Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

Analysis Of Population Data Grouping Using K-Means For Efficiency Data Centralization And Backup

Putri Alda[1*], Rika Nofitri[2], Edi Kurniawan[3]
[1,3]Information System, Faculty Of Computer Science, University Royal, Asahan, North Sumatra, Indonesia
[2]Computer System, Faculty Of Computer Science, University Royal, Asahan, North Sumatra, Indonesia
[1]putrialda0803@gmail.com , [2]nofitririka15@gmail.com , [3]edikurniawan@royal.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 05-04-2026
Accepted: 16-04-2026
Published: 20-04-2026

Keywords:
Data Mining; Clustering; K-Means; Population Data; Data Centralization.

ABSTRACT

Population data management at the village level plays an important role in supporting administrative services, development planning, and data-driven decision-making. However, in many village offices, including the Bandar Sono Village Office, population data is still managed manually and stored in fragmented formats, resulting in inefficiencies in data retrieval, duplication risks, and difficulties in performing structured data backup. This condition indicates the need for a more systematic approach to organizing and analyzing population data. This study aims to analyze and implement the K-Means clustering algorithm to group population data in order to improve the efficiency of data centralization and backup processes. The research method used is quantitative, involving data collection, preprocessing, and clustering analysis using the K-Means algorithm based on attributes such as the number of male residents, female residents, and total population. The results show that the K-Means method successfully groups population data into three clusters, namely small, medium, and large population categories. These clustering results provide more structured and meaningful information, facilitating easier data analysis and improving the effectiveness of data management. In conclusion, the implementation of K-Means clustering contributes to enhancing the efficiency, organization, and reliability of population data management systems, thereby supporting better administrative services and decision-making at the village level.

INTRODUCTION

With the rapid development of information technology, public data management has increasingly become a key priority for government agencies at all levels, including at the village level (Sulistiani et al. 2025). Population data is a vital asset that forms the basis for development planning, the provision of public services, and the allocation of social assistance (Wibowo and Aryanti 2025). However, in many village offices, population data is still stored in a fragmented manner for example, in different Excel files, manual records, or separate computer systems—making the processes of searching, verifying, and backing up data inefficient and prone to data loss (Fazira, Fitri, and Risawandi 2025). This situation results in slow administrative services for residents, difficulties in determining aid recipients, and the risk of data loss due to equipment failure or human error (Iin 2025).

The issue of data centralization becomes increasingly critical as the population grows. The growth in data volume means that traditional backup processes (manual backups to flash drives, external hard drives, or separate files) take longer, are prone to duplication, and can potentially lead to inconsistencies between data sources (Fadhil, Fuadi, and Maryana 2025). Without proper data grouping and normalization mechanisms, efforts to consolidate village databases become difficult and require significant human resources (Rahmawati, Prihartono, and Cirebon 2025). Therefore, a solution is needed that not only centralizes data but also simplifies the understanding of population group characteristics, thereby facilitating backup strategies, service prioritization, and local policies (Homepage et al. 2025).

The Office of the Village Head of Bandar Sono, Nibung Hangus Subdistrict, Batu Bara Regency, is a government agency that plays a vital role in providing various administrative services to the community. This office still manages population data using hard copies and photocopies of Family Cards (KK). As the population grows, this situation results in data that is not yet properly centralized, making data retrieval and backup processes difficult. Therefore, an analysis of population data clustering based on data mining using the K-Means Clustering method is needed to support the efficiency of population data centralization and backup.

However, in practice, data management at the village office still faces challenges due to the lack of a centralized storage system. Population data remains scattered across various physical archives and separate digital files, often leading to duplication, inconsistencies, and slowing down search and verification processes. Reliance on manual record-keeping and a lack of integration among village departments make services less efficient, especially as the volume of data continues to increase year after year.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

In addition to these issues, the growing population of Bandar Sono Village each year presents its own set of challenges. As the volume of data to be managed increases, so does the risk of recording errors or data loss, especially when data processing and storage are still done manually. The data backup process also becomes increasingly difficult and time-consuming due to the lack of a structured and organized system. Reliance on manual storage methods makes data vulnerable to corruption, loss, or incompleteness in the event of technical disruptions such as hardware failure or input errors.

In the context of data analysis, clustering methods are extremely useful tools for grouping entities based on common attributes such as age, occupation, economic status, possession of administrative documents, or address (Alawiyah, Aghnia, and Abdalah 2025) (Rival et al 2024). One of the most widely used clustering algorithms, due to its simplicity and efficiency, is K-Means (Hasim Azari, Dwi Hartanti, and Aprilisa Arum Sari 2024). K-Means works by grouping data into k clusters based on proximity to the centroid; although it has limitations (e.g., sensitivity to initialization and variable scale), recent practices and developments show that with proper data preprocessing (normalization, outlier handling) and cluster metric evaluation, K-Means is capable of producing meaningful clusters for local government needs (Bili, Abineno, and Aha Pekuwali 2024).

The application of K-Means to village-level population data offers several practical benefits (Usino 2024). First, clustering can reveal profiles of resident groups (e.g., clusters of economically vulnerable residents, clusters of young families, clusters of elderly residents), enabling village heads and staff to design more targeted policies. Second, by clustering the data, the backup process can be optimized: for example, frequently changing or at-risk data (family registration records) can be prioritized in periodic backup strategies, while static data can be archived separately, making the backup process more efficient in terms of time and storage space. Third, the results of clustering can be used to detect data inconsistencies or duplications before the centralization process, thereby improving the quality of the centralized database (Susilo et al. 2024).

However, the application of data mining techniques in government data management also requires attention to several aspects, namely the quality and completeness of data attributes, the protection of citizens' privacy, and the readiness of village human resources in terms of system operation. The pre-processing stage (cleaning, normalization, encoding) is critical to the success of clustering (Febriyanti, Harahap, and Masrial 2024). Additionally, developing backup policies that incorporate clustering results requires procedural design and training so that village officials can operate and maintain the centralized system sustainably.

LITERATURE REVIEW

The development of data mining techniques has significantly contributed to improving data processing and analysis across various domains, including government administration. One of the commonly used approaches is clustering, which groups data based on similarities in specific attributes to produce more meaningful information for decision-making (Hendrastuty 2024; (Susilo et al. 2024). Compared to classification methods, clustering does not require predefined labels, making it suitable for exploratory analysis, particularly in population data management where patterns are not yet clearly defined.

Among various clustering methods, K-Means is widely adopted due to its simplicity, computational efficiency, and ability to handle large datasets (Hasim et al 2024; Wibowo and Aryanti 2025). Several studies have demonstrated that K-Means is effective in identifying data patterns, such as grouping student performance, sales data, and regional characteristics (Bili et al 2024: Rahmawati et al 2025). However, these studies also highlight certain limitations, including sensitivity to the number of clusters (K) and dependence on initial centroid selection, which may affect clustering accuracy (Fazira et al 2025 : Febriyanti et al 2024).

Furthermore, previous research tends to focus primarily on clustering as a data analysis technique, with limited integration into real-world information systems that directly support operational processes (Fadhil et al 2025; Rahmawati et al 2025). In contrast, the application of clustering in government data management requires not only accurate grouping but also the ability to integrate results into systems that enhance data centralization, backup efficiency, and service delivery.

Based on these studies, it can be identified that there is still a gap in the implementation of K-Means clustering for population data, particularly in integrating clustering results into practical systems that support administrative efficiency. Therefore, this study proposes the implementation of K-Means clustering within a web-based system to improve population data centralization and backup processes, while also providing structured and actionable information for decision-making.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

METHOD

This study employs a quantitative method with the following steps:

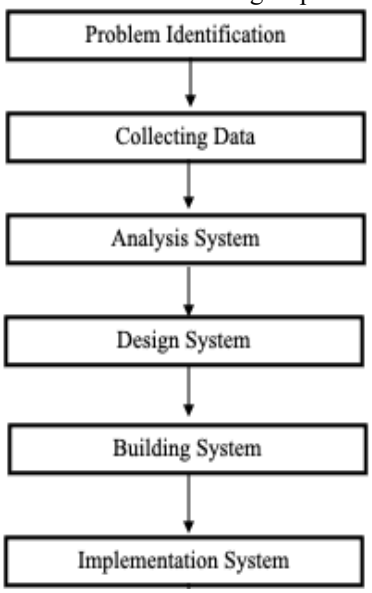

Figure 1. Research Stage

Problem Identification

This phase aims to identify the main issues facing the Bandar Sono Village Office, particularly regarding the management of population data, which continues to grow year after year. This problem identification serves as the basis for the need to implement the K-Means algorithm to assist in clustering population data as a strategic step toward improving the efficiency of centralization and backup processes at the Bandar Sono Village Office.

Collecting Data

Data collection was conducted to obtain accurate information as a basis for the analysis of population data clustering using the K-Means algorithm. This stage is crucial to ensure that the clustering results accurately reflect the actual data conditions and are relevant for supporting efforts to improve the efficiency of data centralization and backup at the Bandar Sono Village Office.

Analysis System

A system analysis was conducted to understand user needs, the workflow for managing population data, and the functions required for the clustering process using the K-Means algorithm. The purpose of this phase is to ensure that the designed system or analytical model can support the data centralization and backup processes more efficiently at the Bandar Sono Village Office.

Design System

The system design phase is conducted to create a blueprint or initial design before the system is built. Activities include:

a. The user interface (UI) design used to create the login screen, admin dashboard, resident data management menu, data centralization feature, and K-Means clustering results display.
b. Creating an ERD and database schema design for managing population data, clustering attributes, and backup data.
c. Creating flowcharts to illustrate the data input process, clustering processing, data storage, and the backup process.

This design ensures that the system developed meets the needs and is capable of improving the efficiency of village administration.

Building System

This phase involves implementing the system design into a functional application. Activities include developing the user interface, creating the database structure, and integrating the population data centralization module and the K-Means algorithm.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

Implementation System

Implementation takes place after the system has been confirmed to be ready based on the results of initial testing. At this stage, the web-based population data management system is installed on the designated server, and the database is configured so that all tables, relationships, and population attributes can be processed properly.

RESULT

Analysis Data

At this stage, the population data was grouped based on the number of males, the number of females, and the total population in each hamlet in Bandar Sono Village using the K-Means Clustering algorithm. The calculations were performed manually using a spreadsheet to ensure the accuracy of the results before they were implemented into the system. The purpose of this grouping was to identify hamlets with small, medium, and large populations, thereby supporting data centralization and backup strategies. The processed data is presented in Table 1 below.

Table 1. Population Data

Label 1	Label 2	Label 3	Total Population	Number of Men	Number of Woman
Dusun I	Dusun I Perbatasan	Dusun I Perbatasan	33	16	17
Dusun II	Dusun II Sei Jawi-Jawi	Dusun II Sei Jawi-Jawi	49	27	22
Dusun III	Dusun III Bunga Tanjung	Jl. Inpres Dusun III Bunga Tanjung	46	24	22
Dusun IV	Dusun IV Kubah Sono	Dusun IV Kub Ah Sono	107	53	54
Dusun V	Dusun V Kedai Ramai	Jl. H. Nongah Dusun V Kedai Ramai	100	53	47
Dusun VI	Dusun VI Sono Tengah	Jl. H. Bahrum Dusun VI Sono Tengah	71	37	34
Dusun VII	Dusun VII Sono Timur	Dusun VII Sono Timur	51	27	24
Dusun VIII	Dusun VIII Alur Naga	Dusun VIII Alur Naga	47	22	25
Dusun IX	Dusun IX Lubuk Rukam	Jl. Lubuk Rukam Dusun IX	62	35	27

Design System

The system design was developed using a use case diagram. The purpose of this design is to explain the requirements in detail (Prasetya, Sintia, and Putri 2022). The use case diagram is shown in Figure 2.

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

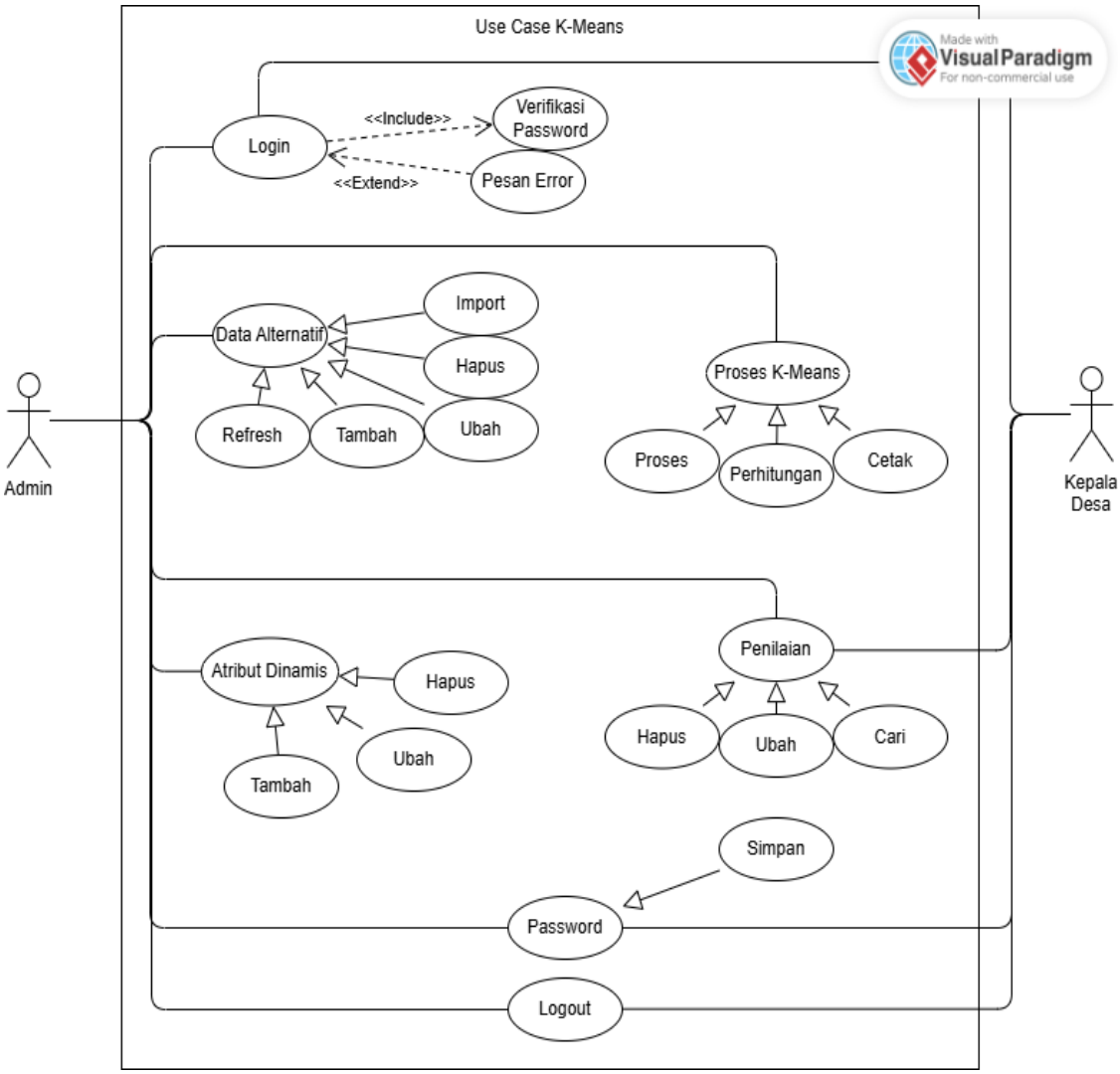

Figure 2. Design System

Figure 2 illustrates the use case diagram of the proposed system, which describes the interaction between users and the system in managing population data. The primary actor in this system is the admin, who has full access to all system functionalities. The admin can perform several main activities, including managing resident data, conducting the clustering process using the K-Means algorithm, viewing clustering results, and generating reports.

In addition, the system provides features to support data centralization and backup processes, ensuring that population data is stored in an organized and structured manner. The clustering feature allows the system to group population data based on predefined attributes, thereby assisting users in identifying data patterns and supporting decision-making. Overall, this diagram represents the functional requirements of the system and serves as a blueprint for system development.

Implementation System

The interface implementation phase is carried out to bring the previously designed system interface to life so that it can be used directly by users.

- **Login Page**

This is the first step before users can access the system. On this page, users are prompted to enter their registered username and password. The system will perform a validation process to ensure that the entered data matches the information stored in the database. If authentication is successful, users will be redirected to the main page based on their access level.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, Number 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

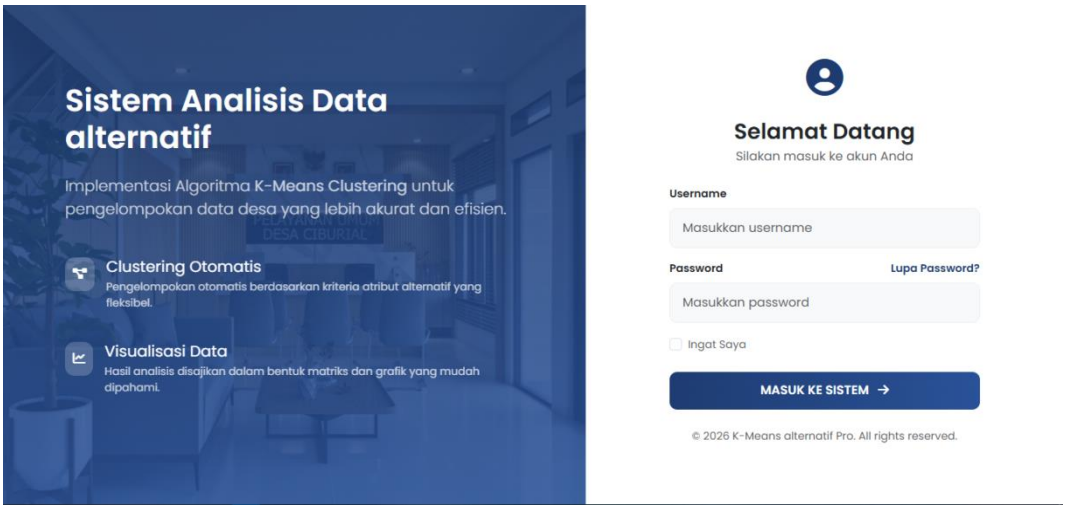

Figure 3. Login Page

- **Main Page**

Displays all menus accessible to the Admin, such as hamlet data management, the clustering process, reports, and account settings. The dashboard interface is designed to present information in a clear and easy-to-understand manner.

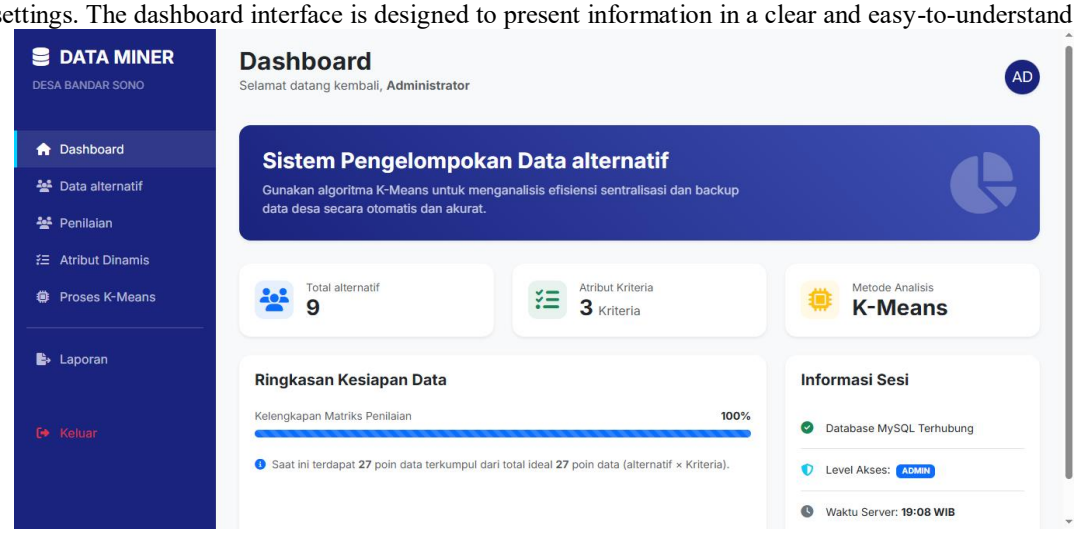

Figure 4. Main Page

- **Population Data Page**

This is a form used to view data and to select, edit, and delete resident data.

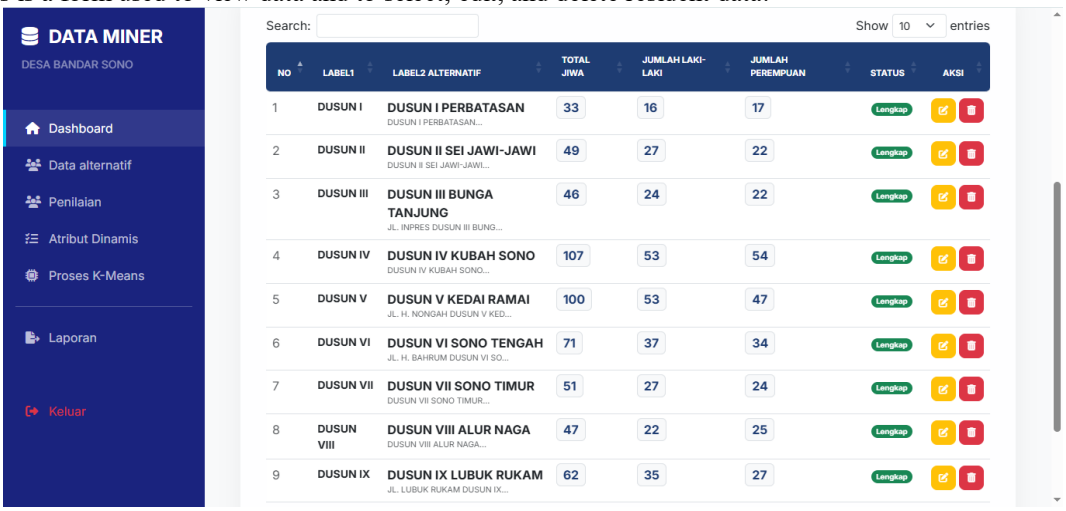

Figure 5. Resident Data Page

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

- **Cluster Page**
  This is a form used to enter population data for each hamlet into the system.

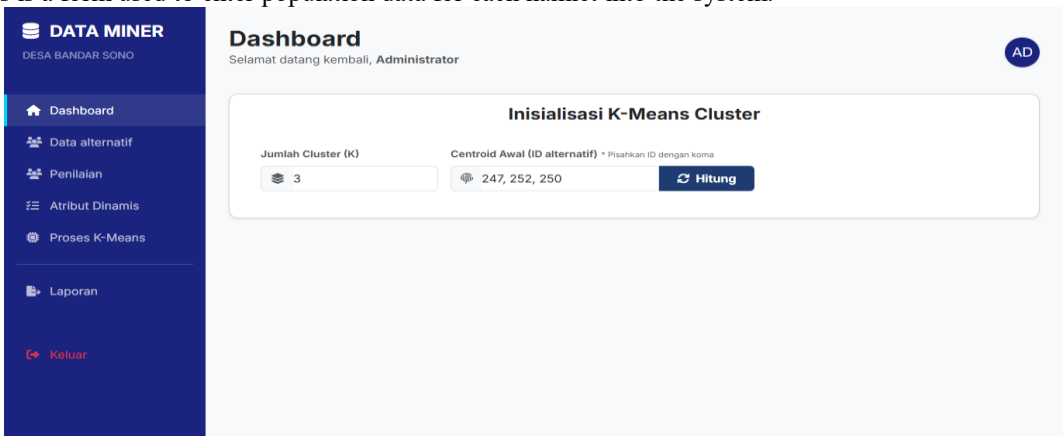

Figure 6. Cluster page

- **Calculation Page**
  This is a form used to process data mining calculations and print the clustering results..

Figure 7. Calculation Page

- **Graphic Page**
  Displays graphical visualization of K-Means clustering results to help users identify population patterns and support decision-making.

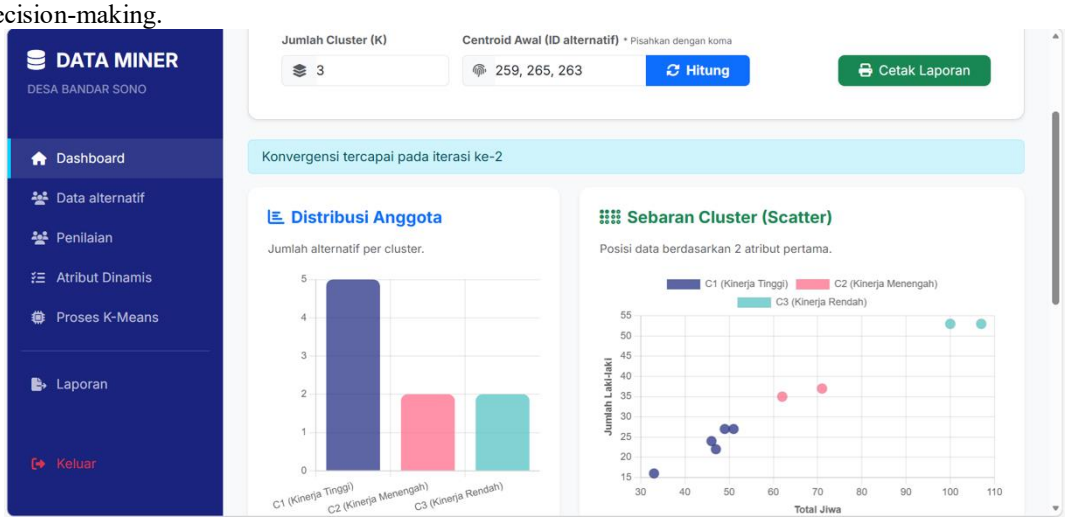

Figure 8. Graphic Page

Brilliance
Research of Artificial Intelligence
E-ISSN : 2807-9035
Volume 6, Number 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

- **Report Page**

Displays clustering results in report form, enabling data documentation, analysis, and supporting decision-making and backup processes.

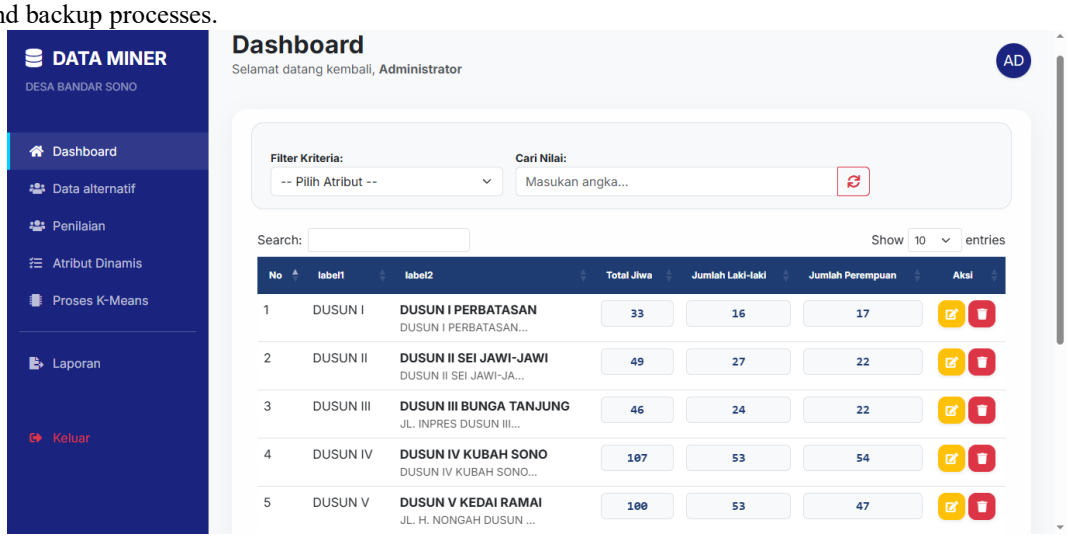

Figure 9. Report Page

DISCUSSION

Before applying the K-Means method, the data used in this study consisted solely of a collection of ungrouped raw data. The available information was still general in nature and could not provide a clear picture of specific patterns or characteristics within the data. This made it difficult to conduct further analysis and make data-driven decisions, due to the lack of grouping that could help identify relevant categories or segments.

After applying the K-Means method, the data was successfully grouped into several clusters based on the similarity of characteristics among the data points. These clustering results provide more structured and easily understandable information, making it easier to identify specific patterns that were previously not visible. Each cluster represents a group of data with similar characteristics, and can thus serve as a basis for further analysis and decision-making.

A comparison of the data before and after applying K-Means demonstrates an improvement in the quality of the information produced. Before clustering, data analysis tended to be general and non-specific; however, after clustering, the data can be analyzed in greater depth based on the formed groups. This allows users to better understand the data distribution and identify groups with specific characteristics.

Thus, the application of the K-Means method not only serves as a data clustering technique but also provides added value in the form of more structured information, thereby supporting more effective decision-making. The resulting clusters can be used to group population data by variable, offering more practical benefits than manual data processing.

CONCLUSION

Based on the research conducted, it can be concluded that the K-Means algorithm is capable of clustering population data based on the variables of the number of males, the number of females, and the total population into three categories: small, medium, and large. The calculation process involves determining centroids, calculating distances using the Euclidean distance method, and iterating until convergence is achieved. The application, developed using PHP and MySQL, can centrally store village data, automatically perform the clustering process, and display the clustering results in the form of tables and graphs. This indicates that the system has met the functional requirements formulated at the beginning of the study, and the information regarding population scale categories for each village provides a more structured overview of data distribution. Consequently, village officials can plan administrative management and data backup strategies in a more systematic and data-driven manner.

REFERENCES

Alawiyah, Ai, Nurul Aghnia, and Frans Fauzan Abdalah. 2025. “Implementasi Clustering Algoritma K-Means Pada Penjualan Beras Di CV Tangguh Bumi Perkasa.” *Jurnal Komisi (Jurnal Komputer Dan Sistem Informasi)* 2(2):17–23.

Bili, Natalia, Reynaldi Thimotus Abineno, and Aha Aha Pekuwali. 2024. “Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Peforma Siswa Pada Pembelajaran Bahasa Indonesia (Studi Kasus : SD Inpress Waingapu 3).” *SATI: Sustainable Agricultural Technology Innovation* 523–37.

E-ISSN : 2807-9035
Volume 6, **Number** 1, May 2026
https://doi.org/10.47709/brilliance.v6i1.8264

Fadhil, Muhammad, Wahyu Fuadi, and Maryana. 2025. "Clustering Tingkat Kecanduan Game Mobile Legends Terhadap Keharmonisan Keluarga Menggunakan Metode K-Means." *RABIT : Jurnal Teknologi Dan Sistem Informasi Univrab* 10(2):981–90.

Fazira, Ira, Zahratul Fitri, and Risawandi. 2025. "Optimasi Jumlah Cluster Pada K-Means Clustering Menggunakan Particle Swarm Optimization Untuk Pengelompokan Ukt Mahasiswa." *RABIT : Jurnal Teknologi Dan Sistem Informasi Univrab* 10(2):874–86.

Febriyanti, Ade Eka, Syaiful Zuhri Harahap, and Masrial Masrial. 2024. "Penerapan Data Mining Untuk Evaluasi Data Penjualan Menggunakan Metode Clustering Dan Algoritma Hirarki Divisive Studi Kasus Toko Sembako Pujo." *INFORMATIKA* 15(1):72–86. doi:10.25130/sc.24.1.6.

Hasim Azari, Dwi Hartanti, and Aprilisa Arum Sari. 2024. "Pengelompokan Produksi Padi Dan Beras Provinsi Jawa Timur Dengan Metode Agglomerative Hierarchical Clustering." *Infotek: Jurnal Informatika Dan Teknologi* 7(2):379–89. doi:10.29408/jit.v7i2.26016.

Hendrastuty, Nirwana. 2024. "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa." *Jurnal Ilmiah Informatika Dan Ilmu Komputer (JIMA-ILKOM)* 3(1):46–56. doi:10.58602/jima-ilkom.v3i1.26.

Homepage, Journal, Arfigo Yahya, Rakhmat Kurniawan, Program Studi, Sistem Informasi, Fakultas Sains, and Dan Teknologi. 2025. "Implementasi Algoritma K-Means Untuk Pengelompokan Data Penjualan Berdasarkan Pola Penjualan." 5(1):350–58.

Iin, Johar Nur. 2025. "Perbandingan Kmeans Dan Hierarchical Clustering Untuk Pemetaan Kawasan Rawan Stunting Di Kabupaten Kolaka." *RABIT : Jurnal Teknologi Dan Sistem Informasi Univrab* 10(2):701–16.

Prasetya, Agung Feby, Sintia, and U. L. D. Putri. 2022. "Perancangan Aplikasi Rental Mobil Menggunakan Diagram UML (Unified Modelling Language)." *Jurnal Ilmiah Komputer Terapan Dan Informasi* 1(1):14–18.

Rahmawati, Ruli, Willy Prihartono, and Kota Cirebon. 2025. "Optimasi Stok Dengan Clustering Data Transaksi Penjualan Menggunakan Algoritma K-Means Di Konter Agung Cell." *JITET (Jurnal Informatika Dan Teknik Elektro Terapan)* 13(2).

Rival, Muhammad, Misriani Misriani, and La Ode Bakrim. 2024. "Penerapan Metode Cluster Dalam Data Mining Mengelompokkan Kenakalan Remaja (Studi Kasus Polda Sultra)." *Simkom* 9(1):79–89. doi:10.51717/simkom.v9i1.375.

Sulistiani, Ahmad Rizky Nusantara Habibi, Adrian Maulana, Hidear Talirongan, Anrom G. Abao, Ahmed Mahmoud Zaki Elmalky, and Asno Azzawagama Firdaus. 2025. "Data Analysis of Student Monitoring Using the K-Means Clustering Method." *Indonesian Journal of Modern Science and Technology* 1(2):50–57. doi:10.64021/ijmst.1.2.50-57.2025.

Susilo, Denis Dwi, Shofa Shofiah Hilabi, Bayu Priyatna, and Elfina Novalia. 2024. "Implementasi Data Mining Dalam Pengelompokan Data Pembelian Menggunakan Algoritma K-Means Pada PT.Otomotif 1." *Jutisi : Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi* 13(1):476. doi:10.35889/jutisi.v13i1.1836.

Usino, Wendi. 2024. "Klasterisasi Tingkat Kelayakan Provinsi Dalam Pembangunan Kawasan Industri Menggunakan Algoritma K-Means." 3(September):324–33.

Wibowo, Eko Andri, and Ririn Aryanti. 2025. "Penerapan Metode Clustering K-Means Menggunakan RapidMiner Untuk Klasifikasi Prestasi Siswa Di Sekolah Swasta." *Journal of Information Technology and Informatics Engineering* 1(1):20–24. https://journal.jci.co.id/jitie/article/view/35.