

Penerapan Data Mining Dalam Pengelompokan Data Nilai Siswa Untuk Penentuan Jurusan Siswa Pada SMA Tamora Menggunakan Algoritma K-Means Clustering

Yohanni Syahra

Program Studi Informasi, STMIK Triguna Dharma

Article Info

Article history:

Received Mei 15th, 2018

Revised June 28th, 2018

Accepted Aug 06th, 2018

Keyword:

Penentuan Jurusan
Clustering
Algoritma K-Means

ABSTRAK

Penjurusan siswa Sekolah Menengah Atas merupakan usaha peningkatan kualitas pendidikan yang langsung berkenaan dengan siswa sebagai salah satu bagian dari peningkatan SDM. Penentuan jurusan yang dilakukan selama ini mempunyai banyak kelemahan, antara lain berdasarkan keinginan siswa tanpa melihat latar belakang nilai akademisnya dan tidak ada acuan yang jelas untuk melakukan pengelompokan data nilai siswa.

Untuk dapat menyelesaikan permasalahan yang ada maka dibuatlah sebuah sistem untuk pengelompokan nilai siswa untuk penentuan jurusan dengan menerapkan salah satu metode Clustering dan algoritma yang digunakan adalah Algoritma K-Means untuk mengidentifikasi objek yang serupa dengan memperhatikan beberapa kriteria.

Dengan demikian hasil dari pengelompokan yang telah dirancang akan membantu pihak SMA Tamora dalam proses Pengelompokan nilai Siswa untuk penentuan Jurusan yang sesuai dengan kriteria, sehingga pengelompokan dapat dilakukan lebih cepat, tepat dan akurat serta terhindar dari kesalahan.

Copyright © 2018 STMIK Triguna Dharma.
All rights reserved.

First Author

Nama : Yohanni Syahra, S.Si, M.Kom
Program Studi : Sistem Informasi STMIK Triguna Dharma
Email :

1. PENDAHULUAN

Penjurusan siswa Sekolah Menengah Atas merupakan usaha peningkatan kualitas pendidikan yang langsung berkenaan dengan siswa sebagai salah satu bagian dari peningkatan SDM. Sesuai kurikulum yang berlaku di seluruh Indonesia, maka siswa kelas X SMA yang naik ke kelas XI akan mengalami penjurusan.

Penjurusan siswa ini bertujuan untuk mengarahkan peserta didik agar lebih fokus mengembangkan kemampuan diri dan minat yang dimiliki. Jurusan yang tidak tepat bisa sangat merugikan siswa dan karirnya di masa mendatang. Dengan penjurusan tersebut diharapkan dapat memaksimalkan potensi, bakat atau talenta individu, sehingga juga akan memaksimalkan nilai akademisnya. Penentuan jurusan yang dilakukan selama ini mempunyai banyak kelemahan, antara lain berdasarkan keinginan siswa tanpa melihat latar belakang nilai akademisnya dan tidak ada acuan yang jelas untuk melakukan pengelompokan data siswa. Sehingga jurusan yang dipilih terkadang menjadi masalah bagi siswa di kemudian hari.

Data Mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (machine learning) untuk menganalisis dan mengekstraksi pengetahuan (knowledge) secara otomatis. Dalam Data Mining terdapat beberapa teknik salah satunya adalah klastering. Klastering digunakan untuk menganalisis pengelompokan data nilai siswa yang memiliki karakteristik tertentu. Algoritma yang akan diterapkan dari klastering ini adalah K-Means.

2. LANDASAN TEORITIS

2.1 Data Mining

Menurut Hermawati (2013:3) "Data Mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (machine learning) untuk menganalisis dan mengekstraksi pengetahuan (knowledge) secara otomatis." Data Mining berisi pencarian trend atau pola yang diinginkan dalam database besar untuk membantu pengambilan keputusan di waktu yang akan datang.

2.2 Algoritma K-Means

Algoritma *K-Means* merupakan metode analisis kelompok yang mengarah pada pemartisian N objek pengamatan ke dalam kelompok (*cluster*) dimana setiap objek pengamatan dimiliki oleh sebuah kelompok dengan *mean* (rata-rata) terdekat, mirip dengan algoritma *Expectation-Maximization* untuk *Gaussian Mixture* dimana keduanya mencoba untuk menemukan pusat dari kelompok dalam data sebanyak iterasi perbaikan yang dilakukan oleh kedua algoritma. (Prasetyo, 2012:178)

Berikut ini adalah langkah-langkah algoritma *K-Means*:

1. Tentukan K sebagai jumlah *cluster* yang dibentuk.
2. Tentukan pusat (*centroid*) *cluster* awal.
digunakan rumus sebagai berikut

$$C_i = \frac{1}{M} \sum_{j=1}^M X_j$$

keterangan :

C_i : *centroid* pada *cluster*

x_j : objek ke-j

M : banyaknya objek/jumlah objek yang menjadi anggota *cluster*

3. Perhitungan

Untuk mengukur jarak antar data dengan pusat *cluster* digunakan *Euclidian Distance*.. Rumusnya adalah :

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

Keterangan :

x_2 : data

x_1 : titik pusat (*centroid*)

$D(x_2, x_1)$: *Euclidian Distance* yaitu jarak antara data x_2 dan x_1

4. Pengelompokan data

setelah sejumlah populasi data tersebut menemukan kedekatan dengan salah satu *centroid* yang ada maka secara otomatis populasi data tersebut masuk kedalam kelas yang memiliki *centroid* yang bersangkutan.

5. Lakukan *iterasi*, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan [2.1]
6. Ulangi langkah 3 apabila masih ada data yang berpindah kelompok, atau ada perubahan nilai *centroid* diatas nilai ambang yang ditentukan, atau juga apabila perubahan nilai pada fungsi objektif yang digunakan masih diatas nilai ambang yang ditentukan.
7. Tentukan K sebagai jumlah *cluster* yang dibentuk.
8. Tentukan pusat (*centroid*) *cluster* awal.

Dalam menentukan n buah pusat *cluster* awal dilakukan pembangkitan bilangan *random* yang merepresentasikan urutan data input. Pusat awal cluster didapatkan dari data sendiri bukan dengan menentukan titik baru, yaitu dengan mengacak (*random*) pusat awal dari data. Kemudian untuk menghitung *centroid cluster* ke-I berikutnya digunakan rumus sebagai berikut

$$C_i = \frac{1}{M} \sum_{j=1}^M X_j$$

keterangan :

C_i : *centroid* pada *cluster*

x_j : objek ke-j

M : banyaknya objek/jumlah objek yang menjadi anggota *cluster*

9. Perhitungan

Untuk mengukur jarak antar data dengan pusat *cluster* digunakan *Euclidian Distance*. Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antar data dengan pusat *cluster*, jarak ini menggunakan bahwa data tersebut berada dalam satu kelompok dengan pusat *cluster* terdekat. Rumusnya adalah :

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

Keterangan :

x_2 : data

x_1 : titik pusat (*centroid*)

$D(x_2, x_1)$: *Euclidian Distance* yaitu jarak antara data x_2 dan x_1

10. Pengelompokkan data setelah sejumlah populasi data tersebut menemukan kedekatan dengan salah satu *centroid* yang ada maka secara otomatis populasi data tersebut masuk kedalam kelas yang memiliki *centroid* yang bersangkutan.
11. Lakukan *iterasi*, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan [2.1]
12. Ulangi langkah 3 apabila masih ada data yang berpindah kelompok, atau ada perubahan nilai *centroid* diatas nilai ambang yang ditentukan, atau juga apabila perubahan nilai pada fungsi objektif yang digunakan masih diatas nilai ambang yang ditentukan.

3. ANALISIS DAN HASIL

Analisa permasalahan bertujuan untuk mengetahui dan mengamati permasalahan yang sedang terjadi, kemudian merumuskan sebuah solusi untuk menyelesaikan permasalahan yang ada. Untuk jurusan yang sesuai dengan minat, maka diperlukan informasi dan pengetahuan tentang informasi nilai rata-rata bidang studi siswa.

Tabel 1 Data Siswa

No	Nama Siswa	Nilai Rata-Rata IPA	Nilai Rata-Rata IPS	Nilai Tes Ujian Umum
1	Abimana Syahputra	80	80	88
2	Agung Ramadhan	75	75	70
3	Agustina Dwi Ningsih	81	88	91
4	Aisyah Aulya	85	80	85
5	Akhir Ramadhan	72	78	86
6	Alysa Syaftia	80	88	77
7	Amalini Hilwa Hadijah	81	91	92
8	Amelia Putri Hartawan	90	87	91
9	Ananda Tiara	79	81	80
10	Andi Syahputra Siregar	78	92	80
11	Andri Hariono	91	80	95
12	Arida Afika Putri	80	88	89
13	Asrul Adam	70	78	85
14	Ayu Mitri	92	93	90
15	Bella Syahfitri	92	82	93
16	Benny Situmorang	77	88	80
17	BoyChandra S	80	78	80
18	Chandra Hutagalung	76	70	91
19	Cici Lestari	85	73	78
20	Citra Kirana	93	80	91
21	Dedek Ridzki Wiranata	81	82	80
22	Deni Yuswanto	78	87	80
23	Desi Kesuma Wardani	86	90	87
24	Desi Purnama	91	88	93
25	Deva Mahendra	75	77	80
26	Devi Paramitha	73	70	80
27	Dewi Suheni	93	78	85
28	Dewi Utami	86	86	82
29	Dian Murni Dewi	76	78	80
30	Dimas Anggara	93	80	85
31	Dimas Candi Gatot	78	78	80
32	Dimas Prasetyo	77	78	80
33	Erik Manalakson	91	70	85
34	Erina Ghasani	88	84	80

35	Erma Wati	85	78	80
35	Erma Wati	85	78	80
36	Ervan Mahardika	92	80	90
37	Esna Surya Putra	78	80	80
38	Faisal Tri Hanafi	80	78	90
39	Fitri Handayani	92	86	87
40	Hakim Fahrezi	94	86	85
..				
..				
..				
127	Trisia Wulandari	80	78	93
128	Ulfa Miftahul	78	84	80

Berikut langkah-langkah clustering menggunakan algoritma k-means.

1. Tentukan jumlah cluster yang diinginkan (misalk=4)
2. Pilih centroid awal secara acak: pada langkah ini secara acak akan dipilih 4 buah data sebagai centroid, misalnya: data {40,20,14,10}
M1=(94,86,85), M2=(93,80,91), M3=(92,93,90),M4=(78,92,80)
3. Hitung jarak dengan centroid(iterasi 1)
Pada langkah ini setiap data akan ditentukan centroid terdekatnya, dan data tersebut akan ditetapkan sebagai anggota kelompok yang terdekat dengan centroid.
Untuk menghitung jarak ke centroid masing-masing cluster pada siswa sbb:
Data: (1,3), centroid M1=(94,86,85), M2=(93,80,91), M3=(92,93,90),M4=(78,92,80)
 $DM1 = \sqrt{(80 - 94)^2 + (80 - 86)^2 + (88 - 85)^2} = 15,52$
 $DM2 = \sqrt{(80 - 93)^2 + (80 - 80)^2 + (88 - 91)^2} = 26,59$
 $DM3 = \sqrt{(80 - 92)^2 + (80 - 93)^2 + (88 - 90)^2} = 14,46$
 $DM4 = \sqrt{(80 - 78)^2 + (80 - 92)^2 + (88 - 80)^2} = 10,82$
Tabel hasil perhitungan jarak selengkapnya antara masing-masing data dengan centroid :

Tabel 2 Hasil Perhitungan Iterasi 1

No	Nama Siswa	Pusat m1	Pusat m2	Pusat m3	Pusat m4	Jarak Terdekat	Klaster
1	Abimana Syahputra	15.52	13.34	17.80	14.56	13.34	2
2	Agung Ramadhan	26.59	28.11	31.83	19.95	19.95	4
3	Agustina Dwi Ningsih	14.46	14.42	12.12	12.08	12.08	4
4	Aisya Aulya	10.82	10.00	15.59	14.76	10.00	2
5	Akhir Ramadhan	23.43	21.68	25.32	16.37	16.37	4
6	Alysa Syaftia	16.25	20.71	18.38	5.39	5.39	4
7	Amalini HilwaHadijah	15.59	16.31	11.36	12.41	11.36	3
8	Amelia Putri	7.28	7.62	6.40	17.03	6.40	3
9	Ananda Tiara	16.58	17.83	20.32	11.05	11.05	4
10	Andi Syahputra	17.80	22.14	17.23	0.00	0.00	4
11	Andri Hariono	12.04	4.47	13.96	23.19	4.47	2
12	Arida Afika Putri	14.70	15.39	13.04	10.05	10.05	4
13	Asrul Adam	25.30	23.85	27.09	16.88	16.88	4
14	Ayu Mitri	8.83	13.08	0.00	17.23	0.00	3
15	Bella Syahfitri	9.17	3.00	11.40	21.56	3.00	2
16	Benny Situmorang	17.83	21.00	18.71	4.12	4.12	4
17	Boy Chandra S	16.88	17.15	21.66	14.14	14.14	4
18	Chandra Hutagalung	24.82	19.72	28.04	24.68	19.72	2
19	Cici Lestari	17.29	16.79	24.35	20.35	16.79	2

20	Citra Kirana	8.54	0.00	13.08	22.14	0.00	2
21	Dedek Ridzki	14.49	16.40	18.49	10.44	10.44	4
22	Deni Yuswanto	16.79	19.87	18.22	5.00	5.00	4
23	Desi Kesuma	9.17	12.85	7.35	10.82	7.35	3
24	Desi Purnama	8.77	8.49	5.92	18.81	5.92	3
25	Deva Mahendra	21.61	21.31	25.40	15.30	15.30	4
26	Devi Paramitha	26.87	24.92	31.46	22.56	22.56	4
27	Dewi Suheni	8.06	6.32	15.84	21.12	6.32	2
28	Dewi Utami	8.54	12.88	12.21	10.20	8.54	1
29	Dian Murni Dewi	20.32	20.35	24.10	14.14	14.14	4
30	Dimas Anggara	6.08	6.00	13.96	19.85	6.00	2
31	Dimas Candi Gatot	18.57	18.71	22.83	14.00	14.00	4
32	Dimas Prasetyo	19.44	19.52	23.45	14.04	14.04	4
33	Erik Manalakson	16.28	11.83	23.56	26.04	11.83	2
34	Erina Ghasani	8.06	12.73	14.04	12.81	8.06	1
35	Erma Wati	13.04	13.75	19.34	15.65	13.04	1
36	Ervan Mahardika	8.06	1.41	13.00	20.98	1.41	2

Pada langkah ini dihitung pula rasio antara besaran *BCV* (*Between Cluster Variation*) dengan *WCV* (*Within Cluster Variation*) :

$$d(m1.m2) = \sqrt{(94 - 93)^2 + (86 - 80)^2 + (85 - 91)^2} = 8,54$$

$$d(m1.m3) = \sqrt{(94 - 92)^2 + (86 - 93)^2 + (85 - 90)^2} = 8,83$$

$$d(m1.m4) = \sqrt{(94 - 78)^2 + (86 - 92)^2 + (85 - 80)^2} = 17,80$$

$$d(m2.m3) = \sqrt{(93 - 92)^2 + (80 - 93)^2 + (91 - 90)^2} = 13,08$$

$$d(m2.m4) = \sqrt{(93 - 78)^2 + (80 - 92)^2 + (91 - 80)^2} = 22,14$$

$$d(m3.m4) = \sqrt{(92 - 78)^2 + (93 - 92)^2 + (90 - 80)^2} = 17,23$$

$$\begin{aligned} BCV &= d(m1.m2)+d(m1.m3)+d(m1.m4)+d(m2.m3)+d(m2.m4)+d(m3.m4) \\ &= 8,54+8,83+17,80+13,08+22,14+17,23 \\ &= 87,63 \end{aligned}$$

Hitung nilai *WCV* (*Within Cluster Variation*) dengan cara memangkatkan jarak terdekat *cluster* dan menjumlahkan setiap nilai *WCV*.

$$\begin{aligned} WCV &= 13,34^2 + 19,95^2 + 12,08^2 \dots + 8,00^2 \\ WCV &= 14908 \end{aligned}$$

Sehingga besar rasio = $BCV/WCV = 87,63/14908 = 0.005$

Karena langkah ini merupakan iterasi 1 maka lanjutkan kelangkah berikutnya. Pusat cluster baru ditentukan berdasarkan pengelompokkan anggota dari masing-masing cluster.

Ulangi iterasi yang dimulai dari langkah 1, sampai rasio tetap atau cluster tidak mengalami perubahan.

Disini sampai dengan iterasi 4, karena bila dibandingkan maka rasio sekarang (0,014) sudah tidak lagi lebih besar dari rasio sebelumnya (0,014) oleh karena itu iterasi dihentikan.

Dari hasil klastering diatas dapat diambil kesimpulan bahwa :

1. kelompok siswa pada klaster 1 merupakan kelompok siswa yang masuk ke jurusan IPA 1 dengan jumlah 33 siswa.
2. Kelompok siswa pada klaster 2 merupakan kelompok siswa yang masuk ke jurusan IPA 2 dengan jumlah 16 siswa.
3. Kelompok siswa pada klaster 3 merupakan kelompok siswa yang masuk ke jurusan IPS 1 dengan jumlah 33 siswa.
4. Kelompok siswa pada klaster 4 merupakan kelompok siswa yang masuk ke jurusan IPS 2 dengan jumlah 46 siswa.

4. KESIMPULAN

Berdasarkan hasil dari perancangan aplikasi penerapan data *mining* dalam pengelompokkan data nilai siswa untuk penentuan jurusan menggunakan algoritma *K-Means*, maka dapat diperoleh beberapa kesimpulan. Adapun kesimpulan tersebut adalah sebagai berikut :

1. *Data mining* dengan algoritma *K-Means* dapat diterapkan pada SMA Tamora untuk menganalisis permasalahan yang ada yang berkenaan dengan pengelompokan data nilai siswa untuk penentuan jurusan.
2. Aplikasi dirancang dengan mengadopsi algoritma *K-Means* yang dapat digunakan dalam pengelompokan data nilai siswa untuk penentuan jurusan pada SMA Tamora.
3. Aplikasi yang dirancang dapat mengimplementasikan dan dapat dijadikan sebagai solusi pemecahan masalah dalam hal pengelompokan data nilai siswa untuk penentuan jurusan siswa pada SMA Tamora.

DAFTAR PUSTAKA

- A.M. Hirin. 2012. VB.Net 2010. ANDI : Yogyakarta
- Hermawati, FA, 2013. Data Mining. ANDI : Yogyakarta.
- Nofriansyah, D, 2015. Algoritma Data Mining Dan Pengujiannya. Deepublish
- Ong Jhon Oscar. 2013. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University. Bekasi : Jurnal
- Prasetyo, Eko, 2012. Konsep Dasar Data Mining Menggunakan Matlab. ANDI : Yogyakarta
- Rosa A.S & M.Shalahudin. 2013. Rekayasa Perangkat Lunak.Informatika : Bandung.
- Sulindawaty dan Fathoni, M, 2010. Pengantar Analisa Perancangan Sistem. (Jurnal SAINTIKOM Vol. 9, No 2) hal 14-17
- Tohari, Hamim, 2014. Analisis Serta Perancangan Sistem Informasi Melalui Pendekatan UML. ANDI : Yogyakarta

BIOGRAFI PENULIS



Yohanni Syahra, S. Si., M. Kom, Perempuan kelahiran Medan, 29 Oktober 1982 saat ini beliau menduduki jabatan Sekretaris Program Studi Sistem Informasi STMIK Triguna Dharma, tamat Universitas Sumatera Utara Bidang Ilmu Kimia, tamat Universitas Putra Indonesia (UPI-YPTK) Konsentrasi Teknologi Komputer, beberapa mata kuliah diampu, diantaranya : Data Mining Dan Data Warehouse, Sistem Pakar, Komputer Multimedia, Komputer Akuntansi, Sistem Basis Data, Aplikasi Finansial, Paket Program Niaga.