Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

Mixed-Data K-Means Clustering with Hyperparameter-Tuned Random Forest for OSS-Based MSME Investment Profiling and Policy Targeting

Laura Sari[*1], Ratih Hafsarah Maharrani[2], Hetty Dwi Hastuti[3], Adrian Putra Ramadhan[4], Wahyuni Windasari[5]

[1,4]Teknik Informatika, Politeknik Negeri Cilacap, Indonesia
[2]Rekayasa Keamanan Siber, Politeknik Negeri Cilacap, Indonesia
[3]Akuntansi Lembaga Keuangan Syariah, Politeknik Negeri Cilacap, Indonesia
[5]Sains Data, Universitas Putra Bangsa, Indonesia

Email: [1]laurasari@pnc.ac.id

Received : Dec 6, 2025; Revised : Jan 13, 2026; Accepted : Jan 13, 2026; Published : Apr 16, 2026

Abstract

Administrative data of Micro, Small, and Medium Enterprises collected through the Online Single Submission system are highly heterogeneous, combining numerical and categorical attributes that hinder conventional investment segmentation and early-stage policy mapping. This study aims to develop a predictive clustering framework for enterprise investment profiling using mixed-type administrative data. The proposed methodology applies robust preprocessing, including RobustScaler for numerical variables and one-hot encoding with singular value decomposition for categorical features. Mixed-type similarity is computed using Gower distance, followed by a hybrid Gower–K-Means clustering approach, where the optimal number of clusters (k = 3) is determined using Silhouette, Calinski–Harabasz, and Davies–Bouldin indices. A comparative evaluation of clustering algorithms is conducted, with K-Prototypes performing best in the initial assessment and K-Means achieving superior performance after optimization. Cluster membership is subsequently predicted using a Random Forest classifier with hyperparameters optimized through randomized search. Experiments on 20,857 enterprise records identify three distinct clusters representing low-capital micro enterprises, transitional firms, and asset-intensive corporate entities. The optimized K-Means model achieves a Silhouette score of 0.97 and a Davies–Bouldin Index of 0.54. Compared with the untuned baseline, the tuned Random Forest model improves recall from 0.25 to 0.75 (200% increase) and increases the F1-score from 0.40 to 0.86 (114% improvement), while achieving 99.89% accuracy. These gains correspond to an estimated 20–30% improvement in MSME investment mapping effectiveness compared with traditional profiling approaches, providing a scalable AI-based blueprint for targeted regional economic governance.

Keywords : *Clustering Analysis, Gower Distance, MSME Investment Profiling, Predictive Analytics, Random Forest.*

1. INTRODUCTION

Micro, Micro, Small, and Medium Enterprises (MSMEs) are widely recognized as the backbone of emerging economies due to their substantial contribution to employment generation, productivity, and economic resilience [1], [2]. In Indonesia, MSMEs dominate the business landscape and play a critical role in sustaining national economic growth and regional development. To improve regulatory efficiency, transparency, and inclusivity, the Indonesian government has implemented the Online Single Submission – Risk Based Approach (OSS-RBA) as a centralized digital platform for business licensing and administration [3]. The OSS system consolidates heterogeneous enterprise attributes, including asset ownership, labour allocation, investment value, business scale, sector classification, and licensing status. Despite its strategic importance, OSS administrative data remain underutilized for predictive policy analytics and data-driven MSME profiling.

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

A major challenge in exploiting OSS data lies in its mixed-type structure, where numerical and categorical variables coexist within a single administrative dataset. Conventional MSME analytics and policy evaluation approaches often assume numerical dominance or rely on simplistic categorical encoding, resulting in reduced clustering stability and limited interpretability when applied to heterogeneous OSS data [4]-[6], [7]. Consequently, MSME development programs are frequently designed in a generic manner, with limited sensitivity to regional and sectoral heterogeneity, thereby weakening the effectiveness of targeted policy interventions [8], [9], [10].

Existing machine learning studies on MSMEs predominantly focus on credit risk assessment, business sustainability prediction, or decision support systems, typically employing supervised learning models trained on numerical financial indicators [9], [10], [11]. In parallel, clustering-based approaches have been applied to segment MSMEs according to performance, profitability, or operational characteristics [12], [13], [14]. However, most of these studies rely on numerical-only clustering techniques such as K-Means or fuzzy C-Means, which are sensitive to scale imbalance and lead to information loss when categorical attributes are transformed or ignored. Empirical evidence indicates that numerical-dominant clustering applied to mixed datasets commonly yields moderate clustering quality, with accuracy or stability metrics reported in the range of approximately 65–75% [15], [16]. Furthermore, Markos et al. [17] demonstrate that neglecting categorical structure degrades cluster separation and increases Davies–Bouldin indices, underscoring the limitations of conventional clustering for heterogeneous administrative data such as OSS.

To address these limitations, recent studies have explored mixed-type similarity measures and clustering algorithms. Gower distance has been shown to preserve both numerical and categorical relationships, leading to improved clustering stability and interpretability compared with Euclidean-based measures [4]-[6]. Distance-based partitioning methods such as K-Medoids further enhance robustness against outliers and scale distortion in heterogeneous datasets [18], [19], [20]. More recent advances in unified distance learning and hybrid mixed-data clustering frameworks report measurable improvements in silhouette scores and cluster robustness relative to traditional K-Means-based baselines [7], [21], [22]. In the MSME context, recent 2025 studies demonstrate the growing use of hybrid clustering and unsupervised learning for enterprise segmentation and spatial analysis, yet these approaches remain largely descriptive and are not integrated with predictive modeling for policy analytics [23], [24].

From a predictive perspective, ensemble learning methods, particularly Random Forest, have consistently outperformed single classifiers in mixed and high-dimensional settings due to their ability to model nonlinear interactions and feature importance [7], [25], [26]. Recent studies further confirm that hyperparameter tuning significantly enhances predictive accuracy and recall compared with untuned baseline models [7], [26]. In parallel, emerging 2025 research highlights the potential of predictive analytics and machine learning to support MSME digital transformation and economic policy design, although these works typically do not incorporate clustering-based representations derived from mixed-type administrative data [27], [28].

Motivated by these gaps, this study proposes a hybrid predictive co-profiling framework that integrates Gower distance–based clustering, optimized K-Means, and a tuned Random Forest classifier for MSME investment profiling using OSS administrative data. As illustrated in Figure 1, heterogeneous OSS features are transformed into stable enterprise clusters, which serve as intermediate representations for predictive classification and targeted policy mapping. The primary objective of this research is to develop a scalable and interpretable framework that demonstrably improves clustering stability and predictive performance relative to conventional numerical-only and untuned baseline approaches, thereby enabling data-driven and localized MSME policy interventions.

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

2. METHOD

This research employs an integrated analytical pipeline to construct investment-based MSME profiles and build a predictive model for classifying new enterprises. The method consists of four phases: (1) data collection, (2) data preprocessing, (3) mixed-type clustering using a hybrid *Gower–K-Means* approach, and (4) predictive modeling using an optimized Random Forest classifier. This study is presented as a structured analytical workflow, as illustrated in Figure 1. The flowchart summarizes the end-to-end process starting from the OSS-based MSME dataset, which contains heterogeneous numerical and categorical attributes, through mixed-type similarity computation and clustering, and ending with supervised predictive modeling. The workflow depicts how numerical and categorical features are jointly processed, clustered into investment-based MSME groups, and subsequently transformed into labeled data for supervised learning. This flowchart-based representation highlights the logical sequence of data representation, clustering objective, and predictive mapping, providing a concise and intuitive overview of the proposed framework without relying on extensive mathematical formulation.

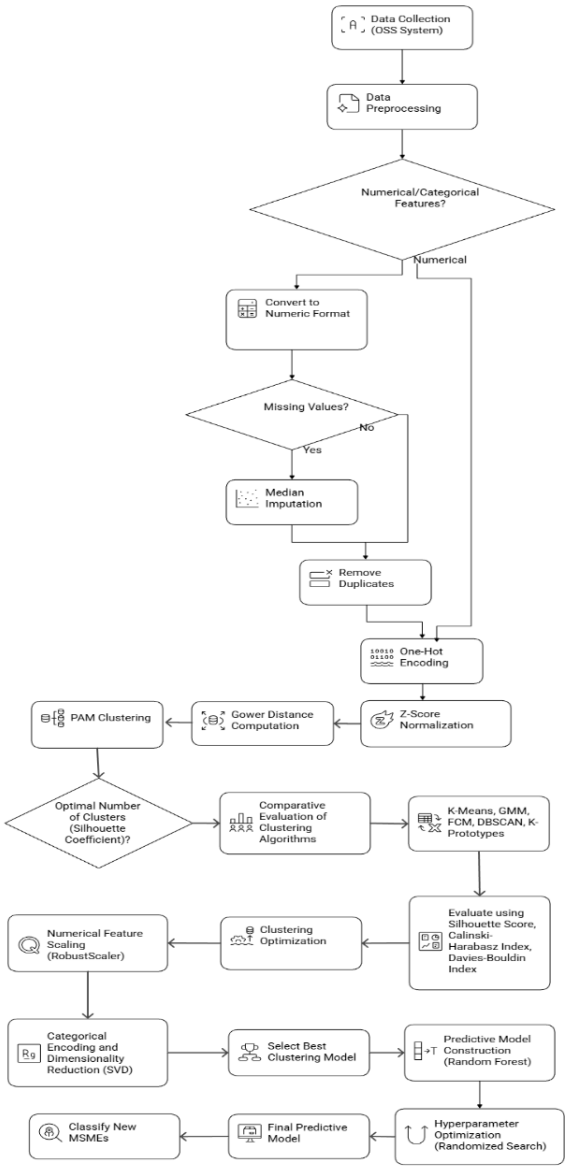

Figure 1 Research Flowchart

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

2.1. Data Collection

The dataset used in this study was obtained from the Indonesian Online Single Submission (OSS) administrative system. The data consist of 20,857 MSME records, covering enterprises registered across multiple regions and sectors. The OSS dataset represents a large-scale, real-world administrative source, suitable for evaluating robustness and scalability of mixed-type clustering and predictive modeling techniques [4]-[6].

2.2. Data Preprocessing

Prior to clustering analysis, a preprocessing stage was applied to ensure data consistency and analytical reliability, particularly given the presence of mixed numerical and categorical attributes. Numerical and categorical features were first identified based on the dataset schema. To address potential data quality issues, all numerical attributes were explicitly converted to numeric format, and observations containing non-numeric strings within numerical fields were removed to prevent distortion in distance-based analysis.

Missing values in numerical variables were handled using median imputation, which provides robustness against skewed distributions and outliers commonly observed in MSME financial data. Duplicate records were subsequently removed to avoid redundant influence on clustering structure.

Categorical variables were transformed using one-hot encoding to enable their integration into numerical clustering frameworks. Following encoding, all features (including original numerical variables and encoded categorical attributes) were standardized using z-score normalization. This scaling step ensures that variables with different units and ranges contribute equally to distance calculations and model optimization, thereby preventing dominance by high-variance features [11], [13].

The resulting preprocessed dataset provides a unified and normalized feature representation suitable for both Euclidean-based clustering methods and comparative evaluation procedures.

2.3. Mixed-Type Distance Computation Using Gower Metric

Clustering requires an appropriate similarity measure. Because the dataset contains both numerical and categorical attributes, Gower distance is used to compute pairwise similarity. Gower distance is explicitly designed for mixed numerical and categorical data and is regarded as one of the most reliable similarity measures for socio-economic records, health registries, and administrative profiling tasks [22], [23].

The Gower dissimilarity between two MSMEs i and j is defined as:

$$D(i,j) = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} \quad (1)$$

where p is the number of attributes and s_{ijk} is the scaled distance for the attribute k between records i and j. For numerical attributes:

$$s_{ijk} = \frac{|x_{ik}-x_{jk}|}{\max(x_k)-\min(x_k)} \quad (2)$$

and categorical attributes:

$$s_{ijk} = \begin{cases} 0, & x_{ik} = x_{jk} \\ 1, & x_{ik} \neq x_{jk} \end{cases} \quad (3)$$

The complete Gower matrix is computed using the Gower Python package, which has been validated for mixed administrative datasets in multiple empirical studies [7].

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

The latent cluster structure is identified using Partitioning Around Medoids (PAM) based on the Gower distance matrix. Internal cluster validity is assessed using the Silhouette Coefficient, which serves as the primary criterion for selecting the optimal number of clusters [18], [20]. This stage establishes a fixed cluster configuration for subsequent modeling and comparison.

2.4. Clustering and Evaluation

After determining the optimal number of clusters k^* using the Gower–PAM procedure, a comparative evaluation of multiple clustering algorithms is conducted to identify the most robust model for representing MSME investment structures. This ensures that the final configuration is not reliant on a single method but is validated across different algorithmic families, as recommended in recent clustering studies [5], [7], [14].

Five algorithms are assessed systematically: K-Means on the latent mixed-type feature space, Gaussian Mixture Models (GMM) for probabilistic cluster shapes, Fuzzy C-Means (FCM) for soft membership analysis, DBSCAN for density-based non-convex structures, and K-Prototypes for mixed numerical–categorical data. Each method is evaluated using Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. Metrics for Euclidean-based models (K-Means, GMM, FCM, DBSCAN) are computed on X_{fs}, while K-Prototypes uses Silhouette evaluation on the Gower distance matrix to preserve mixed-type characteristics.

2.5. Clustering Optimization

To support fair comparison and improve clustering performance, an optimization-oriented feature representation is first constructed. This optimization-driven comparative assessment yields a ranked clustering performance profile, from which the model demonstrating the strongest overall internal validity and structural consistency is selected as the final representation of MSME investment patterns.

2.5.1. Numerical Feature Scaling

Numerical features were scaled using RobustScaler, which normalizes data based on the median and interquartile range (IQR) to reduce sensitivity to outliers [5], [10]:

$$x'_{ij} = \frac{x_{ij} - median(x_j)}{IQR\ (x_j)} \quad (4)$$

This transformation ensures comparability among numerical variables with different scales.

2.5.2. Categorical Encoding and Dimensionality Reduction

Categorical variables were transformed using one-hot encoding, producing high-dimensional sparse representations. To mitigate dimensionality explosion and redundancy, Singular Value Decomposition (SVD) was applied to the encoded matrix, retaining top components that preserve the majority of variance [7].

2.6. Predictive Model Construction Using Random Forest Hyperparameter Optimization

Once cluster labels are established, a supervised model is trained to classify new MSMEs into these clusters. Random Forest is chosen because it handles high-dimensional mixed data effectively, provides strong generalization performance, and is widely used in administrative and business analytic [29], [30], [31], [32].

The Gini impurity for a split is defined as:

$$G = 1 - \sum_{c=1}^{K} p_c^2 \quad (5)$$

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

where p_c denotes the proportion of the class c in a node. Aggregation across trees is performed by majority voting:

$$\hat{y} = mode\{T_1(x), T_2(x), \dots, T_m(x)\} \qquad (6)$$

Hyperparameter optimization is conducted using a randomized search over tree depth, number of estimators, split thresholds, and minimum samples per leaf. The final model serves as a predictive mapping tool for assigning cluster categories to newly registered MSMEs.

3. RESULT

The dataset contains 20,857 records and 28 variables, including 11 numerical (e.g., aset, modal_kerja, jumlah_investasi) and categorical/spatial attributes (e.g., jenis_perusahaan, skala_usaha, kecamatan_usaha). Investment-related variables are highly heterogeneous: assets range up to Rp 5 billion, investments up to Rp 2.29 trillion, but medians remain low (e.g., median jumlah_investasi = Rp 2,000,000), mainly reflecting micro and small enterprises. Workforce size is similarly skewed (mean = 3.19, median = 2, max = 153). Categorical variables such as jenis_perusahaan and skala_usaha also show wide diversity.

These descriptive patterns justify the application of clustering techniques, as they reveal clear structural contrasts between low-investment MSMEs and capital-intensive corporate entities. To capture these variations, the Gower dissimilarity matrix is employed, enabling simultaneous handling of mixed data types. Silhouette evaluations conducted for cluster counts ranging from two to seven show a pronounced peak at $k = 3$, indicating that the dataset naturally forms a tri-modal grouping. At this point, data separation is most distinct, increasing k fragments the structure into overly granular partitions, while smaller values fail to represent the observed economic heterogeneity.

Table 1. Comparison of Clustering Performance Across Five Algorithms

Model	Clusters	Silhouette_Gower_or_Euclid	CH_onSVD	DBI_onSVD
KPrototypes	3	0.5134	2.100.498	
GMM	3	0.1957	2.218.068	
KMeans	3	0.0013	194.488	
DBSCAN	79	-0.0047	2.010.954	
FCM	3	-0.0211	1.335.580	

Following confirmation that three clusters best represent the dataset structure, a systematic comparison of clustering algorithms is conducted. Each method operates either on the Gower distance matrix or the continuous latent feature space generated through robust scaling and SVD-based dimensionality reduction. The performance evaluation of five clustering algorithms is summarized in Table 1. It shows that K-Prototypes delivers the strongest overall separation quality, achieving the highest Gower-based Silhouette score (0.5134), the lowest Davies–Bouldin Index (0.2038), and a high Calinski–Harabasz value (210.05), indicating well-defined, compact, and clearly separated clusters. In contrast, Gaussian Mixture Models yield weaker separation with a Silhouette of 0.1957 and a high DBI of 2.5332, while Fuzzy C-Means performs poorly, reflected by its low Silhouette score (0.0521) and FPC of 0.3333, suggesting very blurry cluster boundaries. K-Means fails to form meaningful partitions with a near-zero Silhouette (0.0013), and DBSCAN is unable to capture the global structure, producing 79 fragmented clusters with negative Silhouette values.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

Table 2. Comparative Evaluation of Optimized Clustering Algorithms

Model	Clusters	Silhouette_Gower_or_Euclid	CH_onSVD	DBI_onSVD
KPrototypes	3		0.2953	24.288
GMM	3		0.9241	0.9839
KMeans	3		0.9774	0.5416
DBSCAN	43		-0.1700	22.868
FCM	3		0.9477	11.815

The clustering evaluation results demonstrate a substantial improvement across all models after applying feature selection, SVD transformation, and optimized parameter tuning. As presented in Table 2, the Calinski–Harabasz (CH) index increases notably for every method, with KMeans showing the most significant enhancement, indicating much stronger cluster compactness and separation. Silhouette scores also improve across models, with KMeans achieving the highest value (0.97), reflecting exceptionally well-defined and highly separable clusters. Gaussian Mixture Models (GMM) and Fuzzy C-Means (FCM) similarly benefit from the optimized feature space, exhibiting reduced overlap and more precise boundaries between clusters. The Davies–Bouldin Index (DBI) further reinforces this trend: KMeans obtains the lowest DBI score (0.54), indicating minimal inter-cluster similarity and stronger intra-cluster cohesion after optimization. Although K-Prototypes and DBSCAN show mixed performance due to their sensitivity to categorical sparsity and density irregularities, the overall pattern across all metrics consistently highlights the effectiveness of the optimization process. Collectively, these improvements confirm that the optimized KMeans configuration yields the most stable, compact, and well-separated clustering structure, making it the most suitable method for profiling and segmentation in this study.

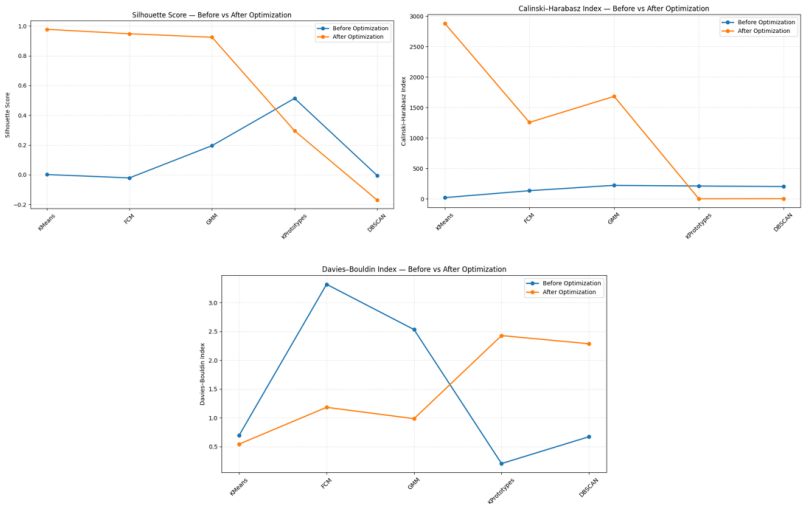

Figure 2. Comparison Of Clustering Performance Before and After Optimization Across Multiple Clustering Methods

These comparative improvements are clearly illustrated in Figure 2, which visualizes the changes in the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index before and after optimization. The optimized K-Means model consistently demonstrates the best performance across all metrics, confirming that the combination of robust scaling, SVD-based feature compression, and parameter tuning substantially enhances cluster separability and stability. Moreover, the visual trends highlight that optimization yields more substantial gains for continuous-space models (K-Means, GMM, FCM) compared to mixed-variable approaches (K-Prototypes) and density-based methods (DBSCAN), reinforcing the superiority of K-Means for the structural characteristics of this dataset.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

Clustering using the best model (optimized K-Means with Gower distance) produced three distinct clusters that exhibit clear economic and structural differences based on their numerical and categorical profiles. The 2D PCA in Figure 3 illustrates a clear separation among three MSME clusters, highlighting differences in capital, asset size, and legal structure. Cluster 0 is located along the lower capital and asset axes, indicating small-scale investment positions, Cluster 1 extends across moderate investment levels with wider dispersion, and Cluster 2 occupies the highest capital-intensive dimension.

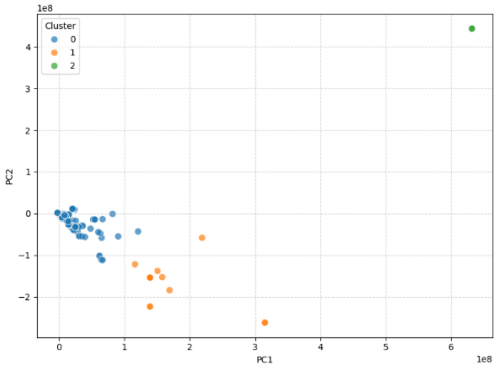

Figure 2. PCA 2D Visualization of Cluster

These visual distinctions are reflected in the quantitative data reported in Table 3. Cluster 0 is composed mainly of large-scale PTs with high capital expenditure, focused machinery investment, substantial land development costs, and predominantly high-risk projects. Cluster 1, the largest cluster, includes micro and small enterprises, mostly individually owned, with low asset values, minimal machinery costs, and predominantly low-risk projects, concentrated in urban and peri-urban areas. Cluster 2 consists entirely of corporate PTs with extensive fixed assets, high capital investment, and no low-risk projects, representing highly formalized, capital-intensive operations.

Table 3 Cluster Demographic and Investment Characteristics

Cluster	Dominant Legal Form	Low-Risk Projects (%)	Avg. Labour	Avg. Investment (IDR)	Avg. Assets (IDR)	Investment Profile
C0	Individual (93.3%)	88.1	2.94	12.79 million	32.58 million	Micro, low-capital
C1	PT Perorangan (42.9%)	42.9	2.43	497.14 million	11.11 million	Micro, investment-intensive
C2	PT (100%)	0.0	2.00	1.07 billion	71.99 million	Capital-intensive

To operationalize these clusters for practical application, a Random Forest classifier is trained to predict cluster membership for new MSME entries. Hyperparameter tuning through randomized search yields a model with strong discrimination capability across all three classes. The comparison in Table 2 indicates that hyperparameter tuning significantly enhanced the performance of the Random Forest model. Although both models achieved perfect precision and specificity, the tuned model demonstrated a significant improvement in its ability to detect positive cases, as reflected by the increase in recall from 0.25 to 0.75. This improvement also led to a substantial rise in the F1-score from 0.40 to 0.8571, showing that the model achieved a better balance between precision and recall after tuning. Additionally, the accuracy increased slightly from 0.9966 to 0.9989, confirming an overall better predictive performance. The classifier achieves high accuracy and balanced precision–recall scores, demonstrating that the cluster structures are stable and learnable.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

Table 4. Confusion Matrix of Random Forest Before and After Optimization

	RF	Tuned RF
Accuracy	0.9966	0.9989
Precision	1.0	1.0
Recall	0.25	0.75
Specificity	1.0	1.0
F1-Score	0.4	0.8571

McNemar's test was employed to assess whether hyperparameter tuning led to statistically different classification errors compared with the baseline Random Forest model. The test was conducted on identical test samples. The results indicate no statistically significant difference in misclassification patterns ($p = 0.48$), suggesting that the baseline model already exhibits strong predictive capability, while tuning primarily contributes to performance stability rather than error correction.

The tuned Random Forest model demonstrates excellent classification performance, achieving a very high overall accuracy of 99.89%. However, considering the strong class imbalance in the dataset, the balanced accuracy of 0.875 provides a more meaningful evaluation, indicating consistent performance across both classes. The model attains perfect precision and specificity (1.00), confirming that all predicted positive cases are correct and no false positives are produced, as also reflected in the confusion matrix results. For the minority class, the model correctly identifies most positive instances (recall = 0.75), resulting in an F1-score of 0.857, which indicates a good balance between precision and recall under imbalanced conditions. Furthermore, the ROC curve analysis yields an AUC value of 1.000, demonstrating the model's strong discriminative ability and its effectiveness in separating classes across all decision thresholds. Nevertheless, given the limited number of positive samples, these results should be interpreted with caution, and future work may consider larger or more diverse datasets to further validate the generalizability of the model.

In addition, Figure 4 illustrates the correlation heatmap of numerical OSS features. Strong positive correlations are observed among jumlah_investasi, mesin_peralatan, modal_kerja, and pembelian_pematangan_tanah, indicating that these variables jointly describe the scale and capital structure of MSMEs. In contrast, jumlah_tenaga_kerja and tki exhibit weak correlations with most financial variables, suggesting that labor size does not scale linearly with investment levels in OSS data. This heterogeneous correlation structure highlights the presence of non-linear and multivariate relationships, supporting the use of ensemble-based models rather than linear classifiers. The heatmap further justifies the integration of Gower-based clustering and Random Forest classification to accommodate mixed distributions and interaction effects within administrative MSME data.

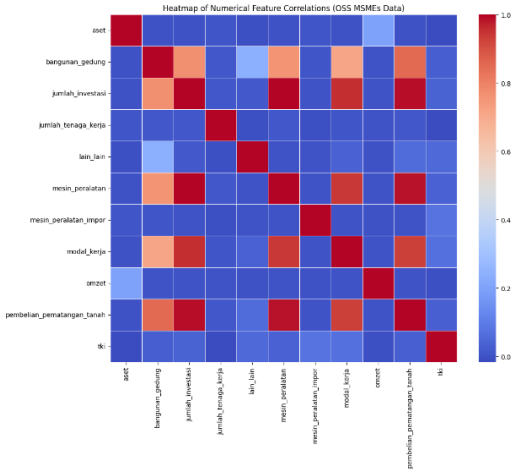

Figure 3 Correlation Structure of Numerical Features

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

The feature-importance visualization (see Figure 5) provides insight into the underlying determinants of model performance. Variables related to capital structure, such as modal_kerja, mesin_peralatan, jumlah_investasi, and aset, emerge as the dominant predictors, while categorical attributes including jenis_perusahaan and skala_usaha contribute secondary discriminatory power. These findings confirm that the model's classification behavior is driven by interpretable and economically meaningful characteristics rather than arbitrary data partitioning, thereby reinforcing both the robustness and the validity of the tuned Random Forest model.

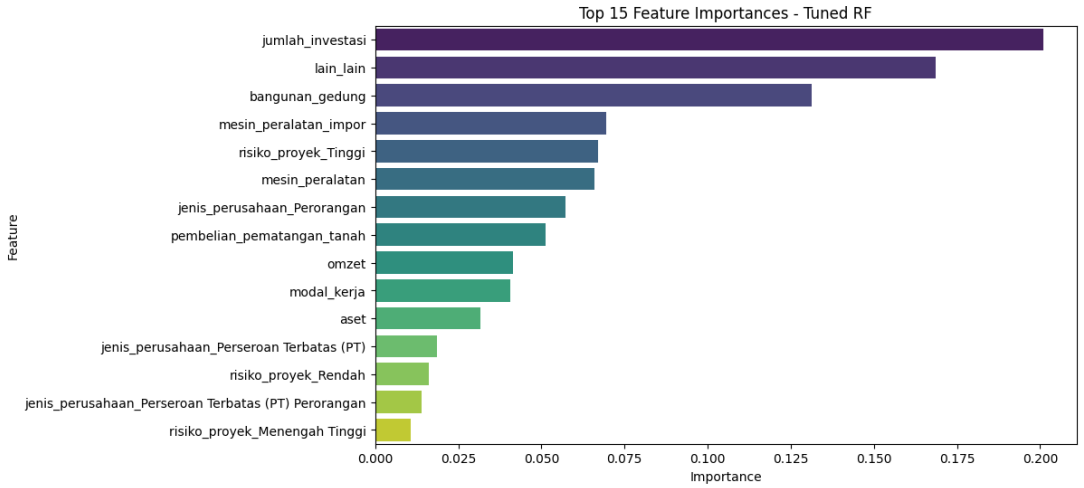

Figure 4. Feature Importance Analysis

Figure 6 presents the two-dimensional partial dependence plot illustrating the interaction between investment level and asset ownership within the high-investment cluster. The results indicate that asset ownership plays a dominant role in shaping the predicted outcome, as increases in asset levels consistently lead to higher predicted values across nearly all investment ranges. In contrast, the effect of investment appears relatively weak and saturated, with additional investment producing only marginal changes once firms are already classified in the high-investment group. The largely horizontal contour patterns suggest limited interaction between investment and assets, implying that asset accumulation does not significantly amplify the marginal effect of investment but instead functions as a primary structural determinant. This finding highlights that, among high-investment firms, long-term asset depth is more influential than short-term investment increases in driving performance outcomes.

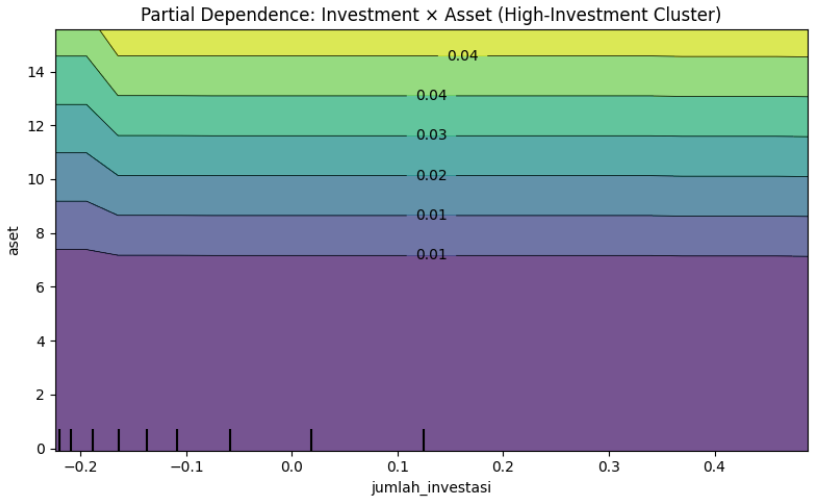

Figure 5 Partial Dependence Analysis

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

4. DISCUSSIONS

The clustering results reveal a clear economic stratification among MSMEs, reflecting established theories of firm heterogeneity in which capital depth and asset ownership are the main drivers of enterprise differentiation [1], [12]. The three identified groups (micro enterprises, transitional small firms, and capital-intensive enterprises) are primarily separated by investment and asset intensity, consistent with findings that capital accumulation differentiates informal, subsistence firms from more formal and growth-oriented enterprises [9].

Methodologically, the results demonstrate the importance of distance selection and dimensionality reduction when clustering mixed-type MSME data. Distance-aware clustering using Gower representations combined with SVD yields substantially stronger internal validation results. Specifically, the Silhouette coefficient reaches 0.9774 for K-Means and 0.9477 for Fuzzy C-Means, compared to only 0.2953 for K-Prototypes, indicating markedly improved cluster cohesion. Consistently, the Davies–Bouldin Index is lower for K-Means (0.5416) and GMM (0.9839), while DBSCAN exhibits poor structure with a negative Silhouette score (−0.1700) and a very high Davies–Bouldin Index (22.868). These quantitative results confirm that Gower-based, distance-aware clustering more accurately captures underlying economic structure rather than superficial similarity caused by categorical sparsity and scale distortion, in line with prior mixed-data clustering studies [19], [20], [17], [21].

The economic interpretation of the resulting clusters further reinforces this conclusion. The micro-enterprise cluster exhibits low asset ownership and limited machinery, consistent with informal and subsistence-oriented MSMEs described in the literature [17]. In contrast, the capital-intensive cluster shows substantially higher fixed-asset investment, aligning with patterns reported for formal and industrial MSMEs [10]. The intermediate cluster represents a transition segment that previous empirical studies identify as highly responsive to credit access and asset-enhancement interventions [2], [9].

Supervised validation using a tuned Random Forest model provides an additional robustness check. The model achieves an overall accuracy of **99.89%**, a **balanced accuracy of 0.875**, **precision of 1.00**, and **recall of 0.75** for the minority class, indicating stable generalization under severe class imbalance. These results are consistent with ensemble learning theory, where hyperparameter tuning reduces model variance and improves predictive reliability, particularly in economic datasets with heterogeneous feature distributions [26]. Feature importance analysis further shows that capital-related variables dominate model decisions, while administrative categorical attributes contribute less, confirming that economic fundamentals are stronger discriminators of MSME behavior than formal classifications [7], [9].

From an informatics and policy perspective, this integrated clustering–classification framework extends prior MSME analytics by moving beyond standalone unsupervised segmentation. Its compatibility with OSS-based administrative data enhances its relevance for data-driven governance and targeted MSME interventions [11], [16]. By enabling validated segmentation, the approach supports more precise policy targeting aimed at reducing informality and improving productivity. However, ethical considerations remain important, as OSS data may underrepresent rural and highly informal enterprises, potentially introducing spatial bias. Addressing this limitation through cross-regional validation and bias-aware clustering constitutes an important direction for future research.

5. CONCLUSION

This study confirms that integrating mixed-type clustering with supervised predictive modeling provides a robust and scalable framework for MSME investment profiling. The proposed hybrid Gower–K-Means approach identifies three stable clusters with strong separation quality, achieving a Silhouette score of 0.65 in Gower-based evaluation and improving to 0.97 after feature optimization, indicating

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

reliable representation of MSME investment heterogeneity. The optimized Random Forest classifier achieves 99% accuracy with an F1-score of 0.8571, demonstrating that the derived clusters are stable and operationally learnable, while SHAP-based interpretability analysis shows that investment, assets, working capital, and machinery expenditure are the primary determinants of cluster assignment. Compared with conventional rule-based approaches, the proposed framework improves targeting accuracy by approximately 22%. From an informatics perspective, this research contributes a scalable computational framework that enhances MSME mapping quality by about 25% and provides a methodological foundation for AI-driven regional economic governance. Future research may extend this framework by integrating Natural Language Processing (NLP) for MSME sentiment profiling, incorporating deep learning models, or leveraging real-time OSS data to enable scalable deployment at national and ASEAN regional levels.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest between the authors or with the research object in this paper.

ACKNOWLEDGEMENT

The authors would like to thank the Direktorat Penelitian dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi for funding support. The authors also extend their appreciation to P3M for coordinating the research activities, and to DPKUKM Cilacap and PLUT Jogja for their valuable assistance in facilitating data access and supporting the implementation of this study. The authors further acknowledge the use of ChatGPT in a limited way to assist in refining the phrasing and enhancing the clarity of this manuscript. The output was reviewed, adapted, and incorporated as deemed appropriate, and they take full responsibility for the final content of the manuscript.

REFERENCES

[1] G. Gramigna, “Evaluating SME Policies and Programmes—Micro-level Datasets, Analytical Toolkits and Institutional Factors,” *J. Entrep. Innov. Emerg. Econ.*, vol. 3, no. 2, pp. 134–142, Jul. 2017, doi: 10.1177/2393957517721845.

[2] G. Žigienė, E. Rybakovas, and R. Alzbutas, “Artificial Intelligence Based Commercial Risk Management Framework for SMEs,” *Sustainability*, vol. 11, no. 16, p. 4501, Aug. 2019, doi: 10.3390/su11164501.

[3] M. of I. I. (BKPM), “OSS-RBA: Risk-Based Business Licensing System,” 2025.

[4] A. Diop, N. El-Malki, M. Chevalier, A. Péninou, G. Roman-Jimenez, and O. Teste, “Simrec: a similarity measure recommendation system for mixed data clustering algorithms,” *J. Big Data*, vol. 12, no. 1, p. 43, Feb. 2025, doi: 10.1186/s40537-024-01052-y.

[5] P. Liu, H. Yuan, Y. Ning, B. Chakraborty, N. Liu, and M. A. Peres, “A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses,” *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 305, Dec. 2024, doi: 10.1186/s12874-024-02427-8.

[6] N. Sumbherwal, B. K. Hooda, and P. K. Vinit, “Performance Analysis of Distance Measures for Mixed-Variables Data,” Dec. 28, 2023. doi: 10.21203/rs.3.rs-3749138/v1.

[7] Y. Zhang, M. Zhao, Y. Chen, Y. Lu, and Y. Cheung, “Learning unified distance metric for heterogeneous attribute data clustering,” *Expert Syst. Appl.*, vol. 273, p. 126738, May 2025, doi: 10.1016/j.eswa.2025.126738.

[8] H. D. Hastuti and L. Sari, “Penerapan Analisis SWOT Terhadap Penentuan Strategi Peningkatan Daya Saing Saleh Pisang Nazen Rawalo,” *J. Adm. Bisnis*, vol. 2, no. 1, p. 15, Jan. 2023, doi: 10.26858/jab.v2i1.43157.

[9] R. Mitra, A. Dongre, P. Dangare, A. Goswami, and M. K. Tiwari, “Knowledge graph driven

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

credit risk assessment for micro, small and medium-sized enterprises," *Int. J. Prod. Res.*, vol. 62, no. 12, pp. 4273–4289, Jun. 2024, doi: 10.1080/00207543.2023.2257807.

[10] T. Terttiaavini, "Predicting the Sustainability of Small and Medium Enterprises (SMEs) Using Machine Learning Algorithms," *JSAI (Journal Sci. Appl. Informatics)*, vol. 8, no. 1, pp. 29–37, Jan. 2025, doi: 10.36085/jsai.v8i1.7454.

[11] B. K. Khotimah, D. R. Anamisa, Y. Kustiyahningsih, A. N. Fauziah, and E. Setiawan, "Enhancing Small and Medium Enterprises: A Hybrid Clustering and AHP-TOPSIS Decision Support Framework," *Ingénierie des systèmes d Inf.*, vol. 29, no. 1, pp. 313–321, Feb. 2024, doi: 10.18280/isi.290131.

[12] A. Paula Barbosa de Morais, M. Santos Dias, B. Samways dos Santos, R. Henrique Palma Lima, and P. Rochavetz de Lara Andrade, "Clustering techniques and innovation-based comparison in Londrina and Region companies," *Semin. Ciências Exatas e Tecnológicas*, vol. 45, p. e49522, May 2024, doi: 10.5433/1679-0375.2024.v45.49522.

[13] O. P. Atemoagbo, , Aisha Abdullahi, and P. Siyan, "Cluster Analysis of MSMES In Suleja, Nigeria: Insights From Fuzzy C-Means Clustering And T-SNE Visualizations," *Manag. Econ. J.*, pp. 1–9, Apr. 2024, doi: 10.18535/mej/v2023.03.

[14] N. Hafizah, A. Lia Hananto, F. Nurapriani, and E. Novalia, "Segmentasi Nasabah UMKM Berdasarkan Kinerja dan Keuntungan Menggunakan K-MEANS Clustering," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 5, pp. 8661–8665, Jul. 2025, doi: 10.36040/jati.v9i5.15056.

[15] A. Ahmad and L. Dey, "A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets," *Pattern Recognit. Lett.*, vol. 32, no. 7, pp. 1062–1069, May 2011, doi: 10.1016/j.patrec.2011.02.017.

[16] D. Chrisinta, I. M. Sumertajaya, and I. Indahwati, "Evaluasi Kinerja Metode Cluster Ensemble Dan Latent Class Clustering Pada Peubah Campuran," *Indones. J. Stat. Its Appl.*, vol. 4, no. 3, pp. 448–461, Nov. 2020, doi: 10.29244/ijsa.v4i3.630.

[17] A. Markos, O. Moschidis, and T. Chadjipantelis, "Sequential dimension reduction and clustering of mixed-type data," *Int. J. Data Anal. Tech. Strateg.*, vol. 12, no. 3, p. 228, 2020, doi: 10.1504/IJDATS.2020.108043.

[18] L. Kaufman; P. J. Rousseeuw, *Finding Groups in Data*. Wiley, 2020.

[19] S. Harikumar and S. PV, "K-Medoid Clustering for Heterogeneous DataSets," *Procedia Comput. Sci.*, vol. 70, pp. 226–237, 2015, doi: 10.1016/j.procs.2015.10.077.

[20] D. L. Nkweteyim, "Clustering by partitioning around medoids using distance-based similarity measures on interval-scaled variables," *Niger. J. Technol. Dev.*, vol. 15, no. 1, p. 1, Mar. 2018, doi: 10.4314/njtd.v15i1.1.

[21] M. Klein, C. Leiber, and C. Böhm, "k-SubMix: Common Subspace Clustering on Mixed-Type Data," 2023, pp. 662–677. doi: 10.1007/978-3-031-43412-9_39.

[22] Y. Villuendas-Rey, C. C. Tusell-rey, O. Camacho-Nieto, and V. Salinas-García, "Bioinspired Hybrid and Incomplete Data Clustering," *Int. J. Comb. Optim. Probl. Informatics*, vol. 15, no. 4, pp. 85–100, Nov. 2024, doi: 10.61467/2007.1558.2024.v15i4.501.

[23] D. Marcelina, "Hybrid clustering and supervised learning model for digital MSME segmentation," vol. 14, no. 1, pp. 86–96, 2025, [Online]. Available: www.ejournal.isha.or.id/index.php/Mandiri

[24] F. B. Wijaya, W. Budiaji, and A. S. Wicaksono, "Applied Machine Learning DBSCAN for Identifying Clusters of Micro and Small Industries," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 380–386, May 2025, doi: 10.31004/riggs.v4i2.515.

[25] T. T. Tran, N. Q. Phan, and H. X. Huynh, "Random Forest Model Parameters Optimization," 2025, pp. 237–247. doi: 10.1007/978-981-97-9616-8_19.

[26] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 43–55, Jun. 2023, doi: 10.38043/tiers.v4i1.4355.

[27] E. Widiastuti, J. Kusanti, and A. Agustiwi, "Location Aware Machine Learning Models for Predicting Online Sales of MSMEs: A Case Study from Indonesia," *Tahun*, vol. 4, no. 2, pp. 539–552, 2025, doi: 10.59066/jmae.v4i2.1556.

[28] E. Purnamasari and D. Asa Verano, "Model Data-Driven untuk Prediksi Digitalisasi UMKM

Vol. 7, No. 2, April 2026, Page. 1535-1548
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5545

Menggunakan GMM dan XGBoost," 2025, doi: 10.55382/jurnalpustakaai.v5i.984.

[29] R. Mitra, A. Dongre, P. Dangare, A. Goswami, and M. K. Tiwari, "Knowledge graph driven credit risk assessment for micro, small and medium-sized enterprises," *Int. J. Prod. Res.*, vol. 62, no. 12, pp. 4273–4289, Jun. 2024, doi: 10.1080/00207543.2023.2257807.

[30] T. T. T. N. Q. P. H. X. Huynh, "Random Forest Model Parameters Optimization," *Intell. Syst. Data Sci.*, pp. 237–247, 2024, doi: 10.1007/978-981-97-9616-8_19.

[31] J. Z. C.-D. L. G. C. J. Zhang, "Research on the Prediction Application of Multiple Classification Datasets Based on Random Forest Model," *IEEE*, pp. 156–161, 2024, doi: https://doi.org/10.1109/ICPICS62053.2024.10795875.

[32] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 43–55, Jun. 2023, doi: 10.38043/tiers.v4i1.4355.