PENERAPAN MODEL *K-MEAN CLUSTERING* UNTUK MENGOPTIMALKAN KELAS DATA *TRAINING* PADA ALGORITMA *K-NN CLASSIFICATION*

Bahar[1], Soegiarto[2]

[1,2]Program Studi Teknik Informatika, STMIK Banjarbaru
[1]082157782162, bahararahman@gmail.com
[2]08125003550, ttsoegiarto.@gmail.com

Abstract

Classification algorithm model that uses standard logic to generate a knowledge base has a weakness, namely the training data sets tends to be forced to enter the class of a particular data set up already, so often an object class of the data did not reflect fully the nature of the existing classes (unnatural), this resulted in the classification process will be less accurate. This article formulates a model of classification using clustering algorithm for the establishment of a naturally training class data as Knowledge Base System, so as to solve the problems on the training data modeling techniques that produces an object class which tends to be imposed.

Key words: *classification, clustering, K-Means algorithm, K-NN algorithm, Knowledge Base*

1. Pendahuluan

Pentingnya pemecahan masalah dalam suatu manajemen bukan didasarkan pada kuantitas waktu yang dihabiskan, akan tetapi pada konsekwensinya, yaitu apakah pemecahan masalah oleh manajemen dapat menekan sebanyak mungkin kemungkinan kerugian atau memperoleh sebesar mungkin kemungkinan keuntungan. Dalam usaha memecahkan suatu masalah, manajer atau pemecah masalah mungkin membuat banyak keputusan. Pengambil keputusan sering menggunakan intuisi dalam proses pengambilan keputusan, padahal dengan intuisi banyak memiliki kekurangan sehingga perlu dikembangkan suatu model pengambilan keputusan yang diambil berdasarkan informasi yang telah diolah dan disajikan dengan dukungan model penunjang keputusan[1].

Ada banyak model penunjang keputusan dalam dunia komputasi yang disesuaikan dengan bidang atau masalah yang akan diselesaikan, misalnya: masalah penetapan skala perioritas, masalah prediksi, masalah klasifikasi dan pengelompokan, masalah asosiasi, dan masalah-masalah lainnya. Umumnya masalah-masalah tersebut berkaitan dengan permasalahan bisnis, misalnya segmentasi pelanggan/produk, asosiasi produk, klasifikasi transaksi bisnis, pengelompokan pelanggan berdasarkan kesamaan karakteristik, prediksi perilaku bisnis di masa mendatang, klasifikasi penerima beasiswa, penetapan skala perioritas pembiayaan, dan lain-lain [2][3][4][5][6]. Model-model penunjang keputusan tersebut memanfaatkan konsep data mining untuk menggali informasi yang berguna, yang tersembunyi dalam tumpukan data suatu organisasi dari tahun ke tahun[7].

Salah satu model penunjang keputusan yang biasa digunakan oleh organisasi adalah prediksi, baik prediksi dalam rentang waktu diskrit (klasifikasi) maupun yang bersifat kontinu (regresi). Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas target yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu (1) pembangunan model berdasarkan data latih sebagai prototip untuk disimpan sebagai memori atau basis pengetahuan sistem, dan (2) penggunaan model atau pengetahuan tersebut untuk melakukan pengenalan/klasifikasi/prediksi suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya [7]. Ada dua cara yang umumnya digunakan pada tahap pembangunan model berdasarkan data latih, yaitu: (1) belajar dari kasus atau kejadian ril yang pernah terjadi di alam nyata berdasarkan hubungan sebab akibat / fakta, dan (2) model dihasilkan dari algoritma model (algoritma pelatihan)[8]. Kedua cara tersebut masing-masing memiliki kelebihan dan kekurangan. Kelebihan dari cara pertama adalah model dihasilkan dengan cara yang cepat karena tidak diperlukan proses pembelajaran (*lazy learner*) [7], namun kelemahannya adalah tidak semua pengetahuan dapat diperoleh berdasarkan kejadian dari alam nyata (fakta), sehingga proses klasifikasi tidak dapat dilakukan pada kasus seperti ini. Kelebihan cara kedua adalah model secara pasti dapat dihasilkan dari suatu teknik pemodelan tertentu (missal: model logika standar If... Then ...), namun kelemahannya adalah data yang dimodelkan berdasarkan model ini cenderung dipaksa untuk masuk pada klas tertentu yang sudah diset sebelumnya, sehingga sering terjadi sebuah objek klas tidak mencerminkan sepenuhnya sifat dari kelas yang ada[2].

Paper ini merumuskan sebuah model Klasifikasi berbasis algoritma *K-NN Classification* menggunakan Algoritma Klastering (*K-Means Clustering*) untuk pembentukan kelas data pembelajaran secara alami sebagai Basis Pengetahuan sistem, sehingga disamping dapat menyelesaikan masalah tidak tersedianya fakta sebagai Basis Pengetahuan sistem, juga dapat menyelesaikan permasalahan pada teknik

pemodelan yang menghasilkan objek kelas yang cenderung dipaksakan.

2. Metodologi

Algoritma *K-NN classification* yang menggunakan algoritma Klastering *K-Means* untuk pembentukan kelas data pelatihan dengan pencarian jarak menggunakan rumus *Euclidian*, sebagai berikut:

1. Menentukan Nilai n buah data awal sebagai basis pengetahuan awal sistem
2. Menentukan Nilai K (tetangga) terdekat
3. Mempersiapkan data training berupa nilai kriteria suatu data baru yang belum diketahui statusnya.
4. Menentukan status (kelas data) setiap data training menggunakan formula *K-Mean Clustering* [3] sebagai basis pengetahuan sistem
5. Menghitung jarak setiap sampel data traning terhadap data yang akan diuji (data uji) berdasarkan persamaan:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad \text{..... (1)}$$

Dengan:

d = jarak
xi = sampel data
yi = data uji
i = variabel data
n = dimensi data

6. Menetapkan status data uji berdasarkan nilai rata-rata K buah sampel data traning terdekat.

3. Pembahasan

Misalkan terdapat data calon mahasiswa penerima beasiswa sebagai data pelatihan seperti pada table 1.

Tabel 1. *Sampel Set Data Pelatihan*

(1)	(2)	(3)	(4)
1	3,55	2.400.000	80
2	2,75	900.000	75
3	3,15	950.000	80
4	2,75	700.000	70
5	3,45	750.000	85
6	3,24	500.000	70
7	3,02	500.000	80
8	3,38	1.500.000	75
9	3,18	850.000	75
10	3,25	900.000	70

Keterangan:
(1) : Mahasiswa
(2) : Indeks Prestasi Kumulatif (IPK)
(3) : Penghasilan Kepala Keluarga (PKK)
(4) : Nilai Keaktipan dalam Berorganisasi (NKB)

Set data pada table 1 akan ditentukan kelas datanya menggunakan algoritma *K-Mean Clustering*. Berdasar parameter IPK, PKK dan NKB ditetapkan status masing-masing calon penerima beasiswa, akan menjadi anggota klaster penerima beasiswa "Berprestasi" atau beasiswa "Kurang Mampu". Dengan menggunakan algoritma *K-Mean*, dihasilkan matriks dengan 2 *centroid* yaitu:

	IPK	PKK	NKB
C1	3,46	1.950.000	77,5
C2	1,08	318.750	28

Misalkan aturan umum yang dijadikan acuan untuk memberikan label pada setiap klaster yang dihasilkan oleh algoritma *K-Mean* adalah: jika Prestasi Akademik (IPK, Keaktipan Berorganisasi) lebih menonjol, maka seorang mahasiswa akan cenderung mendapatkan jenis Beasiswa "Berprestasi". Sebaliknya, jika ketidakmampuan perekonomian keluarga lebih menonjol, maka seorang mahasiswa akan cenderung mendapatkan jenis beasiswa "Kurang Mampu".

Berdasarkan *centroid* yang terbentuk, dapat diinterprestasikan:

(1) Centroid pertama (C1) adalah mahasiswa dengan nilai IPK sekitar 3,46; PKK sekitar Rp.1.950.000 dan nilai NKB sekitar 77,5 yang dapat diinterprestasikan sebagai mahasiswa penerima beasiswa "Berprestasi", yaitu mahasiswa 1 dan 8.

(2) Centroid kedua (C2) adalah mahasiswa dengan nilai IPK sekitar 1,08; PKK sekitar Rp. 318.750 dan nilai NKB sekitar 28 yang dapat diinterprestasikan sebagai mahasiswa penerima beasiswa "Kurang Mampu", yaitu mahasiswa 2, 3, 4, 5, 6, 7, 9 dan 10.

Hasil penentuan kelas data secara lengkap disajikan pada table 2.

Tabel 2. *Hasil Penentuan Kelas Data Menggunakan Algoritma K-MeanClustering*

(1)	(2)	(3)	(4)	(5)
1	3,55	2.400.000	80	Berprestasi
2	2,75	900.000	75	Kurang Mampu
3	3,15	950.000	80	Kurang Mampu
4	2,75	700.000	70	Kurang Mampu
5	3,45	750.000	85	Kurang Mampu
6	3,24	500.000	70	Kurang Mampu
7	3,02	500.000	80	Kurang Mampu
8	3,38	1.500.000	75	Berprestasi
9	3,18	850.000	75	Kurang Mampu
10	3,25	900.000	70	Kurang Mampu

Keterangan:
(1) : Mahasiswa
(2) : Indeks Prestasi Kumulatif (IPK)
(3) : Penghasilan Kepala Keluarga (PKK)
(4) : Nilai Keaktipan dalam Berorganisasi (NKB)
(5) : Jenis Beasiswa Yang Diperoleh (JB)

Data pada table 2 akan menjadi Basis Pengetahuan bagi algoritma *K-NN Classification* dalam proses penentuan

kelas data bagi data baru (data uji) yang belum diketahui status kelasnya.

Misalkan terdapat sebuah data baru (mahasiswa ke-11) yang belum diketahui kelasnya: IPK=2,90; PKK=1.650.000 dan NKB=77,5; akan ditetapkan jenis beasiswa yang akan diperoleh. Dengan menetapkan nilai K tetangga terdekat (K=3), dapat dihitung jarak data uji terhadap setiap data training menggunakan persamaan (1):

Jarak data uji terhadap data training 1 adalah:

$$d1 = \sqrt{(3{,}55 - 2{,}90)^2 + (2.400.000 - 1.650.000)^2 + (80 - 77{,}5)^2}$$

$$= 750.000$$

Jarak data uji terhadap data training 2 adalah:

$$d2 = \sqrt{(2{,}75 - 2{,}90)^2 + (900.000 - 1.650.000)^2 + (75 - 77{,}5)^2}$$

$$= 750.000$$

Demikian seterusnya untuk menghitung jarak data uji terhadap data training ke-3 hingga data training ke-10. Hasil selengkapnya disajikan pada table 3.

Tabel 3. *Hasil Perhitungan Jarak Data Uji terhadap Data Training*

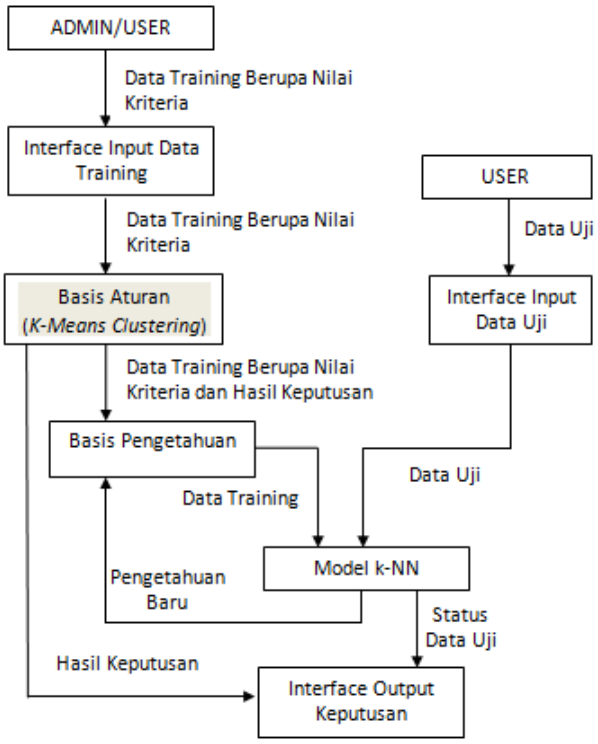
Mhs	IPK	PKK	NKB	JB	Jarak
1	3,55	2.400.000	80	Berprestasi	750,000
2	2,75	900.000	75	Kurang Mampu	750,000
3	3,15	950.000	80	Kurang Mampu	700,000
4	2,75	700.000	70	Kurang Mampu	950,000
5	3,45	750.000	85	Kurang Mampu	900,000
6	3,24	500.000	70	Kurang Mampu	1,150,000
7	3,02	500.000	80	Kurang Mampu	1,150,000
8	3,38	1.500.000	75	Berprestasi	150,000
9	3,18	850.000	75	Kurang Mampu	800,000
10	3,25	900.000	70	Kurang Mampu	750,000

Dengan merujuk pada nilai K=3, maka dapat ditetapkan 3 data dengan jarak terpendek, yaitu:

(1) mahasiswa nomor 3 (jarak=700.000) dengan jenis beasiswa (JB) adalah “Kurang Mampu”;
(2) mahasiswa nomor 1 (jarak = 750.000) dengan jenis beasiswa “Berprestasi”;
(3) mahasiswa nomor 2 (jarak = 750.000) dengan jenis beasiswa “Kurang Mampu”.

Dari 3 data dengan jarak terpendek, terdapat 2 data yang berstatus JB=”Kurang Mampu” dan 1 data berstatus JB=”Berprestasi”. Dengan demikian, status data uji (mahasiswa ke-11) adalah mendapatkan beasiswa “Kurang Mampu”.

Arsitektur Model Klasifikasi yang menggunakan algoritma Klastering untuk pembentukan kelas data pelatihan secara alami sebagai Basis Pengetahuan sistem disajikan pada gambar 1.



Gambar 1. *Arsitektur Model Klasifikasi menggunakan algoritma Klastering untuk pembentukan kelas data pelatihan*

Pada gambar 1, proses dimulai dari Admin atau User menginputkan nilai parameter data yang ditetapkan sebagai data pembelajaran (training) melalui *interface* input data training. Data ini diproses menggunakan algoritma *K-Mean Clustering* pada Basis Aturan untuk menghasilkan kelas data sebagai pengetahuan pada Basis Pengetahuan sistem. Selanjutnya Pengguna dapat menginputkan data baru (data uji) yang akan diketahui statusnya melalui *interface* data uji. Model k-NN memproses data dan menghasilkan output keputusan berupa status data baru yang diuji. Status data baru tersebut juga akan menambah perbendaharaan pengetahuan sistem klasifikasi sebagai pengetahuan baru.

4. Kesimpulan

Penggunaan model Klastering untuk menentukan status kelas pada data pelatihan algoritma Klasifikasi, dapat menyelesaikan permasalahan pemodela data pelatihan yang menghasilkan objek kelas yang cenderung dipaksakan.

Pada artikel ini baru diujicoba satu model klastering, yaitu algoritma *K-Means* Klastering untuk menentukan status kelas pada data pelatihan algoritma klasifikasi. Masih perlu dilakukan ujicoba menggunakan model-model Klastering lainnya dan membandingkan hasilnya dengan penggunaan model *K-Mean* Klastering, agar diperoleh hasil yang lebih baik dalam proses penentuan status kelas data pelatihan pada model Klasifikasi.

Daftar Pustaka

[1] Marimin, *Aplikasi Teknik Pengambilan Keputusan Dalam Manajemen Rantai Pasok*, Bogor*:* IPB Press, 2010.

[2] N. Rosmawanti, Bahar, "Model K-Nearest Neighbor Menggunakan Kombinasi Basis Aturan dan Basis Pengetahuan", dalam *Prosiding Seminar Nasional Ilmu Komputer 2014*, Yogyakarta, 2014.

[3] Kusrini, T.L. Amha, *Algoritma Datamining*, Yogyakarta: Penerbit ANDI, 2008.

[4] R. Tedy, W.I. Ardhitya, P. Wahyu, Sri Kusumadewi, "Sistem Pendukung Keputusan Berbasis Pocket PCSebagai Penentu Status Gizi Menggunakan Metode k_Nearest Neighbor", *Jurnal Teknoin*, vol. 13, no. 2, pp:1-6, 2008.

[5] I. Chusnul, "Penerapan Case Based Reasoning Dengan Algoritma Nearest Neighbor Untuk Analisis Pemberian Kredit di Lembaga Pembiayaan", *Jurnal Manajemen Informatika*, vol. 2, no. 1, pp:11-21, 2013.

[6] K. Nobertus, Helmi, P. Bayu, "Algoritma k-Nearest Neighbor Dalam Klasifikasi Data Produksi Kelapa Sawit Pada PT. Minamas Kecamatan Parindu", *Jurnal Mimaster*, vol.2, no.1, pp:33-38, 2013.

[7] E. Prasetyo, 2014, *Data Mining – Mengelolah Data Menjadi Informasi*, Yogyakarta, Penerbit ANDI, 2014.

[8] B. Santosa, *Data Mining – Teknik Pemanfaatan Data Untuk Keperluan Bisnis*, Yogyakarta: Graha Ilmu, 2007.