Jurnal Mandiri IT
Vol. 14 No. 1, July (2025), pp. 86-96

ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Hybrid clustering and supervised learning model for digital MSME segmentation

Dona Marcelina[1], Terttiaavini[2]
[1]Information Systems Department, Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia
[2]Computer Science Department, Faculty of Computer Science, Universitas Indo Global Mandiri, Palembang, Indonesia

A R T I C L E I N F O

Article history:

Received Jul 1, 2025
Revised Jul 10, 2025
Accepted Jul 19, 2025

Keywords:

Digital Policy;
Hybrid Clustering;
MSMEs Segmentation;
Supervised Learning;
Unsupervised Learning.

ABSTRACT

Digitalization became a key factor in enhancing the competitiveness of Micro, Small, and Medium Enterprises (MSMEs). However, its implementation still faced several challenges, including low levels of technology adoption and inaccurate data segmentation. This study aimed to develop a hybrid approach by combining clustering techniques and supervised learning to conduct segmentation and prediction of MSMEs based on their level of digitalization. Four clustering algorithms were tested: K-Means, Agglomerative, Gaussian Mixture Model, and HDBSCAN. The evaluation results showed that HDBSCAN outperformed the other algorithms, achieving the highest Silhouette Score (0.3501), the lowest Davies-Bouldin Index (0.9557), and the highest Calinski-Harabasz Index (132.38). The segmentation process generated three distinct clusters: Cluster 0 (Traditional – low digitalization, small revenue), Cluster 1 (Semi-Digital – moderate technology adoption, medium revenue), and Cluster 2 (Fully Digital – high technology adoption, large revenue). These cluster results were then used as labels to train six classification algorithms. Among them, XGBoost and Neural Network delivered the best performance, reaching a prediction accuracy of 98.63%. The main contribution of this study was the development of an analytical framework for data-driven segmentation and prediction of MSMEs, providing more precise, targeted, and adaptive support for national digitalization strategies.

This is an open access article under the CC BY-NC license.

Corresponding Author:

Dona Marcelina,
Information Systems Department,
Universitas Indo Global Mandiri,
Sudirman Street No 629 KM 4 Palembang, 30129, Indonesia
Email: donamarcelina@uigm.ac.id

1. INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) played a vital role in the global economy, including in Indonesia, where they made a significant contribution to the Gross Domestic Product (GDP) and employment absorption (Baderi, 2024; Juwitasari, 2023). However, MSMEs often faced challenges in enhancing their competitiveness, especially in the rapidly evolving digitalization era. One of the main challenges was the uneven adoption of digital technology among MSMEs (Godwin et al., 2024). Some businesses had utilized payment platforms and operational management systems, while others still relied on traditional digital methods for marketing (Eliza et al., 2024). This disparity created a competitive gap that could hinder the overall growth of MSMEs.

The segmentation of MSMEs based on their level of digitalization and operational characteristics became essential for designing targeted interventions. Previous studies showed that data-driven segmentation could help identify groups of MSMEs with different needs and potentials

(Mardiana et al., 2024). However, traditional segmentation approaches often relied on simple descriptive analysis or manual classification, which were less accurate and difficult to implement on a large scale. Therefore, more advanced approaches, such as hybrid clustering and supervised learning, were needed to produce more dynamic and predictive segmentation (Baulkani et al., 2024).

Digitalization was believed to be the key to improving MSME productivity. However, several fundamental issues hindered the optimization of its implementation. First, many MSMEs did not understand how to utilize technology optimally, resulting in policy interventions that were often mistargeted (Bahrini & Qaffas, 2019). Second, MSME data were often fragmented and unstructured, making in-depth analysis difficult. Third, conventional segmentation methods such as k-means clustering or traditional demographic analysis often produced less accurate groupings because they did not account for non-linear relationships between variables (Avram et al., 2021; Terttiaavini et al., 2018).

Previous studies on MSME segmentation were generally limited to descriptive statistical approaches or clustering methods without integration with predictive models (Marcelina et al., 2023). In fact, the combination of unsupervised learning techniques (clustering) and supervised learning had the potential to improve segmentation accuracy and enable classification prediction on new data (Alloghani et al., 2020; Heryati et al., 2025). However, few studies have thoroughly explored this hybrid approach specifically for MSME digitalization segmentation and prediction.

Therefore, this study addresses this gap by developing a hybrid framework that leverages the strengths of clustering and machine learning for both segmentation and prediction of MSMEs based on their digitalization levels. Compared to previous approaches, this framework enhances the precision and practical utility of MSME segmentation, enabling more accurate classification on new datasets. This contributes scientifically by advancing methodological integration in MSME research and offers practical benefits by providing policymakers and stakeholders with actionable insights to better target digital transformation initiatives.

The selection of Palembang as the research location was based on its representative characteristics of medium-sized urban centers in Indonesia. As the capital city of South Sumatra Province, Palembang features a diverse MSME landscape that includes both traditional and digitally advanced enterprises across various sectors such as food and beverage, fashion, services, and retail. This diversity allows for the observation of different levels of digital technology adoption. Additionally, Palembang has been involved in several national digital economy programs, making it a suitable microcosm for understanding broader digitalization trends and challenges in the Indonesian MSME context.

This study aimed to: 1) Identify MSME segmentation patterns based on operational characteristics and levels of digitalization using clustering algorithms such as K-Means, Agglomerative Clustering, Gaussian Mixture Model (GMM), and HDBSCAN; 2) Evaluate the performance of each algorithm using metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to determine the most optimal method; 3) Develop supervised learning models including XGBoost, Neural Network, Random Forest, SVM, KNN, and Logistic Regression to predict the segmentation of new MSMEs based on the clustering results; 4) Analyze the dominant factors influencing MSME digitalization levels through feature importance analysis. The results were expected to provide data-driven strategic recommendations for stakeholders in accelerating MSME digital transformation.

To ensure the reliability of clustering labels used in supervised learning, internal validation was conducted using clustering evaluation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Furthermore, the cluster results were analyzed based on their operational and digitalization profiles to ensure that each segment reflected distinct and meaningful MSME characteristics. This study made a significant contribution to the development of scientific knowledge and digital transformation practices in the MSME sector through several key aspects. The researchers produced a hybrid framework that integrated unsupervised learning (clustering) with supervised learning to analyze and predict MSME segmentation more accurately and adaptively (Marcelina et al., 2023; Terttiaavini, 2024). This approach not only enabled grouping based on similarities in operational characteristics and digitalization levels, but also expanded the analytical benefits toward the automatic prediction and classification of new MSME data, thereby supporting real-time data-driven decision-making.

Another key contribution lay in accelerating the digital transformation of the national MSME sector. This study presented an artificial intelligence-based approach that could be used by the government, MSME support institutions, and technology industry players to design more precise and impactful interventions. By leveraging segmentation and prediction results, stakeholders could identify MSMEs with potential for advanced digitalization and design mentoring or funding programs that matched each MSME's digital profile. Therefore, this study not only provided academic contributions but also offered practical value in strengthening the national MSME digital ecosystem and supporting Indonesia's digital economic transformation agenda.

This study used Google Colab as a cloud-based computing platform to perform data analysis, apply machine learning algorithms, and evaluate models efficiently and collaboratively

2. RESEARCH METHOD

Research Approach

This study employed a quantitative approach with an exploratory data analysis method to identify MSME segmentation patterns based on operational characteristics and levels of digitalization. The techniques applied were clustering using algorithms such as KMeans, Agglomerative, Gaussian Mixture, and HDBSCAN, which were capable of grouping unlabeled data based on the similarity of features of each MSME entity.

Subsequently, model testing was conducted using supervised learning methods to evaluate the classification ability based on the segmentation labels obtained from the clustering results. This stage aimed to validate and predict the segmentation of MSMEs on new data by utilizing classification algorithms such as XGBoost, Neural Network, Random Forest, SVM, KNN, and Logistic Regression. To ensure the internal validity of the clustering labels used in the supervised model, validation was conducted using clustering evaluation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to assess the cohesion and separation of clusters. Additionally, construct validity was supported by interpreting the feature distributions and digitalization profiles within each cluster to ensure meaningful and distinguishable segmentation patterns.

Data and Data Sources

Primary data were obtained through the distribution of online questionnaires to MSME actors operating in the Palembang City area, in coordination and with support from the Office of Cooperatives and MSMEs of Palembang City. The data collection process aimed to obtain direct information related to the operational characteristics and levels of digitalization of MSMEs, ensuring that the data acquired were relevant and representative for further analysis in this study. The total number of MSMEs involved in this research was 362.

The questionnaire instrument was developed based on previous studies and relevant literature on MSME digitalization. To ensure its validity, content validation was conducted by experts in the field of digital economy and MSME development. Furthermore, a pilot test involving 30 MSME respondents was performed to assess the reliability of the instrument, resulting in a Cronbach's alpha coefficient of 0.85, indicating good internal consistency. These steps ensured that the questionnaire was a valid and reliable tool for measuring MSME digitalization levels.

Research Variables

The research variables were divided into two main categories: operational variables and digitalization variables. Operational variables covered demographic aspects, business characteristics, and the financial performance of MSMEs, which reflected the overall condition and business activities. The features included in the operational variables consisted of *Education_Level*, *Marital_Status*, *Business_Ownership_Status*, *Subdistrict*, *Business_Scale*, *Business_Type*, *Monthly_Revenue*, *Operating_Costs*, *Profit*, *Avg_Monthly_Production*, and *Products_Sold_Per_Month*. In contrast, the digitalization variables represented the level of digital technology utilization in MSME operational processes, measured through the *Digitalization_Score*. This categorization was essential to comprehensively describe the influence of both traditional and digital aspects on MSME performance in the study.

Exploratory Data Analysis

The data exploration stage was conducted to gain an initial understanding of the distribution, characteristics, and relationships between variables in the MSME dataset used (Ren et

al., 2023). Descriptive statistics of the numerical variables were presented to provide a general overview of the mean, standard deviation, minimum value, quartiles, and maximum value of each analyzed attribute (Milo & Somech, 2020). Table 1 presented the results of the descriptive statistics of the data used in this study.

Table 1. Descriptive statistics results of the research data

Variabel	Mean	Std Dev	Min	25%	50%	75%	Max
Education_Level	3	1	2	2	3	5	5
Marital_Status	1	0	1	1	1	1	2
Business_Ownership_Status	1	1	1	1	1	1	3
Subdistrict	11	6	1	5	10	16	20
Business_Scale	1	0	1	1	1	1	3
Business_Type	3	3	0	1	2	4	15
Monthly_Revenue	6,139,268	7,515,395	40,000	1,637,500	4,000,000	8,000,000	75,000,000
Operating_Costs	2,859,655	7,762,152	0	500,000	1,000,000	3,000,000	130,000,000
Profit	3,144,558	4,440,955	0	1,000,000	2,000,000	4,000,000	50,000,000
Avg_Monthly_Production	9,050	157,728	1	10	30	300	3,000,000
Products_Sold_Per_Month	6,070	105,101	1	10	30	300	2,000,000
Digitalization_Score	4	1	2	3	4	4	4

The data identification process did not reveal the presence of outliers, missing values, or data inconsistencies, so all variables were used in full for the subsequent analysis stage.

Data Analysis Process

Data Preprocessing

The preprocessing stage began with checking for missing values across all attributes. The results showed that no missing values were present, so no imputation or data removal was required. Next, outlier detection was performed using statistical analysis and boxplot visualization. No significant outliers were found, so all data were used without modification (Çetin & Yıldız, 2022; Rashid & Waheed, 2020).

To ensure that the data were on a comparable scale, normalization was applied using the Min-Max Scaler method, which transformed the values of each feature to a range between 0 and 1. This step was important to improve the performance of the clustering and supervised learning algorithms used (Addanki et al., 2022; Safak, 2020).

The data used had highly varied value ranges among features. To address these scale differences and prevent certain features from dominating the modeling process, normalization was performed using the Min-Max Scaler method, with the formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where X was the original value of the feature; X_{min} was the minimum value of that feature in the dataset; X_{max} was the maximum value of that feature in the dataset; and X′ was the normalized feature value scaled between 0 and 1.

This method transformed each attribute value into the range of 0 to 1, ensuring that all features were on a comparable scale and supported the performance of machine learning algorithms sensitive to data scale.

Additionally, correlation analysis among features was conducted using heatmap visualization. This analysis helped to understand the relationships between variables and to identify potential redundancy or strong correlations among features, which could be considered during the modeling process (Gu, 2022). Figure 1 showed the correlation heatmap among the features.

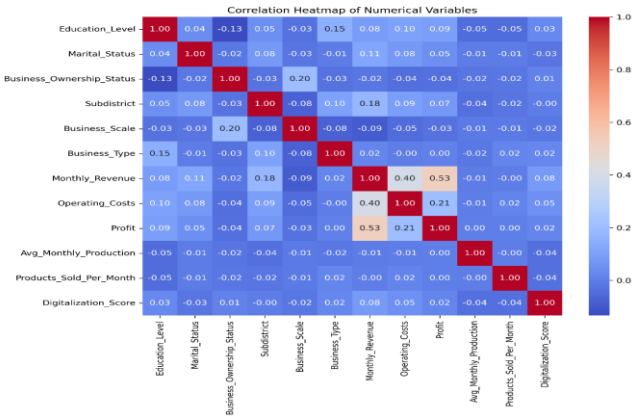

Figure 1. Correlation heatmap of numerical variables

The visualization results of the "Correlation Heatmap of Numerical Variables" showed that strong relationships between variables were indicated by correlation coefficient values close to +1.00 (strong positive correlation) or -1.00 (strong negative correlation). These correlations were visualized using a color gradient, where dark red represented a strong positive correlation and dark blue represented a strong negative correlation.

The variables Monthly_Revenue and OperatingCosts exhibited a very strong positive correlation, approaching a value of 1.00. This relationship indicated that an increase in monthly revenue was generally accompanied by a proportional increase in operating costs.

The variables Avg_Monthly_Production and Products_Sold_Per_Month also showed a very strong positive correlation. This condition reflected that the higher the average monthly production, the greater the number of products sold each month.

The correlation between Monthly_Revenue and Profit was at 0.53, which was categorized as a moderate to strong positive correlation. This indicated that an increase in monthly revenue tended to be directly proportional to an increase in profit.

Meanwhile, Operating_Costs and Profit showed a weaker correlation of 0.21. This value indicated that an increase in operating costs had only a small influence on profit increase.

Overall, correlation values close to +1.00 or -1.00 indicated a strong linear relationship between two variables. This information became an important basis for feature selection and further modeling processes.

Clustering Analysis

Clustering analysis was conducted to group the data into three clusters based on similarity in characteristics, allowing hidden patterns within the data to be better (Sinaga & Yang, 2020; Trento Oliveira et al., 2023). This study employed four clustering algorithms: K-Means, Agglomerative Clustering, Gaussian Mixture Model (GMM), and HDBSCAN, representing partitioning, hierarchical, probabilistic, and density-based approaches, respectively (Avram et al., 2021; Terttiaavini, 2024)

To determine the best algorithm, evaluations were performed using three metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The Silhouette Score assessed the compactness and separation of clusters, the Davies-Bouldin Index evaluated the closeness between clusters, and the Calinski-Harabasz Index measured the variance between and within clusters. The results from these three metrics were used as the basis for selecting the most optimal clustering method in the study. The evaluation metrics for each clustering algorithm were presented in Table 2.

Table 2. Evaluation results of clustering algorithms based on evaluation metrics

Algoritma Clustering	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.3179	1.3456	88.7972
Agglomerative Clustering	0.2810	1.6116	83.3579
Gaussian Mixture	0.2854	1.3639	78.8597
HDBSCAN	0.3501	0.9557	132.38

To test the best performance in determining the optimal number of clusters, the elbow method was applied. This method was used to evaluate the number of clusters based on the inertia value (in the K-Means algorithm), where the elbow point on the graph indicated the appropriate number of clusters specifically when the decrease in inertia values began to slow down significantly (Liu & Deng, 2021). This approach helped ensure that the selected number of clusters provided a balance between model complexity and the quality of data clustering. The graph of the elbow method results for the K-Means algorithm was presented in Figure 2.

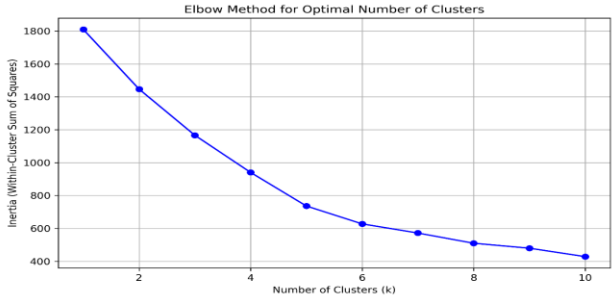

Figure 2. Graph of the elbow method calculation results on the k-means algorithm

The results of the elbow method in Figure 2 showed that the recommended number of clusters for this dataset was 3. This number of clusters provided the best balance between minimizing inertia (creating compact clusters) and preventing the model complexity from becoming excessive.

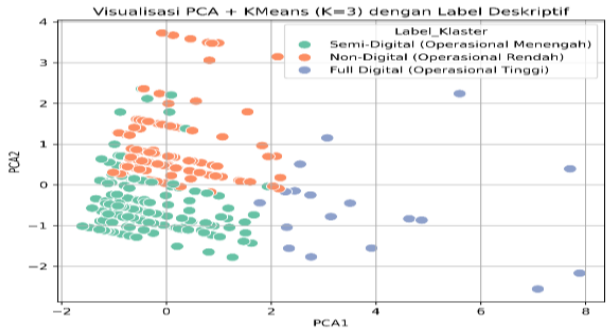

Figure 3. Presented the clustering results for three clusters using K-Means and UMAP

The clustering visualization was performed using the K-Means algorithm because this method allowed the explicit formation of three clusters in accordance with the research objectives. To facilitate the visual interpretation of high-dimensional data, the UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction technique was used. UMAP had advantages in visualizing non-linear data structures better than PCA or t-SNE. The visualization results with UMAP made it easier to interpret the distribution of each cluster and served as a basis for further evaluation using supervised learning algorithms. By using UMAP, the cluster distribution produced by the K-Means process was clearly visualized in a two-dimensional space, making it easier to analyze patterns and relationships among the data.

Prediction Model Testing and Evaluation

After the UMKM segmentation stage using clustering techniques, this study continued with the development and evaluation of predictive models to classify new UMKM into the previously identified digitalization segments. The main objective of this stage was to build accurate and generalizable supervised learning models that could predict UMKM cluster segments based on their respective characteristics. Thus, digitalization strategies could be targeted more efficiently and appropriately (Boateng et al., 2020).

At the initial training stage, several models showed very high prediction accuracy, even reaching 100% on the training data. This indicated the occurrence of overfitting, a condition where

the model adjusted too closely to the training data and lost its ability to accurately predict new data (Montesinos López et al., 2022). To address this issue, the Stratified Split technique was used in dividing the training and test data to ensure that the distribution proportions of each class (cluster segment) remained consistent across both data subsets. This step aimed to prevent bias and improve the model's representativeness and validity.

Additionally, hyperparameter tuning was performed on each model to optimize performance and minimize the risk of overfitting. During the testing and evaluation process, various leading machine learning classification algorithms were implemented, including Random Forest, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Neural Network (Khodabandehlou & Zivari Rahman, 2017). All models were trained using the training set and then evaluated on a separate test set.

Performance evaluation was conducted using accuracy metrics as well as classification reports, which included precision, recall, and F1-score values for each class. The classification model performance evaluation results were presented in Table 3 below.

Table 3. Classification model performance evaluation results

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	9.863	0.99	0.98	0.99
Neural Network	9.863	0.99	0.98	0.99
Random Forest	9.589	0.98	0.95	0.96
SVM	9.315	0.94	0.91	0.91
KNN	9.041	0.92	0.87	0.87
Logistic Regression	8.904	0.92	0.82	0.84

The classification model's prediction performance was further analyzed through the Confusion Matrix shown in Figure 3, which provided a comprehensive overview of the number of correct and incorrect predictions for each class (Susmaga, 2004).

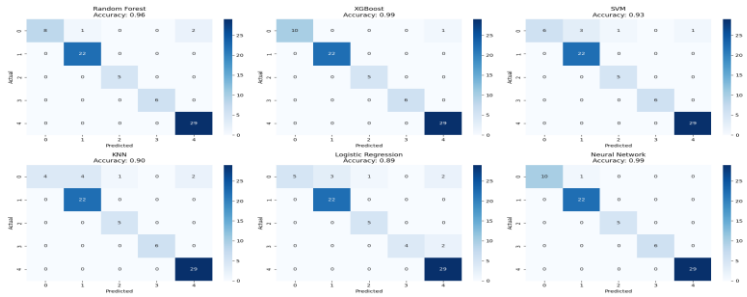

Figure 3. Confusion matrix

The Confusion Matrix presented a visual summary of the classification model's performance in predicting the digitalization segments of UMKM. The matrix clearly showed the number of correct and incorrect predictions for each class (segments 0 to 4). Observations indicated that the model performed very well, especially in classifying UMKM in segments 1, 2, and 3, where all predictions for these classes were accurate. However, minimal errors occurred in segment 0, where 1 out of 11 UMKM was misclassified as segment 1, and in segment 4, where 1 out of 29 UMKM was also misclassified as segment 1. Overall, this Confusion Matrix confirmed the model's high accuracy and strong ability to effectively distinguish between UMKM segments, with only a few misclassifications in some cases.

3. RESULTS AND DISCUSSIONS

Data Preprocessing

The data preprocessing stage was conducted to ensure optimal data quality before entering the modeling phase. The inspection of missing values showed that all attributes were complete, so no imputation or data removal was required. Outlier detection using descriptive statistics and boxplots did not find any significant outliers, therefore all data were used. Due to the wide variation in feature scales, normalization was performed using Min-Max Scaler to bring each feature into the range of 0–1, preventing dominance by certain variables. Additionally, correlation analysis between features was conducted to identify relationships and potential redundancies

among variables. These steps were important to ensure the optimal performance of the clustering and supervised learning algorithms applied.

Hybrid Clustering and Supervised Learning Stages

The clustering analysis was performed to group UMKM based on similarity characteristics, enabling hidden patterns in the data to be identified more clearly. Four clustering algorithms were used in this study: K-Means, Agglomerative Clustering, Gaussian Mixture Model (GMM), and HDBSCAN. The performance evaluation of the algorithms was carried out using three metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The evaluation results showed that the HDBSCAN algorithm had the best performance with the highest Silhouette Score of 0.3501, the lowest Davies-Bouldin Index of 0.9557, and the highest Calinski-Harabasz Index of 132.38. This indicated that the clusters produced by HDBSCAN were more compact, well separated, and had optimal between-cluster variance compared to other algorithms.

To determine the optimal number of clusters, the elbow method was applied to the K-Means algorithm. The elbow plot showed that the optimal number of clusters was three, as at that point the decrease in inertia started to slow down, thus achieving a good balance between model complexity and cluster quality. Based on this clustering result, three distinct UMKM segments were formed: Cluster 0 (blue), consisting of traditional UMKM with low levels of digitalization and relatively small revenue; Cluster 1 (orange), representing semi-digital UMKM with moderate revenue; and Cluster 2 (green), consisting of UMKM that have fully adopted digitalization with high revenue.

Among the three clusters, Cluster 2 was considered the best cluster because it showed full digitalization and high revenue, reflecting UMKM with better business performance and optimal technology adaptation. This cluster became the main target for the development and empowerment of digital UMKM due to its great potential in improving competitiveness and productivity. Meanwhile, Clusters 0 and 1 required special interventions to increase their digitalization levels and revenues in order to progress towards higher-performing clusters.

Supervised Learning for Predicting UMKM Sustainability

After the UMKM segmentation stage using clustering techniques, this study proceeded with the development and evaluation of predictive models to classify new UMKM into previously identified digitalization segments. The main objective of this stage was to build accurate supervised learning models capable of generalization, so they could predict the cluster segment of UMKM based on their characteristics. Thus, digitalization strategies could be targeted more efficiently and precisely.

During the initial training phase, several models showed very high prediction accuracy, even reaching 100% on the training data. This condition indicated the occurrence of overfitting, a critical phenomenon in machine learning where the model learns too specifically from noise and random patterns in the training data, thereby losing its ability to generalize and accurately predict unseen new data. To overcome this overfitting issue and ensure the models had valid predictive capability on real-world data, the Stratified Split technique was used in dividing the training and testing data. This technique ensured that the proportional distribution of each class (cluster segments produced from clustering) remained consistent across both subsets (training and testing). This step was crucial to prevent bias caused by class imbalance and significantly improved the representativeness and validity of the models in real conditions.

Additionally, hyperparameter tuning was conducted on each model to optimize performance and minimize the risk of overfitting. During the testing and evaluation process, several leading machine learning classification algorithms were implemented, including Random Forest, XGBoost, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and Neural Network. All models were trained using the training set and then evaluated on a separate test set. Performance evaluation was conducted using accuracy metrics as well as classification reports, which included precision, recall, and F1-score values for each class. The classification model performance evaluation results were presented in Table 2.

The evaluation results of the six supervised learning models on UMKM data showed that XGBoost and Neural Network were the best models. Both recorded the highest accuracy value of 0.9863, with precision and recall of 0.99 and 0.98 respectively, producing very high F1-scores (0.99). This demonstrated that both models were highly reliable in predicting with high accuracy

and very low classification errors. Random Forest ranked third, with an accuracy of 0.9589 and an F1-score of 0.96, which was still considered good but slightly lower than the top two models.

Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) showed moderate performance, with accuracies of 0.9315 and 0.9041 and F1-scores of 0.91 and 0.87 respectively. This indicated that both models began to lose precision and sensitivity in classification. Logistic Regression had the lowest performance among the tested models, with an accuracy of 0.8904 and an F1-score of 0.84, indicating its limitations in handling the higher complexity of UMKM data. Overall, XGBoost and Neural Network were recommended as the most optimal predictive models for segmentation and classification of UMKM based on digitalization in this study.

Feature Importance Analysis

The feature engineering process was conducted to identify and select the most relevant variables that significantly contributed to clustering analysis and prediction within the machine learning models (Kuhn & Johnson, 2019). This stage began with a correlation analysis among features using a correlation heatmap, which aimed to measure the degree of linear relationships between numerical variables. The analysis results showed that variables such as Monthly Revenue, Operating Costs, Average Monthly Production, Products Sold Per Month, and Profit had fairly strong positive correlations with each other. For example, Monthly Revenue exhibited a very high correlation with Operating Costs, while Average Monthly Production had a close relationship with the number of products sold per month. In addition to these variables, one additional feature representing the digitalization aspect, namely Digitalization_Score, was also selected.

These significant correlations formed the basis for feature selection to be included in the clustering and supervised learning processes. By using these features, the models were expected to identify more accurate patterns and produce more meaningful analyses in assessing the sustainability of UMKM businesses.

Integration of Clustering Results into Supervised Learning

The researchers integrated the clustering results as target labels in the supervised learning process to build classification models capable of predicting UMKM types based on input characteristics. The previous clustering produced three main clusters. These cluster labels were used as target variables in training several supervised learning algorithms such as XGBoost, Neural Network, Random Forest, SVM, KNN, and Logistic Regression. The models were trained using six important features. The objective was for the models to automatically classify new UMKM into the appropriate clusters.

The evaluation results showed that XGBoost and Neural Network delivered the best performance, achieving an accuracy of 98.63% as well as high precision, recall, and F1-score values (0.99). This demonstrated that integrating clustering results into supervised learning was very effective in building an accurate UMKM cluster prediction system. This approach supported data-driven decision-making in the more targeted development and empowerment of UMKM.

4. CONCLUSION

This study successfully developed a hybrid approach that integrated clustering techniques and supervised learning for the segmentation and prediction of UMKM based on their level of digitalization. Among the four clustering algorithms tested, HDBSCAN demonstrated the best performance, as indicated by the highest Silhouette Score (0.3501), the lowest Davies-Bouldin Index (0.9557), and the highest Calinski-Harabasz Index (132.38). HDBSCAN's strength lay in its ability to identify natural clusters without predefining the number of clusters, as well as its effectiveness in handling heterogeneous data.

The resulting segmentation successfully grouped UMKM into three main clusters: Cluster 0 (Traditional UMKM), characterized by low digitalization levels and small revenue; Cluster 1 (Semi-Digital UMKM), which showed partial technology adoption with moderate revenue; and Cluster 2 (Fully Digital UMKM), which had fully adopted digital technology and achieved high revenue. The integration of clustering results into supervised learning models produced highly accurate predictive models, with XGBoost and Neural Network achieving accuracy rates up to 98.63%. These findings demonstrated a strong capability to classify new UMKM based on their digitalization characteristics.

The practical contributions of this research included the development of a machine learning-based analytical framework to design more precise segmentation strategies and interventions. Furthermore, the segmentation results could serve as a foundation for formulating

targeted public policies and accelerating the inclusive and adaptive digital transformation of the UMKM sector. Future developments could focus on integrating real-time data, expanding geographic coverage, and developing automated recommendation systems to support the sustainability of the national UMKM digital ecosystem.

Despite the promising results, this study has several limitations. First, the dataset used was limited in terms of geographic coverage and sectoral diversity, which may affect the generalizability of the segmentation model across all types of MSMEs. Second, the absence of expert judgment in validating the clusters may have impacted the contextual interpretation of each segment. Third, the models were trained on static datasets, which may not capture the dynamic evolution of MSME digitalization over time. For future research, it is recommended to develop an automated system that integrates real-time data streams to continuously update MSME segmentation and predictions. This system could take the form of a digital dashboard or decision-support system for policymakers and stakeholders. Incorporating explainable AI techniques and integrating external data sources (such as social media or financial transactions) would also enhance the robustness and applicability of the system in real-world scenarios.

REFERENCES

Addanki, R., McGregor, A., Meliou, A., & Moumoulidou, Z. (2022). *Improved Approximation and Scalability for Fair Max-Min Diversification*. https://arxiv.org/pdf/2201.06678

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. 3–21. https://doi.org/10.1007/978-3-030-22475-2_1

Avram, A., Matei, O., Pintea, C.-M., Pop, P. C., & Anton, C. A. (2021). Comparative Analysis of Clustering Techniques for a Hybrid Model Implementation. In Á. Herrero, C. Cambra, D. Urda, J. JSedano, H. Quintián, & E. Corchado (Eds.), *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020)* (pp. 22–32). Springer, Cham. https://doi.org/10.1007/978-3-030-57802-2_3

Baderi, F. (2024, December 7). *UMKM Pilar Pemulihan dan Pertumbuhan Ekonomi Nasional*. Harian Ekonomi Neraca. https://www.neraca.co.id/article/209137/umkm-pilar-pemulihan-dan-pertumbuhan-ekonomi-nasional

Bahrini, R., & Qaffas, A. A. (2019). Impact of Information and Communication Technology on Economic Growth: Evidence from Developing Countries. *Economies 2019, Vol. 7, Page 21*, *7*(1), 21. https://doi.org/10.3390/economies7010021

Baulkani, S., Nifasath, P. S., & Priyanga, M. M. (2024). Machine Learning Technologies for Agricultural Prediction to Enhance Economic Growth. *Smart Technologies for Sustainable Development Goals*, 178–195. https://doi.org/10.1201/9781003519010-11

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, *08*(04), 341–357. https://doi.org/10.4236/jdaip.2020.84020

Çetin, V., & Yıldız, O. (2022). A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, *28*(2), 299–312. https://doi.org/10.5505/pajes.2021.62687

Eliza, Hadi, F., & Zefriyenn. (2024). Pengembangan E-Commerce di Era Digitalisasi pada UMKM Produk Kale Kota Padang Panjang. *Jurnal Pengabdian Kepada Masyarakat Nusantara*, *5*(2), 2732–2743. https://doi.org/10.55338/jpkmn.v5I2.3342

Godwin, G., Junaedi, S. R. P., Hardini, M., & Purnama, S. (2024). Inovasi Bisnis Digital untuk Mendorong Pertumbuhan UMKM melalui Teknologi dan Adaptasi Digital. *ADI Bisnis Digital Interdisiplin Jurnal*, *5*(2), 41–47. https://doi.org/10.34306/abdi.v5I2.1172

Gu, Z. (2022). Complex heatmap visualization. *IMeta*, *1*(3), e43. https://doi.org/https://doi.org/10.1002/imt2.43

Heryati, A., Terttiaavini, T., Cahyani, S., Romli, H., & Zaliman, I. (2025). Optimasi Strategi Pemasaran E-Commerce Melalui Prediksi Konversi Berbasis Machine Learning. *JSAI: Journal Scientific and Applied Informatics*, *8*(1), 66–73. https://doi.org/10.36085

Juwitasari, A. (2023, January 7). *Refleksi 2022 dan Outlook 2023, Kemenkop UKM Ungkap Pencapaian dan Rencana Untuk Pelaku UMKM*. https://ukmindonesia.id/baca-deskripsi-program/refleksi-2022-dan-outlook-2023-kemenkop-ukm-ungkap-pencapaian-dan-rencana-untuk-pelaku-umkm

Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, *19*(1/2), 65–93. https://doi.org/10.1108/jsit-10-2016-0061

Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection: A Practical Approach for Predictive Models. *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1–297. https://doi.org/10.1201/9781315108230

Liu, F., & Deng, Y. (2021). Determine the Number of Unknown Targets in Open World Based on Elbow Method. *IEEE Transactions on Fuzzy Systems*, *29*(5), 986–995. https://doi.org/10.1109/tfuzz.2020.2966182

Marcelina, D., Kurnia, A., & Terttiaavini, T. (2023). Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 293–301. https://doi.org/10.57152/malcom.v3i2.952

Mardiana, R., Fahdillah, Y., Kadar, M., Hassandi, I., & Mandasari, R. (2024). Implementasi Transformasi Digital dan Kecerdasan Buatan Sebagai Inovasi Untuk UMKM pada Era Revolusi Industri 4.0. *Jurnal Ilmiah Manajemen Dan Kewirausahaan (JUMANAGE)*, *3*(1). https://doi.org/10.51642/ppmj.v31i04.404

Milo, T., & Somech, A. (2020). Automating Exploratory Data Analysis via Machine Learning: An Overview. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2617–2622. https://doi.org/10.1145/3318464.3383126

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0

Rashid, J., & Waheed, K. (2020). Missing Values and Outliers in Research Data. *Pakistan Postgraduate Medical Journal*, *31*(04), 167–167. https://doi.org/10.51642/ppmj.v31i04.404

Ren, H., Khailany, B., Fojtik, M., & Zhang, Y. (2023). Machine Learning and Algorithms: Let Us Team Up for EDA. *IEEE Design and Test*, *40*(1), 70–76. https://doi.org/10.1109/mdat.2022.3143427

Safak, V. (2020). Min-Mid-Max Scaling, Limits of Agreement, and Agreement Score. *ArXiv*. https://arxiv.org/pdf/2006.12904

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/access.2020.2988796

Susmaga, R. (2004). Confusion Matrix Visualization. *Intelligent Information Processing and Web Mining*, 107–116. https://doi.org/10.1007/978-3-540-39985-8_12

Terttiaavini, T. (2024). A Hybrid Approach Using K-Means Clustering and the SAW Method for Evaluating and Determining the Priority of SMEs in Palembang City. *INSYST: Journal of Intelligent System and Computation*, *6*(1), 46–53. https://doi.org/10.52985/insyst.V6i1.392

Terttiaavini, T., Zamzam, F., Ramadhan, M., K. Rosni, A., Setiawan Saputra, T., Heryati, A., & Dhamayanti. (2018). Clustering Analysis of Premier Research Fields. *International Journal of Engineering & Technology*, *7*(4.44). https://doi.org/10.14419/ijet.v7i4.44.26860

Trento Oliveira, L., Kuffer, M., Schwarz, N., & Pedrassoli, J. C. (2023). Capturing deprived areas using unsupervised machine learning and open data: a case study in São Paulo, Brazil. *European Journal of Remote Sensing*, *56*(1). https://doi.org/10.1080/22797254.2023.2214690