Journal of Technology and Computer (JOTECHCOM)
Vol. 3, No. 3, August 2026, pp. 282~286
ISSN: 3048-0477 (Media Online)

Application of the K-Means Clustering Algorithm to Categorize Pregnant Women's Knowledge and Attitudes Towards Smoking During Pregnancy

Egi Affandi[1], Baginda Harahap[2], Solianna Harahap[3]
[1,2,3]Informatics Study Program, Faculty of Technology, Battuta University
[1]egi.afandi46@gmail.com

Article Info

Article history:

Received August 29, 2026
Revised August 30, 2026
Accepted August 02, 2026

Keywords:

K-Means clustering
Attitude
Pregnant Women
Smoking
Machine Learning

ABSTRACT

Smoking and secondhand smoke exposure during pregnancy remain a significant maternal and child health risk, yet conventional descriptive analysis tends to treat knowledge and attitude scores separately, making it difficult to identify combined risk profiles for targeted health education. This study applies K-Means Clustering, a data mining technique from computer science, to group pregnant women based on the combination of their knowledge and attitude scores toward smoking during pregnancy. Data were collected from 30 respondents through a questionnaire covering 10 knowledge items and 10 Likert-scale attitude items. The knowledge and attitude percentage scores were standardized (Z-score) and used as clustering features. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, both of which pointed to k = 3 as the most interpretable solution, yielding a final Silhouette Score of 0.687. The resulting clusters were labeled Good (mean knowledge 86.0%, mean attitude 86.5%), Moderate (65.0%, 66.0%), and Poor (43.0%, 41.2%), each containing 10 respondents. The Poor cluster was dominated by housewives with the highest average parity, indicating a priority target group for smoking-related health education programs. These findings demonstrate that K-Means Clustering can serve as a practical decision-support tool for prioritizing maternal health interventions based on combined knowledge-attitude profiles, complementing conventional descriptive statistics commonly used in public health research.

Corresponding Author:

Egi Affandi
Battuta University
Email: egi.afandi46@gmail.com

1. INTRODUCTION

Active smoking and exposure to secondhand smoke during pregnancy constitute significant health risk factors that can adversely affect pregnancy outcomes and fetal development. Numerous health studies indicate that maternal exposure to tobacco smoke is associated with an elevated risk of premature delivery, low birth weight (LBW), miscarriage, impaired fetal lung development, and a higher prevalence of maternal anemia [1], [2]. Although these medical risks are widely recognized, the level of public knowledge and attitudes

regarding the dangers of smoking during pregnancy in Indonesia remains heterogeneous. This variation is largely due to the fact that household smoking habits are still culturally normalized by a segment of the population [2].

Pregnant women's knowledge and attitudes toward smoking are interrelated dimensions that serve as a critical foundation for preventing tobacco exposure during gestation. However, these aspects are inherently multidimensional; an expectant mother might possess adequate knowledge but exhibit an unsupportive attitude, or vice versa. Conventional descriptive analytical approaches, which merely present categorical percentages (e.g., good, fair, or poor) for knowledge and attitude variables independently, fail to capture the comprehensive combined patterns of the respondents' characteristics. Consequently, relying solely on such conventional methods makes it challenging to accurately determine target groups for tailored educational interventions.

To address this methodological gap, computer science approaches, particularly data mining, have been widely adopted for data clustering across various fields, including healthcare [3], [4]. The K-Means clustering algorithm is among the most frequently utilized non-hierarchical clustering methods due to its computational simplicity and proficiency in partitioning data into distinct clusters based on shared similarities [5], [6]. The application of K-Means in health data has proven beneficial in uncovering hidden patterns and trends. Previous implementations include mapping regional disease trends using historical health department data [7], categorizing patient medical records by disease type [8], clustering disease data to identify regional disease characteristics [9], and classifying pharmaceutical inventory based on data similarity [6]. These prior studies demonstrate the efficacy of K-Means as a robust, data-driven decision-support tool across healthcare and other application domains [5]-[9].

Building upon these premises, the current study applies the K-Means clustering algorithm to categorize the profiles of pregnant women based on the combination of their knowledge and attitude scores regarding smoking during pregnancy. In contrast to traditional descriptive analyses that treat knowledge and attitudes as isolated variables, this clustering approach yields a nuanced respondent segmentation. This resulting segmentation can be directly interpreted to establish priorities for targeted health education, such as identifying the specific demographic most in need of immediate intervention due to concurrently low scores in both knowledge and attitudes.

2. METHOD

2.1 Research Design and Workflow

This study employed a quantitative approach utilizing an unsupervised machine learning technique, specifically the K-Means clustering algorithm. The research workflow encompassed six primary stages: (1) data collection via questionnaires, (2) data preprocessing, (3) feature selection and data standardization, (4) determination of the optimal number of clusters, (5) execution of the K-Means algorithm, and (6) evaluation and interpretation of the clustering results.

2.2 Population, Sample, and Instrument

Primary data were acquired through a questionnaire assessing pregnant women's knowledge and attitudes regarding smoking during pregnancy, involving a sample size of 30 respondents. The survey instrument consisted of 10 knowledge-based questions (scored dichotomously: 1 for correct, 0 for incorrect) and 10 attitude-related statements measured on a 4-point Likert scale. Additionally, the respondents' demographic characteristics—including age, highest educational attainment, occupation, and parity—were documented.

2.3 Variables and Clustering Features

The primary variables selected as clustering features were the percentage scores of knowledge and attitudes for each respondent, as the conceptual objective was to group the combined patterns of these two metrics. Demographic variables (age, education, occupation, and parity) were intentionally excluded from the clustering features. Instead, they were utilized during the subsequent profiling phase to interpret the demographic characteristics of the resulting clusters.

2.4 Data Preprocessing

The collected questionnaire data underwent a rigorous completeness check, revealing no missing values among the 30 respondents. Given that the K-Means algorithm relies on Euclidean distance and is inherently sensitive to data scaling, both primary features (knowledge and attitude percentages) were standardized using Z-score standardization (StandardScaler) prior to initiating the clustering computation.

2.5 Determination of the Optimal Number of Clusters

The optimal number of clusters (k) was identified utilizing two complementary validation techniques: the Elbow Method and the Silhouette Score. The Elbow Method involves detecting a sharp decline (the "elbow"

point) in the inertia, also known as the within-cluster sum of squares (WCSS). Concurrently, the Silhouette Score evaluates the quality of cluster separation, yielding values ranging from -1 to 1. Both validation methods were systematically tested across a range of k = 2 to k = 6.

2.6 K-Means Algorithm and Evaluation

Upon establishing the optimal cluster count, the K-Means algorithm was executed utilizing the k-means++ initialization method. To mitigate the risk of the algorithm converging at a local minimum, it was run with 10 independent initialization trials (n_init = 10). The final clustering output was re-evaluated using the Silhouette Score. Ultimately, each cluster was comprehensively profiled based on its members' average knowledge score, average attitude score, and demographic distribution.

3. RESULTS AND DISCUSSION

3.1 Respondents' Characteristics

Among the 30 respondents, the majority were aged between 20 and 30 years (n = 19, 63.3%), with most having completed senior high or vocational school as their highest educational level (n = 18, 60.0%), and predominantly identifying as homemakers (n = 16, 53.3%). The participants' total knowledge scores ranged from 3 to 10 (out of a maximum score of 10), yielding a mean of 6.47 (SD = 1.91). Concurrently, total attitude scores spanned from 12 to 36 (out of a maximum of 40), with an average of 25.83 (SD = 7.70). These metrics indicate a substantial variance among the respondents across both measured variables.

3.2 Distribution of Knowledge and Attitude Categories

Based on standardized categorization thresholds (Good ≥76%, Fair 56-75%, Poor <56%), the distribution for both knowledge and attitude levels in this dataset was evenly distributed. Exactly 10 respondents (33.3%) fell into each of the respective classifications: Good/Positive, Fair, and Poor/Negative. This balanced proportional distribution demonstrates that the data possesses adequate variability for further advanced examination utilizing a clustering approach, standing in contrast to datasets that are entirely homogeneous or heavily skewed toward a single category.

3.3 Determination of the Optimal Number of Clusters

The evaluation results derived from applying the Elbow Method and Silhouette Score across a testing range of k = 2 to k = 6 are detailed in Table 1 and illustrated in Figure 1.

Table 1. Summary of Inertia and Silhouette Scores Across k = 2 to k = 6

Number of Clusters (*k*)	Inertia (WCSS)	Silhouette Score
2	17,181	0,614
3	5,046	0,687
4	3,023	0,686
5	1,887	0,679
6	1,064	0,732

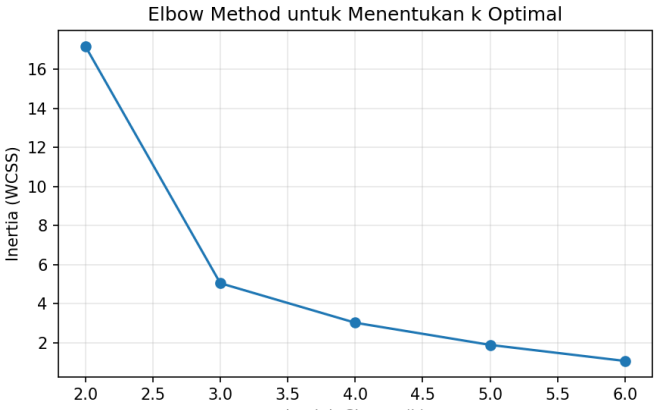

Figure 1. Elbow Method plot.

The Elbow Method plot demonstrates a sharp reduction in inertia from k = 2 to k = 3, which subsequently levels off (exhibiting diminishing returns) from k = 4 onward. This trajectory clearly indicates that the elbow point is located at k = 3. In parallel, the Silhouette Score at k = 3 is relatively robust (0.687),

approaching the maximum value across the tested range, even though k = 6 yielded the highest overall score (0.732). Taking into account the restricted sample size (n = 30) and the necessity for meaningful interpretability—given that k = 3 inherently aligns with the traditional Good/Fair/Poor classification standardly utilized in health knowledge and attitude instruments—this research established k = 3 as the definitive and optimal cluster count.

3.4 K-Means Clustering Results

The execution of the K-Means algorithm configured at k = 3 resulted in the formation of three clusters with a perfectly balanced distribution of members, comprising exactly 10 respondents per cluster. Furthermore, the final computed Silhouette Score of 0.687 signifies a coherent and distinctly separated cluster structure. The centroids for each respective cluster, expressed in their original percentage scales, are detailed in Table 2 and graphically visualized in Figure 2.

Table 2. Cluster Centroid Profiles

Cluster	n	Mean Knowledge (%)	Mean Attitude (%)	Interpretative Label
0	10	65,0	66,0	Poor Knowledge & Attitudes
1	10	43,0	41,2	Fair Knowledge & Attitudes
2	10	86,0	86,5	Good Knowledge & Attitudes

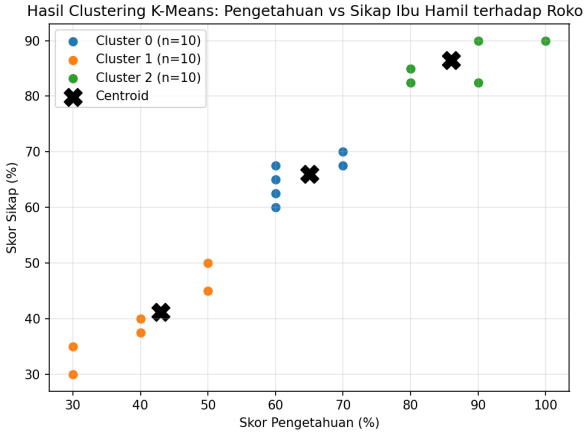

Figure 2. Visualization of K-Means Clustering Results (Knowledge Score vs. Attitude Score)

3.5 Discussion

These findings are consistent with prior research employing the K-Means algorithm within the healthcare sector, corroborating that clustering is highly effective in uncovering latent patterns and facilitating the prioritization of medical services or interventions. Previous applications have ranged from mapping disease incidence trends for regional health planning to categorizing hospital patients for enhanced service management. By applying this methodology to the knowledge and attitudes of expectant mothers, the current study extends the utility of the clustering approach into the realm of data-driven health promotion and behavioral segmentation, moving beyond strictly clinical or administrative datasets.

However, it is imperative to acknowledge that this study utilized a relatively small sample size (n = 30). Therefore, while the clustering outcomes demonstrate robust statistical separation (Silhouette Score = 0.687), these results should primarily be interpreted as exploratory or pilot findings. The consistency of this three-cluster configuration warrants further validation through studies involving larger and more demographically diverse cohorts before these conclusions can be safely generalized to inform broad-scale public health policies.

4. CONCLUSION

This study successfully applied the K-Means clustering algorithm to categorize 30 expectant mothers based on their knowledge and attitude scores regarding smoking during pregnancy. Guided by the Elbow Method and Silhouette Score evaluations, the optimal number of clusters was determined to be k = 3. The final model achieved a Silhouette Score of 0.687, indicating a robust and well-defined separation of the clusters.

The three emergent clusters can be interpreted as follows: (1) the Good Knowledge and Attitudes Cluster (with mean scores of 86.0% and 86.5%, respectively), (2) the Fair Knowledge and Attitudes Cluster (mean scores of 65.0% and 66.0%), and (3) the Poor Knowledge and Attitudes Cluster (mean scores of 43.0%

and 41.2%). Each group contains an evenly distributed subset of exactly 10 respondents. Notably, the cluster exhibiting poor knowledge and attitudes was identified to be predominantly composed of homemakers with the highest average parity. Consequently, this specific demographic is highly recommended as the primary target for prioritized health education programs concerning the perils of tobacco exposure during gestation.

For future research, it is advisable to utilize a substantially larger sample size that is more geographically and demographically diverse. Furthermore, incorporating the respondents' demographic characteristics as additional clustering features, alongside comparing the performance of K-Means with alternative algorithms—such as K-Medoids or Hierarchical Clustering—would be highly beneficial. These subsequent steps are essential to rigorously validate and verify the consistency of the identified grouping patterns.

REFERENCES

[1] M. Alamanda, "Bahaya Merokok pada Wanita Usia Produktif, Ibu Hamil, dan Ibu Menyusui," Alomedika, Feb. 29, 2024. [Online]. Available: https://www.alomedika.com/bahaya-merokok-pada-wanita-usia-produktif-ibu-hamil-dan-ibu-menyusui. [Accessed: Jul. 30, 2026].

[2] Kementerian Kesehatan Republik Indonesia, "Pengaruh Paparan Asap Rokok pada Ibu Hamil," Direktorat Jenderal Kesehatan Lanjutan, Nov. 22, 2022. [Online]. Available: https://keslan.kemkes.go.id/view_artikel/1853/pengaruh-paparan-asap-rokok-pada-ibu-hamil. [Accessed: Jul. 30, 2026].

[3] M. H. Adhikari Egodawele, D. Sedera, and V. Bui, "A Systematic Review of Digital Transformation Literature (2013–2021) and the Development of an Overarching A-Priori Model to Guide Future Research," in Proc. Australasian Conf. Inf. Syst. (ACIS), Melbourne, Australia, Dec. 2022, Paper 89, pp. 1–12. [Online]. Available: https://aisel.aisnet.org/acis2022/89

[4] Y. Kang, N. Choi, and S. Kim, "Searching for New Model of Digital Informatics for Human–Computer Interaction: Testing the Institution-Based Technology Acceptance Model (ITAM)," Int. J. Environ. Res. Public Health, vol. 18, no. 11, Art. no. 5593, 2021, doi: 10.3390/ijerph18115593.

[5] P. Marpaung, I. Febrian, and W. Putri, "Penerapan Data Mining dalam Menentukan Tingkat Kedisiplinan Karyawan Perhotelan Menggunakan Algoritma K-Means Clustering," J. Ilmu Komput. dan Sist. Inf. (JIKOMSI), vol. 7, no. 1, pp. 167–172, 2024, doi: 10.55338/jikomsi.v7i1.2905.

[6] F. P. Pangestu, N. Y. Yasin, R. C. Hasugian, and Yunita, "Penerapan Algoritma K-Means untuk Mengklasifikasi Data Obat," J. Sisfokom (Sist. Inf. dan Komput.), vol. 12, no. 1, pp. 53–62, 2023, doi: 10.32736/sisfokom.v12i1.1461.

[7] S. A. D. Darmawan and Karmilasari, "Penerapan Metode K-Means Clustering dan Simple Moving Average untuk Memprediksi Jenis Penyakit di Provinsi Jawa Timur," J. Teknol. Inf. dan Ilmu Komput., vol. 11, no. 4, pp. 877–886, 2024, doi: 10.25126/jtiik.1148703.

[8] R. Ordila, R. Wahyuni, Y. Irawan, and M. Y. Sari, "Penerapan Data Mining untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit dengan Algoritma Clustering (Studi Kasus: Poli Klinik PT. Inecda)," J. Ilmu Komput., vol. 9, no. 2, pp. 148–153, 2020, doi: 10.33060/jik/2020/vol9.iss2.181.

[9] M. Minarni, E. I. Sari, A. Syahrani, and P. Mandarani, "Klasterisasi Penyakit Menggunakan Algoritma K-Medoids pada Dinas Kesehatan Kabupaten Agam," J. Nas. Pendidik. Tek. Inform. (JANAPATI), vol. 10, no. 3, pp. 137–146, 2021, doi: 10.23887/janapati.v10i3.34904.