PENERAPAN METODE KLASTERING DENGAN ALGORITMA K-MEANS UNTUK PREDIKSI KELULUSAN MAHASISWA PADA PROGRAM STUDI TEKNIK INFORMATIKA STRATA SATU

Gita Premashanti Trayasiwi
Program Studi Teknik Informatika S1, Fakultas Ilmu Komputer
Universitas Dian Nuswantoro
Jalan Nakula 1 no.5-11 Semarang
111201105963@mhs.dinus.ac.id

Abstrak

Lamanya waktu kelulusan mahasiswa tidak selalu dapat diprediksi secara dini oleh pihak mahasiswa maupun perguruan tinggi sehingga dapat berakibat pada waktu lulus yang terlambat dan merugikan kedua belah pihak. Untuk mengatasi permasalahan tersebut diperlukan suatu solusi untuk memprediksi kelulusan mahasiswa. Pada penelitian ini digunakan metode klastering (Clustering) dengan algoritma k-Means. Sebelum dilakukan pengolahan data, dilakukan proses normalisasi data, kemudian data diolah menjadi beberapa klaster. Data yang telah diklasterisasi tersebut menghasilkan kategori prediksi kelulusan mahasiswa berdasarkan lama atau tidaknya waktu kelulusan dan tinggi rendah IPK yang diperoleh mahasiswa pada setiap klaster. Dengan adanya penelitian tersebut, pihak perguruan tinggi dapat mengetahui hasil prediksi kelulusan mahasiswa dan dapat memberikan tindakan preventif untuk mengurangi masalah keterlambatan kelulusan.

Kata kunci: klastering, *k-Means*, mahasiswa

Abstract

The length of time of graduation students can not always be predicted at an early stage by the students and universities so as to result in a late graduation and prejudicial to both sides. To overcome these problems, a solution for predicting student graduation is needed. This study used clustering methods with k-Means algorithm. Before data is processed, the data normalization process is carried out, then the data is processed into multiple clusters. The clustered data generates predictions categories based on length of time of student graduation and the high-low GPA score earned in each cluster. With this study, the college can determine student graduation prediction results and can provide preventive measures to reduce the problem of delay graduation.

Keywords: *clustering*, *k-Means*, student

I. Pendahuluan

Setiap mahasiswa pasti menginginkan lulus tepat waktu. Begitu pula dengan perguruan tinggi. Kelulusan mahasiswa yang tepat waktu akan menguntungkan pihak mahasiswa dan perguruan tinggi. Untuk mahasiswa, semakin cepat lulus maka kesempatan

untuk mengikuti berbagai seleksi dalam mencari pekerjaan semakin banyak. Sedangkan bagi perguruan tinggi, kelulusan mahasiswa dengan tepat waktu dapat memajukan kualitas, meningkatkan reputasi, dan juga berpengaruh pada akreditasi perguruan tinggi tersebut.

Normalnya, mahasiswa (strata-1) lulus dalam jangka waktu 4 tahun, akan tetapi mahasiswa tidak selalu dapat menuntaskan studinya dalam jangka waktu yang ditentukan. Lama waktu kelulusan mahasiswa tidak selalu dapat diprediksi secara dini sehingga mengakibatkan pada waktu lulus yang terlambat dan merugikan kedua belah pihak. Untuk mengatasi masalah tersebut diperlukan adanya analisis *dataset* untuk mengelompokkan data mahasiswa berdasarkan prediksi kelulusan.

Pada umumnya seperti perguruan tinggi-perguruan tinggi di Semarang, Teknik Informatika merupakan program studi bawah naungan Fakultas Ilmu Komputer atau Fakultas Teknik. Berdasarkan data kelulusan suatu perguruan tinggi di Semarang, pada tahun 2011, perguruan tinggi tersebut berhasil meluluskan mahasiswa Teknik Informatika (Strata-1) sebanyak 288 mahasiswa, pada 2012 sebanyak 576, dan pada 2013 sebanyak 595. Sedangkan mahasiswa Teknik Informatika (Strata-1) yang masuk pada tahun 2011 tercatat sebanyak 651 mahasiswa, tahun 2012 sebanyak 744 mahasiswa, dan 2013 sebanyak 600 mahasiswa. Perbedaan jumlah kelulusan dan mahasiswa masuk dari data mahasiswa Teknik Informatika (Strata-1) tahun 2011 hingga 2013 menunjukkan bahwa jumlah kelulusan tidak seimbang dengan jumlah mahasiswa masuk. Pada tahun 2011 dan 2012, jumlah kelulusan yang masih cukup rendah merupakan suatu kendala bagi universitas. Apabila kelulusan dapat diprediksi secara dini, maka jumlah kelulusan dapat ditingkatkan.

Pada penelitian ini akan digunakan metode *clustering* dengan algoritma *k-Means*. *k-Means* merupakan algoritma iteratif. Algoritma ini digunakan untuk mengelompokkan data berdasarkan atribut-atribut ke dalam kelompok *k*. Metode ini digunakan karena algoritma *k-Means* mudah diadaptasi dan dapat digunakan untuk pengolahan *dataset* yang besar [1].

Berdasarkan uraian di atas, maka dilakukan penelitian untuk mengolah data mahasiswa pada program studi Teknik Informatika Strata-1 pada salah satu perguruan tinggi di Semarang sehingga diperoleh solusi alternatif untuk memprediksi kelulusan mahasiswa.

II. Landasan Teori

2.1 Tinjauan Pustaka

2.1.1 Kelulusan Mahasiswa

Mahasiswa adalah aspek penting dalam evaluasi keberhasilan penyelenggaraan program studi pada suatu perguruan tinggi. Sedangkan lulusan merupakan status yang dicapai mahasiswa setelah menyelesaikan proses pendidikan sesuai dengan persyaratan kelulusan yang ditetapkan oleh perguruan tinggi [2]. Kelulusan mahasiswa sangat berpengaruh pada kualitas suatu perguruan tinggi.

Berdasarkan buku II dari Badan Akreditasi Nasional Perguruan Tinggi (BAN-PT) tahun 2011, mahasiswa dan lulusan merupakan aspek penilaian akreditasi.

2.1.2 Data Mining

Data mining atau sering disebut dengan *knowledge discovery in database* (KDD) merupakan kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola hubungan dalam dataset berukuran besar [9]. Menurut [10] data mining merupakan metodologi analisis data yang digunakan untuk mengidentifikasi pola tersembunyi dalam suatu dataset yang besar. Dapat disimpulkan bahwa data mining proses penggalian data tersembunyi dengan menggunakan metodologi analisis data untuk menemukan pola yang unik dan menarik dalam suatu dataset yang besar.

Menurut [1] proses KDD terdiri dari rangkaian iteratif sebagai berikut:

a) *Data cleaning* (untuk menghilangkan *noise* dan data yang tidak konsisten)
b) *Data integration* (penggabungan beberapa sumber data)
c) *Data Selection* (pengambilan data yang relevan untuk analisis)
d) *Data transformation* (data ditransformasikan menjadi bentuk data yang sesuai untuk diproses dalam data mining)
e) *Data mining* (proses yang penting dimana metode diaplikasikan untuk mengekstrak pola data)
f) *Pattern evaluation* (mengidentifikasi pola yang benar-beanr menarik yang mewakili *knowledge* dari pengukuran pola)
g) *Knowledge presentation* (dimana teknik visualisasi dan representasi *knowledge* digunakan untuk menampilkan *knowledge* kepada *user*).

2.1.3 Metode Klastering

Proses pengelompokan satu set objek fisik atau abstrak ke dalam kelas dari objek yang sama disebut *clustering*. Ciri khas *clustering* dalam *data mining* adalah sebagai berikut:

a) Skalabilitas
b) Kemampuan untuk menangani tipe atribut yang berbeda-beda
c) Penemuan klaster dengan bentuk yang berubah-ubah
d) Kebutuhan minimal untuk *domain knowledge* untuk menentukan parameter masukan
e) Kemampuan untuk menangani data *noise*

f) Kenaikan *clustering* dan ketidakpekaan terhadap *record* masukan
g) Dimensionalitas tinggi
h) *Clustering* berdasarkan batasan
i) Mampu untuk diinterpretasi dan digunakan

2.1.4 Algoritma *k-Means*

k-Means merupakan algoritma *clustering* yang berulang-ulang, dimulai dengan pemilihan *k* secara acak. *k* merupakan banyaknya klaster yang akan dibentuk [3]. Metode ini bertujuan untuk meminimkan jumlah jarak antar semua titik dengan pusat klaster.

2.1.5 *Silhouette Index*

Silhouette Index (SI) digunakan untuk memvalidasi sebuah data, klaster tunggal, atau bahkan keseluruhan klaster. Metode ini banyak digunakan untuk memvalidasi klaster yang menggabungkan nilai kohesi dan separasi [4]. Rentang nilai SI adalah -1 hingga +1. Nilai SI mendekati 1 menunjukkan bahwa data tersebut tidak tepat berada pada klaster tersebut. SI bernilai 0 atau mendekati 0 maka posisi data berada pada perbatasan dua klaster.

III. Klasterisasi Data Mahasiswa

3.1 Data yang Digunakan

Data yang digunakan dalam penelitian ini berupa data mahasiswa Teknik Informatika Strata-1 tahun angkatan 2012 berupa data Nomor Induk Mahasiswa (NIM), SKS, dan IPK mahasiswa.

3.2 Normalisasi Data

Normalisasi data sangat diperlukan sebelum proses *data mining* supaya tidak ada parameter yang mendominasi dalam perhitungan jarak antar data [5]. Normalisasi data dapat dihitung dengan persamaan sebagai berikut:

$$Nilai\ Baru = \left(\frac{nilai\ asal - nilai\ min}{nilai\ max - nilai\ min}\right) \quad (1)$$

3.3 Klastering

Langkah-langkah dalam *k-Means* klastering adalah sebagai berikut.

a) Menentukan berapa banyak klaster *k.*
b) Menentukan nilai *centroid*. Nilai awal *centroid* dilakukan secara acak. Sedangkan untuk nilai *centroid* pada saat iterasi, digunakan rumus sebagai berikut:

$$v_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} x_{kj} \quad (2)$$

Dimana,

v_{ij} adalah *centroid* atau rata-rata klaster ke-*i* untuk variabel ke-*j*;

N_i adalah jumlah data yang menjadi anggota klaster ke-*i*;
i,*k* adalah indeks dari klaster;
j adalah indeks dari variabel;
x_{kj} adalah nilai data ke-*k* yang terdapat pada klaster tersebut untuk variabel ke-*j*.

c) Menghitung jarak antara titik *centroid* dengan setiap titik objek. Menghitung jarak menggunakan *euclidean distance*, dengan rumus sebagai berikut:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i} \quad (3)$$

Dimana,
D_e merupakan *euclidean distance*;
i adalah banyaknya objek;
(*x*,*y*) adalah koordinat objek;
(*s*,*t*) adalah koordinat *centroid*.

d) Mengelompokkan objek. Untuk menentukan anggota klaster yaitu dengan menghitung jarak minimum objek. Nilai yang diperoleh dalam keanggotaan data pada jarak matriks adalah 0 atau 1, dimana nilai 1 untuk data yang dialokasikan ke klaster dan nilai 0 untuk data yang dialokasikan ke klaster lain.

e) Kembali lagi ke langkah kedua. Melakukan iterasi hingga hasil *centroid* tetap dan anggota klaster tidak berpindah ke klaster lain.

3.4 Validasi Data Hasil Klastering

Berikut ini adalah beberapa rumus yang digunakan untuk menghitung SI:

$$a_i^j = \frac{1}{m_j - 1} \sum_{r=1}^{m_j} d\left(x_i^j, x_r^j\right), \quad i = 1,2, \dots, mj \quad (4)$$

$$b_i^j = \min\left\{\frac{1}{m_n} \sum_{r=1}^{m_n} d\left(x_i^j, x_r^n\right)\right\}, i = 1,2, \dots, m_n \quad (5)$$

$$SI_i^j = \frac{b_i^j - a_i^j}{max\{a_i^j, b_i^j\}} \quad (6)$$

$$SI_j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (7)$$

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI_j \quad (8)$$

Dimana,
a_i adalah rata-rata jarak data ke-*i* terhadap semua data lainnya dalam satu klaster;
b_i adalah hasil rata-rata jarak data ke-*i* terhadap semua data dari klaster lain, kemudian diambil data yang paling kecil;
$d\left(x_i^j, x_r^j\right)$ adalah jarak data ke-*i* dengan data ke-*r* dalam satu klaster *j*, sedangkan m_j merupakan jumlah data dalam klaster ke-*j*;
k adalah banyaknya klaster;
SI_i^j adalah rumus *Silhouette Index;*
SI_j adalah SI untuk setiap klaster;
SI adalah SI global.

Berikut ini merupakan ukuran nilai silhouette menurut Kaufman dan Rousseeuw [6]. Nilai silhouette coefficient (SC):

a. 0.7 < SC <= 1 *strong structure*
b. 0.5 < SC <= 0.7 *medium structure*
c. 0.25 < SC <= 0.5 *weak structure*
d. SC <= 0.25 *no structure*

IV. Hasil dan Pembahasan

Penelitian ini dilakukan dengan menggunakan data mahasiswa Teknik Informatika Strata Satu tahun angkatan 2012 dengan parameter SKS dan IPK sebanyak 636 *record* data. Sebelum memasuki proses klasterisasi, data dinormalisasi terlebih dahulu supaya tidak ada parameter yang mendominasi dalam perhitungan proses klasterisasi. Tabel data awal dapat dilihat pada tabel 1, sedangkan tabel data normalisasi dapat dilihat pada tabel 2. Untuk contoh perhitungan normalisasi adalah sebagai berikut:

Untuk data ke-1:

SKS→X: $\frac{99-32}{156-32} = \frac{67}{124} = 0.54$

IPK→Y: $\frac{3.04-0}{3.92-0} = \frac{3.04}{3.92} = 0.78$

Untuk data ke-2:

SKS→X: $\frac{84-32}{156-32} = \frac{52}{124} = 0.42$

IPK→Y: $\frac{3.23-0}{3.92-0} = \frac{3.23}{3.92} = 0.82$

Untuk proses klasterisasi digunakan nilai k=3, karena akan dilakukan pengelompokan data menjadi tiga klaster berdasarkan kemiripan data [7]. Kemudian dilakukan proses iterasi menghasilkan titik pusat akhir seperti pada tabel 3 dan hasil posisi anggota klaster pada tabel 4. Mengacu pada kolom titik pusat X dan Y pada tabel 3, prediksi dapat dikategorikan dengan indeks sebagai berikut:

1. Klaster dengan nilai X tertinggi merupakan kategori mahasiswa dengan kelulusan cepat.
2. Klaster dengan nilai X tengah merupakan kategori mahasiswa dengan kelulusan tepat waktu.
3. Klaster dengan nilai X terendah merupakan kategori mahasiswa dengan kelulusan lambat
4. Klaster dengan nilai Y tertinggi merupakan kategori mahasiswa dengan IPK tinggi.
5. Klaster dengan nilai Y tengah merupakan kategori mahasiswa dengan IPK sedang.
6. Klaster dengan nilai Y terendah merupakan kategori mahasiswa dengan IPK rendah.

Hasil analisis klastering mengacu pada tabel 3:

a) Kelompok klaster satu (C1) merupakan kelompok mahasiswa dengan kategori kelulusan tepat waktu dengan jumlah IPK sedang. mahasiswa yang terdapat pada klaster satu sebanyak 292.
b) Kelompok klaster dua (C2) merupakan kelompok mahasiswa dengan kategori kelulusan cepat dengan jumlah IPK tinggi. mahasiswa yang terdapat pada klaster dua sebanyak 274.
c) Kelompok klaster tiga (C3) merupakan kelompok mahasiswa dengan kategori kelulusan lambat dengan jumlah IPK rendah. mahasiswa yang terdapat pada klaster tiga sebanyak 70.

Hasil analisis klastering diatas dapat dilihat pada tabel 5.

Untuk hasil validasi data menggunakan *Silhouette Index* dapat dilihat pada tabel 6. Untuk ukuran nilai SI menurut Kaufman dan Rousseeuw, pada kolom SI klaster, dapat disimpulkan bahwa klaster pertama sebesar 0.53 (C1) termasuk dalam *medium structure*, sedangkan klaster kedua (C2) sebesar 0.46 termasuk dalam *weak structure*, dan klaster ketiga (C3) sebesar 0.31 termasuk dalam *weak structure.* Nilai SI global sebesar 0.43 juga termasuk dalam *weak structure.* Jumlah data hasil klaster yang tidak tepat dari data asli, untuk C1 sebanyak 5 data tidak tepat, C2 sebanyak 13 data tidak tepat, dan C3 sebanyak 9 data tidak tepat.

Tampilan program prediksi kelulusan mahasiswa dapat dilihat pada gambar 1.

V. PENUTUP

5.1 Kesimpulan

Kesimpulan yang didapat berdasarkan penelitian ini adalah:

1. Kelulusan mahasiswa dapat dikelompokkan kedalam tiga klaster berdasarkan jarak terdekat dengan beberapa parameter yaitu SKS dan IPK.
2. Analisa prediksi kelulusan mahasiswa dilakukan dengan tiga kategori pada setiap parameter kelulusan yaitu lulus cepat, lulus lambat, dan lulus tepat waktu, serta IPK tinggi, rendah, dan sedang.
3. Dari hasil percobaan 50 kali replikasi klastering, menggunakan 636 record data diperoleh hasil prediksi: sebanyak 292 mahasiswa termasuk dalam kelompok kelulusan tepat waktu dengan jumlah IPK sedang, 274 mahasiswa termasuk dalam kelompok kelulusan cepat dengan IPK tinggi, 70 mahasiswa termasuk dalam kelompok kelulusan lambat dengan IPK rendah.
4. Hasil dari klasterisasi diuji dengan perhitungan validasi menggunakan Silhouette Index (SI) menghasilkan nilai SI klaster, klaster pertama sebesar 0.53 (C1) termasuk dalam medium structure, sedangkan klaster kedua (C2) sebesar 0.46 termasuk dalam weak structure, dan klaster ketiga (C3) sebesar 0.31 termasuk dalam weak structure. Nilai SI global sebesar 0.43 termasuk dalam weak structure. Jumlah data hasil klaster yang tidak tepat dari data asli, untuk C1 sebanyak 5 data tidak tepat, C2 sebanyak 13 data tidak tepat, dan C3 sebanyak 9 data tidak tepat.
5. Dengan data mahasiswa yang diolah menggunakan metode klastering k-Means tersebut, maka dapat dilakukan pengambilan keputusan sebagai acuan tindakan preventif terhadap mahasiswa dengan kategori kelulusan lambat serta menjadi acuan dalam penerimaan jumlah mahasiswa baru dilihat dari jumlah prediksi kelulusan mahasiswa pada masing-masing kategori kelulusan.

5.2 Saran

Untuk meningkatkan kinerja dan menyempurnakan penelitian, beberapa saran yang dapat digunakan dalam penelitian selanjutnya adalah:

1. Parameter yang digunakan dalam penelitian ini hanya SKS dan IPK. Pada penelitian selanjutnya dapat mengembangkan dan menambah parameter seperti data absensi mahasiswa.
2. Mengembangkan GUI (Graphical User Interface) supaya tampilan lebih menarik dan user friendly.
3. Diharapkan dapat membandingkan dengan algoritma data mining yang lain agar mengetahui algoritma yang lebih baik dalam memprediksi kelulusan mahasiswa.

Daftar Pustaka

[1] M. Subkhan, "Algoritma Clustering," 10 2010. [Online]. Available: http://te.ugm.ac.id/. [Accessed 13 10 2014].

[2] BADAN AKREDITASI NASIONAL PERGURUAN TINGGI, Buku II Akreditasi Institusi Perguruan Tinggi, Jakarta: BAN-PT, 2011.

[3] T. Rismawan and S. Kusumadewi, "Aplikasi k-Means untuk Pengelompokan Mahasiswa Berdasarkan Nilai Body Mass Index (BMI) & Ukuran Kerangka," *Seminar Nasional Aplikasi Teknologi Informasi*, no. 1907-5022, pp. E43-E48, 2008.

[4] E. Prasetyo, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: ANDI, 2014.

[5] N. Atthina and L. Iswari, "Klasterisasi Data Kesehatan Penduduk untuk Menentukan Rentang Derajat Kesehatan Daerah dengan Metode k-Means," *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, no. 1907-5022, pp. B-52, 2014.

[6] L. Kaufman and P. J. Rousseuw, Finding Groups in Data, New York: John Wiley & Sons, 1990.

[7] N. M. Guchi, "Pengelompokan Mahasiswa Potensial Drop Out Menggunakan Metode Clustering Pada Program Studi Strata 1 Ilmu Komputer Dan Teknologi Informasi Universitas Sumatera Utara," Universitas Sumatera Utara, Medan, 2013.

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques Second Edition, San Francisco: Morgan Kaufmann Publishers, 2006.

[9] B. Mirkin, Clustering: A Data Recovery Approach Second Edition, United States of America: CRC Press, 2013.

[10] A. A. A. Hidayat, Pengantar Ilmu Kesehatan Anak Untuk Pendidikan Kebidanan, Jakarta: Salemba Medika, 2008.

Tabel 1. Data awal

Mhs ke-	SKS	IPK
1	99	3.04
2	84	3.23
3	90	2.5
4	84	2.46
5	79	2.62
6	143	3.21
7	91	3.37
8	145	2.9
9	79	2.57
10	142	3.15
.	.	.
.	.	.
632	132	3.11
633	156	3.33
634	90	3.24
635	88	3.27
636	147	3.21

Tabel 2. Data ternormalisasi

Mhs ke-	X	Y
1	0.54	0.78
2	0.42	0.82
3	0.47	0.64
4	0.42	0.63
5	0.38	0.67
6	0.9	0.82
7	0.48	0.86
8	0.91	0.74
9	0.38	0.66
10	0.89	0.8
.	.	.
.	.	.
632	0.81	0.79
633	1	0.85
634	0.47	0.83
635	0.45	0.83
636	0.93	0.83

Tabel 3. Titik pusat akhir

	X (SKS)	Y (IPK)
C1	0.40	0.70
C2	0.47	0.84
C3	0.34	0.39

Tabel 4. Posisi anggota klaster

Mhs ke-	Anggota Klaster
1	2
2	2
3	1
4	1
5	1
6	2
7	2
8	2
9	1
10	2
.	.
.	.
632	2
633	2
634	2
635	2
636	2

Tabel 5. Hasil analisis klastering

Mhs ke-	Anggota Klaster	Kategori kelulusan	IPK
1	2	cepat	tinggi
2	2	cepat	tinggi
3	1	tepat waktu	sedang
4	1	tepat waktu	sedang
5	1	tepat waktu	sedang
6	2	cepat	tinggi
7	2	cepat	tinggi
8	2	cepat	tinggi
9	1	tepat waktu	sedang
10	2	cepat	tinggi

.	.	.	.
.	.	.	.
632	2	cepat	tinggi
633	2	cepat	tinggi
634	2	cepat	tinggi
635	2	cepat	tinggi
636	2	cepat	tinggi

Tabel 6. Hasil *silhouette index*

SI klaster	0.53	0.46	0.31
SI global	0.43		

	C1	C2	C3
Data tidak tepat	5	13	9

Gambar 1. Tampilan program klaster