Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Clustering IT Incidents Using K-Means: Improving Incident Response Time in Service Management

Rini Anggraeni[1)], Farrikh Alzami[2)*], Aris Nurhindarto[3)], Setyo Budi[4)], Rama Aria Megantara[5)] , Ifan Rizqa[6)] , Muslih[7)]
[1,2,3,4,5,6,7)]Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[1)] rinianggraeni62804@gmail.com, [2)] alzami@dsn.dinus.ac.id, [3)] arisnurhindarto@dsn.dinus.ac.id, [4)]setyobudi@dsn.dinus.ac.id , [5)] aria@dsn.dinus.ac.id, [6)] risqa.ifan@dsn.dinus.ac.id, [7)] muslih@dsn.dinus.ac.id

Submitted :May 26, 2025 | **Accepted** : Jun 5, 2025 | **Published** : Jun 13, 2025

Abstract: Incident management is one of the critical processes in Information Technology service management that aims to manage disruptions and minimize the impact of unexpected incidents on business services. This study applies the K-Means algorithm to cluster IT service incidents, aiming to enhance company operational efficiency. Utilizing a dataset from the UCI Machine Learning Repository comprising 141,712 events related to 24,918 incidents, this research analyzes incident patterns and characteristics for optimized handling. The data was analyzed through a series of preprocessing stages, and the elbow and silhouette methods were used to determine the optimal number of clusters. From the results, it was successfully grouped into 4 (four) clusters with a distortion score value of 964264294.569 and 0.52 silhouette score based on incident characteristics, such as urgency, priority, and number of reassignments. From this, the clustering results show that the K-Means algorithm effectively identifies incidents that require further handling, such as those with high urgency and priority, as well as helping the company focus resources to resolve incidents that have the most impact on the business sector. This research provides a data-driven solution to improve incident management and Service Level Agreement (SLA) fulfillment, while offering a framework for more effective and efficient IT incident analysis and resource allocation.

Keywords: incident management, k-means, clustering, service level agreement (SLA), IT Company.

INTRODUCTION

Incident management is the primary process for managing information technology services. It aims to manage disruptions and minimize the impact of unexpected incidents on business services (Baptista & Barata, 2024). Incident management in responding to a service system can go through the stages of detection, analysis, logs, resolution, and closure of incidents if they have been fully resolved (Goel et al., 2024). ServiceNow is one of the cloud platforms that companies widely use to manage IT enterprise service management (Santos & Rodrigues, 2024). The ServiceNow platform can experience incidents that refer to the disruption of the regular operation of a service, which requires immediate action to repair the service (Baptista & Barata, 2024). Incident management is not just about responding quickly to disruptions but also investigating the root cause and preventing recurrence in the future (Gokasar et al., 2023). The importance of incident management lies in its ability to ensure continuity of operations, maintain user satisfaction, and minimize potential losses due to incidents (Palma et al., 2024).

The management of IT incidents, particularly at scale, presents significant challenges for incident management teams. Our study utilizes a dataset from UCI Machine Learning, comprising 141,712 events related to 24,918 incidents (Claudio Amaral, 2018). This substantial volume of data underscores the complexity of incident management in modern IT environments, where each incident may have different characteristics and severity levels. Each of them has different characteristics and severity. Large and complex data availability requires a more effective way to categorize these incidents. Thus, a manager at an IT Company can set the right response priorities and make data-driven decisions (Palma et al., 2024),(Claudio Amaral, 2018).

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

As information technology advances, organizations face an increase in the volume and complexity of such incidents. This problem can be solved by data mining, which performs an Unsupervised Learning modeling process. Unsupervised learning is data that does not have a target or data label. In contrast, Unsupervised Learning can independently determine the structure and patterns of data groups (Fotopoulou, 2024).

Clustering is a grouping of data based on similarities, which can be divided into clusters with similar characteristics (Farou et al., 2024). Here, clustering adopts the K-Means Algorithm, a data clustering algorithm used in machine learning to divide different clusters (Atsa'am et al., 2023). K-means clustering is chosen for its ability to group similar incidents together, allowing for more efficient analysis and management of incidents with similar characteristics.

To overcome this challenge, analysts often use K-means clustering analysis as an effective solution to group incidents based on similar characteristics (Atsa'am et al., 2023). K-means is a widely used clustering algorithm in data management due to its simple and efficient ability to divide data into groups (clusters) based on distance or similarity between data points (Patel & Kushwaha, 2020). This algorithm works by dividing data into K groups or clusters and finding the location of each data point based on its closest distance to the cluster centroid (Nurlaila et al., 2021). In incident management research, this approach allows incidents with similar characteristics, such as similar resolution times or similar causes of problems, to be clustered (Patel & Kushwaha, 2020),(Nurlaila et al., 2021). Finally, it allows management services to develop more specific and efficient resolution strategies.

The advantages of the K-means algorithm are that it can identify hidden patterns in a vast amount of data and divide similar data into the same cluster (Zhang, 2023). Moreover, based on previous research, Dener, the K-Means algorithm is used for clustering in an improved classification prediction performance to help optimize the number of clusters detecting malware on the Android operating system (Dener & Gulburun, 2023). Based on other research by Hamamoto, the K-Means method is used to classify the behavior of Gunma University participants in the information security course on the LMS, which aims to see the final results in a shorter time (Hamamoto et al., 2021).

In addition, the K-Means algorithm is relatively easy to implement. It can be applied to many numerical data, making it suitable for managing descriptive attributes of incidents in the dataset, such as resolution time, duration of repair, or type of problem involved (Yang et al., 2020). However, one of the drawbacks of K-Means is the dependency on determining the optimal number of clusters (K), which requires extensive experimentation and analysis to ensure maximum results (Chen et al., 2024). Therefore, cluster evaluation techniques such as the elbow method or silhouette score are usually used to determine the appropriate K.

This study addresses this gap by applying K-Means clustering to group incidents based on urgency, priority, and reassignment count attributes. Unlike traditional approaches that often focus on reactive analysis, our method provides a proactive framework for understanding incident patterns. Furthermore, we extend the methodological approach by combining K-Means with elbow and silhouette methods for optimal cluster determination, offering a more comprehensive evaluation technique. By linking clustering results to Service Level Agreement (SLA) fulfillment, this study bridges the divide between data analysis and practical implications for IT operational management. Through this approach, we aim to uncover insights that can lead to more efficient incident-handling strategies and improved resource allocation in IT service management.

This study contributes to the field of IT service management by providing a data-driven approach to incident clustering, which can lead to more efficient resource allocation and improved SLA fulfillment.

LITERATURE REVIEW

Previous research F A Syaani et al. (Syaani et al., 2020) Incident Clustering in the Warehouse Workspaces by Using Text Mining. This research uses Incidents in warehouses need to be categorized to identify patterns of occurrence and support preventive action. The data used comes from incident reports in 39 warehouses of PT X during the period January 1 - September 29, 2018, with a total of 6,598 reports. Using text mining techniques, K-Means Clustering and Hierarchical Clustering. Then selected features with Genetic Algorithm to improve accuracy. The results of this study show that, Hierarchical Clustering provides the best results for the Consumer (icd-rate = 0.0005), Retail (icd-rate = 0.790), and SPL (icd-rate = 0.600) sectors. Then K-Means Clustering is more optimal for Life Style (icd-rate = 0.739) and Technology (icd-rate = 0.756) sectors. Feature selection with Genetic Algorithm increases the Silhouette Coefficient value, indicating an increase in cluster quality. The results of the study can assist companies in developing data-driven incident prevention strategies.

Previous research B Poerwanto et al. (Poerwanto, 2021) Evaluating the K-Means Analysis in Clustering Area Based on Estates Productivity in Tana Luwu Using Silhouette Index. This research, has a problem in identifying the productivity of sub-districts in Tana Luwu related to plantation crops. The method used is the K-Means Algorithm with evaluation using the Silhouette Index (SI). The results show that the optimal number of clusters is 2, with an SI value of 0.8068, indicating good clustering performance. The first cluster consists of the 5 most productive sub-districts, located in North Luwu District, while the second cluster includes 40 other less productive sub-districts.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Previous research Fengsui Wang et al. (Wang et al., 2021) Keyframe Generation Method Via Improved Clustering and Silhouette Coefficient for Video Summarization. This paper discusses the problem that the traditional K-Means algorithm is prone to local optimal solutions in video summarization. The method used is a combination of hierarchical clustering algorithm and K-Means. And optimization using silhouette coefficient. The result of the silhouette coefficient is 0.6618 with the value of (K) resulting in optimal clustering at 4. This means that the number of clusters and the number of keyframes that are finally extracted is 4 for this video.

METHOD

This research requires a deep understanding of the problem to solve it. The goal is to identify specific problems, determine impacts, and find effective and efficient solutions. Therefore, a framework is needed to describe the stages of the research. The following is a description of the research stages, which can be seen through the flow diagram below:

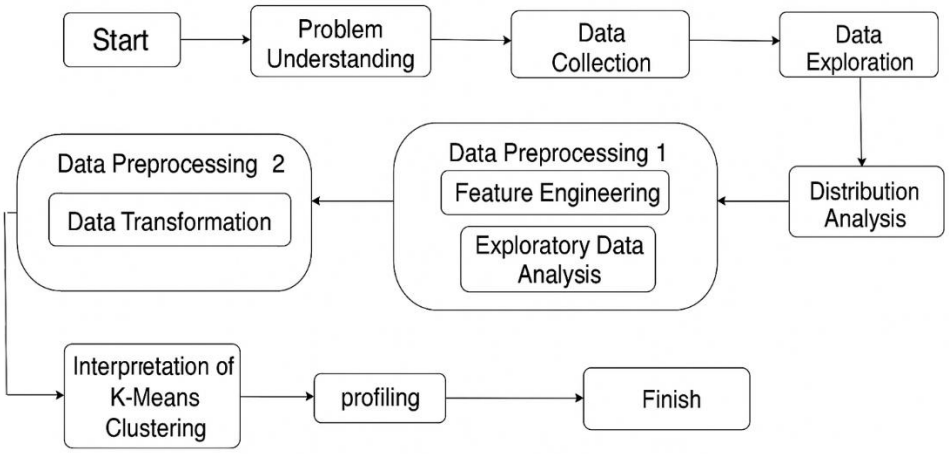

Figure 1 Research Flow Diagram

Data Collection - This research uses a dataset obtained from the UCI Machine Learning Repository. This dataset results from an audit of a ServiceNow platform in an IT Company by adopting an event log. The relational database has a dataset of 141,712 events with 24,918 incidents and 36 different attributes [7]. This dataset was chosen for its comprehensive representation of IT incidents in a real-world setting, providing a robust foundation for our analysis. To find a pattern in this complex data, we used the K-Means algorithm to cluster the data and identify the incidents that most impact the IT company.

Data Exploration - The data exploration stage involves a comprehensive examination of the dataset's structure and content. This includes analyzing the number of rows and columns, identifying data types for each column, detecting missing values, and assessing memory usage. We employed python data science tools (such as pandas, seaborn, scikit-learn) to facilitate this process, enabling a thorough understanding of the dataset's characteristics (Sadeghi et al., 2024). In addition, data exploration is also used to see all descriptive information, including numeric and categorical columns. Data exploration here can also represent missing data or unknown data, so in further understanding, it will pass preprocessing to know whether the data will be deleted or changed to its data type.

Distribution Analysis - After exploring the data, the next step is to view it in the form of column distributions to understand the data patterns and make better decisions in sustainability analysis.

Data Preprocessing - Furthermore, the data preprocessing stage is a crucial step in data analysis, which aims to ensure accurate analysis results and prepare raw data for further processing (Luftensteiner et al., 2024). This research has two data processing methods:

1) **First Data Preprocessing** - In the first processing stage, feature engineering improves model performance and is a creative process for turning raw data into more informative features (Verdonck et al., 2024). It can also help facilitate model interpretation and solve problems with missing values. In feature engineering, the data cleaning process is also carried out to improve the data quality so that it is ready to be analyzed; the aim is to prevent wrong conclusions, facilitate interpretation, and speed up the computational process. Thus, we addressed missing values, removed duplicates, and normalized numerical features to ensure data quality and consistency. The Exploratory Data Analysis stage follows the processing stage, where the results display the data distribution with effective visualization, connecting relationships between variables (Sadeghi et al., 2024).

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2) **Second Preprocessing** - The second processing adopts the data transformation process to transform or manipulate raw data into more relevant data. The data transformation here is to convert categorical to numerical using label-encoding so that it can be used for modeling with its simplicity and ability to retain information in the data well, so that the model can be more effective in recognizing patterns.

K-Means Clustering - The next stage is clustering, which is an analysis method to group data into groups (clusters) that are similar to each other to find data patterns without using already known labels. In the clustering stage, it must be accompanied by an algorithm. This research adopts the K-means algorithm because of its most straightforward and practical interpretation of clustering. In addition, this algorithm can work by dividing data into several clusters based on the similarity of features (Clayman et al., 2020). After that, the cluster first goes through the Hopkins stage to see whether the data will form a good cluster (Crase et al., 2021). Then, before the clustering interpretation stage, PCA (Principal Component Analysis) is appplied. Here, PCA is a multivariate analysis technique that reduces the dimensionality of data in situations where these datasets have many variables, helps reduce complexity, and visualizes them more effectively. K-means provides visualization as an elbow method to determine the optimal K and evaluate by silhouette. The cluster has a calculation in grouping the data; here is the related equation Hopkins statistical (1):

$$H = \frac{\sum_{i=1}^{m} w_i}{\sum_{i=1}^{m} (w_i + u_i)} \quad (1)$$

Equation 1 can be described as 1) Taking random samples from a dataset of m = random samples X and m = random samples Y that are imitations of the same distribution in the data space; 2) Calculating the closest distance for each point to u_i and random distance w_i; 3) H = 1 data tends not to cluster and H = 0 data has a tendency of clustering.

Then, in cluster validation, we use several approaches, such as distortion score in the elbow method and silhouette score. The explanation is described in the latter. The distortion score in the elbow method is used to determine the optimal number of clusters (k) in the K-Means Clustering algorithm. The distortion score measures how far the data in the cluster is from the cluster center point (centroid) (Shi et al., 2021). Moreover, the smaller the distortion, the better the data collected in the cluster. Distortion score equation can be seen as (2):

$$Distortion = \frac{1}{n} \sum_{i=1}^{n} \left(distance(x_i, c_i) \right)^2 \quad (2)$$

Where: n is the number of data; xi is the i data point; ci is the centroid of the cluster where xi is located; Distance (xi, ci) is the Euclidean distance between xi and ci.

Here, we used elbow methods which finds the ideal point, where adding clusters no longer provides a significant reduction in distortion (Shi et al., 2021). The graph uses the elbow method equation to determine the optimal number of clusters. The elbow method has the following steps:

1. Setting up the dataset on the cluster
2. Running K-Means for various values (k) in this study uses the value of k = 1 to k = 10.
3. Calculate the distortion score by calculating the average square of the distance to the centroid.
4. Make a graph of the number of clusters (k) on the x-axis and the distortion score on the y-axis
5. Finally, find the point where the distortion score decreases, which is similar to the shape of an elbow. The number of cluster points is considered optimal.

After finding the optimal k with the elbow method, the author runs clustering on the dataset and evaluates with silhouette score. Silhouette score is an evaluation metric of how well the data in a cluster is grouped. Silhouette has values ranging from -1 to 1. Where a value of 1 has an excellent group, 0 has a border between two clusters, and -1 data that should be in another cluster (Uddin et al., 2024). To calculate the silhouette score evaluation with the following equation (3):

$$silhouette(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad (2)$$

Here: i is each point in the cluster; a(i) is the average distance between point i and points in the same cluster (within-cluster distance); b(i) is the average distance between point i and points in other nearest clusters (nearest-cluster distance)

Where:

1. if a(i) is smaller than b(i), the silhouette value is close to 1.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

2. If a(i) and b(i) are almost equal, then the value is close to 0
3. And if a(i) is more significant than b(i), then it has a negative value

Profiling - The last stage is profiling, which is the process of analyzing the characteristics of a group or individual. Profiling here uses visualization in the form of bar plots, joint plots, and spider plots, aiming to increase efficiency in making better decisions and achieving the main objectives.

RESULTS

The clustering process begins with the data exploration stage of the UCI Machine Learning repository regarding the Incident management process enriched event log. This dataset has 141,712 event rows consisting of 1 incident status attribute, 32 incident description attributes, and two dependent attributes, which are the closing time of the incident. The incidents include 24,918 incidents identified by the incident number in the 'number' column. Each incident consists of several event rows with different statuses. In addition, the author also found '?' values in several rows that indicate the presence of missing values in this dataset. The following is a description of the attributes in Table 1 contained in the Incident management dataset:

Table 1 Attributes Incident Management

No	Attributes	Data Type	No	Attributes	Data Type
1.	number	object	19.	*u_symptom*	object
2.	incident_state	object	20.	*cmdb_ci*	object
3.	active	bool	21.	impact	object
4.	reassignment_count	int64	22.	urgency	object
5.	reopen_count	int64	23.	priority	object
6.	sys_mod_count	int64	24.	*assignment_group*	object
7.	made_sla	bool	25.	*assigned_to*	object
8.	*caller_id*	object	26.	knowledge	bool
9.	*opened_by*	object	27.	u_priority_confirmation	bool
10.	*opened_at*	object	28.	notify	object
11.	*sys_created_by*	object	39.	*problem_id*	object
12.	*sys_created_at*	object	30.	*rfc*	object
13.	sys_updated_by	object	31.	*vendor*	object
14	sys_updated_at	object	32.	*caused_by*	object
15.	contact_type	object	33.	*closed_code*	object
16.	*location*	object	34.	*resolved_by*	object
17.	*category*	object	35.	*resolved_at*	object
18.	*subcategory*	object	36.	closed_at	object

*italics indicate missing values

Of the 36 attributes, 29 features have categorical characteristics, three numeric features, and 4 boolean features, which means that these attributes have missing values in some features. Before starting the data preprocessing stage to analyze clustering, the author performs a distribution analysis step on several attributes that must be done to detect initial patterns and find out the data information to be processed. The following is a figure of the distribution analysis:

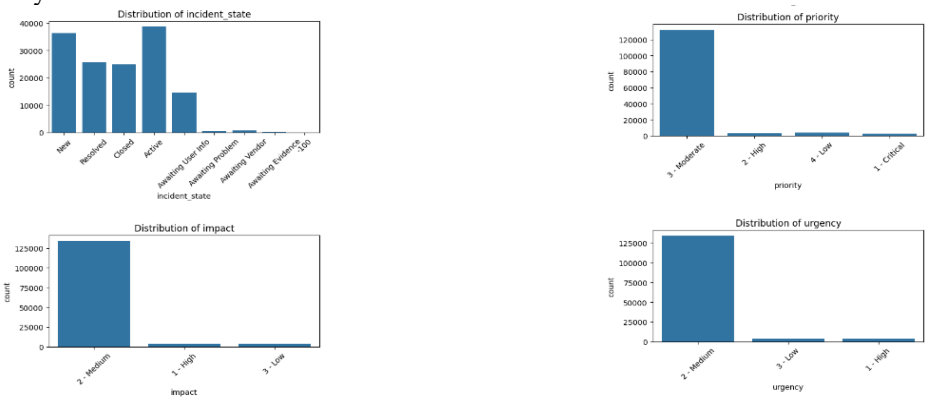

Figure 2. Distribution Analysis

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Based on the diagram in Figure 2, the distribution of incident attributes is uneven in some categorical variables (impact, priority, urgency). Although the imbalance can affect the clustering process by making the model more inclined to the dominant category, we maintained the original distribution to ensure the results were more representative of natural conditions.

To perform incident clustering, we select one event row with a specific status from each incident, ensuring that each incident is represented by only one row of data. This event row selection ensures that each incident is represented consistently in the clustering process. The selected rows are rows with closed incident status because every incident has a closed status, so all incidents can be used for clustering. A new duration feature is added by calculating the difference between the closed_at and opened_at columns, which can increase the variation of the data. Attributes with missing values are handled by deleting attributes with many missing values, filling in the mode for categorical data, and filling in values in datetime-type columns using similar columns. After that, the author represents it in histogram visualization as a form of Exploratory Data Analysis.

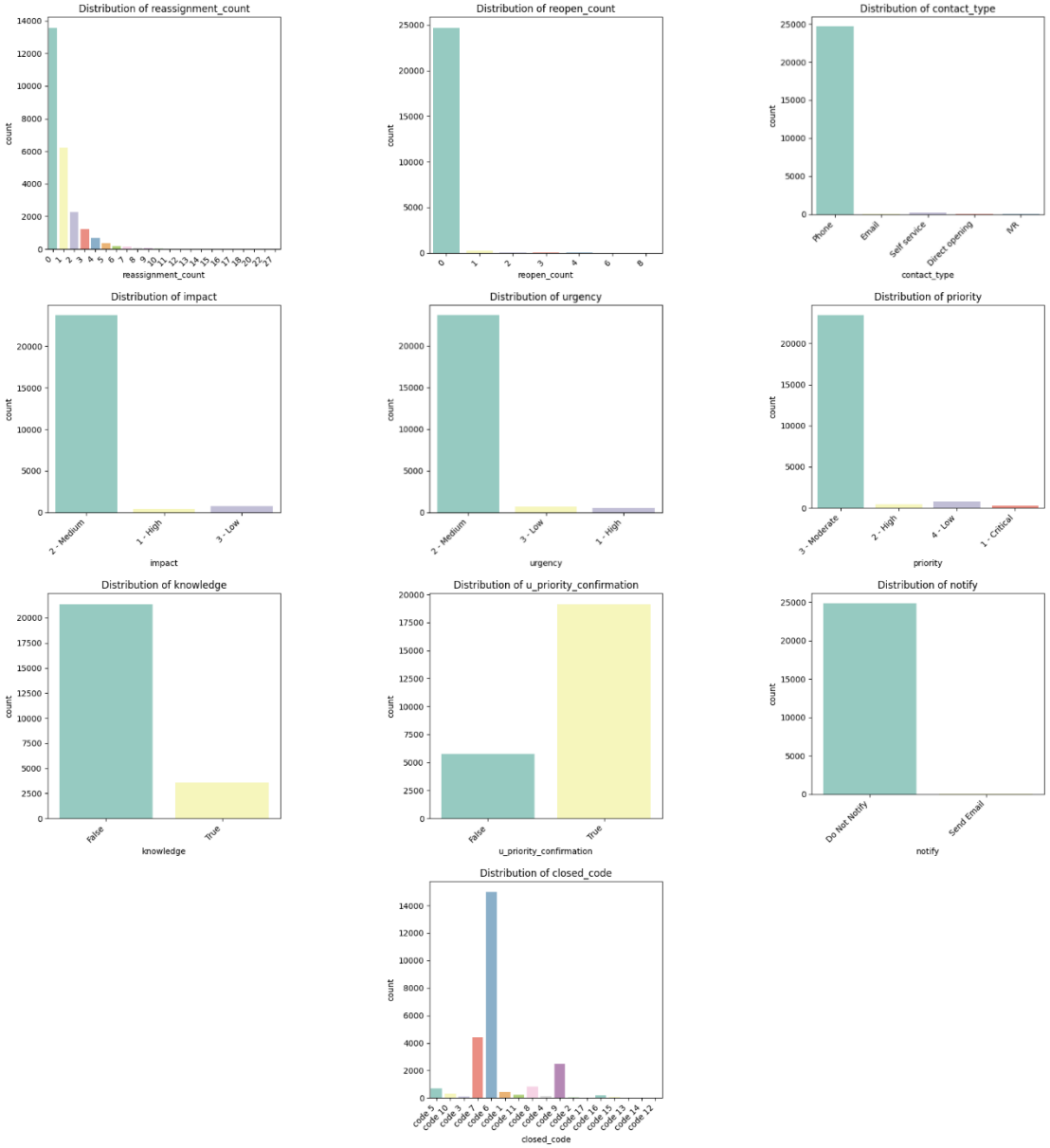

Figure 3. Exploratory Data Analysis using Barplot

Overall, the histogram, as shown in Figure 3, shows that most of the events have relatively standard characteristics, such as not having significant changes in the incident handling process, having a medium level of urgency and priority, and not having a solution-based on existing knowledge. However, a few of these events also have different characteristics, such as experiencing repeated changes in assignments and high urgency.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Because the histogram above has revealed the analysis pattern, the author will use the K-means algorithm to do clustering next to see the frequency distribution of the data. However, before that, we look at the average Hopkins statistics using 100x iterations for closed_data. Its function is to determine whether the data forms a group or not. Hopkins has a value range of 0-1. If it is close to 0, the data will be randomly scattered and not form a cluster. It turns out that the resulting Hopkins value is 0.914181188927481, and it shows that this data can be clustered.

Therefore, the author continued the interpretation with the elbow method to see the optimal K and the distortion score.

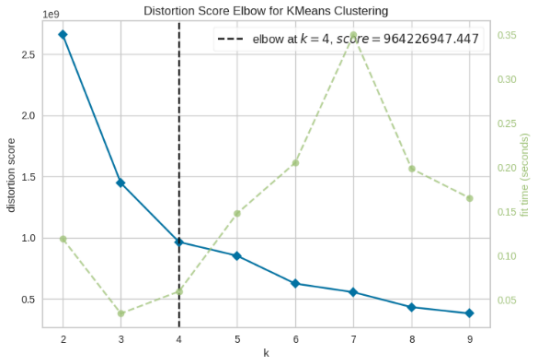

Figure 4. Elbow score diagram

The elbow results shown in Figure 4 show that the optimal number of clusters in K-means is 4. With a distortion score of 964264294.569 from the sum of the squares of the distance between the data and the nearest centroid to the number of clusters. Here X = K and Y = distortion score. In addition, the silhouette score results are also needed to evaluate the cluster quality. The author breaks it into 4 clusters with a silhouette result of 0.5210168104739562, which is quite good.

To visualize the data groups formed from the K-means clustering process with 4 clusters, we perform dimensionality reduction using Principal Component Analysis (PCA) and then present the results in a scatter plot shown in Figure 5. Each color represents a distinct cluster, illustrating the separation and overlap between different incident groups in the feature space.

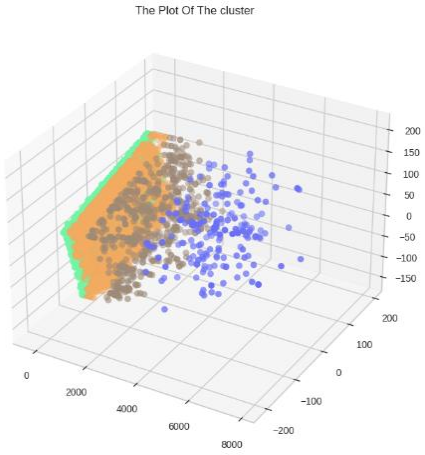

Figure 5. Scatter plot of the clusters

In addition, Figure 6 shows a display using bar plot visualization to compare the distribution of essential variables in each cluster.

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

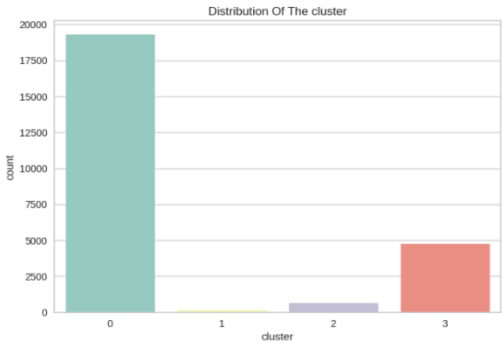

Figure 6. cluster distribution

Figure 6 illustrates the distribution of incidents across the four clusters identified by the K-Means algorithm. Cluster 0, being the largest, represents the most common type of incidents, potentially indicating 'business as usual' scenarios that IT support teams frequently encounter. Clusters 1, 2, and 3, while smaller, likely represent more specialized or complex incident types. This distribution suggests that while most incidents fall into a common category, there's a significant number of incidents that may require specialized handling or resources. This insight could be crucial for IT managers in resource allocation and training decisions.

In order to interpret the clustering results in more depth, the next step is to perform the profiling stage of the members of each cluster. The following is a visualization display with a bar plot:

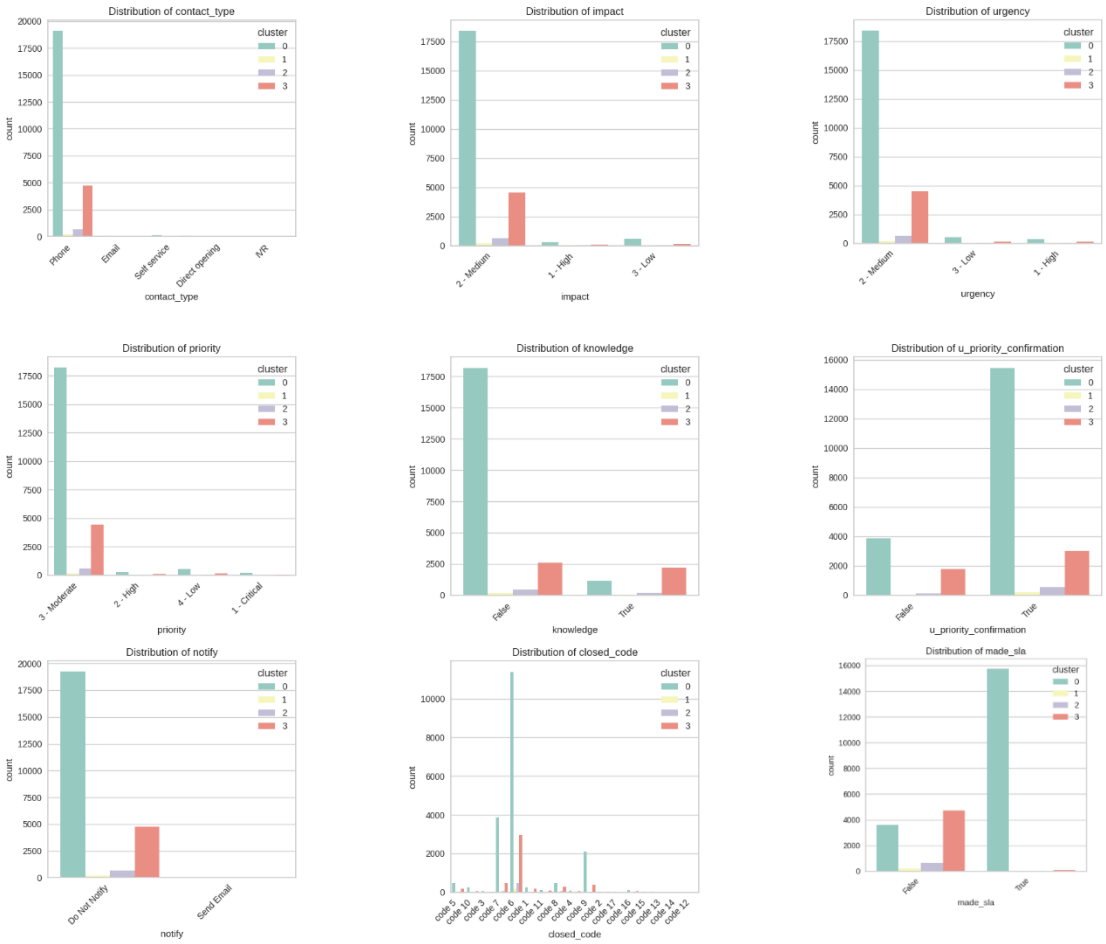

Figure 7. Cluster distribution analysis in some attributes

For more clarity in reading the members in each cluster with the following information:

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

1. Reassignment_count: In cluster 0, most of the data has a very low reassignment_count value of (0), but in cluster 1, there is some data with a higher reassignment_count, possibly problems or difficulties in case handling. Clusters 2 and 3 have a small amount and also have a low reassignment_count value.
2. Reopen_count in cluster 0: Almost all the data in this cluster has a reopen_count of 0, which indicates that most of the terminated cases do not need to be reopened. Cluster 1 has more cases that have a reopen_count value of 1, so some issues have to be handled again. Clusters 2 and 3 have very little data with reopen_count, the same as cluster 0.
3. Contact_type in cluster 0 has the majority of contacts by phone, while cluster 1 has a significant number switching to the email category. Meanwhile, clusters 2 and 3 have smaller numbers across all contact types, with little variation.
4. Impact in cluster 0 is mostly medium, but in cluster 1, there is a combination of medium and high impact, although the number is small. In clusters 2 and 3, it is dominated by medium impact, with very little data in other categories.
5. In cluster 0, there is more data with medium urgency. In contrast, cluster 1 appears well in the medium category, but there is also data with lower urgency. Clusters 2 and 3 have a similar pattern but fewer total incidents.
6. Priority cluster 0 is filled with medium priority. However, cluster 1 is a small part with high priority, and clusters 2 and 3 have very little data, with the dominance of medium priority.
7. Knowledge in cluster 0 most data is false, indicating that the majority do not have clear knowledge, then in cluster 1 is dominated by false, but less, and in clusters 2 and 3 show a small proportion of truth.
8. U_priority_confirmation in cluster 0 shows more incidents with priority confirmation true. However, cluster 1 has a higher number of falls, but not by much. The data from clusters 2 and 3 is similar to the other clusters, but the total number is small.
9. Notify cluster 0 mainly does not perform notifications. Cluster 1 shows a higher proportion in the send email category, and clusters 2 and 3 have a smaller total number, with minor differences.
10. Closed_code cluster 0 dominates with a closed_code of 10, indicating that many cases were handled well. Cluster 1 has more variation in closed codes, with closed_code 7 also having a significant number. Clusters 2 and 3 show some cases with specific closed codes but far fewer than the other clusters.
11. Made_sla cluster 0 mostly shows true, indicating that the SLA (Service Level Agreement) is met. However, cluster 1 has more data showing false, indicating some problems in meeting the SLA. In contrast, clusters 2 and 3 have small data and a pattern closer to cluster 1.

In addition, visualization can also be seen in the form of a spider plot for the final analysis results; here is the visualization display:

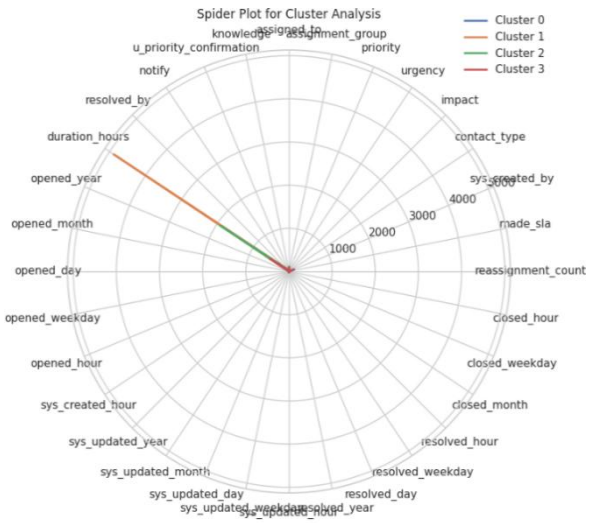

Figure 8. spider plot of the cluster

The display in Figure 8 shows that incident closure refers to the time duration pattern.

DISCUSSIONS

From the results and analysis section, we can see that K-means algorithm can be used in clustering IT service incidents, revealing four distinct incident categories. The resulting clusters, differentiated by urgency, priority, and

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

reassignment frequency, offer crucial insights into incident patterns, potentially revolutionizing IT incident management strategies.

While this research provides valuable insights into IT incident clustering, it is important to acknowledge several limitations:

1. Dataset Specificity: The study utilized data from a single IT company, which may limit the generalizability of results to other organizations or industries with different incident management processes.
2. Temporal Constraints: Our analysis was based on a snapshot of incident data. A longitudinal study might reveal evolving patterns in incident characteristics over time.
3. Clustering Algorithm: While K-Means is widely used, other clustering algorithms (e.g., DBSCAN, Hierarchical Clustering) might produce different results and could be explored in future research.
4. Causality: Our clustering approach identifies patterns but does not establish causal relationships between incident characteristics and outcomes. Further research is needed to determine causality.
5. Human Factors: The study focuses on quantitative data and may not fully capture the qualitative aspects of incident management, such as the expertise level of staff handling the incidents.
6. Dynamic Nature of IT Environments: Rapid changes in IT infrastructure and services may affect the long-term applicability of the identified clusters.

CONCLUSION

This study demonstrates the effective application of the K-Means algorithm in clustering IT service incidents, revealing four distinct incident categories. Utilizing the elbow method, we identified the optimal number of clusters, yielding a distortion score of 964264294.569. The validity of this clustering was confirmed by a robust silhouette score of 0.52. The resulting clusters, differentiated by urgency, priority, and reassignment frequency, offer crucial insights into incident patterns, potentially revolutionizing IT incident management strategies.

The clustering results demonstrate the effectiveness of unsupervised learning techniques in identifying incidents requiring special handling, thereby improving operational efficiency within IT companies. Grouping similar incidents enables organizations to focus resources on resolving the most impactful issues, ensuring better *Service Level Agreement* (SLA) fulfillment. From a practical standpoint, IT companies can leverage these findings to develop targeted strategies for each cluster, potentially reducing response times and improving resource allocation. Additionally, proactive measures can be implemented for high-priority clusters to prevent incident escalation, and SLA policies can be refined based on cluster characteristics to better align with incident complexity and urgency.

This research contributes to the existing literature by demonstrating the application of K-Means clustering in IT incident management. It bridges the gap between data analytics and practical IT service management. The methodology presented here offers a data-driven framework for enhancing incident response strategies

While K-Means proved effective in this context, future research could explore other clustering algorithms, such as DBSCAN or hierarchical clustering, to compare their efficacy in incident categorization. Longitudinal studies could also provide insights into how incident patterns evolve, potentially uncovering seasonal trends or long-term shifts in IT service disruptions. Further areas for investigation include incorporating qualitative data to enrich the clustering analysis, exploring the impact of organizational and environmental factors on cluster formations, and developing predictive models based on cluster membership to forecast incident resolution times.

In conclusion, this study highlights the potential of data mining techniques to revolutionize IT incident management. By providing a more nuanced understanding of incident patterns, our research paves the way for more efficient, proactive, and tailored approaches to maintaining IT service quality and meeting SLA commitments. As organizations continue to grapple with increasing IT complexity, the application of advanced analytics in incident management will likely become not just beneficial, but essential.

REFERENCES

Atsa'am, D. D., Gbaden, T., & Wario, R. (2023). A machine learning approach to formation of earthquake categories using hierarchies of magnitude and consequence to guide emergency management. *Data Science and Management*, *6*(4), 208–213. https://doi.org/10.1016/j.dsm.2023.06.005

Baptista, B., & Barata, J. (2024). Continuously Improving IT Service Management in the Pharmaceutical Industry. *CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2023*, *239*, 923–930. https://doi.org/10.1016/j.procs.2024.06.253

Chen, Y., Tan, P., Li, M., Yin, H., & Tang, R. (2024). K-means clustering method based on nearest-neighbor density matrix for customer electricity behavior analysis. *International Journal of Electrical Power & Energy Systems*, *161*, 110165. https://doi.org/10.1016/j.ijepes.2024.110165

Claudio Amaral, M. F. (2018). *Incident management process enriched event log* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C57S4H

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14822
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Clayman, C. L., Srinivasan, S. M., & Sangwan, R. S. (2020). K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes. *"Complex Adaptive Systems"Malvern, PennsylvaniaNovember 13-15, 2019*, *168*, 97–104. https://doi.org/10.1016/j.procs.2020.02.265

Crase, S., Hall, B., & Thennadil, S. N. (2021). Feature Selection for Cluster Analysis in Spectroscopy. *Computers, Materials and Continua*, *71*(2), 2435–2458. https://doi.org/10.32604/cmc.2022.022414

Dener, M., & Gulburun, S. (2023). Clustering-Aided Supervised Malware Detection with Specialized Classifiers and Early Consensus. *Computers, Materials and Continua*, *75*(1), 1235–1251. https://doi.org/10.32604/cmc.2023.036357

Farou, Z., Wang, Y., & Horváth, T. (2024). Cluster-based oversampling with area extraction from representative points for class imbalance learning. *Intelligent Systems with Applications*, *22*, 200357. https://doi.org/10.1016/j.iswa.2024.200357

Fotopoulou, S. (2024). A review of unsupervised learning in astronomy. *Astronomy and Computing*, *48*, 100851. https://doi.org/10.1016/j.ascom.2024.100851

Goel, D., Husain, F., Singh, A., Ghosh, S., Parayil, A., Bansal, C., Zhang, X., & Rajmohan, S. (2024). X-Lifecycle Learning for Cloud Incident Management using LLMs. *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 417–428. https://doi.org/10.1145/3663529.3663861

Gokasar, I., Simic, V., Deveci, M., & Senapati, T. (2023). Alternative prioritization of freeway incident management using autonomous vehicles in mixed traffic using a type-2 neutrosophic number based decision support system. *Engineering Applications of Artificial Intelligence*, *123*, 106183. https://doi.org/10.1016/j.engappai.2023.106183

Hamamoto, N., Yokoyama, S., Takefusa, A., & Aida, K. (2021). Implemention of Secured Log Analysis Environment for Moodle using Virtual Cloud Provider Service. *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021*, *192*, 3154–3164. https://doi.org/10.1016/j.procs.2021.09.088

Luftensteiner, S., Chasparis, G. C., & Küng, J. (2024). PAS - A Feature Selection Process Definition for Industrial Settings. *5th International Conference on Industry 4.0 and Smart Manufacturing (ISM 2023)*, *232*, 308–316. https://doi.org/10.1016/j.procs.2024.01.030

Nurlaila, I., Irawati, W., Purwandari, K., & Pardamean, B. (2021). K-Means Clustering Model to Discriminate Copper-Resistant Bacteria as Bioremediation Agents. *5th International Conference on Computer Science and Computational Intelligence 2020*, *179*, 804–812. https://doi.org/10.1016/j.procs.2021.01.068

Palma, A., Acitelli, G., Marrella, A., Bonomi, S., & Angelini, M. (2024). A compliance assessment system for Incident Management process. *Computers & Security*, *146*, 104070. https://doi.org/10.1016/j.cose.2024.104070

Patel, E., & Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Third International Conference on Computing and Network Communications (CoCoNet'19)*, *171*, 158–167. https://doi.org/10.1016/j.procs.2020.04.017

Poerwanto, B. (2021). Evaluating the K-Means Analysis in Clustering Area Based on Estates Productivity in Tana Luwu Using Silhouette Index. *Journal of Physics: Conference Series*, *1752*(1), 012014. https://doi.org/10.1088/1742-6596/1752/1/012014

Sadeghi, B., Molayemat, H., & Pawlowsky-Glahn, V. (2024). How to choose a proper representation of compositional data for mineral exploration? *Journal of Geochemical Exploration*, *259*, 107425. https://doi.org/10.1016/j.gexplo.2024.107425

Santos, S. B. M. G., & Rodrigues, N. J. P. (2024). ServiceNow: Implications and Practice within the Business Environment. *CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2023*, *239*, 11–18. https://doi.org/10.1016/j.procs.2024.06.140

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 31. https://doi.org/10.1186/s13638-021-01910-w

Syaani, F. A., Mukarromah, A., & Fithriasari, K. (2020). Incident Clustering in the Warehouse Workspaces by Using Text Mining. *Materials Science and Engineering*.

Uddin, M. A., Talukder, Md. A., Ahmed, Md. R., Khraisat, A., Alazab, A., Islam, Md. M., Aryal, S., & Jibon, F. A. (2024). Data-driven strategies for digital native market segmentation using clustering. *International Journal of Cognitive Computing in Engineering*, *5*, 178–191. https://doi.org/10.1016/j.ijcce.2024.04.002

Verdonck, T., Baesens, B., Óskarsdóttir, M., & Vanden Broucke, S. (2024). Special issue on feature engineering editorial. *Machine Learning*, *113*(7), 3917–3928. https://doi.org/10.1007/s10994-021-06042-2

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Wang, F., Chen, J., & Liu, F. (2021). Keyframe Generation Method via Improved Clustering and Silhouette Coefficient for Video Summarization. *Journal of Web Engineering*. https://doi.org/10.13052/jwe1540-9589.2018

Yang, W., Long, H., Ma, L., & Sun, H. (2020). Research on Clustering Method Based on Weighted Distance Density and K-Means. *Proceedings of the 3rd International Conference on Mechatronics and Intelligent Robotics (ICMIR-2019)*, *166*, 507–511. https://doi.org/10.1016/j.procs.2020.02.056

Zhang, L. (2023). Research on K-means Clustering Algorithm Based on MapReduce Distributed Programming Framework. *3rd International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy*, *228*, 262–270. https://doi.org/10.1016/j.procs.2023.11.030

*name of corresponding author

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.