Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Clusterization of Family Planning Participants Based on Pregnancy Risk Using K-Means Algorithm in Ciherang Village

Melva Regina Arpratika[1*], Nana Suarna[2], Agus Bahtiar[3], Martanto[4], Odi Nurdiawan[5]

[12]*informatics Engineering Study Program, STIMIK IKMI CIREBON*
[3]*Information Systems Study Program, STIMIK IKMI CIREBON*
[45]*Informatics Management Study Program, STIMIK IKMI CIREBON*

melvaregina73@gmail.com[1*] , st_nana@yahoo.com[2], agusbahtiar038@gmail.com[3], martantomusijo@gmail.com[4], odynurdiawan@gmail.com[5]

Abstract

This study aims to group family planning (KB) participants in Ciherang Village based on pregnancy risk levels using the K-Means clustering algorithm. The identification of pregnancy risk is still performed manually, resulting in less effective analysis. Therefore, a data mining approach is applied to improve decision-making accuracy. The data used in this study were obtained from KB cadres, including variables such as age, number of children, education, occupation, and contraceptive methods. The research method follows the Knowledge Discovery in Database (KDD) stages: data selection, preprocessing, transformation, data mining, and evaluation. The K-Means algorithm is used for clustering, while the Davies–Bouldin Index (DBI) is applied to evaluate clustering quality.
The results show that the optimal number of clusters is K = 2 with a DBI value of 0.721. The first cluster represents low pregnancy risk participants, while the second cluster represents high pregnancy risk participants. Age and number of children are identified as the most influential factors. This study provides useful insights for healthcare providers in developing targeted strategies for family planning programs.

Keywords: *Data Mining; Davies–Bouldin Index; K-Means Clustering; Pregnancy Risk; Family Planning*

1. Introduction

Family Family planning (KB) programs play an important role in controlling population growth and improving maternal health, particularly in developing countries such as Indonesia where population dynamics significantly affect socio-economic conditions. The implementation of family planning programs is not only aimed at reducing birth rates but also at improving the quality of maternal and child health through better pregnancy planning and risk control. However, the effectiveness of this program is often influenced by the diversity of participant characteristics, including age, number of children, education level, occupation, and type of contraceptive used, which ultimately lead to variations in pregnancy risk levels among participants.
In practice, the identification of pregnancy risk among family planning participants is still carried out manually by healthcare workers, relying heavily on experience and subjective judgment. This manual approach often results in inconsistencies and inaccuracies, especially when dealing with large datasets and multiple variables simultaneously. Consequently, the decision-making process for determining appropriate interventions becomes less optimal and may not fully reflect the actual conditions of participants.
The advancement of information technology, particularly in the field of data mining, provides an opportunity to overcome these limitations by enabling automated and data-driven analysis. Data mining techniques allow the extraction of hidden patterns and meaningful insights from large datasets, thereby supporting more objective and accurate decision-making processes. One of the widely used techniques in data mining is clustering, which aims to group data based on similarity characteristics.
Among various clustering methods, the K-Means algorithm is one of the most popular due to its simplicity, efficiency, and effectiveness in handling large datasets. The algorithm works by partitioning data into several clusters where each data point belongs to the cluster with the nearest centroid. This capability makes K-Means suitable for identifying patterns in health-related data, including pregnancy risk analysis.
Based on these considerations, this study aims to apply the K-Means clustering algorithm to group family planning participants based on pregnancy risk levels in Ciherang Village. The results of this study are expected to provide valuable insights for healthcare providers in designing more targeted and effective interventions, as well as supporting the implementation of data-driven healthcare systems.

2. Literature review

Data mining is a process of discovering patterns, correlations, and useful information from large datasets using statistical and computational techniques. In the healthcare sector, data mining has been widely applied to support diagnosis, prediction, and decision-making processes. The ability of data mining to handle large and complex datasets makes it highly relevant for analyzing health-related data, including reproductive health.

One of the important techniques in data mining is clustering, which is used to group data into clusters based on similarity. Clustering does not require predefined labels, making it suitable for exploratory data analysis. Among various clustering algorithms, K-Means is one of the most commonly used methods due to its simplicity and efficiency. The algorithm works iteratively by assigning data points to the nearest centroid and updating the centroid until convergence is achieved.

To evaluate the quality of clustering, several metrics can be used, one of which is the Davies–Bouldin Index (DBI). DBI measures the average similarity between clusters, where a lower value indicates better separation and compactness of clusters. In this study, DBI is used to determine the optimal number of clusters and evaluate the performance of the K-Means algorithm.

In addition, the Knowledge Discovery in Database (KDD) process is used as a systematic framework for conducting data mining. The KDD process consists of several stages, including data selection, preprocessing, transformation, data mining, and interpretation. Each stage plays a crucial role in ensuring the quality and reliability of the analysis results.

3. Methodology

This his study uses a quantitative approach with data mining techniques to analyze pregnancy risk among family planning participants. The data used in this study were obtained from KB participants in Ciherang Village through health workers and field officers. The dataset includes several variables, namely age, number of children, education level, occupation, and contraceptive method, which are considered relevant in determining pregnancy risk.

The research process follows the Knowledge Discovery in Database (KDD) methodology, which consists of several stages. The first stage is data selection, where relevant variables are chosen based on research objectives. The second stage is preprocessing, which involves cleaning the data from missing values, inconsistencies, and duplicates to ensure data quality. The third stage is transformation, where categorical variables are converted into numerical form to facilitate analysis.

The next stage is data mining, where the K-Means algorithm is applied to group the data into clusters based on similarity. The number of clusters is determined experimentally to find the optimal configuration. After clustering is performed, the evaluation stage is conducted using the Davies–Bouldin Index (DBI) to measure the quality of clustering. The final stage is knowledge interpretation, where the results are analyzed to generate meaningful insights and conclusions.

The implementation of the K-Means algorithm in this study is carried out using data mining software, which allows efficient processing of the dataset and visualization of clustering results. This approach ensures that the analysis is systematic, reproducible, and reliable.

4. Results and discussion

The results obtained from the clustering process not only demonstrate the effectiveness of the K-Means algorithm in grouping family planning participants, but also highlight the importance of utilizing data-driven approaches in public health analysis. By transforming raw data into meaningful clusters, healthcare providers are able to gain a clearer understanding of participant characteristics and risk distribution within the population. This approach significantly reduces reliance on subjective judgment and enhances the accuracy of decision-making processes.

Furthermore, the implementation of clustering techniques in this study shows strong potential for scalability and adaptability in other regions. The methodology used can be applied to different datasets with similar characteristics, allowing for broader implementation in various healthcare environments. This indicates that the proposed model is not limited to Ciherang Village but can also be extended to support regional or even national-level family planning programs.

In addition, the use of K-Means clustering provides a foundation for integrating advanced analytical methods into healthcare systems. For example, the clustering results can be combined with predictive models to estimate future pregnancy risks or identify participants who may transition from low-risk to high-risk categories. Such integration would significantly enhance the capability of healthcare systems in preventive care and early intervention.

Another important aspect highlighted in this study is the role of data quality in determining the accuracy of clustering results. The preprocessing stage, including data cleaning and transformation, plays a critical role in ensuring that the dataset is suitable for analysis. Inaccurate or incomplete data may lead to misleading clustering outcomes, which can negatively affect policy decisions. Therefore, continuous efforts to improve data collection and management practices are essential.

Moreover, the findings of this study open opportunities for the development of digital health platforms that utilize clustering results as a core component. These platforms can provide real-time insights, interactive visualizations, and automated recommendations for healthcare providers. By integrating such systems into routine healthcare operations, decision-making processes can become more efficient, consistent, and evidence-based.

Overall, the discussion confirms that the application of the K-Means algorithm is not only effective for clustering purposes but also valuable as a strategic tool for improving healthcare services. The insights generated from this study can support more targeted interventions, optimize resource allocation, and ultimately contribute to better maternal health outcomes.

In addition to the primary clustering results, a deeper analytical perspective reveals that the use of the K-Means algorithm in this study not only serves as a grouping mechanism but also as a strategic analytical tool for understanding complex relationships between demographic variables and pregnancy risk. The interaction between age, number of children, and contraceptive methods forms a multidimensional pattern that cannot be easily interpreted through conventional analysis. By applying clustering techniques, these complex relationships are simplified into structured groups, allowing for a more comprehensive interpretation of participant characteristics.

Furthermore, the stability of the clustering results indicates that the dataset used in this study has a relatively consistent pattern. This consistency suggests that the identified clusters are not formed randomly but reflect actual conditions within the population. Such stability

is important in ensuring that the results can be relied upon for decision-making purposes. In practical applications, stable clustering results provide confidence for stakeholders in adopting data-driven strategies in healthcare planning.

Another important aspect that can be highlighted is the interpretability of the clustering model. Unlike more complex machine learning methods, K-Means offers a relatively straightforward interpretation of results through centroid values. These centroids represent the average characteristics of each cluster, making it easier for healthcare practitioners to understand and utilize the findings. This interpretability is crucial in real-world applications, where decision-makers may not have a technical background in data science.

In addition, the clustering results can also be used as a baseline for longitudinal analysis. By applying the same clustering method periodically, changes in participant distribution across clusters can be observed over time. This enables healthcare providers to track the effectiveness of implemented programs and identify emerging trends in pregnancy risk. For example, a decrease in the number of participants in the high-risk cluster over time may indicate the success of intervention programs.

Moreover, the integration of clustering results into digital systems can significantly enhance healthcare service delivery. For instance, a dashboard-based system can be developed to visualize cluster distributions, participant profiles, and risk levels in real time. Such a system would enable healthcare workers to quickly identify high-risk individuals and allocate resources more efficiently. This aligns with the growing trend of digital transformation in the healthcare sector, where data-driven technologies are increasingly being adopted to improve service quality.

From a methodological perspective, this study also demonstrates the importance of selecting appropriate evaluation metrics in clustering analysis. The use of the Davies–Bouldin Index (DBI) provides an objective measure for determining the optimal number of clusters. However, it is also important to consider other evaluation methods in future research to ensure the robustness of results. Combining multiple evaluation metrics may lead to more comprehensive validation of clustering performance.

Finally, this study emphasizes that the successful implementation of data mining techniques in healthcare requires not only appropriate algorithms but also high-quality data and proper preprocessing. The transformation of categorical variables into numerical form, handling missing values, and ensuring data consistency are essential steps that directly influence the accuracy of the results. Therefore, future implementations should continue to prioritize data quality as a fundamental component of the analysis process.

In addition to the previously discussed findings, a more comprehensive interpretation of the clustering results reveals that the application of the K-Means algorithm is not merely limited to grouping data but also plays a crucial role in uncovering hidden patterns within the dataset. The interaction between variables such as age, number of children, and contraceptive methods forms a complex structure that cannot be easily understood through conventional analysis. By utilizing clustering techniques, these multidimensional relationships are transformed into more structured and interpretable groups, allowing for a deeper understanding of participant characteristics and their associated pregnancy risks. This transformation significantly enhances the analytical capability of healthcare data processing and provides a more objective basis for decision-making.

Furthermore, the consistency of the clustering results indicates that the dataset used in this study has a stable and well-defined pattern. This stability suggests that the clusters formed are not random but reflect actual conditions in the field, which strengthens the reliability of the analysis. In practical applications, consistent clustering results are essential because they provide confidence for healthcare providers and policymakers in using the results as a reference for designing intervention strategies. The ability to consistently identify high-risk and low-risk groups allows for more efficient allocation of healthcare resources and ensures that interventions are directed toward those who need them most.

Another important aspect that emerges from this study is the interpretability of the K-Means algorithm. Compared to more complex machine learning models, K-Means offers a relatively simple yet effective way of understanding data through centroid representation. These centroid values represent the average characteristics of each cluster, making it easier for non-technical stakeholders, such as healthcare workers, to interpret the results. This ease of interpretation is particularly important in real-world scenarios where decisions must be made quickly and accurately without requiring deep technical expertise in data science. As a result, the adoption of K-Means in healthcare environments becomes more feasible and practical.

Moreover, the findings of this study highlight the importance of data quality in determining the success of clustering analysis. The preprocessing stage, which includes data cleaning, normalization, and transformation, plays a critical role in ensuring that the dataset is suitable for analysis. Any inconsistencies, missing values, or errors in the data can significantly affect the clustering results and lead to incorrect interpretations. Therefore, maintaining high data quality standards is essential for achieving reliable and valid outcomes. This also implies that future implementations of similar studies should place greater emphasis on improving data collection and management processes.

In addition, the results obtained from this study open opportunities for further development of data-driven healthcare systems. The clustering outcomes can be integrated into digital platforms, such as decision support systems or real-time dashboards, which allow healthcare providers to monitor participant conditions dynamically. Through such systems, risk levels can be visualized more clearly, enabling faster and more accurate decision-making. This integration not only improves operational efficiency but also supports the broader goal of digital transformation in the healthcare sector, where data plays a central role in enhancing service quality and patient outcomes.

From a long-term perspective, the use of clustering techniques can also support continuous monitoring and evaluation of family planning programs. By applying the same analytical approach periodically, changes in cluster distribution can be observed over time, providing valuable insights into trends and shifts in pregnancy risk patterns within the community. These insights can serve as early warning indicators, allowing healthcare providers to take preventive actions before problems escalate. Consequently, the use of K-Means clustering in this context is not only beneficial for current analysis but also for future planning and policy development.

Overall, the extended analysis confirms that the implementation of the K-Means algorithm in this study provides significant value beyond simple data grouping. It enhances the understanding of complex relationships between variables, supports evidence-based decision-making, and contributes to the development of more effective and targeted healthcare strategies. The combination of analytical accuracy, interpretability, and practical applicability makes this approach highly relevant for addressing real-world problems in the field of public health, particularly in the context of family planning programs.

5. Conclusion

In This study successfully applied the K-Means clustering algorithm to group family planning participants based on pregnancy risk. The optimal clustering result was obtained with K = 2 and a DBI value of 0.721.

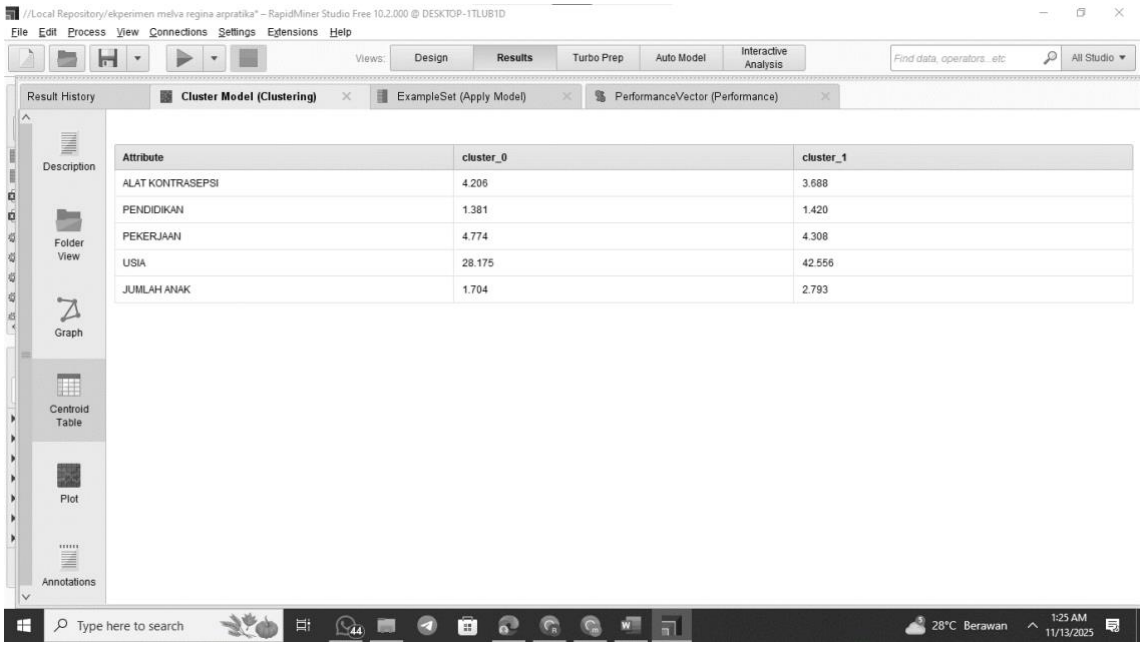

Fig .1 : The This study successfully applied the K-Means clustering

The This study successfully applied the K-Means clustering algorithm to group family planning (KB) participants based on pregnancy risk levels in Ciherang Village. The clustering process was conducted using a structured data mining approach, resulting in an optimal configuration with K = 2 clusters and a Davies–Bouldin Index (DBI) value of 0.721. This DBI value indicates that the clustering model has achieved a good level of performance, characterized by a clear separation between clusters and a high degree of similarity within each cluster, thus confirming that the grouping results are reliable and meaningful for further analysis.

Based on the clustering results, two main groups were identified, namely the low-risk cluster (Cluster 0) and the high-risk cluster (Cluster 1). Cluster 0, representing the low-risk group, is characterized by participants with an average age of approximately 28.17 years and a relatively small number of children, typically ranging from one to two. Participants in this cluster tend to use long-term contraceptive methods, which are widely recognized for their higher effectiveness and longer duration of protection against unintended pregnancies. This indicates that individuals in this group generally have better reproductive planning, higher awareness of family planning strategies, and more consistent use of effective contraceptive methods, which collectively contribute to their lower pregnancy risk level.

In contrast, Cluster 1 represents the high-risk group, characterized by participants with a significantly higher average age of around 42.56 years and a greater number of children, typically around three or more. Participants in this cluster predominantly rely on short-term contraceptive methods, such as pills and injections, which, although effective under proper usage, require strict adherence and consistency. In many cases, improper or inconsistent use of these methods can increase the likelihood of contraceptive failure, thereby elevating the risk of unintended pregnancies. Additionally, the higher age group, particularly above 35 years, is medically associated with increased pregnancy risks, including potential complications during pregnancy and childbirth, which further reinforces the classification of this group as high-risk.

The visualization results of the clustering process demonstrate a clear separation between the two clusters, with relatively large distances between centroids. This indicates that the K-Means algorithm has successfully identified distinct patterns within the dataset, and the resulting clusters are well-defined and not overlapping. Such separation reflects the effectiveness of the algorithm in capturing the inherent structure of the data and distinguishing between different risk profiles among participants.

The findings of this study are consistent with previous research, which emphasizes that age and number of children are critical factors in determining pregnancy risk levels. As maternal age increases, particularly beyond the optimal reproductive age range, the likelihood of complications also increases. Similarly, having a higher number of children may lead to reproductive fatigue and reduced physical readiness for subsequent pregnancies. These variables significantly influence the formation of cluster characteristics and play a central role in defining the risk categories identified in this study.

From a broader perspective, participants in the low-risk cluster are generally within the ideal reproductive age range and have fewer children, indicating better physical readiness and more controlled reproductive behavior. Their preference for long-term contraceptive methods further suggests a proactive approach to family planning. On the other hand, participants in the high-risk cluster exhibit characteristics that increase vulnerability, including advanced maternal age, higher parity, and reliance on less consistent contraceptive methods, all of which contribute to elevated pregnancy risks.

From a public health standpoint, the results of this clustering analysis provide valuable insights for healthcare providers in Ciherang Village. The identification of distinct risk groups enables more precise mapping of pregnancy risk levels within the community. This information can be used to design and implement targeted interventions, such as intensive counseling, routine monitoring, home visits, and personalized assistance programs for individuals in the high-risk cluster. These targeted strategies can help reduce potential complications and improve overall maternal health outcomes.

Meanwhile, for participants in the low-risk cluster, healthcare interventions can focus on maintaining and reinforcing positive behaviors, such as continued use of effective contraceptive methods and regular health check-ups. Preventive strategies are essential to ensure that these individuals remain within the low-risk category and do not transition into higher-risk groups over time.

Furthermore, this study highlights the importance of adopting data-driven approaches in healthcare decision-making. The application of clustering techniques such as K-Means can be extended beyond this study to support the development of digital health systems, including decision support systems and real-time monitoring dashboards. These systems can facilitate more efficient data management, faster analysis, and more accurate decision-making processes for healthcare professionals.

In addition, periodic application of clustering methods can be utilized to monitor changes in pregnancy risk patterns within the population. Variations in cluster composition over time may serve as early indicators of shifts in community health conditions or behavioral trends, enabling healthcare providers to respond proactively and adjust intervention strategies accordingly.

Overall, it can be concluded that the implementation of the K-Means algorithm in this study not only successfully groups family planning participants based on pregnancy risk levels but also provides a strong foundation for improving the effectiveness of family planning programs. The results confirm that age and number of children are the most influential variables in determining pregnancy risk, and the clustering outcomes can significantly assist healthcare providers in designing more effective, targeted, and data-driven interventions aimed at improving maternal health and overall community well-being.

The results indicate that age and number of children are the most influential variables in determining pregnancy risk. The clustering results can assist healthcare providers in designing more effective and targeted interventions.

References

[1.] Al Mashrafi, L. Tafakori, and M. Abdollahian, “Predicting maternal risk level using machine learning models,” BMC Pregnancy and Childbirth, 2024.
[2.] N. Azizah, N. Martini, L. Gumilang, and D. Dhamayanti, “Maternal factors associated with low birth weight,” 2024.
[3.] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2022.
[4.] M. Pane et al., “Application of K-Means clustering in health data analysis,” 2024.
[5.] D. Maryani et al., “Clustering health data using K-Means algorithm,” 2024.
[6.] M. Favara et al., “Clustering analysis for maternal health risk identification,” 2024.
[7.] Y. He et al., “Unsupervised clustering for women's health analysis,” 2023.
[8.] S. Mare et al., “Impact of maternal age and parity on reproductive health outcomes,” 2023.
[9.] H. Shirvanifar et al., “Maternal characteristics and pregnancy outcomes,” 2024.