




# How can entrepreneurs improve digital market segmentation? A comparative analysis of supervised and unsupervised learning algorithms

Laura Sáez-Ortuño<sup>1</sup>  · Ruben Huertas-García<sup>1</sup> · Santiago Forgas-Coll<sup>1</sup> · Eloi Puertas-Prats<sup>2</sup>

Accepted: 1 August 2023 / Published online: 5 August 2023  
© The Author(s) 2023

## Abstract

The identification of digital market segments to make value-creating propositions is a major challenge for entrepreneurs and marketing managers. New technologies and the Internet have made it possible to collect huge volumes of data that are difficult to analyse using traditional techniques. The purpose of this research is to address this challenge by proposing the use of AI algorithms to cluster customers. Specifically, the proposal is to compare the suitability of supervised algorithms, XGBoost, versus unsupervised algorithms, K-means, for segmenting the digital market. To do so, both algorithms have been applied to a sample of 5 million Spanish users captured between 2010 and 2022 by a lead generation start-up. The results show that supervised learning with this type of data is more useful for segmenting markets than unsupervised learning, as it provides solutions that are better suited to entrepreneurs' commercial objectives.

**Keywords** Digital marketing · Clusters · AI algorithms · Unsupervised algorithms · Supervised algorithms · XGBoost · K-means

---

✉ Laura Sáez-Ortuño  
laurasaez@ub.edu

Ruben Huertas-García  
rhuertas@ub.edu

Santiago Forgas-Coll  
Santiago.forgas@ub.edu

Eloi Puertas-Prats  
epuertas@ub.edu

<sup>1</sup> Business Department, Universitat de Barcelona, Avda. Diagonal, 690, Barcelona 08034, Spain

<sup>2</sup> Maths and Computer Science Department, Universitat de Barcelona, Gran Via, 585, Barcelona 08007, Spain

## Introduction

In recent decades, new developments in information technologies, the Internet, social networks, blogs, and so on have led to the emergence of new forms of purchasing (Audretsch et al., 2020; Balioukas et al., 2022; Desai, 2019). This ecosystem of digital connections facilitates the generation and dissemination of information both between consumers and between consumers and companies (Hartmann et al., 2019). Until recently, online information was mostly recorded and transmitted in written form, which made it easy for market research companies to collect in order to explore developments in the market, and to extract ideas and knowledge about both traditional and online products and services (González-Padilla et al., 2023). The gathering of this kind of data not only serves to study online marketplaces such as eBay, Taobao, Über and AirBnB (Tadelis, 2016), but also to complement data collected through more traditional market research (Netzer et al., 2012).

The entrepreneurship literature considers value-creating propositions to emerge after exploring business opportunities (Amit & Zott, 2012; Dahle et al., 2018; Guerola-Navarro et al., 2022) and, furthermore, that the entrepreneurial process follows a systematic sequence (Dahle et al., 2023). Dahle et al. (2023) classify this sequence into seven stages: (1) project purpose, (2) resource identification, (3) exploring the business idea, (4) testing the business model, (5) goal setting, (6) task specification, and (7) project development forecasting. This study focuses on the third activity, that of exploring the business idea and trying to define what the project will achieve (Dahle et al., 2018). At this stage, the entrepreneur focuses on the customer and, to do so, must identify his or her target audience. In other words, he or she must collect relevant information on consumer preferences to segment the market into homogeneous groups and adjust the design or offer to the segment that enables the co-creation behaviour of greatest value (Cossio-Silva et al., 2013; Dahle et al., 2023). The segmentation process and the selection of the target audience are among the most studied activities in the academic literature and the most applied solutions by practitioners (DeSarbo & Grisaffe, 1998; Wedel & Kamakura, 2000). In fact, Smith (1956, p.6) used these terms in his definition of market segmentation: “Market segmentation, on the other hand, consists of viewing a heterogeneous market (one characterized by divergent demand) as a number of smaller homogeneous markets in response to differing product preferences among important market segments.”

Consumer grouping techniques are often used to explore possible segments. The most widely used is undoubtedly clustering (Wedel & Kamakura, 2000), a technique that begins with the collection of data (usually through surveys) on certain criteria, such as demographic attributes, purchasing habits and consumer preferences. Using multivariate techniques, these are then clustered into groups that are as homogeneous as possible around centroids, while maintaining sufficient distance to ensure that they can be considered distinct groups (Wedel & Kamakura, 2000). However, the Internet environment offers the possibility of collecting enormous amounts of unstructured data, which poses a challenge to clustering algorithms that need to analyse and cluster increasingly larger and noisier databases (Ali et al., 2023).

Several different techniques and methodologies are used for consumer clustering. Milligan and Cooper (1988) consider the five dominant ones to be Forgy’s method,

Jancey's method, MacQueen's method (K-means), the convergence method and the Exchange algorithm. However, it should be noted that not all of these can work with large databases. In fact, they are usually classified into hierarchical (either top-down (divisive) or bottom-up (agglomerative), and partitional, the latter forming clusters by partitioning the data, but without imposing a hierarchical structure (Jain et al., 1999). Consequently, partitional techniques are the most suitable for clustering large datasets. Undoubtedly, the most widely used is K-means because of its conceptual simplicity, versatility, and ease of implementation (Jain et al., 1999; Jain, 2010). However, the increasing computational capacity of data processing systems has raised the possibility of using different algorithms, such as supervised learning algorithms (Tukey, 1962; Kaufman & Rousseeuw, 2009). One such method is XGBoost. Proposed by Chen and Guestrin (2016), this open-source software library is not specifically a clustering algorithm, but it is characterised by working with defined labels and can be used to group subjects with labelled data (Liang et al., 2019). It recently attracted much attention when it dominated Kaggle machine learning competitions due to its speed of execution and performance (Poongodi et al., 2022). Provided the algorithm fits well with the data provided, it can generate valuable information to help entrepreneurs to explore their business ideas and improve their offer for target segments (Dahle et al., 2018). Segmentation theory argues that the market is made up of consumers with different needs, and knowledge of its heterogeneous nature and the ability to extract profiles of similar consumers is an essential aspect of the design of any offer or market strategy (Cossío-Silva et al., 2013). Since there is no universal method that works with all applications and databases, any proposal for a new method would require thorough comparison with other methods to find the best fit with the analysed dataset. As this would consume a lot of time and resources, comparative studies of a limited number of algorithms are usually performed (Fernández-Delgado et al., 2014; Hartmann et al., 2019).

Comparative studies of classification methods are far more common in the computer science literature than in market research literature. The few examples of the latter include Hartmann et al. (2019) and Liu et al. (2010). However, there are no references on the best methods for grouping databases generated by lead capturing start-ups (Sáez-Ortuño et al., 2023a). Lead capture is the set of actions focused on obtaining contact details from potential customers (individuals or companies) via the Internet to nurture databases to nurture databases that can be used in-house or sold on to other companies (Sáez-Ortuño et al., 2023b).

This research attempts to fill this gap by comparing the adequacy of the supervised XGBoost algorithm for classifying lead data into groups of suitable consumers for value propositions (Amit & Zott, 2012; Dahle et al., 2018), with respect to the more popular K-Means system, although the latter is unsupervised. In other words, this paper studies how well these two automated cluster algorithms aggregate a database of over 5 million Spanish leads to explore commercial opportunities. The data comes from users who registered to participate in sweepstakes and online tests, and is provided by the start-up CoRegistros, S.L.U.

The rest of the document is organized as follows. First, a conceptual framework focusing on cluster algorithmics in marketing is presented. Second, the research methods and results are described. Following a discussion of those results, the impli-

cations for academia and management are addressed. The study concludes by summarising the key themes that emerged from the results, discussing their limitations, and suggesting avenues for future research.

## Theoretical framework

The digital ecosystem, made possible by the Internet, has become a complex web of social relations through which abundant information of interest to both consumers and companies circulates. In order to develop successful projects, marketing experts and entrepreneurs need to know their way around this labyrinth and be able to identify the sources of information that generate the greatest impact on consumers (González-Padilla et al., 2023). Given that segmentation theory proposes that one of the basic objectives of marketing is to design and promote specific products and services for a target audience, the only way to achieve this is to extract patterns of customers that can be used to classify them and, thus, to study and try to understand their needs. When this is done digitally, it is called digital marketing (Bala & Verma, 2018).

### Supervised and unsupervised methods for clustering consumers into segments

According to Wedel and Kamakura (2000), there are more than 50 data clustering methods that could be used for market segmentation, although most of these were developed in fields outside marketing. One of the earliest and most popular examples was ISODATA, proposed by Ball and Hall (1967), which was widely used in geoscience applications. This very practical method for clustering multivariate data was used to find patterns in complex interactions, resulting in a set of cluster centroids that tend to minimise the sum of the squared distances of each piece of information from the nearest centroid (Memarsadeghi et al., 2007).

However, to process large databases, these mathematical models describing a function must be associated with a particular learning algorithm (Sathya & Abraham, 2013) that helps to generate efficient heuristic methods (Liu et al., 2010). For example, the computational burden involved in obtaining clusters based on hierarchical shape trees, and the biases associated with centroid selection, recommend the use of non-hierarchical methods (Wedel & Kamakura, 2000) or supervised machine learning algorithms (Chen & Guestrin, 2016).

Machine learning algorithms are often classified as supervised and unsupervised (Memarsadeghi et al., 2007). Supervised learning requires labelled input and output data during the training phase, and often, since most available data is usually raw (unlabelled), it is generally labelled by the researcher or some expert in the domain. Thus, in order to be trained, an algorithm requires a feature vector (or instance) describing the event/object and a label indicating the type of output generated (Zhou, 2018).

In contrast, unsupervised learning refers to the ability to train the model with raw, unlabelled data and without a ground truth to evaluate the possible solution (Sathya & Abraham, 2013). Although at first glance the algorithm's lack of direction might make the procedure seem uncontrolled, this can sometimes be advantageous because

patterns can be found that were not previously considered (Kohonen et al., 1996). However, although clustering algorithms do not require as much human intervention, they do require the researcher to set the parameters of the model, such as the number of cluster groups (Liu et al., 2010).

It should be emphasised that a precondition for the implementation of algorithms is the availability of quality databases. In general, data taken from the Internet are often unstructured, in free format, such as in text form, which can make them very difficult to manage (Hartmann et al., 2019). However, the information collected through leads is fully structured because the consumers have filled in specific fields on a form. But they are still very noisy due to the amount of false data provided, hence a screening process is required that can often be costly in both time and effort (Sáez-Ortuño et al., 2023b).

Although numerous algorithms can be used to group data into clusters, the literature has not been able to find any one technique that generally dominates over the rest (Arabie et al., 1996; Boone & Roehm, 2002; Hartmann et al., 2019; Wedel & Kamakura, 2000). For example, Vriens et al. (1996) compared nine segmentation methods from conjoint metrics using a Monte Carlo study and found that the differences in predictive accuracy were small. That is, each method has its own strengths and limitations (Dayan et al., 2021), and depending on the information that a database contains, one technique might perform better than another.

### **Comparative analysis between K-means and XGBoost**

This study tests two algorithms, K-Means and XGBoost, that may take different conceptual approaches but are both of major relevance to market research as segmentation tools. Their performance has already been widely proven in other disciplines, such as in the development of tools to detect anomalous behaviour in unknown or potential security attacks (Henriques et al., 2020), and the results obtained suggest that both K-Means and XGBoost are among the highest performing methods due to their versatile structures (Ibrahim & Abdulazeez, 2021; Liu et al., 2021). However, given the methodological diversity of the two algorithms used in this study and the commercial goal pursued by clustering, the two groups are expected to be different, with relatively little correlation between them.

However, as Hartmann et al. (2019) point out, there is no such thing as a free lunch, in the sense that each algorithm has its own characteristics and, therefore, the choice will depend on the particularities of the dataset and the specific needs of each problem.

Thus, the following research question is proposed:

RQ1. Which of the two algorithms, supervised or unsupervised, will achieve the best groupings of consumers captured through leads to achieve the commercial objectives?

## Unsupervised and non-hierarchical machine learning techniques: the K-means algorithm

The K-means algorithm is the most popular unsupervised, non-hierarchical clustering machine learning technique for dividing a dataset into  $k$  clusters (or groups) of similar cases. It was proposed by MacQueen (1967) and has been widely used in many different applications, such as comparative studies (Boone & Roehm, 2002), due to its simplicity and efficiency (Kuo et al., 2002). However, it requires the number of clusters to be specified a priori, which may lead to suboptimal results if the data have complex shapes or if outliers are present.

It is an iterative algorithm, represented by the function  $J$ , which aims to minimize in each iteration the within-cluster variance, or the quadratic error function for all points and for each cluster (see Eq. 1).

$$J = \sum_m \sum_{i=1}^{k=1} w_{ik} |x_i - \mu_k|^2$$

where  $w_{ik}$  equals 1 if point  $x_i$  belongs to cluster  $k$ , and 0 in any other case and  $\mu_k$  is the centroid for cluster  $k$ . The K-means algorithm works by randomly assigning  $k$  initial centroids, and then assigning each data point to the nearest cluster based on its Euclidean distance. In a second iteration, the centroids are recalculated as the mean of the data points assigned to the cluster, and the data points are re-assigned. This process is repeated until the centroids no longer change or until a certain number of iterations is reached (Lloyd, 1982). The value of  $k$  can be determined using the elbow method (Syakur et al., 2018).

The K-means algorithm is fast and easy to implement as it does not require a model training phase, and it is assumed that the clusters are in the shape of a circle, which can be a drawback as it may not work well for clusters of other shapes. However, as noted above, there are no conclusively dominant techniques in this field as the performance of the algorithm varies depending on the database to be clustered (Arabie et al., 1996; Boone & Roehm, 2002; Wedel & Kamakura, 2000).

## Supervised and hierarchical machine learning techniques: the XGBoost algorithm

Supervised machine learning algorithms are often used for clustering (Mitchell & Frank, 2017; Gultom et al., 2018). One of the most popular, albeit outside the marketing literature, is eXtreme Gradient Boosting (XGBoost), which is based on gradient boosting decision trees, and was developed for the sole purpose of improving model performance and computational speed (Liu et al., 2021). This algorithm can be used for both classification and estimation by regression (Chen & Guestrin, 2016), and has already been applied to find solutions in the marketing field (Liang et al., 2019).

The algorithm works by creating a set of decision trees, but instead of averaging independent trees, it builds them sequentially (as what are known as Classification and Regression Trees (CART) (Ahsan et al., 2021), where trees are created by learning and using the prediction errors or residuals of the previous tree model until the

error can no longer be corrected (what is known as “gradient downward” (Basu et al., 2022; Hastie et al., 2009)).

The algorithm begins by constructing a decision tree in which each node is split into sub-nodes based on a specific feature and assigned a score, as well as a pruning threshold to predict the target variable. A new tree is then generated that attempts to reduce the error made in the previous step. The initial tree is combined with the second one and a new tree is generated where the mean square error will be smaller than that of the initial tree (Liang et al., 2019).

The objective function (Eq. 2) that the XGBoost algorithm minimizes in each iteration is as follows:

$$J^{(t)} = \sum_n^{i=1} j \left( y_i, y_i^{(t-1)} + f_t(x_i) \right) + \Omega(f_t)$$

where  $y_i$  is the target label of point  $x_i$  known from the dataset and  $y_i'$  is the predicted label. We can observe that the objective function  $J$  of the XGBoost algorithm is a function of functions  $j$ , which in turn is a differentiable convex loss function that measures the difference between the prediction  $y_i'$  and the target  $y_i$ . The regularized term  $\Omega(f_t)$  penalizes the complexity of the model (set of trees) and helps to smooth the learned final weights to avoid overfitting. It tends to intuitively select a model using simple and predictive functions. To optimize this function of functions, this needs to be done iteratively. Therefore, we must calculate the function  $J^{(t)}$  at iteration  $t$  from the prediction of labels  $y_i^{(t-1)}$  in the previous iteration ( $t-1$ ) and greedily add the tree  $f_t(x_i)$  to the model in such a way that it improves it.

Although the XGBoost algorithm can process a large volume of data with multiple features, it also has some drawbacks as it consumes a lot of computational resources when working with large databases. Hence, it is advisable to define the most relevant variables and only use these in the construction of the model. Also, it only works with numerical vectors and the parameters of the algorithm (e.g., the extent of the tree) need to be determined by the researcher (Chen & Guestrin, 2016; Liang et al., 2019).

## Overview of the empirical study

One of the essential phases of an entrepreneurship process is the exploration of the business idea (Kraus et al., 2020). This implies the need to understand the heterogeneity of consumers and their needs (Dahle et al., 2018), and involves determination of the target consumer profile and matching of the offer to their needs, while also creating value for the company (Liu et al., 2010).

Therefore, business objectives are more effectively achieved if the right instrument can be found that can best group consumers into homogeneous sets based on the commercial offer. However, it is not only the instrument that is important, but also its fit with the database. In other words, context matters, and it is essential to test several algorithms to determine the most appropriate one for any given context (Hartmann et al., 2023).



To address our research question, the two proposed algorithms, K-means and XGBoost, will be compared in terms of their effectiveness at achieving the required objectives and performance in the process of clustering a large database into homogeneous segments with commercial value. Specifically, the information was on 5,185,857 participants in online sweepstakes and contests in Spain and was gathered by a lead generation start up between 2010 and 2022. There is a long tradition of comparison methods in marketing research, which have set standards for comparable empirical research (Bremer et al., 2017; Hartmann et al., 2023).

## Data set

The database was obtained after signing a confidentiality agreement with the company CoRegistros, S.L.U., and contains registered information on more than 5 million consumers collected over twelve years (2010–2022). According to the company's own information, the data was recorded by participants in online sweepstakes (96%) and in self-assessment questionnaires on topics such as intelligence, geography, and cooking, among others (4%), as a condition for receiving a prize or seeing the result of the questionnaire. The sample, which consists of 10.7% of the Spanish population, is somewhat asymmetric in structure with respect to stated sex (65% female vs. 35% male) and age (Generation Z, 0.2%; Millennials, 40%; Generation X, 52%; Baby Boomers, 7%; Silent Generation, 0.8%). In addition, as false information is frequently recorded, in order to reduce noise as much as possible, the data were screened and cross-checked with official databases and, as a result, about 6% of registered users were removed (for further information on this process, see Sáez-Ortuño et al., 2023b).

The data matrix contains 37 fields (columns), and these are grouped into five blocks: (1) Users, (2) Marketing, (3) Conversions, (4) Ads, and (5) Sweepstakes. The user block contains descriptive data about the consumers. The marketing block describes how the user provided their information. The conversion block contains the history of users who have become buyers by purchasing a subsequently promoted product. The campaigns block contains the marketing actions in which the user has participated. Finally, the sweepstakes block shows information linked to the sweepstakes column in the user table in the form of the variable *id\_prom*, and it is also linked to the marketing table through *id\_prom*, since each marketing campaign is assigned a sweepstake (the same sweepstake may be assigned to different marketing campaigns). Finally, the variable that identifies the user is *id\_user*. Table 1 describes the items that correspond to each block.

To apply the clustering algorithm, the users block was used as the base, which corresponds to the data to be clustered, while the remaining blocks represent the dimensions. After analysing the supplied database, we decided it would be useful to have a description of the different types of marketing campaigns that had been carried out, as well as the type of sweepstake/prize and more information about conversions into purchases. That is, information on the objectives pursued by the grouping of consumers. Hence, new information was requested from the company, which was delivered in three more blocks: (1) *ads\_tipo.csv*, (2) *clasificacion\_sorteos.csv*, and (3) *clasificacion\_conversions.csv*. Table 2 shows the items contained in the blocks.



**Table 1** Description of the tables provided by the company for the study

Table name in the database	Description of the content of the database table
<b>1. users</b>	Master table of users. Contains all fields with descriptive information about the user.
<b>2. marketing</b>	Master table of marketing campaigns through which users are registered. It relates to the users table through the <i>id_m</i> field.
<b>3. conversions</b>	Master table of conversions. Contains the historical data of users who have converted to a product in the past. It relates to the users table through the <i>id_user</i> field.
<b>4. ads</b>	Master table of client campaigns. These campaigns are sent to users who are registered in the database with the aim of converting them to the offered product. It relates to the conversions table through the <i>id_ad</i> field.
<b>5. sweepstakes</b>	Master table of sweepstakes. It relates to the Sweepstake column in the users table through the <i>id_prom</i> column. It also relates to the marketing table through the <i>id_prom</i> field since each marketing campaign is assigned a sweepstake (the same sweepstake can be assigned to different marketing campaigns).

**Table 2** Description of the auxiliary tables provided by the company for the study

Name of file	Description of block content
<b>1. ads_type.csv</b>	After analyzing the ads table, the need for a description of different campaign types was identified. To solve this issue, the <i>ads_type</i> file was created as a master of campaign descriptions (with a tab as a separator). This file is related to the ads table through <i>ad_type</i> .
<b>2. clasification_sweepstakes.csv</b>	After analyzing the sweepstake table, the need to classify sweepstakes according to the prize was identified. To solve this issue, the <i>clasification_sweepstakes</i> file was created (with a tab as a separator). This file is related to the sweepstake table through <i>id_prom</i> . The created categories are: beauty, content, electronics, home, iPhone, leisure, test, and travel.
<b>3. clasification_conversions.csv</b>	After analyzing the conversions table, the need to classify the clients' campaigns ( <i>id_ad</i> ) that appear in that table (i.e., campaigns that have resulted in at least one conversion) according to the final product to which each user converted was identified. To solve this issue, the <i>clasification_conversions</i> file was created. This file is related to the conversions and ads tables through <i>id_ad</i> . The created categories are: hearing aids, energy, finance, games, NGO, insurance, and telcos.

Finally, to complete the data, information on some external variables was sought from public sources. For example, postcodes were used to incorporate the geographic longitude and latitude in the database. With all this information, the company was asked to perform an Extract-Transform-Load (ETL) to transform the table to its final format before applying the algorithms. ETL is a process for moving data from different sources to a target file. It involves three stages: (i) extraction of data from different sources (in our study, official public sources), (ii) cleaning and transformation of the data (filtering, sorting, conversion into appropriate format and aggregation), (iii) loading into the corresponding database (Vassiliadis, 2009). Finally, the variables to be considered in the models were selected and transformed into Boolean logic. Table 3 shows the list of variables, the type of variable (string, Boolean, and interval) and the variables used in the comparative study, which are marked with an X.

**Table 3** List of final columns of the users table

Index	Column	Type	K-Means	XGB
1	<b>producto_conv</b>	String	-	-
2	<b>id_producto_conv</b>	Int (*)	-	X
3	<b>id_user</b>	Int (*)	X	X
4	<b>email</b>	String	-	-
5	<b>dominio_email</b>	String	-	-
6	<b>id_dominio_email</b>	Int (*)	-	-
7	<b>sexo</b>	String	-	-
8	<b>id_sexo</b>	Bool	X	X
9	<b>nombre</b>	String	-	-
10	<b>edad</b>	Int	X	X
11	<b>codigopostal</b>	String	-	-
12	<b>latitute</b>	Float	X	X
13	<b>longitute</b>	Float	X	X
14	<b>telefono</b>	Int (*)	-	-
15	<b>comp_telf</b>	String	-	-
16	<b>grupo_comp_telf</b>	String	-	-
17	<b>valido</b>	Bool	X	X
18	<b>finaliza</b>	Bool	X	X
19	<b>espactividad</b>	Bool	X	X
20	<b>estado_telf</b>	Bool	X	X
21	<b>cla_sorteo</b>	String	-	-
22	<b>id_cla_sorteo</b>	Int (*)	-	-
23	<b>dominio_email_gmail</b>	Bool	X	X
24	<b>dominio_email_hotmail</b>	Bool	X	X
25	<b>dominio_email_outlook</b>	Bool	X	X
26	<b>dominio_email_yahoo</b>	Bool	X	X
27	<b>dominio_email_live</b>	Bool	X	X
28	<b>dominio_email_msn</b>	Bool	X	X
29	<b>dominio_email_otros</b>	Bool	X	X
30	<b>cla_sorteo_belleza</b>	Bool	X	X
31	<b>cla_sorteo_contenido</b>	Bool	X	X
32	<b>cla_sorteo_electronica</b>	Bool	X	X
33	<b>cla_sorteo_hogar</b>	Bool	X	X
34	<b>cla_sorteo_iphone</b>	Bool	X	X
35	<b>cla_sorteo_ocio</b>	Bool	X	X
36	<b>cla_sorteo_test</b>	Bool	X	X
37	<b>cla_sorteo_viajes</b>	Bool	X	X

### Methodological study of the unsupervised algorithm: K-means

To test the unsupervised K-means algorithm with the sample data, the following steps were followed: (1) selection of the data set, (2) standardization of the data (mean=0 and variance=1), (3) selection of centroid, (4) application of the algorithm, and (5) validation and estimation of the effectiveness and efficiency of the algorithm.

First, 24 variables (2 int, 2 float, and 20 Boolean) were selected from the set of 37. Although most of the variables are Boolean, they were standardized using the Python StandarScaler library (Zamri et al., 2022). This is a recommended procedure

in clustering algorithms for segmentation processes (Milligan & Cooper, 1988; Stead et al., 2007). Although there is no need for binary Boolean variables, in this case there were four non-Boolean variables (Chakraborty et al., 2009). The combination of binary variables with scales or ratios is not recommended, as one of them might contain higher variances than the others, and may erroneously dominate them (Stead et al., 2007).

Application of the K-means algorithm necessarily requires prior definition of the number of target groups or clusters, represented by the  $k$  variable. Based on the study by Kodinariya and Makwana (2013), it was considered that the number of groups should be related to the users collected by the lead who had become buyers of some product ( $id\_producto\_conv \neq 0$ ). Then, to determine the number of clusters, the elbow rule was applied, and five groups were considered (Likas et al., 2003). Figure 1 illustrates the process, and the elbow would be the parabola curve that is generated by the crossing of the two lines (solid and dashed). While the solid line captures the sum of clustering errors as the number of clusters increases, the dashed line captures the time complexity of finding the optimal point. The intersection helps to find the approximate point of 5 (Sujatha & Sona, 2013).

Once the number of centroids had been selected, the K-means algorithm from the Python library was applied, which works as follows: (1) The  $k$  centroids are initialized at random coordinates; (2) The distance of the users to each centroid is calculated, and they are grouped around the nearest centroid based on the minimum distance between the points and the centroid; (3) The centroids are updated, recalculating their new position, and steps (2) and (3) are repeated; (4) The process is stopped when the centroids no longer change (Likas et al., 2003).

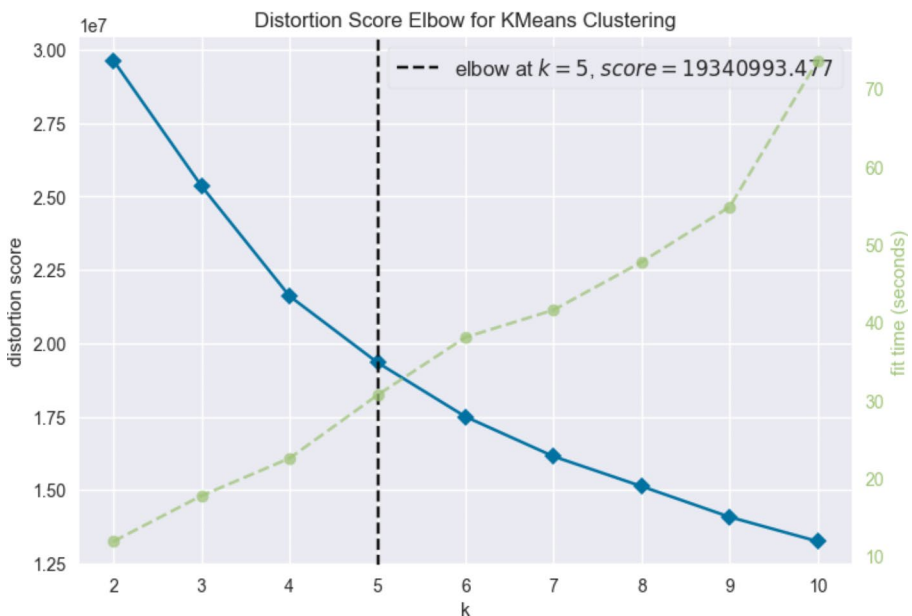


Fig. 1 Elbow method applied to a reduced sample of the dataset (1,000,000 users)

The results were then analysed for validity and effectiveness. Since labels were available, we were able to calculate accuracy to determine whether K-means had “matched” the labels of the target class. In general, these groups will be considered good if  $k$  has been chosen correctly.

## Results

The results obtained by applying K-Means suggest that this algorithm was not able to generate commercially valuable segments from the dataset used (Kamthania et al., 2018). The algorithm generated five rather heterogeneous groups, with overlapping attributes (e.g., the Insurance variable participates in all five groups and, in four of them, is the dominant one) and which are also highly unbalanced. Table 4; Fig. 2 show the weight of conversions into buyers of the sponsored products in each of the clusters (Jain et al., 1999). While games dominate in cluster 1, insurance dominates in all other clusters. In other words, it is not possible to determine which of the generated segments would be the most suitable for the promotion of the sponsored products.

Figures 3 and 4 graphically represent the clusters in two dimensions, without highlighting product conversions and highlighting product conversions (Strehl & Ghosh, 2003). The representation of clusters in 2D is a technique used to visualise the distribution of data in which each point represents an observation and is coloured according to the cluster to which it belongs (Strehl & Ghosh, 2003). It was noted that the 2D representation may not fully capture the structure of the data in a high-dimensional space.

The K-Means algorithm is an unsupervised learning clustering method, which means that it learns based on variances in the data (Lloyd, 1982). Since the target audience of this analysis is consumers who become buyers of the promoted products, such variance is not associated with conversions (MacQueen, 1967). So, although the clusters are found to be informative, the solution is not entirely satisfactory for commercial purposes (Alonso-González et al., 2020; Jain, 2010). Finally, to corroborate this result, the K-means algorithm was run again with the same data, and the results changed significantly, generating inconsistent and unstable results, corroborating the unsuitability of the algorithm for that dataset (Murray et al., 2017).

## Methodological study of the supervised algorithm: XGBoost

To apply the XGBoost algorithm, a supervised learning method based on decision trees (Chen & Guestrin, 2016), the following steps were followed: (1) selection of the dataset; (2) application of the algorithm to train and fit the model; (3) selection of hyperparameters; (4) application of the algorithm and evaluation of the obtained performance; (5) visualization of the results (through graphics such as the learning curve and predictor variable importance); and (6) cross-validation.

**Table 4** Characteristics of each cluster

Cluster	Number of users	Average age	Product	N° Conversions	Conversions rate (%)
1	2,760,626	44.45	Hearing aids	732	7.42
			Energy	50	0.51
			Finance	103	1.04
			<b>Games</b>	<b>4729</b>	<b>47.9</b>
			NGO	173	1.75
			Insurance	4068	41.2
			Telcos	8	0.08
2	2,158,265	45.23	Hearing aids	2740	21
			Energy	258	1.98
			Finance	2	0.01
			Games	503	3.86
			NGO	630	4.84
			<b>Insurance</b>	<b>8643</b>	<b>66.4</b>
			Telcos	237	1.82
3	174,126	42.38	Hearing aids	19	11.6
			Energy	1	0.61
			Finance	5	3.07
			Games	3	1.84
			NGO	1	0.61
			<b>Insurance</b>	<b>132</b>	<b>81</b>
			Telcos	2	1.22
4	229,770	52.13	Hearing aids	93	11.5
			Energy	10	1.24
			Finance	9	1.11
			Games	212	26.2
			NGO	21	2.59
			<b>Insurance</b>	<b>458</b>	<b>56.6</b>
			Telcos	6	0.74
5	41,356	46.18	Hearing aids	12	11
			Energy	0	0
			Finance	1	0.91
			Games	5	4.59
			NGO	5	4.59
			<b>Insurance</b>	<b>86</b>	<b>78.9</b>
			Telcos	0	0

(1) First, 25 variables (3 int, 2 float, and 20 Boolean) were selected from the set of 37 vectors. Before applying the algorithm, the dataset had to be divided into several subsets as shown in Table 5.

(2) Next, the algorithm was applied to train and fit the model. This second stage should involve two parts: the training of the algorithm using the  $X_{train}$  function, and the evaluation of its performance using the  $X_{test}$ , which estimated the level of confidence.  $X_{predict}$  was then applied. For the training process, only

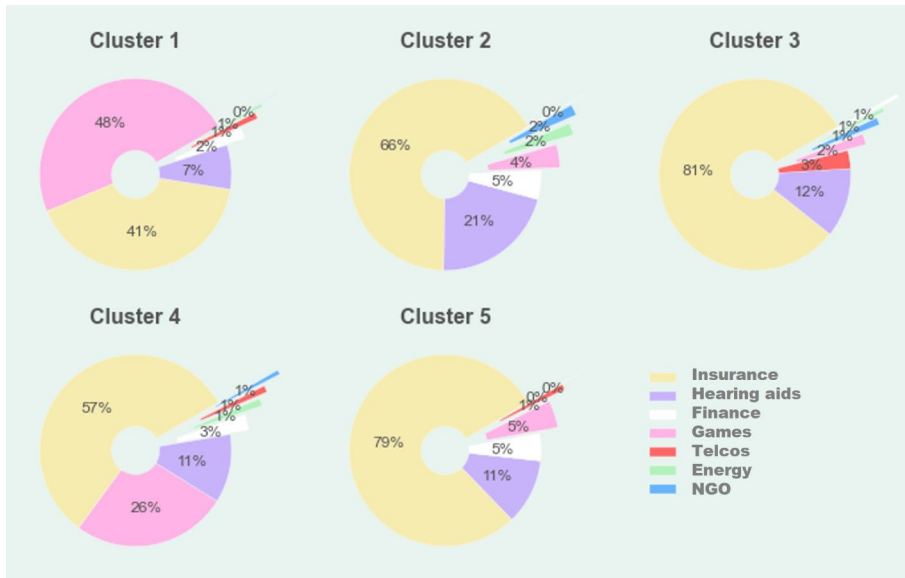


Fig. 2 Distribution of conversions by cluster

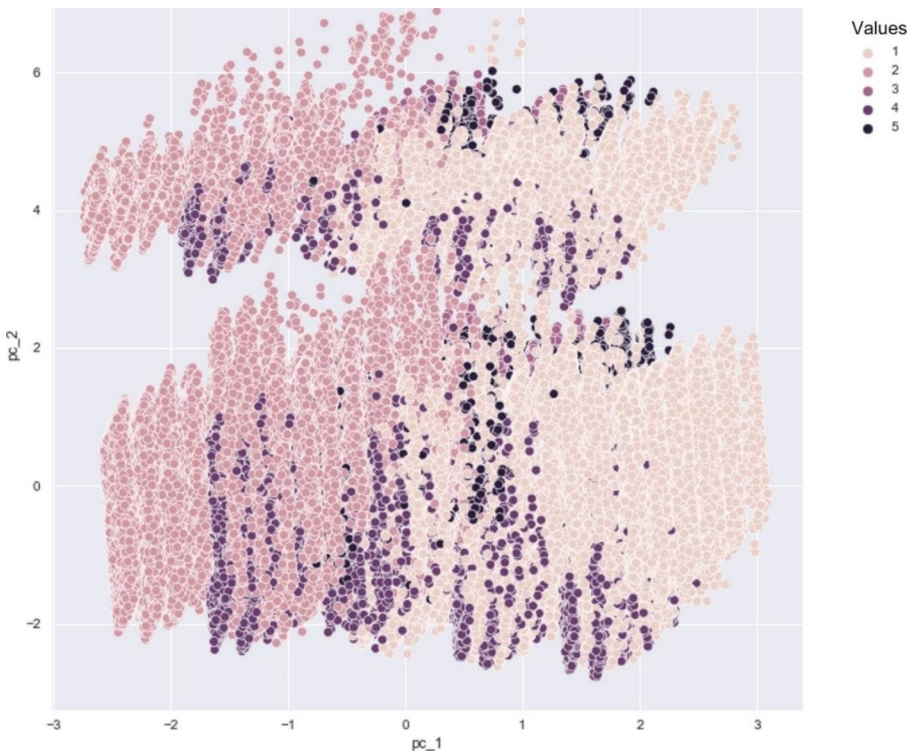


Fig. 3 Distribution of conversions by cluster

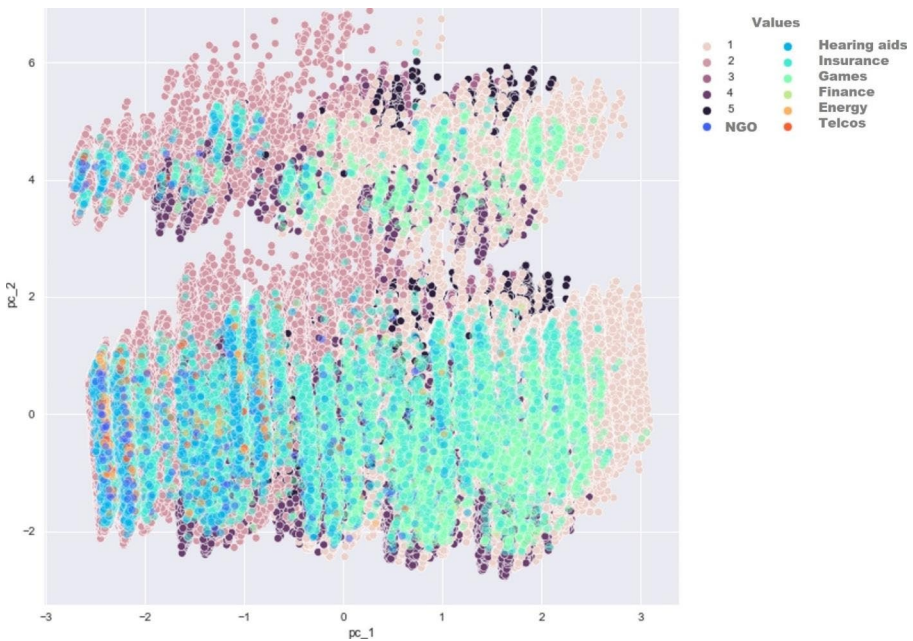


Fig. 4 Representation of the clusters in 2D highlighting the users with conversions

Table 5 Subsets of data

Subsets of data.	Description	Size
<b>X</b>	Users with conversions. The variables in this matrix are those indicated in Sect. 3.6 Final Structure, with the exception of <code>id_producto_conv</code> and <code>id_user</code> .	$[25.612 \times 23]$
<b>Y</b>	<code>id_producto_conv</code> corresponds to the users of <b>X</b> .	$[25.612 \times 1]$
<b>X_train</b>	90% of users with conversions.	$[23.050 \times 23]$
<b>y_train</b>	<code>id_producto_conv</code> corresponds to the users of <b>X_train</b> .	$[23.050 \times 1]$
<b>X_test</b>	10% of users with conversions.	$[2.562 \times 23]$
<b>y_test</b>	<code>id_producto_conv</code> corresponds to the users of <b>X_test</b> .	$[2.562 \times 1]$
<b>X_predict</b>	Users without conversions.	$[5.160.245 \times 23]$
<b>y_predict</b>	A value that is unknown at the beginning of the study ( <code>id_producto_conv=0</code> ) and will be predicted after applying this algorithm	$[5.160.245 \times 1]$

users who had previously become buyers of one of the promoted products were considered, that is, users with conversions (Bishop & Nasrabadi, 2006). Hence, the training data needed to contain some information about the correct response or about the target variable of the study (Hastie et al., 2009). Thus, the learning algorithm found patterns in *X\_train*, assigned the input data attributes to the target (*Y\_train*), and generated a Machine learning (ML) model that captured those patterns (Jordan & Mitchell, 2015). Figure 1 shows the structure of the training.



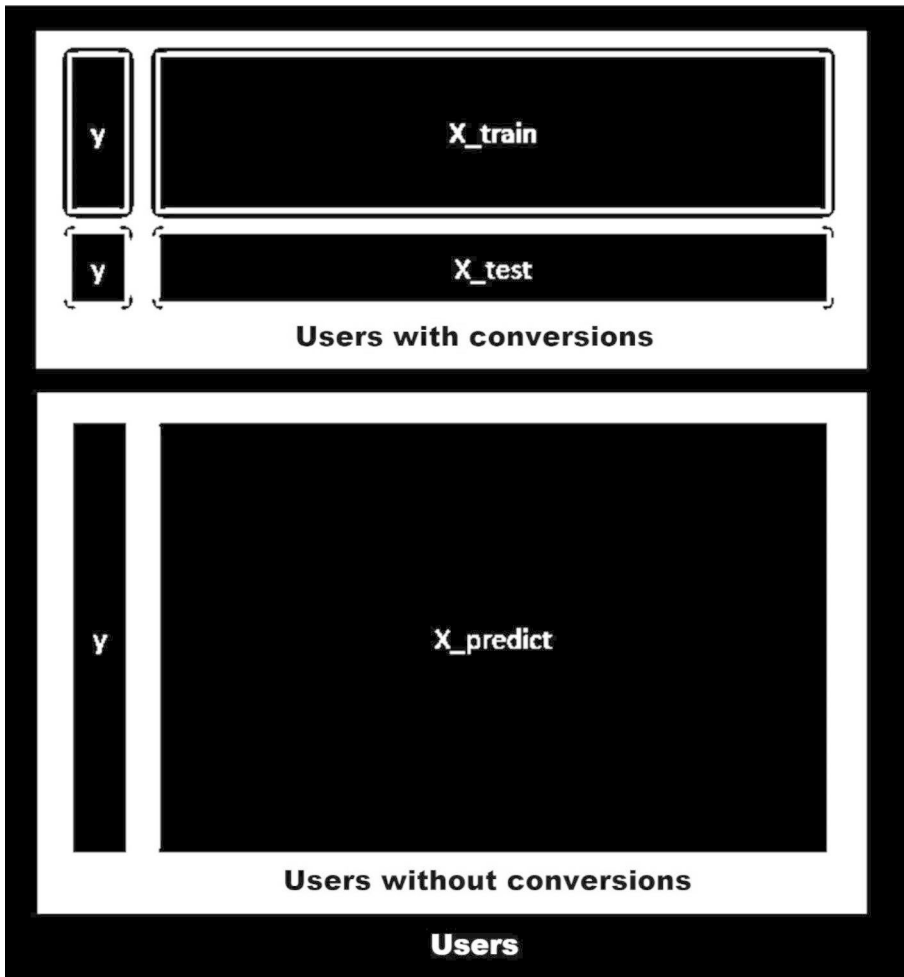


Fig. 5 Schema of the different subsets of data

(3) Subsequently, the hyperparameters were defined, which were the number of iterations ( $n\_estimators=100$ ) and the maximum depth of each tree ( $max\_depth=8$ ). To do this, several permutations were analysed ( $n\_estimators=50, 100, 200, 500, 750, 1,000$  and  $max\_depth=4, 6, 8, 10, 15, 20$ ). In addition, given that the frequencies of the products from the conversions were imbalanced, the weights of the parameters that weighted each group had to be adjusted. To calculate these weights, the Python library *class\_weight* was used, which performed the following calculation internally:

$$product\ weight_i = \frac{y_{train}^n}{number\ of\ products \cdot Frequency\ of\ product_i \in y_{train}^n}$$

(4) Afterwards, the algorithm was applied, and its performance evaluated. After adjusting the parameters and training the model with  $X\_train$ , the resulting algorithm was applied to  $X\_test$  and the values of  $y\_test$  were predicted in the form of prob-

**Table 6** Extraction of the probability table along with the true value of  $y_{test}$ 

Index	id_user	Hearing aids	Energy	Finance	Games	NGO	Insurance	Telcos	$y_{test}$
1	11,188,636	0.0000	0.0002	0.0001	<b>0.9635</b>	0.0000	<b>0.0356</b>	0.0003	Games
2	13,810,831	0.0003	0.0043	0.0000	0.0054	<b>0.1140</b>	<b>0.8605</b>	0.0152	Insurance
3	17,242,446	<b>0.4853</b>	0.0058	0.0000	0.0024	0.0108	<b>0.4919</b>	0.0035	Hearing aids
4	17,242,871	<b>0.7443</b>	0.0004	0.0000	0.0022	0.0390	<b>0.2135</b>	0.0003	Hearing aids

**Table 7** Conversion frequencies and accuracy percentages by product

Product	Frequencies totals	Frequencies $y_{train}$	Frequencies $y_{test}$	Correct predictions rate (%)
(1) Hearing aids	5,363	4,848	515	61.74
(2) Energy	318	286	32	3.12
(3) Finance	116	102	14	21.42
(4) Games	5,447	4,885	562	79.89
(5) NGO	831	757	74	0.00
<b>(6) Insurance</b>	13,286	11,938	1,348	73.88
(7) Telcos	251	234	17	0.00

**Table 8** Confusion matrix

Product	Hearing aids	Energy	Finance	Games	NGO	Insurance	Telcos	Totals
<b>Hearing aids</b>	<b>318</b>	0	0	12	1	184	0	<b>515</b>
<b>Energy</b>	4	<b>1</b>	0	1	0	26	0	<b>32</b>
<b>Finance</b>	0	0	<b>3</b>	10	0	1	0	<b>14</b>
<b>Games</b>	16	0	1	<b>449</b>	0	96	0	<b>562</b>
<b>NGO</b>	28	0	0	3	<b>0</b>	43	0	<b>74</b>
<b>Insurance</b>	232	1	0	113	5	<b>996</b>	1	<b>1,348</b>
<b>Telcos</b>	1	0	0	0	0	16	<b>0</b>	<b>17</b>
<b>Totals</b>	599	2	4	588	6	1,362	1	<b>2,562</b>

abilities of a user becoming a buyer of a given product (Ravikumar et al., 2010). Moreover, as the last  $y_{test}$  column in Table 6 shows, the product with the highest estimated probability is assigned (Kohavi & Provost, 1998). Sometimes the estimates clearly point to a consumer preference ( $id\_user$  11,183,636 when choosing Games) or ( $id\_user$  13,810,831 when choosing Insurance), but in others, when assigning similar probabilities, it is not so clear, ( $id\_user$  17,242,446, between Hearing Aids and Insurance).

To determine the degree of fit of the resulting predictions, certain goodness-of-fit metrics were calculated: the percentage of correct predictions of the  $y_{test}$  values, which was estimated at approximately 70%; and the percentage of correct predictions per product, where there is major variability from 0 to about 80% (see Table 7). To provide more detail, the confusion matrix was extracted, where in addition to displaying the hits (collected on the diagonal), it also shows the errors when confusing one class with another (Murray et al., 2017), which means different types of error can be worked on separately (Kohavi & Provost, 1998). As shown in Table 7, each

column represents the number of predictions made for each product, while each row shows their correspondence with the actual class. For example, in the Hearing Aids category, 318 records were hits and the remaining 197 (12+1+184) were misses.

Once the parameters had been adjusted and the confidence of the defined algorithm was known, the algorithm was trained again, but this time with 100% of the records with conversions. Once trained, it was applied to  $X_{predict}$  (the users who had not converted into purchases) to make an estimate of the types of products most likely to convert into purchases for each user and, in addition, to select the two most important ones. All these results are collected in the  $y_{predict}$  matrix. Consequently, once the XGBoost algorithm had been trained to cluster users who became buyers of specific products, probabilities of becoming buyers were assigned to users who did not become buyers and they are grouped according to their probability of becoming buyers of specific products (Hearing Aids, Energy, Finance, Games, NGO, Insurance, Telcos). In other words, the clustering technique sought to identify groups of users with significant differences in terms of the products they purchase, with the aim of maximising the probability of them becoming buyers.

(5) The results were then represented by the learning curve and the degree of importance of predictor variables. The learning curve shows how the accuracy of a machine learning model improves as the size of the training dataset increases. Meanwhile, the importance of predictor variables refers to how much they influence the outcome of the model (Tufté, 2001).

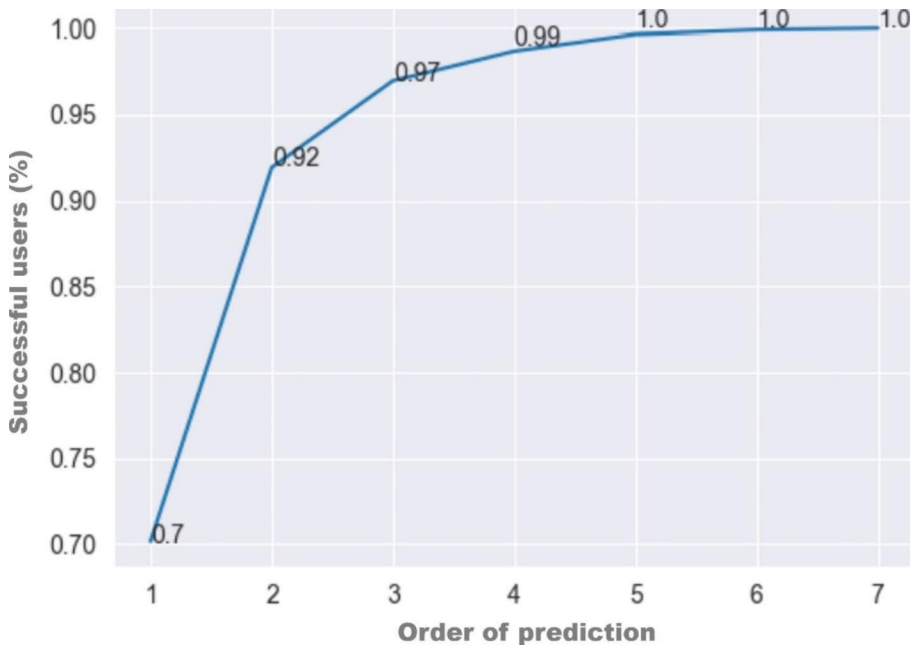
(6) Cross-validation was used as a criterion for the experiment. Specifically, this was ten-fold cross-validation, which means that the model worked with 90% of the records of users who became buyers, and from the model fitting, the behaviour of the remaining 10% of users was predicted (Kohavi, & Provost, 1998).

## Results

The results obtained by the supervised algorithm, XGBoost, after being trained, are collected in the following output  $y_{predict}$  (the prediction of  $id_{product\_conv}$ ) of those users who had not become buyers of any sponsored product or service in the past (Chen & Guestrin, 2016). It also provides the products that best suit each user. The  $y_{predict}$  file replicates the way that  $y_{test}$  imitates the results obtained. That is, the probability matrix of each user becoming a buyer of one of the sponsored products and several recommendations. The clusters will therefore be formed by users according to their highest predisposition to purchase one of the recommended products, in this study seven groups ((1) Hearing Aids, (2) Energy, (3) Finance, (4) Games, (5) NGO, (6) Insurance, (7) Telcos). However, as shown in the accuracy percentages matrix (Table 7), each of the groups has different degrees of reliability, with groups (4) Games, (6) Insurance and (1) Hearing Aids exceeding 60%, but the rest of the groups being highly unreliable (range between 0 and 21.4%). To illustrate one of the least successful groups, Table 9 shows a sample of four users from the “telcos” group ( $id_{product\_conv}=7$ ), which consists of those users who became buyers of telcos plus those who are assigned the product by the algorithm between the first two recommendations ( $id_{product\_conv}=7$  OR  $id_{pro\_recommender\_1}=7$

**Table 9** Sample of the content of the ‘recommendation’ table for “telcos” product (id\_product\_conv=7)

id_user	id_pro_recomendation_1	id_pro_recomendation_2	pb_recomendation_1	pb_recomendation_2
154,063	7	5	0.6797	0.3158
287,605	6	2	0.6833	0.2454
329,118	3	5	0.9552	0.0329
473,911	7	4	0.9027	0.0493

**Fig. 6** Cumulative accuracy percentage

OR  $id\_pro\_recommender\_2=7$ ), which could not occur simultaneously. Table 9 shows the probability of becoming buyers of two recommended products (shown in columns  $id\_pro\_recommender\_1$  and  $id\_pro\_recommender\_2$ ). Two recommendations were selected because, if only the product to which the algorithm assigns a higher probability is taken, an average hit rate of 70% is reached, but if those of the second product are added, an average hit rate of 92% is reached. Table 9 shows the average probabilities of each recommendation.

Although the clustering process and performance of both unsupervised K-Means and supervised XGBoost algorithms are not directly comparable, it is common in the literature to compare different algorithm profiles as new methodologies are developed (Hartmann et al., 2019, 2023). In addition, due to its popularity, K-Means is often used as a base algorithm in the analysis of new classification methods (Boone & Roehm, 2002; Liu et al., 2010). The K-Means algorithm forms user groups based on distances that are configured by measuring attribute ratings and, therefore, does not respond to any intentionality (Sommer & Haug, 2011). However, when XGBoost

is supervised, researchers label and assign priority to one of the attributes, with which they train the algorithm and thereby generate highly focused user groups (Chen & Guestrin, 2016). Although neither of the two generate perfect clusters, their results can be used to try to answer the research question as to which of the two types of algorithms, unsupervised K-Means, or supervised XGBoost, is most suitable and efficient for clustering lead data. The findings of this study suggest that the supervised method, due to the greater focus of the algorithm, offers clearly superior results to the unsupervised one.

## Conclusions

The development of information technologies and the Internet has given rise to an ecosystem where information from different sources (Websites, Blogs, social media, etc.) is constantly flowing between consumers and companies about different products, brands, and companies (González-Padilla et al., 2023). However, the process of capturing and analysing these enormous sources of data represents a major challenge for digital marketing management (Bala & Verma, 2018), which has led to the development of new tools based on artificial intelligence and the emergence of leads capture start-ups to provide quality databases (Sáez-Ortuño et al., 2023b).

Although the field of market research is witnessing a boom in the study of algorithms for capturing and analysing unstructured information (Hartmann et al., 2023; Ordabayeva et al., 2022; Timoshenko & Hauser, 2019), structured information is still relevant for clustering consumers into groups despite getting less consideration in the literature (one exception is Liu et al., 2010). The use of algorithms for the classification of consumers into segments is still a common practice among customer-focused entrepreneurs (Cossío-Silva et al., 2013; Liao et al., 2022; Wedel and Kamakura, 2000). In fact, entrepreneurs can use these instruments to estimate the segments that make up the digital market and to select their target group on which to focus their commercial offer considering value co-creation. If the chosen algorithm makes the best use of the gathered data, this helps them to understand customer preferences, focus the offer, improve customer satisfaction, and ultimately drive business success. Last, but not least, business success translates into employment opportunities, wealth, and economic growth (Cossío-Silva et al., 2013; Dahle et al., 2023). The division of consumers into similar groups of buying behaviour, attitudes, or demographic characteristics, etc., helps to develop marketing strategies focused on specific segments with the aim of maximising the effectiveness of communication and marketing efforts (James et al., 2013).

This research compares the suitability of two algorithms, K-Means (unsupervised learning) and XGBoost (supervised learning), for clustering large, structured databases obtained by lead acquisition.

As Memarsadeghi et al. (2007) point out, the analysis of databases composed of millions of pieces of information can only achieve efficient solutions in a reasonable time (N-problem) in special cases with  $k$  defined in a few groups, but there is no efficient solution for a general value of  $k$  (NP-strong problem). NP-strong problems imply the existence of algorithms that could solve the given task but in a polynomial

time that depends on the algorithm's input data. That is, the algorithm might find a solution, but not in a reasonable time (e.g., it might require several months of computation) (Gary & Johnson, 1978). To delimit the combinatorial range of the algorithms, the elbow rule was used in this study to define a priori the number of clusters (five) in the case of K-Means, and the length of the tree was restricted to 8 branches and the interactions to 100 in the case of XGBoost.

The K-Means algorithm was used because it is one of the most common algorithms for clustering consumers in market research and is often used as a basis for comparative studies (Hung et al., 2019; Liu et al., 2010). In turn, XGBoost is a decision tree-based machine learning algorithm that is often used in classification and regression problems. Although there are some exceptions, e.g. Liang et al. (2019), unlike K-Means it is not common in consumer clustering studies in marketing or entrepreneurial research. Market research is usually interested in segmenting the market with the aim of determining the most suitable target audience for a commercial offer. Therefore, economic relevance is a crucial factor when choosing a consumer clustering method (Hartmann et al., 2019).

## Theoretical implications

The findings of this study confirm the conjecture that every database requires the most appropriate algorithm for classification. In this comparative study, XGBoost offers better predictive performance with respect to the proposed commercial objectives to achieve a greater volume of transformations than K-Means, in line with Chen and Guestrin (2016). The tree structure of XGBoost means it can work with many features and unbalanced data (Natekin & Knoll, 2013) and detect interactions between variables, unlike K-Means which uses a Euclidean distance measure that can only represent linear patterns, meaning the former is better at representing complex relationships between data (Ke et al., 2017). XGBoost is also more robust to the presence of noise and outliers, thanks to the optimisation of the objective function, while in K-Means, outliers can skew cluster centroids (Raschka & Mirjalili, 2019; Sculley, 2010).

Undoubtedly, the different nature of the two algorithms makes them suitable for different objectives. While XGBoost requires labelled data to train the model, it is better at predicting the target variable, such as the probability of a user purchasing a product or service (Chen & Guestrin, 2016). In contrast, K\_Means, which can group users according to their characteristics and preferences, does not seem well suited to the objectives of the study. How easy it is to interpret the results is also important. The five groups formed by K-Means (Table 4; Fig. 2) are not easy to interpret, since conversion into one of the products (insurance) dominates in four of the five clusters. Meanwhile, the decision trees constructed by XGBoost based on the assigned labels, and based on the setting of a target variable, may be easier to interpret by providing a direct assignment to one of the seven clusters ((1) Hearing Aids, (2) Energy, (3) Finance, (4) Games, (5) NGO, (6) Insurance, (7) Telcos), which makes it easier to understand the relationship between the features and the target variable (Breiman, 2017). In summary, the choice between XGBoost and K-means in the analysis of

online sweepstakes and test data will depend on the objective of the study and the nature of the data. XGBoost provides greater predictive capability, detects more complex patterns in the data and shows greater robustness to noise compared to K-Means when analysing structured lead data. The outcome is more accurate and decision-useful clustering models.

### Managerial implications

The use of K-Means and XGBoost may also have managerial implications. The application of K-Means to databases that combine multiple variables, even if they are of a structured nature, does not seem to be a very good solution. In this study, a couple of undesirable consequences of compensatory clustering processes using distance algorithms, as previously noted by DeSarbo et al. (2005), have been obtained: (1) the resulting clusters are dominated by one or two variables over the rest (the Insurance variable dominates in four clusters), and (2) the obtained solution is not useful for management. In contrast, XGBoost has the advantage that it directly addresses the multivariate nature of the database, as it organises it sequentially in a tree structure. That is, it does not require aggregation or trade-offs between variables for the formation of consumer segments. Decision makers and entrepreneurs often prefer to do their analysis after looking at the full spectrum of solutions (both compatible and non-compatible solutions), rather than analysing aggregate results where the model has already performed a compensatory analysis between variables (Liu et al., 2010). This holistic view provides decision-makers with major flexibility and insights that are lacking in K-Means. However, the predictive estimates of market segmentation made by XGBoost need to be taken with caution as the hit rates of the trained model vary greatly between products. Ultimately, the challenge for any company is to be able to understand the desires and needs of its customers in order to group them according to their similarities, and thus be able to make an offer of value to the customer so that the company can meet its objectives. In addition, market segmentation by grouping homogeneous consumers enables the design of customised communication, which can improve customer receptiveness and transform them into purchasers, thus boosting customer satisfaction and loyalty (Cossío-Silva et al., 2013; Eskerod, 2020). However, it is well known that this convergence of interests often fails to occur, either because the customer is not interested in the offer that he/she receives, or because the person receiving it generates little value to the company (Cossío-Silva et al., 2013). Therefore, it is essential to generate segments based on business objectives that allow for convergence of interests.

### Limitations and future lines of research

In this study we have analysed a dataset from a Spanish consumer lead capture start-up. Although the data represents a real-world problem, it may differ from databases from other countries or cultures, or for other purposes than participation in online games and sweepstakes.

In line with Liu et al. (2010), the chosen algorithms should be used with caution as they often fail to withstand a comparative analysis of cross-solutions. That is, random



variation in clustering processes can sometimes cause consumers to be assigned to one cluster, but in a re-analysis, they are assigned to another cluster. For example, Table 6 illustrates the probability distribution of four consumers becoming purchasers of the sponsored products and their assignment to one of the clusters by XGBoost. In some cases, where the measures are markedly different, as in *id\_user* 11,188,636 who is assigned a probability of 0.96 of becoming a purchaser of Games and 0.03 of Insurance and has been assigned to the Games cluster by the algorithm, it is very likely that this user will withstand a second reevaluation. But in other cases, where the measures are very similar, such as *id\_user* 17,242,446, who has been assigned a 0.48 probability of purchasing a Hearing aid and a 0.49 probability of purchasing Insurance, and to whom the algorithm assigned Insurance, the user does not seem to withstand a second reevaluation. That is, the algorithm always places consumers in a cluster despite the similarity of probabilities assigned to various descriptive attributes.

Another concern is the ethical implications arising from the application of AI algorithms. In this study, the researchers verified that the data provided by the company was collected via a process that complied with current regulations (for more details see Sáez-Ortuño et al., 2023b). However, as Tsamados et al. (2021) point out, AI algorithms are not ethically neutral, as they depend on the sources consulted, the data collected, the statistical analyses applied and the interpretation of their results. One of these ethical dangers is mechanistic objectivity, which can lead researchers to assert inconclusive evidence or misconceptions, sometimes even against their own experienced assessments (Buhmann et al. 2019). This danger of misinterpretation is often exacerbated when the researcher applies algorithms without understanding how they generate their results (Tsamados et al., 2021). Thus, in the classification algorithms used in this study, including those using causal regression models, in some cases it was found that the available data are not sufficient to produce an accurate diagnosis and therefore to advise the decision-maker on the correct action (Olhede & Wolfe, 2018). Moreover, there is an ethical risk of profiling consumers by discriminating against gender, racial and other minorities due to structural inequalities in the databases that are collected and fed to algorithms, which simply replicate the structure of the data provided and are rarely corrected. More data alone does not lead to greater representativeness (Tsamados et al., 2021). Everyone is unique and a more in-depth understanding of their needs and preferences is necessary before marketing decisions can be made (Lloyd, 1982).

The results of this research lie halfway between market research and computer science studies. While the former aims to help entrepreneurs or managers with their decision making, the latter are much more interested in technical accuracy, so the same interests are not always shared (Hartmann et al., 2019). However, this study points out in its comparison between the older unsupervised algorithm and the modern supervised algorithm that the clusters resulting from the latter are more suitable for marketing interests than the former. Hence, there is a commonality of goals. But this is not necessarily always the case, so comparisons must continue to be made with new algorithms, because the degrees of accuracy achieved are not yet entirely satisfactory.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11367-023-00893-3>.

[org/10.1007/s11365-023-00882-1](https://doi.org/10.1007/s11365-023-00882-1).

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9(3), 52. <https://doi.org/10.3390/technologies9030052>.
- Ali, I., Balta, M., & Papadopoulos, T. (2023). Social media platforms and social enterprise: Bibliometric analysis and systematic review. *International Journal of Information Management*, 69(April), 102510. <https://doi.org/10.1016/j.ijinfomgt.2022.102510>.
- Alonso-González, M. J., Hoogendoorn-Lanser, S., van Oort, N., Cats, O., & Hoogendoorn, S. (2020). Drivers and barriers in adopting mobility as a service (MaaS)—A latent class cluster analysis of attitudes. *Transportation Research Part A: Policy and Practice*, 132, 378–401. <https://doi.org/10.1016/j.tra.2019.11.022>.
- Amit, R., & Zott, C. (2012). Creating value through business model innovation. *MIT Sloan Management Review*, 53(3), 41–49.
- Arabie, P., Hubert, L., & De Soete, G. (Eds.). (1996). *Clustering and classification*. World Scientific Publishing, NJ.
- Audretsch, D. B., Belitski, M., Caiazza, R., & Lehmann, E. E. (2020). Knowledge management and entrepreneurship. *International Entrepreneurship and Management Journal*, 16(2), 373–385. <https://doi.org/10.1007/s11365-020-00648-z>.
- Bala, M., & Verma, D. (2018). A critical review of digital marketing. *International Journal of Management IT & Engineering*, 8(10), 321–339.
- Balioukas, P., Llopis, J., Gasco, J., & Gonzalez, R. (2022). Implementing turnaround strategies as an entrepreneurial process. *International Entrepreneurship and Management Journal*. <https://doi.org/10.1007/s11365-022-00810-9>.
- Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral science*, 12(2), 153–155. Doi: <https://doi.org/10.1002/bs.3830120210>.
- Basu, D., Sinha, R., Sahu, S., Malla, J., Chakravorty, N., & Ghosal, P. S. (2022). Identification of severity and passive measurement of oxidative stress biomarkers for  $\beta$ -thalassemia patients: K-means, random forest, XGBoost, decision tree, neural network based novel framework. *Advances in Redox Research*, 5, 100034. <https://doi.org/10.1016/j.arres.2022.100034>.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (4 vol., p. 738). Springer, 4.
- Boone, D. S., & Roehm, M. (2002). Retail segmentation using artificial neural networks. *International Journal of Research in Marketing*, 19(3), 287–301. Doi: [https://doi.org/10.1016/S0167-8116\(02\)00080-0](https://doi.org/10.1016/S0167-8116(02)00080-0).
- Breiman, L. (2017). *Classification and regression trees*. Routledge.
- Bremer, L., Heitmann, M., & Schreiner, T. F. (2017). When and how to infer heuristic consideration set rules of consumers. *International Journal of Research in Marketing*, 34(2), 516–535. Doi: <https://doi.org/10.1016/j.ijresmar.2016.10.001>.
- Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse. *Journal of Business Ethics*, 163(2), 265–280. <https://doi.org/10.1007/s10551-019-04226-4>.

- Chakraborty, H., Moore, J., Carlo, W. A., Hartwell, T. D., & Wright, L. L. (2009). A simulation based technique to estimate intracluster correlation for a binary variable. *Contemporary clinical trials*, 30(1), 71–80. Doi: <https://doi.org/10.1016/j.cct.2008.07.008>.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
- Cossío-Silva, F. J., Revilla-Camacho, M. A., & Vega-Vázquez, M. (2013). Heterogeneity of customers of personal image services: a segmentation based on value co-creation. *International Entrepreneurship and Management Journal*, 9, 619–630. Doi: <https://doi.org/10.1007/s11365-013-0266-3>.
- Dahle, Y., Reuther, K., Steinert, M., & Supphellen, M. (2023). Towards a systemic entrepreneurship activity model. *International Entrepreneurship and Management Journal*, 1–28. Doi: <https://doi.org/10.1007/s11365-023-00874-1>.
- Dahle, Y., Duc, A. N., Steinert, M., & Chizhevskiy, R. (2018, June). Resource and competence (internal) view vs. environment and market (external) view when defining a business. In 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1–9). IEEE.
- Dayan, N., Twitto, M., Rochman, Y., Beitler, U., Zion, I. B., Bortnikov, E., & Rabinovich, N. (2021). The end of Moore's law and the rise of the data processor. Proceedings of the VLDB Endowment, 14(12), 2932–2944.
- Desai, V. (2019). Digital marketing: A review. *International Journal of Trend in Scientific Research and Development*, 5(5), 196–200. Doi: <https://doi.org/10.31142/ijtsrd23100>.
- DeSarbo, W. S., & Grisaffe, D. (1998). Combinatorial optimization approaches to constrained market segmentation: An application to industrial market segmentation. *Marketing Letters*, 9, 115–134. Doi: <https://doi.org/10.1023/A:1007997714444>.
- DeSarbo, W. S., Di Benedetto, A., Song, C. M., & Sinha, I. (2005). Revisiting the Miles and Snow strategic framework: uncovering interrelationships between strategic types, capabilities, environmental uncertainty, and firm performance. *Strategic Management Journal*, 26(1), 47–74. Doi: <https://doi.org/10.1002/smj.431>.
- Eskerod, P. (2020). A stakeholder perspective: Origins and core concepts. In Oxford Research Encyclopedia of Business and Management.
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 3133–3181.
- Garey, M. R., & Johnson, D. S. (1978). "Strong" NP-Completeness results: Motivation, examples, and implications. *Journal of the Association for Computing Machinery (JACM)*, 25(3), 499–508.
- González-Padilla, P., Navalpotro, F. D., & Saura, J. R. (2023). Managing entrepreneurs' behavior personalities in digital environments: A review. *International Entrepreneurship and Management Journal*, 1–25. Doi: <https://doi.org/10.1007/s11365-022-00823-4>.
- Guerola-Navarro, V., Gil-Gomez, H., Oltra-Badenes, R., & Soto-Acosta, P. (2022). Customer relationship management and its impact on entrepreneurial marketing: A literature review. *International Entrepreneurship and Management Journal*. <https://doi.org/10.1007/s11365-022-00800-x>.
- Gultom, S., Sriadi, S., Martiano, M., & Simarmata, J. (2018, September). Comparison analysis of K-means and K-medoid with Euclidean distance algorithm, Chanberra distance, and Chebyshev distance for big data clustering. In IOP Conference Series: Materials Science and Engineering (Vol. 420, No. 1, p. 012092). IOP Publishing.
- Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1), 20–38. Doi: <https://doi.org/10.1016/j.ijresmar.2018.09.009>.
- Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1), 75–87. Doi: <https://doi.org/10.2139/ssrn.3489963>.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2 vol., pp. 1–758). Springer.
- Henriques, J., Caldeira, F., Cruz, T., & Simões, P. (2020). Combining k-means and xgboost models for anomaly detection using log datasets. *Electronics*, 9(7), 1164. <https://doi.org/10.3390/electronics9071164>.
- Hung, P. D., Ngoc, N. D., & Hanh, T. D. (2019, February). K-means clustering using RA case study of market segmentation. In Proceedings of the 2019 5th International Conference on E-Business and Applications (pp. 100–104).

- Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. *Journal of Applied Science and Technology Trends*, 2(01), 10–19. Doi: <https://doi.org/10.38094/jastt20179>.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. Doi: <https://doi.org/10.1016/j.patrec.2009.09.011>.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264–323. Doi: <https://doi.org/10.1145/331499.331504>.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (112 vol., p. 18). Springer.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>.
- Kamthania, D., Pawa, A., & Madhavan, S. S. (2018). Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business. *Journal of computing and information technology*, 26(1), 57–68. Doi: <https://doi.org/10.20532/cit.2018.1003863>.
- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 90–95.
- Kohavi, R., & Provost, F. (1998). Confusion matrix. *Machine learning*, 30(2–3), 271–274.
- Kohonen, T., Oja, E., Simula, O., Visa, A., & Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 84(10), 1358–1384.
- Kraus, S., Breier, M., & Dasí-Rodríguez, S. (2020). The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, 16(3), 1023–1042. <https://doi.org/10.1007/s11365-020-00635-4>.
- Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. *Computers & Operations Research*, 29(11), 1475–1493. Doi: [https://doi.org/10.1016/S0305-0548\(01\)00043-0](https://doi.org/10.1016/S0305-0548(01)00043-0).
- Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019, August). Product marketing prediction based on XGboost and LightGBM algorithm. In *Proceedings of the 2nd international conference on artificial intelligence and pattern recognition* (pp. 150–153).
- Liao, Y. K., Nguyen, V. H. A., & Caputo, A. (2022). Unveiling the role of entrepreneurial knowledge and cognition as antecedents of entrepreneurial intention: A meta-analytic study. *International Entrepreneurship and Management Journal*, 18(4), 1623–1652. <https://doi.org/10.1007/s11365-022-00803-8>.
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2), 451–461. Doi: [https://doi.org/10.1016/S0031-3203\(02\)00060-2](https://doi.org/10.1016/S0031-3203(02)00060-2).
- Liu, Y., Ram, S., Lusch, R. F., & Brusco, M. (2010). Multicriterion market segmentation: a new model, implementation, and evaluation. *Marketing Science*, 29(5), 880–894. Doi: <https://doi.org/10.1287/mksc.1100.0565>.
- Liu, J., Wu, J., Liu, S., Li, M., Hu, K., & Li, K. (2021). Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *Plos one*, 16(2), e0246306. Doi: <https://doi.org/10.1371/journal.pone.0246306>.
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2), 129–137. Doi: <https://doi.org/10.1109/TIT.1982.1056489>.
- MacQueen, J. (1967, June). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability* (pp. 281–297). Los Angeles LA USA: University of California.
- Memarsadeghi, N., Mount, D. M., Netanyahu, N. S., & Le Moigne, J. (2007). A fast implementation of the ISODATA clustering algorithm. *International Journal of Computational Geometry & Applications*, 17(1), 71–103. Doi: <https://doi.org/10.1142/S0218195907002252>.
- Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5, 181–204. Doi: <https://doi.org/10.1007/BF01897163>.
- Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Computer Science*, 3, e127. <https://doi.org/10.7717/peerj-cs.127>.
- Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering*, 109, 233–252. Doi: <https://doi.org/10.1016/j.cie.2017.04.017>.

- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7, 21. Doi: <https://doi.org/10.3389/fnbot.2013.00021>.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543. Doi: <https://doi.org/10.3389/fnbot.2013.00021>.
- Olhede, S. C., & Wolfe, P. J. (2018). The growing ubiquity of algorithms in society: Implications, impacts and innovations. *Philosophical transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128), 20170364. <https://doi.org/10.1098/rsta.2017.0364>.
- Ordabayeva, N., Cavanaugh, L. A., & Dahl, D. W. (2022). The upside of negative: Social distance in online reviews of identity-relevant brands. *Journal of Marketing*, 86(6), 70–92. Doi: <https://doi.org/10.1177/00222429221074704>.
- Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., & Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. *International Journal of System Assurance Engineering and Management*, 1–12. Doi: <https://doi.org/10.1007/s13198-021-01130-x>.
- Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.
- Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using  $\ell_1$ -regularized logistic regression. *Ann. Statist.* 38(3): 1287–1319. Doi: <https://doi.org/10.1214/09-AOS691>.
- Sánchez-Ortuño, L., Forgas-Coll, S., Huertas-García, R., & Sánchez-García, J. (2023a). What's on the horizon? A bibliometric analysis of personal data collection methods on social networks. *Journal of Business Research*, 158, 113702. Doi: <https://doi.org/10.1016/j.jbusres.2023.113702>.
- Sánchez-Ortuño, L., Forgas-Coll, S., Huertas-García, R., & Sánchez-García, J. (2023b). Online cheaters: Profiles and motivations of internet users who falsify their data online. *Journal of Innovation & Knowledge*, 8(2), 100349. <https://doi.org/10.1016/j.jik.2023.100349>.
- Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2), 34–37. <https://doi.org/10.14569/IJARAI.2013.020206>.
- Sculley, D. (2010, April). Web-scale k-means clustering. In Proceedings of the 19th international conference on World wide web (pp. 1177–1178).
- Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, 21(1), 3–8. Doi: <https://doi.org/10.2307/1247695>.
- Sommer, L., & Haug, M. (2011). Intention as a cognitive antecedent to international entrepreneurship—understanding the moderating roles of knowledge and experience. *International Entrepreneurship and Management Journal*, 7(1), 111–142. <https://doi.org/10.1007/s11365-010-0162-z>.
- Stead, M., Gordon, R., Angus, K., & McDermott, L. (2007). A systematic review of social marketing effectiveness. *Health education*, 107(2), 126–191. Doi: <https://doi.org/10.1108/09654280710731548>.
- Strehl, A., & Ghosh, J. (2003). Relationship-based clustering and visualization for high-dimensional data mining. *INFORMS Journal on Computing*, 15(2), 208–230. Doi: <https://doi.org/10.1287/ijoc.15.2.208.14448>.
- Sujatha, S., & Sona, A. S. (2013). New fast k-means clustering algorithm using modified centroid selection method. *International Journal of Engineering Research & Technology (IJERT)*, 2(2), 1–9.
- Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing. Doi: <https://doi.org/10.1088/1757-899X/336/1/012017>.
- Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8, 321–340. Doi: <https://doi.org/10.1146/annurev-economics-080315-015325>.
- Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1), 1–20. Doi: <https://doi.org/10.1287/mksc.2018.1123>.
- Tsamados, A., Aggarwal, N., Cowsls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The Ethics of Algorithms: Key Problems and Solutions. In: Floridi, L. (eds) Ethics, Governance, and Policies in Artificial Intelligence. Philosophical Studies Series, vol 144. (97–123) Springer, Cham. Doi: [https://doi.org/10.1007/978-3-030-81907-1\\_8](https://doi.org/10.1007/978-3-030-81907-1_8).
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.
- Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1), 1–67.
- Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(3), 1–27. <https://doi.org/10.4018/jdwm.2009070101>.

- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Kluwer Academic Publishers Group.
- Zamri, N., Pairan, M. A., Azman, W. N. A. W., Abas, S. S., Abdullah, L., Naim, S., & Gao, M. (2022). A comparison of unsupervised and supervised machine learning algorithms to predict water pollutions. *Procedia Computer Science*, 204, 172–179. Doi: <https://doi.org/10.1016/j.procs.2022.08.021>.
- Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53. Doi: <https://doi.org/10.1093/nsr/nwx106>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.