BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Pada penelitian Rismawan dan Kusumadewi (2008) mengelompokkan mahasiswa berdasarkan status gizi *Body Mass Index* (BMI) dan ukuran kerangka. Dari data yang dilatih, diperoleh 3 kelompok BMI dan ukuran kerangka, yaitu cluster 1 BMI normal dan kerangka besar, cluster 2 BMI obesitas sedang dan kerangka sedang, cluster 3 BMI obesitas berat dan kerangka kecil.

Pada penelitian Hastuti (2013) mengelompokkan mahasiswa berdasarkan kriteria - kriteria yang telah ditentukan, antara lain IPK, jumlah tanggungan keluarga, dan penghasilan total orang tua. Pengelompokan dilakukan dengan menerapkan algoritma K-Means Clustering. Pengelompokan akan menunjukkan siapa saja yang akan masuk ke dalam masing-masing kelompok. Jumlah cluster yang digunakan pada kasus tersebut adalah tiga (3) berdasarkan perhitungan validasi cluster optimal, yaitu mahasiswa yang direkomendasikan menerima beasiswa, dipertimbangkan menerima beasiswa, dan tidak menerima beasiswa.

Penelitian Asroni dan Adrian (2015) menggunakan algoritma *K-Means* untuk menentukan cluster yang terbaik. Cluster terbaik ini dipergunakan untuk pemilihan mahasiswa-mahasiswa terbaik yang dapat diikutkan lomba. Sehingga peluang untuk mendapatkan juara dalam lomba bisa semakin besar.

Pada penelitian Setiawan (2015) mengelompokan berdasarkan mata kuliah yang dominan ke masing-masing konsentrasi meliputi 12 mata kuliah yang akan dikelompokan dan dicari rata-rata nilainya (Rata-rata Sistem Informasi dan Enterprise, Rata-rata Jaringan Komputer , dan Rata-rata Rekayasa Perangkat Lunak). Kemudian di-cluster menjadi 3 group anggota data yang masuk cluster 1 maka akan direkomendasikan untuk masuk ke kelas konsentrasi Jaringan Komputer dan Multimedia, untuk anggota cluster 2 maka akan direkomendasikan untuk masuk ke kelas konsentrasi Rekayasa Perangkat Lunak dan Animasi sedangkan untuk anggota cluster 3 akan rekomendasikan untuk masuk ke kelas konsentrasi Sistem Informasi dan Enterprise.

Pada penelitian Sukamdi (2015) menganalisis pola alumni mahasiswa dengan mengunakan teknik *clustering* berdasarkan nilai UAN saat mendaftar dan IPK menjadi 3 kelompok, yaitu kelompok 1, kelompok 2, kelompok 3. Data yang digunakan dalam penelitian berasal dari sistem akademik STMIK AKAKOM Yogyakarta. Algoritma *clustering* yang digunakan adalah k-means.

Pada Usulan (2016) penelitian ini memiliki kriteria nilai fleksibel, dalam kebutuhan mengelompokkan calon pegawai berdasarkan nilai wawancara dan psikotes Pengelompokan akan menunjukkan siapa saja yang akan masuk ke dalam masing-masing kelompok, Kemudian di-cluster menjadi 3 kelompok yaitu cluster 1 tidak dipertimbangkan, cluster 2 dipertimbangkan dan cluster 3 disarankan.

Dalam penelitian ini mengunakan tinjauan pustaka sebagai berikut :

Tabel 2.1 Tinjauan Pustaka

No	Penulis	Objek	Metode	Kriteria	Bahasa Pemrograman
1.	Rismawan dan Kusumadewi (2008)	Pengelompokan Mahasiswa Body Mass Index (BMI) & Ukuran Kerangka	K-Means	Body Mass Index (BMI) & Ukuran Kerangka	Java
2.	Hastuti (2013)	Penentuan Penerimaan Beasiswa	K-Means	Indeks Prestasi Kumulatif (IPK), Jumlah Tanggungan Keluarga dan Penghasilan Total Orang Tua (penghasilan ayah ditambah dengan penghasilan ibu)	PHP
3.	Asroni dan Adrian (2015)	Clustering Mahasiswa Berdasarkan Nilai Akademik Jurusan Teknik Informatika UMM	K-Means	nilai matakuliah Algoritma dan Pemrograman, Fisika Dasar, Kalukulus dan Indek Prestasi Komuliatif (IPK)	Java
4.	Setiawan (2015)	Konsentrasi bagi Mahasiswa Infromatika UMS	K-Means	Rerata SKS yang diambil mahasiswa	Java
5.	Sukamdi (2016)	Klasifikasi data mahasiswa STMIK AKAKOM	K-Means	rata-rata UAN dan IPK	Java
6.	Usulan (2016)	Klasifikasi seleksi calon pegawai	K-Means	kriteria nilai fleksibel	Java

2.2. Dasar Teori

2.2.1. Definisi Sistem

Suatu sistem dapat terdiri dari beberapa subsistem atau sistem-sistem bagian. Komponen-komponen atau subsitem dalam suatu sistem tidak dapat berdiri lepas sendiri-sendiri. Komponen-komponen dan subsistem saling berinteraksi dan saling berhubungan membentuk satu kesatuan sehingga tujuan atau sasaran dapat tercapai. Menurut Sutarman (2009:5), “sistem adalah kumpulan elemen yang saling berhubungan dan saling berinteraksi dalam satu kesatuan untuk menjalankan suatu proses pencapaian suatu tujuan utama”.

Menurut Jimmy L.Goal (2008:9), “sistem adalah hubungan satu unit dengan unit-unit lainnya yang saling berhubungan satu sama lainnya dan yang tidak dapat dipisahkan serta menuju satu kesatuan dalam rangka mencapai tujuan yang telah ditetapkan. Apabila suatu unit macet atau terganggu, unit lainnya pun akan terganggu untuk mencapai tujuan yang telah ditetapkan tersebut”.

2.2.2. Data Mining

Data mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan. Data mining ini juga dikenal dengan istilah pattern recognition (Santosa, 2007).

2.2.3. Clustering

Clustering adalah suatu metode untuk penggelompokan dokumen dimana dokumen dikelompokan dengan konten untuk mengurangi ruang pencarian yang diperlukan dalam merespon suatu *query*. Misalnya koleksi dokumen yang berisi dokumen-dokumen medis dan hukum dapat dikelompokkan sedemikian rupa sehingga semua dokumen medis ditempatkan dalam satu *cluster* dan semua dokumen hukum ditempatkan dalam satu *cluster* hukum (Grossman, David A. dan Ophir Frieder, 2004, h.105).

2.2.4. K-Means

Metode K-Means pertama kali diperkenalkan oleh MacQueen JB pada tahun 1976. Metode ini adalah salah satu metode *non hierarchi* yang umum digunakan. Metode ini termasuk dalam teknik penyekatan (*partition*) yang membagi atau memisahkan objek ke k daerah bagian yang terpisah. Pada K-Means, setiap objek harus masuk dalam kelompok tertentu, tetapi dalam satu tahapan proses tertentu, objek yang sudah masuk dalam satu kelompok, pada satu tahapan berikutnya objek akan berpindah ke kelompok lain.

K-means merupakan salah satu metode data klustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster / kelompok. Metode ini mempartisi ke dalam cluster / kelompok sehingga data yang memiliki karakteristik yang sama (High intra class similarity) dikelompokkan ke dalam satu cluster yang sama dan yang memiliki

karakteristik yang berbeda (Law inter class similarity) dikelompokkan pada kelompok yang lain (Giyanto, Heribertus. 2008).

Proses klustering dimulai dengan mengidentifikasi data yang akan dikluster, Xij (i=1,...,n; j=1,...,m) dengan n adalah jumlah data yang akan dikluster dan m adalah jumlah variabel. Pada awal iterasi, pusat setiap kluster ditetapkan secara bebas (sembarang), Ckj (k=1,...,k; j=1,...,m). Kemudian dihitung jarak antara setiap data dengan setiap pusat kluster. Untuk melakukan penghitungan jarak data ke-i (xi) pada pusat kluster ke-k (ck), diberi nama (dik), dapat digunakan formula Euclidean (Han, Jiawei; & Kamber, Micheline. 2001) seperti pada persamaan (2.1), yaitu

$$d_{ik} = \sqrt{\sum_{j=1}^{m} (c_{ij} - c_{kj})^2} \quad \text{.....(2.1)}$$

Suatu data akan menjadi anggota dari kluster ke-k apabila jarak data tersebut ke pusat kluster ke-k bernilai paling kecil jika dibandingkan dengan jarak ke pusat kluster lainnya. Hal ini dapat dihitung dengan menggunakan persamaan (2.2) Selanjutnya, kelompokkan data-data yang menjadi anggota pada setiap kluster.

$$\text{Min} \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m} (c_{ij} - c_{kj})^2} \quad \text{.....(2.2)}$$

Nilai pusat kluster yang baru dapat dihitung dengan cara mencari nilai rata-rata dari data-data yang menjadi anggota pada kluster tersebut, dengan menggunakan rumus pada persamaan (2.3):

$$c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad \text{.....(2.3)}$$

Dimana xij ∈ kluster ke-k

p = banyaknya anggota kluster ke k

Algoritma dasar dalam *k-means* adalah *Clustering* menggunakan metode *K-Means* secara umum dilakukan dengan algoritma sebagai berikut:

1. Tentukan jumlah *cluster* (*k*), yang ingin dibentuk.
2. Bangkitkan *k centroid* (titik pusat *cluster*) awal secara random.
3. Hitung jarak setiap data ke masing-masing *centroid*. Setiap data memilih *centroid* yang terdekat.
4. Tentukan posisi *centroid* baru dengan cara menghitung nilai rata-rata dari data yang memilih pada *centroid* yang sama.
5. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

2.2.5. Contoh Perhitungan Algoritma *K-Means*

Ada 5 data,variabel yang digunakan dalam pengelompokan adalah *x* dan *y*. Jarak yang digunakan adalah *Euclidean distance* dengan Jumlah *cluster (K)* adalah 3, data dan nilai dari variabel dapat dilihat di tabel 2.2.

Tabel 2.2 Data Contoh

Data ke-i	***x***	***y***
1	2	1
2	1	2
3	4	3
4	5	4
5	3	1

1. Terdapat 2 *cluster* (*k*).
2. Bangkitkan *k centroid* (titik pusat *cluster*) awal secara random, dapat dilihat di tabel 2.3.

Tabel 2.3 *Centroid*

Data ke-i	**x**	**y**
1	2	1
3	4	3

3. Hitung jarak setiap data ke masing-masing *centroid*. Setiap data memilih *centroid* yang terdekat. Misalkan untuk menghitung jarak data pertama dengan pusat *cluster* pertama adalah

$$d_{11} = \sqrt{(2-2)^2 + \ (1-1)^2} \ = 0{,}00$$

Jarak data pertama dengan pusat *cluster* kedua :

$$d_{12} = \sqrt{(4-2)^2 + (3-1)^2} = 2{,}83$$

Hitung jarak setiap data ke masing-masing *centroid*, dapat dilihat di tabel 2.4.

Tabel 2.4 Hitung jarak iterasi 0

Data ke-i	x	y	C1	C2	Minimal	Kelompok
1	2	1	0,00	2,83	0,00	C1
2	1	2	1,41	3,16	1,41	C1
3	4	3	2,83	0,00	0,00	C2
4	5	4	4,24	1,41	1,41	C2
5	3	1	1,00	2,24	1,00	C1

4. Tentukan posisi *centroid* baru dengan cara menghitung nilai rata-rata dari data yang memilih pada *centroid* yang sama. Untuk *centroid* pertama

   C11= (2 + 1 + 3) / 3 = 2,0

   C12= (1 + 2 + 1) / 3 = 1,3

   Untuk *centroid* Kedua :

   C21= (4 + 5) / 2 = 4,5

   C22= (3 + 4) / 2 = 3,5

Jadi berhenti di iterasi 1 dengan centroid sama, dapat dilihat di tabel 2.5.

Tabel 2.5 *Centroid1*

Data ke-i	x	y
1	2,0	1,3
3	4,4	3,5

Perhitungan jarak setiap data ke masing-masing *centroid*, dapat dilihat di tabel 2.6.

Tabel 2.6 Hitung jarak iterasi 1

Data ke-i	x	y	C1	C2	Minimal	Kelompok
1	2	1	0,09	3,54	0,09	C1
2	1	2	1,22	3,81	1,22	C1
3	4	3	2,62	0,71	0,71	C2
4	5	4	4,04	0,71	0,71	C2
5	3	1	1,04	2,92	1,04	C1