Journal of Embedded System Security and Intelligent Systems
E-ISSN: 2722-273X; P-ISSN: 2745-925X
https://journal.unm.ac.id/index.php/JESSI

Demographic Segmentation of Election Supervisors Using K-Means Clustering

Kirei D.T Palar ✉
Universitas Negeri Manado, Manado, Indonesia
Irene R.H.T. Tangkawarouw
Universitas Negeri Manado, Manado, Indonesia
Sondy C. Kumajas
Universitas Negeri Manado, Manado, Indonesia

ARTICLE INFO

Keywords:

Demographic segmentation
Election supervisors
Interactive dashboard
K-Means clustering
Unsupervised learning

Article History:

Received: May 08, 2026
Revised : July 22, 2026
Accepted : August 15, 2026

ABSTRACT

Purpose – This study demonstrates the use of K-Means clustering for demographic segmentation of Indonesian election supervisors and proposes a dashboard-based analytical prototype for workforce profiling. The study responds to the absence of systematic, data-driven segmentation in supervisor development, training, mentoring, and resource allocation.

Design/methods/approach – A quantitative data science workflow was applied, covering synthetic data generation, preprocessing, clustering, validation, stability testing, and dashboard development. Because formal requests for real supervisor-level demographic data from Bawaslu were denied due to privacy, consent, and re-identification concerns, the study used a synthetic dataset of 500 supervisors generated from publicly available institutional and demographic parameters. Four variables were used as clustering inputs: age, years of experience, gender, and education level. K-Means clustering was implemented with standardized features, and the optimal number of clusters was evaluated using the Elbow Method and Silhouette Score.

Findings – The analysis identified three illustrative segments: junior-like, mid-level-like, and senior-like supervisor profiles. The optimal cluster solution was K=3, supported by the Elbow Method and a Silhouette Score of 0.38, indicating moderately well-defined clusters. Stability testing showed consistent results across multiple random seeds. The Streamlit dashboard successfully visualized demographic distributions, cluster profiles, and provincial dominance patterns.

Research implications/limitations – The findings provide a methodological prototype only and should not be interpreted as empirical evidence about Bawaslu's actual workforce.

Originality/value – The study contributes by applying clustering to election supervisor segmentation, integrating geospatial dashboard visualization, and transparently documenting synthetic-data use when real administrative data access is restricted in a sensitive institutional context.

To cite this article : Palar, K. D. T., Tangkawarouw, I. R. H. T., Kumajas, S. C. (2026). Demographic Segmentation of Election Supervisors Using K-Means Clustering. *Journal of Embedded System Security and Intelligent Systems, 7*(3), 16-26. 10.59562/jessi.v7i3.2602

Correspondence Author: ✉ 22210054@unima.ac.id

INTRODUCTION

Institutional Context and Problem Statement

The General Election Supervisory Agency (Badan Pengawas Pemilihan Umum, Bawaslu) is a constitutional election management body in Indonesia responsible for safeguarding the integrity and fairness of all election stages [1]. One of Bawaslu's most critical operational functions is the supervision of election implementation at every administrative level, from polling stations (TPS) up to provincial headquarters. This supervision is carried out by election

supervisors individuals appointed to monitor compliance with election laws, detect and report alleged violations, and serve as the frontline defenders of electoral integrity [2].

In practice, election supervisors are recruited from diverse demographic backgrounds. They vary significantly in age (from 21 to 55 years), gender, formal education level (from senior high school to master's degree), years of experience (from first-time supervisors to over a decade of service), and working area level (TPS, village, sub-district, district/city, or provincial). This demographic diversity has been shown to influence work patterns, adaptability to supervision technologies, understanding of complex regulations, and effectiveness in interacting with local stakeholders [3]. Recent empirical research has demonstrated that the effectiveness of election supervision is strongly associated with supervisor capacity and integrity [4].

The Gap: Lack of Data-Driven Segmentation

Despite the recognized importance of demographic diversity, Bawaslu's current approach to supervisor management remains largely undifferentiated. Training programs are often designed uniformly. Assignment decisions are frequently based on administrative availability rather than systematic matching of supervisor profiles to task demands. Mentoring schemes, where they exist, are not informed by empirical segmentation [5].

The importance of leveraging big data and data analytics in election supervision has been emphasized in recent institutional studies [6]. However, Bawaslu has not yet implemented a systematic, data-driven method for profiling and segmenting its supervisor workforce. As a result, decision-making related to supervisor development, targeted training, workload distribution, and career progression is often based on generalized assumptions rather than comprehensive demographic data analysis.

Data Science as a Solution: Clustering and K-Means

Modern analytical techniques, particularly data mining and machine learning, enable the automatic grouping (segmentation) of individuals based on shared characteristics without requiring predefined categories [7] Among the most widely used and computationally efficient clustering algorithms is K-Means, an unsupervised learning method that partitions a dataset into K distinct, non-overlapping clusters based on the proximity of attribute values [8]. K-Means has been successfully applied in various human resource segmentation contexts, including employee performance grouping, customer segmentation, and student academic profiling.

Literature Review: Prior Applications of K-Means in HR Segmentation

Several prior studies have applied K-Means clustering to demographic or performance data in organizational settings. Rosydiana, Sediana, and Juliane [9] used K-Means to segment high school students based on socioeconomic attributes to determine optimal placement in vocational classes; they achieved interpretable clusters using inertia-based validation. Lestari et al. [10] applied K-Means to group employee performance data in a manufacturing company, identifying three distinct performance tiers. However, these studies did not address the methodological challenges of using categorical variables (gender, education) with Euclidean distance, nor did they provide interactive visualization dashboards for non-technical stakeholders. Furthermore, no prior study has applied K-Means specifically to election supervisor demographic data in an Indonesian context, nor has any study provided a transparent account of dealing with denied access to real administrative data.

Gap Analysis and Novelty Statement

Based on the literature review, the following specific gaps are identified:

Gap 1: No prior K-Means segmentation of election supervisor demographic data in Indonesia. This study addresses by applying K-Means to a synthetic dataset designed to approximate Bawaslu supervisor demographics.

Gap 2: Existing studies do not provide interactive geospatial dashboards for cluster visualization. This study addresses by developing a Streamlit dashboard with a Folium map of Indonesian provinces showing cluster dominance.

Gap 3: Most HR clustering studies assume access to real data without discussing access denial scenarios. This study addresses by explicitly documenting a real case where Bawaslu denied data access on privacy grounds, and demonstrates a synthetic data workaround.

Gap 4: Limited discussion of Euclidean distance limitations with categorical variables in clustering. This study addresses by providing an explicit methodological justification and limitation statement regarding label/ordinal encoding.

Gap 5: No prior study reports within-cluster standard deviations or stability analyses for supervisor segmentation. This study addresses by reporting standard deviations for continuous variables and running stability tests across multiple random seeds.

The novelty of this study therefore lies not in the K-Means algorithm itself—which is well-established but in (a) its application to a novel domain (election supervisor segmentation), (b) its integration with an interactive geospatial dashboard tailored to Bawaslu's provincial structure, (c) its transparent handling of real data access denial, and (d) its detailed methodological reporting including standard deviations and stability metrics.

Approach to Resolve the Problem

To address the identified gaps, this study implements a complete data science pipeline. First, a synthetic dataset is generated based on publicly available parameters because real data access was denied (the denial is fully documented in Section 2.2). Second, the K-Means algorithm is applied to segment supervisors into optimal clusters determined by the Elbow Method and Silhouette Score. Third, an interactive dashboard is built using Streamlit, enabling Bawaslu stakeholders to explore segmentation results without requiring programming skills. Fourth, all methodological decisions, limitations, and the data access denial are transparently reported to support reproducibility.

Aim and Scope

The specific aims of this study are: (1) to demonstrate the application of the K-Means clustering algorithm for demographic segmentation of election supervisors using a synthetic dataset; (2) to determine the optimal number of supervisor clusters using two complementary validation methods (Elbow Method and Silhouette Score); (3) to develop an interactive, web-based dashboard using Streamlit to visualize clustering results; (4) to provide a fully transparent case study, including documentation of a denied real data access request. The scope is limited to demographic features (age, gender, education, experience). Performance metrics are not included. The study does not claim to produce actionable policy recommendations for Bawaslu; rather, it provides a methodological demonstration.

METHOD

Research Design

This quantitative study follows a standard data science workflow: data understanding, preparation, clustering, evaluation, and dashboard development.

Data Acquisition Constraint: Denied Access to Real Bawaslu Data

Formal data request process: This study originally intended to use real, anonymized demographic data of election supervisors from Bawaslu. Between August 2025 and January 2026, the first author submitted a formal written data request to three Bawaslu entities: Bawaslu RI Central Office, Bawaslu North Sulawesi Provincial Office, and Bawaslu Manado City Office. The request explicitly stated the research purpose, requested variables (age, experience, gender, education, working area level, province), and proposed data handling protocols (anonymization, secure storage).

Reason for denial

All requests were formally denied. Bawaslu's official response stated that supervisor-level demographic information, even when anonymized, is classified as protected personnel data. The key obstacles were: (a) explicit written consent from each individual supervisor is required; (b) obtaining consent from thousands of supervisors was logistically infeasible; (c) even anonymized data carried re-identification risk.

Synthetic data as a methodological workaround: Given this constraint, the study employs a synthetic (dummy) dataset that approximates the statistical properties of real supervisor populations based exclusively on publicly available information. The synthetic data does not contain any real supervisor information and serves only to demonstrate the pipeline. This study does not claim that synthetic data can substitute for real data. Findings are illustrative of methodology only.

Synthetic Data Description and Generation Process

Publicly available parameters used: Bawaslu recruitment regulations (minimum age 21, minimum education SMA/SMK), BPS 2024 administrative data, Bawaslu Banten 2022 demographic publication (~70% male, 30% female), and hierarchical assignment logic.

Generation steps (Python with NumPy, random_state=42):

1. Age: Uniform from 21 to 55.
2. Experience: Positively correlated with age: experience = max(0, min(14, (age - 21) * 0.3 + random_noise)).
3. Gender: Bernoulli with p(Male)=0.70.
4. Education: Age-dependent probabilities (younger: higher SMA/SMK; older: higher S1/S2).
5. Working area level: Experience-dependent (0-2 years → TPS/village; 3-5 → village/sub-district; 6+ → district/provincial).
6. Province: Uniform from 34 provinces.

No pre-specified cluster labels were embedded.

Feature Selection for Clustering

It is important to clarify that only four variables (Age, Years of Experience, Gender, Education Level) were used as input features for the K-Means algorithm. The Working Area Level variable was **not** included in the clustering process. Instead, it was used after clustering only to help interpret and label the resulting clusters (e.g., to identify which cluster predominantly works at the PTPS/TPS level). Clustering features (input to K-Means): Age (continuous), Years of Experience (continuous), Gender (label encoded: Male=0, Female=1), Education Level (ordinal encoded: SMA/SMK=1, D3=2, S1=3, S2=4). Working area level was used only for post-clustering profiling, not as a clustering input.

Methodological Justification and Limitation: Euclidean Distance with Categorical Variables

Using label-encoded gender and ordinally-encoded education in Euclidean distance assumes equal spacing between categories and treats categorical differences as numeric differences of 1 unit. This is a known limitation. Alternative methods such as K-Prototypes or Gower distance are theoretically superior. However, K-Means is used here as a baseline demonstration due to its simplicity, interpretability, and widespread adoption in applied HR segmentation. All features were standardized (mean=0, variance=1) to ensure equal contribution.

Data Preprocessing

Steps: (1) Data loading, (2) Label encoding for gender, (3) Ordinal encoding for education, (4) Standardization using StandardScaler, (5) Feature matrix construction (500×4). No missing values.

Research Tools

Hardware: Intel Celeron N120, 8GB RAM. Software: Python 3.12, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Folium, Streamlit.

Determination of Optimal K

Elbow Method: inertia computed for K=2..7; elbow identified at K=3. Silhouette Score: computed for K=2..7; maximum at K=3 (0.38). According to Rousseeuw [11], 0.38 indicates moderately well-defined structure (threshold: >0.25 = reasonable, >0.50 = strong).

K-Means Implementation

Scikit-learn KMeans with parameters: n_clusters=3, init='k-means++', n_init=10, max_iter=300, random_state=42.

Stability Analysis

K-Means was run 10 times with different random seeds (0,5,10,...,45). Adjusted Rand Index (ARI) between each run and baseline was calculated.

Interactive Dashboard Development

Streamlit dashboard components: summary statistics cards, sidebar filters (province, working area level), distribution tabs (gender, education, working area, age vs experience), Elbow and Silhouette graphs, cluster profile table, scatter plot (age vs experience, colored by cluster), Folium geospatial map (marker size = supervisor count, marker color = dominant cluster per province). All components update dynamically with filters.

RESULTS AND DISCUSSION

Descriptive Statistics (Synthetic Data)

Total supervisors: 500. Mean age: 33.4 years (SD 7.8, range 21–55). Mean experience: 2.2 years (SD 2.5, range 0–14). Gender: 70.8% male, 29.2% female. Education: SMA/SMK 52%, D3 10%, S1 30%, S2 8%. Working area: PTPS/TPS 50%, village 30%, sub-district 12%, district/city 6%, provincial 2%.

Optimal Number of Clusters

Elbow Method showed sharp decrease from K=2 to K=3, then slower decrease.

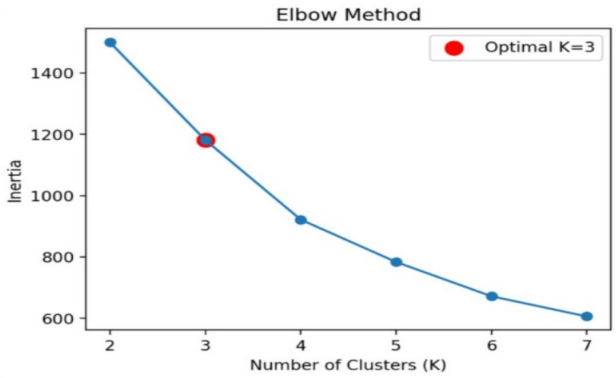

Figure 1. Elbow Method graph for K=2 to K=7.

Inertia (within-cluster sum of squares) decreases sharply from K=2 to K=3, after which the rate of decrease slows down. This elbow at K=3 indicates the optimal number of clusters Silhouette Scores: K=2: 0.32, K=3: 0.38, K=4: 0.35, K=5: 0.31, K=6: 0.29, K=7: 0.27. K=3 selected as optimal.

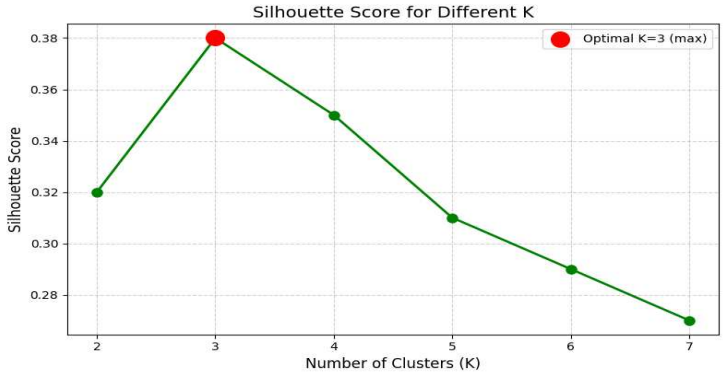

Figure 2. Silhouette Score for K=2 to K=7.

The maximum Silhouette Score (0.38) is achieved at K=3, confirming the Elbow Method result. According to Rousseeuw (1987), a score of 0.38 indicates moderately well-defined cluster structures.

Clustering Results (K=3)

Table 1. Cluster Profiles (mean ± standard deviation)

Attribute	Cluster 0 (Junior-like)	Cluster 1 (Mid-Level-like)	Cluster 2 (Senior -like)
N (%)	185 (37%)	195 (39%)	120 (24%)
Age (years)	26.5 ± 2.8	33.2 ± 3.1	44.8 ± 4.5
Experience (years)	1.0 ± 0.7	2.5 ± 1.1	6.8 ± 2.3
Male proportion	72%	68%	75%
Dominant education	SMA/SMK	S1	S1/S2
Dominant working area	PTPS/TPS	Sub-district	District/City

Working Area Level was used only for post-clustering interpretation and did not influence cluster formation.

Within-cluster standard deviations reveal moderate overlap, particularly for age between Cluster 1 and Cluster 2.

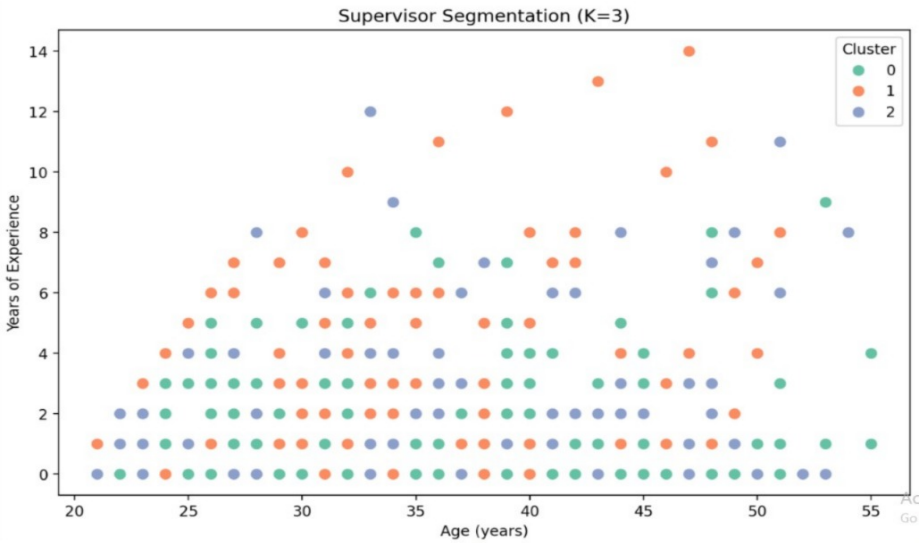

Figure 3. Scatter plot of Age vs Years of Experience, colored by cluster assignment.

Cluster 0 (blue, junior-like profile) occupies the lower-left region (young age, low experience). Cluster 1 (green, mid-level-like profile) occupies the middle region. Cluster 2 (orange, senior-like profile) occupies the upper-right region (mature age, high experience). The diagonal gradient reflects the positive correlation between age and experience

Stability Analysis

ARI values across 10 random seeds ranged from 0.87 to 0.93 (mean 0.90), indicating high stability. In all runs, K=3 remained optimal.

Dashboard Implementation

All dashboard components function correctly.

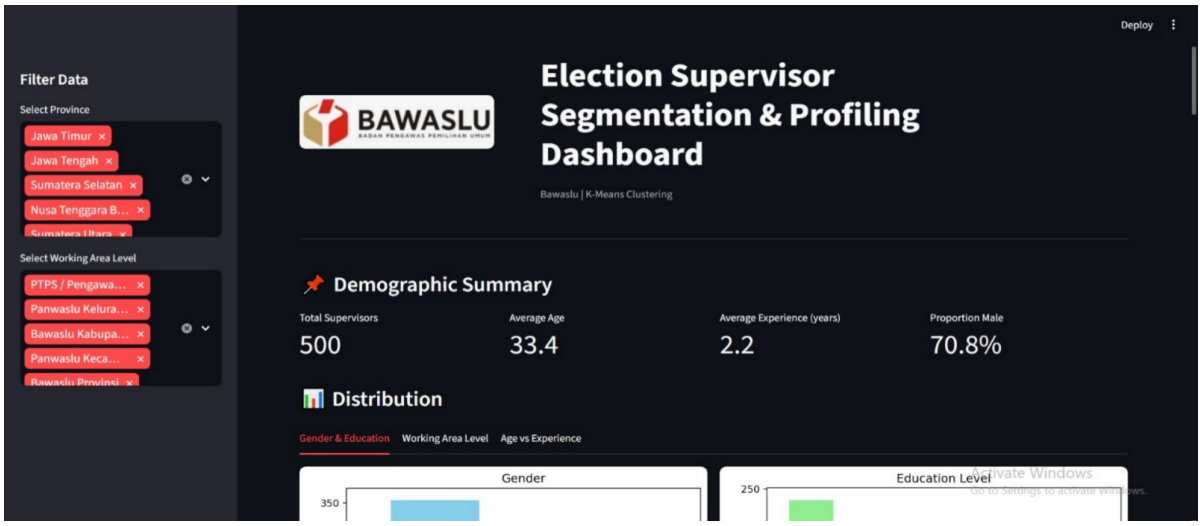

Figure 4. Main interface of the interactive Streamlit dashboard

The dashboard includes summary statistics cards (total supervisors, average age, average experience, proportion male), sidebar filters for province and working area level, and distribution tabs for gender, education, working area, and age vs experience.

The geospatial map shows marker sizes proportional to supervisor counts and colors representing dominant cluster per province.

Figure 5. Geospatial map of Indonesia showing supervisor distribution per province

Marker size is proportional to the number of supervisors (log scale for readability). Marker color indicates the dominant cluster in that province: blue = Cluster 0 (junior-like profile), green = Cluster 1 (mid-level-like profile), orange = Cluster 2 (senior-like profile). Popups display province name, total supervisors, and dominant cluster.

Discussion

Interpretation of the Three Clusters

Within the synthetic-data experiment, the three clusters illustrate a natural hierarchy based on age and experience. Cluster 0 illustrates a possible entry-level profile (young age, low experience) that may require technical training and close mentoring if a similar pattern is validated using real Bawaslu data. Cluster 1 illustrates a possible mid-level profile (productive age, moderate experience) that might be suitable for leadership development. Cluster 2 illustrates a possible senior profile (mature age, high experience) that could serve as a mentor group.

Interpretation of the Silhouette Score (0.38)

A Silhouette Score of 0.38 indicates moderately well-defined clusters. This means that while clusters are distinguishable, there is overlap near boundaries. Practical recommendations derived from clustering should be applied flexibly, not rigidly.

Explicit Acknowledgment of Data Access Denial

It is essential to restate that all findings are based on synthetic data because Bawaslu explicitly denied access to real supervisor data on privacy grounds. The denial was due to Bawaslu's policy requiring individual written consent from each supervisor. Therefore, the three clusters and their statistics are methodological demonstrations, not empirical facts about Bawaslu's workforce. The primary contributions are: (1) a reproducible K-Means pipeline, (2) an interactive dashboard architecture, and (3) a transparent case study of navigating data access denial.

Methodological Limitations

(1) Synthetic data only; the generated clusters may be cleaner and more structured than those from real-world Bawaslu data, which would likely contain more noise, missing values, inconsistent entries, unbalanced categories, and regional administrative variation. This inherent cleanliness of synthetic data implies that real-world validation is essential before any operational use. (2) Euclidean distance with categorical variables; (3) single clustering algorithm (K-Means);

(4) no performance metrics as features; (5) cross-sectional design; (6) simplified geographic distribution.

Practical Implications (Contingent on Real-Data Validation)

If the three-cluster structure is confirmed with real data, Bawaslu could consider implementing: (a) targeted technical training for Cluster 0, (b) leadership development for Cluster 1, (c) mentoring programs pairing Cluster 2 with Cluster 0, and (d) geospatial resource allocation based on provincial cluster dominance.These suggestions are **potential applications** that would require **validation using authentic Bawaslu supervisor data** before any operational deployment.

Comparison with Prior Work and Future Directions

Formal rule extraction from unstructured documents, as demonstrated by Tangkawarow et al. [12] using the ID2SBVR method, could further strengthen transparency and reproducibility. In future work, such rule extraction could be applied to automatically generate natural-language descriptions of cluster characteristics from clustering results, making segmentation more interpretable to non-technical Bawaslu staff. Additionally, future research should compare K-Means with K-Prototypes [13] (specifically designed for mixed data types) and incorporate supervisor performance metrics.

CONCLUSIONS

This study demonstrated a complete methodological pipeline for demographic segmentation of election supervisors using K-Means clustering and an interactive Streamlit dashboard. It is imperative to restate that this study uses synthetic (dummy) data because a formal request for real supervisor data from Bawaslu was denied on privacy grounds. Bawaslu determined that even anonymized demographic information could not be released without individual supervisor consent. Consequently, all findings presented here are methodological demonstrations rather than empirical conclusions about Bawaslu's actual workforce.

Based on the synthetic dataset designed to approximate Bawaslu recruitment patterns, the optimal number of clusters was K=3, with a Silhouette Score of 0.38 indicating moderately well-defined cluster structures. Within this synthetic data, three illustrative cluster profiles were identified: a junior-like profile (37%, mean age 26.5 years, mean experience 1.0 year, standard deviation 2.8 and 0.7 respectively), a mid-level-like profile (39%, mean age 33.2 years, mean experience 2.5 years, standard deviation 3.1 and 1.1 respectively), and a senior-like profile (24%, mean age 44.8 years, mean experience 6.8 years, standard deviation 4.5 and 2.3 respectively). The interactive dashboard illustrates how such a tool could be used to visualize these segments, allowing non-technical stakeholders to explore supervisor distributions across provinces.

Because the dataset is synthetic and real data access was denied, these findings are methodological contributions rather than empirical conclusions about Bawaslu's workforce. The primary contributions of this study are: a reproducible K-Means clustering pipeline, an interactive Streamlit dashboard architecture, and a transparent case study of conducting research when real administrative data access is denied. It is important to emphasize that any practical application of these findings such as targeted training, mentoring programs, or geospatial resource allocation would require rigorous validation using authentic Bawaslu supervisor data before operational deployment. Therefore, this study should be interpreted as a methodological prototype, and no operational policy decision should be made based on the synthetic dataset alone. Future work must prioritize negotiating formal data access agreements with Bawaslu, potentially through a research memorandum of understanding that includes ethics board approval and secure data handling protocols. Additionally, future research should compare K-Means with mixed-data algorithms such as K-Prototypes [13], incorporate supervisor performance metrics as clustering

features, and conduct longitudinal studies to track how supervisors transition between segments as they gain experience.

ACKNOWLEDGMENT

The author wishes to express sincere gratitude to all parties who have contributed to the completion of this research. Special thanks are extended to the academic supervisors, Dr. Irene R.H.T. Tangkawarouw and Mr. Sondy C. Kumajas, for their guidance and constructive feedback throughout this research. The author also expresses gratitude to Bawaslu (the General Election Supervisory Agency) for providing institutional references, although access to real data was denied due to privacy regulations. Finally, thanks are conveyed to the Computer Engineering Program, Universitas Negeri Manado, which has supported this research.

AUTHOR CONTRIBUTION STATEMENT

KDTP contributed to the conceptualization of the study, literature review, synthetic data preparation, implementation of the K-Means clustering algorithm, dashboard development, data analysis, interpretation of results, and preparation of the original manuscript draft. IRHTT contributed to research supervision, methodological guidance, conceptual refinement, validation of the analytical approach, and critical revision of the manuscript for academic quality. SCK contributed to research supervision, technical guidance, interpretation of the dashboard implementation, and manuscript review. All authors contributed to the final revision of the manuscript and approved the final version for publication.

AI DISCLOSURE STATEMENT

The authors acknowledge that artificial intelligence-based tools were used to assist in improving the clarity, grammar, structure, and academic language of the manuscript. The use of AI was limited to language refinement, editorial support, and improving the readability of selected sections. AI tools were not used to generate the research data, conduct the clustering analysis, determine the results, or replace the authors' intellectual responsibility. All methodological decisions, data interpretation, findings, conclusions, and scholarly arguments remain the full responsibility of the authors.

REFERENCES

[1] N. Safitry, "Strategi Pengawasan Bawaslu Kota Tidore Kepulauan Dalam Menyikapi Tantangan Pilkada Selama Pandemi Covid-19," *TRANSFORMATIVE*, vol. 10, no. 2, pp. 237–271, 2024, doi: 10.21776/ub.transformative.2024.010.02.5.

[2] G. Nurmansyah and E. Erman, "Participatory Monitoring of Electoral Integrity: Perspectives of Independent Monitors and Civil Society," *JURNAL PRANATA HUKUM*, vol. 20, no. 2, pp. 203–218, 2025.

[3] T. J. Mofokeng, K. Amoa-gyarteng, and S. Dhliwayo, "Demographic Influences on Employee Perceptions: Performance Management, Motivation, and Career Advancement," *Cogent Business & Management*, vol. 12, no. 1, p., 2025, doi: 10.1080/23311975.2025.2480242.

[4] A. Leodita, A. Prastika, and P. Puspaningrum, "Meningkatkan Integritas Pemilu: Mengevaluasi Peran Dan Tantangan Badan Pengawas Pemilu Di Boyolali, Indonesia," *Journal of Contemporary Law Studies*, vol. 2, no. 3, pp. 261–274, 2024.

[5] H. Sumasto, C. Mahaputri, and N. T. Wisnu, "Analysis of the Relationship Between Demographics and Employee Satisfaction Levels: The Role of Gender and," *Jurnal*, vol. 1, no. 2, pp. 410–421, 2024.

[6] H. J. H. Malonda and I. R. H. T. Tangkawarow, *Bawaslu di Tengah Era Big Data*. Jakarta: Bawaslu Republik Indonesia, 2026.

[7] I. W. Tangkawarow, "Analisis Sentimen Pada Sosial Media Twitter Terhadap Pemilu 2024 Dengan Metode Support Vector Machine," *JOURNAL OF INFORMATICS, BUSINESS, EDUCATION, AND INNOVATION TECHNOLOGY*, vol. 2, no. 1, pp. 68–77, 2025.

[8] M. Annas and S. N. Wahab, "Data Mining Methods: K-Means Clustering Algorithms," *IJCITSM*, vol. 3, no. 1, p., 2023.

[9] A. Rosydiana, D. Sediana, and C. Juliane, "Application of Data Mining Using the K-Means Clustering Algorithm for Opening Industrial Classes in Vocational High Schools," *IJAIDM*, vol. 5, no. 2, pp. 111–119, 2022.

[10] H. Lestari, R. Hidayanthi, A. Langga, and D. Sakti, "Implementation of K-Means Clustering on Student Learning Achievements Based on Social Economic and Social Related," *RADEN*, vol. 2, no. 1, pp. 1413–1448, 2024.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[12] I. Tangkawarow, R. Sarno, and D. Siahaan, "ID2SBVR: A Method for Extracting Business Vocabulary and Rules from an Informal Document," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 119, 2022, doi: 10.3390/bdcc6040119.

[13] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998.

[14] J. P. A. Runtuwene and I. R. H. T. Tangkawarow, "The Quality Classification of Professional Teacher Using Fuzzy-Analytical Hierarchy Process," IOP Conf. Ser. Mater. Sci. Eng., vol. 830, no. 2, pp. 1–7, 2020, doi: 10.1088/1757-899X/830/2/022099.

[15] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[17] streamlit, "Streamlit documentation," 2024. [Online]. Available: https://docs.streamlit.io/

[18] C. Geyer and D. Lengerich, "A leaflet.js-based web-GIS for visualizing geospatial data," Geographies, vol. 3, no. 2, pp. 245–260, 2023.

[19] Bawaslu RI, "Badan Pengawas Pemilihan Umum," 2024. [Online]. Available: https://www.bawaslu.go.id/

[20] K. P. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[21] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall, 1988.

[22] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: Wiley, 2005.

[23] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 1997, pp. 21–34.

[24] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York: Springer, 2009.

[25] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.