

Clustering Analysis of Civil Servant Data in Binjai City Based on Regional Work Unit and Age Using the K-Means Clustering Method

Sapriko^{1*}, Novriyenni², I Gusti Prahmana³

^{1,2,3} College of Informatics and Computer Management (STMIK) Kaputama, Binjai
sapriko17@gmail.com^{1*}, Novriyenni.sikumbang@gmail.com², igustiprahmana4@gmail.com³

Abstract

Civil Servants (Aparatur Sipil Negara/ASN) play a strategic role in government administration and public services in Binjai City. However, the management of ASN data has thus far remained largely administrative and descriptive, failing to adequately reveal employee distribution patterns based on regional work unit, age, and rank. This study aims to cluster ASN data in Binjai City using the K-Means Clustering algorithm based on the variables of regional work unit, age, and rank. The dataset consists of 498 civil servants obtained from the Personnel and Human Resource Development Agency (BKPSDM) of Binjai City, after removing anomalous data belonging to Rank I/d. The clustering process was carried out using MATLAB through the stages of data selection, transformation into numerical form, normalization, Euclidean Distance calculation, and iteration until the centroids stabilized. Testing was conducted on three cluster scenarios, namely $k=3$, $k=4$, and $k=5$, evaluated using Sum of Squared Error (SSE), the Davies-Bouldin Index, the Silhouette Score, and Variance analysis. The results show that $k=4$ produced the best clustering outcome, with a Silhouette Score of 0.5705, a Davies-Bouldin Index of 0.9309, an SSE of 29.9369, and a more balanced Variance distribution among clusters. The clustering results yielded four distinct groups of civil servants, one of which was identified as consisting predominantly of senior employees from the Education Office, requiring particular attention in succession planning. The findings of this study are expected to serve as a basis for policy-making in the management of human resources within the Binjai City Government.

Keywords: Age, Civil Servants, Data Mining, K-Means Clustering, Rank, Regional Work Unit

1. Introduction

Civil Servants (ASN) are a core element in government administration and public services at the regional level, including in the Binjai City Government. ASN in Binjai City are distributed across various regional work units with diverse age characteristics. This diversity represents both an opportunity and a challenge for human resource management, particularly in workforce planning, competency development, and succession planning within each regional work unit.

To date, ASN data management in the Binjai City Government has remained largely administrative and descriptive, such as simple employee headcount records, so that in-depth analysis of ASN distribution patterns based on age and regional work unit has not been carried out optimally. This condition risks producing inaccurate decision-making, uneven ASN distribution across regional work units, and suboptimal succession planning.

One approach that can be leveraged to uncover patterns in large ASN datasets is data mining, particularly clustering techniques. The K-Means Clustering algorithm was chosen for its simplicity, computational speed, and effectiveness in grouping large datasets. Prior research by Bisri and Kurniawan showed that K-Means is effective for clustering civil servant data based on regional work unit and age, producing clusters of good quality [1]. Building on this, the present study applies K-Means Clustering to group ASN data in Binjai City based on the variables of regional work unit, age, and rank, with the aim of providing a clear picture of ASN distribution patterns and recommendations for human resource planning within the Binjai City Government.

2. Literature Review

2.1. Related Work

Several prior studies have applied K-Means Clustering across various domains. Putri, Safwandi, and Fitri applied K-Means to cluster population density data, evaluated using the Davies-Bouldin Index [2]. Kesuma and Hamidani applied K-Means to cluster scholarship recipient data, showing that the method could perform grouping quickly and accurately [3]. Fatkhudin et al. clustered student data based

on academic attributes to support academic mentoring strategies [4]. Handayanna and Sunarti applied K-Means to cluster ICT skill levels based on age [5], while Mubarak et al. applied the same algorithm to cluster job vacancy data [6]. These studies demonstrate the effectiveness of K-Means across various domains; however, its application specifically to civil servant (ASN) data based on regional work unit and age remains limited, motivating the present study as an extension with a different object and objective.

2.2. Data Mining and Clustering

Data mining is the process of extracting valuable knowledge from large and complex datasets, encompassing the stages of data selection, preprocessing, transformation, algorithm application, and interpretation/evaluation of results [7]. One function of data mining is clustering, the process of grouping data based on the degree of similarity of characteristics without a target variable, where data within a cluster exhibit high similarity while data across clusters exhibit low similarity [7].

2.3. K-Means Clustering Algorithm

K-Means is a non-hierarchical clustering algorithm that partitions data into a number of clusters based on the distance to a cluster centroid, performed iteratively until no significant change occurs in either the centroid positions or cluster membership [8]. The steps of the K-Means algorithm are as follows: (1) determine the number of clusters (k); (2) determine the initial centroids randomly; (3) calculate the distance of each data point to each centroid using Euclidean Distance; (4) assign each data point to the nearest cluster; (5) recompute the centroids as the mean of the members of each cluster; and (6) repeat the process until cluster membership stabilizes [8]. The Euclidean Distance equation used is as follows:

$$d(X, Y) = \sqrt{[(x1-y1)^2 + (x2-y2)^2 + (x3-y3)^2]} \quad (1)$$

where X denotes the data coordinates (Regional Work Unit, Age, Rank) and Y denotes the coordinates of the cluster centroid.

3. Research Method

3.1. Research Stages

This study uses a quantitative, data mining-based approach carried out through six stages: problem identification, literature review, data collection, data analysis, testing and implementation, and evaluation, as illustrated in Figure 1.

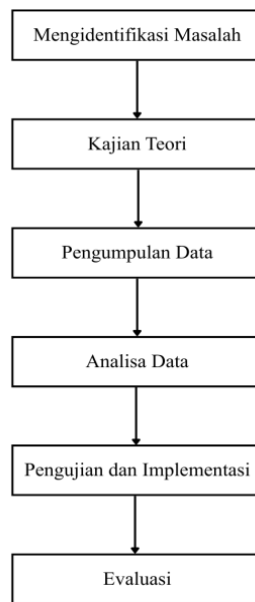


Fig. 1 : Research workflow

3.2. Research Data

The data used are ASN data from Binjai City for the year 2025, obtained from the Personnel and Human Resource Development Agency (BKPSDM) of Binjai City, covering the attributes of name, age, rank, and regional work unit. After removing anomalous data in Rank I/d whose count was very small and could act as noise during clustering — 498 ASN records were retained for use in this study.

3.3. Data Transformation and Normalization

Categorical attributes were transformed into numerical form so that they could be processed by the K-Means algorithm. Regional work unit was encoded into 42 unique codes (1–42) corresponding to each work unit; age was grouped into four coded ranges referring to the productive-age classification of Statistics Indonesia (BPS) [9]; and rank (II/a to IV/b) was encoded into numerical values ranging from 2.1 to 4.2, as shown in Tables 1 and 2. Rank I was excluded because its population in current ASN data is extremely limited.

Table 1 : Age range encoding

No	Age Range	Code
1	20 – 30 years	1
2	31 – 40 years	2
3	41 – 50 years	3

No	Age Range	Code
4	> 50 years	4

Table 2 : Rank encoding

No	Rank	Code
1	II/a – II/d	2.1 – 2.4
2	III/a – III/d	3.1 – 3.4
3	IV/a – IV/b	4.1 – 4.2

All variables were then normalized using min-max normalization to the range [0,1] so that each variable contributes proportionally to the distance calculation, given that the scales of regional work unit, age, and rank differ.

3.4. Implementation and Evaluation

The clustering process was implemented in MATLAB using the kmeans function, covering data reading, anomaly removal, variable encoding, normalization, determination of the number of clusters (k), Euclidean Distance calculation, and iteration until the centroids converged. Testing was carried out for three scenarios, namely k=3, k=4, and k=5, evaluated using Sum of Squared Error (SSE), the Davies-Bouldin Index (DBI), the Silhouette Score, and inter-cluster Variance analysis to determine the most optimal number of clusters.

4. Results and Discussion

4.1. Cluster Number Testing Results

Testing was carried out on 498 ASN records from Binjai City for three cluster scenarios. A summary of the evaluation results is presented in Table 3.

Table 3 : Comparison of cluster evaluation results

k	SSE	Davies-Bouldin Index	Silhouette Score
3	38.9543	1.0965	0.5297
4	29.9369	0.9309	0.5705
5	24.7317	0.8807	0.5560

The highest Silhouette Score was obtained at k=4 (0.5705), higher than both k=3 (0.5297) and k=5 (0.5560). The Davies-Bouldin Index at k=4 (0.9309) also indicates fairly good cluster separation, although the lowest SSE was obtained at k=5, which is expected since SSE generally decreases as the number of clusters increases.

4.2. Inter-Cluster Variance Evaluation

To further support the determination of the optimal number of clusters, a Variance analysis was also performed for each scenario. A comparison of the minimum, maximum, and average Variance across clusters is presented in Table 4.

Table 4 : Comparison of Variance test results

Number of Clusters	V minimum	V maximum	Average Cluster Variance
k = 3	0.053919	0.123692	0.085625
k = 4	0.042269	0.098475	0.069121
k = 5	0.022537	0.109370	0.055756

The average cluster Variance decreased consistently as the number of clusters increased. However, the maximum Variance (V maximum) at k=5 (0.109370) was actually higher than at k=4 (0.098475), indicating that one cluster in the k=5 scenario still exhibited a fairly large and less balanced spread of data. In contrast, the Variance distribution at k=4 was relatively more stable and balanced across clusters.

4.3. Determination of the Optimal Cluster

Considering the highest Silhouette Score, a favorable Davies-Bouldin Index, and a more balanced Variance distribution, k=4 was selected as the best number of clusters in this study. The three-dimensional visualization of the clustering results for the 498 ASN records under the k=4 scenario is shown in Figure 2, with cluster membership of 192 records in Cluster 1, 160 records in Cluster 2, 68 records in Cluster 3, and 78 records in Cluster 4.

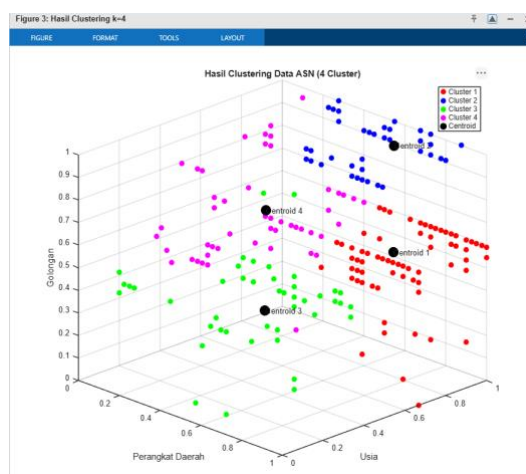


Fig. 2 : Clustering results of ASN data with k=4

4.4. Cluster Characteristics

The clustering results with $k=4$ reveal four groups of civil servants with distinct characteristics. Cluster 2 is the most senior group, with an average age of about 53.1 years and an average rank of 4.18, dominated by employees from the Education Office, and therefore requires priority attention in succession planning. Cluster 1 (192 employees, average age 48.9 years) is likewise dominated by Education Office employees with mid-to-upper ranks. Cluster 3 is the youngest group, with an average age of 33.3 years and the lowest average rank, dominated by employees from the Health Office, reflecting a group of employees in the early stages of their careers who represent potential succession candidates. Cluster 4 is a middle-aged group (average age 44.5 years) concentrated among health workers at RSUD Dr. R. M. Djoelham. These findings provide a concrete picture of ASN distribution based on regional work unit and age, which can serve as a basis for workforce planning, competency development, and succession planning within the Binjai City Government.

5. Conclusion

The K-Means Clustering algorithm proved effective for clustering 498 ASN records from Binjai City based on the variables of regional work unit, age, and rank, after removing anomalous data in Rank I/d. Based on testing across three cluster scenarios, $k=4$ produced the best results, with a Silhouette Score of 0.5705, a Davies-Bouldin Index of 0.9309, an SSE of 29.9369, and a more balanced inter-cluster Variance distribution compared to $k=3$ and $k=5$. The clustering results yielded four distinct groups of civil servants, with Cluster 2 identified as the most senior group, dominated by Education Office employees, requiring particular attention in succession planning. The findings of this study can serve as a basis for consideration by the Binjai City Government in workforce planning, competency development, and succession planning for civil servants. Future research is recommended to incorporate additional variables such as educational background, position, or tenure, and to compare K-Means results with other clustering methods such as Hierarchical Clustering or DBSCAN to further validate the resulting clustering patterns.

Acknowledgement

The author would like to thank the Personnel and Human Resource Development Agency (BKPSDM) of Binjai City for providing the data used in this study, and STMIK Kaputama Binjai for its guidance and facilities that supported the completion of this research.

References

- [1] M. Bisri and A. Kurniawan, "Analysis of civil servant data clustering based on regional work unit and age using the K-Means Clustering method," *Jurnal Teknologi Informasi dan Data Mining*, vol. 5, no. 1, pp. 45–52, 2024.
- [2] N. Putri, Safwandi, and Y. Fitri, "Application of the K-Means Clustering algorithm for clustering population density data using the Davies-Bouldin Index," *Jurnal Data Mining dan Analisis Sistem*, vol. 6, no. 1, pp. 30–38, 2024.
- [3] D. Kesuma and R. Hamidani, "Implementation of data mining using the K-Means Clustering algorithm for clustering scholarship recipient data," *Jurnal Ilmu Komputer dan Informatika*, vol. 10, no. 2, pp. 88–96, 2024.
- [4] A. Fatkhudin, T. Suryanto, and R. Pratama, "Application of the K-Means Clustering algorithm for clustering student data," *Jurnal Informatika dan Sistem Informasi*, vol. 9, no. 1, pp. 55–63, 2023.
- [5] F. Handayanna and S. Sunarti, "Application of K-Means Clustering for clustering information and communication technology skill levels," *Jurnal Teknologi dan Sistem Komputer*, vol. 12, no. 1, pp. 25–33, 2024.
- [6] A. Mubarak, N. Hidayat, and R. Firmansyah, "Implementation of the K-Means algorithm for clustering job vacancy data," *Jurnal Sistem Informasi dan Teknologi*, vol. 11, no. 1, pp. 40–48, 2025.
- [7] R. Delima, "Application of data mining in data analysis for decision-making," *Jurnal Teknologi Informasi*, vol. 7, no. 2, pp. 101–110, 2023.
- [8] M. Ma'ady, A. Rahman, and D. Putra, "Utilization of data mining in processing large-scale data to support decision-making," *Jurnal Sains Data dan Analitika*, vol. 5, no. 1, pp. 15–24, 2024.
- [9] Statistics Indonesia (Badan Pusat Statistik), *Indonesian Labor Force Indicators 2023*. Jakarta: Badan Pusat Statistik, 2023.
- [10] Republic of Indonesia, *Law Number 5 of 2014 concerning Civil Servants (Aparatur Sipil Negara)*, 2014.
- [11] Syahabuddin et al., "Analysis of regional work units in the management of civil servant human resources," *Jurnal Administrasi Publik*, 2021.
- [12] A. Fatwa et al., "Utilization of MATLAB in numerical computation analysis and engineering programming," *Jurnal Informatika*, 2022.