BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Penelitian Terdahulu

Sebelum melakukan penelitian penulis terlebih dahulu melakukan tinjauan pustaka dari penelitian yang lain yang berkaitan dengan bantuan siswa miskin, penelitian ini bukanlah pertama kalinya. Sebelumnya sudah ada penelitian yang membahas mengenai penerimaan bantuan menggunakan algoritma K-Means, seperti di bawah ini:

Tabel 2.1 penelitian Terdahulu

Peneliti	Judul	Metode	Tahun
Pramudi Utomo	"Analisis Kontribusi Pemberian Beasiswa Terhadap Prestasi Akademik Mahasiswa"	Sampling	2011
Anggoro E. Wicaksono	"Implementasi Data Mining Dalam Pengelompokan Peserta Didik di Sekolah untuk Memprediksi Calon Penerima Beasiswa Dengan Menggunakan Algoritma K-Means",	Algorithma K-Means	2016

Irfan Ajmal Khan and Jin Tak Choi	"An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction"	Algorithma C4.5	2015
Nadiya Hijriana	"Penerapan Metode Decision Tree Algoritma C4.5 Untuk Seleksi Calon Penerima Beasiswa Tingkat Universitas"	Algorithma C4.5	2017
Dina Maurina	"Penerapan data mining untuk rekomendasi beasiswa pada sma muhammadiyah menggunakan algoritma C4.5"	algoritma C4.5	2015

- Penelitian oleh (Dina Maurina)Tahun 2015

Penelitian yang pertama adalah penelitian yang dilakukan oleh Dina Maurina yang berjudul "Penerapan data mining untuk rekomendasi beasiswa pada sma muhammadiyah menggunakan algoritma C4.5", Data yang digunakan yaitu data jurusan, kelas, jumlah nilai, penghasilan orangtua, dan jumlah saudara kandung. Proses data mining pada data training akan menghasilkan pohon keputusan atau rule. Metode evaluasi yang dilakukan dalam penelitian ini

yaitu menggunakan confusion matrix dan nilai akurasi, untuk sekali pengujian tingkat akurasi yang dihasilkan yaitu 77%. hal ini membuktikan bahwa algoritma C4.5 cukup akurat dalam menentukan rekomendasi beasiswa pada SMA Muhammadiyah.

- Penelitian oleh (Anggoro E. Wicaksono)Tahun 2016

Penelitian yang kedua adalah penelitian yang dilakukan oleh Anggoro E. Wicaksono yang berjudul "Implementasi Data Mining Dalam Pengelompokan Peserta Didik di Sekolah untuk Memprediksi Calon Penerima Beasiswa Dengan Menggunakan Algoritma K-Means", Algoritma ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda dikelompokan ke dalam cluster yang lain. Aplikasi data mining ini menampilkan hasil berupa masing-masing data yang dikelompokan berdasarkan nilai dan gaji orang tua sebagai pertimbangan untuk mendapatkan beasiswa , pengelompokan data peserta didik berdasarkan nilai dan gaji orang tua dengan menggunakan algoritma K-Means telah berhasil dibuat. Dengan menggunakan algoritma ini membantu guru dalam memproses data nilai akademik dan gaji orang tua peserta didik sehingga dalam menentukan calon penerima beasiswa dapat dengan mudah diprediksi, dan menghindari kesalahan dalam proses seleksinya.

Setelah dilihat dari beberapa sumber dalam penelitian terkait, maka penulis akan melakukan penelitian menggunakan algoritma K-Means dan perangkat lunak pengujiannya adalah rapid miner sebagai acuannya. Dengan menggunakan algoritma tersebut peneliti dapat melihat nilai akurasi, yang hasilnya akan digunakan untuk rekomendasi penerima bantuan siswa miskin. Hal ini akan mempermudah petugas dalam menentukan siapa yang berhak mendapatkan beasiswa dengan proses selektif dan tepat sasaran .

2.2 Data Mining

a. 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 Knowladge Discovery in Database (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data. KDD terdiri dari berberapa langkah (Santosa, 2007) yaitu:

- Pembersihan data (membuang noise dan data yang tidak konsisten).
- Integrasi data (penggabungan data dari berberapa sumber).
- Seleksi data (memilih data yang relevan yang akan digunakan untuk analisa).
- Data mining.
- Evaluasi model.
- Presentasi pengetahuan dengan teknik visualisasi.

b. Metode Data Mining

Pada umumnya metode data mining dapat dikelompokkan kedalam 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:

- 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.

- Clustering

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

- Association

Tujuan dari metode ini yaitu untuk menghasilkan sejumlah rule yang menjelaskan sejumlah data yang terhubung kuat dengan yang lainya. Sebagai contoh association analysis dapat digunakan untuk menentukan produk yang datang dibeli secara bersamaan oleh banyak pelanggan, atau bisa juga disebut dengan marketbase dan analysis.

- Regression

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

- Forecasting

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

- Sequence Analysis

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

- Deviation analysis

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

2.3 Algoritma K-Means

Algoritma 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 berberapa teknik clustering yang paling sederhana dan umum dikenal adalah clustering K-Means. Dalam teknik ini kita ingin mengelompokan 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 berberapa jumlah cluster yang paling tepat. Secara detail kita biasa menggunakan ukuran tidak miripan untuk mengelompokkan obyek kita. Tidak miripan bisa diterjemahkan dalam konsep jarak. Jika jarak dua obyek atau dua titik cukup dekat maka dua obyek itu mirip. Semakin dekat berarti semakin tinggi kemiripanya. Semakin tinggi jarak semakin tinggi tidak miripanya. Dalam penelitian yang mengungkapkan langkah-langkah pengerjaan algoritma K-Means (Santosa, 2007) yaitu:

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. Perhitungan jarak dengan pusat cluster untuk mengukur jarak antara data dengan pusat cluster digunakan rumus Euclidiandistance. Algoritma perhitungan jarak data dengan pusat cluster Ambil nilai data dan nilai pusatcluster. itung Euclidiandistance data dengan tiap pusat cluster.

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:

- Ambil nilai jarak tiap pusat cluster dengan data.
- Cari nilai jarak terkecil.
- Kelompokan data dengan pusat cluster yang memiliki jarak terkecil.

  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 itersi 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 :

Cari jumlah anggota tiap cluster.

Hitung pusat baru dengan rumus. Pusat cluster baru=$x1+x2+x3,\ldots xn$

$jumlah+1$

Dimana :

x1,x2,x3,…..xn adalah anggota cluster.

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

a. Contoh kasus pengelompokan dengan metode K-Means

Pengelompokan 10 yang dapat dilihat pada Tabel 2.2. 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 2.2. Data set contoh pengelompokan dengan metode K-Means

Data ke-i	Fitur x	Fitur y
1	1	1
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, misal dipilih data ke-2, data ke-4, dan data ke-6

Tabel 2.3. 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

$$D(x_1, x_1) = || x2 - x1||_2 = \sqrt{\sum_{j=1}^{p} | x_{2j} - x_{1j} |^2} \quad \text{.....2.1}$$

Jarak ke setiap centroid pada data ke-1

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

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

Tabel 2.4. 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
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 dikuti. Kemudian dilakuakan perhitungan rata-rata pada masing-masing cluster menggunakan Persamaan 2.2

$$j\frac{1}{Nk}$$

Nk adalah jumlah data yang tergabung dalam sebuah cluster

Tabel 2.5. 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

Pada table 2.5 merupakan data dari centroid 1 pada iterasi 1

Tabel 2.6. 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

Pada table 2.6 merupakan data dari centroid cluster 2 pada iterasi 1

Tabel 2.7. 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.

$$\boldsymbol{d(xi, \mu j) = \sqrt{\sum(xi - \mu j)^2}} \qquad (1)$$

Tabel 2.8. 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

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 2.9. 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

Pada tabel 2.9 Merupakan data ke centroid dan cluster yang diikuti pada iterasi 2

Tabel 2.10. Centroid cluster 1 pada iterasi 2

Data anggota	**Fitur x**	**Fitur y**
2	4	1
3	6	1
6	5	3
7	2	5
Nk	Jumlah x	Jumlah y
3	15	5
Data anggota	Fitur x	Fitur y
Rata-rata	5.0000	1.6667

Pada Tabel 2.10 merupakan centroid cluster 1 pada iterasi 2

Tabel 2.11. 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

Pada table 2.11 merupakan centroid cluster 2 pada iterasi 2

Tabel 2.12. 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 merupakan centroid cluster 3 pada iterasi 2

Tabel 2.13.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
7	0	5.3125	0
8	0	0	1.8889

9	0	0	0.5556
10	0	0	2.8889

J = 19.75

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 2.14 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 merupakan Jarak data ke Centroid dan Cluster yang diikuti pada iterasi 3

Tabel 2.15. 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

Pada tabel 2.15 merupakan centroid cluster 1 pada iterasi 3

Tabel 2.16 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

Pada Tabel 2.16 merupakan Centroid cluster 2 pada iterasi 3

Tabel 2.17 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

Pada tabel 2.17 merupakan centroid cluster 3 pada iterasi 3

Tabel 2.18 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

Tabel 2.19 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
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 merupakan jarak data ke centroid dan cluster yang diikuti

Tabel 2.20 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 merupakan centroid cluster 1 pada iterasi 4

Tabel 2.21 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 merupakan centroid cluster 2 pada iterasi 4

Tabel 2.22 Centroid cluster 3 pada iterasi 4

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 2.22 merupakan Centroid cluster 3 pada iterasi 4

Tabel 2.23 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 2.24 Data yang ikut cluster 1

Data ke	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3

Tabel 2.24 merupakan data yang ikut kedalam cluster 1

Tabel 2.25 Data yang ikut cluster 2

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

Table 2.25 merupakan data yang ikut kedalam cluster 2

Tabel 2.26 Data yang ikut cluster 3

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

Table 2.26 merupakan data yang ikut kedalam cluster 3

Tabel 2.27 Centroid

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

Tabel 2.27 merupakan tabel dari centroid

2.4 Integration Data

Integration data adalah pengubahan data menjadi format yang sesuai untuk digunakan dalam proses data mining. Contoh : Setelah menemukan batu-batu yang cocok, selanjutnya penambang akan mulai mengkombinasikan untuk dijadikan batangan emas atau bentuk emas lainnya, Dalam data mining, data yang berhasil dibersihkan juga akan diintegrasi.

2.5 Cleaning Data

Data cleaning adalah serangkaian proses untuk mengidentifikasi kesalahan pada data dan kemudian mengambil tindakan lanjut, baik berupa perbaikan ataupun penghapusan data yang tidak sesuai. Prosedur data cleaning dilakukan untuk memastikan kualitas data yang digunakan, untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Caranya adalah dengan mendeteksi adanya eror atau corrupt pada data, kemudian memperbaiki atau menghapus data jika memang diperlukan.

Terkadang, saat Anda menggabungkan beberapa data sources sekaligus, ada kemungkinan data terduplikasi atau bahkan salah label. Situasi seperti ini juga memerlukan data cleaning agar tidak muncul masalah yang lebih rumit.

2.6 Tool yang Digunakan

2.6.1 Rapid Minner

Rapid Minner berkembang sejak tahun 2001, sebelumnya disebut dengan nama YALE (YetAnotherLearningEnvironment). Software ini dikembangkan oleh Ralf Klinkenberg, Ingo Mierswa, serta Simon Fischer pada Unit Artificial Intelligence dari Technical University of Dortmund. Rapid Miner adalah platform analisis modern yang meliputi data mining, mechine learning, analisis prediktif, text mining dan analisis bisnis [12]. Software ini digunakan untuk mengukur kinerja algoritma dan untuk menemukan algoritma terbaik yang akan berguna untuk klasifikasi, prediksi dan teknik lainya di data mining. RapidMiner merupakan software yang userfriendly dan memiliki GUI.

(Graphic User Interface) yang efektif yang digunakan untuk bekerja dengan mudah. RapidMiner memberikan mechine learning dan data prosedur termasuk loading data dan transformasi (Extract, Transform, Load (ELT)), data preprocessing dan pemodelan statistik, visualisasi dan analisis prediktif, penyebaran dan evaluasi. Terdapat sifat-sifat yang dimiliki oleh Rapid Miner, yakni :

1. Penulisan menggunakan bahasa Java. Hal ini memungkinkan RapidMiner bisa berjalan pada sistem operasi yang berbeda-beda.
2. Proses menemukan pengetahuan dituangkan dalam model operator trees.
3.Merepresentasikan XML internal guna memungkinkan format standar pertukaran data.
4. Penggunaan bahasa scripting yang memungkinkan untuk eksperimen dalam skala besar dan pengotomatisasian eksperimen.
5. Konsep multi-layer yang menjadikan tampilan data menjadi efisien serta memastikan penanganan data.
6. Mempunyai GUI (Graphic User Interface), command line mode, serta Java API yang bisa dipanggil melalui program lain

- **Fitur Rapid Minner**

Adapun fitur-fitur yang terdapat pada RapidMiner adalah sebagai berikut :

1. Terdapat banyak algoritma data mining seperti decision tree dan self-orga nization map.
2. Bentuk grafis yang handal seperti tumpang tindih diagram historgram, tree chart dan 3d Scatter plots.
3. Memiliki banyak variasi plugin, seperti text plugin yang dapat digunakan untuk melakukan analisis teks.
4. Tersedia prosedur data mining dan machine learning termasuk ELT, data preprocessing, visualisasi, modelling dan evaluasi.
5. Proses data mining disusun berdasarkan operator-operator yang nestable, dideskripsikan dengan XML, dan dibangun dengan GUI.
6. Mengintegrasikan proyek data mining Weka dan statistik R.

2.7 Bantuan Siswa Miskin

Bantuan Siswa Miskin adalah pemberian berupa bantuan keuangan yang diberikan kepada perorangan yang bertujuan untuk digunakan demi keberlangsungan pendidikan yang ditempuh. Bantuan Siswa Miskin dapat diberikan oleh Lembaga pemerintah, perusahaan ataupun yayasan. Beasiswa merupakan suatu bantuan untuk membantu pelajar atau mahasiswa yang masih sekolah atau kuliah supaya mereka bisa menyelesaikan tugasnya dalam mencari ilmu pengetahuan sampai selesai. Beasiswa dalam bentuk bantuan dapat berupa dana sebagai penunjang biaya yang harus dikeluarkan oleh pelajar atau mahasiswa selama menempuh masa Pendidikan di tempat belajar. Bantuan Siswa Miskin di SMKN SUKOHARJO diperuntukan bagi siswa yang rentan atau miskin dan yatim piatu.

2.8 Davies Bouldin Indeks

Davies Bouldin Indeks merupakan salah satu metode evaluasi internal yang mengukur evaluasi cluster pada suatu metode pengelompokan yang didasarkan pada nilai kohesi dan separasi. Dalam suatu pengelompokan , kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap centroid dari cluster yang diikuti. Sedangkan separasi didasarkan pada jarak antar centroid dari clusternya, semakin kecil nilai DBI yang diperoleh non-negatif = 0, maka semakin baik cluster yang di peroleh dari pengelompokan K-Means yang digunakan.