KLASTERISASI CALON MAHASISWA BARU MENGGUNAKAN ALGORITMA K-MEANS

CLASTERIZATION OF PROSPECTIVE STUDENTS USING K-MEANS ALGORITHM

Adi Sucipto[1]
Sekolah Tinggi Multi Media "MMTC" Yogyakarta[1]
adi.sucipto@kominfo.go.id[1]

Abstract

Information Technology is rapidly developing and its use increasingly widespread. Data processing for decision making can be done with the help of Information Technology devices. Like wise the decision making of new student's admission. STMM use 2 (two) types of selection's system, namely CBT and interviews that is conducted at the same day. Selection by combining 2 (two) score of the examination with certain procentage, so that the rank of prospective student is obtained. Based on these problems, this article present a result of student clusterization by using K-Means Algorithm. Clustering by both of scores of examination. This article's purpose is to describe the pattern of cluster of students's scores. As the results, there are 4 clusters, in the first cluster with a total of 1059 prospective student data, there are 132 who passed the selection or 22.147% of the total of student that pass the selection. Then in the second cluster, out of 424 members, 12 pass the selection or around 2% of the total. Then in the third cluster, out of 659 members, 447 pass the selection or 75% of the total. And in the fourth cluster, out of 547 members, only 5 pass the selection or 0.8% of the total.

Keywords: *K-Means, clusterization, database, information.*

Abstrak

Teknologi Informasi berkembang dengan pesat dan penggunaannya juga semakin luas. Pengolahan data hingga pengambilan keputusan dapat dilakukan dengan bantuan perangkat Teknologi Informasi. Demikian juga dengan pengambilan keputusan untuk penerimaan mahasiswa baru. Sistem seleksi penerimaan mahasiswa baru di STMM menggunakan 2 (dua) jenis ujian yaitu CBT (Computer Based Test) dan wawancara yang dilaksanakan pada hari yang sama. Seleksi dengan menggabungkan 2 (dua) nilai tersebut dengan prosentase tertentu, sehingga didapatkan rangking calon mahasiswa baru. Berdasar pada permasalahan tersebut penulis pada kesempatan ini akan membuat klasterisasi data calon mahasiswa baru tersebut berdasar nilai CBT dan wawancara menggunakan algoritma K-Means. Tujuan dari penelitian ini adalah untuk mengetahui pola pengelompokan nilai CBT dan wawancara calon mahasiswa baru STMM Yogyakarta. Hasilnya Ada empat klaster, klaster pertama dengan anggota sebanyak 1059 data calon mahasiswa, terdapat 132 yang lolos seleksi atau 22,147% dari total lulus. Kemudian pada klaster kedua, dari 424 anggota terdapat 12 yang lolos seleksi atau sekitar 2% dari total lulus. Kemudian pada klaster ketiga, dari 659 anggota terdapat 447 yang lolos seleksi atau 75% dari total lulus . Dan pada klaster keempat, dari 547 anggota hanya 5 yang lolos seleksi atau 0,8% dari total lulus.

Kata Kunci: K-Means, klasterisasi, database, informasi.

PENDAHULUAN

Teknologi Informasi berkembang dengan pesat dan penggunaannya juga semakin luas. Pengolahan data hingga pengambilan keputusan dapat dilakukan dengan bantuan perangkat Teknologi Informasi. Demikian juga dengan pengambilan keputusan untuk penerimaan mahasiswa baru. Pemerintah menggunakan model seleksi CAT (*Computer Assisted Test*) untuk mendapatkan nilai calon mahasiswa yang kemudian dapat digunakan untuk mendaftar di Perguruan Tinggi Negeri tanpa harus mengikuti ujian ulang, cukup dengan mendaftarkan hasil CAT tersebut ke Perguruan Tinggi Negeri yang diinginkan (Kemenristekdikti, 2018).

STMM yang dulu dikenal sebagai lembaga diklat penyiaran ahli madya multi media MMTC (Multi Media Training Center) mulai menyelenggarakan pendidikan vokasi dibidang penyiaran pada tahun 2014 (STMM, 2015). Sekolah Tinggi Multi Media saat ini belum tergabung dalam seleksi yang diadakan oleh Ristekdikti. Sehingga untuk seleksi penerimaan mahasiswa baru menggunakan sistem CBT sendiri sejak tahun 2014. Sistem seleksi penerimaan mahasiswa baru di STMM menggunakan 2 (dua) jenis ujian yaitu CBT (*Computer Based Test*) dan wawancara yang dilaksanakan pada hari yang sama. Seleksi dengan menggabungkan 2 (dua) nilai tersebut dengan prosentase tertentu, sehingga didapatkan rangking calon mahasiswa baru .

Data yang terkumpul dalam database pada dasarnya masih dapat untuk diolah dan menjadi sumber informasi baru. Pengolahan data dalam database yang biasa dikenal dengan KDD (Knowledge Discovery in Database) adalah untuk mencari informasi baru yang tersembunyi dalam berlimpahnya data (Asroni, Fitri, & Prasetyo, 2018). Pencarian informasi baru tersebut dapat menggunakan metode clustering, yaitu mengelompokkan data dengan parameter yang sejenis sehingga berbeda dengan pengelompokan yang lain.

Klasterisasi yang diharapkan adalah beberapa klaster yang dapat menggambarkan pola kombinasi antara nilai CBT dan wawancara. Variabel nilai CBT dan wawancara dipilih karena kedua variabel tersebut digunakan sebagai pertimbangan untuk kelulusan seleksi. Konsistensi data kelulusan saat ini belum dapat digambarkan secara jelas, karena saat ini STMM tidak menggunakan ‘batas nilai kelulusan’ sebagai acuan dalam seleksi. Sehingga nilai calon mahasiswa akan beragam tiap periode dan juga untuk tiap program studi. Solusi klasterisasi adalah untuk melihat ragam nilai dan kecenderungan kelulusan seleksi.

Data calon mahasiswa baru beserta nilainya saat ini tersimpan di database STMM. Data yang terkumpul ini dapat digunakan untuk dasar pengelompokan calon mahasiswa berdasar nilai CBT dan wawancara yang kemudian dapat digunakan untuk membantu sistem seleksi penerimaan mahasiswa baru di STMM. Dan berdasar pada permasalahan tersebut penulis pada kesempatan ini akan membuat klasterisasi data calon mahasiswa baru tersebut berdasar nilai CBT dan wawancara menggunakan algoritma K-Means. Tujuan dari penelitian ini adalah untuk mengetahui pola pengelompokan nilai CBT dan wawancara calon mahasiswa baru STMM Yogyakarta.

METODE PENELITIAN

1. Klasterisasi

Klasterisasi adalah salah satu teknik yang digunakan dalam data mining. Pengertian klasterisasi dalam data mining adalah suatu teknik untuk mengelompokkan data kedalam suatu klaster tertentu yang memungkinkan data dalam klaster tersebut memiliki kesamaan dan memiliki perbedaan yang jelas dengan data pada klaster lainnya. Perbaikan model klasterisasi

masih menjadi kajian tersendiri para ilmuan.

Klasterisasi dan klasifikasi adalah 2 (dua) hal yang berbeda, karena dalam klasifikasi sebuah objek dipilah dalam kelas yang telah ditentukan, sedangkan pada klasterisasi kelas dibuat saat memilah objek (Rohma, Ismail, & Waluyo, 2018). Sehingga pemilahan data nilai mahasiswa akan lebih sesuai dengan metode klaterisasi, karena tidak ada penentuan kelas diawal pemilahan data nilai calon mahasiswa. Metode klasterisasi telah banyak digunakan dalam klasterisasi data mahasiswa. Variabel yang digunakan untuk klasterisasi dapat berupa nilai UAN, jenis sekolah (SMA,SMK dan lainnya) dan program studi pilihan (Yunita, 2018), kemudian ada yang menggunakan variabel pilihan program studi (Asroni et al., 2018), dan ada juga yang menggunakan variabel program studi, nilai hasil tes, dan nama sekolah (Chasanah, 2017).

Masing-masing klasterisasi tersebut adalah untuk mencari pola pengelompokan data. Sehingga pada akhirnya hasil pengelompokan data tersebut dapat digunakan untuk mendukung pengambilan keputusan bagi manajemen lembaga itu sendiri. Dan berlimpahnya data calon mahasiswa di STMM juga dapat digunakan sebagai pertimbangan pengambilan keputusan utamanya untuk seleksi penerimaan mahasiswa baru.

2. K-Means

Algoritma klasterisasi yang digunakan pada artikel ini adalah K-Means yaitu sebuah algoritma yang dapat mendefinisikan objek dalam suatu pusat kelompok data yang biasanya menjadi titik-tengah dari kelompok data tersebut (Wahyudi & Jananto, 2013). Algoritma ini akan melakukan perulangan untuk mendapatkan titik-tengah yang optimal. Dan K-Means ini akan membagi klaster sesuai dengan jumlah klaster yang telah ditentukan atau diinisiasi diawal saat menjalankan algoritma ini.

Algoritma K-Means mengikuti alur sebagai berikut :

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

Penghitungan jarak antara data dengan centroid menggunakan *euclidean distance* dengan persamaan sebagai berikut:

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

Kemudian untuk menentukan centroid berikutnya pada langkah 5 yaiu dengan persamaan :

$$C(i) = \frac{x_1+x_2+x_3+\ldots+x_n}{\sum x} \quad (2)$$

dengan C adalah *centroid* baru yang dihasilkan berdasarkan data pada klaster yang terbentuk.

3. Pengumpulan Data

Data yang dikumpulkan adalah data calon mahasiswa STMM yang terdaftar pada SIPMB STMM. Data meliputi seluruh calon mahasiswa yang telah mengikuti tahapan pendaftaran sampai dengan registrasi, tetapi pada penulisan ini hanya akan dibatasi untuk periode 2018/2019. Daftar calon mahasiswa baru ini tidak dipisahkan sesuai periode pendaftaran ataupun program studi yang dipilih. Data yang terkumpul sebanyak 5560 calon mahasiswa, yaitu data calon mahasiswa yang mendaftar melalui SIPMB STMM Yogyakarta.

4. Preprocessing Data

Data calon mahasiswa yang masih utuh ini kemudian akan diolah untuk mengambil data yang lengkap penilaiannya, sehingga data yang tidak lengkap dapat diabaikan dan dihapus dari dataset. Variabel yang dipilih adalah nilai CBT dan nilai wawancara, sehingga kedua kolom ini harus dipastikan ada nilainya. Setelah proses ini didapatkan sebanyak 2689 data, yaitu data calon mahasiswa yang mempunyai nilai CBT dan nilai wawancara.

Pada tabel 1 dan tabel 2 adalah dataset sebelum dan sesudah memastikan kolom nilai CBT dan Wawancara harus terisi. Pada tabel 1 pada baris 1,3 dan 4 menunjukkan bahwa nilai pada kolom CBT dan wawancara berisi NaN atau tidak ada nilainya. Sehngga setelah menghapus baris itu maka pada tabel 2 dapat dilihat bahwa baris itu sudah dihapus dan digantikan oleh baris dibawahnya.

Table 1. Dataset Sebelum Hapus Baris

No	No Test	CBT	Wawancara	Total nilai
0	201805040	23.75	76.0	44.65
1	201802006	NaN	NaN	NaN
2	201802396	42.50	88.0	60.70
3	201804696	NaN	NaN	NaN
4	201801328	NaN	NaN	NaN

Table 2. Dataset Setelah Hapus Baris

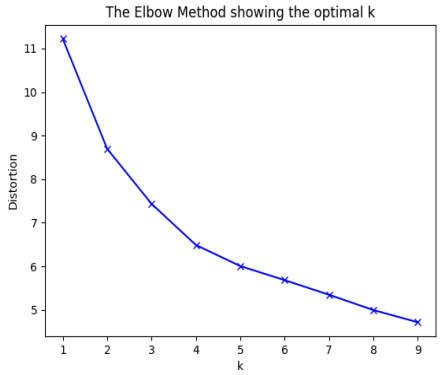
No	No Test	CBT	Wawancara	Total nilai
0	201805040	23.75	76.0	44.65
2	201802396	42.50	88.0	60.70
7	201803140	51.25	80.0	62.75
8	201804996	37.50	78.0	53.70
9	201803475	55.00	81.0	65.40

K-Means adalah algoritma yang cocok untuk mengolah dataset numerik, sehingga data yang akan diolah dipastikan berupa numerik. Nilai CBT dan Wawancara adalah 2 (dua) variabel yang ada dalam dataset dan secara default adalah numerik. Dan pengolahan data dapat dilanjutkan tanpa harus mengubah data pada kolom tersebut.

5. Penentuan Jumlah Klaster

Pada penggunaan Algoritma K-Means hal yang paling penting adalah penentuan jumlah klaster. Inisiasi awal klaster menjadi penting karena algoritma akan menginisasi centroid tiap klaster secara random saat algoritma dijalankan. Jumlah klaster optimal dapat ditentukan menggunakan *elbow methods* (Python Tutorial, 2017).

Metode ini akan menghitung semua kemungkinan jumlah klaster dari 1 – 10, pada metode ini dapat kita lihat seberapa optimal jumlah klaster digambarkan pada grafik membuat siku seperti pada Gambar 1 . Metode ini membandingkan nilai fungsi cosinus antara distorsi (sumbu Y) dan k (jumlah klaster) pada sumbu X. Dan lekukan membentuk siku (*elbow*) inilah yang dianggap sebagai jumlah klaster optimal. Nilai k optimal terlihat pada lekukan (siku) menunjuk pada k=4, dan inilah jumlah klaster optimal pada dataset tersebut.



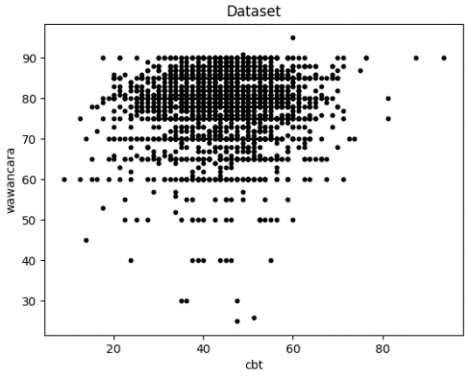
Gambar 1. Elbow Methods dan K Optimal

HASIL DAN PEMBAHASAN

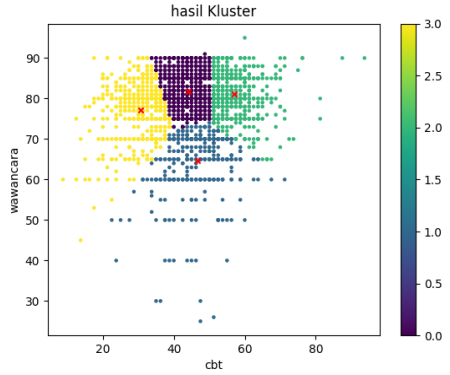
Implementasi algoritma K-Means menggunakan bahasa pemrograman python dan library yang digunakan adalah sklearn. Library sklearn telah memiliki modul untuk klasterisasi K-Means, sehingga implementasi lebih cepat. Jumlah klaster yang digunakan adalah 4 klaster, sesuai dengan penghitungan elbow methods. Pengaturan perulangan untuk pencarian centroid diatur pada maksimal perulangan 300 kali hingga ditemukan centroid yang tetap.

Pada gambar 3 dapat dilihat sebuah klaster tunggal yang merepresentasikan sebaran nilai CBT dan wawancara dari calon mahasiswa. Data tersebut adalah dataset dari 2689 calon mahasiswa yang mengikuti tahapan seleksi. Rentang nilai CBT pada sumbu X dan nilai wawancara pada sumbu Y. Sebuah klaster besar dari data nilai yang belum dapat dibaca secara detail karena belum terlihat pembeda klasternya.

Setelah klasterisasi dijalankan didapatkan 4 (empat) buah klaster dengan centroidnya. Klaster pertama dengan centroid di (43.98253069 81.84135977), klaster kedua dengan centroid di (46.62146226 64.75471698), klaster ketiga dengan centroid di (56.97647951 81.08801214) dan klaster keempat dengan centroid di (30.72212066 77.22486289). Pada gambar 4 semua titik centroid ditandai dengan huruf ‘x’.



Gambar 2. Sebaran Data Sebelum Klaterisasi



Gambar 3. Klaster Terbetuk

Pada klaster pertama beranggotakan 1059 buah data, klaster kedua beranggotakan 424 buah data, klaster ketiga beranggotakan 659 buah data, dan klaster keempat beranggotakan 547 buah data. Rentang nilai pada masing-masing klaster adalah yang mempunyai jarak terdekat pada centroid klaster tersebut. Sehingga dapat dilihat pada klaster-klaster tersebut terpisah dengan jelas menurut warna masing-masing.

Kemudian untuk melihat hubungan tiap klaster dengan kelulusan seleksi calon mahasiswa dapat dilihat pada tabel 3 . Berdasar pada data kelulusan secara total seleksi PMB STMM tahun 2018/2019 adalah sebanyak 596 calon mahasiswa dinyatakan lulus. Berikut adalah jumlah kelulusan ditiap klaster. Pada klaster pertama dengan anggota sebanyak 1059

data calon mahasiswa, terdapat 132 yang lolos seleksi atau 22,147% dari total lulus. Kemudian pada klaster kedua, dari 424 anggota terdapat 12 yang lolos seleksi atau sekitar 2% dari total lulus. Kemudian pada klaster ketiga, dari 659 anggota terdapat 447 yang lolos seleksi atau 75% dari total lulus. Dan pada klaster keempat, dari 547 anggota hanya 5 yang lolos seleksi atau 0,8% dari total lulus.

Dari data kelulusan tiap klaster, maka pada klister pertama dan ketiga yang memberikan data paling relevan untuk kelulusan seleksi PMB di STMM. Pada kedua klaster tersebut menunjukkan rentang nilai yang tinggi untuk nilai CBT dan juga wawancaranya. Sedangkan pada klaster kedua dan keempat cenderung menunjukkan nilai CBT yang relatif rendah, ataupun nilai wawancara yang relatif rendah dan tidak berimbang antara nilai CBT dan wawancaranya.

Table 2. Klaster dan Lulus Seleksi

Cluster	Jumlah	Lulus
0	1059	132
1	424	12
2	659	447
3	547	5

KESIMPULAN

Berdasarkan *elbow methods* ada 4 (empat) klaster optimal untuk klasterisasi data calon mahasiswa yang mengikuti ujian di STMM. Dari ke empat klaster pada tabel 3 dapat dilihat bahwa ada 2 (dua) klaster yang menunjukkan jumlah diterimanya signifikan yaitu pada klaster pertama dan klaster ketiga. Sehingga dapat diambil kesimpulan bahwa klasterisasi dapat dijadikan pedoman untuk seleksi calon mahasiswa baru.

DAFTAR PUSTAKA

Asroni, A., Fitri, H., & Prasetyo, E. (2018). 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). Semesta Teknika, 21(1), 60–64.

Chasanah, T. T. (2017). Penentuan Strategi Promosi Penerimaan Mahasiswa Baru Dengan Algoritma Clustering K-Means. IC-Tech, XII(1), 6.

Kemenristekdikti. (2018, October 22). Siaran Pers No: 198/SP/HM/BKKP/X/2018. Retrieved September 18, 2019, from Kementerian Riset, Teknologi, Dan Pendidikan Tinggi Republik Indonesia website: https://ristekdikti.go.id/kabar/skema-baru-seleksi-masuk-ptn-2019/.

Python Tutorial. (2017, July 10). kmeans elbow method | Python Tutorial. Retrieved September 19, 2019, from Python Tutorial website: https://pythonprogramminglanguage.com/kmeans-elbow-method/.

Rohma, F. F., Ismail, I. E., & Waluyo, Y. S. (2018). Implementation Of Kmeans Clustering On SIPP-KLING Dashboard Applications. MULTINETICS, 4(2), 38–42. https://doi.org/10.32722/multinetics.Vol4.No.2.2018.pp.38-42

STMM. (2015). Sekolah Tinggi Multi Media “MMTC” Yogyakarta. Retrieved September 19, 2019, from Sekolah Tinggi Multi Media website: https://mmtc.ac.id/index.php/menu/Menu/index/625/Sekilas%20STMM.

Wahyudi, E. N., & Jananto, A. (2013). Final Report Penilaian Kinerja

Dosen oleh Mahasiswa pada Satu Periode Tahun Akademik menggunakan Teknik Klustering (Studi Kasus : Universitas Stikubank Semarang). Dinamik, 18(2). Retrieved from https://www.unisbank.ac.id/ojs/index.php/fti1/article/view/1698.

Yunita, F. (2018). Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru. SISTEMASI, 7, 238. https://doi.org/10.32520/stmsi.v7i3.388.