BAB II
LANDASAN TEORI

2.1 Pengertian Data Mining

Data mining adalah tindakan yang menggabungkan bermacam-macam penggunaan dan historis untuk menentukan keteraturan, pola atau hubungan dalam dataset yang sangat besar. Salah satu tugas utama data mining adalah mengelompokkan clustering data yang belum memiliki contoh kelompok dikumpulkan. Data mining yang dapat juga disebut sebagai knowledge discovery in database (KDD). KDD adalah Tindakan yang menggabungkan bermacam-macam penggunaan dan historis untuk menentukan keteraturan, pola atau hubungan dalam dataset yang sangat besar (Santosa, 2007).

Data mining adalah tindakan mengetahui pola yang menarik dari banyak data, data dapat disimpan dalam database, data warehouse, atau penyimpanan data lainnya. Data mining berkaitan dengan bidang ilmu yang berbeda misalnya database sistem, data warehousing, statistic, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh berbagai ilmu misalnya neural network, pengenalan pola, spatial data analysis, image database, signal processing (Han, 2006).

2.2 Tahap – tahap Data Mining

Sebagai perkembangan rangkaian proses, data mining dibagi menjadi beberapa tahap. Tahap tersebut bersifat interaktif, pemakai pemakai knowledge base.

Tahapan data mining ada 6 yaitu :

1. Pembersihan Data

Pembersihan data adalah proses menghilangkan keributan dan informasi yang bertentangan atau berlebihan. Secara umum data diperoleh, dari hasil eksperimen atau database perusahaan, memiliki bagian yang kurang sempurna contohnya seperti data yang hilang, data yang salah ketik maupun yang tidak valid. Demikian pula, ada juga atribut data yang kurang relevan dengan hipotesa yang dimiliki data mining. data yang kurang relevan lebih baik dibuang. Pembersihan data juga akan

mempengaruhi penyajian darik metode data mining dengan alasan bahwa data yang diurus akan berkurang jumlah dan kompleksitasnya

2. Integritas Data

Integritas data adalah campuran data dari berbagai database ke dalam satu database baru. Tidak jarang data yang diperlukan data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau record teks. Integritas data dilakukan pada atribut nama, jenis barang, nomor pelanggan dan lain-lain. Integritas data harus dilakukan secara cermat karena kesalahan pada integritas data akan menghasilkan hasil yang berbeda dan bahkan salah pengambilan tindakan nantinya. Misalnya integritas data dilihat dari jenis produk berakhir dengan menggabungkan produk dari berbagai produk yang berbeda untuk mendapatkan hubungan antara produk yang sebenarnya tidak ada.

3. Seleksi Data

Data dalam database seringkali tidak semuanya digunakan, oleh karena uty hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Misalnya, untuk situasi yang melihat kecenderungan orang untuk membeli dalam kasus market basket analisis, cukup mengambil nomor id nya saja, tidak perlu mengambil nama pelanggan..

4. Transformasi Data

Data diubah atau digabungkan menjadi format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining yang dibutuhkan format data yang khusus sebelum diaplikasikan. Misalnya, beberapa strategi standar, seperti pemeriksaan afiliasi dan pengelompokan dapat mengakui masukan informasi yang jelas. Misalnya, beberapa strategi standar seperti menganalisis asosiasi dan clustering hanya dapat menerima input data katagorikal. Dengan car aini data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini disebut transformasi data.

5. Proses Mining

Proses mining adalah proses utanma Ketika metode diterapkan untuk menemukan pengetahuan tersembunyi dan berharga dari data.

6. Evaluasi Pola

Pola data yang dihasilkan dari proses data mining harus ditampilkan dalam struktur yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahapan ini adalah bagian dari proses KDD yang disebut juga interpretation. Tahapan ini mencakup pemeriksaan apakah contoh data atau pola ditemukan bertentangan dengan kenyataan saat ini atau hipotesa yang ada sebelumnya. (Sunjana,2010)

2.3 Teknik Data Mining

Berikut ini adalah teknik-teknik data mining (Bala., et al, 2012) :

1. Analisis asosiasi

   Analisis asosiasi adalah pengungkapan aturan asosiasi yang menggambarkan kondisi atribut nilai yang biasanya terjadi bersamaan dalam satuan data tertentu. Analisis asosiasi umumnya digunakan untuk Analisa data pasar dan transaksi.

2. Klasifikasi dan Prediksi

   Klasifikasi merupakan pemrosesan untuk menemukan sebuah model yang menjelaskan dan mencirikan konsep atau kelas data untuk kepentingan tertentu yang dapat menggunakan pemodelan untuk memprediksi kelas obyek yang labelnya belum diketuahui. Model yang didapat mungkin diwakili dalam berbagai format aturan klasifikasi IF-THEN, pohon keputusan, formula matematika, atau jaringan syaraf tiruan pengklasifikasian dapat digunakan guna memprediski label kelas data obyek.

3. Analisis Custering

   Clustering merupakan menganalisis obyek data tanpa mengkonsultasikan label kelas yang dikenal. Pada umumnya nilai kelas tinggi tidak diperoleh dalam penanganan data dasar karena mereka tidak tahu bagaimana memulainya. Clustering bisa dimanfaatkan untuk megenerate label. Obyek yang dikelompokkan bergantung pada aturan memaksimalkan persamaan dalam kelas meminimalkan kesamaan antar kelas. Sehingga cluster terhadap obyek dibentuk sedemikian rupa sehingga obyek dalam cluster memiliki persamaan

yang tinggi dalam perbandingan dengan obyek lainnya, tapi sangat berbeda dengan obyek dari cluster lain.

4. Analisis Outlier

   Sebuah database bisa jadi berisi obyek yang tidak sesuai dengan kebiasaan umumnya dari data yang disebut outlier. Analisa terhadap outlier mungkin membantu dalam mendeteksi kesalahan dan nilai-nilai abnormal.

2.4 Clustering

Clustering adalah proses pengelompokkan satu set benda abstrak ke dalam obyek yang sama. (Han and Kamber, 2006).

Baskoro (2010) menyatakan bahwa :

Clustering adalah salah satu perangkat untuk data mining yang mempunyai fungsi mengelompokkan obyek–obyek ke dalam cluster–cluster. Cluster adalah sekelompok atau sekumpulan obyek-obyek data yang similar satu dengan lainnya dalam cluster yang sama dan disimiliar cluster terhadap obyek-obyek yang berbeda cluster. obyek akan dikelompokkan ke dalam satu atau lebih cluster sehingga obyek-obyek yang berada pada satu cluster mempunyai kesamaan yang tinggi antara satu dan lainnya. Obyek dikelompokkan sesuai aturan memaksimalkan kesamaan obyek pada kelompok yang sama dan memaksimalkan ketidaksamaan pada kelompok yang berbeda. Kesamaan obyek pada umumnya didapat dari kualitas dari nilai-nilai atribut yang menjelaskan obyek data, sedangkan obyek-obyek umumnya diperesentasikan sebagai sebuah titik dalam ruang berlapis lapis.

Dengan memanfaatkan clusterisasi, kita dapat menemukan daerah yang tebal, menemukan pola-pola distribusi keseluruhan , dan menemukan keterkaitan yang menarik antar atribut data. Dalam data mining diusahakan berfokus pada metode penemuan untuk cluster ada basis data ang memiliki ukuran besar secara spesifik dan efektif. Beberapa kebutuhan clusterisasi dalam data mining meliputi skalabilitas, mampu menangani dimensional yang tinggi, menangani data yang mempunyai noise, kemampuan untuk menangani tipe atribut yang berbeda, dan dapat diterjemahkan dengan mudah. Adapun tujuan data clustering ini adalah untuk membatasi objective function yang dirancang dalam proses clustering, yang

sebagian besar bertujuan untuk membatasi varietas di dalam suatu cluster dan meningkatkan varietas antar cluster.

2.5 Algoritma K-Means

K-Means adalah perhitungan clustering yang berulang-ulang. Algoritma K-Means dimulai dengan menetukan secara acak K, K disini merupakan banyaknya cluster yang akan dibentuk. Lalu dipilih nilai K secara acak, kemudian nilai itu akan menjadi pusat cluster untuk sementara yang biasa disebut centroid, mean atau "means" Hitung jarak setiap data terhadap masing-masing centroid dengan menggunakan rumus Euclidian sampai ditemukan jarak yang paling dekat dari setiap data dengan centroid. Kelompokan data berdasarkan kedekatannya dengan centroid. Lakukan cara ini sampai nilai centroid tidak berubah (stabil) [SAN07].

Dari beberapa metode klastering yang paling dasar dan paling umum dikenal adalah klastering K-means. Dalam metode ini ini kita perlu mengelompokkan obyek kedalam K kelompok atau klaster. Untuk melakukan klaustering, nilai K harus diputuskan terlebih dahulu. Umumnya user atau pemakai sudah mempunyai data awal tentang obyek yang sedang dipelajari, , termasuk beberapa jumlah klaster yang paling cocok. Secara detail kita bisa memanfaatkan ukuran ketidakmiripan untuk mengelompokkan obyek kita. Ketidakmiripan bisa diartikan dalam konsep jarak. Jika jarak dua obyek atau data titik cukup dekat maka dapat diartikan dua obyek itu mirip. Semakin dekat maka semakin tinggi kemiripannya. Semakin tinggi nilai jarak maka semakin tinggi ketidakmiripannya.

Algoritma K-means :

1) Menentukan nilai K sebagai jumlah cluster yang akan dibentuk. Jumlah cluster yang akan dibentuk ditentukan oleh pengguna sistem
2) Memproduksi K centroid (titik pusat cluster) secara acak. Dalam menyimpulkan K buah pusat cluster awal dilakukan pembangkitan secara acak yang mempresentasikan urutan data input. Pusat cluster awal didapatkan dari data itu sendiri bukan dengan memutuskan titik lain, yaitu dengan mengacak pusat yang mendasari data.
3) Menghitung jarak data dengan pusat cluster menggunakan rumus Euclidean distance, Manhattan Distance dan Minkowski Distance.

4) Cari jarak terpendek dan masukkan X ke dalam cluster sesuai centroid tersebut. Hasil penentuan jarak akan dilihat dan jarak terkecil antara data dan pusat cluster akan dipilih. Jarak ini menunjukkan bahwa data itu ada dalam satu cluster dengan pusat cluster terdekat. Perhitungan pengelompokan data:

   1. Ambil nilai jarak tiap pusat dengan cluster.
   2. Temukan nilai jarak terkecil.
   3. Kumpulkan data dengan pusat cluster yang mempunyai jarak paling kecil.

5) Tentukan tempat centroid baru dengan menghitung rata-rata dari data yang terpilih pada centroid yang sama. Untuk mendapatkan pusat cluster baru, dapat ditentukan dengan rata-rata nilai anggota cluster yang baru.

   Perhitungan penentuan pusat cluster baru:

   1. Cari jumlah anggota tiap cluster.
   2. Hitung pusat baru dengan rumus :

$$v_{ij} = \frac{1}{N}\sum_{k}^{N} X_{kj} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

   Dimana :

   $v_{ij} = rata - rata\ cluster\ ke - i\ untuk\ variabel\ ke - j$

   $N_i\ = jumlah\ data\ yang\ menjadi\ anggota\ cluster\ ke - i$

   $i, k\ = indeks\ dari\ cluster$

   $j\ = indeks\ dari\ variabel$

   $X_{kj} = nilai\ data\ ke - k\ yang\ ada\ didalam\ cluster\ tersebut\ untuk\ variabel\ ke - j$

6) Lakukan tahap 3 - 5 sampai posisi anggota cluster baru dengan cluster lama tidak berubah. Pusat cluster yang baru digunakan untuk melakukan perhitungan iterasi berikutnya, jika hasil yang didapatkan belum konvergen, dan data akan berhenti jika hasilnya yang dicapai sudah konvergen (pusat cluster baru sama dengan pusat cluster lama) atau jika terdapat perubahan nilai centroid diatas nilai ambang atau nilai pada fungsi objektif yang telah ditentukan. Dimana nilai ambang (threshold) adalah 0.0000 < 1.

2.6 Euclidean Distance

Pengukuran jarak yang sering digunakan dalam beberapa penelitian tentang clustering adalah Euclidean Distance. Euclidean Distance dianggap sebagai distance matrix yang mengadopsi prinsip Phytagoras. Hal ini karena pada pola perhitungannya menggunakan aturan pangkat dan akar kuadrat. Euclidean akan memberikan hasil jarak yang relatif kecil karena menggunakan aturan akar kuadrat. Rumus persamaan pengukuran jarak euclidean distance sebagai berikut:

$$D(x_{2,}x_1) = ||x_2 - x_1|| = \sqrt{\sum_{i=1}^{n} |x_{2i} - x_{1i}|^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.2)$$

Keterangan :

$D(x_{2,}x_1) = Jarak\ antara\ data\ dengan\ pusat\ cluster$

$x_{2i} \quad = Nilai\ data\ dua\ ke - i$

$x_{1i} \quad = Nilai\ data\ pertama\ ke - i$

2.7 Manhattan Distance

Manhattan distance atau jarak Manhattan yang juga dikenal sebagai Taxicab distance atau City Block distance adalah metrik ukur yang biasanya digunakan untuk menghitung jarak antara dua titik data yang terletak dalam jalur yang mirip dengan grid. Jarak Manhattan dihitung sebagai jumlah dari perbedaan mutlak antara dua vektor. Adapun rumus dari Manhattan distance adalah:

$$D(x_2, x_1) = \sum_{i=1}^{n} |x_{2i} - x_{1i}| \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.3)$$

Keterangan :

$D(x_{2,}x_1) = Jarak\ antara\ data\ dengan\ pusat\ cluster$

$x_{2i} \quad = Nilai\ data\ dua\ ke - i$

$x_{1i} \quad = Nilai\ data\ pertama\ ke - i$

2.8 Minkowski Distance

Minkowski Distance adalah metrik ukur yang digunakan untuk menghitung jarak antara dua vektor bernilai bilangan riil. Metrik ini adalah generalisasi dari jarak Euclidean dan jarak Manhattan, tetapi memiliki parameter yang disebut "order" atau p, yang memungkinkan penghitungan jarak yang berbeda ketika dihitung. Berikut ini adalah rumus Minkowski Distance:

$$D(x_{2,}x_1) = (\sum_{i=1}^{n} |x_2 - x_1|^p)^{(\frac{1}{p})} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.4)$$

Keterangan :

$D(x_{2,}x_1) = Jarak\ antara\ data\ dengan\ pusat\ cluster$

$x_{2i} \quad = Nilai\ data\ dua\ ke - i$

$x_{1i} \quad = Nilai\ data\ pertama\ ke - i$

$p \quad = Parameter\ Order$

Ketika p diatur ke 1, perhitungannya sama dengan jarak Manhattan. Ketika p diatur ke 2, akan sama dengan jarak Euclidean.

p=1: Jarak Manhattan.

p=2: Jarak Euclidean.

2.9 Contoh Perhitungan Metode K-Means

Sebuah dataset mempunyai 4 obyek sebagai titik data pelatihan dan setiap obyek mempunyai 2 atribut. setiap koordinat dari obyek, yaitu :

Obyek atribut 1 (x) : bobot indeks

Obyek atribut 2 (y) : pH

Tabel 2.1 Data Kasus

Obyek	Atribut 1 (x) :Bobot indeks	Atribut 2 (y) : pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Untuk mengatasi permasalahan ini, kita bisa lakukan beberapa tahap, yaitu :

1. Menentukan jumlah cluster

   Dengan melihat data yang ada, kita dapat mengelompokkan obyek menjadi dua (cluster 1 dan cluster 2) sesuai dengan atributnya. Masalahnya adalah bagaimana cara mengetahui medicine tersebut adalah anggota cluster 1 atau cluster 2. Dari data yang didapat, dapat disimpulkan bahwa 4 obyek tersebut memiliki 2 atribut (Bobot indeks dan pH), dimana masing-masing medicine mewakili satu titik dengan 2 atribut (x,y).

2. Menentukan nilai awal centroid

   Untuk menentukan nilai yang mendasari centroid dilakukan secara acak, dalam contoh kasus ini titik koordinat medicine A adalah cluster 1 (C1) dan medicine B (C2) sebagai nilai centroid awal.

   a. C1 = (1,1)
   b. C2 = (1,2)

3. Menghitung jarak antar titik centroid menggunakan jarak Manhattan :

   Medicine A = (1,1) dengan C1 = (1,1)

   a. $|1-1|+|1-1| = 0$

      dengan C2 = (2,1)

   b. $|1-2|+|1-1| = 1$

      Medicine B = (2,1) dengan C1 = (1,1)

   c. $|2-1|+|1-1| = 1$

      dengan C2 = (2,1)

   d. $|2-2|+|1-1| = 0$

      Medicine C = (4,3) dengan C1 = (1,1)

   e. $|4-1|+|3-1| = 5$

      dengan C2 = (2,1)

   f. $|4-2|+|3-1| = 4$

      Medicine D = (5,4) dengan C1 = (1,1)

   g. $|5-1|+|4-1| = 7$

      dengan C2 = (2,1)

   h. $|5-2|+|4-1| = 6$

Dari perhitungan tersebut, maka didapatkan hasil jarak matriksnya, yaitu :

$$D^0 \begin{pmatrix} A & B & C & D \\ 0 & 1 & 5 & 7 \\ 1 & 0 & 4 & 6 \end{pmatrix} \begin{matrix} \\ \rightarrow \\ \rightarrow \end{matrix} \begin{matrix} \\ Cluster\ 1 \\ Cluster\ 2 \end{matrix}$$

4. Menghitung jarak antar titik centroid menggunakan jarak Minkowski dengan nilai P = 3 :

   Medicine A = (1,1) dengan C1 = (1,1)

   a. $\sqrt[3]{(1-1)^3 + (1-1)^3} = 0$

   dengan C2 = (2,1)

   b. $\sqrt[3]{(1-2)^3 + (1-1)^3} = 1$

   Medicine B = (2,1) dengan C1 = (1,1)

   c. $\sqrt[3]{(2-1)^3 + (1-1)^3} = 1$

   dengan C2 = (2,1)

   d. $\sqrt[3]{(2-2)^3 + (1-1)^3} = 0$

   Medicine C = (4,3) dengan C1 = (1,1)

   e. $\sqrt[3]{(4-1)^3 + (3-1)^3} = 29$

   dengan C2 = (2,1)

   f. $\sqrt[3]{(4-2)^3 + (3-1)^3} = 10$

   Medicine D = (5,4) dengan C1 = (1,1)

   g. $\sqrt[3]{(5-1)^3 + (4-1)^3} = 67$

   dengan C2 = (2,1)

   h. $\sqrt[3]{(5-2)^3 + (4-1)^3} = 30$

   Dari perhitungan tersebut, maka didapatkan hasil jarak matriksnya, yaitu :

$$D^0 \begin{pmatrix} A & B & C & D \\ 0 & 1 & 29 & 67 \\ 1 & 0 & 10 & 30 \end{pmatrix} \begin{matrix} \\ \rightarrow \\ \rightarrow \end{matrix} \begin{matrix} \\ Cluster\ 1 \\ Cluster\ 2 \end{matrix}$$

5. Pada contoh kasus ini menggunakan jarak Euclidean. Berikut adalah cara untuk menghitung jarak dari tiap obyek :

   Medicine A = (1,1) dengan C1 = (1,1)

   a. $\sqrt{(1-1)^2 + (1-1)^2 = 0}$

   dengan C2 = (2,1)

b. $\sqrt{(1-2)^2+(1-1)^2=1}$

Medicine B = (2,1) dengan C1 = (1,1)

c. $\sqrt{(2-1)^2+(1-1)^2}=1$

dengan C2 = (2,1)

d. $\sqrt{(2-2)^2+(1-1)^2=0}$

Medicine C = (4,3) dengan C1 = (1,1)

e. $\sqrt{(4-1)^2+(3-1)^2=3{,}61}$

dengan C2 = (2,1)

f. $\sqrt{(4-2)^2+(3-1)^2=2{,}83}$

Medicine D = (5,4) dengan C1 = (1,1)

g. $\sqrt{(5-1)^2+(4-1)^2=5}$

dengan C2 = (2,1)

h. $\sqrt{(5-2)^2+(4-1)^2=4{,}24}$

Dari perhitungan tersebut, maka didapatkan hasil jarak matriksnya, yaitu :

$$D^0\begin{pmatrix} A & B & C & D \\ 0 & 1 & 3{,}61 & 5 \\ 1 & 0 & 2{,}83 & 4{,}24 \end{pmatrix}\begin{matrix} \\ \rightarrow \\ \rightarrow \end{matrix}\begin{matrix} \\ Cluster\ 1 \\ Cluster\ 2 \end{matrix}$$

6. Pengelompokkan Obyek

Setelah menentukan jarak matriks, kami menyimpulkan anggota cluster menurut jarak minimum dari centroid. Dengan melihat jarak matriks, medicine A termasuk cluster 1, sedangkan medicine B, C dan D termasuk cluster 2. Hal ini dapat dilihat pada skor yang diperoleh sebagai berikut :

$$G^0\begin{pmatrix} A & B & C & D \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}\begin{matrix} \\ \rightarrow \\ \rightarrow \end{matrix}\begin{matrix} \\ Cluster\ 1 \\ Cluster\ 2 \end{matrix}$$

7. Pada iterasi 1

a. Menentukan centroid baru

Himpunan yang dibentuk pada iterasi sebelumnya, telah diketahui anggota tiap cluster. Untuk cluster 1 memiliki anggota medicine A saja, sedangkan cluster 2 memiliki anggota medicine B, C dan D. Dari data tersebut, hitung ulang centroid untuk menentukan centroid baru. Karena pada cluster 1 hanya

memiliki 1 anggota, maka untuk centroid baru masih pada C1 = (C1). Sedangkan di C2 dengan menghitung nilai rata-rata, nilai centroid baru dapat diperoleh, untuk lebih spesifiknya sebagai berikut :

$$C2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) \qquad C2 = \left(\frac{11}{3}, \frac{8}{3}\right)$$

b. Menghitung jarak antara titik centroid baru dengan tiap titik obyek. Pada tahap menghitung jarak antara obyek dengan centroid baru. Hampir sebanding dengan tahap 3, khususnya menghitung jarak dengan C2

$$C2 = \left(\frac{11}{3}, \frac{8}{3}\right)$$

Melalui perhitungan serupa pada tahap 3, maka diperoleh jarak matriksnya, yaitu :

$$D^1 \begin{pmatrix} A & B & C & D \\ 0 & 1 & 3{,}61 & 5 \\ 3{,}14 & 2{,}36 & 0{,}47 & 1{,}89 \end{pmatrix} \begin{matrix} \rightarrow \\ \rightarrow \end{matrix} \begin{matrix} C1 = (1{,}1) \\ C2 = \left(\frac{11}{3}, \frac{8}{3}\right) \end{matrix}$$

c. Pengelompokkan obyek

Hampir setara dengan tahap 4, yang menentukan individu kelompok atau anggota cluster dengan memastikan jarak minimum tiap objek dengan centroid baru. Hasil yang diperoleh adalah :

$$G^1 \begin{pmatrix} A & B & C & D \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} \rightarrow \\ \rightarrow \end{matrix} \begin{matrix} Cluster\ 1 \\ Cluster\ 2 \end{matrix}$$

8. Pada iterasi 2

a. Menentukan Centroid Baru

Tahap ini mengulangi tahap 5, yaitu menghitung centroid baru. Dari cluster 1 yang mempunyai 2 anggota yaitu medicine A dan B, dan cluster 2 mempunyai anggota yaitu medicine C dan D, hasil centroid baru yang didapat adalah :

$$C1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) \qquad C1 = \left(\frac{3}{2}, 1\right)$$

$$C2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) \qquad C2 = \left(\frac{9}{2}, \frac{7}{2}\right)$$

b. Menghitung jarak antara titik centroid baru dan setiap titik obyek. Tahap ini juga hampir sama dengan tahap 3 yaitu menghitung jarak dengan centroid baru

$C1 = \left(\frac{3}{2}, 1\right)$ $C2 = \left(\frac{9}{2}, \frac{7}{2}\right)$

Dengan cara perhitungan yang sama seperti pada tahap 3, maka diperoleh hasil jarak matriksnya, yaitu

$$D^2 \begin{pmatrix} A & B & C & D \\ 0{,}5 & 0{,}5 & 3{,}20 & 4{,}61 \\ 4{,}30 & 3{,}54 & 0{,}71 & 0{,}71 \end{pmatrix} \begin{matrix} \rightarrow \\ \\ \rightarrow \end{matrix} \begin{matrix} C1 = \left(\frac{3}{2}, 1\right) \\ \\ C2 = \left(\frac{9}{2}, \frac{7}{2}\right) \end{matrix}$$

c. Pengelompokkan Obyek

Hampir setara dengan tahap 4, yang menentukan individu kelompok atau anggota cluster dengan memastikan jarak minimum tiap objek dengan centroid baru. Hasil yang diperoleh adalah :

$$G^2 \begin{pmatrix} A & B & C & D \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} \rightarrow \\ \rightarrow \end{matrix} \begin{matrix} Cluster\ 1 \\ CLuster2 \end{matrix}$$

Berdasarkan hasil anggota cluster yang didapatkan tetap sama antara G^1 dan G^2 maka iterasi dihentikan

Tabel 2.2 Hasil Clustering

Obyek	Atribut 1 (x) :Bobot indeks	Atribut 2 (y) : pH	Cluster (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

2.10 Davies-Bouldin Index

Davies-Bouldin Index dikenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979. Sum-of square within cluster (SSW) sebagai matrik kohesi dalam sebuah cluster. Separasi dengan Sum-of-square-between-cluster (SSWB) dengan cara mengukur antara centroid C_i dan C_j. $R_{i,j}$ merupakan ukuran rasio seberapa bagus nilai perbandingan antara cluster ke-i dan cluster ke-j.

Rumus SSW adalah : Rumus SSB :

$$SSW = \frac{1}{N}\sum_{i=1}^{N} ||x_i - C_{pi}||^2 \qquad SSB = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,j\neq 1}^{M} ||C_i - C_j||^2$$

Rumus R dan DBI :

$$R_{i,j} = \frac{SSW_{i+}SSW_j}{SSB_{i,j}} \qquad DBI = \frac{1}{K}\sum_{i=1}^{K} \underset{i \neq j}{max}(R_{i,j})$$

Dapat ditunjukkan bahwa ketika nilai SSW semakin kecil, maka hasil clustering yang didapat akan lebih baik. Secara esensial, DBI menginginkan nilai sekecil (non-negatif ≥ 0) mungkin untuk menilai baiknya cluster yang didapat.

2.11 Penelitian Sebelumnya

Beberapa artikel digunakan sebagai acuan pembelajaran, berikut artikel yang digunakan sebagai bahan wacanan antara lain :

a. Penelitian yang dilakukan Christofer Satria dan Anthony Anggrawan (2021) Pada penelitian ini algoritma K-Means digunakan untuk mewujudkan pengelompokan kelas belajar berdasarkan nilai dan prestasi siswa baru sehingga diperoleh klasifikasi kelas unggulan. Metode penelitian yang dipakai adalah perhitungan k-means dengan program aplikasi berbasis web. Konsekuensi dari tinjauan ini menunjukkan bahwa perhitungan k-means cocok untuk menentukan pilihan dan pembagian kelas-kelas unggulan untuk siswa baru yang direncanakan sebagaimana ditunjukkan oleh nilai kemampuan siswa. Penerapan kelas yang unggulan jelas mempengaruhi perkembangan pendidikan lebih lanjut.

b. Penelitian yang dilakukan Imam Amirulloh (2019) Pada penelitian ini algoritma k-means digunakan untuk membuat kelompok kerja siswa dengan kualitas masing – masing kelompok merata, pengukuran kualitas siswa dapat dilakukan dengan metode clustering K-Means, dengan uji coba pada 37 siswa yang membentuk 7 kelompok kerja sehingga kualitas siswa terbagi pada 5 cluster, cluster 0 sebanyak 20 orang, cluster 1 sebanyak 3 orang, cluster 2 sebanyak 7 orang, cluster 3 sebanyak 3 orang, cluster 4n sebanyak 4 orang.
c. Penelitian yang dilakukan Penda Sudarto Hasugian dan Jijon Raphita Sagala (2022). Pada penelitian ini Proses Data mining dengan menerapkan perhitungan K-Means dilakukan untuk mengelompokkan data kedalam setidaknya satu kelompok atau lebih, dimana data yang memiliki reprensentative persamaan dikelompokkan dalam satu kelompok dan data yang memiliki perbedaan masuk kedalam kelompok lainnya. Pengelompokan data siswa dilakukan untuk memudahkan sekolah dalam memfasilitasi siswa berdasarkan perbedaan kemampuannya dalam belajar dan mengikuti pembelajaran yang terdiri dari kelompok atau kelas siswa unggulan, kelompok sedang dan rendah. Hasil penerapan metode k-means diuji menggunakan aplikasi rapid miner dari data nilai siswa adalah sama dimana Kelompok 1 (C1) berhasil dengan 2 siswa yaitu siswa 002 dan siswa 003. Kelompok 2 (C2) yang merupakan kelompok Sedang, dimana ada 4 data yang terdapat untuk kelompok ini, yaitu Siswa006, Siswa008, Siswa009, dan Siswa010. Kelompok 3 (C3) merupakan kelompok rendah dimana terdapat 4 data yang terdapat pada kelompok ini, yaitu Siswa001, Siswa004, Siswa005 dan Siswa007. Proses klasterisasi memberikan hasil klasifikasi pengelompokan data yang efektif. Sehingga dapat menghemat waktu dalam melakukan klasterisari kelas siswa.
d. Penelitian yang dilakukan Yani Prihati, Suwarno, Alexander Dharmawan. Pada penelitian ini algoritma k-means digunakan untuk mengelompokkan data dengan memaksimalkan kemiripan data dalam satu klaster dan meminimalkan kemiripan data antar klaster. Dengan menggunakan Metode Algoritma K-Means Clustering, dapat mementukan pengelompokan prestasi siswa tinggi, menengah dan rendah. Atribut yang digunakan pada penelitian ini yaitu 15

atribut dan menghasilkan 3 cluster. Cluster_0 rendah terdapat 2 items, cluster_1 menengah terdapat 6 items dan cluster_2 tinggi terdapat 13 items.

e. Penelitian yang dilakukan Mardalius. Pada penelitian ini algoritma k-means digunakan untuk menentukan kelas kelompok belajar tambahan. Tambahan belajar yang dilakukan oleh siswa dimaksudkan untuk memperluas pemahaman dan pengembangan materi suatu mata pelajaran. Tujuan ini terkait dengan perencanaan seorang siswa untuk menghadapi ujian di sekolah, baik ujian tengah semester, ujian semester akhir maupun ujian akhir nasional. Jumlah tes data yang akan digunakan adalah 26 siswa jurusan IPA. Hasil yang didapat yaitu C0 memiliki 2 anggota yang diartikan bahwa kelompok pertama adalah kategori kemampuan siswa pintar. C1 memiliki 22 orang yang berarti kelompok selanjutnya adalah kategori kemampuan siswa sedang. C2 memiliki 2 orang yang artinya kelompok ketiga adalah kelas kemampuan siswa yang kurang cerdas dan siswa tersebut akan diberikan pembelajaran ekstra.
f. Penelitian yang dilakukan Bagus Yayang Fatkhurrahman. Pada penelitian ini algoritma k-means digunakan untuk membagi kelompok belajar siswa dalam menghadapi Ujian Nasional (UN) dan Ujian Akhir Sekolah (UAS). Jumlah data yang digunakan berjumlah 20 Siswa. Hasil dari pembagian kelompok yaitu Kelompok 1 ada siswa 9 siswa, kelompok 2 ada 4 siswa, kelompok 3 ada 7 siswa.
g. Penelitian yang dilakukan Melissa Triandini, , Sarjon Defit , Gunadi Widi Nurcahyo(2021). Pada penelitian ini algoritma k-means digunakan untuk mengelompokkan siswa guna mengukur sejauh mana kemampuan siswa dalam menjalani proses pembelajaran serta menjadi acuan dan bahan evaluasi bagi pihak sekolah dalam keberhasilan para pendidik saat melaksanakan belajar mengajar. Dari hasil yang didapat dalam pengelompokan yang didapat, data dikelompokan menjadi 3 cluster yaitu, cluster 1 yang bernilai kurang aktif sebanyak 12 siswa, cluster 0 yang bernilai aktif sebanyak 10 siswa, dan cluster 2 yang bernilai sangat aktif sebanyak 17 siswa.