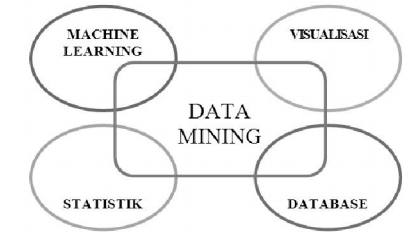
**BAB II**

**LANDASAN TEORI**

1. **Data Mining**

Menurut Turban dalam bukunya yang berjudul “Decision Support Systems and Intelligent System”, Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam basis data. Data Mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan, dan machine learning untuk mengekstaksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar.



Gambar 2.1 Data Mining merupakan irisan dari berbagai disiplin.

1. **Pengelompokan Data Mining**

Menurut Larose dalam bukunya yang berjudul ”Discovering Knowledge in Data: An Introduction to Data Mining”, data mining dibagi menjadi beberapa kelompok berdasarkan tugas/pekerjaan yang dapat dilakukan, yaitu:

1. Deskripsi

Terkadang peneliti dan anilis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola kecenderungan sering memberikan kemunginan penjelasan untuk suatu pola atau kecenderungan.

1. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variable target estimasi lebih ke arah numeric dari pada ke arah kategori. Model dibangun menggunakan baris data (record)

Lengkap yang menyediakan nilai dari variable target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi dari variable target dibuat berdasarkan nilai variable target dibuat berdasarkan nilai variable prediksi.

1. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

1. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

1. Pengklasteran (Clustering)

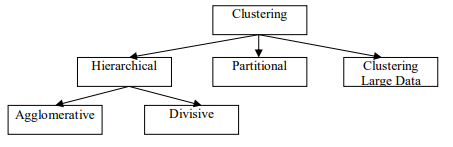
Pengklasteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas obyek-obyek yang memiliki kemiripan. Klaster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan record dalam klaster yang lain. Berbeda dengan klasifikasi, pada pengklasteran tidak ada variabel target. Pengklasteran tidak melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target, akan tetapi, algoritma pengklasteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

1. Asosiasi

Tugas asosiasi dalam data mining adalah untuk menemukan atribut yang muncul dalam satu waktu. Salah satu implementasi dari asosiasi adalah market basket analysis atau analisis keranjang belanja, sebagaimana yang akan dibahas dalam tesis ini.

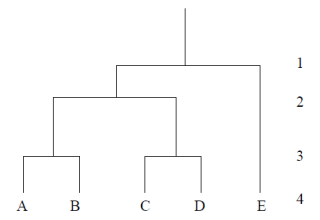
1. **Algoritma *Clustering***

Secara umum pembagian algortima *clustering* dapat digambarkan sebagai berikut:



Gambar 2.2 Kategori Algoritma *Clustering*

1. *Hierarchical Clustering* menentukan sendiri jumlah cluster yang dihasilkan. Hasil dari metode ini adalah suatu struktur data berbentuk pohon yang disebut dendogram dimana data dikelompokkan secara bertingkat dari yang paling bawah dimana tiap instance data merupakan satu cluster sendiri, hingga tingkat paling atas dimana keseluruhan data membentuk satu cluster besar berisi cluster-cluster seperti gambar dibawah ini:

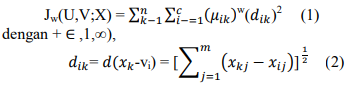


Gambar 2.3 *Dendogram*

1. *Divisive Hierarchical Clustering* mengelompokkan data dari kelompok yang terbesar hingga ke kelompok yang terkecil, yaitu masing-masing instance dari kelompok data tersebut. Sebaliknya, agglomerative hierarchical clustering mulai mengelompokkan data dari kelompok yang terkecil hingga kelompok yang terbesar. Beberapa algoritma yang menggunakan metode ini adalah: *RObust Clustering Using LinKs (ROCK), Chameleon, Cobweb, Shared Nearest Neighbor (SNN).*
2. *Partitional Clustering* yang mengelompokkan data ke dalam k cluster dimana k adalah banyaknya cluster dari input user. Kategori ini biasanya memerlukan pengetahuan yang cukup mendalam tentang data dan proses bisnis yang memanfaatkannya untuk mendapatkan kisaran nilai input yang sesuai. Beberapa algoritma yang masuk dalam kategori ini antara lain: *K-Means, Fuzzy C-Means, Clustering Large Aplications (CLARA), Expectation Maximation (EM), Bond Energy Algorithm (BEA), algoritma Genetika, Jaringan Saraf Tiruan.*
3. **Algoritma *Fuzzy Clustering C-Means* (FCM)**

Pada proses pengklasteran (*clustering*) secara klasik (misalnya pada algoritma *Clustering K-Means*), pembentukan partisi dilakukan sedemikian rupa sehingga setiap obyek berada tepat pada satu partisi. Namun, adakalanya tidak dapat menempatkan suatu obyek tepat pada suatu partisi, karena sebenarnya obyek tersebut terletak di antara 2 atau lebih partisi yang lain. Pada logika *fuzzy*, metode yang dapat digunakan untuk melakukan pengelompokan sejumlah data dikenal dengan nama *fuzzy clustering*. *Fuzzy Clustering* lebih alami jika dibandingkan dengan pengklasteran secara klasik. Suatu algoritma *clustering* dikatakan sebagai *fuzzy clustering* jika algoritma tersebut menggunakan parameter strategi adaptasi secara *soft competitive*. Sebagian besar algoritma *fuzzy clustering* didasarkan atas optimasi fungsi obyektif atau modifikasi dari fungsi obyektif tersebut.

Salah satu teknik *fuzzy clustering* adalah *Fuzzy C-Means* (FCM). FCM adalah suatu teknik pengklasteran data yang keberadaan tiap-tiap data dalam suatu cluster ditentukan oleh nilai/derajat keanggotaan tertentu. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981. Berbeda dengan teknik pengklasteran secara klasik (dimana suatu obyek hanya akan menjadi anggota suatu klaster tertentu), dalam FCM setiap data bisa menjadi anggota dari beberapa *cluster*.Batas-batas *cluster* dalam FCM adalah lunak (*soft*). Konsep dasar FCM, 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 data memiliki derajat keanggotaan untuk tiap-tiap *cluster*. Dengan cara memperbaiki pusat cluster dan nilai keanggotaan tiap-tiap data secara berulang, maa akan terlihat bahwa pusat cluster akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif. Fungsi Obyektif yang digunakan pada FCM adalah:



X adalah data yang akan diklaster:



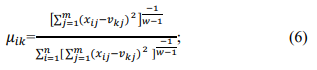
Dan v adalah matriks pusat cluster :



Nilai Jw terkecil adalah yang terbaik, sehingga:



Jika 𝑑𝑖𝑘 > 0, ∀𝑖, 𝑘 ; 𝑤 > 1 dan X setidaknya memiliki m elemen, maka (u, v) ∈ 𝑀𝑓𝑚 > 0,∀𝑖 , 𝑘 ; 𝑤 dapat meminimasi Jw hanya jika:



Dan



Algoritma *Fuzzy C-Means* (FCM) diberikan sebagai berikut :

1. Menentukan data yang akan di cluster X, berupa matriks berukuran n x m (n=jumlah sampel data, m = atribut setiap data). Xij=data sampel ke-I (i=1,2,…,n), atribut ke-j (j=1,2,…,m).
2. Menentukan :

* Jumlah cluster = c
* Pangkat = w
* Maksimum iterasi = MaxIter
* Error terkecil yang diharapkan = ξ
* Fungsi objektif awal = P0 = 0
* Iterasi awal = t = 1

1. Membangkitkan bilangan random µik, i=1,2,3 …,n; k=1,2,3…c; sebagai elemenelemen matriks partisi awal U. Menghitung jumlah setiap kolom :



Dengan j=1,2,…..n.

Menghitung :



1. Menghitung pusat cluster ke-k: Vkj, dengan k=1,2,…c; dan j=1,2,…m.



1. Menghitung fungsi objektif pada iterasi ke-t :



1. Mengitung perubahan matriks partisi :



1. **Proses *Clustering* menggunakan *Fuzzy C-Means***

Tahap ini akan diterapkan metode Fuzzy C-Means untuk mengelompokkan data. Hasil pengelompokkan ini kemudian akan digunakan untuk pertimbangan menentukan mahasiswa yang berhak menerima beasiswa. Adapun algoritma C-Means Clustering pada penerima beasiswa adalah sebagai berikut :

1. Menetapkan matriks partisi awal U berupa matriks berukuran n x m (n adalah jumlah sampel data, yaitu=75, dan m adalah parameter/atribut setiap data, yaitu=2). Xij=data sampel ke-i (i=1,2,…,n), atribut ke-j (j=1,2,..,m).
2. Menentukan Nilai Parameter Awal :

* Jumlah cluster (c) = 2
* Pangkat/bobot (w) = 2
* Maksimum interasi (MaxIter) = 100
* Error terkecil yang diharapkan (ξ)= 10
* Fungsi Objektif awal (Po) = 0
* Interasi awal (t) = 1

1. Membangkitkan bilangan random µik, i=1,2,…c; sebagai elemen-elemen matriks partisi awal (U).