VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8271

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

HYBRID DATA MINING METHODS TO SUPPORT MSME SUSTAINABILITY IN RURAL AND URBAN AREAS

Krisantus Jumarto Tey Seran[1*]; Debora Chrisinta[1]; Yasinta Oktaviana Legu Rema[1]; Hevi Herlina Ullu[1]; Budiman Baso[1]

Information Technology[1]
Universitas Timor, Kefamenanu, North Central Timor, Indonesia[1]
https://unimor.ac.id/[1]
krisantusteyseran@unimor.ac.id*, deborachrisinta@unimor.ac.id, rema.ivana@gmail.com, heviherlina@unimor.ac.id, budimanbaso@gmail.com

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *This study aims to analyze the factors influencing the sustainability of MSMEs in North Central Timor Regency by utilizing the K-Mode clustering method and Naive Bayes classification. The data used includes 550 MSMEs in North Central Timor Regency, East Nusa Tenggara Province, classified based on attributes such as location, product price, financial condition, innovation, technology utilization, and sustainability. The K-Mode method was employed to group MSMEs based on categorical similarities after the data was segmented by location attributes, while the Naive Bayes method was applied to classify MSME sustainability following clustering. The results indicate that in rural areas, MSMEs tend to dominate with high product prices and good financial conditions but show low levels of innovation and technology utilization. In contrast, MSMEs in urban areas are generally more innovative and technology-driven despite facing infrastructure challenges. The application of Naive Bayes demonstrated that a data training ratio of 70:30 yielded the best accuracy. Accordingly, the resulting model can be utilized to monitor the sustainability conditions of MSMEs. This study provides insights into sustainability patterns of MSMEs in both rural and urban areas and opens opportunities for further research on external factors affecting sustainability and barriers to technology adoption in rural areas.*

Keywords: *Data Mining, K-Mode, MSME Sustainability, Naive Bayes, Categorical.*

Intisari— *Penelitian ini bertujuan untuk menganalisis faktor-faktor yang mempengaruhi keberlanjutan UMKM di Kabupaten Timor Tengah Utara dengan memanfaatkan metode klaster K-Mode dan klasifikasi Naive Bayes. Data yang digunakan mencakup 550 UMKM di Kabupaten Timor Tengah Utara, Provinsi Nusa Tenggara Timur, yang diklasifikasikan berdasarkan atribut seperti lokasi, harga produk, kondisi keuangan, inovasi, pemanfaatan teknologi, dan keberlanjutan. Metode K-Mode digunakan untuk mengelompokkan UMKM berdasarkan kesamaan kategorikal setelah data disegmentasi berdasarkan atribut lokasi, sedangkan metode Naive Bayes diterapkan untuk mengklasifikasikan keberlanjutan UMKM setelah pengelompokan. Hasil penelitian menunjukkan bahwa di daerah pedesaan, UMKM cenderung didominasi dengan harga produk tinggi dan kondisi keuangan yang baik namun menunjukkan tingkat inovasi dan pemanfaatan teknologi yang rendah. Sebaliknya, UMKM di daerah perkotaan umumnya lebih inovatif dan berbasis teknologi meski menghadapi tantangan infrastruktur. Penerapan Naive Bayes menunjukkan bahwa rasio pelatihan data 70:30 menghasilkan akurasi terbaik. Dengan demikian, model yang dihasilkan dapat digunakan untuk memantau kondisi keberlanjutan UMKM. Penelitian ini memberikan wawasan tentang pola keberlanjutan UMKM di daerah pedesaan maupun perkotaan serta membuka peluang penelitian lebih lanjut mengenai faktor eksternal yang mempengaruhi keberlanjutan dan hambatan adopsi teknologi di daerah pedesaan.*

Kata Kunci: *Penambangan Data, K-Mode, Keberlanjutan UMKM, Naive Bayes, Kategorikal.*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8271

INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) play a crucial role in Indonesia's economy, particularly in generating employment and fostering economic growth in both rural and urban areas [1]. However, MSMEs in North Central Timor Regency, East Nusa Tenggara Province, face numerous challenges, such as limited market access, technological constraints, and inadequate support for data driven decision making [2], [3], [4], [5], [6], [7]. These challenges often make it difficult for MSMEs to survive and grow sustainably. This research is significant as it aims to provide solutions to the challenges faced by MSMEs through a technology based approach. Factors such as location, product pricing, financial conditions, innovation, and technology adoption are crucial components influencing the success and sustainability of MSMEs. By thoroughly understanding these factors, more effective strategies can be developed to enhance the competitiveness of MSMEs.

Preceding research have discussed the use of data mining methods to support data driven analysis and decision making in the MSME sector [8], [9]. For instance, [10] demonstrated that clustering methods could provide optimal recommendations for determining MSME marketing strategies. Similarly, [11] utilized data mining techniques to facilitate grouping, simplifying the process of coaching and assistance for MSME development. Additionally, a study by [12] confirmed the effectiveness of hybrid data mining methods in analyzing categorical data to identify customer behavior patterns. However, most of these studies were conducted in urban areas with adequate technological access, leaving a gap in the application of these methods for MSMEs in remote regions such as North Central Timor Regency.

One of the approaches used in this study is the implementation of hybrid data mining methods, specifically a combination of K-Mode and Naive Bayes algorithms [13], [14]. While this combination is established in data mining literature, this research contributes a domain specific framework tailored for the MSME sector, where data is predominantly categorical and often unstructured. Rather than treating the algorithms as a generic pipeline, this study introduces a specialized feature engineering process that aligns MSME behavioral indicators with the clustering output of K-Mode, which then serves as a refined prior for the Naive Bayes classification. This applied methodology enables a more robust analysis of market segmentation and key success factors, positioning the research as a strategic applied study that bridges the gap between theoretical data mining and practical business intelligence for small scale enterprises.

To elucidate the positioning of this study's contribution, it should be emphasized that the novelty does not lie in the methodological development of the K-Mode or Naive Bayes algorithms individually, as both are already well-established in the data mining literature. Instead, the primary contribution of this research is applied contribution: (1) the development of a feature engineering framework tailored to the context of MSMEs in regions characterized by limited structured data; (2) to the best of the authors' knowledge, the first application of a hybrid K-Mode and Naive Bayes approach to an MSME dataset from a disadvantaged region such as North Central Timor Regency; and (3) the separation of the analysis by location (rural vs urban) to generate distinct policy recommendations. Thus, readers should not claim that this study introduces a new algorithm; rather, it demonstrates how existing methods can be effectively adapted to address real-world challenges in the rural and urban MSME sectors.

Prior research has shown that hybrid data mining methods can enhance accuracy and efficiency in data analysis. For example, the combination of K-Mode and Naive Bayes has been successfully applied in various fields, including healthcare, education, and business management. However, specific applications to support the sustainability of MSMEs in remote areas like North Central Timor Regency remain rare, creating opportunities for further exploration. This study offers novelty in applying hybrid data mining methods to analyze and support MSME sustainability, particularly within unique geographical and socio-economic contexts. The findings are expected to not only provide new insights into data driven MSME management but also offer strategic recommendations that can be implemented by stakeholders to support the sustainable growth of MSMEs in both rural and urban settings.

To guarantee a more detailed and relevant analysis, data from rural and urban MSMEs will be analyzed separately [15]. This separation is necessary due to significant differences in the social, economic, and infrastructure characteristics between rural and urban MSMEs. Rural MSMEs often face challenges such as limited access to technology and markets, while urban MSMEs are more influenced by intense competition and the need for innovation. By separating the analyses, this

study can provide more specific and tailored recommendations for each region's needs.

Thus, this study seeks to answer the following question: How can hybrid data mining methods be effectively applied to support MSME sustainability in North Central Timor Regency? This research not only contributes to the advancement of scientific knowledge but also offers practical and impactful solutions for the development of MSMEs in the region.

MATERIALS AND METHODS

This study employs a quantitative approach with an exploratory design to analyze the sustainability of MSMEs in North Central Timor Regency. The data utilized in this research encompasses 550 MSMEs, classified based on several attributes, namely location, product pricing, financial condition, innovation, technology utilization, and MSME sustainability. To provide better transparency, the distribution of the target variable (MSME sustainability) was analyzed. The results indicate that the proportion of sustainable and non-sustainable MSMEs is relatively balanced, minimizing the risk of classification bias in the model.

The dataset was obtained from official records of the Cooperatives and SMEs Agency of North Central Timor Regency. The data collection process involved administrative documentation and field verification conducted by the agency. Therefore, the dataset is considered reliable for representing MSME conditions in the region. However, several potential limitations and biases should be acknowledged. First, the data may contain reporting bias, as it relies on self-reported information from MSME actors. Second, the dataset is limited to a specific geographical region, which may affect the generalizability of the findings. Third, some attributes are categorical and simplified, which may not fully capture the complexity of MSME conditions. Despite these limitations, the dataset remains suitable for exploratory analysis and provides valuable insights into MSME sustainability patterns in rural and urban areas. The following are the details of the attributes used:

Table 1. Definition of Research Attributes

Attribute	Category	Information
Location	Village, City	The geographical location where MSMEs operate.
Product Price	Low, Medium, High	Categories based on the price level of products sold by MSMEs.
Financial Conditions	Good, Bad	The financial condition of MSMEs is based on income and financial stability.
Innovation	Yes, No	The existence of innovation in MSME products, services, or business processes.
Utilization of Technology	Yes, No	The level of technology adoption in MSME operations.
Sustainability of MSMEs	Yes, No	Indicators of whether MSMEs can continue to operate sustainably

Source: (Research Results, 2025)

The data were collected at the Office of the Cooperatives and SMEs Agency of North Central Timor Regency. This process involved gathering relevant data, such as MSME locations, product prices, financial conditions, levels of innovation, technology utilization, and sustainability status. A total of 550 MSMEs were included, serving as the foundation for the research analysis. Once the data were collected, they were entered into a notebook using the Python programming language for preprocessing. missing data checks and label encoding The preprocessing stage aimed to ensure data quality before analysis. In addition to handling missing values and applying label encoding additional steps were conducted [16], [17], several additional steps were conducted. First, data consistency checking was performed to identify and correct inconsistent or invalid entries.

Duplicate records were also removed to avoid redundancy in the dataset. Furthermore, the distribution of the target variable was analyzed to ensure that the dataset is relatively balanced, thereby reducing potential bias in the classification model. Since the dataset consists primarily of categorical attributes, outlier detection was not applied, as the concept of extreme values is not directly relevant in categorical data. Instead, emphasis was placed on ensuring the validity and consistency of category labels. In addition, feature relevance was reviewed to ensure that each selected attribute contributes meaningfully to the analysis of MSME sustainability.

However, advanced preprocessing techniques such as feature selection algorithms and data resampling methods were not implemented in this study. These aspects are acknowledged as limitations and provide opportunities for future research to further improve model performance. Missing data checks involved examining whether any values were missing or empty [18]. If identified, handling methods such as row deletion or mode imputation were applied to maintain dataset integrity. Meanwhile, label encoding converted all categorical data into numerical representations. For instance, attributes like Rural and Urban in the location data were assigned labels 0 and 1, respectively. This step prepared the data for the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8271

Naive Bayes algorithm, which only accepts numerical input for target attributes [19]. Following data preprocessing, the data were categorized based on the location attribute into two categories: Rural and Urban. This categorization aimed to identify specific characteristics of MSMEs in each location before further analysis. Subsequently, clustering and classification methods were applied to the data separated by location attributes. The clustering method employed the K-Mode algorithm, which is designed for categorical data. While the K-Means algorithm is commonly used for numerical data, it is unsuitable for categorical data due to its reliance on Euclidean distance. K-Mode replaces Euclidean distance with a distance measure appropriate for categorical data. The equation for clustering objects using the K-Mode algorithm is as follows:

$$Cost(C) = \sum_{i=1}^{n} \sum_{j=1}^{m} \delta(X_{ij}, C_{kj}) \quad (1)$$

Where X_{ij} represents the value of the j-th attribute in the i-th data point, C_{kj} denotes the mode of the k-th cluster, and $\delta(X_{ij}, C_{kj})$ is the distance function that equals 1 if $x \neq y$ and 0 if $x = y$. The steps for object clustering using the K-Mode algorithm are as follows [20]:

1. Initialize the values of k cluster centers randomly.
2. Assign each data point to the nearest cluster based on the categorical distance (δ).
3. Calculate the mode of the data within each cluster to determine the new centers.
4. Repeat the steps until the cluster centers remain unchanged or the maximum number of iterations is reached.

A parameter of $k = 2$ is utilized in this study to partition the dataset into two distinct clusters: one that aligns more closely with sustainability characteristics (cluster 1) and another that is less sustainable (cluster 0). Subsequently, the data within each cluster (0 and 1) is divided into training and testing sets using three different ratios: 90:10, 80:20, and 70:30 [21]. The use of these three ratios aims to ensure the consistency of analytical results across varying proportions of training and testing data and to identify the optimal ratio for balancing generalization and model performance [22]. The classification method applied is the Naive Bayes algorithm, based on the probability equations of categorical attributes, as follows [23], [24]:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (2)$$

Where $P(C|X)$ represents the probability of class C given data X, $P(X|C)$ denotes the probability of the occurrence of data X within class C, $P(C)$ indicates the prior probability of class C, and $P(X)$ signifies the total probability of data X. he stages of the classification process using categorical attributes consist of:

1. Categorical data attributes are processed to ensure that each category can be analyzed probabilistically. For instance, the attribute "Location" (Village, City) is transformed into the frequency distribution of each category within each class.
2. The prior probability $P(C)$ is calculated as the proportion of data within each class:

$$P(C) = \frac{Number\ of\ data\ points\ in\ class\ C}{Total\ number\ of\ data\ points} \quad (3)$$

3. For each categorical attribute X_i, the likelihood probability $P(X|C)$ is determined based on the frequency distribution of attribute values within the class:

$$P(X_i|C) = \frac{Number\ of\ data\ points\ with\ value\ Xi\ in\ class\ C}{Total\ number\ of\ data\ points\ in\ class\ C} \quad (4)$$

4. The posterior probability P(C | X) is calculated as:

$$P(C|X) \propto P(C) \times \prod_{i=1}^{n} P(X_i|C) \quad (5)$$

where n is the number of attributes.

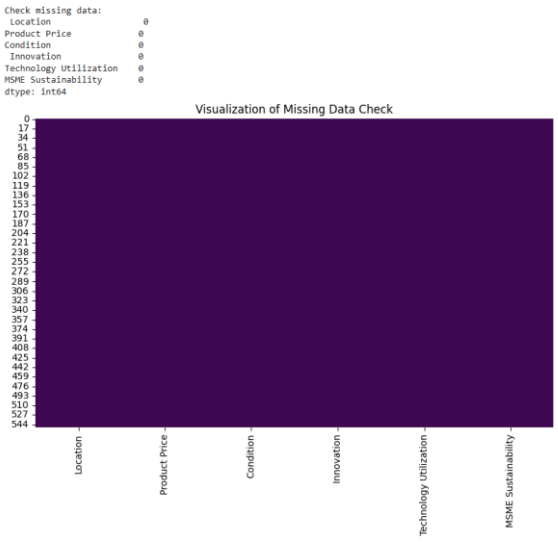
5. The data X is classified into the class C with the highest posterior probability:

$$C = \underset{C}{argmax}\, P(C|X) \quad (6)$$

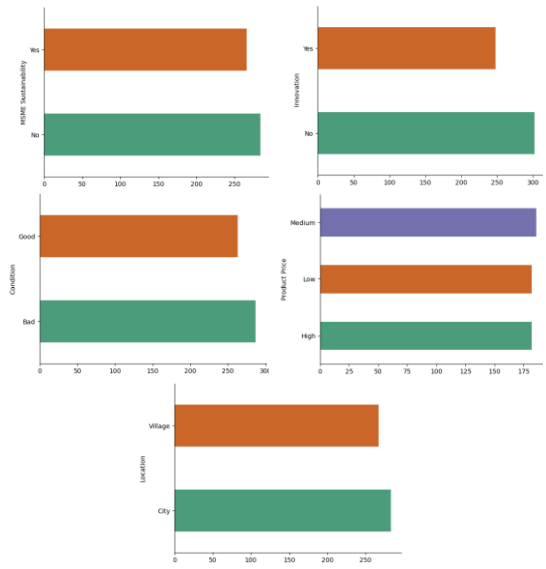
RESULTS AND DISCUSSION

Based on the visualization of predictor attributes in Figure 1. Conducted as an initial data exploration step, the distribution of each category within the research attributes shows a relatively balanced proportion. In the MSME Sustainability attribute, the “Yes” (sustainable) and “No” (unsustainable) categories have almost the same number. The Innovation attribute indicates that the

number of SMEs with innovation “Yes” is slightly lower than those without innovation “No”, but the difference is not significant. For the Condition attribute, the "Good" category is slightly more dominant than "Bad”, though the balance is still relatively maintained. In the Product Price attribute, the "High", "Medium", and "Low" categories are well distributed, although the "High" category is slightly higher. Meanwhile, in the Location attribute, SMEs in the city slightly outnumber those in the village. This balance in distribution is crucial to prevent bias in the analytical model and to produce more accurate results.

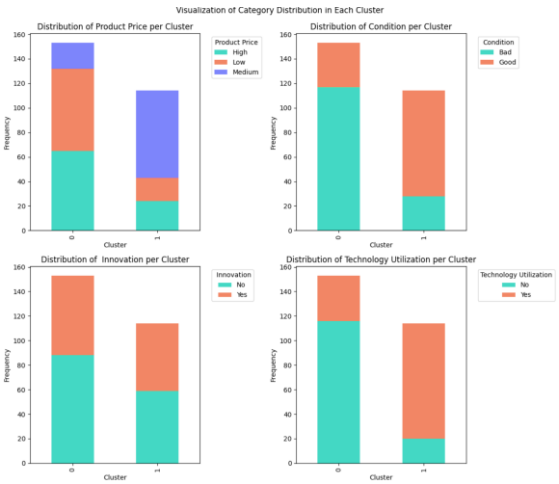


Source: (Research Results, 2025)
Figure 1. Results of Missing Data Check



Source: (Research Results, 2025)
Figure 2. Visualization of Category Distribution in Predictor Attributes

The data is separated based on the Location attribute into two subsets, namely Village and City, in order to analyze the characteristics of MSMEs in each region more specifically. Subsequent to the separation, the K-Mode Clustering method with the number of clusters $k = 2$ is applied to each subset. This method is chosen because the data is categorical, and K-Mode is effective for clustering such data based on category similarity. The initialization of $k = 2$ is performed to divide the MSMEs into two main groups, such as based on differences in innovation patterns, financial conditions, or technology utilization. The clustering process involves initializing the cluster centers [25], determining the clusters based on category distance, and updating the mode until convergence. The results of this clustering yield two groups in each location, reflecting the main patterns of MSME characteristics, which are then used for further analysis using the Naïve Bayes algorithm. The following is the clustering result for the village category Location attribute:



Source: (Research Results, 2025)
Figure 3. Distribution of Village Data Predictor Attribute Categories Based on Cluster Formation

Evident from the graph of the distribution of predictor attribute categories in the MSME data from the village with the formation of clusters $k = 2$, the following is a complete explanation for each attribute:

Table 2. Category Characteristics in Predictor Attributes Based on the Formation of Village Data Clusters

Atribut	Cluster 0	Cluster 1	Conclusion
Product Price	It is dominated by the "High" price category, followed by "Low" and	The distribution is more evenly distributed with fewer	Cluster 0 represents MSMEs that focus on high-end

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8271

Atribut	Cluster 0	Cluster 1	Conclusion
	"Medium." The majority of MSMEs sell products at high prices.	"High" categories than "Low" and "Medium." MSMEs tend to offer products at more affordable prices.	products, while Cluster 1 is more focused on affordable product pricing.
Condition	The financial condition of "Good" is more dominant than "Bad," indicating that MSMEs in this cluster are more financially stable.	The proportion of "Bad" financial conditions is greater than "Good," indicating that more MSMEs are facing financial challenges.	Cluster 0 is composed of more financially stable MSMEs, while Cluster 1 faces greater financial challenges, with a higher proportion of MSMEs experiencing financial difficulties.
Innovation	MSMEs without innovation "No" are more dominant than MSMEs with innovation "Yes". Innovation is not the top priority in this cluster.	More MSMEs with innovation "Yes" than cluster 0, although still fewer than those without innovation.	Cluster 0 shows limited focus on innovation, while Cluster 1 has a higher tendency to adopt innovation, although still not widespread.
Technology Utilization	Dominated by MSMEs that do not utilize technology "No", indicating the lack of use of technology in operations.	The proportion of MSMEs that utilize technology "Yes" is higher than cluster 0, indicating better technological adaptation.	Cluster 0 shows a significant gap in technology utilization, whereas Cluster 1 shows better technology adaptation, though it still needs improvement overall.

Source: (Research Results, 2025)

In general, it can be said that Cluster 0 tends to consist of SMEs with high product prices, good financial conditions, but lacking innovation and rarely utilizing technology. In contrast, Cluster 1 predomi-nantly includes SMEs with lower or medium product prices, more challenging financial conditions, but with greater use of innovation and technology. This distribution provides an overview of the characteristic patterns of SMEs in villages, which can be used for more specific policy making. With the clustering results established, the next step is to apply the Naive Bayes method to further analyze each cluster.

The main goal of applying Naive Bayes to the clustering results is to build a classification model that can predict the sustainability of SMEs based on predictor attributes in each cluster. By dividing the data based on clusters, the analysis becomes more specific, as the characteristics of each clus-ter have unique patterns that can help improve prediction accuracy. Below is an explanation of how the Naive Bayes method is applied to the clustering results and how the results provide deeper insights. The application of the Naive Bayes method on data grouped using the K-Mode clustering method shows varied performance based on training and testing data ratios (90:10, 80:20, and 70:30) as well as within each cluster (Cluster 0 and Cluster 1). Below is the analysis based on evaluation metrics:

Table 3. Naive Bayes Performance Evaluation Metrics on Village Data

Cluster	Data Ratio	Precision	Recall	F1-score	Accuracy
0	90:10	77,72%	76,32%	76,73%	76,32%
	80:20	89,44%	89,47%	89,43%	89,47%
	70:30	91,56%	92,84%	91,35%	92,86%
1	90:10	42,22%	60%	48,57%	60%
	80:20	68,90%	70%	68,96%	70%
	70:30	85,23%	83,33%	80,32%	83,33%

Source: (Research Results, 2025)

and In cluster 0, with a 90:10 ratio, the precision and recall values are balanced, with a relatively high F1-score, indicating that the model performs well in detecting both positive and negative data. The 80:20 ratio shows better performance compared to the 90:10 ratio, with precision, recall, and accuracy values very close to one another, demonstrating a good balance. The 70:30 ratio yields the best results, with accuracy reaching 92.86% and 83.33%. This finding suggests that increasing the proportion of training data improves the model's ability to generalize and predict unseen data, which is consistent with recent studies on data splitting strategies in machine learning [26].

In this study, model evaluation is conducted using accuracy, precision, recall, and F1-score, which remain widely used metrics for classification performance evaluation, especially in probabilistic models such as Naive Bayes. These metrics are considered effective for assessing classification performance across different class distributions. However, it is acknowledged that additional evaluation techniques such as confusion matrix analysis, Receiver Operating Characteristic (ROC)

curves, Area Under the Curve (AUC), and k-fold cross-validation can further enhance evaluation robustness and model reliability. These approaches are particularly useful for analyzing classification thresholds, model stability, and generalization performance. Due to the exploratory nature of this study and its focus on applied analysis, these advanced evaluation techniques were not implemented. Nevertheless, their inclusion is recommended for future research to provide a more comprehensive evaluation framework.

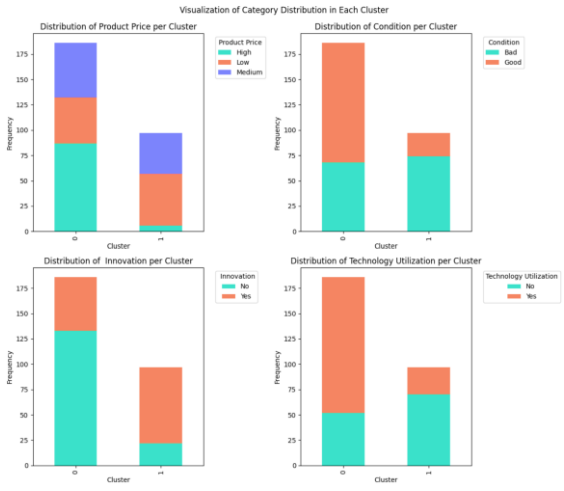
Regarding cluster 1, the 90:10 ratio shows low precision, indicating many false positive predictions. However, recall is relatively high, suggesting the model is still good at detecting positive data, even with many incorrect predictions. Model performance improves at the 80:20 ratio, with more balanced precision and recall. Accuracy also improves, but it still lags behind cluster 0, suggesting the model may be more difficult to predict accurately for this cluster. At the 70:30 ratio, the model achieves better results, with relatively balanced precision and recall. Accuracy significantly improves compared to previous ratios, showing that more training data helps enhance model performance.

Overall, cluster 0 tends to deliver better results in terms of precision, recall, F1-score, and accuracy, with performance improving as the proportion of training data increases (from 90:10 to 70:30). Cluster 1 shows relatively lower performance compared to cluster 0, indicating that the model may face greater challenges in capturing patterns within this cluster. Based on the observed evaluation metrics, the 70:30 data split consistently produces higher performance values across most indicators. This suggests a tendency for improved model performance when a larger portion of the data is used for training.

However, it should be noted that this conclusion is based on descriptive comparison rather than statistical significance testing. Statistical validation methods such as t-test or ANOVA are commonly recommended to assess whether performance differences are statistically significant. These methods were not applied in this study, which is considered a limitation. Therefore, the findings should be interpreted as indicative trends rather than definitive conclusions. Future research is encouraged to incorporate statistical testing and repeated experimental runs to provide more robust validation of model performance across different data splitting strategies cluster.

The application of the K-Mode algorithm and Naive Bayes to MSME data in the village shows that K-Mode clustering divides the data into two main clusters. Cluster 0, which is dominated by MSMEs with high product prices, good financial conditions, but low innovation and rare use of technology. Cluster 1, which contains MSMEs with more affordable product prices, challenging financial conditions, but a higher adoption of innovation and technology. The application of Naive Bayes to each cluster demonstrates better performance at a 70:30 training ratio, achieving an accuracy of 92.86% in cluster 0 and 83.33% in cluster 1. Therefore, the Naive Bayes model can be used to check the conditions of MSMEs in the village.

Following the analysis of MSME clusters data in the village using the K-Mode algorithm, the next step is to analyze the results obtained from applying K-Mode to city data. As with the village data, the main goal of this clustering is to identify patterns in the characteristics of MSMEs that can provide deeper insights into the differences in needs and challenges faced by MSMEs in urban environments. Below is the visualization of the data distribution within each cluster:



Source: (Research Results, 2025)

Figure 4. Distribution of City Data Predictor Attribute Categories Based on Cluster Formation

Based on the visualization in Figure 4, the analysis of the city data clustering results by category, present-ed in detail, is provided in the following table:

Table 4. Category Characteristics in Predictor Attributes Based on the Formation of City Data Clusters

Atribut	Cluster 0	Cluster 1	Conclusion
Product Price	The product price distribution is more evenly distributed with significant	It is dominated by the "High" category with a smaller	Cities in cluster 1 have product prices that tend to be

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8271

Atribut	Cluster 0	Cluster 1	Conclusion
	proportions for the "High," "Medium," and "Low" categories.	proportion of other categories.	higher than cluster 0.
Condition	Dominated by "Bad" conditions, although there are some data with "Good" conditions.	The proportion of "Good" conditions is larger compared to cluster 0.	Cities in cluster 1 have better infrastructure or conditions.
Innovation	The majority of data shows that innovation has not been implemented "No".	The proportion of cities that adopt innovation "Yes" is higher.	Cities in cluster 1 are more innovative than cluster 0.
Technology Utilization	The majority of cities have not made much use of technology "No".	The use of technology "Yes" is more dominant, although there are still some who have not taken advantage of it.	Cities in cluster 1 are more advanced in the application of technology.

Source: (Research Results, 2025)

Generally, cluster 0 represents cities with more underdeveloped characteristics, including greater var-iation in product prices, poorer conditions, low innovation levels, and minimal technology utilization. In contrast, cities in cluster 1 exhibit more advanced characteristics, with higher product prices, better condi-tions, higher innovation levels, and more significant technology utilization. It can be stated that cluster 1 represents cities with greater potential to become centers for technology and innovation development, while cluster 0 requires more attention for infrastructure improvement and technology adoption. Upon implementing the clustering process on the city data to group cities based on specific characteristics, the next step is to apply the Naive Bayes method. Below are the metric results from the Naive Bayes method on the clustering data:

Table 5. Naive Bayes Performance Evaluation Metrics on City Data

Cluster	Data Ratio	Precision	Recall	F1-score	Accuracy
0	90:10	77,72%	76,32%	76,73%	76,32%
	80:20	89,44%	89,47%	89,44%	89,47%
	70:30	91,56%	92,86%	91,35%	92,86%
1	90:10	52,22%	70%	58,57%	70%
	80:20	78,90%	80%	78,96%	80%
	70:30	85,23%	83,33%	80,32%	83,33%

Source: (Research Results, 2025)

For Cluster 0, model performance improved consistently with more training data. At the 90:10 ratio, accuracy reached only 76.32%. Increasing to 80:20 boosted accuracy to 89.47%. The best results emerged at 70:30, achieving 92.86% accuracy alongside 91.56% precision and 92.86% recall. Cluster 1 proved more challenging. The 90:10 ratio yielded low precision (52.22%) indicating many false positives despite moderate recall (70%). Adding training data (80:20) raised accuracy to 80%. Peak performance occurred at 70:30, with 83.33% accuracy, 85.23% precision, and 83.33% recall. Still, all metrics for Cluster 1 lagged behind those of Cluster 0, suggesting greater heterogeneity or class overlap within this cluster.

Across both clusters, enlarging the training set from 90:10 to 70:30 consistently enhanced precision, recall, F1-score, and accuracy. Accordingly, the 70:30 data ratio is recommended as it yields a more optimal model. Based on the clustering results and the application of the Naive Bayes method on city data, it can be concluded that the clustering successfully grouped cities based on key characteristics such as product prices, infrastructure conditions, innovation levels, and technology adoption. Cities in cluster 1 exhibit more advanced characteristics, making them potential centers for innovation and technology development, while cluster 0 represents cities facing greater challenges in infrastructure and technology adoption.

The application of Naive Bayes shows that increasing the proportion of training data (70:30 ratio) consistently improves the model's performance in predicting patterns within each cluster [27] [28], [29]. Therefore, using a larger data ratio is recommended for more accurate predictions, especially for clusters with more complex characteristic patterns, such as cluster 1.
The analysis of MSME data in rural and urban areas reveals significant differences in the characteristics of the formed clusters. In rural data, cluster 0 is dominated by MSMEs with high product prices, good financial conditions, but limited innovation and technology adoption, while cluster 1 has more affordable product prices, challenging financial conditions, but a higher rate of innovation and technology adoption.

On the other hand, in urban data, cluster 0 consists of cities with varying product prices, poor infrastructure conditions, low innovation levels, and minimal technology adoption, while cluster 1 consists of cities with higher product prices, better infrastructure, high innovation levels, and more significant technology utilization. The application of the Naive Bayes method to both datasets show that

using a 70:30 training data ratio results in the best prediction performance. This suggests a need for different policy approaches for rural and urban areas, focusing on innovation and technology in rural areas, and infrastructure improvement and technology adoption in urban areas.

Despite the promising results of the hybrid K-Mode and Naive Bayes approach in analyzing MSME sustainability, this study has several methodological limitations that warrant acknowledgment. First, no baseline model comparison was conducted with simpler classifiers such as logistic regression, decision trees (C4.5 or CART), or k-nearest neighbors (k-NN) on the same dataset. Consequently, it remains uncertain whether the observed accuracy (92.86% for the rural cluster) is statistically superior to less complex alternatives. Second, model validation relied exclusively on single data splits (90:10, 80:20, and 70:30) without employing cross-validation techniques (k-fold or repeated cross-validation). This approach increases the risk of overfitting to specific data partitions and limits the generalizability of the results. Future studies should therefore incorporate systematic benchmarking against multiple algorithms using identical evaluation metrics, along with statistical significance tests (McNemar's test or paired t-tests). Additionally, adopting repeated cross-validation would provide more robust performance estimates by ensuring that every data point contributes to both training and testing phases.

Third, the dataset is confined to 550 MSMEs from a single regency (North Central Timor), which restricts the generalizability of the findings to other regions with differing socio-economic and infrastructural characteristics. Future research should expand the geographical scope and integrate external predictor variables, including access to credit, government support programs, and digital infrastructure availability. Moreover, longitudinal studies that track changes in MSME sustainability status over time would be particularly valuable for understanding the dynamics of technology adoption and innovation, especially in rural contexts. By explicitly acknowledging these limitations, the authors intend for this study to serve as a replicable foundation for more robust decision-support systems and to encourage collaborative efforts between data mining researchers and MSME practitioners in underserved regions.

CONCLUSION

This study successfully addressed the question of the factors influencing the sustainability of MSMEs in rural and urban areas, and identified differences in the characteristics of MSMEs in each region. In rural areas, MSMEs tend to have higher product prices and better financial conditions but are less innovative and tend not to utilize technology, whereas in urban areas, MSMEs are more innovative and use technology more frequently despite facing challenges in terms of infrastructure and financial conditions. The application of the K-Means method for clustering and Naive Bayes for classification showed that a 70:30 training ratio yielded the best accuracy in predicting SMEs sustainability. This study opens opportunities for more in-depth studies, such as analyzing external factors that influence SMEs competitiveness and the application of other machine learning methods [30], Furthermore, longitudinal research can be conducted to monitor changes in SMEs sustainability over time and investigate barriers to technology adoption in rural areas.

REFERENCE

[1] R. R. Bakrie, S. Atikah Suri, S. A. Nabila, V. H Pratama, and Firmansyah., "Pengaruh Kreativitas UMKM Serta Kontribusinya Di Era Digitalisasi Terhadap Perekonomian Indonesia," *Jurnal Ekonomi Dan Bisnis*, vol. 16, no. 2, pp. 82–88, Jul. 2024, doi: 10.55049/JEB.V16I2.308.

[2] S. Widaningsih, W. Muhamad, R. Hendriyanto, and H. Nugroho, "An ID3 Decision Tree Algorithm-Based Model for Predicting Student Performance Using Comprehensive Student Selection Data at Telkom University," *Ingénierie des Systèmes d'Information*, vol. 28, no. 5, pp. 1205–1212, 2023, doi: 10.18280/isi.280508.

[3] M. D. Pangastuti and F. W. Nalle, "Determinan Kinerja Pelaku Usaha Kecil Menengah (UMKM) Pasar Perbatasan Kabupaten Timor Tengah Utara–Timor Leste," *Ekonomi dan Bisnis*, vol. 11, no. 2, pp. 109–134, 2024, doi: 10.35590/jeb.v11i2.7334.

[4] E. Aminullah et al., "Interactive Components of Digital MSMEs Ecosystem for Inclusive Digital Economy in Indonesia," *Journal of the Knowledge Economy*, vol. 15, pp. 487–517, 2024, doi: 10.1007/s13132-022-01086-8.

[5] L. Anatan and Nur, "Micro, Small, and Medium Enterprises' Readiness for Digital Transformation in Indonesia," *Economies*, vol. 11, no. 6, p. 156, 2023, doi: 10.3390/economies11060156.

[6] F. Faiz, V. Le, and E. K. Masli, "Determinants of Digital Technology Adoption in

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8271

Innovative SMEs," *Journal of Innovation & Knowledge*, vol. 9, no. 4, p. 100610, 2024, doi: 10.1016/j.jik.2024.100610.

[7] T. Yuwono, A. Suroso, and W. Novandari, "Information and Communication Technology in SMEs: A Systematic Literature Review," *Journal of Innovation and Entrepreneurship*, vol. 13, art. 31, 2024, doi: 10.1186/s13731-024-00392-6.

[8] S. Chanmee and K. Kesorn, "Semantic Decision Trees: A New Learning System for The ID3-Based Algorithm using a Knowledge Base," *Advanced engineering informatics*, vol. 58, p. 102156, 2023, doi: 10.1016/j.aei.2023.102156.

[9] K. J. T. Seran, D. Chrisinta, and Y. O. L. Rema, "Sistem Pendukung Keputusan Keberlanjutan UMKM Menggunakan Algoritma ID3 di Wilayah Perbatasan RI-RDTL," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 5, no. 1, pp. 41–53, 2025, doi: 10.51454/decode.v5i1.946.

[10] Terttiaavini, "A Hybrid Approach Using K-Means Clustering and the SAW Method for Evaluating and Determining the Priority of SMEs in Palembang City," *Journal of Intelligent System and Computation*, vol. 6, no. 1, pp. 46–53, 2024, doi: 10.52985/insyst.v6i1.392.

[11] Y. Kustiyahningsih, E. Rahmanita, A. Khozaimi, Y. D. P. Negara, B. K. Khotimah, and J. Purnama, "SCM-SCOR and K-Means Approaches for Clustering MSME Batik Bangkalan Madura Indonesia," *AIP Conference Proceedings*, vol. 3250, p. 050012, 2025, doi: 10.1063/5.0240736.

[12] D. E. Putri and E. P. W. Mandala, "Hybrid Data Mining berdasarkan Klasterisasi Produk untuk Klasifikasi Penjualan," *Jurnal KomtekInfo*, vol. 9, no. 2, pp. 68–73, Jun. 2022, doi: 10.35134/KOMTEKINFO.V9I2.279.

[13] G. Dwilestari and T. A. Afifah, "Perbandingan Kinerja Algoritma Naive Bayes Dan Decision Tree Dalam Klasifikasi Kanker Paru-Paru," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, pp. 801–807, 2025, doi: 10.36040/jati.v9i1.12463.

[14] M. Z. Haq, C. S. Octiva, A. Ayuliana, U. W. Nuryanto, and D. Suryadi, "Algoritma Naïve Bayes untuk Mengidentifikasi Hoaks di Media Sosial," *Jurnal Minfo Polgan*, vol. 13, no. 1, pp. 1079–1084, Jul. 2024, doi: 10.33395/JMP.V13I1.13937.

[15] A. Holl and R. Rama, "Spatial Patterns and Drivers of SME Digitalisation," *Journal of the Knowledge Economy*, vol. 15, pp. 5625–5649, 2024, doi: 10.1007/s13132-023-01257-1.

[16] V. Çetin and O. Yıldız, "A Comprehensive Review on Data Preprocessing Techniques in Data Analysis," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.

[17] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, p. 257, 2025, doi: 10.3390/ai6100257.

[18] D. Chrisinta and J. Eduardo Simarmata, "Eksplorasi Teknik Web Scraping pada Data Mining: Pendekatan Pencarian Data Berbasis Python," *Faktor Exacta*, vol. 17, no. 1, pp. 1979–276, 2024, doi: 10.30998/faktorexacta.v17i1.22393.

[19] O. Peretz, M. Koren, and O. Koren, "Naive Bayes Classifier—An Ensemble Procedure for Recall and Precision Enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108972, 2024, doi: 10.1016/j.engappai.2024.108972.

[20] X. Ye, Y. Zhu, S. Zhang, H. Deng, and P. Yu, "Non-Interactive K-Mode Clustering of High-Dimensional Categorical Data under Local Differential Privacy," *Inf. Sci. (N. Y).*, vol. 718, p. 122417, 2025, doi: 10.1016/j.ins.2025.122417.

[21] B. Supri, B. Mawadah, and H. Ali, "Asian Stock Index Price Prediction Analysis Using Comparison of Split Data Training and Data Testing," *JEMSI (Jurnal Ekonomi, Manajemen, Dan Akuntansi)*, vol. 9, no. 4, pp. 1403–1408, Aug. 2023, doi: 10.35870/JEMSI.V9I4.1339.

[22] I. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," in *7th International Mardin Artuklu Scientific Researches Conference*, 2022, pp. 496–504.

[23] S. S. Prasetiyowati and Y. Sibaroni, "Unlocking the Potential of Naive Bayes for Spatio Temporal Classification: A Novel Approach to Feature Expansion," *Journal of Big Data*, vol. 11, art. 106, 2024, doi: 10.1186/s40537-024-00958-x.

[24] D. Chrisinta, J. S.-K. J. S. Komputer, and undefined 2023, "Analisis Sentimen Penilaian Masyarakat Terhadap Pejabat Publik Menggunakan Algoritma Naïve Bayes Classifier," *ojs.unikom.ac.idD Chrisinta, JE SimarmataKomputika: Jurnal Sistem*

Komputer, 2023•ojs.unikom.ac.id, vol. 12, no. 1, p. 2020, 2023, doi: 10.34010/komputika.v12i1.9638.

[25] J. E. Simarmata, D. Chrisinta, and M. Purnomo, "Implementation of K-Means Clustering to Human Development Indicators in East Nusa Tenggara," *Journal of Research in Mathematics Trends and Technology*, vol. 6, no. 2, pp. 46–56, Sep. 2024, doi: 10.32734/jormtt.v6i2.17066.

[26] A. Lal, A. Sharan, K. Sharma, A. Ram, D. K. Roy, and B. Datta, "Scrutinizing Different Predictive Modeling Validation Methodologies and Data-Partitioning Strategies: New Insights Using Groundwater Modeling Case Study," *Environmental Monitoring and Assessment*, vol. 196, art. 623, 2024, doi: 10.1007/s10661-024-12794-w.

[27] B. Supri, Rudianto, Abdurohim, Badriatul Mawadah, and Helmi Ali, "Asian Stock Index Price Prediction Analysis Using Comparison of Split Data Training and Data Testing," *JEMSI (Jurnal Ekonomi, Manajemen, dan Akuntansi)*, vol. 9, no. 4, pp. 1403–1408, 2023, doi: 10.35870/jemsi.v9i4.1339.

[28] D. El-Shahat et al., "Machine Learning and Deep Learning Models Based Grid Search Cross Validation for Short-Term Solar Irradiance Forecasting," *Journal of Big Data*, vol. 11, art. 134, 2024, doi: 10.1186/s40537-024-00991-w.

[29] T. Q. A'yuni, B. N. Febriati, L. I. Effendie, M. Muhajir, and R. Yotenka, "MSME Sales Clustering Based on Business Aid Distribution Priority Using K-Affinity Propagation," *Enthusiastic: International Journal of Applied Statistics and Data Science*, vol. 3, no. 1, pp. 111–124, 2023, doi: 10.20885/enthusiastic.vol3.iss1.art10.

[30] M. Conciatori, A. Valletta, and A. Segalini, "Improving the Quality Evaluation Process of Machine Learning Algorithms Applied to Landslide Time Series Analysis," *Computers & Geosciences*, vol. 184, p. 105531, 2024, doi: 10.1016/j.cageo.2024.105531.