

PEMETAAN SISWA BERPRESTASI MENGGUNAKAN METODE *K-MEANS CLUSTERING*

Fitri Larasati Sibuea¹, & Andy Sapta²

^{1,2}Sistem Informasi, STMIK Royal

email: ¹bukmus.inaction@gmail.com, ²saptaandy@gmail.com

Abstract: The high level of student success and the low level of student failure is a quality of the education world. The world of education is currently required to have the ability to compete by utilizing all resources owned. In addition to facilities, infrastructure and human resources, information systems are one of the resources that can be used to improve competency skills. Data mining is a process of data analysis to find a dataset of data set. Data mining is able to analyze large amounts of data into information that has meaning for decision supporters. One process of data mining is clustering. Attributes used in the grouping of student achievement are Name, Extracurricular, Value which include Task Value, Uts Value, Value of Uses, total absenteeism, and Attitude value. The case study of 20 students with distance calculation using manhattan distance, chbychep distance and euclidian distance yielded 67% accuracy.

Keywords: data mining, clustering, k-means, student achievement

Abstrak: Tingginya tingkat keberhasilan siswa dan rendahnya tingkat kegagalan siswa merupakan cemin kualitas dunia pendidikan. Dunia pendidikan saat ini dituntut untuk memiliki kemampuan bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Selain sumber daya sarana, prasarana dan manusia, sistem informasi merupakan salah satu sumber daya yang dapat digunakan untuk meningkatkan kemampuan bersaing. Data mining merupakan proses analisa data untuk menemukan suatu pola data kumpulan data. Data mining mampu menganalisa jumlah data yang besar menjadi informasi yang mempunyai arti bagi pendukung keputusan. Salah satu proses data mining adalah clustering. Atribut yang digunakan dalam pengelompokan prestasi siswa adalah Nama, Ekstrakurikuler, Nilai yang meliputi Nilai Tugas, Nilai Uts, Nilai Uas, jumlah ketidakhadiran siswa (absensi), dan Nilai sikap. Studi kasus pada 20 siswa dengan perhitungan jarak menggunakan manhattan distance, chbychep distance dan euclidian distance menghasilkan akurasi sebesar 67%.

Kata kunci: data mining, *clustering*, *k-means*, prestasi siswa

PENDAHULUAN

Dewasa ini pengolahan data elektronika telah menjadi kebutuhan yang sangat utama. Perkembangan pesat dalam

teknologi informasi yang menjadikan semua informasi dapat disimpan dalam jaringan komputer telah membuat munculnya sistem basis data yang sangat besar. Dalam hitungan detik, data-data

dalam berbagai basis data akan senantiasa terbaru, baik dikarenakan adanya *update* maupun penambahan data baru. Permasalahan yang kemudian muncul adalah bagaimana mengetahui informasi yang terdapat dalam basis data yang sangat besar.

Knowledge discovery in database (KDD) didefinisikan sebagai ekstraksi informasi potensial, implisit dan tidak dikenal dari sekumpulan data. Proses *knowledge discovery* melibatkan hasil dari proses data mining (proses mengekstrak kecenderungan pola suatu data), kemudian mengubah hasilnya secara akurat menjadi informasi yang mudah dipahami.

Ada beberapa macam pendekatan berbeda yang diklasifikasikan sebagai teknik pencarian informasi/pengetahuan dalam KDD. Ada pendekatan kuantitatif, seperti pendekatan probalistik and statistik. Beberapa pendekatan memanfaatkan teknik visualisasi, pendekatan klasifikasi seperti logika induktif, pencarian pola, dan analisis pohon keputusan. Pendekatan yang lain meliputi deviasi, analisis kecendrungan, algoritma genetik, jaringan syaraf tiruan dan pendekatan campuran dua atau lebih dari beberapa pendekatan yang ada.

Tingginya tingkat keberhasilan siswa dan rendahnya tingkat kegagalan siswa merupakan cermin kualitas dunia pendidikan. Dunia pendidikan saat ini dituntut untuk memiliki kemampuan bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Selain sumber daya sarana, prasarana dan manusia. Sistem informasi dapat digunakan untuk meningkatkan kemampuan bersain. Sistem informasi dapat digunakan untuk menunjang kegiatan operasional sehari-hari sekaligus menunjang kegiatan pengambilan keputusan strategis.

Metode pengambilan keputusan konvensional yang ada, tidak dapat menangani data dalam jumlah yang sangat besar. Hal ini mendorong munculnya cabang ilmu baru untuk mengatasi masalah penggalan informasi

yang penting dari data kumpulan data, yang disebut dengan data mining.

Secara umum penilaian prestasi siswa yaitu mata pelajaran baik teori dan praktek, penilaian ekstrakurikuler kehadiran dan ketidakhadiran siswa dan pada saat mengikuti ekstrakurikuler, yang dikategorikan dalam nilai pengetahuan, nilai keterampilan dan penilaian sikap. Evaluasi dan penilaian terhadap prestasi siswa dilakukan dengan memberi nilai oleh pengajar kepada semua siswa yang mengikuti pelajaran yang diajarkan dan ekstrakurikuler yang diikutinya.

Seiring dengan terus bertambahnya jumlah data siswa setiap tahun, maka jumlah data yang siswa yang terus meningkat sehingga penumpukan data yang belum diolah dengan optimal untuk menggali informasi dan pengetahuan baru melalui pola-pola yang terbentuk dari penumpukan data tersebut. Jumlah data yang terus meningkat ini merupakan beberapa teknik ataupun metode untuk mengolahnya menjadi sebuah informasi dan pengetahuan yang dapat digunakan sebagai bahan pertimbangan pendidik dalam proses pengambilan kebijakan dan keputusan juga sebagai peringatan dini (*early warning*) bagi siswa tertentu yang berdasarkan hasil pengelompokan prestasi rendah yang berpotensi terhadap ketidaklulusan siswa.

Beberapa penelitian yang telah dilakukan melakukan pemodelan mengenai pemodelan aturan dalam memprediksi akademik siswa, mengevaluasi kinerja akademik mahasiswa pada tahun ke-2 dan diklasifikasikan dalam kategori mahasiswa yang dapat lulus tepat waktu atau tidak, konsep pengclasteran dengan pola data yang sudah diatur, tehnik cluster membuat pengelompokan data iklim tropis di samudra hindia bagian utara.

Data Mining

Data mining merupakan salah satu disiplin ilmu yang digunakan untuk menentukan suatu informasi tertentu dalam sekumpulan data sebagai pen-

dukung pengambilan keputusan. Data mining juga sebagai salah satu proses untuk memperoleh informasi yang memiliki nilai guna dari sekumpulan data (Tan, 2006).

Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database. Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan machine learning untuk menginteraksi mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terikat dari berbagai database besar (Kusrini, 2009).

Data Mining merupakan proses menemukan kolerasi baru yang bermanfaat, pola dan trend dengan manmbang sejumlah repositori data dalam jumlah besar, menggunakan teknologi pengenalan pola seperti statisik dan teknik matematika. Data Mining semakin menyebar dan berkembang dengan pesat belakangan ini karena kemampuannya dalam menambang pola bermanfaat dan trend dari basis data yang sudah ada. Perusahaan-perusahaan telah menghabiskan dana milyaran untuk mengumpulkan data dalam jumlah megabytes atau terabytes tapi tidak mendapatkan keuntungan yang bernilai didalamnya, padahal didalamnya terbatat informasi yang berharga namun tersembunyi pada repositori data. (Larose, 2005)

Menurut Prastyo (2012), data mining terbagi menjadi beberapa kelompok yaitu sebagai berikut:

1. Model Prediksi
Pemodelan yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang telah didapat.
2. Analisis Kelompok
Mengelompokkan sekumpulan data ke dalam kelompok (*cluster*) tertentu berdasarkan karakteristik yang sama pada masing-masing data.
3. Analisis Asosiasi

Menentukan pola yang menggambarkan kekuatan hubungan fitur dalam data.

4. Deteksi Anomaly
Pengamatan suatu data dari sekumpulan data yang secara signifikan memiliki karakteristik yang berbeda dari sisa data yang lain. (Larose, 2005)

Data mining telah dipakai diberbagai bidang seperti ilmu sains, bisnis dan industri, teknik, kesehatan, serta pertahanan dan keamanan. Menurut (Ayub, 2007) Ketersediaan data yang berlimpah yang dihasilkan dari penggunaan teknologi informasi di hampir semua bidang kehidupan, menimbulkan kebutuhan untuk dapat memanfaatkan informasi dan pengetahuan yang terkandung didalam limpahan data tersebut, yang kemudian melahirkan data mining. data mining merupakan proses untuk menemukan pengetahuan (*knowledge discovery*) yang ditambang dari sekumpulan data yang volumenya sangat besar.

Algoritma *K-Means*

K-Means merupakan suatu algoritma pengklasteran yang cukup sederhana yang mempartisi databest kedalam beberapa clasteran *k*. Algoritma cukup mudah untuk diimplementasikan dan dijalankan, relatif cepat, mudah disesuaikan dan banyak digunakan (Wu & Kumar, 2009). Prinsip utama dari teknik ini adalah menyusun *K* buah partisi/pusat massa (*centroid*)/rata-rata (*mean*) dari sekupulan data. Algoritma *K-Means* dimulai dengan pembentukan partisi klaster diawal kemudian secara iteraktif partisi *cluster* ini diperbaiki hingga tidak terjadi perubahan yang signifikan pada partisi *cluster* (Written, 2011).

Langkah pertama algoritma *k-means* adalah menentukan jumlah cluster, padga penelitian ini ditetapkan 3 *cluster*, yang dipilih secara random. dengan variabel jumlah ekstrakulikuler, nilai rata-rata, Absen.

Ada banyak metode yang dapat digunakan dalam pengelompokan contohnya metode K-means. Pengelompokan yang dapat digunakan seperti pengelompokan non hierarki yang membagi data kedalam bentuk dua atau lebih kelompok. K-means merupakan metode analisis kelompok yang mengarah pada pembagian N objek pengamatan kedalam K kelompok (*cluster*) dan setiap objek pengamatan dimiliki oleh suatu kelompok dengan rata-rata (*mean*) terdekat (Prasetyo, 2012).

Dalam menentukan nilai *centroid* untuk awal iterasi, nilai awal *centroid* dilakukan secara acak. Sedangkan jika menentukan nilai *centroid* yang merupakan tahap dari iterasi, maka digunakan rumus sebagai berikut:

$$\bar{v}_j = \frac{1}{N_j} \sum_{k=0}^{N_j} x_{kj}$$

1. Menghitung jarak antara titik *centroid* dengan titik tiap objek.

$$D_o = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

2. Pengelompokan object untuk menentukan anggota cluster adalah dengan memperhitungkan jarak minimum objek.
3. Kembali ke tahap ke-2, lakukan perulangan hingga nilai *centroid* yang dihasilkan tetap dan anggota cluster tidak berpindah ke cluster lain.

Metode *K-Means Clustering* hanya bisa mengolah data dalam bentuk angka, maka untuk data yang berbentuk nominal harus diinisialisasikan terlebih dahulu ke dalam bentuk angka. Langkahnya adalah:

1. Urutkan data berdasarkan frekuensi kemunculannya.
2. Inisialisasikan data tersebut mulai dari data tertinggi dengan nilai 1. Kemudian data selanjutnya 2, 3 dan seterusnya.

Kelemahan *K-Means*:

1. Bila jumlah data tidak terlalu banyak, mudah untuk menentukan cluster awal.
2. Jumlah *cluster*, sebanyak K , harus ditentukan sebelum dilakukan perhitungan.
3. Tidak pernah mengetahui real cluster dengan menggunakan data yang sama, namun jika dimasukkan dengan cara yang berbeda mungkin dapat memproduksi cluster yang berbeda jika jumlah datanya sedikit.
4. Tidak tahu kontribusi dari atribut dalam proses pengelompokan karena dianggap bahwa setiap atribut memiliki bobot yang sama.

RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *RapidMiner* memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data *preprocessing* dan visualisasi. *RapidMiner* merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. *RapidMiner* ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

METODOLOGI

Clustering merupakan salah satu teknik dari salah satu fungsionalitas data mining, algoritma *clustering* merupakan algoritma pengelompokan sejumlah data menjadi kelompok-kelompok data tertentu (*cluster*).

Setiap *cluster* memiliki *centroid* yang merupakan suatu besaran yang

dihitung dari rata-rata nilai tiap items dari suatu *cluster* dan juga memiliki *medoid* yang merupakan item yang letaknya paling tengah. Sementara jarak antar *cluster* didefinisikan dengan menggunakan beberapa metode-metode untuk menentukan *cluster* mana yang berdekatan. (Kusrini, 2009)

1. *Single Link* adalah jarak terkecil antara satu elemen *cluster* dalam suatu *cluster* dengan elemen dalam suatu *cluster* dengan elemen lain di *cluster* yang berbeda.
2. *Complete Link* adalah jarak terbesar antara satu elmen dalam suatu *cluster* dengan elemen lain di *cluster* yang berbeda.
3. *Average* adalah jarak rata-rata antara satu elemen dalam suatu *cluster* dengan elemen lain di *cluster* yang berbeda.
4. *Centroid* adalah jarak anatar *centroid* dari tiap *cluster* dengan *centroidcluster* lainnya.
5. *Medoid* adalah jarak antara *medoid* dari tiap *cluster* dengan *medoid cluster* lainnya.

HASIL DAN PEMBAHASAN

Data yang diolah dalam penelitian ini merupakan sampel yang diambil dari data raport peserta didik SMK Yapim Simpang Kawat kelas X AK2 tahun ajaran 2016/2017. Dataset siswa terdiri dari atribut data induk dan data raport. Atribut yang dijadikan variable dalam algoritma ini di sederhanakan menjadi atribut yang terdiri dari NIS, Nama, Ekstrakurikuler, Nilai Tugas, Uts , Uas dan Jumlah ketidakhadiran. Data sampel yang akan diuji cobakan terdiri dari 10 peserta didik.

Selanjutnya dicoba mengelompokkan data diatas menjadi 3 kelompok. Dengan menggunakan algoritma K-means, berikut langkah-langkah penyelesaiannya:

1. Menentukan Jumlah *cluster*, jumlah *cluster* merupakan jumlah kelompok yan akan dihasilkan. Dalam penelitian ini jumlah *cluster* yang akan digunakan adalah sebanyak 3 *cluster*.
2. Menentukan *centroid* awal, *centroid* awal diperoleh secara acak. *Centroid* awal merupakan titik pusat *cluster* pertama. *Centroid* awal dari penelitian ini adalah :
 $C1 = (4, 83.00, 80.08, 87.67, 2, 75)$
 $C5 = (1, 86.67, 84.33, 93.67, 2, 65)$
 $C10 = (3, 83.67, 90.07, 70.97, 1, 78)$

Tabel 1. Data Peserta Didik SMK YAPIM

No	Kode Siswa	Ekstra-kurikuler	Tugas	UTS	UAS	Absen	Nilai Sikap
1	AAA	4	83.00	80.08	87.67	2	75
2	BBB	2	80.67	83.33	88.00	0	65
3	CCC	2	83.67	85.67	82.33	1	85
4	DDD	1	76.67	86.33	86.67	1	75
5	EEE	1	86.67	84.33	93.67	2	65
6	FFF	0	87.67	80.00	87.00	0	86
7	GGG	1	83.33	80.67	80.67	0	77
8	HHH	2	85.67	83.67	76.67	1	66
9	III	1	86.33	76.67	86.67	0	88
10	JJJ	3	83.67	90.07	70.97	1	78

3. Menghitung jarak setiap data yang ada terhadap setiap pusat cluster. Berikut perhitungannya dengan menggunakan persamaan Euclidean Distance Space:

a. Jarak antara data siswa pertama dengan pusat cluster pertama.

$$c1 \sqrt{(4-4)^2 + (83.00-83.00)^2 + (80.08-80.08)^2 + (87.67-87.67)^2 + (2-2)^2 + (75-75)^2} = 0$$

b. Jarak antara data siswa pertama dengan pusat cluster ke dua

$$c2 \sqrt{(1-4)^2 + (86.67-83.00)^2 + (84.33-80.08)^2 + (93.67-87.67)^2 + (2-2)^2 + (65-75)^2} = 13.29$$

c. Jarak antara data siswa pertama dengan pusat cluster ke tiga

$$c3 \sqrt{(3-4)^2 + (83.67-83.00)^2 + (90.07-80.08)^2 + (70.97-87.67)^2 + (1-2)^2 + (78-75)^2} = 19.75$$

Adapun hasil dari perhitungan dari keseluruhan data terhadap tiap pusat cluster awal disajikan pada tabel 2.

Tabel 2. Hasil Perhitungan Jarak Data Siswa pada Masing-Masing Centroid.

C1	C2	C3
0	13.29	19.75
11.14	8.61	22.70
12.85	23.27	14.09
9.49	15.94	17.96
13.29	0	27.04
12.78	22.59	21.17
8.15	18.48	13.73
15.07	17.13	14.91
14.34	25.31	23.20
19.75	27.04	0

4. Pengelompokan data cluster, setelah perhitungan jarak data pada centroid, langkah selanjutnya adalah pengelompokan data. Berikut adalah hasil pengelompokannya.

Perhitungan dapat dilakukan secara manual dengan menggunakan *microsoft excel* maksud dari angka 2 adalah menunjukkan daerah yang

paling sedikit mendapatkan prestasi, angka 1 menunjukkan daerah yang sedang, dan angka 0 menunjukkan daerah yang paling sering mendapatkan prestasi.

Tabel 3. Hasil Pengelompokan Data.

C1	C2	C3	C1	C2	C3
0	13.29	19.75	2	1	0
11.14	8.61	22.70	1	2	0
12.85	23.27	14.09	2	0	1
9.49	15.94	17.96	2	1	0
13.29	0	27.04	1	2	0
12.78	22.59	21.17	2	0	1
8.15	18.48	13.73	2	0	1
15.07	17.13	14.91	1	0	2
14.34	25.31	23.20	2	0	1
19.75	27.04	0	1	0	2

5. Setelah semua data ditempatkan kedalam cluster yang terdekat, kemudian hitung kembali pusat cluster yang baru berdasarkan rata-rata anggota ada pada cluster tersebut.

Tabel 4. Centroid Baru

C1	C2	C3
1.50	1.50	2.50
83.45	83.67	84.67
81.57	83.83	86.87
85.17	90.84	73.82
0.67	1.00	1.00
81.00	65.00	72.00

Setelah didapat titik pusat baru dari tiap cluster, hitung kembali data dengan pusat cluster yang baru sampai didapat pola terakhir yang sudah tidak berpindah, dalam penelitian ini. Data dihitung ulang sampai iterasi ke 3, dimana setiap cluster tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu cluster ke cluster lainnya.

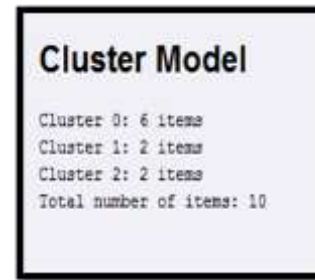
6. Menghitung jarak data ke *centroid* yang baru untuk iterasi berikutnya. Dengan menggunakan nilai *centroid* yang baru, jarak ke masing-masing *centroid* dihitung sampai group baru sama dengan group lama atau group sebelumnya.
7. Menghitung jarak data ke *centroid* yang baru untuk iterasi berikutnya. Dengan menggunakan nilai *centroid* yang baru, jarak ke masing-masing *centroid* dihitung sampai group baru sama dengan group lama atau group sebelumnya.

Tabel 5. Hasil dan Pola Terakhir Jarak Antara *Centroid* dan Pusat *Cluster*

C1	C2	C3	C1	C2	C3
2	1	0	2	1	0
1	2	0	1	2	0
2	0	1	2	0	1
2	1	0	2	1	0
1	2	0	1	2	0
2	0	1	2	0	1
2	0	1	2	0	1
0	1	2	0	1	2
2	0	1	2	0	1
1	0	2	1	0	2

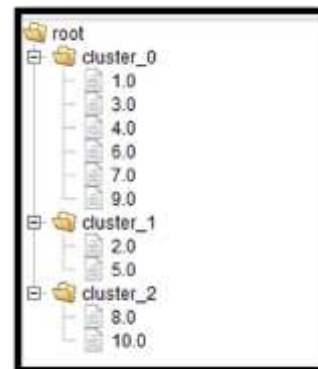
Data yang dikelompokkan pada cluster 1 berjumlah 6 orang, di cluster 2 berjumlah 2 orang, dan di cluster 3 berjumlah 2 orang. Jika hasil antara iterasi sudah sama dengan iterasi sebelumnya, maka dapat dikatakan hasil penelitian telah selesai.

8. Pengujian dengan aplikasi yang telah ditentukan. Hasil dari analisis diuji lagi dengan menggunakan *Rapid Miner* yang sudah ada. Dengan memasukkan data hasil analisis ke *Rapid Miner*. Pada *Text View* maka akan menampilkan *cluster* model. *Cluster* model yang diperoleh dari hasil pengujian terhadap data menggunakan *k-means clustering* terlihat pada gambar 1.



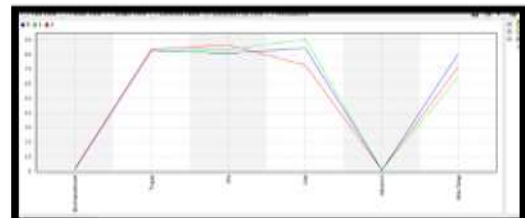
Gambar 1. Cluster Model.

Dapat dilihat *cluster* model yang dihasilkan terdiri dari *cluster* 0 yaitu 6 items, *cluster* 1 terdiri dari 2 items, dan *cluster* 2 terdiri dari 2 items. Dari total jumlah 10 items. Dalam bentuk *root* memiliki 3 *cluster* yaitu folder *cluster* 0, *cluster* 1 dan *cluster* 2. Untuk memudahkan melihat anggota yang dimiliki oleh setiap *folder cluster*, lihat gambar 2 yang menampilkan *membership* dari masing-masing *cluster membership*.



Gambar 2. Tampilan Member Masing-Masing *Cluster*

Untuk melihat *cluster* mana yang paling tinggi bisa dilihat dari gambar 3:



Gambar 3. Diagram Penggambaran *Cluster*.

SIMPULAN

1. Dengan menggunakan Metode Algoritma *K-Means Clustering*, dapat menentukan pengelompokan prestasi siswa tinggi, menengah dan cukup.
2. Dengan menginput data yang digunakan dalam penelitian ini yaitu berupa kegiatan ekstrakurikuler yang diikuti, Nilai dari Tugas, UTS, UAS, Absensi dan Nilai sikap. *Output*

ditampilkan dari data yang digunakan yaitu berupa pengelompokan data berdasarkan algoritma *K-Means*.

3. Dengan adanya *software Rapid Miner* dalam penelitian ini maka keakuratan data akan cukup baik terhadap permasalahan yang terjadi terkait dengan prestasi siswa dengan hasil 70% dapat mengenali data pada 10 data yang digunakan sebagai sampel.

DAFTAR PUSTAKA

Kusrini. (2007). *Konsep Dan Aplikasi Sistem Pendukung Keputusan*. Yogyakarta: Andi

Kusuma, S. & Purnomo, H. (2006). *Aplikasi Logika Fuzzy untuk*

Pendukung Keputusan. Yogyakarta: Graha Ilmu.

Purnomo & Kusumadewi. (2010). *Aplikasi Logika Fuzzy Untuk Pendukung Keputusan*. Yogyakarta: Graha Ilmu.

