Penerapan Algoritma *K-Means* dalam *Data Mining* untuk Peminatan Jurusan Bagi Siswa Kelas X (Studi Kasus: SMA Negeri 29 Jakarta)

Nurhayati [1], Luigi Ajeng Pratiwi [2]
[1), 2)] Program Studi Teknik Informatika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta
email: nurhayati@uinjkt.ac.id [1)], luigi.pratiwi@gmail.com[2)]

Abstrak

Sesuai dengan Surat Keputusan (SK) Dirjen Mandikdasmen Departemen Pendidikan Nasional Nomor 12/C/KEP/TU/2008 dalam Juknis Penulisan Laporan Hasil Belajar Peserta Didik diketahui bahwa penentuan jurusan sangat diperlukan bagi siswa Sekolah Menengah Atas (SMA), begitu juga di SMA Negeri 29 Jakarta. Peminatan dilakukan pada saat siswa berada di kelas X (sepuluh) dan akan naik ke kelas XI (sebelas). Proses penentuan jurusan bagi siswa SMA akan terjadi secara berulang setiap tahun. Jumlah data yang banyak tersebut maka sangat mungkin terjadi kesalahan ketika proses peminatan jurusan dilakukan. Hal tersebut dapat diminimalisir dengan penggunaan teknologi *Data Mining*. Pada penelitian ini, peneliti menerapkan algoritma *K-Means* dalam teknologi *Data Mining* sehingga diharapkan dapat diketahui jumlah siswa yang masuk jurusan IPA atau IPS secara akurat. Hasil implementasi algoritma K-Means pada penelitian ini menghasilkan tingkat akurasi yang tinggi yaitu sebesar 75%.

Kata Kunci :*Data Mining, K-Means*

1. PENDAHULUAN

Untuk mendapatkan informasi diperlukan data serta cara yang tepat. Hal ini sejalan dengan Banyaknya data yang diolah dapat menimbulkan kesulitan dalam hal pengelompokkan data. Namun dengan perkembangan Teknologi Informasi (TI) terdapat berbagai macam solusi untuk mengatasi kesulitan tersebut, salah satunya adalah dengan menggunakan teknologi *data mining*. Teknologi data mining merupakan salah satu alat bantu untuk penggalian data pada basis data berukuran besar dan dengan spesifikasi tingkat kerumitan yang telah banyak digunakan pada banyak domain aplikasi seperti perbankan maupun bidang telekomunikasi [6]. Larose(2005) berpendapat bahwa *data mining* terbagi atas beberapa kelompok berdasarkan tugas yang dapat dilakukan, salah satunya adalah clustering atau pengklasteran [8].

Algoritma *K-Means* merupakan algoritma teknik klustering yang berulang-ulang. Algoritma ini mempartisi data 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 [3].

Penggalian informasi pada sebuah data yang berukuran besar (mempunyai jumlah record dan jumlah field yang cukup banyak) tidak dapat dilakukan dengan mudah sehingga penerapan teknologi *data mining* diperlukan untuk membantu sekolah dalam proses peminatan jurusan bagi siswa kelas X.

Dari hasil wawancara yang dilakukan terhadap guru SMA Negeri 29 Jakarta diketahui bahwa saat ini SMA Negeri 29 Jakarta memiliki kesulitan untuk proses peminatan yang masih menggunakan sistem manual dalam mengolah data akademik siswa yang berjumlah 253 orang sehingga membutuhkan waktu yang lama untuk mengetahui hasil peminatan siswa. Hal ini menarik untuk diselesaikan agar bisa diperoleh informasi mengenai siapa saja siswa yang masuk peminatan IPA dan siapa saja

siswa yang masuk peminatan IPS serta jumlah siswa yang berada pada dua program peminatan tersebut. Jumlah siswa dan nilai siswa yang ada, akan diolah dengan menerapkan algoritma *K-Means* dalam teknologi data mining sehingga diharapkan dapat diketahui jumlah siswa yang masuk jurusan IPA atau IPS secara akurat.

2. KAJIAN LITERATUR

2.1. Data Mining

Menurut [1, 5] *Data Mining* merupakan proses iteratif dan interaktif untuk menemukan pola atau model baru yang dapat digeneralisasi untuk masa yang akan datang, bermanfaat dan dapat dimengerti dalam suatu database yang sangat besar.

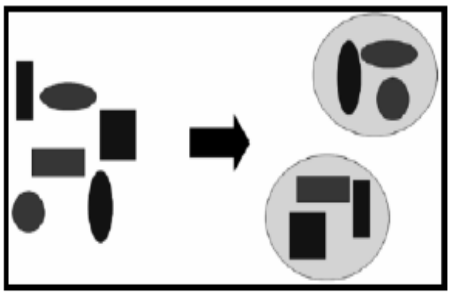
Menurut [16] *Data Mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. *Data Mining* [10] juga diartikan sebagai analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya.

Menurut [11, 13, 14, & 15] Definisi *Data Mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. *Data Mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah data mining kadang disebut juga *knowledge discovery*. Salah satu teknik yang dibuat dalam data mining adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data lainnya yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Pada data mining, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil. Semua hal tersebut bertujuan mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan diharapkan dapat tercapai.

2.2. Clustering

Clustering adalah salah satu alat bantu pada data mining yang bertujuan mengelompokkan obyek-obyek ke dalam *cluster-cluster*. *Cluster* adalah sekelompok atau sekumpulan obyek-obyek data yang similar satu sama lain dalam *cluster* yang sama dan *dissimilar* terhadap obyek-obyek yang berbeda *cluster*. Obyek akan dikelompokkan ke dalam satu atau lebih *cluster* sehingga obyek-obyek yang berada dalam satu *cluster* akan mempunyai kesamaan yang tinggi antara satu dengan lainnya [2, 12].

Obyek-obyek tersebut dikelompokkan berdasarkan prinsip memaksimalkan kesamaan obyek pada *cluster* yang sama dan memaksimalkan ketidaksamaan pada *cluster* yang berbeda. Kesamaan obyek biasanya diperoleh dari nilai-nilai atribut yang menjelaskan obyek data, sedangkan obyek-obyek data biasanya direpresentasikan sebagai sebuah titik dalam ruang multidimensi [2, 12].



Gambar 1. Ilustrasi *Clustering*

2.3. Algoritma *K-Means*

Secara umum, *K-Means* merupakan algoritma yang umum digunakan untuk clustering dokumen. Prinsip utama *K-Means* adalah menyusun k prototype atau pusat massa (centroid) dari sekumpulan data berdimensi n. [4, 7].

Menurut [9] *K-Means* merupakan metode klasterisasi yang paling terkenal dan banyak digunakan di berbagai bidang karena sederhana, mudah diimplementasikan, memiliki kemampuan untuk mengklaster data

Nurhayati [1)], Luigi Ajeng Pratiwi [2)]
[1), 2)] Program Studi Teknik Informatika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta

yang besar, mampu menangani data outlier, dan kompleksitas waktunya linear O(nKT) dengan n adalah jumlah dokumen, K adalah jumlah kluster, dan T adalah jumlah iterasi. K-means merupakan metode pengklasteran secara partitioning yang memisahkan data ke dalam kelompok yang berbeda.Dengan partitioning secara iteratif, *K-Means* mampu meminimalkan rata-rata jarak setiap data ke klasternya.

Algoritma *K-Means* memiliki 5 cara dalam mengklaster, menurut [12] yaitu :

1. Penentuan pusat awal *cluster*
2. Hitung jarak setiap data ke pusat kluster menggunakan persamaan Euclidean.
3. Kelompokkan data ke dalam kluster yang dengan jarak yang paling pendek menggunakan persamaan

$$Min\sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j}^{m}(Cij - Ckj)^2}$$

Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat *cluster*, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat.

4. Hitung pusat *cluster* yang baru menggunakan persamaan

$$C_{kj} = \frac{\sum_{i=1}^{p} C_{ij}}{P}$$

Dimana Cij merupakan elemen dari kluster ke-k dan P adalah banyak anggota *cluster* k.Setelah diketahui anggota tiap-tiap *cluster* kemudian pusat cluster baru dihitung berdasarkan data anggota tiap-tiap *cluster* sesuai dengan rumus pusat anggota *cluster*.

5. Ulangi langkah 2 sampai dengan 4 hingga sudah tidak ada lagi data yang berpindah ke *cluster* yang lain.

3. METODE PENELITIAN

Berikut merupakan tahapan penelitian yang dilakukan:

1. Pengumpulan data yang dibutuhkan pada penelitian ini dilakukan dengan wawancara, observasi, dan studi pustaka.
2. Pembuatan aplikasi peminatan siswa dengan menggunakan algoritma *K-Means.*
3. Membandingkan hasil aplikasi dengan hasil nyata untuk mendapatkan tingkat akurasi *K-Means* pada peminatan siswa.

4. HASIL DAN PEMBAHASAN

Berikut merupakan contoh data nilai siswa yang digunakan pada penelitian ini:

NO	NIS	NAMA	MTK	Fisika	Biologi	Kimia	Sejarah	Geografi	Ekonomi	Sosiologi
1	16705	ABRAR SUMANGERUKKA	79	75	75	85	76	78	76	80
2	16706	ADZKIA RAHMATILLAH	84	76	79	77	76	77	75	81
3	16707	AHMAD RISQI SABILAL RASYAD	77	84	78	85	92	89	77	82
4	16708	AMBAR PURWANTININGSIH	78	86	84	77	78	77	75	75
5	16709	ANDIKA SYAHALAM	82	82	81	91	90	82	79	91
6	16515	ANGGIE SURYO PRAKOSO	75	75	70	82	75	75	79	75
7	16710	ANISAH DIYANTI	77	75	75	89	80	80	75	75
8	16711	ARIANSYAH PRABOWO	77	76	70	77	77	80	75	80
9	16712	AUDI NADYA	79	76	75	84	77	81	76	77
10	16717	DIMAS MURYA PERDANA	80	75	75	75	75	78	77	79
11	16718	ELISABETH GULTOM	76	71	75	75	77	81	79	84
12	16719	GALUH ADE YOSHE	80	77	75	78	78	77	77	80

Seperti yang telah dijelaskan pada kajian literature bahwa algoritma *K-Means* mempunyai 5 langkah, Berikut merupkan hasil dari setiap langkah tersebut.

1. Untuk penentuan awal di asumsikan:
   - Diambil data ke- 2 sebagai pusat Cluster Ke-1: (84, 76, 79, 77, 76, 77, 75, 81) dengan batas nilai minimal 75 untuk mata pelajaran Matematika, Fisika, Biologi, Kimia.
   - Diambil data ke- 5 sebagai pusat Cluster Ke-2: (82, 82, 81, 91, 90, 82, 79, 91) dengan batas nilai minimal 75 untuk mata pelajaran Sejarah, Geografi, Ekonomi, Sosiologi.
2. Setelah dilakuan perhitungan jarak dari data ke-1 hingga ke n terhadap pusat cluster maka kemudian akan didapatkan matrik jarak sebagai berikut:

Nurhayati [1)], Luigi Ajeng Pratiwi [2)]
[1), 2)] Program Studi Teknik Informatika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta

1	2	3	4	5	6	7	8	9	10	11	12	
10.4	0	24.1	14.1	24.02	15.6	16.4	11.8	11,1	6,8	12,2	6,5	C1
21.7	24.02	14.5	26.03	0	28.8	22.1	26.2	22,4	27,09	25,8	22,9	C2

3. Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat cluster, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat cluster terdekat. Berikut ini akan ditampilkan data matriks pengelompokkan group, nilai 1 berarti data tersebut berada dalam group.

1	2	3	4	5	6	7	8	9	10	11	12	
1	1	0	1	0	1	1	1	1	1	1	1	C1
0	0	1	0	1	0	0	0	0	0	0	0	C2

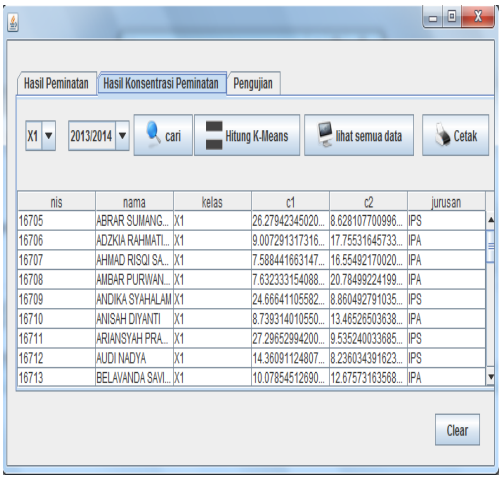
4. C1 dengan 10 anggota menjadi C1=(79.5; 83; 79.5; 88; 91; 85.5; 78; 86.5) dan C2 dengan 2 anggota menjadi C2=(78.5, 76.2, 75.3, 79.9, 76.9, 78.4, 76.4, 78.6).
5. Langkah selanjutnya sama dengan langkah pada nomer 3 jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat cluster, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat cluster terdekat. Berikut ini akan ditampilkan data matriks pengelompokkan group, nilai 1 berarti data tersebut berada dalam group. Berikut hasil pengelompokkan Group 2.

1	2	3	4	5	6	7	8	9	10	11	12	
1	1	0	1	0	1	1	1	1	1	1	1	C1
0	0	1	0	1	0	0	0	0	0	0	0	C2

Karena G2 = G1 memiliki anggota yang sama maka tidak perlu dilakukan iterasi/perulangan lagi. Hasil clustering telah mencapai stabil dan konvergen.

Pengujian algoritma *K-Means* dilakukan menggunakan matriks konfusi. Pengujian dilakukan dengan membandingkan hasil peminatan algoritma K-Means terhadap hasil peminatan manual. Dari hasil pengujian tersebut diketahui bahwa algoritma K-Means dalam peminatan jurusan siswa kelas X SMA memiliki tingkat akurasi sebesar 75%. Laju kesalahan atau laju error sebesar 25%. Laju kesalahan disebabkan karena rentang nilai antara IPA dan IPS terlalu dekat sehingga menyebabkan hasil penghitungan algoritma K-Means bernilai samar. Siswa bisa masuk ke dalam peminatan IPA dan IPS karena rentang nilainya terlalu dekat.

Berikut ini tampilan aplikasi peminatan siswa.



Gambar 2. Tampilan Aplikasi Hasil Peminatan Siswa

5. KESIMPULAN

Dari pembahasan yang telah diuraikan penulis, maka dapat diambil kesimpulan Algoritma *K-Means* dapat di implementasikan dalam teknologi data mining untuk sistem peminatan jurusan bagi siswa kelas X SMA dengan tingkat akurasi sebesar 0.753 atau 75.3% atau 75% terhadap 253 data sampel. Laju kesalahan sebesar 0.247 atau 24.7% atau 25%.

Dengan tingkat akurasi sebesar 75% aplikasi peminatan jurusan cukup efektif untuk membantu pihak sekolah dalam menentukan minat siswa kelas X sesuai dengan nilai akademik yang dimiliki masing-masing siswa.

Nurhayati [1)], Luigi Ajeng Pratiwi [2)]
[1), 2)] Program Studi Teknik Informatika, Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta

6. REFERENSI

[1] Dunham dan Margaret H. 2003. *Data Mining Introductory and Advanced Topics*. Prentice Hall, New Jersey.

[2] Ediyanto, M.N. Mara, N. Satyahadewi. 2013. Pengklasifikasian Karakteristik Dengan Metode *K-Means Cluster Analysis*. *Buletin Ilmiah Mat.Stat. dan Terapannya (BIMASTER)*. 2(2): 133-136.

[3] Giyanto, H. 2008. Penerapan algoritma *Clustering K-Means, K-Medoid*, Gath Geva. *Tesis*. Universitas Gajah Mada, Yogyakarta.

[4] Han, J., M. Kamber, dan J. Pei,. 2011. *Data Mining Concept and Techniques Third Edition*. Morgan Kaufmann Publishers, San Francisco.

[5] Hermawati, F. A. 2013. *Data Mining*. Andi,Yogyakarta.

[6] Jananto, A. 2010. Memprediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining (Studi kasus data akademik mahasiswa UNISBANK). *Tesis*. Universitas Gajah Mada,Yogyakarta.

[7] Kantardzic, M. 2003. *Data Mining: Concepts, Models, Methods, and Algorithms*. IEEE & Wiley Inter-Science, New York.

[8] Kusrini, T. Luthfi, dan Emha. 2009. *Algoritma Data Mining.* CV. Andi Offset,Yogyakarta.

[9] MacQueen, J.B. 1967. *Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley.

[10] Prasetyo, E. 2012. *Data Mining – Konsep dan Aplikasi Menggunakan MATLAB*. ANDI,Yogyakarta.

[11] Rismawan, T dan Kusumadewi, S. 2008. Aplikasi *K-Means* untuk Pengelompokkan Mahasiswa Berdasarkan Nilai *Body Mass Index* (BMI) & Ukuran Kerangka. Seminar Nasional Aplikasi Teknologi Informasi 2008 (SNATI 2008). 21 Juni 2008. Yogyakarta. E.43-E.48.

[12] Santoso, B. (2007), *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu,Yogyakarta.

[13] Tan, P.N., Stainbach, M., dan Kumar, V. 2005, *Introduction to Data Mining*, Addison Wesley Publisher, New York.

[14] Vercellis, C. 2009. *Business intelligence: Data Mining and Optimization for Decision Making*. John Wiley & Sons,Chichester.

[15] Witten, I.H. dan Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publishers, San Fransisco, California, U.S.A

[16] Pramudiono, I. 2007. *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data*. http://www.ilmukomputer.org/wpcontent/uploads/2006/08/iko-datamining.zip Diakses pada tanggal 24 September 2013 pukul 2.38PM