PROSIDING Seminar Nasional SISFOTEK ISSN 2597-3584 (Media Online)
Sistem Informasi dan Teknologi
website : www.seminar.iaii.or.id | email : seminar@iaii.or.id

Clustering Data Penerimaan Mahasiswa Baru Universitas Handayani Makassar Menggunakan Algoritma K-Means

A. Ade Rosali Saputra[1], Samsart Deandi Palumery[2], Hazriani[3], Yuyun[4]
[1,2,3]Pascasarjana, Ilmu Komputer, Universitas Handayani, Makassar, Indonesia
[4]Badan Riset dan Inovasi Nasional (BRIN)
andi.adherosalisapoetra@gmail.com

Abstract

This research aims to group new student data for the 2022 academic year at Handayani University Makassar by utilizing a data mining process using the K-Means Clustering algorithm. Implementation using Rapidminer software is used to help determine accurate values. The data used in carrying out this research consists of 4 (four) attributes, namely, school origin, average UAS (final semester exam) score, gender and chosen study program. The research process begins by selecting data and then transforming the data into a numerical group. It is hoped that the results of this research can help universities in improving appropriate promotional strategies in each study program at Handayani University, Makassar.

Keywords: data mining, clustering, k-means, handayani university, computer systems.

Abstrak

Penelitian ini bertujuan untuk mengelompokkan data Mahasiswa Baru tahun akademik 2022 pada Universitas Handayani Makassar dengan memanfaatkan proses data mining menggunakan algoritma K-Means Clustering. Implementasi menggunakan software Rapidminer digunakan untuk membantu menetukan nilai akurat. Data yang digunakan dalam mengerjakan penelitian ini terdiri dari 4 (empat) atribut yaitu, asal sekolah, nilai rata-rata UAS (ujian akhir semester), jenis kelamin dan program studi yang dipilih. Proses penelitian dimulai dengan cara seleksi data lalu mentransformasikan data tersebut kedalam sebuah kelompok numerik. Hasil dari penelitian ini diharapkan dapat membantu Perguruan Tinggi dalam meningkatkan strategi promosi yang tepat dalam setiap program studi yang dimiliki oleh Universitas Handayani Makassar.

Kata kunci: data mining, clustering, k-means, universitas handayani, sistem komputer.

1. Pendahuluan

Didalam Pendidikan tinggi data sangat diperlukan untuk melengkapi setiap proses administrasi Mahasiswa mulai dari diterimanya sebagai Mahasiswa aktif sampai pada akhirnya menjadi alumni. Tentunya dari data-data tersebut bisadiketahui beberapa Profil penting dari setiap Mahasiswa dan kesempatan bagi pihak Universitas ketika mendapatkan data-data informasi dari setiap Mahasiswa, selain sebagai bentuk arsip bagi pihak Universitas, dari data-data tersebut bisa digunakan untuk menganalisis kondisi dari setiap mahasiswa sebelum mendaftar pada Universitas tertentu. Sebagai contoh yaitu Nama Mahasiswa, Nomor Formulir, Nama Maba, Asal Sekolah, Daerah/Kota Sekolah, Tanggal dan Tahun Lahir, Nilai Ujian Akhir Semester (UAS), Jenis Kelamin dan terakhir Jurusan atau Prodi yang dipilih ketika masuk ke Perguruan Tinggi. Ketika pihak kampus mendapatkan data ini, maka akan semakin mudah bagi pihak manajemen universitas untuk mencari sekolah-sekolah yang mana saja yang paling efektif untuk dikunjungi dalam hal mempromosikan Universitas[1].

Didalam Pendidikan tinggi data sangat diperlukan untuk melengkapi setiap proses administrasi Mahasiswa mulai dari diterimanya sebagai Mahasiswa aktif sampai pada akhirnya menjadi alumni. Tentunya dari data-data tersebut bisa diketahui beberapa Profil penting dari setiap Mahasiswa dan kesempatan bagi pihak Universitas ketika mendapatkan data-data informasi dari setiap Mahasiswa, selain sebagai bentuk arsip bagi pihak Universitas, dari data-data tersebut bisa digunakan untuk menganalisis kondisi dari setiap mahasiswa sebelum mendaftar pada Universitas tertentu. Sebagai contoh yaitu Nama Mahasiswa, Nomor Formulir, Nama Maba, Asal Sekolah, Daerah/Kota Sekolah, Tanggal dan Tahun Lahir, Nilai Ujian Akhir Semester (UAS), Jenis Kelamin dan terakhir Jurusan atau Prodi yang dipilih ketika masuk ke Perguruan Tinggi. Ketika pihak kampus mendapatkan data ini, maka akan semakin mudah bagi pihak manajemen universitas untuk mencari sekolah-sekolah yang mana saja yang paling efektif untuk dikunjungi dalam hal mempromosikan Universitas[1].

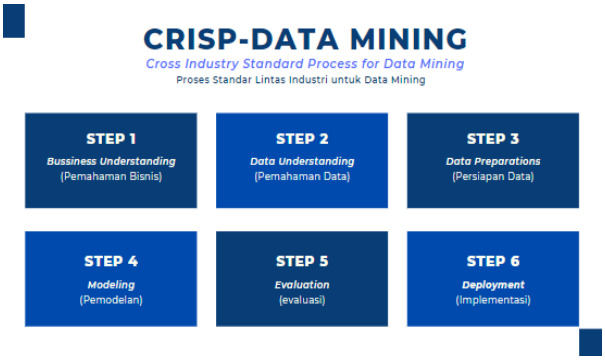
2. Metode Penelitian

2.1. Data

Data yang digunakan dalam penelitian ini bersumber dari rekapitulasi data penerimaan MABA (Mahasiswa Baru) tahun akademik 2022 Universitas Handayani Makassar, dimana mahasiswa yang di terima menjadi mahasiswa baru pada Universitas ini sebanyak 200 data MABA.

2.2. Pendekatan Proses Mining

CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah metodologi panduan fleksibel yang membantu tim proyek untuk mengelola dan menjalankan proyek analisis data dengan efektif [4]. Urutan pendekatan CRISP-DM dapat dilihat pada gambar 1.



Gambar 1. CRISP Data Mining

Pada gambar 1 telihat urutan Pemahaman Bisnis: Memahami tujuan bisnis dan tujuan proyek. Identifikasi masalah atau peluang yang ingin diatasi. Pemahaman Data: Mengumpulkan dan memahami data yang tersedia. Evaluasi kualitas, relevansi, dan ketersediaan data. Persiapan Data: Membersihkan, mentransformasi, dan mempersiapkan data untuk analisis. Langkah ini mencakup pemilihan variabel yang relevan dan penggabungan data dari sumber yang berbeda. Modeling: Memilih metode analisis dan membangun model yang sesuai dengan tujuan proyek. Ini melibatkan eksperimen dengan berbagai algoritma dan pendekatan untuk menghasilkan model yang efektif. Evaluasi: Mengukur kinerja model menggunakan metrik yang relevan dengan tujuan bisnis. Model dievaluasi dan dianalisis untuk memastikan bahwa ia memenuhi kriteria keberhasilan yang telah ditetapkan. Implementasi: Solusi yang dikembangkan diterapkan dalam lingkungan bisnis sesuai dengan rencana implementasi yang telah dirancang.

Pendekatan CRISP-DM bersifat siklus, yang berarti bahwa langkah-langkah ini tidak selalu berlangsung dalam urutan linear. Ada kemungkinan untuk kembali ke langkah sebelumnya jika diperlukan perbaikan atau penyesuaian. CRISP-DM metodologi yang digunakan untuk mengatasi proyek analisis data dan data mining[5].

2.3. Metode Tahap Persiapan

Membersihkan, mentransformasi, dan mempersiapkan data untuk analisis. Langkah ini mencakup pemilihan variabel yang relevan atau yang disebut *fitur-selection*. Hal ini bertujuan untuk menyesuaikan karakteristik model proses mining yang akan digunakan. Pada penelitian ini terdiri dari dua metode yang digunakan yaitu seleksi data dana transformasi data.

2.3.1. Seleksi Data

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam Knowledge Discovery in Database (KDD) dimulai [6]. Atribut (kolom) yang terdapat pada dataset ini terdiri dari 8 atribut atau kolom, namun atribut yang akan diambil untuk dilakukan klasterisasi hanya 4 atribut yaitu kolom asal sekolah, nilai rata-rata UAS, jenis kelamin dan prodi (program studi) pada tabel 1.

Tabel 1. Data Seleksi

No. FORMULIR	ASAL SEKOLAH	NILAI RATA-RATA UAS	JENIS KELAMIN	PRODI
0001	SMAN 5	83.83	Perempuan	(S1) Teknik Informatika
0002	SMA 9	81.53	Laki-laki	(S1) Teknik Informatika
0003	SMAN 15	87.15	Laki-laki	(S1) Teknik Informatika
0004	SMAN 3	90.83	Perempuan	(S1) Teknik Informatika
0005	MA Al-Munaawwarah	79.76	Laki-laki	(S1) Teknik Informatika
0006	SMKN 5	83.85	Laki-laki	(S1)Sistem Komputer
0007	SMKN 2	79.82	Perempuan	(S1)Sistem Komputer
0008	SMAN 2	89.99	Perempuan	(S1)Sistem Informasi
0009	MA BPII Pamboang	90.52	Perempuan	(S1)Sistem Informasi
0010	MA Nurul As'adiyah	88.45	Laki-laki	(D3) Manajemen Informatika

2.3.2. Transformasi Data

Tranformasi data dilakukan untuk mengubah data tujuannya agar data dapat diolah dengan menggunakan metode K-means Clustering. Adapun variabel yang digunakan pada data mahasiswa baru yaitu data asal sekolah, nilai rata- rata UAS, jenis kelamin, prodi. Untuk asal sekolah dikelompokan menjadi 3 kelompok yaitu; SMA di transformasi menjadi nilai 1, SMK di transformasi menjadi nilai 2 dan MA menjadi nilai 3. Untuk atribut nilai rata-rata UAS dikelompokan menjadi 3kelompok yaitu; nilai <80 di transformasi menjadi nilai 1, nilai >80 dan <=90 di transformasi menjadi nilai 2, dan nilai >=90 menjadi nilai 3. Atribut jenis kelamin dikelompokan menjadi 2 kelompok yaitu; kelamin laki-

laki ditransformasi menjadi nilai 1 sedangkan perempuan menjadi nilai 2. Untuk atribut prodi dikelompokan menjadi 4 kelompok yaitu; prodi (D3) Manajemen Informatika di transformasi menjadi nilai 1, (S1) Sistem Informasidi transformasi menjadi nilai 2, (S1) Sistem Komputer menjadi nilai 3 dan (S1) Teknik Informatika nilai 4.

2.4. Modeling

Pada penelitian ini menggunakan algoritma K-Means, dimana model proses mining *unsupervised-learning* yaitu klasterisasi. Klasterisasi dengan algoritma K-Means ini dipilih berdasarkan karakteristik dataset. Pada penelitian ini klasterisasi K-Means dilakukan untuk menghitung jarak *Euclidean* dari dataset yang telah dinormalisasi dan tidak dinormalisasi[7].

Algoritma klastering K-Means adalah salah satumetode umum dalam analisis data yang digunakan untuk mengelompokkan data menjadi beberapa kelompok atau klaster berdasarkan kesamaan fitur atau atribut. Tujuan dari algoritma K-Means adalah untuk meminimalkan total varian dalam klaster dengan mengalokasikan titik data ke klaster yang sesuai [8]. Secara umum, langkah-langkah algoritma K-Means adalah Inisialisasi yaitu Pilih jumlah klaster yang diinginkan (K) dan tentukan titik awal (pusat) untuk setiap klaster secara acak atau menggunakan metode lain; Assign Points to Clusters yaitu untuk setiap titik data, hitung jarak antara titik data dan pusat klaster; Update Cluster Centers yaitu hitung pusat baru untuk setiap klaster dengan menggunakan rata-rata dari semua titik data yang di-assign ke klaster tersebut; Iterasi yaitu ulangi langkah 2 dan 3 sampai kondisi berhenti terpenuhi. Kondisi berhenti bisa berupa jumlah iterasi yang telah ditentukan,atau perubahan yang kecil dalam posisi pusat klaster. Formula K-Means untuk menghitung jarak antara dua titik data (x dan y) dalam ruang fitur yangmemiliki n dimensi dapat dihitung menggunakan jarak Euclidean ditunjukkan pada persamaan 1.

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}}\,(x_i - y_i)^2 \qquad (1)$$

Pada rumus di atas d (x, y) adalah jarak antara titik data x dan y; n adalah jumlah dimensi (fitur) dari data dan x_i dan y_i adalah nilai fitur ke-i dari titik data x dan y. Selama proses iteratif, pusat klaster baru dihitung dengan mengambil rata-rata dari semua titik data dalam klaster. Jadi, jika C_k adalah klaster ke-k, maka pusat baru m_k dapat dihitung dengan formula yang di tunjukkan pada persamaan 2.

$$m_k = \frac{1}{N_k}\sum X \epsilon C_k \;\; X \qquad (2)$$

Pada rumus diatas m_k adalah pusat baru klaster ke-k. Nk adalah jumlah titik data dalam klaster ke-k dan $\sum x \epsilon c_k x$ adalah penjumlahan dari semua titik datadalam klaster ke-k. algoritma K-Means adalah algoritma iteratif, yang berarti bahwa langkah 2dan 3 akan diulang hingga konvergensi atau berhenti setelah jumlah iterasi tertentu. Setelah menghitung jarak antara dua titik maka selanjutnya adalah mengukur seberapa baik klaster yang dihasilkan oleh algoritma klastering[9] dengan memisahkan kelompok data yang berbeda dan mendekati pusat klasternya. Semakin rendah nilai Davies-Bouldin Index, semakin baik klastering yang dihasilkan. Rumus Davies-Bouldin Index untuk menghitung kualitas klastering antara dua klaster C_i dan C_j adalah yang di tunjukkan pada persamaan 3.

$$R_{ij} = \frac{s_{i+}s_j}{d_{I.j}} \qquad (3)$$

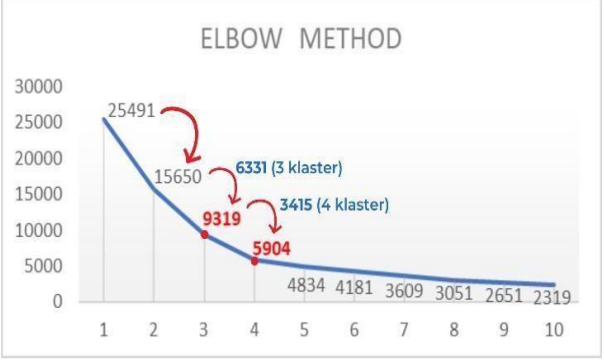
R_{ij} adalah nilai Davies-Bouldin Index antaraklaster C_i dan C_j. S_i adalah ukuran dispersi atau sebaran data dalam klaster C_i, yang dapat dihitung dengan variasi atau jarak rata-rata antara titik data dalam klaster dengan pusat klaster m_i. S_j adalah ukuran dispersi atau sebaran data dalam klaster C_j, yang dapat dihitung dengan variasi atau jarak rata-rata antara titik data dalam klaster dengan pusat klaster

m_i. Langkah-langkah umum untuk menghitung Davies-Bouldin Index adalah d_{ij} adalah jarak antara pusat klaster m_i dan m_j. Hitung pusat klaster m_i dan m_j untuk masing- masing klaster C_i dan C_j. Hitung ukuran dispersi S_i untuk klaster C_i danS_j untuk klaster C_j. Ini dapat dilakukan dengan menghitung jarak rata-rata antara titik-titik data ideal untuk dataset tersebut dalam klaster dengan pusat klaster. Hitung jarak d_{ij} antara pusat klaster m_i dan m_j. Gunakan rumus Davies-Bouldin index untuk menghitung nilai R_{ij} antara klaster C_i dan C_j. Ualngi langkah 1 hingga 4 untuk setiap pasangan klaster.

3. Hasil dan Pembahasan

3.1. Analisis Perbandingan Penentuan Jumlah klaster

Dalam penelitian ini, pertama-tama kami membandingkan hasil perhitungan metode elbow dalam menentukan jumlah klaster yang akan dilakukan proses mining dengan algoritma K-Means. Pada gambar 2 merupakan visualisasi hasil dari metode elbow.



Gambar 2. Grafik Elbow

Grafik diatas menujukkan bahwa secara kasat mata terdapat dua siku yang paling dominan, dari nilai k=1 sampai dengan nilai k=10, didapati titik siku pada nilai k = 3 dan k = 4, disini perlu dianalisis secara mendalam terkait dengan hasil tersebut, sebab apabila penentuan nilai k hanya ditentukan berdasarkan penglihatan (secara kasat mata), ini akan mempengaruhi hasil dari evaluasi klaster.

Prinsip dasar dari metode elbow yaitu apabila titik di grafik menunjukkan penurunan varians yang tajam dan menyerupai bentuk siku *(elbow)*, titik inilah yang merupakan indikasi jumlah klaster yang optimal. Berdasarkan prinsip tersebut maka dapat diketahui bahwa penurunan varians yang signifikan terdapat pada klaster 2 ke 3 dengan nilai selisihnya sebesar 6331 sedangkan selisih penurunan varians dari klaster 3 ke 4 yaitu 3415. Maka seharusnya jumlah klaster ideal adalah 3 klaster.

3.2. Perhitungan Nilai SSW

Untuk mengetahui kohesi dalam sebuah cluster ke-I adalah dengan menghitung nilai dari Sum of Square Within-cluster (SSW). Kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat cluster dari sebuah cluster yang diikuti.

Tabel 2 merupakan rekapan nilai centroid pada iterasi terakhir terhadap kedua klaster tersebut yang telah konvergen untuk selanjutnya dihitung nilai SSW.

Tabel 2. Centroid Konvergen 3 Klaster

Centroid	Asal Sekolah	Nilai UAS	Jenis Kelamin	Prodi
Titik Pusat C1	3.40741	2.03704	1.44444	3.59259
Titik Pusat C2	1.44444	3.00000	1.55556	3.62222
Titik Pusat C3	1.37500	1.72656	1.52344	3.48438

4. Kesimpulan

Hasil evaluasi khususnya evaluasi internal tanpa menggunakan *ground-truth* sebagai parameter sangat bergantung pada tahap persiapan data serta penentuan jumlah klaster diawal proses mining. Persiapan data bergantung pada karakteristik dan kondisi data. Catatan berikut, Sangat penting kiranya jika dalam menetukan jumlah klaster kemudian yang digunakan ialah metode elbow maka disarankan agar tidak menetukan secara kasat-mata atau dengan mengandalkan penilaian tanpa menghitung penurunan nilai varian apabila terdapat dua atau lebih titik siku pada grafik elbow. Setelah dilakukan pengelompokkan data penerimaan Mahasiswa baru menggunakan metode K-Means terbentuk 3 klaster yaitu, klaster 1 dengan total 27 item, klaster 2 45 item, dan klaster 3 jumlah 128 item, total item keseluruhan 200 item dataset.

Daftar Rujukan

[1] Damanik, N, dan Sigiro, M. (2021). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Pada Penerimaan Mahasiswa Baru Sebagai Metode Promosi. Jutisal Jurnal Teknik Informatika Universal, 4, 33–43.

[2] Muliono, R., & Sembiring, Z. (2019). DataMining Clustering Menggunakan Algoritma K- Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. CESS (Journal of Computer Engineering, System and Science), 4(2), 2502– 2714.

[3] Prakoso, Y. (2019). Sistem Pendukung Keputusan Penerimaan Calon Mahasiswa Baru Menggunakan Metode K-Means Clustering Di Universitas Negeri Surabaya. Jurnal Manajemen Informatika, 9(2), 79–86.

[4] Budiman, I., Prahasto, T., & Christyono, Y. (2012). Data Clustering Menggunakan Metodologi Crisp-Dm Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma. In Seminar NasionalAplikasi Teknologi Informasi.

[5] Ardiada, I. M. D., Sudarma, M., & Giriantari, D. (2019). Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour. Majalah Ilmiah Teknologi Elektro, 18(1), 55. https://doi.org/10.24843/mite.2019.v18i01.p08

[6] Apriliani, A., Zainuddin, H., Hasanuddin, Z. B., Handayani Makassar, S., & Pembangunan Nasional Veteran Jawa Timur, U. (2020). Peramalan Tren Penjualan Menu Restoran Menggunakan Metode Single Moving Average. 7(6), 1161–1168. https://doi.org/10.25126/jtiik.202072732

[7] Lestari, W., Bina, S., & Kendari, B. (2019). Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari). In SIMKOM (Vol. 4, Issue 2). http://e-jurnal.stmikbinsa.ac.id/index.php/simkom35

[8] Mukrodin1), R. T. D. S. R. E. (2022). Data MiningClustering Data Obat-Obatan Menggunakan Algoritma K-Means pada RSU An Ni'Mah Wangon. JIKA (Jurnal Informatika), 7, 165–172.

[9] Suyanto. 2017. Data Mining untuk Klasifikasi danKlasterisasi Data. Edisi Pertama. Bandung:Penerbit Informatika, 2017. p. 342. 978-602-6232-36-6