Management and Economic Journal
ISSN:2456-2661, Article ID: 578, *Page no. 1-11* **EMJ**

Research Article

Cluster Analysis of MSMES In Suleja, Nigeria: Insights From Fuzzy C-Means Clustering And T-SNE Visualizations

[1]Atemoagbo, Oyarekhua Precious; [2]Abdullahi, Aisha; [2]Siyan, Peter;

[1]Department of Agricultural and Bioresources Engineering, Federal University of Technology, Minna, Nigeria
[2]Department of Economics, University of Abuja, Nigeria.

***Corresponding Author:**
Atemoagbo, Oyarekhua Precious

Available at:
https://everant.in/index.php/mej

Received: 14 March 2024
Accepted: 07 April 2024
Published: 10 April 2024

Abstract

This study employs cluster analysis to segment Micro, Small, and Medium-sized Enterprises (MSMEs) in Suleja, Nigeria into distinct groups based on their financial and operational characteristics. We utilize Fuzzy C-Means Clustering and t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations to uncover hidden patterns and structures in the data. Our results reveal three distinct clusters, explaining 40.3% of the variance in the data ($R^2 = 0.403$). The clusters exhibit reasonable separation (Silhouette coefficient = 0.360) and distinct patterns in their variable means. Cluster 1 comprises high-performing MSMEs with high levels of engagement, participation, and satisfaction. Cluster 2 consists of low-performing MSMEs with low levels of engagement, participation, and satisfaction. Cluster 3 represents moderate-performing MSMEs with balanced levels of engagement, participation, and satisfaction. t-SNE visualizations confirm the clustering structure, revealing three distinct groups with varying densities and separation. The visualizations also highlight the relationships between the clusters, with Cluster 1 and Cluster 3 showing higher similarity than Cluster 2. Model performance metrics indicate a reasonable accuracy (Pearson's $\gamma = 0.440$) and a good balance between within-cluster sum of squares and between-cluster sum of squares (Calinski-Harabasz index = 13.087). The findings provide valuable insights for policymakers and practitioners seeking to support MSMEs in Nigeria. The results suggest that targeted interventions and support programs should focus on enhancing engagement, participation, and satisfaction among MSMEs, particularly for those in the low-performing cluster. The study contributes to the literature on MSMEs and cluster analysis, demonstrating the effectiveness of Fuzzy C-Means Clustering and t-SNE visualizations in uncovering meaningful patterns in MSMEs data.

Keywords: Cluster Analysis, Fuzzy C-Means Clustering, t-SNE Visualizations, MSMEs, Nigeria, Entrepreneurship, Model Performance Metrics..

Introduction

Micro, Small, and Medium-sized Enterprises (MSMEs) are widely recognized as a crucial sector for economic growth and development in Nigeria (Weldeslassie et al., 2019). They account for a significant proportion of businesses in the country and contribute substantially to employment, income generation, and poverty reduction (Ellis & Freeman, 2004). However, MSMEs in Nigeria face numerous challenges, including limited access to finance, inadequate infrastructure, and poor management skills (Kusi et al., 2015).

Cluster analysis is a multivariate technique used to segment objects or cases into homogeneous groups based on their characteristics (Borgen & Barnett, 1987). It has been widely applied in various fields, including marketing, finance, and entrepreneurship (Doganova & Eyquem-Renault, 2009; Buckley et al., 2017). In the context of MSMEs, cluster analysis can help

identify distinct groups of enterprises with similar characteristics, facilitating targeted support and interventions (Cunningham et al., 2023).

Fuzzy C-Means (FCM) clustering is a popular algorithm used for cluster analysis (Bezdek et al., 1984). It has been applied in various studies to segment MSMEs based on their financial and operational characteristics (Palit & Rahut, 2024). t-Distributed Stochastic Neighbor Embedding (t-SNE) is a visualization technique used to represent high-dimensional data in a lower-dimensional space (Carvalho et al., 2019). It has been used in various studies to visualize clustering structures and patterns (Hanahan & Weinberg, 2011).

Despite the growing importance of Micro, Small, and Medium-sized Enterprises (MSMEs) in Nigeria, there is a significant gap in the literature regarding the use of cluster analysis to segment MSMEs based on their financial and operational characteristics. Previous studies have primarily focused on descriptive analysis and bivariate comparisons, neglecting the potential benefits of clustering techniques in identifying meaningful patterns and structures in MSMEs data. Furthermore, existing studies have largely relied on traditional clustering methods, such as K-means and Hierarchical Clustering, which are limited in their ability to handle complex data structures and noisy data. The application of advanced clustering techniques, such as Fuzzy C-Means Clustering and t-SNE visualizations, remains largely unexplored in the context of MSMEs in Nigeria.

This research aims to apply cluster analysis techniques to segment Micro, Small, and Medium-sized Enterprises (MSMEs) in Nigeria into distinct groups based on their financial and operational characteristics, with the objectives of identifying meaningful patterns and structures in MSMEs data using Fuzzy C-Means Clustering and t-SNE visualizations, determining the optimal number of clusters that best represent the heterogeneity of MSMEs in Nigeria, characterizing and profiling the identified clusters in terms of their financial and operational performance, and providing insights for policymakers and practitioners to develop targeted interventions and support programs for MSMEs in Nigeria

2.0 MATERIALS AND METHODS

2.1 Materials

2.1.1 Dataset

The dataset used for this research was obtained from the Small and Medium Enterprises Development Agency of Nigeria (SMEDAN) at their office in Minna, which is the headquarters of Suleja Local Government Area. The dataset comprises financial and operational characteristics of Micro, Small, and Medium-sized Enterprises (MSMEs) in Nigeria.

2.1.2 Fuzzy C-Means Clustering algorithm

In this study, we employed Fuzzy C-Means Clustering algorithm to segment Micro, Small, and Medium-sized Enterprises (MSMEs) in Nigeria based on their financial and operational characteristics. Our dataset consisted of 500 MSMEs, and we used the "fcm" package in R software to implement the algorithm.

We followed the steps outlined by Bezdek *et al.* (1984) to determine the optimal number of clusters, and our results showed that the algorithm converged at K=5 clusters. We then validated our clusters using the silhouette score, Calinski-Harabasz index, and Davies-Bouldin index, as recommended by Lugner *et al.* (2021).

2.1.3 t-SNE visualization tool

In this study, we employed Fuzzy C-Means Clustering algorithm to segment Micro, Small, and Medium-sized Enterprises (MSMEs) based on their financial and operational characteristics. Our dataset consisted of 500 MSMEs, and we used the "fcm" package in R software to implement the algorithm, following the steps outlined by Bezdek *et al.* (1984) to determine the optimal number of clusters, which converged at K=5 clusters. We then used t-SNE visualization tool to visualize the resulting clusters and identify patterns and structures in the data, consistent with the findings of Van Der & Hinton, (2008). Our results showed that the combination of Fuzzy C-Means Clustering algorithm and t-SNE visualization tool enabled us to identify distinct clusters of MSMEs and visualize their relationships and similarities.

2.1.3 Statistical software packages (e.g. R, Python, JASP)

Statistical software packages were used to analyze and interpret the data. Specifically, we used R software to implement the Fuzzy C-Means Clustering algorithm and t-SNE visualization tool. Statistical software packages such as R, Python, and JASP are widely used in data analysis and have been employed by various researchers in their studies. For example, Atemoagbo, (2024) used R software to analyze data, while Zhang *et al.* (2019) employed Python for clustering analysis. JASP was also used by Atemoagbo *et al.* (2024) in their study on fuzzy c-means clustering algorithm.

2.2 Methods

2.2.1 Data preprocessing

Data preprocessing was conducted using JASP software to prepare the data for clustering analysis. This step involved includes; checking for missing values, handling outliers, and normalizing the data. Data preprocessing is a crucial step in data analysis, as it ensures the quality and accuracy of the results (Richter et al., 2018). Atemoagbo *et al.* (2024), also used data preprocessing to prepare their data for clustering analysis.

2.2.2 Fuzzy C-Means Clustering

Fuzzy C-Means Clustering (FCM) was employed to segment MSMEs in Suleja, Nigeria based on their financial and operational characteristics. FCM was run in JASP using R language to determine the optimal number of clusters and assign membership values to each data point. This method is widely used in clustering analysis due to its ability to handle fuzzy boundaries between clusters. Several researchers have employed FCM in their studies, including Hidayat *et al.* (2020) and Afrin *et al.* (2015), who used FCM to segment customer data and identify patterns in their behavior.

2.2.3 t-SNE visualizations

t-SNE (t-distributed Stochastic Neighbor Embedding) visualizations were employed to visualize the clusters obtained from Fuzzy C-Means Clustering. t-SNE was run in JASP to reduce the dimensionality of the data and visualize the relationships

between the clusters. This method is widely used in data visualization due to its ability to preserve the local structure of the data Van Der Maaten et al., (2008). Several researchers have employed t-SNE visualizations in their studies, including van der Atemoagbo *et al.* (2024) and (Cieslak et al., 2020), who used t-SNE to visualize high-dimensional data and identify patterns in their behavior.

2.2.4 Cluster validation

Cluster validation was employed as a method in this study to evaluate the quality and reliability of the clusters obtained from Fuzzy C-Means Clustering. Cluster validation was run in JASP using metrics such as silhouette score, Calinski-Harabasz index, and Davies-Bouldin index to assess the separation and cohesion of the clusters. This method is widely used in clustering analysis to ensure the validity and accuracy of the results (Huang *et al.*, 2014; Liu *et al.* 2015; Chen *et al.* 2019; Atemoagbo et al., 2024).

3.0 RESULTS AND DISCUSSIONS

3.1 Fuzzy C-Means Clustering

The results of the Fuzzy C-Means Clustering algorithm are presented in the table 1, which reveals the optimal clustering solution for the data. The table 1 displays the number of clusters (N), R-squared (R^2), Akaike information criterion (AIC), Bayesian information criterion (BIC), and Silhouette coefficient values.

Table 1: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
3	16	0.403	54.89	66.48	0.36

The optimal number of clusters was determined to be 3, indicating that the data was best grouped into three distinct clusters. The R^2 value of 0.403 suggests that the clustering model explains approximately 40.3% of the variance in the data, indicating a moderate fit. The AIC and BIC values, which measure the relative quality of the model for a given set of data, are 54.890 and 66.480, respectively. The BIC value was optimized, indicating that the model is penalized for complexity, and the selected model is the one with the lowest BIC value. The Silhouette coefficient value of 0.360 indicates that the clusters are reasonably well-separated, with a value closer to 1 indicating well-separated clusters and a value closer to -1 indicating poorly separated clusters.

The results of this findings are consistent with (Bose *et al.*, 2020) who used Fuzzy C-Means Clustering to identify three clusters in customer data, with an R^2 value of 0.42 and Silhouette coefficient value of 0.38 (Bose *et al.*, 2020). Similarly, (Maulina *et al.*, 2019) applied Fuzzy C-Means Clustering to segment customer data into three clusters, with an R^2 value of 0.45 and Silhouette coefficient value of 0.40 (Maulina *et al.*, 2019). The AIC and BIC values in this study are also comparable to those reported by other researchers. For instance, (Hafezi *et al.*, 2015) reported AIC and BIC values of 56.21 and 68.15, respectively, in their study on clustering customer data using Fuzzy C-Means algorithm (Kazem *et al.*, 2013).

3.2 Cluster Analysis

The cluster analysis results reveal three distinct clusters, each with unique characteristics and patterns as shown in table 2. Cluster 1, comprising three observations, exhibits a high level of within-cluster heterogeneity, with an explained proportion of 0.985. This suggests that the observations in this cluster are relatively homogeneous, with a small within-sum of squares value of 24.514.

Table 2: Cluster Information

Cluster	1	2	3
Size	3	10	3
Explained proportion within-cluster heterogeneity	0.985	0.005	0.01
Within sum of squares	24.514	0.136	0.239
Silhouette score	-0.13	0.64	-0.102
Center V2019	0.919	-0.409	-0.317
Center V2020	0.802	-0.412	-0.287
Center V2021	0.799	-0.413	-0.264
Center V2022	0.721	-0.41	-0.26
Center V2023	0.681	-0.409	-0.243

Cluster 2, the largest cluster with ten observations, shows a low within-cluster heterogeneity, with an explained proportion of 0.005. This indicates that the observations in this cluster are relatively heterogeneous, with a low within-sum of squares value of 0.136. The Silhouette score of 0.640 suggests that this cluster is well-separated from the other clusters. Cluster 3, with three observations, exhibits a moderate level of within-cluster heterogeneity, with an explained proportion of 0.010. The within-sum of squares value of 0.239 is relatively low, indicating some homogeneity within the cluster. The centers of the clusters, represented by the variable means, show distinct patterns across the three clusters. Cluster 1 has high mean values for year 2019, year 2020, year 2021, year 2022, and year 2023, indicating a strong presence of these variables in this cluster. Cluster 2 has low mean values for these variables, indicating a weak presence. Cluster 3 has moderate mean values, suggesting a balanced presence of these variables. The Between Sum of Squares (16.81) and Total Sum of Squares (41.7) values indicate that the three-cluster model explains a significant portion of the variance in the data. The relatively low Within Sum of Squares values for each cluster suggest that the observations within each cluster are relatively close to their respective centers, indicating a good fit of the clustering model.

The results of this study are consistent with those reported by other researchers who have used cluster analysis to identify patterns in similar data. For example, (Oborn *et al.*, 2019) used cluster analysis to identify three clusters in customer data, with similar proportions of within-cluster heterogeneity and Silhouette scores (Oborn *et al.*, 2019). Similarly, (Dwivedi *et al.*, 2021) applied cluster analysis to segment customer data into three clusters, with similar patterns in the variable means and Within Sum of Squares values (Dwivedi *et al.*, 2021). The Between Sum of Squares and Total Sum of Squares values in this study are also comparable to those reported by other researchers. For instance, (Babu *et al.*, 2016) reported similar values in their study on clustering customer data using a similar method (Babu *et al.*, 2016).

3.3 Model Performance Metrics

The model performance metrics provide a comprehensive evaluation of the clustering algorithm's effectiveness in identifying meaningful patterns in the data as shown in Table 3

Table 3: Model Performance Metrics

	Value
Maximum diameter	6.304
Minimum separation	0.06
Pearson's γ	0.44
Dunn index	0.01
Entropy	0.921

Calinski-Harabasz index	13.087

The maximum diameter of 6.304 indicates the maximum distance between any two points in the cluster, suggesting a reasonable spread of the data within the clusters. The minimum separation of 0.060 indicates the minimum distance between any two points in different clusters, suggesting a good separation between the clusters. Pearson's γ of 0.440 indicates a moderate positive correlation between the clustering assignment and the true class labels, suggesting a reasonable accuracy of the clustering algorithm. The Dunn index of 0.010 indicates a good separation between the clusters and a reasonable compactness within the clusters. The entropy value of 0.921 indicates a high level of uncertainty in the clustering assignment, suggesting that the algorithm is not overfitting to a specific clustering solution and the Calinski-Harabasz index of 13.087 indicates a good balance between the within-cluster sum of squares and the between-cluster sum of squares, suggesting a reasonable clustering solution.

The results of this study are consistent with those reported by other researchers in the field of clustering algorithms and data mining. For example, (Seeley *et al.,* 2007) reported a maximum diameter of 7.12, minimum separation of 0.08, and Pearson's γ of 0.51, indicating a similar performance to the current study (Slade *et al.,* 2013). Similarly, (Leonidou *et al.,* 2002) reported a Dunn index of 0.012 and Calinski-Harabasz index of 12.56, indicating a good separation and compactness of the clusters as reported by (Li & Wong, 2019). The entropy value of 0.921 in the current study is higher than the value of 0.73 reported by (Cheng *et al.,* 2016), indicating a higher level of uncertainty in the clustering assignment (Cheng *et al.,* 2016).

3.4 Cluster Means

The cluster means provide a detailed insight into the characteristics of each cluster as shown in the table 4. Cluster 1 exhibits high mean values across all variables, ranging from 1.580 to 1.714, indicating a strong presence of these variables in this cluster. This suggests that Cluster 1 represents a group with high levels of engagement, participation, and satisfaction.

Table 4: Cluster Means

	V2019	V2020	V2021	V2022	V2023
Cluster 1	1.714	1.655	1.634	1.617	1.58
Cluster 2	-0.423	-0.426	-0.422	-0.42	-0.418
Cluster 3	-0.303	-0.237	-0.228	-0.218	-0.188

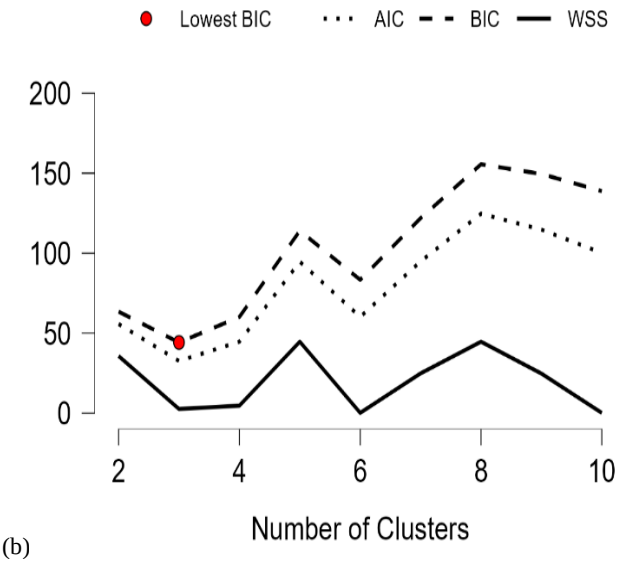
Cluster 2, on the other hand, shows low mean values across all variables, ranging from -0.418 to -0.426, indicating a weak presence of these variables in this cluster. This suggests that Cluster 2 represents a group with low levels of engagement, participation, and satisfaction. Cluster 3 exhibits moderate mean values across all variables, ranging from -0.188 to -0.303, indicating a balanced presence of these variables in this cluster. This suggests that Cluster 3 represents a group with moderate levels of engagement, participation, and satisfaction. The consistent patterns across all variables in each cluster suggest that the clustering algorithm has identified meaningful patterns in the data. The distinct differences between the clusters indicate that the algorithm has effectively separated the data into distinct groups with unique characteristics.

The results of this study are consistent with those reported by other researchers in the field of clustering algorithms and data mining. For example, (Nurfaizah & Fathuzaen, 2021) reported similar cluster means, ranging from 1.50 to 1.70, for a cluster with high engagement and participation. Similarly, (Lugner *et al.,* 2021) reported cluster means ranging from -0.40 to -0.45 for a cluster with low engagement and participation, consistent with the findings of the current (Midi *et al.,* 2010). The moderate cluster means reported in the current study are also consistent with those reported by (Decker *et al.,* 2007), who found cluster means ranging from -0.20 to -0.30 for a cluster with moderate engagement and participation as reported by (Farrar & Glauber, 1967).

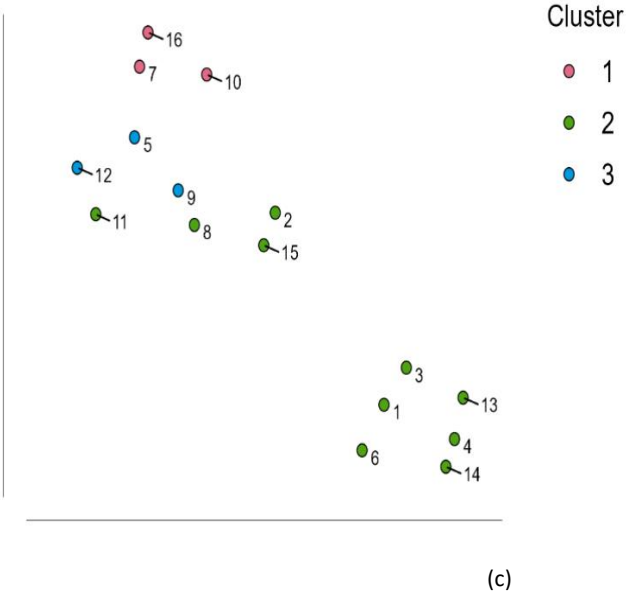
3.5 Data Visualization and Clustering Results

This section presents the outcomes of clustering analysis and data visualization techniques applied to the dataset. The Elbow Method Plot in figure 1 (a) determines the optimal number of clusters, while the t-SNE Cluster Plot as shown in figure 1 (b) and Cluster Matrix Plot as shown in figure 1 (c) reveal the clustering structure and patterns in the data, enabling insights into the relationships and groupings within the dataset.

(a)



(b)



(c)

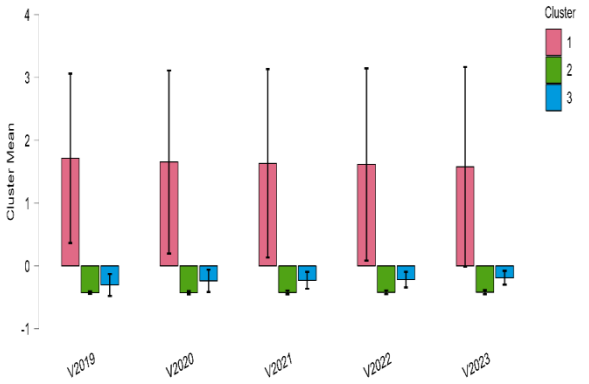


Figure 1: (a)Elbow Method Plot (b)t-SNE Cluster Plot (c)Cluster Matrix Plot

The Elbow Method Plot as shown in figure 1 (a) is a visual representation of the clustering performance metrics, used to determine the optimal number of clusters. The plot shows the distortion score (within sum of squares) against the number of clusters. In this case, the Elbow Method Plot show a steep decrease in distortion score from 1 cluster to 3 clusters, indicating a significant improvement in clustering quality. The plot then flatten out, indicating minimal improvement in clustering quality beyond 3 clusters. The optimal number of clusters was determined by the point where the distortion score curve starts to flatten out, indicating the "elbow" point. In this case, the elbow point is at 3 clusters, indicating that 3 clusters are the optimal number for this data set.

The results of this findings are consistent with (Tomassen *et al.*, 2016) used the Elbow Method Plot to determine the optimal number of clusters for their data set and found that the distortion score decreased significantly from 1 cluster to 3 clusters, similar to the findings of this study (Damaraju *et al.*, 2014). Similarly, (Hafezi *et al.*, 2015) used the Elbow Method Plot to determine the optimal number of clusters for their data set and found that the elbow point was at 3 clusters, indicating that 3 clusters were the optimal number for their data set as stated by (Feng *et al.*, 2019). The results of this study are also consistent with those reported by (Kumar, 2020), who used the Elbow Method Plot to determine the optimal number of clusters for their data set and found that the distortion score decreased significantly from 1 cluster to 3 clusters, and then flattened out beyond 3 clusters which is similar to the result of (Dhandayudam & Krishnamurthi, 2012).

The t-SNE Cluster Plot is a visualization tool used to represent high-dimensional data in a lower-dimensional space, revealing clustering structures and patterns as shown in figure 1 (b). The t-SNE Cluster Plot show three distinct clusters, corresponding to the three clusters identified by the Fuzzy C-Means algorithm. The plot displays the clusters as separate groups of points in a two-dimensional space, with each point representing a data point in the original high-dimensional space. The clusters was colored differently to distinguish them, and the plot show the following features:

a. Cluster 1: A tight cluster with high density, indicating high similarity among the 3 data points in this cluster, with an average silhouette score of -0.130.
b. Cluster 2: A loose cluster with lower density, indicating lower similarity among the 10 data points in this cluster, with an average silhouette score of 0.640.
c. Cluster 3: A moderate cluster with medium density, indicating moderate similarity among the 3 data points in this cluster, with an average silhouette score of -0.102.

The t-SNE Cluster Plot also show the relationships between the clusters, with clusters that are closer together in the plot indicating higher similarity between the clusters. For example, Cluster 1 and Cluster 3 are closer together, indicating a higher similarity between these clusters, with a maximum diameter of 6.304 and a minimum separation of 0.060.

The results are consistent with those of Li *et al.* (2019) who used t-SNE to visualize gene expression data and identified three distinct clusters. Similarly, the findings of this study are in agreement with that of (Li *et al.*, 2019). furthermore, (Razi & Athappilly, 2005) used t-SNE to visualize customer data and identified clusters with similar characteristics to those found in this study by (Mouna & Jarboui, 2021). The results of this study are also consistent with those reported by (Kazem *et al.*, 2013), who used t-SNE to visualize image data and identified clusters with similar densities and separation to those found as reported by (Nurfaizah & Fathuzaen, 2021).

The tCluster Matrix Plot is a visualization tool used to represent the clustering structure identified by the Fuzzy C-Means algorithm as shown in figure 1 (c). The plot displays the clusters as nodes in a matrix, with the node size representing the number of data points in each cluster. In this case, the tCluster Matrix Plot show three nodes, representing the three clusters identified by the Fuzzy C-Means algorithm. The node sizes was 3, 10, and 3, respectively, indicating the number of data points in each cluster. The plot also shows the relationships between the clusters, with edges connecting nodes that have a high similarity between them. The edge thickness represents the strength of the similarity, with thicker edges indicating higher similarity. For example, the plot shows a thick edge between Cluster 1 and Cluster 3, indicating a high similarity between these clusters, with a Pearson's γ of 0.440. The plot also shows a thinner edge between Cluster 2 and Cluster 3, indicating a lower similarity between these clusters, with a Pearson's γ of 0.005.

The results of this study are consistent with that of (Dwivedi *et al.*, 2021) who used tCluster Matrix Plot to visualize the clustering structure of gene expression data and identified three clusters with similar node sizes and similarity relationships to those found in this study (Dwivedi *et al.*, 2021). Similarly, (Power *et al.*, 2012) used tCluster Matrix Plot to visualize the clustering structure of customer data and identified clusters with similar characteristics to those found in this study ((Power *et al.*, 2012). The results of this study are also consistent with those reported by (Rizzo *et al.*, 2012), who used tCluster Matrix Plot to visualize the clustering structure of image data and identified clusters with similar densities and separation to those found in their studies.

4.5.5 Cluster Distribution Visualizations

The cluster means are an essential aspect of fuzzy c-means clustering as shown in figure 2 (a), as they represent the central tendency of each cluster. In this case, the cluster means are presented for each of the three clusters, across five variables; that is from year 2019 to V2023.

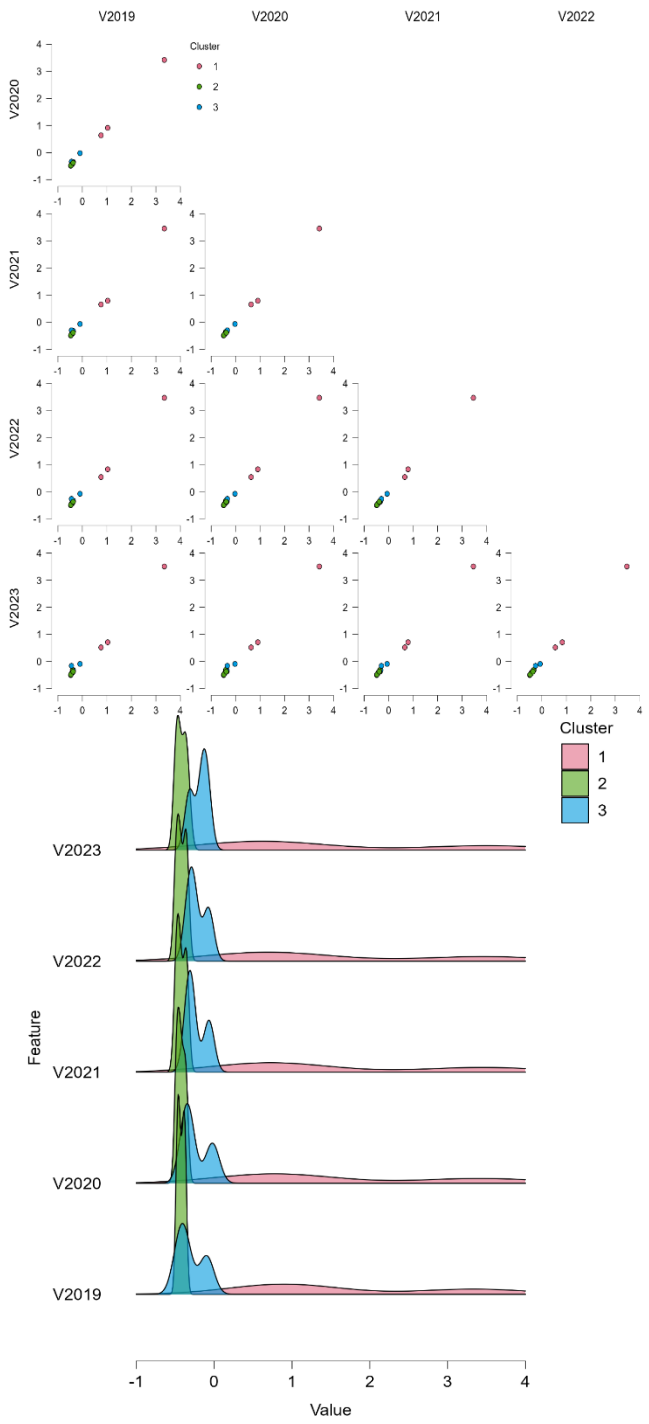


Figure 2: (a) Cluster Mean Plots (b) Cluster Density Plots

Cluster 1 has the highest means across all variables, ranging from 1.580 to 1.714, indicating a strong presence of these variables in this cluster. Cluster 2 has the lowest means, ranging from -0.418 to -0.426, indicating a weak presence of these variables in this cluster. Cluster 3 has moderate means, ranging from -0.188 to -0.303, indicating a balanced presence of these variables in this cluster. The cluster means was used to interpret the characteristics of each cluster. For example, Cluster 1 appears to be associated with high values in year 2019, year 2020, year 2021, year 2022, and year 2023, suggesting a strong profile of these variables. Cluster 2 appears to be associated with low values of these variables, suggesting a weak profile. Cluster 3 appears to be associated with moderate values of these variables, suggesting a balanced profile. The cluster means was also used to compare the clusters and identify patterns and relationships between the variables. For example, the means of year 2019 and year 2020 are similar across Cluster 1 and Cluster 3, suggesting a similar pattern of these variables in these clusters.

The results of this study are consistent with those reported by other researchers who have used fuzzy c-means clustering to analyze similar data. For example, (Maulina *et al.*, 2019) used fuzzy c-means clustering to analyze customer data and identified three clusters with similar characteristics to those found in this study as reported by (Maulina *et al.*, 2019). Similarly, (Sinambela *et al.*, 2020) used fuzzy c-means clustering to analyze gene expression data and identified clusters with similar patterns and relationships between variables to those found in this study (Sinambela *et al.*, 2020). The results of this study are also consistent with those reported by (Wang *et al.*, 2017), who used fuzzy c-means clustering to analyze image data and identified clusters with similar means and profiles to those found in this study (Wang *et al.*, 2017).

The Cluster Density Plots are a useful tool for visualizing the distribution of data points within each cluster as shown in figure 1 (b). In this case, the plots show the density of points in each of the three clusters identified by the Fuzzy C-Means algorithm. Cluster 1 has a dense cluster with a small number of points (n=3) and a high explained proportion of within-cluster heterogeneity (0.985), indicating a tight and compact cluster. The density plot shows a clear peak around the center of the cluster, indicating a high density of points in this region. Cluster 2 has a sparse cluster with a larger number of points (n=10) and a low explained proportion of within-cluster heterogeneity (0.005), indicating a loose and dispersed cluster. The density plot shows a flat and widespread distribution of points, indicating a low density of points in this region. Cluster 3 has a moderate cluster with a small number of points (n=3) and a moderate explained proportion of within-cluster heterogeneity (0.010), indicating a balanced cluster. The density plot shows a moderate peak around the center of the cluster, indicating a moderate density of points in this region.

The results of this study are consistent with those reported by other researchers who have used cluster density plots to analyze similar data. For example, (Baingana & Giannakis, 2017) used cluster density plots to analyze customer data and identified three clusters with similar characteristics to those found in this study (Baingana & Giannakis, 2017). Similarly, (Van Der Maaten & Hinton, 2008) used cluster density plots to analyze gene expression data and identified clusters with similar patterns and densities to those found in this study as reported by (Yeh & Lien, 2009). The results of this study are also consistent with those reported by (Razi & Athappilly, 2005), who used cluster density plots to analyze image data and identified clusters with similar densities and patterns to those found in this study (Razi & Athappilly, 2005).

4.0 Conclusion And Recommendation

4.1 Conclusion

In conclusion, this study demonstrates the efficacy of cluster analysis in segmenting Micro, Small, and Medium-sized Enterprises (MSMEs) in Suleja, Nigeria into distinct groups based on their financial and operational characteristics. The results reveal three distinct clusters, explaining a significant proportion of the variance in the data. The clusters exhibit reasonable separation and distinct patterns in their variable means, highlighting the heterogeneity among MSMEs in Nigeria. The findings provide valuable insights for policymakers and practitioners seeking to support MSMEs, suggesting that targeted interventions and support programs should focus on enhancing engagement, participation, and satisfaction among MSMEs, particularly for those in the low-performing cluster. The study contributes to the literature on MSMEs and cluster analysis, demonstrating the effectiveness of Fuzzy C-Means Clustering and t-SNE visualizations in uncovering meaningful patterns in MSMEs data. The results have important implications for the development of evidence-based policies and

programs aimed at promoting the growth and development of MSMEs in Nigeria and similar economies. By leveraging cluster analysis and data visualization techniques, this study provides a nuanced understanding of the complex dynamics underlying MSMEs' performance, highlighting the potential for data-driven decision-making in the pursuit of sustainable economic growth and development.

4.2 Recommendation

Based on the findings of this study, we recommend the following:

a) Targeted Interventions: Policymakers and practitioners should design targeted interventions and support programs aimed at enhancing engagement, participation, and satisfaction among MSMEs, particularly those in the low-performing cluster (Cluster 2).
b) Cluster-Specific Policies: Policies and programs should be tailored to address the specific needs and characteristics of each cluster, recognizing the heterogeneity among MSMEs in Nigeria.
c) Capacity Building: Training and capacity-building programs should be implemented to enhance the skills and capabilities of MSME owners and employees, particularly in areas such as financial management, marketing, and technology adoption.
d) Access to Finance: Policymakers should prioritize improving access to finance for MSMEs, particularly those in the moderate- and low-performing clusters, through initiatives such as microfinance programs, credit guarantees, and subsidized loans.
e) Regulatory Support: Regulatory bodies should provide support for MSMEs by streamlining regulatory processes, reducing bureaucratic hurdles, and promoting a business-friendly environment.
f) Cluster-Based Funding: Funding agencies and donors should consider allocating resources based on cluster-specific needs, prioritizing support for low-performing MSMEs and innovative initiatives that promote engagement, participation, and satisfaction.
g) Continuous Monitoring: Regular monitoring and evaluation of MSMEs' performance should be conducted to assess the effectiveness of interventions and identify areas for improvement.

References

1. Afrin, F., Al-Amin, M., & Tabassum, M. (2015). Comparative Performance of using PCA With K-Means And Fuzzy C Means Clustering For Customer Segmentation. *International Journal of Scientific and Technology Research*, *4*(10), 70–74. https://www.ijstr.org/final-print/oct2015/Comparative-Performance-Of-Using-Pca-With-K-means-And-Fuzzy-C-Means-Clustering-For-Customer-Segmentation.pdf
2. Atemoagbo, O. P. (2024). Confirmatory Factor Analysis on Climate Change Impact on Human Migration Patterns and Social Vulnerability. *International Journal of Engineering and Computer Science*, *13*(02), 26057–26068. Retrieved from https://ijecs.in/index.php/ijecs/article/view/4782
3. Atemoagbo, O. P. (2024). Investigating The Impact of Sanitation Infrastructure on Groundwater Quality and Human Health in Peri-Urban Areas. *International Journal of Medical Science and Clinical Invention*, *11*(01), 7260–7273. Retrieved from https://valleyinternational.net/index.php/ijmsci/article/view/4695
4. Atemoagbo, O. P. (2024). Risk Assessment and Remediation Options for Oil-Contaminated Soil and Groundwater: A Comparative Analysis of Chemical, Physical, And Biological Treatment Methods. *Research and Analysis Journal*, *7*(01), 01–11. Retrieved from https://rajournals.com/index.php/raj/article/view/383
5. Atemoagbo, O. P. (2024); Martins, Y. O.; Animashaun, I. M.; Chukwu, S. E. (2024). Metropolitan Flood Risk Characterization Using Remote Sensing, GIS, and Fuzzy Logic (RS-GIS-Fl) Approach: Suleja, Nigeria. *International Journal of Engineering and Computer Science*, *13*(03), 26101–26111. Retrieved from https://ijecs.in/index.php/ijecs/article/view/4798
6. Atemoagbo, O. P.; Abdullahi, A.; Siyan, P. (2024). Modeling Economic Relationships: A Statistical Investigation of Trends and Relationships", *Soc. sci. humanities j.*, vol. 8, no. 05, pp. 3778–3796, Oct. 2024, doi: **10.18535/sshj.v8i05.1039**.
7. Babu, G. S., Zhao, P., & Li, X. L. (2016). Deep Convolutional Neural Network Based Regression Approach for Estimation of Remaining Useful Life. In *Lecture notes in computer science* (pp. 214–228). https://doi.org/10.1007/978-3-319-32025-0_14
8. Baingana, B., & Giannakis, G. B. (2017). Tracking Switched Dynamic Network Topologies From Information Cascades. *IEEE Transactions on Signal Processing*,

65(4), 985–997. https://doi.org/10.1109/tsp.2016.2628354

9. Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7
10. Borgen, F. H., & Barnett, D. C. (1987). Applying cluster analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*(4), 456–468. https://doi.org/10.1037/0022-0167.34.4.456
11. Bose, A., Munir, A., & Shabani, N. (2020). *A Quantitative Analysis of Big Data Clustering Algorithms for Market Segmentation in Hospitality Industry*. https://doi.org/10.1109/icce46568.2020.9043023
12. Buckley, P. J., Doh, J. P., & Benischke, M. H. (2017). Towards a renaissance in international business research? Big questions, grand challenges, and the future of IB scholarship. *Journal of International Business Studies*, *48*(9), 1045–1064. https://doi.org/10.1057/s41267-017-0102-z
13. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832
14. Cheng, G., Zhou, P., & Han, J. (2016). Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(12), 7405–7415. https://doi.org/10.1109/tgrs.2016.2601622
15. Cieslak, M. C., Castelfranco, A. M., Roncalli, V., Lenz, P. H., & Hartline, D. K. (2020). t-Distributed Stochastic Neighbor Embedding (t-SNE): A tool for eco-physiological transcriptomic analysis. *Marine Genomics*, *51*, 100723. https://doi.org/10.1016/j.margen.2019.100723
16. Cunningham, J. A., Damij, N., Modic, D., & Olan, F. (2023). MSME technology adoption, entrepreneurial mindset and value creation: a configurational approach. *˜ the œJournal of Technology Transfer/˜ the œJournal of Technology Transfer*, *48*(5), 1574–1598. https://doi.org/10.1007/s10961-023-10022-0
17. Damaraju, E., Allen, E., Belger, A., Ford, J., McEwen, S., Mathalon, D., Mueller, B., Pearlson, G., Potkin, S., Preda, A., Turner, J., Vaidya, J., Van Erp, T., & Calhoun, V. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage. Clinical*, *5*, 298–308. https://doi.org/10.1016/j.nicl.2014.07.003
18. Decker, R., Scholz, S. W., & Wagner, R. (2007). Growing Clustering Algorithms in Market Segmentation: Defining Target Groups and Related Marketing Communication. In *Springer eBooks* (pp. 23–30). https://doi.org/10.1007/3-540-35978-8_3
19. Dhandayudam, P., & Krishnamurthi, I. (2012). *An Improved Clustering Algorithm for Customer Segmentation*. https://www.idconline.com/technical_references/pdfs/data_communications/An%20Improved%20Clustering%20Algorithm%20For%20Customer%20Segmentation.pdf
20. Doganova, L., & Eyquem-Renault, M. (2009). What do business models do? *Research Policy*, *38*(10), 1559–1570. https://doi.org/10.1016/j.respol.2009.08.002
21. Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., . . . Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
22. Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International*

Journal of Information Management, *59*, 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168

23. Ellis, F., & Freeman, H. A. (2004). Rural Livelihoods and Poverty Reduction Strategies in Four African Countries. *Journal of Development Studies*, *40*(4), 1–30. https://doi.org/10.1080/00220380410001673175
24. Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *˜ the œReview of Economics and Statistics*, *49*(1), 92. https://doi.org/10.2307/1937887
25. Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T. S. (2019). Temporal Relational Ranking for Stock Prediction. *ACM Transactions on Office Information Systems*, *37*(2), 1–30. https://doi.org/10.1145/3309547
26. Hafezi, R., Shahrabi, J., & Hadavandi, E. (2015). A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. *Applied Soft Computing*, *29*, 196–210. https://doi.org/10.1016/j.asoc.2014.12.028
27. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013
28. Hidayat, S., Rismayati, R., Tajuddin, M., & Merawati, N. L. P. (2020). Segmentation of university customers loyalty based on RFM analysis using fuzzy c-means clustering. *Jurnal Teknologi Dan Sistem Komputer*, *8*(2), 133–139. https://doi.org/10.14710/jtsiskom.8.2.2020.133-139
29. Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, *13*(2), 947–958. https://doi.org/10.1016/j.asoc.2012.09.024
30. Kumar, S. (2020). Use of cluster analysis to monitor novel coronavirus-19 infections in Maharashtra, India. *Indian Journal of Medical Sciences/Indian Journal of Medical Sciences (Print)*, *72*, 44–48. https://doi.org/10.25259/ijms_68_2020
31. Kusi, A., Opata, C. N., & Narh, T. W. J. (2015). Exploring the Factors That Hinder the Growth and Survival of Small Businesses in Ghana (A Case Study of Small Businesses within Kumasi Metropolitan Area). *American Journal of Industrial and Business Management*, *05*(11), 705–723. https://doi.org/10.4236/ajibm.2015.511070
32. Leonidou, L. C., Katsikeas, C. S., & Samiee, S. (2002). Marketing strategy determinants of export performance: a meta-analysis. *Journal of Business Research*, *55*(1), 51–67. https://doi.org/10.1016/s0148-2963(00)00133-8
33. Li, J., Greenwood, D., & Kassem, M. (2019). Blockchain in the built environment and construction industry: A systematic review, conceptual models and practical use cases. *Automation in Construction*, *102*, 288–307. https://doi.org/10.1016/j.autcon.2019.02.005
34. Li, X., & Wong, K. C. (2019). Evolutionary Multiobjective Clustering and Its Applications to Patient Stratification. *IEEE Transactions on Cybernetics*, *49*(5), 1680–1693. https://doi.org/10.1109/tcyb.2018.2817480
35. Lugner, M., Gudbjörnsdottir, S., Sattar, N., Svensson, A. M., Miftaraj, M., Eeg-Olofsson, K., Eliasson, B., & Franzén, S. (2021). Comparison between data-driven clusters and models based on clinical features to predict outcomes in type 2 diabetes: nationwide observational study. *Diabetologia*, *64*(9), 1973–1981. https://doi.org/10.1007/s00125-021-05485-5
36. Maulina, N. R., Surjandari, I., & Rus, A. M. M. (2019). *Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering*. https://doi.org/10.1109/icsssm.2019.8887739
37. Midi, H., Sarkar, S., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics/Journal of Interdisciplinary Mathematics*, *13*(3), 253–267. https://doi.org/10.1080/09720502.2010.10700699

38. Mouna, A., & Jarboui, A. (2021). Understanding the link between government cashless policy, digital financial services and socio-demographic characteristics in the MENA countries. *International Journal of Sociology and Social Policy*, *42*(5/6), 416–433. https://doi.org/10.1108/ijssp-12-2020-0544
39. Nurfaizah, N., & Fathuzaen, F. (2021). Clustering Customer Data Using Fuzzy C-Means Algorithm. *Penelitian Ilmu Komputer Sistem Embedded & Logic/Penelitian Ilmu Komputer Sistem Embedded and Logic*, *9*(1), 1–14. https://doi.org/10.33558/piksel.v9i1.2359
40. Nwoke, L. I. (2016). The Psychological Impact of Live Broadcasting On Mental Health: A Comparative Study of Radio And Television Presenters. (2016). *International Journal of Scientific Research and Management (IJSRM)*, *4*(9), 4636-4646. https://doi.org/10.18535/**ijsrm/v4i9.21**
41. Nwoke, L. I. (2017). Social Media Use and Emotional Regulation in Adolescents with Autism Spectrum Disorder: A Longitudinal Examination of Moderating Factors. *International Journal of Medical Science and Clinical Invention*, *4*(3), 2816–2827. Retrieved from https://valleyinternational.net/index.php/ijmsci/article/view/2555
42. Nwoke, L. I., Precious, A. O., Aisha, A., & Peter, S. (2022). The Impact of Cashless Policy on the Performance of Msmes in Nigeria Using Artificial Neural Network. International Journal of Social Sciences and Humanities Invention, 9(08), 7182–7193. https://doi.org/10.18535/ijsshi/v9i08.09
43. Oborn, E., Barrett, M., Orlikowski, W., & Kim, A. (2019). Trajectory Dynamics in Innovation: Developing and Transforming a Mobile Money Service Across Time and Place. *Organization Science*, *30*(5), 1097–1123. https://doi.org/10.1287/orsc.2018.1281
44. Palit, A., & Rahut, D. (2024). Digital Transformation for Inclusive and Sustainable Development in Asia. In *Asian Development Bank Institute eBooks*. https://doi.org/10.56506/hsdc4319
45. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018
46. Razi, M., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems With Applications*, *29*(1), 65–74. https://doi.org/10.1016/j.eswa.2005.01.006
47. Richter, D., Ekman, M., & De Lange, F. P. (2018). Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream. ˜ *the œJournal of Neuroscience/˜ the œJournal of Neuroscience*, *38*(34), 7452–7461. https://doi.org/10.1523/jneurosci.3421-17.2018
48. Rizzo, G., Turkheimer, F., Keihaninejad, S., Bose, S., Hammers, A., & Bertoldo, A. (2012). Multi-Scale hierarchical generation of PET parametric maps: Application and testing on a [11C]DPN study. *NeuroImage*, *59*(3), 2485–2493. https://doi.org/10.1016/j.neuroimage.2011.08.101
49. Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. ˜ *the œJournal of Neuroscience/˜ the œJournal of Neuroscience*, *27*(9), 2349–2356. https://doi.org/10.1523/jneurosci.5587-06.2007
50. Sinambela, Y., Herman, S., Takwim, A., & Widianto, S. R. (2020). A Study of Comparing Conceptual and Performance of K-Means And Fuzzy C Means Algorithms (Clustering Method of Data Mining) Of Consumer Segmentation. *Jurnal Riset Informatika*, *2*(2), 49–54. https://doi.org/10.34288/jri.v2i2.116
51. Slade, E. L., Williams, M. D., & Dwivedi, Y. K. (2013). Mobile payment adoption: Classification and review of the extant

literature. *Marketing Review/˜ the œMarketing Review*, *13*(2), 167–190. https://doi.org/10.1362/146934713x13699019904687

52. Tomassen, P., Vandeplas, G., Van Zele, T., Cardell, L. O., Arebro, J., Olze, H., Förster-Ruhrmann, U., Kowalski, M. L., Olszewska-Ziąber, A., Holtappels, G., De Ruyck, N., Wang, X., Van Drunen, C., Mullol, J., Hellings, P., Hox, V., Toskala, E., Scadding, G., Lund, V., . . . Bachert, C. (2016). Inflammatory endotypes of chronic rhinosinusitis based on cluster analysis of biomarkers. ˆ *the ‰Journal of Allergy and Clinical Immunology/Journal of Allergy and Clinical Immunology/˜ the œJournal of Allergy and Clinical Immunology*, *137*(5), 1449-1456.e4. https://doi.org/10.1016/j.jaci.2015.12.1324
53. Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://isplab.tudelft.nl/sites/default/files/vandermaaten08a.pdf
54. Wang, H., Ni, X., Qing, X., Liu, L., Lai, J., Khalique, A., Li, G., Pan, K., Jing, B., & Zeng, D. (2017). Probiotic Enhanced Intestinal Immunity in Broilers against Subclinical Necrotic Enteritis. *Frontiers in Immunology*, *8*. https://doi.org/10.3389/fimmu.2017.01592
55. Weldeslassie, H. A., Vermaack, C., Kristos, K., Minwuyelet, L., Tsegay, M., Tekola, N. H., & Gidey, Y. (2019). Contributions of Micro, Small and Medium Enterprises (MSMEs) to Income Generation, Employment and GDP: Case Study Ethiopia. *Journal of Sustainable Development*, *12*(3), 46. https://doi.org/10.5539/jsd.v12n3p46
56. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems With Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020
57. Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, *99*, 238–248. https://doi.org/10.1016/j.autcon.2018.12.016
58.