BAB 2
LANDASAN TEORI

2.1 Data Mining

Menurut Tan (2006) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengestrakan informasi baru yang diambil dari bongkahan data yang besar yang membantu dalam pengambilan keputusan. Istilah data mining sering disebut juga knowledge discovery

Salah satu teknik yang dibuat dalam data minig adalah bagaimana menelusuri data yang ada untuk membangun sebuah model. Kemudian menggunakan model tersebut agar dapat mengenali pola data yang lainnya

Selanjutnya perbedaan data mining dengan data werehouse dapat dilihat bahwa data mining adalah bidang yang sepenuhnya menggunakan apa yang dihasilkan oleh data werehouse, bersama dengan bidang yang menangani masalah pelaporan dan menejemen data. Sementara data werehouse sendiri bertugas untuk menarik data dari basis data mentah untuk memberikan hasil data ynag nantinya digunakan oleh bidang yang menangani menejemen,pelaporan dan data mining (Prasetyo,2013)

2.2 Pengertian Database

Menurut Fuazi (2010) Kata “database” diambil dari bahasa inggris, terdiri dari dua kata yaitu *base* dapat diartikan sebagai dasar, landasan, atau tempat berkumpul. *Data* adalah representasi fakta dunia nyata yang mewakili suatu objek seperti manusia, barang, peristiwa, konsep, keadaan dan sebagainya, yang direkam dalam bentuk angka, huruf, simbol, teks, gambar, bunyi atau kombinasinya.

Database dapat didefenisikan dalam sejumlah sudut pandang seperti:

- Himpunan kelompok data (arsip) yang saling berhubungan yang diorganisasi sedemikian rupa agar kelak dapat dimanfaatkan kembali dengan cepat dan mudah.

- Kumpulan data yang saling berhubungan yang disimpan secara bersama sedemkian rupa dan tanpa redundansi yang tidak perlu untuk memenuhi berbagai kebutuhan.

•Kumpulan file/tabel/arsip yang saling berhubungan yang disimpan dalam media penyimpanan elektronis.

Dari beberapa definisi-definisi tersebut, dapat dikatakan bahwa database mempunyai berbagai sumber data dalam pengumpulan data, bervariasi derajat interaksi kejadian dari dunia nyata, dirancang dan dibangun agar dapat digunakan oleh beberapa pengguna untuk berbagai kepentingan.

2.3 Sistem Database

Sistem adalah sebuah tatanan (keterpaduan) yang terdiri atas sejumlah komponen fungsional dengan fungsi khusus yang saling berhubungan dan secara bersama-sama bertujuan untuk memenuhi suatu proses tertentu. Database hanyalah sebuah objek yang pasif. Database tidak akan pernah berguna jika tidak ada penggeraknya. Yang menjadi penggeraknya secara langsung adalah DBMS (Database Management System) yaitu aplikasi yang menangani semua (Fauzi,Rahmad.2010).

2.4 Pengelompokan Database

Pada penelitian ini, penulis menggunakan metode Fuzzy C-Means untuk mengelompokan keterampilan database mahasiswa yang ada di Universitas Muhammadiyah Gresik jurusan Teknik Informatika angkatan 2010, konsep data yang digunakan adalah sebagai berikut :

a.Data primer dari kuesioner yang berisi kumpulan data mahasiswa Universitas Muhammadiyah Gresik jurusan Teknik Informatika angkatan 2010.

b. Pelabelan yang digunakan untuk pengelompokan keterampilan database mahasiswa antara lain : sangat tinggi, tinggi, sedang, rendah. Pada penelitian ini, penulisannya menggolongkan sebagian mahsiswa Universitas Muhammadiyah Gresik jurusan Teknik Informatika angkatan 2010 yang tergolong rendah untuk dikelompokan kedalam cluster mahasiswa rendah yang tergolong sedang dikelompokan kedalam cluster mahasiswa sedang, sedangkan mahasiswa yang tergolong tinggi dikelompokan kedalam cluster tinggi, dan mahasiswa yang tergolong sangat tinggi dikelompokan kedalam cluster sangat tinggi dalam hal keterampilan database. Oleh karena itu pada penelitian ini, penulis melakukan pengolahan data dari semua mahasiswa UMG jurusanTeknik Informatika angkatan 2010, baik yang tergolong terampil dan tidak terampil database, dan output dari fuzzy C-means

digunakan sebagai data pengelompokan untuk menentukan mahasiswa dengan keterampilan database sangat tinggi, tinggi, sedang, rendah.

2.5 Fuzzy Logic

Logika fuzzy adalah salah satu cara yang tepat untuk memetakan suatu ruang input ke dalam suatu ruang output (Kusumadewi,2004).sebagai contoh:

1. Manajer pergudangan mengatakan pada manajer produksi seberapa banyak persediaan barang pada akhir minggu ini.kemudian manajer produksi akan menetapkan jumlah barang yang harus di produksi esok hari.
2. Pelayanan restoran memberikan pelayanan terhadap tamu, kemudian tamu akan memberikan tip yang sesuai atas baik tidaknya pelayanan yang diberikan.

2.5.1 Alasan digunakan logika fuzzy

Ada beberapa alasan orang menggunakan logika fuzzy, diantaranya :

a. Konsep logika fuzzy mudah dimengerti. Konsep matematis yang mendasari penalaran fuzzy sangat sederhana dan mudah dimengerti.
b. Logika fuzzy bersifat sangat fleksibel.

2.6 Fuzzy C-Means

Fuzzy clustering adalah salah satu teknik untuk menentukan *cluster* optimal dalam suatu ruang *vektor* yang didasarkan pada bentuk normal untuk jarak antar vektor. *Fuzzy clustering* sangat berguna bagi pemodelan *fuzzy* terutama dalam mengindentifikasi aturan-aturan *fuzzy*. Metode *clustering* merupakan pengelompokan data beserta parameternya dalam kelompok – kelompok sesuai kecenderungan sifat dari masing-masing data tersebut (kesamaan sifat). Ada beberapa algoritma *clustering* data, salah satu diantaranya adalah *FuzzyC-Means. Fuzzy C-Means* adalah suatu teknik peng-*cluster*-an yang mana keberadaannya tiap-tiap titik data dalam suatu *cluster* ditentukan oleh derajat keanggotaan.

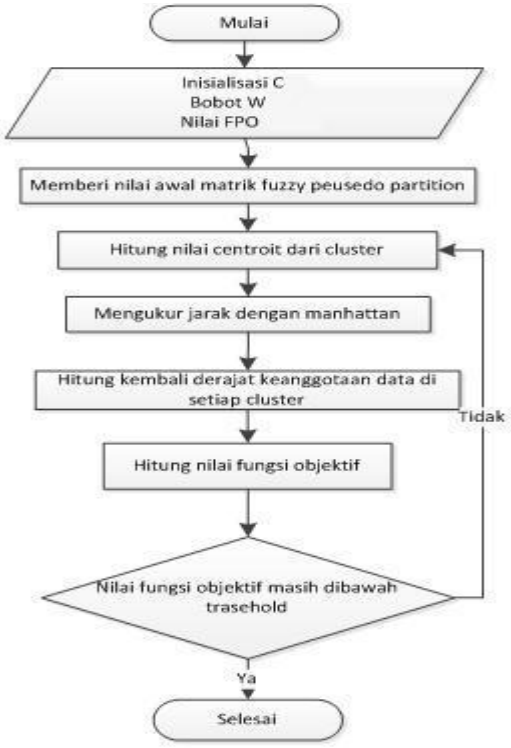
Clustering dengan metode fuzzy C-means (FCM) didasarkan pada teori logika fuzzy, teori ini pertama kali diperkanalkan oleh LotfiZadeh (1965) dengan nama himpunan fuzzy (fuzzy set). Dalam teori fuzzy, keanggotaan sebuah data tidak diberikan nilai secara tegas dengan nilai 1 (menjadi anggota) dan 0 (tidak menjadi anggota) melainkan dengan suatu nilai derajat keanggotaanya yang jangkauan nilainya 0 sampai 1. Nilai keanggotaan suatu dalam sebuah himpunan menjadi 0 ketika sama sekali tidak menjadi anggota, dan menjadi 1 ketika menjadi anggota secara penuh dalam himpunan. Umumnya nilai

keanggotaan antara 0 dan 1. Semakin tinggi nilai keanggotaanya maka semakin tinggi derajat keanggotaanya,dan semakin kecil maka semakin rendah derajat keanggotaanya. kaitanya dengan K-means, Sebenarnya FCM merupakan versi fuzzy dari K-Means dengan beberapa modifikasi yang membedakanya dengan K-Means (Prasetyo, 2013).

Konsep dari *Fuzzy C-Means* pertama kali adalah menentukan pusat *cluster*, yang akan menandai lokasi rata-rata untuk tiap-tiap *cluster*. Pada kondisi awal, pusat *cluster* ini masih belum akurat. Tiap – tiap titik data memiliki derajat keanggotaan untuk tiap-tiap *cluster*.Dengan cara memperbaiki pusat *cluster* dan derajat keanggotaan tiap – tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimalisasi fungsi objektif yang menggambarkan jarak dari titik data yang diberikan kepusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut.(Kusumadewi dan Purnomo,2010).

2.7 Flowchart Fuzzy C-Means

alir yang akan digunakan dalam penelitian ini secara umum dapat dilihat pada gambar 2.4 berikut.



Gambar 2.1 Flowchart Fuzzy clustering Means.

Pada gambar 2.4 digambarkan secara umum proses yang terjadi adalah:

1. Menginputkan data yang dicluster x, berupa matrik berukuran n x m berfungsi untuk menentukan jumlah data dan atribut setiap data yang akan dipergunakan:

n = jumlah sampel data

m = atribut setiap data

X ij = data sampel ke-i (i=1,2,..n) atribut ke-j (j=1,2…,m).

X = $\begin{bmatrix} x_1 & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$…… ……..(2.1)

2. Menentukan :

Jumlah cluster = c;

Bobot pangkat = w;

Maksimum Iterasi = MaxIter;

Eror terkecil yang diharapkan = ε

Fungsi Objektif awal = Po = 0;

Iterasi awal = t=1;

3. Membangkitkan bilangan random matriks pseudo awal μ_{ik}, i=1,2,…n; k = 1,2,…

Menghitung jumlah tiap kolom (atribut) :

$Q_j = \sum_{k=1}^{c} \mu_{ik}$

Hitung

$\mu_{ik} = \frac{\mu_{ik}}{Q_j}$

4. Menghitung pusat cluster ke –k: V_{kj},dengan k=1,2,…c; dan j=1,2,….,m; penentuan pusat cluster digunakan untuk menandai lokasi rata-rata untuk tiap cluster dengan kondisi awal tidak akurat.

$V_{ij} = \frac{\sum_{k=1}^{n}((\mu_{ik})^w * Xkj)}{\sum_{k=1}^{n}(\mu_{ik})w}$…… (2.2)

V = $\begin{bmatrix} V_{11} & \cdots & V_{1m} \\ \vdots & \ddots & \vdots \\ V_{c1} & \cdots & V_{cm} \end{bmatrix}$

5. Menghitung jarak data ke pusat cluster dengan mengunakan manhattan (Rectilinear) distance kemudian akan didapatkan matrik jarak sebagai berikut :

$d_{ik} = d(x_k - v_i) = \sum_{j=1}^{m} | x_{kj} - v_{ij} |$..... (2.3)

6. Menghitung perubahan matriks partisi, penghitungan ini berfungssi sebagai nilai awal matriks jika mengalami perulangan dan agar lokasi cluster bisa berada pada posisi yang benar.

$\mu_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(w-1)}\right]^{-1}$…… (2.4)

7. Menghitung fungsi objektif pada iterasi ke = t, Pt : perhitungan fungsi objektif digunakan untuk menggambarkan jarak dari titik data yang diberikan ke pusat cluster yang berbobot oleh derajat keanggotaan titik data tersebut.

   Pt= $\sum_{i=1}^{c}\sum_{j=1}^{m}$ ⍰$_{ik}{}^{w}\left\|x_{kj}-v_{ij}\right\|^{2}$

8. Cek Kondisi berhenti :
   - Jika ([Pt-Pt$_{-1}$]<ε) atau (t < maxIter) maka berhenti;
   - Jika tidak : t = t+1, ulangi langkah ke-4;

   Langkah ketujuh berfungsi sebagai pengkodisian perhitungan terhadap data, apakah suatu cluster yang telah dihasilkan, sudah memenuhi syarat atau perlu dilakukan iterasi selanjutnya agar lokasi cluster yang dihasilkan bisa berada pada posisi yang benar.

2.8 JarakManhattan

Jarak manhattan merupakan jarak yang diukur mengikuti jalur yang tegak lurus. Disebut dengan jarak manhattan, mengingatkan jalan-jalan manhattan yang membentuk garis parallel dan saling tegak lurusan satu jalan dengan jalan lainya. Pengukuran dengan jarak manhattan sering digunakan karena mudah perhitunganya, mudah dimengerti dan untuk beberapa masalah lebih sesuai, misalnya untuk menentukan jarak antar kota, jarak antar fasilitas dimana peralatan pemindahan bahan hanya dapat bergerak searah tegak lurus (Setya Yaniar.Nimas.2012).

Dalam pengukuran jarak manhattan digunakan notasi sebagai berikut :

$d(x,y) = |Xi - Yi|$

d (x,y) = *Manhattan distance* yaitu jarak antara data pada titik x dan titik y

x = Data

y = pusat cluster

2.8.1 Kelebihan Manhattan Distance

Berikut ini adalah keunggulan mengunakan manhattan distance disbanding dengan algoritma pengukuran jarak lainya adalah

1. Semua jalur – jalur dapat ditemukan (masalah dapat dipecahkan).
2. Hal ini disebabkan karena pada setiap penambahan nilai g (n), pada perhitungan nilai heuristic-nya terjadi pula perubahan pada nilai d nya. Sehingga dengan penambahan nilai g(n) , tidak mempengaruhi pencarian jalur.
3. Dengan mengunakan fungsi heuristic manhattan distance, didapatkan nilai iterasi

dan jumlah langkah yang paling kecil dibanding dengan mengunakan fungsi heuristic yang lain (Setya Yaniar, Nimas.2012).

2.9 Indeks XB (Xie dan Beni)

Untuk menentukan banyak kelompok dapat dilakukan dengan menghitung Indeks XB (Xie dan Beni). Indeks ini ditemukan oleh Xie dan Beni dan pertama kali dikemukakan pada tahun 1991. Indeks XB dituliskan sebagai berikut

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{m} \square_{ik}{}^{w} \left\| x_{kj} - v_{ij} \right\|^2}{n(min\left\| x_{kj} - v_{ij} \right\|^2)}$$

dimana i = banyak kelompok, $\square_{ik}$ = derajat keanggotaan, j = jarak pengamatan dengan dan n =banyak objek yang akan dikelompokkan, serta kriteria banyak kelompok yang optimum ditunjukkan dengan nilai indeks XB yang minimum pada lembah pertama. Penggunaan indeks XB untuk menentukan jumlah kelompok yang optimum pada metode *fuzzy C-means* menyatakan bahwa indeks XB memiliki ketepatan dan keandalan yang tinggi untuk memberikan jumlah kelompok yang optimum pada metode *fuzzy C-means (Sandhika Jaya.Tri.2012).*

2.10 Penelitian Sebelumnya

Penelitian sebelumnya dilakukan oleh Narwati dengan judul ”PENGELOMPOKAN MAHASISWA MENGGUNAKAN ALGORTIMA K-MEANS”. Pada penelitian ini dibangun bertujuan untuk menghasilkan pola dari prestasi mahasiswa. Mahasiswa saat masuk dari sejumlah 936 mahasiswa adalah sejumlah 116 mahasiswa atau sebesar 12,39% masuk kluster 1, 363 (38,782%) mahasiswa masuk kluster 2 dan 457 (48,852) mahasiswa masuk kluster 3. Kesimpulan dari penelitian ini adalah hasil kluster untuk nilai tes prosentase terbesar adalah mahasiswa yang masuk kluster 3 dengan kemampuan tinggi berdasarkan hasil test.

Penelitian kedua dilakukan oleh Wahyu Oktri Widyanto dengan judul ”CLUSTERING DATA NILAI MAHASISWA UNTUK PENGELOMPOKAN KOSENTRASI JURUSAN MENGGUNAKAN FUZZY C-MEANS. Penelitian ini bertujuan untuk menganalisa data mahasiswa menurut bobot mata kuliah tertentu menggunakan konsep FCM. Pada penelitian ini menguji sebanyak 126 data. Hasil clustering ini menghasilkan sejumlah kelompok kosentrasi dengan jumlah masing-masing mahasiswa sesuai dengan cluster yang ada. Mahasiswa yang tercluster kedalam cluster 1 (Multimedia) sebanyak

28 mahasiswa, cluster 2 (WEB) sebanyak 70 mahasiswa dan cluster 3 (Pemogramanan) sebanyak 85 mahasiswa.

Penelitian ketiga dilakukan oleh Irma Irandha dengan judul ”ANALISA KELUARGA MISKIN DENGAN MENGUNAKAN METODE FUZZY C-MEANS ”. Penelitian ini bertujuan untuk memberikan pelabelan data keluarga miskin dengan katagori sangat miskin, mendekati miskin dan mampu. Atribut yang digunakan ada 6 yaitu jumlah ART, jenis pekerjaan, kesehatan, pendidikan, perumahan dan lingkungan ekonomi. Hasil clustering ini menghasilkan sejumlah kategori yang meliputi hampir mendekati miskin, hampir sangat miskin, miskin dan sangat miskin