Application of Data Mining to Measure Student Intelligence Level Using The K-Means Method

Lukman Hakim Ardiansyah[a1], Zaehol Fatah[a2]
[a]Information Systems Study Program, Faculty of Science and Technology, Ibrahimy University, East Java, Indonesia - 68374
e-mail: [1]lukmannhakimm1900@gmail.com, [2]zaeholfatah@gmail.com

Abstrak

Kecerdasan merupakan kemampuan individu untuk memahami, belajar, dan berpikir untuk memecahkan masalah yang kompleks. Kecerdasan manusia meliputi kecerdasan intelektual, emosional, spiritual, dan multiple intelligences. Pada dasarnya, seseorang memiliki kecerdasan yang berbeda-beda dalam banyak bidang disiplin ilmu. Penelitian ini berfokus pada pengukuran tingkat kecerdasan intelektual siswa di Sekolah Takhassus Abu Hurairah Sukorejo menggunakan metode K-Means Clustering yang diimplementasikan menggunakan aplikasi Rapidminer. Metode K-Means dipilih karena prosesnya sederhana dan relatif mudah diterapkan pada dataset yang besar. Siswa dikelompokkan menjadi tiga cluster berdasarkan tingkat kecerdasan berdasarkan jarak terdekat dengan centroid. Hasil penelitian menunjukkan bahwa cluster dengan jarak centroid terdekat memiliki karakteristik kecerdasan tertinggi, sedangkan cluster dengan jarak centroid terjauh menunjukkan kecerdasan yang lebih rendah. Dengan nilai Davies-Bouldin sebesar 0,599 menunjukkan bahwa pengelompokan yang diterapkan cukup efektif dan optimal. Hasil penelitian ini dapat menjadi acuan bagi sekolah untuk menentukan strategi pembelajaran yang tepat bagi setiap kelompok siswa sehingga dapat meningkatkan efektivitas pembelajaran di sekolah.

Kata kunci: *Clustering, Data Mining, Kecerdasan, K-Means, Rapidminer*

Abstract

Intelligence is an individual's ability to understand, learn, and think to solve complex problems. Human intelligence includes intellectual, emotional, spiritual, and multiple i intelligences. asically, a person has different intelligence in many fields of discipline. This study focuses on measuring the level of intellectual intelligence of students at Takhassus Abu Hurairah Sukorejo School using the K-Means Clustering method implemented using the Rapidminer application. The K-Means method was chosen because the process is simple and relatively easy to apply to large datasets. Students are grouped into three clusters based on intelligence levels based on the closest distance to the centroid. The results show that the cluster with the closest centroid distance has the highest intelligence characteristics, while the cluster with the furthest centroid distance shows lower intelligence. With a Davies-Bouldin value of 0.599, it shows that the grouping applied is quite effective and optimal. The results of this study can be a reference for schools to determine the right learning strategy for each group of stustudents that it can increase the effectiveness of learning in schools.

Keywords : *Clustering, Data Mining, Intelligence, K-Means, RapidMiner*

1. Introduction

The rapid development of information technology in the current era has brought major changes in various fields, including education. The education process begins when humans are still in kindergarten until they finally graduate from high school or college. The factor that most influences students at the level of education is the level of intelligence, often referred to as the intelligence quotient (IQ). Intelligence is the excellence or perfection of the development of reason, such as cleverness, accuracy, and sharpness of mind. In English, two terms are used that have the same meaning, namely intelligence and quotient. The first term, for example, is used in a combination of emotional intelligence or emotional intelligence. The second, for example, is used in a combination of adversity quotient or intelligence of resilience, tenacity, toughness, or intelligence in facing challenges [1].

Grouping students based on IQ/intelligence levels is very necessary in order to find out the right learning methods for students. Students with lower IQ/intelligence levels are different from students with high IQ levels; most of them need more attention in order to follow and understand the lessons well. Traditionally, measuring student intelligence is usually done by adding up all the midterm and final semester exam scores using a manual calculator, or what is commonly known as a calculator. The highest or largest score is considered a student with a high IQ and vice versa, but this method of measurement is prone to errors, such as entering the wrong number of scores, calculation errors, and so on. Therefore, a more comprehensive approach is needed to identify the potential intelligence and strengths of each student. Data generated from the learning process, such as exam results, psychological test results, and student participation in academic work, can now be analyzed and processed in more detail using data mining techniques.

Data mining is a process of extracting valuable knowledge or information from large and complex datasets [2]. In other words, data mining is a method that allows users to access large amounts of data in a relatively short time. Or, in other words, data mining is a tool and application that uses statistical analysis on data through a process of extracting or mining previously unknown data and information. Simply put, data mining is a process of mining data that leads to the discovery of the latest information by searching for certain patterns or rules from a very large amount of data, so that the way data mining actually works is to examine large databases to find new patterns or forms that are useful in the decision-making process [3]. The author uses the K-Means Clustering algorithm data mining method in grouping existing student data. Simply put, K-means clustering works by dividing data into several clusters or groups depending on the similarity of the feature patterns used. This process makes it possible to identify groups of students with different support priorities, such as students with high and low needs [4].

Several previous studies have shown that data mining can be used to improve the learning process and student assessment. Research conducted by students at Universitas Puter Batam found that the use of clustering techniques can group students based on their IQ scores using the Rapidminer application, so that teachers who teach can provide appropriate learning according to the IQ level of each student [5]. In addition, there is also research conducted by students of the STMIK Triguna Dharma

information system who also succeeded in grouping new student data for the 2022/2023 academic year using the K-Means algorithm, which produced two clusters. The first cluster has 47 students, while the second cluster has 23 students. Which results can be used to assist in making decisions about school promotion strategies [6]. From this study, it is proven that the clustering method with the K-Means algorithm is effective for grouping data so that it can find information or patterns that are useful for analysts.

This study aims to apply data mining with the K-Means Clustering Algorithm Method in measuring the level of intelligence of students at Takhassus Abu Hurairah Sukorejo School using the Rapidminer application. The data used in this study is the data of students' final exam scores. The purpose of applying data mining using the K-Means Clustering algorithm in measuring the level of intelligence of students is to identify patterns that may be hidden in the learning outcome data, then to divide students into groups based on the level of intelligence and learning achievement of each student, and provide valuable insights to teachers in educational institutions. It is hoped that the results of this study can help schools or institutions in making policies and taking appropriate steps to improve the quality of education in schools.

2. Research Method

This research focuses on the application of the clustering method with the K-Means algorithm in grouping student intelligence. The following are the stages used in this research:

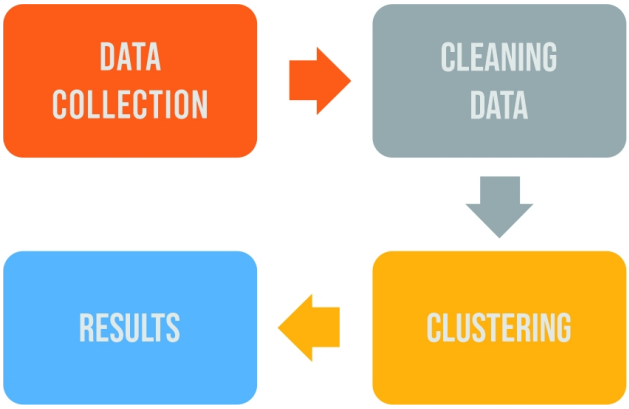

Figure 1. Research Stages

This stage begins with the collection of relevant data, then continues with the data cleaning stage, where the data is cleaned, integrated, and prepared for further analysis. Next, the data modeling stage is carried out, where the K-means clustering data mining algorithm technique is applied to identify hidden patterns, trends, and relationships in the data. Then, the results of the modeling will be evaluated and validated to ensure their accuracy and validity.

3. Literature Study

3.1. Data collection

An important initial step in developing a comprehensive understanding of students' learning progress and achievement throughout the learning process is to collect data on their IQ. Collected data for analysis is the first step. The data for this study came from the final exams of the 3rd batch of Takhassus Abu Hurairah school for the 2023–2024 academic year. This step is important because collecting high-quality

data on students' IQ is used to communicate not only individual student progress but also the efficacy of the process, learning methods, and educational program as a whole.

3.2. Cleaning Data

Data cleaning is the process of detecting and correcting (or removing) corrupted or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant data and then replacing, modifying, or deleting the corrupted data. Data cleaning involves identifying and correcting errors in data. Data cleaning issues involve incomplete (or missing) data being replaced by inserting values, duplicate records being removed, and inconsistent data values being corrected. Data cleaning is important because it ensures that data is accurate and complete [7].

3.3. Clustering

Clustering is a method of grouping data. According to Tan, clustering is a process of grouping data into several clusters or groups so that data in one cluster has a maximum level of similarity and data between clusters has a minimum similarity. Clustering is the process of partitioning a set of data objects into subsets called clusters. Objects in a cluster have similar characteristics between each other and are different from other clusters. Partitioning is not done manually but with a clustering algorithm. Therefore, clustering is very useful and can find unknown groups or groups in data [8]. The main purpose of clustering is to find hidden structures in unstructured or semi-structured data. By grouping data based on similar features, clustering helps identify patterns and relationships that are not visible to casual observation. This is especially important in big data, where the volume and complexity of data exceed the capabilities of manual analysis [9].

3.4. K-Means

In this study, the clustering algorithm method used is the K-Means algorithm. K-Means was first published by Stuart Lloyd in 1984 and is a widely used clustering algorithm. According to MacQueen J.B., K-Means is the most famous and widely used clustering method in various fields because it is simple, easy to implement, and has the ability to cluster large data. K-Means is a partitioning clustering method that separates data into different groups. With iterative partitioning, K-Means is able to minimize the average distance of each data point to its cluster [10].

Some of the advantages of the K-Means algorithm over other algorithms in the clustering method are as follows:

1. Very easy to understand and implement.
2. If we have many variables, K-means will be faster than hierarchical clustering.
3. In recalculating the centroid, an instance can change the cluster.
4. Tighter clusters are formed with K-means compared to hierarchical clustering [11].

3.5. Davies Bouldin

The Davies-Bouldin Index is one of the metrics used to evaluate the quality of cluster analysis in clustering analysis. The Davies-Bouldin Index was introduced by researchers David L. Davies and Donald W. Bouldin. The Davies-Bouldin Index aims to measure the extent to which each cluster is clearly separated and has internal consistency. This index assesses how well the clusters are formed by calculating the average similarity between clusters. The lower the Davies-Bouldin Index value, the better the clustering quality, because it shows clusters that are more separated and have less overlap [12].

3.6. Rapidminer

The tools or tools used to simplify the calculations or clustering used are the Rapidminer application. Rapidminer was previously known as YALE (Yet Another Learning Environment) and was developed in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer from the artificial intelligence unit of the Technical University of Dortmund. Rapidminer is open-source software. Rapidminer is a solution for analyzing data mining, text mining, and predictive analysis. Rapidminer has approximately 500 data mining operators, including operators for input, output, data processing, and visualization. Rapidminer is written using the Java language so that it can work on all operating systems [13].

4. Result and Discussion

4.1. Data Collection

The first step is to collect data. Regarding the data that will be used for analysis, it is important to note that the data comes from open sources. The data used is data from the Takhassus Abu Hurairah school data archive and is the final exam data for Takhassus Abu Hurairah Sukorejo school students, class 3, 2023/2024 academic year. The data includes several data attributes, namely: Name, Read Yellow Book, Read Al-Quran, Imla ', Taqrirotul Mawaddah, Final Exam, Attendance, Fiqh , Morals, and Activeness. An example of the data used can be seen in Figure 2 below.

No	Nama	Nilai Baca Kitab	Nilai Al-qur'an	Imla'	Nilai Taqrirotul Mawaddah	Nilai Ujian Akhir Sanah	Kehadiran	Praktek Fiqh	Akhlaq	Keaktifan	Jumlah
1	Khairul Maulana	94	54	65	80	60	80	80	80	80	673
2	Muhammad Aflah Mubarok	73	75	80	90	97	85	80	80	80	740
3	Ahmad Salman rafa	60	60	60	80	60	80	80	80	80	640
4	Azkar Junlansah	60	60	60	80	91	80	80	80	80	671
5	Rifky Alfian Nuri	60	60	60	80	60	80	80	80	80	640
6	Nurwan Hadi Medal	60	60	60	80	60	80	80	80	80	640
7	Maelky Zainullah	69	85	85	80	60	80	80	85	80	704
8	Muhammad Aidil Anwar Aisy	80	60	75	80	100	85	80	80	80	720
9	Alfan Fadlan	60	60	60	80	60	85	80	80	80	645
10	Moch Sultan Maghrobi	60	75	60	80	60	90	90	80	75	670
11	Ahmad Baihaqi	60	60	60	80	60	85	80	80	80	645
12	Moch Herullah	60	85	75	80	60	80	90	90	80	700

Figure 2. Final Exam Result Data

4.2. Cleaning Data

Cleaning data, or data cleaning, is an important stage in the data mining process that involves preparing and cleaning raw data before further analysis. This stage includes several things, namely eliminating duplicate data, correcting data errors, deleting unnecessary data, adding incomplete data, and so on. By carrying out these stages, it is expected to produce attributes that are clean and good enough to be able to proceed to the next stage, namely clustering. This stage is usually carried out in the Microsoft Excel application because the tools provided are easier to use, such as for arranging columns and rows and correcting input errors in the data.

4.3. Clustering

Data mining techniques are presented at this stage. The author uses the K-Means clustering data mining algorithm to carry out the data grouping process. The process of this algorithm is as follows:

1. The first step is to input the final exam results data of Takhassus Abu Hurairah School into the Rapidminer application. The data inputted is data in the form of a Microsoft Excel file, which is then saved in the.csv format presented in Figure 2 above.
2. After the pre-processing process, namely cleaning and inputting data, the next step is to create a clustering model using the K-Means method, which will produce the desired number of clusters. In this study, clusters or groups were determined to be 3 clusters with 10 iterations. The K-means clustering operator can be seen in Figure 3.

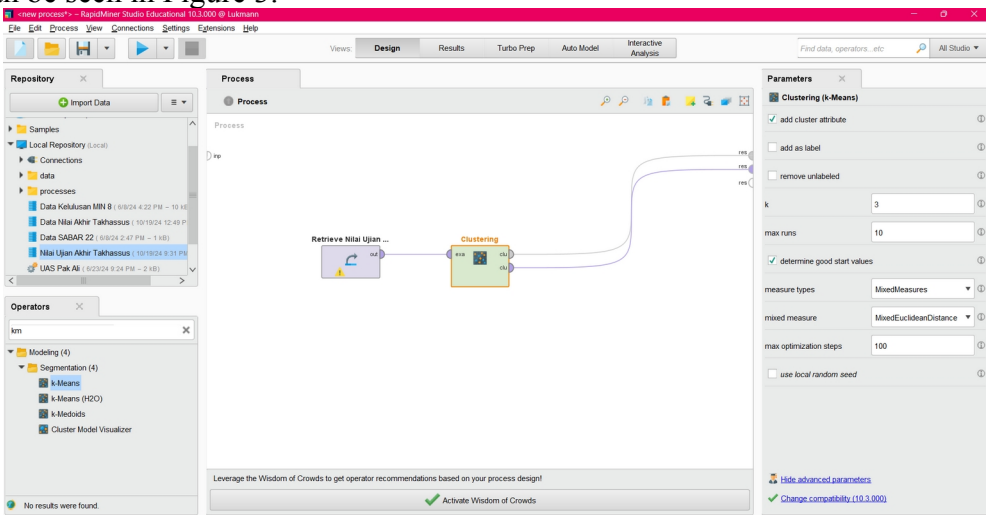

Figure 3. K-Means Operator

3. After the testing process with the K-Means operator using the Rapidminer application, from 50 Takhassus Abu Hurairah student data, the grouping results were obtained consisting of 3 clusters, namely cluster 0, cluster 1, and cluster 2. The following clusters of the process data are presented in Figure 4.

Figure 4. Clustering Results

4. The following is a display of members from each cluster after testing with the K-Means algorithm using the Rapidminer application:

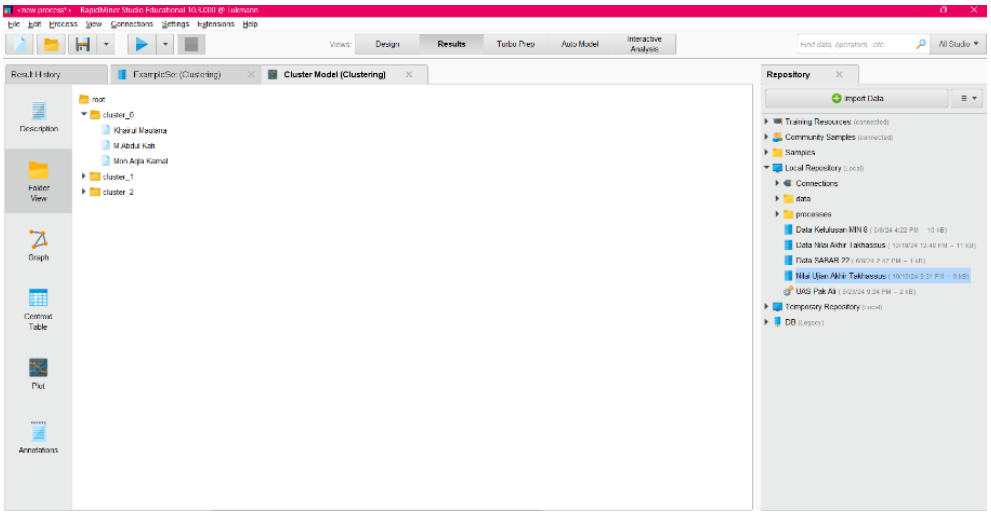

Figure 5. Cluster Member 0 on Rapidminer

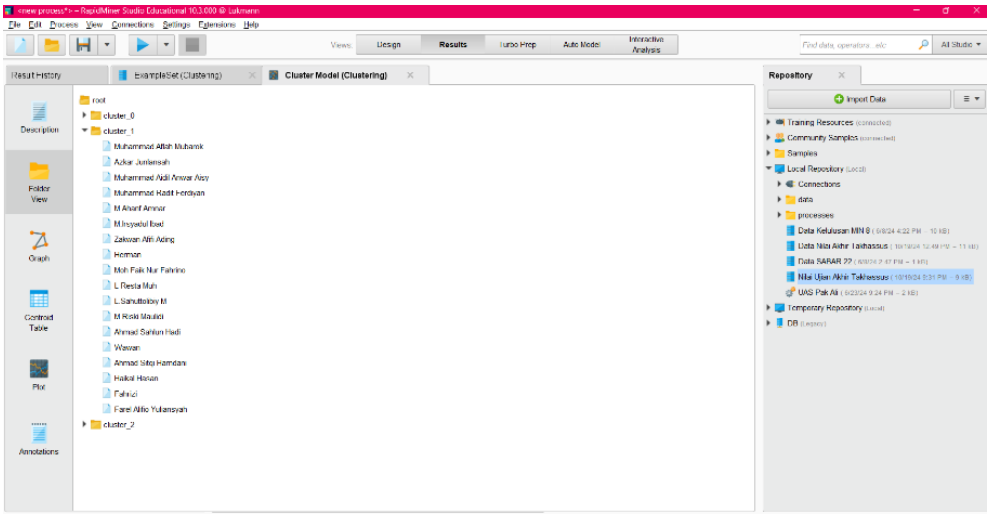

Figure 6. Cluster Member 1 on Rapidminer

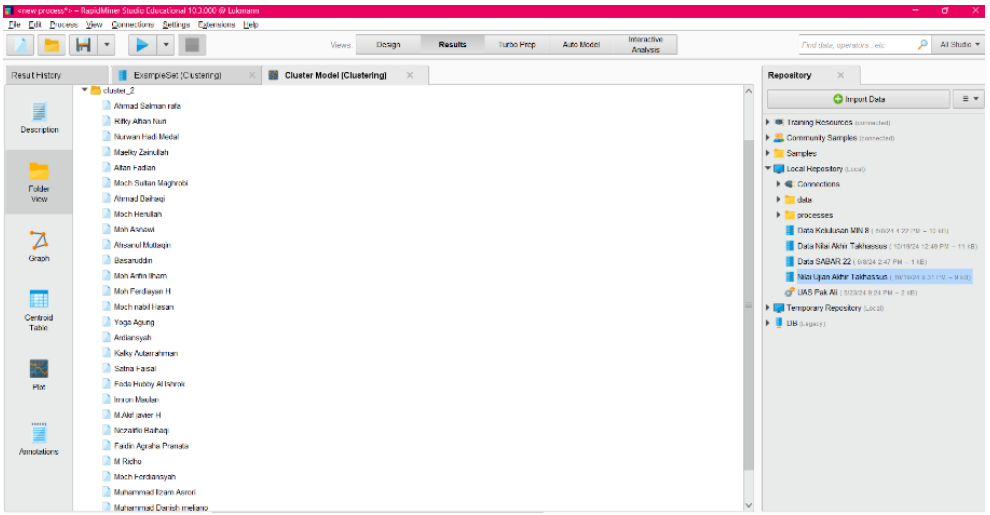

Figure 7. Cluster Member 2 on Rapidminer

5. The next step is to calculate the Davies Bouldin value in the cluster in the Rapidminer application using the Cluster Distance Performance operator which aims to determine how effective the clustering is and to evaluate the cluster in general. The results of the calculation can be seen in Figure 8 below.

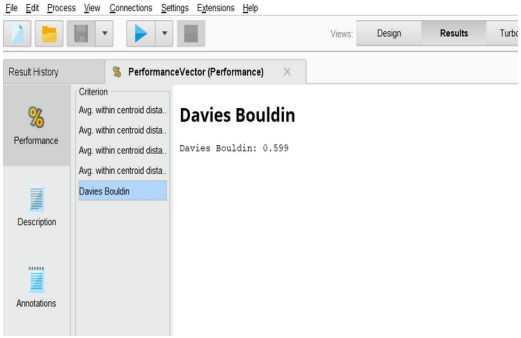

Figure 8. Davies Bouldin Cluster Values

4.4. Implementation and Results

From 50 student data points that have been tested using the Rapidminer application, the following results were obtained: Cluster 0 consists of 3 students, Cluster 1 consists of 18 students, and Cluster 2 consists of 29 students. With the performance vector:

1. Avg. within centroid distance: 176.298
2. Avg. within centroid distance_cluster_0: 36.444
3. Avg. within centroid distance_cluster_1: 269.59
4. Avg. within centroid distance_cluster_2: 132.859
5. Davies Bouldin: 0.599

Based on the results of clustering using the K-Means algorithm, we can provide an interpretation of the level of student intelligence based on the average distance to the centroid (Avg. within centroid distance). The smaller the distance to the centroid, the more similar the cluster members are to each other, which can be interpreted as an indication that students in the cluster have more uniform characteristics or can be called students with higher levels of intelligence. Here is the analysis:

1. Cluster 0: 3 students, with an average distance from the centroid of 36,444. With an average distance value of 36,444, this cluster is the closest to the centroid. This demonstrates how similar the kids in this cluster are to one another and suggests that they may share highly intelligent traits, or that they are a group of students with the highest IQs.
2. Cluster 2: 29 students, with an average distance from the centroid of 132.859. With a distance value of 132.859, this cluster is quite distant from the centroid on average. This indicates that while the kids in this cluster are intelligent, they are not as intelligent as those in Cluster 0, which is thought to be a cluster of children with a moderate IQ.
3. Cluster 1: 18 students, with an average distance from the centroid of 269.593. This cluster has the largest average distance to the centroid with a distance value of 269.593, indicating that there is more variation among students in this cluster. It is likely that students here have lower intelligence than other clusters.

After testing using the Cluster Distance Performance Operator, the Davies-Bouldin value was 0.599, indicating that the clustering results were quite good, with good cohesion between members in one cluster and clear separation between one cluster

and another. This value also shows that the grouping carried out was optimal in identifying variations in students' intelligence levels.

5. Conclusion

The test results using the RapidMiner application with the K-Means method show that out of 50 students of Takhassus Abu Hurairah Sukorejo School, they can be grouped into three clusters with a Davies-Bouldin value of 0.599, which indicates that the grouping is quite good. Cluster 0 contains 3 students with high intelligence, Cluster 2 contains 29 students with moderate intelligence, and Cluster 1 contains 18 with lower intelligence. The use of the K-Means Clustering method provides a deeper picture of the level of student intelligence, which allows schools to design more effective learning strategies and support optimal student development. These findings also confirm the great potential of data mining as an effective analysis tool in the education sector.

6. Acknowledgment

There is nothing we can say other than our deepest gratitude to all parties who have helped carry out this research and for all the efforts that have been made so that this research can be carried out properly. Especially to our parents, who always pray for us from afar. In addition, our gratitude to the supervising lecturers who have guided, helped, spent time and ideas, and provided very valuable input in terms of data mining science. In addition, we also thank the Takhassus Abu Hurairah school for granting permission to carry out this research and providing the facilities and data that made it easier for us to complete this research. Without them, this research would not have gone well.

References

[1] *Password Menuju Sukses*. Esensi. [Online]. Available: https://books.google.co.id/books?id=CQrowO-qOAAC

[2] A. Wasik *et al.*, "Implementasi data mining untuk memprediksi penjualan accessoris handphone dan handphone terlaris menggunakan metode k-nearest neighbor (k-nn) 1," vol. 1, no. 2, pp. 469–479, 2024.

[3] Y. Ardilla *et al.*, *DATA MINING DAN APLIKASINYA*. Penerbit Widina, 2021. [Online]. Available: https://books.google.co.id/books?id=53FXEAAAQBAJ

[4] M. S. Iskandar and Z. Fatah, "Gudang Jurnal Multidisiplin Ilmu Implementasi Metode Algoritma K-Means Clustering Untuk Menentukan Penerima Program Indonesia Pintar (PIP)," vol. 2, no. November, pp. 1–8, 2024.

[5] A. W. Aranski and K. Handoko, "Data Mining Dalam Pengelompokan Nilai Iq Siswa," *J. Teknol. Dan Open Source*, vol. 2, no. 2, pp. 13–22, 2019, doi: 10.36378/jtos.v2i2.347.

[6] M. Norshahlan, H. Jaya, and R. Kustini, "Penerapan Metode Clustering Dengan Algoritma K-means Pada Pengelompokan Data Calon Siswa Baru," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 2, no. 6, p. 1042, 2023, doi: 10.53513/jursi.v2i6.9148.

[7] D. Jollyta, A. Hajjah, E. Haerani, and M. Siddik, *Algoritma Klasifikasi untuk Pemula Solusi Python dan RapidMiner*. Deepublish, 2023. [Online]. Available: https://books.google.co.id/books?id=y84TEQAAQBAJ

[8] Z. Setiawan *et al.*, *BUKU AJAR DATA MINING*. PT. Sonpedia Publishing Indonesia, 2023. [Online]. Available: https://books.google.co.id/books?id=1nLVEAAAQBAJ

[9] P. W. Rahayu *et al.*, *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia,

2024. [Online]. Available: https://books.google.co.id/books?id=vCruEAAAQBAJ

[10] *PEMODELAN K- MEANS ALGORITMA DAN BIG DATA ANALYSIS (PEMETAAN DATA MUSTAHIQ)*. Pascal Books, 2022. [Online]. Available: https://books.google.co.id/books?id=_bJmEAAAQBAJ

[11] E. A. Novia, W. I. Rahayu, and C. Prianto, *SISTEM PERBANDINGAN ALGORITMA K-MEANS DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRIORITAS PEMBAYARAN TAGIHAN RUMAH SAKIT BERDASARKAN TINGKAT KEPENTINGAN*. Kreatif. [Online]. Available: https://books.google.co.id/books?id=MND9DwAAQBAJ

[12] D. Davies and D. Bouldin, "A Cluster Separation Measure," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. PAMI-1, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[13] S. T. M. K. Yahya, *Data Mining*. CV Jejak (Jejak Publisher), 2022. [Online]. Available: https://books.google.co.id/books?id=0J2mEAAAQBAJ