Journal of
Information Systems Engineering
and Business Intelligence

Vol.10, No.1, February 2024
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover

Ni Putu Sutramiani [1)*], I Made Teguh Arthana [2)], Pramayota Fane'a Lampung [3)], Shana Aurelia [4)], Muhammad Fauzi [5)], I Wayan Agus Surya Darma [6)]

[1)2)3)4)5)] *Department of Information Technology, Faculty of Engineering, Universitas Udayana, Denpasar, Indonesia*

[1)]sutramiani@unud.ac.id, [2)]teguh.arthana005@student.unud.ac.id, [3)]pramayotalampung@student.unud.ac.id,
[4)]aurelia_na037@student.unud.ac.id, [5)]masfauzi001@student.unud.ac.id

[6)] *Department of Informatics, Faculty of Technology and Informatics, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*

[6)]surya@instiki.ac.id

Abstract

Background: This study focuses on the latest knowledge regarding Micro, Small and Medium Enterprises (MSMEs) as a current central issue. These enterprises have shown their significance in providing employment opportunities and contributing to the country's economy. However, MSMEs face various challenges that must be addressed to optimize their outcomes. Understanding the characteristics of this group was crucial in formulating effective strategies.
Objective: This study proposed to cluster or combine micro, small, and medium enterprises (MSMEs) data in a particular area based on asset value and turnover. As a result, this study aimed to gain insights into the MSME landscape in the area and provided valuable information for decision-makers and stakeholders.
Methods: This study utilized two methods, namely the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method and the K-Means method. These methods were chosen for their distinct capabilities. DBSCAN was selected for its ability to handle noisy data and identify clusters with diverse forms, while K-Means was chosen for its popularity and ability to group data based on proximity. The study used a dataset containing MSME information, including asset values and turnover, collected from various sources.
Results: The outcomes encompassed identifying clusters of MSMEs based on their closeness in the feature space within a specific region. Optimizing the clustering outcomes involved modifying algorithm parameters like epsilon and minimum points for DBSCAN and the number of clusters for K-Means. Furthermore, this study attained a deeper understanding of the arrangement and characteristics of MSME clusters in the region through a comparative analysis of the two methodologies.
Conclusion: This study offered perspectives on clustering MSMEs based on asset value and turnover in a specific region. Employing DBSCAN and K-Means methodologies allowed researchers to depict the MSME landscape and grasp the business attributes of these enterprises. These results could aid in decision-making and strategic planning concerning the advancement of the MSME sector in the mentioned area. Future study may investigate supplementary factors and variables to deepen comprehension of MSME clusters and promote regional growth and sustainability.

Keywords: Asset Value, Clustering, DBSCAN, K-Means, Turnover

Article history: Received 19 July 2023, first decision 29 September 2023, accepted 9 December 2023, available online 28 February 2024

I. INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) are often regarded as one of the main pillars in supporting the economic stability of a region. MSMEs play a crucial and strategic role in developing the national economy. They contribute to distributing the development process's benefits, economic growth, and job creation [1]. A survey conducted by PricewaterhouseCoopers found that 74% of MSMEs do not have access to financing [2]. It is a significant challenge for MSMEs, which comprise 99.99% of the total business population and employ 96.9% of workforce [3]. MSME entrepreneurs may have numerous business ideas to develop their ventures, but they often get

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.10.1.13-24

stuck due to the lack of additional funding. Many MSMEs struggle to obtain additional capital from financial institutions due to unfulfilled requirements.

The categorization of MSMEs can aid the government in formulating marketing strategies and maximizing the potential of MSMEs. Based on this background introduction, the study proposed the clustering method by grouping the MSMEs into several clusters based on their similar characteristics [4]. Clustering is a data analysis method that seeks to organize data into multiple clusters, grouping similar objects together within the same cluster while separating dissimilar ones into distinct clusters [5]. The primary goal of clustering is to maximize homogeneity within a cluster and heterogeneity among different clusters [6].

The role of clustering in information analysis is continually evolving in tandem with the rapid advancements in the fields of data mining, machine learning, artificial intelligence, and various other disciplines [7]. Clustering has been widely utilized in various fields, such as hotspot data clustering, customer segmentation, customer behavior analysis, and many more [8]. This algorithm employs iterative techniques to group data set samples into categories that contain similar features [9]. As one of the unsupervised classification techniques, clustering aims to group sample objects into different clusters without prior knowledge, so that objects within the same cluster will exhibit higher similarity to each other compared to objects in different clusters [4].

DBSCAN was first proposed in the 1990s as a density-based clustering algorithm. There are three main characteristics of DBSCAN, namely its ability to discover clusters with various shapes, accurately identify noise, and not require a pre-determined number of clusters [9]. In DBSCAN, n query regions are needed for a dataset consisting of n objects, where each query region consists of the local neighbor calculations of a specific object within the dataset. The process of expanding a cluster that contains the current object (or creating a new cluster based on the current object) depends on the condition that the environment around the current object must contain at least a certain number of objects [9]. DBSCAN has higher homogeneity and diversity when it performs personalized clustering on data sets of non-uniform density with broad values and gradually sparse forwards [10]. DBSCAN implemented uses a concurrent execution of Reverse Nearest Neighbor graph traversals using density-based clustering [11]. Other study applied heuristic based iterative DBSCAN clustering algorithm [12].The study proposed MinPts, the new approach allows for variable grouping of data with varying densities. Specifically, the new clustering is done based on the MinPts after the first phase of estimating the values of MinPts based on user area densities. In this way, too dense clusters are avoided. The revised DBSCAN algorithm's clustering effect is assessed in the trials using purity and silhouette coefficient. It is shown that the suggested approach performs better at clustering all types of data and is more flexible in handling multi-density data.

Our study is not limited to one country; it addresses a universal concern. The challenges faced by MSMEs in accessing financing resonate globally. We believe that the methods and insights derived from our study can benefit the international community. Based on previous studies, our study takes shape, building upon previous studies which highlighted the vulnerability of MSMEs, particularly during the COVID-19 pandemic, using innovative clustering techniques [13]. Our goal is to leverage advanced data analysis methods to categorize MSMEs and offer solutions to the financing challenges faced by these enterprises, contributing to a global understanding of this critical issue. This categorization will empower the government with insights to formulate effective marketing strategies and unlock the potential of these enterprises by employing advanced data analysis techniques and filling a critical research gap. The research gap based on previous research is that it has not carried out a comparison of clustering models to find out the most optimal model for MSME data. So this study discusses a comparison of two methods to determine the performance of cluster models that are relevant to the data used.

II. Literature Review

Clustering algorithms play a crucial role in various domains, including data analysis, network threat analysis, intrusion detection, and more. The clustering process involved utilizing a data analysis technique to categorize data into various clusters, wherein comparable entities were grouped in a single cluster. Meanwhile, dissimilar ones were segregated into separate clusters. The comparison and evaluation of these algorithms have been extensively explored in the literature. Abdullah et al [14], focused on clustering Covid-19 cases worldwide. They compared the performance of DBSCAN and K-Means algorithms in this context. The results showed that both algorithms effectively grouped Covid-19 cases, but DBSCAN performed better in handling noise and outliers. The study highlighted the importance of considering density-based approaches like DBSCAN for complex and noisy datasets. Similarly, Cinderatama [15] applied clustering algorithms, including K-Means, DBSCAN, and Meanshift, to network threat analysis in intrusion detection systems. Their study aimed to identify patterns and anomalies in network traffic data. The authors found that DBSCAN outperformed K-Means and Meanshift in detecting network threats, indicating its potential to enhance the security of intrusion detection systems.

The K-Means method was previously proposed by Wahyuri et al [5]. The study's findings were derived from an extensive dataset comprising five distinct attributes: drug name, therapeutic class, district/city, sample category, and drug surveillance evaluation. Through analysis, the data unveiled distinct drug distribution patterns, resulting in the identification of three primary clusters. The first cluster comprised 522 drug items spanning eight therapeutic classes, distributed across ten districts. In contrast, the second cluster included 1,542 drug items within five therapeutic classes, dispersed across five districts. The third cluster, encompassing 503 drug items across eleven therapeutic classes, was observed across nine districts.

The DBSCAN method was previously proposed by Cui et al [6] was employed to cluster grain temperature statistical parameters in this research study for identifying distinct grain inventory modes, specifically focusing on empty storage and aeration. By utilizing the DBSCAN algorithm with statistical grain temperature parameters gathered over a year from 27 grain warehouses in China, the study successfully grouped these parameters. Subsequent parameter analysis and experimental investigations were conducted to detect the grain inventory modes. The analysis revealed that parameters such as temperature differences between adjacent layers and the ratio of grain temperature aggregation from multiple layers effectively identified empty warehouses with a precision and recall rate of 100%. Furthermore, changes in grain temperature rates and standard deviation rates were identified as indicative of aeration periods, achieving a recall rate of approximately 85.4% and a precision rate of around 97.4%, as determined through parameter analysis.

Clustering is vital for developing MSMEs as it facilitates analysis and decision-making to enhance efficiency and competitiveness. It also offered benefits, such as identifying MSME characteristics, devising effective marketing strategies, improving product quality, fostering collaboration, enhancing information exchange, and providing better access to resources. Kusmiati et al [16] proposed stratifying micro, small, and medium enterprises (MSMEs) within the Garut Regency, utilizing a clustering approach grounded in common attributes and factors. This encompassed ownership of taxpayer identification and business licenses, overall net worth and capital, employee count, business volume, product certification, and licensing, as well as brand presence. Employing the K-Means clustering technique with RapidMiner 5.3 software, the study delineated three discrete clusters among MSMEs, shedding light on their inherent distinctions and groupings within the region's economic landscape. Cluster 3 displayed the most significant characteristics, including a larger number of MSMEs, higher total net worth, and ownership of product and brand licenses. Cluster 2 had the largest percentage of MSMEs, while Cluster 1 had more MSMEs with business license ownership. These results could inform local government policies in the Garut Regency regarding cluster formation for MSMEs.

The literature also included identification works that laid the foundation for clustering algorithms. In recent years, environmental sustainability has gained prominence in Australian cities, with Virtual Power Plant (VPP) emerging as a promising tool for urban sustainability. One critical aspect of VPP modeling was understanding local conditions, particularly the electricity demand profiles. However, data privacy limitations often hinder the availability of detailed energy demand profiles. To overcome this challenge, Wang et al [17] proposed a clustering technique to extract electrical patterns from high-dimensional data in an Australian city. The study applied K-Means and DBSCAN clustering algorithms, each with its unique advantages and disadvantages, and the choice depended on the specific clustering objectives. These algorithms provided valuable insights for VPP modeling. Meanwhile, in the educational context, Aulia and Nurahman [18] proposed a comparison of two clustering techniques to categorize education levels in regions. The algorithms used are K-Means and K-Medoids. K-Means demonstrated superior performance with a lower Davies Bouldin Index (DBI) value, enabling the categorization of education levels into six clusters. These findings were instrumental in informing educational policy decisions in the region.

Furthermore, the literature included a comparison process for the K-Means and DBSCAN algorithms. Andriyani and Puspitarani [19] proposed a comparative analysis of the K-Means and DBScan clustering algorithms in the context of product review clustering. The objective was to determine which algorithm performs better in terms of accuracy. These two algorithms were selected based on their distinct clustering methodologies: K-Means utilized centroid-based clustering, while DBScan employed density-based clustering. The outcomes of text clustering have practical applications in e-commerce platforms, marketplaces, and product review platforms, aiding customers in making informed product choices. Customers often face challenges in assessing products due to the sheer volume of reviews and difficulty extracting relevant information. Text clustering simplified this process by enabling quicker and more accessible product evaluation based on reviews. The study utilized product review data from the Female Daily website, subjecting it to text preprocessing before applying the K-Means and DBScan algorithms for clustering. The results demonstrated that DBScan outperformed K-Means, achieving an accuracy rate of 99.80%, while K-Means achieved an accuracy rate of 99.50%. This finding was significant for the literature review, particularly in the context of text-based clustering for product reviews, highlighting the superior performance of DBSCAN in this specific application.

III. METHODS

A. General Process

Data mining is a term used to describe the process of discovering knowledge within a dataset [20]. The process employed statistical methodologies, mathematical principles, artificial intelligence, and machine learning to extract and recognize valuable information and pertinent knowledge from extensive databases. Alternatively, data mining explored significant associations, patterns, and trends. It involved examining a vast data collection stored in a repository using pattern recognition methods like statistics and mathematics. Data mining emerged due to the abundance of data overload experienced by various institutions, companies, or organizations.

It was caused by the accumulation of transactional data recorded over the years. Essentially, data mining serves the purpose of specifying the patterns that need to be discovered in the data mining process. Generally, the task can be divided into four stages. There were six stages in the data mining process, where the first three stages were also referred to as data preprocessing (including data cleaning, data integration, and data transformation), which typically take up approximately 60% of the entire process. It was followed by creating data mining models, pattern evaluation, and knowledge presentation [20]. The data mining process can be illustrated as shown in Fig. 1.

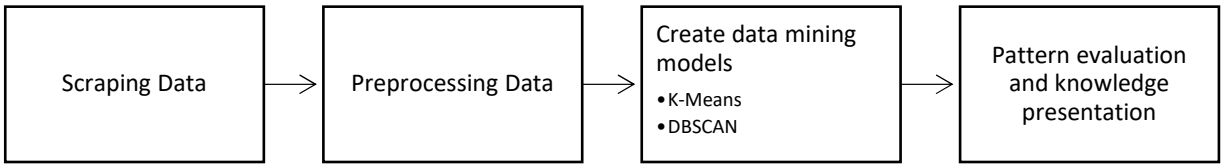

Fig. 1 General process

B. Data Collection

The dataset used consisted of scraped data on MSMEs in Aceh province, which comprised 7002 rows and 10 columns (UMKM Name, UMKM Sector, Business Sector, District, Subdistrict, Village, Business Type, Workforce, Assets, and Revenue). Clustering (grouping) was performed based on the values of Assets and Revenue using the K-Means and DBSCAN algorithms. The data scraping process was conducted using the Selenium library in Python, and the scraped data was stored in the CSV format. The dataset consists of 10 fields, i.e., the names of the MSMEs, business sectors, MSME sectors (Micro or Small), districts, sub-districts, villages, types of businesses, number of employees, turnover, and assets.

TABLE 1
ACEH MSMES DATASET

	Business Name	Business Sector	MSME Sector	Regency	Sub-District	Village	Type of Business	Work force	Turnover	Assets
0	Mutmainnah	Trade	Small	Banda Aceh	Baiturahman	Jl. Seulawah	Photocopy	2	110.0	60.000
1	Dibaharuddin	Trade	Micro	Banda Aceh	Baiturahman	Pasar Aceh	Clothing Sales	2	75.0	45.000
2	H. Mahyuddin MT	Trade	Small	Banda Aceh	Baiturahman	Pasar Aceh	Coffee Sales	2	25.0	37.500
3	H. Mahyuddin MT	Trade	Small	Banda Aceh	Baiturahman	Pasar Aceh	Phone Credits	2	15.0	22.500
4	Fajri Abdullah	Trade	Small	Banda Aceh	Baiturahman	Pasar Aceh	Coffee Sales	2	20.0	30.000
…	…	...	...	…	…	…	...	…	…	…
5095	Raymosn Tailor	Industry	Micro	Banda Aceh	Jaya Baru	Lamteumen Barat	Clothing Tailor	2	37.500	25.000
5096	Rekha Gorden	Industry	Micro	Banda Aceh	Jaya Baru	Seutui	Curtain Tailor	2	37.500	25.000
5097	Ria Gorden	Industry	Micro	Banda Aceh	Jaya Baru	Seutui	Curtain Sewing	2	37.500	25.000
5098	Indomer Utama	Industry	Small	Banda Aceh	Jaya Baru	Lamteumen	Printing	3	38.000	75.000
5099	Aceh Beudoh Makmur	Industry	Small	Banda Aceh	Jaya Baru	Stui	Printing	3	39.000	100.000

The data collection process on the website of the Small and Medium Enterprises (UMKM) website in Aceh Province involved a total of 7002 data. Based on the available data, various aspects that aided in understanding the Micro, Small, and Medium Enterprises (MSMEs) could be identified. The data encompassed information regarding the names of the MSMEs, business sectors, MSME sectors (Micro or Small), districts, sub-districts, villages, types of businesses, number of employees, turnover, and assets. The sectors included trade and industry, while the MSME sector was divided into Micro and Small. The types of businesses vary, such as photocopying, clothing retail, coffee sales, tailoring, printing, etc. The MSMEs were distributed in Banda Aceh District with specific details on sub-district and village locations. Furthermore, there was a variation in the number of employees, turnover, and assets for each MSME. Analyzing the grouping of MSMEs based on business sectors, MSME sectors, or types of businesses would provide a deeper understanding of the characteristics and distribution of MSMEs in each cluster. Table 1 shows the detailed MSMEs dataset.

C. Pre-Processing

Data preprocessing was performed in every analysis method. The preprocessing stages were removing duplicate data, fixing the data type on the variable, filling in missing values, and scaling data, which will be further discussed in the results and discussion section. However, a numeric distance is added in hierarchical analysis to determine the Euclidean distance in the numerical columns. The stages include as follows:

1) Remove duplicate data

In this preprocessing stage, this study focused on identifying and removing data entries that exhibit identical or highly similar values within a dataset. Data duplication could manifest in various types of datasets, including tabular data, text, or even image data.

2) Fix the data type on the variable

This stage involved adjusting or rectifying the data type assigned to a variable within a computer program. In programming, variables were utilized to store specific values or data, and each variable was associated with a data type that dictates the type of values it could accommodate. This study encountered several variables with data types that did not align with our expectations. For instance, variables like “assets” and “turnover” still had string data types, which could hinder the modeling process in the DBSCAN and K-Means models. Therefore, it was imperative to convert them to integer data types.

3) Fill in missing value

In this data preprocessing step, the study addressed the task of substituting or filling in missing or empty values within a dataset with appropriate values. When encountering an incomplete dataset or data gaps, inputting missing values became a crucial step in preparing the data before proceeding with further analysis or model development.

4) Scaling the data

This stage pertained to standardizing the data scale for each feature (variable) within a dataset. The objective was to harmonize the range of values for each feature, ensuring they shared a similar or proportionate scale. Scaling data was of paramount importance in data analysis and machine learning, as many algorithms and methods required data to be uniformly scaled to achieve optimal results. In this study, we applied data scaling to the “asset” and “turnover” variables since these two variables had dissimilar value ranges. “Assets” ranged from 20,000 rupiahs to 10,000,000 million rupiahs, while “turnover” ranged from 110 rupiahs to 5000,000 million rupiahs. This study aimed to expedite the modeling process, reduce computational complexity, and prevent biases arising from varying scale values among variables by standardizing the data.

The first step is removing duplicate data. This is a stage in data preprocessing to remove data with twins or the same values. In the process of eliminating duplicate data, 1326 data had the same value, so it needed to be removed. After removing the available dataset, there were 6257 data. After eliminating duplicates in the data, the next step was to correct unappropriated data types. For example, the variables “assets” and “Turnover” had numeric values but used string data types. Thus, it needed to be corrected to make it easier when carrying out the modeling process.

The next step after correcting the data type was to continue by filling in the empty values for each variable. In the dataset used, there was 1 row of data that had an empty value, namely the variable “type of business”, so it was filled in using the mode value from the dataset with the variable “type of business”. Variable normalization is a transforming or adjusting data step to have a consistent scale. In this study, we utilized the min-max scaler normalization to transform the range of values for the variables “Turnover” and “Asset” into the range of 0 to 1. Table 2 shows the normalization results using the Min-Max Scaler on the “Turnover” and “Asset” variables. As can be seen, both variables have been successfully normalized and converted to the range of 0 to 1.

,

TABLE 2
THE RESULT OF MIN-MAX SCALER NORMALIZATION ON THE "TURNOVER" AND "ASSET" VARIABLES

	Business Name	Business Sector	MSME Sector	Regency	Sub-District	Village	Type of Business	Workforce	Turnover	Assets
0	Mutmainnah	Trade	Small	Banda Aceh	Baiturahman	Jl. Seulawah	Photocopy	2	0.000545	0.000028
1	Dibaharuddin	Trade	Micro	Banda Aceh	Baiturahman	Pasar Aceh	Clothing Sales	2	0.000370	0.000021
2	H. Mahyuddin MT	Trade	Small	Banda Aceh	Baiturahman	Pasar Aceh	Coffee Sales	2	0.000120	0.000017
3	H. Mahyuddin MT	Trade	Small	Banda Aceh	Baiturahman	Pasar Aceh	Phone Credits	2	0.000070	0.000010
4	Fajri Abdullah	Trade	Small	Banda Aceh	Baiturahman	Pasar Aceh	Coffee Sales	2	0.000095	0.000014

D. Algorithm

1) DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering technique operated by expanding regions with a significant density into clusters, enabling the identification of clusters with arbitrary shapes within a spatial database that might include noise [21]. DBSCAN defined a cluster as a maximum set of density-connected points. Any objects that did not belong to any cluster were considered noise.

DBSCAN, a density-based clustering approach, examined clusters by analyzing the ε-neighborhood of each data point. When the ε-neighborhood of a point p contained more objects than the specified MinPts threshold, a new cluster was established with p as a core object. Subsequently, DBSCAN progressively gathered density-reachable objects from the core object, potentially resulting in the merging of multiple density-reachable clusters. The fundamental principles of this method could be summarized as follows in the paragraph below.

Firstly, the ε-neighborhood of a data object refers to the proximity within a radius Θ. Secondly, an object is categorized as a core object if its ε-neighborhood contains at least the minimum required number of other objects, denoted as MinPts. Thirdly, in a set of objects D, an object p is considered density reachable from object q concerning ε and MinPts if there exists a sequence of objects p1, p2, ..., pn, where p1 = q and pn = p. Each consecutive pair pi and pi+1 in the sequence is directly density reachable from each other concerning ε and MinPts, for $1 \leq i \leq n$, with all pi being members of D. Lastly, in a set of objects D, an object p is regarded as density connected to object q concerning ε and MinPts if there exists an object o, which is a member of D, such that both p and q are density reachable from o concerning ε and MinPts.

2) K-Means

K-Means clustering is a widely used method for clustering data based on their proximity. The algorithm functioned by iteratively assigning data points to their closest centroids and updating the centroids until convergence was achieved [21]. The K-Means algorithm process involved several steps. First, select the desired number of clusters, denoted as **k**. Next, k centroids were randomly placed within the feature space as an initial step. Then, each data point was assigned to the nearest centroid using the Euclidean distance as a measure. Afterward, the centroids were updated by calculating the average of all data points assigned to each centroid. These assignments and updates continued iteratively until the centroids experienced minimal changes or the maximum specified number of iterations was reached.

This clustering approach aimed to minimize the total sum of squared distances within clusters, often referred to as inertia or distortion. The algorithm could handle numeric data and was widely used for various clustering tasks. It did not require pre-defined density thresholds or assumptions about cluster shapes, making it a versatile and efficient method. To apply the K-Means algorithm to the specific data and determine the optimal number of clusters, we need to adapt the mentioned steps and consider the characteristics and requirements of the dataset.

E. Modeling

Fig. 2 shows the flowchart of the modeling stage. The modeling stage involved creating K-Means and DBSCAN models with the best parameters to obtain the DBI score. After obtaining the lowest DBI score, the next step was to analyze the clustered data.

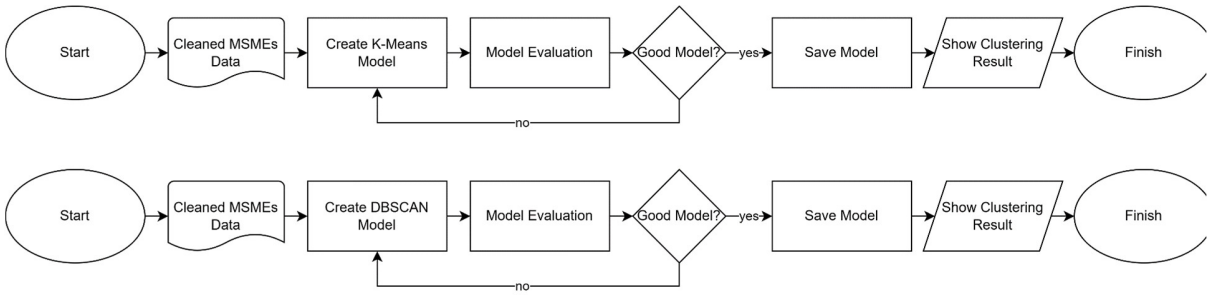

Fig. 2 Modeling process on K-Means and DBSCAN

F. Evaluation

Evaluation refers to assessing and measuring the quality of the results or models generated from the data mining process. Evaluation is conducted to understand the extent to which the results or models can be relied upon, are relevant, and align with the objectives of the analysis or the problem being addressed. The clustering results evaluation metrics used to obtain the model with the best clustering results were the Davies Bouldin Index (DBI) metrics. The Davies-Bouldin Index (DBI) was an evaluation metric used to measure the quality of clustering results in data analysis. Lower DBI values indicated that clusters tend to be close to each other and homogeneous. Higher DBI values indicated that clusters tended to be spread further apart and were heterogeneous. The following is the formula for the Davies Bouldin Index.

$$\text{DBI} = \left(\frac{1}{\text{n}}\right) x \sum max \frac{(R_i + R_j)}{(d(C_i, C_j))} \quad (1)$$

Equation (1) refers to; n as the number of clusters, R_i as the average distance between objects in cluster i to the center of cluster i, and $d(C_i, C_j)$ as the distance between the center of cluster i and the center of cluster j.

IV. RESULTS

A. Modeling

In Fig. 3(a), each color represents a specific cluster, and the yellow color with the '*' marker represents outliers. It indicates that certain data points did not belong to any cluster. In Fig. 3(a), the result of clusters formed using the DBSCAN algorithm, namely 5 clusters are formed consisting of Cluster 0 with 6217 data points, cluster 1 with 8 data points, cluster 2 with 8 data points, and cluster 3 also has 8 data points. In cluster 4, there are 8 data points and 8 outlier data points.

In Fig. 3(b), each color represents a specific cluster, and the yellow color with the 'X' marker represents the center of each cluster. The data points in cluster 0 amount to 4279, and there are 90 data points in cluster 1, 1578 data points in cluster 2, 305 data points in cluster 3, 5 data points in cluster 4.

Fig. 4 shows DBI Score and clusters analysis comparison results of DBSCAN and K-Means algorithm. It shows the comparison of the DBI score values and the number of clusters between after comparing the K-Means and DBSCAN algorithms are presented. It could be seen that the K-Means algorithm produced a higher DBI score compared to DBSCAN. It showed that K-Means outperformed DBSCAN in terms of clustering quality and produced a greater number of clusters. In contrast to the K-Means algorithm, DBSCAN effectively identified outliers because they remained unassigned to any cluster. However, in the DBSCAN algorithm, outliers could be separated from clusters even though the DBSCAN algorithm produced a worse DBI score than K-Means.

B. Evaluation

The DBSCAN clustering method was tested using one year of data divided into several clusters. In the DBSCAN method, the number of clusters was not predetermined by the user. Still, the user determined the number of epsilon (distance) values and minimum points (minimum neighbors so that data could be in a cluster). Then, the calculation was carried out using the DBSCAN algorithm by looking for the lowest DBI (Davies-Bouldin Index) value and saving the cluster results from the conducted experiment. The determination of the number of epsilons, the minimum number of points, the resulting DBI value, and the number of clusters formed in the experiments that have been carried out can be seen in Table 3.

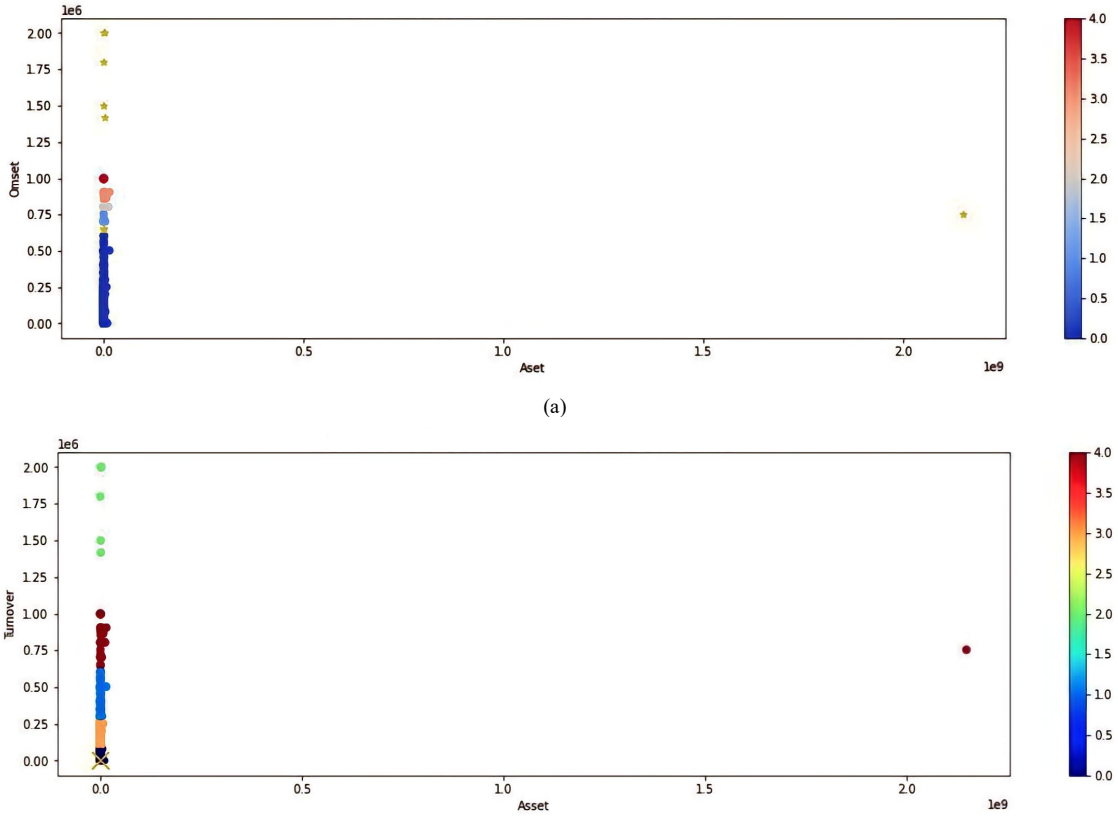

Fig. 3 Plot view of clustering results using (a) DBSCAN algorithm and (b) K-Means

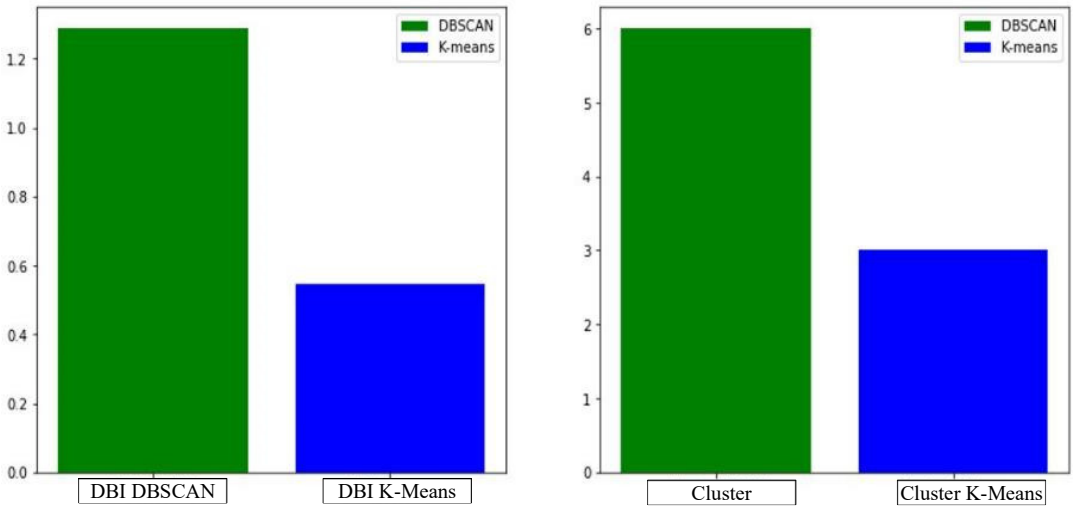

Fig. 4 DBI *Score* and clusters analysis comparison results of DBSCAN and K-Means algorithm

Table 3 shows a DBI score of 1.28 with five clusters using an epsilon value of 0.02 and a minimum point value of 3. In the data analyzed, there were five different clusters. Cluster 0 mainly consisted of the MSME sector with trade and agricultural businesses, including the micro MSME category, which usually employed an average of two people. The average asset value of this cluster was 110,396, with an average turnover of 52,406. Moving to Cluster 1 was

dominated by the MSME sector with a focus on trading businesses, including in the medium MSME category, which employs an average of 3 workers. This cluster's average asset value and turnover were respectively 744,125 and 708.75. Likewise, Cluster 2 was also dominated by the MSME sector, which operated in the trading business sector, especially classified as medium MSMEs, had an average workforce of three people, with an average asset value of 2,049,750, and an average turnover of 800,000. In Cluster 3, dominated by the MSME sector with trading businesses included in the medium MSME category, the average number of workers was seven people, the average asset value was 3,247,250, and the average turnover was 862,500. Finally, Cluster 4 was dominated by the MSME sector, which operated in the trade and industrial sectors included in the micro-medium MSME category. This cluster tended to have a larger average workforce of 12 people and showed an average asset value of 440,252 and an average turnover of 1,000,000.

TABLE 3
DBSCAN CLUSTERING ANALYSIS

eps	Min eps	cluster	DBI Score
0.02	3	5	1.28
0.02	4	5	1.28
0.02	5	5	1.28
0.35	3	2	10
0.35	4	2	10
0.23	5	2	10
0.05	3	2	10
0.05	4	2	10
0.05	5	2	10
0.65	3	2	10
0.65	4	2	10
0.65	5	2	10
0.08	3	1	10
0.08	4	1	10
0.08	5	1	10

TABLE 4
K-MEANS CLUSTERING ANALYSIS RESULTS

Cluster	N_init (Iteration)	Inertia Value	DBI Score
3	5	5	0.54
4	5	4	0.47
5	5	3	0.51
6	5	2	0.41
7	5	1	0.40
8	5	0.8	0.39

Table 4 shows the number of clusters tested is in the range 3 to 8, with random initialization of cluster centers (n_init) carried out 5 times using the Python programming language. Fig. 5 shows the most optimal number of clusters was based on observations from the elbow method. It showed a decrease in value, starting to stabilize, namely 5 clusters with a DBI value of 0.51. The following is the cluster center formed using the K-Means algorithm ([8.33725914e-03,3.79890721e-05],[3.17163252e-01,1.16088200e-02],[4.79540310e-02,5.54438269e-05],[1.23229649 e-01, 1.60372622e-04],[8.71 817359e-01, 5.64562160e-04]). The clustering results obtained with the k-Means algorithm are as follows below:

There were five main clusters in the analyzed data. Cluster 0 was dominated by the MSME sector with trade and agricultural businesses, in the micro and small MSME category with an average workforce of two people. It had an average asset value of 83,894 and an average turnover of 24,713. Cluster 1 was dominated by the MSME sector with trade and industrial businesses, categorized as micro and small MSMEs, with an average workforce of five people. It had an average asset value of 605,489 and an average turnover of 402,204. Meanwhile, Cluster 2 was dominated by the MSME sector with trade and agricultural businesses, included in the small MSME category, with an average workforce of three people. It had an average asset value of 1,212,400 and an average turnover of 1,743,636. Meanwhile, Cluster 3 was dominated by the MSME sector with trade and agricultural businesses, included in the small MSME category, with an average workforce of three people. It had an average asset value of 174,283 and an average turnover of 141,121. Meanwhile, Cluster 4 was dominated by the MSME sector with trade and agricultural businesses, included in the small MSME category, with an average workforce of six people. It had an average asset value of 6,284,088 and an average turnover of 834,285.

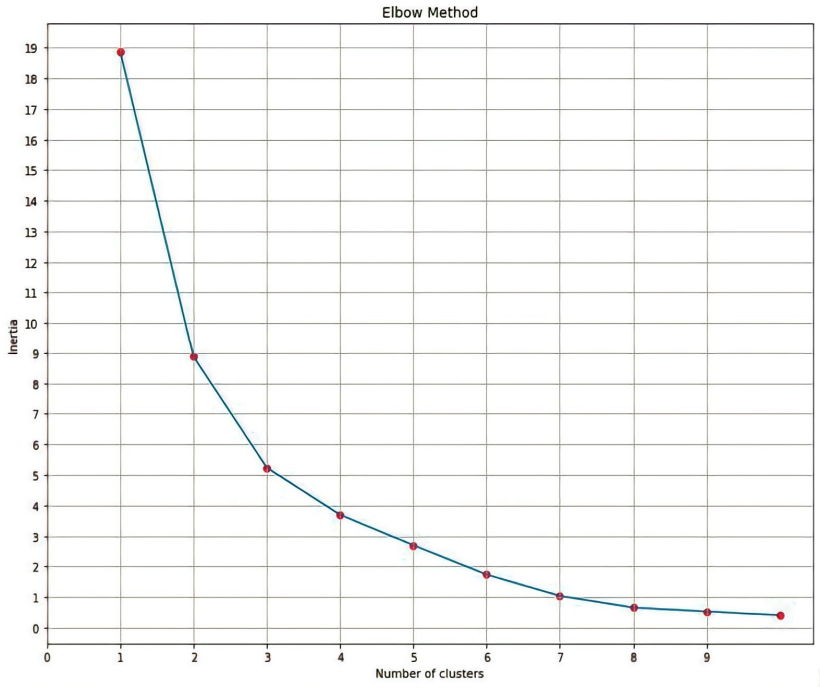

Fig. 5 Elbow method

V. DISCUSSION

In this section, we investigated into a comprehensive analysis of the comparative performance of the DBSCAN and K-Means clustering algorithms based on our experimental results. Our focus extended beyond the mere assessment of DBI scores and the number of resulting clusters. this study aimed to provide a nuanced understanding of the strengths and limitations inherent in each clustering approach.

Our experimental findings revealed that the K-Means algorithm outperformed DBSCAN regarding DBI scores. The lower DBI score obtained by K-Means signified its superior ability to form internally cohesive clusters that were well-separated from each other. This observation underscored K-Means' effectiveness in capturing the underlying structure of the dataset and generating clusters that closely align with the inherent patterns. Table 5 shows the comparison of dbi scores from both methods.

TABLE 5
COMPARISON OF DBI SCORES FROM BOTH METHODS

Algorithm	DBI Score	Clusters
K-Means	0.39	8
DBSCAN	1.39	5

While K-Means exhibited commendable performance in optimizing DBI scores, it was essential to recognize that the DBSCAN algorithm possessed a notable advantage in handling outliers. Unlike K-Means, which assign every data point to a cluster without considering its distance from the centroid, DBSCAN could effectively identify and segregate outliers from clusters. This capability was particularly valuable in real-world scenarios where outliers were prevalent and demanded separate treatment. The isolation of outliers by DBSCAN contributed to an overall improvement in clustering quality and higher DBI scores.

Another notable advantage of DBSCAN was its ability to automatically determine the number of clusters, eliminating the need for user-defined specifications as required in K-Means. DBSCAN assessed the data point density, ensuring the correct number of clusters without relying on prior assumptions about the dataset's structure. This inherent flexibility enabled DBSCAN to adapt to diverse datasets with varying characteristics, establishing it as a versatile algorithm for data clustering tasks.

Comparing our findings with previous studies [22], [23], we observed a trend of consistently superior performance of K-Means in terms of DBI scores. These studies reinforced the reliability and effectiveness of K-Means in scenarios involving relatively less dense or complex data. It was worth noting that K-Means' widespread used and well-

documented performance in various clustering scenarios add credence to its suitability for a broad range of applications. Despite its impressive performance, K-Means did have its limitations. One critical aspect was the selection of appropriate values for cluster parameters and the number of iterations (n_inits), which could significantly influence clustering results. Optimal parameter choices demanded careful consideration and domain expertise to ensure the best possible outcomes.

Additionally, we acknowledged potential threats to the validity of our study. While our comparative analysis offered valuable insights, variations in dataset characteristics, size, and distribution could impact the generalizability of our findings. Future study endeavors should consider these factors to provide a more comprehensive understanding of clustering algorithm performance.

In summary, our comparative analysis of DBSCAN and K-Means clustering algorithms highlighted their respective strengths and weaknesses. K-Means excelled in optimizing DBI scores, while DBSCAN offered advantages in handling outliers and automatically determining the number of clusters. Researchers and practitioners should carefully consider these characteristics when selecting an appropriate clustering algorithm for specific applications, bearing in mind the potential limitations and the need for parameter tuning in K-Means.

VI. Conclusions

DBSCAN and K-Means produced the same number of clusters, namely five clusters, but the DBSCAN algorithm produced four clusters with 1 outlier. Thus, the DBSCAN algorithm could identify more complex data variations and effectively separate points that were not included in the main cluster (outliers). On the other hand, K-Means tended to produce more centralized and homogeneous clusters. DBSCAN provided a higher DBI score than K-Means. DBI scores were used to measure the quality of a grouping, and lower scores indicated better separation between groups. In this case, K-Means showed better separation in classifying data into relevant groups based on comparison of DBI scores. The clusters produced by both methods were dominated by the MSME sector with a focus on the trade sector. However, there were differences in the MSME sector categories and other characteristics between the clusters produced by DBSCAN and K-Means. These differences showed that these two methods had different approaches to grouping data. However, the selection of clustering method must also be considered based on the characteristics of the data and specific study objectives. In conclusion, this study provided empirical evidence that supports the effectiveness of the DBSCAN algorithm for data clustering tasks. Even though it had a worse DBI score than K-Means, DBSCAN could handle outliers, making DBSCAN the right choice, primarily when the number of clusters was previously unknown, and the data exhibited complex data structures. The results of this research can be used as a reference in advanced analysis to categorize MSMEs and provide solutions to the financing challenges faced by MSMEs.

Future study can focus on further exploring DBSCAN parameter optimization techniques to improve its performance on different data sets. Additionally, investigating the implementation of other clustering algorithms and conducting comparative studies with DBSCAN and K-Means can provide valuable insights into their performance in various scenarios.

Author Contributions: *Ni Putu Sutramiani*: Conceptualization, Review, Editing. *I Made Teguh Arthana*: Data Curation, Conceptualization, Methodology, Investigation, Writing - Original Draft, Supervision. *Pramayota Fane'a Lampung*: Resources, Investigation, Writing - Original Draft, Visualization. *Shana Aurelia*: Project Administration, Methodology, Investigation, Writing - Original Draft, Writing - Review & Editing. *Muhammad Fauzi*: Data Curation, Investigation, Writing - Original Draft, Writing - Review & Editing. *I Wayan Agus Surya Darma*: Review, Editing.

Funding: No specific grant was received from any funding agency for this study.

Acknowledgments: The authors express their gratitude to the Udayana University's Faculty of Engineering and the Department of Information Technology for granting research facilities.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The data cannot be openly shared for the protection of study participant privacy.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Ni Putu Sutramiani: https://orcid.org/0000-0002-4302-0445
I Made Teguh Arthana: https://orcid.org/0009-0002-5097-2927
Pramayota Fane'a Lampung: https://orcid.org/0000-0003-0143-193X
Shana Aurelia: https://orcid.org/0009-0007-0873-9108
Muhammad Fauzi: https://orcid.org/0009-0006-2818-0420
I Wayan Agus Surya Darma: https://orcid.org/0000-0002-3507-4654

REFERENCES

[1] H. Gunawan, B. L. Sinaga, and W. P. Sigit Purnomo, "Assessment of the readiness of micro, small and medium enterprises in using E-money using the unified theory of acceptance and use of technology (UTAUT) method," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 316–323.
[2] Subianto and D. Wake, "Indonesia's Fintech Lending Potential," Indonesia's Fintech Lending.
[3] Organisation for Economic Cooperation and Development (OECD), "Key facts on SME financing: Indonesia," OECD iLibrary.
[4] T. Boonchoo, X. Ao, Y. Liu, W. Zhao, F. Zhuang, and Q. He, "Grid-based DBSCAN: Indexing and inference," *Pattern Recognit*, vol. 90, pp. 271–284, Jun. 2019.
[5] W. Wahyuri, U. Athiyah, I. Puspitasari, and Y. Nita, "Clustering of Drug Sampling Data to Determine Drug Distribution Patterns with K-Means Method : Study on Central Kalimantan Province, Indonesia," *Journal of Information Systems Engineering and Business Intelligence*, vol. 5, no. 2, p. 208, Oct. 2019.
[6] H. Cui, W. Wu, Z. Zhang, F. Han, and Z. Liu, "Clustering and application of grain temperature statistical parameters based on the DBSCAN algorithm," *J Stored Prod Res*, vol. 93, Sep. 2021.
[7] G. Armano and M. R. Farmani, "Multiobjective clustering analysis using particle swarm optimization," *Expert Syst Appl*, vol. 55, pp. 184–193, Aug. 2016.
[8] S. Monalisa and F. Kurnia, "Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 17, no. 1, pp. 110–117, Feb. 2019.
[9] S. F. Galán, "Comparative evaluation of region query strategies for DBSCAN clustering," *Inf Sci (N Y)*, vol. 502, pp. 76–90, Oct. 2019.
[10] D. Deng, "DBSCAN Clustering Algorithm Based on Density," *Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, pp. 949–953, Sep. 2020.
[11] R. Dhivya and N. Shanmugapriya, "An Efficient DBSCAN with Enhanced Agglomerative Clustering Algorithm," *2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings*, pp. 1322–1327, 2023.
[12] L. Ma, "An improved and heuristic-based iterative DBSCAN clustering algorithm," *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 2709–2714, 2021.
[13] R. E. Caraka *et al.*, "Micro, small, and medium enterprises' business vulnerability cluster in Indonesia: An analysis using optimized fuzzy geodemographic clustering," *Sustainability (Switzerland)*, vol. 13, no. 14, Jul. 2021.
[14] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual Quant*, vol. 56, no. 3, pp. 1283–1291, Jun. 2021.
[15] T. A. Cinderatama *et al.*, "Implementasi Metode K-Means, Dbscan, dan Meanshift Untuk Analisis Jenis Ancaman Jaringan Pada Intrusion Detection System," *Jurnal Inovtek Polbeng - Seri Informatika*, vol. 7, no. 1, 2022.
[16] E. Kusmiati, D. Turipanam Alamanda, and F. Fahru Roji, "MSME Clusterization Using K-Means Clustering in Garut Regency, Indonesia," *Review of Integrative Business and Economics Research*, vol. 12, p. 199, 2023.
[17] K. Wang, R. Yang, C. Liu, T. Samarasinghalage, and Y. Zang, "Extracting Electricity Patterns from High-dimensional Data: A comparison of K-Means and DBSCAN algorithms," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics.
[18] D. Dwi Aulia and N. Nurahman, "Comparison Performance of K-Medoids and K-Means Algorithms In Clustering Community Education Levels," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 12, no. 2, pp. 273–282, Jul. 2023.
[19] F. Andriyani and Y. Puspitarani, "Performance Comparison of K-Means and DBScan Algorithms for Text Clustering Product Reviews," *SinkrOn*, vol. 7, no. 3, pp. 944–949, Jul. 2022.
[20] IBM, "Data Mining Process," IBM. Accessed: Oct. 09, 2023.
[21] S. Dang and P. H. Ahmad, "Performance Evaluation of Clustering Algorithm Using Different Datasets Computer Science and Management Studies Performance Evaluation of Clustering Algorithm Using Different Datasets," 2015.
[22] K. Nurmayanti, W. P. Aini, S. R. Amrullah, and L. M. Sya'roni, "Comparison of Algorithms K-Means and DBSCAN for Clustering Student Cognitive Learning Outcomes in Physics Subject," *Kappa Journal*, vol. 7, no. 1, pp. 251–255, 2023.
[23] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer Segmentation using K-means Clustering," *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems*, 2018.