Article

Using Clustering for Customer Segmentation from Retail Data

Henrique José Wilbert , Aurélio Faustino Hoppe , Andreza Sartori , Stefano Frizzo Stefenon [*] ,
Luís Augusto Silva , Valderi Reis Quietinho Leithardt

Article

Using Clustering for Customer Segmentation from Retail Data

Henrique José Wilbert [1], Aurélio Faustino Hoppe [1], Andreza Sartori [1],
Stefano Frizzo Stefenon [2,3,*] Luís Augusto Silva [4] and Valderi Reis Quietinho Leithardt [5,6]

[1]   Information Systems and Computing, Regional University of Blumenau, Blumenau 89030-903, Brazil
[2]   Digital Industry Center, Fondazione Bruno Kessler, Trento 38123, Italy
[3]   Department of Mathematics, Computer Science and Physics, University of Udine, Udine 33100, Italy
[4]   Expert Systems and Applications Lab, Faculty of Science, University of Salamanca, Salamanca, 37008, Spain
[5]   Instituto Superior de Engenharia de Lisboa (ISEL), Instituto Politécnico de Lisboa, 1959-007 Lisbon, Portugal
[6]   VALORIZA, Research Center for Endogenous Resources Valorization. Instituto Politécnico de Portalegre, 7300-555 Portalegre, Portugal
*   Correspondence: sfrizzostefenon@fbk.eu

Abstract: While there are several ways to identify customer behaviors, few extract this value from information already in a database, much less extract relevant characteristics. This paper presents the development of a prototype using the recency, frequency, and monetary attributes for customer segmentation of a retail database. For this purpose, the standard K-means, K-medoids, and MiniBatch K-means were evaluated. The standard K-means clustering algorithm was more appropriate for data clustering than other algorithms as it remained stable until solutions with 6 clusters. The evaluation of the clusters' quality was obtained through the internal validation indexes: Silhouette, Calinski Harabasz, and Davies Bouldin. Once consensus was not obtained, three external validation indexes were applied: global stability, stability per cluster, and segment-level stability across solutions. Six customer segments were obtained, identified by their unique behavior: Lost customers, disinterested customers, recent customers, less recent customers, loyal customers, and best customers. Their behavior was evidenced and analyzed, indicating trends and preferences.

Keywords: Retailing, customer behavior, clustering, segmentation, external validation indices.

---

1. Introduction

With the evolution of information technology in the 1990s [1], large companies adopted management systems in the form of enterprise resource planning (ERP) software [2]. This software helps in their routines at the operational level, whether in inventory control, tax, financial, transactions, and even human resources [3]. As a result of this, a level of efficiency never conceived was reached, since records previously made on paper and pen began to be produced automatically. In parallel with the computerization of these processes, there was also a growth in the amount of data stored relating to products, customers, transactions, expenses, and revenues [4].

In this context, direct marketing tactics were also advanced, such as sending catalogs by mail, up to highly targeted offers to selected individuals whose transaction information was present in the database. The focus of company-customer relationships then turned to customers who already have a record with the company, since the cost of acquiring a new customer through advertising is much higher than the cost of nurturing an existing relationship [5].

With the increase in the amount of data and the manual work required for segmentation [6], Oyelade et al. [7] state that the automation of this process has become indispensable, and one of its main techniques is clustering. This technique consists of categorizing unlabeled data into groups called clusters, whose members are similar to each other and different from members of other clusters, based on the characteristics analyzed. The cluster methods are increasingly being used for several applications [8–11]. As presented in [12] and [13], state-of-the-art clustering methods can be inspired by the behaviors of animals.

Among the clustering algorithms, the K-means algorithm is one of the most popular, being simple to implement and having extensive studies on its behaviors. In the context of evaluation, Hämäläinen, Jauhiainen, and Kärkkäinen [14] point out that the quality of a solution can be measured through validation indices, which consider the compactness of the cluster data and its separation with other clusters, allowing a higher degree of certainty to be obtained when evaluating a segmentation result coming from a clustering algorithm.

Given the importance of customer segmentation, and extracting their behavioral characteristics effectively, this paper presents the creation of a prototype that uses the attributes of the RFM model together with the K-means clustering algorithm. It automatically extracts information from a real retail database to identify different customer segments based on their behavior. To validate the number of clusters three internal indexes and three external indexes: global stability, stability per cluster, and segment-level stability across solutions (SLSa) were used to highlight the quality of the solutions obtained [15].

The main contribution of this work concerns the application of external validation algorithms since internal validation algorithms proved to be incoherent in their suggestions for the database used. Furthermore, in the research process of references for the work, few works were found that use even one external validation algorithm, and this work used three, establishing a line of reasoning between the results presented by the indices. Part of the value of this work resides in the evolution process between the choices for validating the number of clusters, as well as its application in a set of data coming from customers and real purchases.

This paper demonstrated an alternative for when the internal indexes fail in their congruence, presenting the aspect related to replicability and stability of K-means clustering (external indexes), focusing on the analysis of the optimal number of clusters. The K-medoids and MiniBatch K-means were compared to the standard K-means.

The K-medioids algorithm differs from K-means on the issue of centroids for calculating the center point of the cluster. K-medoids assign an existing point to represent the center, while K-means assign it to an imaginary point by averaging the distances of the points contained in the current cluster. The MiniBatch K-means algorithm is an attempt to reduce the computational expense of the original algorithm, where each iteration is applied to parcels or subsets (batches) of the original data, their constituents being chosen randomly in each iteration.

2. Related Works

According to Reinartz, Thomas, and Kumar [16], when companies treat spending between customer acquisition and retention, allocating fewer resources to retention will result in lower profitability in the long term, compared with lower investments in customer acquisition. According to the authors, the concept of retention relationships places great emphasis on customer loyalty and profitability, where loyalty is the customer's tendency to buy from the company, and profitability is the general measure of how much profit a customer brings to the company through his or her purchases.

The use of artificial intelligence models with fuzzy logic for data segmentation can be a promising alternative, being superior to deep learning models [17]. Techniques based on fuzzy logic have been increasingly used for their high-performance results for insulator fault forecasting [18], prediction of the safety factor [19], power forecasting [20]. There is a growing trend to use simpler models in combination to solve difficult tasks, such as fault [21], price [22], contamination [23], load [24], and/or signal forecasting [25].

Despite this trend, many authors still use deeper layer models to solve more difficult tasks, such as fault classification [26], epidemic prediction [27], classification of defective components [28], emotion recognition [29], power generation evaluation [30], and attitude determination [31]. The use of hybrid models is still overrated in this context [32]. Been possible to apply in cryptocurrencies [33].

In addition to applications for data segmentation [34], applications for internet of things (IoT) [35], classification (using deep neural networks [36], k-nearest neighbors [37], and multilayer perceptron

[38]), optimization [39], and prediction [40–42] stand out. According to Nguyen, Sherif, and Newby [43], with the advancement of customer relationship management, new ways were opened through which customer loyalty and profitability can be cultivated, attracting a growing demand from companies, since the adoption of these means allows organizations to improve their customer service.

Different tools end up being used, such as recommendation systems that, usually in e-commerce branches, consider several characteristics pertinent to the customer's behavior, building a profile of their own that will be used to make a recommendation for a product that may be of interest. Another tool relevant to profits and loyalty is segmentation, which aims to separate a single mass of customers into homogeneous segments in terms of behavior, allowing the development of campaigns, decisions, and marketing strategies specialized to each group according to their characteristics [44].

Roberts, Kayande, and Stremersch [45] state that segmentation tools have the greatest impact among available marketing decisions, indicating a high demand for such tools over the next decade. Dolnicar, Grün, and Leisch [46] inquire that customer segmentation presents many benefits if implemented correctly, among the main ones is the introspection by the company about the types of customers it has, and consequently, their behaviors and needs. On the other hand, Dolnicar, Grün, and Leisch [46] also point out that if segmentation is not applied correctly, the implementation of the practice in its entirety generates a waste of resources, since the failure returns segments that are not consistent with the actual behavior, leaving the company that applied it with no valid information about the customers it has.

In relation to customer segmentation, some metrics become relevant in the contexts in which they are inserted. According to Kumar [47], the recency frequency monetary (RFM) model is used in companies that sell by catalog, while high-tech companies tend to use a share of wallet (SOW) to implement their marketing strategies. The past customer value (PCV) model, on the other hand, is generally used in financial services companies. Among the models mentioned above, RFM is the easiest to apply in several areas of commerce, retail, and supermarkets, since only transactions data (sales) of customers are required, from which the attributes of recency (R), frequency (F), and monetary (M) are obtained.

Based on this data, according to Tsiptsis and Chorianopoulos [48], it is possible to detect good customers from the best RFM scores. If the customer has recently made a purchase, his *R* attribute will be high. If he buys many times during a given period, his F attribute will be higher. Finally, if his total spending is significant, he will have a high *M* attribute. By categorizing the customer within these three characteristics, it is possible to obtain a hierarchy of importance, with customers who have high RFM values at the top, and customers who have low values at the bottom.

Despite these possibilities for segmentation, the original standard model is somewhat arbitrary, segmenting customers into quintiles, five groups with 20% of the customers, and not paying attention to the nuances and all the interpretations that the customer base can have. In addition, the method can also produce many groups (up to 125), which often do not significantly represent the customers of an establishment. Table 1 summarizes the main characteristics listed from related works. Gustriansyah, Suhandi, and Antony [49] grouped products from a database using the standard RFM model. Peker, Kocyigit, and Eren [50] opted for the development of a new model, considering the periodicity (LRFMP). Tavakoli et al. [51] also developed a new model, to which the recency feature was modified and separated (R+FM).

Table 1. Comparison between related works.

Related / Characteristics	Gustriansyah, Suhandi, and Antony [49]	Peker, Kocyigit, and Eren [50]	Tavakoli et al. [51]
Clustering target	Products	Customers	Customers
Model used	RFM	LRFMP	R+FM
Targeting objective	Inventory management	Customer relationship management	Customer relationship management
Clustering algorithm used	K-means	K-means	K-means
Methodological focus	Optimization of k with different metrics	Formulation of a new model and analysis of results	Formulation of a new model and offer campaign
Number of data (customers/products)	2.043	16.024	3.000.000
Number of indices for k validation	8 (Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski-Lai Index)	3 (Silhouette, Calinski-Harabasz and Davies-Bouldin)	Not applicable
Number of generated clusters	3	5	10
Inferences about the data	not applicable	Yes	Yes
Using external indexes	No	No	No

Gustriansyah, Suhandi, and Antony [49] aimed to improve inventory management, valuing a more conclusive segmentation of products, since the standard RFM model arbitrarily defines segments without adapting to the peculiarities of the data, while the model applied through K-means achieved a segmentation with highly similar data in each cluster. On the other hand, Peker, Kocyigit, and Eren [50] and Tavakoli et al. [51] aimed at managing customer relationships through strategies focused on segments, aiming to increase the income they provide to the company. All authors used the K-means algorithm, as it is reliable and widely used. It is noteworthy that in the work by Gustriansyah, Suhandi, and Antony [49], the algorithm had a greater methodological focus, since 8 validation indexes were used for k clusters, aiming to optimize the organization of the segments.

The amount of segmented data varied greatly between the three works due to the different application contexts. Gustriansyah, Suhandi, and Antony [49] had 2,043 products in the database to segment, resulting in 3 clusters. They had a record of 16,024 customers of a bakery chain, with 5 segments specified, obtained through analysis by three validation indices (Silhouette, Calinski-Harabasz, and Davies-Bouldin). Finally, Tavakoli et al. [51] grouped data from 3 million customers belonging to a Middle East e-commerce database, resulting in 10 clusters, 3 belonging to the recency characteristic, and the other 7 distributed between frequency and monetary characteristics. It is noteworthy that Tavakoli et al. [51] tested the model in production, setting up a campaign that focused on the active customer segment, primarily aiming to increase the company's profits, also using a control group and comparison of income before and after the campaign.

Gustriansyah, Suhandi, and Antony [49] demonstrated the possibility of applying RFM outside the conventional use of customer segmentation and acquired clusters with an average variance of 0.19113. In addition, the authors suggested other forms of data comparison, such as particle swarm optimization, medioids, or even maximizing expectancy. Peker, Kocyigit, and Eren [50] segmented customers from a market network in Turkey into "high contribution loyal customers", "low contribution loyal customers", "uncertain customers", "high spending lost customers" and "lost customers". low cost". In this way, the authors provided visions and strategies (promotions, offers, perks) to increase income on customer behavior, but limited themselves to applying it to a specific market segment.

Tavakoli et al. [51] grouped customers of an e-commerce company based on their recency, resulting in "Active", "Expiring" and "Expiring" customers, and from these segments, successively separated into groups of "High", "Medium" and "Low" values, subsequently validating the segmentation through an offer campaign for customers in the "Active" group.

Related works presented by other authors implemented the RFM model in a context of clustering by K-means, using either internal indexes or no index to assert the quality of the clusters. In addition to using internal indexes, this work applied three external index techniques, bringing a totally different approach from the others. Additionally, other authors like Łukasik et al. [52] have introduced pioneering techniques such as text-mining for assortment optimization, effectively identifying identical products in competitor portfolios, and successfully matching items with incomplete and inconsistent descriptions."

The external indices were used in this work due to the uncertainty generated by the internal validation indices in contrast to real data (variability in the results indicative of the number of suggested clusters), making it necessary to acquire other views on the data set, so that it could be possible to ensure the ideal number of clusters (meaningful, coherent, and stable).

For this, we used (i) global stability measure based on the Adjusted Rand Index (ARI), (ii) cluster stability measure based on the Jaccard index, and (iii) Segment Level Stability across method solutions (SLSa) from the entropy measure, which are the differentials of this work in relation to its correlates.

3. Prototype Description

In this section, the most relevant aspects of the developed prototype are described. The requirements specification and the metrics used to measure stability are presented.

A record has its own identification number (ID) referring to its ID number in the original database. It also records its recency, representing the number of days since the last purchase. The frequency counts the purchases made during the given period. Finally, the "monetary" information represents the total spent in R$ within the period considered.

Each of the RFM attributes was obtained from the extraction of all sales made per cash front for a given customer (trade). Recency was acquired by calculating the difference in days between the date of the last purchase and the end date of the period established to obtain the data (12/01/21).

In the case of the customer in Table 2, his recency is 139 days since his last purchase (07/15/21). The frequency was acquired by totaling the number of sales made to the customer in the given period. For the customer in Table 2, his frequency accounts for 65 purchases in the period from 01/01/2016 to 01/12/21. Finally, the monetary attribute was created from the sum of the totals of each sale, resulting in a value of R$37,176.00.

Table 2. Structure of the obtained data.

ID	Recency	Frequency	Monetary
38	139	65	37176

3.1. Data Handling

In this step, procedures were performed to remove inconsistent data such as empty sales, unsuitable transaction types such as credit sales receipts and payments, and to remove customers with no sales in the store. With these operations, 97 customers were removed, resulting in a total of 1748 customers in the base. Next, a normalization of the attributes was applied, since the K-means uses a distance measure, and the value range of the attributes varies according to their nature (monetary can present values in the thousands, while the other attributes are distributed in hundreds), which can negatively influence the results. Advanced data handling techniques have been exploited to improve the capability of artificial intelligence models [53–55].

The Min-Max method was used to normalize the attributes. According to Saranya and Manikandan [56], normalization by the Min-Max method performs a linear change in the data, which

is transformed into a new interval. As exposed in Equation 1, having a value v of an attribute A from interval $[min_A, max_A]$, it is transformed to the new interval $[n(min_A), n(max_A)]$, which in the case of this application is between 0 and 1, considering:

$$\frac{v - min_A}{max_A - min_A} \left(n(max_A) - n(min_A) \right) + n(min_A). \tag{1}$$

By applying this method, the registers shown in Table 2 have their values converted and represented in Table 3, with a maximum value of 1 and a minimum value of 0.

Table 3. Structure of the data obtained after normalization.

ID	Recency	Frequency	Monetary
38	0.0074928	0.71910112	0.43890863

Values close to 1 indicate that the attribute of the customer in question is high relative to all other customers, and values close to 0 indicate that the attribute is low relative to the others. In other words, the customer with the best frequency attribute will have 1 as its value, otherwise, it will have 0. An exception is the recency attribute, which due to the format in which it was acquired, ends up having inverse values, having a good recency if the value is close to 0 and a bad one if it is close to 1 (fewer days since the last purchase, the better). For reasons of simplicity and consistency of measures, a simple transformation of the recency values was applied, subtracting the value from 1. This way, low values become high values, and vice-versa, contributing to better analysis, since now customers with a good RFM performance tend to have all attributes close to 1.

After manipulating the data, it is possible to display each customer in a 3D graph, with each axis representing an attribute as in Figure 1. It is possible to identify that although the data does not provide a natural cluster distribution, it does present a structure of its own, with many customers clustered in the left corner of the graph indicating a low frequency, distributed over several recency intervals, with few high monetary attribute customers.
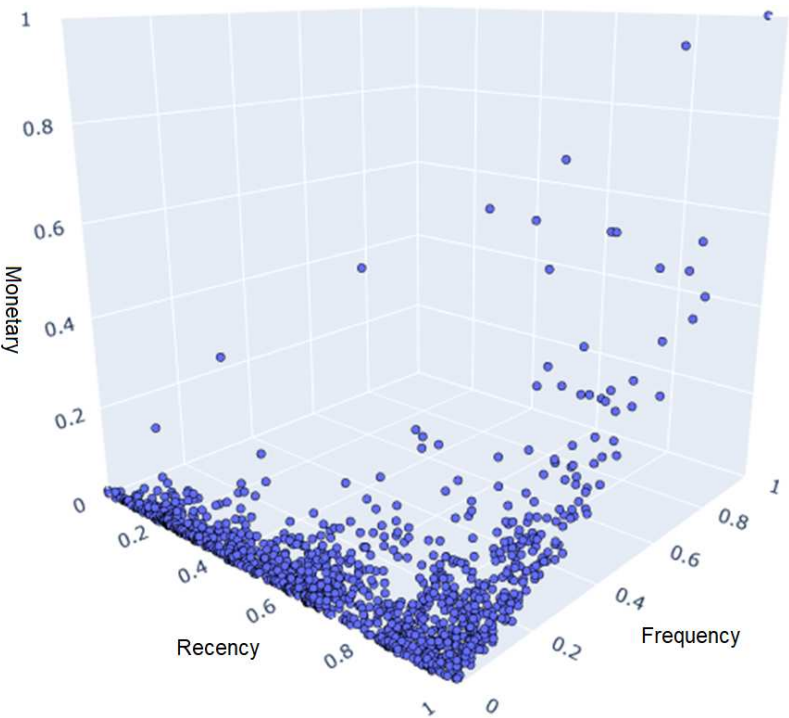

Figure 1. Representation of customers.

The segmentation and validation steps were performed in parallel since the algorithm used, K-means, requires the specification of the desired number of clusters. Internal and external validations were available to assist in the decision. Arbelaitz et al. [57] conclude in their study through statistical analysis that of the 30 internal indexes researched, 10 prove to be recommendable for use. At the top of this list are the Silhouette, Calinski-Harabasz, and Davies-Bouldin indexes.

According to Rousseeuw [58], to generate the Silhouette index for data, only two things are needed: the clusters obtained and the set of distances between all observed data, and for each i its respective Silhouette index $s(i)$ is calculated. The average dissimilarity of the distances of i with the rest of the data in the cluster of i, denoted by $a(i)$, is also calculated.

In the next step, the minimum value between the distances of i and any other cluster is obtained (the neighboring cluster of i is then discovered, i.e. the cluster with which i would most fit if it were not in its original cluster), denoted by $b(i)$. This process can be summarized by Equation 2, which results in a number between -1 and 1, where -1 is a bad categorization of the object i (not matching its current cluster) and 1 is an optimal categorization. To obtain the quality of the clustering in general, the average of $s(i)$ is obtained for all objects i in the dataset, where $s(i)$ is given by:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), \ b(i)}. \tag{2}$$

For Caliński and Harabasz [59], their variance rate criterion (VRC) index, shown in Equation 3, considers: the between group sum of squares (BGSS) that portrays the variance between clusters taking into account the distance from their centroids to the global centroid; and the within-group sum of squares (WGSS) that portrays the variance within clusters taking into account the distances from the points in a cluster to its centroid. It also considers the number of observations/data n and the number of clusters k. When this index is used, an attempt is made to maximize the result as the value of k is changed.

$$VRC = \frac{BGSS}{k-1} \bigg/ \frac{WGSS}{n-k}. \tag{3}$$

Davies and Bouldin [60] denote that the goal of their index is to define a cluster separation measure $R(S_i, S_j, M_{ij})$ that allows the computation of the average similarity of each cluster with its most similar (neighboring) cluster, the lowest possible value would be the optimal result. With S_i being the dispersion measure of cluster i, S_j being the dispersion measure of cluster j, and M_{ij} being the distance between clusters i and j, according to:

$$R_{ij} \equiv \frac{S_i + S_j}{M_{ij}} \quad \bar{R} \equiv \frac{1}{N} \sum_{i=1}^{N} R_i. \tag{4}$$

According to Davies and Bouldin [60], first R_{ij} is obtained for all clusters, that is, the ratio of inter- and intra-cluster distances between cluster i and j. After that, R_i (the highest value of R_{ij}) is obtained by identifying for each cluster, the neighboring cluster to which it is most similar. Finally, the index itself is calculated $\bar{R}$, being this, the total sum of the similarities of N clusters with their closest neighbors.

Eight segmentation solutions were generated with the K-means algorithm, starting from $k = 3$ to $k = 10$. After that, the best results among the k solutions according to each index were obtained. According to Figure 2, the Silhouette index suggested 4 clusters, while Calinski-Harabasz suggested 8 and Davies-Bouldin 3. It should be noted that in the interpretation of the Silhouette and Calinski-Harabasz index, the highest value is chosen, while in the Davies-Bouldin index the lowest value is selected.
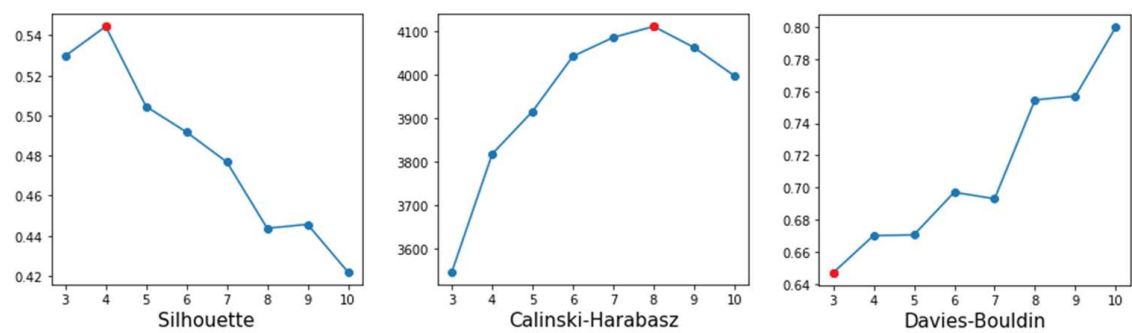
Figure 2. Internal validation indices.

The suggestions for the number of clusters from the indices showed high variability, causing great uncertainty in detecting the number of clusters. This result is common among datasets that do not have naturally occurring clusters. According to Dolnicar, Grün, and Leisch [46], consumer data typically does not contain natural segments, making it difficult to obtain the optimal number of clusters from internal validation indices.

Furthermore, various features in the data distribution can affect the internal validation indices, Liu et al. [61] point out that different densities, noise, arbitrary shapes, and very close clusters limit the results and introduce additional challenges to the process of estimating the number of clusters. More specifically, the Silhouette and Davies-Bouldin indices suffer from close clusters, and Calinski-Harabasz performs poorly on unequal-size distributions. All these cited characteristics are present when viewing the distribution of the data in Figure 3, which in addition to showing different sizes in the possible clusters, demonstrates a clustering of data on a specific side of the distribution and a low-density in areas of a high monetary attribute.

3.2. Specification

For the description of the prototype functions, the functional requirements (FR) and non-functional requirements (NFR) are shown in Table 4.
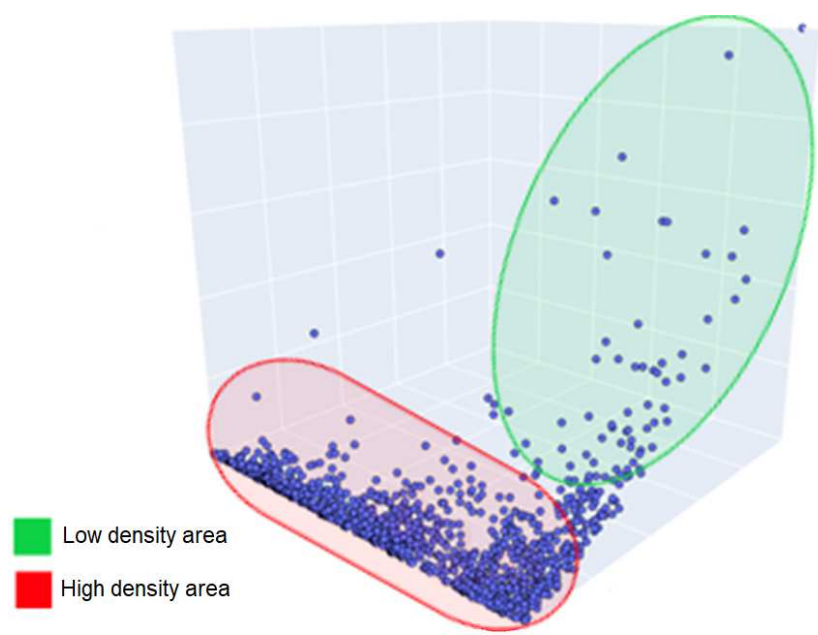

Figure 3. Densities of customer distribution.

Table 4. Requirements.

Functional Requirements	RF01	Acquire the transactions data of customers from a database.
	RF02	Filter out customers with irregular information.
	RF03	Extract the characteristics used in the RFM model from the customers.
	RF04	Normalize the data to avoid disparities in attribute scales.
	RF05	Display on a 3D graph the location of the customers from the RFM feature scores.
	RF06	Segm. into clusters the cust. based on the RFM attributes.
Non Functional Requirements	RNF01	Use the K-means clustering algorithm for segm. of clients.
	RNF02	Apply the Silhouette, Calinski-Harabasz and Davies-Bouldin internal validation indexes to val. the quality of the clusters.
	RNF03	Apply the external validation index of global stability, stability per cluster, and SLSa stability.
	RNF04	Use the Python language for prototype development.

The first step refers to obtaining the customers and pertinent information from a commercial management software database. The market segment of the database in question is focused on the sale of men's and women's clothing. A total of 1845 customers were extracted with information from the period between 01/01/2016 and 01/12/2021, consolidating data from sales tables to fit with Table 1, which represents the structure needed to perform the segmentation based on RFM attributes.

3.3. Global stability

With the uncertainty generated by the internal validation indexes, it is necessary to acquire other views on the data set, so that it is possible to ensure an optimal number of clusters with a good margin of certainty. External validation indices were applied. Since there are no "true" clusters or test data with a priori categories to make the external comparison, a global stability measure was used, defined by Ernst and Dolnicar [62] where the external information is composed of solutions with different amounts of clusters. This measure uses two main concepts: bootstrapping for random sample selection, and the adjusted rand index for the similarity measure between two solutions.

According to Roodman et al. [63], bootstrapping methods consist of generating samples that represent the original dataset and applying evaluations on such samples so that an evaluation of a sample is generally equivalent to an evaluation of the original dataset. Bootstrapping methods can also use the concept of replacement, where a sample may contain repeated data from the original set, creating a character of randomness and variance in the sample set.

According to Robert, Vasseur, and Brault [64], the rand index (RI) is a measure of similarity between two clustering solutions z and z', and can be defined by Equation 5. Where a is the number of element pairs that were assigned to the same cluster in z and z'. b is the number of element pairs that were assigned to the same cluster in solution z, but in different clusters in solution z'. c is the number of element pairs that were assigned to different clusters in solution z, but in the same clusters in solution z', and finally, d is the number of element pairs that are in different clusters in both z and z'.

Robert, Vasseur, and Vincent [64] point out that a and d can be interpreted as concordances, while b and c can be interpreted as disagreements between the evaluated solutions z and z'. The RI results in

a value between 0 and 1, with 1 being a perfect agreement of the clusters between the solutions and 0 being a total disagreement.

$$RI = \frac{a+d}{a+b+c+d}. \tag{5}$$

According to Santos and Embrechts [65], RI has some known problems, such as not always presenting a value of 0 for completely random solutions and varying positively as the number of clusters in the solutions increases. Different measures have been created to correct such problems, one of these measures is the adjusted rand index (ARI), given by:

$$ARI = \frac{Index - Expected\ Index}{Max\ Index - Expected\ Index}. \tag{6}$$

With Index being the result of RI, Expected Index being the expected RI in an occasion when observations are randomly assigned to different clusters, and Maximum Index being the maximum possible value of RI. The ARI index ranges between -1 and 1, with -1 being a value for high dissimilarity and 1 being a value for high similarity between two solutions.

From the two concepts presented, it is possible to apply the global stability measure suggested by Ernst and Dolnicar [62], divided into the following steps:

- a) create 50 pairs of bootstrap samples with replacement from the data;
- b) perform the clustering of each pair of samples with k clusters;
- c) calculate the ARI value of the clustered pair, generating a value from -1 to 1;
- d) repeat steps "b" and "c" until the desired number k is reached.

After finishing this algorithm, we have 50 ARI values for each k analyzed. It is then possible to represent the values in a boxplot chart as shown in Figure 4, where the horizontal axis represents the solutions with different numbers of clusters and the vertical axis represents the ARI index value, the box shapes represent 50% of the values and the outer dashes represent the other 50%.

Outlier values (values outside the standard distribution) are represented by circles outside the outer part and the orange dash indicates the average of the values. With this graph, a concrete view of the stability of each solution with k clusters is obtained. The ARI value tends to decrease as k is increased, indicating a greater variation in the possible differences between the clusters of each solution, that is, the higher the value of k, the different runs of K-means will result in totally different clustering solutions.
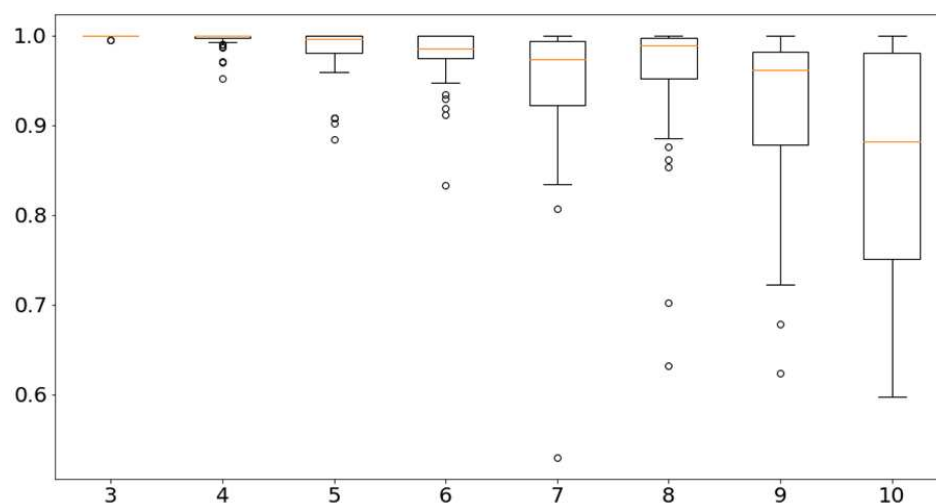

Figure 4. Boxplot of ARI for each number of clusters.

After analyzing the boxplot, it is evident that after 6 clusters the ARI value between solutions constantly varies negatively. Therefore, solutions with 4, 5 and 6 clusters become viable, since they

have desirable stability in relation to solutions with k larger numbers, and still, allow a more detailed analysis of each cluster. The option with $k = 3$ was not considered because it has few clusters, aggregating different customers in the same group, making the solution more generalized and with few discernible details in each cluster.

3.4. Stability by Cluster

Global stability allows analyzing the solutions with respect to their change according to several executions of a clustering algorithm but does not allow a detailed analysis of the specific structure of the solutions, i.e., the clusters. After selecting three segmentation candidates ($k = 4, k = 5$ and $k = 6$), it is possible to calculate the cluster stability described by Hennig [66], which is similar to the previous method, but with a focus on clusters instead of entire solutions.

This stability allows the detection of unstable clusters within stable solutions and vice-versa, helping later in descriptive analyses and selection of the solutions themselves, since it provides a view by cluster, facilitating the choice of a potential customer segment. The method uses bootstrapping and Jaccard's index to calculate stability.

According to Lee et al. [67], the Jaccard index (J) measures the similarity between two data sets A and B, considering the union and intersection of these sets, as expressed in Equation 7. The upper part represents the intersection of A with B, thus containing values common to both sets.

The lower part represents the union of A with B, containing all the values of A and all the values of B, then subtracting the values common to both sets to avoid their duplication. The Jaccard index returns a value between 0 and 1, with 1 being a value that represents the similarity between the two sets, and 0 representing the total dissimilarity between the sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{7}$$

In short, the method described by Hennig [66] performs a bootstrap sampling of the original set, comparing through Jaccard's index each cluster belonging to the original solution with its bootstrap representation, generating an index for each cluster. The algorithm can be described with the following steps:

- (a) create 100 bootstrap samples with replacement from the original solution data;
- b) perform the sample clustering;
- c) extract the common data between the original solution and the sample (remembering that the sample may contain repeated data or not contain some data from the original set);
- d) for each cluster of the original set, calculate the maximum Jaccard's index between it and each cluster of the bootstrap sample
- e) repeat step "b" for the remaining samples.

Running the algorithm results in 100 values in a range from 0 to 1 for each cluster, which can then be displayed in a boxplot. The horizontal axis of each graph represents the different clusters contained in a solution, while the vertical axis represents the value of Jaccard's index, allowing you to intuitively visualize the stability of each cluster within a solution.

As there are three candidates for the solution ($k = 4, k = 5$ and $k = 6$), the algorithm was applied to each one resulting in Figure 5, where it is possible to compare the solutions with respect to the stability of their clusters. It can be observed in the solutions with $k = 4$ and $k = 5$ that the last cluster has great instability, reaching Jaccard values close to 0.4. In the solution with $k = 6$, the stability of the last cluster varies with less intensity. The solution still presents an instability in the first cluster, indicating a possible division of a large and unstable cluster into two smaller and more stable clusters.
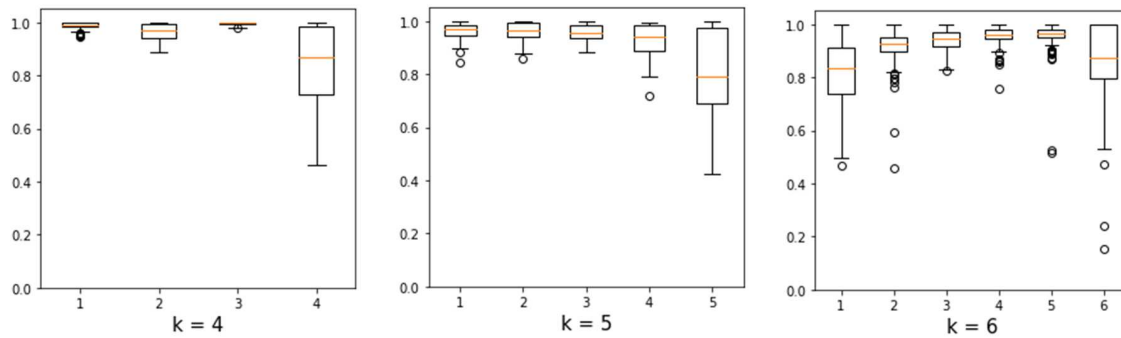
Figure 5. Jaccard boxplot for each cluster of each solution.

3.5. SLSa Stability

Another method for analyzing possible solutions with respect to the number of clusters is the SLSa, presented by Dolnicar and Leisch [68], which evaluates the cluster-level stability over several solutions and allows identifying changes in cluster structures such as joins and splits, providing information about the history of a cluster regarding its composition. This method applies the concept of relabeling and uses the entropy measure formulated by Shannon [69].

Dolnicar and Leisch [68] denote that to effectively implement the SLSa algorithm, the application of relabeling is required, which refers to the consistent naming of clusters across possible solutions. More specifically, it is the act of identifying identical clusters belonging to different solutions and assigning the same name to them so that their tracking becomes possible. For the dataset used in this work, although the focus is on candidate solutions with $k = 4$, $k = 5$, and $k = 6$, it was chosen to apply relabeling on solutions with $k = 2$ through $k = 9$ for a better understanding of the cluster formation process.

According to Dolnicar and Leisch [68], the entropy measure represents the uncertainty in a probability distribution $(p_1, p_2, p_3, ... p_k)$. It is described by Equation 8, where p_j is the probability distribution in question. The maximum entropy value consists of a probability distribution where all values are equal, resulting in an entropy value $H = 1$. The minimum entropy value consists of a probability distribution where only one of the values is 1, $[0, 0, 1, 0, 0]$ for example, resulting in an entropy value $H = 0$ and, in the context of the algorithm, signaling that all the data in one cluster in a solution is the same as all the data in another cluster in a previous solution.

$$H = -\sum_{i=1}^{n} p_j \log p_j. \tag{8}$$

To apply the SLSa calculation of Dolnicar and Leisch [68], it is necessary to calculate the entropy measure H of each cluster S_l^i (cluster l belonging to the solution i) with respect to all clusters of the previous solution S_j^{i-1} (clusters $j = 1, ... k_{i-1}$ belonging to the previous solution $i - 1$). Therefore, the SLSa value of a segment l belonging to a solution with k_i segments is defined by Equation 9, where a minimum value of 0 represents the worst possible stability, while 1 indicates the best possible stability. In short, a cluster with $SLSa = 1$ is equivalent to a cluster that was not formed from other clusters, but rather persisted throughout the k-segment solutions, while a cluster with $SLSa = 0$ was created from two or more clusters in the previous $k - 1$ solution.

$$SLSa(S_l^i) = -1\frac{H}{\log{(k_{i-1})}}. \tag{9}$$

Based on the results of SLSa the standard K-means [70] is compared to the K-medoids [71] and MiniBatch K-means [72]. Then the most appropriate clustering model is standardized for the following analyses.

4. Results and Discussion

This section will present the results and discussions concerning the application of the proposed method, considering the initial calculations to define the stability criteria. After calculating the SLSa for each cluster of each solution up to $k = 9$, it is possible to represent the values in a graph (see Figure 6), starting from a solution with two clusters in the left corner and ending with a solution with nine clusters in the right corner. Clusters with low SLSa values are colored with a shade of red according to their instability.

The black lines represent the total number of clients belonging to one cluster that is assigned to another cluster in the next solution, thick lines indicate a larger amount, and many lines to the left of a cluster indicate that it was generated from several others.

Cluster number 3 in the three-cluster solution has a high level of instability since it was created from the data in clusters 1 and 2 in the previous solution (effectively representing half of each cluster in the previous solution). Other clusters follow the same behavior, more specifically, the clusters created from a new solution (the last clusters in each column) are most often the product of joining parts of other clusters.
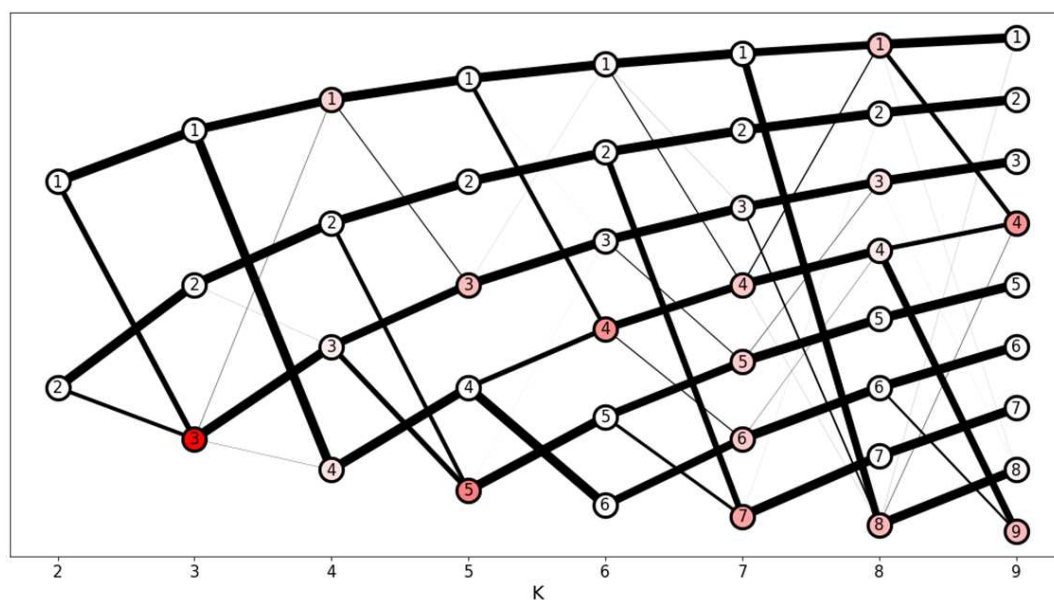

Figure 6. SLSa solutions of K-means with k=2 to k=9.

After solution 6, almost all the clusters in the following solutions present some amount of instability, being formed from two or more clusters in previous solutions with a few exceptions. Of the candidate solutions (4, 5, and 6) only solution 6 presents a satisfactory distribution of stable clusters, with five clusters having only one parent in the previous solution.

For better visualization, Figure 7 exposes the transition of clusters along the different candidate solutions. Cluster 5 (in orange) was created in solution 5 from data coming from clusters 2 (in red) and 3 (in green). Similarly, cluster 6 (in cyan) in solution 6 was created from half of the data from cluster 4 (in purple), which consequently was shifted towards cluster 1 (in purple), resulting in the apparent "junction" between two halves of clusters.
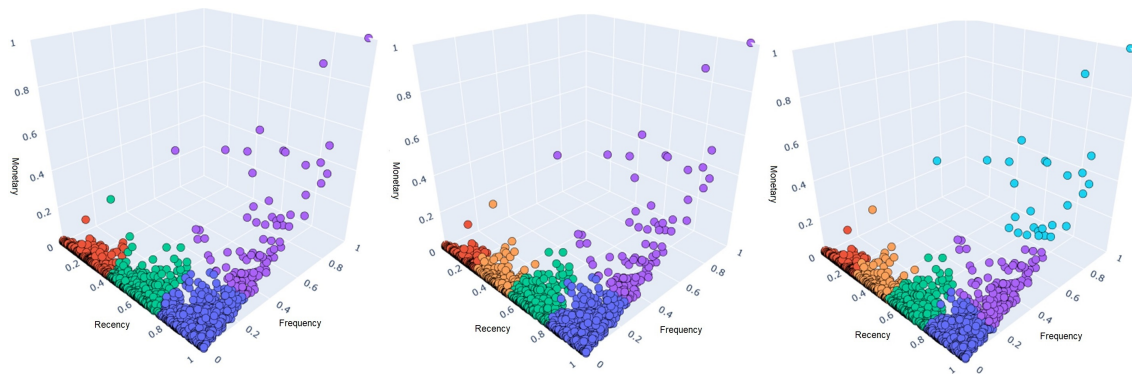
Figure 7. Graph of the clusters of solutions 4, 5, and 6.

Customers in cluster 6 (in cyan) are completely absorbed into another cluster in the smaller solutions, despite having unique characteristics such as having all three RFM attributes high compared to the rest of the clusters. Therefore, the solution with 6 clusters was chosen because it satisfactorily represents all customer types present in the dataset, as well as having an acceptable overall stability (above 0.95) and a tolerable instability per cluster (only one cluster formed from shifts).

4.1. Comparison to other algorithms

When applying K-medoids and MiniBatch K-means to the same database, there were variations from the standard K-means, which will be explained here. With internal indices, there was a recommendation of 4 clusters by the Silhouette index, 8 or 9 clusters by the Calinski-Harabasz index, and 3 clusters by the Davies-Bouldin index. In this case, the internal validation indexes did not provide an equivalent value in the number of clusters. Therefore, unlike the result of standard K-means, K-medoids, and MiniBatch K-means performed lower overall stability starting at 4 clusters. For standard K-means, stability remained high until solutions with 6 clusters.

The lower stability in other algorithms occurs because they suffer more from the repeated iterations and initializations required by stability methods. For example, K-medoids take as centroids the very points present in the data set and may suffer multiple divergences over too many runs of the algorithm, because as the data set presents many points, the initialization, and subsequent execution may vary.

In the case of MiniBatch, the algorithm randomly obtains a subset of the data to perform cluster assignment, further increasing the variability between solutions, and contributing to lower overall stability. With the overall stability reduced, the stability per cluster follows this trend, showing more variation in most solutions. Considering K-medoids and MiniBatch K-means, there were few clusters that remain with high stability across all solutions.

The SLSa results presented in Figure 8 show the stability drop by demonstrating the history of each cluster in each solution. Note that in comparison with the result referring to K-means (Figure 6), the two algorithms presented many more "Splits" and "Joins" among the members of each solution, contributing to a larger number of clusters with an inadequate entropy level, while K-means exhibits this behavior only after the amount of 6 clusters.
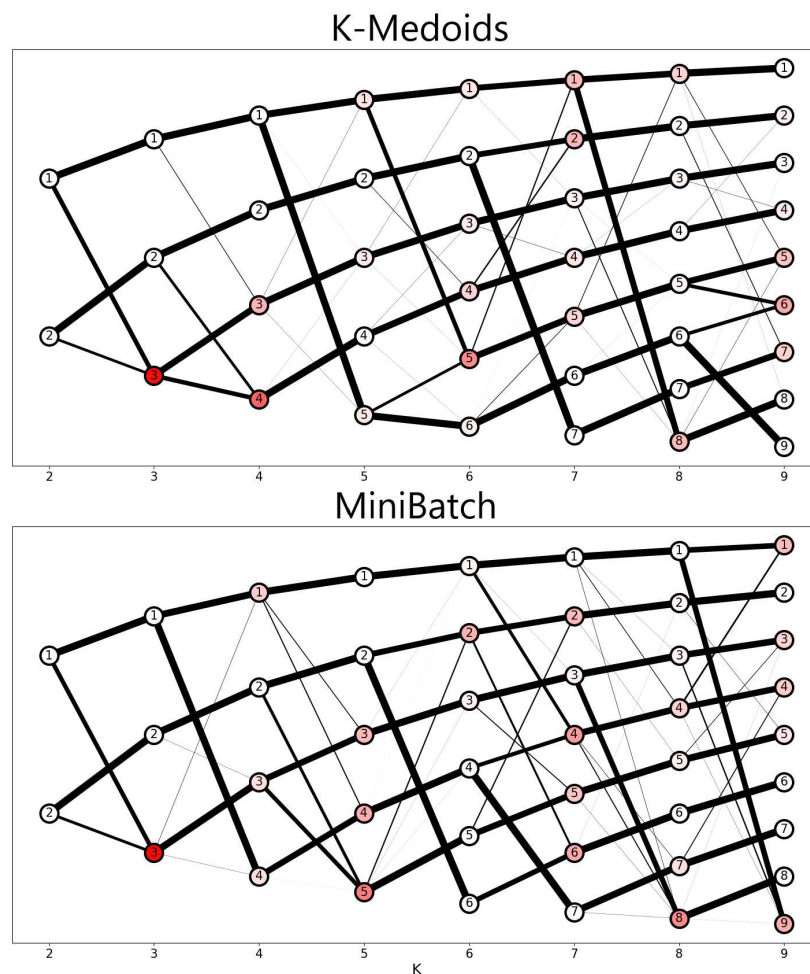
Figure 8. SLSa results of K-medoids and MiniBatch K-means.

Based on this comparison, it is shown that the external validation indexes prove to be not only a good tool for choosing the number of clusters but also a good tool for comparing different clustering algorithms. Overall, K-means demonstrated higher reliability compared to K-medoids and MiniBatch K-means, showing higher stabilities throughout the analysis process.

4.2. Cluster Profile

Once the desired solution is obtained, it is necessary to analyze the clusters contained therein, so that their profile is easily understood, and which characteristics are really relevant. Witschel, Loo, and Riesen [73] state that before benefiting from the results, an analyst needs to understand the essence of each cluster, that is, what are the characteristics shared among the customers of a cluster that differentiate them from others.

A bar chart was created that presents the average of each RFM characteristic of each cluster contained in the $k = 6$ solution, presented in Figure 9. Each bar represents an RFM attribute, and its height is defined by the average of the attribute in question in the cluster. In this representation, each attribute has a black dot referring to the average of the entire solution, allowing you to compare whether the attribute of the cluster stands out in relation to all the others.
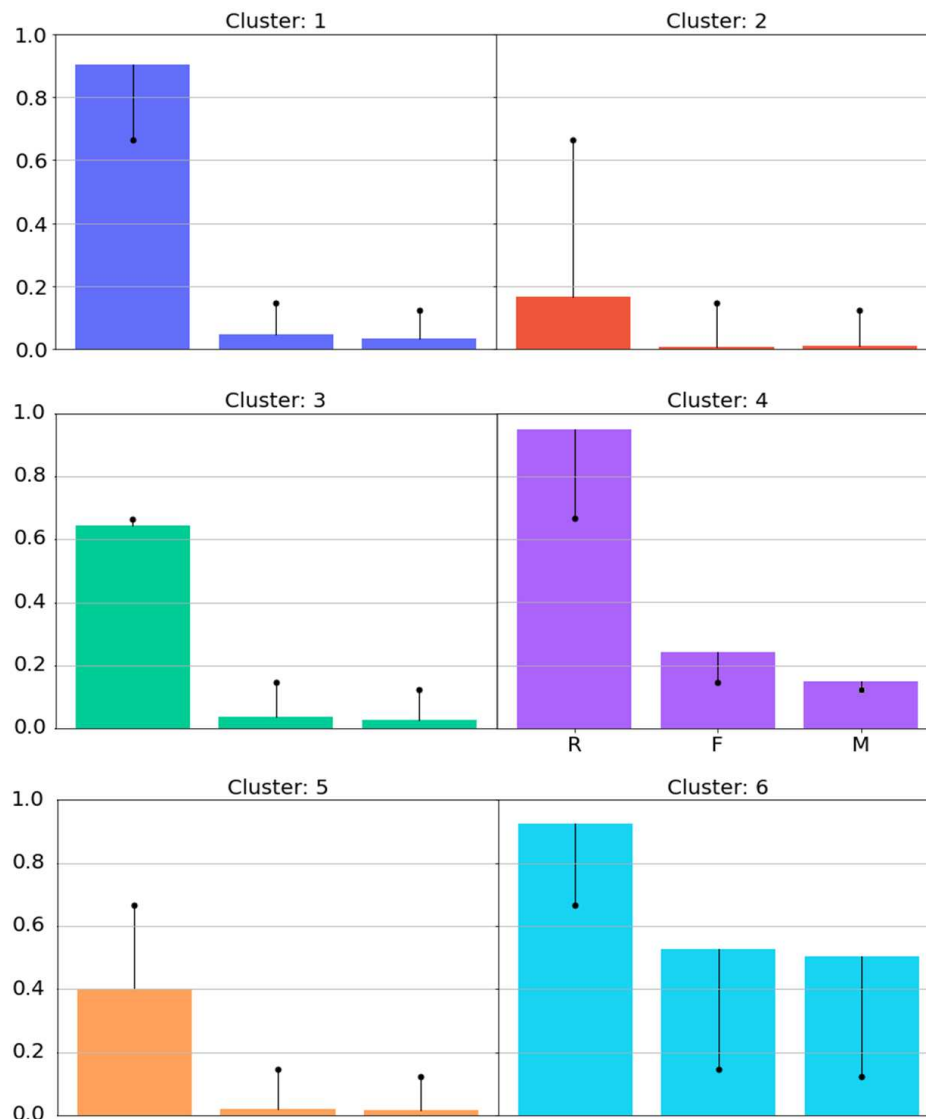
Figure 9. Profile plot of the clusters of the k=6 solution.

By analyzing Figure 9 considering each RFM attribute, it is possible to have the following interpretations:

- (a) clusters 2 and 5 have a recency, frequency, and monetary attribute below the overall average, possibly indicating a type of customer who no longer frequents the store (cluster 2) or is in the process of stopping frequenting (cluster 5). Clusters 2 and 5 have 390 and 338 customers respectively, representing about 41% of all registered customers;
- b) Clusters 1 and 3 have high recency, but low frequency and monetary, indicating a new type of customer who is not yet familiar with the store, or is in the process of developing a frequent visiting relationship, or even an old customer who frequented the store recently. Either way, these clusters may represent the flow of customers who have recently purchased from the store. Clusters 1 and 3 have 474 and 379 customers respectively, representing about 48% of all registered customers;
- c) clusters 4 and 5 have above-average RFM attributes, indicating loyal customers who buy frequently and spend high total money relative to others. Cluster 6 has the highest values among all clusters, representing the store's best customers. Its RFM attributes are expressively higher, yet this cluster contains only 28 customers. Cluster 4 also has fewer customers than the other

clusters, with 139 in total. The two clusters together represent a total of 167 customers, about 11% of all registered customers.

With the information generated by the cluster profiles, it is possible to obtain a succinct summary of the types of customers who frequent the company, these being: lost customers (with low recency, frequency, and monetary), customers in the process of being lost (with below average recency, low frequency and monetary), recent customers (with high recency but the low frequency and monetary), less recent customers (with high recency but lower than recent customers, and a lower frequency and monetary than recent customers), loyal customers (high recency, frequency and monetary) and finally the best customers (best possible RFM attributes).

4.3. Cluster Description

From the analysis of segmentation variables, it becomes feasible to implement promotional campaigns, incentive actions, and even methods to rescue lost customers. However, the analysis is not necessarily finished, according to Dolnicar, Grün, and Leisch [46], one of the important steps after obtaining the cluster profiles is the description process. Cluster description consists of the individual analysis of the clusters from variables external to the clustering process, called descriptive variables. These variables can contain information such as: age, gender, location, buying pattern, information from questionnaires, and other characteristics pertinent to the scope of the company.

As the database has several eligible information, five were chosen for the cluster description process: age, sex, time of contact with the store, number of purchases per season, and rate of returns. After extracting the descriptive data, mosaic charts were used for display. This type of chart is similar to the bar chart but displays the information in cells that have their size relative to the amount of information observed and may vary in width according to the number of customers/purchases in a cluster, and in height according to the percentage of the variable observed compared to the percentage of other variables.

Another concept pertinent to the mosaic chart is the statistical model applied called bimodal distribution, which displays abnormal variations in the distribution of values based on an assumption of independence of variables. In this way, higher-than-expected values (above two standard deviations, or outside the 95% value limit) are displayed in red shades of greater intensity, lower-than-expected values are displayed in blue shades of greater intensity, and normal values take on green. With this view, it is possible to observe unique characteristics of clusters that have abnormal variations.

Regarding the descriptive information used, the age variable was transformed into an ordinal variable. This variable starts from 18 to 24 years old, considering age intervals of six years onwards for each category, with the penultimate one being for customers over 40 years old and the last one for a category representing a lack of information in the register. The gender variable available in the database consists of the categories "male" and "female".

The result of the graphs applied to these variables can be seen in Figure 10, which presents the age graph on the left side and the sex graph on the right side, each graph displays on the vertical axis the categories of the descriptive variables analyzed and on the horizontal axis the clusters. Since the distribution of the cells occurs according to the observed variable and the number of observations in the cluster, the size of each varies in width and height. Taking cluster 6 (C6) as an example, its width is thin due to the low number of customers it has, and the height of each cell belonging to it depends on the percentage that each category represents in relation to the other categories in the same cluster, if a category has 99% of customers, it will occupy 100% of the cell, as in cluster 6 (C6) in the graph on the right side.
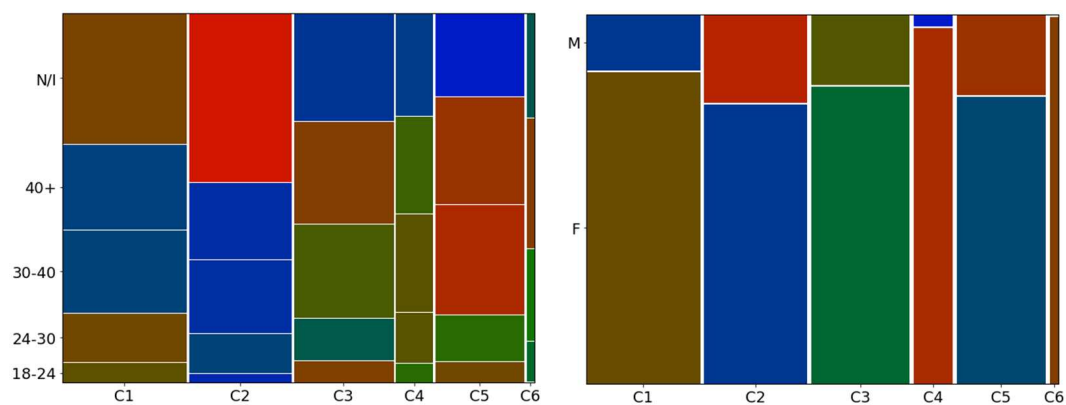
Figure 10. Mosaic chart of ages (left) and sex (right).

By analyzing Figure 10, it is possible to have the following interpretations:

- a) Regarding age (left graph), the cluster of recent customers (C1) has a lower-than-expected number (cells in blue) of adult and elderly customers and higher concentrations of young adults and customers with no information, indicating that there may be a flow of young people being attracted by the store. The cluster of lost customers (C2) has a higher-than-expected number of customers who did not inform their age, indicating a certain resistance to filling out registrations. The cluster of customers being lost (C5) has a higher than expected amount of customers over 30 years old, indicating a possible dissatisfaction with the products offered to this age group, information that is corroborated by the fact that the flow of recent customers (C1) has more young people than expected.
- b) Regarding gender (right chart), the most important customers (belonging to clusters C4 and C6) are mostly women and are in larger numbers than expected, even though the store offers male lines, indicating a female preference for the clothes offered. This information is corroborated by the fact that the clusters with customers lost or in the process of being lost (C2 and C5) have a larger number of men than women, indicating a possible lack of male engagement with the options offered.

The two other variables that allow for a mosaic display are: the number of years since a customer's registration with the company and the number of purchases made during each season. For the first, the interval was established as less than a year (<1), one, two, three, and more than four years. For the second the interval is composed of the four seasons (summer, autumn, winter, and spring). The graphs generated are shown in Figure 11, which follows the same structure as the previous figure.
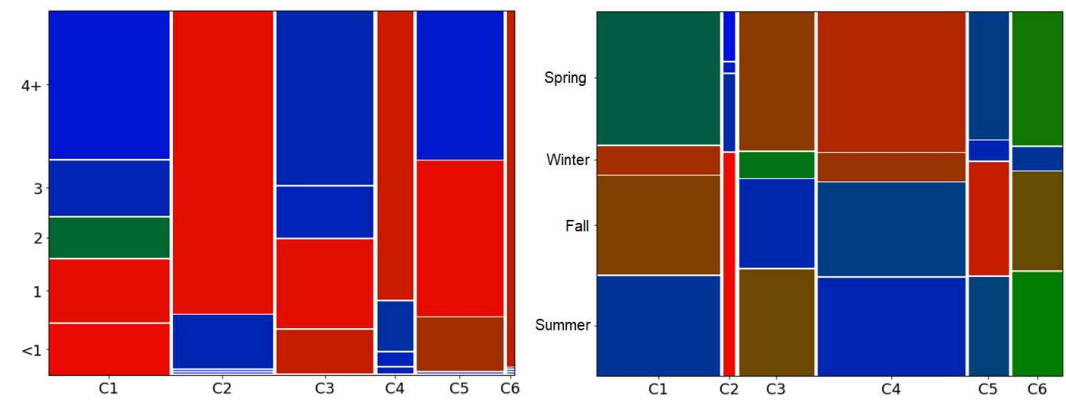

Figure 11. Mosaic chart of registration time (left) and total purchases per season (right).

From the analysis of Figure 11, it is possible to make the following interpretations:

- a) In relation to the time of customers' registration (graph on the left side), it is possible to identify that the clusters with recent customers (C1 and C3) have a higher amount of newly registered customers than normal, as well as customers with one year of registration, allowing to identify that these clusters present a flow of new customers. The clusters with the best customers (C4 and C6) have many customers registered for more than four years (in cluster C6 it is all customers), indicating that customers with good RFM performance are rarely new customers, requiring a long relationship with the store. Finally, clusters C2 and C5, which represent customers lost or in the process of being lost, present many customers registered for more than four years, which justifies the characteristic of lost customers.
- b) Regarding the number of purchases per season (right chart), the chart shows the preferences of each cluster in relation to specific seasons, showing a general preference for the winter, summer, and fall collections. The cluster of lost customers (C2) shows a high rate of purchases made during the summer, possibly indicating a certain dissatisfaction with this season's line, since customers in this cluster no longer frequent the store. The cluster with the second best RFM performance (C4), presents the highest number of purchases of all the other clusters (denoted by the width of the cells), of these sales, higher than normal was the frequency of purchases in spring, indicating a preference of this group for the line of this season.

The last variable analyzed, purchase returns rate (transaction or sale that contains at least one return), was obtained through the ratio between the number of returns in a cluster and its total sales quantity. Thus, Table 5 displays the percentages of returns for each cluster.

Table 5. Rate of returns per cluster.

Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5	Clust. 6
9.09%	6.93%	8.11%	11.81%	8.68%	17.50%

Based on the percentages presented, the clusters with the best RFM performance (clusters 4 and 6) have the highest return rates (11.81% and 17.50% respectively), indicating a high selectivity among their customers. The cluster of lost customers (cluster 2) has the lowest return rate (6.93%), indicating that a dissatisfied customer rarely makes a return, and simply does not frequent the store anymore instead of exchanging the product and trying to buy again.

5. Conclusion

Customer segmentation allows an in-depth analysis of a company's customer behavior. With the right data, previously obscure profiles can be identified, based on information sometimes considered useless beyond the operational layer of a company's sales and registration. This work had as its initiative the numbering and identification of these profiles, for which the database of a real retail clothing company was used, containing registration and transaction information from 1845 customers. Each customer was assigned characteristics based on the RFM model, and then the data was cleaned and manipulated to fit the clustering algorithm used, K-means.

To validate the cluster solution as well as its quantity, three internal validation indexes were used, and when they were not conclusive enough to define the quantity, the following external validation indexes were used: global stability measure based on the ARI index, stability measure per cluster based on the Jaccard index, and the SLSa method from the entropy measure. After selecting three candidate solutions (with 4, 5, and 6 clusters) based on the global stability, the stability per cluster presented a better result in the solution with 6 clusters, being then confirmed and detailed from the SLSa method, demonstrating the process of dividing, and joining clusters throughout the iterations with different numbers for the k parameter of the K-means algorithm.

Thus, the solution with 6 clusters was chosen, and its clusters were presented in a chart containing their RFM characteristics so that their profiles could be detected based on inferences made from their attributes. With the profiling of the clusters, six segments were named based on their peculiarities:

lost customers (with low recency, frequency, and monetary), customers in the process of being lost (with below average recency, low frequency, and monetary), recent customers (with high recency, but the low frequency and monetary), less recent customers (with high recency, but lower than recent customers, and a lower frequency and monetary than recent customers), loyal customers (high recency, frequency and monetary) and finally the best customers (best possible RFM attributes).

After highlighting the profile of each segment through the RFM segmentation variables, an analysis was performed from descriptive variables based on the data available in the database. The segments were evaluated through mosaic graphs and tables based on their age, gender, registration time, purchases per season, and returns, pointed out particularities present in each descriptive variable, such as possible trends of the segments, abnormal flows, and non-standard amounts, among others.

The objective of identifying different customer segments based on their behavior was achieved. Although the internal validation indexes do not present a consensus among the number of natural clusters, it was possible to obtain a guarantee of the stability of the segments through the external indexes. That said, it is clear that despite the absence of natural clusters, it was still possible to obtain significant segments, containing distinguishable characteristics that differentiate them from each other, allowing further insights into the types of customers who frequent the establishment, extrapolating to customer types in general in the retail industry.

Furthermore, this work contributes to the academic community, by applying models (RFM), indexes (three internal and three external), methods (Min-Max normalization, bootstrapping, Jaccard Index, and ARI), and K-means algorithm, in a real database, analyzing its influence on data with a different distribution of training data (whose characteristics commonly present well-defined clusters, unlike a database with real data). A conclusion derived from applying such techniques to this dataset is that internal validation indices do not always present a consensus on the number of clusters requiring the use of other types of validation. In addition, it has been shown that valuable information for the apparel retail industry and possibly other industries can be extracted from a database of transactional and registration information, indicating the intrinsic value of data that is often only stored and rarely analyzed in the context of customer clusters.

Given the above, the present work can be complemented by the following proposals: use of the RFM method in conjunction with K-means applied to a database of a different retail branch, such as supermarkets, dealerships, real estate agents, among others; application of different internal and external indexes for the validation of the quality of clusters under different visions; use of other descriptive variables, such as time spent per purchase, lines of products most purchased and quantity of products per purchase; application of questionnaires, to use in conjunction with the analysis of the profiles, crossing the variables based on the questioned cluster.

References

1. Rolim, C.O.; Rossetto, A.G.; Leithardt, V.R.Q.; Borges, G.A.; Geyer, C.F.R.; dos Santos, T.F.M.; Souza, A.M. Situation awareness and computational intelligence in opportunistic networks to support the data transmission of urban sensing applications. *Computer Networks* **2016**, *111*, 55–70. https://doi.org/10.1016/j.comnet.2016.07.014.

2. Mahmood, F.; Khan, A.Z.; Bokhari, R.H. ERP issues and challenges: A research synthesis. *Kybernetes* **2020**, *49*, 629–659. https://doi.org/10.1108/K-12-2018-0699.

3. de Oliveira, J.R.; Stefenon, S.F.; Klaar, A.C.R.; Yamaguchi, C.K.; da Silva, M.P.; Salvador Bizotto, B.L.; Silva Ogoshi, R.C.; Gequelin, E.d.F. Planejamento de recursos empresariais e gerenciamento de relacionamento com o cliente através da gestão da cadeia de fornecimento. *Interciencia* **2018**, *43*, 784–791.

4. dos Santos, R.P.; Fachada, N.; Beko, M.; Leithardt, V.R.Q. A Rapid Review on the Use of Free and Open Source Technologies and Software Applied to Precision Agriculture Practices. *Journal of Sensor and Actuator Networks* **2023**, *12*, 28. https://doi.org/10.3390/jsan12020028.

5. Srivastava, S.K.; Chandra, B.; Srivastava, P. The impact of knowledge management and data mining on CRM in the service industry. In *Nanoelectronics, circuits and communication systems*; Springer, 2019; pp. 37–52. https://doi.org/10.1016/0377-0427(87)90125-7.

6. Souza, B.J.; Stefenon, S.F.; Singh, G.; Freire, R.Z. Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV. *International Journal of Electrical Power & Energy Systems* **2023**, *148*, 108982. https://doi.org/10.1016/j.ijepes.2023.108982.

7. Oyelade, J.; Isewon, I.; Oladipupo, F.; Aromolaran, O.; Uwoghiren, E.; Ameh, F.; Achas, M.; Adebiyi, E. Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology insights* **2016**, *10*, BBI–S38316. https://doi.org/10.4137/bbi.s38316.

8. Kowalski, P.A.; Jeczmionek, E. Parallel complete gradient clustering algorithm and its properties. *Information Sciences* **2022**, *600*, 155–169. https://doi.org/10.1016/j.ins.2022.03.087.

9. Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hussien, A.G.; Khasawneh, A.M.; Alshinwan, M.; Houssein, E.H. Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis. *Algorithms* **2020**, *13*, 12. https://doi.org/10.3390/a13120345.

10. Lai, D.T.C.; Sato, Y. An Empirical Study of Cluster-Based MOEA/D Bare Bones PSO for Data Clustering. *Algorithms* **2021**, *14*, 11. https://doi.org/10.3390/a14110338.

11. Valdez, F.; Castillo, O.; Melin, P. Bio-Inspired Algorithms and Its Applications for Optimization in Fuzzy Clustering. *Algorithms* **2021**, *14*, 4. https://doi.org/10.3390/a14040122.

12. Trzciński, M.; Kowalski, P.A.; Łukasik, S. Clustering with Nature-Inspired Algorithm Based on Territorial Behavior of Predatory Animals. *Algorithms* **2022**, *15*, 43. https://doi.org/10.3390/a15020043.

13. Kowalski, P.A.; Łukasik, S.; Charytanowicz, M.; Kulczycki, P., Nature Inspired Clustering – Use Cases of Krill Herd Algorithm and Flower Pollination Algorithm. In *Interactions Between Computational Intelligence and Mathematics Part 2*; Kóczy, L.T.; Medina-Moreno, J.; Ramírez-Poussa, E., Eds.; Springer International Publishing: Cham, 2019; pp. 83–98. https://doi.org/10.1007/978-3-030-01632-6_6.

14. Hämäläinen, J.; Jauhiainen, S.; Kärkkäinen, T. Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms* **2017**, *10*, 105. https://doi.org/10.3390/a10030105.

15. Hajibaba, H.; Grün, B.; Dolnicar, S. Improving the stability of market segmentation analysis. *International Journal of Contemporary Hospitality Management* **2020**, *32*, 1393–1411. https://doi.org/10.1108/IJCHM-02-2019-0137.

16. Reinartz, W.; Thomas, J.S.; Kumar, V. Balancing acquisition and retention resources to maximize customer profitability. *Journal of marketing* **2005**, *69*, 63–79. https://doi.org/10.1509/jmkg.69.1.63.55511.

17. Sopelsa Neto, N.F.; Stefenon, S.F.; Meyer, L.H.; Ovejero, R.G.; Leithardt, V.R.Q. Fault Prediction Based on Leakage Current in Contaminated Insulators Using Enhanced Time Series Forecasting Models. *Sensors* **2022**, *22*, 6121. https://doi.org/10.3390/s22166121.

18. Stefenon, S.F.; Freire, R.Z.; Coelho, L.S.; Meyer, L.H.; Grebogi, R.B.; Buratto, W.G.; Nied, A. Electrical Insulator Fault Forecasting Based on a Wavelet Neuro-Fuzzy System. *Energies* **2020**, *13*, 484. https://doi.org/10.3390/en13020484.

19. Safa, M.; Sari, P.A.; Shariati, M.; Suhatril, M.; Trung, N.T.; Wakil, K.; Khorami, M. Development of neuro-fuzzy and neuro-bee predictive models for prediction of the safety factor of eco-protection slopes. *Physica A: Statistical Mechanics and its Applications* **2020**, *550*, 124046. https://doi.org/10.1016/j.physa.2020.124046.

20. Stefenon, S.F.; Kasburg, C.; Freire, R.Z.; Silva Ferreira, F.C.; Bertol, D.W.; Nied, A. Photovoltaic power forecasting using wavelet Neuro-Fuzzy for active solar trackers. *Journal of Intelligent & Fuzzy Systems* **2021**, *40*, 1083–1096. https://doi.org/10.3233/JIFS-201279.

21. Branco, N.W.; Cavalca, M.S.M.; Stefenon, S.F.; Leithardt, V.R.Q. Wavelet LSTM for Fault Forecasting in Electrical Power Grids. *Sensors* **2022**, *22*, 8323. https://doi.org/10.3390/s22218323.

22. Klaar, A.C.R.; Stefenon, S.F.; Seman, L.O.; Mariani, V.C.; Coelho, L.d.S. Structure optimization of ensemble learning methods and seasonal decomposition approaches to energy price forecasting in Latin America: A case study about Mexico. *Energies* **2023**, *16*, 3184. https://doi.org/10.3390/en16073184.

23. Stefenon, S.F.; Ribeiro, M.H.D.M.; Nied, A.; Mariani, V.C.; Coelho, L.S.; Leithardt, V.R.Q.; Silva, L.A.; Seman, L.O. Hybrid Wavelet Stacking Ensemble Model for Insulators Contamination Forecasting. *IEEE Access* **2021**, *9*, 66387–66397. https://doi.org/10.1109/ACCESS.2021.3076410.

24. Ribeiro, M.H.D.M.; da Silva, R.G.; Ribeiro, G.T.; Mariani, V.C.; dos S. Coelho, L. Cooperative ensemble learning model improves electric short-term load forecasting. *Chaos, Solitons & Fractals* **2023**, *166*, 112982. https://doi.org/10.1016/j.chaos.2022.112982.

25. Stefenon, S.F.; Bruns, R.; Sartori, A.; Meyer, L.H.; Ovejero, R.G.; Leithardt, V.R.Q. Analysis of the Ultrasonic Signal in Polymeric Contaminated Insulators Through Ensemble Learning Methods. *IEEE Access* **2022**, *10*, 33980–33991. https://doi.org/10.1109/ACCESS.2022.3161506.

26. Stefenon, S.F.; Yow, K.C.; Nied, A.; Meyer, L.H. Classification of distribution power grid structures using inception v3 deep neural network. *Electrical Engineering* **2022**, *104*, 4557–4569. https://doi.org/10.1007/s00202-022-01641-1.

27. Fernandes, F.; Stefenon, S.F.; Seman, L.O.; Nied, A.; Ferreira, F.C.S.; Subtil, M.C.M.; Klaar, A.C.R.; Leithardt, V.R.Q. Long short-term memory stacking model to predict the number of cases and deaths caused by COVID-19. *Journal of Intelligent & Fuzzy Systems* **2022**, *6*, 6221–6234. https://doi.org/10.3233/JIFS-212788.

28. Singh, G.; Stefenon, S.F.; Yow, K.C. Interpretable visual transmission lines inspections using pseudo-prototypical part network. *Machine Vision and Applications* **2023**, *34*, 41. https://doi.org/10.1007/s00138-023-01390-6.

29. Vieira, J.C.; Sartori, A.; Stefenon, S.F.; Perez, F.L.; de Jesus, G.S.; Leithardt, V.R.Q. Low-Cost CNN for Automatic Violence Recognition on Embedded System. *IEEE Access* **2022**, *10*, 25190–25202. https://doi.org/10.1109/ACCESS.2022.3155123.

30. Stefenon, S.F.; Kasburg, C.; Nied, A.; Klaar, A.C.R.; Ferreira, F.C.S.; Branco, N.W. Hybrid deep learning for power generation forecasting in active solar trackers. *IET Generation, Transmission & Distribution* **2020**, *14*, 5667–5674. https://doi.org/10.1049/iet-gtd.2020.0814.

31. dos Santos, G.H.; Seman, L.O.; Bezerra, E.A.; Leithardt, V.R.Q.; Mendes, A.S.; Stefenon, S.F. Static Attitude Determination Using Convolutional Neural Networks. *Sensors* **2021**, *21*, 6419. https://doi.org/10.3390/s21196419.

32. Stefenon, S.F.; Freire, R.Z.; Meyer, L.H.; Corso, M.P.; Sartori, A.; Nied, A.; Klaar, A.C.R.; Yow, K.C. Fault detection in insulators based on ultrasonic signal processing using a hybrid deep learning technique. *IET Science, Measurement & Technology* **2020**, *14*, 953–961. https://doi.org/10.1049/iet-smt.2020.0083.

33. Morais, R.; Crocker, P.; Leithardt, V. Nero: A Deterministic Leaderless Consensus Algorithm for DAG-Based Cryptocurrencies. *Algorithms* **2023**, *16*, 38. https://doi.org/10.3390/a16010038.

34. Wang, E.K.; Chen, C.M.; Hassan, M.M.; Almogren, A. A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain. *Future Generation Computer Systems* **2020**, *108*, 135–144. https://doi.org/10.1016/j.future.2020.02.054.

35. Leithardt, V.; Santos, D.; Silva, L.; Viel, F.; Zeferino, C.; Silva, J. A Solution for Dynamic Management of User Profiles in IoT Environments. *IEEE Latin America Transactions* **2020**, *18*, 1193–1199. https://doi.org/10.1109/TLA.2020.9099759.

36. Stefenon, S.F.; Corso, M.P.; Nied, A.; Perez, F.L.; Yow, K.C.; Gonzalez, G.V.; Leithardt, V.R.Q. Classification of insulators using neural network based on computer vision. *IET Generation, Transmission & Distribution* **2021**, *16*, 1096–1107. https://doi.org/10.1049/gtd2.12353.

37. Corso, M.P.; Perez, F.L.; Stefenon, S.F.; Yow, K.C.; Ovejero, R.G.; Leithardt, V.R.Q. Classification of Contaminated Insulators Using k-Nearest Neighbors Based on Computer Vision. *Computers* **2021**, *10*, 112. https://doi.org/10.3390/computers10090112.

38. Sopelsa Neto, N.F.; Stefenon, S.F.; Meyer, L.H.; Bruns, R.; Nied, A.; Seman, L.O.; Gonzalez, G.V.; Leithardt, V.R.Q.; Yow, K.C. A Study of Multilayer Perceptron Networks Applied to Classification of Ceramic Insulators Using Ultrasound. *Applied Sciences* **2021**, *11*, 1592. https://doi.org/10.3390/app11041592.

39. Stefenon, S.F.; Seman, L.O.; Schutel Furtado Neto, C.; Nied, A.; Seganfredo, D.M.; Garcia da Luz, F.; Sabino, P.H.; Torreblanca González, J.; Quietinho Leithardt, V.R. Electric Field Evaluation Using the Finite Element Method and Proxy Models for the Design of Stator Slots in a Permanent Magnet Synchronous Motor. *Electronics* **2020**, *9*, 1975. https://doi.org/10.3390/electronics9111975.

40. Medeiros, A.; Sartori, A.; Stefenon, S.F.; Meyer, L.H.; Nied, A. Comparison of artificial intelligence techniques to failure prediction in contaminated insulators based on leakage current. *Journal of Intelligent & Fuzzy Systems* **2022**, *42*, 3285–3298. https://doi.org/10.3233/JIFS-211126.

41. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing* **2020**, *90*, 106181. https://doi.org/10.1016/j.asoc.2020.106181.

42. Kasburg, C.; Stefenon, S.F. Deep Learning for Photovoltaic Generation Forecast in Active Solar Trackers. *IEEE Latin America Transactions* **2019**, *17*, 2013–2019. https://doi.org/10.1109/TLA.2019.9011546.

43. Nguyen, T.H.; Sherif, J.S.; Newby, M. Strategies for successful CRM implementation. *Information management & computer security* **2007**. https://doi.org/10.1108/09685220710748001.

44. Ziafat, H.; Shakeri, M. Using data mining techniques in customer segmentation. *Journal of Engineering Research and Applications* **2014**, *4*, 70–79. https://doi.org/10.9790/9622-04917079.

45. Roberts, J.H.; Kayande, U.; Stremersch, S. From academic research to marketing practice: Exploring the marketing science value chain. In *How to Get Published in the Best Marketing Journals*; Edward Elgar Publishing, 2019. https://doi.org/10.1016/j.ijresmar.2013.07.006.

46. Dolnicar, S.; Grün, B.; Leisch, F. *Market segmentation analysis: Understanding it, doing it, and making it useful*; Springer Nature, 2018.

47. Kumar, V. *Managing customers for profit: Strategies to increase profits and build loyalty*; Prentice Hall Professional, 2008.

48. Tsiptsis, K.K.; Chorianopoulos, A. *Data mining techniques in CRM: inside customer segmentation*; John Wiley & Sons, 2011.

49. Gustriansyah, R.; Suhandi, N.; Antony, F. Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science* **2020**, *18*, 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477.

50. Peker, S.; Kocyigit, A.; Eren, P.E. LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning* **2017**, *35*, 544–559. https://doi.org/10.1108/mip-11-2016-0210.

51. Tavakoli, M.; Molavi, M.; Masoumi, V.; Mobini, M.; Etemad, S.; Rahmani, R. Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques: a case study. In Proceedings of the 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE). IEEE, 2018, pp. 119–126. https://doi.org/10.1109/ICEBE.2018.00027.

52. Łukasik, S.; Michałowski, A.; Kowalski, P.A.; Gandomi, A.H. Text-Based Product Matching with Incomplete and Inconsistent Items Descriptions. In Proceedings of the Computational Science – ICCS 2021; Springer International Publishing: Cham, 2021; pp. 92–103.

53. Borré, A.; Seman, L.O.; Camponogara, E.; Stefenon, S.F.; Mariani, V.C.; Coelho, L.d.S. Machine Fault Detection Using a Hybrid CNN-LSTM Attention-Based Model. *Sensors* **2023**, *23*, 4512. https://doi.org/10.3390/s23094512.

54. Stefenon, S.F.; Seman, L.O.; Aquino, L.S.; dos Santos Coelho, L. Wavelet-Seq2Seq-LSTM with attention for time series forecasting of level of dams in hydroelectric power plants. *Energy* **2023**, *274*, 127350. https://doi.org/10.1016/j.energy.2023.127350.

55. Klaar, A.C.R.; Stefenon, S.F.; Seman, L.O.; Mariani, V.C.; Coelho, L.d.S. Optimized EWT-Seq2Seq-LSTM with attention mechanism to insulators fault prediction. *Sensors* **2023**, *23*, 3202. https://doi.org/10.3390/s23063202.

56. Saranya, C.; Manikandan, G. A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology (IJET)* **2013**, *5*, 2701–2704.

57. Arbelaitz, O.; Gurrutxaga, I.; Muguerza, J.; Pérez, J.M.; Perona, I. An extensive comparative study of cluster validity indices. *Pattern recognition* **2013**, *46*, 243–256. https://doi.org/10.1016/j.patcog.2012.07.021.

58. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **1987**, *20*, 53–65.

59. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* **1974**, *3*, 1–27. https://doi.org/10.1080/03610927408827101.

60. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1979**, *PAMI-1*, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

61. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of internal clustering validation measures. In Proceedings of the 2010 IEEE international conference on data mining. IEEE, 2010, pp. 911–916. https://doi.org/10.1109/ICDM.2010.35.

62. Ernst, D.; Dolnicar, S. How to avoid random market segmentation solutions. *Journal of Travel Research* **2018**, *57*, 69–82. https://doi.org/10.1177/0047287516684978.

63. Roodman, D.; Nielsen, M.O.; MacKinnon, J.G.; Webb, M.D. Fast and wild: Bootstrap inference in Stata using boottest. *The Stata Journal* **2019**, *19*, 4–60. https://doi.org/10.1177/1536867X19830877.

64. Robert, V.; Vasseur, Y.; Brault, V. Comparing high-dimensional partitions with the Co-clustering Adjusted Rand Index. *Journal of Classification* **2021**, *38*, 158–186. https://doi.org/10.1007/s00357-020-09379-w.

65. Santos, J.M.; Embrechts, M. On the use of the adjusted rand index as a metric for evaluating supervised classification. In Proceedings of the International conference on artificial neural networks. Springer, 2009, pp. 175–184. https://doi.org/10.1007/978-3-642-04277-5_18.

66. Hennig, C. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis* **2007**, *52*, 258–271. https://doi.org/10.1016/j.csda.2006.11.025.

67. Lee, S.; Jung, W.; Kim, S.; Kim, E.T. Android malware similarity clustering using method based opcode sequence and jaccard index. In Proceedings of the 2019 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2019, pp. 178–183. https://doi.org/10.1109/ICTC46691.2019.8949745.

68. Dolnicar, S.; Leisch, F. Using segment level stability to select target segments in data-driven market segmentation studies. *Marketing Letters* **2017**, *28*, 423–436. https://doi.org/10.1007/s11002-017-9423-8.

69. Shannon, C.E. A mathematical theory of communication. *The Bell system technical journal* **1948**, *27*, 379–423. https://doi.org/10.1145/584091.584093.

70. Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* **2020**, *9*, 1295. https://doi.org/10.3390/electronics9081295.

71. Yu, D.; Liu, G.; Guo, M.; Liu, X. An improved K-medoids algorithm based on step increasing and optimizing medoids. *Expert Systems with Applications* **2018**, *92*, 464–473. https://doi.org/10.1016/j.eswa.2017.09.052.

72. Xiao, B.; Wang, Z.; Liu, Q.; Liu, X. SMK-means: an improved mini batch k-means algorithm based on mapreduce with big data. *Computers, Materials & Continua* **2018**, *56*. https://doi.org/10.3970/cmc.2018.01830.

73. Witschel, H.F.; Loo, S.; Riesen, K. How to support customer segmentation with useful cluster descriptions. In Proceedings of the Industrial Conference on Data Mining. Springer, 2015, pp. 17–31. https://doi.org/10.1007/978-3-319-20910-4_2.