Customer Segmentation in Food Retail Sector: An Approach from Customer Behavior and Product Association Rules

Chapter · January 2023

DOI: 10.1007/978-3-031-24985-3_18

2 authors, including:

Juan Carlos Llivisaca
University of Cuenca
15 PUBLICATIONS **32** CITATIONS

SEE PROFILE

Customer segmentation in food retail sector: an approach from customer behavior and product association rules

Juan Llivisaca[1,2][0003-2154-3277] and JonnatanAvilés-González[3][0001-9868-8488]

[1] Universidad Politécnica Estatal del Carchi, Dirección de Posgrado, Av. Universitaria y Antisana, Tulcán, Ecuador.
[2] Department of Applied Chemistry and Systems of Production, Faculty of Chemical Sciences, University of Cuenca, 010107 Cuenca, Ecuador.
[3] Universidad del Azuay, Facultad de CCTT, Centro de Investigación en Ingeniería de Producción y Operaciones, Av. 24 de mayo 7-77 y Hernán Malo. Apartado 01.01.981, Cuenca, Ecuador
lncs@springer.com

Abstract. In competitive markets, customer segmentation improves customer loyalty and business performance, but in practice, these analyses are carried out using simple relationships in dashboard, or Microsoft Excel' sheets, which do not show customer behavior. Data segmentation in the era of big data has changed this paradigm with some techniques that try to decrease bias. In this research, four segmentation techniques are tested with a large set of data from a retail store. CLARA (Clustering Large Applications Algorithm) and Random Forest algorithms both were the best. Through the RFM (Recency, Frequency, Monetary) approach, eight customer segments were found, where *Champions* customers spend more money and return frequently to the retail store. In addition, each segment of customer buys following a model, this was demonstrated with the a priori algorithm. Finally, some strategies are given into which products should go together and how to distribute them so that customers can find them, as well as the best-selling products.

Keywords: Data mining, Clustering algorithm, Random Forest, Retail, A priori.

1 Introduction

The retail sector is dedicated to the retail sale of many products such as food, household appliances, jewelry, etc. The main advantage it has is the flexibility to adapt to customer changes; so, any company that is dedicated to this business should know that managing a good diversification of products, customer portfolio, and a good store along with an adequate level of inventories can lead to increase market and competitive. Retail customers seek these companies for their variety of products and for the ease of purchasing a product. Therefore, having a physical store is one of the critical factors involved in the customer's purchase decision, since it provides the customer

with a comfortable place to make their purchase, which ultimately ensures that the customer will return, and thus increase sales [1].

The store design considers which customers will be served and in what quantity. The number of customers entering a store is used to segment customers. For this purpose, retail stores have adopted the use of barcode scanners to obtain information about the product purchased, which has allowed the retailer to use the information to make decisions on procurement, shelf space and decision in advertising campaigns. The data obtained from each payment is used to define the consumer's buying pattern and thus organize the products according to the consumer's preference [1]. However, this technique is limited, because it only categorizes customers using information such as age, gender, or education level. But information about how often the customer buys, what kind of products the customer buys, or how much money the customer spends is not considered if no customer segmentation is performed. Customer segmentation tries to find clusters that have similarities and differences between groups. Generally, this segmentation is based on historical data, and on observing customers' buying behaviors, to evaluate them and discover high-value customers [2]. In competitive markets, customer segmentation improves customer loyalty and business level, but in practice, these long-lasting relationship analyses are carried out by applying profitable customer databases and without a holistic analysis [3]. Therefore, segmenting customers to find potential customers plays a significant role in a retailer's expansion and competitiveness.

In customer segmentation research, some clustering algorithms have been determined such as decision trees, fuzzy c-means, genetic algorithms, particle swarm K-means, affinity propagation algorithms, spectral clustering algorithms, and Gaussian mixtures, K-means, Mahalanobis distances, among others [2, 4–6]. Each of these proposals can be applied considering the number of clusters, bandwidth, scalability, and geometric distance. It is known that fuzzy c-means algorithms, and genetic algorithms, are more used in environments where the size of groups is equal, while affinity propagation algorithms, spectral clustering algorithms, Gaussian mixtures, are used more in non-planar distance geometry and with groups that have connectivity constraints. Despite this variety of algorithms, the two most used algorithms are K-means algorithm and decision trees. These two algorithms are applied in many situations and with a few clusters these work quite well. Khalili-Damghani [3] proposed a decision tree that follows a flowchart approach, where each internal node in the tree contains a question about a particular feature and each branch shows the result of the experiment. In this study, this algorithm gave good classification results due to the classification rules that were placed, and it was possible to classify a group of data based on economic indicators. While in the study of Espinosa and Zúñiga [7], it can be noted that the decision trees facilitated the classification of a group of data giving a performance of 91.38% which shows the robustness of this algorithm. On the other hand, the K-means algorithm is very popular because of its ease of use. This technique selects the initial cluster centroids using sampling based on an empirical probability distribution. The distance to the centroids is based on euclidean distance and associates the members of each cluster considering this variable. Anitha and Patil [4], suggest a customer segmentation considering K-means, with this they found a method to

identify profitable and high-value customers. While Saraf et al [6], using K-means managed to obtain four clusters of different customers that help to know the buying patterns of customers as well as how they bill each cluster so that strategies can be applied to serve these customers and positively lengthen the customer life cycle.

On the other hand, in customer segmentation, it is common to take an approach with variables such as *Recency, Frequency, and Monetary*, which are: the customer's most recent purchase (R), frequency of purchases (F), and money spent (M). RFM can be defined as the segmentation of customer analysis that not only provides information about the customer's frequent purchase pattern but also about the recent purchase and the profit earned [8]. A high measure of RFM indicates the presence of a high-value customer and conversely with a low measure. Parikh and Abdelfattah [2], in their research, performed data segmentation and found five defined clusters. Cluster 5 was identified as the highest spender but had a low frequency. While cluster 3 spent the least money, they are the ones who made the most recent purchases. From this work, it can be noted that the customers who spend the most are not the ones who buy the most or the most recent purchases, in addition to which customers the business should concentrate on in order to formulate different strategies. In [5], their study did an RFM analysis of customers, and determined that only eight transactions are enough to classify a customer this is interesting because in a retail business there are many customers but few transactions. However, in [9] not only transactions were considered for their RFM, but they proposed one more metric to the RFM, and that was to measure the trend of customers dynamically.

Many retailers distribute their products intuitively, without any study to validate their decisions. So, in data mining, one of the main tasks is to find association rules and discover useful patterns in large volumes of data [10]. These associations are not evident with a simple descriptive analysis of the data, and usually, different association algorithms must be used to find these patterns. The *A priori* algorithm proposed in 1993 has been the most widely used to find these associations, the power of an association rule being measured by support and confidence. Moreover, for an association rule to be considered interesting, these two parameters must be together [11]. Support measures the number of transactions in which the items are present according to the association rule, i.e., whether the items are together in the data relative to the total number of transactions. Confidence represents the actual probability that a set of items coexists with another set of items in a data set, it measures the accuracy of the association rule used. This measure indicates the percentage of transactions containing the ascending term and the consequent term in relation to the number of transactions containing the antecedent part [10].

From the previous paragraphs, retail store optimization and customer categorization in retail stores becomes an important research topic, because there is a big problem in identifying high-value customers while having an adequate distribution of products in the stores. Something that small retail stores do not have. Therefore, the following research question is posed: *Could customer categorization be based on customers' consumption behavior, considering which products they purchase?*

This paper is organized as follows: section II presents the methodology of work, and describes the data set used in this research, section III describes the results to

apply four clustering algorithms and rules associations, and section IV is a discussion of results and comparison of others research, finally, section V presents the conclusion and future work.

2 Methodology

2.1 Exploratory Analysis and Data Preprocessing

The first step was to review scientific articles, which resulted in the recognition of different studies with models and methods oriented to segment customers, and associate products in cases of retail stores selling food products. Meanwhile, the data obtained are from a small retail store in Cuenca - Ecuador. The database used was part of a research on retail trade in the city of Cuenca. Data set includes all transactions carried out in the retail trade corresponding to September, October, November, and December 2020, it was 60,596. The database has nine features, these are, Issue Date, Consumer ID, Transaction Number, Product Code, Quantity Purchased, Unit Price, Product Name, Category, and Unit of Measurement. The first six features are numeric and the last three are string features. As for the customers, there were 1,001 customers at the beginning of the database. Exploratory Analysis and Data Preprocessing (EAD) consisted of cleaning the data in the database, and statistical graphs were used to find anomalies and thus use the appropriate statistical techniques. In addition, a coding and text mining of the variables was performed in order to have a uniform description of the products, the string features must be converted to numeric features. Likewise, all the product codes were revised so that they have values of product price or quantity purchased. Meanwhile, the products purchased by customers were placed in four categories. Category 1 included products such as beef, chicken, pork, seafood, and sausages. Category 2 includes products such as fruits, vegetables, legumes, and eggs. Category 3 includes dairy products, bakery products, confectionery, canned goods, rice, sugar, salt, oils, alcoholic and non-alcoholic beverages, condiments, and sauces. In Category 4, there is any type of product that is not in the previous categories. It is important to note that the brands of the products were not relevant at the time of the study.

2.2 Segmentation of customers

The RFM approach was used to segment customers, and different classification algorithms were used to evaluate the performance of each in the case study. The RFM values (refers to Recency, Frequency, Monetary value), complaints registered, their purchases, products, and date since the last store visit, contribute to highly accurate segmentation [6]. The calculation of the RFM value corresponds to the following equation [12]:

$$RFM\ score = (rs \times rw) + (fs \times fw) + (ms \times mw) \quad (1)$$

where rs = recency score and rw = recency weight, fs = frequency score and fw = frequency weight, ms = monetary score and mw = monetary weight.

In this section, eight categories have been placed for customers which are: *"Champions", "Loyal Customers", "Potential Loyalist", "New Customers", "Promising", "Need Attention", "About to Sleep", "At Risk", "Can't Lose Them", and "Lost"*. These data categories were placed according to descending RFM scores.

Because of the literature review, it was decided to use four classification algorithms. The first uses the centroid method (K-means) to know the effect of variability in the data, the second uses the K-medoids clustering method (CLARA) which is more robust, and the third is a hierarchical algorithm that allows knowing the effect of not assigning a cluster number previously and finally Random Forest which is a machine learning algorithm for large volumes of data. K-means algorithm Clusters the data by attempting to separate the samples into n groups of equivalent variances. Unlike some of the other clustering algorithms, K-means requires that the number of k clusters is provided. In order to define the number of clusters [6] describe in their publication that the Elbow method, which is the most popular heuristic. In this case, the Elbow Method was taken, which for any unsupervised learning algorithm, determining the number of clusters is elementary. The Elbow method is the most popular method to determine the optimal value of k, and the silhouette method was used. The objective of the K-means is to segment the existing data into K- clusters, such that the total Euclidean distance between the cluster centroid and the respective data points is minimized. Here is a mathematical equation to represent the same:

$$\min_{B_j C_j} \sum_{i=L}^{k} \sum_{x_i \in B} \left\| x_i - B_j \right\|_{L2}^{2} \qquad (2)$$

Where: B_j is K clusters, y C_j centroids. L2 is euclidean distance.

When there is a large amount of data or the presence of outliers there are sophisticated algorithms that handle this problem, for example, K-medoids with the use of Partitioning Around Medoids (PAM), but the drawback of this algorithm is its high run time cost [13]. To overcome this, the Clustering Large Applications Algorithm (CLARA) is presented, which combines the idea of K-medoids with resampling in order to work with large volumes of data more efficiently [14]. The algorithm CLARA repeatedly applies PAM on a subsample with $n' \ll n$ objects. Afterwards the remaining objects are assigned to their closest medoid. The CLARA method considered a sample of 60 data in order to perform the clustering and the distance to the medoids was performed considering the Manhattan method. As with the previous algorithm, Silhouette width and Dunn were used to internally validate the Clusters generated.

Hierarchical clustering is a general family of clustering algorithms that create nested clusters by successively merging or splitting them [6]. This algorithm does affinity clustering, and the clusters can be of different sizes. The algorithm looks for hierarchies in the data and generates the Clusters based on this, this is represented using dendrograms. In the hierarchical algorithm, the agglomerative method was used.

While the selected distance, the available methods were analyzed, such as complete distance, average distance, single distance, Ward.D2, euclidean distance, Mcquitty, etc. To choose the appropriate method, the correlation index was used and the highest one was selected. Since this will better represent the distance of the data to the centroids. In addition, an internal validation of clusters generated with the use of the algorithm was performed, for this, the Silhouette width, and Dunn were used [15].

The Random Forest algorithm is a supervised learning technique whose purpose is to generate a decision on a set of data, using for this purpose a segment of these data (training) usually chosen by bootstrapping [7]. The main advantages of the Random Forest algorithm according to Cánovas-García [16] are: that it can be used for classification, for this each tree chooses a class and the result of the model is the class with the highest number of "votes" in all trees. Besides, this is an easier training model, in order to compare against complex techniques, but with similar performance. Finally, it maintains its accuracy with large missing data. For the Random Forest algorithm, 70% of the data was taken for training the model, and 30% for testing. As for classification, we tested whether the eight categories of the selected RFM clients were able to be reached using this algorithm. The algorithm requires the number of trees and the number of random variables to be found as candidates in each branch and the maximum number of nodes selected from the hyperparameters analyzed.

2.3 Association rules

Association rules are widely used in machine learning. In this step of the methodology, the products that are most purchased by each of the identified customer segments were identified. Once this was done, with the *A priori* algorithm, the respective association rules were obtained. This algorithm is included in the rules association rules package of the R statistical software. For data analysis, it was necessary to create a matrix with the transactions of each customer segment where each row represents a transaction and each column an item. Then, diagrams were created with the five best-selling products for each segment. The confidence level and support value for each rule were adjusted to 0.01 and 0.6 respectively so that the results obtained were the products that are frequently sold together, and the rules are as reliable as possible. It was considered necessary to obtain, in addition to the indicators that the algorithm yields by default, the indicators of leverage and conviction, which help to determine the direction of the rule and to choose the optimal product mix. Also in rules, a process of verification was developed to know whether there were duplicate and redundant rules and remove them from the study. The rule that has more products always has lower support and is considered redundant since it does not provide additional information. Likewise, two rules are identical when they have the same antecedent and consequent [1].

3 Results

3.1 Exploratory Analysis and Data Preprocessing

The results described below were developed with the R program (version 4.2.1). In EAD, from the database, we had 60,596 transactions corresponding to 1,001 different customers. Each customer has an identification code and those customers who purchase products were considered. Customers who did not have complete information were eliminated; in the end, only 853 customers were analyzed. From the verification of the quantities of the variables price and quantity paid, three data that have a value of zero were eliminated. The analysis of outliers, in this case, was not performed, because we want to know how these data affect customer segmentation, on the other hand, in the case study there may be customers who buy very high quantities of products and do so frequently. However, all outliers were identified.

3.2 RFM analysis

In the RFM analysis, the starting point is the refined dataset of EAD. To start the RFM analysis, the final analysis date was set as December 31, 2020, since at that date the company performs a quarterly evaluation of its annual business strategy. A preprocessing of the RFM was performed using the R program, with the "rfm" library and "ggplot2" for the graphs (Fig.1). To understand the behavior of customers, 101 reference days were counted (09/21/2020 to 12/31/2020). Customers as active are those who in the mentioned period have visited the retailer at least four times in the quarter, sleeping is those who have visited the retailer between 4 to 2 times in the quarter and inactive customers are those who have visited the retailer less than 2 times in the quarter. In Figure 1, it can be seen that the majority are inactive customers, followed by sleeping customers, and finally active customers. This is common in small retailers. If we analyze the frequency of transactions, only 4,094 of the visits are made by registered customers. While the total number of transactions analyzed is 60,596, this implies that it is only capturing about 6.76% of the transactions made by returning customers. It can be noted that many customers are inactive, but to look at this segment in more detail, in the RFM analysis, eight categories were proposed which represent more detail of the retail situation. The data was distributed as follows: *Champions* (9.50%), *Loyal customer* (17.94%), *Potential Loyal* (18.41%), while *Lost* (8.32%), *At Risk* (1.29%), *About to sleep* (5.04%), *Need Attention* (0.82%) and *Others* (38.69%) have a low percentage, which is good for the retail business. In terms of average monetary spending, the *Champions* segment is the one that spends the most (USD 43 average per transaction). On the other hand, the customer with ID 102000000000, is the one who has visited the retail business the most, while the customer with ID 301231767 is the one who spends the most in the retail business (Table1).

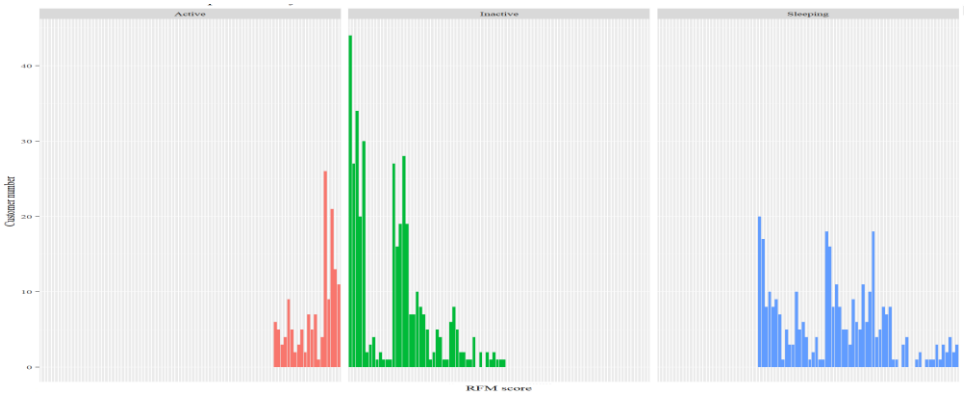


Fig. 1. Pre-processing of transactions.

Table 1. Top 5 customer in RFM.

Customer ID	Recency	Frequency	Monetary
102000000000	21.0	114	18.60
301231767	24.0	21.0	41.3
102914041	22.0	20.0	34.1
103547758	26.0	14.0	37.2
103584082	27.0	12.0	46.1

3.3 Algorithms of segmentations

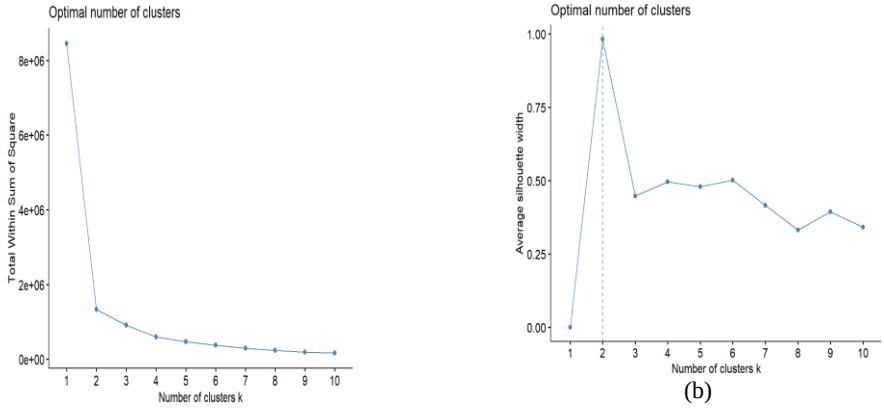
In this section, four segmentation algorithms were tested. The first algorithm was K-means, which took as input data the results of the RFM. From the data, a statistical normality test was run for the Recency, Monetary, and Frequency variables. It is desired to know if these variables have any correlation, so it was necessary to calculate a correlation coefficient, therefore, a normality test was run on the variables. The results are shown in Table 2, and it is found that the variables do not follow a normal distribution. With these data, a Spearman's test can be performed to see if there is a correlation between the variables. The relationship between Recency and Frequency is the highest and most significant (-0.57), between Recency and Monetary it does not exist (-0.017) as well as between Frequency and Monetary (0.039). This indicates that K-means clustering can be performed since the groups formed are representatives.

Table 2. Normality test.

Variable	Kolmogorov-Smirnov	P-value (*)
Recency	0.13612	2.2 e -16
Frequency	0.47005	2.2 e -16
Monetary	0.22957	2.2 e -16

(*) α=0.01

K-means is an iterative clustering algorithm that initially randomly assigns data points to clusters, unlike hierarchical clustering. The number of clusters is always a problem to determine in this algorithm. But as suggested by many authors, Elbow Method can be used for any unsupervised learning algorithm, therefore determining the number of clusters is elementary, likewise, the silhouette method which analyzes how well the resulting clusters are separated can be used [6]. From Figure 2, it can be seen that the optimal number of Clusters was 2, while in Table 3, the results are summarized, and based on these it is decided to take 3 clusters.



(a) (b)

Fig. 2. a) Elbow method, b) Silhouette method.

Table 3. Resultados con K=2 y K=3

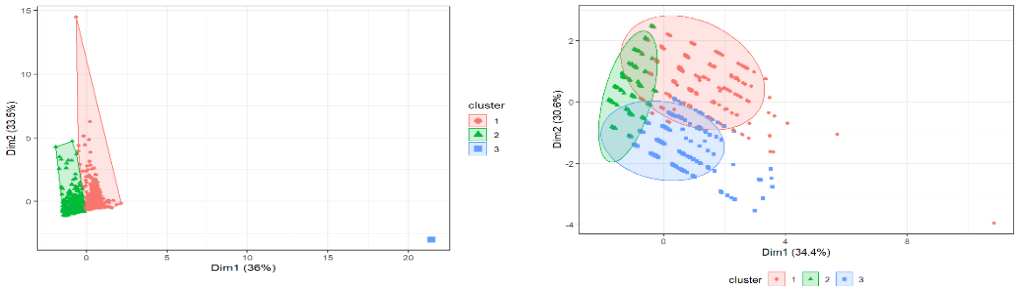
Cluster	Recency	Frequency	Monetary
1	51.73	4.804	20.81
2	21	2675	12.34
1	21	2675	12.34
2	79.86	1.503	20.21
3	33.41	6.953	21.2

Note: average values

As can be seen with only two clusters, the segmentation is not good since the second cluster collects the information with more clients. Therefore, these values are improved by considering the three clusters, since the clustering is improved, and more information is obtained. In the clustering based on K-means, it was performed with the "stats" package of R. The result is shown in Figure 3, where the three clusters formed can be seen. However, it is noticeable that due to the presence of data with very high frequencies (outliers), there are few clients in cluster 3. Moreover, cluster 1 has good recency, but a low frequency and high monetary. In this group are the category of customers *Champions, Potential Loyal, and Loyalty*. In cluster 2, it can be seen that the Recency is low, frequency is high, as well as the monetary is low. In this group, there are customers *Need Attention and At risk*, in the last cluster 3, it is noticed that there are low values of recency and frequency, but high monetary, these are

customers who have been few times to the business but have spent a high monetary value, the category of *Lost*, *About to sleep and Others* are in this group.

The second algorithm used to improve clustering was the CLARA algorithm. This algorithm is a non-hierarchical algorithm that handles a lot of data, and its resolution is based on the K-medoids proposal. In addition, to improve the clustering, the Manhattan distance method is used, as this helps to avoid overlapping the groups. The number of clusters K=3 is placed, a sample (60 data) is selected for clustering and the following results are obtained (Fig 3). This Clustering shows that the centroids are not affected by very large or small values. The clusters are formed dynamically. However, the overlapping of the Clusters is noticed.



(a) (b)

Fig. 3. a) Results of K-means k =3, b) Results of CLARA.

In this algorithm, cluster 1 is formed by *Loyal* customers, its average dissimilarity is 3.415, its maximum dissimilarity is 29.449 and it is composed of 294 customers. cluster 2 is formed by *Potential Loyal* customers, its average dissimilarity is 2.941, its maximum dissimilarity is 5.915 and it is composed of 301 customers. Cluster 3 is formed by *Other* customers, its average dissimilarity is 2.876, its maximum dissimilarity is 12.894 and it is composed of 258 customers. The uniform size of the clusters can be noticed by partitioning the data, while the dissimilarity, it is seen that cluster 1 has the highest dissimilarity. This implies that the *Loyal* customer segment differs from the other clusters.

On the other hand, the third algorithm that was used is the hierarchical Clustering algorithm. In this algorithm to perform the clusters, the best possible distance is chosen, and the complete method was the one that best fits the data. This distance is calculated, as the distance between all pairs formed by one observation of cluster A and one of cluster B. The average value of all of them is selected as the distance between the two clusters (mean inter-cluster dissimilarity). The hierarchical algorithm does not require calculating the number of clusters and when the respective dendrogram is available, a cut of the dendrogram is performed. The cut was made at h=5 so that three groups defined in the clustering can be counted. For better visualization of the dendrogram, it can be seen in **Appendix 1**. The Clustering by the hierarchical algorithm, there are two groups marked (blue and green) and cluster 1 (red) is not noticeable. In cluster 1 are the data with higher recency and higher frequency, in this group

are the *Lost* group, while in cluster 2 are the customers of type *Champions Potential Loyal and Loyalty*, and cluster 3 are the rest of the groups such as *Need Attention, At risk, and About to sleep*.

To validate if the number of clusters placed is adequate, the silhouette method was calculated to know if the classification is correct. The average silhouette method is like the Elbow method, but instead of minimizing the total inter-cluster sum of squares, the average silhouette coefficient or silhouette index (SI) is maximized. silhouette quantifies how well an observation has been assigned. The SI coefficient varies between -1 to 1, if the values are close to 1, it can be mentioned that the observation was performed correctly, and if it is less than zero the observation was poorly performed [17]. The silhouette coefficients for this algorithm were: cluster 1 (0.34), cluster 2 (0.35), and cluster 3 (0.32), likewise the data that are misclassified in cluster 1 (2 customers), cluster 2 (21 customers), and cluster 3 (16 customers).

The fourth algorithm tested for classification was the random forest. For this algorithm, a training sample (597 data) and a test sample (256) were taken. The number of trees was 100, which is sufficient for classification [7]. A Depth of 10 and mtry max of 10 were also determined. With this algorithm, the highest classification in each category was achieved. With the training group it was possible to classify all the categories and there were no errors in the classification. The results are *About-To-Sleep* (9 customers), *At Risk* (2 customers), *Champions* (19 customers), *Lost* (23 customers), *Loyal Customers* (43 customers), *Need Attention* (0 customers), *Others* (111 customers), and *Potential Loyalist* (49 customers). This shows that this algorithm had the best-ranking design.

3.4 Associations rules

In the association rules, the data was prepared in such a way that it can be read by the algorithm a priori. Products are listed by the invoice numbers for association rule mining and coded transactions made by customers. *A priori* algorithm is used to find high and low-value customers. These customers are employed to retrieve the products that customers may purchase, after purchasing a relevant product. For that, the *A priori* algorithm finds the associations that exist among the products purchased by customers in the clusters. After that, *A priori* algorithm was applied to identify the associate rules under the minimum support of 0.01 and the minimum confidence of 0.6. The top 3 rules are depicted in **Appendix 2.** All these rules were ordered by the confidence indicator, which indicates the probability that a product A is purchased, given that a product B was purchased.

From the analysis of the best-selling products, it can be noted that most customers buy up to five products, being these products the best sellers: avocado, ripe plantain, tomatoes, philippine banana, eggs, and ginger syrup. For the best-selling product we have the following rule {pickle} => {avocado}, this means that if a customer buys a pickle, he/she is very likely to buy avocado.

In the analysis by segments, the customer segment (*Champions*) is composed of 81 customers, who spend an average of USD 43 on each purchase. In this group, it can be noted that most of them buy category 2, which is vegetables and fruits, and most of

them buy one product or a maximum of three products from this category. For example, tomato has been purchased 2,636 times. So, it can be understood that in this segment there are people who buy food for the home, and they do it frequently. For example, if they buy an avocado and peeled pearl onion, they have a very high probability of buying tomatoes. As for the segment known as *Loyal customers*, the association rule, with more confidence, indicates that onion and tamarillo are purchased in conjunction with tomato. This group also buys fruits and vegetables, as well as chicken and meat (categories 1 and 2). Products such as beef, tamarillo and tomatoes were purchased 1,580 times. As for the *Potential Loyal* segment, only 6 rules were created, and the behavior of the two groups described above is very marked. This group of customers is the highest in the segmentation (157 customers) and together with the *Loyal* (153 customers) is the most representative group. In the *At risk* and *Need Attention* groups, customers who buy category 3 products, e.g., eggs, and milk has been purchased about 741 times. This group is the one to watch out for as they are the people who can get lost. As for the Lost group, it can be noted that they are customers who buy products that are not edible (category 4), but they buy them only once or after some time. It is made up of 71 customers, none of whom make frequent purchases. Most of them buy beer and products such as snacks. Likewise, the segment of customers *About to sleep* has the most preferred for the purchase of products of category 3 and chocolates. This group is composed of 43 customers. The *Other* group was not analyzed since it does not show a marked trend.

4 Discussion

In the proposed research in the retail store, the large data is a problem when analyzing the information. However, many proposals have been developed to deal with this problem. Some of these techniques were used in this work. The RFM analysis contributed to knowing the consumer preferences, and as mentioned by [1], this behavior was searched. In Figure 1, many customers are inactive or are sleeping in the retail business, so it is possible to capture these customers by giving them the possibility of discounts or promotions such as membership cards with benefits. The business should consider a rewards system to increase the percentage of returning customers. As in [2], the RFM study found eight categories of customers, in contrast to the aforementioned research, the *Champions* group is the one that spends the most on average (USD 43) and returns the soonest (10 days). While the *At risk* is the second group that most spends on average (USD 26) and they are a considerable number of customers, therefore, the retail business must consider a marketing campaign for this group so that they are not lost.

On the other hand, as for the ranking algorithms that were used, Table 4 summarizes the comparison of these. It is known that the average Silhouette width uses the average distance of observation and by comparing with all others in the nearest group. In addition, once misclassification is suspected, it is necessary to know how many observations have fallen into these false positives. As for the Dunn indicator, which measures the ratio of the minimum distance of the observations to the maximum dis-

tance between the observations. Therefore, high Dunn index numbers are preferred since this implies having compact and well-separated clusters because the minimum separation between clusters will be high and the maximum separation between clusters will be low. Considering this, for the K-means algorithm, a value of 0.134 was obtained, for the CLARA algorithm a value of 8.338393e-05 was obtained, while for the Hierarchical agglomerative it is 0.0511.

Table 4. Comparison between clustering methods

Clustering method	Measurement	Cluster 1	Cluster 2	Cluster 3
K-means	Silhouette	0.47	0	0.57
	False positive	1	1	1
CLARA	Silhouette	0.39	0.44	0.36
	False positive	3	0	1
Hierarchical Agglomerative	Silhouette	0.34	0.35	0.32
	False positive	2	21	16

Note: False positive is misclassified data.

As can be seen, the numeric data show the best algorithm for classification is K-means, followed by CLARA, and finally the Hierarchical method. These results are similar to those obtained in [2, 4–6], while it was the opposite since in the latter study the Hierarchical method was the best, but in our research, it was the one with the lowest score. For the choice of any of these three algorithms, it is considered that in the presence of data that has a lot of bias, it is necessary to opt for a robust method. Therefore, for the case study, the CLARA method can be chosen, due to the high number of average Silhouette width and low false positives, although the Dunn index is lower. In these classification algorithms, Random Forest was considered, which is a supervised method and requires the use of samples and a defined number of trees. As mentioned by [3],], a good choice for classification is classification trees and a robust method is Random Forest as in the works of [7, 16], the method gave good results. In the case of this research, the classification of this algorithm coincides remarkably where it can be evidenced that there is a perfect classification.

Finally, as for the analysis of the rules of association based on the market basket, the top of these rules can be noted (Appendix 2). It can be noted that customers purchase products from category 1 (beef, chicken and pork, seafood, and sausages) and category 2 (fruits, vegetables, legumes, and eggs) regardless of the segment to which they belong since they are considered elementary in the basic market basket. As for the *Lost* group, according to the rules analyzed, they are one-time customers and non-commodities products. As for the *Others* group, no rules were found that define this group, i.e., the data do not follow any purchase pattern, this could be because many of these customers are people who do not register data or have bought in the store only once, and their purchase was perhaps influenced by the loss of the pandemic caused by Covid-19. With the use of the rules of association, the retail business is given an idea of how to distribute its products and what items should not be missing for its customers, as well as the possibility of creating promotions that include these prod-

ucts. In addition, with these association rules, the retail business can control the flow of customers by placing each product as close as possible or as far away as possible, depending on the objective of the business, whether it is to sell more products (the customer will spend more time in the store, thus increasing the probability of purchasing more products) or to attract customers (the customer will be able to find everything he/she is looking for immediately).

5 Conclusion

The research question posed at the beginning of this research was: Could customer categorization be based on customers' consumption behavior, considering which products they purchase? And it has been shown that customer categorization does follow a pattern based on their recency, frequency, and monetary behavior. In addition, customer clustering is affected by data with high variability, so robust algorithms should be used to avoid misclassifications. On the other hand, the store design is a critical factor in retail store management, which is associated with high investment and maintenance costs. Therefore, having an analysis of the consumer's purchase history and knowing which products are purchased the most and whether they are purchased individually or in groups is important so that shelf space and other equipment are occupied by items that will benefit both the customer and the company. If a business has the necessary products distributed optimally, the customer will likely return and go from being a champion or loyal customer.

For future research, it is recommended that the horizon of analysis might be extended, and the purchase consumer´s behavior, outside the pandemic period, can be compared in order to contrast whether consumption remains the same. In addition, the analysis of brands is important; in this study they were not considered since there was no catalog of business brands. Besides, the seasonality of the data is not considered since the case study, strategy to acquire products is done in a controlled manner and has a product acquisition model that contemplates seasonality.

Appendix

Appendix 1: https://n9.cl/appendix-1
Appendix 2: https://n9.cl/appendix-2

References

1. Joe, T., Sreejith, R., Sekar, K.: Optimization of store layout using market basket analysis. Int. J. Recent Technol. Eng. 8, 6459–6463 (2019). doi:10.35940/ijrte.B2207.078219
2. Parikh, Y., Abdelfattah, E.: Clustering Algorithms and RFM Analysis Performed on Retail Transactions. In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2020. pp. 506–511. Institute of Electrical and Electronics Engineers Inc. (2020)

3. Khalili-Damghani, K., Abdi, F., Abolmakarem, S.: Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. Appl. Soft Comput. J. 73, 816–828 (2018). doi:10.1016/j.asoc.2018.09.001
4. Anitha, P., Patil, M.M.: RFM model for customer purchase behavior using K-Means algorithm. J. King Saud Univ. - Comput. Inf. Sci. 34, 1785–1792 (2022). doi:10.1016/j.jksuci.2019.12.011
5. Rahim, M.A., Mushafiq, M., Khan, S., Arain, Z.A.: RFM-based repurchase behavior for customer classification and segmentation. J. Retail. Consum. Serv. 61, 102566 (2021). doi:10.1016/j.jretconser.2021.102566
6. Saraf, E., Pradhan, S., Joshi, S., Sountharrajan, S.: Behavioral Segmentation with Product Estimation using K-Means Clustering and Seasonal ARIMA. In: 2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings. pp. 1641–1648. Institute of Electrical and Electronics Engineers Inc. (2022)
7. Espinosa Zúñiga, J.J.: Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ing. Investig. y Tecnol. 21, 1–16 (2020). doi:10.22201/fi.25940732e.2020.21.3.022
8. Hu, Y.H., Yeh, T.W.: Discovering valuable frequent patterns based on RFM analysis without customer identification information. Knowledge-Based Syst. 61, 76–88 (2014). doi:10.1016/j.knosys.2014.02.009
9. Yoseph, F., Heikkila, M.: Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. In: Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018. pp. 77–82. Institute of Electrical and Electronics Engineers Inc. (2019)
10. Dahbi, A., Jabri, S., Balouki, Y., Gadi, T.: A new method to select the interesting association rules with multiple criteria. Int. J. Intell. Eng. Syst. 10, 191–200 (2017). doi:10.22266/ijies2017.1031.21
11. Cheng, W.Z., Li Xia, X.: Fast algorithm for mining association rules. Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS. 513–516 (1994). doi:10.1109/ICSESS.2014.6933618
12. Yoseph, F., Heikkila, M.: Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018. 77–82 (2019). doi:10.1109/iCMLDE.2018.00029
13. Schubert, E., Rousseeuw, P.J.: Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In: Amato, G., Gennaro, C., Oria, V., and Miloš, R. (eds.) Similarity Search and Applications. pp. 171–187. Springer International Publishing, Cham (2019)
14. Everitt, B., Landau, S., Leese, M., Stahl, D.: Cluster Analysis. Wiley, Hoboken, NJ, Estados Unidos (2011)
15. Paranavithana, I.R., Rupasinghe, T.D., Prior, D.D.: Unsupervised Learning and Market Basket Analysis in Market Segmentation. In: Lecture Notes in Engineering and Computer Science. pp. 122–127. Newswood Limited (2021)
16. Cánovas-García, F., Alonso-Sarría, F., Gomariz-Castillo, F., Oñate-Valdivieso, F.: Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. Comput. Geosci. 103, 1–11 (2017). doi:10.1016/j.cageo.2017.02.012
17. Amat Rodrigo, J.: Clustering y heatmaps: aprendizaje no supervisado, https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps