SEMESTA TEKNIKA
Vol. 21. No. 1, 60-64 , Mei 2018
DOI: 10.18196/st.211211

Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik)

(Application of Clustering Method With K-Means Algorithm in Data Grouping Prospective New Students at Universitas Muhammadiyah Yogyakarta (Case Study: Faculty of Medicine and Nursing, and Faculty of Social and Political Sciences)

ASRONI, HIDAYATUL FITRI, EKO PRASETYO

ABSTRACT

The increasing new prospective students in a University to make the stack more and more data, departing from it then conducted a search for new knowledge with data mining. Grouping data for prospective new students will be made by the method Clustering and used the algorithm k-means. In this penmaru there are 5 data attributes are used i.e., hometown, gender, status to qualify for selection, driveways, and majors. This analysis is performed using WEKA software and the source data taken from admissions data (penmaru) in the form of a data warehouse. Class from the use of this method is the attribute of the majors. Iteration performed as many as 3 times and the number of a cluster at the Faculty of medicine and health sciences, i.e. 4 clusters, Faculty of social and political science 3 clusters. Method Clustering can be applied to the classification of data for prospective new students. Another thing that can be analyzed from the results of the grouping candidate data, promotion strategies from each Department to increase the quantity and quality.

Keywords: data mining, k-means, clustering, penmaru, WEKA.

PENDAHULUAN

Setiap tahun ajaran baru peminat calon mahasiswa baru yang mendaftar di Universitas Muhammadiyah Yogyakarta sangat banyak. Tahun 2010 sampai 2015 pendaftar di Universitas Muhammadiyah Yogyakarta (UMY) tercatat 101.217 orang . Adapun calon mahasiswa baru tersebut tidak hanya dari kawasan Jawa Tengah saja, melainkan dari seluruh daerah di Indonesia bahkan dari luar negeri. Banyaknya calon mahasiswa baru di Universitas Muhammadiyah Yogyakarta akan berakibat semakin banyak pula data yang masuk dalam *server* database di Biro Sistem Informasi UMY. Semakin banyak data yang masuk, maka semakin menumpuk pula data tersebut, sehingga data yang berlimpah bertahun-tahun dapat diolah untuk menemukan informasi yang tersembunyi, tentunya informasi tersebut sangat berguna bagi pihak universitas.

Metode dalam menangani penumpukan volume data pada penerimaan mahasiswa baru di UMY adalah dengan penerapan *data mining*. Teknik *data mining* dapat mengolah data yang berlimpah menjadi informasi yang penting biasanya disebut *knowledge discovery database* (KDD). Adapun metode yang digunakan dalam pengelompokan data calon mahasiswa baru di UMY adalah metode *clustering*.

Menurut Narwati (2010) *clustering* merupakan cara untuk menemukan kelompok objek yang memiliki kemiripan dan dapat menemukan pola penyebaran dan pola hubungan dalam kumpulan data yang besar. Dalam proses *clustering* yang terpenting adalah mengumpulkan pola ke kelompok yang sesuai untuk menemukan persamaan dan perbedaan agar menghasilkan kesimpulan yang berharga.

Metode *clustering* menggunakan algoritma *k-means* dalam pengelompokan data pada calon mahasiswa baru di Universitas Muhammadiyah Yogyakarta. Metode *k-means cluster analysis* bisa menjadi solusi untuk pengklasifikasian karakteristik dari objek. Algoritma *k-means* memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma ini relatif

lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar. Selain itu algoritma *k-means* tidak terpengaruh dengan adanya urutan objek. (*Aranda* , 2016).

Asroni (2015) melakukan pengujian data yang telah ada di *data warehouse* UMM Magelang untuk mencari 5 orang mahasiswa jurusan Teknik Informatika dalam melakukan penyeleksian untuk mengikuti lomba. Adapun lomba yang akan diikuti adalah kompetisi *event Cyberjawara* yang diselenggarakan oleh *indonesia security incident response team on internet infrastructure* (ID SIRTII) Kementerian Komunikasi dan Informatika RI. Dengan penerapan Metode K-Means bisa diperoleh 1 *cluster* dengan IPK tertinggi untuk memilih 5 mahasiswa untuk mewakili lomba.

1. Rumusan Masalah

Berdasarkan latar belakang tersebut, permasalahan yang harus diselesaikan dalam penelitian adalah untuk mengetahui kesenjangan jumlah mahasiswa baru yang diterima terhadap jumlah pendaftar serta penerapan metode *clustering* dengan menggunakan algoritma *k-means* pada pengelompokkan data calon mahasiswa baru di Universitas Muhammadiyah Yogyakarta dengan studi kasus di Fakultas Kesehatan dan Ilmu Keperawatan dan Fakultas Ilmu Sosial dan Ilmu Politik. Dengan diketahui nilai kesenjangan tersebut akan menentukan kebijakan pimpinan dalam melakukan promosi pada masing-masing Fakultas.

2. Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini adalah:

a. Mengimplementasikan metode *clustering* dengan algoritma *k-means* dalam pengelompokkan data calon mahasiswa baru di Universitas Muhammadiyah Yogyakarta.
b. Mengetahui pengelompokkan data berdasarkan jurusan bagi calon mahasiswa baru Universitas Muhammadiyah Yogyakarta dari sistem pengambilan keputusan pada calon mahasiswa baru dengan algoritma *k-means*.

Data Mining

Menurut Fayyad dalam buku (Kusrini, 2009) Istilah *data mining* dan *knowledge discovery in database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Narwati, 2010):

1. *Data selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data dari hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing* atau *Cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang *inkosisten*, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*). juga dilakukan proses *enrichement*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

Coding adalah transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat bergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

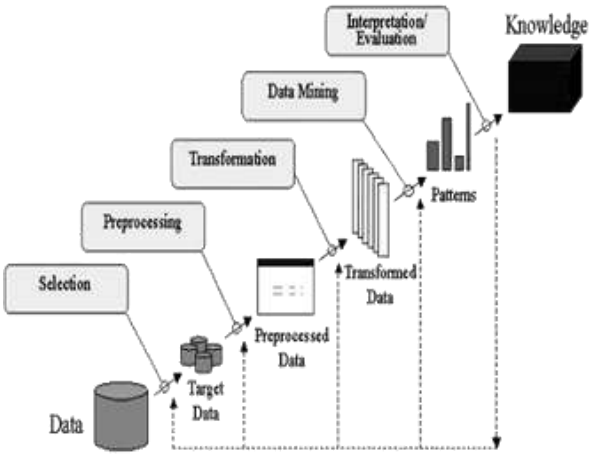
4. *Data mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation* atau *Evaluation*

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*.

Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.



GAMBAR 1. Proses *data mining* (Nasari, 2015).

Clustering

Menurut Eko Prasetyo (2012:6), pengelompokkan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada.

Pengelompokkan data dibedakan menurut struktur kelompok, keanggotaan data dalam kelompok, dan kekompakan data dalam kelompok. Menurut struktur, pengelompokkan dibagi dua, yaitu hierarki dan *partitioning*. Dalam hierarki, satu data tunggal bisa dianggap sebuah kelompok, dua atau lebih. Pengelompokkan *partitioning* membagi setiap data hanya menjadi anggota satu kelompok. Menurut keanggotaan data dalam kelompok, dibagi menjadi dua, yaitu ekslusif dan tumpang-tindih. Dalam kategori ekslusif, sebuah data bisa dipastikan hanya menjadi anggota satu kelompok dan tidak menjadi anggota kelompok yang lain. Sedangkan kategori tumpang-tindih adalah metode pengelompokkan yang membolehkan sebuah data menjadi anggota lebih dari satu kelompok. Menurut kategori kekompakan, pengelompokkan terbagi menjadi dua, yaitu komplet dan parsial. Jika semua data bisa bergabung menjadi satu, bisa dikatakan semua data kompak menjadi satu kelompok. Apabila ada satu atau dua data yang tidak ikut bergabung dalam kelompok mayoritas, data tersebut dikatakan mempunyai perilaku menyimpang. Data yang menyimpang dikenal dengan sebutan *outlier*, *noise* atau *uninterested background* (Eko Prasetyo, 2012:177).

K-Means

K-means merupakan salah satu metode pengelompokkan data *non-hierarki* yang mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang di *set* dalam suatu kelompok dan memaksimalkan variasi antar kelompok (Eko Prasetyo, 2012:178).

Metode *k-means* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam suatu kelompok mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dasar algoritma *k-means* adalah sebagai berikut:

1. Tentukan nilai *k* sebagai jumlah klaster yang ingin dibentuk.
2. Inisialisasi *k* sebagai *centroid* yang dapat dibangkitkan secara *random*.
3. Hitung jarak setiap data ke masing-masing *centroid* menggunakan persamaan *Euclidean Distance* yaitu sebagai berikut:

$$d(P,Q) = \sqrt{\sum_{j=1}^{p} (x_j(P) - x_j(Q))^2}$$

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroidnya*.
5. Tentukan posisi *centroid* baru (*k*).
6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama

METODE PENELITIAN

Penelitian ini menggunakan algoritma *k-means* untuk pengelompokkan data calon mahasiswa baru berdasarkan jurusan pada Fakultas Kedokteran dan Ilmu Keperawatan dan Fakultas Ilmu Sosial dan Ilmu Politik. Data yang diolah adalah data calon mahasiswa baru Universitas Muhammadiyah Yogyakarta (UMY).

Lokasi Penelitian

Penelitian ini dilaksanakan di Biro Sistem Informasi Universitas Muhammadiyah Yogyakarta. Data penerimaan mahasiswa baru di Universitas Muhammadiyah Yogyakarta telah dibangun sistem *data warehouse*.

HASIL DAN PEMBAHASAN

Pengumpulan Data

Sumber data utama yang digunakan dalam penelitian ini dari *data warehouse* penmaru UMY tahun 2010 sampai 2014.

Preprocessing Data

Setelah data diseleksi sesuai dengan atribut yang digunakan maka dilakukan *preprocessing* data, tujuannya adalah agar tidak adanya duplikasi data, tidak *missing value* dan memperbaiki kesalahan-kesalahan yang ada pada *dataset*. Tahap ini data akan dilakukan *cleaning* atau pembersihan data, sehingga data trsebut dapat diolah dan dilakukan proses *data mining*.

Algoritma K-Means

Dalam menggunakan algoritma *k-means* akan melakukan pengulangan tahapan hingga terjadi kestabilan. Adapun tahapannya sebagai berikut.

1. Menentukan jumlah *cluster* dan menentukan koordinat titik tengah *cluster*. Penentuannya berdasarkan frekuensi kurang, frekuensi sedang dan frekuensi tinggi secara acak seperti tabel 1.
2. Penentuan nilai dari *cluster* untuk dijadikan acuan dalam melakukan perhitungan jarak objek ke *centroid*, perhitungan jarak mengacu pada rumus *euclidean*. Hasil perhitungan antar *centroid* dapat dilihat pada tabel 2.

$$d(P,Q) = \sqrt{\sum_{j=1}^{p} (x_j(P) - x_j(Q))^2}$$

3. Dilakukan pengelompokkan *centroid* sesuai dengan hasil dari jarak antar *centroid* tersebut. Hasilnya digunakan untuk penentuan kelompok *clustering*.

Pengujian dengan Software WEKA

Pengujian data dengan *menggunakan software WEKA* menghasilkan data sebagai berikut:

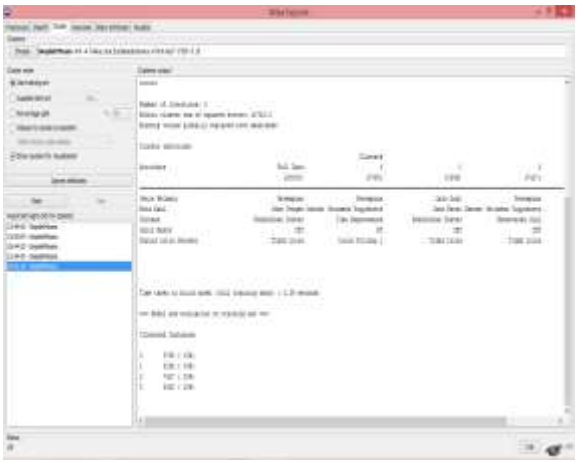
1. Nilai *cluster centroid* seperti pada gambar 2.
2. Grafik *clustering* dari calon mahasiswa baru dalam memilih jurusan seperti pada gambar 3.

TABEL 1. Penentuan frekuensi dalam menentukan jarak *centroid*.

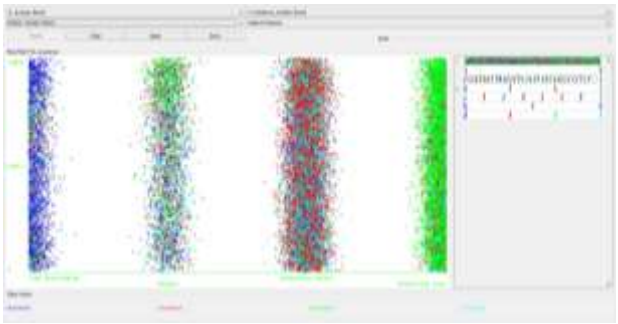
Centroid					
Centroid 1	1.2	1.8	1.3	1.1	1.1
Centroid 2	1.5	16.7	2.2	3.6	1.9
Centroid 3	1.7	18.3	2.6	5.3	2.5
Centroid 4	1.9	34.5	3.8	7.7	2.8

TABEL 2. Jarak antara *centroid*.

Centroid 1	Centroid 2	Centroid 3	Centroid 4
3.06	14.1	16.03	32.26
1.48	14.03	16.05	32.38
27.24	12.45	11.32	8.32
27.23	12.5	11.41	8.59
2.40	13.8	15.63	32.02
2.40	13.8	15.63	32.02
8.23	7.98	9.56	25.86
5.91	10.25	12.27	28.36
2.93	13.77	15.55	31.91
10.26	5.47	7.80	23.61



GAMBAR 2. Hasil *cluster centroid* dan *cluster instances* dengan *WEKA*.



GAMBAR 3. Hasil grafik *clustering*.

Berdasarkan hasil *cluster instances* menggunakan *software WEKA* data penerimaan calon mahasiswa baru Fakultas Kedokteran dan Ilmu Keperawatan pada setiap *cluster* adalah sebagai berikut.

1. *Cluster* 0 dengan jurusan Ilmu Keperawatan, sebanyak 3795 pendaftar dari jumlah 25000 calon mahasiswa baru (15%).
2. *Cluster* 1 dengan jurusan Pendidikan Dokter, sebanyak 8296 pendaftar dari jumlah 25000 calon mahasiswa baru (33%).
3. *Cluster* 2 dengan jurusan Kedokteran Gigi, sebanyak 7427 pendaftar dari jumlah 25000 calon mahasiswa baru (30%).
4. *Cluster* 3 dengan jurusan Pendidikan Dokter, sebanyak 5482 pendaftar dari jumlah 25000 calon mahasiswa baru (22%).

KESIMPULAN

Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa Pendidikan Dokter dan Ilmu Hubungan Internasional menjadi jurusan pilihan calon mahasiswa baru Universitas Muhammadiyah Yogyakarta. Hasil penelitian ini menjadi acuan pihak universitas untuk melakukan strategi promosi ke calon mahasiswa baru.

DAFTAR PUSTAKA

Aranda, J., Natasya, WAG. 2016. "Penerapan Metode K-Means Cluster Analysis Pada Sistem Pendukung Keputusan Pemilihan Konsentrasi Untuk Mahasiswa *International Class* STMIK AMIKOM Yogyakarta" dalam Jurnal Karya Ilmiah Teknik Informatika. Volume 4, No 1.

Asroni., Adrian, R. 2015. " Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang" dalam Jurnal Ilmiah Semesta Teknika. Volume 18. No 1.

Fadlika Dita Nurjanto. 2013. *Tahap-tahap K-Means Clustering*. https://fadlikadn.wordpress.com/2013/06/14/tahap-tahap-k-means-clustering/, 24 Agustus 2016.

Hermawati, F. A. Data Mining. 2013. *Andi: Yogyakarta*.

Kusrini, E. T. L. (2009). Algoritma Data Mining. *Yogyakarta: Andi Offset*.

Narwati. 2010. "Pengelompokkan Mahasiswa Menggunakan Algoritma K-Means'' dalam jurnal Dinamika Informatika. Volume 2, No 2.

Nasari, F., & Darma, S. (2013). Penerapan K-Means Clustering pada Data Penerimaan Mahasiswa Baru (Studi Kasus: UNIVERSITAS POTENSI UTAMA). *SEMNASTEKNOMEDIA ONLINE*, *3*(1), 2-1.

Ong, J. O. (2013). Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University.

Prasetyo, E. Data Mining: Konsep Dan Aplikasi Menggunakan MATLAB. 2012. *Penerbit ANDI. Yogyakarta*.

PENULIS:

Asroni

Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Yogyakarta, Yogyakarta.

Email: asroni@umy.ac.id

Hidayatul Fitri

Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Yogyakarta, Yogyakarta.

Email: hida.unayaa@gmail.com

Eko Prasetyo

Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Yogyakarta, Yogyakarta.

Email: eko.prasetyo@umy.ac.id