Penentuan Jurusan Siswa SMA Menggunakan Metode *K-Means++*

Pratama Agung Rizaldi[1], Maftahatul Hakimah[2], Tutuk Indriyani[3]
Institut Teknologi Adhi Tama Surabaya
e-mail: hakimah.mafta@itats.ac.id

ABSTRACT

The increasing development of technology has a major influence on the education system in schools. This is because schools need a very fast program to process student grouping data for the majors of specialization offered. Choosing the right major is very important for students because it will affect their way of learning. One solution to minimize students choosing the wrong major is to analyze the data for the major entrance test according to the ability of the students. Therefore, the purpose of this research is how to cluster students based on the criteria of the available majors. The benefits of this research are expected to help minimize errors in the selection of majors. The grouping of students was obtained by applying the k-means ++ method. This method is the development of the k-means method to overcome the centroid initialization problem. The test results show that the k-means ++ method is able to group students into specialization majors at SMA Antarctica Sidoarjo. The validation of the cluster results using SSE showed that k-means was better than k-means++ with a difference of 279.95. Meanwhile, based on the Silhoutte Coefficient, k-means ++ can increase the value of confidence in the membership of each cluster compared to the standard k-means by 10%.

Keywords*: Data mining, Student data, Clustering, K-means++, Silhouette Coefficient, Antartika Sidoarjo High School*

ABSTRAK

Perkembangan teknologi yang semakin meningkat mempunyai pengaruh besar terhadap sistem pendidikan di sekolah. Hal ini dikarenakan sekolah membutuhkan program yang sangat cepat untuk mengolah data pengelompokkan siswa terhadap jurusan peminatan yang ditawarkan. Pemilihan jurusan yang tepat merupakan hal yang sangat penting bagi siswa karena akan berpengaruh pada cara belajar mereka. Salah satu solusi untuk meminimalisir siswa salah memilih jurusan adalah dengan melakukan analisis data ujian tes masuk jurusan sesuai kemampuan para siswa. Oleh karena itu, tujuan penilitian ini adalah bagaimana mengklasterkan siswa berdasarkan kriteria jurusan yang tersedia. Manfaat penilitian ini diharapkan membantu meminimalisir kesalahan dalam pemilihan jurusan. Pengelompokan siswa diperoleh dengan menerapkan metode *k-means ++*. Metode ini merupakan pengembangan metode *k-means* untuk mengatasi permasalahan inisialisasi *centroid*. Hasil pengujian menunjukkan bahwa metode *k-means ++* mampu mengelompokkan siswa tehadap jurusan peminatan di SMA Antartika Sidoarjo. Validasi hasil klaster menggunakan SSE menunjukkan *k-means* lebih baik daripada *k-means*++ dengan selisih sebesar 279,95. Sedangkan berdasarkan Silhoutte Coefficient, *k-means ++* bisa meningkatkan nilai kepercayaan terhadap keanggotaan setiap klasternya dibandingkan *k-means* standar sebesar 10%.

Kata kunci: *Data mining; Klastering; K-means++; Silhoutte Coeffcient; SMA Antartika Sidoarjo*

PENDAHULUAN

Penentuan Jurusan bidang studi di SMA sebaiknya direncanakan dengan matang oleh pihak sekolah maupun siswa. Hal ini dikarenakan tujuan dari pengambilan jurusan tersebut agar siswa lebih fokus dan terarah dalam mengembangkan kemampuan diri terutama kemampuan akademik. Jika jurusan yang dipilih sesuai dengan kemampuan siswa, harapannya pemahaman terhadap materi yang disampaikan akan lebih maksimal sehingga nilai akademis bisa meningkat[1]. Sebaliknya, dampak kedepannya jika bidang studi jurusan tidak sesuai dengan kemampuan diri

siswa adalah menurunnya semangat belajar siswa dan bisa menimbulkan tingkat stress yang tinggi. Kondisi salah dalam memilih jurusan yang tepat bagi siswa SMA sering terjadi. Faktor penyebab umumnya dikarenakan siswa lebih cenderung mengikuti pilihan teman terdekatnya, bahkan dikarenakan budaya pemahaman Jurusan di lingkungan sekolah tersendiri. Salah satu cara untuk meminimalisir siswa salah dalam memilih jurusan adalah dengan mengenali kemampuan minat dan bakat terutama dalam bidang akademis. Penelitian ini akan membantu mengatasi permasalahan pemilihan jurusan bidang studi ini dengan mengelompokkan kemampuan akademis siswa berdasarkan nilai Mata Pelajaran. Penelitian ini dilakukan di SMA Antartika Sidoarjo. Di SMA ini akan menerapkan 4 Jurusan yaitu IPA Olahraga, IPA Bahasa, IPS Olahraga, IPS Bahasa dari sebelumnya hanya ada IPA dan IPS saja. Pemilihan jurusan sebelumnya dilakukan berdasarkan nilai ujian sekolah saat mendaftar kemudian dilihat berdasarkan nilai tertinggi dari Mata Pelajaran yang bersesuaian dengan Jurusan IPA atau IPS.

Pemilihan Jurusan di SMA Antartika ini akan diselesaikan menggunakan konsep data mining. Prinsip klastering akan diterapkan untuk mengelompokkan siswa dalam 4 Jurusan berdasarkan nilai IPA, IPS, Bahasa dan Olahraga. Metode klastering yang sering digunakan adalah *k-means*. Algoritma *k-means* sangat mudah untuk diterapkan dalam permasalahan klastering. Namun, salah satu kelemahan dari *k-means* adalah penentuan centroid awal yang dipilih secara random. Pemilihan centroid awal ini sangat berpengaruh pada hasil klaster yang diperoleh. Banyak penelitian mengusulkan perbaikan *k-means* dalam mengatasi penentuan centroid awal [2], [3],[4], [5]. Salah satu metode perbaikan *k-means* adalah *k-means* ++. Metode *k-means* ++ memilih centroid awal berdasarkan probabilitas objek yang mempunyai jarak terjauh antara centroid 1 dengan centroid yang lain [5].

Penelitian sebelumnya telah memanfaatkan algoritma klastering menggunakan k-*means* untuk pengelompokan jurusan siswa SMA 1 Karangmojo [6]. Pengelompokan dilakukan dengan mempertimbangkan nilai rapor SMP, nilai Ujian Nasional (UN), serta nilai tes penempatan *(placement test)* dengan hasil tingkat akurasi 75,52%. Penulis merekomendasikan untuk perbaikan pada k-*means*. Dengan tema penelitian yang sama, metode klastering diimplementasikan untuk menentukan jurusan pada SMA. Penulis menggunakan metode Fuzzy C-Means untuk menentukan akurasi ketepatan pemilihan jurusan oleh siswa. Perbandingan metode ini juga dilakukan terhadap metode *k-means* dengan kesimpulan metode *k-means* belum bisa mengelompokkan siswa dalam Jurusan yang tepat secara akurat[7].

Dengan latar belakang tersebut diatas, penelitian ini akan mengelompokkan siswa berdasarkan kemampuan akademisnya ke Jurusan yang sesuai menggunakan *k-means* ++. Hal ini diharapkan bisa meningkatkan motivasi belajar siswa di SMA ANTARTIKA SIDOARJO.

TINJAUAN PUSTAKA

Data Mining

Data Mining atau juga penambangan data merupakan teknik yang cukup cepat serta mudah untuk menemukan informasi, pola dan/atau relasi antar data, secara otomatis. Dengan menggabungkan empat ilmu komputer seperti pada definisi di atas, pengetahuan bisa ditemukan dalam lima proses berurutan: seleksi, prapemrosesan, transformasi, data mining, serta interpretasi/evaluasi [8].

Algoritma *k-means*

K-means merupakan metode data mining yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan melakukan pengelompokan data dengan sistem partisi [9],[10]. Langkah algoritma *k-means* [11]:

1. Pilih sembarang centroid awal c_i, i =1, …, k dimana k merupakan banyaknya *klaster* yang akan dibentuk.
2. Untuk setiap $i \in \{1, \ldots, k\}$, Pilih objek (titik data) yang lebih dekat ke centroid c_i dibandingkan dengan c_j untuk semua $j \neq i$. Kedekatan objek x dan y bisa diukur menggunakan jarak *Euclidean* yang diberikan Persamaan (1) berikut.
$$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$
3. Untuk setiap $i \in \{1, \ldots, k\}$, cari centroid c_i terbaru dari perhitungan rata-rata setiap klaster C_i yang terbentuk dengan Persamaan (2).
$$c_i = \frac{1}{|C_i|}\sum \boldsymbol{x} \quad (2)$$
dengan, $\boldsymbol{x}$ merupakan anggota klaster C_i.
4. Ulangi Langkah 2 dan 3 sampai tidak ada anggota klaster yang berpindah.

Algoritma *K-Means++*

K-Means ++ merupakan metode pengembangan dari *k-means* untuk mengatasi inisialisasi centorid. Algoritma *K-means++* merupakam algoritma yang diusulkan pertama kali oleh David Arthur dan Sergei Vassilvitskii [11]. Peningkatan kualitas klastering dari algoritma *K-means++* adalah memastikan inisialisasi centroid yang lebih cerdas [12]. Berikut adalah tahapan algoritma *k-means ++*[11]:

1. Pilih 1 centroid secara acak dari himpunan titik data.
2. Tentukan centoid baru c_i dari himpunan titik data x yang belum terpilih sebelumnya berdasarkan probabilitas terbesar jarak kuadrat titik data x ke pusat klaster yang lama. Probabilitas tersebut dihitung dengan Persamaan (3).
$$K = \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \quad (3)$$
dengan, $D(x)$ merupakan jarak yang di berikan pada Persamaan (1).
3. Ulangi langkah 2 sampai *k centroid* telah dipilih.
4. Lanjutkan proses klastering menggunakan algoritma *k-means* standar.

Algoritma secara bertahap memilih satu set K pusat *klaster* dengan mengambil sampel pusat berikutnya dari distribusi dimana setiap titik memiliki probabilitas yang sebanding dengan jarak kuadratnya ke pusat terdekat saat ini [13].

Silhoutte Coefficient

Metode ini merupakan metode evaluasi klaster yang menggabungkan metode cohessian dan separation. Cohessian diukur dengan menghitung seluruh objek yang terdapat dalam sebuah klaster dan separation diukur dengan menghitung jarak rata-rata setiap objek dalam sebuah klaster dengan klaster. Persamaan (4) berikut ini merupakan rumus *silhouette width* [14].
$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$
dengan, $s(i)$ merupakan *silhouette width*, $a(i)$ = rata-rata jarak suatu titik data x_i di klaster C_i sedangkan $b(i)$ merupakan nilai minimum dari rata-rata jarak titik data x_i di klaster C_i dengan semua objek di Klaster C_j dengan $j \neq i$. nilai $s(i)$ merupakan indikator kepercayaan untuk keanggotaan sampel di Klaster C_i [14]. Perhitungan *silhoutte coefficient* (SC) diberikan oleh Persamaan (5) sebagai berikut [15].
$$SC = \frac{\sum_{i=1}^{n} s_i}{n} \quad (5)$$
Nilai rata-rata yang dimiliki oleh *silhouette coefficient* dari masing-masing data objek dalam suatu klaster menunjukan seberapa layak data tersebut dimasukkan dalam klaster.

Sum of Squared Errors (SSE)

SSE merupakan salah satu teknik untuk validasi klastering. Kualitas klastering menggunakan SSE diukur berdasarkan jumlahan kuadrat jarak Euclidean antara setiap titik objek x_i terhadap centroid klasternya[16]. Perhitungan SSE diberikan oleh Persamaan (6) sebagai berikut.

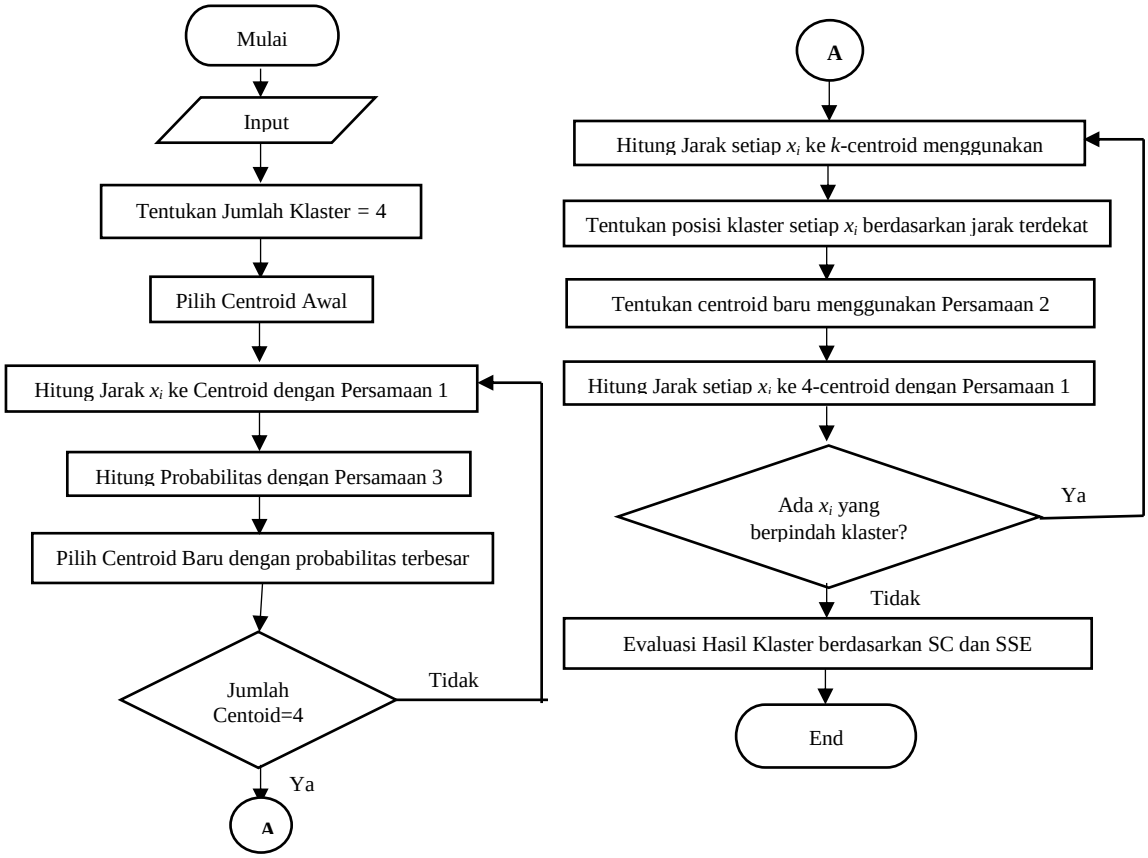
$$SSE(X, C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} \sqrt{(x_j - c_i)^2} \quad (6)$$

DATA PENELITIAN

Data yang digunakan di objek penelitian penelitian ini adalah data sekunder yang diperoleh dari bagian Kesiswaan SMA Antartika Sidoarjo. Data yang diambil Tahun 2020/2021 yakni nilai IPA, nilai IPS, nilai Bahasa, serta nilai Olahraga di salah satu kelas X SMA.

METODE

Tujuan dari penelitian ini akan di selesaikan menggunakan metode yang yang diberikan pada Gambar 1.



Gambar 1. Metode Penelitian

HASIL DAN PEMBAHASAN

Pengelompokan Jurusan di SMA Antartika dibagi menjadi 4 yaitu IPA Bahasa, IPA Olahraga, IPS Bahasa dan IPS Olahraga. Oleh karena itu jumlah klaster ditentukan sebanyak 4. Kriteria

pengelompokan didasarkan pada 4 nilai Mata Pelajaran yaitu Nilai IPA, Nilai IPS, Nilai Bahasa dan Nilai Olahraga. Sampel data penelitian disajikan pada Tabel 1 sebagai berikut.

Tabel 1. Sampel Data Penelitian

No.	Objek	IPA	IPS	BAHASA	OLAHRAGA
1	Siswa 1	60	75	85	70
2	Siswa 2	78	80	95	85
3	Siswa 3	80	90	92	85
4	Siswa 4	85	80	80	80
5	Siswa 5	75	88	78	80

Klastering siswa dengan sampel data pada Tabel 1 diproses menggunakan *k-means* ++ dan metode *k-means* secara terpisah. Langkah ini untuk mengevaluasi hasil klastering metode k-means ++ terhadap *k-means* standar. Tabel 2 berikut ini merupakan evaluasi hasil klastering menggunakan metode *k-means* ++ dan *k-means*.

Tabel 2. Evaluasi Hasil Klaster

Metode	SSE	Silhoutte Coefficient
k-means ++	1800,173	0,187
k-means	1520,223	0,170

Tabel 2 menunjukkan perbandingan hasil klastering menggunakan *k-means* ++ dan *k-means*. Berdasarkan pengukuran SSE, selisih kuadrat antara anggota klaster dengan centroidnya lebih baik hasil klastering dari *k-means* daripada *k-means* ++. Selisih SSE antara kedua metode tersebut sebesar 279,95. Sementara itu, besarnya nilai SC pada *k-means* ++ memberikan hasil yang lebih baik dengan peningkatan sebesar 10% dari *k-means* standar. Nilai SC menunjukkan nilai kepercayaan terhadap keanggotaan di setiap klaster. Hasil proses klastering ini tidak memberikan panduan untuk label dari setiap klaster. Sehingga, validasi hasil klastering pada penelitian ini dilakukan dengan melakukan konfirmasi ke pihak SMA Antartika. Tabel 2 merupakan hasil klastering *k-means* ++ disertai labelnya.

Tabel 2. Hasil Klasterisasi menggunakan *k*-means ++

Klaster	Label Klaster	Jumlah Anggota
1	IPA Olahraga	13
2	IPS BAHASA	4
3	IPS Olahraga	1
4	IPA BAHASA	2

KESIMPULAN

Penentuan Jurusan Bidang Studi untuk SMA telah diselesaikan menggunakan Metode Klastering *k-means* ++. Proses *k-means* ++ memberikan panduan untuk mendapatkan centroid awal. Berdasarkan hasil pengujian SSE, *k-means* ++ mencapai nilai 1800,173 yang menunjukkan hasilnya tidak lebih baik daripada *k-means* standar dengan nilai SSE-nya sebesar 1520,223. Namun pada pengukuran nilai Silhoutte Coefficient, *k-means* ++ mempu meningkatkan nilai keanggotaan setiap titik data terhadap klaster yang diperoleh sebesar 10% dari i-*means*. Hasil

klastering ini bisa menjadi rekomendasi awal untuk SMA Antartika Sidoarjo dalam membantu siswanya menentukan Jurusan Bidang Studi kedepannya.

DAFTAR PUSTAKA

[1] Y. S. Nugroho, "Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 1, no. 1, p. 1, 2015, doi: 10.23917/khif.v1i1.1175.

[2] R. Nainggolan and E. Purba, "Perbaikan Performa Cluster *K-means* Menggunakan Sum Squared Error (Sse) Pada Analisis Online Customer Review Terhadap Produk Toko Online," *J. TIMES*, vol. VIII, no. 2, pp. 1–8, 2019.

[3] A. Aprilia, W. M. Rahmawati, and M. Hakimah, "Penentuan Kategori Status Gizi Balita Menggunakan Penggabungan Metode Klasterisasi Agglomerative Dan *K-means*," *Semin. Nas. Sains dan Teknol. Terap. VII - Inst. Teknol. Adhi Tama Surabaya*, pp. 595–600, 2019.

[4] H. Zhao, "Research on Improvement and Parallelization of *K-means* Clustering Algorithm," in *IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, 2021, pp. 57–61.

[5] N. Daoudi, S., Anouar Zouaoui, C. M., El-Mezouar, M. C., & Taleb, "Parallelization of the *K-means* ++ Clustering Algorithm," vol. 26, no. 1, 2021, [Online]. Available: https://web.s.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=16331311&AN=149636209&h=osAPpJOx8kOx7fkxakuC2PTz3GnvsfcGZnYp8w4tTQJFWa%2BMikzsrpvvtFEaic7LwQulnr0Qc%2BGrG8YPD92Y8g%3D%3D&crl=c&resultNs=AdminWebAuth&resultLoca.

[6] M. E. Sulistiyani, B. Soedijono, and S. A. Syahdan, "Sistem Penentuan Jurusan Sekolah Menengah Atas Negeri 1 Karangmojo," *Semnasteknomedia Online*, vol. 3, no. 1, pp. 2-2–247, 2015, [Online]. Available: http://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/819/785.

[7] H. K. Candra, M. Bahit, and B. Sabella, "Penerapan Metode Klustering Fuzzy C-Means Untuk Penentuan Peminatan Pemilihan Jurusan Pada Sekolah Menengah Tingkat Atas," *POSITIF J. Sist. dan Teknol. Inf.*, vol. 7, no. 2, pp. 108–119, 2021, doi: 10.31961/positif.v7i2.1106.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," vol. 17, no. 3, 1996, doi: https://doi.org/10.1609/aimag.v17i3.1230.

[9] F. Nasari and S. Darma, "Penerapan *k-means* clustering pada data penerimaan mahasiswa baru (studi kasus: universitas potensi utama)," *Semnasteknomedia Online*, vol. 3, no. 1, 2013, [Online]. Available: https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/837/801.

[10] et al. Muhima, Rani Rotul, *KUPAS TUNTAS ALGORITMA CLUSTERING: KONSEP, PERHITUNGAN MANUAL, DAN PROGRAM*. Andi, 2022.

[11] D. Arthur and S. Vassilvitskii, "*k-means* ++: The advantages of careful seeding.," 2006.

[12] C. A. Sri Fastaf and Y. Yamasari, "Analisa Pemetaan Kriminalitas Kabupaten Bangkalan Menggunakan Metode *K-means* dan *K-means* ++," *J. Informatics Comput. Sci.*, vol. 3, no. 04, pp. 534–546, 2022, doi: 10.26740/jinacs.v3n04.p534-546.

[13] S. Lattanzi and C. Sohler, "A better *k-means* ++ algorithm via local search," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 6521–6530, 2019.

[14] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus External cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27--34, 2011, [Online]. Available: http://w.naun.org/multimedia/UPress/cc/20-463.pdf.

[15] H. Řezanková, "Different approaches to the silhouette coefficient calculation in cluster evaluation," *21st Int. Sci. Conf. AMSE*, no. September, pp. 1–10, 2018.

[16] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, "The Clustering Validity with Silhouette and Sum of Squared Errors," pp. 44–51, 2015, doi: 10.12792/iciae2015.012.