke *centroid* baru. Seperti langkah 2, dihasilkan matrik jarak di iterasi ke 1 yaitu:

$$d^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} c_1 = (1,1) \text{ kelompok 1} \\ c_2 = (\frac{11}{3}, \frac{8}{3}) \text{ kelompok 2} \end{matrix}$$

3. Pengelompokan data objek (object clustering), menempatkan objek pada jarak minimum. Dan berdasarkan pada matrik jarak baru, dan memindahkan obat B ke kelompok 1 dan sedangkan semua objek lainnya tetap. Bentuk matriknya seperti berikut:

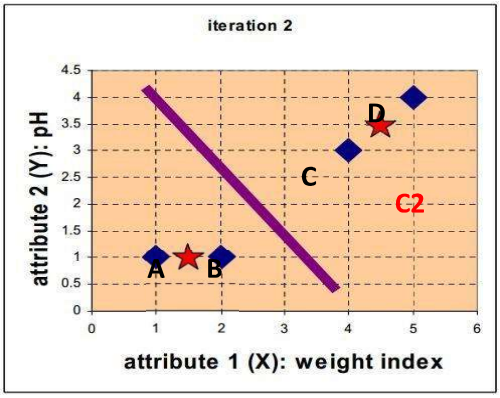
$$G^1 = \begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{matrix} \begin{matrix} \\ \text{kelompok 1} \\ \text{kelompok 2} \end{matrix}$$

- Iterasi ke 2

1. Menghitung jarak titik koordinat *centroid* baru berdasarkan pengelompokan iterasi sebelumnya. Kelompok 1 dan juga kelompok 2, kedua-duanya memiliki jumlah dua anggota, sehingga mendapatkan *centroid* yang baru

$$c^1 = \left(\frac{1+2}{2} \cdot \frac{1+1}{2}\right) = (1\tfrac{1}{2}, 1) \text{ dan } c^2 = \left(\frac{4+5}{2} \cdot \frac{3+4}{2}\right) = (4\tfrac{1}{2}, 3\tfrac{1}{2})$$

Berikut posisi *centroid* baru dan partisi kelompok.



Gambar 2. 4 Ruang Fitur pada Iterasi Ke 2
Sumber: (Larose, 2005:153)

2. Menghitung jarak dari semua objek ke *centroid* baru maka didapatkanmatriks jarak baru pada iterasi ke 2 ini seperti berikut:

$$d^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad c^1 = (1\frac{1}{2}, 1) \text{ Kelompok 1}$$

$$c^2 = (4\frac{1}{2}, 3\frac{1}{2}) \text{ Kelompok 2}$$

3. Pengelompokan objek, menempatkan setiap objek berdasarkan jarak minimum. Pada iterasi kedua ini diperoleh matriks kelompok sebagai berikut:

$$G^2 = \begin{matrix} A & B & C & D \\ \end{matrix}$$
$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} \text{Kelompok 1} \\ \text{Kelompok 2} \end{matrix}$$

Hasil yang diperoleh bahwa $G^1 = G^2$. iterasi ini menunjukkan objek tidak berpindah kelompok lagi. Dengan demikian, perhitungan dari pengelompokan *algoritma k-means* telah stabil dan tidak diperlukan iterasi lanjutan lagi. Sehingga dihasilkan hasil akhir sebagai berikut :

Tabel 2. 2 Hasil Iterasi Terakhir

Objek	Atribut 1 (X): Indeks berat badan	Atribut 2 (Y):pH	Kelompok(hasil)
Obat A	1	1	1
Obat B	2	1	1
Obat C	4	3	2
Obat D	5	4	2

Sumber: (Peneliti, 2023)

2.4 Algoritma *K-Means*

Cluster awal yang dibuat oleh *algoritma K-Means* dipusatkan pada sejumlah anggota populasi yang berbeda. Pada titik ini, pusat klaster dipilih secara acak dari sekumpulan data populasi. *K-Means* menguji data *swarm* pada langkah

berikut untuk mengetahui jarak minimum antara setiap komponen dan setiap pusat *cluster* yang telah ditentukan sebelumnya. Posisi pusat klaster dihitung ulang setiap kali komponen data baru perlu ditetapkan ke pusat klaster tertentu.

Isi buku (Wahyudi et al. 2020: 6) memberikan penegasan bahwa dengan menggunakan algoritma pengelompokan *K-means*, data bisa dibagi menjadi beberapa kelompok berdasarkan seberapa jauh data satu sama lain dalam kelompok-kelompok tersebut. Algoritma menggunakan fungsi untuk membandingkan beberapa kesamaan yang dimiliki dataset. Data akan diurutkan berdasarkan kedekatan kemudian disimpan.

Dibawah ini merupakan langkah-langkah untuk pengelompokkan data (Witanto et al., 2019: 704)

1. Menentukan berapa banyak cluster yang dibutuhkan.
2. Kemudian menentukan pusat *cluster* secara random dan melakukan inisialisasi data agar data mudah diolah.
3. Dalam mennetukan jarak dari objek satu dengan objek yang lain, posisi dari setiap titik data bisa diletakkan pada *cluster* yang terdekat. Dalam tahap ini, *Euclidean* berperan sebagai penghitung jarak dan menentukan perbedaan dan kesamaan data. Dengan rumus sebagai berikut:

Rumus 2. 2 Rumus *Euclidean*

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

dimana:

“$d(x,y)$ = ukuran ketidaksamaan

$x_i = (x_1, x_2, x_3, \dots, x_n)$ adalah ukuran ketidaksamaan

$y_i = (y_1, y_2, y_3, \dots, y_n)$ adalah variabel titik pusat"

4. Rata-rata objek pada cluster digunakan dalam perhitungan pusat cluster. Median bisa juga bisa digunakan dalam menghasilkan perhitungan.
5. Agar proses pengklasteran selesai, maka harus melakukan penghitungan ulang pada jarak diantara objek dan juga pusat *cluster* sampai dengan hasil *cluster* sama atau stabil.

2.5 Software Pendukung

Software Microsoft Excel 2016 dan *Software WEKA 3.9.6* merupakan software pendukung dalam penelitian ini karena diperlukan untuk melakukan proses pengujian data.

2.5.1 *Microsoft Excel 2016*

Gambar 2. 5 *Software Microsoft Excel 2016*
Sumber: (Peneliti, 2023)

Microsoft Excel adalah program pengolah data yang mengotomatiskan manipulasi data dengan memanfaatkan alat yang sudah ada sebelumnya untuk membuat perhitungan dasar, manipulasi data, pembuatan tabel, dan manajemen data grafis, dan tugas-tugas lainnya, agar menjadi lebih sederhana. Dalam bidang-bidang seperti akuntansi, teknik, statistik, dan lain-lain yang membutuhkan

perhitungan cepat dan tepat, *Microsoft Excel* unggul dalam memproses berbagai tipe data.

Pada lokasi penelitian ini digunakan *Microsoft Excel* untuk mengolah data, dan pengetikan manual digunakan untuk meringkas data pengeluaran barang menjadi satu *file*. Setelah itu, tanggal, jumlah barang, proyek, dan nomor transaksi akan dimasukkan dan disusun menjadi kartu stok berdasarkan barang barang.

Pada penelitian ini, *Microsoft Excel* akan digunakan sebagai pengolah data awal untuk membersihkan data, memeriksa kesalahan, menghapus duplikat, dan sebagainya, sehingga *Software WEKA* dapat mengolah data tersebut.

2.5.2 *WEKA 3.9.6*

Gambar 2. 6 Tampilan Utama *WEKA*
Sumber: (Peneliti, 2023)

University of Waikato di New Zealand menciptakan sistem data *mining Knowledge Analysis (WEKA)*, yang menggunakan algoritma data *mining* (Wahyudi et al., 2020). *WEKA* adalah kumpulan algoritma pembelajaran mesin terkait penambangan data. *Algoritma* dapat dipanggil dari kode *Java* itu sendiri atau

diterapkan langsung ke kumpulan data. Pra-pemrosesan data, klasifikasi, regresi, pengelompokan, aturan asosiasi, dan visualisasi semuanya didukung oleh *WEKA*. Strategi pembelajaran mesin baru juga dapat dikembangkan menggunakan *WEKA (www.cs.waikato.ac.nz). WEKA* menawarkan penerapan algoritma pembelajaran yang mudah digunakan untuk kumpulan data. Selain itu, implementasinya mencakup alat untuk pra-pemrosesan dataset, menganalisis klasifikasi yang dihasilkan dan kinerjanya tanpa menulis kode program, mengubah dataset, dan menyediakan skema pembelajaran. *WEKA* dapat digunakan untuk menerapkan teknik pembelajaran pada kumpulan data dan memeriksa hasilnya untuk menyelidiki data lebih lanjut. Sebagai model pembelajaran, juga dapat digunakan untuk memprediksi kasus baru. Aplikasi ini diuji pada sejumlah studi yang berbeda dan membandingkan seberapa metode belajar yang diinginkan dari menu di tampilan utama. Lembar editor objek adalah tempat parameter yang selaras untuk banyak metode dapat ditemukan. Semua classifier dievaluasi menggunakan modul evaluasi yang sama. *WEKA* dapat digunakan untuk menerapkan teknik pembelajaran pada kumpulan data dan memeriksa hasilnya untuk menyelidiki data lebih lanjut. Sebagai model pembelajaran, juga dapat digunakan untuk memprediksi kasus baru. Aplikasi ini diuji pada sejumlah studi berbeda dan membandingkan seberapa baik kinerjanya. Satu studi dipilih untuk digunakan untuk prediksi. Anda dapat memilih metode pembelajaran yang disukai dari menu di tampilan utama. Lembar editor objek adalah tempat parameter yang selaras untuk banyak metode dapat ditemukan. Semua kinerja pengklasifikasi dievaluasi menggunakan modul evaluasi tunggal.

WEKA 3.9.6 memiliki Lima tabmenu utama, yaitu :

1. *Explore*

   Explore berarti pilihan untuk menjelajahi data dan kemudian muntuk dilakukan proses olah data. *Explore* memiliki enam subtab dengan tugas sebagai berikut

   a. *Preprocess*

      Merupakan tab bidang pemilihan dataset dan memodifikasi dengan berbagai macam cara.

   b. *Classify*

      Merupakan tab pelatihan skema yang melaksanakan tugas klasifikasi atau regresi dan juga evaluasinya.

   c. *Cluster* merupakan pembelajaran pengelompokan untuk dijadikan *dataset*.

   d. *Associate* merupakan aturan untuk mengasosiasikan data dan juga evaluasinya.

   e. *Select attributes* merupakan pemilihan dari aspek yang paling relevan dalam *dataset*.

   f. *Visualize* merupakan tampilan plot dari dua dimensi yang berbeda dari data tersebut dan interaksinya.

2. *Experimenter*

   Eksperimen dapat digunakan untuk melakukan perbandingan statistik antar skema. Pengguna *Eksperimen* dapat melakukan eksperimen skala besar, dan proses analisis kinerja dilakukan secara statistik pada hasil yang

diperoleh selama *eksperimen.*

3. *Knowledge Flow*

Knowledge Flow Pengetahuan adalah pemilihan bidang dengan antarmuka seret dan lepas yang menyediakan fitur dasar yang sama seperti penjelajahan. Salah satu salah satu manfaatnya. Pengguna terhubung ke grafik terarah yang memproses dan menganalisis data dalam tampilan utama aliran pengetahuan ini, di mana mereka bisa menyaksikan dimana tata letak kinerja dari proses yang mereka lakukan. Dengan cara yang tidak bisa dijelajahi, bagian ini memberikan penjelasan yang jelas tentang cara kerja data dalam suatu sistem.

4. *Workbench*

Workbench memiliki antarmuka pengguna grafis dan kumpulan alat visualisasi dan algoritma untuk pemodelan prediktif dan analisis data.

5. *Simple CLI*

Simple CLI memberikan tampilan perintah sederhana yang berguna untuk memungkinkan perintah langsung di eksekusi. Selain *explore, experimenter, knowledge flow* pada *WEKA* terdapat juga fungsi dasar yang bisa digunakan secara langsung pada tampilan perintah. Tampilan perintah ini terdapat pada tab *simple CLI*, pada tampilan utama *WEKA* panel *simple CLI* terletak di sebelah kanan bawah.

2.6 Penelitian Terdahulu

Dalam penelitian ini, penelitian terdahulu menjadi acuan bagi penulis agar teori yang digunakan dalam penelitian dapat direplikasi. Penulis dapat

mengidentifikasi penelitian dengan judul yang sama dengan penelitiannya sendiri berdasarkan penelitian sebelumnya. Oleh karena itu, peneliti menggabungkan temuan penelitian sebelumnya sebagai berikut:

1. Hasil Penelitian Inna Alvi Nikmatun, Indra Waspada. (Nikmatun & Waspada, 2019). dari Universitas Diponegoro, E-ISSN: 2549-3108, Vol. 10 No. 2, berjudul ***"Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor"***

   Penelitian ini merupakan penelitian yang menggunakan metode *Algoritma K-Nearest Neighbor.* Penelitian ini memiliki variable yaitu mata kuliah pilihan Dan masa studi mahasiswa.

   Berdasarkan penelitian yang dilakukan oleh peneliti ini, maka ditarik kesimpulan mengenai pengelompokkan masa studi mahasiswa menggunakan algoritma K-Nearest Neighbor adalah sebagai berikut:

   a. Dengan mengacu proses data mining Knowledge Discovery Databases telah dibangun sebuah perangkat lunak yang dapat melakukan klasifikasi masa studi mahasiswa.

   b. Dari enam skenario percobaan yang telah dilakukan diperoleh nilai akurasi tertinggi pada skenario yang menggunakan atribut mata kuliah pilihan yaitu 75.95%.

2. Hasil Penelitian Yunita, Sukrina Herman, Ahsani Takwim, Septian Rheno Widianto, (Sinambela et al., 2020). dari STMIK LIKMI Bandung, E-ISSN: 2656-1735, Vol. 2, No. 2. Berjudul ***"A Study Of Comparing Conceptual And Performance Of Kmeans And Fuzzy C Means Algorithms (Clustering***

Method Of Data Mining) Of Consumer Segmentation" Penelitian ini merupakan penelitian yang menggunakan metode *Kmeans And Fuzzy C Means Algorithms (Clustering Method Of Data Mining).*

Berdasarkan penelitian yang dilakukan oleh peneliti, maka dapat disimpulkan bahwa, Segmentasi konsumen merupakan dasar strategi pemasaran. Untuk mendukung proses pengelompokan hasil konsumen atau segmentasi konsumen ini maka dukungan data mining sangatlah penting. Algoritma data mining yang paling tepat dan sering digunakan untuk segmentasi adalah K-Means Clustering dan Fuzzy C Means. Pembandingan kinerja kedua algoritma tersebut adalah melakukan pengelompokan untuk melakukan penggabungan algoritma clustering data pelanggan (K mean Clustering dan Fuzzy C-Means) dengan beberapa algoritma data mining seperti Classification, Association, dan CPV matrix untuk mendapatkan nilai potensial dari setiap kluster. Atribut yang digunakan untuk proses mining pada segmentasi konsumen adalah data pelanggan, produk, demografi, perilaku konsumen, transaksi, RFMDC, RFM (Recency, Frequency Monetary) dan LTV (Life Time Value).

3. Hasil Penelitian Ari Fadli*), Mulki Indana Zulfa, Yogi Ramadhani. (Fadli et al., 2018). dari Universitas Jenderal Soedirman, E-ISSN: 2338-0403, vol. 6, no. 4 Berjudul ***"Perbandingan Unjuk Kerja Algoritma Klasifikasi Data Mining dalam Sistem Peringatan Dini Ketepatan Waktu Studi Mahasiswa"*** Penelitian ini merupakan penelitian yang menggunakan metode *Analisis CRISP-DM.* dan menggunakan Variabel Mhasiswa dan Studi Mahasiswa.

Berdasarkan penelitian yang dilakukan oleh peneliti, maka dapat disimpulkan bahwa, Hasil perbandingan unjuk kerja algoritme decision tree, ANN dan SVM yang menggunakan data akademik mahasiswa aktif di FT Unsoed menunjukkan bahwa algoritme SVM memberikan nilai terbaik, yaitu accuracy sebesar 90,55% dan AUC sebesar 0,959. Performansi model dengan algoritma SVM sudah baik, sehingga dapat digunakan untuk mendapatkan informasi dari database mahasiswa dan memasukkannya ke dalam early warning system untuk akurasi studi mahasiswa. Dengan demikian, pembuat kebijakan dapat memantau masa studi mahasiswa dan memetakan mahasiswa yang mungkin harus menunda studinya.

4. Hasil Penelitian Gustientiedinaa , M.Hasmil Adiyaa , Yenny Desnelitab. (Gustientiedina et al., 2019). dari Sekolah Tinggi Ilmu Komputer Pelita Indonesia, E-ISSN 2476-8812 , VOL. 05 NO. 01, berjudul ***"Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru"*** , Penelitian ini merupakan penelitian yang menggunakan metode *Algoritma K-Means Clustering.* dan menggunakan variable obat-obatan.

   Berdasarkan penelitian yang dilakukan oleh peneliti, maka dapat disimpulkan bahwa,

   Dari hasil clusterisasi pada data obat – obatan dapat ditarik kesimpulan bahwa kelompok obat yang termasuk pemakaian sedikit rata rata permintaan obat setiap tahunnya kurang dari 18000 buah, dan obat yang termasuk pemakaian sedang rata rata permintaan obat setiap tahunnya diantara 18000–70000 buah,

sedangkan obat yang masuk kedalam kelompok obat yang pemakaian tinggi rata – rata permintaan obat setiap tahunnya diatas 70000 buah. Dari analisa cluster diatas mungkin perlu dilakukan lagi penelitian lanjutan agar clusterisasi data obat dapat dilakukan secara lebih valid dengan menetapkan nilai centroid terbaik.

5. Hasil Penelitian Allfanisa Annurrrullah Fajrin, Koko Handoko.(Fajrin & Handoko, 2018). dari Universitas Putera Batam, E-ISSN: 2615-1049, Vol. 06, No. 02, berjudul ***''Penerapan Data Mining Untuk Mengolah Tata Letak Buku Dengan Metode Association Rule''*** Penelitian ini merupakan penelitian yang menggunakan metode *Association Rule*.

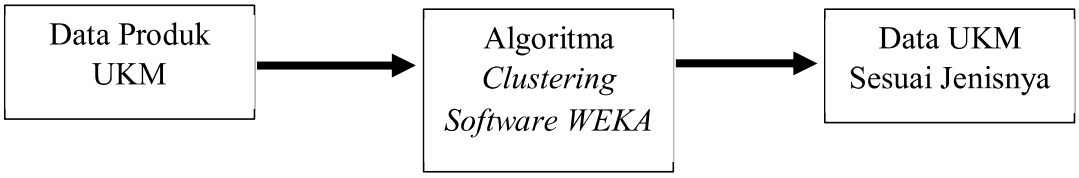
2.7 Kerangka Pemikiran

Kerangka pemikiran adalah penjelasan sementara terhadap sesuatu yang menjadi objek permasalahan. Kerangka pemikiran ini disusun berdasarkan pada tinjauan pustaka dan hasil penelitian yang relafan.

Seperti yang terlihat dari *flowchart*, langkah awal dari penelitian ini adalah:

1. Pengambilan data produk UKM yang kemudian dilakukan penyeleksian data dan pembersihan data yang berguna untuk mengetahui variable apa saja yang akan diolah.
2. Berdasarkan data-data yang diperoleh kemudian dilakukan proses Algoritma *Clustering* dengan menggunakan *Software WEKA*, untuk melakukan penggalian atau pengumpulan data.

3. Hasil temuan dari penelitian ini adalah aturan untuk mengelompoka data UKM yang masih mentah menjadi data UKM yeng sudah berpola, dan juga sudah sesuai dengan jenisnya masing-masing.



Gambar 2. 7 Kerangka Pemikiran
Sumber: (Peneliti, 2023)