

## Klasterisasi Mahasiswa Berdasarkan Performa Akademik Menggunakan Algoritma *K-Means* pada *RapidMiner*: Studi Kasus dengan *Dataset Student Academic Performance*

<sup>1</sup>Athaya Rima Hariyanto, <sup>2</sup>Berliani Salsabiilah, <sup>3</sup>Siti Khodijah, <sup>4</sup>Winona Septi Aulia, <sup>5</sup>Maulana Fansyuri

<sup>1</sup>Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

<sup>2</sup>Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

<sup>3</sup>Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

<sup>4</sup>Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

<sup>5</sup>Sistem Informasi, Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

<sup>1</sup>[athaya.rima@gmail.com](mailto:athaya.rima@gmail.com), <sup>2</sup>[berlianisalsabiilah04@gmail.com](mailto:berlianisalsabiilah04@gmail.com), <sup>3</sup>[skhodijah369@gmail.com](mailto:skhodijah369@gmail.com),

<sup>4</sup>[winonaseptiau@gmail.com](mailto:winonaseptiau@gmail.com), <sup>5</sup>[dosen02359@unpam.ac.id](mailto:dosen02359@unpam.ac.id)

### Abstract

*One of the primary indicators used to convey the effectiveness of the learning process and to create more efficient teaching methods is student academic Performance. This study uses the RapidMiner application to use the K-Means Clustering method in order to group students according to their academic Performance. The synthetic data, which includes details about student involvement, attendance rates, and academic grades, is taken from the Kaggle platform. This study was carried out in a number of steps, including Cluster quality assessment, attribute selection, algorithm application, and data pre-processing. Based on the results, three student groups with characteristics of high, medium, and low academic Performance were examined. The Davies-Bouldin Index examination indicated that the Clustering results were optimal. These findings are expected to serve as a guide for educational institutions to develop more appropriate and successful teaching strategies.*

*Keywords: K-Means, Clustering, Academic Performance, RapidMiner, Data Mining.*

### Abstrak

Prestasi akademik siswa sangat penting untuk mengevaluasi keberhasilan pembelajaran dan mengembangkan metode pendidikan yang lebih baik. Algoritma *K-Means Clustering* digunakan dalam aplikasi *RapidMiner* untuk mengelompokkan siswa berdasarkan prestasi akademik mereka. Data sintetis yang digunakan dari platform *Kaggle* mencakup informasi tentang nilai akademik, tingkat kehadiran, dan partisipasi siswa. Penelitian ini melakukan pra-pemrosesan data, pemilihan atribut, penerapan algoritma, dan evaluasi kualitas kluster. Hasil menunjukkan bahwa tiga kelompok siswa menunjukkan prestasi akademik yang rendah, sedang, dan baik. Evaluasi dengan indeks *Davies-Bouldin* digunakan untuk menentukan hasil klasterisasi yang ideal. Temuan ini diharapkan dapat digunakan oleh institusi pendidikan untuk mengembangkan metode pembelajaran yang lebih sesuai dan efisien.

**Kata Kunci:** *K-Means*, Pengelompokan, Performa Akademik, *RapidMiner*, Penambangan Data

## A. PENDAHULUAN

Pendidikan menentukan kualitas sumber daya manusia. Akademik siswa menjadi salah satu indikator utama dalam menilai keberhasilan pembelajaran di perguruan tinggi. Namun, banyak lembaga pendidikan masih kesulitan menemukan dan memahami karakteristik siswa berdasarkan data akademik mereka.

Salah satu masalah utama yang dihadapi oleh mitra institusi pendidikan adalah tidak dapat memanfaatkan data akademik dengan benar saat merancang strategi pembelajaran yang sesuai dengan kebutuhan siswa. Metode yang digunakan saat ini seringkali seragam dan mengabaikan keberagaman capaian belajar siswa. Akibatnya, proses pembelajaran tidak efektif, terutama untuk siswa yang memerlukan pendekatan khusus.

Di era digital saat ini, banyak data akademik dibuat, mulai dari nilai mata kuliah, kehadiran, hingga partisipasi dalam kegiatan kelas. Potensi besar ini belum dimanfaatkan sepenuhnya karena keterbatasan analisis data di institusi mitra. Oleh karena itu, dibutuhkan teknik yang dapat mengekstraksi informasi penting dari kumpulan data.

Metode data mining, khususnya metode klasterisasi, adalah solusi. Sebuah metode yang dikenal sebagai klasterisasi memungkinkan pengelompokan data berdasarkan kesamaan karakteristik tertentu. Akibatnya, siswa dapat dikelompokkan berdasarkan pola belajar mereka secara objektif. Hal ini akan membantu organisasi menentukan strategi pembelajaran yang lebih khusus untuk masing-masing kelompok.

Algoritma *K-Means* dapat digunakan untuk menyusun klaster dengan cepat dan akurat, dan merupakan salah satu teknik klasterisasi yang efektif. Algoritma ini dapat mengelompokkan siswa ke dalam kategori performa tinggi, sedang, dan rendah berdasarkan nilai akademik, kehadiran, dan partisipasi mereka di kelas. Kelebihan lain dari algoritma *K-Means* adalah bahwa itu dapat menangani *Dataset* berukuran sedang hingga besar, yang membuatnya cocok untuk analisis data pendidikan. Selain itu, algoritma ini mudah digunakan dan dapat diintegrasikan dengan banyak platform analisis data, seperti *RapidMiner* atau Python. Karena hasilnya mudah dipahami, algoritma ini dapat menjadi alat bantu dalam pengambilan keputusan yang berbasis data, khususnya dalam meningkatkan kualitas pembelajaran di sekolah.

Perangkat lunak *RapidMiner*, yang memiliki antarmuka berbasis drag-and-drop, digunakan untuk mendukung implementasi algoritma tersebut. Hal ini memungkinkan pengguna, termasuk pihak institusi mitra, melakukan proses analisis data tanpa harus memiliki keahlian pemrograman yang lebih lanjut. Penelitian ini mencakup langkah-langkah seperti pemilihan atribut, *preprocessing*, klasterisasi, dan evaluasi hasil.

Penelitian ini menggunakan data sintetis *Student Academic Performance* yang diperoleh dari platform *Kaggle*. Data ini mencakup informasi tentang nilai mata pelajaran, kehadiran, dan tingkat partisipasi siswa. Untuk tetap objektif, analisis tidak memasukkan atribut demografis seperti usia atau jenis kelamin.

Penelitian ini diharapkan dapat membantu institusi mitra memahami kondisi akademik siswa. Hasilnya akan mencakup model performa siswa dalam klasterisasi kelompok serta rekomendasi untuk strategi pembelajaran yang dapat digunakan berdasarkan temuan ini. Selain itu, tujuan dari upaya ini adalah untuk melibatkan mitra secara langsung dalam proses melakukan analisis, dari pemilihan data hingga interpretasi hasilnya. Ini akan memungkinkan adopsi dan penerapan rencana yang ditawarkan secara berkelanjutan.

## B. TEORI DAN METODELOGI

Dalam penelitian kuantitatif ini, metode eksploratif berbasis data sekunder digunakan. Tujuan utama penelitian ini adalah untuk menggunakan algoritma *Clustering K-Means*, yang ditemukan dalam perangkat lunak *RapidMiner*, untuk mengelompokkan siswa berdasarkan prestasi akademik mereka.

Penelitian ini menggunakan data publik *Student Academic Performance*, yang diperoleh melalui platform *Kaggle*. *Dataset* sintetis ini dibuat untuk memenuhi kebutuhan pembelajaran dan mempelajari teknik data mining, terutama untuk menganalisis prestasi akademik siswa di sekolah menengah. *Dataset* terdiri dari seratus entri data siswa yang memiliki nilai matematika, IPA, Inggris, partisipasi kelas, dan persentase kehadiran. Selain itu, atribut tambahan seperti usia, jenis kelamin, dan tingkat kelas dapat digunakan; namun, untuk menjaga fokus pada aspek akademik dan menghindari bias, atribut ini tidak digunakan dalam analisis.

Algoritma *K-Means* banyak digunakan dalam pengelompokan data dan merupakan salah satu metode belajar tanpa pengawasan yang efektif, yang membuatnya menjadi pilihan yang tepat. Algoritma ini membagi data ke dalam berbagai klaster berdasarkan kedekatan nilai antar atributnya. Algoritma ini memiliki beberapa fungsi. Ini termasuk menentukan jumlah klaster (*k*), mengaktifkan *centroid* secara acak, mengatur data berdasarkan jarak ke *centroid* terdekat, menghitung ulang posisi *centroid*, dan mengulang proses hingga *centroid* mencapai kondisi konvergen. Dalam penelitian ini, ada tiga klaster yang ditentukan, dengan asumsi bahwa siswa dapat dikategorikan ke dalam kelompok tinggi, sedang, dan rendah berdasarkan kinerja akademik mereka.

Sebelum algoritma digunakan, data diproses untuk memenuhi kebutuhan analisis. *RapidMiner* memulai proses dengan menggunakan operator *Read Excel* untuk

membaca file. Selanjutnya, operator Select Attributes digunakan untuk memilih atribut yang terkait langsung dengan prestasi akademik. Ini termasuk skor Matematika, Skor Sains, Skor Inggris, Rata-rata Kehadiran, dan Kehadiran. Operator *K-Means* digunakan untuk menerapkan algoritma *K-Means* setelah data selesai, dengan parameter jumlah klaster (*k*) diatur ke angka tiga.



Gambar 1 Proses Klasterisasi Mahasiswa Menggunakan Algoritma *K-Means* di *RapidMiner*

Setelah proses klasterisasi selesai, evaluasi kualitas hasil klasterisasi dilakukan dengan menggunakan operator *Cluster Distance Performance*. Dua metrik utama yang digunakan dalam evaluasi ini adalah *Average Within Centroid Distance*, yang mengukur homogenitas anggota dalam klaster, dan *Davies-Bouldin Index (DBI)*, yang menunjukkan tingkat pemisahan antar klaster. Nilai DBI yang lebih rendah menunjukkan hasil klasterisasi yang lebih baik dan terpisah secara jelas.



Gambar 2 Proses Klasterisasi Disertai Evaluasi Performa Klaster di *RapidMiner*

Seluruh proses analisis dilakukan menggunakan *RapidMiner Studio*. *RapidMiner Studio* adalah perangkat lunak visual berbasis drag-and-drop yang digunakan untuk menyelesaikan analisis secara keseluruhan. Perangkat lunak ini memungkinkan pemodelan data tanpa perlu menulis kode secara manual, dan memiliki fitur evaluasi visual yang interaktif dan menyeluruh.

Penelitian ini diharapkan dapat menghasilkan klasterisasi siswa berdasarkan prestasi akademik mereka secara objektif. Hasil klasterisasi ini juga diharapkan dapat memberikan wawasan yang dapat digunakan untuk membantu proses pengambilan keputusan tentang strategi pembelajaran berbasis data.

## C. HASIL DAN PEMBAHASAN

### Hasil Proses Klasterisasi

Data dikelompokkan menggunakan algoritma *K-Means* dengan tiga klaster ( $k = 3$ ). *Dataset* yang dianalisis terdiri dari seratus data siswa yang memiliki metrik utama yang menunjukkan prestasi akademik mereka. Mereka termasuk skor Matematika, skor Ilmu Pengetahuan Alam, skor Bahasa Inggris, dan skor partisipasi kelas. Perangkat lunak yang disebut *RapidMiner* digunakan untuk melakukan analisis.

Setelah proses klasterisasi, terbentuk tiga kelompok siswa: Klaster 0 terdiri dari 32 siswa, Klaster 1 terdiri dari 29 siswa, dan Klaster 2 terdiri dari 39 siswa. Nilai *centroid* dari masing-masing atribut menunjukkan bahwa setiap klaster memiliki karakteristik akademik unik.

| Attribute     | cluster_0 | cluster_1 | cluster_2 |
|---------------|-----------|-----------|-----------|
| Math_Score    | 65.594    | 56.621    | 60.436    |
| Science_Score | 70.594    | 66.769    | 51.769    |
| English_Score | 70.594    | 67.414    | 71.282    |
| Absence_Rate  | 84.455    | 79.582    | 81.009    |
| Participation | 1.875     | 2.689     | 2.006     |

Tabel 1 Nilai *Centroid* dari Masing-Masing Klaster

### Interpretasi klaster:

- **Klaster 0** memiliki siswa dengan nilai Matematika dan Bahasa Inggris yang cukup baik (68,63 dan 75,63), kehadiran baik (83,46), tetapi partisipasi relatif rendah (1,875). Ini menunjukkan bahwa siswa memiliki performa akademik yang baik tetapi partisipasi kelas sedang.
- **Klaster 1** memiliki nilai IPA tertinggi (84,76), partisipasi aktif (2,689), dan kehadiran yang sangat baik (91,76), tetapi mereka memiliki nilai matematika yang rendah (60,31).
- **Klaster 2** Klaster 2 rata-rata (Matematika 69,82, Sains 71,19, dan Kehadiran 79,62), tetapi partisipasi sangat tinggi (3,00), menunjukkan bahwa siswa aktif dengan performa moderat.

### Evaluasi Hasil Klasterisasi

*RapidMiner* menggunakan dua metrik: *Average Within Centroid Distance* dan *Davies-Bouldin Index*. Nilai DBI yang tinggi menunjukkan bahwa klaster saling terpisah dengan baik, sedangkan nilai DBI yang rendah menunjukkan bahwa data dalam satu klaster sangat mirip. Akibatnya, hasil evaluasi menunjukkan bahwa klasterisasi memiliki kualitas yang baik dan layak digunakan sebagai dasar pengambilan keputusan berbasis data.

```
PerformanceVector:  
Avg. within centroid distance: -576.146  
Avg. within centroid distance_cluster_0: -554.583  
Avg. within centroid distance_cluster_1: -565.250  
Avg. within centroid distance_cluster_2: -601.941  
Davies Bouldin: -1.337
```

### Analisis dan Implikasi

Hasil klasterisasi siswa ini memberi institusi pendidikan pengetahuan praktis. Pertama, intervensi dapat diberikan kepada siswa dalam klaster berisiko rendah, yang memiliki nilai dan kehadiran rendah. Kedua, siswa dengan nilai rata-rata dan partisipasi tinggi dapat diarahkan untuk mengembangkan soft skills. Ketiga, siswa dengan prestasi tinggi dapat diberi program tantangan lanjutan untuk mempertahankan dan meningkatkan prestasinya.

Secara keseluruhan, temuan ini dapat membantu guru dan lembaga akademik dalam membuat pendekatan pembelajaran yang lebih sesuai dengan kebutuhan siswa dan berbasis pada profil mereka.

### Kelebihan dan Keterbatasan Metode

Algoritma *K-Means* lebih efisien dan mudah digunakan. Hasilnya didasarkan pada nilai *centroid*, sehingga mudah dipahami dan sangat cocok untuk analisis eksploratif pada tahap awal klasifikasi data. Namun, metode ini memiliki kelemahan. Sejak awal, jumlah klaster harus ditentukan; jika nilai *k* salah dipilih, hasilnya bisa menyesatkan. Selain itu, agar atribut tidak mendominasi hasil, *K-Means* memerlukan normalisasi data karena sensitif terhadap outlier. *K-Means* juga tidak dapat menemukan klaster dengan distribusi yang kompleks atau bentuk tidak bundar.

### D. PENUTUP

#### Kesimpulan

Penelitian ini menunjukkan bahwa algoritma *K-Means* dapat digunakan secara efektif untuk mengatur siswa berdasarkan prestasi akademik mereka. Dengan menggunakan perangkat lunak *RapidMiner*, siswa dibagi ke dalam tiga kelompok besar berdasarkan prestasi akademik mereka: kelompok tinggi, sedang, dan rendah. Dengan menggunakan metrik seperti nilai dalam mata pelajaran Matematika, Ilmu Pengetahuan Alam, Bahasa Inggris, dan tingkat kehadiran dan partisipasi dalam kelas, Anda dapat mendapatkan profil akademik yang lengkap dari siswa.

Hasil evaluasi yang dilakukan menggunakan metrik *Davies-Bouldin Index* menunjukkan bahwa klaster yang terbentuk memiliki kualitas yang baik; ini ditunjukkan oleh sebaran data yang seragam dalam setiap klaster dan jarak yang jelas di antara kelompok. Sekolah dapat menggunakan temuan ini untuk membuat metode pembelajaran yang lebih tepat sasaran dan menemukan siswa yang mungkin mengalami penurunan prestasi akademik sejak dini.

### Saran

Agar hasil klasterisasi lebih aplikatif untuk pengembangan ke depan, penelitian harus menggunakan data asli dari institusi pendidikan. Untuk mencegah bias yang disebabkan oleh perbedaan skala antar atribut, proses normalisasi data juga penting. Sebaiknya gunakan metode evaluasi seperti *Elbow Method* atau *Silhouette Score* untuk menentukan jumlah klaster. Selain itu, sebagai dasar kebijakan akademik berbasis data, algoritma tambahan seperti DBSCAN dan *Clustering Hierarchical* dapat dipertimbangkan untuk hasil yang lebih luas.

### Ucapan Terima Kasih

Penulis berterima kasih kepada semua orang yang membantu dalam proses membuat dan menyelesaikan penelitian ini, terutama pengelola platform *Kaggle* yang menyediakan *Dataset* terbuka untuk keperluan pembelajaran dan penelitian.

### E. DAFTAR PUSTAKA

- Bawono, R. (2020). Perlindungan data pribadi di era digital. *Jurnal Ilmiah Komputer dan Informatika (JIKI)*, 5(1), 22–29.
- Bayu, D., & Aditya, H. (2022). Edukasi penggunaan internet aman untuk remaja. *Jurnal Abdimas Teknologi*, 6(2), 110–117.
- Daniel, D. T. (2023). Pengelompokan penerimaan mahasiswa baru dengan algoritma *K-Means* untuk meningkatkan potensi pemasaran. *Bulletin of Information Technology (BIT)*, 4(3), 294–298.
- Han, J., & Pei, J. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.
- Jefri, J., & Fatah, Z. (2025). Klasifikasi data mining untuk memprediksi kelulusan mahasiswa menggunakan metode Naive Bayes. *Jurnal Ilmiah Multidisiplin Ilmu*, 2(1), 29–37.
- Kaur, H., & Verma, P. (2021). Performance analysis of *K-Means Clustering* algorithm with different Distance metrics. *International Journal of Scientific & Technology Research*, 10(2), 164–169.
- Linoff, G. S., & Berry, M. J. A. (2022). *Data mining techniques: For marketing, sales, and customer relationship management* (3rd ed.). Wiley.
- Oktarina, R. (2021, July). Determine the *Clustering* of cities in Indonesia for disaster management using *K-Means* by Excel and *RapidMiner*. In *IOP Conference Series: Earth and Environmental Science* (Vol. 794, No. 1, p. 012094). IOP Publishing.
- Prasetyo, V. R., Lazuardi, H., Mulyono, A. A., & Lauw, C. (2021). Penerapan aplikasi *RapidMiner* untuk prediksi nilai tukar rupiah terhadap US dollar dengan

- metode regresi linier. *Jurnal Nasional Teknologi dan Sistem Informasi (TEKNOSI)*, 7(1), 8–17.
- Purnama, Y. W., & Marito, W. (2023). Penerapan algoritma *K-Means* untuk pengelompokan data akademik mahasiswa. *Jurnal Teknologi Informasi dan Pendidikan*, 16(1), 45–52.
- Putra, A. R. (2023). The implementation of data mining techniques for predicting student study period using the C4.5 algorithm: Penerapan teknik data mining terhadap prediksi masa studi mahasiswa menggunakan algoritma C4.5. *Indonesian Journal of Electrical Engineering and Renewable Energy (IJEERE)*, 3(2), 96–100.
- Rahmadayanti, F., & Rahayu, R. (2023). Penerapan metode data mining pada kasus kriminalitas Indonesia. *Jurnal Teknologi Informasi Mura*, 15(1), 52–61.
- Rahmadini, R., Lubis, E. E. L., Priansyah, A., RWN, Y., & Meutia, T. (2023). Penerapan data mining untuk memprediksi harga bahan pangan di Indonesia menggunakan algoritma K-Nearest Neighbor. *Jurnal Mahasiswa Akuntansi Samudra*, 4(4), 223–235.
- Sabna, E., & Arrafi, A. (2024). Community service: Increasing lecturer competence through data mining training using *RapidMiner* tools in the Master of Public Health Study Program, Faculty of Health, Hang Tuah University, Pekanbaru. *RECORD: Journal of Loyalty and Community Development*, 1(3), 143–150.
- Safnita, F., & Dewi, R. (2024). Penerapan algoritma *K-Means* dalam pengklasteran hasil evaluasi akademik mahasiswa. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, 5(2), 513–520.
- Sudarsono, B. G., Leo, M. I., Santoso, A., & Hendrawan, F. (2021). Analisis data mining data Netflix menggunakan aplikasi *RapidMiner*. *JBASE: Journal of Business and Audit Information Systems*, 4(1).
- Suraya, S., & Sari, R. (2023). Penerapan metode *Clustering* dengan algoritma *K-Means* pada pengelompokan indeks prestasi akademik mahasiswa. *SKANIKA: Sistem Komputer dan Teknik Informatika*, 6(1), 51–60.
- Susetyo, A., & Nugraha, A. (2022). Analisis pengelompokan data menggunakan algoritma *K-Means* dalam sistem informasi akademik. *Jurnal Ilmu Komputer dan Aplikasi*, 8(2), 78–84.
- Wahap, A. L., & Putra, A. P. (2024). Implementasi algoritma *K-Means Clustering* pendeteksian dini performa siswa pada pembelajaran Bahasa Indonesia. *Prosiding Seminar Nasional SATI*, 3(1), 581–592.
- Watratana, A. F., & Moeis, D. (2020). Implementasi algoritma Naive Bayes untuk memprediksi tingkat penyebaran COVID-19 di Indonesia. *Journal of Applied Computer Science and Technology*, 1(1), 7–14.