Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering

Nouri Hicham, Sabri Karim
Research Laboratory on New Economy and Development (LARNED)
Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University of Casablanca, Morocco

***Abstract*—Customer segmentation is key to a corporate decision support system. It is an important marketing technique that can target specific client categories. We create a novel consumer segmentation technique based on a clustering ensemble; in this stage, we ensemble four fundamental clustering models: DBSCAN, K-means, Mini Batch K-means, and Mean Shift, to deliver a consistent and high-quality conclusion. Then, we use spectral clustering to integrate numerous clustering findings and increase clustering quality. The new technique is more flexible with client data. Feature engineering cleans, processes, and transforms raw data into features. These traits are then used to form clusters. Adjust Rand Index (ARI), Normalized Mutual Information (NMI), Dunn's Index (DI), and Silhouette Coefficient (SC) were utilized to evaluate our model's performances with individual clustering approaches. The experimental analysis found that our model has the best ARI (70.14%), NMI (71.75), DI (75.15), and SC (72.89%). After retaining these results, we applied our model to an actual dataset obtained from Moroccan citizens via social networks and email boxes between 03/06/2022 and 19/08/2022.**

Keywords*—Machine learning; customer segmentation; marketing; clustering ensemble; spectral clustering

I. Introduction

It is possible to win the competition in the market and enhance corporate earnings by better understanding the client's needs. Companies can develop successful marketing strategies if they are aware of the wants and needs of their target audiences. While the requirements and expectations of each customer are unique, many customers share identical or quite similar qualities. Customer segmentation is one method that may be used to put together multiple different consumers who share similar qualities. Improving the quality of the connections with your customers also requires proper consumer segmentation. Marketing intelligence is conducting information analysis to comprehend better a target market and its consumers' demographics [1], [2].

In marketing, it is common for analysts to categorize customers into comparable customer groups to understand better how to advertise to each group of customers. Therefore, segmentation is a collection of approaches that might be useful in categorizing different types of customers. Customers' existing relationships with a company are the primary focus of most direct marketing operations. The more you know about your customer's needs, desires, and purchasing habits, the easier it is to tailor marketing programs to their needs and desires and how they buy things [3].

Marketers can determine the approach that will be most successful in communicating with each unique consumer by segmenting their customer base. Marketers can zero in on particular demographic, behavioral, and other characteristics of their target audience by conducting in-depth analyses of vast amounts of data about existing and prospective clients [4], [5].

A frequent objective in marketing is to increase the worth of each consumer (revenue and profit). To achieve success in this aspect of the marketing mix, it is vital to understand how a particular marketing action influences customers' behaviour. Regarding customer segmentation, and "action-centric" approach prioritizes the impact that marketing activities will have on a customer's lifetime value (CLV) over the value that marketing activities will have in the short term. This is in contrast to a "short-term value" approach. As a direct result, consumers ought to be divided into distinct categories according to the amount of money they will spend throughout their lifetime [6].

In this paper, we ensemble four basic clustering models (DBSCAN, K-means, MiniBatch K-means, and MeanShift) to develop a novel consumer segmentation strategy based on a clustering ensemble, which yields a more consistent and high-quality result than any of the individual clustering techniques. We then use spectral clustering to combine the findings of different clustering methods to increase the overall quality of our clustering results. After the retention of these results, we applied our model to a real dataset, which was collected from Moroccan residents using a questionnaire sent via social networks and email boxes between 03/06/2022 and 25/07/2022. As for the rest of this paper, it is structured as follows: The literature review is described in Section II, the methodology and proposed model are presented in Section III, the findings are covered in Section IV, as well as the findings, interpretations of the study and directions for future research are presented in Section V.

II. Background

Knowing customer behavior in today's highly competitive and ever-changing business environment is crucial. Customers must be categorized according to their demographics and the products they buy. This is an essential component of client segmentation that enables marketers to target certain target groups more precisely with their promotional, marketing, and product development strategies. The relevance of data-driven marketing, a crucial component in client segmentation, is growing due to the growth of data sources, particularly social networks. Because of their immense size and complexity, modern databases make data-driven marketing and customer segmentation exceedingly challenging in the retail industry. Until recently, traditional cluster analysis approaches were employed on retail databases. Nevertheless, because there are so many different kinds of customers today, statistical clustering algorithms find it harder and harder to evaluate and understand what customers do [7].

In the past five years, due to the recent development of machine learning techniques and data science, many one-of-a-kind algorithms in these two fields have been developed for customer relationship management, specifically for customer segmentation. These algorithms have been developed for customer relationship management (CRM) software [8]. An integrated strategy employing the Apriori algorithm and the CRM method with associated mining is used in customer segmentation [9]. This strategy brings the benefits of both methodologies to bear in solving this challenge. For consumer segmentation, [10] utilized two primary methods: the LRFM (Length, Recency, Frequency, and Monetary) techniques and an extended model known as the LRFM-Average Item (AI) model. Both of these methods utilized LRFM techniques. The authors concluded that adding simple parameters and averages did not improve customer segmentation and did not show a significant change in results. This means that complex parameters are needed for better customer segmentation results.

As demonstrated in [11], an investigation into silent customers was carried out since silent consumers are a category of customers that a company runs the risk of quickly losing. As a result, it is essential to research the characteristics of these clients to arrive at the most appropriate business decisions. This research came up with a K-means method for customer segmentation that focused on silent customers to help the company make more money in the telecom industry.

About the Yunnan Electricity Market, an algorithm for market segmentation was developed in [12] that was primarily concerned with density-based spatial clustering of applications with noise (DBSCAN) and the K-means technique. In [13], the K-means clustering algorithm and the SPSS Software were used to construct a real-time and online system for predicting seasonal sales in annual cycles. This system integrated an important complex feature of temporal spikes in the sales of particular items, making it capable of predicting seasonal sales on annual cycles. The study [14] explains how the technique of unsupervised machine learning can be used to tackle the difficult problem of consumer segmentation by analyzing purchase data from credit cards used by African customers. The objective of [15] customer segmentation with a multi-layer perceptron (MLP) neural network is to categorize customers into separate groups according to the characteristics of those categories.

Furthermore, in [16], The Mini Batch K-means technique is implemented to group sediment samples. The clustering result will be verified using a set of four typical evaluation indices. Using this approach, simulations show that the Sample network may be divided into three sedimentary clusters: fluvial, marine, and lacustrine. Researchers have found that experimental results on sediment particle size have an accuracy of up to 0.92254367, suggesting that this technique of studying sedimentary environment by grain size works exceptionally well and precisely. On the other hand, in [17], Using the proposed automatic selection of nearest neighbors for density gradients, it is proven to identify the number, position reliably, and form of non-elliptical clusters in multivariate data analysis and picture segmentation using mean shift.

Recent years have seen an increase in ensemble learning (EL), which has emerged as a successful teaching strategy. EL utilizes a meta-classifier to integrate the results of different classification techniques [18]. The complete training set is utilized for training the base classifiers, and the outputs of the base-level model are used as features in the learning process for the meta-classifier. Another thing to add is that EL is superior to other approaches because it combines the most accurate components of numerous machine learning to provide more accurate predictions than any algorithm in the ensemble can produce. This makes EL the method of choice [19].

III. Methodology

This study offered a new customer segmentation strategy based on the Clustering ensemble technique. DBSCAN, K-means, MiniBatch K-Means, and MeanShift algorithm are used in the suggested method, shown in Fig. 1. The outputs are combined with a consensus function. Stacking ensemble learning allows us to use each model's structural and functional benefits while increasing overall performance. The consensus function and unsupervised machine learning models will be discussed in greater detail in the following paragraphs.

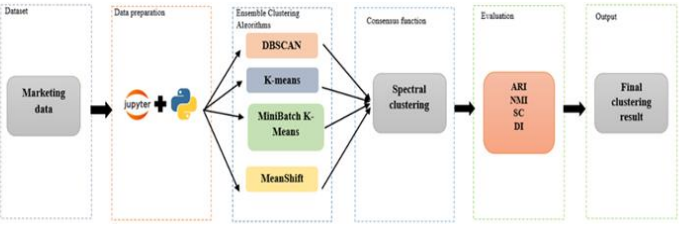

Fig. 1. The Global Architecture.

The general architecture structure is made up of the five fundamental steps listed below:

TABLE I. THE USED DATASET

Variable	Data Type	Range	Description
ID	numerical	Integer	Displays a customer's ID.
Sex	categorical	{0,1}	Customer's gender. 0=male/1=female
Marital status	categorical	{0,1}	Customer's marital status. 0=single/1=non-single
Age	numerical	Integer	The customer's age in years
Education	categorical	{0,1,2,3}	Customer’s education. 1=high school, 2=college, 3=graduate
Income	numerical	Real	Customer's self-reported annual US income.
Occupation	categorical	{0,1,2}	Customer'sprofession.0=unemployed/1=employee/official/2=management or self-employed
Settlement size	categorical	{0,1,2}	Customer's city size. 0=small/1=midsize/2=big

TABLE II. SOME DESCRIPTIONS OF DATASET

	ID	**Sex**	**Marital status**	**Age**	**Education**	**Income**	**Occupation**	**Settlement size**
count	35000	35000	35000	35000	35000	35000	35000	35000
mean	1.00	0.457	0.496500	35.909	1.038	120954.419000	0.810500	0.739000
std	5.77e+02	0.4982	0.500113	11.7194	0.599	38108.824679	0.638587	0.812533
min	1.00	00	00	18.00	0.00	35832.00	00	00
25%	1.00	00	00	27.00	1.00	97663.250	00	00
50%	1.00	00	00	33.00	1.00	115548.50	1.00	1.00
75%	1.00	1.00	1.00	42.00	1.00	138072.250	1.00	1.00
max	1.00	1.00	1.00	76.00	3.00	309364.00	2.00	2.00

A. *Dataset*

The data for this study, which involves 35,000 clients and eight different characteristics, was collected from a supermarket mall. This database includes essential customer information, such as the customer's identification number (Customer ID), annual income, gender, age, and expenditure score. In order to make sense of the marketing team and develop a suitable strategy for the situation, we need to have a solid understanding of these clients, such as who the target customers are. The full dataset description is provided in Table I and II.

1) Data preparation: Data selection, preprocessing, and transformation are the three stages of the process that are involved in getting data suitable for an algorithm that does machine learning [20] [21].

2) Data selection: At this point, we narrow down all of the data we have access to and are utilizing by selecting a subset of it to work with. Consider the data we already own, the data we do not possess, and the data we can get rid of.

3) Data preprocessing: The preprocessing of data is necessary because we frequently receive a large amount of raw data that machine learning algorithms cannot use. It is essential to process the raw data entirely before incorporating it into the various machine learning algorithms. In order to compile our selected data, we first formatted it, then cleaned it, and last, we took some samples from it. The data's poor quality hinders several attempts to process the data.

4) Data transformation: Data processing can be transformed through a series of procedures known as "data transformation." Another name for this practice is "feature engineering." The extraction of features from our data is a time-consuming process, but the benefits of machine learning may be worth the wait. The following are the three most frequent methods in which data is altered:

Depending on the amount being measured, the properties of the pre-processed data may have been given a variety of scales. Each of the characteristics must be the same size.

- Aggregation: It is possible that some features can be combined to make a single feature that fits the issue that we are attempting to solve better.

- Decomposition: It is possible that intricate elements are easier to understand if they are broken up into chunks. Decomposing the subject into its parts may be helpful. For illustration, a feature that displays the time and date as a long string can be simplified such that it only displays the current hour of the day.

B. *Ensemble Clustering Algorithms*

The DBSCAN, K-means, MiniBatch K-means, and MeanShift basis clustering algorithms will be the topics of discussion in this article section.

- DBSCAN

Density-Based Clustering is a term that refers to different unsupervised learning approaches that identify different clusters in the data. These approaches are premised on the idea that a group in data space is a sector that contains a significant number of points and is partitioned into smaller regions that include a less considerable amount of points [22].

The abbreviation for a technique that uses density-based clustering as its basis is DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise. Using large amounts of data with outlying values and noise, clusters of different shapes and sizes can be found [23].

The DBSCAN method makes use of two different parameters, which are as follows [24]:

minPts is the minimum amount of clustered points (also known as a threshold) that are necessary for a region to be considered dense.

Eps() is an abbreviation for the distance measure used to determine which points are located in close proximity to a specific point.

- K-means

Unsupervised machine learning K-means clustering may be used on a dataset to determine the data object groupings within it. There are other alternative data grouping methods, but the k-means clustering method is one of the oldest and simplest to comprehend [25]. The "K-Means algorithm" parameter "k" represents the number of data clusters. We have two ways of picking the right number of clusters: the Elbow method (which graphs random values to achieve the best k value) (see Fig. 5) and the Silhouette score approach (the value which has the greatest score will be considered as the optimal k value) (the value which has the highest score will be taken as the best k value). Using these two, we find the optimum k value is three and train the model using three [26]. According to the Euclidean distance, clustering goals are chosen which reduce the total of squares of all types in a given data set X, which contains n multi-dimensional data points.

$$d = \sum_{k=1}^{k}\sum_{i=1}^{n} \mathrm{II}(x_i - u_k)\,\mathrm{II}^2$$

K denotes the centers of the first K clusters, u_k is shorthand for the kth cluster center, and xi is shorthand for the ith data point. The answer to the problem of the centroid uk can be found as follows:

$$\frac{\delta}{\delta uk} = \frac{\mathrm{\delta y}}{\mathrm{\delta uk}}\sum_{k=1}^{k}\sum_{i=1}^{n}(x_i - u_k)^2$$

$$= \sum_{k=1}^{k}\sum_{i=1}^{n}\frac{\delta y}{\delta uk}(x_i - u_k)^2$$

$$= \sum\nolimits_{i=1}^{n} 2(x_i - u_k)$$

Let the second equation be zero. Then $u_k = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} xi$

- MiniBatch K-means

Clustering on enormous datasets can also be accomplished using the Mini-batch K-means clustering algorithm as an alternative to the K-means technique. It frequently outperforms the standard K-means algorithm when working with big datasets since it does not cycle over the entire dataset. This is because it does not cycle over the entire dataset. It starts by making random data batches to store them in memory. Then, for each loop, it collects a random data batch to update the clusters [27] [29].

The key benefit of mini-batch K-means is that it reduces the processing required to identify clusters. The K-means algorithm may be more to your liking, but when dealing with a big dataset, the mini-batch method is the way to go [28].

- Mean Shift algorithm

The Mean Shift Clustering Algorithm is a technique based on the centroid that is useful in various applications, including unsupervised learning. It is one of the most successful algorithms for a variety of machine learning applications, including clustering, which is one of those applications. Each individual data point is then relocated in the direction of the centroids of the region, which are determined by taking the average of all the other places. A different name for this technique is the mode-seeking algorithm. The benefit of the method is that it disperses groups of objects according to the data without automatically estimating the number of clusters depending on bandwidth. This is an advantage over competing algorithms.

C. Consensus Function

Clustering ensemble (CE) is a method that has emerged as an essential tool for improving the overall quality of clustering solutions. This method merges the multiple clustering results obtained through DBSCAN, K-means, MiniBatch K-means, and MeanShift. This research project describes a novel approach to cluster ensembles predicated on spectral clustering. This function is the first step in the CE algorithm, and it is possible to improve the outcomes of individual clustering algorithms because it is the major step in the algorithm [30] [31]. The final consensus partition is found, which is the result of any CE technique that has been used.

D. Evaluation

Adjust Rand Index (ARI), Normalised Mutual Information (NMI), Dunn's Index (DI), and Silhouette Coefficient (SC) were the performance indicators that we used to compare the results obtained by the proposed method with those obtained by base classifiers (DBSCAN, K-means, MiniBatch K-means, and MeanShift). This allowed us to determine whether or not the method that was proposed was effective. We applied it to the dataset that was described above.

- Adjust Rand Index (ARI)

The adjusted Rand index (ARI) is a measurement that is frequently utilized in the field of cluster analysis to determine the level of agreement that exists between two data partitions. Since the index's inception, there has been a growing amount of interest in investigating cases involving extreme agreement and disagreement under various conditions. This investigation aims to gain a deeper understanding of the index [32], [33].

The following denotes the ARI:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/(\binom{n}{2})}{1/2[\sum_i\binom{a_i}{2} + \sum_j\binom{b_j}{2}] - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/(\binom{n}{2})}$$

Where n_{ji}, a_i and b_j are values from the contingence table.

- Normalised Mutual Information (NMI)

Normalized mutual information, also known as NMI, is a measurement that is frequently utilized to compare different community identification approaches. More recently, the necessity of adjusting information theory-based measures have been advocated because of the so-called selection bias problem. This problem is that these kinds of measures tend to pick clustering solutions that include more communities [34], [35]. The Mutual Information (MI) score is normalized to produce the Normalized Mutual Information (NMI), which scales the results between 0 and 1 (1 perfect correlation).

The following formula shows the NMI:

$$NMI(y,c) = \frac{2 * \mathrm{I(y,c)}}{[H(y) + H(c)]}$$

where, 1) y = class names 2) c = group identifiers 3) H(.) = Entropy 4). The formula for the amount of mutual information between y and c is I(y; c).

- Silhouette Coefficient (SC)

A statistic that can be used to evaluate a clustering technique is called the silhouette score. The silhouette score is the sum of two separate scores, referred to as a and b. ‘a’ indicates the average range between a sampling site and every other point that is part of the same cluster. [36] In contrast, ‘b’ shows the typical mean distance between a sample and all other points that are part of the cluster that is the next closest to it [37]. The following formula is used to determine a sample's score for the silhouette category:

$$s = \frac{b - \mathrm{a}}{\max(\mathrm{a,b})}$$

The average of the silhouette scores obtained for each sample constitutes a set's overall silhouette score. The score for the silhouette might range anywhere from -1 to +1. A silhouette score of -1 indicates that the clustering was inaccurate, whereas a score of +1 indicates that the clustering was correct and highly dense. A score of 0 for the silhouette indicates that the groups are overlapping.

- Dunn’s Index(DI)

Another statistic might be employed when evaluating clustering techniques, like DI [38]. The formula for calculating DI is dividing the shortest distance between clusters by the most significant size possible. A more significant DI indicates that the clustering is done more effectively. It works under the assumption that more effective clustering results in clusters that are both closely packed and physically distinct from one another [39]. The following formula can be used to determine the value of the Dunn's Index:

$$DI = \frac{min_{1 \le i < j \le n} \mathrm{d(i,j)}}{max_{1 \le k \le n} \mathrm{d'(k)}}$$

where I j, and k are indices for groups, d represents the distance between clusters, and d' measures the difference between clusters within the same group.

IV. Results and Discussions

The following information can be found in this section: The research results of our tests assess the effectiveness of the most common diverse analytical and unsupervised techniques and offer a new customer segmentation strategy based on the Clustering ensemble technique. The suggested strategy uses the algorithms DBSCAN, K-means, MiniBatch K-means, and MeanShift.

Before beginning the presentation and discussion of the results obtained, it seems like it would be of some use to us to have a better understanding of the potential connections that exist between the many different variables that are contained within our dataset. The bivariate analysis of the relationships between the various factors is depicted in Fig. 2.

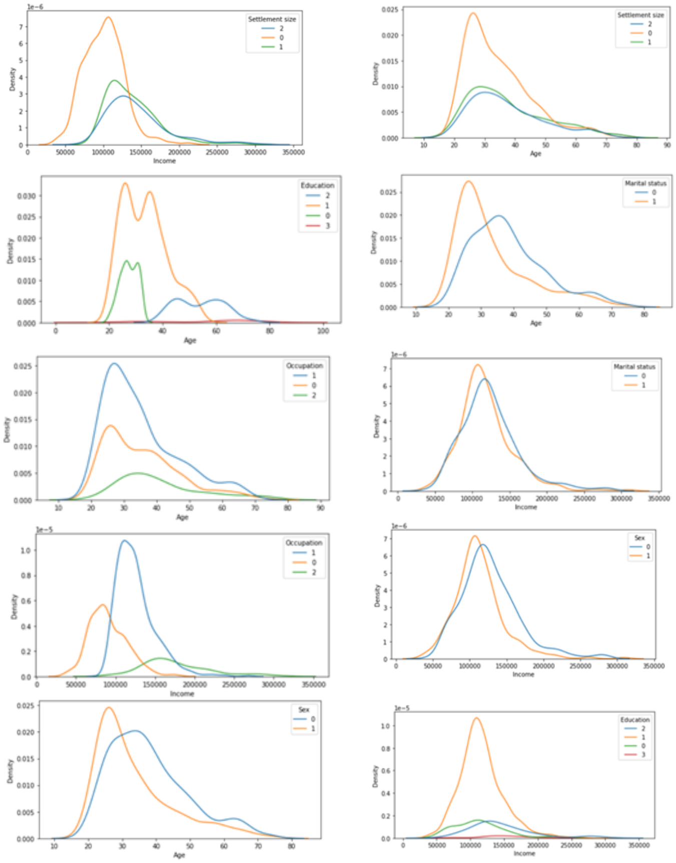

Fig. 2. Bivariate Analysis between the Various Variables.

The analysis of the data reveals several interesting correlations, including the following (Fig. 3):

1) People with an occupation of '0' (unemployed) are more likely to reside in smaller cities closer to the customers.

2) Married customers are more likely to have post-graduate or high school-level education.

3) Married people are more likely to reside in less populous cities.

4) A woman on the client list is more likely to be married than a man.

5) Males make up a larger proportion of the customer list than females. On the other hand, the dataset has a disproportionately high number of unemployed women.

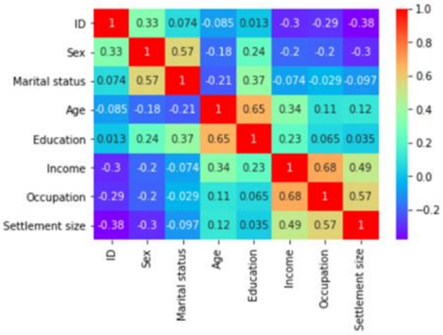

Fig. 3. Correlation between the Various Variables.

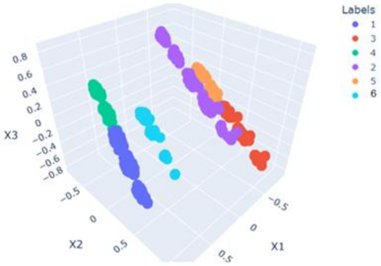

Fig. 4. Clusters Visualization.

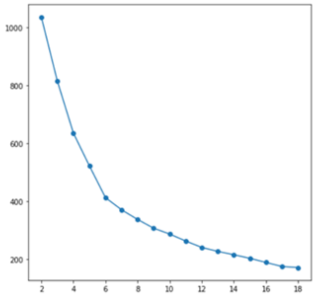

Fig. 5. Elbow Method (k=6).

To begin, we should note that the clusters are incredibly distinct from one another and that the algorithm performs the cluster separation task accurately, given that the clusters' boundaries appear pretty distinct.

In the following, we present the statistical characteristics of the different clusters generated by our model (Fig. 4).

This cluster comprises unemployed, middle-income, single men residing in small cities. As we will see later in the analysis, this also occurs in other clusters. Therefore we conclude that age and education are good cluster separators for clusters 2 and 4 in our dataset.

This cluster consists of married women with a high school diploma or higher and a moderate salary. They are either unemployed or work as employees/officials and reside in small cities. The age falls within the same range as before, so we will disregard it.

Summary statistics of Clusters 1 to 6 are shown below in Tables III to VIII.

TABLE III. SUMMARY STATISTICS OF CLUSTER 1

	count	mean	std	min	25%	50%	75%	max
Sex	9048.0	00	00	0.0	0.0	0.0	0.0	0.0
Marital status	9048.0	00	00	0.0	0.0	0.0	0.0	0.0
Age	9048.0	40.00	11.94	20.0	31.0	37.0	47.0	75.0
Education	9048.0	0.82	0.61	0.0	0.0	1.0	1.0	2.0
Income	9048.0	145373	38286	82398	119276	136323	159757	287247
Occupation	9048.0	1.26	0.48	0.0	1.0	1.0	2.0	2.0
Settlement size	9048.0	1.52	0.50	0.0	1.0	2.0	2.0	2.0
Labels	517.0	00	00	0.0	0.0	0.0	0.0	0.0

TABLE IV. SUMMARY STATISTICS OF CLUSTER 2

	count	mean	std	min	25%	50%	75%	max
Sex	5355.0	1.00	0.0	1.0	1.0	1.0	1.0	1.0
Marital status	5355.0	1.00	0.0	1.0	1.0	1.0	1.0	1.0
Age	5355.0	34.28	13.00	18.0	25.0	29.0	40.0	76.0
Education	5355.0	1.33	0.57	1.0	1.0	1.0	2.0	3.0
Income	5355.0	136536	38103	88800	108455	126778	155107	309364
Occupation	9048.0	1.26	0.48	0.0	1.0	1.0	2.0	2.0
Settlement size	9048.0	1.26	0.48	0.0	1.0	1.0	2.0	2.0
Labels	5355.0	1.18	0.40	0.0	1.0	1.0	1.0	2.0

TABLE V. SUMMARY STATISTICS OF CLUSTER 3

	count	mean	std	min	25%	50%	75%	max
Sex	5460.0	1.00	0.0	1.0	1.00	1.0	1.00	1.0
Marital status	5460.0	1.00	0.0	1.0	1.00	1.0	1.00	1.0
Age	5460.0	32.41	10.91	18.0	25.00	28.5	36.00	71.0
Education	5460.0	1.21	0.46	1.0	1.00	1.0	1.00	3.0
Income	5460.0	102142	25801	35832	86281	102323	120459	207262
Occupation	5460.0	0.42	0.49	0.0	0.00	0.0	1.00	1.0
Settlement size	5460.0	0.01	0.10	0.0	0.00	0.0	0.00	1.0
Labels	5460.0	2.00	0.0	2.0	2.00	2.0	2.00	2.0

Cluster 3 comprises non-single females with at least a high school diploma and a high level of education and income. Large to medium-sized cities are where they reside. This cluster's distribution of Age is also identical; hence this trait does not provide any further information.

If we look at the median Age of the people in clusters 2 and 3, we obtain 28.5 and 29 years, respectively, whereas the median Age of the people in clusters 1 and 4 is significantly older (Clusters 0 and 3 have a median value of 36, while clusters 3 have a median value of 37.). However, the change is not negligible, as we can see from the next cell.

In Table VI, the cluster represented is single men with higher incomes and managerial or self-employed employment. They reside in medium to big urban centers. Education is comparable to cluster one, with the majority holding a high school diploma or less. Ages appear to fall within the same range as cluster one; hence they will not be considered.

Cluster 5 is made up of married or cohabiting men who have completed high school or higher levels of education but have a low to medium income. The vast majority of them hold jobs as employees or officials. It is important to note that this consumer base is equally represented in small, medium, and large cities; hence, we have chosen to ignore that fact in this research. Age suffers from the same issue as the other clusters, which is why it is not taken into consideration.

Cluster 6 is made up of single females of low education. The vast majority of them have completed their high school education, reside in small cities, and are either unemployed or working for someone else.

The performance of this model, which is based on clustering ensemble and spectral clustering, is presented together with its comparison to the performance of other classical models in Table VIII, which allows us to evaluate the effectiveness of our model.

As a result of the findings acquired at the ARI, NMI, SC, and DI levels, respectively, compared to other traditional models, our model demonstrates the highest level of performance across all four evaluation levels.

TABLE VI. .SUMMARY STATISTICS OF CLUSTER 4

	count	mean	std	min	25%	50%	75%	max
Sex	4497.0	0.0	0.0	0.0	0.00	0.0	0.00	0.0
Marital status	4497.0	0.0	0.0	0.0	0.00	0.0	0.00	0.0
Age	4497.0	37.56	10.63	21.0	29.75	36.0	42.00	74.0
Education	4497.0	0.73	0.57	0.0	0.00	1.0	1.00	2.0
Income	4497.0	102566	26584	43684	81804	103618	120396	219319
Occupation	4497.0	0.36	0.50	0.0	0.00	0.0	1.00	2.0
Settlement size	4497.0	0.06	0.23	0.0	0.00	0.0	0.00	1.0
Labels	4497.0	3.0	0.0	3.0	3.00	3.0	3.00	3.0

TABLE VII. SUMMARY STATISTICS OF CLUSTER 5

	count	mean	std	min	25%	50%	75%	max
Sex	3115.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Marital status	3115.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Age	3115.0	33.96	10.75	18.0	26.0	31.0	40.0	67.0
Education	3115.0	1.26	0.50	1.0	1.0	1.0	1.0	3.0
Income	3115.0	122976	38529	62263	96769	115369	146519	280570
Occupation	3115.0	0.93	0.63	0.0	1.0	1.0	1.0	2.0
Settlement size	3115.0	0.918288	0.827468	0.0	0.0	1.0	2.0	2.0
Labels	3115.0	4.0	0.0	4.0	4.0	4.0	4.0	4.0

TABLE VIII. SUMMARY STATISTICS OF CLUSTER 6

	count	mean	std	min	25%	50%	75%	max
Sex	7525.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Marital status	7525.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Age	7525.0	35.14	9.75	19.0	27.0	34.5	41.0	70.0
Education	7525.0	0.93	0.55	0.0	1.0	1.0	1.0	3.0
Income	7525.0	97997	21702	36760	80892	101511	113265	143321
Occupation	7525.0	0.37	0.50	0.0	0.0	0.0	1.0	2.0
Settlement size	7525.0	0.07	0.26	0.0	0.0	0.0	0.0	1.0
Labels	7525.0	5.00	0.0	5.0	5.0	5.0	5.0	5.0

Then, we put our model to the test by applying it to an actual database collected from Moroccan residents using a questionnaire sent via social networks and email boxes between the dates of 03/06/2022 and 19/08/2022. This authentic database had 1357 individuals with eight distinguishing traits (the same as the last database). We used the programming language python and its library to manipulate and process the collected data. The results that were obtained are displayed in the figure that follows (Fig. 6).

Comparative analysis of several performance metrics for various models is shown in Table IX.

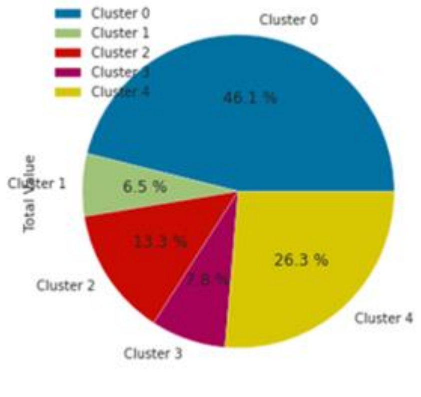

Fig. 6. Percentage of the 5 Clusters.

TABLE IX. COMPARATIVE ANALYSIS OF SEVERAL PERFORMANCE METRICS FOR VARIOUS MODELS

	DBSCAN	K-means	MiniBatch K-means	MeanShift	**The proposed model**
ARI	0.6953	0.6917	0.6252	0.6164	0.7014
NMI	0.7110	0.7035	0.6937	0.6839	0.7175
SC	0.7215	0.7172	0.6991	0.6927	0.7289
DI	0.7461	0.7201	0.7104	0.7063	0.7515

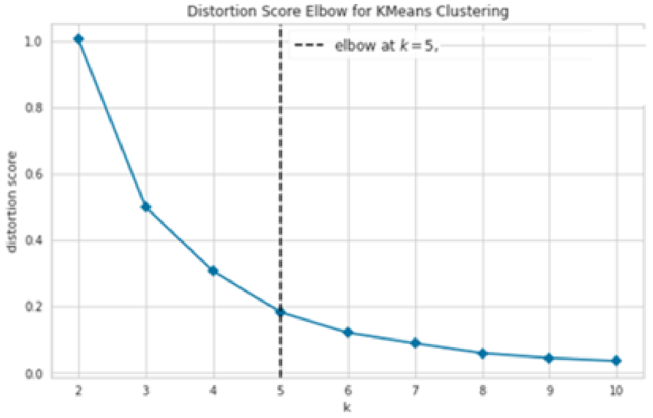

Fig. 7. Elbow Method (k=5).

We can clearly see in Fig. 7 that the elbow method reports five different clusters and that the distribution of these clusters differs from each other; however, cluster 0 has the highest percentage of 46.10%, followed by cluster 4 with a value of 26.30%, while clusters 1, 2 and 3 have a total of 27.60%.

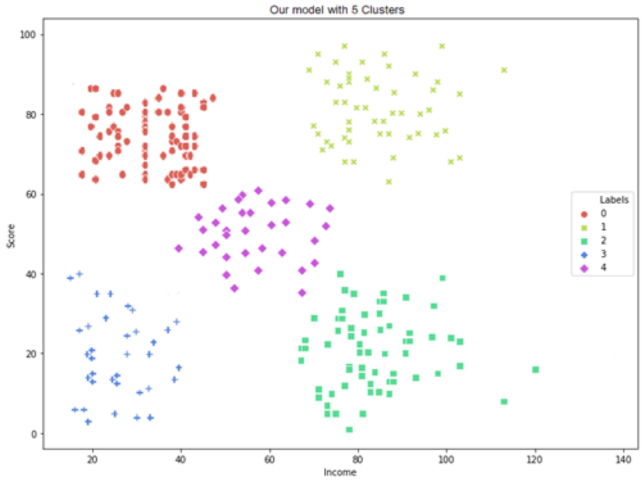

Fig. 8. Presentation of the Clusters According to our Model.

The clusters generated in this real dataset are characterized by the following (Fig. 8):

- Cluster 0 consists of middle-class single men who are highly educated and/or self-employed. We could be inclined to claim they are between twenty and forty years old.
- People in Cluster 1 are evenly split between the sexes; on average are 56 years old and have all earned at least a bachelor's degree.
- Cluster 2 consists of women who are married and have higher incomes, as well as management or self-employed employment. They have education up to a high school diploma and live in medium to large cities.
- Single people of either gender with a middle-class income and education equal to or higher than that of a high school graduate make up cluster 3. They reside in relatively tiny towns and either do not have jobs or work as employees or officials.
- Cluster 4 includes married or cohabiting women who have graduated from high school or above and have a high income. They are either employed or run their businesses. They call somewhat large to sizable cities home.

V. CONCLUSION AND FUTURE DIRECTIONS

This paper proposes a novel clustering ensemble approach; based on a clustering ensemble; in this step, we used four essentials clustering models; DBSCAN, K-means, Mini Batch K-means, and Mean Shift, to provide a superior-conclusion, in terms of consistency and quality, to that produced by the individual clustering algorithms. After that we utilize spectral clustering to merge the multiple clustering results to improve the overall quality of clustering solutions. After the retention of these results, we applied this model to a real dataset, which was collected from Moroccan residents using a questionnaire sent via social networks and email boxes between 03/06/2022 and 19/08/2022. Therefore, the research can involve deep learning models and other performance indicators. The model can also be compared to other datasets.

REFERENCES

[1] S. Das et J. Nayak, « Customer Segmentation via Data Mining Techniques: State-of-the-Art Review », in Computational Intelligence in Data Mining, Singapore, 2022, p. 489-507.

[2] X. Xiahou et Y. Harada, « B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM », J. Theor. Appl. Electron. Commer. Res., vol. 17, no 2, p. 458-475, avr. 2022, doi: 10.3390/jtaer17020024.

[3] C.-L. Wu et N. K. Ma, « The impact of customised mobile marketing on passenger shopping behaviour in the airport terminal », J. Retail. Consum. Serv., vol. 66, p. 102941, mai 2022, doi: 10.1016/j.jretconser. 2022.102941.

[4] A. Griva, « "I can get no e-satisfaction". What analytics say? Evidence using satisfaction data from e-commerce », J. Retail. Consum. Serv., vol. 66, p. 102954, 2022, doi: https://doi.org/10.1016/j.jretconser.2022 .102954.

[5] H. Jin, « The effect of overspending on tariff choices and customer churn: Evidence from mobile plan choices », J. Retail. Consum. Serv., vol. 66, p. 102914, mai 2022, doi: 10.1016/j.jretconser.2022.102914.

[6] School of Management, Hospitality and Tourism (ESGHT), Universidade do Algarve, Portugal et al., « Hotel customer segmentation and sentiment analysis through online reviews: An analysis of selected European markets », Tour. Manag. Stud., vol. 18, no 1, p. 29-40, janv. 2022, doi: 10.18089/tms.2022.180103.

[7] S. P. Nguyen, « Deep customer segmentation with applications to a Vietnamese supermarkets' data », Soft Comput., vol. 25, no 12, p. 7785-7793, juin 2021, doi: 10.1007/s00500-021-05796-0.

[8] M. Carnein et H. Trautmann, « Customer Segmentation Based on Transactional Data Using Stream Clustering », in Advances in Knowledge Discovery and Data Mining, vol. 11439, Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, et S.-J. Huang, Éd. Cham: Springer International Publishing, 2019, p. 280-292. doi: 10.1007/978-3-030-16148-4_22.

[9] B. Kaur et P. K. Sharma, « Implementation of Customer Segmentation using Integrated Approach », vol. 8, no 6, p. 3, 2019.

[10] [10] P. P. Pramono, I. Surjandari, et E. Laoh, « Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method », in 2019 16th International Conference on Service Systems and Service Management (ICSSSM), Shenzhen, China, juill. 2019, p. 1-5. doi: 10.1109/ICSSSM.2019.8887704.

[11] Y. Qiu, P. Chen, Z. Lin, Y. Yang, L. Zeng, et Y. Fan, « Clustering Analysis for Silent Telecom Customers Based on K-means++ », in 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, vol. 1, p. 1023-1027. doi: 10.1109/ITNEC48623.2020.9084976.

[12] X. Wang et al., « Electricity Market Customer Segmentation Based on DBSCAN and k-Means : —A Case on Yunnan Electricity Market », in 2020 Asia Energy and Electrical Engineering Symposium (AEEES), 2020, p. 869-874. doi: 10.1109/AEEES48850.2020.9121413.

[13] K. R. Kashwan et C. Velu, « Customer Segmentation Using Clustering and Data Mining Techniques », Int. J. Comput. Theory Eng., p. 856-861, 2013.

[14] E. Umuhoza, D. Ntirushwamaboko, J. Awuah, et B. Birir, « Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa », SAIEE Afr. Res. J., vol. 111, no 3, p. 95-101, 2020, doi: 10.23919/SAIEE.2020.9142602.

[15] Ş. Ozan et L. O. Iheme, « Artificial Neural Networks in Customer Segmentation », in 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019, p. 1-4. doi: 10.1109/SIU.2019.8806558.

[16] Q. Su, Y. Zhu, Y. Jia, P. Li, F. Hu, et X. Xu, « Sedimentary Environment Analysis by Grain-Size Data Based on Mini Batch K-Means Algorithm », Geofluids, vol. 2018, p. 1-11, déc. 2018, doi: 10.1155/2018/8519695.

[17] T. Duong, G. Beck, H. Azzag, et M. Lebbah, « Nearest neighbour estimators of density derivatives, with application to mean shift clustering », Pattern Recognit. Lett., vol. 80, p. 224-230, sept. 2016, doi: 10.1016/j.patrec.2016.06.021.

[18] K. Sarkar, « A Stacked Ensemble Approach to Bengali Sentiment Analysis », in Intelligent Human Computer Interaction, Cham, 2020, p. 102-111.

[19] S. Ardabili, A. Mosavi, et A. R. Várkonyi-Kóczy, « Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods », MATHEMATICS & COMPUTER SCIENCE, preprint, août 2019. doi: 10.20944/preprints201908.0203.v1.

[20] L. Berti-Equille, « Learn2Clean: Optimizing the Sequence of Tasks for Web Data Preparation », in The World Wide Web Conference on - WWW '19, San Francisco, CA, USA, 2019, p. 2580-2586. doi: 10.1145/3308558.3313602.

[21] R. Heinrich, « Structured Data Preparation Pipeline for Machine Learning-Applications in Production », p. 6.

[22] D. Ienco et G. Bordogna, « Fuzzy extensions of the DBScan clustering algorithm », Soft Comput., vol. 22, no 5, p. 1719-1730, mars 2018, doi: 10.1007/s00500-016-2435-0.

[23] D. Deng, « DBSCAN Clustering Algorithm Based on Density », in 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, sept. 2020, p. 949-953. doi: 10.1109/IFEEA51475.2020.00199.

[24] D. Devarapalli, A. S. Virajitha, B. T. Keerthi, et A. P. Devi, « Analysis of RFM Customer Segmentation Using Clustering Algorithms », Int. J. Mech. Eng., vol. 7, no 1, p. 8, 2022.

[25] K. P. Sinaga et M.-S. Yang, « Unsupervised K-Means Clustering Algorithm », IEEE Access, vol. 8, p. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[26] C. Yuan et H. Yang, « Research on K-Value Selection Method of K-Means Clustering Algorithm », J, vol. 2, no 2, p. 226-235, juin 2019, doi: 10.3390/j2020016.

[27] M. M. Chavan, A. Patil, L. Dalvi, et A. Patil, « Mini Batch K-Means Clustering On Large Dataset », p. 3.

[28] S. C. Hicks, R. Liu, Y. Ni, E. Purdom, et D. Risso, « mbkmeans: fast clustering for single cell data using mini-batch k-means », p. 29.

[29] Y. Ren, U. Kamath, C. Domeniconi, et G. Zhang, « Boosted Mean Shift Clustering », in Machine Learning and Knowledge Discovery in Databases, vol. 8725, T. Calders, F. Esposito, E. Hüllermeier, et R. Meo, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, p. 646-661. doi: 10.1007/978-3-662-44851-9_41.

[30] D. Tang, J. Man, L. Tang, Y. Feng, et Q. Yang, « WEDMS: An advanced mean shift clustering algorithm for LDoS attacks detection », Ad Hoc Netw., vol. 102, p. 102145, mai 2020, doi: 10.1016/j.adhoc.2020.102145.

[31] N. Nguyen et R. Caruana, « Consensus Clusterings », in Seventh IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, USA, oct. 2007, p. 607-612. doi: 10.1109/ICDM.2007.73.

[32] J. E. Chacón et A. I. Rastrojo, « Minimum adjusted Rand index for two clusterings of a given size ». arXiv, 9 décembre 2020. Consulté le: 15 juin 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2002.03677.

[33] J. M. Santos et M. Embrechts, « On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification », in Artificial Neural Networks – ICANN 2009, vol. 5769, C. Alippi, M. Polycarpou, C. Panayiotou, et G. Ellinas, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, p. 175-184. doi: 10.1007/978-3-642-04277-5_18.

[34] A. Amelio et C. Pizzuti, « Correction for Closeness: Adjusting Normalized Mutual Information Measure for Clustering Comparison: Correction For Closeness: Adjusting NMI », Comput. Intell., vol. 33, no 3, p. 579-601, août 2017, doi: 10.1111/coin.12100.

[35] S. Romano, J. Bailey, N. X. Vinh, et K. Verspoor, « Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance », p. 9.

[36] R. Hidayati, A. Zubair, A. H. Pratama, et L. Indana, « Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering », Techno.Com, vol. 20, no 2, p. 186-197, mai 2021, doi: 10.33633/tc.v20i2.4556.

[37] H. Řezanková, « Different Approaches to the Silhouette Coefficient Calculation in Cluster Evaluation », p. 10, 2018.

[38] C.-E. Ben Ncir, A. Hamza, et W. Bouaguel, « Parallel and scalable Dunn Index for the validation of big data clusters », Parallel Comput., vol. 102, p. 102751, mai 2021, doi: 10.1016/j.parco.2021.102751.

[39] S. Mahallati, J. C. Bezdek, D. Kumar, M. R. Popovic, et T. A. Valiante, « Interpreting Cluster Structure in Waveform Data with Visual Assessment and Dunn's Index », in Frontiers in Computational Intelligence, vol. 739, S. Mostaghim, A. Nürnberger, et C. Borgelt, Éd. Cham: Springer International Publishing, 2018, p. 73-101. doi: 10.1007/978-3-319-67789-7_6.