

BAB II

LANDASAN TEORI

2.1. Data Mining

2.1.1. Pengertian Data Mining

Data mining adalah salah satu teknik penelusuran data untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Dalam *data mining*, pengelompokan data juga dilakukan. Tujuannya adalah agar penulis dapat mengetahui pola dan tindak lanjut yang diambil. Semua hal tersebut bertujuan untuk mendukung kegiatan evaluasi agar sesuai dengan yang diharapkan (Prasetyo, 2012).

Data mining merupakan sebuah proses untuk menemukan pola atau pengetahuan yang bermanfaat secara otomatis atau semi otomatis dari sekumpulan data dalam jumlah besar. *Data mining* hadir dianggap sebagai bagian dari *Knowledge Discovery in Database* (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data. KDD terdiri dari beberapa langkah (Santosa, 2007) yaitu :

1. Pembersihan data (membuang *noise* dan data yang tidak konsisten).
2. Integrasi data (penggabungan data dari beberapa sumber).
3. Seleksi data (memilih data yang relevan yang akan digunakan untuk analisa).
4. Data mining.
5. Evaluasi model.
6. Presentasi pengetahuan dengan teknik visualisasi.

2.1.2. Metode Data Mining

Pada umumnya metode *data mining* dapat dikelompokkan ke dalam dua kategori, yaitu : *deskriptif* dan *prediktif*. Metode *deskriptif* untuk mencari pola yang dapat dimengerti oleh manusia yang menjelaskan karakteristik dari data. Metode *clustering* menggunakan ciri-ciri tertentu dari data untuk

melakukan pengelompokan data, (Kusrini & Emha. 2009). Metode-metode dalam *data mining* adalah sebagai berikut :

1. *Classification*

Klasifikasi merupakan proses untuk menemukan sekumpulan model yang menjelaskan kelas-kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui pada sebuah obyek. Untuk mendapatkan model, kita harus melakukan analisis terhadap data latih (*training set*). Sedangkan data uji (*test uji*) digunakan untuk mengetahui tingkat akurasi dari model yang telah dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai dari suatu obyek data.

2. *Clustering*

Pengelompokan (*clustering*) merupakan proses untuk melakukan segmentasi. Digunakan untuk melakukan pengelompokan secara alami terhadap atribut suatu set data. Termasuk kedalam *undersurvised task*. Metode inilah yang digunakan dalam tugas akhir ini.

3. *Association*

Tujuan dari metode ini yaitu untuk menghasilkan sejumlah rule yang menjelaskan sejumlah data yang terhubung kuat dengan yang lainnya. Sebagai contoh *association analysis* dapat digunakan untuk menentukan produk yang dating dibeli secara bersamaan oleh banyak pelanggan, atau bias juga disebut dengan *market based analysis*.

4. *Regression*

Regression mirip dengan klasifikasi. Perbedaan utamanya adalah terletak pada atribut yang diprediksi berupa nilai yang *kontinyu*.

5. *Forecasting*

Prediksi (*forecasting*) berfungsi untuk melakukan prediksi kejadian yang akan datang berdasarkan data sejarah yang ada.

6. *Sequence Analysis*

Tujuan dari metode ini adalah untuk mengenali pola dari *data diskrit*, sebagai contoh adalah menemukan kelompok gen dengan tingkat ekspresi yang mirip.

7. *Deviation Aanalysis*

Tujuan dari metode ini adalah untuk menemukan penyebab perbedaan antara data yang satu dengan data yang lain dan biasa disebut sebagai *outlier detection*. Sebagai contoh adalah apakah sudah terjadi penipuan terhadap pengguna kartu kredit dengan melihat catatan transaksi yang tersimpan dalam basis data perusahaan tersebut.

2.2. *Clustering*

Analisis kluster atau *clustering* merupakan proses membagi data dalam suatu himpunan kedalam beberapa kelompok yang kesamaan datanya dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok yang lain.(Kusrini & Emha, 2009).

Potensi *clustering* dapat digunakan untuk mengetahui struktur dalam data yang dapat dipakai lanjut dalam berbagai aplikasi secara luas seperti klasifikasi, pengolahan gambar dan pengenalan pola. *Clustering* dapat diterapkan kedalam data yang *kuantitatif* (numerik), *kualitatif* (kategorikal) atau kombinasi keduanya, (Kusrini & Emha. 2009). Data dapat merupakan hasil pengamatan dari suatu prose. *Cluster* secara umum merupakan wujud himpunan bagian dari suatu himpunan data dan metode *clustering* dapat diklasifikasi berdasarkan himpunan bagian yang dihasilkan : apakah *fuzzy* atau *crisp (hard)*.

Pada penelitian tugas akhir ini akan dibahas mengenai penggunaan metode *Fuzzy K-Means* dengan pengelompokan santri.

2.3. *Algoritma K-Means*

K-Means merupakan algoritma *clustering* yang berulang-ulang. Algoritma K-Means dimulai dengan pemilihan secara acak N, N disini merupakan banyaknya *cluster* yang ingin dibentuk. Kemudian tetapkan nilai-nilai N secara *random*, untuk

sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan centroid, mean atau “means”. Hitung jarak setiap data yang ada terhadap masing-masing centroid menggunakan rumus Euclidian hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Lakukan langkah tersebut hingga nilai centroid tidak berubah, (Santosa, 2007).

Dari beberapa teknik *clustering* yang paling sederhana dan umum dikenal adalah *clustering K-Means*. Dalam teknik ini kita ingin mengelompokkan obyek kedalam N kelompok atau *cluster*. Untuk melakukan *clustering*, nilai N harus ditentukan terlebih dahulu. Biasanya *user* atau pemakai sudah mempunyai informasi awal tentang obyek yang sedang dipelajari, termasuk beberapa jumlah *cluster* yang paling tepat. Secara detail kita bias menggunakan ukuran ketidakmiripan untuk mengelompokkan obyek kita. Ketidakmiripan bisa diterjemahkan dalam konsep jarak. Jika jarak dua obyek atau dua titik cukup dekat maka dua obyek itu mirip. Semakin dekat berarti semakin tinggi kemiripannya. Semakin tinggi jarak semakin tinggi ketidakmiripannya.

Dalam penelitiannya mengungkapkan langkah-langkah pengerjaan algoritma K-Means (Santosa, 2007) yaitu :

1. Penentuan pusat *cluster* awal

Dalam penentuan nilai N buah pusat *cluster* awal dilakukan pembangkitan bilangan *random* yang mempresentasikan urutan data input pusat awal *cluster* didapatkan dari data sendiri bukan dengan menentukan titik baru, yaitu dengan merandom pusat awal dari data.

2. Perhitungan jarak dengan pusat *cluster*

Untuk mengukur jarak antara data dengan pusat cluster digunakan rumus *Euclidian distance*. Algoritma perhitungan jarak data dengan pusat cluster

- a. Ambil nilai data dan nilai pusat cluster.
- b. Hitung *Euclidian distance* data dengan tiap pusat cluster.

3. Pengelompokan data

Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat cluster, jarak ini akan menunjukkan data tersebut

berada dalam suatu kelompok dengan pusat cluster terdekat. Algoritma pengelompokan data :

1. Ambil nilai jarak tiap pusat cluster dengan data.
2. Cari nilai jarak terkecil.
3. Kelompokkan data dengan pusat cluster yang memiliki jarak terkecil.
4. Penentuan pusat *cluster* baru

Untuk mendapatkan pusat cluster baru bisa dihitung dari rata-rata nilai anggota cluster dan pusat cluster. Pusat cluster yang baru digunakan untuk melakukan iterasi selanjutnya, jika hasil yang didapatkan belum konvergen. Proses iterasi ini akan berhenti jika telah memenuhi maksimum iterasi yang dimasukkan oleh user atau hasil dicapai sudah konvergen (pusat cluster baru sama dengan pusat cluster lama).

Algoritma penentuan pusat cluster baru yaitu:

- a. Cari jumlah anggota tiap cluster.
- b. Hitung pusat baru dengan rumus.

$$\text{Pusat cluster baru} = \frac{x_1 + x_2 + x_3, \dots, x_n}{\text{jumlah} + 1}$$

Dimana :

$x_1, x_2, x_3, \dots, x_n$ adalah anggota cluster.

Berikut ini adalah uraian dari perancangan algoritma K-means untuk menentukan pengelompokan potensi atau nilai siswa.

2.3.1. Contoh kasus pengelompokan dengan metode K-Means

Pengelompokan 10 yang dapat dilihat pada Tabel 2.1. yang menggunakan dua dimensi yakni x dan y. pengukuran jarak menggunakan Euclidean distance. Jumlah cluster (K) yang dipakai 3 dan threshold (T) yang digunakan untuk perubahan fungsi objektif adalah 0.1.

Tabel 2.1. Data set contoh pengelompokan dengan metode K-Means

| Data ke-i | Fitur x | Fitur y |
|-----------|---------|---------|
| 1 | 1 | 1 |

| Data ke-i | Fitur x | Fitur y |
|-----------|---------|---------|
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |
| 6 | 5 | 3 |
| 7 | 2 | 5 |
| 8 | 3 | 5 |
| 9 | 2 | 6 |
| 10 | 3 | 8 |

Inisialisasi :

- Jumlah cluster (K) = 3
- Treshold (T) perubahan fungsi objektif = 0.1.
- Pemilihan K data sebagai centroid awal, missal dipilih data ke-2, data ke-4, dan data ke-6

Tabel 2.2. Jarak data ke Centroid dan Cluster yang diikuti

| Data ke-i | Fitur x | Fitur y | Centroid |
|-----------|---------|---------|------------|
| 1 | 1 | 1 | |
| 2 | 4 | 1 | Centroid 1 |
| 3 | 6 | 1 | |
| 4 | 1 | 2 | Centroid 2 |
| 5 | 2 | 3 | |
| 6 | 5 | 3 | Centroid 3 |
| 7 | 2 | 5 | |
| 8 | 3 | 5 | |
| 9 | 2 | 6 | |
| 10 | 3 | 8 | |

Untuk nilai fungsi objektif awal, karena data belum masuk dalam cluster, maka nilai fungsi objektif diberi nilai awal yang besar, misalnya 1000.

Iterasi 1

Menghitung jarak setiap data ke centroid terdekat menggunakan Persamaan Euclidean yang dapat dilihat pada Persamaan 2.1

$$D(x_1, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^p |x_{2j} - x_{1j}|^2} \dots\dots\dots 2.1$$

Jarak ke setiap centroid pada data ke-1

$$d(x_1, c_1) = \sqrt{\sum_{i=1}^r (x_{1i} - c_{1i})^2} = \sqrt{(1-4)^2 + (1-1)^2} = \sqrt{9+0} = 3$$

$$d(x_1, c_2) = \sqrt{\sum_{i=1}^r (x_{1i} - c_{2i})^2} = \sqrt{(1-1)^2 + (1-2)^2} = \sqrt{0+1} = 1$$

$$d(x_1, c_3) = \sqrt{\sum_{i=1}^r (x_{1i} - c_{3i})^2} = \sqrt{(1-5)^2 + (1-3)^2} = \sqrt{16+4} = 4.4721$$

Tabel 2.3. Jarak data ke Centroid dan Cluster yang diikuti pada iterasi 1

| Data ke-i | Jarak ke Centroid | | | Terdekat | Cluster yang diikuti |
|--------------|-------------------|--------|--------|----------|----------------------|
| | 1 | 2 | 3 | | |
| 1 | 3.0000 | 1.0000 | 4.4721 | 1.0000 | 2 |
| 2 | 0.0000 | 3.1623 | 2.2361 | 0.0000 | 1 |
| 3 | 2.0000 | 5.0990 | 2.2361 | 2.0000 | 1 |
| 4 | 3.1623 | 0.0000 | 4.1231 | 0.0000 | 2 |
| 5 | 2.8284 | 1.4142 | 3.0000 | 1.4142 | 2 |
| 6 | 2.2361 | 4.1231 | 0.0000 | 0.0000 | 3 |
| 7 | 4.4721 | 3.1623 | 3.6056 | 3.1623 | 2 |

| Data ke-i | Jarak ke Centroid | | | Terdekat | Cluster yang diikuti |
|--------------|-------------------|--------|--------|----------|----------------------|
| | 1 | 2 | 3 | | |
| 8 | 4.1231 | 3.6056 | 2.8284 | 2.8284 | 3 |
| 9 | 5.3852 | 4.1231 | 4.2426 | 4.1231 | 2 |
| 10 | 7.0711 | 6.3246 | 5.3852 | 5.3852 | 3 |

Kemudian setelah mendapatkan nilai jarak terdekat maka dilakukan pengelompokan sesuai cluster yang diikuti. Kemudian dilakukan perhitungan rata-rata pada masing-masing cluster menggunakan Persamaan 2.2

$$c_j = \frac{1}{Nk} \sum_{l=1}^{Nk} x_{jl} \dots\dots\dots 2.2$$

Nk adalah jumlah data yang tergabung dalam sebuah cluster

Tabel 2.4. Centroid cluster 1 pada iterasi 1

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| Nk | Jumlah x | Jumlah y |
| 2 | 10 | 2 |
| Rata-rata | 5.0000 | 1.0000 |

Tabel 2.5. Centroid cluster 2 pada iterasi 1

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 1 | 1 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |
| 7 | 2 | 5 |
| 9 | 2 | 6 |
| Nk | Jumlah x | Jumlah y |

| Data anggota | Fitur x | Fitur y |
|--------------|---------|---------|
| 5 | 8 | 17 |
| Rata-rata | 1.6000 | 3.4000 |

Tabel 2.6. Centroid cluster 3 pada iterasi 1

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 6 | 5 | 3 |
| 8 | 3 | 5 |
| 10 | 3 | 8 |
| Nk | Jumlah x | Jumlah y |
| 3 | 11 | 16 |
| Rata-rata | 3.6667 | 5.3333 |

Setelah didapatkan nilai rata-rata pada setiap kelompok maka dilakukan perhitungan nilai objektif menggunakan Persamaan 2.3.

$$J = \sum_{i=1}^N \sum_{l=1}^K a_{ic} d(x_i, c_l)^2 \dots\dots\dots 2.3$$

Tabel 2.7. Nilai fungsi objektif pada iterasi 1

| Data ke-i | Cluster 1 | Cluster 2 | Cluster 3 |
|-----------|-----------|-----------|-----------|
| 1 | 0 | 6.1200 | 0 |
| 2 | 1.0000 | 0 | 0 |
| 3 | 1.0000 | 0 | 0 |
| 4 | 0 | 2.3200 | 0 |
| 5 | 0 | 0.3200 | 0 |
| 6 | 0 | 0 | 7.2222 |
| 7 | 0 | 2.7200 | 0 |
| 8 | 0 | 0 | 0.5556 |
| 9 | 0 | 6.9200 | 0 |
| 10 | 0 | 0 | 7.5556 |

$$J = 35.7333$$

$$\text{Perubahan } J = 1000 - 35.7333 = 964.2667$$

Masih Ada data yang pindah cluster atau perubahan $J > T$ maka dilanjutkan ke iterasi berikutnya.

Iterasi 2

Tabel 2.8. Jarak data ke Centroid dan Cluster yang diikuti pada iterasi 2

| Data ke-i | Jarak ke Centroid | | | Terdekat | Cluster yang diikuti |
|-----------|-------------------|--------|--------|----------|----------------------|
| | 1 | 2 | 3 | | |
| 1 | 4.0000 | 2.4739 | 5.0881 | 2.4739 | 2 |
| 2 | 1.0000 | 3.3941 | 4.3461 | 1.0000 | 1 |
| 3 | 1.0000 | 5.0120 | 4.9216 | 1.0000 | 1 |
| 4 | 4.1231 | 1.5232 | 4.2687 | 1.5232 | 2 |
| 5 | 3.6056 | 0.5657 | 2.8674 | 0.5657 | 2 |
| 6 | 2.0000 | 3.4234 | 2.6874 | 2.0000 | 1 |
| 7 | 5.0000 | 1.6492 | 1.6997 | 1.6492 | 2 |
| 8 | 4.4721 | 2.1260 | 0.7454 | 0.7454 | 3 |
| 9 | 5.8310 | 2.6306 | 1.7951 | 1.7951 | 3 |
| 10 | 7.2801 | 4.8083 | 2.7487 | 2.7487 | 3 |

Tabel 2.9. Centroid cluster 1 pada iterasi 2

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| 6 | 5 | 3 |
| Nk | Jumlah x | Jumlah y |
| 3 | 15 | 5 |

| Data anggota | Fitur x | Fitur y |
|--------------|---------|---------|
| Rata-rata | 5.0000 | 1.6667 |

Tabel 2.10. Centroid cluster 2 pada iterasi 2

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 1 | 1 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |
| 7 | 2 | 5 |
| Nk | Jumlah x | Jumlah y |
| 4 | 6 | 11 |
| Rata-rata | 1.5000 | 2.7500 |

Tabel 2.11. Centroid cluster 3 pada iterasi 2

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 8 | 3 | 5 |
| 9 | 3 | 6 |
| 10 | 3 | 8 |
| Nk | Jumlah x | Jumlah y |
| 3 | 8 | 19 |
| Rata-rata | 2.6667 | 6.3333 |

Tabel 2.12. Nilai fungsi objektif pada iterasi 2

| Data ke-i | Cluster 1 | Cluster 2 | Cluster 3 |
|-----------|-----------|-----------|-----------|
| 1 | 0 | 3.3125 | 0 |
| 2 | 1.4444 | 0 | 0 |
| 3 | 1.4444 | 0 | 0 |
| 4 | 0 | 0.8125 | 0 |
| 5 | 0 | 0.3125 | 0 |
| 6 | 1.7778 | 0 | 0 |

| Data ke-i | Cluster 1 | Cluster 2 | Cluster 3 |
|-----------|-----------|-----------|-----------|
| 7 | 0 | 5.3125 | 0 |
| 8 | 0 | 0 | 1.8889 |
| 9 | 0 | 0 | 0.5556 |
| 10 | 0 | 0 | 2.8889 |

$$J = 19.75$$

$$\text{Perubahan } J = 35.75 - 19.75 = 15.98$$

Masih Ada data yang pindah cluster atau perubahan $J > T$ maka dilanjutkan ke iterasi berikutnya.

Iterasi 3

Tabel 2.13. Jarak data ke Centroid dan Cluster yang diikuti pada iterasi 3

| Data ke-i | Jarak ke Centroid | | | Terdekat | Cluster yang diikuti |
|-----------|-------------------|--------|--------|----------|----------------------|
| | 1 | 2 | 3 | | |
| 1 | 4.0552 | 1.8200 | 5.5877 | 1.8200 | 2 |
| 2 | 1.2019 | 3.0516 | 5.4975 | 1.2019 | 1 |
| 3 | 1.2019 | 4.8283 | 6.2893 | 1.2019 | 1 |
| 4 | 4.0139 | 0.9014 | 4.6428 | 0.9014 | 2 |
| 5 | 3.2830 | 0.5590 | 3.3993 | 0.5590 | 2 |
| 6 | 1.3333 | 3.5089 | 4.0689 | 1.3333 | 1 |
| 7 | 4.4845 | 2.3049 | 1.4907 | 1.4907 | 3 |
| 8 | 3.8873 | 2.7042 | 1.3744 | 1.3744 | 3 |
| 9 | 5.2705 | 3.2882 | 0.7454 | 0.7454 | 3 |
| 10 | 6.6416 | 5.4601 | 1.6997 | 1.6997 | 3 |

Tabel 2.14. Centroid cluster 1 pada iterasi 3

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| 6 | 5 | 3 |
| Nk | Jumlah x | Jumlah y |
| 3 | 15 | 5 |
| Rata-rata | 5.0000 | 1.6667 |

Tabel 2.15. Centroid cluster 2 pada iterasi 3

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 1 | 1 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |
| Nk | Jumlah x | Jumlah y |
| 3 | 4 | 6 |
| Rata-rata | 1.3333 | 2.0000 |

Tabel 2.16. Centroid cluster 3 pada iterasi 3

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 7 | 5 | 3 |
| 8 | 3 | 5 |
| 9 | 3 | 6 |
| 10 | 3 | 8 |
| Nk | Jumlah x | Jumlah y |
| 4 | 10 | 24 |
| Rata-rata | 2.5000 | 6.0000 |

Tabel 2.17. Nilai fungsi objektif pada iterasi 3

| Data ke-i | Cluster 1 | Cluster 2 | Cluster 3 |
|-----------|-----------|-----------|-----------|
| 1 | 0 | 1.1111 | 0 |
| 2 | 1.4444 | 0 | 0 |
| 3 | 1.4444 | 0 | 0 |
| 4 | 0 | 0.1111 | 0 |
| 5 | 0 | 1.4444 | 0 |
| 6 | 1.7778 | 0 | 0 |
| 7 | 0 | 0 | 1.2500 |
| 8 | 0 | 0 | 1.2500 |
| 9 | 0 | 0 | 0.2500 |
| 10 | 0 | 0 | 4.2500 |

$J = 14.3333$

Perubahan $J = 19.75 - 14.33 = 5.24$

Masih Ada data yang pindah cluster atau perubahan $J > T$ maka dilanjutkan ke iterasi berikutnya.

Iterasi 4

Tabel 2.18. Jarak data ke Centroid dan Cluster yang diikuti

| Data ke-i | Jarak ke Centroid | | | Terdekat | Cluster yang diikuti |
|-----------|-------------------|--------|--------|----------|----------------------|
| | 1 | 2 | 3 | | |
| 1 | 4.0552 | 1.0541 | 5.2202 | 1.0541 | 2 |
| 2 | 1.2019 | 2.8480 | 5.2202 | 1.2019 | 1 |
| 3 | 1.2019 | 4.7726 | 6.1033 | 1.2019 | 1 |
| 4 | 4.0139 | 0.3333 | 4.2720 | 0.3333 | 2 |
| 5 | 3.2830 | 1.2019 | 3.0414 | 1.2019 | 2 |
| 6 | 1.3333 | 3.8006 | 3.9051 | 1.3333 | 1 |

| Data ke-i | Jarak ke Centroid | | | Terdekat | Cluster yang diikuti |
|-----------|-------------------|--------|--------|----------|----------------------|
| | 1 | 2 | 3 | | |
| 7 | 4.4845 | 3.0732 | 1.1180 | 1.1180 | 3 |
| 8 | 3.8873 | 3.4319 | 1.1180 | 1.1180 | 3 |
| 9 | 5.2705 | 4.0552 | 0.5000 | 0.5000 | 3 |
| 10 | 6.6416 | 6.2272 | 2.0616 | 2.0616 | 3 |

Tabel 2.19. Centroid cluster 1 pada iterasi 4

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 2 | 4 | 1 |
| 3 | 6 | 1 |
| 6 | 5 | 3 |
| Nk | Jumlah x | Jumlah y |
| 3 | 15 | 5 |
| Rata-rata | 5.0000 | 1.6667 |

Tabel 2.20. Centroid cluster 2 pada iterasi 4

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 1 | 1 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |
| Nk | Jumlah x | Jumlah y |
| 3 | 4 | 6 |
| Rata-rata | 1.3333 | 2.0000 |

Tabel 2.21. Centroid cluster 3 pada iterasi 4

| Data anggota | Fitur x | Fitur y |
|--------------|---------|---------|
| 7 | 5 | 3 |
| 8 | 3 | 5 |

| Data anggota | Fitur x | Fitur y |
|--------------|----------|----------|
| 9 | 3 | 6 |
| 10 | 3 | 8 |
| Nk | Jumlah x | Jumlah y |
| 4 | 10 | 24 |
| Rata-rata | 2.5000 | 6.0000 |

Tabel 2.22. Nilai fungsi objektif pada iterasi 4

| Data ke-i | Cluster 1 | Cluster 2 | Cluster 3 |
|-----------|-----------|-----------|-----------|
| 1 | 0 | 1.1111 | 0 |
| 2 | 1.4444 | 0 | 0 |
| 3 | 1.4444 | 0 | 0 |
| 4 | 0 | 0.1111 | 0 |
| 5 | 0 | 1.4444 | 0 |
| 6 | 1.7778 | 0 | 0 |
| 7 | 0 | 0 | 1.2500 |
| 8 | 0 | 0 | 1.2500 |
| 9 | 0 | 0 | 0.2500 |
| 10 | 0 | 0 | 4.2500 |

$J = 14.3333$

Perubahan $J = 14.33 - 14.33 = 0$

Tidak ada data yang pindah cluster atau perubahan $J < T$ maka iterasi clustering dihentikan.

Hasil Clustering

Tabel 2.23. Data yang ikut cluster 1

| Data ke | Fitur x | Fitur y |
|---------|---------|---------|
| 2 | 4 | 1 |

| Data ke | Fitur x | Fitur y |
|---------|---------|---------|
| 3 | 6 | 1 |
| 6 | 5 | 3 |

Tabel 2.24. Data yang ikut cluster 2

| Data ke | Fitur x | Fitur y |
|---------|---------|---------|
| 1 | 1 | 1 |
| 4 | 1 | 2 |
| 5 | 2 | 3 |

Tabel 2.25. Data yang ikut cluster 3

| Data ke | Fitur x | Fitur y |
|---------|---------|---------|
| 7 | 2 | 5 |
| 8 | 3 | 5 |
| 9 | 2 | 6 |
| 10 | 3 | 8 |

Tabel 2.26. Centroid

| Data ke | Fitur x | Fitur y |
|---------|---------|---------|
| 1 | 5.0000 | 1.6667 |
| 2 | 1.3333 | 2.0000 |
| 3 | 2.5000 | 6.0000 |

2.4. Sekolah Hafizh Qur'an (SHQ)

Program Sekolah Hafizh Qur'an (SHQ) adalah program pembinaan tahfizh pada anak usia 3-6 dan 7-12 tahun dengan metode *talaqqi*. Anak bisa menghafal sebelum mampu membaca Al-Quran, proses pembelajaran dilakukan oleh muhaffizh berpengalaman. Menekankan munculnya kecintaan anak dalam menghafal, sehingga proses *talaqqi* dan menghafal dikondisikan dalam suasana

menyenangkan. Tidak menitik beeratkan pada target-target tertentu, tetapi lebih kepada bagaimana melahirkan kenikmatan dalam menghafal.

2.5. Variabel yang digunakan

Kriteria yang digunakan dalam uji kemampuan ada;ah sebagai berikut :

1. Usia

Pada parameter usia dibedakan menjadi 2 kelompok yakni usia 3-6 tahun dan 7 – 12 tahun.

2. Mad (Panjang/Pendek)

Menurut lughat, arti *mad* adalah memanjangkan atau menambahkan, sedangkan menurut istilah *tadjwid*, mad adalah memanjangkan suara dengan salah satu huruf *mad* atau huruf *liin*.

Mad terbagi menjadi dua bagian yaitu :

- a. Mad Asli : *Madd-i Thabii*
- b. Mad Far'i : *Madd-i Wajib Muttasil*
Madd-i Jaiz Munfashil
Madd-i Lazim
Madd-I Aridz Lis-Sukun
Madd-I Lien

3. Izhar (Kejelasan)

Izhar adalah memisahkan antara dua huruf sambil menjauhkannya (*nun-u* sakin dan huruf *izhar*). Ketika membaca *izhar* yang perlu diperhatikan adalah menampakkan *dzat tanwin* atau *nun-u sakin* dan tidak melebihi pada sukunya.

Sebab dilakukakn *izhar* adalah *makhraj* ُ dengan huruf *izhar* yang merupakan huruf *khalq* berkejauhan sehingga perlu jarak yang diberikan pada ucapan mereka karena ُ merupakan huruf lisan sedangkan huruf *izhar* merupakan huruf tenggorokan.

4. Ikhfa (Samar)

Ikhfa adalah sebuah hal antara *izhar* dan *idgham*, dalam keadaan terbebas dari *tasydid* dan bacaannya disertai dengan *ghunnah*.

Sebab dilakukan *ikhfa* adalah *makhraj* \cup dengan *makhraj* huruf *ikhfa* tidak dekat dan tidak pula jauh sehingga kita membaca *nun* atau *tanwin* dengan *ikhfa*.

5. *Ghunnah* (Dengung)

Apabila ada huruf *mim* dan *nun* yang *bertydid*, maka cara membacanya harus dibaca dengan suara dengung. Tempat membunyikan *ghunnah* adalah pada janur hidung sehingga terkesan bunyinya seperti suara orang yang bindeng. Untuk lebih jelas lagi yaitu dengan sedikit menutup lubang hidung, sehingga akan terasa selalu getaran suara dengung pada janur hidung.

6. Jumlah Hafalan

Pada parameter jumlah hafalan calon santri diuji dengan surat Al-Qur'an yang telah dihafalkan. Pada parameter ini surat dikonversi menjadi jumlah ayat.

7. Kelancaran Menghafal

Pada parameter ini calon santri diuji untuk mengetahui kelancaran menghafal.

2.6. Evaluasi

Terdapat beberapa metode atau indeks evaluasi yang dapat digunakan untuk mengukur kualitas sebuah algoritma *clustering*. Hasil yang didapat dari beragam metode tersebut dapat berbeda karena pendekatan yang dipakai pun berbeda. Ada beberapa pendekatan kriteria yang dapat digunakan dalam melakukan evaluasi, yaitu pendekatan dengan kriteria eksternal dan pendekatan dengan kriteria internal (Karhendana, 2008).

Salah satu metode evaluasi yang bias digunakan adalah metode *Davies Bouldin Index* (DBI). DBI diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979 adalah metric untuk mengevaluasi hasil algoritma *clustering*. Pendekatan pengukuran ini untuk memaksimalkan jarak *inter-cluster* diantara *Cluster* C_i dan C_j dan pada waktu yang sama mencoba untuk meminimalkan jarak antara titik dalam sebuah *cluster*. Jarak *intra-cluster* $s_c(Q_k)$ dalam *cluster* Q_k dapat dilihat pada Persamaan 2.4.

$$s_c(Q_k) = \frac{\sum_i \|x_i - c_k\|}{N_k} \dots\dots\dots 2.4$$

Dengan N_k adalah banyak titik yang termasuk dalam *cluster* Q_k dan C_k adalah centroid dari *Cluster* Q_k . Rumus jarak *Inter-cluster* dapat dilihat pada Persamaan 2.5.

$$d_k = ||C_k - C_l|| \dots\dots\dots 2.5$$

Dengan C_k dan C_l ialah *centroid Cluster k dan cluster l*. Dilain pihak, indeks

Davies Bouldin dapat dilihat pada Persamaan 2.6.

$$DB(nc) = \frac{1}{n} \sum_{k=1}^{nc} \max \left\{ \frac{S_c(Q_k) + S_c(Q_l)}{d_{kl}(Q_k, Q_l)} \right\} \dots\dots\dots 2.6$$

Dengan nc ialah banyak *cluster*.

2.7. Penelitian Sebelumnya

Berikut ada beberapa paper yang digunakan sebagai refrensi pembelajaran yaitu sebagai berikut :

1. “Aplikasi k-means untuk pengelompokan mahasiswa berdasarkan nilai body mass index (BMI) dan ukuran kerangka” oleh Tedy Rismawan dan Sri Kusumadewi. Dalam penelitian ini penulis menjelaskan bahwa tentang masalah masalah BMI dan ukuran kerangka seseorang. Apabila seseorang telah mengetahui nilai BMInya, orang tersebut maka dapat mengontrol berat badan sehingga dapat mencapai berat badan normal yang sesuai dengan tinggi badan. Sedangkan apabila orang tersebut mengetahui ukuran kerangka tubuhnya maka orang tersebut dapat mengontrol berat badanya agar dapat selalu berada dalam keadaan ideal.

Metode yang digunakan adalah metode K-Means, hasil dari penelitian ini yaitu algoritma klasifikasi k-means dapat digunakan untuk mengelompokkan mahasiswa berdasarkan status gizi dan ukuran kerangka. (Rismawan & Kusumadewi, 2012).

2. “Pengelompokan potensi akademik siswa RA TARBIYATUL AULAD dengan metode K-Means” oleh Ba’natus Sa’adah. Penulis menelaskan masalah yang terjadi dalam penelitian ini adalah yakni perkembangan sekolah dasar yang melakukan seleksi bagi siswa TK maupun RA, hal ini dilakukan untuk

mengetahui kemampuan dan pengetahuan siswa tersebut sehingga sekolah dasar dapat mengelompokkan kemampuan siswa yang berbeda-beda.

Metode yang digunakan adalah metode K-Means, hasil dari penelitian ini menghasilkan bahwasanya metode K-Means dapat digunakan untuk mengelompokkan potensi akademik siswa RA TARBIYATUL AULAD. (Sa'adah, 2014).