



# Intelligent System Comparing Clustering Algorithms to Recommend Sales Strategies

Gianella Arévalo-Huaman, Jose Vallejos-Huaman, and Daniel Burga-Durango<sup>(✉)</sup>

Universidad Peruana de Ciencias Aplicadas (UPC), Lima San Isidro 15076, Peru  
{u2016c716,u201622094,pcsidbur}@upc.edu.pe

**Abstract.** The context of the pandemic has accelerated the growth of electronic commerce in recent years. Consequently, there is intense competition among companies to boost sales and achieve success in a market environment where the failure rate stands at 80%. Motivated by this reason, an Intelligent System is proposed to recommend a sales campaign strategy within an e-commerce platform, automating the analysis of customer data by employing machine learning algorithms to segment (K-means) customers into groups based on their information. Additionally, the system recommends (Decision Tree) a specific sales strategy for each group. Therefore, the objective of this study is to analyze all the relevant aspects that arise in the relationship between an e-commerce business and its customers, as well as the effectiveness of generating strategies based on specific groups through Customer Segmentation. As a result, the system achieved a significant increase in Web Traffic, Click-through Rate, and Sales Revenue by 14%, 5%, and 10%, respectively, indicating a monetary growth and improved engagement after the utilization of our tool.

**Keywords:** Customer Segmentation · E-commerce · K-means · Marketing Strategies

## 1 Introduction

According to [1], it is estimated that 85% of e-commerce businesses with an investment of less than 10,000 euros fail to surpass two years of activity. The expenditure in ecommerce increased by 44% in 2020, with a similar trend in 2021, as reported by Forbes in [www.elmundo.es](http://www.elmundo.es) [2]. However, the initiative of new companies entering the market in such a competitive context result in 90% of them failing within the first 4 months. To reduce the risk of user loss and strengthen their trust in products within an e-commerce platform, customer analysis should be conducted using techniques such as clustering to understand users and their needs, exploring potential customer groups [3].

In [3], the authors suggest that e-commerce platforms have the capacity to collect a large amount of data daily, which should be utilized for analyzing customer behavior and creating products for different groups of potential users. However, [4] mentions that for data to become an important resource for a company, it must be effectively utilized; otherwise, it can become a burden. This involves the utilization of technologies

such as artificial intelligence, statistics, and databases, among others [4]. Consequently, smaller e-commerce companies find themselves in a disadvantaged position as they lack specialized technology and data analysis teams to compete with larger companies.

Therefore, aligned to [5], this study proposes a System that enables the generation of Sales Campaign Strategies for each customer segment in the e-commerce platform. To achieve this, the more precise Clustering algorithm will be used to classify customers for retail e-commerce, choosing between k-means and Hierarchical Clustering based on the collected company data. Then, these data will be fed into the Decision Tree algorithm to generate personalized sales strategies that the company can apply in their marketing plan, aiming to improve their performance in the market [6].

## 2 Related Work

In [7], a class model was developed to segment customers into groups and classify them based on their income. This proposal provided access to key information for maintaining good customer relationships and evaluating them in the long term [7]. The segmentation was based on identifying customer groups using characteristics of their own behavior. It was discussed how the obtained results could be relevant for implementing a loyalty program and improving future studies that also address customer quality and lifestyle as an additional variable, combining survey data and purchase history.

In [8], research was conducted to find patterns of customer behavior using data mining techniques and the support of the K-Means clustering algorithm to identify the best-selling products and the payment method used. RapidMiner was used to facilitate the use of necessary tools for analysis, and the Davies-Bouldin index (DBI) was used to evaluate the quality of the segmentation algorithm.

In [9], data mining was employed to discover patterns in customer behavior. It was found that the best-selling products belonged to category 503–505 and that the most used payment method was credit. It is important to adapt to changes in the virtual market and personalize strategies to achieve better results in audience relationships.

In [10], the authors propose a comprehensive and simplified system from data pre-processing to visualization, suitable for small businesses. This system identifies the popularity of each product over a period and targets potential customers based on that information. The purpose of customer segmentation is to divide the user base into smaller groups that can be targeted with specialized content and offers. Customer segmentation allows businesses to efficiently target each specific group of customers. The study implements customer segmentation using a hierarchical clustering algorithm with a small dataset. Additionally, a credit card dataset is utilized. The agglomerative hierarchical clustering method is performed using the `hclust` function from the `Cluster` package in R. The study concludes with the authors' perspective. However, it is noted that this method can be slow and hardware-dependent.

### 3 Method

#### 3.1 Algorithm Comparison

To compare the K-means and HC algorithms, we conducted a benchmarking process. We used two tables (Table 1 and Table 2) to evaluate the algorithms in terms of scalability and five other dimensions based on e-commerce sales datasets, both public and private. Then, we compared the results in Table 3 to determine the accuracy of each algorithm in customer segmentation. We used this information to develop an effective customer segmentation system.

**Table 1.** Algorithm Benchmarking.

Criteria	Hierarchical Clustering	K-Means
Initial condition	No	Yes
Final condition	Regular	Accurate
Random value	Not required	Numeric attribute
Impact on the size of the datasets	Regular	Good
Granularity	Flexible	K and starting point
Dynamic data management	No	Yes
Implementation	Simple	Simple

**Table 2.** Algorithm Benchmarking.

Criteria	%	Hierarchical Clustering		K-Means	
		Score	Results	Score	Results
Initial condition	10%	0	0	1	0.1
Final condition	15%	0	0	1	0.15
Random value	10%	1	0.1	0	0
Impact on the size of the datasets	25%	0	0	1	0.25
Granularity	10%	1	0.1	0	0
Dynamic data management	20%	0	0	1	0.2
Implementation	10%	1	0.1	1	0.1
<b>TOTAL</b>	<b>100%</b>		<b>0.3</b>		<b>0.8</b>

In Table 1, we evaluated the algorithms across five different dimensions. In Table 2, we assigned weights to each dimension to identify which one had the greatest impact on customer segmentation. Based on the results obtained in these two tables, we determined that K-means was the best algorithm for our project.

**Table 3.** Comparative Table of Cluster Accuracy for K-means and HC.

Datasets	K-Means	Hierarchical Clustering
Dataset 1	0.420	0.145
Dataset 2	0.456	0.409
Dataset 3	0.532	0.129
Dataset 4	0.603	0.257
Dataset 5	0.131	0.051

Finally, in Table 3, we display the results of the accuracy comparison between K-means and HC across five different companies. The results indicate that K-means achieved higher accuracy than HC in the division of 4 customer clusters. This information was crucial for selecting the appropriate algorithm for our customer segmentation application.

### 3.2 Decision-Making Algorithm

In the second stage of executing our solution, a decision tree algorithm was used to establish the best sales strategy for the e-commerce platform for each of the customer clusters generated as the output of the K-means segmentation algorithm. The algorithm implementation was based on creating a supervised learning model, which requires a labeled dataset from which the created model can make the correct decision based on what it has learned.

In our project, the decision was made to create a column called “Strategy,” which was used as the label to determine the most accurate value. Therefore, we used the clusters created in the previous step, which have average characteristics of the customers belonging to each cluster, as the input for the decision tree algorithm. This way, the algorithm can determine the most suitable strategy to apply. To accomplish this, it was necessary to gather a dataset of 200 clusters to train the decision tree with specific strategies labeled in the designated column based on the characteristics of each cluster. We had the professional support of a marketing specialist who helped us label the best strategy.

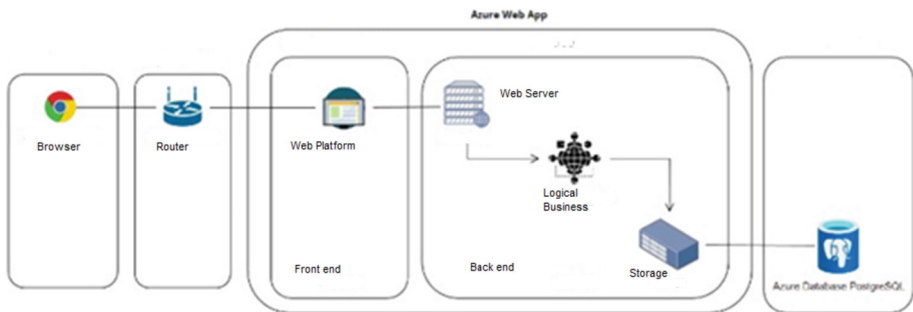
In this manner, the intelligent system can provide the company with precise and personalized sales strategies, which contributes to improving its performance in the market and increasing profitability [4].

### 3.3 Tool Implementation

Once the K-means algorithm has been selected as the preferred approach for segmenting customers of small e-commerce businesses, an intelligent system was developed to generate personalized sales campaign strategies for each customer segment identified in the e-commerce platform using the data collected by the company. To classify retail customers, the K-means clustering algorithm will be employed. Then, this data will

be fed into a decision tree algorithm to generate customized sales strategies that the company can apply in its marketing plan with the aim of improving its performance in the market.

The logical structure of the system consists of various layers: client, access, presentation, business, and data. First, the client layer is established, where the web browser plays a fundamental role as a means of communication with the user. To achieve successful connection, it is essential to have an internet gateway located in the access layer. The presentation and business layers are adjacent to the access layer and are encapsulated in the service provided by Azure Web App, where our solution will be deployed. The presentation layer focuses on the visual aspect and user interface of our platform, being the visible point of contact for the user. On the other hand, the business layer hosts all the necessary logic to carry out the functionalities of our solution. As for data storage, we propose directing the data generated by our intelligent system to an external cloud database using the Azure Database PostgreSQL service, which is in a separate layer called the Data Layer.



**Fig. 1.** Logical Architecture

The proposed physical architecture outlines how the connection will be established between the user, in this case, a marketing employee of an e-commerce business, and our solution. It also details the necessary technologies for the proper development of the intelligent system.

The user's connection will be made through a laptop or desktop computer, either via a Wi-Fi or direct Ethernet connection. Once the user has internet access, they can access our platform deployed on the Azure Web App service.

The technology distribution is divided into two sections: the front-end part of our platform, where direct visual interaction with the user will be established. For this, we will use HTML, CSS, and JavaScript, and enhance the user experience using Angular, taking advantage of its facilities for implementing modules and views.

In the back-end part, the main language will be Python, along with the Django framework, to carry out the necessary logical functionalities in the development of our algorithms and data storage. For data storage and connection with the external Azure Database PostgreSQL service, we will utilize the SQLAlchemy library, which will facilitate the creation of the data model and transactional rules to be applied in our solution.

For the creation and development of the clustering algorithms, we will employ the scikit-learn library, which offers various functions related to major clustering algorithms and simplifies their implementation.

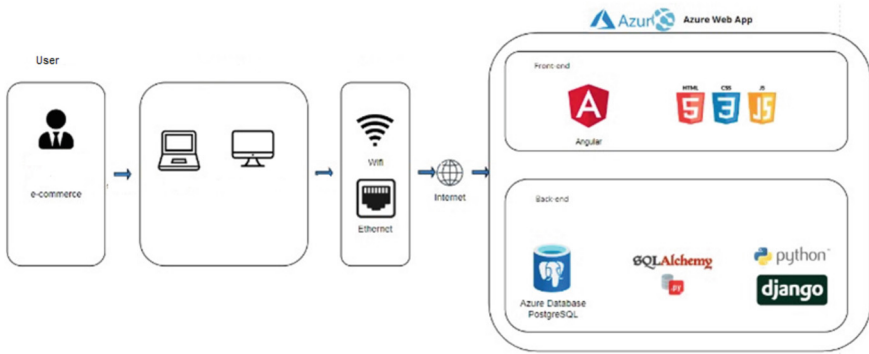


Fig. 2. Physical Architecture

The intelligent system under development offers a set of functionalities that improve its usefulness and efficiency. These features include: system login for secure access; registration of new users to create individual accounts; the ability to modify the user’s profile to update personal information; access to different views of the system through an intuitive navigation menu; system dataset management, including loading, modifying, and deleting datasets; running the analysis tool, which allows you to select a specific dataset for analysis; and finally, the execution of the analysis itself, using a segmentation model to obtain relevant and significant results. These combined functionalities offer users a complete and efficient experience when using the intelligent system. The success of the development of these functionalities can be seen in Figures 1, 2 and 3 of the tool (Figs. 4 and 5).

## 4 Validation

The selected e-commerce company carried out a comprehensive evaluation of our platform and its key features over an approximate period of 1 month. The objective of this process was to obtain quantifiable results that can be compared with historical data, in order to validate any observed variations after using our tool [11]. In this way, the aim is to determine how our solution can support the growth of an e-commerce business by effectively leveraging customer data and providing effective support in that regard. The success or failure in e-commerce will depend on the positive response to the established indicators [12].

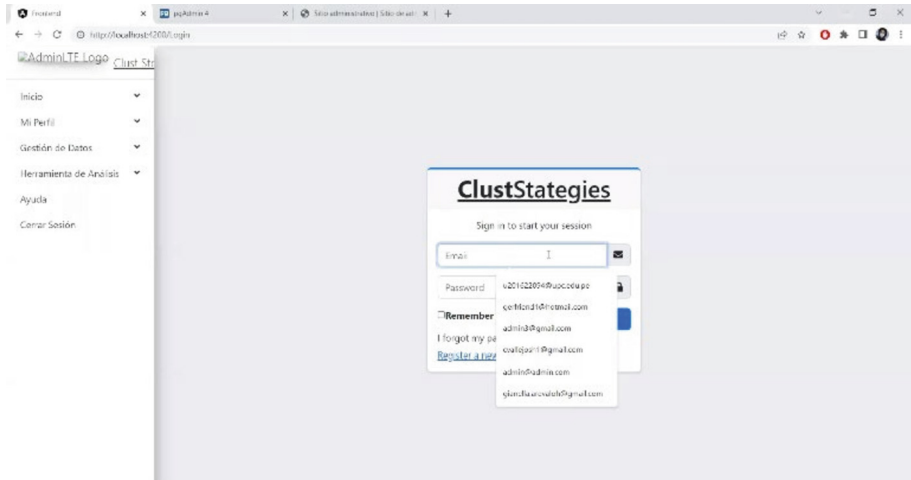


Fig. 3. Login

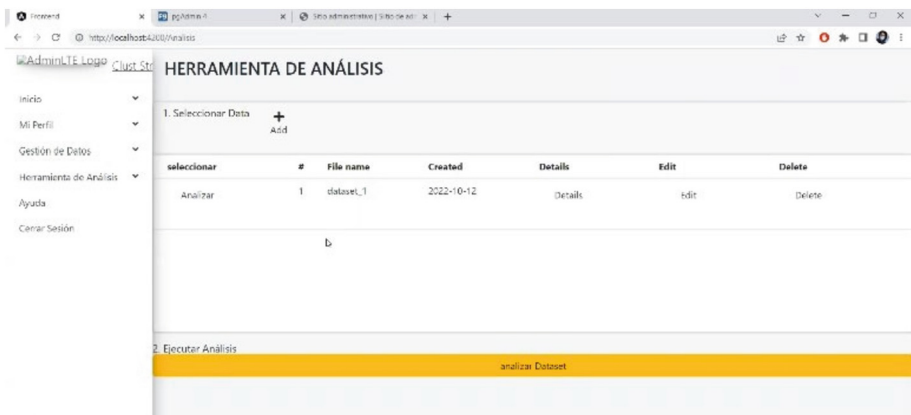


Fig. 4. Dataset Loading

## 4.1 Evaluation Indicators

The validation of the results obtained after the company has used our service will serve as support to evaluate four relevant indicators that contribute to improving business growth and increasing productivity. This validation will also help prevent the inclusion of the company in the alarming failure rate that prevails among small e-commerce businesses (Table 4).

The proposed indicators encompass several key aspects to evaluate the performance and growth of an e-commerce business. These indicators include Web Traffic, Click-through Rate, Customer Acquisition, and Sales Revenue. Both the visit and page click

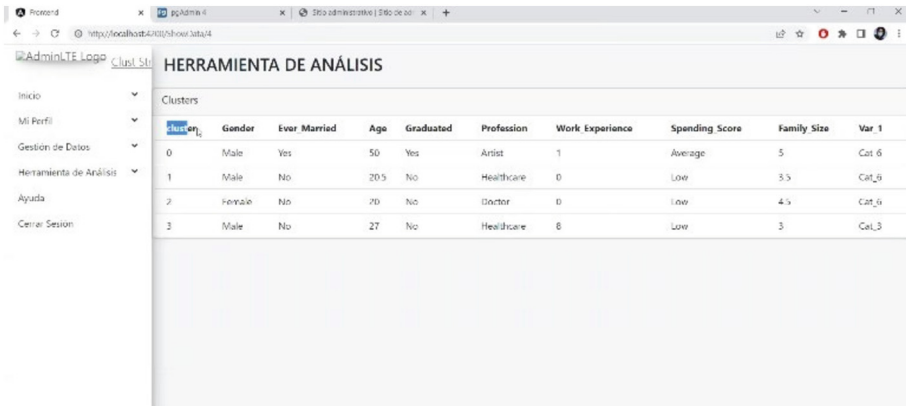


Fig. 5. Dataset Analysis and Cluster Results

Table 4. Evaluation Indicators.

Indicator	Description
Website Traffic (WT)	A visit is recorded each time a user accesses the website
Click-through Rate (CTR)	It is calculated by dividing the number of clicks received by the number of impressions or views of the item in question
Customer Acquisition (CA)	It is calculated by counting the number of customers who make their first purchase
Sales Revenue (SR)	It is calculated by adding the monetary value of all sales transactions made in that period

indicators are useful for analyzing the progress of interaction and behavior of customers and potential customers on the website. These indicators allow us to determine if there has been growth after implementing the strategies defined by our solution.

Additionally, it is important to consider that increased interaction can lead to the conversion of potential customers into actual customers. This aspect will also be evaluated to measure growth in the buyer base. Lastly, the primary indicator to observe will be the increase in sales through the platform, as this will provide a substantial outcome on the growth of the e-commerce business.

In summary, the selected indicators such as page visits, new customers, page clicks, and sales are fundamental elements to evaluate the growth and success of the e-commerce business. These indicators will provide us with a comprehensive view of performance and allow us to measure the impact of the implemented strategies on e-commerce growth.

#### 4.2 Pre-implementation State

To validate the obtained results, it is necessary to have access to the historical information of the e-commerce business related to our indicators, in order to make a comparison



and determine if there has been an improvement and growth with the use of our tool. The selected company uses Shopify as the platform for website management and organization, along with Google Analytics for page interaction ratios, which has facilitated the extraction of various metrics related to their customers, generated in their information panel of both tools. These metrics have been aligned with our indicators, including Web Traffic, Click-through Rate, Customer Acquisition, and Sales Revenue. The comparison will be performed using data from the last six months, prior to the implementation of our solution, as shown in Table 5.

**Table 5.** Six-Month Metrics per Indicator.

Indic	Dec	Jan	Feb	Mar	Apr	May
WT	406	356	363	322	337	340
CTR	1766	1758	1769	1651	1712	1680
CA	10	7	8	3	1	3
SR	S/4.756,89	S/4.449,49	S/4.410,50	S/4.268,36	S/3.916,89	S/4.131,67

Then, Formula 1 will be used to assess the variation of these indicators over the 6-month study period.

$$VarVVatVvon = \frac{\text{results ffor the month}_a - \text{results ffor the month}_{a-1}}{\text{results ffor the month}_{a-1}} \quad (1)$$

In formula (1), “a” is the index of the month being evaluated and “a – 1” is the index of the previous month.

### 4.3 Pos-implementation State

With the objective of evaluating the indicators, an estimated period of 1 month (4 weeks) was established during which a weekly marketing plan was implemented based on the recommendations provided by our platform. This plan sought to obtain results that reflected the impact and benefits of our solution on the potential development of the electronic commerce business and to compare the growth percentages between the previous months without the use of the tool, compared to the growth of the latter. Month. Once the 1-month period was over, we proceeded to evaluate the results of the four key indicators (WT, CTR, CA and SR). We made a comparison between the values prior to the implementation of the solution and the values after the application of the strategies recommended by our system. To carry out this comparison, we considered the results of one month prior to the use of the tool in comparison with the results obtained using our tool.

Next, we define the formula used to determine the variation in the indicators and that gives us an objective answer about the growth or not of the indicators:

$$VarVVatVvon = \frac{\text{scenar}\% \text{ with tool} - \text{no tool scenar}\%}{\text{no-tool scenar}\%} \quad (2)$$

## 5 Results

The metrics from the e-commerce Shopify platform were collected, and the results of the 4 indicators were analyzed considering the periods mentioned in Table 6 (Table 7).

**Table 6.** Sampling Periods.

Range	Period
1 month before	15/04/2023–14/05/2023
1 month after	15/05/2023–14/06/2023

**Table 7.** Two-Month Metrics per Indicator.

Indicator	May (no tool)	Jun (with tool)
WT	340	420
CTR	1680	1781
CA	3	4
SR	<i>S/</i> .4131,67	<i>S/</i> .4557,49

We applied Formula 1 and Formula 2 to evaluate a comparison of both scenarios for each of the indicators.

**Table 8.** Indicators Variation.

Indicator	Previous scenario	Current scenario
WT	4,73%	14,41%
CTR	1,21%	5,95%
CA	25,85%	33,33%
SR	0.94%	10.31%

Table 8 displays the Variation of the 4 indicators (WT, CTR, CA, and SR), showing a comparison of the average growth before the use of the tool, contrasted with the growth of the metrics with the use of our solution. This allows for a more objective and effective comparison of the percentages.

## 6 Conclusions

The analysis of the results concludes that there is a significant increase in web traffic, click-through rate, and total sales compared to the previous scenario where the tool was not used and marketing strategies were not planned based on data. It is important to

consider that December, January, and February are the months with the highest sales in the year. Despite these factors, it is evident that the use of the proposed solution allows for the growth of the indicators. It is also important to highlight that there is a notable increase in sales in monetary terms, despite the implementation of discounts and promotions on the products.

In conclusion, the sales strategies recommended by the tool have positively influenced the company both in monetary terms and in engagement.

## References

1. We Need To Talk About E-Commerce: Too Many Digital Brands Are Failing, And We Need New Tools To Help Keep Founders On Track. <https://www.forbes.com/sites/forbestechcolumnist/2021/07/22/we-need-to-talk-about-e-com-merce-too-many-digital-brands-are-failing-and-we-need-new-tools-to-help-keep-founders-on-track/?sh=736f25871c9c>. Accessed 21 Apr 2023
2. La burbuja de las tiendas ‘online’: el 85% cierra antes de cumplir 2 años. <https://www.elmundo.es/comunidad-valen-ciana/2016/05/21/57401de422601d252e8b4589.html>. Accessed 21 Apr 2023
3. Zhang, B., Wang, L., Li, Y.: Precision marketing method of e-commerce platform based on clustering algorithm. *Complexity* **2021**, Article no. 5538677, 10 p. (2021). <https://doi.org/10.1155/2021/5538677>
4. Fang, C., Liu, H.: Research and application of improved clustering algorithm in retail customer classification. *Symmetry* **13**(10), 1789 (2021). <https://doi.org/10.3390/sym13101789>
5. Zhan, M., Gao, H., Liu, H., Peng, Y., Lu, D., Zhu, H.: Identifying market structure to monitor product competition using a consumer-behavior-based intelligence model. *Asia Pac. J. Mark. Logist.* **33**(1), 99–123 (2020). <https://doi.org/10.1108/APJML-08-2019-0497>
6. Lam, H. Y., Tsang, Y.P., Wu, C.H., et al.: Data analytics and the P2P cloud: an integrated model for strategy formulation based on customer behaviour. *Peer Peer Netw. Appl.* **14**, 2600–2617 (2021). <https://doi.org/10.1007/s12083-020-00960-z>
7. Shen, B.: E-commerce customer segmentation via unsupervised machine learning. In: *The 2nd International Conference on Computing and Data Science (CONF-CDS 2021)*, pp. 1–7. Association for Computing Machinery, New York, Article no. 45 (2021). <https://doi.org/10.1145/3448734.3450775>
8. Punhani, R., Arora, V.P.S., Sabitha, S., Kumar Shukla, V.: Application of clustering algorithm for effective customer segmentation in e-commerce. In: *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, pp. 149–154 (2021). <https://doi.org/10.1109/ICCIKE51210.2021.9410713>
9. Zhou, J., Wei, J., Xu, B.: Customer segmentation by web content mining. *J. Retail. Consum. Serv.* **61**, 02588 (2021). ISSN 0969-6989. <https://doi.org/10.1016/j.jretconser.2021.102588>
10. Hung, P.D., Lien, N.T.T., Ngoc, N.D.: Customer segmentation using hierarchical agglomerative clustering. In: *Proceedings of the 2nd International Conference on Information Science and Systems (ICISS 2019)*, pp. 33–37. Association for Computing Machinery, New York (2019). <https://doi.org/10.1145/3322645.3322677>
11. Alvarez Intriago, V., Agreda Fernández, L., Cevallos Gamboa, A.: Análisis de la estrategia de marketing digital mediante herramientas de analítica web. *Investigatio* (7), 81–97 (2021). <https://doi.org/10.31095/investigatio.2016.7.5>
12. Fernández-Planells, A.: Guía para la definición y creación de sitios web de calidad: evaluación y análisis comparativo a través del análisis experto. *Revista Española De Documentación Científica* **39**(4), e158. Recuperado a partir de (2016). <https://redc.revistas.csic.es/in-dex.php/redc/article/view/962>