Indonesian Journal of Modern Science and Technology Modern (IJMST)
Vol. 1, No. 2, May 2025, pp. 50-57
E-ISSN: 3090-3831 | DOI: doi.org/10.64021/ijmst.1.2.50-57.2025

Data Analysis of Student Monitoring Using the K-Means Clustering Method

Sulistiani [1,*], Ahmad Rizky Nusantara Habibi [2], Adrian Maulana [3], Hidear Talirongan [4], Anrom G. Abao [5], Ahmed Mahmoud Zaki Elmalky [6], Asno Azzawagama Firdaus [7]
[1,2,3,7] Department of Computer Science, Universitas Qamarul Huda Badaruddin, Indonesia
[4] Misamis University, College of Computer Studies, Ozamiz City, Misamis Occidental, Philippines
[5] Eazy Pte Ltd, Singapore
[6] Quality Coordinator of Hospital Morbidity and Mortality Review Committee (Clinical Outcome Review and Improvement Committee), King Saud University, Saudi Arabia

ARTICLE INFO

Article history:

Received March 20, 2025
Revised April 19, 2025
Published May 10, 2025

Keywords:

Data Mining; Pre-processing; Student Monitoring; K-Means Clustering.

ABSTRACT

This study aims to group student monitoring data by focusing on two main variables, namely anxiety level and mood score, using the K-Means Clustering method. The research data was obtained from the Kaggle platform, which contains 1000 rows of data with nine attributes, including Student ID, Date, Class Time, Attendance Status, Stress Level, Sleep Hours, Anxiety Level, Mood Score, and Risk Level. The research process involved several stages, from problem identification, data collection, data cleaning and preprocessing, to the application of the K-Means algorithm. The analysis results showed that the data could be divided into two main groups: Cluster 1 consists of students with low to moderate anxiety levels and high mood scores, while Cluster 2 includes students with high anxiety and low mood scores. These findings provide relevant information for schools or campuses to design more effective psychological support and emotional monitoring programs. Additionally, this clustering method can serve as a foundation for developing an early detection system for psychological issues among students.

Corresponding Author:

Sulistiani, Department of Computer Science, Universitas Qamarul Huda Badaruddin, Indonesia
Email: tiani6902@gmail.com

1. INTRODUCTION

The process of extracting data into previously unreported information, with the right data mining processes and techniques, will yield optimal results [1]. Data mining is a series of processes in searching for patterns, relationships, and extracting added value from large amounts of data and information in the form of knowledge with the aim of finding relationships and simplifying data to obtain information that is easy to understand and useful [2]. One of the classification techniques based on consumer needs that can be used in data mining is the clustering method. Clustering is a method used in data mining that works by searching for data and grouping data that has similar characteristics between one data set and another that has been obtained [3].

One clustering method that can be used is the K-means method, because K-means is one of the algorithms in data mining that can be used to cluster data. In the clustering process, it generally attempts to minimize the variation within a group and maximize the variation between groups [4].

n this research discussion, the clustering method was chosen because it can group several data/objects into a cluster, so that each cluster contains similar data. Clustering produces similar objects that are close to each other in one cluster and produces the greatest possible distance between clusters [5]. This study chose the k-means algorithm because it is detailed, simple, and widely known. In this technique, objects are grouped into clusters or several groups.

Research on data clustering has been extensively developed by researchers. There are various machine learning methods that have been studied. One of them is in the research conducted by Wahyu Andi Prasyabudi et al. (2024), who used the K-means method to cluster 1000 student monitoring data points with 9 attributes and obtained results with a market share of around 42%, which is the main market segment for implementing strong and targeted marketing strategies [6]. This study applies K-means in clustering training data. The results of this study are clusters that form the data into two clusters, namely high and low. The data in these clusters can inspire users in determining high and low values [7].

The purpose of this study was conducted because there were still some missing or empty data in the monitoring data, so it had to be preprocessed first to be clustered using the K-means algorithm. Therefore, this study aimed to determine the monitoring data clusters based on anxiety level and mood score, using the K-means clustering algorithm method.

2. METHODS

In data grouping, there are several steps that can be taken, as follows:

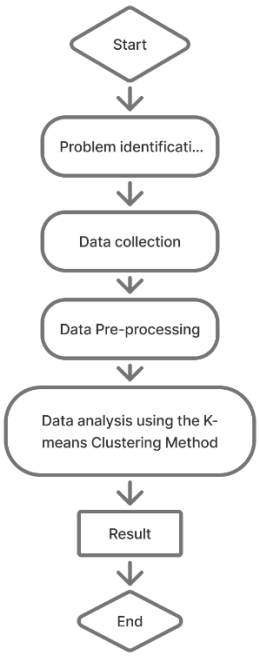

Figure 1. *Research stages*

2.1. Problem Identification

The initial stage of the research is to identify the problem. Problem identification aims to describe and explain the problem to be studied and develop it in the research object so that the background of this research can be identified with how to overcome Risk Level grouping based on Anxiety Level & Mood Score.

2.2. Data Collection

The data used in this study was obtained from Kaggle (https://www.kaggle.com), an open platform that provides various datasets for research and data analysis model development purposes. The dataset used is titled

"Student Health and Attendance Data." This dataset consists of 1000 data entries with 9 attributes, namely Student ID, Date, Class Time, Attendance Status, Stress Level (GSR), Sleep Hours, Anxiety Level, Mood Score, and Risk Level. The data used is secondary data, meaning it was not collected directly by the researcher but obtained from a reliable public source [8].
Here is an example of data obtained from Kaggle:

	Student ID	Date	Class Time	Attendance Status	Stress Level (GSR)	Sleep Hours	Anxiety Level	Mood Score	Risk Level
0	1	01/12/2024	9:00-15:00	Late	0.92	7.6	6.0	6.0	Low
1	1	02/12/2024	8:00-16:00	Late	1.17	6.0	NaN	2.0	Medium
2	1	03/12/2024	11:00-14:00	Late	4.56	6.3	4.0	8.0	High
3	1	04/12/2024	11:00-16:00	Late	3.07	9.0	2.0	10.0	Low
4	1	05/12/2024	9:00-13:00	Absent	3.93	7.4	9.0	4.0	High

Figure 2. Data Student Monitoring

From this data, there are some missing or incomplete data. Before further analysis, this dataset must first be processed through preprocessing stages to ensure data validity. This process includes handling missing values [9].

2.3. Data Pre-processing

Data preprocessing is the first step in the data mining process, which aims to convert raw data into a more efficient and informative format. Raw data from various sources generally contains errors, missing values, and inconsistencies. Therefore, the process of cleaning and adjusting the data format is necessary to ensure that the results of data mining are more accurate and reliable [10]. Effective data preprocessing plays a crucial role in shaping data mining outcomes, enhancing the accuracy and clarity of newly generated insights. This step involves the process of data cleaning [11]. Data cleaning is a process used to remove inconsistent and noisy data from various databases that may have different formats or platforms, which are then integrated into a single database [12]. The reason data must be cleaned is because there is incorrect or incomplete data, certain columns are missing, or there is a lot of missing data [13].
Here are the results of the data analysis after preprocessing the data, such as removing missing values:

	Student ID	Date	Class Time	Attendance Status	Stress Level (GSR)	Sleep Hours	Anxiety Level	Mood Score	Risk Level
0	1	01/12/2024	9:00-15:00	Late	0.92	7.6	6.0	6.0	Low
1	1	03/12/2024	11:00-14:00	Late	4.56	6.3	4.0	8.0	High
2	1	04/12/2024	11:00-16:00	Late	3.07	9.0	2.0	10.0	Low
3	1	05/12/2024	9:00-13:00	Absent	3.93	7.4	9.0	4.0	High
4	1	06/12/2024	8:00-14:00	Present	4.96	6.6	5.0	9.0	High

Figure 3. Result cleaning data

2.4. K-Means Clustering

K-means clustering [14] is a widely used unsupervised machine learning technique that is a data partitioning method that assigns observations into different groups or clusters based on their similarities. This technique has been widely used in various fields, including data analysis, image processing, and bioinformatics [15].
The K-means algorithm operates based on the principle of iterative partitioning [16]. This algorithm begins by randomly selecting K initial cluster centers, where K represents the number of clusters that has been determined beforehand. Next, data points are assigned to the cluster whose center is closest, which is usually measured using Euclidean distance. The centers are then recalculated as the average of the points in the cluster, and this process is repeated until convergence is achieved.
Essentially, the K-means algorithm minimizes the sum of squared distances between data points and each cluster centroid. The objective function can be expressed as:

$$J = \sum_{i}^{k} = 1 \sum_{j}^{n} = 1 \left| X_j^{(i)} - \mu i \right|^2 \quad (1)$$

Where J is the objective function, K is the number of clusters, ni is the number of data points in *cluster* i, $X_j^{(i)}$ represent j data points in cluster i, μi is the center of cluster i. The K-Means algorithm is a method used to cluster data by dividing it into several groups, where data with similarities are placed in the same group, while data with differences are placed in different groups [3].
The steps of the k-means algorithm are:

- Determine the number of *cluster*s.
- Initialize the initial *centroid* values of each *cluster* randomly.
- Calculate the distance of each data point to the *cluster* based on the closest distance to the *cluster* center.

- For each *cluster*, determine the new *centroid* value based on the mean of each data point in the *cluster*.

Repeat steps 2 and 3 until the *centroid* value equals the average of the items in the *cluster*. The distance calculation uses Eucledian distance to calculate the distanxe between points. The flow og the K-means algorithm is shown in figure 4.

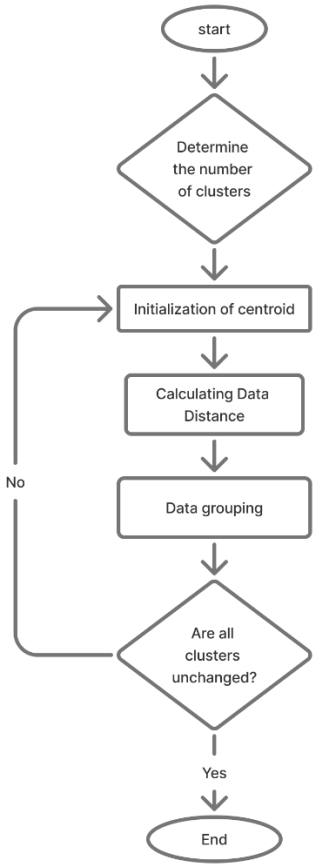

Figure 4. *Flowchart K-Means*

Figure 4 shows the K-means flowchart, which begins with determining the number of *cluster*s. Once the value is. Determined, the next step is to determine the *cluster* centers, followed by calculating the distance of each object from each *cluster* center. The distance between each data point and each *centroid* is calculated using Euclidean distance until the shortest distance between each data point and *centroid* is determined. Next, the objects are grouped based on the minimum distance to the *cluster* center*[17]*. The center of the *cluster* is then temporarily designated as the *cluster* center, or *centroid*. If there are still objects that need to be moved to other *cluster*s, the process is repeated, but if not, the process is complete[18].

Each *cluster* formed will improve the partition criteria, such as the difference function based on distance, so that objects within a *cluster* become similar, and objects in different *cluster*s are found to be dissimilar in terms of dataset attributes. Euclidean distance is used as a measure of distance in the K-means approach to highlight the similarity between each *cluster* with the smallest distance and the highest similarity. The Euclidean distance between point α=(α1,α2,…,αx) and point "b = (b1,b2,…,bn)" can be calculated using Formula (2).

$$d(b_i, a_t) = \sqrt{\sum_{j=1}^{l}(b_{ij} - a_{tj})^2} \quad (2)$$

Where:

d= distance between data value and cluster value

bi = data value, i = 1, 2, ….n,n = number of data

αt = cluster center value, t=1,2 …… K, K = number of clusters

l = number of attributes of dimensions

The K-means algorithm is as follows:

a. Determine the number of *clusters* to be created, for example K= 2.

b. Create the initial *centroid* or *cluster* center point, fot example C1: (5, 4), *centroid* C2: (1, 8).

c. Uding the Eucledian distance between two objects, calculate the distance of each data point *bi* to each *centroid* a_t. Suppose the data used is a student named Andi (A) who has a score (9, 2), then the distance ot the data from each *centroid* is:

d. $d(A, C1) = \sqrt{(9-5)^2 + (2-4)^2}$

$d(A, C1) = \sqrt{20 = 4{,}47}$

$d(A, C2) = \sqrt{(9-1)^2 + (2-8)^2}$

$d(A, C1) = \sqrt{100 = 10}$

e. Sort the data into groups based on the shortest distance between the data and each centroid [19]. Table 1 shows the results of calculating the distance between each data point and the *centroid* value.

Table 1. Determining data membership to *cluster*s

Name	Calculation of data distance to *centroid*		Closest distance	
Andi	C1	C2	C1	C2
	4, 47	10	√	

f. Calculate the average value of the data in the same *cluster* to obtain the new *centroid* position. After all students have been calculated in terms of their distance to all *centroid*s and their *cluster* membership has been determined, the new *centroid* position is calculated[20]. Table 2 shows the results of data updating/processing using the above *centroid* values by calculating the average value in each *cluster*.

Table 2. *new centroid value*

Cluster 1	6,64	4,23
Cluster 2	3,25	7,36

g. If the new *centroid* position differs from the previous *centroid*, return to step c. if not iteration is complete.

3. RESULTS AND DISCUSSION

Based on the results of the analysis using the K-means method, the optimal number of *cluster*s (K) is 2. After applying the K-Means algorithm to the processed data, two main groups of students were obtained based on their psychological conditions and behavior. [21]. The results of student monitoring data processing using the K-means *cluster*ing method and Visual Studio Code software assistance, with a total of 1000 student monitoring data points, were performed nine times so that the *cluster* results remained unchanged. This can be seen in Figure 5.

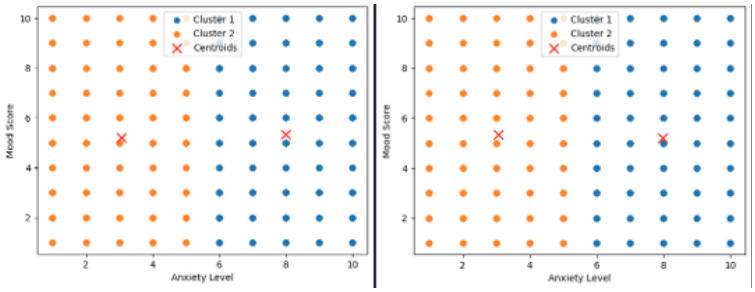

Figure 5 . *Cluster*ing results using K-means

Student monitoring data processing using K-means *cluster*ing with 1000 data samples and 2 *cluster*s, defining the K value as *cluster*_1 = 495 items and *cluster*_2 = 504 items, as shown in Figure 2.

```
Selesai !!
Total Iterasi: 9
Waktu eksekusi: 2.10 detik
Cluster 1: 495
Cluster 2: 504
```

Figure 6. *Total number of clusters*

The following (Figure 6) is the result of K-means *cluster*ing modeling showing the *cluster*_1 and *cluster*_2 graphs from a sample of 1000 data points using Visual Studio Code.

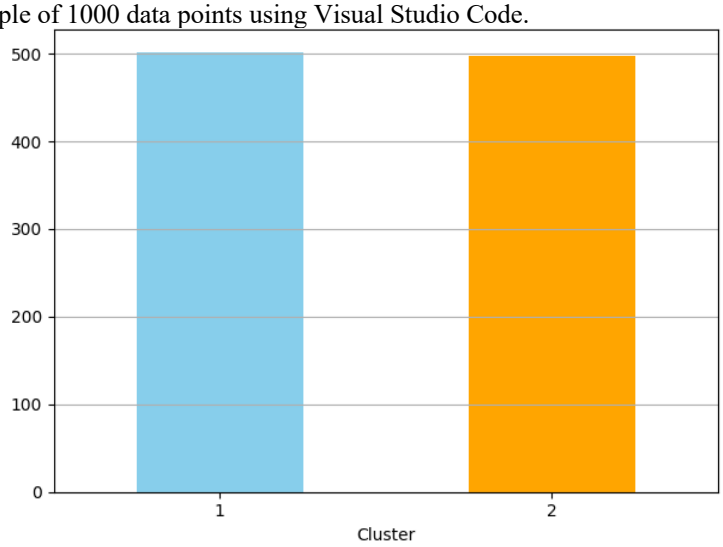

Figure 7. *Graph showing the result for each cluster*

The modeling results show that student monitoring data can be grouped into two *cluster*s based on the main attributes of Anxiety Level and Mood Score [22]. *Cluster*_1 is marked in blue, representing students with low to moderate anxiety levels and high mood scores. This group shows a tendency toward stable psychological conditions. *Cluster*_2 is marked in orange, representing students with high anxiety levels and low mood scores, indicating potential emotional stress and requiring further attention [23].

The following table shows the grouping based on the proximity of the *centroid* to each student data based on these two attributes.

Table 3. *Cluster analysis results*

Cluster	Number of students	General Characteristics
Cluster 1	495 students	Anxiety Level low-currently, Mood Score tinggi
Cluster 2	504 students	Anxiety Level high, Mood Score low

From the *cluster*ing results, most students are in a relatively stable psychological condition. However, around 37% of students fall into the risk category based on their anxiety levels and low mood scores [24].

These results can be used by academics to follow up with preventive counseling, psychological needs mapping, and activity planning that can help improve students' emotional well-being. With this data analysis, educational institutions can develop more targeted psychological support strategies and build a more effective periodic monitoring system [25].

4. CONCLUSION

This study proves that the K-Means *Cluster*ing algorithm can be effectively applied to group student monitoring data based on two main indicators, namely Anxiety Level and Mood Score. From the analysis results, the ideal number of *cluster*s was determined to be two groups. The first *cluster* includes students with low to moderate anxiety levels and relatively high mood scores, while the second *cluster* consists of students with high anxiety levels and low mood scores. This division provides valuable insights into the psychological condition of students, which can be used by the campus or academic authorities as a basis for designing counseling programs, psychological interventions, and more relevant and data-driven emotional support policies. While this study makes an important contribution to the context of monitoring students' mental health, there are several limitations that need to be considered. This study only utilized two main variables, thus not considering other factors such as stress level, sleep hours, and attendance status, which may also influence students' psychological conditions. This study also did not include an in-depth evaluation of the quality of the

cluster results using measures such as the Silhouette Score or Davies-Bouldin Index, so the accuracy of the *cluster* division cannot be quantitatively confirmed. As input for further research, the following points can be used as a reference: 1) Adding other relevant attributes to make the *cluster*ing results more in-depth and comprehensive, 2) Using *cluster*ing evaluation methods to measure the quality of the *cluster*s formed, 3) Comparing K-Means with other *cluster*ing algorithms such as DBSCAN, Hierarchical *Cluster*ing, or Gaussian Mixture Model to identify the most suitable approach, 4) Implementing real-time data monitoring or developing a technology based monitoring application system to enable continuous and adaptive analysis processes.

REFERENCES

[1] A. Koyalil and S. Rajalingam, "Enhanced Multi-level K-means Clustering and Cluster Head Selection Using a Modernized Pufferfish Optimization Algorithm for Lifetime Maximization in Wireless Sensor Networks," Results in Engineering, p. 105836, Jun. 2025, doi: 10.1016/j.rineng.2025.105836.

[2] R. K. Dinata, S. Safwandi, N. Hasdyna, and N. Azizah, "Analisis K-Means Clustering pada Data Sepeda Motor," INFORMAL: Informatics Journal, vol. 5, no. 1, p. 10, Apr. 2020, doi: 10.19184/isj.v5i1.17071.

[3] A. Al Fahrozi, F. Insani, E. Budianita, and I. Afrianty, "Implementasi Algoritma K-Means Clustering dalam Menentukan di Badan Pelatihan KesehatanPekanbaru," vol. 1, pp. 474–492, 2023.

[4] M. Sadeghi, P. Casey, E. J. M. Carranza, and E. P. Lynch, "Principal components analysis and K-means clustering of till geochemical data: Mapping and targeting of prospective areas for lithium exploration in Västernorrland Region, Sweden," Ore Geol Rev, vol. 167, p. 106002, Apr. 2024, doi: 10.1016/j.oregeorev.2024.106002.

[5] I. Nuryani and D. Darwis, "ANALISIS CLUSTERING PADA PENGGUNA BRAND HP MENGGUNAKAN METODE K-MEANS," 2021.

[6] W. A. Prastyabudi, A. N. Alifah, and A. Nurdin, "Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering," Procedia Comput Sci, vol. 234, pp. 96–105, 2024, doi: 10.1016/j.procs.2024.02.156.

[7] M. Balcilar, A. H. Elsayed, and S. Hammoudeh, "Financial connectedness and risk transmission among MENA countries: Evidence from connectedness network and clustering analysis," Journal of International Financial Markets, Institutions and Money, vol. 82, p. 101656, Jan. 2023, doi: 10.1016/j.intfin.2022.101656.

[8] T. Liddell, A. S. Boser, S. Orofino, T. Mangin, and T. Carleton, "stagg:: A data pre-processing R package for climate impacts analysis," Environmental Modelling & Software, vol. 183, p. 106202, Jan. 2025, doi: 10.1016/j.envsoft.2024.106202.

[9] S. Wang et al., "Advances in Data Preprocessing for Biomedical Data Fusion: An Overview of the Methods, Challenges, and Prospects," Information Fusion, vol. 76, pp. 376–421, Dec. 2021, doi: 10.1016/j.inffus.2021.07.001.

[10] S. Khalighi, L. Ma, S. Ren, and A. Varveri, "Evaluating the impact of data pre-processing methods on classification of ATR-FTIR spectra of bituminous binders," Fuel, vol. 376, p. 132701, Nov. 2024, doi: 10.1016/j.fuel.2024.132701.

[11] W. Cardoso, J. V. Roque, J. J. Jansen, S. Y. Teng, and R. F. Teófilo, "Combinatorial Order Pre-processing Search (COPS): A new pre-processing strategy for large-scale interpretable data analysis in process analytical technologies," Comput Chem Eng, vol. 192, p. 108892, Jan. 2025, doi: 10.1016/j.compchemeng.2024.108892.

[12] R. Hasanah et al., "Play Store Data Scrapping and Preprocessing done as Sentiment Analysis Material," Indonesian Journal of Modern Science and Technology, vol. 1, no. 1, pp. 16–21, Jan. 2025, doi: 10.64021/ijmst.1.1.16-21.2025.

[13] S. Selvakumaran et al., "Improving operational use of post-disaster damage assessment for Urban Search and Rescue by integrated graph-based multimodal remote sensing data analysis," Progress in Disaster Science, vol. 25, p. 100404, Jan. 2025, doi: 10.1016/j.pdisas.2025.100404.

[14] F. Liu et al., "Use of latent profile analysis and k-means clustering to identify student anxiety profiles," BMC Psychiatry, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12888-021-03648-7.

[15] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[16] E.-M. Papia and A. Kondi, "Quantifying subtle color transitions in Mark Rothko's abstract paintings through K-means clustering and Delta E analysis," J Cult Herit, vol. 72, pp. 194–204, Mar. 2025, doi: 10.1016/j.culher.2025.02.005.

[17] W. Song, M. Zhao, and J. Yu, "Price distortion on market resource allocation efficiency: A DID analysis based on national-level big data comprehensive pilot zones," International Review of Economics & Finance, p. 104128, Apr. 2025, doi: 10.1016/j.iref.2025.104128.

[18] C. Gilga et al., "Legal and ethical considerations for demand-driven data collection and AI-based analysis in flood response," International Journal of Disaster Risk Reduction, vol. 122, p. 105441, May 2025, doi: 10.1016/j.ijdrr.2025.105441.

[19] E. A. Saputra and Y. Nataliani, "Analisis Pengelompokan Data Nilai Siswa untuk Menentukan Siswa Berprestasi Menggunakan Metode Clustering K-Means," Journal of Information Systems and Informatics, vol. 3, no. 3, 2021, [Online]. Available: http://journal-isi.org/index.php/isi

[20] A. L. Maukar, F. Marisa, A. A. Widodo, N. Kamilaningtyas, D. Novian, and D. Nugraha, "ANALISIS DATA PENERIMAAN MAHASISWA BARU BERBASIS K-MEANS," Jurnal Informatika dan Komputer), vol. 6, no. 2, pp. 142–147, 2022.

[21] M. Abidin et al., "Classification of Heart (Cardiovascular) Disease using the SVM Method," Indonesian Journal of Modern Science and Technology, vol. 1, no. 1, pp. 9–15, Jan. 2025, doi: 10.64021/ijmst.1.1.9-15.2025.
[22] L. Sans, I. Vallvé, J. Teixidó, J. M. Picas, J. Martínez-Roldán, and J. Pascual, "La era del big data: análisis del lenguaje natural mediante la aplicación de folksonomía," Nefrología, vol. 42, no. 6, pp. 680–687, Nov. 2022, doi: 10.1016/j.nefro.2021.09.006.
[23] A. Restrepo Román, D. J. Villegas, C. Rodriguez, A. Cogollo, I. D. Bedoya, and A. A. Amell Arrieta, "Implementation of a hierarchical cluster model to analyze wind and solar availability in the department of Antioquia, Colombia," Case Studies in Chemical and Environmental Engineering, vol. 10, p. 101006, Dec. 2024, doi: 10.1016/j.cscee.2024.101006.
[24] A. R. Nusantara Habibi et al., "Diabetes Mellitus Disease Analysis using Support Vector Machines and K-Nearest Neighbor Methods," Indonesian Journal of Modern Science and Technology, vol. 1, no. 1, pp. 22–27, Jan. 2025, doi: 10.64021/ijmst.1.1.22-27.2025.
[25] A. Maulana et al., "Classification of Stunting in Toddlers using Naive Bayes Method and Decision Tree," Indonesian Journal of Modern Science and Technology, vol. 1, no. 1, pp. 28–33, Jan. 2025, doi: 10.64021/ijmst.1.1.28-33.2025.