DOI: https://doi.org/10.70656/ijcse.v3i02.900

Optimization of Constraints in Valins New Installation Data Validation in Monitoring ODP Installation Progress Using the K-Means Algorithm (Case Study: PT Telkom Akses Bengkulu)

1st Maharanti Putri Cempaka
Information System Study Program
Department of Engineering
University of Bengkulu
Bengkulu, Indonesia
maharantip.c@gmail.com

2nd Andang Wijanarko, M.Kom
Information System Study Program
Department of Engineering
University of Bengkulu
Bengkulu, Indonesia
andang@unib.ac.id

3rd Dirgan Wisang
PT Telkom Akses Bengkulu
Bengkulu, Indonesia
dirganwisang24@gmail.com

Abstract*—The Valins PSB data validation process played a crucial role in monitoring the progress of ODP installation at PT Telkom Akses Bengkulu. Accurate and comprehensive validation was necessary for the monitoring system to provide reliable information. However, in practice, there were still challenges such as incomplete data, data inconsistencies, limitations in the use of QR barcodes, and disruptions to the validation bot, which resulted in some data remaining unvalidated and affected the effectiveness of monitoring. This study aimed to analyze validation challenges, identify data patterns, and provide optimization recommendations. The method used was K-Means Clustering with 1,647 Valins PSB data points that underwent selection, cleaning, and transformation before the clustering process. The results showed that the data could be grouped into several clusters representing different levels of validation success, thereby helping identify patterns of challenges based on location. This study was expected to support improvements in validation quality and monitoring effectiveness at PT Telkom Akses Bengkulu.

Keywords- Data Mining, K-Means Clustering, Valins PSB, Data Validation, Monitoring

I. INTRODUCTION

The Valins PSB system was used at PT Telkom Akses Bengkulu to support the installation data validation process in monitoring the progress of Optical Distribution Points (ODPs). Through this system, data validation was expected to ensure that installation data was accurate and could be used effectively. However, in practice, there were still various challenges, such as incomplete data, data inconsistencies, limitations in the use of QR codes, and disruptions to the validation bot. These issues resulted in some data remaining unvalidated, thereby affecting the effectiveness of monitoring.

To address these issues, an analytical approach was needed that could identify patterns of validation challenges and support optimization efforts. One approach that was used was data mining, specifically clustering techniques using the K-Means Clustering algorithm. This algorithm was used to group data based on similarity, thereby providing a clearer understanding of the underlying patterns.

K-Means clustering is a data mining algorithm that divides data into several clusters based on proximity to the centroid. This process is performed iteratively until stable clusters are formed [1]. This algorithm was chosen because it is simple, efficient, and capable of grouping data without requiring initial labels, making it suitable for analyzing validation data with various characteristics. Previous research has also shown that K-Means is effective in identifying patterns and supporting data analysis [1] [2]. Additionally, this algorithm is capable of processing large amounts of data with fast computation and is easy to implement [2].

In this study, the analysis process was conducted using Google Colaboratory (Google Colab), a cloud-based platform that allowed users to run Python code through a browser. Google Colab supported various libraries such as Pandas, Matplotlib, and NumPy, which were used for data processing and visualization. Using this platform allowed for efficient analysis without the need for complex local installations, and facilitated the presentation of analysis results in visual form [3] [4].

Conceptually, this study integrated data mining, clustering, and data visualization into a single analytical process. Data mining was used to process data, clustering was used to group data based on similarity, and visualization was used to present the analysis results in an informative manner. These three components complemented one another in aiding data interpretation and decision-making [1] [2] [3] [4].

The implementation of K-Means clustering in this study offers substantial operational benefits for PT Telkom Akses Bengkulu, particularly in addressing the limitations of manual validation monitoring. In conventional practice, identifying which data points remain unvalidated requires technicians and supervisors to manually inspect each entry across multiple operational areas (STAGRO), a process that is time-consuming, error-prone, and difficult to scale as the volume of Valins PSB data grows. By applying K-Means clustering, this study enables the automated grouping of data into distinct validation clusters, allowing monitoring personnel to rapidly identify areas with high concentrations of unvalidated data without the need for exhaustive manual review. This capability transforms the monitoring workflow from a reactive, record-by-record inspection into a pattern-driven, cluster-level analysis, thereby improving both the speed and accuracy of validation oversight. Ultimately, the clustering

approach supports more targeted decision-making, enabling PT Telkom Akses Bengkulu to prioritize intervention in areas where validation failures are most prevalent and to allocate technical resources more efficiently.

Therefore, this study aimed to identify the challenges in validating Valins PSB data, cluster the data based on validation patterns using K-Means clustering, and provide optimization recommendations to improve the validation process and the effectiveness of monitoring at PT Telkom Akses Bengkulu.

II. RESEARCH METHOD

This study used a data mining approach with the K-Means algorithm to analyze validation constraints in Valins PSB data for monitoring ODP installation progress at PT Telkom Akses Bengkulu.

A. Research Flow

This study employed a data mining workflow consisting of data collection, preprocessing, clustering, and data analysis. Data mining techniques are widely used to discover patterns and insights from large datasets [2] [5]. The implementation was carried out using Google Colab with the Python programming language and libraries such as Pandas, Scikit-learn, and Matplotlib.



Fig. 1. Research Research Flowchart

This figure illustrated the research process from data collection to clustering analysis.

B. Data Collection

The dataset was derived from Valins PSB operational data at PT Telkom Akses Bengkulu, consisting of 1,647 data points. The data included attributes such as STAGRO, SN, SA, VALINS STATUS, VALINS ID, and TECHNICIAN. The use of real-world data was crucial in data mining to produce accurate and meaningful analyses [2].

C. Data Preparation

The data preparation stage was an important step in data mining to improve data quality before further analysis [5] [6]. In this study, data preparation was carried out using Google Colab by importing datasets through the Pandas library. Cleaning procedures confirmed that no missing values were present in the primary clustering attributes, and duplicate records were removed alongside text standardization across categorical fields. All 1,647 records were retained for analysis, as rigorous data verification is critical to ensure unbiased centroid computation in K-Means clustering.

D. Data Transformation

After data preparation, transformation was carried out to convert categorical data into numerical form so that it could be processed by the K-Means algorithm [6].

In this study:

STATUS VALINS was converted into numerical values (1 = validated, 0 = not validated)

STAGRO was transformed using a category encoding technique

This transformation enabled the data to be used effectively in the clustering process.

E. Normalization

Normalization was performed to ensure that all variables were on the same scale before the clustering process. In this study, the StandardScaler method was used to standardize the data [5]. The purpose of this process was to prevent any single variable from dominating the clustering results. The normalized data was then used as input for clustering and visualization processes.

F. Clustering using K-Means

K-Means is a clustering algorithm that divides data into several groups based on similarity, where each group (cluster) is represented by a centroid [1]. The algorithm works by determining the distance between each data point and the nearest centroid, grouping the data accordingly, and updating the centroid positions iteratively until a stable condition (convergence) is reached [1] [6]. This method is widely used in data mining because it has a simple process and is efficient in handling large amounts of data [2].

The number of clusters in this study was determined as K = 2, which represented two main conditions, namely validated data and unvalidated data. This selection was based on analytical needs and data characteristics to simplify the interpretation of clustering results. The steps of the K-Means algorithm included determining the number of clusters, initializing centroids, calculating distances, grouping data, updating centroids, and iterating until stability is achieved. Distance calculation used the Euclidean distance formula:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

As a simple example, data (0,1) with centroid (0,2) produced a distance of 1. In actual implementation, all distance calculations were performed automatically using Python.

G. Data Analysis

After the clustering process was completed, the results were analyzed using cluster distributions, cross-tables between clusters and validation status, and area-based analysis (STAGRO) [2] [6]. Additionally, data visualization was performed using graphs generated in Google Colab to illustrate cluster distributions and validation patterns. This visualization helped in understanding the results more clearly and supports the decision-making process. This analysis was used to identify validation constraints and support process optimization.

III. RESULT AND DISCUSSIONS

This section presented the results of the clustering process using the K-Means algorithm on Valins PSB data along with the analysis of the obtained results.

A. Clustering Evaluation

This study used the Elbow method to determine the optimal number of clusters by analyzing the relationship between the number of clusters (K) and the inertia value. The inertia value indicated the level of closeness of data within a cluster, where the smaller the value, the better the clustering quality. Based on the results shown in Figure 2, there was a

significant decrease in inertia from K=1 to K=2, after which the decrease tended to flatten for subsequent K values.

This indicated the presence of an "elbow" point at K=2, which represented the optimal number of clusters in capturing the main pattern of the data. Although larger K values such as K=3 also show a decrease in inertia, the selection of K=2 in this study was based on analytical needs that focus on two main conditions, namely validated data and unvalidated data. Therefore, the selection of the number of clusters considers not only technical aspects but also relevance to the research context.

To substantiate the cluster selection quantitatively, the inertia values obtained across successive cluster configurations were analyzed. At K=1, the inertia was approximately 3300; at K=2, it decreased substantially to approximately 1650; and at K=3, it further declined marginally to approximately 1000. The proportional reduction in inertia from K=1 to K=2 was considerably larger than the reduction observed from K=2 to K=3, indicating a clear inflection point commonly referred to as the "elbow" at K=2. Beyond this point, increasing the number of clusters yields diminishing returns in terms of inertia reduction, while potentially reducing the interpretability of the two main operational conditions: validated and unvalidated data.

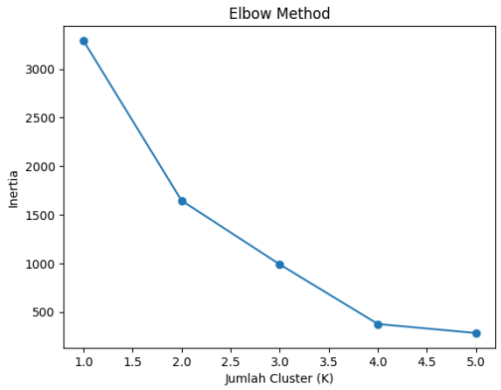


Fig 2. Elbow Method Graph for Determining the Number of Clusters

Based on these results, the number of clusters used in this study was K=2. This result also indicated that the selected number of clusters was appropriate with the data distribution pattern.

B. Clustering Results

The clustering process was carried out using the K-Means algorithm with the number of clusters set to K = 2. These clusters represented two main conditions, namely validated data and unvalidated data. The results of the clustering process were stored in a new column named "cluster", where each data point was grouped based on its proximity to the centroid. This grouping process allowed data with similar characteristics to be placed in the same group, making it easier to identify patterns in validation data [1] [2] [7].

C. Cluster Distribution

The results showed that cluster 0 consisted of 849 data points, while cluster 1 consisted of 798 data points. This distribution indicated that the data was relatively balanced across the two clusters, meaning that both validated and unvalidated conditions had a significant proportion in the dataset. The difference in the number of data points in each cluster reflected variations in validation conditions. This showed that the K-Means algorithm was capable of grouping data based on similar characteristics and revealing patterns within the data [8].

D. Area-Based Analysis (STAGRO)

Further analysis was conducted using STAGRO to identify areas with validation issues. The results showed that in cluster 0, the highest number of data points was found in CRP (246 data points), followed by MNA (118 data points) and BTH (103 data points). Meanwhile, in cluster 1, the highest counts were found in BNC (183 data points) and CRP (147 data points). This indicated that each area had different validation patterns across clusters. For example, CRP appeared in both clusters with relatively high values, indicating varied validation conditions. On the other hand, BNC was more dominant in cluster 1, which could indicate higher validation constraints in that area. These findings showed that clustering was able to reveal area-based patterns and helped identify areas that required further attention, thereby supporting the role of data mining in decision-making for optimization [9]. In addition, differences in data distribution across areas could be influenced by operational factors such as the number of technicians, geographical conditions, and the level of difficulty in network installation. This indicated that validation constraints were influenced not only by the system but also by field conditions that needed to be considered in evaluation and optimization processes.

E. Crosstab Analysis

Crosstab analysis was conducted to examine the relationship between clusters and validation status. The results showed that cluster 0 consisted of 849 validated data points with no unvalidated data, while cluster 1 consisted of 798 unvalidated data points with no validated data. These results indicated that the clustering process successfully separated the data into two clearly distinct groups, where cluster 0 represented validated data and cluster 1 represents unvalidated data. This showed that the K-Means algorithm was able to effectively group data according to validation conditions [9]. Furthermore, the clear separation indicated that the variables used in the clustering process had a strong relationship with validation status. In other words, the attributes used were able to significantly represent the condition of the data.

F. Data Visualization

Data visualization was used to present clustering results in graphical form to facilitate interpretation. Figure 3 showed the distribution of data based on area (STAGRO) in each cluster, where differences in the number of data points could be observed across operational areas. Areas such as BNC and MKO were more dominant in cluster 1 (unvalidated), while areas such as CRP and MNA were dominated by cluster 0 (validated). This indicated that validation constraints were not evenly distributed across all areas.

Figure 4 showed the distribution of clusters based on validation status. It could be seen that cluster 0 consisted entirely of validated data, while cluster 1 consisted entirely of unvalidated data. This indicates that the clustering process successfully separated the data into two clearly distinct groups without overlap.

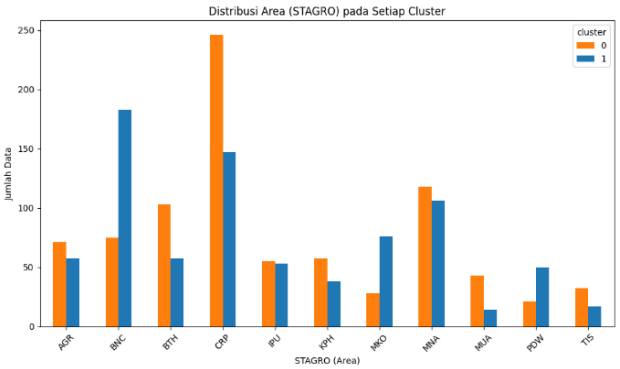


Fig 3. Distribution of Clusters by Area (STAGRO)

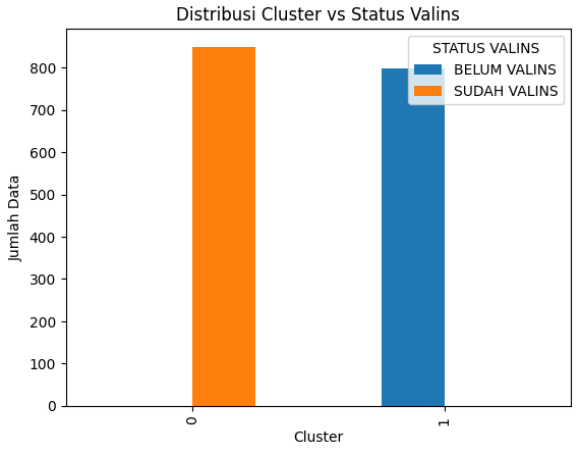


Fig 4. Distribution of Clusters by Validation Status

Visualization played an important role in simplifying complex data patterns and supporting decision-making. Based on these results, unvalidated data tended to be concentrated in certain areas, indicating that the validation process has not been evenly implemented across all operational regions. Therefore, optimization efforts were needed by focusing improvements on areas with a high number of unvalidated data points to enhance the effectiveness of the validation process and monitoring [9].

IV. CONCLUSION

This study was conducted to address validation constraints in Valins PSB data in monitoring the progress of ODP installation at PT Telkom Akses Bengkulu. The contribution of this research was to provide a data-driven approach using the K-Means Clustering algorithm to identify validation patterns and support decision-making in monitoring processes. The results showed that the data could be clearly grouped into two clusters, namely validated data and unvalidated data. STAGRO-based analysis indicated that validation constraints were not evenly distributed across operational areas, where some areas had a higher number of unvalidated data points. In addition, data visualization helped clarify these patterns. These findings indicated that clustering not only functioned as a data grouping method but also served as an analytical approach to understanding operational problems and as a basis for systematically evaluating the validation process [1] [2].

Optimization in this study was carried out in a targeted manner based on clustering results by focusing improvements on areas with a high number of unvalidated data points. Improvements were made through enhancing validation quality and more focused monitoring, thereby increasing the effectiveness of ODP installation progress monitoring. This was in line with previous research stating that K-Means could support data grouping and analysis in data-driven systems [10]. However, this study was limited to the use of specific data and the K-Means method, so future research could develop other methods or include more complex variables to obtain more optimal results.

V. ACKNOWLEDGMENTS

The author would like to express sincere gratitude to PT Telkom Akses Bengkulu for providing the opportunity to carry out the internship and for supporting the data used in the preparation of this article. The author also extends appreciation to the academic supervisor for guidance and support throughout the internship and the writing of this article. In addition, the author would like to thank all parties who have contributed to the completion of the internship and this article.

REFERENCES

[1] K. Mustaqim, G. Al Raffi, N. S. Fitri, and N. Riza, “Penerapan Metode K-Means Clustering Untuk Klasifikasi Tingkat Literasi Terhadap Penyebaran Hoaks Di Provinsi Jawa Barat,” vol. 5, no. 2, pp. 1–9, 2025.

[2] S. I. Safitri, C. Suhery, S. Bahri, J. Rekayasa, and S. Komputer, “IMPLEMENTASI ALGORITMA K – MEANS UNTUK CLUSTERING SENTIMEN PADA OPINI KUALITAS PELAYANAN,” vol. 09, no. 02, pp. 186–197, 2021.

[3] R. G. Guntara, “Visualisasi Data Laporan Penjualan Toko Online Melalui Pendekatan Data Science Menggunakan Google Colab,” vol. 2, no. 6, pp. 2091–2100, 2023.

[4] B. M. Suparwanto, “Visualisasi Data Dari Data Ketidakhadiran Mahasiswa Menggunakan Pemrograman Python,” J. Teknol. Inf., vol. 02, no. 02, pp. 103–113, 2023.

[5] M. B. Uddin and Z. Fatah, “Penerapan Data Mining Clustering K-Means Dalam Mengelompokkan Data Penduduk Penyandang Disabilitas,” Gudang J. Multidisiplin Ilmu, vol. 2, no. November, pp. 86–94, 2024.

[6] A. Fatkhudin et al., “Implementasi Algoritma Clustering K-Means Dalam Pengelompokan Mahasiswa Studi Kasus (Prodi Manajemen Informatika),” vol. 12, pp. 777–783, 2023.

[7] N. R. Setri Dwi Prasetiani, “Penerapan Data Mining Untuk Clustering Menu Favorit Menggunakan Algoritma K-Means (Studi Kasus Kedai Expo),” J. Inform. dan ilmu Komput., vol. 03, pp. 278–286, 2022.

[8] A. Alif, R. Mulyana, A. R. Juwita, A. M. Siregar, and A. Fauzi, “Penerapan Algoritma K-means Clustering dan Hierarchical Clustering dalam Mengelompokkan Data Pengangguran di Karawang,” no. 2021, 2024, doi: 10.33364/algoritma/v.21-2.2155.

[9] N. Hendrastuty, “Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa,” vol. 3, pp. 46–56, 2024.

[10] B. Ariansah et al., “Penerapan K-Means Clustering untuk Pengelompokan Data Industri Kecil Menengah di Provinsi Jambi,” vol. 6, no. September, pp. 359–371, 2025, doi: 10.35957/jtsi.v6i2.13553.