

ANALISIS PENGELOMPOKAN TINGKAT PELANGGARAN SANTRI MENGUNAKAN ALGORITMA K-MEANS CLUSTERING DI PONDOK PESANTREN AL-MA'RIFAH

Abdul Rohman¹, Nana Suarna², Agus Bahtiar³, Ade Irma Purnamasari⁴.

Program Studi Teknik Informatika¹²⁴
Program Studi Sistem Informasi³

STMIK IKMI Cirebon
<https://ikmi.ac.id/page/18/?lang=de>
adung0224@gmail.com

(*) Corresponding Author : adung0224@gmail.com
Published : 30 Juni 2026

Abstract—The application of machine learning technology in managing student behavior opens up new opportunities for improving the effectiveness of guidance systems in educational settings, including Islamic boarding schools. This study aims to analyze patterns of student misconduct at the Al-Ma'rifah Islamic Boarding School using the K-Means Clustering algorithm. The research data consists of records of student violations that include the type of violation, frequency, severity level, and total violation points. The analysis process includes data preprocessing, determining the optimal number of clusters using the Elbow Method, applying the K-Means algorithm, and evaluating the clustering results using the Silhouette Score. The research results indicate that student violation data can be grouped into three main clusters: minor, moderate, and severe violations. A Silhouette Score of 0.33 indicates a reasonably good clustering quality and can serve as a basis for identifying student behavior patterns. These findings provide practical contributions to Islamic boarding schools in designing more objective, measurable, and data-driven guidance strategies. Additionally, this study reinforces the evidence that the data mining approach using K-Means is effective in helping educational institutions gain a deeper understanding of students' disciplinary behavior.

Keywords: K-Means Clustering; data mining, student violations, machine learning, Islamic boarding school.

Abstrak—Penerapan teknologi *machine learning* dalam pengelolaan perilaku peserta didik membuka peluang baru dalam meningkatkan efektivitas sistem pembinaan di lingkungan pendidikan, termasuk pesantren. Penelitian ini bertujuan untuk menganalisis pola pelanggaran santri di Pondok Pesantren Al-Ma'rifah menggunakan algoritma *K-Means Clustering*. Data penelitian terdiri dari catatan pelanggaran santri yang mencakup jenis pelanggaran, frekuensi, tingkat keparahan, serta total poin pelanggaran. Proses analisis meliputi *data preprocessing*, penentuan jumlah cluster optimal menggunakan Elbow Method, penerapan algoritma K-Means, serta evaluasi hasil klusterisasi menggunakan nilai Silhouette Score. Hasil penelitian menunjukkan bahwa data pelanggaran santri dapat dikelompokkan ke dalam 3 cluster utama, yaitu pelanggaran ringan, sedang dan berat. Nilai Silhouette Score sebesar 0,33 menunjukkan kualitas pengelompokan yang cukup baik dan dapat digunakan sebagai dasar identifikasi pola perilaku santri. Temuan ini memberikan kontribusi praktis bagi pihak pesantren dalam merancang strategi pembinaan yang lebih objektif, terukur, dan berbasis data. Selain itu, penelitian ini memperkuat bukti bahwa pendekatan data mining melalui K-Means efektif untuk membantu lembaga pendidikan dalam memahami perilaku disiplin peserta didik secara lebih mendalam.

Kata Kunci: K-Means Clustering, data mining, pelanggaran santri, machine learning, pesantren.

INTRODUCTION

Perkembangan teknologi informasi saat ini memungkinkan berbagai institusi pendidikan untuk memanfaatkan data sebagai dasar dalam pengambilan keputusan. Salah satu pendekatan yang banyak digunakan adalah teknik data mining yang mampu mengekstraksi informasi

penting dari kumpulan data yang besar. Dalam bidang pendidikan, penerapan data mining dikenal dengan istilah Educational Data Mining (EDM), yaitu proses analisis data pendidikan untuk menemukan pola dan informasi yang dapat mendukung peningkatan kualitas pembelajaran dan manajemen pendidikan.

Salah satu permasalahan yang sering terjadi di lingkungan pendidikan adalah pelanggaran disiplin yang dilakukan oleh siswa atau santri. Pelanggaran tersebut dapat berupa pelanggaran ringan, sedang, maupun berat yang biasanya dicatat dalam sistem administrasi kedisiplinan. Data pelanggaran tersebut sebenarnya memiliki potensi untuk dianalisis lebih lanjut guna mengetahui pola perilaku siswa atau santri. Namun dalam banyak kasus, data pelanggaran hanya digunakan sebagai arsip administrasi tanpa dilakukan analisis yang lebih mendalam.

Di lingkungan pesantren, kedisiplinan santri merupakan salah satu aspek penting dalam pembentukan karakter dan proses pendidikan. Oleh karena itu, pengelolaan data pelanggaran santri menjadi hal yang penting untuk membantu pihak pesantren dalam memahami pola perilaku santri serta menentukan strategi pembinaan yang lebih tepat. Dengan memanfaatkan teknik analisis data, data pelanggaran santri dapat diolah untuk menemukan kelompok santri berdasarkan tingkat pelanggaran yang mereka lakukan.

Salah satu metode data mining yang dapat digunakan untuk melakukan pengelompokan data adalah *K-Means Clustering*. Metode ini merupakan salah satu algoritma clustering yang populer karena mampu mengelompokkan data berdasarkan kemiripan karakteristik antar data. Dengan menggunakan metode *K-Means*, data pelanggaran santri dapat dikelompokkan ke dalam beberapa cluster yang merepresentasikan tingkat pelanggaran santri.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menerapkan metode *K-Means Clustering* dalam menganalisis data pelanggaran santri sehingga dapat mengidentifikasi kelompok santri berdasarkan tingkat pelanggaran yang dilakukan. Hasil penelitian ini diharapkan dapat memberikan gambaran pola pelanggaran santri serta menjadi bahan pertimbangan bagi pihak pesantren dalam melakukan pembinaan dan pengelolaan kedisiplinan santri secara lebih efektif.

Perkembangan teknologi *machine learning* dan *educational data mining* (EDM) telah merevolusi cara lembaga pendidikan memahami, menganalisis, dan memprediksi perilaku peserta didik. EDM tidak hanya berfungsi untuk menilai hasil belajar, tetapi juga mampu mendeteksi pola perilaku yang berpotensi mengarah pada pelanggaran atau penurunan kinerja akademik [1]. Pendekatan ini memungkinkan lembaga pendidikan untuk mengidentifikasi siswa berisiko melalui temuan

pola tersembunyi pada data akademik dan non-akademik, sehingga dapat dilakukan intervensi secara lebih dini dan efektif. [2]menambahkan bahwa penerapan *knowledge discovery* berbasis *data mining* mampu membantu lembaga pendidikan mengenali faktor-faktor yang memengaruhi perilaku dan retensi siswa.

Berbagai penelitian terkini memperkuat relevansi penerapan *machine learning* di bidang pendidikan. [3]menunjukkan bahwa analisis lintasan pembelajaran mahasiswa menggunakan visualisasi dan *data mining* memberikan pemahaman yang lebih mendalam terhadap perilaku akademik individu. [4]mengembangkan sistem cerdas berbasis *Light GBM* untuk memprediksi performa siswa, [5] membuktikan efektivitas algoritma *tree-based* dalam mengidentifikasi faktor signifikan yang memengaruhi hasil belajar. Meskipun sebagian besar penelitian tersebut berfokus pada performa akademik, prinsip dan metode yang digunakan dapat diadaptasi untuk memahami perilaku disiplin peserta didik di lembaga pendidikan lainnya.

Dalam konteks pendidikan Islam, khususnya di pondok pesantren, perilaku santri merupakan aspek fundamental yang menentukan keberhasilan proses pembinaan spiritual dan akademik. Kedisiplinan dan kepatuhan terhadap tata tertib menjadi ukuran keberhasilan pendidikan karakter di lingkungan pesantren. Namun, pendekatan tradisional dalam menilai dan menangani pelanggaran santri sering kali bersifat subjektif dan reaktif, bergantung pada penilaian individu pembina tanpa dukungan analisis data yang terstruktur. Kondisi ini menyebabkan pengelolaan perilaku santri kurang efisien dan sulit dievaluasi secara objektif.

Penelitian-penelitian sebelumnya menunjukkan bahwa analisis perilaku siswa melalui teknik *data mining* dapat membantu mengidentifikasi pola pelanggaran secara lebih sistematis. [6] menggunakan *association rule mining* untuk mengungkap hubungan antara kebiasaan belajar dan pelanggaran disiplin siswa. [7] menegaskan bahwa kombinasi metode *clustering* dan *rule-mining* dapat memperkuat sistem manajemen perilaku mahasiswa dalam konsep *smart campus*. [8]juga menunjukkan bahwa analisis perilaku berbasis *embedded systems* mampu memberikan umpan balik real-time kepada pendidik, sementara [9]menyoroti peran *learning analytics* berbasis *big data* dalam memprediksi perilaku siswa serta membangun strategi pembinaan yang terpersonalisasi.

Algoritma *K-Means Clustering* merupakan salah satu metode *unsupervised learning* yang paling banyak digunakan dalam pengelompokan data

berdasarkan kesamaan karakteristik. Metode ini bekerja dengan membagi data ke dalam sejumlah kelompok (*cluster*) berdasarkan jarak terdekat terhadap pusat kelompok (*centroid*). [10] menjelaskan bahwa keunggulan K-Means terletak pada kemampuannya melakukan segmentasi data dalam waktu komputasi yang efisien serta menghasilkan pola pengelompokan yang mudah diinterpretasikan. Oleh karena itu, metode ini sangat sesuai digunakan untuk mengelompokkan data pelanggaran santri berdasarkan atribut seperti jenis pelanggaran, tingkat keparahan, dan frekuensi kejadian.

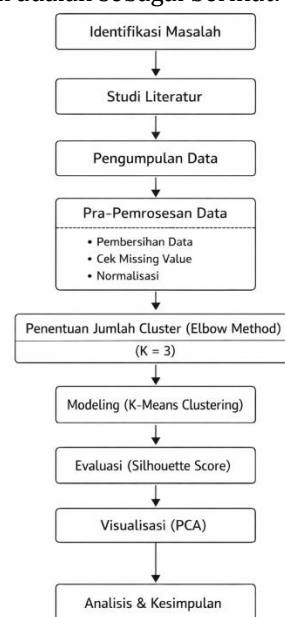
Penerapan algoritma K-Means dalam konteks pesantren memungkinkan terwujudnya pengelolaan disiplin santri berbasis data. Dengan mengelompokkan data pelanggaran ke dalam kategori tertentu misalnya pelanggaran ringan, sedang, dan berat pihak pesantren dapat mengenali pola perilaku santri secara objektif. Hasil pengelompokan tersebut dapat digunakan sebagai dasar dalam merancang strategi pembinaan yang lebih tepat sasaran, efektif, dan preventif. Dengan demikian, penelitian ini berupaya mengintegrasikan pendekatan ilmiah berbasis *data mining* dalam sistem pembinaan santri, untuk mendukung transformasi pengelolaan perilaku dari pendekatan reaktif menuju pendekatan proaktif yang berbasis bukti (*evidence-based discipline management*). Meskipun berbagai penelitian telah menerapkan teknik data mining untuk menganalisis perilaku siswa, sebagian besar penelitian berfokus pada performa akademik atau prediksi nilai siswa. Penelitian yang secara khusus menganalisis pola pelanggaran disiplin santri di lingkungan pesantren masih sangat terbatas. Selain itu, sebagian besar penelitian belum memanfaatkan pendekatan clustering untuk mengidentifikasi kelompok perilaku pelanggaran secara sistematis. Oleh karena itu, penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan menerapkan algoritma K-Means Clustering untuk mengelompokkan tingkat pelanggaran santri berdasarkan pola pelanggaran yang terjadi.

MATERIALS AND METHODS

Penelitian ini merupakan penelitian kuantitatif yang menggunakan pendekatan data mining untuk menganalisis pola tingkat pelanggaran santri di Pondok Pesantren Al-Ma'rifah. Pendekatan kuantitatif dipilih karena data yang digunakan berupa angka (numerik) yang dapat diolah secara matematis untuk menemukan pola tertentu melalui proses

clustering. Teknik analisis utama yang digunakan adalah algoritma K-Means Clustering, yaitu algoritma pengelompokan (*unsupervised learning*) yang bertujuan untuk mengelompokkan objek ke dalam beberapa cluster berdasarkan tingkat kemiripannya.

Penelitian ini menggunakan metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*) sebagai kerangka kerja dalam proses analisis data. CRISP-DM merupakan metode standar yang digunakan dalam proyek data mining yang terdiri dari enam tahapan utama, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*. Adapun tahapan CRISP-DM yang diterapkan dalam penelitian ini adalah sebagai berikut:



Gambar 1. Alur Penelitian

Business Understanding

Tahap ini bertujuan untuk memahami permasalahan yang terjadi pada objek penelitian. Dalam penelitian ini permasalahan yang diidentifikasi adalah belum adanya analisis berbasis data untuk mengetahui pola pelanggaran santri di lingkungan pesantren. Oleh karena itu diperlukan analisis menggunakan teknik data mining untuk mengelompokkan tingkat pelanggaran santri.

Data Understanding

Tahap ini dilakukan dengan mengumpulkan dan memahami dataset yang akan digunakan dalam penelitian. Data yang digunakan merupakan data arsip pelanggaran santri yang diperoleh dari bagian kedisiplinan pesantren. Dataset tersebut berisi informasi mengenai jumlah pelanggaran ringan, pelanggaran sedang, pelanggaran berat, serta total poin pelanggaran yang dimiliki oleh santri.

Data preparation

Pada tahap ini dilakukan proses persiapan data sebelum dilakukan pemodelan. Proses ini meliputi validasi dataset, pembersihan data, serta normalisasi data menggunakan metode *MinMaxScaler* untuk menyesuaikan rentang nilai pada setiap variabel sehingga proses perhitungan jarak pada algoritma K-Means dapat dilakukan secara optimal.

Modeling

Tahap modeling dilakukan dengan menerapkan algoritma *K-Means Clustering* untuk mengelompokkan data pelanggaran santri berdasarkan kemiripan karakteristik data. Penentuan jumlah cluster dilakukan menggunakan *Elbow Method* dengan rentang nilai K dari 1 hingga 10 untuk memperoleh jumlah cluster optimal.

Evaluation

Tahap evaluasi dilakukan untuk menilai kualitas model clustering yang dihasilkan. Evaluasi model dilakukan menggunakan *Silhouette Score* untuk mengukur tingkat pemisahan dan kekompakan antar cluster. Selain itu, hasil clustering juga divisualisasikan menggunakan *Principal Component Analysis (PCA)* untuk mempermudah interpretasi distribusi cluster.

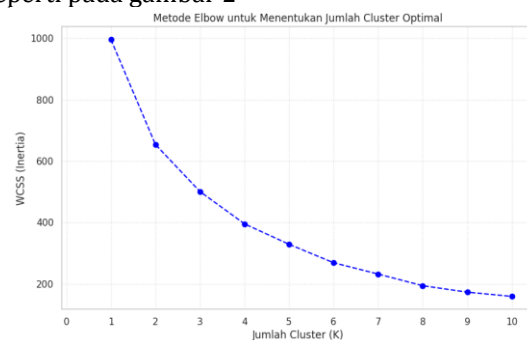
Deployment

Tahap deployment merupakan tahap penyajian hasil analisis kepada pihak yang berkepentingan. Dalam penelitian ini hasil clustering digunakan untuk memberikan gambaran mengenai pola pelanggaran santri sehingga dapat menjadi bahan pertimbangan bagi pihak pesantren dalam melakukan pembinaan dan pengelolaan kedisiplinan santri.

RESULTS AND DISCUSSION

Penentuan jumlah cluster merupakan tahap penting sebelum menerapkan algoritma *K-Means Clustering*, karena nilai K yang tepat akan menghasilkan struktur pengelompokan yang optimal. Pada penelitian ini, penentuan jumlah cluster dilakukan menggunakan *Elbow Method*, yaitu metode yang mengukur nilai *Within-Cluster Sum of Squares (WCSS)* untuk berbagai nilai K. WCSS merupakan total jarak kuadrat antara setiap data dengan centroid cluster-nya. Semakin kecil nilai WCSS, semakin baik data dikelompokkan dalam cluster

tersebut. Namun, peningkatan nilai K secara terus-menerus akan membuat WCSS semakin kecil tanpa memberikan peningkatan signifikan dalam kualitas cluster. Oleh karena itu, dipilih titik dimana penurunan WCSS mulai melambat tajam, yang disebut sebagai *elbow point*. Pada penelitian ini, perhitungan WCSS dilakukan untuk K = 1 hingga K = 10, dan hasilnya divisualisasikan dalam bentuk grafik *Elbow*. Grafik tersebut menunjukkan adanya titik siku (*elbow*) yang jelas pada K = 3. Titik ini menandakan bahwa penambahan jumlah cluster setelah K = 3 tidak memberikan pengurangan WCSS yang signifikan, sehingga tidak meningkatkan kualitas pengelompokan secara berarti. Dengan demikian, jumlah cluster optimal yang digunakan dalam penelitian ini adalah K = 3. Nilai ini selanjutnya digunakan dalam proses penerapan algoritma *K-Means Clustering* untuk mengelompokkan data pelanggaran santri ke dalam empat kategori berdasarkan karakteristik tingkat pelanggarannya. Adapun Penentuan jumlah cluster seperti pada gambar 2



Gambar 2 Penentuan jumlah cluster

Evaluasi Model Clustering

Evaluasi model dilakukan untuk menilai kualitas hasil pengelompokan data pelanggaran santri yang dihasilkan oleh algoritma K-Means. Pada penelitian ini evaluasi model dilakukan menggunakan *Elbow Method* dan *Silhouette Score*.

Evaluasi Menggunakan *Elbow Method* *Elbow Method* digunakan untuk menentukan jumlah cluster yang paling optimal dalam proses clustering. Metode ini dilakukan dengan menghitung nilai *Sum of Squared Error (SSE)* pada berbagai jumlah cluster dengan rentang nilai K = 1 sampai K = 10. Berdasarkan hasil perhitungan *Elbow Method*, diperoleh titik siku (*elbow point*) pada nilai K = 3. Hal ini menunjukkan bahwa jumlah cluster yang paling optimal untuk mengelompokkan data pelanggaran santri adalah 3 cluster. Dengan demikian, proses clustering pada penelitian ini menggunakan K = 3 sebagai jumlah cluster yang digunakan dalam algoritma K-Means

Evaluasi Menggunakan Silhouette Score Selain menggunakan Elbow Method, evaluasi model juga dilakukan menggunakan *Silhouette Score* untuk mengukur tingkat kualitas cluster yang terbentuk. *Silhouette Score* digunakan untuk menilai sejauh mana suatu data berada dalam cluster yang tepat dengan membandingkan jarak antar data dalam satu cluster dan jarak dengan cluster lainnya. Hasil perhitungan menunjukkan bahwa model K-Means dengan jumlah cluster $K = 3$ menghasilkan nilai *Silhouette Score* sebesar 0.3329. Nilai tersebut menunjukkan bahwa cluster yang terbentuk memiliki tingkat pemisahan dan kekompakan yang cukup baik untuk data perilaku pelanggaran santri.

Visualisasi Hasil Clustering Untuk mempermudah interpretasi hasil clustering, dilakukan visualisasi menggunakan metode *Principal Component Analysis (PCA)* yang mereduksi dimensi dataset menjadi dua komponen utama. Visualisasi ini membantu memperlihatkan distribusi data pada masing-masing cluster sehingga pola pengelompokan santri berdasarkan tingkat pelanggaran dapat terlihat dengan lebih jelas.

Hasil clustering ini memberikan gambaran komprehensif mengenai pola pelanggaran santri, yang dapat dijadikan dasar dalam penyusunan strategi pembinaan, pengawasan, dan evaluasi kedisiplinan di Pondok Pesantren Al-Ma'rifah. Pada tahap ini dilakukan analisis terhadap rata-rata total poin pelanggaran yang terdapat pada setiap cluster. Visualisasi berikut menunjukkan perbandingan rata-rata poin pelanggaran di antara cluster 0 hingga 3. Adapun Gamabr rata rata total poin pelanggaran per cluster seperti pada gambar 3



Gambar 3 Rata rata total poin pelanggaran

Berdasarkan Gambar 3 menunjukkan distribusi jumlah santri yang masuk ke masing-masing cluster setelah proses pengelompokan

menggunakan algoritma K-Means. Terdapat tiga cluster yang terbentuk (Cluster 0, Cluster 1, dan Cluster 2), dengan jumlah santri yang relatif seimbang. Cluster 0 memiliki jumlah santri tertinggi, yaitu 87 santri. Cluster 1 berisi 78 santri, menjadi cluster dengan jumlah paling sedikit. Cluster 2 berisi 85 santri, hampir setara dengan Cluster 0.

Distribusi yang seimbang ini menunjukkan bahwa model K-Means berhasil mengelompokkan data dengan proporsi yang relatif merata, tanpa adanya cluster yang terlalu dominan atau terlalu kecil. Hal ini membantu memastikan bahwa hasil pengelompokan lebih stabil dan representatif terhadap pola pelanggaran santri.

Hasil ini memperkuat bahwa penerapan algoritma K-Means mampu memisahkan santri berdasarkan tingkat keparahan pelanggaran secara jelas dan sistematis. Dengan visualisasi ini, pihak pesantren dapat lebih mudah menentukan prioritas intervensi pembinaan sesuai tingkat pelanggaran masing-masing kelompok. Pada tahap ini dilakukan proses implementasi algoritma *K-Means Clustering* dengan jumlah cluster optimal $K = 3$ yang diperoleh dari hasil analisis Elbow Method sebelumnya. Pada proses ini, data yang telah melalui tahap pra-pemrosesan dan standardisasi kemudian dikelompokkan berdasarkan kemiripan pola pelanggaran masing-masing santri. Hasil pengelompokan ini menghasilkan label cluster untuk setiap santri, yang menunjukkan posisi mereka dalam kelompok pelanggaran tertentu. Berikut ditampilkan sebagian hasil pengelompokan K-Means yang menunjukkan nama santri, cluster yang diperoleh, serta total poin pelanggaran. Adapun Penerapan Algoritma k-means seperti pada gambar 4

--- TAHAP 4: PENERAPAN ALGORITMA K-MEANS CLUSTERING ---

DataFrame asli telah ditambahkan kolom 'Cluster' dengan 3 kelompok.
5 Data teratas dengan label Cluster:

NO	Nama Santri	Jumlah Pelanggaran	Pel Ringan	Pel Sedang	Pel Berat	Usia	Usia Mndk	Total Poin Pelanggaran	Cluster	
0	1	M.Saf'i	8	5.0	3	1	14	2.0	75	1
1	2	Dzikron Fagih	5	2.0	2	1	15	3.0	50	1
2	3	Ragata Hikmal	4	2.0	0	2	15	3.0	50	2
4	5	Kurniasan	5	4.0	1	0	17	3.0	30	2
5	6	Ardil F	9	5.0	2	2	15	3.0	85	0

Gambar 4 Implementasi k-means

Gambar 4 menampilkan lima data teratas setelah proses clustering menggunakan algoritma K-Means dengan jumlah cluster = 3. Pada tabel tersebut terlihat bahwa dataframe asli telah ditambahkan kolom baru bernama "Cluster", yang menunjukkan setiap santri termasuk ke dalam kelompok mana berdasarkan pola pelanggaran. Setiap cluster merepresentasikan tingkat pelanggaran yang berbeda sesuai karakteristik total poin dan jenis pelanggaran:

Cluster 0 umumnya berisi santri dengan tingkat pelanggaran sedang hingga tinggi. Cluster 1 cenderung berisi santri dengan tingkat pelanggaran

menengah. Cluster 2 berisi santri dengan tingkat pelanggaran relatif ringan.

Dari contoh 5 data pertama pada tabel, terlihat bahwa setiap santri memperoleh label cluster yang berbeda sesuai hasil pengelompokan. Ini menunjukkan bahwa algoritma K-Means berhasil memisahkan santri berdasarkan kemiripan pola pelanggaran, sehingga dapat digunakan untuk memahami kelompok mana yang membutuhkan pembinaan lebih intensif.

CONCLUSION

Berdasarkan hasil penelitian yang telah dilakukan mengenai Analisis Pengelompokan Tingkat Pelanggaran Santri Menggunakan Algoritma K-Means Clustering di Pondok Pesantren Al-Ma'rifah, dapat diperoleh beberapa kesimpulan penting sebagai berikut:

Penelitian berhasil mengelompokkan santri ke dalam empat cluster tingkat pelanggaran berdasarkan data numerik yang meliputi Pel_Ringan, Pel_Sedang, Pel_Berat, dan Total_Poin_Pelanggaran.

Hasil Elbow Method menunjukkan bahwa titik siku (elbow point) berada pada $K = 3$, sehingga jumlah cluster optimal yang digunakan dalam penelitian ini adalah empat kelompok. Setiap cluster menggambarkan perbedaan intensitas dan jenis pelanggaran yang dilakukan oleh santri.

Proses pra-pemrosesan data (data cleaning, pengecekan duplikat dan missing value, serta normalisasi dengan *MinMaxScaler*) berhasil membuat dataset berada dalam kondisi optimal untuk dianalisis.

Algoritma K-Means Clustering mampu memberikan struktur pengelompokan yang cukup baik untuk data sosial-keperilaku.

Hal ini ditunjukkan oleh nilai *Silhouette Score* sebesar 0.3329, yang menggambarkan bahwa kualitas cluster berada dalam kategori "cukup baik", mengingat data pelanggaran santri cenderung memiliki tumpang tindih (overlap) antar individu. Setiap cluster memiliki karakteristik yang berbeda sehingga dapat memberikan wawasan mendalam bagi pihak pesantren.

Visualisasi PCA 2D membantu memperjelas persebaran data dalam dua dimensi dan menunjukkan posisi centroid dari masing-masing cluster. PCA berhasil merangkum pola clustering sehingga lebih mudah dipahami secara visual, terutama bagi pihak pesantren yang membutuhkan gambaran cepat mengenai tingkat pelanggaran santri. Hasil pengelompokan dapat menjadi bahan

pertimbangan dalam merancang strategi pembinaan, monitoring, dan evaluasi kedisiplinan santri secara lebih terstruktur dan berbasis data (*data-driven decision making*).

REFERENCE

- [1] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, pp. 1-14, 2022, doi: 10.1186/s40561-022-00192-z.
- [2] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge discovery for higher education student retention based on data mining: machine learning algorithms and case study in Chile," *Entropy*, vol. 23, no. 4, p. 485, 2021, doi: 10.3390/e23040485.
- [3] R. Maqsood, P. Ceravolo, M. Ahmad, and M. S. Sarfraz, "Examining students' course trajectories using data mining and visualization approaches," *International Journal of Educational Technology in Higher Education*, vol. 20, p. 55, 2023, doi: 10.1186/s41239-023-00423-4.
- [4] A. Abukader, A. Alzubi, and O. R. Adegboye, "Intelligent system for student performance prediction: an educational data mining approach using meta-heuristic-optimized LightGBM with SHAP-based learning analytics," *Applied Sciences*, vol. 15, no. 20, p. 10875, 2025, doi: 10.3390/app152010875.
- [5] F. Aydın, "Educational data mining: construction of a tree-based model to predict students' performance," *Bartın University Journal of Faculty of Education*, vol. 14, no. 1, pp. 181-195, 2025, doi: 10.14686/buefad.1390209.
- [6] T. Wang, B. Xiao, and W. Ma, "Student Behavior Data Analysis Based on Association Rule Mining," *International Journal of Computational Intelligence Systems*, vol. 15, p. 32, 2022, doi: 10.1007/s44196-022-00087-4.
- [7] J. Zhang, Y. Kuang, and J. Zhou, "Data Mining Technology for Smart Campus in Behavior Association Analysis of College Students," *Scalable Computing: Practice and Experience*, vol. 25, no. 6, 2024, doi: 10.12694/scpe.v25i6.2187.
- [8] L. Li, C. P. Chen, L. Wang, and W. Bao, "Exploring Artificial Intelligence in Smart Education," *Sustainability*, vol. 15, no. 10, p. 7940, 2023.

- [9] X. Luo, "Learning analytics based on big data: student behaviour prediction and personalized educational strategy formulation," *Applied and Computational Engineering (ACE)*, vol. 116, 2024, doi: 10.54254/2755-2721/116/20251732.
- [10] M. Zubair, B. Alqadri, and F. Artina, "Kesadaran Mahasiswa Terhadap Penyebaran Paham Radikalisme Melalui Media Literasi Online," *Pemikiran dan Penelitian Ilmu-Ilmu Sosial, Hukum & Pengajarannya*, vol. XVII, no. 1, pp. 115–127, 2022.