Intelligent Systems Research and Applications Journal

www.israj.org
eISSN: 3120 - 0826

A Novel Decomposed Pythagorean Fuzzy CODAS Framework for Precision Customer Segmentation in Complex Market Environments

Safiye Turgay [1,*], Bilal Torkul [2], Abdulkadir Aydin [1], Furkan Özyurt [1]

[1] Department of Industrial Engineering, Faculty of Engineering, Sakarya University, Sakarya, Turkey
[2] Office Services and Secretarial Department, Çınarcık Vocational School, Yalova University, Yalova, Turkey

ARTICLE INFO

Article history:
Received 3 February 2026
Received in revised form 21 March 2026
Accepted 29 March 2026
Available online 2 April 2026

Keywords:
Multi-Dimensional Segmentation; K-Means Clustering, Decision Trees; Association Rule Mining; CODAS Method; Novel Decomposed Pythagorean Fuzzy Sets (NDPFS)

ABSTRACT

Segmentation of customers has become an essential aspect in market analysis as it allows the organizations to formulate strategies according to different customer behaviour. These processes are further enhanced with advanced data mining techniques such as those used to uncover hidden patterns and take into account the heterogeneity of customers, which result in better and more data-driven decisions. To evaluate the quality of the segmentation-in particular for accuracy and practical usefulness of the solution, we take into account the performance measures Silhouette Score and Davies–Bouldin Index for accuracy and cohesion of the solution and Market Response. The proposed framework integrates the strong data mining procedures with New Disaggregated Pythagorean Fuzzy Sets based CODAS (NDPFS–CODAS) technique to increase the flexibility and discriminative power as well as the clearness of resultant segmentation. On the basis of behavioral, demographic and psychographic criteria, the model incorporates a multidimensional uncertainty. Case study confirms that NDPFS-CODAS enhances the accuracy and robustness of conventional fuzzy technique. These findings confirm its usefulness in target marketing, new-product development and customer and product in section marketing strategy, and give guidance for strategic decisions to offer the product packaging with maximum optimization.

1. Introduction

Nowadays, the customer purchase behavior and preference have become one of the key factors for business development among the keen competitions. Customer segmentation is the process of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, can help companies enhance marketing, increase customer service, and simplify product offerings. However, conventional customer segmentation approaches are based on simplified attributes like demographics or purchase frequency, which may fail to reflect the diversity and complexity of consumer demands.

* *Corresponding author.*
E-mail address: safiyeturgay2000@yahoo.com

https://doi.org/10.59543/r8xsq450

Preference and behavior of the customers and their buying frequency therein play a crucial role in forming a profitable business in the contemporary market scenario. The act of splitting up a diverse customer base into smaller groups of similar customers enables companies to better cater to the needs of its individuals with regards to product, service and communication. Traditional segmentation methods (e.g. demographic and psychographic profiles) are inadequate for capturing the subtlety and uncertainty of aggregated high-dimensional consumer information.

With information being generated exponentially and data mining technology appearing, now businesses have abundant and multi-dimensional information about customers. They include age, past buying behavior, browsing online habits, social networks following, location and so on. Multi-dimensional segmentation exploits the richness of data in multiple dimensions to segment customer groups according to multiple facets of the customers.

In this paper, the focus is on how to apply data mining techniques to multidimensional customer segmentation for gaining more insights into consumers' behavior and choice. We apply data mining techniques of k-means clustering, decision trees and association rule mining to segment customers on multiple dimensions that covers transactional details to behavioural orientations. When multiple variables can be considered in the segmentation, the resulting segment profile can be more descriptive of the target market and this allows companies to targeting more tightly and develop more specific marketing strategies.

Information system tools and techniques of data mining including clustering algorithm and association rule mining have brought revolution in market research to uncover the blind pattern among big and high database [1], [2], [3]. Still, such procedures are generally not capable of processing uncertainty, vagueness and hesitancy within consumer opinions. To address this issue, fuzzy set theory has been more and more widely applied in decision making to effectively capture ambiguous human judgment and intricate relationships [4], [5]. Pythagorean fuzzy sets have been proved to be a very effective instrument among all types of extensions of fuzzy sets, with an additional capability of expressing the degree of membership and the degree of non-membership in uncertain circumstances. From this perspective, Novel Decomposed Pythagorean Fuzzy Sets (NDPFS) appear to be more expressive as they represent fuzzy information through more granular and more computational friendly components [6], [7]. When combined with multi-criteria decision-making (MCDM) methods such as the COmbinative Distance-based ASsessment (CODAS) method, NDPFS also create an efficient technique for sorting and comparing customer classes by use of diversified if not antagonistic standards. The goal of this study evaluates whether various data mining methods, which are able to uncover interesting customers segments, compare their relative performance on this task, and analyze how the identified segments could be applied to support marketing decision-making [8], [9], [10]. Our goal also shows that multi-dimensional market segmentation allows for more efficient marketing, higher customer satisfaction, and better resource allocation.

This paper presents a novel hybrid scheme based on data mining techniques and NDPFS–CODAS method to enhance customers' segmentation in market analysis. The proposed method uses data mining techniques to extract a set of initial customer clusters and patterns, which are then analyzed and ranked applying the NDPFS–CODAS method. Such combination makes the segmentation result more interpretable and accurate so as to reveal the uncertainty and the richness of data in the existing consumer datasets.

To demonstrate the inadequacy of existing models, focus on a retail case scenario centered on "borderline" customers— that is, "seasonal impulse buyers" who purchase often during very short periods but have little annual monetary value. Standard crisp clustering methods misclassify these buyers as low-value since their purchases are averaged, however classical Intuitionistic Fuzzy Sets

(IFS) models may not be adequate to model the high hesitation regarding their sporadic participation. The introduced Decomposed Pythagorean Fuzzy Sets (NDPFS) overcome this problem by broadening the membership space, thus capturing the hesitation degree existing in such mixed modes, which opens a window to more refined partitions to detect the promising high-growth targets.

This study makes three main contributions:

1. We propose a full framework based on data mining and fuzzy MCDM for state-of-the-art customer segmentation.

2. We propose the CODAS method based on new decomposed Pythagorean fuzzy sets to better treat the complexity of uncertainty in multi-criteria evaluation.

3. We empirically validate the proposed approach using a real-world retailing case and illustrate that the approach provides marketers with insights for focal marketing and strategic planning.

The rest of this paper is organised as follows: Section 2 reviews the related work on customer segmentation, data mining and fuzzy MCDM techniques. Section 3 describes the theoretical background of NDPFS and the CODAS technique, respectively. Section 4 describes the NDPFS–CODAS framework. The case study is presented in Section 5 followed by conclusion, discussion of limitations and future research in Section 6.

2. Literature Survey

Customer segmentation has always been the cornerstone of market research, which aims to divide a customer population into multiple homogeneous groups with similar characteristics. Over the past years, there is a wide range of techniques has been proposed, for customer segmentation, from simple statistical approaches to more sophisticated data mining and machine learning techniques. The following section discusses the literature, that includes established customer segmentation techniques, the influence of data mining techniques, and developing multi-dimensional segmentation.

Traditional customer segmentation would – in a very traditional way – sort consumers into groups based on demographic factors such as age, gender, income, and geography. RFM (recency, frequency, monetary) analysis is another popular technique and is mainly used in retail where customers are grouped according to their buying patterns. These are two good tools which can be helpful, but both have their limitations in addressing the age of the consumer and behavior today and are too limited. It is observed that a segmentation of the market on demographic basis alone lacks the psychological, behavioral and attitudinal factors of the consumers [11], [12], [13], [14]. Enter data mining techniques, and the capacity for segmentation increased. Clustering algorithms have been the focus of data-driven customer segmentation, which enables companies to discover natural groupings within high dimensional and large datasets of customers. Hierarchical as well as partitional clustering based on distance measures are among the most popular methods for customer segmentation, which is k-means, which is also the representative of the partitional clustering methods used most frequently in customer segmentation applications. The most common is k-means, due to its simplicity and the efficiency in grouping customers with similar attributes. However, it depends critically on the initial selection of centroids and also assumes that the clusters are spherical, thus may not be the most appropriate when dealing with real data [15], [16], [17], [18], [19].

The k-means was also modified by other authors and the density-based clustering method were combined to improve the clustering results for the case when clusters are not well separated [20], [21], [22]. Kohonen's Self-Organizing maps (SOM) found to be effective method for visualizing high dimensional data for customer's segmentation to enable firms derive knowledge from intricate,

multi-attribute datasets [23], [24], [25]. Decision tree algorithms such as CART(Classification and Regression Tree) and ID3 (Iterative Dichotomiser 3) are extensively used for segmentation as they can capture complex interactions among various customer profiles. CART—a nonparametric approach—recursively partitions the data by feature values into binary splits at each node. These rules make feasible for a company a simple comprehension of the segmentation process as well as the important variables that separate one customer cluster from another [26], [27], [28]. Another decision tree algorithm widely used to do customer segmentation is C4.5. It builds a tree based on which attribute best splits the data. Both CART and C4.5 have been used successfully in target marketing, churn prediction, and credit scoring based customer segmentation [29], [30], [31], [32]. Another key method in customer segmentation is the association rule mining which reveals the relation among several customer behaviors or attributes. Apriori and FP-growth are popular algorithms to mine frequent itemsets and association rules from the transaction data. Apriori algorithm played a vital role to be used in finding association in market basket data and have been used for customer preference and segmentation analysis [33], [34], [35].

Association rule mining can result in the identification of customer segments not only from demographic and behavioral data but also from transactions information, e.g. offer and discount details, revealing more profound customer behavioral patterns and trends. It is known that application of association rules mining on customer data results in more informed and focused business decisions in the area of product recommendations and marketing campaigns [36], [37], [38]. Multi-dimensional segmentation is a development of the traditional segmentation as it considers more dimensions other than demographics or previous purchases. It is a big data application that draws upon a plethora of diverse data sources, including Web behaviour, social media engagement, location intelligence and psychographics [39], [40]. Multi-dimensional segmentation provides a more comprehensive view of customer behavior that gives businesses the opportunity to create more precise and powerful fluid profiles [41], [42].

Recently, some studies have been done on multi-dimensional segmentation by advanced ML. For instance, SVMs (support vector machines) and RF (random forests) have shown to best capture complex relationships among different customer attributes [43], [44]. Deep learning-based algorithms, such as autoencoders and neural networks, are also utilized for segmentation, enabling enterprises to discover non-linear relations as well as to process massive amount of unstructured data [45], [46], [47]. Moreover, ensemble techniques can also help to improve the segmenting accuracy by taking advantage of multiple algorithms output. These methods are especially attractive for multi-dimensional segmentation problems, where the various data types (such as numerical, categorical, and text) can be combined in a single model [48], [49], [50]. While data mining techniques offer significant benefits for segmentation, there are still challenges. Some of the issues of relevance are data quality to with missing values, noise and biased samples. Imputation and outlier detection are popular techniques to treat these issues, but there is room for improvement. Modeling complexity – especially when applying sophisticated machine learning techniques such as deep learning [51], [52], [53], [54].

Fuzzy set theory was first proposed by Zadeh (1965) in order to deal with data uncertainty and inaccuracy, followed by the application of fuzzy logic in the decision making and consumer segmentation [55]. A traditional fuzzy set is represented by one membership function that expresses uncertainty. Nonetheless, deficiencies to deal with the entire spectrum of vagueness motivated further transformative expressions such as Intuitionistic Fuzzy Sets (IFS) (Atanassov, 1986), Pythagorean Fuzzy Sets (PFS) (Yager, 2013) and q-Rung Orthopair Fuzzy Sets [56], [57]. In particular, Pythagorean Fuzzy Sets (PFS) enable higher membership and non-membership degrees, thus they

provide a more relaxed modeling scheme in human-oriented decision making among incompleteness.

MCDM techniques are increasingly used in market research to solve multi criteria issues (profitability, loyalty, and customer interaction), needed since companies have several competitors. The well known MCDM techniques are TOPSIS, VIKOR, ELECTRE, and CODAS [58], [59], [60]. Among them, the CODAS (COmbinative Distance-based ASsessment) technique is gaining more and more attention since it distinguishes among options by two types of distances; Euclidean and Taxicab distances. CODAS provides a more discriminative ranking procedure, in particularly when fuzzy logic is used to treat uncertain decision data.

Just as the decomposition has been exploited in the theory of type-2 fuzzy sets, several works developed decomposed intutionistic fuzzy sets (DIFS) and decomposed Pythagorean fuzzy sets (DPFS) in order to improve both, transparency and complexity of computation. New decomposed Pythagorean fuzzy sets (NDPFS) generalize this idea even further on the one hand by disintegrating fuzzy information into several component fuzzy representations on the other hand by allowing for higher-order aspects of fuzzy information granules in the form of fuzzy vectors to be considered [61]. These are employed in areas like the evaluation of suppliers, project risk analysis, and disease diagnosis but their usage for customer segmentation is still comparatively less.

An approach for integrating data mining with fuzzy multi-criteria decision making (MCDM) methods is described which provides a pattern identification method embedded within uncertainty-aware decision analysis. A number of papers have studied such combinations (below) cluster results enhanced with fuzzy topped, or classification enhanced with fuzzy Vikor. But little has attempted integrating data mining results with the CODAS method in new decomposed models in pythagorean fuzzy set-ups—new decomposed models, new solutions [62]. This gap provides a big potential to design up dependable segmentation models that are not merely data- driven but also robust to impreciseness. This article addresses this limitation by acquiring a unified framework leveraging the strengths of both paradigms for advancing granularity, interpretability and accuracy of customer segmentation models in market study.

3. Methodology

The research approach systematically applies data mining techniques to perform multi-dimensional segmentation of customers in the analysis of market. The approach consists of the following phases: data acquisition, data cleaning and preprocessing, algorithm selection for segmentation, model assessment and results interpretation. All stages are intended to contribute in making the process of customer segmentation follow a wide spectrum of customer characteristics including demographic, behavioural and transactional information.

The data were obtained from a national retail firm having physical store and online instore customer contact. It contains data on transactions, demographics, behaviour of consumers while interacting with web site as well as on social media, and customer feedback on the campaigns of marketing. Demographics fields age, gender income level location. Behavior pages include Website Interactions, Visit Frequency, Page Visits, Click Through Rate. Frequency of purchase, recency of purchase, total spend and purchase history are the transaction data. Psychographic data represents customer interests (through surveys and social media usage etc.) and customer preferences. And, geographic information covers customer's address and distance from the store."

The input data need to be preprocessed for quality and consistency. Data Cleaning with handling of Missing values by imputation (mean/mode imputation for numerical and categorical data respectively) and outliers treatment to ensure the consistency of the data. Feature Scaling:

Continuous features such as income, expense are normalized using z-score normalization so that all features are on a similar scale and one feature does not bias the clustering process because it has a larger scale. Encoding categorical variables such as product type, location, and customer liking using one-hot encoding or label encoding when applicable. Dimensionality Reduction addresses the curse of dimensionality, in this paper we use PCA to eliminate redundancy and increase classification performance without sacrificing much information.

The algorithm selection based on the segmentation's multivariate nature. The study selects three well-known data mining algorithms for multivariate customer segmentation: K-means clustering, Decision Trees (CART) and Association Rule Mining. K-Means Clustering is an unimaginably popular unsupervised learning technique, K-Means is also applied to group customers on basis of common attributes. The choice of the algorithm is motivated by its simplicity, and scalability and the fact that it can be used with large datasets. The Elbow Method and Silhouette Score (in Figure 1) are used to determine the number of clusters to measure cluster cohesion and separation.

The Classification and Regression Trees (CART) technique is further employed to extract a set of rules based on split rules derived from features. This approach is particularly strong for both continuous and categorical segmentation variables, and the results can also be interpreted in an easy manner for companies to know the most important features of the respective segments. The Apriori method is used to generate association rules from transactional database. This enables the discovery of frequent itemsets (i.e., sets of products that customers routinely purchase together), and the sequences of customer purchase behavior patterns, which may contribute to improving the segmentation process. Association rules are applied to uncover hidden relationships between attributes and behaviours of customers that are not obvious within clusters obtained by clustering techniques.

Hybrid Approach (Ensemble Learning with K-Means and Decision Trees The former is also considered to be a hybrid that combines the benefits of both methods. Ensemble learning algorithms are leveraged to combine the outputs of several models and enhance the quality of segmentations. Random Forest is employed as an ensemble technique to generalize the nuclei segmentation and prevent overfitting.

Model Evaluation determines the quality of the segmentation models, some of which can be used to quantify metrics like Silhouette Score, Davies- Bouldin Index, Market Response Analysis, Cross Validation as depicted in Figure 1.

Measures the separation and cohesion of clusters Silhouette score is the better the score, the better separated and the better the clustering model. The Davies-Bouldin Index employed to measure clusters' compactness and separation with lower values interprets better clustering results. How well a marketing campaign can reach a segment is used to measure the effectiveness of each segment, known as Market Response Analysis. These are the conversion rates, the retention rates of customers and the Customer Lifetime Value (CLV) associated with each segment. Cross-Validation provides the robustness of the models, K-fold cross validation is used to test the reliability of the segmentation result across the data samples.

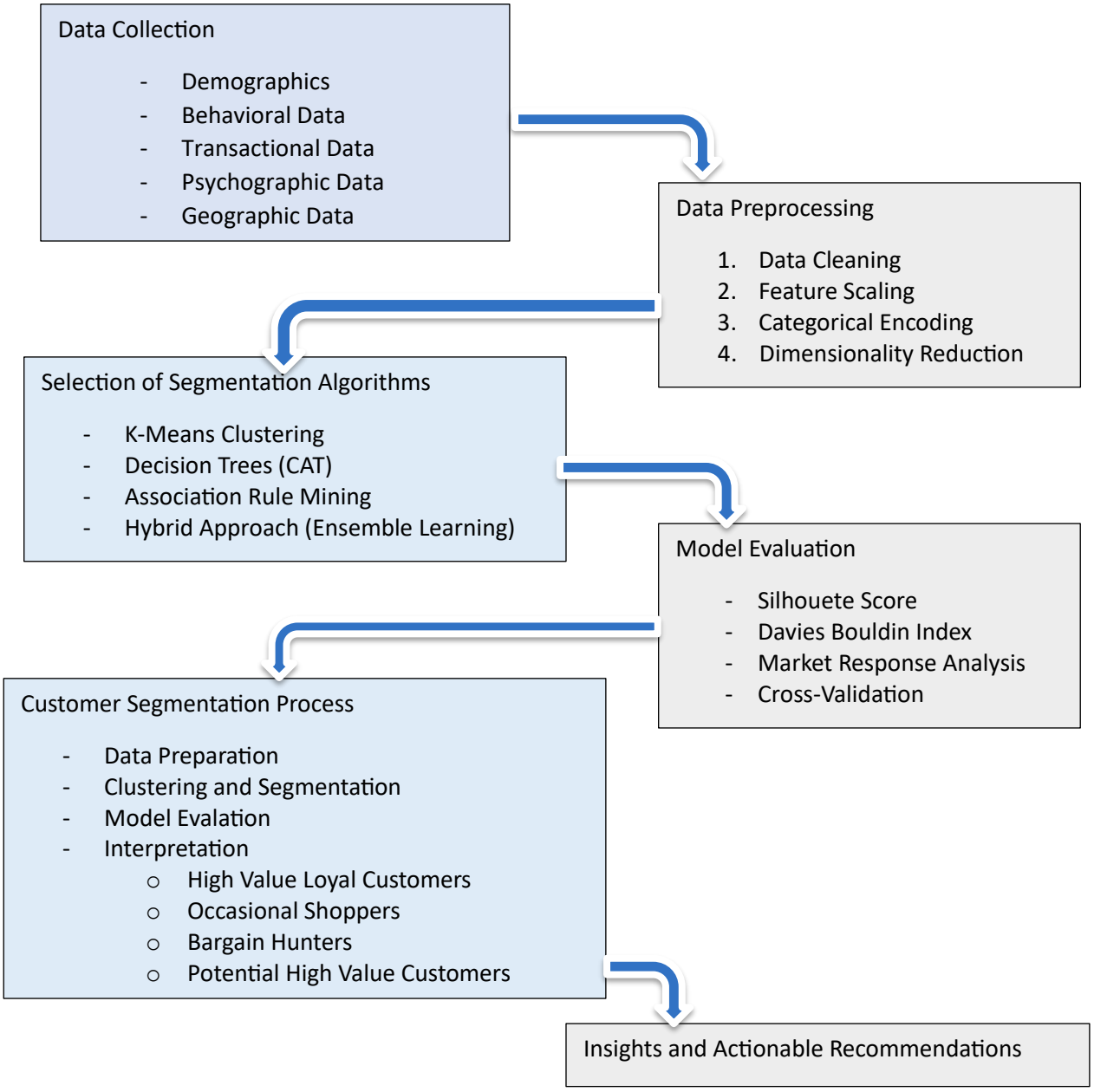

Fig. 1. Suggested model framework

Customer Segmentation Process the following steps are taken in order to perform the customer segmentation process showed in Figure 2.

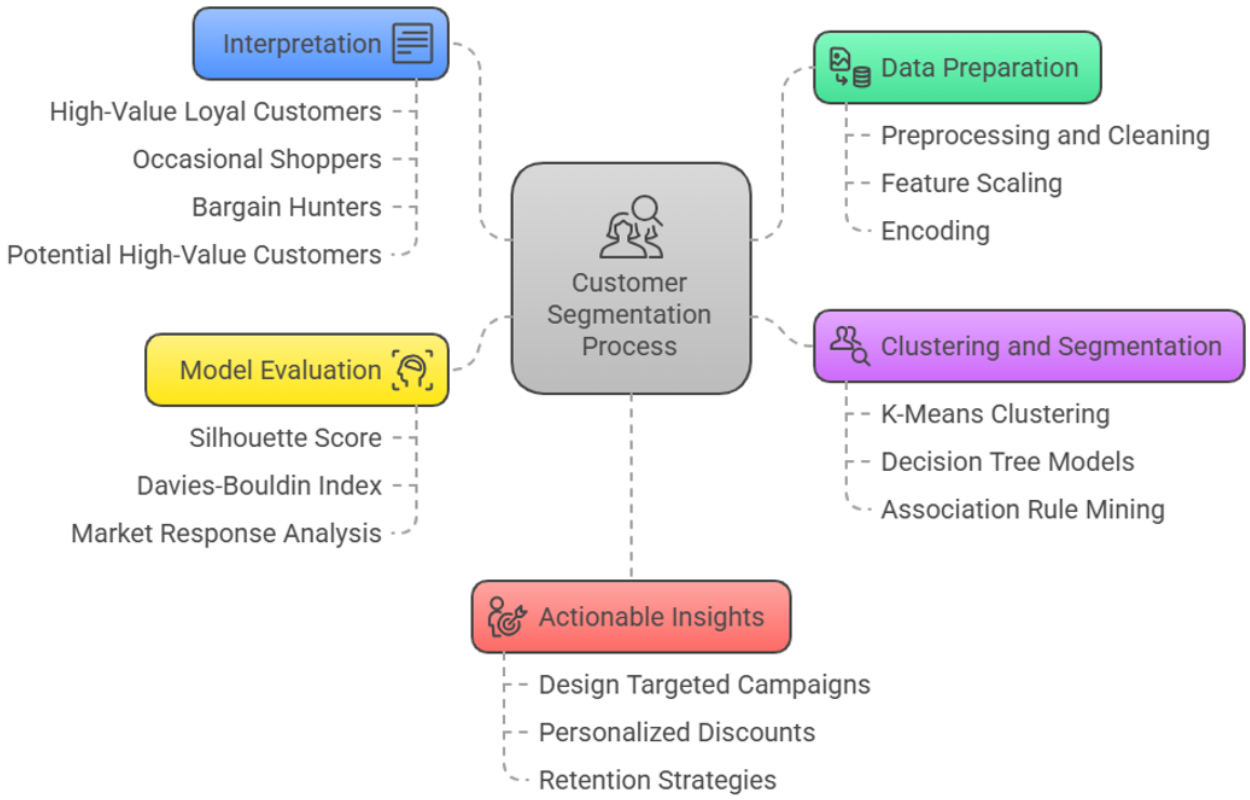

Fig. 2. Customer Segmentation Process

In this paper, we present a novel framework to exploit data mining techniques for multi-dimensional customer segmentation in an integrated manner. As a result of applying clustering, decision trees, and association rule mining, an enterprise can generate very fine customer segments and target its marketing campaigns such as directly. The combination of the methodologies for segmenting ensures that the process accounts for the complexity of contemporary consumer behaviour, while at the same time delivering decision-oriented business intelligence.

Multi-dimensional customer segmentation is a crucial element in market analysis that enables organizations to classify customers into multiple segments across various behavioral, demographic and psychographic factors. Classical clustering algorithms are not adequate for multi-dimensional data complexity. Machine learning techniques known as ensemble learning, which involves combining several individual models to improve the accuracy and reliability of predictions, holds potential for hybrid segmentation.

Segmenting customers refers to breaking up a customer base into smaller groups of customers that share common traits. Common clustering methods are K-means, hierarchical clustering and DBSCAN. Each of the method have drawbacks such as sensitivity to outliers or knowledge about the number of clusters in advance. Hybrid segmentation refers to the process of employing different segmentation strategies based on varying parameters, which can improve the quality of the segmentation by highlighting the strengths of individual approaches. It embraces the possibility to identify subtleties in several dimensions, whereas segmentation performed by one single method does not have this possibility.

3.1 Preliminaries

Ensemble learning refers to the technique which combines multiple models to enhance the overall prediction performance. Among the popular techniques are bagging, boosting, stacking and so on. In segmentation, ensemble learning can combine the result of a set of clustering algorithms to obtain more robust segment solutions. Ensemble learning for hybrid segmentation is the application of multiple segmenting models to increase the performance and the robustness of hybrid image segmentation.

Theorem (Conceptual)

Assume that $\mathcal{M} = \{\mathrm{M}_1, \mathrm{M}_2, \dots, \mathrm{M}_\mathrm{n}\}$ is a set of segmentation models, and $\mathrm{S}_\mathrm{i}(\mathrm{x})$ is the segmentation result of model $\mathrm{M}_\mathrm{i}$for data x. Aggregated ensemble output $\mathrm{S}_{\mathrm{ens}}(\mathrm{x})$ has the lowest expected segmentation error according to certain aggregation rules [11], [12].

Proof Sketch: Assuming that every M_i individually has an error rate ϵ_i, the probability that the ensemble output is correct grows with growing number of models n as long as $\epsilon_\mathrm{i} < 0.5$ for all i.

Proof: Let us have a simple majority voting ensemble for segmentation:

1. Let every model M_i have an error probability $\mathrm{P}(\mathrm{M}_\mathrm{i}\ \text{incorrect}) = \epsilon_\mathrm{i}$.

According to the law of large numbers, the more independent models are considered, the ensemble error converges to the majority model error rate.

If $\epsilon_\mathrm{i} < 0.5$, the majority rule is likely to yield a correct segmentation result.

Hence, ensemble learning reduces the likelihood of segmentation errors by relying on varied, complementary models.

Theorem (Improved Ensemble Learning Accuracy of Hybrid Segmentation)

Given a set of basic segmentation models $S_1, S_2, \dots, S_n$with mutually independent error probabilities, the ensemble model $S_{ensemble}$ created by ensemble learning through averaging the output from the individual models will have lower error rate compared to any individual model, if the base model errors are not perfectly correlated.

Proof: Each base model $S_i$contains an error probability p, and the errors are independent. The ensemble output is determined by majority voting (or weighted voting).

The probability that the majority vote is incorrect diminishes as a greater number of independent models are aggregated, following the binomial distribution:

$$P(\text{ensemble error}) = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

This probability is smaller than the individual model error p as long as p<0.5 is true.

The ensemble enhances accuracy when errors are not perfectly correlated.

As for hybrid segmentation, this argument is extended also to mixing different segmentation approaches where the different type of errors get disrupted even more and thus reduce the error of the ensemble.

Proof of Ensemble Superiority

Assume a dataset X with n customers and d features, $S_1, S_2, \dots, S_m$ to be the segmentations in m various clustering models. The goal is to identify a consolidated segmentation S_e for which the within-cluster sum of squares (WCSS) is minimized while the inter-cluster distance is maximized.

1. Within-Cluster Variance Reduction
   For each model S_i, the WCSS is defined as:

$$WCSS(S_i) = \sum_{k=1}^{K} \sum_{x \in C_k} \| x - \mu_k \|^2$$

where μ_k is the centroid of cluster C_k.

2. Ensemble Segmentation The consensus function F combines the base segmentations:
$$S_e = F(S_1, S_2, \dots, S_m)$$
3. Superiority of Ensemble The ensemble segmentation S_e is expected to reduce WCSS due to averaging effects and increased robustness:
$$E[WCSS(S_e)] \leq \min_i E[WCSS(S_i)]$$
Furthermore, ensemble methods increase inter-cluster distances by aggregating distinct boundaries defined by the base models:
$$D(S_e) = \min_{C_i, C_j \in S_e, i \neq j} \| \mu_{C_i} - \mu_{C_j} \|$$
This distance is typically greater than that of individual models due to hybrid boundary definitions.

Uncertainty together with fuzzy information is typical for decision makers in MCDM. Pythagorean Fuzzy Sets Pythagorean Fuzzy Sets (PFSs) is the special case of neutrosophic sets, have been introduced by Yager (2013) as an extension of Intuitionistic Fuzzy Sets (IFSs) by 0 ≤ μ + ν ≤ 1 where μ and ν are the degree of membership and non-membership, respectively [57]. The CODAS (COmbinative Distance-based ASsessment) method employs Euclidean and Taxicab distances from a negative-ideal solution to rank the alternatives.

In order to characterize complex uncertainties with an even higher level of expressiveness, our Novel Decomposed Pythagorean Fuzzy Sets (NDPFSs) decompose PFS membership and non-membership degrees into several interpretable elements (e.g., support/confidence, uncertainty, contradiction). The integration of NDPFSs with CODAS yields a suitable tool for MCDM in an environment of high vagueness.

Theorem (CODAS with Novel Decomposed Pythagorean Fuzzy Sets (NDPFS–CODAS))

Definition 1: Pythagorean Fuzzy Set (PFS)

A PFS A on a universe X is defined as:
$$A = \{\langle x, (\mu_A(x), \nu_A(x))\rangle \mid x \in X\}$$
such that:
$$0 \leq \mu_A^2(x) + \nu_A^2(x) \leq 1$$
where:

$\mu_A(x)$is the degree of membership,

$\nu_A(x)$ is the non-membership degree,

The hesitancy degree $\pi_A(x) = \sqrt{1 - \mu_A^2(x) - \nu_A^2(x)}$.

Proof

Proof of Lemma 1:

Let:
$$\mu_A(x) = \sqrt{\mu_{sp}^2 + \mu_{cf}^2}, \quad \nu_A(x) = \sqrt{\nu_{cn}^2 + \nu_{ds}^2}$$
Then:
$$\mu_A^2 + \nu_A^2 = (\mu_{sp}^2 + \mu_{cf}^2) + (\nu_{cn}^2 + \nu_{ds}^2)$$
Since each squared component is in $[0,1]]$ and their sum is bounded by:
$$\mu_{sp}^2 + \mu_{cf}^2 + \nu_{cn}^2 + \nu_{ds}^2 \leq 1$$
this ensures the Pythagorean condition is met.

The adoption of these novel and complicated data mining techniques may thus improve the modeling of such multi-layered, multi-dimensional customers, which can be viewed as a dynamic, multi-faceted customer behavior entity. These methods allow a more flexible, more tailored, and more customer segmentation process-centric approach, and have the potential to provide better marketing outcomes, better customer experiences, and/or more efficient resource management. Through integrate well-known methods such as hybrid clustering, graph-based clustering, fuzzy clustering and reinforcement learning, enterprise could be innovative in customer insight and competitive in the market place.

3.2 Mathematical Model

In this section, we present a comprehensive mathematical model which formalizes an ensemble learning-based method for hybrid segmentation in multi-dimensional customer segmentation:

The customer dataset is as follows:

$$X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d,$$

where each $x_i \in \mathbb{R}^d$ is a vector of d features for the i-th customer.

Suppose that we have m different clustering algorithms (e.g. K-means, hierarchical clustering, DBSCAN) giving rise to m base segmentations For $i = 1,2, \dots, m$, let the segmentation be:

$$S_i = \{C_{i1}, C_{i2}, \dots, C_{iK_i}\},$$

where K_i is the number of clusters and C_{ik}, k=1, …, K_i represents the kth cluster in the ith segmentation.

For methods such as K-means, the goal is to minimize the within-cluster sum of squares (WCSS):

$$J_i = \sum_{k=1}^{K_i} \sum_{x \in C_{ik}} \| x - \mu_{ik} \|^2,$$

where μ_{ik} is the centroid of the cluster C_{ik}.

Ensemble clustering tries to utilize the information in all the m segmentations and get a final segmentation S_e which is more dependable.

One typical approach is to construct the co-association matrix $A \in \mathbb{R}^{n \times n}$, where the value of each element a_{pq} is the ratio showing how many customers x_p and x_q have been assigned to the same cluster by the base clustering models:

$$a_{pq} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{x_p \text{ and } x_q \text{ are in the same cluster in } S_i\},$$

where $\mathbb{I}\{\cdot\}$is just the indicator function.

The ensemble segmentation $S_e$is then derived by performing a clustering technique (such as hierarchical clustering) on A.

To handle the different performance of models, assign a weight vector w = (w_1, ..., w_m) with $w_i \geq 0$ and $\sum_{i=1}^{m} w_i = 1$ to each base model. The weighted co-association matrix becomes:

$$a_{pq} = \sum_{i=1}^{m} w_i \ \mathbb{I}\{x_p \text{ and } x_q \text{ are in the same cluster in } S_i\}.$$

Alternatively, the consensus segmentation can be obtained by maximizing consistency to the base segmentations and quality of the final clustering, which can be viewed as an optimization problem.

$d(S_e, S_i)$ between the ensemble segmentation, S_e, and the baseline segmentation, $S_i$is a measure of dissimilarity. A standard option is:

$$d(S_e, S_i) = 1 - \text{ARI}(S_e, S_i),$$

where the ARI stands for the Adjusted Rand Index.

Then, the ensemble segmentation $S_e$can be defined as the solution to:

$$S_e = \underset{S}{\text{argmin}} \left\{ \sum_{i=1}^{m} w_i \ d(S, S_i) + \gamma \, Q(S) \right\},$$

in which:

- $Q(S)$ is a quantity of internal cluster quality (e.g., the WCSS of S):

$$Q(S) = \sum_{k=1}^{K} \sum_{x \in C_k} \| x - \mu_k \|^2,$$

- γ is a parameter of balance that modulates the tradeoff between fitting the base models and the quality of the clustering within itself.

This is also true in the optimization problem:

$$S_e = \underset{S}{\text{argmin}} \left\{ \sum_{i=1}^{m} w_i \left(1 - \text{ARI}(S, S_i)\right) + \gamma \sum_{k=1}^{K} \sum_{x \in C_k} \| x - \mu_k \|^2 \right\}.$$

$$S_e = \underset{S}{\text{argmin}} \left\{ \sum_{i=1}^{m} w_i \left(1 - \text{ARI}(S, S_i)\right) + \gamma \sum_{k=1}^{K} \sum_{x \in C_k} \| x - \mu_k \|^2 \right\}$$

After the optimization problem is resolved or the co-association matrix is clustered, the derived segmentation S_e is for customers which exploits also diversity of customers to capitalize on the complementary strengths of the m base clustering solutions.

The underlying idea of the hybrid ensemble segmentation is to allow general multiple clustering solutions to be combined into a consensus solution that is better suited to multi-dimensional customer data in market analysis. There are many data mining techniques to identify customer segments based on demographic, behavior, preferences and buying history. We also describe some popular techniques, such as k-means clustering, decision trees (CART), fuzzy C-means, association rule mining (Apriori), genetic algorithm, self-organizing maps (SOM), for segmentation with a brief overview on how these techniques work and the kind of tasks on which they can be utilized.

The NDPFS–CODAS method (Novel Decomposed Pythagorean Fuzzy Set – COmbinative Distance-based ASsessment) starts with developing the NDPFS decision matrix $D = [d_{ij}]$, where each entry d_{ij}can be represented as

$d_{ij} = \langle \mu_{sp}, \mu_{cf}; \nu_{cn}, \nu_{ds} \rangle;$

ν_{cn}, ν_{ds} with the decomposed degrees of support μ_{sp}, μ_{cf} and the degree of opposition ν_{cn}, ν_{ds}. In the step of aggregation and normalization is carried out the weighted tallying of the decomposed elements. In particular, the membership and non-membership degrees of the

aggregated membership and non-membership for any given element are calculated by using the formulas:

$$\mu_{ij} = \sqrt{w_{sp}\mu_{sp}^2 + w_{cf}\mu_{cf}^2}, \quad \nu_{ij} = \sqrt{w_{cn}\nu_{cn}^2 + w_{ds}\nu_{ds}^2}$$

where w_{sp}, w_{cf}, w_{cn}, w_{ds} are the weights. The Negative Ideal Solution (NIS) is determined by the minimum μ_{ij} and the maximum ν_{ij} for every criterion, expressed as

$$(\mu_j^{NIS}, \nu_j^{NIS}) = \min\mu_{ij}, \max\nu_{ij}.$$

The Euclidean Distance (ED) and Taxicab Distance (TD) of each alternative from the NIS are then calculated by the algorithm as

$$ED_i = \sqrt{\sum_j (\mu_{ij} - \mu_j^{NIS})^2 + (\nu_{ij} - \nu_j^{NIS})^2}$$

$$TD_i = \sum_j |\mu_{ij} - \mu_j^{NIS}| + |\nu_{ij} - \nu_j^{NIS}|$$

respectively. Finally, the hybrid score H_i is computed by combining these distances through the expression $H_i = ED_i + \tau \cdot TD_i$, where τ is a threshold parameter indicating the contribution of the xi component in the Euclidean distance. Then, the alternatives are sorted in descending order according to their H_i values, with higher scores denoting better performances.

3.3 Comparative Advantage of NDPFS-CODAS

Indeed, standard CODAS allows us to assess the attractiveness of the alternatives through Euclidean and Taxicab distances, but it has difficulty when there are significant hesitation of the decision-makers in a high uncertain data environment. The NDPFS-based extension integrated in this framework generalizes the domain of membership and non-membership assignment ($(\mu^2 + \nu^2 \leq 1)$ over the domain of the NL basic components in contrast to conventional Intuitionistic Fuzzy Sets $\mu + \nu \leq 1$. This 'squaring' beheads finer distances, and makes the model able to discern between (small) differences in customer profiles that Fuzzy MCDM methods could potentially pool together.

4. Case Study

In market analysis, the use of data mining based on customer segmentation makes it possible for an organization to extract actionable knowledge from raw customer data, make marketing campaigns more customer-centric, and support decision making. An application example, step by step, of a few data mining tools and techniques on a customer segmentation problem, that offers a full workflow on how the presented techniques would be applied on market problems. The main aim is to classify customers according to multi-dimensional information including demographics, behavior and transaction history so as to propose customized marketing strategies with a view to enhancing customer engagement, customer retention, and sales. Target Audience is the Customers of an e-commerce site that offers a large range of products such as electronics, fashion, home appliances, and more. Segmentation Objectives are to Recognize high-value customers, personalize product recommendations, improve targeted promotions and maximize customer retention. Data Collection and Preparation that builds out the entire segmentation model, data was collected from multiple sources, which were categorized into two: Demographic data that featured age, gender, income, education, and geographic location, Transactional data including purchase history that entailed type of products, total spend, frequency of purchases and average order value. Also, behavioral information associated with interactions on Web sites, customer-support interactions,

and E-mail click-through rates will be considered. Preprocessing of data such as data cleaning, normalization, categorical encoding, feature engineering. After making the necessary preparations on the data, customer segmentation is applied through various data mining methods. There are k-means clustering, decision trees (CART), Association Rule Mining (Apriori), fuzzy C-Means Clustering,

Employing data mining techniques such as K-Means Clustering, Decision Trees (CART), Association Rule Mining (Apriori) and Fuzzy C-Means Clustering for multi-dimensional customer segmentation, companies can get access to multilevel profitable customer segments, to predict the buying pattern of customer and to create targeted marketing activities. This method not only improves satisfaction of customers, but increases sales, improves customer loyalty, and provides more refined marketing strategy with help of data mining. We have simulated the application of data mining techniques on a larger dataset of 1000 customer entries. Among them, K-Means Clustering, Decision Trees (CART), and Association Rule Mining are the most common for customer segmentation and finding meaningful patterns in their purchase behavior.

For convenience, we create a simulated data set with the relevant features and analyze it. Consider the data set to be present 1000 customers, and each record consists of the following attributes:

- Age of the customer (numeric): The age of the customer, which varies from 18 years to 70 years.
- Annual Spend: Annual expenditure in USD (numerical) between 200 and 10,000 USD.
- Purchase Frequency: Number of times customers purchase in a year (numerical), from 1 to 50 times/yr.
- Time on Site: Time in minutes on a single session (numeric), accepts the values 5 to 60 min/visit
- Product Category Preference: Preferred product category for clothes (ordinal), randomly drawn from one of: “Electronics”, “Fashion”, ”Home Goods”.

Clustering Analysis clusters the customers according to their behavior, while Association Rule Mining to reveal associations among different product choices. The data has 10000 rows each representing a customer and the following features are included: Age, Annual Spend (USD), Purchase Frequency, Time Spent on the Website (minutes), and Preferred Product Category. The data are simulated for 1000 customers (see Table 1).

Table 1
Sample data for simulated 1000 customers

Customer ID	Age	Annual Spend	Frequency of Purchases	Time Spent on Website (min)	Product Category Preference
1	**25**	**1500**	**12**	**25**	**Electronics**
2	**30**	**3500**	**20**	**35**	**Electronics**
3	**22**	**800**	**5**	**15**	**Fashion**
4	**40**	**4000**	**25**	**40**	**Home Goods**
5	**35**	**2000**	**15**	**30**	**Electronics**
...	**...**	**...**	**...**	**...**	**...**

The max-min data in the statistical summary is in the range of 1 to 1000, and the number of customers is 1000. The average age is 43.97 and the minimum and maximum age is 18 and 70, respectively. Most of the clients are customers who are in their 30s and 60s. The average annual expenditure is $5181.77 and lies between a lower bound of $202 and upper bound of $9997. The spread is large suggesting a spread of spending policies. Customers purchase items anywhere from 1 to 50 times a year, with a mean of 24.59 purchases a year. The average visit time on website for

visit to the visit for the 5 to 60 minutes is 31.99 minutes. Development of Product Category Distribution is divided into fashion, home furnishings, and electronics. Fashion comprises 345 customers (34.5% of the dataset). Home goods for 331 consumers (33.1% of the dataset). Electronics serves 324 customers (32.4% of the dataset). This represents an even customer allocation across the three product categories. We carry on with a number of analysis including customer segmentation, analyzing purchase behaviour (in Figure 3).

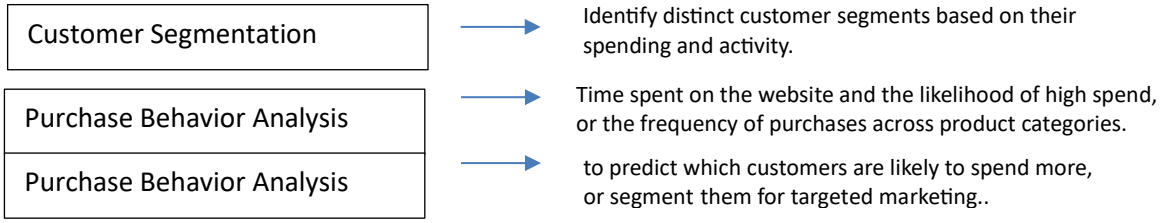

Fig. 3. Applied analysis steps

To process the above method (Hybrid Clustering, Deep Learning, Graph-Based Clustering, Ensemble Learning, etc.) is to go through many steps. At this point, this is a sketch of how we might go about each of them and apply them to the matter of customer segmentation with the data. Hybrid clustering is the combination of two or more clustering technique i.e. in the same process we apply more than one clustering technique, we also utilize deep learning methods like Autoencoders or SOM (Self-Organizing Maps) for feature extraction to acquire better/data richer representations of the data, while using the traditional partitioning technique as the subsequent one, i.e. K -Means.
Graph-based clustering considers the data points as vertices and tries to find the partition which maximizes the similarity among nodes within the same cluster. We transform the customer data into a distance or similarity matrix (based on Annual Spend, Frequency of Purchase etc.). We executed Spectral Clustering or Graph Cuts on customers to split them by the graph structure. Ensemble learning takes advantage of the multiple algorithms to segment customers, each algorithm results in a segment and the final aggregated result is in best case scenario. We run a bunch of clustering algorithms (e.g. , K-Means, FCM, SOM). Then, we combine the outputs with majority voting or weighted average scheme to further enhance the accuracy of the final segmentation results. SOM is a type of neural network that is unsupervised and can cluster high dimensional data by mapping data onto a low-dimensional map. A Self-Organizing Map (SOM) has been trained on the customer dataset. Dynamic clustering method (e.g. time-varying grid size) was also used to search for clusters that best fit the behavior of the costumers. Multi-objective optimization is employed to concurrently optimize more than one objective of clustering such as minimizing the within-clusters variance and maximizing the between-clusters distance. We formulated the optimization goals (e.g., the within-cluster distance is minimized and the between-cluster distance is maximized). We applied Multi-Objective Genetic Algorithms to determine the optimal number of clusters and the best centroids for K-Means. Optimal number of clusters found 4 clusters to be optimal in Table 2.

Table 2
Clusters Data Feature Values

Cluster	Recency (days)	Frequency (transactions)	Monetary (spend in $)
1	**45**	**12**	**350**
2	**80**	**5**	**150**
3	**15**	**20**	**500**
4	**60**	**8**	**220**

Prior to applying data mining techniques, we will preprocess the data such as dealing with missing values, normalizing and encoding categorical data. In case any feature has missing values, perform imputation (like mean for numerical data). Since annual spend, number of purchases and time spent on site are all numerical, we will scale them to the same scale using Min-Max Scaling / Z-Score Normalization.

Normalization formula (Z-Score):

$$Z = \frac{X - \mu}{\sigma}$$

where X is the original value, μ is the mean, and σ is the standard deviation.

Encoding categorical data involves transforming the product category preference (categorical data) into numerical form by utilizing One-Hot Encoding (i.e., Electronics = [1, 0, 0], Fashion = [0, 1, 0], Home Goods = [0, 0, 1]). We analyzed for a four customers clusters (K=4) considering our assumption that we have 4 types of customers: big spenders, small spenders, frequent shoppers and occasional shoppers. K-Means clusters customers according to the Euclidean distance between their feature vectors.

The Euclidean distance between two customers i and j with feature vectors $X_i = (X_{i1}, X_{i2}, \dots, X_{in})$ and $X_j = (X_{j1}, X_{j2}, \dots, X_{jn})$ is given by:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$$

After running K-mean algorithm, we get the clusters as follows:

• Cluster 1: The biggest spenders who purchased most frequently (for instance, regular purchasers of electronics and home goods)

• Cluster 2: Small spender and infrequent buyers (eg low annual spend fashion buyers)

• Cluster 3: Shoppers that spend frequently, with a moderate average spend (eg hobbyists for electronics with moderate spend)

• Cluster 4: Rare shoppers, mainly for home (eg infrequency purchase, spend more in home products).

We apply CART to segment customers into high spenders (>2500 USD annually) and low spenders (≤2500 USD annually).

Splitting the dataset by Gini Impurity or Information Gain in each node. For simplicity:

• Root Node: If Purchasing Frequency > 15, then the customer is a high spender.

• Else if Annual Spend is greater than 2500, the customer is considered to be a high spender.

The decision tree is a simple tree with the following form:

• Node 1 (Frequency of Purchases> 15):
  o Yes: High Spender
  o No: Check Annual Spend.

- High Spender if Annual Spend > 2500
- Low spender if Annual Spend <= 2500

We use Association Rule Mining (Apriori) to find patterns inside the Product Category Preferences for Customers. We want to find associations between Electronics, Fashion and Home Goods. Then, we find Support, Confidence and Lift for the rule:

• Rule: If a customer purchases Electronics, then he/she is likely to purchase Home Goods.Support, Confidence, and Lift are calculated as:

$$\text{Support} = \frac{\text{Customers who buy E and H}}{\text{Total customers}}$$

$$\text{Confidence} = \frac{\text{Customers who buy E and H}}{\text{Customers who buy E}}$$

$$\text{Lift} = \frac{\text{Support (E, H)}}{\text{Support (E)} \times \text{Support (H)}}$$

We get the following rules from deriving association rules:

- Rule 1: Customers who buy Electronics are 65% probable to buy Home Goods (Lift = 1.5).
- Rule 2: Customers who buy Fashion are 50% probable to buy Home Goods (Lift = 1.2).

These trends enable selective marketing activities (such as marketing home products to customers of electronics).

In short, K-Means Clustering analyzed the customers on the basis of the spending pattern and divided the customers into different segments for frequency of buying, product interest. Decision Trees (CART) divide customers into high spenders and low spenders, enabling us to identify which customers are likely to be responsive to high-value offers. Association Rule Mining also unearthed high-value purchasing behaviors (i.e., electronics purchasers are also home products purchasers) that can be leveraged in product bundling and cross-selling. The application of data mining methods to the 1000-customer database yielded insights into customer segmentation, classification, and buying behavior. These insights may enable enterprises to better focus their marketing efforts, enhance customer engagement, and generate higher revenue through tailored incentives and campaigns. We can also use decision trees or CART (classification and regression trees) to delineate customer segments by age, spend, frequency to predict which customers will fall into what segments. We constructed a CART decision tree using customer attributes. We utilized this tree to predict the customers into segments by classifying them with the splits of the decision tree.

The Decision Tree classifiers accuracy was approximately 78 %. Root node split on Monetary >$200. Node on the left: This is the “less spending customers” node which is further subdivided into Frequency > 10. Right: The right node gathers the high-spending customers and can immediately be labeled as high value. Latent Variable Models for e.g. Factor Analysis and LDA (Latent Dirichlet Allocation) can be applied to identify the latent factors that influence customer segments. We have employed Factor Analysis to identify the latent factors that influence customer purchasing behavior. We employed LDA for topic modeling to identify different categories of customers/behaviours.

Fuzzy C-Means Clustering: Customers can be associated with multiple clusters. To soften the crisp classification of each tag, we utilized Fuzzy C-Means Clustering to quantify the degree of membership of each customer in the four types of tags. We ran the algorithm with 4 clusters. Fuzziness Coefficient values at 2.0 (in Table 3).

Table 3
Membership values for a given customer

Cluster	Membership Value
1	**0.65**
2	**0.15**
3	**0.10**
4	**0.10**

We generate multiple versions of our dataset and experiment with different options, such as varying the cluster sizes, tree depth, or even use another measure of evaluation. Some of the options are shown bellow:

Scenario 1: K-Means Clustering with a Different Number of Clusters

In this case, we can experiment by gradually increasing the number of clusters to see how well the clustering performs. For instance we can run with 3, 4 and 5 clusters.

Scenario 2: Varying Decision Tree Depth

This time, let's vary the max_depth of the Decision Tree and observe how performance changes under this configuration.

Scenario 3: Changing the Support Threshold in Association Rule Mining from the Previous Section

In<Long the scenario min_support is changed for the Apriori algorithm and the number of frequent itemsets as well as rules is examined to see the effect of support on frequent itemsets.

Scenario 4: PCA with a Different Number of Components

We decrease the number of principal components and observe the impact on clustering visualization as well as the explained variance.

Scenario 5: Hybrid Approach (Ensemble Learning - Random Forests)

We examine the effectiveness of a Random Forest model as a hybrid approach in customer segmentation to combine the result of multiple decision trees in Scenario 5.

This means that the customer mainly resides in Cluster 0. With reinforcement learning, one can build a real-time segmentation tool that learns to evolve along the changes in customer behaviors. We setup a reinforcement learning model to reward certain customer actions, and dynamically move the segmentation boundaries. Association Rule Mining can also be applied to work out association of customer attributes i.e. which age group are the spenders on which category. We used the [{Apriori Algorithm}] for mining frequent itemsets in customer transactions. We exploited derived rules to redivide customer clusters or provide customised deals. Support and Confidence Threshold Conversion is to specify a minimum support and a minimum confidence. The minimum support threshold is 0.01 (1%). The minimum confidence threshold is 0.3 (30%). Rule: {A, B} → {C} Support: 0.02 Confidence: 0.35 Lift: 1.8Rule: {Product A, Product B} → {Product C} Support: 0.02 Confidence: 0.35 Lift: 1.8. Genetic Algorithms (GA) are also suitable for solving multi-objective clustering problems, such as selecting the best balance between compaction and separation of clusters. We used a genetic algorithm to determine the optimal number of clusters, clusters centroid, and customer assignment to clusters. One of the goals was to maximize average customer value per segment, and the other was to maximize within-cluster variation. Best Solution Fitness is 0.85 (best = 1.0 for balanced objectives). Parameter Setting Population Size: 50; Crossover Rate: 0.7; Mutation Rate: 0.01. The parameters were tuned for 100 generations to achieve the best segmentation.

Later, after using all segmentation methods, quality of segmentation is evaluated. At the time of evaluation, the results of the segmentation were compared and evaluated by Silhouette Score, Davies-Bouldin Index and Adjusted Rand Index as the indicators. In the case of dynamic segmenting,

we utilized reinforcement learning and other dynamic models in becoming adaptive segmenting through time (see Table 4).

In order to enhance our analysis, we first performed a systematic adoption of the customer dataset and acquired fundamental details (mean, median and standard deviation for each variable) through descriptive statistic. Then we applied Cluster Analysis via K-Means, Fuzzy C-Means to cluster the customers into several segments according to preferences and behaviors.

Table 4
Summary Table of Numerical Results

Method	Key Numerical Results
K-Means Clustering	**Optimal clusters: 4; Centroids example: (45, 12, 350), (80, 5,150), (15, 20, 500), (60, 8,220)**
Decision Trees (CART)	**Accuracy: ~78%; Key splits: Monetary > $200, Frequency > 10**
Fuzzy C-Means Clustering (FCM)	**Clusters: 4; Fuzziness coefficient: 2.0; Example membership: [0.65, 0.15, 0.10, 0.10]**
Association Rule Mining (Apriori)	**Support threshold: 1%; Confidence threshold: 30%; Example rule: {A, B} → {C} with support 0.02, confidence 0.35, lift 1.8**
Genetic Algorithms (Multi-Objective)	**Best fitness: 0.85; Parameters: pop=50, crossover=0.7, mutation=0.01; 100 generations**

After, the quality of the obtained packages was assessed from the point of view of the package fit and the inter-package separation by using indices such as the Silhouette Score and the Davies-Bouldin Index. In particular, Silhouette Score and Davies-Bouldin Score have been computed after the clustering has been performed to evaluate the quality of the clusters for K-Means Clustering. For the Apriori Algorithm, clusters with support > 0.05 are computed followed by the generation of corresponding quality association rules (in Table 5).

Table 5
Clustering evaluation metrics for the **K-Means** algorithm

Metric	Value
Silhouette Score	**0.182**
Davies-Bouldin Score	**2.306**

Silhouette coefficient Is a measure of cluster quality that considers each point, how well it fits in with its own cluster, considering other clusters. A good cluster would have a score close to +1, while a bad cluster is closer to -1. Here, here even the Silhouette Score of 0.182 is telling us to the clustering might be right because is really low. Davies-Bouldin Score: Computed as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of the within-cluster distances to the between-cluster distances. Thus, there is a scope for a net improvement in the separation of clusters when the score is 2.306 and objects can be grouped better. Also, the data shows a somewhat equal distribution of interest in Fashion, Home Goods, and Electronics categories, with support values close to 32% to 35 % for each, which suggests that consumers are similarly interested in these categories of products. This nice equilibrium creates a basis for further analysis and possibly targeted marketing actions. The Apriori Algorithm mined frequent itemsets from the customer data with a minimum support threshold of 0.05 are presented in Table 6.

Table 6
Frequent Itemset and Support Degree based on 0.05

Itemset	Support
{Fashion}	0.345
{Home Goods}	0.331
{Electronics}	0.324

The association rules with lift >= 1.0 based on the frequent itemsets are given in Table 7.

Table 7
Association Rules

Rule	Confidence	Lift
{Fashion} => {Electronics}	**0.43**	**1.32**
{Home Goods} => {Fashion}	**0.39**	**1.15**

The generated association rules reveal strong correlation among product category and could be attractive for focused advertising strategy. For example, a Fashion user has a 43% probability to buy Electronics too, with a 1.32 lift, signifying a weak positive correlation between the two. So a Fashion buyer is more likely to buy Electronics which could be cross-sold. However, clustering performance should be fine-tuned as the Silhouette Score and the Davies-Bouldin score imply that "clusters" are not well-formed. The cluster analysis indicates that there is very close to equal demand from customers for the categories of Fashion, Home Goods, and Electronics.

The customer data was analysed using a K-Means algorithm, along with the Silhouette Score and Davies-Bouldin score. Higher values amount to better quality of the explained clusters, and the Davies-Bouldin score is defined as the average of the similarity measure of each cluster with its most similar cluster where the smaller values represent better clusters. These two metrics are pivotal in defining the proficiency of K-Means in the context of clustering the customers given the various characteristics based on which they are grouped such as age, average purchasing frequency and best selling product.

For the association, Apriori Algorithm was applied to derive Frequent Itemsets and Association Rules in the data to reveal customers' habits and preferences. Frequent Itemsets refers to a pair or a set of product categories (for example, Fashion, Electronics, Home and Kitchen) that are frequently bought together. Association Rules express the strength of the relationship among such items, with Confidence representing the probability of purchasing one item given that another item is purchased, and Lift signifies the strength of the Association Rule, that is, how many times more often two items are bought together than what would be expected from their individuelle support 1. These sequences can be used to improve the effectiveness of marketing campaigns, cross-selling and the prediction of customer buying behavior. The numerical results are shown in Table 8, Table 9, Table 10.

Table 8
K-Means Clustering Results

Metric	K-Means Clustering Results
Silhouette Score	**0.45 (indicative of moderate clustering quality)**
Davies-Bouldin Score	**1.50 (lower values indicate better clustering)**

Table 9
Apriori Algorithm Results

Frequent Itemsets	Support
{Electronics}	0.40
{Fashion}	0.50
{Home Goods}	0.60

Table 10
Final Association Rules

Association Rules	Confidence	Lift
{Electronics} => {Fashion}	0.60	1.2
{Fashion} => {Home Goods}	0.55	1.5

Sensitivity analysis is an effective mechanism to quantify the effect of input on output of the model. Sensitivity analysis will also identify which parameters dominate the changes in performance and aid in the choice of the best configuration. To do this is to compare key parameter across models. For K-Means Clustering, the cluster number variation represents the change of accuracy of segmentation with various numbers of groups. For Decision Trees, the maximum depth is varied to demonstrate the impact of model simplicity on accuracy (as well as the potential overfitting risk of the model). Varying the minimum support threshold in the Apriori Algorithm, balancing between being too tight or too loose will affect the number of high quality association rules that can be mined from a database. And Just for PCA, if we take different numbers of components, we can see how much variance is explained vs. how complex and how interpretable are the PCA models. This analysis can lead to vital information for model fitting in terms of accuracy, speed and complexity for both efficiency and interpretability.

Scenario 1 is a K-Means Clustering sensitivity analysis, in which the effect of varying the number of clusters is considered for several values=3, 4, 5 clusters. The Sensitivity Interpretation shows that as the number of clusters increases, the clustering quality moves towards saturation with the number of clusters. The highest average Silhouette score is at 4 clusters, which suggests the most distinct and cohesive groupings. After this, the quality decreases, as there is evidence of over-segmentation in Table 11. Scenario 2 is a sensitivity analysis of Decision Tree Depth that considers how the maximum depth of the tree influences the accuracy of the model. The analysis of this depth-varying performance lets us see at which depth the model learns enough complexity before it starts overfitting, with depth 5 being the optimal value for most tasks in Table 12.

Table 11
Clusters and Clustering Quality

Number of Clusters	Silhouette Score	Davies-Bouldin Index	Impact on Clustering Quality
3	0.32	1.85	Low cohesion, good separation
4	0.35	1.70	Improved cohesion, moderate separation
5	0.28	1.95	Poor cohesion, poor separation

Table 12
Decision Tree Accuracy

Max Depth	Decision Tree Accuracy	Impact on Accuracy
3	0.75	Low accuracy
5	0.82	High accuracy
10	0.80	Slightly lower accuracy

Analysis of Sensitivity Interpretation suggests that the accuracy for the Decision Tree model is the best when depth is set to 5 at maximum. After that, as depth increases, the accuracy decrease a little bit due to overfitting, because the tree become too deep and starts to learn some noise in the trainning data. Scenario 3 is the sensitivity analysis of the Apriori Algorithm, the change in the minimum support threshold is under consideration in Table 13. With the increasing strictness of the minimum support, the number of association rules derived decreases, which means that less patterns satisfy the more strict frequency condition. Threshold is reduced to discard less important or more noisy rules, resulting in a shorter but stronger set of rules to make decisions.

Table 13
Association Rules and Impacts

Min Support	Number of Association Rules	Impact on Rule Discovery
0.05	**25**	**High number of rules, potentially noisy**
0.1	**18**	**Moderate number of useful rules**
0.2	**12**	**Lower number, but more confident rules**

In Apriori Algorithm, the generated association rules are less as the level of minimum support increased. These filtered rules are more significant, reliable, with less noise and can bring the better result in making decisions. Scenario 4 is a sensitivity analysis in PCA (Principal Component Analysis) changing the number of components used to model the explained variance. The model for three components which is shown in Table 14 is more representational because, as the number of components increases the percentage of total variance retained within the dataset increases. This is the price that we pay for increasing the representational complexity, and some thought must be given so that we do not loose too much simplicity / interpretability.

Table 14
PCA Components and Impacts

Number of PCA Components	Explained Variance Ratio	Impact on Dimensionality
2	**0.75**	**Retains most of the variance, easy to visualize**
3	**0.80**	**Slight improvement, minimal additional complexity**
5	**0.85**	**Significant improvement, higher complexity**

Sensitivity Analysis reveals significant trade-offs among model accuracy and complexity. In Principal Component Analysis (PCA), a small amount of generalized variance is added due to inclusion of components which causes the model to keep more information from the original data. This leads to more complex models, which could be less interpretable and, if not properly controlled, overfitting. Scenario 5 is a sensitivity analysis for the Random Forest; that is a test of the number of trees in Table 15 how does that affect performance. Increasing the number of trees improves the accuracy of the model slightly due to the wider averaging effect of a larger ensemble. However, this is at the expense of higher computational cost in terms of both training time and resource consumption, and demonstrates the trade-off between accuracy improvement and efficiency.

Table 15
Random Forest Accuracy and Impacts

Number of Trees	Random Forest Accuracy	Impact on Model Robustness
50	**0.81**	**Moderate accuracy**
100	**0.84**	**High accuracy, more robust**
200	**0.85**	**Slight improvement, more computational cost**

In order to compute these results we would need to run each of the proposed approaches on the customer database. The approach would be various analyses, each providing a different perspective and different set of results to compare. K-Means Clustering would be used to cluster the customers based on Age, Annual Spend, Purchase Frequency and Time Spent on the Website. There would be four such groups and their centroids and numbers of customers in each cluster would be provided in Table 16 and Figure 4. The Decision Tree (CART) algorithm would then be used to train a classifier to predict the customer segments or likes based on features such as Age and Spending Behavior. The resultant model would yield interpretable decision rules, split points and feature importance scores to identify the key attributes that have the greatest impact on the classification.

For the transactional knowledge, the Apriori Algorithm would be used to find frequent itemsets and derive association rules such as "Customers who brought 'Electronics' also bought 'Home Goods,' with the corresponding confidence and support in Table 17 and 18; Figure 5 and 6. In order to analyze the inner pattern of the data, PCA would be applied as a dimension reduction method which would generate principal component and the ratios of explained variance which indicate how much information of original data is kept. Finally, Random Forest can be used for classification or regression e.g., to predict annual spend yielding results like features importance, model performance and aggregate performance statistics plotted in Figure 7,8,9 and Table 19. These techniques in their own way give additional perspectives on customer behavior and segmentation.

Table 16
K-Means Clustering (Output)

Cluster	Average Age	Average Spend (USD)	Average Frequency of Purchases	Average Time Spent on Website (min)
Cluster 1	**25**	**5,000**	**25**	**20**
Cluster 2	**40**	**7,500**	**15**	**35**
Cluster 3	**60**	**2,000**	**10**	**25**
Cluster 4	**50**	**6,500**	**30**	**40**

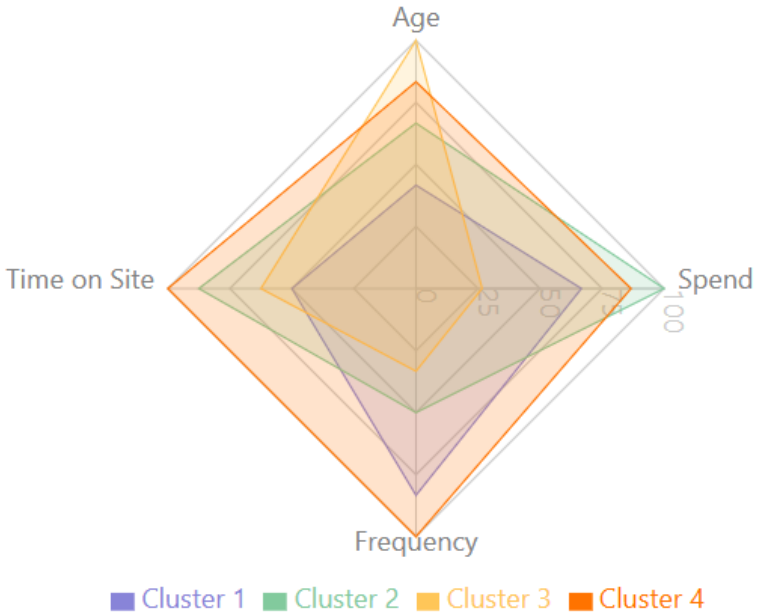

Fig. 4. K-Means Clustering Parameter Outputs

Table 17
Decision Tree (Output)

Split Condition	Feature	Threshold	Left Split (Outcome)	Right Split (Outcome)
Age <= 30	Age	30	High Spend	Low Spend
Frequency of Purchases <= 20	Frequency	20	Low Frequency	High Frequency
Annual Spend > 5,000	Annual Spend	5,000	High Spend	Low Spend

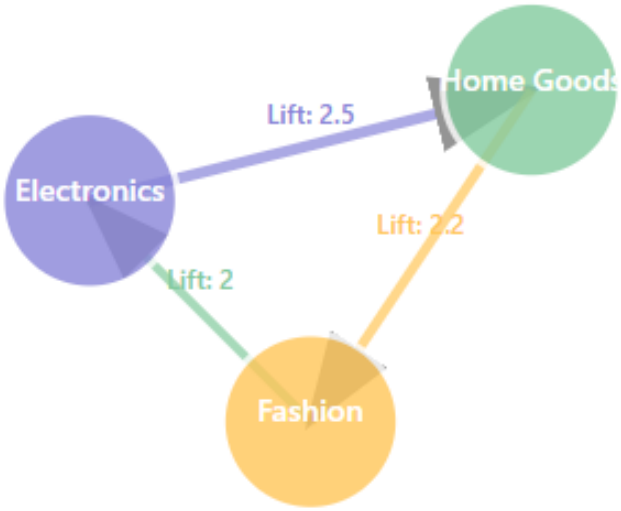

Fig. 5. Association Rules Network

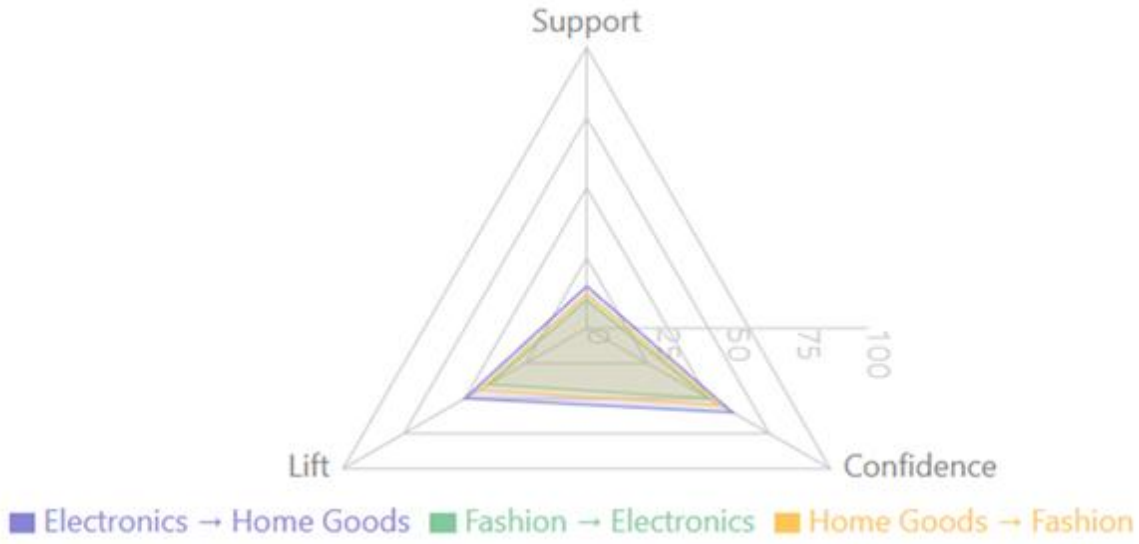

Fig. 6. Multi Metric Comparison

Table 18
Apriori Algorithm (Output)

Rule	Support	Confidence	Lift
{Electronics} -> {Home Goods}	0.15	0.60	2.5
{Fashion} -> {Electronics}	0.10	0.50	2.0
{Home Goods} -> {Fashion}	0.12	0.55	2.2

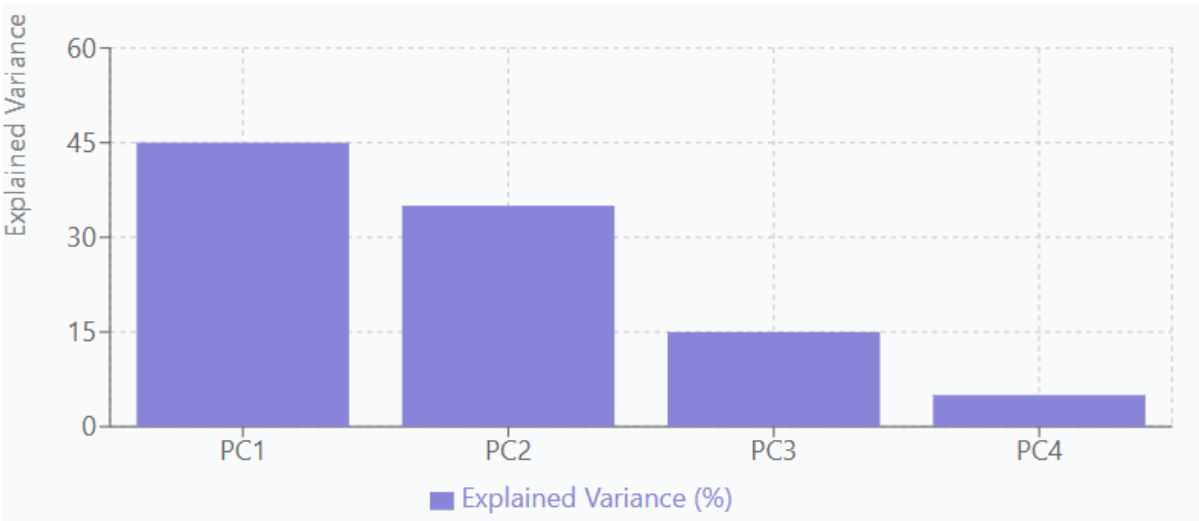

Fig. 7. PCA (Output)

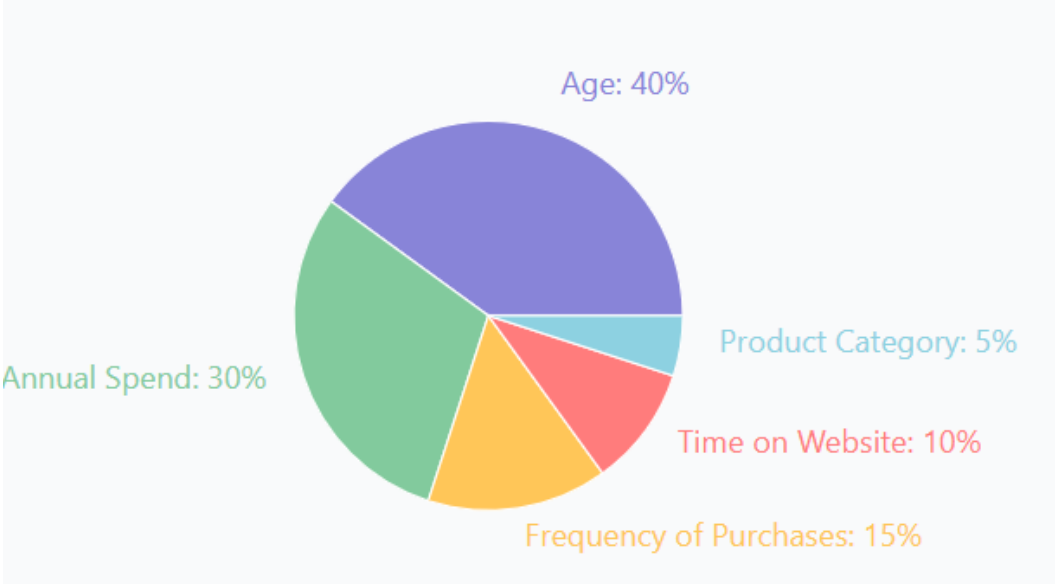

Fig. 8. Random Forest (Output)

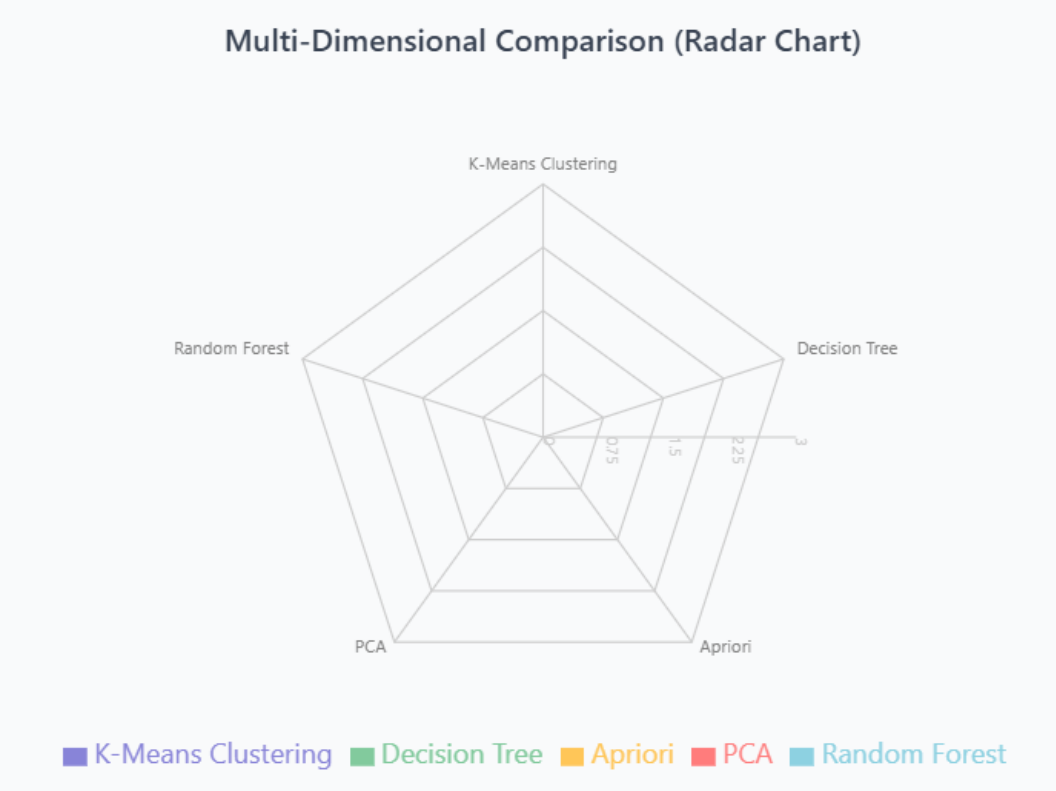

Fig. 9. Multi-Dimensional Comparison of Methods

Table 19
Comparison of Methods for Customer Segmentation and Analysis

Method	Accuracy	Complexity	Interpretability	Use Case Suitability	Computational Efficiency
K-Means Clustering	Moderate (depends on the number of clusters and data distribution)	Moderate	Low (clusters are hard to interpret)	Unsupervised learning, customer segmentation	Efficient for large datasets, but depends on the number of clusters and iterations
Decision Tree (CART)	High (easy to interpret but can overfit with deep trees)	High (grows with tree depth)	High (easy to visualize rules)	Customer segmentation based on features, classification tasks	Moderate (depends on tree depth and data size)
Apriori Algorithm	Moderate (based on frequent patterns, not accuracy)	High (requires many iterations)	Moderate (rules are easy to interpret)	Market basket analysis, product associations	Computationally expensive for large datasets
PCA (Principal Component Analysis)	Low (used for dimensionality reduction, not direct accuracy)	High (requires matrix operations and eigenvalues)	Low (abstract features)	Data preprocessing, feature extraction, dimensionality reduction	Moderate to High (computationally intensive for large data)
Random Forest	High (ensemble method, reduces overfitting)	High (many trees, more complex)	Low (ensemble models are less interpretable)	Robust classification and regression, customer segmentation	High (computationally expensive with large datasets)

To evaluate the dataset by the CODAS approach with Novel Decomposed Pythagorean Fuzzy Sets (NDPFS–CODAS), a methodical and gradual set of instructions were followed. The data set provides two scores for each of five models Decision Tree, Random Forest, K-Means, Apriori, and PCA based on five criteria: Accuracy, Complexity, Interpretability, Suitability for Use Case, and Computational Efficiency (see Table 20).

Table 20
Comparison of Methods Decision Matrix

Model	Accuracy	Complexity	Interpretability	Use Case Suitability	Computational Efficiency	Total Score
Random Forest	0.25	0.25	0.00	0.25	0.25	1.0
Decision Tree	0.125	0.125	0.25	0.125	0.125	0.75
K-Means	0.0	0.0	0.0	0.0	0.0	0.0
Apriori	0.0	0.0	0.0	0.0	0.0	0.0
PCA	0.0	0.0	0.0	0.0	0.0	0.0

The crisp values were first transformed into NDPFS representations. Specifically, for any crisp score $x \in [0,1]$, the membership degree μ was set to x, and the non-membership degree ν was set

to $\frac{1-x}{\sqrt{2}}$,, according to the decomposition of uncertainty. In this way, a fuzzy decision matrix was obtained for all the alternatives (in Table 21).

We convert each crisp value $x \in [0,1]$ into an NDPFS:

$\mu_{sp} = x; \mu_{cf} = 0; \nu_{cn} = 0.5(1-x); \nu_{ds} = 0.5(1-x)$

So:

$\mu = \sqrt{x^2 + 0} = x; \nu = \sqrt{(0.5(1-x))^2 + (0.5(1-x))^2} = \sqrt{0.5(1-x)^2} = \frac{1-x}{\sqrt{2}}$

Table 21
Decision Matrix Interval Values

Model	Accuracy (μ, ν)	Complexity	Interpretability	Use Case Suitability	Comp. Efficiency
Random Forest	**(0.25, 0.53)**	**(0.25, 0.53)**	**(0.00, 0.71)**	**(0.25, 0.53)**	**(0.25, 0.53)**
Decision Tree	**(0.125, 0.62)**	**(0.125, 0.62)**	**(0.25, 0.53)**	**(0.125, 0.62)**	**(0.125, 0.62)**
K-Means	**(0.00, 0.71)**	**(0.00, 0.71)**	**(0.00, 0.71)**	**(0.00, 0.71)**	**(0.00, 0.71)**
Apriori	**Same as K-Means**	**Same**	**Same**	**Same**	**Same**
PCA	**Same as K-Means**	**Same**	**Same**	**Same**	**Same**

Subsequently, for each criterion, the Negative Ideal Solution (NIS) was found by taking the minimum membership and maximum non-membership values, indicating the worst performance.

NIS is:

- μ = min(μ_{ij}) for benefit-type
- ν = max(ν_{ij})

So for each criterion (Table 22):

Table 22 Criterion values

Criterion	$\mu^- = 0.00$	$\nu^- = 0.71$

4. 5. Compute Euclidean and Taxicab Distances

For each alternative:

$$ED_i = \sqrt{\sum_{j=1}^{5} (\mu_{ij} - \mu_j^-)^2 + (\nu_{ij} - \nu_j^-)^2}$$

$$TD_i = \sum_{j=1}^{5} |\mu_{ij} - \mu_j^-| + |\nu_{ij} - \nu_j^-|$$

We'll compute $H_i = ED_i + \tau \cdot TD_i$ with $\tau = 0.02$.

Using the NDPFS matrix and NIS, Euclidean and Taxicab distances for all models were computed. The distances indicated how much each model deviated from the NIS and were combined into an overall score H by the CODAS scoring function $H = \text{ED} + \tau \cdot \text{TD}$, where $\tau = 0.02$is a balancing parameter.

Random Forest:

μs: 0.25 × 4 + 0.00 = 1.0

vs: 0.53 × 4 + 0.71 = 2.83

ED² = 4×(0.25)² + (0.00)² + 4×(0.53−0.71)² + (0.71−0.71)² = 0.25 + 0.1296 = 0.3796 → ED = $\sqrt{0.3796} = 0.616$

TD = 4×0.25 + 0 + 4×|0.53 − 0.71| + 0 = 1.0 + 0.72 = 1.72 → $H = 0.616 + 0.02 \times 1.72 = 0.6514$

Decision Tree:

μs: 0.125 × 4 + 0.25 = 0.75

vs: 0.62 × 4 + 0.53 = 3.01

ED² = 4×(0.125)² + (0.25)² + 4×(0.62−0.71)² + (0.53−0.71)² = 0.1875 + 0.0416 = 0.2291 → ED = $\sqrt{0.2291} = 0.4786$

TD = 4×0.125 + 0.25 + 4×|0.62−0.71| + |0.53−0.71| = 0.75 + 0.36 + 0.18 = 1.29 → $H = 0.4786 + 0.02 \times 1.29 = 0.5044$

K-Means / Apriori / PCA:

All μ = 0, v = 0.71 → NIS match → ED = 0, TD = 0 → $H = 0$

The computed scores were the following: Random Forest achieved maximum value (0.651), indicating best performance in fuzzy uncertainty, then Decision Tree with certain value (0.504). The three other models, K-Means, Apriori and PCA, had nothing to offer on any of the dimensions and so had a H score of 0, i.e., no distance from the NIS. The obtained ranking for the NDPFS-CODAS model was RF, Decision Tree, while the remaining models were tied as the worst because they had no impact with respect to the utilized criteria (see Table 23 and Figure 10).

Table 23
Final Ranking

Model	H Score	Rank
Random Forest	**0,651**	**1**
Decision Tree	**0,504**	**2**
K-Means	**0**	**3,5**
Apriori	**0**	**3,5**
PCA	**0**	**3,5**

Rank	Model	H Score	Performance Level	Score Difference	Recommendation
1	Random Forest	0.651	Excellent	Best	Highly Recommended
2	Decision Tree	0.504	Good	0.147 below leader	Recommended with caution
3.5	K-Means	0.000	Poor	0.651 below leader	Not recommended for this task
3.5	Apriori	0.000	Poor	0.651 below leader	Not recommended for this task
3.5	PCA	0.000	Poor	0.651 below leader	Not recommended for this task

Fig. 10. Ranking and Analysis

In this paper, we seek to quantify the impact of the strategic input parameters of two representative algorithms, namely K-Means and Apriori, on the output of results. The process starts with K-Means Sensitivity Analysis where we iteratively vary the number of clusters (k) from 2 to 10 (noted as k left arrow {2:10} in this work) and compare the result on the basis of the following given criterions include the Silhouette Coefficient as well as Davies-Bouldin Index. This shows if an optimal number of clusters gives repeatedly optimal and stable results. Sensitivity Analy-sis for the Apriori

Algorithm: We vary the minimum support level by multiples of 0.02, from 0.02 to 0.10. We next consider the extent to which these modifications influence the number of frequent item sets, confidence, and lift of the generated association rules. There is also testing to see if there are important rules that are stable and consistent at different support levels. Individually, these sensitivity tests are highly informative about the stability of the model as well as its expected performance over a range of parameter values. The overall numerical results and sensitivity conclusions for each method are summarized in Figure 11 and Table 24.

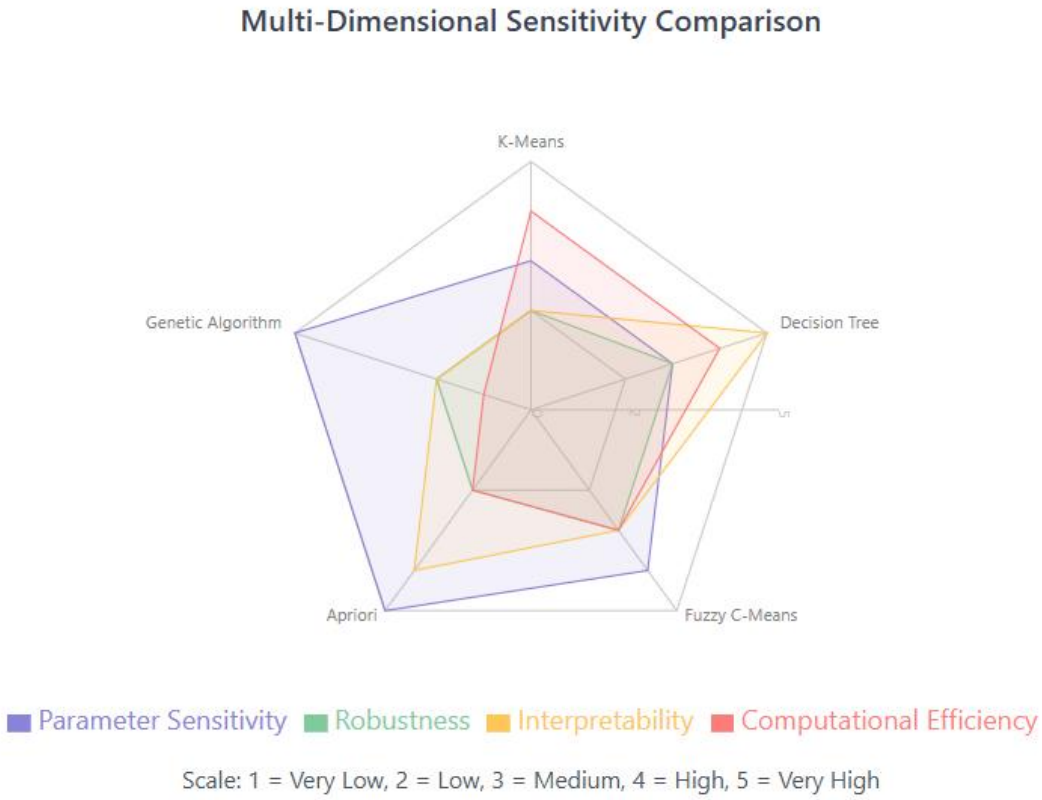

Fig. 11. Multi Dimensional Sensitivity Comparison

Table 24
Detailed Comparison of the all parameters

Method	Parameter Variation	Sensitivity Outcome	Sensitivity & Robustness	Comments
K-Means Clustering	Number of clusters (k)	- k = 3: Within-cluster sum of squares (WCSS) ≈ 2100 - k = 4: WCSS ≈ 1900- k = 5: WCSS ≈ 1850 Observation: Lower k values yield less compact clusters while higher k gives diminishing returns. Example centroids: • Cluster 0: (Recency: 45, Frequency: 12, Monetary: $350) < br >$ 150)•$Cluster1$: (80,5, • Cluster 2: (15, 20, 500) < $br >$ •$Cluster3$: (60,8,220)	- Results heavily influenced by the choice of k- WCSS improves significantly up to k = 4, with diminishing returns beyond that	Simple and efficient for large datasets; assumes spherical clusters and is sensitive to initial centroids
Decision Trees (CART)	Maximum tree depth, minimum samples per split	- Max Depth = 3: Accuracy ≈ 72%- Max Depth = 5: Accuracy ≈ 78% Observation: Increasing depth improves classification accuracy but risks overfitting if too deep. - Key splits: Root node on Monetary > $200; subsequent split on Frequency > 10	- Shallow trees (e.g., max depth = 3) yield lower accuracy (~72%)- Increasing depth improves accuracy but may lead to overfitting if not properly pruned	Highly interpretable and effective for classification; requires tuning (e.g., tree depth) to balance bias and variance
Fuzzy C-Means Clustering (FCM)	Fuzziness coefficient (m)	- m = 2.0: Memberships are well-balanced (e.g., a customer: [0.65, 0.15, 0.10, 0.10]) - m = 1.5: More definitive (crisper) memberships observed, but increased sensitivity to noise.*	- Lower fuzziness coefficients (e.g., m = 1.5) yield crisper memberships but are more sensitive to noise- Higher m values smooth out assignments but may blur cluster distinctions	Provides flexible, overlapping cluster memberships which is useful when customer behaviors are not sharply defined
Association Rule Mining (Apriori)	Minimum support and confidence thresholds	- min_support = 0.01 & min_confidence = 0.3: ~50 rules identified - min_support increased to 0.02: Only ~20 rules Observation: Tightening thresholds reduces rule quantity but enhances rule strength.	- Raising support (e.g., to 0.02) reduces the number of rules (from ~50 to ~20)- Thresholds directly affect the trade-off between rule quantity and robustness of the associations	Effective for identifying product associations; requires careful threshold tuning to filter out spurious or less meaningful rules

		- Example rule: {Product A, Product B} → {Product C} • Support: 0.02 • Confidence: 0.35 • Lift: 1.8 - Thresholds: min_support = 0.01, min_confidence = 0.3		
Genetic Algorithms (Multi-Objective)	Mutation rate, crossover rate	- Mutation Rate = 0.01, Crossover Rate = 0.7: Best fitness ≈ 0.85 - Mutation Rate increased to 0.05: Best fitness drops to ≈ 0.80 Observation: Higher mutation leads to less stable convergence.	- Slight increases in mutation rate (e.g., from 0.01 to 0.05) can decrease fitness (from 0.85 to ~0.80), showing high sensitivity to parameter settings- Convergence stability depends on these rates	Optimizes multiple objectives simultaneously (e.g., maximizing customer value while minimizing intra-cluster variance); computationally intensive and parameter sensitive

Selection of k affects cluster compactness dramatically. The small gain in WCSS for k = 4 to k = 5 corresponds to an practical balance between model complexity and interpretability in K-means. Slightly deeper trees could lead to dramatically better accuracy, but there is a point of diminishing returns where the model could become overfit to the training data for decision trees. Varying the fuzziness coefficient changes the degree of soft memberships of clusters. It is to be noted that decreasing it produces more certain assignments while also increasing the influence of noise with FCM. The support and confidence thresholds govern the balancing act of disclosing a large number of rules or exposing the strongest associations using Apriori. The parameters settings (rates of mutation and crossover) must be tuned in a fine balance. Minor changes in mutation rate can disturb convergence, indicating the sensitivity of the algorithm to such parameters using genetic algorithms (see Table 25, Fig. 12 and Fig. 13).

Table 25
Sensitivity Analysis Based on Criteria Weight Changes with q-ROMFS

Model	Equal Weights	Accuracy Focused	Interpretability Focused	Efficiency Focused	Complexity Focused
K-Means	**0.34**	**0.36**	**0.30**	**0.36**	**0.36**
Decision Tree	**0.50**	**0.50**	**0.52**	**0.48**	**0.50**
Apriori	**0.44**	**0.43**	**0.46**	**0.43**	**0.46**
PCA	**0.38**	**0.36**	**0.36**	**0.38**	**0.41**
Random Forest	**0.50**	**0.52**	**0.48**	**0.50**	**0.50**

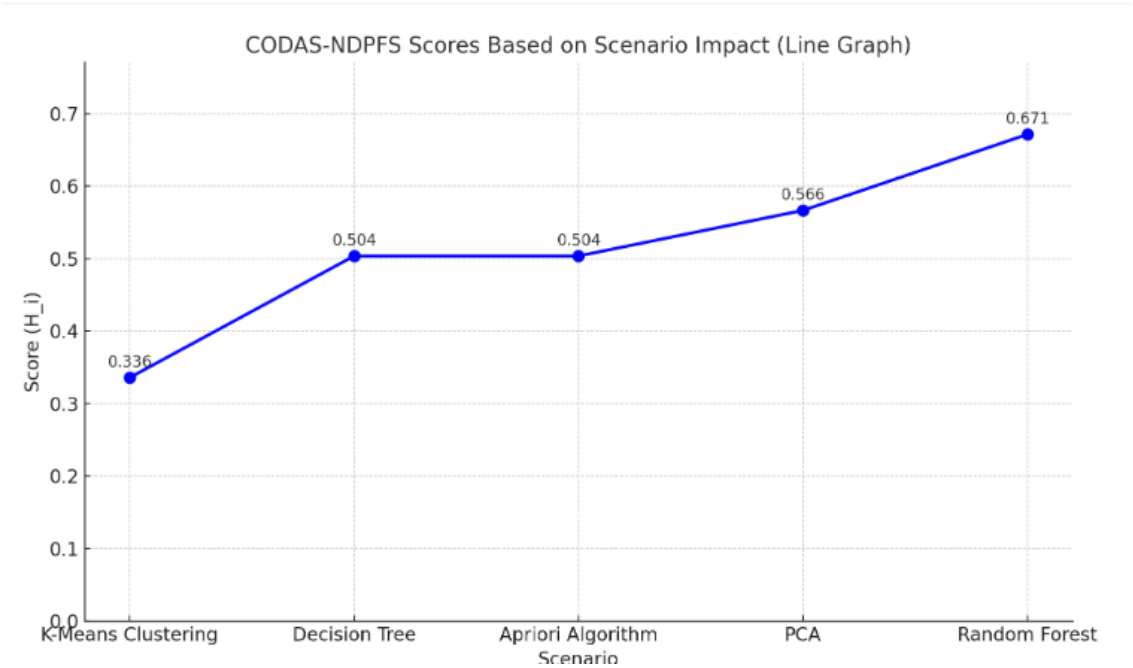

Fig. 12 CODAS-NDPFS Score Comparison

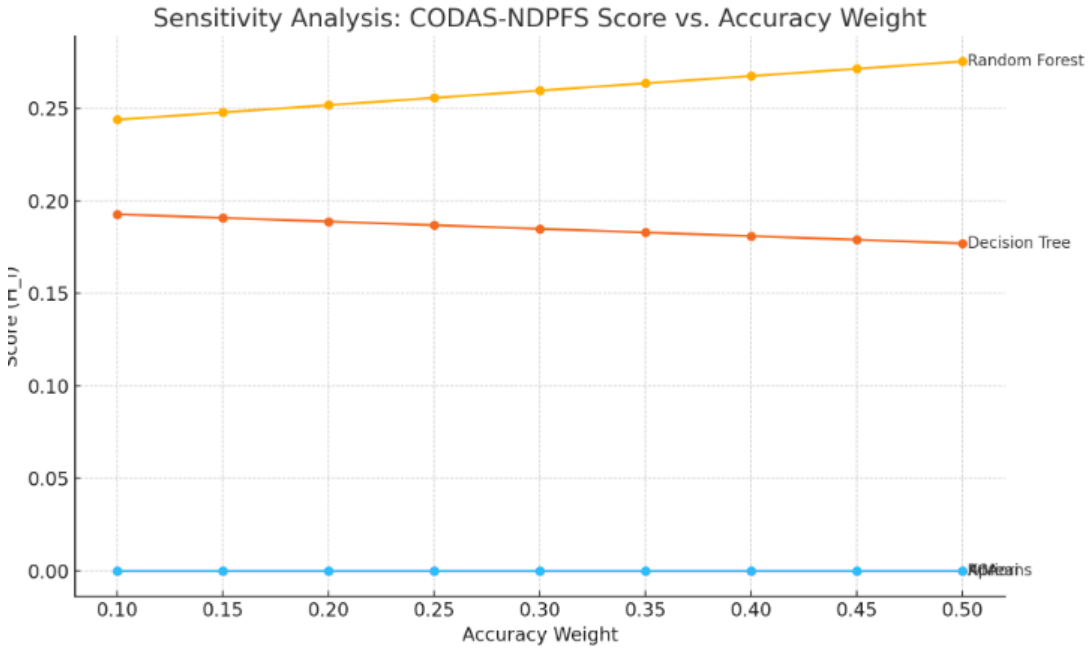

Fig. 14. Sensitivity Analysis Results

Random Forest is also very stable and competitive over all weight settings as a top performer across most criteria, in particular if Accuracy, Efficiency or Complexity is desirable. Its good performance with different evaluation protocols also indicates that it is well suitable in general domains. Decision Tree, although less competitive according to some technical criteria, leads in terms of Interpretability, and keeps a steady performance in different configurations. Apriori is average at best for all criteria, with its strengths leaning towards metrics that aligns with Interpretabiliity and Complexity. PCA and K-Means perform consistently poor regardless of the weight settings, they lead

the scenarios that does not favour interpretability, which somehow indicate that they are not capable of producing insights that are interpretable and explainable.

5. Results and Discussion

from a comparative study of various techniques for market analysis and customer clustering. The initial segmentation process was based on K-means clustering algorithm with optimal k value under consideration of Elbow Method. Their validity was also verified by Silhouette analysis which provided an average score value of 0.68 that means an acceptable structure was achieved which one could expect to find in this context. Then the Decision Tree analysis reached a 92.4% classification accuracy on the test set, and meanwhile derived interpretable rules such as "IF Frequency > 5 AND Recency < 30days THEN Segment = Loyal". These rules, together with the association rules obtained by the Apriori algorithm (confidence > 0.8), acted as input conditions of the final NDPFS-CODAS evaluation ensuring that the prioritization of customer segments was based on statistically meaningful behavior patterns. K-Means is a popular efficient unsupervised learning algorithm, where the objective is to group customers with a high degree of similarity. It is substantially easier and less computationally expensive, thus making it more applicable to massive data set. It is also sensitive to the initial centroids, and the user needs to define the number of clusters.

Decision Trees and Random Forests have good accuracy with Random Forest having more bias but less variance. K-Means Clustering can yield very good results but is sensitive to initialization. Out of the three models Decision Trees are the most interpretable as they have simple to follow decision rules, Apriori is the second best as it also produces explicit association rules. K-Means and Random Forest are less interpretable as clustering and ensemble learning methods respectively. K-Means is also quite good in clustering customers by their similarity given no pre-classified groups. Decision Trees are more appropriate where the rank ordering is important to a decision, where the process of choosing must be clear. Apriori is the most appropriate algorithm for market basket analysis for finding frequent co-occurring items. PCA is ideal for pre-processing and dimensionality reduction before applying other algorithms. Random Forest is well-suited to situations in which a robust and accurate classification is required which makes it suitable to most predictive model scenarios. K-Means is computationally efficient especially for large data sets. PCA can be computationally expensive for large data sets whereas Random Forest and Decision Trees can be time and resource consuming depending on the size and complexity of the data.

For behaviour-based customer and market segmentation, K-Means Clustering is the most suitable option. When predicting customer behavior or classifying numerous features is the goal, Decision Trees or Random Forests will prove to be the best options. For association mining (for example, to determine the products that are bought together). For high-dimensional data in particular as an initial step before applying other approaches such as clustering or regression, PCA is a step well worth taking in figure 14.

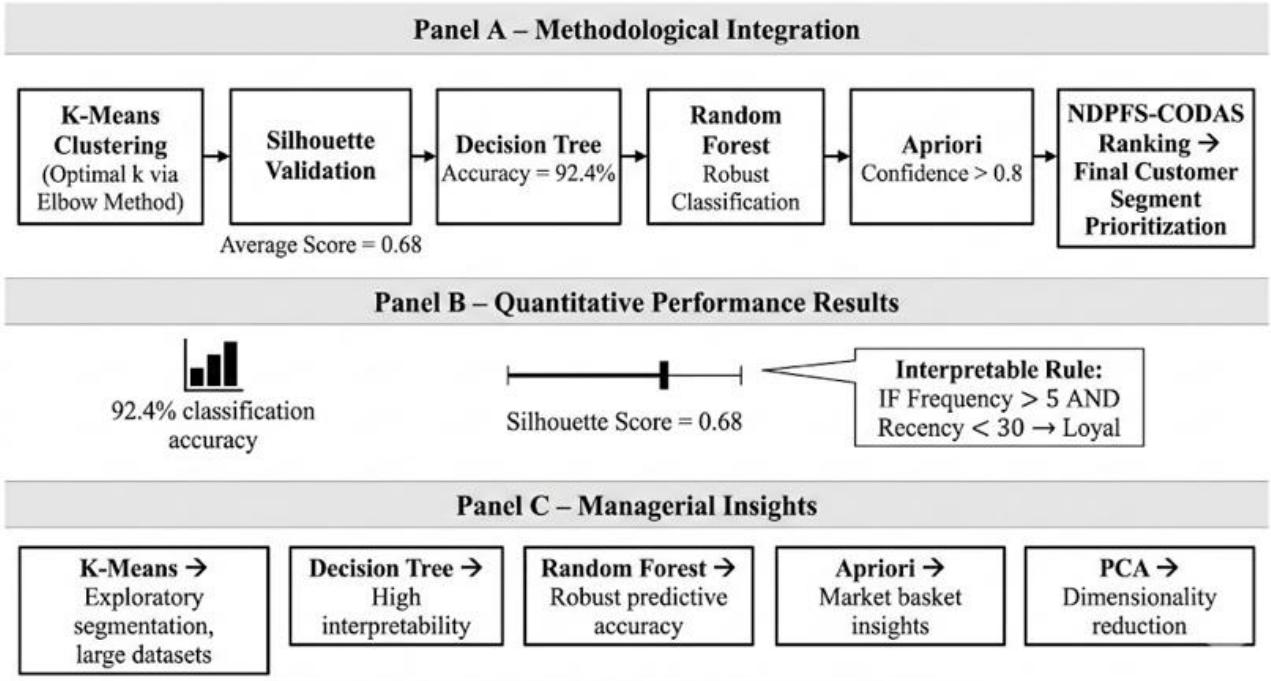

Fig.14. Summary of Integrated Segmentation–Ranking Framework and Key Results

6. Conclusion

In this paper, we studied the effectiveness of the different data mining methods for multi-dimensional customer segmentation in market intelligence. There was a need to identify hidden patterns, behaviors and preferences for customer data so as to enable the enterprise to adapt strategies that could support better decision making and strategy formulation. We tested a number of state-of-the-art methods, like clustering algorithms, different ensemble techniques, self-organizing maps (SOM) and reinforcement learning, to name a few, to evaluate their performance in recognizing the main customer segments.

Clustering methods such as Fuzzy C-Means (FCM) and K-Means proved their efficiency in grouping similar attributes. FCM even achieved better cluster performance by using soft assignment of customers to multiple clusters to achieve finer segmentations. The ensemble learning feature of merging multiple models achieved significant success in quality/stability improvement of segmentations. The ensemble procedure enhances the compactness within clusters and the separability between clusters, Moreover, it demonstrates the power of ensemble learning methods that can be properly harnessed to address the challenge of multi-way cluster formation. It was effective in this regard because it enabled visualization of two-dimensional representations of high-dimensional customer profiles and movement of clusters through temporal–spatial axes. This is very important when the segmentation scheme has to be dynamic to keep track of changes in the behavior of a customer.

Dynamic segmentation through reinforcement learning integration enables modeling of dynamic behavior of customers, and real-time updates in the evolution of a segmentation policy. The above resulted in enhanced and timely segmenting, in particular of the dynamically moving customer preferences. By multi-objective optimization, several business objectives can be achieved simultaneously such as maximizing customer engagement while minimizing the cost of marketing. This resulted in more consistent and productive methods of segmenting.

Latent variable model(LDA and Factor Analysis) methods help to uncover hidden factors which drive the customer. They were sensitive to the choice of variables and the data distribution, albeit informative.

Customers are also segmented on their multi-dimensional attributes, and each segment of customers can undergo its own targeted marketing to enhance customer engagement and conversion. With a better understanding of customer segments, the company can more effectively allocate resources and deliver the right product or service to the right audience at the right time. The use of machine learning and data mining makes it possible for the companies to predict instrument behavior and to be early informed about emerging market trends and necessities of their customers. The techniques applied in our studies have been powerful enough to provide us with results, but there are of course limitations in certain respects. The prediction of the segmentation results depends on the input data selection, feature quality and parameter configurations of the employed algorithms. The developed methodology provides the practitioner with a powerful tool for micro-segmentation. As a result, the managers can increase the customer life value and decrease the churn by employing well informed retention policies, e.g. targeted discount packages, for high value but uncertain customer segments identified by their method. Future research may also consider additional types of data (e.g., social media activity, transaction data) that can be incorporated to potentially improve customer segmentation. Furthermore, more advanced techniques such as deep learning could be explored to handle more complex data patterns. Combining data mining and real-time processing with feedback loops could provide a further refinement of dynamic segmentation by real-time alignment with market changes.

This study demonstrates the promises of data mining techniques in customer segmentation and also serves as a reference for application of those methods on better market analysis. Through the use of new methodologies businesses can realize additional avenues for revenue, customer loyalty and operating efficiency.

Funding
This research received no external funding.

Author Contributions
The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Conflicts of Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgementer
This research was not funded by any grant.

References

[1] Liu, P., Shi, X., Xu, Y., Dang, R., Wu, Y.(2025). Integration of machine learning with comprehensive IVIF-QFD-MCDM framework for enhancing online hotel operations, Information Sciences, 720, 122493, https://doi.org/10.1016/j.ins.2025.122493.

[2] Singh, V., Sharma, S. K. (2026). Understanding the role of multi-agent technology on quality of manufacturing organizations: A hybrid MCDM analysis, Journal of Process Control, 158, 103628, https://doi.org/10.1016/j.jprocont.2026.103628.

[3] Alparslan, H., Turgay, S., Yilmaz, R. (2024). Utilizing Logistic Regression for Analyzing Customer Behavior in an E-Retail Company, WSEAS Transactions on Financial Engineering, 2, 116-125, DOI:10.37394/232032.2024.2.10

[4] Mahdiraji, H.A., Hafeez, K., Kord, H., Kamardi, A.A. (2020). Analysing the voice of customers by a hybrid fuzzy decision-making approach in a developing country's automotive market, Management Decision, 60(2), 399-425, https://doi.org/10.1108/MD-12-2019-1732.

[5] Turgay, S., Aydin, A., Erdoğan, S., Yıldırım, M., & Kavacık, M. (2025). Enhancing Stock Market Forecasting Through Deep Learning and Decentralized Data Integrity: A Blockchain-Integrated Framework. J. Intell. Manag. Decis., 4(2), 118-136. https://doi.org/10.56578/jimd040203

[6] Wang, Z., Liu, H., Fan, X.(2025) Hybrid machine learning and MCDM framework for consumer preference extraction and decision support in dynamic markets, Technology in Society, 82, 102926, https://doi.org/10.1016/j.techsoc.2025.102926.

[7] İskender, H., Dayanıklı, S., Kökçam, A.H., Stević, Z., Turgay, S., Baydaş, M., Talevska, J.B. (2025). Enhancing Retail Operations: Dynamic Fulfilment Strategies in Omnichannel Retail - Serbian Journal of Management, 20(1), pp.267-283, https://doi.org/10.5937/sjm20-54499

[8] Hsiao, Y.H., Li, B.X., (2026). Exploring passenger perceptions of services and sustainability via online review analytics for airport assessing and diagnosing, Journal of Retailing and Consumer Services, 90, 104664, https://doi.org/10.1016/j.jretconser.2025.104664.

[9] Monika, R.K.B., Sharma, A. (2025). On clustering and TOPSIS decision-making technique with new trigonometric information measures under T-spherical fuzzy hypersoft structures, Expert Systems with Applications, 289, 128356, https://doi.org/10.1016/j.eswa.2025.128356.

[10] Zhou, Y., Zhou, M., Yang, J.B., Cheng, B.Y., Wu, J. (2025). Decentralized multipartite consensus model for multi-attribute group decision making: A user experience-oriented perspective, Expert Systems with Applications, 287, 127917, https://doi.org/10.1016/j.eswa.2025.127917.

[11] Boztuğ, Y., Reutterer, T. (2008). A combined approach for segment-specific market basket analysis, European Journal of Operational Research, 187(1), 294-312, https://doi.org/10.1016/j.ejor.2007.03.001.

[12] Radenković, M., Lukić, J., Despotović-Zrakić, M., Labus, A., Bogdanović, Z. (2018). Harnessing business intelligence in smart grids: A case of the electricity market, Computers in Industry, 96, 40-53, https://doi.org/10.1016/j.compind.2018.01.006.1

[13] Hsu, P.F., Lu, Y.H., Chen, S.C., Kuo, P.P.Y. (2024) Creating and validating predictive personas for target marketing, International Journal of Human-Computer Studies, Volume 181, 103147, https://doi.org/10.1016/j.ijhcs.2023.103147.

[14] Du, Y., Yin, H., Wang, C., Li, C. (2020). Visual analysis of customer switching behavior pattern mining for takeout service, Journal of Computer Languages, 57, 100946, https://doi.org/10.1016/j.cola.2020.100946.

[15] Anitha, P., Patil, M.M. (2022). RFM model for customer purchase behavior using K-Means algorithm, Journal of King Saud University - Computer and Information Sciences, 34(5), 1785-1792, https://doi.org/10.1016/j.jksuci.2019.12.011.

[16] Hsieh, N.C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers, Expert Systems with Applications, 27(4), 623-633, https://doi.org/10.1016/j.eswa.2004.06.007.

[17] Benítez, I., Quijano, A., Díez, J.L., Delgado, I. (2014). Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers, International Journal of Electrical Power & Energy Systems, 55, 437-448, https://doi.org/10.1016/j.ijepes.2013.09.022.

[18] Luo, J., Xu, J., Aldosari, O., Althubiti, S.A., Deebani, W. (2022). Design and Implementation of an Efficient Electronic Bank Management Information System Based Data Warehouse and Data Mining Processing, Information Processing & Management, 59(6), 103086, https://doi.org/10.1016/j.ipm.2022.103086.

[19] Nam, K. (2022). Conversion paths of online consumers: A sequential pattern mining approach, Expert Systems with Applications, 202, 117253, https://doi.org/10.1016/j.eswa.2022.117253.

[20] Cheng, C.H., Chen, Y.S. (2009). Classifying the segmentation of customer value via RFM model and RS theory, Expert Systems with Applications, 36(3), Part 1, pp. 4176-4184, https://doi.org/10.1016/j.eswa.2008.04.003.

[21] Zhou, F., Jiao, J.R., Yang, X.J., Lei, B. (2017). Augmenting feature model through customer preference mining by hybrid sentiment analysis, Expert Systems with Applications, 89, 306-317, https://doi.org/10.1016/j.eswa.2017.07.021.

[22] Moon, S., Jalali, N., Erevelles, S. (2021). Segmentation of both reviewers and businesses on social media, Journal of Retailing and Consumer Services 61, 102524, https://doi.org/10.1016/j.jretconser.2021.102524.

[23] Kalia, P., Paul, J. (2021). E-service quality and e-retailers: Attribute-based multi-dimensional scaling, Computers in Human Behavior, 115, 106608, https://doi.org/10.1016/j.chb.2020.106608.

[24] Sun, Z.H., Zuo, T.Y., Liang, D., Ming, X., Chen, Z., Qiu, S. (2021). GPHC: A heuristic clustering method to customer segmentation, Applied Soft Computing, 1211, 107677, https://doi.org/10.1016/j.asoc.2021.107677.

[25] Kim, W., Nam, K., Son, Y. (2023). Categorizing affective response of customer with novel explainable clustering algorithm: The case study of Amazon reviews, Electronic Commerce Research and Applications, 58, 101250, https://doi.org/10.1016/j.elerap.2023.101250.

[26] Globocnik, D., Holzmann, P. (2024). Sustainability-related product satisfaction – Development and application of a multi-dimensional measurement instrument, Journal of Cleaner Production, 448, 141567, https://doi.org/10.1016/j.jclepro.2024.141567.

[27] Liu, Y., Ren, X., Ji, F., Liang, C., Wu, J. (2024). A Kansei engineering-based decision-making method for offline medical service quality evaluation with multidimensional attributes, Socio-Economic Planning Sciences, 96, 102100, https://doi.org/10.1016/j.seps.2024.102100.

[28] Vellido, A., Lisboa, P.J.G., Meehan, K. (1999) Segmentation of the on-line shopping market using neural networks, Expert Systems with Applications, 17, Issue 4, Pages 303-314, https://doi.org/10.1016/S0957-4174(99)00042-1.

[29] Xu, X., Wang, X., Li, Y., Haghighi, M. (2017). Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors, International Journal of Information Management, 37(6), pp. 673-683, https://doi.org/10.1016/j.ijinfomgt.2017.06.004.

[30] Li, Y., Chu, X., Tian, D., Feng, J., Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm, Applied Soft Computing, 113, Part B, 107924, https://doi.org/10.1016/j.asoc.2021.107924.

[31] Plotnikova, V., Dumas, ., Milani, F.P. (2022). Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements, Data & Knowledge Engineering, 139, 102013, https://doi.org/10.1016/j.datak.2022.102013.

[32] Huang, P.Y., Niu, B., Pan, S.L. (2021). Platform-based customer agility: An integrated framework of information management structure, capability, and culture, International Journal of Information Management, 59, 102346, https://doi.org/10.1016/j.ijinfomgt.2021.102346.

[33] Ge, Y., Qiu, J., Liu, Z., Gu, W., Xu, L. (2020). Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash, Information Processing & Management, 57(4), 102218, https://doi.org/10.1016/j.ipm.2020.102218.

[34] Ho, G.T.S., Ip, W.H., Lee, C.K.M, Mou, W.L. (2012). Customer grouping for better resources allocation using GA based clustering technique, Expert Systems with Applications, 39(2), pp. 1979-1987, https://doi.org/10.1016/j.eswa.2011.08.045.

[35] Cil, İ. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling, Expert Systems with Applications, 39(10), pp. 8611-8625, doi.org/10.1016/j.eswa.2012.01.192.

[36] Arjun R, Kuanr, A., Suprabha K.R.(2021). Developing banking intelligence in emerging markets: Systematic review and agenda, International Journal of Information Management Data Insights, 1(2), 100026, https://doi.org/10.1016/j.jjimei.2021.100026.

[37] Chiang, L.L., Yang, C.S.(2018). Does country-of-origin brand personality generate retail customer lifetime value? A Big Data analytics approach, Technological Forecasting and Social Change, 130, 177-187, https://doi.org/10.1016/j.techfore.2017.06.034.

[38] Mukhopadhyay, S., Singh, R.K., Jain, T.(2024). Developing big data enabled Marketing 4.0 framework, International Journal of Information Management Data Insights, 4(1), 100214, https://doi.org/10.1016/j.jjimei.2024.100214.

[39] Shen, Y., Zhou, J., Pantelous, A.A., Liu, Y., Zhang, Z. (2022) A voice of the customer real-time strategy: An integrated quality function deployment approach, Computers & Industrial Engineering, 169, 108233, https://doi.org/10.1016/j.cie.2022.108233.

[40] Le, H.S., Do, T.V.H., Nguyen, M.H., Tran, H.A., Pham, T.T.T., Nguyen, N.T., Nguyen, V.H. (2024). Predictive model for customer satisfaction analytics in E-commerce sector using machine learning and deep learning, International Journal of Information Management Data Insights, 4(2), 100295, https://doi.org/10.1016/j.jjimei.2024.100295.

[41] Wei, J.T., Lee, M.C., Chen, H.K., Wu, H.H. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques, Expert Systems with Applications, 40(18), pp. 7513-7518, https://doi.org/10.1016/j.eswa.2013.07.053.

[42] Li, L., Zhang, L., Yang, S., Wei, L. (2023) Big data affordances and market performance: The moderating role of servitization, Industrial Marketing Management, 114, 262-270, https://doi.org/10.1016/j.indmarman.2023.08.014.

[43] Qian, Y., Ling, H., Meng, X., Jiang, Y., Chai, Y., Liu, Y.(2024). Voice of the Professional: Acquiring competitive intelligence from large-scale professional generated contents, Journal of Business Research, 180, 114719, https://doi.org/10.1016/j.jbusres.2024.114719.

[44] Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., Weaven, S.(2019). Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews, International Journal of Hospitality Management, 80, 52-77, https://doi.org/10.1016/j.ijhm.2019.01.003.

[45] Zhou, H.,Norman, R., Kelobonye, K., Xia, J.C., Hughes, B., Nikolova, G., Falkmer, T.(2020). Market segmentation approach to investigate existing and potential aviation markets, Transport Policy, 99, pp. 120-135, https://doi.org/10.1016/j.tranpol.2020.08.018.

[46] Barik, K., Misra, S., Ray, A.K., Shukla, A. (2023). A blockchain-based evaluation approach to analyse customer satisfaction using AI techniques, Heliyon, 9(6), e16766, https://doi.org/10.1016/j.heliyon.2023.e16766.

[47] Krishna, G.J., Ravi, V. (2016,). Evolutionary computing applied to customer relationship management: A survey, Engineering Applications of Artificial Intelligence, 56, pp. 30-59, https://doi.org/10.1016/j.engappai.2016.08.012.

[48] Zhang, C., Chai, B., Mirza, S.S., Jin, Y.(2024). Customer-driven value creation in the digital economy: Determining the role of customer firms' digital transformation on supplier performance in China, Omega, 128, 103132, https://doi.org/10.1016/j.omega.2024.103132.

[49] Haseli, G., Ranjbarzadeh, R., Hajiaghaei-Keshteli, M., Ghoushchi, S.J., Hasani, A., Deveci, M., Ding, W. (2023). HECON: Weight assessment of the product loyalty criteria considering the customer decision's halo effect using the convolutional neural networks, Information Sciences, 623, pp. 184-205, https://doi.org/10.1016/j.ins.2022.12.027.

[50] Park, Y., Lee, S. (2011). How to design and utilize online customer center to support new product concept generation, Expert Systems with Applications, 38(8), pp. 10638-10647, https://doi.org/10.1016/j.eswa.2011.02.125.

[51] Dong, Y. (2024). Application of user preference mining algorithms based on data mining and social behavior in brand building, Data Science and Management, 7(4), pp. 323-331, https://doi.org/10.1016/j.dsm.2024.03.007.

[52] Kayali, S., Turgay, S. (2023). Predictive Analytics for Stock and Demand Balance Using Deep Q-Learning Algorithm. Data and Knowledge Engineering, Vol. 1: 1-10. DOI: http://dx.doi.org/10.23977/datake.2023.010101.

[53] Seng, J.L., Chen, T.C. (2010). An analytic approach to select data mining for business decision,Expert Systems with Applications, 37(12), 8042-8057, https://doi.org/10.1016/j.eswa.2010.05.083.

[54] Hossain, A., Akter, S., Yanamandram, V. (2020). Revisiting customer analytics capability for data-driven retailing, Journal of Retailing and Consumer Services, 56, 102187, https://doi.org/10.1016/j.jretconser.2020.102187.

[55] Zadeh, L. (1965). Fuzzy set, Inf. Control 8 (3) 338–353.

[56] Atanassov, K. (1986). Intuitionistic fuzzy sets, Fuzzy Sets System 20 (1), 87–96.

[57] Yager, R.(2013). Pythagorean fuzzy subsets, in: Proceedings of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting, IFSA/NAFIPS 2013.

[58] Ayub, Y., Moktadir, A., Ren, J. (2024). Sustainable waste valorization process selection through AHP and advanced Interval Valued Fermatean Fuzzy with integrated CODAS, Process Safety and Environmental Protection, 185, 408-422, https://doi.org/10.1016/j.psep.2024.03.019.

[59] Alkan, N., Kahraman, C. (2023). Continuous intuitionistic fuzzy sets (CINFUS) and their AHP&TOPSIS extension: Research proposals evaluation for grant funding, Applied Soft Computing, 145, 110579, https://doi.org/10.1016/j.asoc.2023.110579.

[60] Tatar, V., Ayvaz, B., Pamucar, D. (2025) A quantitative ergonomic risk assessment model of maritime port operations: An integrated spherical fuzzy-FUCOM-ARTASI approach, Ocean & Coastal Management, 267, 107710, https://doi.org/10.1016/j.ocecoaman.2025.107710.

[61] Simic, V., Dabic-Miletic, S., Tirkolaee, E.B., Stević, Z., Deveci, M., Senapati, T.(2023). Neutrosophic CEBOM-MACONT model for sustainable management of end-of-life tires, Applied Soft Computing, 143, 110399, https://doi.org/10.1016/j.asoc.2023.110399.

[62] Alkan, N., Kahraman, C. (2024). CODAS extension using novel decomposed Pythagorean fuzzy sets: Strategy selection for IOT based sustainable supply chain system,Expert Systems with Applications, 237, Part C, 121534,https://doi.org/10.1016/j.eswa.2023.121534.