eSystems Engineering Society EST. 2007

International Journal of Data Science and Advanced Analytics

Journal homepage: www.ijdsaa.com

Leveraging Advanced Machine Learning Algorithms for Optimized Supplier Selection: A Multi-Criteria Decision-Making Approach

Nikhil Nan [1], Shatha Ghareeb[2] and Snehal Patel[2]
[1]Liverpool John Moores University, UK [2]Teesside University, UK

[1]nikhil.2019ex@iimu.ac.in, [2]s.ghareeb@tees.ac.uk, [2]e4202433@tees.ac.uk

Abstract- This study investigates the application of K-means clustering, a machine learning technique, to enhance supplier selection and management within large organizations. By segmenting suppliers based on key performance metrics, the research aims to establish more strategic supplier relationships. The methodology leverages purchase order data transformed for clustering analysis. K-means clustering is chosen for its effectiveness in grouping suppliers with similar characteristics. The analysis identifies four distinct supplier segments: high-ranked, steady growth, high spending, and rapid growth. These segments become the foundation for developing targeted supplier management strategies. The study explores practical implications like retention programs for high-ranked suppliers, growth support for steady performers, cost management for high spenders, and investment opportunities for rapidly growing suppliers. Additionally, it discusses supplier rationalization and the advantages of data-driven decision-making in supplier relationship management. The concluding remarks emphasize the effectiveness of K-means clustering for supplier segmentation. The research offers a framework for optimizing supplier selection and management, with the potential to support improved efficiency, stronger supplier relationships, and better business outcomes through more targeted and data-driven supplier management.

Keywords- Supplier selection; Supplier management; Clustering algorithms; K-means clustering; Unsupervised learning; Decision support systems; Supply chain analytics; Data-driven decision making.

1. Introduction

Supplier Relationship Management (SRM) is central to procurement performance and supply chain resilience, particularly in large organizations where supplier portfolios are extensive and heterogeneous [1]. Effective supplier selection and engagement influence operational reliability, cost control, and long-term relationship outcomes, and empirical research has shown that structured supplier evaluation positively affects relationship success and firm performance [1]. However, traditional supplier selection approaches often rely on manual assessments and heuristic scoring systems,

https://doi.org/10.69511/ijdsaa.v7i2.333
Received 21 June 2025; Received in revised form 14 August 2025; Accepted 27 September 2025.
Available online 03 October 2025.

which can introduce subjectivity, inconsistency, and scalability challenges as transaction volumes increase.

To address these limitations, multi-criteria decision-making (MCDM) frameworks have been widely adopted to incorporate multiple performance dimensions such as quality, delivery, cost, and sustainability into supplier evaluation processes [2, 3]. Although MCDM methods provide structured analytical support, they frequently depend on expert-driven weighting schemes and repeated manual assessments, which may limit efficiency and introduce bias in large-scale environments. Concurrently, machine learning (ML) techniques have emerged as powerful tools for supplier-related decision support, including supervised classification and ranking models for supplier evaluation [4, 5]. Systematic reviews indicate a growing integration of quantitative analytics in supplier management; however, practical implementation challenges remain, particularly regarding scalability, consistency, and automation across large supplier bases [2].

Supplier segmentation offers a complementary strategy by grouping suppliers with similar behavioral and performance characteristics to enable differentiated SRM policies. Portfolio-based segmentation models have long been discussed in the literature [6], and taxonomic research highlights how segmentation supports strategic alignment between supplier management approaches and market conditions, relational factors, and performance attributes [7]. Despite these theoretical advancements, segmentation practices in many organizations remain qualitative and judgment driven. Prior studies demonstrate that ML can enhance objectivity in supplier segmentation by automating categorization using transactional and e-invoice data [8, 9].

Motivated by these developments, this study applies K-means clustering (KM) as an unsupervised learning method to segment suppliers using supplier-level features derived from procurement transactions. By transforming routine procurement data into analytically meaningful variables and evaluating cluster validity using internal metrics, the research demonstrates how scalable, data-driven segmentation can support systematic differentiation across supplier groups. The study contributes to existing supplier evaluation and segmentation research by operationalizing unsupervised clustering for SRM and by providing an empirically grounded framework that directly links procurement analytics to actionable supplier management strategies.

2. Materials and Methods

This section describes in detail the procedures followed for data acquisition, preprocessing, feature engineering, clustering implementation, and cluster interpretation.

2.1 Research Design

The study adopts a quantitative analytical design supported by structured interpretation of clustering outcomes. The primary objective is to segment suppliers using an unsupervised ML approach and subsequently translate the resulting segments into actionable SRM strategies. The methodological workflow consists of the following stages: (i) data acquisition, (ii) data cleaning and preprocessing, (iii) feature engineering, (iv) clustering using KM, (v) cluster validation, and (vi) interpretation and strategic mapping. The use of ML in supplier-related decision support is supported in prior research in classification and evaluation contexts [1, 4, 10].

2.2 Data Source and Study Context

The dataset used in this study was obtained from the State Contract and Procurement Registration System (SCPRS). SCPRS provides centralized procurement records for purchases exceeding $5,000 and contains supplier-level and transaction-level information. The threshold of $5,000 is a type of characteristic of the SCPRS data source as highlighted in previous studies. By design, lower-value transactions are excluded from the registry, ensuring that the dataset primarily captures higher-impact purchasing behavior relevant to supplier relationship management. The dataset covers three fiscal years: 2012–2013, 2013–2014, and 2014–2015. The raw data includes supplier identifiers, department information, United Nations Standard Products and Services Code (UNSPSC) classifications, and purchase order transaction values. Four primary tables were used: Supplier, Department, UNSPSC, and Purchase Order. Purchase orders containing multiple UNSPSC codes were simplified by retaining only the first code to maintain consistent categorization. This ensured uniform aggregation across suppliers.

2.3 Data Preparation and Cleaning

Data preparation was conducted to ensure analytical suitability and reduce noise introduced by formatting inconsistencies or missing values.

The following steps were implemented:

- Missing values in the “Total Price” field were replaced with zero prior to aggregation [9].
- Non-numeric characters (e.g., commas) were removed from monetary fields before converting them into numeric format.
- Supplier-level aggregation was performed using pivot tables to compute total spend per supplier.
- Additional pivot tables were generated to compute spend per segment title and fiscal year.

Data normalization was applied to continuous variables to ensure comparability across features. Total spend and percentage growth variables were standardized to zero mean and unit variance. The binary variable “Is Number One Supplier” was scaled consistently to prevent dominance during clustering. The importance of structured preprocessing and feature preparation in supplier analytics has been emphasized in prior ML applications in procurement [1, 4].

2.4 Feature Engineering

Supplier-level features were derived from aggregated procurement data to capture financial importance, category breadth, dominance, and temporal change. These variables provide a financially grounded representation of supplier importance and enable the identification of strategic patterns within procurement behaviour.

The engineered features include:
- Total Spend: Aggregated monetary value per supplier across the study period.
- Unique Segment Titles: Number of distinct procurement categories associated with each supplier.
- Is Number One Supplier: Binary indicator identifying whether a supplier is the top spender within its primary category.
- Total Increase % (Year 1 to Year 3): Percentage change in supplier spend from the first to the third fiscal year.
- Rank: Supplier rank within its highest-spend category.

These features were selected based on their analytical relevance and consistency with prior supplier evaluation and segmentation research [1, 4, 9, 10].

2.5 Clustering Method

Supplier segmentation was performed using the KM clustering algorithm. KM was selected due to its computational efficiency, interpretability, and suitability for partitioning suppliers based on numerical similarity.

The clustering procedure involved:
- Constructing a supplier-level feature matrix.
- Standardizing all continuous features.
- Running K-means for multiple values of K.
- Determining the optimal number of clusters using the Elbow Method.

The Elbow Method evaluates within-cluster sum of squared errors (SSE) across increasing K values. The inflection point observed at K = 4 indicated an appropriate balance between compactness and interpretability.

2.6 Cluster Quality Evaluation

Cluster validity was assessed using three internal evaluation metrics:
- Inertia (SSE): Measures within-cluster compactness.
- Silhouette Score: Evaluates cohesion and separation.
- Davies–Bouldin Index: Measures average inter-cluster similarity.

These metrics collectively ensure that clusters are compact, well-separated, and structurally meaningful. The use of quantitative validation measures aligns with best practices in machine learning-supported supplier evaluation research [10].

2.7 Interpretation and Strategic Mapping

Cluster interpretation was conducted using centroid analysis. Centroids represent the mean standardized feature values for suppliers within each cluster and enable direct comparison across segments.

The interpretation phase involved:
- Examining dominant features within each cluster.
- Comparing relative differences in rank, spend, and growth.
- Mapping cluster characteristics to SRM strategy categories.

This structured interpretation enables the transition from analytical segmentation to managerial action, consistent with prior supplier segmentation frameworks [9].

2.8 Tools and Implementation Environment

All data processing, feature engineering, clustering, and validation procedures were implemented using Python 3.0. ML operations were executed using Scikit-learn, while data manipulation and visualization were performed using Pandas and Matplotlib. All steps were documented to ensure transparency and reproducibility.

3. Analysis

This section presents a structured analytical examination of the procurement dataset and the application of clustering techniques to segment suppliers. The objective of this analysis is to transform raw transactional procurement data into meaningful supplier-level insights that support systematic supplier segmentation and strategic SRM.

Supplier segmentation has been widely recognized as a strategic tool for aligning supplier management practices with performance characteristics and relational priorities [6, 7]. However, segmentation approaches in practice are often qualitative and based on managerial judgment.

Recent research emphasizes the increasing role of ML techniques in supplier evaluation and segmentation to enhance objectivity and scalability [4, 8, 9]. In line with these developments, this paper applies KM clustering to identify natural groupings within the supplier base.

3.1. Data Exploration

Data exploration was conducted to understand the structure, characteristics, and suitability of the procurement dataset for clustering analysis. The dataset was obtained from SCPRS and covers fiscal years 2012–2013, 2013–2014, and 2014–2015. It includes purchase order records categorized using the UNSPSC.

3.2 From Original Data to Derived Dataset

To prepare the data for clustering, several preprocessing and transformation steps were performed:

- Data Loading: The procurement dataset was loaded into a structured data frame for processing.
- Handling Missing Values: Null values in the “Total Price” column were replaced with zero to ensure completeness during aggregation.
- Data Type Conversion: The “Total Price” field, originally stored as text due to formatting characters, was converted into a numeric format to enable accurate computation.
- Aggregation and Pivot Tables: Pivot tables were constructed to compute: Total spend per supplier; total spend per supplier per segment title; supplier-level summaries across fiscal years.
- Metric Derivation: The following supplier-level features were derived: Total Spend; Unique Segment Titles; Is Number One Supplier; Rank within top spend category; Total Increase % (Year 1 to Year 3). These metrics align with supplier evaluation dimensions emphasized in prior supplier selection research, including financial importance, performance dominance, and growth trajectory [2, 3, 4].
- Fiscal Year Mapping: Fiscal years were standardized to Year 1, Year 2, and Year 3 to allow consistent temporal comparison.
- Spend Growth Calculation: Year-over-year and cumulative spend increase percentages were computed to capture dynamic changes in supplier importance.

The resulting derived dataset provided a structured supplier-level feature matrix suitable for clustering analysis.

3.3 Clustering Analysis

Clustering analysis was conducted using the KM algorithm to segment suppliers based on similarity across selected performance features. KM was selected due to its computational efficiency, scalability, convergence properties, and interpretability in segmentation contexts [5, 9].

3.3.1 Determining the Optimal Number of Clusters

The Elbow Method was applied to determine the appropriate number of clusters. The SSE was computed for increasing values of K and plotted to identify the point at which marginal improvement diminished. As illustrated in Figure 1, a noticeable inflection point occurs at four clusters, where the reduction in SSE begins to stabilize. This indicates that K = 4 provides an appropriate balance between model compactness and interpretability.

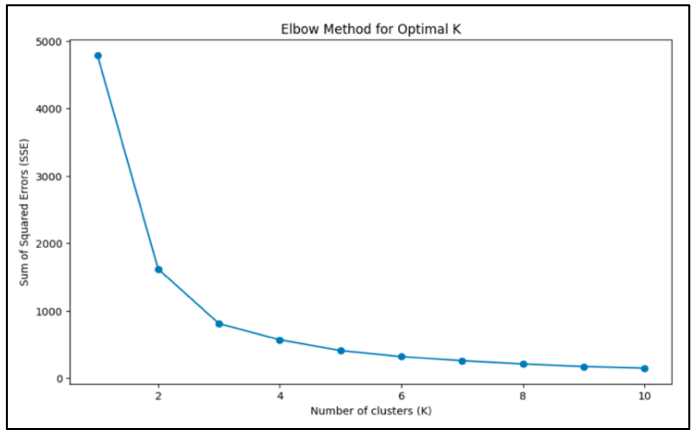

Fig. 1 Elbow method graph of SSE against number of clusters.

3.3.2. Cluster Assignment

Using K = 4, the KM algorithm was applied to the standardized supplier feature matrix. Each supplier was assigned to one of four clusters based on minimum Euclidean distance to cluster centroids. The clustering process produced four distinct supplier groups characterized by differences in rank, spend growth, and financial importance.

3.3.3. Comparative Clustering Analysis

To further validate this choice, a post-hoc comparative evaluation was conducted applying four clustering algorithms to the same standardised feature set (n = 13,625 suppliers with complete data) (Table 2).

Table 1 Comparative clustering algorithm evaluation

Algorithm	Silhouette Score	Davies-Bouldin Index	No. of Clusters
K-Means (K=4) *	0.69	0.61	4

Hierarchical Clustering (Ward, K=4)	0.42	0.80	4
Gaussian Mixture Model (K=4)	0.34	2.00	4
DBSCAN (eps=0.8, min_samples=5)	0.79	0.18	2 (+ noise)

* The algorithm selected for this study. Higher Silhouette Score indicates better-defined clusters (range: -1 to 1). Lower Davies-Bouldin Index indicates greater cluster separation. Metrics computed on a random sample of n = 3,000 suppliers. DBSCAN metrics are computed on non-noise points only.

K-Means (K=4) achieves the strongest balance of clustering quality and practical utility among the algorithms evaluated for four segments, with a Silhouette Score of 0.69 and Davies-Bouldin Index of 0.61. Hierarchical Clustering (Ward) and Gaussian Mixture Models produced notably weaker separation metrics (Silhouette: 0.42 and 0.34 respectively), with GMM's performance attributed to the highly right-skewed, non-Gaussian distribution of procurement spending values. While DBSCAN achieved strong internal metrics, it produced only two clusters across all well-parameterised configurations, which is insufficient to support the four-tier SRM strategy framework. These results confirm that K-Means with K=4 is the most appropriate choice for this dataset and research objective [5, 9, 12].

3.4 Evaluation of Cluster Quality

To assess the robustness and validity of the clustering solution, three internal evaluation metrics were used:

- Inertia (SSE)
- Silhouette Score
- Davies–Bouldin Index

These metrics are commonly used to evaluate cluster compactness and separation in segmentation studies [5].

The values used for SSE, Silhouette score, and Davies-Bouldin Index were 568.67, 0.51 and 0.59 respectively. These metrics confirm that the clustering solution is compact, well-separated, and analytically meaningful. The segmentation therefore provides a reliable structural representation of supplier behavior within the dataset.

3.5 Analysis of Cluster Centroids

Cluster centroids were computed to understand the defining characteristics of each supplier segment. The centroid values represent the mean standardized feature values within each cluster. The centroid visualization presented in Figure 2 highlights differences across clusters for total spend, rank, supplier dominance, and spend growth.

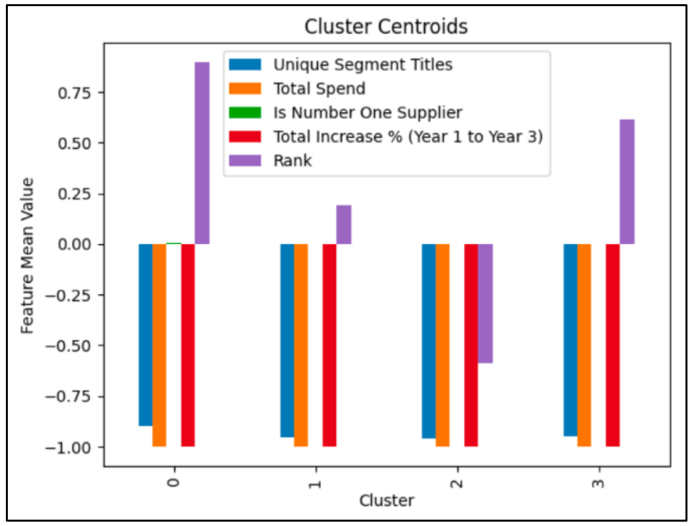

Fig. 2 Cluster centroids for each of the four clusters.

In this respect, cluster 0 showed high positive rank, negative spend growth, average total spend and fewer unique segment titles. Cluster 1 demonstrated slightly positive rank, moderate performance and near-average growth. Cluster 2 showed positive spend growth, moderate rank and financially relevant suppliers. In addition, cluster 3 showed the highest spend growth, high rank and strong positive performance trend.

4. Results and Discussion

The results highlight characteristics of each cluster and explore their practical and strategic implications. The results demonstrate how clustering can be an effective tool for supplier segmentation and rationalization, ultimately leading to more informed and strategic decision-making in supplier management.

4.1 Analysis of Clusters

The characteristics of each cluster as depicted in the centroid graph are analyzed. Further, the practical implications and strategic supplier management strategies for each cluster are explored. The centroid values indicate the average feature values for suppliers within each cluster, providing insights into the defining parameters and how they can be utilized for strategic decision-making.

Key defining parameters for clusters are rank, total spend, and the following features 'unique segment titles, is number one supplier and total increase %'. The rank is significantly high for clusters 0 and 1 and moderate for clusters 2 and 3. The total spend is high for cluster 2, moderate for clusters 0 and 3, and low for cluster 1. The remaining features including 'unique segment titles, is number one supplier and total increase %' are average or high for cluster 3; yet, average or low for clusters 0, 1 and 2.

The clusters' practical implications for clusters 0, 1, 2 and 3 were likely delivering high quality service, having potential for growth, having substantial spending, and significant increase in spending and capabilities.

Utilization and strategies across clusters varied between clusters. Cluster 0 shows retention and development, strategic partnerships, and performance monitoring. Cluster 1 demonstrates growth support, incentive programs and collaborations. Cluster 2 highlights cost management, value extraction, and performance improvement. Cluster 3 demonstrates investment and support, strategic alliance, and growth monitoring.

4.3 Overall Implications

The analysis of the clusters provides understanding of the supplier base, allowing businesses to tailor their strategies for each cluster. By leveraging the insights gained from clustering, businesses can optimize supplier selection processes, enhance SRM, and achieve cost savings. Key strategies include:

- Retention and Development: Prioritize high-performing suppliers for retention and development.
- Incentive Programs: Implement incentive programs to encourage growth and improvement in lower-performing clusters.
- Cost Management: Focus on cost management and value extraction from high-spending clusters.
- Strategic Partnerships: Form strategic partnerships and alliances to leverage the strengths of each cluster.

By adopting these strategies, businesses can strengthen supplier management practices through more differentiated and evidence-informed decision-making. While the present study does not directly measure post-implementation business outcomes, the identified clusters demonstrate meaningful structural differences that provide a practical basis for targeted supplier relationship management. This interpretation is consistent with prior research showing that structured supplier selection and buyer–supplier engagement are associated with improved relationship performance and stronger firm-level outcomes [1]. In a similar vein, supplier segmentation studies have shown that data-driven categorisation can improve procurement focus and managerial prioritisation [9]. Accordingly, the present framework should be understood as supporting more effective SRM and operational decision-making, while direct longitudinal validation of business outcomes remains an important avenue for future research.

This paper provides an in-depth analysis of the clustering results, examining the characteristics and strategic implications of each cluster. The KM clustering algorithm allows segmenting the suppliers into distinct groups, each with unique features and performance metrics.

Four clusters with distinct characteristics are identified. Cluster 0 is defined by high ranks, indicating top-performing suppliers. Cluster 1 shows a steady increase in spending, suggesting potential for growth. Cluster 2 is characterized by substantial spending, making these suppliers significant from a financial perspective. Cluster 3 demonstrates significant growth in spending, indicating expansion and increasing capabilities.

The analysis shows that clustering can be a powerful tool for supplier segmentation and rationalization. By understanding the key characteristics of each cluster, businesses can develop targeted strategies for supplier management. High-performing clusters can be prioritized for retention and development, while lower-performing clusters can be considered for renegotiation or elimination.

Additionally, the evaluation metrics confirm the robustness and validity of the clusters, providing a reliable basis for strategic decision-making. The practical applications discussed, such as cost management, value extraction, and strategic partnerships, highlight the benefits of a data-driven approach to supplier management.

5. Conclusions

The clustering analysis performed in this study yields significant insights into the characteristics and performance metrics of suppliers, allowing for their segmentation into meaningful groups. The key findings from this analysis show that suppliers are segmented into four distinct clusters, each with unique characteristics and strategic implications. The evaluation metrics confirm the validity and reliability of the clusters, providing a solid foundation for strategic decision-making in supplier management. By understanding the key features and implications of each cluster, businesses can develop targeted strategies to optimize supplier selection, enhance SRM, and achieve cost savings. The insights gained from this analysis are instrumental in guiding future research and practical applications in supplier management and provide a structured basis for more effective supplier segmentation, targeted SRM strategies, and improved managerial focus.

References

[1] V. R. Kannan and K. Tan, "Buyer–supplier relationships: The impact of supplier selection and buyer-supplier engagement on relationship and firm performance," *International Journal of Physical Distribution & Logistics Management,* vol. 36, no. 10, pp. 755-775, 2006.

[2] W. Ho, X. Xu and P. K. Dey, "Multi-criteria decision making approaches for supplier evaluation and selection: A literature review.," *European Journal of Operational Research,* vol. 202, no. 1, pp. 16-24, 2010.

[3] B. Aslani, M. Rabiee and M. Tavana, "An integrated information fusion and grey multi-criteria decision-making framework for sustainable supplier selection," *International Journal of Systems Science: Operations and Logistics,* vol. 8, no. 4, pp. 348-370, 2021.

[4] M. R. Ali, S. M. A. Nipu and S. A. Khan, "Ali, M. R., Nipu, S. M. A., and Khan, S. A. (2023). A decision support system for classifying supplier selection criteria using machine learning and random forest approach. Decision Analytics Journal. https://doi.org/10.1016/j.dajour.2023.100238," *Decision Analytics Journal,* 2023.

[5] M. S. Kıran, E. Eşme, B. Torğul and T. Paksoy, "Supplier Selection with Machine Learning Algorithms," in *Logistics 4.0*, CRC Press, 2020, pp. 103-125.

[6] H. F. Lajimi and S. Majidi, "Supplier segmentation: A systematic literature review," *Journal of Supply Chain Management Science,* vol. 2, no. 3-4, pp. 138-158, 2021.

[7] M. Day, G. M. Magnan and M. M. Moeller, "Evaluating the bases of supplier segmentation: A review and taxonomy," *Industrial Marketing Management,* vol. 39, no. 4, pp. 625-639, 2010.

[8] J.-s. Hong, H. Yeo, N.-W. Cho and T. Ahn, "Identification of core suppliers based on e-invoice data using supervised machine learning," *Journal of Risk and Financial Management,* vol. 11, no. 4, p. 70, 2018.

[9] B. Bahameish, "Supplier Segmentation Method Using Supervised Machine Learning: A Case Study of Qatar Foundation," Doctoral dissertation, Hamad Bin Khalifa University, 2023, Qatar.

[10] M. M. Khan, I. Bashar, G. M. Minhaj, A. I. Wasi and N. U. I. Hossain, "Resilient and sustainable supplier selection: an integration of SCOR 4.0 and machine learning approach," *Sustainable and Resilient Infrastructure,* pp. 1-17, 2023.

[11] J. H. Dyer, D. S. Cho, and W. Chu, "Strategic supplier segmentation: The next 'best practice' in supply chain management," *California Management Review,* vol. 40, no. 2, pp. 57–77, 1998.

[12] F. Rahiminia, A. Doniavi, and M. Jafari, "Cluster-based supplier segmentation: a sustainable data-driven approach," *Modern Supply Chain Research and Applications,* vol. 5, no. 3, pp. 209–225, 2023.