

Segmentation and Prediction of the Development of Small and Medium Industries (SMEs) in North Sumatra Using K-Means and Random Forest Algorithms Based on OSS RBA Data

Siska Mayasari Rambe, Zulham Sitorus

Abstract

Small and Medium Industries (SMEs) have a strategic role in increasing regional economic growth, creating jobs, and encouraging the competitiveness of the national industrial sector. Along with the implementation of the Online Single Submission Risk-Based Approach (OSS RBA) system, business licensing data is available that can be used to support data-based decision-making. However, the use of OSS RBA data in identifying the characteristics and development of SMEs is still limited. This study aims to segment the development of SMEs in North Sumatra Province using the K-Means algorithm and build a prediction model using the Random Forest algorithm. The research dataset comes from OSS RBA data for the 2022-2025 period which has gone through the stages of data cleaning, feature engineering, data aggregation, and normalization. The number of clusters was determined using the Elbow Method, Silhouette Score, and Davies-Bouldin Index, then the segmentation process was carried out using the K-Means algorithm. The results of the segmentation are then used as the classification target in the Random Forest model. The results showed that the optimal number of clusters used was five clusters (K=5). The Random Forest model is able to predict segmentation results with excellent performance based on the results of the model evaluation. The Feature Importance analysis shows that the variables of Year, Growth of SMEs, Number of SMEs, and several characteristics of the industrial sector are the most influential factors in the formation of clusters. The results of this research are expected to be the basis for the formulation of policies for fostering and developing SMEs based on the characteristics of the region and industrial sector in North Sumatra Province.

Keywords: *Small and Medium Industries, OSS RBA, K-Means, Random Forest, Segmentation, Prediction.*

Introduction

Small and Medium Industries (SMEs) are one of the strategic sectors that play an important role in increasing national economic growth through job creation, increasing product added value, and strengthening the national industrial structure [1]. This sector has the characteristics of being a labor-intensive industry that is able to absorb a large number of workers and is spread almost throughout Indonesia so as to contribute to the equitable distribution of economic development. In addition, SMEs also have a high level of flexibility in dealing with changes in economic conditions so that they become one of the main drivers of the national economy[2].

Siska Mayasari Rambe

Master of Information Technology, Universitas Pembangunan Panca Budi, Indonesia

e-mail: mayasarisiska5@gmail.com

Zulham Sitorus

e-mail: zulhamsitorus@dosen.pancabudi.ac.id

3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)

Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS

<https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF>

In North Sumatra Province, the SME sector has shown significant development. Based on data on the Number of SMEs, Workforce, and Value by Regency/City in North Sumatra Province in 2024, there are 70,865 SME business units that are able to absorb 114,131 workers with a total investment value of around IDR 2.26 trillion. The distribution of SMEs in North Sumatra also shows diverse characteristics, where Medan City has the largest number of business units of 25,073 business units, followed by Deli Serdang Regency with 17,818 business units, Langkat Regency with 10,690 business units, Binjai City with 3,228 business units, and Asahan Regency with 1,838 business units. The difference in the number of business units between regions shows that there is a variation in the level of development of SMEs that require an analytical approach to identify the characteristics of each region so that it can be the basis for the formulation of industrial development policies that are more targeted [3].

Along with the digital transformation of business licensing services, the government has implemented the Online Single Submission Risk-Based Approach (OSS RBA) system as an integrated platform in the process of issuing risk-based business licenses. Through this system, the entire licensing process is carried out electronically so as to produce an administrative database that contains information about the identity of business actors, business locations, business fields based on the Indonesian Business Field Standard Classification (KBLI), investment status, number of workers, and the time of issuance of business licenses. RBA OSS data has great potential to be used as a source of information in analyzing the characteristics of SME development because it is able to describe the condition of business actors in more detail than conventional statistical data. However, the use of OSS RBA data is still more widely used as administrative data and has not been optimally utilized as a source of machine learning-based analysis in supporting decision-making and policy formulation for SME development [4].

The use of machine learning techniques in industrial data analysis has grown rapidly because it is able to process large amounts of data and find patterns that are difficult to identify through conventional analysis. The K-Means algorithm is one of the unsupervised learning methods that is widely used to segment data based on the level of similarity of characteristics [5], while the Random Forest algorithm is a supervised learning method that has high classification capabilities, is stable for large-dimensional data, and is able to identify the most influential variables through feature importance analysis [6]. The integration of the two algorithms allows for segmentation and prediction processes so as to produce more comprehensive information to support data-based decision-making.

Research by Yuliana et al. (2026) applied the integration of K-Means and Random Forest algorithms to classify unemployment patterns in 34 provinces in Indonesia. The study used K-Means to form four groups of unemployment patterns based on data time series, then use the Random Forest to identify the socioeconomic factors that have the most influence on cluster formation. The results of the study show that the two-stage approach is able to provide a better interpretation than the use of one algorithm alone. However, the object of the study still focuses on unemployment patterns so that it has not discussed the segmentation of the development of Small and Medium Industries or the use of RBA OSS data as the main data source [7]. Another study was conducted by Laura Sari et al. (2026) who developed a framework *predictive clustering* using K-Means and Random Forest on OSS administrative data to map the investment profile of MSMEs. The study showed that a combination of processes *clustering* and classification are able to produce very high accuracy in identifying the investment characteristics of business actors. However, the study is oriented towards the investment profile of MSMEs in general and has not specifically analyzed the development of SMEs based on indicators of the number of businesses, SME growth, and regional characteristics in North Sumatra Province [8].

Based on this description, there are still several research gaps that need to be studied further. Research on the use of OSS RBA data as the main source to analyze the development

of Small and Medium Industries (SMEs) at the provincial level is still relatively limited, even though the data has complete characteristics and is able to describe the conditions of business actors in more detail. In addition, previous studies generally have not integrated indicators of SME development, such as the number of business units, the growth of the number of SMEs (Growth IKM), and the average workforce as the main variables in the segmentation process. On the other hand, most studies still apply the K-Means algorithm for segmentation or Random Forest for classification separately, so it has not utilized the two algorithms in a single integrated analytical framework. Therefore, this study combines the K-Means and Random Forest algorithms to produce segmentation as well as a model for predicting the development of SMEs. The determination of the number of clusters was carried out using a combination of the Elbow Method, Silhouette Score, and Davies-Bouldin Index to make the cluster formation process more objective, while the Feature Importance analysis in the Random Forest was used to identify the variables that had the most influence on the segmentation results. This approach is expected to be able to provide more comprehensive information as the basis for the formulation of SME development policies in North Sumatra Province.

Based on the problems that have been described, this study aims to segment the development of Small and Medium Industries (SMEs) in North Sumatra Province using OSS RBA data for the 2022-2025 period and build a prediction model for the results of the segmentation. This research is expected to provide an overview of the characteristics of the development of SMEs in each business group, identify factors that affect the formation of the group, and produce information that can be used as a basis for the formulation of policies for fostering and developing SMEs that are more effective, adaptive, and data-based. In addition to making a practical contribution to local governments, this research is also expected to enrich the application of machine learning methods in the analysis of RBA OSS administrative data, especially in the Small and Medium Industry sector.

Research Methodology

1. Research Design

This study uses a quantitative approach with the *data mining* to segment and predict the development of Small and Medium Industries (SMEs) in North Sumatra Province. The research process consists of two main stages, namely *unsupervised learning* using the K-Means algorithm to form groups (*cluster*) SMEs based on characteristics that have a level of similarity, and *supervised learning* uses the Random Forest algorithm to build a prediction model on the segmentation results obtained [9][10]. Overall, the stages of research carried out are shown in Figure 1.

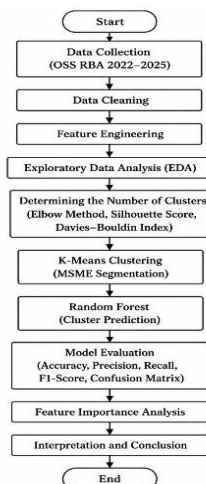


Figure 1. Research Stages

Figure 1 shows the stages of research starting from the collection of OSS data for the RBA of North Sumatra Province for the 2022-2025 period, followed by the process *data cleaning*, *feature engineering*, and *exploratory data analysis* (EDA). Furthermore, the optimal number of clusters was determined using the Elbow Method, Silhouette Score, and Davies-Bouldin Index before the segmentation process using the K-Means algorithm [11][12][13]. The segmentation results are then used as targets in the Random Forest algorithm to build a prediction model. The final stage includes the evaluation of the model using *Confusion Matrix*, *Accuracy*, *Precision*, *Recall*, and *F1 Score*, as well as analysis *Feature Importance* to obtain the interpretation of the results of the research [14].

2. Data Sources and Characteristics

The data used is secondary data obtained from the Online Single Submission Risk-Based Approach (OSS RBA) system of North Sumatra Province for the 2022-2025 period. The dataset contains information about Small and Medium Industry (IKM) players who already have a Business Identification Number (NIB).

3. Segmentation Using the K-Means Algorithm

The K-Means algorithm is used to group SME data based on the similarity of the characteristics of each observation. Before the clustering process, all numerical variables are normalized using the StandardScaler to have a comparable scale.

The determination of the number of clusters was carried out using three evaluation methods, namely:

- Elbow Method to observe value changes *Within Cluster Sum of Squares* (WCSS).
- Silhouette Score to measure the degree of proximity of objects in the same cluster compared to other clusters [12].
- Davies-Bouldin Index (DBI) to evaluate the quality of separation between clusters [13].

The number of clusters used in this study was selected based on the combination of the three evaluation methods so as to produce more optimal segmentation.

4. Prediction Using the Random Forest Algorithm

The segmentation results obtained from the K-Means algorithm are used as the target variable (*Label*) in the classification process using the Random Forest algorithm. All categorical variables were first transformed using the One-Hot Encoding technique, while the dataset was divided into training data (*training data*) and test data (*testing data*) using an 80:20 ratio. Random Forest was chosen because it is able to handle high-dimensional data, has good generalization capabilities, and is relatively resistant to *overfitting* through the formation of a number of *decision tree* Randomly [15].

5. Model Evaluation

The performance of the model was evaluated using several classification metrics, namely:

- Accuracy, to measure the accuracy of the model's predictions.
- Precision, to measure the accuracy of the model in identifying each class.
- Recall to measure the model's ability to find all the data in each class.
- F1-Score, as a harmonious average between precision and recall.
- Confusion Matrix, to evaluate the results of the classification based on the number of true and false predictions in each cluster.

In addition, Feature Importance analysis is used to determine the variables that have the greatest contribution to the classification process so that it can provide interpretation of factors that affect the development of SMEs.

Results

3.1 Characteristics of Research Datasets

The initial stages of the research began with the formation of a research dataset derived from the Online Single Submission Risk-Based Approach (OSS RBA) system of North Sumatra

Province for the 2022-2025 period. The dataset contains information about Small and Medium Industry (IKM) actors who have obtained a Business Identification Number (NIB) as the legality of their business activities. The RBA OSS data was chosen because it is an official data source that records risk-based business licensing activities electronically so that it is able to describe the development of SMEs more comprehensively in each observation period.

Initial data were obtained in the form of four separate datasets representing each observation year, namely 2022, 2023, 2024, and 2025. Each dataset still has a structure that is not yet uniform, such as the existence of a document title line, inconsistent attribute naming, and some values that are not needed in the analysis process. Therefore, before the segmentation and prediction process is carried out, the entire dataset goes through the data *cleaning* stage to produce data that has better quality and consistency.

The *data cleaning* process is carried out through several stages, including merging all datasets based on the observation period, deleting rows that do not contain observation information, standardizing attribute names, adjusting data types, handling missing *values*, and adding Year attributes as the identity of the business license issuance period. This stage aims to ensure that all data has a uniform structure so that it can be used in *the feature engineering process*, segmentation using the K-Means algorithm, and the development of prediction models using the Random Forest algorithm.

The results of *the data cleaning* process produced an integrated dataset consisting of 156,723 observations with 13 attributes. This dataset is then used as the main dataset in all stages of analysis.

To obtain an overview of the composition of the data in each observation period, an analysis of the distribution of the amount of data by year was carried out. The results of the distribution are presented in Table 1.

Table 1. Dataset Distribution by Year

Year	Amount of Data	Percentage (%)
2022	4.550	2,90
2023	31.684	20,22
2024	49.624	31,66
2025	70.865	45,22
Total	156.723	100,00

In addition to the table, the distribution of the amount of data by year can also be visualized to show the change in the number of SMEs recorded in the OSS RBA system during the study period.

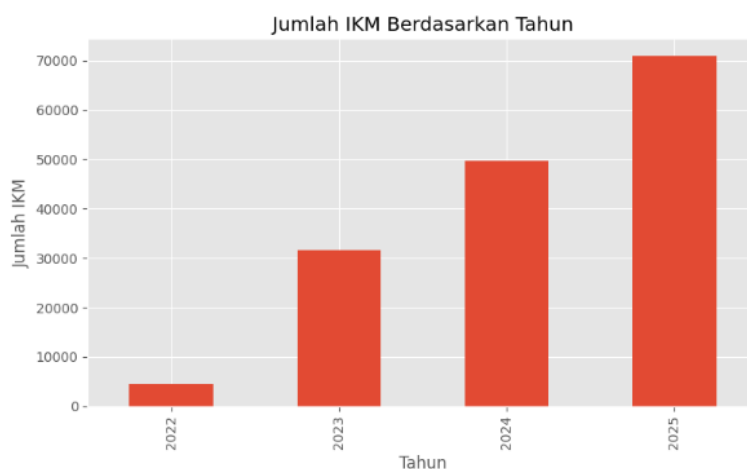


Figure 2. Distribution of the Number of IKM Data by Year

Based on Table 1 and Figure 2, it can be seen that the amount of SME data recorded in the RBA OSS system has increased in each observation period. In 2022, it had the least number

of observations, which was 4,550 data or 2.90% of the total dataset. This number increased to 31,684 data (20.22%) in 2023, then increased to 49,624 data (31.66%) in 2024, and reached the highest number in 2025 of 70,865 data (45.22%).

The increase in the amount of data every year shows that more and more SMEs are registering their businesses through the RBA OSS system. This condition indicates an increase in business legalization activities and the wider use of the risk-based licensing system in the business licensing administration process. From the perspective of data analysis, the variation in the number of observations in each period provides important information in the formation of the IKM Growth variable, so that the model built not only considers the number of business units, but also pays attention to the dynamics of the development of SMEs from year to year.

The cleaning dataset is then processed through *the feature engineering* stage to form analysis variables, such as the Number of SMEs, SME Growth, Average Workforce, Industry Group, and other supporting attributes used in the segmentation process using the K-Means algorithm and the development of prediction models using the Random Forest algorithm. This stage is the basis for the formation of patterns of IKM characteristics which will be analyzed in the next subchapter through the process of determining *the optimal* number of clusters.

3.2 Determination of the Number of Clusters

Determining the number of *clusters* is one of the important stages in the segmentation process using the K-Means algorithm. The number of *clusters* selected will affect the quality of the grouping results because it is related to the level of homogeneity of data within *the cluster* and the level of heterogeneity between *clusters*. Selecting too few *clusters* can cause objects with different characteristics to be in the same group, while too many *clusters* can result in segmentation that is difficult to interpret. Therefore, this study uses three evaluation methods, namely Elbow Method, Silhouette Score, and Davies-Bouldin Index (DBI), to obtain *the most optimal* number of clusters.

a. Evaluation Using the Elbow Method

The Elbow Method is used to identify the optimal number of *clusters* through observation of changes in the Within Cluster Sum of Squares (WCSS) values. The WCSS value represents the square of the distance between each data and the center of *the cluster* (*centroid*). The smaller the WCSS value, the higher the level of similarity between members in a *cluster*. However, increasing the number of *clusters* will always lower the WCSS value, so it is necessary to identify the point of change (*elbow point*) as the basis for selecting *the most efficient* number of clusters.

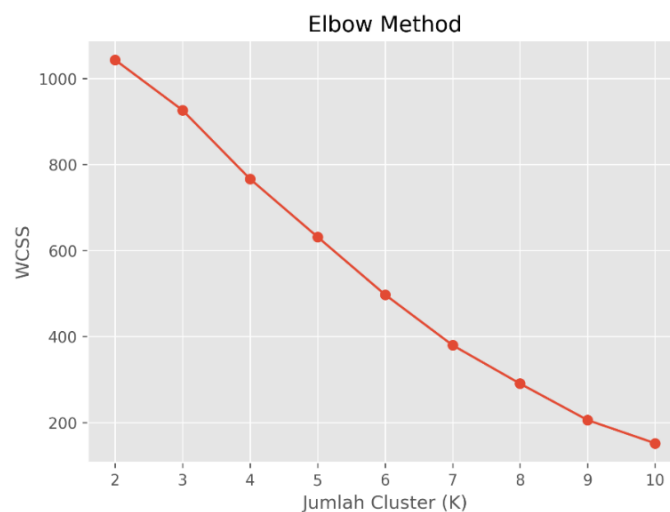


Figure 3. Results of Determining the Number of Clusters Using the Elbow Method

Based on Figure 3, it can be seen that the WCSS value has decreased significantly at a low number of *clusters*. The decline then began to slope as the number of *clusters* increased, suggesting that the addition of *clusters* no longer provided a significant reduction in variation in *clusters*. This pattern indicates that the number of *clusters* worth considering is in the range of four to five clusters.

Although the Elbow Method provides an initial indication of the optimal number of *clusters*, it is visual so the results need to be validated using other evaluation methods. Therefore, this study continues the evaluation using the Silhouette Score.

b. Evaluation Using Silhouette Score

Silhouette Score is used to measure the quality of segmentation results by comparing the degree of proximity of each object to members in the same cluster (*cohesion*) and its distance to the nearest cluster (*separation*). The Silhouette Score value is in the range -1 to 1 . A value that is getting closer to 1 indicates that the object is in the right cluster, while a value that is closer to 0 indicates that there is overlap between *clusters*. A negative value indicates that an object tends to be closer to another *cluster* than the *cluster* in which it is located.

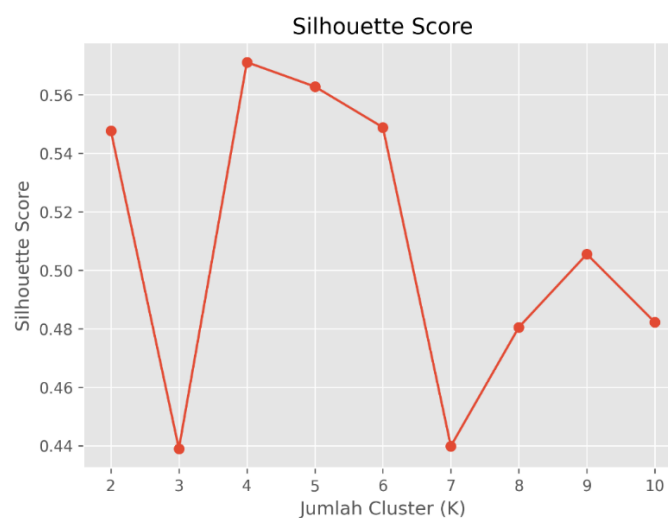


Figure 4. Evaluation Results Using Silhouette Score

Based on Figure 4, the highest Silhouette Score value was obtained at $K = 4$ with a value of around 0.57 , while at $K = 5$ a value of around 0.56 . The difference between the two values is relatively small, indicating that the quality of segmentation in both clusters is almost equal. Although the highest value is obtained at $K=4$, the use of $K=5$ still provides a good level of cohesiveness and cluster separation.

In addition, the use of five *clusters* results in more detailed segmentation so that the characteristics of the development of SMEs can be better explained than if only four groups are used. Therefore, in terms of interpretation of the results, $K = 5$ is considered more representative of the variation in the characteristics of SMEs in North Sumatra Province.

c. Evaluation Using the Davies-Bouldin Index

As the final validation stage, this study uses the Davies-Bouldin Index (DBI). In contrast to the Silhouette Score, the Davies-Bouldin Index measures the quality of segmentation based on the ratio between the spread of data within the *cluster* and the distance between the *cluster's* centers. The smaller the DBI value, the better the quality of the segmentation resulting because it shows that each *cluster* has a high degree of cohesiveness and a clear separation of the other *clusters*.

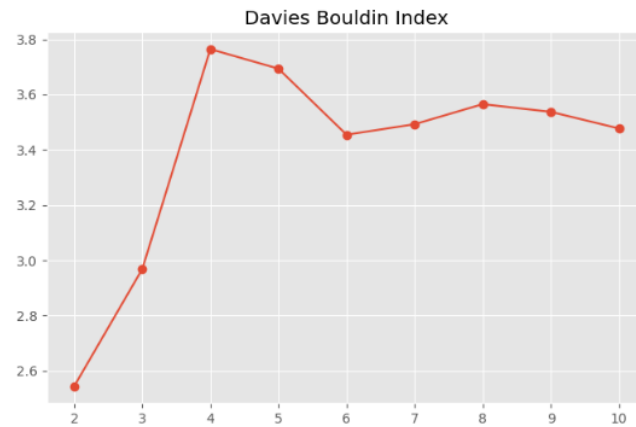


Figure 5. Evaluation Results Using the Davies-Bouldin Index

The results of the evaluation using the Davies-Bouldin Index show that the DBI value at $K = 5$ is still in the good category and supports the results of the evaluation using the Elbow Method and Silhouette Score. This shows that the five *clusters* are able to produce a balance between the level of homogeneity within the *cluster* and the level of heterogeneity between *clusters*.

Based on the combination of the three evaluation methods, this study determined $K = 5$ as the number of *clusters* used in the segmentation process using the K-Means algorithm. This decision not only considers the value of statistical evaluation, but also aspects of the interpretation of the results. The use of five *clusters* allows the identification of the characteristics of the development of SMEs in more detail so that each group has a more specific representation and can be used as a basis for the preparation of strategies for fostering and developing SMEs in North Sumatra Province.

3.3 Segmentation Results Using the K-Means Algorithm

After obtaining the optimal number of clusters as many as five clusters ($K = 5$), the segmentation process was carried out using the K-Means algorithm on the feature engineering dataset. Before the clustering process, all numerical variables are normalized using the StandardScaler method to eliminate the difference in scale between variables. The normalization stage is carried out so that each variable has a balanced contribution to the Euclidean distance calculation process used by the K-Means algorithm in determining the center of the cluster (*centroid*).

The results of the segmentation process resulted in five groups of SMEs that have different development characteristics. Each *cluster* represents a business group with a level of similarity based on the variables of Number of SMEs, SME Growth, Average Workforce, Industry Group, Regency/City, Investment Status, and Year. Thus, each group is not only differentiated based on a single variable, but is the result of a combination of all the characteristics used in the clustering process.

To provide a visual overview of the segmentation results, dimension reduction was carried out using Principal Component Analysis (PCA). The PCA technique is used to project high-dimensional data into two main components so that the distribution pattern of each *cluster* can be observed more easily.

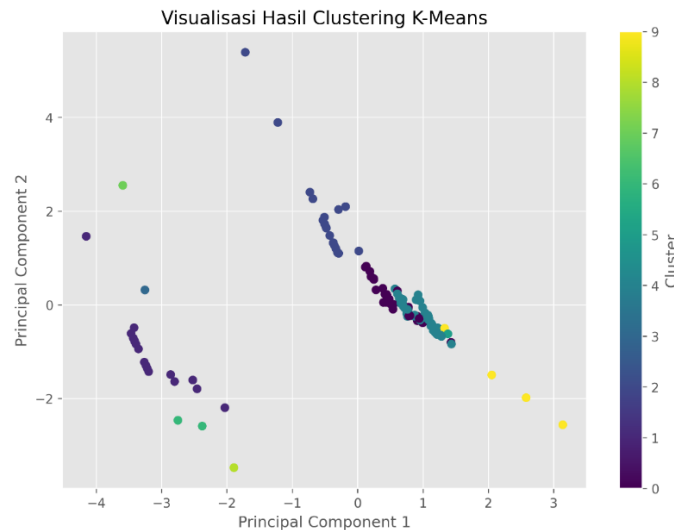


Figure 6. Visualization of segmentation results using the K-Means algorithm

Based on Figure 6, it can be seen that the K-Means algorithm managed to group the data into five *clusters* that have different distribution patterns. Most observations form relatively separate groups, although there are still some points in the transition area between clusters. This condition shows that the characteristics of SMEs in North Sumatra Province have a high enough level of diversity that not all data can be separated perfectly. However, the results of the visualization show that the segmentation process is able to form groups that have a good level of internal similarity and quite clear differences with other groups.

To understand the characteristics of each *cluster*, an analysis was carried out on the average value of the main variables of the *feature engineering results*. A summary of the characteristics of each *cluster* is presented in Table 2.

Table 2. Profile of each cluster characteristics

Cluster	Number of SMEs	Average Workforce	Growth IKM
0	134,79	3,19	552,71
1	34,86	2,73	351,47
2	1,06	48,12	297,78
3	32,99	3,03	337,10
4	38,20	2,80	241,79

Based on Table 2, each *cluster* shows different characteristics so that it can be used to identify patterns of SME development in North Sumatra Province.

Cluster 0 is the group with the highest level of development. This cluster has an average number of SMEs of 134.79 units, much higher than all other *clusters*. In addition, the IKM Growth value of 552.71 indicates that this group experienced the fastest growth in the number of businesses during the observation period. The average workforce of 3.19 indicates that most businesses are still on a small industrial scale, but have high growth dynamics. These characteristics show that Cluster 0 can be categorized as a rapidly growing group of SMEs, which has the potential to be the driving force for the growth of the small and medium industrial sector in North Sumatra Province.

Cluster 1 has an average number of IKM of 34.86 units, with an IKM Growth of 351.47 and an average workforce of 2.73 people. These values indicate that this group has a fairly good growth rate, although not as high as Cluster 0. This characteristic indicates that Cluster 1 consists of SMEs that develop steadily and have opportunities to be improved through business capacity building programs, access to financing, and strengthening marketing networks.

Cluster 2 has the most different characteristics compared to other *clusters*. The average number of SMEs is only 1.06 units, but it has an average workforce of 48.12 people, which is the highest score among all clusters. The IKM Growth Value of 297.78 shows that the increase in the number of businesses is not too high, but each business unit has a relatively large workforce capacity. This condition indicates that Cluster 2 is dominated by labor-intensive SMEs, namely businesses with a relatively small number of units but able to absorb a large number of workers. These characteristics make Cluster 2 important in the context of job creation even though its contribution to the growth of the number of businesses is not as large as other clusters.

Cluster 3 has an average number of SMEs of 32.99 units, an average of 337.10 SMEs, and an average workforce of 3.03 people. This value shows that this group has relatively balanced characteristics between the number of businesses, business growth, and labor capacity. This cluster can be interpreted as a group of SMEs with a medium level of development that still has the potential to be improved through coaching programs, product innovation, and productivity improvement.

Cluster 4 has an average number of SMEs of 38.20 units, with an average workforce of 2.80 people and an IKM Growth of 241.79, which is the lowest growth value compared to all clusters. This condition shows that although the number of business units is relatively higher than Cluster 1 and Cluster 3, the growth rate is still relatively low. Therefore, Cluster 4 can be prioritized as a group that requires policy intervention through business assistance, increasing competitiveness, and expanding access to financing and markets.

Overall, the results of the segmentation show that the development of SMEs in North Sumatra Province is not homogeneous, but is divided into five groups with different characteristics. These differences are reflected in the variation in the number of business units, growth rates, and labor capacity in each *cluster*. These findings show that a uniform coaching approach is not appropriate to be applied to all SMEs. On the other hand, the results of segmentation provide a more objective basis for local governments in designing policies that are tailored to the characteristics of each group. Thus, coaching resources can be allocated more effectively and on target according to the needs of each *cluster*.

3.4 Prediction Results Using Random Forest Algorithm

After the segmentation process using the K-Means algorithm resulted in five groups of SMEs, the next stage is to build a prediction model using the Random Forest algorithm. In this study, the *cluster* labels generated by the K-Means algorithm were used as the *target variables*, while the *feature engineering* results were used as predictor variables. This approach aims to evaluate the ability of the Random Forest algorithm to study the characteristic patterns of each segmentation result so that the resulting model is able to predict the SME group based on its characteristics.

Before the model training process was carried out, all categorical variables were transformed using the One-Hot Encoding technique, while the numerical variables had gone through the normalization process at the previous stage. Furthermore, the dataset is divided into 80% training data and 20% test data (*test set*). This data sharing aims to test the model's ability to classify data that has never been used during the training process so that the model's performance can be evaluated objectively.

The evaluation of model performance was carried out using several classification metrics, namely accuracy, precision, recall, and F1-score. The four metrics were chosen because they were able to provide a comprehensive picture of the level of accuracy of classification both as a whole and in each segmentation result.

a. Model Performance Evaluation

The results of the evaluation of the Random Forest model are presented in Table 3.

Table 3. Random Forest Model Evaluation Results

Metric	Value
Accuracy	0,9881
Precision	0,9882
Recall	0,9881
F1 Score	0,9881

Based on Table 3, the Random Forest model obtained an accuracy value of 0.9881 or 98.81%, which indicates that almost all of the test data was successfully classified into the appropriate clusters. The high accuracy value indicates that the model is able to effectively learn the characteristic patterns of segmentation results so that it has excellent predictive capabilities.

A precision value of 0.9882 indicates that most of the data predicted to be a member of a *cluster* actually comes from that *cluster*. In other words, the rate of false *positives* produced by the model is relatively low. Meanwhile, a recall value of 0.9881 shows that almost all members in each *cluster* were successfully recognized by the model, so the false *negative rate* is also relatively small.

In addition, the model obtained an F1-score of 0.9881, which is a harmonious average between *precision* and *recall*. This value shows that the model is able to maintain a balance between prediction accuracy and the ability to recognize all members in each *cluster*. Overall, the four evaluation metrics showed that the Random Forest algorithm had an excellent classification performance in predicting the segmentation results obtained from the K-Means algorithm.

b. Confusion Matrix Analysis

To obtain a more detailed picture of the classification results, an evaluation was carried out using the Confusion Matrix.

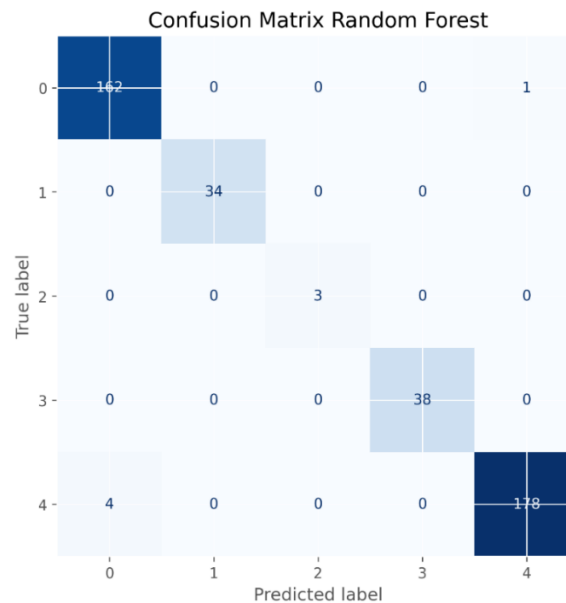


Figure 8. Confusion Matrix Model Random Forest

Based on Figure 8, most of the data resides on the main diagonal of the matrix, which shows that the model managed to classify most of the data into the correct clusters. The dominance of the values on the main diagonal indicates that the rate of *misclassification* is relatively low. Prediction errors only occur in a small number of observations that have relatively similar characteristics to other clusters. This shows that the model is able to effectively distinguish the characteristics of each *cluster* resulting in high classification performance.

c. Analisis Classification Report

In addition to using the *Confusion Matrix*, model evaluation is also carried out through a Classification Report to find out the classification performance of each *cluster*. The results of the evaluation are presented in Table 6.

Tabel 6. Classification Report Model Random Forest

Cluster	Precision	Recall	F1-Score	Support
0	0,9759	0,9939	0,9848	163
1	1,0000	1,0000	1,0000	34
2	1,0000	1,0000	1,0000	3
3	1,0000	1,0000	1,0000	38
4	0,9944	0,9780	0,9862	182
Accuracy	0,9881	0,9881	0,9881	420
Macro Average	0,9941	0,9944	0,9942	420
Weighted Average	0,9882	0,9881	0,9881	420

Based on Table 6, the Random Forest model shows excellent classification performance across the entire *cluster*. This is indicated by the precision, recall, and F1-score values which are in the range of 0.9759 to 1.0000.

In Cluster 0, the model obtained a precision of 0.9759, a recall of 0.9939, and an F1-score of 0.9848 with a total of 163 observations. A higher recall value than *precision* indicates that almost all members of Cluster 0 are successfully recognized by the model. A small difference between the two metrics indicates a small amount of data from other *clusters* that are predicted to be members of Cluster 0. Nevertheless, an F1-score value of 0.9848 indicates that the model still has excellent classification performance in this group.

In Cluster 1 and Cluster 3, the model produces precision, recall, and F1-score values of 1.0000. The results show that all data in both *clusters* were correctly classified without prediction errors. This condition indicates that the characteristics of the two groups have a very clear degree of separation so that they are easily recognized by the Random Forest algorithm.

A similar thing was also obtained in Cluster 2, which resulted in a precision, recall, and F1-score value of 1.0000. However, it should be noted that this *cluster* consists of only 3 observations (*support* = 3). Therefore, even if the model manages to correctly classify all the data, the results need to be interpreted carefully because the relatively small sample size does not fully represent the variation in characteristics that may arise when the data is larger. Thus, the perfect evaluation value of Cluster 2 does not necessarily indicate that this group is more predictable than other *clusters*, but it is also influenced by a very limited amount of data.

In Cluster 4, the model obtained a precision of 0.9944, a recall of 0.9780, and an F1-score of 0.9862 with a total of 182 observations. A higher *precision* value than recall indicates that almost all of the data predicted to be a member of Cluster 4 is a correct member. However, there are still a small number of Cluster 4 members who are predicted to be members of other *clusters*, resulting in a slightly lower *recall* value. Nonetheless, an F1-score value of 0.9862 indicates that the model still has excellent classification capabilities in this group.

In addition to the evaluation of each *cluster*, the macro average results showed an F1-score value of 0.9942, while the weighted average produced a value of 0.9881. The difference between the two values is relatively small, suggesting that the model's performance remains consistent even though the distribution of the amount of data in each *cluster* is not completely balanced. This condition indicates that the Random Forest model has good generalization capabilities and does not experience a significant decrease in performance due to the difference in the number of members in each *cluster*.

Overall, the evaluation results show that the Random Forest algorithm is able to learn the segmentation result patterns formed by the K-Means algorithm with a very high level of accuracy. The high accuracy, precision, recall, and F1-score values show that the integration of

the K-Means and Random Forest algorithms is an effective approach in segmenting as well as predicting the development of Small and Medium Industries (SMEs). In addition to producing a model with high classification performance, this approach also provides a solid basis to support data-driven decision-making in the formulation of SME development policies in North Sumatra Province.

3.5 Analisis *Feature Importance*

In addition to evaluating the accuracy level of the model, this study also analyzed the contribution of each variable to the classification process using the Feature Importance method in the Random Forest algorithm. This analysis aims to identify the variables that exert the greatest influence in the process of forming the classification model so that the main factors that distinguish the characteristics of each *cluster* can be known.

In the Random Forest algorithm, the *feature importance* value is calculated based on the accumulated decrease in the *impurity* value (Gini Impurity) generated by each variable in the entire *decision tree*. The greater the value of *feature importance*, the greater the contribution of the variable in improving the model's ability to distinguish the segmentation result classes.

The results of the *feature importance* analysis are shown in Figure 9.

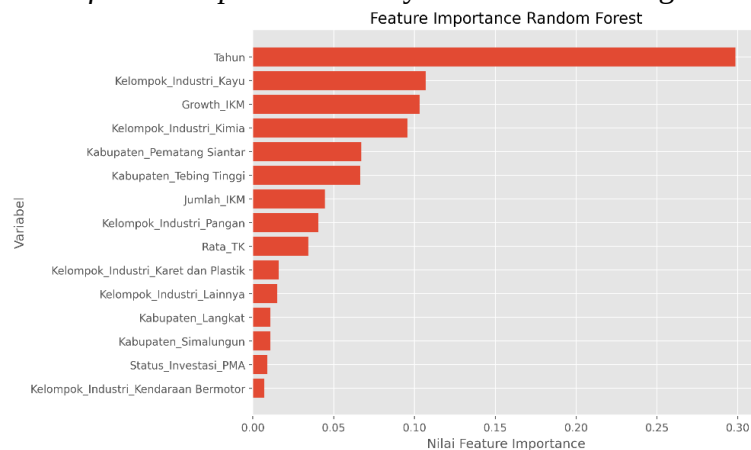


Figure 9. Random *Forest Model* Feature Importance Value

To provide a more detailed picture of the contribution of each variable, the *feature importance value* can also be presented in the form of a table.

Table 7. Variable Ranking Based on *Feature Importance Value*

Peringkat	Variabel	Importance
1	Year	0,2986
2	Kelompok_Industri_Kayu	0,1069
3	Growth_IKM	0,1032
4	Kelompok_Industri_Kimia	0,0959
5	Kabupaten_Pematang Siantar	0,0673
6	Kabupaten_Tebing Height	0,0664
7	Jumlah_IKM	0,0446
8	Kelompok_Industri_Pangan	0,0405
9	Rata_TK	0,0344
10	Kelompok_Industri_Furnitur	0,0252

Based on Figure 9 and Table 7, it can be seen that each variable has a different level of contribution to the classification process. The Year variable is the factor that contributes the most with a *feature importance value* of 0.2986. This value shows that the time dimension is the main indicator in distinguishing the characteristics of the development of SMEs in North

Sumatra Province. The high contribution of this variable indicates that changes in the condition of SMEs during the 2022-2025 period provide very important information for the model in recognizing the pattern of segmentation results. In other words, the dynamics of business development from year to year are the dominant factors that affect the formation of *clusters*.

The next largest contributor was the Timber Industry Group with a *feature importance value* of 0.1069, followed by SME Growth of 0.1032, and the Chemical Industry Group of 0.0959. These results show that in addition to the time factor, the characteristics of the industrial sector also play an important role in distinguishing the SME group. The high contribution of the wood and chemical industry sectors indicates that the two industry groups have quite different development characteristics compared to other industrial sectors so that they are often used by models in the decision-making process. The SME Growth variable which is ranked third shows that the growth rate of the number of SMEs is one of the main indicators in the classification process. These findings are in line with the purpose of the research which focuses on the analysis of the development of SMEs. The *high feature importance value* shows that changes in the number of SMEs between periods are able to provide significant information in distinguishing the characteristics of each *cluster*.

In addition to sectoral characteristics and business growth, the regional aspect also contributes significantly to the model. This is shown by the emergence of the variables of Pematang Siantar Regency and Tebing Tinggi Regency in the fifth and sixth ranks with *feature importance values* of 0.0673 and 0.0664, respectively. The findings show that the location of the business also affects the characteristics of the development of SMEs. Thus, there are variations in development patterns between regions that are successfully recognized by the Random Forest algorithm in the classification process. The variable Number of SMEs obtained a *feature importance value* of 0.0446, while the Average Workforce obtained a value of 0.0344. Although the contribution is lower than the time, industry sector, and region variables, they still play a role in improving the model's ability to distinguish the characteristics of each *cluster*. This shows that the number of business units and labor capacity remain relevant indicators in describing the development conditions of SMEs.

Interestingly, several variables in the category of *One-Hot Encoding* process results, such as the Food Industry Group and the Furniture Industry Group, are also included in the ten variables with *the highest feature importance values*. These results show that the characteristics of the industrial sector have a considerable influence on the classification process. Thus, the development of SMEs is not only influenced by quantitative aspects, such as the number of businesses and growth, but also by the characteristics of the business fields carried out.

Overall, the results of *the feature importance* analysis show that the Random Forest model not only produces a high level of classification accuracy, but is also able to identify the main factors that influence the formation of segmentation results. The dominance of the variables of Year, Growth of SMEs, industrial groups, and regions shows that the development of SMEs in North Sumatra Province is influenced by a combination of temporal, sectoral, and spatial aspects. This information provides added value to this research because it not only produces accurate prediction models, but also provides an analytical basis that can be used by local governments in formulating more targeted SME development policies. For example, regions or industrial sectors that have a major contribution to the formation of clusters can be prioritized in coaching programs, business capacity building, innovation development, and the provision of access to financing so that the policies implemented become more targeted and data-based.

Conclusion

This study successfully integrated the K-Means and Random Forest algorithms to segment and predict the development of Small and Medium Industries (IKM) in North Sumatra Province using OSS RBA data from 2022 to 2025. Following data cleaning and feature engineering, the analysis incorporated variables representing IKM characteristics, including the number of IKM units, IKM growth, average workforce, industry group, regency/city, investment status, and year. The K-Means results identified five optimal clusters ($K = 5$) based on the Elbow Method, Silhouette Score, and Davies–Bouldin Index. The resulting clusters exhibited distinct characteristics in terms of business units, IKM growth, workforce capacity, industrial sectors, and geographical distribution, indicating that IKM development in North Sumatra is heterogeneous. The Random Forest model demonstrated excellent predictive performance, achieving an accuracy of 98.81%, precision of 98.82%, recall of 98.81%, and F1-score of 98.81%. Feature Importance analysis revealed that Year, Wood Industry Group, IKM Growth, Chemical Industry Group, and several geographical variables were among the most influential factors in cluster classification. These findings indicate that IKM development is influenced by a combination of temporal, sectoral, and geographical factors. Overall, the integration of K-Means and Random Forest provides an effective approach for segmenting and predicting IKM development and can support more targeted, objective, and data-driven policymaking for IKM development in North Sumatra Province.

References

- [1] A. Irene Silitonga, A. Akbar Lubis, J. Darma, F. Indra H Sinaga, and Y. Simamora, “Klasterisasi Kabupaten Dan Kota Di Provinsi Sumatera Utara Berdasarkan Industri Kecil Dan Mikro Menggunakan Algoritma K-Means Clustering,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 3, pp. 4878-4885, 2025, doi: 10.36040/jati.v9i3.13797.
- [2] L. Hasna, R. Buaton, and S. Syahputra, “DATA MINING PENGELOMPOKAN INDUSTRI KECIL DAN MENENGAH BERDASARKAN HASIL PRODUKSI MENGGUNAKAN METODE CLUSTERING DI KABUPATEN,” *J. Tek. Inform. Kaputama*, vol. 6, no. 2, pp. 876-884, 2022.
- [3] B. P. S. P. S. U. (2024). P. I. M. dan K. P. S. U. 2023. B. P. S. P. S. Utara., “Profil Industri Mikro dan Kecil Provinsi Sumatera Utara 2024.” [Online]. Available: https://sumut.bps.go.id/id/publication/2024/12/31/3958146f1f61cb8ba1550b1d/profile-of-micro-and-small-industries-in-sumatera-utara-province-2023.html?utm_source=chatgpt.com
- [4] M. Arifullah, J. T. Hardinata, and Y. Pranayama Purba, “Implementasi Metode K-Means Dalam Klasifikasi Desa/Kelurahan Menurut Jenis Industri Kecil Mikro,” *J. Informatics Manag. Inf. Technol.*, vol. 2, no. 1, p. 42, 2022, [Online]. Available: <https://hostjournals.com/>
- [5] Z. Sitorus, I. Syahputra, C. Indra Angkat, and D. Sartika, “Implementation of K-Means Clustering for Inventory Projection,” *Int. J. Sci. Technol. Manag.*, vol. 5, no. 3, pp. 673-678, 2024, doi: 10.46729/ijstm.v5i3.856.
- [6] C. A. Fatikasari, A. M. Rahmania, L. Laili, R. A. Mega Saputri, and M. Arifin, “Analisis Segmentasi dan Prediksi Pola Pembelian IC Label Gamis Menggunakan Hybrid K-Means Random Forest,” *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 5, no. 1, pp. 355-364, 2026, doi: 10.62712/juktisi.v5i1.1007.
- [7] Yuliana, Rima Fazri Ramadhania, Asep Abdul Latip, and Muhammad Farhan Harahapa, “Analysis of Unemployment Patterns in Indonesia Using K-Means Clustering and Identification of Dominant Factors Using Random Forest,” *J. Inotera*, vol. 11, no. 1, pp. 219-231, 2026, doi: 10.31572/inotera.vol11.iss1.2026.id574.
- [8] L. Sari, R. H. Maharrani, H. D. Hastuti, A. P. Ramadhan, and W. Windasari, “Mixed-Data K-Means Clustering with Hyperparameter-Tuned Random Forest for OSS-Based

- MSME Investment Profiling and Policy Targeting,” *J. Tek. Inform.*, vol. 7, no. 2, pp. 1535-1548, 2026, doi: 10.52436/1.jutif.2026.7.2.5545.
- [9] L. N. Hapsari, I. Fannani, Y. Rahmawati, and A. Muhariya, “Explainable Machine Learning untuk Prediksi Risiko Penyakit Jantung Menggunakan Random Forest dan Analisis SHAP,” *Remik*, vol. 10, no. 1, pp. 190-199, 2026, doi: 10.33395/remik.v10i1.15766.
- [10] S. W. Maghfiroh and Supriyono, “PEMODELAN PREDIKTIF POTENSI UMKM DI KABUPATEN PONOROGO MENGGUNAKAN RANDOM FOREST,” *Pros. Semin. Nas. Teknol. dan Sist. Inf.*, no. November, pp. 326-338, 2025.
- [11] Abdussalam Amrullah, Intam Purnamasari, Betha Nurina Sari, Garno, and Apriade Voutama, “Analisis Cluster Faktor Penunjang Pendidikan Menggunakan Algoritma K-Means (Studi Kasus: Kabupaten Karawang),” *J. Inform. dan Rekayasa Elektron.*, vol. 5, no. 2, pp. 244-252, 2022, doi: 10.36595/jire.v5i2.701.
- [12] M. D. Triyadi, M. S. Novelan, and R. F. Wijaya, “Penerapan Data Mining Dan Machine Learning Dalam Menentukan Strategi Bisnis Umkm Berdasarkan Tren Penjualan Dan Perilaku Konsumen,” *Instal J. Komput.*, vol. 18, no. 05, pp. 62-69, 2026.
- [13] B. Nurul and L. Riska, “Klasifikasi Keberhasilan Usaha Mikro , Kecil , dan Menengah (UMKM) di Indonesia Menggunakan Algoritma C4 . 5 dengan Pendekatan Data Mining,” vol. 5, no. 1, pp. 74-83, 2026.
- [14] F. Riskiyono and D. Mahdiana, “Implementation of Random Forest Algorithm for Graduation Prediction,” *Sinkron*, vol. 8, no. 3, pp. 1662-1670, 2024, doi: 10.33395/sinkron.v8i3.13750.
- [15] A. Ernawati, Khairul, Z. Sitorus, M. Iqbal, and D. Nasution, “Penerapan Data Mining Untuk Klasifikasi Penduduk Miskin Di Kabupaten Labuhanbatu Menggunakan Random Forest Dan K-Nearest Neighbors,” *Bull. Inf. Technol.*, vol. 6, no. 2, pp. 71-78, 2025, doi: 10.47065/bit.v5i2.1783.