Implementation of the K-Means Clustering Algorithm in City and Regency Clustering in North Sumatra Province Based on Small and Micro Industries

Agnes Irene Silitonga [1], Ali Akbar Lubis [2], Jufri Darma [3],
Ferry Indra Sakti H. Sinaga [4], Yoakim Simamora [5]

agnesirenesilitonga@unimed.ac.id [1], aliakbarlubis@unimed.ac.id [2], jufridarma@unimed.ac.id [3],
ferryindrasakti@unimed.ac.id [4], yoakimsimamora@unimed.ac.id [5]

Universitas Negeri Medan, North Sumatera, Indonesia [1,2,3,4,5]

Abstract. Small and micro industries have a strategic role in the regional economy, especially in improving people's welfare and driving economic growth. This study aims to cluster regencies and cities in North Sumatra Province using the K-Means Clustering algorithm. The data used include the number of business units, workforce, and bank capital loans in each region. The K-Means Clustering method was chosen because of its ability to cluster data based on similar characteristics so that it can provide an overview of the classification of regions with similar small and micro industry potential. The results of the study show that regencies and cities in North Sumatra Province can be clustered into three clusters, namely low, medium, and high clusters. The results of this clustering are expected to be the basis for local governments in designing more effective policies for the development of small and micro industries in each region, such as the allocation of capital assistance, training, and strengthening the industrial supply chain.

Keywords: K-Means Clustering Algorithm, Small and Micro Industries, Clustering, Machine Learning.

1 Introduction

The increasingly dynamic transformation of the Indonesian economy demands a deep understanding of the productivity of small and micro industries, especially in North Sumatra Province, which has significant economic potential. This is because small and micro industries play an important role in the Indonesian economy. This sector is not only a major contributor to economic growth, but can also absorb labor and reduce poverty rates at the local level. More than 64 million micro, small, and medium enterprises (MSMEs) contribute around 60% of the national Gross Domestic Product (GDP) and absorb more than 97% of the workforce [1]. In North Sumatra Province, small and micro industries play a significant role in supporting regional economic resilience, especially amidst the global situation full of uncertainty.

Despite having great potential, small and micro industries are often not optimized to their full potential. Small and micro industries face a number of challenges that can hinder their growth [2]. These challenges include access to financial resources, limitations in product marketing, and low adoption of modern technology. In addition, the lack of information on the segmentation or specific characteristics of small and micro industries in each region is one of the obstacles in formulating effective policies.

ICIESC 2025, September 16, Medan, Indonesia

DOI 10.4108/eai.16-9-2025.2361141

North Sumatra is one of the provinces with a high level of economic diversity. The small and micro industry sector in this province covers various fields, such as food and beverages, handicrafts, textiles, and creative services [3]. Geographically, North Sumatra has heterogeneous characteristics, ranging from urban areas such as Medan to rural areas in Tapanuli or Nias. This diversity creates additional complexity in the management of small and micro industries.

Areas with high concentrations of small and micro industries in North Sumatra often show patterns of inequality in terms of infrastructure support, market access, and quality of human resources [4]. For example, small and micro industries in urban areas tend to have better access to technology and financing than small and micro industries in rural areas. This phenomenon requires a data-based approach to identify the needs and opportunities of each group of small and micro industries.

Advances in information technology and data analysis open up new opportunities in the management of small and micro industries, one of which is the use of machine learning. One use of machine learning that has been widely used is data-based decision making. In the context of clustering small and micro industries in North Sumatra Province, machine learning is used to find hidden patterns or structures in industrial data using the unsupervised learning approach, which is a learning method without data labels.

One of the machine learning algorithms in the field of clustering is the K-Means Clustering algorithm, which is able to group data into homogeneous groups based on certain characteristics. K-Means Clustering can be used to group small and micro industry data based on various factors, such as geographic location, product type, and business scale (reference). Similar studies have been conducted in various countries to support the development of small and micro industries. For example, research conducted in India used K-Means Clustering to group MSMEs based on industrial sector and business scale [5]. The results of this study help the Indian government in designing specific fiscal policies for each MSME group. In Indonesia, research by [4] shows that clustering can increase the efficiency of small and micro industry assistance programs by up to 40%. With this algorithm, regencies / cities in North Sumatra Province can be clustered based on small and micro industry productivity into three clusters, namely high, medium, and low productivity. Through the application of K-Means Clustering in North Sumatra Province, it is hoped that information will be obtained on regencies / cities that have low small and micro industry productivity. For example, regencies / cities that have a large workforce tend to have a large number of small and micro industries as well.

In addition, through the application of K-Means Clustering, it can be seen which regencies/cities require access to technology based on the productivity clusters of small and medium industries. Access to technology is often a differentiating factor between rapidly growing and stagnant small and micro industries [2]. By clustering small and micro industries, development policies can be focused on improving technology in certain clusters. Clustering can also reveal differences in needs between small and micro industry groups. For example, small and micro industries in small-scale clusters may require basic financial management training, while more advanced clusters can focus on digital marketing.

In North Sumatra, the diversity of business types and geographical conditions adds to the complexity of managing small and micro industries in each regency / city. Therefore, this study aims to cluster regencies / cities in North Sumatra Province based on the number of small and micro industries using the K-Means Clustering algorithm so that it is expected to provide a significant contribution by adapting a similar algorithm. By using local data and

relevant variables, the clustering results can help formulate more focused policies that have a long-term impact on the development of small and micro industries.

2 Literature Review

2.1 Machine Learning

Machine learning is a subset of artificial intelligence that examines and constructs models capable of learning from data or environmental stimuli to anticipate unknown data or make decisions based on environmental information [6]. Machine learning is defined as a computer science that provides the ability for computer systems to learn automatically from previous experiences (historical data), without having to be explicitly programmed [7]. This system uses algorithms to find patterns or trends in data so that it can make predictions or decisions. Machine learning has been applied in various fields such as information technology, health, business, and automotive.

2.2 Maintaining the Integrity of the Specifications

In machine learning, there are various techniques used to extract patterns from data. One important technique in the unsupervised learning group is the K-Means Clustering algorithm, which is an unsupervised learning algorithm that aims to group data into several clusters based on similarities between data. This algorithm groups data by minimizing the distance between data points and the cluster center point (centroid) [8]. The result of this process is the division of the dataset into several groups that have similar patterns or characteristics.

In real-world implementations, K-Means Clustering is widely used because of its speed and ability to handle large datasets. K-Means Clustering has been applied in fields such as nutrition / stunting [9], base transceiver stations [10], distribution of MSMEs [11], and others.

K-Means Clustering implementation often uses the Python programming language with the help of libraries such as scikit-learn.

```
from sklearn.cluster import KMeans
import numpy as np
  X = np.array([[1,2], [1,4], [1,0], [10,2], [10,4], [10,0]])
  kmeans = KMeans(n_clusters=2)
  kmeans.fit(X)
      print(kmeans.cluster_centers_)
```

Machine learning and K-Means Clustering are closely related, especially in the unsupervised learning category. Machine learning provides a framework for this clustering algorithm, namely the ability to learn from data without any specific labels. As an unsupervised method, K-Means Clustering helps in exploring and identifying hidden patterns in complex datasets [7].

3 Research Methods

This study uses the K-Means Clustering algorithm because of its ability to cluster data into several homogeneous clusters based on similar characteristics (references). The following are the steps of the K-Means Clustering algorithm:

1. Data preparation involves the integration of all variables (x1, x2, …, xn) into a structured database to ensure consistency in subsequent analyses. Prior to clustering, data cleaning is conducted to remove invalid or incomplete values that can potentially bias the results. In addition, normalization or standardization procedures are applied, where necessary, to ensure that all variables operated on a comparable scale, thereby preventing dominance of variables with larger numerical ranges.
2. The next step is the determination of the number of clusters (K). In this study, the number of clusters is predefined as three, representing the categories of low, medium, and high. The selection of k=3 is further justified through cluster validation approaches, such as the Elbow method and the Silhouette coefficient, which provide statistical support for the appropriateness of the chosen number of clusters.
3. Initialization of the centroids is conducted by randomly selecting three data points from the dataset to serve as the starting centroids. This is followed by the initial grouping process, in which the Euclidean distance between each observation (regency/city) and each centroid is calculated. Each observation is then assigned to the cluster corresponding to the nearest centroid.
4. Subsequently, the recalculation of centroids is performed. Once all data points have been assigned to their respective clusters, a new centroid is computed as the mean of all data points within each cluster. This step ensures that the centroids represented the updated positions of the clusters.
5. The process then entered the iteration and convergence stage. The steps of distance calculation, cluster reassignment, and centroid recalculation were repeated iteratively. The algorithm was considered to have converged when there was no further reassignment of data points between clusters or when the maximum number of iterations had been reached. At this stage, the final, stable cluster configuration was obtained.
6. Finally, the cluster interpretation phase is conducted to analyze the characteristics of each resulting cluster. For example, clusters that exhibited higher values across the variables (x_1, x_2, …, x_n) are classified as high clusters, those with moderate values are categorized as medium clusters, and those with lower values are classified as low clusters. This interpretation provided meaningful insights into the distribution and differentiation of the observed data.

The object and scope of the research are regencies/cities in North Sumatra Province. There are nine variables used in this study with data sourced from the Central Statistics Agency of North Sumatra Province. Data analysis techniques used in this study include:

1. **Descriptive Analysis**. Presenting the initial profile of each regency and city, such as the number of small and micro industries, the number of workers, and capital loans on average.
2. **Cluster Analysis.** Using the K-Means Clustering algorithm, then validating the results by checking the average intra-cluster and inter-cluster distances, and can use the Silhouette Score to see how well the clusters are separated.
3. **Drawing Conclusions.** Explaining policy implications based on clusters, such as regencies and cities that are included in high clusters receive different policy support than low clusters.

The following is the clustering process of regencies and cities in North Sumatra Province based on small and micro industries with K-Means Clustering:

1. Before clustering, the data is normalized to a scale of 0–1 (if necessary). This ensures that each variable (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9) has a balanced weight in calculating distance. The variables used in this study are the number of small and micro businesses in each regency/city, the number of workers, the number of industries that receive bank capital loans, the number of industries establishing partnerships, the number of workers who completed at least elementary school education, the number of industries that have business permits, the number of industries that have received guidance / training / counseling, the number of industries that did innovations, and the number of industries that have certificates / patents / copyrights / intellectual property rights.
2. Determination of the number of clusters as many as 3 clusters (k=3). The reason for choosing three clusters is to facilitate interpretation: (1) low cluster, (2) medium cluster, and (3) high cluster. Empirically, the Elbow method can also be used to see the bending point on the SSE (Sum of Squared Errors) graph which indicates that adding clusters no longer provides significant error reduction.
3. Determination of three initial centroids, which are selected randomly from the dataset. Following this initialization, the distance of each regency or city to the centroids is calculated using the selected distance metric, after which each observation is assigned to the cluster with the nearest centroid. Subsequently, the centroid of each cluster is recalculated by computing the mean value of all data points contained within the cluster. This iterative procedure continues, with recalculations and reassignments performed in successive steps, until a convergence criterion is met. The algorithm is considered to have converged when there are no significant changes in the composition of cluster members or when the centroid positions become stable, thereby producing the final cluster structure.

4 Result and Discussion

As an initial step, the data used in this study is data from 33 districts/cities in North Sumatra Province as in Table 1.

Table 1. Small And Micro Business Data By Regency/City In North Sumatra Province

No	City / Regency	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	Nias	955	1741	0	0	1671	189	0	0	0
2	Mandailing Natal	3342	6435	376	221	5924	168	51	974	97
3	Tapanuli Selatan	4635	7650	282	17	7565	278	10	101	16
4	Tapanuli Tengah	5250	8498	1434	0	7569	305	15	521	30
5	Tapanuli Utara	6681	7974	132	0	7793	261	111	406	0
6	Toba	3258	6055	560	780	5955	255	389	272	68
7	Labuhan Batu	4614	9748	1363	365	9159	463	36	85	12
8	Asahan	8258	23198	1417	0	22554	348	318	932	575
9	Simalungun	9051	19415	855	1043	18674	919	722	1125	345
10	Dairi	1859	2872	84	57	2768	279	55	337	101
11	Karo	1875	3172	463	51	3009	169	131	24	48
12	Deli Serdang	8556	23282	1290	342	22915	1196	325	1780	525
13	Langkat	10001	23615	1754	1537	21812	546	254	381	322
14	Nias Selatan	1826	3450	25	0	2890	1	0	61	0
15	Humbang Hasundutan	2066	2741	202	0	2595	92	0	170	0

16	Pakpak Bharat	266	360	28	6	264	29	0	14	0
17	Samosir	6151	7951	1019	0	7390	216	161	1201	11
18	Serdang Bedagai	6002	14699	1330	134	13645	557	4	48	103
19	Batu Bara	4706	11276	1197	13	10004	194	67	77	25
20	Padang Lawas Utara	1169	2414	14	58	2331	38	0	161	2
21	Padang Lawas	4017	8685	361	1	8038	166	0	157	31
22	Labuhan Batu Selatan	908	1794	135	1	1650	137	0	8	0
23	Labuhan Batu Utara	2034	3585	158	6	3442	98	208	16	28
24	Nias Utara	586	1253	0	0	0	773	0	0	0
25	Nias Barat	336	389	0	0	365	3	0	0	0
26	Sibolga	650	1298	159	0	1257	14	0	2	0
27	Tanjung Balai	2592	4064	74	55	3893	281	134	27	80
28	Pematang Siantar	3383	5582	290	42	5319	453	39	157	90
29	Tebing Tinggi	2718	6611	408	112	6369	637	103	96	474
30	Medan	12300	30946	1418	438	30491	2223	230	661	348
31	Binjai	2785	6559	161	151	6368	295	144	155	152
32	Padangsidimpuan	1802	3932	337	47	3839	122	50	85	14
33	Gunungsitoli	2275	3915	108	5	3473	296	104	112	36

The first step is determining the number of clusters as many as 3 clusters, namely low (K1), medium (K2), high (K3). The determination of this cluster is based on the productivity of small and micro industries in districts / cities in North Sumatra Province. Furthermore, the center point of each cluster is determined, namely Pakpak Bharat Regency as the center point of the low cluster, Toba Regency as the center point of the medium cluster, and Medan City as the center point of the high cluster. After determining the center point of each cluster, the distance of each data to the center point of each cluster is calculated using the Euclidean Distance formula as in Equation 1.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

where:

$(x_i - y_i)$ = the difference in the value of the i-th attribute between two objects
$\sum_{i=1}^{n}(\dots)$ = all attribute (component) differences are added together after being squared
$\sqrt{\dots}$ = square root of the sum of the squares of the attribute differences

The calculation of the distance from each data to the cluster center point is done repeatedly. The Euclidean Distance calculation process will stop if the cluster center value is stagnant or does not change. The Euclidean Distance calculation used in this study is as in Equation 2.

$$d(x,y) = \sqrt{\begin{matrix}(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + \\ (x_4 - y_4)^2 + (x_5 - y_5)^2 + (x_6 - y_6)^2 + \\ (x_7 - y_7)^2 + (x_8 - y_8)^2 + (x_9 - y_9)^2\end{matrix}} \qquad (2)$$

The calculation results using Euclidean Distance in the first iteration can be seen in Table 2.

Table 2. First Iteration Calculation Results

No	City / Regency	Distance to Centroid		
		K1 (Low)	K2 (Medium)	K3 (High)
1	Nias	2094.79	6982.24	42651.89
2	Mandailing Natal	8917.58	1128.03	35918.26
3	Tapanuli Selatan	11210.33	2473.43	33656.80

4	Tapanuli Tengah	12113.69	3617.63	32910.04
5	Tapanuli Utara	12491.67	4395.87	32869.14
6	Toba	8656.11	1133.81	36170.65
7	Labuhan Batu	13721.03	4781.17	31097.79
8	Asahan	32948.38	23900.06	11966.85
9	Simalungun	27993.27	18946.14	16911.60
10	Dairi	3911.98	5241.19	40885.56
11	Karo	4273.37	4871.14	40510.77
12	Deli Serdang	33373.30	24324.92	11511.87
13	Langkat	33253.19	24209.68	11771.43
14	Nias Selatan	4345.18	4781.47	40437.03
15	Humbang Hasundutan	3794.91	5425.81	41049.61
16	Pakpak Bharat	0.00	9066.70	44738.32
17	Samosir	12061.86	4023.88	33241.92
18	Serdang Bedagai	20483.21	11446.94	24305.97
19	Batu Bara	15334.38	6347.56	29480.73
20	Padang Lawas Utara	3054.74	6019.09	41692.59
21	Padang Lawas	11998.41	2990.17	32775.67
22	Labuhan Batu Selatan	2100.64	6971.48	42640.46
23	Labuhan Batu Utara	4867.43	4250.74	39905.79
24	Nias Utara	1234.54	8595.68	44198.52
25	Nias Barat	132.79	8961.27	44633.90
26	Sibolga	1425.10	7650.34	43319.95
27	Tanjung Balai	5691.40	3525.13	39122.22
28	Pematang Siantar	7925.77	1575.09	36897.07
29	Tebing Tinggi	9117.33	543.63	35635.25
30	Medan	44738.32	35677.19	0.00
31	Binjai	9066.70	0.00	35677.19
32	Padangsidimpuan	5292.70	3790.48	39452.56
33	Gunungsitoli	5203.00	3958.90	39587.62

Next, the shortest distance of the data to the initial cluster center that was previously selected randomly is determined. Determining the shortest distance to the cluster center by choosing the smallest value. The results of determining the shortest distance to the cluster center that have been tabulated can be seen in Table 3.

Table 3. Binary Tabulation of Closest Data Distance To Clusters

No	City / Regency	Distance to Centroid		
		K1 (Low)	K2 (Medium)	K3 (High)
1	Nias	1	0	0
2	Mandailing Natal	0	1	0
3	Tapanuli Selatan	0	1	0
4	Tapanuli Tengah	0	1	0
5	Tapanuli Utara	0	1	0
6	Toba	0	1	0
7	Labuhan Batu	0	1	0
8	Asahan	0	0	1
9	Simalungun	0	0	1
10	Dairi	1	0	0
11	Karo	1	0	0
12	Deli Serdang	0	0	1

13	Langkat	0	0	1
14	Nias Selatan	1	0	0
15	Humbang Hasundutan	1	0	0
16	Pakpak Bharat	1	0	0
17	Samosir	0	1	0
18	Serdang Bedagai	0	1	0
19	Batu Bara	0	1	0
20	Padang Lawas Utara	1	0	0
21	Padang Lawas	0	1	0
22	Labuhan Batu Selatan	1	0	0
23	Labuhan Batu Utara	1	0	0
24	Nias Utara	1	0	0
25	Nias Barat	1	0	0
26	Sibolga	1	0	0
27	Tanjung Balai	0	1	0
28	Pematang Siantar	0	1	0
29	Tebing Tinggi	0	1	0
30	Medan	0	0	1
31	Binjai	0	1	0
32	Padangsidimpuan	1	0	0
33	Gunungsitoli	0	1	0

Based on Table 3, districts/cities can be grouped into low, medium, and high clusters as in Table 4.

Table 4. First Iteration Clustering Results

Cluster	Cluster Member	Number of members
K1 (Low)	1,10,11,14,15,16,20,22,23,24,25,26	12
K2 (Middle)	2,3,4,5,6,7,17,18,19,21,27,28,29,31,32,33	16
K3 (High)	8,9,12,13,30	5

The next step, the determination of the new cluster center point is carried out by calculating the average of each cluster for each variable. The results of the average calculation can be seen in Table 5.

Tabel 5. New Centroids

K_1	X_1	1210.83
	X_2	2089.08
	X_3	105.67
	X_4	14.92
	X_5	1853.50
	X_6	151.83
	X_7	32.83
	X_8	66.08
	X_9	14.92
K_2	X_1	4013.19
	X_2	7477.13

	X_3	589.50
	X_4	121.44
	X_5	7018.94
	X_6	309.19
	X_7	88.63
	X_8	279.63
	X_9	77.44
	X_1	8297.00
	X_2	20069.20
	X_3	1181.40
	X_4	613.80
K_3	X_5	19308.00
	X_6	841.00
	X_7	331.00
	X_8	624.60
	X_9	327.60

Based on Table 5, a new cluster center point is obtained which is used as a reference in the second iteration, with the following values:

1. K_1 = [1210.83; 2089.08; 105.67; 14.92; 1853.50; 151.83; 32.83; 66.08; 14.92]
2. K_2 = [4013.19; 7477.13; 589.50; 121.44; 7018.94; 309.19; 88.63; 279.63; 77.44]
3. K_3 = [8297.00; 20069.20; 1181.40; 613.80; 19308.00; 841.00; 331.00; 624.60; 327.60]

Next, the calculation is carried out for the second iteration until the results of the cluster center point do not change in value. When the cluster center point does not change in value, the clustering process is complete. In this study, the clustering process stops at the fourth iteration because the value of the cluster center point at the fourth iteration is the same as the value of the cluster center point at the third iteration. The calculation results at the fourth iteration can be seen in Table 6.

Table 6. Results of The Fourth Iteration Calculation

No	City / Regency	Distance to Centroid		
		K1 (Low)	K2 (Medium)	K3 (High)
1	Nias	1033.93	9631.74	32349.00
2	Mandailing Natal	5836.03	2946.41	25592.89
3	Tapanuli Selatan	8109.60	846.25	23361.49
4	Tapanuli Tengah	9031.89	1196.49	22604.89
5	Tapanuli Utara	9443.09	2356.11	22621.37
6	Toba	5577.52	3191.66	25851.52
7	Labuhan Batu	10632.50	2157.41	20786.80
8	Asahan	29876.62	21369.89	2049.75
9	Simalungun	24897.24	16321.23	6637.32
10	Dairi	864.71	7834.15	30577.52
11	Karo	1203.74	7452.66	30201.48
12	Deli Serdang	30298.94	21789.65	1657.96
13	Langkat	30164.09	21592.85	2021.39
14	Nias Selatan	1271.01	7393.01	30126.12
15	Humbang Hasundutan	813.25	7969.50	30740.55
16	Pakpak Bharat	3105.84	11713.19	34431.57

17	Samosir	9015.34	2071.69	22954.93
18	Serdang Bedagai	17397.28	8863.38	14001.28
19	Batu Bara	12250.36	3783.49	19163.62
20	Padang Lawas Utara	336.35	8678.98	31383.72
21	Padang Lawas	8903.25	756.26	22464.41
22	Labuhan Batu Selatan	1024.58	9618.24	32334.92
23	Labuhan Batu Utara	1773.17	6859.10	29597.69
24	Nias Utara	2741.12	11201.46	33903.25
25	Nias Barat	2999.67	11606.78	34327.44
26	Sibolga	1710.00	10299.92	33012.69
27	Tanjung Balai	2596.25	6059.47	28819.09
28	Pematang Siantar	4822.90	3826.03	26595.87
29	Tebing Tinggi	6045.12	2833.39	25335.37
30	Medan	41652.65	33103.53	10369.88
31	Binjai	5986.18	2820.07	25370.12
32	Padangsidimpuan	2223.17	6447.42	29143.77
33	Gunungsitoli	2100.47	6530.07	29281.62

Next, the shortest distance to the cluster center is determined by selecting the smallest value. The tabulated results can be seen in Table 7.

Table 7. Binary Tabulation of Closest Data Distance To Clusters

No	City / Regency	Distance to Centroid		
		K1 (Low)	K2 (Medium)	K3 (High)
1	Nias	1	0	0
2	Mandailing Natal	0	1	0
3	Tapanuli Selatan	0	1	0
4	Tapanuli Tengah	0	1	0
5	Tapanuli Utara	0	1	0
6	Toba	0	1	0
7	Labuhan Batu	0	1	0
8	Asahan	0	0	1
9	Simalungun	0	0	1
10	Dairi	1	0	0
11	Karo	1	0	0
12	Deli Serdang	0	0	1
13	Langkat	0	0	1
14	Nias Selatan	1	0	0
15	Humbang Hasundutan	1	0	0
16	Pakpak Bharat	1	0	0
17	Samosir	0	1	0
18	Serdang Bedagai	0	0	1
19	Batu Bara	0	1	0
20	Padang Lawas Utara	1	0	0
21	Padang Lawas	0	1	0
22	Labuhan Batu Selatan	1	0	0
23	Labuhan Batu Utara	1	0	0
24	Nias Utara	1	0	0
25	Nias Barat	1	0	0
26	Sibolga	1	0	0
27	Tanjung Balai	1	0	0

28	Pematang Siantar	0	1	0
29	Tebing Tinggi	0	1	0
30	Medan	0	0	1
31	Binjai	0	1	0
32	Padangsidimpuan	1	0	0
33	Gunungsitoli	1	0	0

Based on Table 7, districts/cities can be grouped into low, medium, and high clusters as in Table 8.

Table 8. Last Iteration Clustering Results

Cluster	Cluster Member	Number of members
K1 (Low)	1,10,11,14,15,16,20,22,23,24,25,26	12
K2 (Middle)	2,3,4,5,6,7,17,18,19,21,27,28,29,31,32,33	16
K3 (High)	8,9,12,13,30	5

Based on the last iteration clustering, the following results were obtained:

1. Cluster 1 (Low) with a centroid value of K1 = [1413.27; 2465.33; 119.13; 19.07; 2229.80; 168.07; 45.47; 67.80; 20.60]: Nias, Dairi, Karo, South Nias, Humbang Hasuduntan, Pakpak Bharat, North Padang Lawas, South Labuhan Batu, North Labuhan Batu, North Nias, West Nias, Sibolga, Tanjung Balai, Padangsidimpuan, Gunungsitoli.
2. Cluster 2 (Medium) with a centroid value of K2 = [4426.31; 8286.38; 685.62; 141.23; 7776.77; 326.77; 86.92; 326.92; 85.31]: Mandailing Natal, South Tapanuli, Central Tapanuli, North Tapanuli, Toba, Labuhan Batu, Samosir, Serdang Bedagai, Batu Bara, Padang Lawas, Pematang Siantar, Tebing Tinggi, Binjai
3. Cluster 3 (High) with centroid point value K3 = [9633.20; 24091.20; 1346.8; 672; 23289.2; 1046.4; 369.8; 975.8; 423]: Asahan, Simalungun, Deli Serdang, Langkat, Medan.

The results of this clustering support the concept that small and micro industries in Indonesia have heterogeneous characteristics, mainly influenced by access to capital, number of workers, and infrastructure support.

The K-Means Clustering algorithm has proven effective in capturing patterns of similarities between regions, in line with previous studies that applied this algorithm for MSME segmentation in various regions [5].

From a public policy perspective, clustering can be a relevant reference for local governments in distributing assistance according to cluster needs and the effectiveness of MSME assistance programs will increase if preceded by a data-based clustering process. Thus, the results and discussion of this study emphasize the need for grouping small and micro industry data before formulating policies. The three clusters formed (low, medium, and high) can be an initial guideline for determining intervention priorities, both in terms of training, providing capital, and building supporting infrastructure. Through this approach, efforts to improve the welfare of small and micro industry actors in North Sumatra Province can be carried out in a targeted and sustainable manner.

5 Conclusion

Overall, clustering shows that data-based grouping is able to provide a picture of the heterogeneity of small and micro industries in North Sumatra Province. The use of diverse variables greatly influences the cluster structure that is formed. Clustering of small and micro businesses can help the government distribute aid programs more efficiently and on target.

From this study, there are several policy consequences obtained that can be taken by the government. In low-cluster areas**,** development strategies should prioritize the reinforcement of fundamental business capacities. One essential approach is to expand access to financing through low-interest credit schemes, enabling small entrepreneurs to overcome capital constraints. Equally important is the provision of basic training in entrepreneurship and financial management, which can enhance managerial skills and foster greater business sustainability. Furthermore, the establishment of marketing facilities and distribution networks is necessary to improve market penetration, thereby increasing sales opportunities for local products and enhancing their overall competitiveness.

In medium-cluster areas**,** the focus should be directed toward advancing technological capacity and stimulating innovation. Capacity building may be achieved by providing training in the use of modern machinery and digital tools, which facilitates efficiency and productivity improvements. Moreover, business actors should be encouraged to pursue product innovation and engage in collaborative initiatives with surrounding enterprises, thereby strengthening local business ecosystems. Market expansion is also a critical component in this stage of development, and can be promoted through participation in trade exhibitions as well as training in e-commerce, which provides access to broader consumer bases and global markets.

In high-cluster areas**,** the emphasis shifts toward fostering research-driven innovation and global competitiveness. Facilitating research and development is crucial to the creation of high value-added products that can compete in international markets. At the same time, strengthening global networks through export facilitation and venture capital participation is essential to ensuring financial sustainability and international market integration. Finally, strategies for product diversification and business line expansion should be adopted to reduce dependency on a single sector, thereby promoting resilience and long-term growth within small and micro-industrial sectors.

References

[1] Kementerian Perindustrian, "Statistik Industri Kecil dan Mikro 2023", Jakarta: Kementerian Perindustrian, 2023.

[2] M.U. Dewi, A. Mekaniwati, Y. Nurendah, P. Cakranegara, & A. S. Arief, "Globalization Challenges of Micro Small and Medium Enterprises", *Eur. J. Mol. Clin. Med*, vol. 7, no. 11, pp. 1909-1915, 2020.

[3] Badan Pusat Statistik, "Profil Industri Mikro dan Kecil Provinsi Sumatera Utara 2023", Medan: BPS, 2023.

[4] R.S. Nugraha, W.S. Jatiningrum, & R. D. Astuti, "Cluster analysis to determine business strategy for MSMEs in Yogyakarta". In *IOP Conference Series: Materials Science and Engineering*, vol. 1072, no. 1, p. 012011, IOP Publishing, 2021.

[5] I. H. Witten and E. Al, *Data Mining: Practical Machine Learning Tools and Techniques*, Amsterdam: Morgan Kaufmann, 2017.

[6] W. Wang, *Principles of Machine Learning*, Springer Nature, 2025.

[7] E. Alpaydin, *Introduction to Machine Learning*, Cambridge, Massachusetts: The Mit Press, 2020.
[8] J. Wu, *Advances in K-means Clustering a Data Mining Thinking*, Berlin Springer Berlin, 2014.
[9] A. I. Silitonga, Z. A. Nabila, C. R. Z. Lubis, N. Safitri, and Haryadi, "Klasterisasi Gizi Buruk Dan Stunting di Provinsi Sumatera Utara Menggunakan K-Means Clustering," *METHODIKA: Jurnal Teknik Informatika dan Sistem Informasi*, vol. 10, no. 2, pp. 13–18, Sep. 2024, doi: https://doi.org/10.46880/mtk.v10i2.3147.
[10] A. I. Silitonga, M. A. Nasution, A. Fitri, K. S. Rizwinie, A. Hidayatullah, and Y. Simamora, "Klasterisasi Penyebaran Base Transceiver Station Menggunakan K-Means Clustering," *Semnas Ristek (Seminar Nasional Riset dan Inovasi Teknologi)*, vol. 9, no. 1, pp. 344–351, Jan. 2025, doi: https://doi.org/10.30998/semnasristek.v9i1.794.
[11] Azzam, A. I. Purnamasari, and I. Ali, "Implementasi Algoritma K-Means Clustering Untuk Analisis Persebaran Umkm Di Jawa Barat," *JATI (Jurnal Mahasiswa Teknik Informatika),* vol. 8, no. 3, pp. 3062–3070, May 2024, doi: https://doi.org/10.36040/jati.v8i3.8450.