Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Implementation of Data Mining Using the K-Means Algorithm to Group Students Based on Academic Performance

Tatang Sujana[1], Rini Astuti[2], Willy Prihartono[3], Ryan Hamonangan[4]

[1,3,4]STMIK IKMI Cirebon
[2]STMIK LIKMI Bandung
Dusun Nagri RT 01 Rw 07, Des. Cibungur, Kec. Rancakalong, Kab Sumedang, Jawa Barat 45361
Tatangsujana710@yahoo.com[1], riniastuti@likmi.ac.id[2], willy@ikmi.ac.id[3], ryanhamonangan@gmail.com[4]*

Abstract

Data clustering is a critical technique in data mining that identifies patterns or groups within large datasets. This study applies the K-Means algorithm to cluster students from an Islamic boarding school based on their academic performance. The K-Means algorithm was chosen due to its ability to divide data into homogeneous clusters, facilitating a better understanding of academic characteristics for each group. Data from students' test scores—including written tests, oral exams, and classical Islamic book comprehension—were analyzed using Python. The analysis included data collection, preprocessing, determining the optimal number of clusters (K), implementing the K-Means algorithm, and validating clustering outcomes using the Davies-Bouldin Index (DBI). Results demonstrated that students could be grouped into ten clusters, with key insights to improve teaching strategies.
Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to interact with, identify useful information, and extract knowledge from various large databases.[1] Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to interact with, identify useful information, and extract knowledge from various large databases. [2] The purpose of this research is to group data of outstanding class students so that in the learning process at school, it is easier to facilitate education according to the students' abilities.[3]

Keywords: *Data Mining, K-Means, Clustering, Academic Performance, Islamic Boarding School*

1. Introduction

Education is something very important and a necessity that must be fulfilled by every individual, as each individual will eventually become a morally upright generation of the nation. School is a formal educational institution tasked with providing quality services.[4] Student achievement is an accomplishment as well as a benchmark and reference in the knowledge obtained from education at school, which includes formal assessments such as test scores, mastery of subjects, and attitudes determined through grades or numbers given by teachers to their students.[5] Student achievements are important to examine considering that student achievements can be used as a reference in future decision-making, among others as follows.[6]

The rapid advancement of information and communication technology has significantly impacted various aspects of life, including education. In Islamic boarding schools, large amounts of data are generated, which necessitate effective methods to process and analyze this data. Data mining, a branch of informatics, offers solutions for uncovering hidden patterns and insights within large datasets. Among various techniques, the K-Means algorithm is highly effective for clustering data into meaningful groups.[7] Datamining is the process of discovering new, useful correlations, patterns, and trends by mining large amounts of data repositories, using pattern recognition technologies such as statistics and mathematical techniques.[8]

In the K-Means Algorithm, there are several approaches to developing clustering methods, the two main approaches being the Partial Algorithm and the Hierarchical Algorithm.[9] Data mining is also known as Knowledge Discovery in Database (KDD) in the process of collecting information from a dataset. The KDD process is interactive and iterative, including several steps that involve users in decision-making and can be repeated in two steps.[10]

2. Methodology

The research methodology comprised the following steps:

1. Data Collection:
   The first stage in the KDD process is data selection. In the first step, we select a subset of the data that will be used for the clustering process. This data selection process is important to ensure that the data we select is relevant to the problem we want to solve. Here, we select certain columns related to relevant test scores for clustering, such as 'written_test', 'spoken_test', 'book_test'.

```
[6] # Drop non-numeric columns and scale the data for clustering
    selected_columns = ['tes_tulis', 'tes_lisan', 'tes_kitab']
    data = df[selected_columns]
```

Figure 1: Selection Columns Clustering

2. Data Pre-processing
   At this stage, we clean the data by removing missing values (NaN). This is very important because many algorithms, including K-Means, cannot handle missing data well. To overcome this, we delete rows that have NaN values to ensure the quality of the data used. Data cleaning is an important step to ensure the data used does not have problems that could interfere with the analysis results. By removing missing values, we ensure that the clustering algorithm can run smoothly.

```
# Menghapus nilai NaN
df_clean = df.dropna()

# Menampilkan informasi setelah pembersihan data
print(df_clean.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 232 entries, 0 to 231
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   nama_lengkap  232 non-null    object
 1   tes_tulis     232 non-null    int64
 2   tes_lisan     232 non-null    int64
 3   tes_kitab     232 non-null    int64
dtypes: int64(3), object(1)
memory usage: 7.4+ KB
None
```

Figure 2: Data Pre-processing

3. Data Transformation
   At this stage, we normalize the data using StandardScaler. This normalization process changes the data so that it has a mean of 0 and a standard deviation of 1. Normalization is very important in clustering, especially if the data has different scales, because the K-Means algorithm is sensitive to feature scale. This data transformation ensures that each feature has a balanced contribution to the distance calculation. Without normalization, larger scale features can dominate the clustering and influence the results.

```
[9] from sklearn.preprocessing import StandardScaler

    # Normalisasi data dengan StandardScaler
    scaler = StandardScaler()
    normalized_data = scaler.fit_transform(df_clean[selected_columns])
```

Figure 3: Data normalization with StandardScaler

4. Data Mining
   After carrying out the data transformation stage, the next stage is implementing data mining. At the stage of implementing data mining, the author uses Python tools on Google Colab with the k-means algorithm method. At this data mining stage the researcher carried out two stages, namely the first stage the author carried out a clustering process using the Kmeans algorithm with the operator used, namely the Clustering operator (K-Means) and the second stage the writer carried out testing and evaluating the results of clustering using the Cluster Distance Performance operator with the method used. The Davies Bouldin Index (DBI) evaluation was used.

Stages of K-Means Clustering

The first stage is that the author carries out a clustering process using K-Means for various numbers of clusters (from 2 to 14) and calculates the Davies-Bouldin Index (DBI) value for each experiment. Exploration *K* to 14 was carried out to determine the optimal number of clusters in the K-Means algorithm using the Davies-Bouldin Index (DBI) as an evaluation metric. The *K* range allows identification of the best cluster division, with the optimal *K* determined based on the lowest DBI value.

The maximum limit of *K*-14 was chosen to accommodate data complexity without risking overfitting, ensuring clustering results are more meaningful and representative of the data structure. A lower DBI indicates better clustering, and using this code, you can determine the optimal number of clusters by selecting the number of clusters that provides the lowest DBI. Data mining modeling for grouping students at the Kebon Kelapa Al-Ma'rifah Islamic boarding school using the K-means algorithm can be seen in Figure 4.

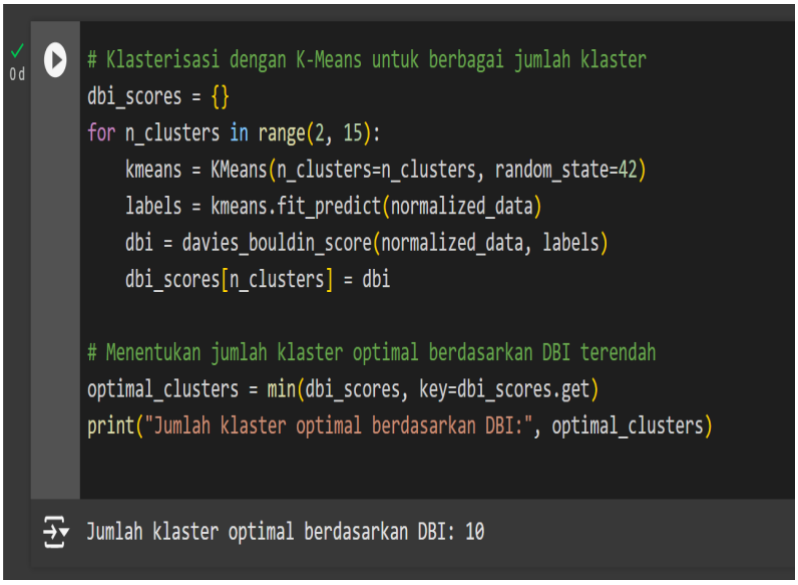

Figure 4: K-means algorithm

In Figure 4, it is explained that researchers carried out clustering using K-Means for various numbers of clusters (from 2 to 14) and calculated the Davies-Bouldin Index (DBI) value for each experiment. A lower DBI indicates better clustering, and the result is 10 clusters.

Evaluasi Davies Bouldin Index (DBI)

The second stage is testing the clustering results by applying the DBI value to the Cluster Distance Performance parameter. In DBI testing, the cluster that has the smallest DBI value or close to 0 is used as the best cluster. To find the smallest cluster, the author conducted experiments from cluster 2 to 15 recapitulations. The number of clusters resulting from the Davies Bouldin Index value can be seen from figure 5 below

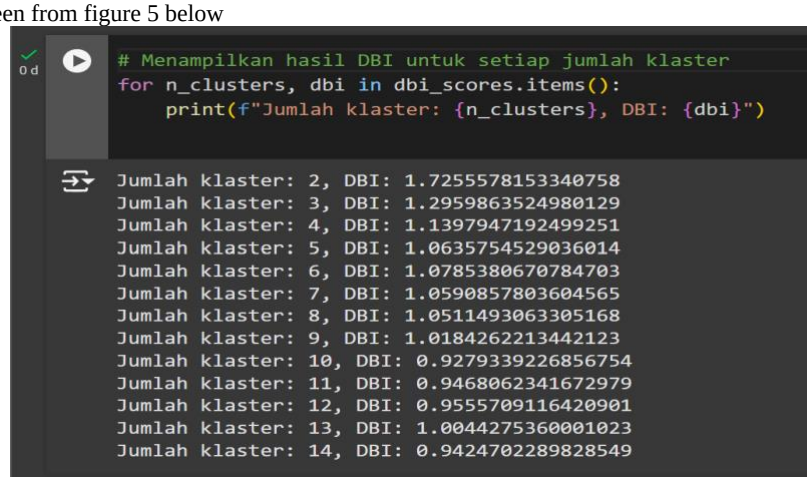

Figure 5: DBI results

In Figure 5 we can see the DBI results for each cluster, and in cluster 10 the DBI score is closest to 0 with a score of 0.9279339226856754.

5. Evalution:
   After the clustering model is applied, we use the Davies-Bouldin Index (DBI) to evaluate the clustering results. DBI is a metric that measures how well clusters are separated and how far they overlap with each other. Lower DBI values indicate better and more separated clusters. Model evaluation via DBI allows us to assess the quality of the resulting clusters. Using this metric helps us choose the optimal number of clusters, which will provide more meaningful cluster results and make it easier to understand the data. Based on table 4.5, it shows that the Tsanawi level student value data cluster using the Davies

Bouldin Index, the value closest to 0 with the cluster 2 to cluster 15 experiment produced the best K value in cluster 10, namely 0.9279339226856754 with the number of members in Cluster 0: 24 students. Cluster 1: 36 students. Cluster 2: 27 students Cluster 3: 51 students Cluster 4: 6 students Cluster 5: 14 students Cluster 6: 25 students Cluster 7: 18 students Cluster 8: 15 students Cluster 9: 16. It can also be seen in figure 6 below.

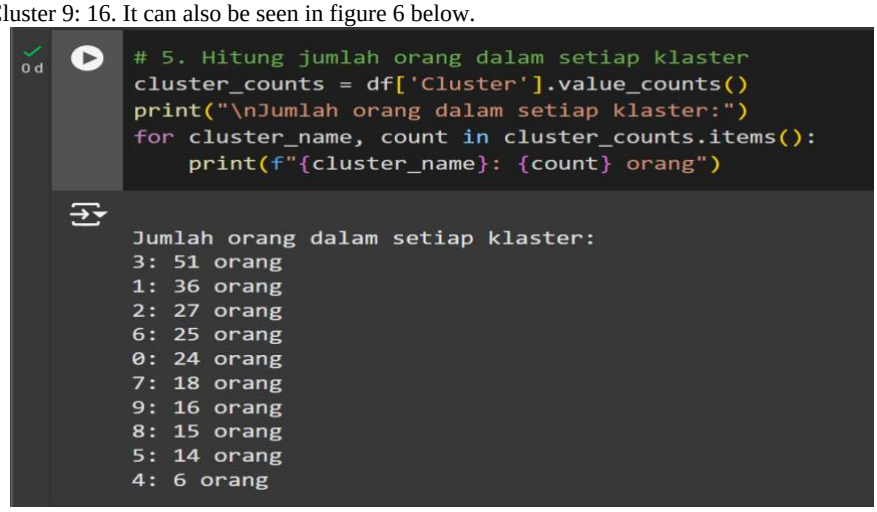

Figure 6: Students Cluster

Based on the evaluation results with the Silhouette Score and Davies-Bouldin Index (DBI) provided, the following is an explanation of the results:

1. Silhouette Score: 0.32

The value of 0.32 in this model indicates that the clustering carried out is quite good, but can still be improved. This value indicates that although much of the data is well grouped within its respective clusters, there is some data that may be in ambiguous areas, which means there is room for improvement in the selection of the number of clusters or the quality of the clusters.

2. Davies-Bouldin Index (DBI): 0.93

With a DBI of 0.93, this model shows fairly good cluster quality. In general, a DBI value close to 0 indicates an ideal cluster, but a value around 1 can still be considered quite good. In this case, even though it is not optimal, the model provides clusters that are separated quite clearly and do not overlap too much.

3. Results and Discussion

The study identified ten clusters based on student performance. These clusters provided insights into academic strengths and weaknesses:

- High Performers: Students with consistently high scores across all assessments.
- Low Performers: Students requiring additional support in specific areas.
- Balanced Performers: Students with moderate and consistent performance across subjects.

The clustering process was visualized using DBI values to determine the optimal number of clusters.

Figure 7: Davies-Bouldin Index (DBI) Analysis for Cluster Optimization

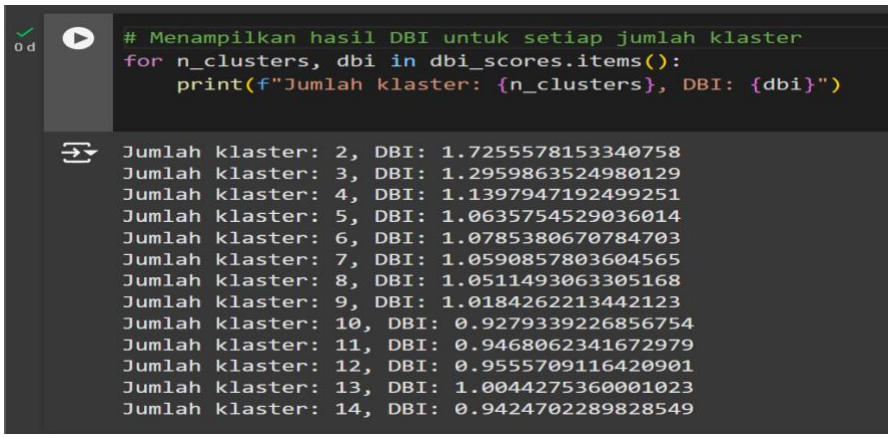

Figure 7: DBI Analysis

The cluster characteristics revealed targeted interventions for student groups, enhancing the educational experience. For instance, students in Cluster 5 displayed excellent performance in written tests but struggled with oral exams, highlighting the need for tailored communication skill programs.

```
Statistik rata-rata setiap cluster:
         tes_tulis  tes_lisan  tes_kitab
Cluster
0        80.208333  91.458333  85.416667
1        70.555556  81.944444  82.222222
2        78.148148  68.148148  86.296296
3        83.725490  79.215686  79.607843
4        65.000000  65.000000  65.000000
5        89.285714  65.714286  79.285714
6        85.400000  84.600000  67.600000
7        67.222222  81.388889  69.166667
8        92.000000  83.000000  86.333333
9        79.062500  69.375000  68.437500
```

Figure 8: Statistical Distribution of Test Scores Across Clusters

4. Conclusion

The implementation of the K-Means algorithm successfully grouped students into ten meaningful clusters based on academic performance. The findings enable targeted teaching strategies, addressing the specific needs of each group. Future research could expand the dataset to include demographic or behavioral variables for deeper analysis.

Acknowledgement

The author extends gratitude to STMIK IKMI Cirebon and the Islamic boarding school for providing the data and resources necessary for this research. Special thanks to academic supervisors and peers for their guidance and support.

References

[1] S. Natalia, B. Sembiring, H. Winata, and S. Kusnasari, "Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means," vol. 1, pp. 31–40, 2022.
[2] T. Widyanti and E. N. , Shofa Shofiah Hilabi, Agustia Hananto, Tukino, "Implementasi K-Means dan K-Nearest Neighbors pada Kategori Siswa Berprestasi," *J. Inf. dan Teknol.*, vol. 5, no. 1, pp. 75–82, 2023, doi: 10.37034/jidt.v5i1.255.
[3] J. Hutagalung, "Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 1, pp. 606–620, 2022, doi: 10.35957/jatisi.v9i1.1516.
[4] C. Satria and A. Anggrawan, "Aplikasi K-Means berbasis Web untuk Klasifikasi Kelas Unggulan," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 21, no. 1, pp. 111–124, 2021, doi: 10.30812/matrik.v21i1.1473.
[5] A. S. Muhammad Qusyairi1*, Zul Hidayatullah2, "Penerapan K-Means Clustering Dalam Pengelompokan Prestasi Siswa Dengan Optimasi Metode Elbow," vol. 7, no. 2, pp. 500–510, 2024.
[6] F. N. R. F. J. Aziz, B. D. Setiawan, and I. Arwani, "Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 6, pp. 2243–2251, 2018.
[7] R. Ishak and A. Bengnga, "Clustering Tingkat Pemahaman Mahasiswa Pada Perkuliahan Probabilitas Statistika Dengan Metode K-Means," *Jambura J. Electr. Electron. Eng.*, vol. 4, no. 1, pp. 65–69, 2022, doi: 10.37905/jjeee.v4i1.11997.
[8] F. Nasari and C. J. M. Sianturi, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat," *CogITo Smart J.*, vol. 2, no. 2, pp. 108–119, 2016, doi: 10.31154/cogito.v2i2.19.108-119.
[9] R. P. 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*, vol. 5, no. 1, p. 88, 2021, doi: 10.24269/jkt.v5i1.686.
[10] S. Haviyola, S. Susilawati, and M. Jajuli, "Pengelompokan Prestasi Siswa Guna Kualifikasi Beasiswa Berdasarkan Data Nilai Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2786–2791, 2024, doi: 10.36040/jati.v7i4.7200.