JOISTECH: Journal of Information System and Technology
ISSN 3063-9778 (Media Online)
Vol 01, No 01, Juni 2024, Hal 25-31

Optimizing Student Extracurricular Classification: RapidMiner Based K-Means Clustering Study at Darunnajah High School

Supriadi Panggabean[1*], Angga Kautsar Ibrahim[2], Zakiyanto[3], Azrul Azmani[4]

[1]Faculty of Science and Technology, System and Information Technology, Darunnajah University, Jakarta, Indonesia
[234] Magister of Management Study Program, Budi Luhur University, Jakarta, Indonesia
Email: [1*]supriadipanggabean@darunnajah.ac.id, [2]anggakautsar@gmail.com, [3]zakiyanto@gmail.com, [4]azrulazmani@darunnajah.ac.id

Abstract− This scientific journal discusses the grouping of students' extracurricular activities at Darunnajah School using the K-Means method based on RapidMiner. This research aims to uncover patterns and preferences in the realm of extracurricular involvement, providing valuable insights for optimizing school programs. Using the K-Means clustering technique with RapidMiner, groups of students with similar extracurricular interests can be identified. This research contributes to a deeper understanding of student engagement in non-academic activities, forming the basis for tailored planning and enhanced extracurricular offerings. The implementation of the RapidMiner platform has proven to be very efficient in analyzing and grouping student data, creating a more dynamic and interesting extracurricular environment at Darunnajah School.

Keywords: Clustering, grades, Darunnajah School, K-Means, Davies Bouldin Index

1. INTRODUCTION

Darunnajah Islamic Boarding School is a private (non-government) Islamic educational institution. Pioneered in 1942, the Islamic Boarding School was founded on April 1 1974, located on Jalan Ulujami Raya, number 86, Ulujami Village, Pesanggrahan District, South Jakarta City, DKI Jakarta Province. The location of the Islamic boarding school is very advantageous because it is in the capital, which makes communication easier, both with government agencies and with the wider community (Manaf, 2023).

Supported by a beautiful environment, the Darunnajah Islamic Boarding School strives to produce people who are muttafaqoh fiddin to become cadres of leaders of the people/nation, always striving to create education for students who have a spirit of sincerity, simplicity, independence, ukhuwah Islamiyah, freedom of thought and behavior based on Al- Quran and Sunnah of Rasulullah SAW to increase devotion to Allah SWT (Rosadhi, 2024).

Education has a very important role in shaping the future of the younger generation. In the current era of information technology, the use of data and analytical methods is becoming increasingly relevant to increasing the effectiveness of the education system (Muchamad Bachram Shidiq et al., 2023). One critical aspect in this context is student classification, which allows schools to provide a more personalized and efficient approach to learning and extracurricular programs. This study aims to optimize the extracurricular classification of students, especially at Darunnajah High School, through the application of the RapidMiner-based K-Means clustering method (Simangunsong et al., 2023).

The importance of student classification lies in the ability to recognize student learning behavior patterns and provide appropriate support. The K-Means method, as a clustering algorithm, allows identifying groups of students with similar characteristics. By integrating RapidMiner technology in data analysis, it is hoped that student groups that are more homogeneous can be found in terms of learning preferences and level of understanding of extracurricular subject matter (Alkhalifi et al., 2020).

Darunnajah High School, as an educational institution committed to the quality of learning, feels the need to improve existing student classification methods. In line with the development of information technology, the application of clustering algorithms has become a promising alternative to achieve this goal. Through more optimal student classification, it is hoped that schools can provide a more adaptive educational approach, according to the learning needs and preferences of each student in choosing extracurricular activities at school (Khodijah & Mukminin, 2024).

Clustering is a data mining method that is capable of grouping objects into clusters. A cluster is a group of data objects that are similar to each other, but are in the same group (Panggabean et al., 2022). Clustering techniques have been widely used to solve problems related to data separation(Fauzan et al., 2024). Clustering techniques can be used as a way to group texts that have similar content. Previous studies stated that with the clustering method, we can carry out the process of grouping dense areas, and ultimately find distribution patterns, as well as gain knowledge about the relationships between data attributes (Kusuma et al., 2023).

In this study, the goal is to optimize the student classification process at Darunnajah High School through a RapidMiner-based K-Means clustering study. By identifying homogeneous groups of students, it is hoped that the application of this technique can increase efficiency in preparing extracurricular learning programs that can provide maximum benefits for each student. Optimizing Student Extracurricular Classification: RapidMiner Based K-Means Clustering Study at Darunnajah High School

One method that can be used to solve the student identification problem is using the K-Means method. K-Means is a data mining algorithm that is capable of partitioning and separating data into different groups (Cesar et al., 2023). If K-Means groups data into several clusters within the same group, it will have different characteristics from other groups (Trisnapradika et al., 2023). The purpose of grouping the data is so that the objective function can be minimized. In other words, variation within a group will be minimized and variation between existing groups will be maximized (Lalu et al., 2024).

The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering in a dataset. It measures the average similarity ratio of each cluster with respect to its most similar cluster, considering both the dispersion within clusters and the separation between clusters (Ashari et al., 2023). A lower DBI value indicates better clustering, as it signifies that clusters are compact and well-separated (Umagapi et al., 2023).

DBI is calculated by averaging the maximum similarity ratios for each cluster. The similarity ratio for a cluster is determined by dividing the sum of within-cluster scatter and between-cluster distance by the between-cluster distance (Idham et al., 2022). This index helps in assessing the effectiveness of different clustering algorithms or configurations, guiding the selection of the most appropriate clustering method for a given dataset(Tarigan, 2023).

One advantage of using the Davies-Bouldin Index is its simplicity and ease of computation, which makes it a popular choice for quickly evaluating clustering results(Nozomi, 2023). Unlike some other clustering evaluation metrics, DBI does not require a predefined number of clusters, making it flexible and useful in various scenarios. Additionally, the index provides a clear and interpretable score that can be used to compare different clustering solutions directly(Kristanto et al., 2023).

However, the Davies-Bouldin Index also has some limitations. It may not always capture the true quality of clusters in cases where the clusters are not spherical or have varying densities. Furthermore, the index relies on the definition of a "distance" measure, which can affect the results depending on the choice of metric used (Budi et al., 2022). Despite these limitations, the DBI remains a valuable tool for initial clustering evaluation and comparison, especially when used in conjunction with other metrics and domain-specific knowledge to ensure robust clustering performance(Fathurrahman et al., 2023).

2. RESEARCH METHODOLOGY

This research utilizes data sourced from Pondok Pesantren Darunnajah. In this case, the method to be used is KDD (Knowledge Discovery in Database), which aims to identify new relevant patterns or knowledge from existing data. The data obtained includes information about students' academic achievements, participation in extracurricular activities, and student preferences towards learning programs. The KDD method will assist in extracting insights from large and complex data, allowing the identification of patterns and relationships among various relevant variables. Thus, this research will make a significant contribution to a deeper understanding of the factors influencing student achievement in the Islamic boarding school environment.

2.1. Modelling

The proposed model for grouping is to use the K-Means algorithm.

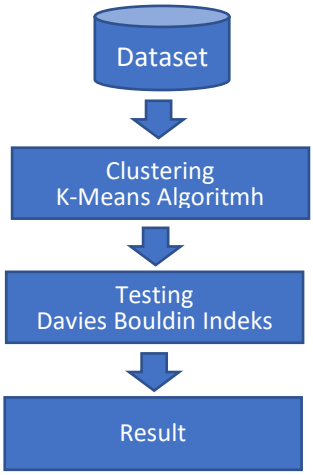

Figure 1. Proposed Research

2.2. Knowledge Discovery in Databases(KDD)

KDD is a method for obtaining new knowledge from collected data, as well as looking for related relationships in interconnected database tables and obtaining new knowledge from a series of mining processes for the basis of decision making and strategy. The KDD process has several stages starting from data selection to knowledge discovery (Llatas et al., 2024).

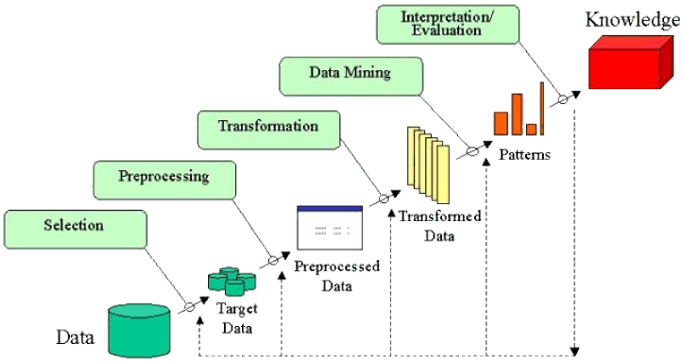

Figure 2. KDD Process in Data Science

2.1.1. Data Selection

It is very necessary to carry out the data selection process from the collected data so that the modeling form is appropriate to the method used. The data selection process is a dataset processed by the proposed algorithm, and the results can be stored in a separate file from the initial database. In this study, the data used is data from Darunnajah High School students for the 2023-2024 academic year. From this data there are several variables including, the student's name, the student's age, the student's grade level, the student's academic grades and the income of the student's parents. However, the data selected for the application of data mining is the student's academic grades and the income of the student's parents. Because based on the results of interviews with the management of the Darunnajah Islamic Boarding School, the income level of the students' parents is the most appropriate variable that can represent the academic values of the students, especially at Darunnajah High School.

2.1.2. Preprocessing/Cleaning

The cleaning process is carried out before the data mining stage is carried out. The need for a cleaning process aims to process the data that is focused on in KDD, as well as eliminating data duplication, resolving inconsistent data issues, and correcting errors in the data. On the other hand, enrichment also needs to be carried out because it is useful for processing existing information to make it more relevant. In this research, unnecessary data and incomplete data will be cleaned so that optimal grouping results can be obtained.

2.1.4. Transformation

The data transformation process is carried out so that the data becomes more suitable for use in the data mining process. The coding process is carried out as a response to a process that depends on the type of information to be sought from the database.

2.1.5. Data Mining

Data mining is the process of looking for interesting patterns or knowledge in previously selected data. Data mining has various techniques, methods and algorithms in carrying out the process. The choice of data mining method or algorithm depends on the goals to be achieved and the overall KDD process. In this process, research was carried out using RapidMiner to manage data using the K-Means Clustering technique.

2.1.6. Evaluation

The testing stage is needed to ensure that the model formed is the best. This stage includes checking whether the facts resulting from the modeling process contradict the hypothesis or previously existing facts. This stage will include a comparison of data generated by Excel and RapidMiner.

2.1.7. Knowledge

It is at this stage that the information pattern or structure is presented to the user. At this stage the knowledge gained can be understood by everyone and can be used as a reference for decision making. Data that has been processed will produce knowledge which will then be visualized using images.

2.2. Data Visualization

The process of presenting data to make it easier to accept and understand is through visualization. This condition developed since visual symbols were able to represent the meaning of something presented. Visualization transforms data into information that can be understood universally [8].

2.3. K-Means Algorithm Steps

The K-Means algorithm is an iterative clustering algorithm that works by partitioning the dataset into a desired number of K clusters [9]. This method will divide the data into several groups where the groups have the same traits or characteristics. The aim of this grouping is to minimize diversity within a group and maximize the types within the group [10]. The working method of the K-Means algorithm is as follows:

a. Set the value of k as the number of clusters you want to generate,
b. Setting the cluster center,
c. Calculate the distance of each data to the cluster center using the Euclidean equation:

$$d_{ik} = \sqrt{\sum_{j}^{m} \left(C_{ij} - C_{kj}\right)^2}$$

d. Group data into clusters with the shortest distance using the equation:

$$min \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j}^{m} \left(C_{ij} - C_{kj}\right)^2}$$

e. Calculate the new cluster center using the equation:

$$C_{kj} = \frac{\sum_{i=1}^{p} X_{ij}}{P}$$

Where the kth cluster and P is the number of members of the kth cluster

f. Iterate from steps 2 to 4 until there are no more movements of cluster members formed.

3. RESULTS AND DISCUSSION

3.1. Implementation of the K-Means Algorithm

In the initial stage of the study, a manual simulation was carried out to implement the K-means algorithm, with a number of Clusters of 5, a total of 351 data (students), and a number of attributes of 2 (average student grades and extracurricular choices). This manual simulation provided a foundational understanding of the algorithm's functionality and allowed for a hands-on exploration of the clustering process. Additionally, it served as a benchmark for evaluating the performance of automated clustering techniques later in the study.

3.2. Data Visualization

Data visualization is the graphical representation of data and information to present complex datasets in a clear, concise, and understandable manner. It involves creating visual elements such as charts, graphs, and maps to help users interpret and comprehend large volumes of data quickly and efficiently. Data visualization plays a crucial role in data analysis, as it allows analysts and decision-makers to identify patterns, trends, and relationships within the data that may not be apparent from raw numbers alone.

By visually representing data, organizations can gain valuable insights and make data-driven decisions more effectively. For example, through interactive dashboards, executives can monitor key performance indicators (KPIs) and track progress towards organizational goals in real-time. Similarly, data visualization can aid researchers in communicating their findings more clearly to a wider audience, enhancing understanding and facilitating knowledge dissemination.

Moreover, data visualization tools and techniques continue to evolve, allowing for the creation of increasingly sophisticated visualizations. These tools often incorporate features such as interactivity, drill-down capabilities, and predictive analytics, enabling users to explore data from multiple angles and gain deeper insights. From simple bar charts to complex network diagrams, data visualization empowers users to explore data in meaningful ways, leading to better decision-making and improved understanding of complex phenomena.

3.3. Result Clusterization

3.3.1 K-Means Clustering on RapidMiner

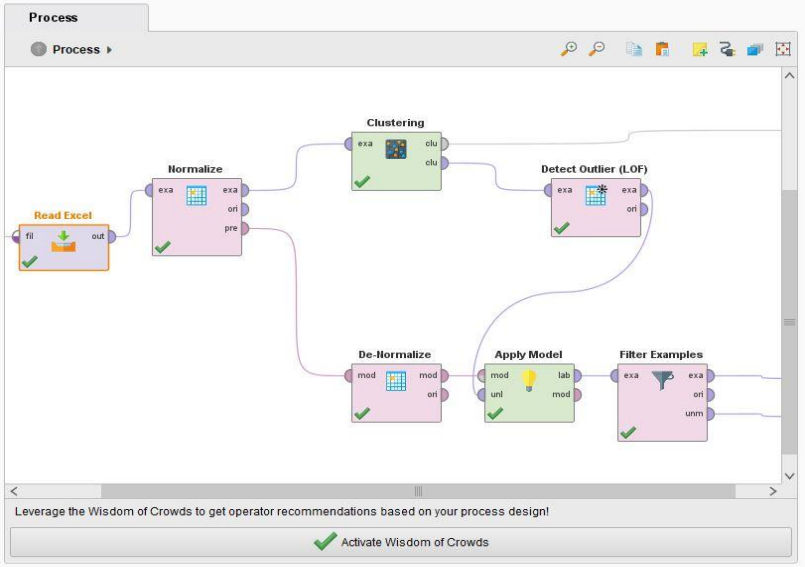

Figure 3. K-Means Algorithm Modeling

3.3.2. Cluster population average

The study focused on examining the relationship between cluster population average achievement and extracurricular selection among students. Initially, a manual simulation was conducted to implement the K-means algorithm, dividing the student population into 5 clusters based on their average grades and extracurricular choices. This simulation aimed to understand how different student profiles clustered together based on these two attributes. The results of the simulation revealed interesting patterns in the cluster population average achievement with extracurricular selection. It was found that certain clusters exhibited higher academic performance levels compared to others, while also displaying distinct preferences in extracurricular activities. For instance, clusters with higher average achievement tended to show preferences for academic clubs or leadership roles, whereas clusters with lower average achievement leaned towards recreational or social extracurriculars.

Further analysis indicated that the relationship between cluster population average achievement and extracurricular selection was not solely determined by academic performance. Instead, it was influenced by a combination of factors, including students' interests, aspirations, and socio-economic backgrounds. Students in clusters with higher average achievement often engaged in extracurricular activities that complemented their academic goals, while those in lower-achieving clusters may have prioritized activities that provided social support or relaxation. Understanding these patterns can provide valuable insights for educational institutions in designing extracurricular programs tailored to students' diverse needs and interests. By recognizing the preferences and characteristics of different clusters, schools can better support student development and academic success by offering relevant and engaging extracurricular opportunities. Additionally, this research underscores the importance of considering holistic approaches to student engagement and achievement beyond academic performance alone.

Figure 4. Cluster population average achievement with extracurricular selection

The visualization data used in this research is the Rapidminer application, which aims to provide a bubble view image based on a distribution map of students' extracurricular choices based on students' average achievement scores. The first visualization depicts the distribution of extracurricular choices for average student achievement in each cluster with different colors adjusting the cluster so that it looks like the picture 4.

3.3.3. Student Grade Dsitribution Graph

In the initial stage of the study, a manual simulation was carried out to implement the K-means algorithm, with a number of Clusters of 5, a total of 351 data points (students), and a number of attributes of 2 (average student grades and extracurricular choices). The Student Grade Distribution Graph was then used to visualize the distribution of grades among the students, providing a clear depiction of how grades were spread across the different clusters.

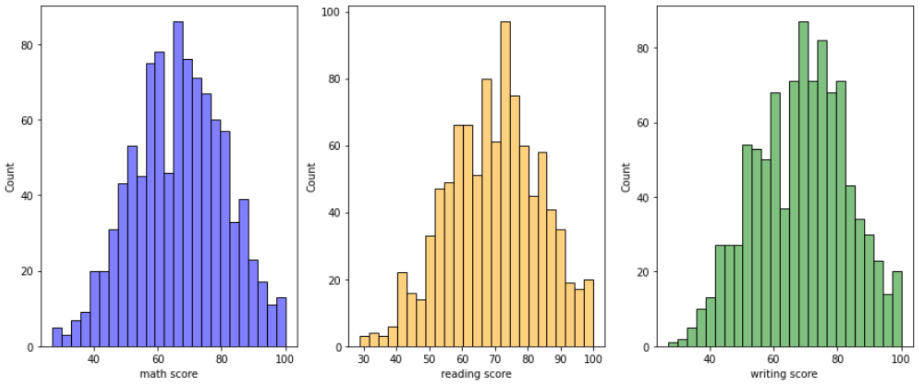

Figure 5. Student Grade Dsitribution Graph

3.3.3. Davies Bouldin Index Testing

The validation test used is the Davies-Bouldin Index (DBI) validation test. Based on the DBI test, the quality of the clustering can be determined, with smaller or minimal test values indicating better clustering results. In this test, the validation was conducted once. After performing the test with a model of k=2, the obtained DBI or Davies-Bouldin Index value was:

Davies Bouldin

```
Davies Bouldin: 0.653
```

Figure 6. Davies Bouldin Index Testing

4. CONCLUSION

The research conducted, utilizing both manual analysis and the RapidMiner application, led to the classification of student extracurricular types into 5 distinct clusters, with each cluster exhibiting notable similarities. The K-Means method can be used to determine the appointment of permanent employees by using assessment attributes, a calculation process with the k-means algorithm with the Euclidean Distance formula. The DBI validation result of 0,653 can be categorized as quite good. This study demonstrates the effectiveness of employing K-Means clustering through the RapidMiner application, providing organizations and institutions with a practical means to identify and analyze data clusters efficiently. Moreover, the application offers convenience in monitoring and evaluating students, particularly in the process of assigning classrooms based on their extracurricular activities. By leveraging this technology, institutions can streamline their decision-making processes and better understand student preferences and behaviors, ultimately enhancing organizational efficiency and student satisfaction. The findings of this research contribute to the growing body of knowledge in data analysis techniques and their practical applications in educational settings.

REFERENCES

Alkhalifi, Y., Gata, W., Prasetyo, A., & Budiawan, I. (2020). Analisis Sentimen Penghapusan Ujian Nasional pada Twitter Menggunakan Support Vector Machine dan Naïve Bayes berbasis Particle Swarm Optimization. *CoreIT*, *6*(2), 71–78. http://ejournal.uin-suska.ac.id/index.php/coreit/article/view/9723

Ashari, I. F., Dwi Nugroho, E., Baraku, R., Novri Yanda, I., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. *Journal of Applied Informatics and Computing*, *7*(1), 89–97. https://doi.org/10.30871/jaic.v7i1.4947

Budi, S., Gata, W., Noor, M., Pangabean, S., & Rahayu, C. S. (2022). News Portal Website Measurement Analysis Using Iso/Iec 25010 and Mccall Methods (Analisis Pengukuran Website Portal Berita Menggunakan Metode Iso/Iec 25010 dan McCall). *Journal of Applied Engineering and Technological Science*, *4*(1), 273–285.

Cesar, W., Saputra, R. R., & Wibowo, A. (2023). Klasterisasi Kepadatan Pegawai dengan Metode K-Means untuk Prediksi Kebutuhan CASN Instansi Pemerintah. *JATISI (Jurnal Teknik …*, *10*(2), 340–354. https://jurnal.mdp.ac.id/index.php/jatisi/article/view/4593

Fathurrahman, F., Harini, S., & Kusumawati, R. (2023). Evaluasi Clustering K-Means Dan K-Medoid Pada Persebaran Covid-19 Di Indonesia Dengan Metode Davies-Bouldin Index (Dbi). *Jurnal Mnemonic*, *6*(2), 117–128. https://doi.org/10.36040/mnemonic.v6i2.6642

Fauzan, A. S., Irma, A., Sari, P., & Ali, I. (2024). ANALISIS PERBANDINGAN ALGORITMA DECISIOIN TREE DAN NAÏVE UNTUK MENGEVALUASI PRESTASI BELAJAR SISWA STUDI KASUS : SMK AL-MUSYAWIRIN. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 741–747.

Idham, I., Ghudafa Taufik Akbar, M., Panggabean, S., & Noor, M. (2022). Perbandingan Prediksi Harga Saham Dengan Menggunakan LSTM GRU Dengan Transformer. *Smart Comp: Jurnalnya Orang Pintar Komputer*, *11*(1), 44–47. https://doi.org/10.30591/smartcomp.v11i1.3185

Khodijah, S., & Mukminin, U. (2024). *P engaruh Pembelajaran Aqidah Akhlak Terhadap Perilaku Peserta Didik Kelas V M IS Darunnajah 2 Cipining*. 2–5.

Kristanto, B., Turmudi Zy, A., & M. Fatchan. (2023). Analisis Penentuan Karyawan Tetap Dengan Algoritma K-Means Dan Davies Bouldin Index. *Bulletin of Information Technology (BIT)*, *4*(1), 112–120. https://doi.org/10.47065/bit.v4i1.521

Kusuma, M. R., Windu Gata, Sigit Kurniawan, Dedi Dwi Saputra, & Supriadi Panggabean. (2023). Software Defect Prediction For Quality Evaluation Using Learning Techniques Ensemble Stacking. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(2), 1–13. https://doi.org/10.35585/inspir.v13i2.58

Lalu, Q., Serta, Q., & Grafik, D. (2024). *PEMETAAN DAN ANALISIS KANDUNGAN NUTRISI PADA MINUMAN STARBUCKS DENGAN METODE K-MEANS Mapping and Analysis of Nutritional Content in Starbucks Beverages Using The K-Means Methode industri makanan dan minuman menjadikan tahunan Komisi Pengawasan Persaingan Usaha melakukan pengukuran Indeks Persaingan Usaha (IPU) dengan sistem skor 1-7 dengan Target Nasional yang tertuang dalam Rencana Pembangunan Jangka Menengah Tahun 2024 , Persaingan usaha yang berada di makanan dan minuman mengadakan berbagai*. *1*(2).

Llatas, C., Soust-Verdaguer, B., Torres, L. C., & Cagigas, D. (2024). Application of Knowledge Discovery in Databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design. *Journal of Building Engineering*, *91*(January), 109546. https://doi.org/10.1016/j.jobe.2024.109546

Manaf, S. (2023). Peran Kepala Sekolah Dalam Meningkatkan Kompetensi Guru Di Pondok Pesantren Darunnajah Jakarta. *MUDIR (Jurnal Manajemen Pendidikan)*, *5*(1), 49–54. https://doi.org/10.55352/mudir

Muchamad Bachram Shidiq, Gata, W., Kurniawan, S., Saputra, D. D., & Panggabean, S. (2023). Time Effort Prediction Of Agile Software Development Using Machine Learning Techniques. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(2), 39–48. https://doi.org/10.35585/inspir.v13i2.57

Nozomi, I. (2023). Penerapan Data Mining Untuk Peringatan Dini Banjir Menggunakan Metode Klastering K-Means (Studi Kasus Kota Padang). *Jurnal Sains Informatika Terapan*, *2*(2), 39–44. https://doi.org/10.62357/jsit.v2i2.165

Panggabean, S., Gata, W., & Setiawan, T. A. (2022). Analysis of Twitter Sentiment Towards Madrasahs Using Classification Methods. *Journal of Applied Engineering and Technological Science*, *4*(1), 375–389. https://doi.org/10.37385/jaets.v4i1.1088

Rosadhi, A. H. (2024). *Strategi Pemasaran Jasa Pendidikan Di Pesantren Darunnajah 2 Cipining*. *1*.

Simangunsong, B. N., Manalu, M. R., & Medan, U. I. (2023). *Testing the K-Means Clustering Algorithm in Processing Student Assignment Grades Using the RapidMiner Application*. *1*, 51–60.

Tarigan, D. A. (2023). Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification. *Media Online)*, *4*(1), 545–552. https://doi.org/10.30865/klik.v4i1.964

Trisnapradika, G. A., Ghozi, W., & Yuminah, Y. (2023). Rekomendasi Paket Mata Pelajaran Pilihan (MPP) pada SMA Negeri 1 Kebumen Menggunakan Algoritma K-means. *Jurnal Informatika Dan Rekayasa Perangkat Lunak*, *5*(2), 96. https://doi.org/10.36499/jinrpl.v5i2.8514

Umagapi, I. T., Umaternate, B., Komputer, S., Pasca Sarjana Universitas Handayani, P., Kepegawaian Daerah Kabupaten Pulau Morotai, B., & Riset dan Inovasi, B. (2023). Uji Kinerja K-Means Clustering Menggunakan Davies-Bouldin Index Pada Pengelompokan Data Prestasi Siswa. *Seminar Nasional SISFOTEK*, 303–308.