

Analisis Data Akademis dengan Menerapkan Algoritme K-Means dan K- Means++

Cepi Ramdani^{1*}, Naura Safadila²

^{1,2}*Sistem Informasi, Institut Teknologi Telkom Purwokerto, Indonesia*

*Corresponding Author: cepi@ittelkom-pwt.ac.id

Abstract

The Study Program is the implementing unit of education in Higher Education. Study Program has the mandate to carry out quality, standardized educational activities, and always make improvements. In an effort to improve the quality of education can be done by evaluating the results of educational activities that have been implemented. Student academic data can be used in this regard. Every semester, student academic data is collected and growing. Processing and analysis of this data can produce useful information for Study Programs in efforts to improve the quality of education. In this study, data processing and analysis was carried out using the K-Means clustering method. The data used is data on student learning outcomes for the 2017 – 2019 class of 332 students, in the form of Semester Grade Point Average (GPA) data and Course Credits from semester 1 to semester 5. The data is clustered into 3 clusters which represent academic performance categories low, medium, and high students. Clustering is done by utilizing the sklearn library with the Python programming language. The K-Means++ algorithm is used in the process of initializing the initial cluster center values (centroids), this is to reduce the instability that occurs in K-Means so as to provide stable and better clustering results. The results of the clustering found that most students fall into the category of high academic performance.

Keywords: clustering, course credits, k-means, semester gpa, study program

Abstrak

Program Studi (Prodi) adalah unit pelaksana pendidikan di Perguruan Tinggi. Prodi mengemban amanat untuk menyelenggarakan kegiatan pendidikan yang bermutu, berstandar, dan selalu melakukan perbaikan. Dalam upaya perbaikan kualitas pendidikan dapat dilakukan dengan melakukan evaluasi terhadap hasil kegiatan pendidikan yang telah dilaksanakan. Data akademis mahasiswa dapat digunakan dalam hal ini. Setiap semester data akademis mahasiswa dikumpulkan dan semakin bertambah. Pengolahan dan analisis terhadap data tersebut dapat menghasilkan informasi yang bermanfaat bagi Program Studi dalam upaya peningkatan kualitas pendidikan. Pada penelitian ini dilakukan pengolahan dan analisis data dengan metode klasterisasi *K-Means*. Data yang digunakan adalah data hasil belajar mahasiswa Angkatan 2017 – 2019 sebanyak 332 mahasiswa, berupa data Indeks Prestasi Semester (IPS) dan Beban Satuan Kredit Semester (SKS) dari semester 1 hingga semester 5. Data tersebut diklasterisasi kedalam 3 klaster yang merepresentasi kategori kinerja akademis mahasiswa rendah, sedang, dan tinggi. Klasterisasi dilakukan dengan memanfaatkan pustaka *sklearn* dengan bahasa pemrograman *Python*. Algoritme *K-Means++* digunakan dalam proses inialisasi nilai pusat klaster (*centroids*) awal, hal ini untuk mengurangi ketidakstabilan yang terjadi pada *K-Means* sehingga memberikan hasil klasterisasi yang stabil serta lebih baik. Hasil klasterisasi didapati bahwa sebagian besar mahasiswa masuk kedalam kategori berkinerja akademik tinggi.

Kata Kunci: beban sks, indeks prestasi semester, klasterisasi, k-means, program studi

I. INTRODUCTION

Program Studi (Prodi) sebagai unit pelaksana pendidikan di Perguruan Tinggi, memiliki tugas untuk menyelenggarakan pendidikan yang bermutu, berstandar dan dituntut untuk terus menerus melakukan

perbaikan. Dalam upaya perbaikan kualitas pendidikan secara berkelanjutan, dapat dilakukan dengan melakukan evaluasi terhadap hasil kegiatan pendidikan yang telah dilaksanakan. Produk dari kegiatan pembelajaran salah satunya adalah hasil belajar mahasiswa. Data hasil belajar mahasiswa setiap semester dikumpulkan dan jumlahnya terus bertambah. Pengolahan terhadap data tersebut dapat menghasilkan informasi yang bernilai untuk menjadi acuan dalam perbaikan kualitas pendidikan.

Pada penelitian ini dilakukan klusterisasi data akademis mahasiswa program studi S1 Sistem Informasi Institut Teknologi Telkom Purwokerto dengan menggunakan algoritme *K-Means*. Algoritme *K-Means* merupakan salah satu algoritme klusterisasi yang sederhana, dapat digunakan pada data yang berukuran besar, dan sangat efisien[1]. Salah satu kelemahan dari algoritme *K-Means* adalah penentuan nilai pusat kluster awal yang dilakukan secara acak, nilai pusat kluster awal dapat mempengaruhi hasil klusterisasi. Untuk menutupi kelemahan tersebut maka diterapkan algoritme *K-Means++*, sehingga mengurangi ketidakstabilan yang terjadi pada *K-Means* dan penentuan nilai pusat kluster awal lebih terarah dan cerdas[2][3]. Adapun data akademis mahasiswa yang digunakan adalah data hasil belajar mahasiswa 3 angkatan yakni 2017, 2018, dan 2019 sebanyak 332 mahasiswa. Data tersebut berupa nilai Indeks Prestasi Semester (IPS) dan Beban Satuan Kredit Semester (SKS) semester 1 - 5 sehingga terdapat 10 atribut.

Proses klusterisasi memanfaatkan pustaka *sklearn* dengan bahasa pemrograman *Python*, pengaturan parameter klusterisasi *K-Means* diantaranya jumlah kluster (k) = 3, toleransi kesalahan = $1e^{-7}$, maksimal iterasi = 300, dan $init = "k-means++"$. Pemilihan jumlah kluster sebanyak 3, hal ini merepresentasikan kategori kinerja akademis mahasiswa rendah, sedang, dan tinggi. Untuk mengukur kualitas hasil klusterisasi digunakan Indeks *Davies-Bouldin*.

II. LITERATURE REVIEW

Penelitian yang dilakukan ini mengacu pada beberapa hasil penelitian sebelumnya diantaranya, penelitian[4] melakukan klusterisasi data hasil belajar dengan menggunakan algoritme *K-Means*. Data hasil belajar yang diklusterisasi adalah data Indeks Prestasi Kumulatif (IPK) mahasiswa Angkatan 2017-2018. Jumlah kluster yang dihasilkan sebanyak 3 kluster. Disamping itu dilakukan perbandingan hasil klusterisasi data IPK dengan klusterisasi data nilai setiap mata kuliah, disimpulkan hasil klusterisasi data IPK lebih baik daripada hasil klusterisasi data nilai setiap mata kuliah.

Penelitian[5] melakukan klusterisasi gaya belajar mahasiswa dengan aplikasi WEKA. Algoritme klusterisasi yang digunakan adalah *K-Means*. Data yang digunakan adalah data gaya belajar mahasiswa yang didapatkan dari *log e-learning* dan data nilai mahasiswa. Terdapat 7 gaya belajar yang diuraikan yakni, *Physical*, *Verbal*, *Solitary*, *Aural*, *Visual*, *Social* dan *Logical*. Kesimpulan dari penelitian yang dilakukan yakni mahasiswa yang rajin mengakses video dan modul pembelajaran cenderung memiliki nilai yang baik.

Penelitian[6] melakukan klusterisasi mahasiswa berdasarkan data akademis untuk penentuan kelompok kelas mata kuliah peminatan. Penelitian tersebut dilatarbelakangi karena jumlah lulusan yang masih kurang sehingga diperlukan adanya strategi dalam penentuan mata kuliah peminatan. Algoritme klusterisasi yang digunakan adalah *K-Means* dan *K-Means KNN*. Jumlah kluster yang dihasilkan sebanyak 2 kluster. Kesimpulan dari penelitian yang dilakukan yakni Algoritme *K-Means KNN* memberikan hasil yang lebih baik daripada *K-Means*.

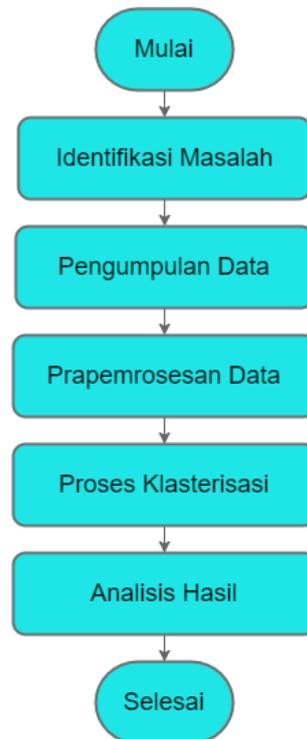
Penelitian[7] melakukan klusterisasi mahasiswa kemudian memprediksi mahasiswa yang mungkin berprestasi. Algoritme yang digunakan dalam proses klusterisasi yakni *K-Means* sedang algoritme yang digunakan untuk melakukan prediksi adalah *C4.5*. Data yang digunakan sebanyak 100 data mahasiswa dengan 6 atribut yakni penghasilan orang tua, jarak tempat tinggal dengan kampus, kelengkapan perangkat belajar, sosialisasi mahasiswa dengan lingkungan kampus, sosialisasi mahasiswa dengan lingkungan tempat tinggalnya, dan Indeks Prestasi Kumulatif (IPK). Hasil dari klusterisasi dan prediksi menunjukkan bahwa 27% mahasiswa akan berprestasi dan sisanya yakni 73% mahasiswa diprediksi tidak akan berprestasi.

Penelitian[8] melakukan komparasi algoritme klusterisasi *K-Means* dengan *K-Medoids*. Data yang digunakan adalah dokumen skripsi. Tujuan klusterisasi adalah mengkategorikan dokumen skripsi

berdasarkan topik penelitian. Hasil dari penelitian yang dilakukan yakni berdasarkan hasil evaluasi dengan Indeks *Davies-Bouldin* didapati Algoritme *K-Medoids* lebih baik daripada Algoritme *K-Means*, namun dari sisi waktu pemrosesan, Algoritme *K-Means* lebih cepat daripada Algoritme *K-Medoids*.

III. RESEARCH METHOD

Alur penelitian pada penelitian ini dapat dilihat pada gambar 1, dimulai dengan proses identifikasi masalah, pengumpulan data, prapemrosesan data, proses klusterisasi, dan analisis hasil.



Gambar 1. Alur Penelitian

1. Identifikasi Masalah

Pada tahap ini, dilakukan pencarian permasalahan penelitian dengan cara observasi dan mengkaji berbagai literatur ilmiah diantaranya jurnal, buku, dan forum diskusi. Topik yang diangkat mengenai pemanfaatan data akademis mahasiswa. Data akademis mahasiswa masih jarang digunakan dalam upaya peningkatan kualitas proses pendidikan yang dilaksanakan oleh Prodi.

2. Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data penelitian. Data yang dikumpulkan merupakan data sekunder berupa nilai Indeks Prestasi Semester (IPS) dan beban Satuan Kredit Semester (SKS) setiap semesternya. Dimulai dari semester 1 hingga semester 5. Data tersebut dikumpulkan secara manual dari Sistem Informasi Akademis. Jumlah data yang dikumpulkan sebanyak 332 data mahasiswa yang terdiri dari 10 atribut yakni SKS 1, SKS 2, SKS 3, SKS 4, SKS 5, IP 1, IP 2, IP 3, IP 4, dan IP 5. Tabel 1, menunjukkan sebagian data akademis mahasiswa yang digunakan dalam penelitian ini.

TABEL 1
DATA AKADEMIS MAHASISWA

SKS 1	SKS 2	SKS 3	SKS 4	SKS 5	IP 1	IP 2	IP 3	IP 4	IP 5
20	20	17	20	17	2.45	2.4	2.82	2.25	2.50
20	20	24	23	23	3.7	3.75	3.67	3.72	3.74
20	20	24	23	23	3.5	3.5	3.44	3.00	3.26
20	20	24	23	20	3.05	3	3.38	2.46	3.53
20	20	24	23	23	3.55	3.5	3.35	3.22	3.80
20	20	21	20	21	2.6	2.9	3.05	2.50	3.19
20	20	24	23	23	3.75	3.4	3.56	3.76	3.50
20	20	17	20	20	2.4	2.5	2.91	2.35	2.83
20	20	21	20	20	2.85	2.55	2.88	2.43	3.08
20	20	24	23	23	3.75	3.85	3.92	3.80	3.61

3. Prapemrosesan Data

Pada tahap ini data yang telah dikumpulkan dilakukan pembersihan, data yang tidak lengkap dihapus, sehingga hanya terdapat data yang lengkap saja.

4. Proses Klasterisasi

Pada tahap ini dilakukan proses klasterisasi data akademis mahasiswa dengan menerapkan algoritme *K-Mean* dan *K-Means++*. Penerapan *K-Means* memanfaatkan pustaka *sklearn* dengan bahasa pemrograman *Python* dan lingkungan pemrograman *Jupyter Lab*. Pengaturan parameter klasterisasi *K-Mean* sebagai berikut, jumlah klaster (k) = 3, jumlah maksimal iterasi = 300, toleransi kesalahan = $1e^{-7}$, dan $init = "k-means++"$. Pemilihan jumlah klaster sebanyak 3, hal ini bermaksud mengkategorikan kinerja akademis mahasiswa kedalam kategori rendah, sedang, dan tinggi. Sementara penerapan algoritme *K-Means++* digunakan untuk penetapan nilai pusat klaster awal agar lebih terarah dan cerdas.

5. Analisis Hasil

Pada tahap ini dilakukan pengukuran kualitas hasil klasterisasi dan analisis hasil klasterisasi. Pengukuran kualitas hasil klasterisasi dengan menggunakan Indeks *Davies-Bouldin*. Indeks *Davies-Bouldin* mengukur kedekatan jarak antar klaster dan mengukur kedekatan setiap anggota klaster pada pusat klaster yang diikuti[9]. Semakin kecil nilai indeks ini, mendekati 0 (nol) maka semakin baik hasil klasterisasi[10].

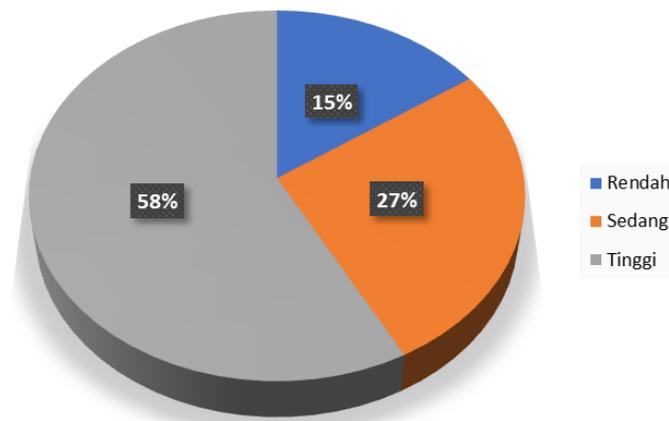
IV. RESULTS AND DISCUSSION

Hasil klasterisasi memiliki nilai Indeks *Davies-Bouldin* sebesar 0.926, ini merupakan nilai indeks terkecil yang didapatkan dari proses perulangan klasterisasi. Nilai Indeks *Davies-Bouldin* mendekati 0, hal ini dapat diasumsikan jika hasil klaster cukup baik. Sementara untuk nilai akhir pusat klaster yang dihasilkan dapat dilihat pada tabel 2. Data dibagi kedalam 3 klaster, dapat diasumsikan jika klaster 0 merupakan kategori mahasiswa dengan kinerja akademis rendah, klaster 1 merupakan kategori mahasiswa dengan kinerja akademik sedang, dan klaster 2 adalah mahasiswa dengan kinerja akademik tinggi.

TABEL 2
PUSAT KLASTER AKHIR

Pusat Klaster Akhir		
0	1	2
20	20	20
20	20	20
19.92157	20.35955	23.29167
20.05882	22.46067	22.64583
20	23.47191	22.40625
2.798039	3.056742	3.403646
2.677451	3.091011	3.503125
2.933446	3.413284	3.587108
2.78123	3.571221	3.668567
2.863449	3.580441	3.710349

Berdasarkan pusat klaster akhir tersebut, didapati jumlah mahasiswa yang masuk kedalam kategori kinerja akademis rendah (klaster 0) sebanyak 51 mahasiswa atau sebesar 15 persen, kategori kinerja akademis sedang (klaster 1) sebanyak 89 mahasiswa atau sebesar 27 persen, dan kategori kinerja akademis tinggi (klaster 2) sebanyak 192 atau sebesar 58 persen. Hasil klasterisasi dapat dilihat pada gambar 2.



Gambar 2. Hasil Klasterisasi

Tabel 3, menunjukkan sebagian pemetaan hasil klasterisasi terhadap data penelitian. Pada kolom ujung kanan menunjukkan keterangan mahasiswa yang masuk kategori kinerja akademis rendah, sedang, atau tinggi.

TABEL 3
PEMETAAN HASIL KLASTERISASI

SKS1	SKS2	SKS3	SKS4	SKS5	IP1	IP2	IP3	IP4	IP5	Kluster	Keterangan
20	20	24	23	23	3.75	3.4	3.56	3.76	3.50	2	Tinggi
20	20	17	20	20	2.4	2.5	2.91	2.35	2.83	0	Rendah
20	20	21	20	20	2.85	2.55	2.88	2.43	3.08	0	Rendah
20	20	24	23	23	3.75	3.85	3.92	3.80	3.61	2	Tinggi
20	20	21	23	23	3.2	3.05	3.14	2.98	3.24	1	Sedang
20	20	21	23	24	2.95	2.9	3.26	3.04	3.33	1	Sedang

20	20	24	23	20	3.75	3.35	3.71	3.30	3.63	2	Tinggi
20	20	21	20	20	2.95	2.35	2.36	1.43	0.50	0	Rendah
20	20	21	23	21	3.2	2.75	3.26	2.63	2.95	1	Sedang
20	20	24	23	23	3.75	3.6	3.71	3.67	3.80	2	Tinggi

V. Conclusion

Hasil klasterisasi data akademis mahasiswa Angkatan 2017-2019 sebanyak 332 mahasiswa didapati sebanyak 58 persen mahasiswa masuk kedalam kategori memiliki kinerja akademis tinggi. Hal tersebut menunjukkan kualitas pendidikan yang dilakukan selama ini sudah cukup baik namun tetap perlu perbaikan dan peningkatan karena masih terdapat mahasiswa yang memiliki kinerja akademis rendah dengan persentase sebesar 15 persen.

REFERENCES

- [1] G. Abdurrahman, "Clustering Data Ujian Tengah Semester (UTS) Data Mining," *J. Sist. Teknol. Inf. Indones.*, vol. 1, no. 2, pp. 71–79, 2016.
- [2] A. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms," in *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, 2017, pp. 1–6, doi: 10.1109/CICT.2017.7977272.
- [3] P. A. Rizaldi, M. Hakimah, and T. Indriyani, "Penentuan Jurusan Siswa SMA Menggunakan Metode K-Means ++," *Semin. Nas. Sains dan Teknol. Terap. X*, pp. 1–7, 2022.
- [4] B. J. M. Putra and D. A. F. Yuniarti, "Analisis Hasil Belajar Mahasiswa Dengan Clustering Menggunakan Metode K-Means," *Poros Tek.*, vol. 12, no. 2, pp. 49–58, 2020, [Online]. Available: <https://ejournal.poliban.ac.id/index.php/porosteknik/article/view/927>.
- [5] B. J. M. Putra and D. A. F. Yuniarti, "Analisis Gaya Belajar terhadap Nilai Mahasiswa dengan Menggunakan Metode k-Means," *Techno.Com*, vol. 21, no. 2, pp. 343–354, 2022, doi: 10.33633/tc.v21i2.5837.
- [6] S. Rustam and H. Annur, "Akademik Data Mining (ADM) K-Means Dan K-Means K-NN Untuk Mengelompokan Kelas Mata Kuliah Kosentrasi Mahasiswa Semester Akhir," *Ilk. J. Ilm.*, vol. 11, no. 3, pp. 260–268, 2019, doi: 10.33096/ilkom.v11i3.487.260-268.
- [7] A. Ramadhanu, S. Defit, and S. W. Kareem, "Hybrid Data Mining with the Combination of K-Means Algorithm and C4.5 to Predict Student Achievement," *Int. J. Artif. Intell. Res.*, vol. 5, no. 2, pp. 180–189, 2021, doi: 10.29099/ijair.v6i1.225.
- [8] S. Ramadhani, D. Azzahra, and T. Z., "Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 13, no. 1, pp. 24–33, 2022, doi: 10.31849/digitalzone.v13i1.9292.
- [9] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012128.
- [10] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, pp. 95–100, 2021, [Online]. Available: <https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/10428/4839>.