Process Mining on Customer Complaint Analysis using Inductive Miner and k-Means Clustering

Rahmalia Rahadi
School of Computing
Telkom University
Bandung, Indonesia
rahmarahadi@student.telkomuniversity.ac.id

Angelina Prima Kurniati
School of Computing
Telkom University
Bandung, Indonesia
angelina@telkomuniversity.ac.id

***Abstract*— UWV, a Dutch agency responsible for employee insurance and labor market services, provided a data set of customer complaints collected over eight months. Managing complex process flows is essential for companies to avoid delays, rework, and waste of resources. This research combines process mining with k-means to improve insights from clustering customer complaint data. Using the Inductive Miner algorithm for process mining and clustering the data with k-means based on similarity provides deeper insights and optimizes the analysis beyond what process mining alone can achieve. These insights allow UWV to customize its service strategy more effectively. The results showed that integrating process mining with k-means clustering significantly improved the precision of the process model, making it a more accurate representation of real-world processes. The model fitness improved by 6%, and the precision increased by 21% after clustering. However, this approach also resulted in a 37% decrease in generalization and a 1% decrease in simplicity. It indicates that while the clustered models are more detailed and precise, they become less adaptable to unobserved variation and slightly more challenging to interpret.**

Keywords*— *Customer Complaint, Inductive Miner, K-means Clustering, Process Mining

I. INTRODUCTION

In modern business management, managing complex process flows is essential for companies. This difficulty can cause various problems, such as delays, rework, and resource waste [1]. Business process management methods have been developed to ensure that business strategies align with the wishes of customers and interested stakeholders. Improvements that can be made include lowering failure rates, reducing costs, and improving process times [2].

UWV is an agency of the Ministry of Social Affairs and Employment (SZW) tasked with managing employee insurance and providing labor market services and data in the Netherlands [3]. UWV sought insights into customers' experiences through data collected over eight months from several sources, including complaint data showing when customers made complaints [3]. Currently, the data set is open to the public. It can be used for research purposes for process mining, including during the 2016 Business Process Intelligence Challenge (BPIC) as one of the five data sets combined for analysis.

Customer complaint data has yet to be analyzed further. Previous research using this data tends to focus on comparing the performance of process mining algorithms rather than gaining insight from the analyzed data [4]. One of the main challenges in working with customer complaint data is getting insight into the problem for customers and how to find out their complaints precisely so that the treatment is targeted. The most common way to deal with this problem is to cluster based on customer similarity [5]. Split traces based on similarities can be done using plugins or other tools in process mining. However, this is risky in handling extensive data, and no metric measures how accurately the data is split based on its characteristics.

With its unique position, process mining bridges the gap between data mining and business process modeling. The process mining approach uses event logs for business process analysis. By combining event logs and process models, process mining techniques provide insight into how processes run in an organization [6]. Through this analysis, organizations can identify patterns and trends in process execution. These techniques include three main classifications: Process Discovery, Conformance Checking, and Process Enhancement [7]. The process discovery algorithm used is Inductive Miner because this algorithm is very efficient in handling large logs with unique behavior [6]. Inductive Miner is designed to improve Alpha Miner's and Heuristics Miner's performance and ensure the resulting process model has good fitness values [8]. However, this algorithm is still weak in being sensitive to incomplete event logs or unusual data [9]. So, additional data preprocessing and complementary algorithms are needed to handle incomplete logs. The process discovery results in a detailed process model representing the business processes. Conformance analysis is then evaluated using four dimensions of process model quality: fitness, precision, generalization, and simplicity [10].

In this research, we propose to cluster data based on customer attributes. K-means clustering techniques can be applied to similar processes in groups based on their attributes, such as gender, customer age, or other relevant metrics. This helps identify patterns and trends in customer behavior, which can inform business process improvements [11]. Previous research combining process mining with k-means clustering improved the process model's performance by reducing the initial model's complexity [2], [5]. K-means have some disadvantages relevant to this research, as k-means are not optimal for categorical data because they rely on Euclidean distance [12]. However, this problem can be overcome using one-hot encoding, label encoder, or K-modes that use Hamming distance.

This paper is written with the following systematics: Section 1 is the Introduction, Section 2 is related studies, Section 3 contains the Research Methods, Section 4 describes the Results and Discussion, and Section 5 closes this paper with the Conclusion.

979-8-3503-6808-6/24/$31.00 ©2024 IEEE

II. Related Works

The following is research related to the mining process: Pramudia et al. [4]. This study uses the same data, namely UWV complaints for the mining process with a focus on comparing two algorithms, namely alpha miner and inductive miner. Alpha miner produces fitness of 0.48, precision of 0.79, and generalization of 0.83, while inductive miner produces fitness of 0.93, precision of 0.60, and generalization of 0.52.

Cirne et al [2]. This paper proposes the use of α Algorithm for process discovery, which is affected by noise, along with k-means clustering technique to improve the trace fitness value by reducing the initial model's complexity. The result is that the fitness function of the initial model shows superiority in conformance checking, using the token-based replay method, ranging from 6% to 30%.

Kurniati et al [5]. This research focuses on patient clustering using SimpleKMeans in process mining disease trajectory analysis can help to improve insights from sequence patterns in disease trajectories. Trace fitness has been increased by 0.0006, precision has been increased by 0.2600, generalization has decreased by 0.0754, and simplicity has decreased by 0.007.

In this study, we used the same data as previous research [4], but with a different approach by combining clustering with process mining [2]. We applied clustering before the process discovery stage using the inductive miner algorithm, which is more effective at handling noise to enhance process mining performance [5]. While using a similar approach to the research [5], this study is different because it fills a gap in the field of business management, with the goal of understanding customer complaints and business processes.

III. Methodology

Figure 1 illustrates the stages of the method used in this research. It starts with data exploration and understanding the content of the UWV customer complaint dataset. This is followed by data preprocessing to transform the data into an event log with only four variables. Then, experiments and a mining process with and without k-means will be conducted. Finally, the performance of the resulting model will be analyzed.

A. Data Understanding

UWV (Employee Insurance Agency) provided the dataset used in this study and covered eight months of data [3]. This data is collected from several sources, including complaint data that records when a customer makes a complaint. The dataset, also known as the event log, proved usable in process mining because it fulfills the four main attributes and was used in BPIC 2016. To maintain confidentiality, fields containing sensitive information have been masked. The event log consists of 289 rows and 18 columns or attributes. The existing event log data is still raw and requires processing before being analyzed. This preprocessing is of two types: one designed for process mining to extract event sequences and another for k-means clustering.

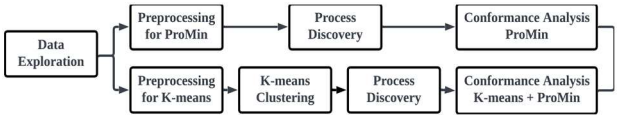


Fig. 1. Research Methodology

TABLE I. Event Log Parameter Column Definition

No.	Dataset Column	Attribute Parameter Set
1.	CustomerID	Case
2.	ComplaintTopic_EN	Activity
3.	ContactDate	Timestamp
4.	Gender	Resource
5.	AgeCategory	Other

TABLE II. Dataset Description

Dataset Column	Data Type	Description	Count
CustomerID	Integer	Unique customerID number	226
ComplaintTopic_EN	Object	Topic of the complaint	70
ContactDate	Object	Date the complaint was submitted	131
Gender	Object	M (male)	128
		F (female)	161
AgeCategory	Object	(18-29)	46
		(30-39)	85
		(40-49)	60
		(50-65)	98

TABLE III. Data Before and After Preprocessing

Before			
CustID	**Age**	**Gender**	**Topic**
1945807	50-65	M	(Revision) Decision not/too late taken
1202227	30-39	F	(Multiple) requests, to little avail
1203383	40-49	F	irregular payment
…	…	…	…
After			
0.333333	1.000000	0.0	0.014493
0.062222	0.333333	1.0	0.000000
0.066667	0.666667	1.0	0.565217
…	…	…	…

B. Data Preprocessing

a) Preprocessing for Process Mining

Preprocessing is the first step in preparing event log to be converted into quality data for process mining, which involves cleaning the data by removing attributes irrelevant to the research objectives [4]. In this stage, event logs are filtered to reduce complexity by determining the four main components of process mining: case ID, timestamp, resource, and activity [14]. Out of 18 attributes, only 5 are needed for analysis, as shown in Table I. This process is done using the Disco tool, which facilitates efficient identification and assignment of attributes [15]. After that, continue to filter the activities that will be displayed using the filter plugin in ProM 6.13 so that the resulting process model is easy to view. This preprocessing results in a clean and structured event log in XES format.

b) Preprocessing for k-means Clustering

In k-means, preprocessing is carried out to ensure the data is in optimal condition so that the clustering results are more accurate [16], [17]. The first step is to remove unnecessary columns and clean the data to overcome missing values and invalid data. Table II contains data description after the irrelevant columns are removed. The selected columns contain categorical data, so it is necessary to encode the data because k-means can only work with numerical data. We use Scikit-learn label encoder because the results are more suitable for representing the data. It encodes the target label

with a value between 0 and n_classes-1. For example, gender M is converted to 0 and V to 1. One approach to handling outlier values is data normalization, which rescales the dataset to (0.0-1.0) so that attributes with higher values do not dominate attributes with lower values [16]. The normalization methods used are Min Max Normalization. Finally, Principal Component Analysis (PCA) reduces dimensionality, feature extraction, and data visualization [18]. The tool used at this stage is Python, which has libraries such as Pandas, Scikit-learn, and NumPy. Table III shows the results of data preprocessing, which involves changing categorical data into numerical data. By converting the data, the k-means algorithm can more easily calculate the distance between data. In addition, this preprocessing helps reduce data complexity so that the algorithm can work more efficiently.

C. K-means Clustering

Cluster analysis aims to accurately group customers to address their complaints more effectively through personalization, commonly using the mathematical method known as k-means clustering [19]. Once the pre-processing stage is complete and the categorical data has been successfully converted into numerical data, the k-means algorithm works as follows [19]:

1) Randomly determine the number of centroids within the data.
2) Cluster the data into 'k' groups by associating each data point to the nearest centroid.
3) Calculate the average position of all objects in each cluster and move the centroid to that average position.
4) Repeat steps 2 and 3 until the assignment of data points to clusters does not change in the next iteration.

Next, we select the candidate 'k' and evaluate the optimal value. Evaluation of clustering output is important for data modeling [18]. There are different steps for clustering validation; the most used are [20]:

- *Silhouette coefficient (SC)*: The Silhouette coefficient measures how well the data is grouped into clusters. The range of Silhouette coefficient values is from -1 to 1. The higher the Silhouette coefficient value, the better the clusters are defined.
- *Davies-Bouldin (DB) index*: The Davies-Bouldin Index evaluates how well a clustering algorithm separates data into different clusters. The algorithm that produces the lowest Davies-Bouldin Index value is considered the best at separating clusters.

The well-clustered data will be exported in CSV format. Each cluster formed will be analyzed by process mining to understand the process of the resulting model and identify the effect of k-means on conformance analysis.

D. Process Mining

a) Process Discovery

Discovering and analyzing an organization's business processes is crucial for identifying key problem areas and enhancing the performance of existing processes. Process Discovery is fundamental to improving the quality of business process management with business process modeling [21]. Process discovery techniques focus on creating process models from event logs using various process mining algorithms, such as alpha miner, heuristics miner, and inductive miner. The resulting process model can be visualized in various notations, such as process tree, Petri Net, BPMN, and heuristics net [7].

This research applies an inductive miner algorithm to discover the process model of customer complaints. The algorithm was selected because it effectively manages large event logs and handles infrequent activities. [22]. This process uses the ProM tool as an established framework for academic process mining projects by selecting the "Mine Petri Net with Inductive Miner" plugin [23]. This plugin accepts XES format files as input and outputs the process model as a Petri Net. In this research, the input data for the discovery process consists of the original customer complaint event log and the event logs for each cluster generated by K-means. The event log must be properly filtered to make the Petri Net clear and easily visible.

b) Conformance Analysis

Conformance analysis relates events in the event log to activities in the process model to find similarities and differences between the log and the model. This technique is the final stage of this research, which is evaluated based on four quality dimensions. [10]:

1) Fitness measures how much of the behavior in the event log can be produced by the process model. The fitness value ranges from 0 to 1, where one indicates that the process model can completely replay every trace present in the event log through the generated process model and vice versa. The formulation is as follows [10]:

$$Q_f = 1 - \frac{|T_E \cap T_M|}{|C_E|} \quad (1)$$

T_E : traces from event log
T_M : traces from process model
$|T_E \cap T_M|$: number of traces in event log and process model
C_E : unique cases of events in the event log.

2) Precision measures how the model can capture the process behavior described in the event logs without oversimplifying it. Its value ranges from 0 to 1, where one indicates that the model has a high level of precision. This can be formulated as [10]:

$$Q_p = 1 - \frac{|T'_E \cap T_M|}{|C_E|} \quad (2)$$

$|T'_E \cap T_M|$: number of traces exclusively in process model and not the event log

3) Generalization measures the model's ability to handle variations the logs may not observe. A value close to 1 means the model can handle all possible variations. This can be calculated with the following formula:

$$Q_g = 1 - \frac{\sum_{a \in M_A} \sqrt{Ex(a)^{-1}}}{|M_A|} \quad (3)$$

a : denotes activities
M_A : unique activities in the process model
$Ex(a)$: number of executions of each activity referred

4) Simplicity measures how easily humans can understand the model, referred to as model complexity. The values range from 0 to 1, where 1 indicates a complex model with many events, while lower values indicate a simpler model. Simplicity is

calculated as follows:

$$Q_s = \frac{|M_D| + |E_A - M_A|}{|M_A| + |E_A|} \quad (4)$$

M_D : duplicate activities in the model
E_A : unique activities in the event log

$|E_A - M_A|$: number of missing events in the model

Conformance analysis was done using the ProM 6.13 tool. There are many different plugins for conformance or performance analysis in ProM [23]. "Replay Log on Petri Net for Conformance/Performance Analysis, Measure Precision /Generalization" by Adriansyah and "Multi-perspective Process Explorer" by Mannhardt can provide fitness, precision, and generalization values. Meanwhile, the simplicity value can be obtained by using PM4PY (a process mining library for Python) [24], as there is currently no ProM plugin that specifically provides simplicity values. After calculating the four quality dimensions for each cluster in the experiments with k-means, the average values were computed to compare with the results obtained from process mining without k-means.

IV. Result and Discussion

In implementing the k-means algorithm, a candidate number of clusters 'k' is randomly selected to evaluate the most optimal number using two evaluation metrics: Silhouette Score and Davies-Bouldin Index. The evaluation results for various 'k' candidates are listed in Table IV. The table shows that 8 clusters are the most optimal because they have the closest Silhouette Score (0.80) value to 1 and the smallest Davies-Bouldin Index (0.29) value, which means that the algorithm works well in separating data into different clusters. By choosing 8 clusters, the k-means algorithm can effectively group data points with similar characteristics.

Next, we analyzed the characteristics of each cluster. Table V presents information regarding each cluster. The resulting 8 clusters help determine the most frequent customer complaints based on age and gender. Further analysis yields the following points:

- The men aged 40-65 years and women aged 18-39 years often complain about incorrect or inconsistent information provided by UWV.
- The men aged 30-39 feel disinterested or need more attention from UWV. This indicates that customers in this group need to be more cared for or better served in their service processes.
- Women aged 50-65 years and men aged 18-29 years complained of serious problems related to payments not received by UWV.
- Women and men aged 18-29 years complained of insufficient information regarding insurance products or services.
- The women aged 40-49 feel disrespected or not taken seriously by UWV. Customers in this group may feel that their complaints or needs need to get the attention they deserve from customer service.

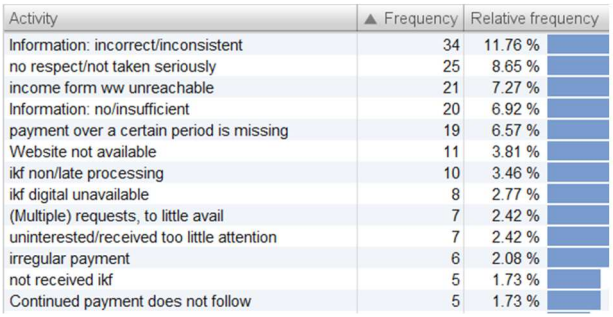
This cluster-based analysis can help UWV understand the most common complaints experienced by their various customer groups and enable them to adjust service and communication strategies to improve the overall customer experience.

TABLE IV. Metric Evaluation Results

Candidate 'k'	Silhouette Score	Davies-Bouldin Index
4	0.67	0.49
6	0.68	0.42
8	0.80	0.29
9	0.77	0.32

TABLE V. Cluster Characteristics

Cluster	Age	Gender	Most Complaint
1	50-65	M	Information: incorrect/inconsistent
2	30-39	F	Information: incorrect/inconsistent
3	30-39	M	uninterested/received too little attention
4	50-65	F	payment over a certain period is missing, income from ww unreachable, Website not available
5	18-29	F	Information: incorrect/inconsistent, Information: no/insufficient
6	40-49	F	no respect/not taken seriously
7	40-49	M	Information: incorrect/inconsistent
8	18-29	M	payment over a certain period is missing, Information: no/insufficient, not/hard to reach



Activity	Frequency	Relative frequency
Information: incorrect/inconsistent	34	11.76 %
no respect/not taken seriously	25	8.65 %
income form ww unreachable	21	7.27 %
Information: no/insufficient	20	6.92 %
payment over a certain period is missing	19	6.57 %
Website not available	11	3.81 %
ikf non/late processing	10	3.46 %
ikf digital unavailable	8	2.77 %
(Multiple) requests, to little avail	7	2.42 %
uninterested/received too little attention	7	2.42 %
irregular payment	6	2.08 %
not received ikf	5	1.73 %
Continued payment does not follow	5	1.73 %

Fig. 2. 20% of the Activities Frequently Complained About by Customers

In the discovery process, creating a process model directly from the event log produces a complicated model. This condition shows the importance of simplifying visualization to increase process understanding. Figure 3 shows 13 of the customer complaint dataset's most frequent activities or complaints. These frequent activities can be used as a filter to simplify the process model. The simplified process model is shown in Figure 2. This process model illustrates the activities or complaints customers submit over eight months. For example, as observed in Figure 2, some customers have only complained once, while others have complained four times on different topics.

The resulting complete model process allows us to analyze the model's quality. Table VI presents the conformance analysis model values from the original event log before clustering. The fitness value shows that the process model can reproduce 93% of the activity sequences in the event log. The precision value shows that 76% of the behavior produced by the model occurs in the event log.

TABLE VI. Comparison of Conformance Analysis

	Conformance Analysis			
	Fitness	Precision	Generalization	Simplicity
Without K-means	0.93	0.76	0.57	0.56
With K-means	0.99 (+0.06)	0.97 (+0.21)	0.20 (-0.37)	0.55 (-0.01)

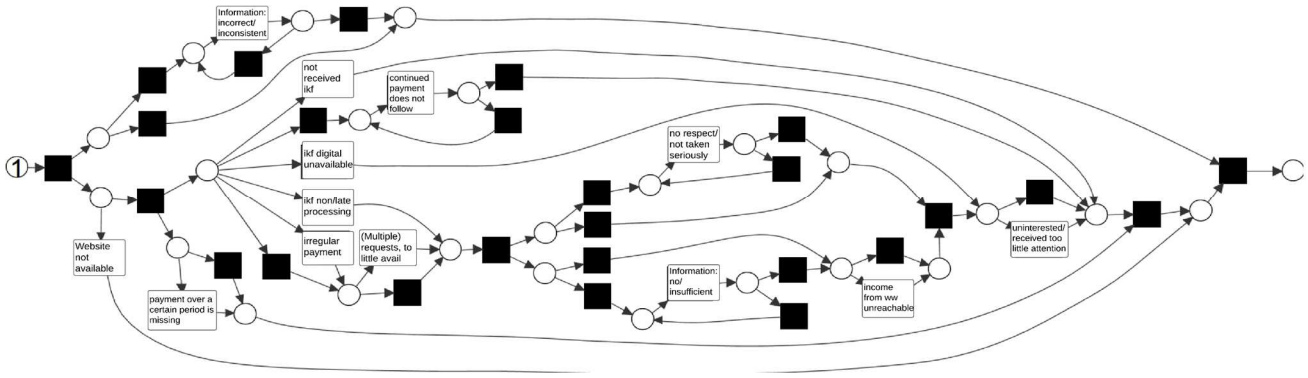


Fig. 3. Simplified Process Model (20% Activity) using Inductive Miner

So, the model is irrelevant if used for the new event log variance. The simplicity of 0.56 indicates that the model is not too complex but not too simple, reflecting a balance between detail and ease of understanding. Usually, the values of generalization and simplicity are the opposite. Low values on both may indicate that the model is too complex and not flexible enough to handle new variations in the data.

The precision value is less than satisfactory, where 24% of the model may still produce some irrelevant behavior or does not occur in the real system. We experimented to see whether the use of K-means affected increasing or decreasing this value. Experiments were carried out by dividing the data into several clusters using the k-means algorithm and then creating a process model for each cluster. This process makes it possible to understand whether a more segmented process model can provide a more accurate and relevant representation. After that, we calculated the conformance analysis of the eight resulting process models and averaged the results. The graph in Figure 4 shows the results of this experiment.

The 6% increase in fitness indicates that the more segmented model through clustering can capture more variability in the data and reflect business processes more accurately. The increase in precision shows that the resulting model after clustering is 21% more likely to match the actual log data, reducing previous errors and inaccuracies. There was a 37% decrease in generalization and 1% decrease in simplicity. This happens because the segmented model is too specific to the data in each cluster, so the resulting model can only handle other variations seen in the cluster. The decrease in generalization and simplicity values is a limitation of this study. The same happened in a similar study using larger data [5]. No study increases the generalization and simplicity values after clustering. More advanced clustering methods can further improve performance.

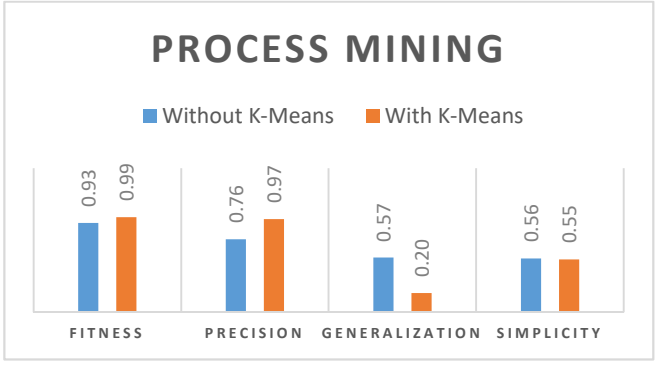


Fig. 4. Comparison Graph of K-means Clustering Usage

V. Conclusion

In this study, we propose customer clustering to improve process mining performance. By clustering customers, we found that it is applicable to improve the performance and conformance value and gain essential insights into customer characteristics. K-means shows good performance in clustering customers as measured by evaluation metrics. The approach in this study can also be used on different data that meet the minimum attributes of process mining in other fields, such as healthcare, retail, and banking, that need insights about their customers to improve the service system.

Further research shows that technical improvements can be made by improving and applying the clustering method to more complex case studies. Conclusion: K-means has shown promising results. More advanced clustering methods will improve performance further. On the other hand, customer service improvements can be made by analyzing the complaint aspects of the data and using the findings to support customer service improvement recommendations.

References

[1] T. Benedict *et al.*, *BPM CBOK Version 4.0: Guide to the Business Process Management Common Body Of Knowledge*. 2019.

[2] R. Cirne, C. Melquiades, R. Leite, E. Leijden, A. MacIel, and F. B. D. L. Neto, 'Data Mining for Process Modeling: A Clustered Process Discovery Approach', in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 587–590. doi: 10.15439/2020F95.

[3] M. Dees and B. F. (B. van Dongen, "BPI Challenge 2016." UWV, Apr. 22, 2016, doi: 10.4121/UUID:360795C8-1DD6-4A5B-A443-185001076EAB.

[4] R. G. Pramudia, R. Ariandi, F. S. Salma, and R. Andreswari, 'Process Mining Analysis and Implementation on Customer Complaints Dataset', in *2022 International Conference of Science and Information Technology in Smart Administration, ICSINTESA 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 30–34. doi: 10.1109/ICSINTESA56431.2022.10041581.

[5] A. P. Kurniati *et al.*, 'Patient Clustering to Improve Process Mining for Disease Trajectory Analysis Using Indonesia Health Insurance Dataset', in *2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, May 2024, pp. 88–93. doi: 10.1109/ICAIBD62003.2024.10604436.

[6] W. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. doi: 10.1007/978-3-662-49851-4.

[7] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19345-3.

[8] A. Bogarín, R. Cerezo, and C. Romero, 'Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs).', *Psicothema*, vol. 30, no. 3, pp. 322–329, Aug. 2018, doi: 10.7334/psicothema2018.116.

[9] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, 'Discovering Block-Structured Process Models from Event Logs - A Constructive Approach', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7927 LNCS, 2013, pp. 311–329. doi: 10.1007/978-3-642-38697-8_17.

[10] A. T. S. Ireddy and S. V. Kovalchuk, 'An Experimental Outlook on Quality Metrics for Process Modelling: A Systematic Review and Meta Analysis', *Algorithms*, vol. 16, no. 6, p. 295, Jun. 2023, doi: 10.3390/a16060295.

[11] L. Al-Dabbas, H. Al-Tarawneh, and T. A. Al-Rawashdeh, 'Customer Personality Segmentation Using K-Means Clustering', in *2023 International Conference on Information Technology (ICIT)*, IEEE, Aug. 2023, pp. 537–543. doi: 10.1109/ICIT58056.2023.10225996.

[12] K. S. Dorman and R. Maitra, 'An efficient K-modes algorithm for clustering categorical datasets', *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 1, pp. 83–97, Feb. 2022, doi: 10.1002/sam.11546.

[13] H. N. Prasetyo, R. Sarno, R. Budiraharjo, and K. R. Sungkono, 'The Effect of Duration Heteroscedasticity to the Bottleneck in Business Process Discovered by Inductive Miner Algorithm', in *Proceedings - 2021 IEEE Asia Pacific Conference on Wireless and Mobile, APWiMob 2021*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 52–58. doi: 10.1109/APWiMob51111.2021.9435199.

[14] R. Rahmawati, R. Andreswari, and R. Fauzi, 'Analysis and Exploratory of Lecture Preparation Process to Improve the Conformance using Process Mining', in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Jan. 2022, pp. 0461–0466. doi: 10.1109/CCWC54503.2022.9720762.

[15] P. Porouhan, 'Optimization of Overdraft Application Process with Fluxicon Disco', in *2022 20th International Conference on ICT and Knowledge Engineering (ICT&KE)*, IEEE, Nov. 2022, pp. 1–12. doi: 10.1109/ICTKE55848.2022.9983238.

[16] D. Usman and F. S. Stores, 'On Some Data Pre-processing Techniques For K-Means Clustering Algorithm', *J Phys Conf Ser*, vol. 1489, no. 1, p. 012029, Mar. 2020, doi: 10.1088/1742-6596/1489/1/012029.

[17] M. Zulkifilu and A. Yasir, 'About Some Data Precaution Techniques For K-Means Clustering Algorithm', *UMYU Scientifica*, vol. 1, no. 1, pp. 12–19, Sep. 2022, doi: 10.56919/usci.1122.003.

[18] A. Abdulhafedh, 'Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation', *Journal of City and Development*, vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.

[19] A. Kumar, 'Customer Segmentation of Shopping Mall Users Using K-Means Clustering', in *Advancing SMEs Toward E-Commerce Policies for Sustainability*, IGI Global, 2023, ch. 13, pp. 248–270. doi: 10.4018/978-1-6684-5727-6.ch013.

[20] G. Liu, 'A New Index for Clustering Evaluation Based on Density Estimation', Jul. 2022.

[21] W. M. P. van der Aalst, 'Business Process Management: A Comprehensive Survey', *ISRN Software Engineering*, vol. 2013, pp. 1–37, Feb. 2013, doi: 10.1155/2013/507984.

[22] M. W. Wibisono, A. P. Kurniati, and G. A. A. Wisudiawan, 'Process Mining using Inductive Miner Algorithm to Determine the actual Business Process Model', *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, p. 1128, Aug. 2022, doi: 10.30865/jurikom.v9i4.4769.

[23] F. Yasmin, R. Bemthuis, M. Elhagaly, and F. Wijnhoven, 'A Process Mining Starting Guideline for Process Analysts and Process Owners: A Practical Process Analytics Guide using ProM', Jul. 2020. [Online]. Available: www.processmining.org.

[24] A. Berti, S. van Zelst, and D. Schuster, 'PM4Py: A process mining library for Python[Formula presented]', *Software Impacts*, vol. 17, Sep. 2023, doi: 10.1016/j.simpa.2023.100556.