Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Application of the K-Means Clustering Algorithm for Classifying High-Achieving Students at MI Irsyadul Athfal Depok

Ilham Mubarok [1*], Rita Wahyuni Arifin [2]

[1]*Informatics engineering, Universitas Bina Insani, Indonesia.*
[2]*Informatics management,Universitas Bina Insani, Indonesia.*
ilhammubarok22@gmail.com[1], ritawahyuni@binainsani.ac.id[3]*

Abstract

Education is an important foundation in the development of knowledge, skills, and attitudes in accordance with the values of life. This process aims to improve people's standard of living. In this context, Madrasah Ibtidaiyah (MI) Irsyadul Athfal faces a challenge in selecting outstanding students which is still done subjectively, with the main focus on academic scores without considering other aspects. To improve objectivity and efficiency in assessment, this research proposes the use of machine learning technology with the K-Means clustering algorithm. This research aims to develop a prediction model for outstanding students based on academic and non-academic data, such as attendance, summative, mid- and end-of-semester assessments, and extracurricular activities. The K-Means algorithm was chosen because of its advantages in clustering data that is fast, simple, and flexible. The research was conducted at MI Irsyadul Athfal with observation data from local students. The results of the research are expected to provide accurate predictions of outstanding students, support more objective decision making, and increase student motivation in developing their potential. In addition, this research is also a reference for the development of similar prediction systems in other educational institutions.

Keywords: *Education, outstanding student prediction, K-Means clustering algorithm, machine learning, academic and non-academic data.*

1. Introduction

Education is a process of learning knowledge, skills, and habits passed from one generation to the next through various methods, aiming to foster better lives with attitudes and behaviors aligned with core values. Prediction, on the other hand, is the activity of estimating the most likely events in the future based on past and present information [1]. According to the Indonesian Dictionary (KBBI), prediction is the result of estimating or forecasting future outcomes using historical data. It is a systematic process to minimize the gap between predicted results and actual outcomes [2].

Madrasah Ibtidaiyah (MI) is a formal educational institution providing general education with an Islamic character. MI Irsyadul Athfal, a private MI located in Depok, West Java, plans to select its best students this year to recognize academic and extracurricular achievements [3]. Currently, the student evaluation system is largely subjective, relying only on academic performance such as report card grades, without considering other aspects of student development [4].

To address this issue, an innovative solution is to implement machine learning using the K-Means clustering algorithm. By analyzing data from students in grades 4, 5, and 6 in 2024—including academic and non-academic aspects like grades, attendance, and extracurricular participation—the algorithm can group students based on similarities [5]. This approach allows objective identification of high-achieving students while capturing patterns in both academic and non-academic performance [6].

This study aims to develop a predictive model for high-achieving students at MI Irsyadul Athfal using K-Means clustering. The model not only clusters students by performance levels but also provides a ranking feature based on combined indicators [7]. It is expected to help curriculum administrators select outstanding students more objectively and motivate students to improve both academic and non-academic skills, fostering balanced development and maximizing their potential [8].

2. Research Methodology

2.1. Cross Industry Standard Process for Data Mining (CRISP-DM)

This system was developed using the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, a widely adopted standard for planning, managing, and executing data mining projects. CRISP-DM is iterative and flexible, making it adaptable across various industries and research fields. Its application in this system ensures a systematic approach to developing a predictive model for high-achieving students, covering problem understanding, data processing, model building, evaluation, and implementation of predictions

into a web-based application [9]. This methodology provides a structured and flexible framework, supporting the entire data mining development process efficiently.

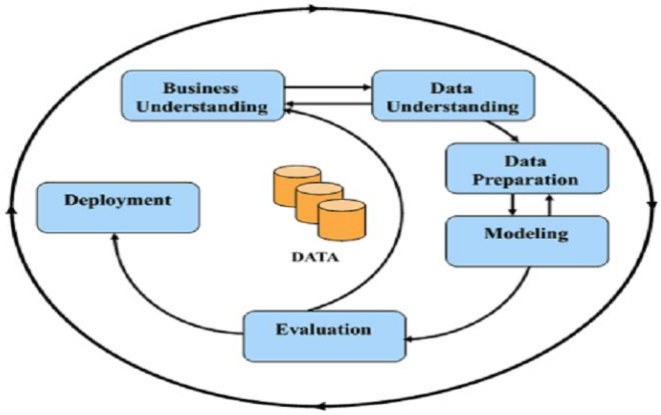

Fig. 1: Cross Industry Standard Process for Data Mining (CRISP-DM)

The stages of the CRISP-DM method are explained as follows:

1. Business Understanding:
   The system objectives were identified to help the madrasa, particularly the vice principal for curriculum, in objectively identifying high-achieving students. The main issue was the absence of a system capable of automatically grouping students based on both academic and non-academic data.
2. Data Understanding:
   This stage involved collecting, exploring, and analyzing initial student data. The dataset included report card grades, attendance, and extracurricular scores from grades 4, 5, and 6 of MI Irsyadul Athfal in the 2024 academic year. The goal was to understand the structure, distribution, and quality of the data and ensure its relevance for modeling.
3. Data Preparation:
   Collected data were cleaned and transformed, including handling missing values, normalizing data, and creating a new dataset ready for modeling. Weightings were assigned to each variable (grades 50%, attendance 30%, extracurricular 20%) to support the clustering process.
4. Modeling:
   The K-Means Clustering algorithm was applied to group students into three main clusters: High-achieving, Competent, and Needs Guidance. K-Means was chosen for its effectiveness in identifying patterns in unsupervised data and grouping students based on similar characteristics.
5. Evaluation:
   The clustering results were evaluated to ensure they reflected the initial objectives. The distribution of data in each cluster was examined, and 2D and 3D visualizations were used along with a ranking feature based on the average of the three indicators (grades, attendance, extracurricular). This ensured the model's accuracy and usefulness for decision-making at the madrasa level.
6. Deployment:
   The final model was implemented in a web-based application. Users can input student data, automatically perform clustering, visualize results, and export reports in PDF format. The system serves as a practical and efficient tool for selecting high-achieving students at MI Irsyadul Athfal.

2.2. K-Means Clustering Algorithm

The method used in this study is the K-Means Clustering algorithm, an unsupervised learning technique for grouping data into clusters based on similarity. The process begins by determining the number of clusters (k), followed by randomly initializing k centroids as the starting points. Each data point is then assigned to the nearest centroid using Euclidean distance, after which the centroids are recalculated as the mean of all data points in their respective clusters. This assignment and update process is repeated iteratively until the centroids stabilize (converge) or the maximum number of iterations is reached, resulting in optimal cluster formation [10].

$$d(x, c) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots + (x_n - c_n)^2}$$

Fig. 2: Euclidean Distance Formula

3. Result and Discussion

The student selection system at Madrasah Ibtidaiyah Irsyadul Athfal serves as a mechanism to recognize students with outstanding achievements in both academic and non-academic fields. Currently, the selection process is carried out manually through direct nomination based on teacher interviews, which limits efficiency and accuracy [11]. Moreover, the absence of analytical technology in predicting high-achieving students reduces the objectivity of the results. To address these challenges, a web-based application with predictive features using the K-Means Clustering algorithm is proposed. This system is expected to improve both efficiency and accuracy in identifying outstanding students, particularly assisting the Vice Principal for Student Affairs [12].

The proposed application integrates several key features to support effective identification and analysis of student data. It includes student data management for adding, updating, and deleting records such as grades, attendance, and extracurricular participation, along with bulk data import through Excel or CSV files [13]. The K-Means Clustering algorithm is applied to group students into categories based on similar characteristics, with the number of clusters determined manually or via the Elbow Method. Results are visualized in scatter plots,

and the Elbow Method graph assists in selecting the optimal cluster number. All clustering outcomes are stored in a database for further analysis, while an interactive dashboard provides easy access to data, features, and comprehensive summaries of the analysis results [14].

3.1. Functional and Non-Functional Requirements

In developing a web-based application for predicting high-achieving students at Madrasah Ibtidaiyah Irsyadul Athfal, it is necessary to define system requirements to ensure the application functions effectively and efficiently. The requirements are categorized into two types: functional requirements, which describe the core features that the system must provide, and non-functional requirements, which specify the hardware and software environment needed to support the system. These requirements serve as the foundation for system design and development to meet the goals of accuracy, usability, and efficiency in the student selection process.

Table 1: Functional Requirements

No	Functional Requirements	Description
1	Dashboard	Provides tools for preparing student data, running predictions, and visualizing results in graphical form.
2	Data Import	Allows users to upload student data in bulk (Excel or CSV format) to simplify data entry.
3	Prediction	Implements K-Means Clustering to identify high-achieving students based on academic and non-academic data.
4	Data Export	Generates reports of predicted high-achieving students, exportable in formats such as PDF.

Table 2: Non-Functional Requirements

No	Non-Functional Requirements	Description
1	Operating System	Minimum Windows 10 for system compatibility.
2	Memory (RAM)	At least 4 GB RAM to ensure smooth performance.
3	Web Browser	A modern browser is required, with Google Chrome recommended for optimal use.

3.2. Data Collection and Preprocessing

The data collection and preprocessing stage serves as the foundation for developing the student achievement prediction system. The dataset, consisting of 228 records, was obtained from the administration office of Madrasah Ibtidaiyah Irsyadul Athfal and includes academic performance, attendance, extracurricular activities, and other relevant information. After collection, the data underwent preprocessing steps such as cleaning, handling missing values, normalization, and selecting relevant attributes. These steps were carried out to ensure data consistency and quality before applying the K-Means clustering algorithm [15].

3.3. Algorithm Testing and Evaluation Results

At this stage, testing was conducted to evaluate the algorithm's ability to group students based on the defined characteristics. Furthermore, clustering results were assessed to measure the accuracy and relevance of the clusters, both statistically and through domain-specific interpretation. Below is a screenshot of the source code for the algorithm testing and evaluation results.

```
# Elbow Method
inertia = []
for k in range(1, 6):
    kmeans = KMeans(n_clusters=k, random_state=0)
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)

plt.figure()
plt.plot(range(1, 6), inertia, marker='o')
plt.title('Elbow Method')
plt.xlabel('Jumlah Cluster')
plt.ylabel('Inertia')
plt.savefig('static/elbow.png')

# Clustering dengan 3 cluster
kmeans = KMeans(n_clusters=3, random_state=0)
df['cluster'] = kmeans.fit_predict(X)
```

Fig. 3: Source code for Euclidean

Based on the code shown above, the formula used to calculate Euclidean distance is:

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Explanation:

- $\sum_i$: summation over index i
- x_i and y_i: the i-th components of each vector

Fig 4. Formula Euclidean

Table 3: Manual Calculations for the Five Student Samples

No	Full Name	Report Card Score	Attendance	Extracurricular Score
1	FADHIL	75	85	80
2	RAASYIKA	40	65	59
3	ALFARO	90	80	80
4	RHEVINKA	85	75	85
5	KAYLA	75	80	75

Table 4: Centroid Value Examples

centroid	y_1	y_2	y_3
m_1	50	65	70
m_2	75	75	75
m_3	85	85	85

Table 5: Results of Manual Euclidean Distance Calculation

No	Full Name	m1	m2	m3	Min m	(Min m)^2
1	FADHIL	33.541	11.180	11.180	11.180	125
2	RAASYIKA	14.866	39.762	55.687	14.866	221
3	ALFARO	43.875	16.583	8.660	8.660	75
4	RHEVINKA	39.370	14.142	10.000	10.000	100
5	KAYLA	29.580	5.000	15.000	5.000	25

Based on the calculation results from the table above, each individual data point has three distances to the centroids (m1, m2, and m3). The smallest value (Min m) among them is selected and then squared to obtain $(\text{Min m})^2$. From these calculations, the output clusters can be observed after the iteration process, as shown in the following data.

Table 6: Clusters

C1	: Raasyika
C2	: Kayla
C3	: Fadhil, Alfaro, Rhevinka

3.4. Dashboard

It displays student data including name, academic score, attendance, extracurricular score, cluster classification, and average score. The system categorizes students into the "Outstanding" cluster based on their performance, with average scores ranging above 91. The interface also provides features such as adding new data, uploading files, processing clustering, exporting data to PDF, and filtering clusters. Additionally, each student entry includes action buttons to edit or delete data, making the system efficient for managing and analyzing student achievement records.

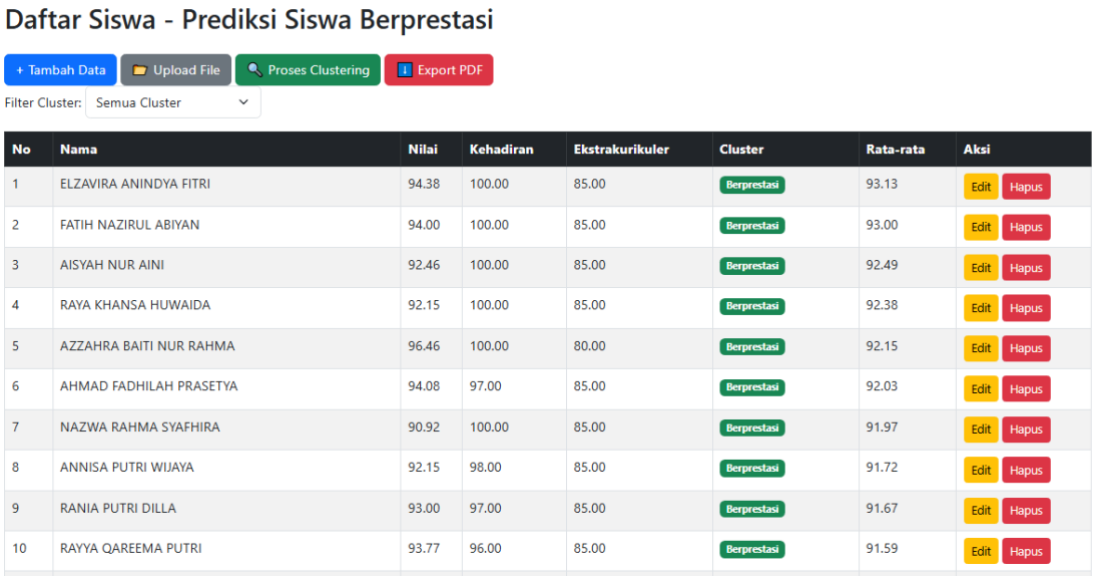

Fig. 5: Dashboard

4. Conclusion

After completing the design of the outstanding student prediction system at Madrasah Ibtidaiyah Irsyadul Athfal using the K-Means Clustering algorithm in a web-based application, the study concludes that the system was successfully developed to integrate both academic and non-academic data, including report card grades, attendance, and extracurricular participation. This system provides a more objective and comprehensive alternative compared to the conventional manual selection method. The implementation of K-Means Clustering effectively grouped students into clusters based on data patterns, making it easier for schools to identify top-performing students. Testing with eight respondents indicated that the application is user-friendly, provides relevant clustering accuracy, and supports more effective and efficient decision-making.

The research has successfully developed a real-time student prediction system using data from Madrasah Ibtidaiyah Irsyadul and applied the K-Means Clustering algorithm to predict outstanding students. However, the development and implementation process faced limitations due to time constraints. For future improvements, it is recommended to integrate additional methods such as classification or weighting techniques (e.g., SAW or TOPSIS) to provide more detailed ranking or recommendations. Further research could also include broader testing with more respondents and involving multiple madrasahs to evaluate the system's generalization and scalability.

References

[1] D. Puspita and S. Aminah, "Implementasi Naive Bayes Untuk Sistem Prediksi Mahasiswa Berprestasi," *Jurnal Ilmiah Teknosains*, vol. 8, no. 2, pp. 14–19, 2022.

[2] D. Aditiya and U. Latifa, "Uji Efektivitas Penerapan Machine Learning Classification Untuk Survey Kepuasan Pelanggan Maskapai Penerbangan X," *Barometer*, vol. 8, no. 1, pp. 9–18, 2023, doi: 10.35261/barometer.v8i1.6566.

[3] Y. Yudiana, A. Yulia Agustina, and dan Nur Khofifah, "Prediksi Customer Churn Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan," *IJIEB: Indonesian Journal of Islamic Economics and Business*, vol. 8, no. 1, pp. 1–20, 2023, [Online]. Available: http://e-journal.lp2m.uinjambi.ac.id/ojp/index.php/ijoieb

[4] Asmana, Y. A. Wijaya, and Martanto, "Clustering Data Calon Siswa Baru Menggunakan Metode K-Means Di Sekolah Menengah Kejuruan Wahidin Kota Cirebon," *Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 2, pp. 552–9, 2022, [Online]. Available: https://www.smkwahidincrb.sch.id/

[5] D. Apriandi, R. M. Sari, and M. I. Sarif, "Analisis Clustering Untuk Menentukan Siswa Berprestasi di SMK Swasta TI Panca Dharma Stungkit Menggunakan Metode K-Means," *Jurnal Minfo Polgan*, vol. 13, no. 1, pp. 1117–1129, 2024, doi: 10.33395/jmp.v13i1.13959.

[6] R. Putra Primanda, A. Alwi, and D. Mustikasari, "Data Mining Seleksi Siswa Berprestasi Untuk Menentukan Kelas Unggulan Menggunakan Metode K-Means Clustering (Studi Kasus Di MTS Darul Fikri)," *KOMPUTEK : Jurnal Teknik Universitas Muhammadiyah Ponorogo*, vol. 5, no. 1, pp. 88–101, 2021, [Online]. Available: http://studentjournal.umpo.ac.id/index.php/komputek

[7] V. Saputra Ginting and E. Taufiq Luthfi, "Penerapan Algoritma C4.5 Dalam Memprediksi Keterlambatan Pembayaran Uang Sekolah Menggunakan Python," *Jurnal Teknologi Informasi*, vol. 4, no. 1, pp. 1–6, 2020.

[8] R. Sovia, E. Praja, W. Mandala, S. Mardhiah, J. Raya, and L. Begalung, "Algoritma K-Means dalam Pemilihan Siswa Berprestasi dan Metode SAW untuk Prediksi Penerima Beasiswa Berprestasi," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 6, no. 2, pp. 181–187, 2020.

[9] S. Asyuti and A. A. Setyawan, "Data Mining Dalam Penggunaan Presensi Karyawan Denga Cluster Means," *Jurnal Ilmiah Sains Teknologi Dan Informasi*, vol. 1, no. 1, pp. 1–10, 2023.

[10] I. Maulana and U. Rosalina, "Clustering Data Nilai Ujian Akhir Semester Menggunakan Algoritma Data Mining K-Means," *PERISKOP (Jurnal Sains dan Ilmu Pendidikan)*, vol. 1, no. 2, pp. 76–85, 2021.

[11] F. Pramataning Dewi, P. Siwi Aryni, and Y. Umaidah, "Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 7, no. 2, pp. 111–121, 2011.

[12] R. Fardiani, C. Responden, and K. Kunci Pendidikan, "Penerapan Algoritma C4.5 Untuk Prediksi Penerima Beasiswa Siswa Berprestasi Di Mi Al-Ishlah Ciganitri," *JIKA (Jurnal Informatika) Universitas Muhammadiyah Tangerang P*, vol. 8, no. 4, pp. 402–410, 2024.

[13] N. Nyoman, P. Pinata, M. Sukarsa, N. Kadek, and D. Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *JURNAL ILMIAH MERPAT*, vol. 8, no. 3, pp. 188–196, 2020.

[14] I. M. Faiza and W. Andriani, "Tinjauan Pustaka Sistematis: Penerapan Metode Machine Learning Untuk Deteksi Bencana Banjir," *Jurnal Minfo Polgan*, vol. 1, no. 2, pp. 59–63, 2022, doi: 10.33299/jpkop.22.2.1752.

[15] S. Supardi *et al.*, "Peran Data Mining dalam Memprediksi Tingkat Penjualan Sepatu Adidas Menggunakan Metode Algoritma Regresi Linear Sederhana," *Jurnal Ekonomi Manajemen Sistem Informasi (JEMSI)*, vol. 4, no. 5, pp. 883–890, 2023, doi: 10.31933/jemsi.v4i5.