Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 8 Issue 2, February - 2022

"Optimization in the assignment of routes using the K-means algorithm and the Elbow method"

Yareli Mireles Moreno[1]
University of Guanajuato, Multidisciplinary Studies Department
Yuriria, Guanajuato.
y.mirelesmoreno@ugto.mx

Dr. Roberto Baeza Serrato[2]
University of Guanajuato, Multidisciplinary Studies Department
Yuriria, Guanajuato
r.baeza@ugto.mx

***Abstract*— This article aims to design and develop the K-means algorithm, as well as the Elbow method, which were applied in a SME in the southern state of Guanajuato so that, through these two tools, the company can achieve significant savings in I count to fuel. The proposed K-means algorithm was developed with K = 3, K = 4, K = 5, K = 6 and K = 7 where first the coordinates of the company were obtained, as well as the coordinates of 50 of its clients, later it was calculated the Euclidean distance to obtain the smallest distances and group them with the corresponding centroid, finally, the results were plotted on a two-dimensional axis and then made the comparison with the Google maps. The Elbow method in this research was used to know with better precision how many clusters the K-means algorithm works best with, resulting in this case that the K-means algorithm works better with K = 4.**

***Keywords*— K-means algorithm, Elbow method, Euclidian distance.**

I. INTRODUCTION

The analysis of SMEs is a topic of interest which has become more relevant in recent years, generally SMEs are family businesses where knowledge is empirical and the business can pass from generation to generation, this being the way in which the one that these economic entities have operated over the years. SMEs are the ones that have the greatest impact on the country's economy because despite the belief that only large companies can help economic growth and stability, the growth of SMEs has been of great help for economic development, as well. as for the competitiveness of the country [1]. According to figures obtained by INEGI in 2018, there are a total of 111 thousand 958 small and medium-sized companies which represent 2.7% of market share, small and medium-sized companies, according to Inegi data, contribute 42% of the gross and general domestic product 78% of employment [2].

On the other hand, the retail or retail sector in Mexico encompassing the companies that constitute the sector of frequent consumer products where supermarkets and hypermarkets are located is one of the most important sectors in Mexico, not only in terms of the commercial activities that they develop, but also because of your participation in the gross domestic product, since according to figures from the national association of self-service and department stores they have represented about 3.1% on average in recent years [3].

However, the problem of grouping objects according to their attributes has been widely studied due to its applications in areas such as machine learning, data mining, knowledge discovery, recognition, and classification of patterns. The goal of grouping is to partition a group of objects such that the patterns in each group are similar. One of the most popular and widely used grouping methods is K-means, particularly because its implementation is relatively simple [4].

As [5] mentions, the Elbow algorithm is a method that helps researchers select an optimal number of clusters for the K-means algorithm, adjusting the model with a range of values for K, where K is the number of clusters. A graph is made in which the sum of the distances between each point and each cluster center is bought against the number of clusters, if the line in the resulting graph looks like an arm, then the elbow or the inflection point of the curve is a good indication that the algorithm and its data fit better with that cluster number.

Having said the above, this research corresponds to a study focused on Optimization in the allocation of routes for an SME through the design of the K-means method and the Elbow method through an Excel spreadsheet. The research was developed through two stages, the first of them being the development of the K-means method where first the coordinates of the 50 clients under study were collected and later the algorithm was developed with K = 3, K = 4, K = 5, K = 6 and finally with K = 7, later the data were grouped in A, B, C, D, E, F, G finally the results are plotted on a two-dimensional axis which are compared with a Google maps. the second stage is given by the Elbow method in which the sum of the assigned distances is made and subsequently the results obtained from K = 3, K = 4, K = 5, K = 6 and K = 7 are graphed to identify how many centroids the K-means algorithm works best with.

II. LITERATURE REVIEW

In this section a review of the state of the art on the topics of the K-means algorithm and the Elbow method in different applications is carried out to know the approaches and trends that the different authors have worked on these two tools having some perspectives and points of view on the behavior and results of the application of both methods.

[6] in their research made use of semi-supervised and unsupervised clustering algorithms to group similar sequences of enzymes, based on the k-means method with different values, as well as the implementation in Spark of four algorithms that group the enzymes agree on their function. These are based on transformations of existing methods such as global logic, K-means, and cluster assembly. With the proposal of four algorithms for clustering, 6 Clusters corresponding to the enzymatic activity were obtained. [7] presented the results of an empirical evaluation of two unsupervised algorithms to perform metagenomic binnin tasks, these being the EM vs K-means. These algorithms were tested for long and short sequences of a data set, the results obtained show that the K-means algorithm in general has a better performance than the EM algorithm. For their part, [8] in their research carried out a non-hierarchical K-means cluster analysis which generated two teacher profiles: constructivists and behaviorists, this because the objective of their research was to visualize the teaching of the teachers and their relationship with the use of the Moodle platform. [9] applied the K-means algorithm to analyze the data obtained from the results of the PLANEA 2017 test in upper secondary schools, obtaining through the algorithm 3 groups (K = 3) classifying the results as satisfactory, indispensable, and insufficient.

[10] implemented the K-means algorithm in thermal images, detecting that the implementation of this algorithm facilitates the detection of hot spots in this type of images. In his work [11] used the K-means algorithm together with the S3 algorithm to face the high dimension of the data set in the unsupervised classification of images, with the result that when using the K-means and S3 algorithms they can be performed classification processes in medical images in a relatively short time. [12] use the K-means algorithm for processing agro-industrial images, obtaining promising results which represent a theoretical and practical advance in the area. In the same way, [13] propose the K-means algorithm and the swarm intelligence algorithm as the method of segmentation of early blight disease in a tomato leaf, resulting in the performance of the segmentation method that uses the K-means, and the swarm algorithm significantly improves the results. For their part, [14] used the K-means algorithm to divide the data set into corresponding groups, with the result that with the application of the algorithm the data were grouped into 6 different clusters.

[15] apply the K-medias algorithm to determine the linguistic proximity between different national and international press media, obtaining as a result 4 different clusters. [16] in their work describe that the use of the K-means algorithm to classify brain activity in Norvegicus Wistar rats and that the clusters obtained are consistent with respect to the frequency and regularity attributes of the waves. For their part, [17] in their work analyzed five IR estimation methods using the K-means method classification on a population of 119 adults, with the result that the population was divided into two clusters C-N and C-RI. Added to this. [18] in their study used the K-means algorithm to classify the similarities between 23 countries of the European Union to classify occupational accidents based on data from forestry and logging. [19] use the K-means algorithm to determine the main geological parameters of tunneling, obtaining as results the grouping of data with K = 3. For their part, [20] propose the K-means algorithm to find if the values of their study in plants should be optimized for the best growth of the plant, obtaining precise results compared to a traditional method. [21] implemented the K-means algorithm to investigate the diagnostic heats of ultrasound, obtaining K = 2 clusters which were segmented into groups a and b, thus obtaining very precise data. [22] uses the K-means algorithm to group and discover the hidden characteristics of users of telecommunications fraud, obtaining as a result that thanks to the implementation of this algorithm they can identify fraudulent telephone numbers, as well as distinguish fraud facts. [23] implemented the K-means algorithm to observe the evolution of MERCOSUR based on the years between 1983 and 2015, obtaining as results that the 3 clusters efficiently comply with the objective of the research, being coherent and consistent results in terms of to the description made by specialists.

[24] used the K-means algorithm to analyze cities and rural areas with different latitudes to control fires in Australia, obtaining as results different grouping groups according to their characteristics. On the other hand, [25] makes a comparison between the K-means method and a self-organized neural network to determine which method is more effective in the grouping and classification of SO2 patterns, obtaining as a result that the K-means algorithm has an efficiency of 44% while the network of 56%. [26] use the K-means algorithm for customer segmentation according to their characteristics and behavior, with the results that the developed algorithm works quite well for a large data set. [27] propose in their work the implementation of the K-means algorithm together with the Spark algorithm to perform customer segmentation in e-commerce companies, obtaining as a result four different types of customers. Finally, [28] propose an unsupervised learning algorithm such as K-means so that it is free of initializations without parameter selection and that it can simultaneously find the optimal number of clusters, the results obtained are that the development algorithm shows good appearance regarding the grouping obtained.

[29] in their research propose the use of the Elbow method to determine the best number of clusters and determination of centroids based on the mean and median data, obtaining as results that when using the Elbow method, the iterations required are less than using the number of random clusters. On the other hand, [30] mentioned in his research that the value of

K is difficult to determine, and the initial center of the group is also complex to find, which is why he uses the Elbow method to find the most appropriate value of K obtained as a result that the proposed algorithm works properly with 3 centroids. [31] propose the Elbow method to determine the number of clusters in the K-means algorithm. The results obtained are two precise numbers of groups thanks to the implementation of the Elbow method. In addition to this, [32] applied the Elbow method to determine the optimal number of centers for each sample of input data, they obtained as a result that the number of centers obtained through the Elbow method shows better performance. [33] in their research used the Elbow method to improve the efficiency and performance of the K-means algorithm where the Elbow method searches for the best number of clusters to use, obtaining as a result that the number of optimal clusters for their research is K = 3. On the other hand, [34] analyzed four algorithms for selection of values for K, this is the Elbow method, Gap statistics, silhouette, and canopy coefficient, obtaining that the Elbow method finds K = 2 for the optimization of the algorithm. K-means likes the other methods used. For their part, [35] used the Elbow method to determine the number of optimal centroids in the K-means algorithm, having as results that the algorithm that uses the Elbow method has better performance than the one that does not use it anymore. that the calculated distances are much smaller.

III. METHODOLOGY

This section shows the methodology used for the case study; the methodology is made up of two sections. The first part contains the conceptualization of the case study made up of eight stages. The second part is made up of the development of two tools: the K-means algorithm which is made up of 5 stages and the Elbow method, which is made up of Finally, 3 stages have the conclusions.

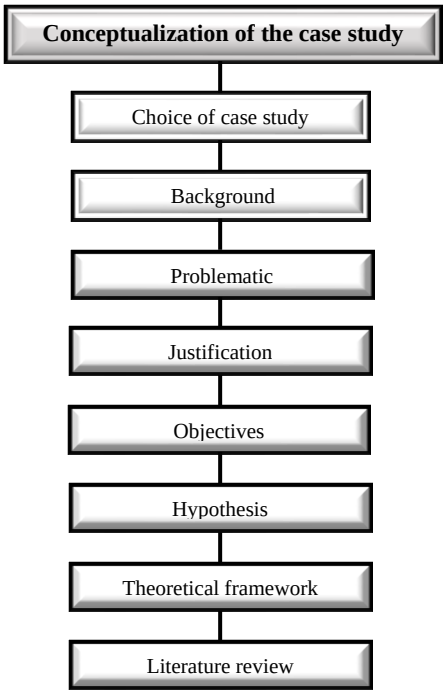

Fig. 1. Research methodology

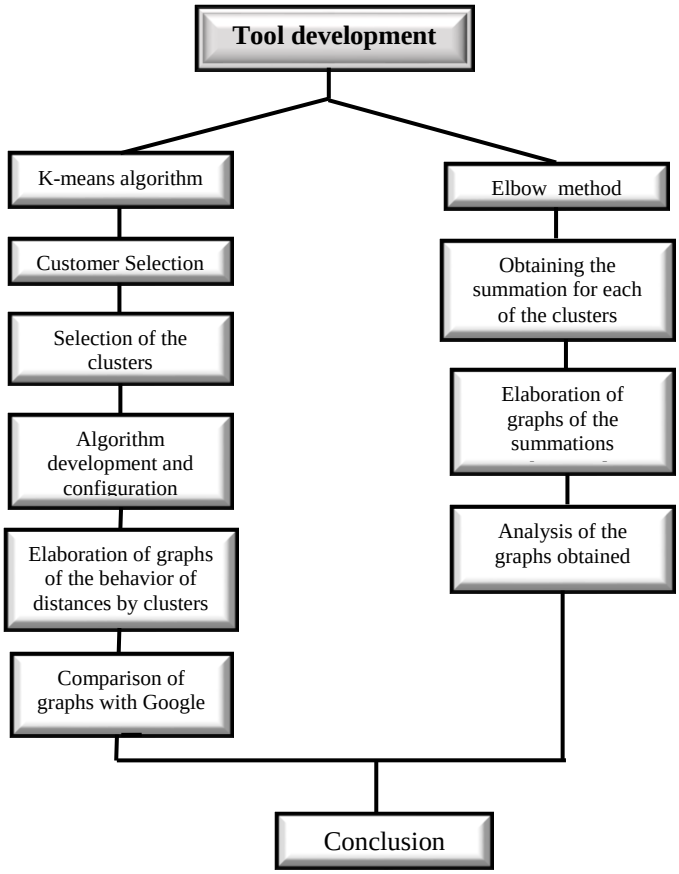

Fig. 2. Methodology of the tools used

IV. RESULTS AND DISCUSSION

This section shows the results obtained from the development of the K-means algorithm with 3,4,5,6 and 7 centroids, as well as the results obtained from the development of the Elbow method.

For the development and configuration of the K-means algorithm, first the centers or centroids were arbitrarily chosen for this case the 3 centroids were chosen, our first centroid belongs to client number 1 which has the following coordinates (20.08141, -101.23411), while our second centroid belongs to client number 20 with the following coordinates. (20.11504, -101.20694) finally the centroid number 3 belongs to client 25 with the coordinates (20.12617, -10120317).

After selecting the 3 centroids, the Euclidean distance is calculated, which helps us to calculate the distance between two points, in this case it is calculated between each of our clients with respect to the position of the centroid. The Euclidean distance is given by the following formula:

$$Euclidean\ distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Where:
x_2 It is the client´s coordinate
x_1 It is the coordinate of the centroid
y_2 It is the client´s coordinate
y_1 It is the coordinate of the centroid

According to the Euclidean distance formula we have the following examples of calculations corresponding to customer 5 with respect to centroid 1. In (1) it can be observed that $X_1 = (-101.2908)$ It corresponds to the length coordinate of our client number 5 while $X_2 = (-101.23411)$ corresponds to the longitude coordinate of our centroid number 1 and this result will be raised to the most square $Y_1 = (20.06886)$which corresponds to the latitude of the coordinate of our client number 5 minus $Y_2 = (20.08141)$ which corresponds to the latitude of the coordinate of our centroid 1 this result is also raised to the table to later add it with the first result obtained and finally take the square root obtained lastly a distance of 0.05806254.

$$Euclidean\ distance = \sqrt{((-101.2908) - (-101.23411))^2 + ((20.06886 - 20.08141))^2} = 0.05806254 \quad (1)$$

Calculation of client 20 with respect to centroid 2.

For this case our $X_1 = (-101.20443)$ corresponds to the longitude coordinate of our client number 19 while $X_2 = (-101.20694)$ corresponds to the longitude coordinate of our centroid number 2 and this result will be raised to the most square $Y_1 = (20.11313)$ which corresponds to the latitude of the coordinate of our client number 19 minus $Y_2 = (20.11504)$ which corresponds to the latitude of the coordinate of our centroid 2 this result is also raised to the table to later add it with the first result obtained and finally take the square root obtained finally, a distance of 0.003154077 as shown in (2).

$$Euclidean\ distance = \sqrt{((-101.20443) - (-101.20694))^2 + ((20.11313 - 20.11504))^2} = 0.003154077 \quad (2)$$

Calculation of customer number 45 whit respect to centroid 3

Finally the $X_1 = (-101.210311)$ corresponds to the longitude coordinate of our client Number 45 while $X_2 = (-101.20317)$ corresponds to the longitude coordinate of Centroid Number 3 and this result Will be squared plus $Y_1 = (20.111071$ corresponds to the latitude of the coordinate of our client Number 45 less $Y_2 = (20.12617)$ which corresponds to the latitude of the coordinate of our centroid 3 this result is also squared to later add it to the first result obtained and finally take the square root obtained, finally a distance of 0.016702505 as shown in (3).

$$Eucliadian\ distance = \sqrt{((-101.210311) - (-101.20317))^2 + ((20.111071 - 20.12617))^2} = 0.016702505 \quad (3)$$

In table number 1 you can see first the number of clients, as well as the longitudes and latitudes of each of our corresponding clients. At X and Y, you can then see the calculations of the Euclidean distance from centroids 1,2 and 3 for each of the 50 clients.

Table 1. Start of the K-means algorithm whit 3 centroids

			Initiation					
			Centroid 1		Centroid 2		Centroid 3	
Client	X	Y	-101.23411	20.08141	-101.20694	20.11504	-101.20317	20.12617
1	-101.23411	20.08141	0	1	0.04323408	0	0.05441269	0
2	-101.23688	20.0806	0.002886	1	0.04563461	0	0.05668323	0
3	-101.25076	20.07578	0.0175761	1	0.05883485	0	0.06931061	0
4	-101.25146	20.07608	0.01815025	1	0.05916005	0	0.06957681	0
5	-101.2908	20.06886	0.05806254	1	0.09573449	0	0.10470651	0
6	-101.29291	20.06483	0.06109285	1	0.09955845	0	0.1087008	0
7	-101.30724	20.0532	0.0783824	1	0.11783156	0	0.12710305	0
8	-101.30891	20.05586	0.0790433	1	0.11789891	0	0.12698206	0
9	-101.34132	20.05229	0.11109437	1	0.14830896	0	0.15666422	0
10	-101.34236	20.05119	0.1123891	1	0.14971773	0	0.15810078	0
11	-101.20423	20.11138	0.04232039	0	0.00455409	1	0.01482794	0
12	-101.20612	20.11103	0.04075272	0	0.00409298	1	0.01542472	0

Initiation								
Client	X	Y	Centroid 1		Centroid 2		Centroid 3	
13	-101.20815	20.11151	0.03974835	0	0.00373162	1	0.01548276	0
14	-101.20848	20.11117	0.03927537	0	0.00416515	1	0.01591214	0
15	-101.20819	20.11231	0.04033183	0	0.00300257	1	0.0147411	0
16	-101.20473	20.11286	0.0430382	0	0.00310427	1	0.01340111	0
17	-101.20615	20.11287	0.04208911	0	0.00230933	1	0.01362976	0
18	-101.20799	20.11244	0.04056002	0	0.00280401	1	0.01455147	0
19	-101.20443	20.11313	0.04344031	0	0.00315408	1	0.01310073	0
20	-101.20694	20.11504	0.04323408	0	0	1	0.01175116	0
21	-101.21121	20.10208	0.03084897	0	0.01364531	1	0.02539625	0
22	-101.21061	20.10261	0.03164949	0	0.01296047	1	0.02470682	0
23	-101.20109	20.12860	0.05759528	0	0.01476808	0	0.00319864	1
24	-101.20445	20.12985	0.0567992	0	0.01501786	0	0.00389625	1
25	-101.20317	20.12617	0.05441269	0	0.01175116	0	0	1
26	-101.20574	20.11769	0.04605535	0	0.00290904	1	0.00886089	0
27	-101.20744	20.12110	0.04781825	0	0.00608059	1	0.00662856	0
28	-101.210724	20.110390	0.03723903	0	0.00599509	1	0.01749489	0
29	-101.210692	20.109765	0.03677511	0	0.00647326	1	0.01804729	0
30	-101.210215	20.109325	0.03674532	0	0.00658687	1	0.01825886	0
31	-101.209903	20.108650	0.03644169	0	0.00704354	1	0.01876922	0
32	-101.209429	20.109151	0.03713105	0	0.00639339	1	0.01813343	0
33	-101.209168	20.108755	0.03701152	0	0.00666822	1	0.01841896	0
34	-101.209989	20.109967	0.03738081	0	0.00591876	1	0.01757942	0
35	-101.209993	20.109554	0.03706365	0	0.0062783	1	0.01796232	0
36	-101.209013	20.108919	0.03723714	0	0.0064625	1	0.01821367	0
37	-101.209439	20.109151	0.0371244	0	0.00639729	1	0.01813689	0
38	-101.208881	20.111268	0.03908967	0	0.00424211	1	0.01595886	0
39	-101.208886	20.111480	0.03924863	0	0.00405716	1	0.01576289	0
40	-101.20877	20.111031	0.03898231	0	0.00440609	1	0.01614085	0
41	-101.208793	20.111167	0.03906955	0	0.00429345	1	0.01602211	0
42	-101.209849	20.111167	0.03839369	0	0.0048438	1	0.01642252	0
43	-101.209715	20.111631	0.03883845	0	0.00439567	1	0.01594426	0
44	-101.210987	20.111293	0.03778448	0	0.00551527	1	0.01680567	0
45	-101.210311	20.111071	0.03802851	0	0.00520736	1	0.01670251	0
46	-101.20713	20.112814	0.04140207	0	0.00223409	1	0.0139307	0
47	-101.20711	20.11815	0.04559416	0	0.00311464	1	0.00893555	0
48	-101.20582	20.11792	0.04618771	0	0.00309011	1	0.00866516	0
49	-101.20826	20.10977	0.03837333	0	0.0054328	1	0.01717172	0
50	-101.21058	20.10258	0.0316517	0	0.0129808	1	0.02472643	0

In figure number 3 you can see the seed centroids, as well as the different clients that the company has, the centroids were assigned to each of the closest or similar data groups because theoretically the method says so.

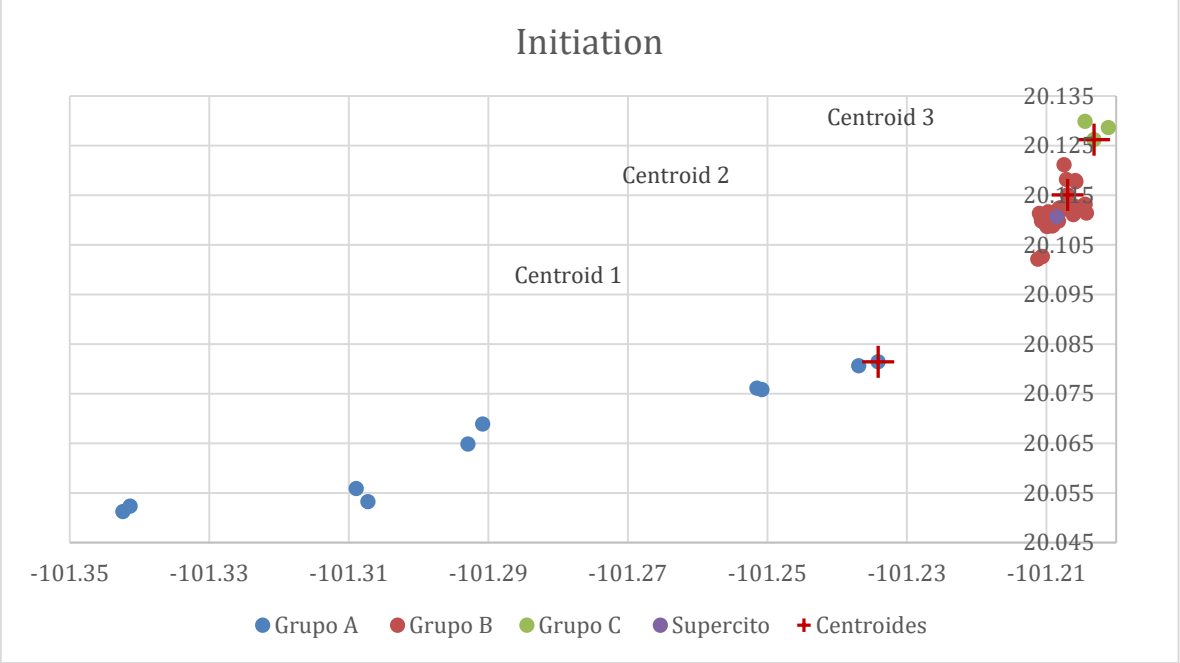

Fig. 3. Start of the K-means algorithm on a two-dimensional axis.

Iteration 1: It is necessary to remember the operation of the K-means method which tells us that it generates a partition of a set of n observations into k groups, each group is represented by the average of the points that compose it. The representative of each group is called the centroid, the k-means method begins with k randomly located centroids and this assigns each observation to the closest centroid. After the assignment, the centroids are moved back to the average location of the data assigned to it and the points are reassigned according to the new positions of the centroids.

In accordance with the above, centroid 1 will be calculated by the average of the smallest quantities from the beginning, more specifically, centroid 1 will be made up of the customer's average from 1 to 10 since these are the smallest distances obtained while our centroid number two will be confirmed by the average of the clients of the 11, 12,14,15,16,17,18,19,20,21,22,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50, Finally, centroid 3 is formed by the average of clients 23, 24 and 25, in this way the new centroids are obtained in a more graphic way, obtaining the average is given as follows:

Formula for X

$$\bar{X} = \frac{\Sigma X}{N}$$

Where:

ΣX: It is the sum of all the lengths of all our clients

N: It is the Number of customers whit the shortest distance

In (4) you can see the calculation to obtain the length of our first centroid. The sum of the lengths from 1 to 10 gives us a total of -1012.85675, this amount is divided by the number of clients, in this case it is between 10, resulting in a length of -101.285675.

$$\bar{X} = \frac{-1012.85675}{10} = -101.285675 \tag{4}$$

To obtain our latitude we have the following

Formula for Y

$$\bar{X} = \frac{\Sigma X}{N}$$

Where:

ΣX: It is the sum of all the lengths of all our clients

N: It is the Number of customers whit the shortest distance

In this case, the sum of the data with the least amount is 200.6601, said amount will be divided by 10 since it is the number of customers who obtained the shortest distance, obtaining a latitude of 20.06601 as shown in (5).

$$\bar{X} = \frac{200.6601}{10} = 20.06601 \tag{5}$$

According to the above, our first centroid is made up of a longitude of -101.285675 and a latitude of 20.06601. This procedure is carried out to compute each of our new centroids. Subsequently, the calculations of the Euclidean distance are carried out for each of our clients and for each of our 3 centroids, thus obtaining table number 2 in which we can observe very important things, the first one is an assignment by row that is to say, that the smaller distances obtained in each assignment will be assigned a number 1 which helps us to visualize to which centroids each of our clients is assigned, then a validation stage can be observed in which it is sought that all the data coincide by making a comparison with the previous iteration if the data coincide the algorithm will assign a 1 in case there is a data that does not match the algorithm will assign a 0 which also indicates that if there is a zero in the assignment then is not the same as the previous one.

Table 2. Iteration 1 of the K-means algorithm with 3 centroids

Iteration 1						
Centroid 1		Centroid 2		Centroid 3		Validation
-101.285675	20.06601	-101.208488	20.1111402	-101.202903	20.1282067	
0.053815511	0	0.03924744	1	0.05624752	0	0
0.05092956	0	0.0416989	1	0.05848768	0	0
0.036256174	1	0.05511123	0	0.07098462	0	1
0.035666106	1	0.05545983	0	0.07123861	0	1
0.005864139	1	0.09253564	0	0.10605589	0	1
0.007330595	1	0.09628952	0	0.11008089	0	1
0.025082769	1	0.11449446	0	0.12849957	0	1
0.025355231	1	0.11463173	0	0.12834116	0	1
0.057311469	1	0.14528464	0	0.15786866	0	1
0.058590286	1	0.14668224	0	0.15931017	0	1
0.093229421	0	0.00426499	1	0.01687888	0	1
0.091410056	0	0.00237081	1	0.01747526	0	1
0.08989091	0	0.00050113	1	0.01750161	0	1
0.089434298	0	3.0877E-05	1	0.01792616	0	1
0.090264141	0	0.00120718	1	0.0167527	0	1
0.093525481	0	0.00413303	1	0.015455	0	1
0.092304308	0	0.00290851	1	0.01567655	0	1
0.090502509	0	0.00139198	1	0.01656689	0	1
0.093920415	0	0.00451979	1	0.01515376	0	1
0.092753119	0	0.00419585	1	0.01377156	0	1
0.082741049	0	0.00946023	1	0.02741539	0	1
0.08351236	0	0.00879016	1	0.02673167	0	1
0.105224191	0	0.01896252	0	0.0018555	1	1
0.103310436	0	0.0191406	0	0.00225671	1	1
0.102109258	0	0.01594294	0	0.00205405	1	1
0.095186273	0	0.00710297	1	0.01089252	0	1
0.095685021	0	0.01001477	0	0.00843125	1	0

Iteration 1						
Centroid 1		Centroid 2		Centroid 3		Validation
0.087104746	0	0.00235828	1	0.01945755	0	1
0.086815611	0	0.00259766	1	0.02001895	0	1
0.08700805	0	0.00250535	1	0.02024791	0	1
0.086945762	0	0.00286406	1	0.02077158	0	1
0.087604785	0	0.00220048	1	0.02014206	0	1
0.087638211	0	0.00248021	1	0.02043559	0	1
0.087524788	0	0.00190493	1	0.01956763	0	1
0.087314633	0	0.00218643	1	0.01995458	0	1
0.08785354	0	0.00228239	1	0.0202322	0	1
0.087596082	0	0.00220477	1	0.02014531	0	1
0.089138123	0	0.00041301	1	0.01796249	0	1
0.089241646	0	0.00052311	1	0.01776439	0	1
0.089115527	0	0.00030033	1	0.01814932	0	1
0.089162753	0	0.00030593	1	0.01802882	0	1
0.088253821	0	0.00136102	1	0.01840088	0	1
0.088606982	0	0.00132128	1	0.0179207	0	1
0.087343274	0	0.00250342	1	0.01874614	0	1
0.087807894	0	0.00182407	1	0.01866828	0	1
0.091432661	0	0.00215552	1	0.01596242	0	1
0.094292305	0	0.00714397	1	0.01090104	0	1
0.09524426	0	0.00728592	1	0.01069217	0	1
0.088927048	0	0.00138912	1	0.01919908	0	1
0.083526187	0	0.00881211	1	0.02675177	0	1

In figure number 4 you can see how the centroids have moved since they no longer have the same position as the beginning. Let us remember that the K-means algorithm mentions that these centers move until they find the shortest distance and when this happens our centroids will no longer move. On the other hand, we can also observe how there is a reassignment of our clients in the three groups since the grouping in iteration 1 has changed as shown in graph 2.

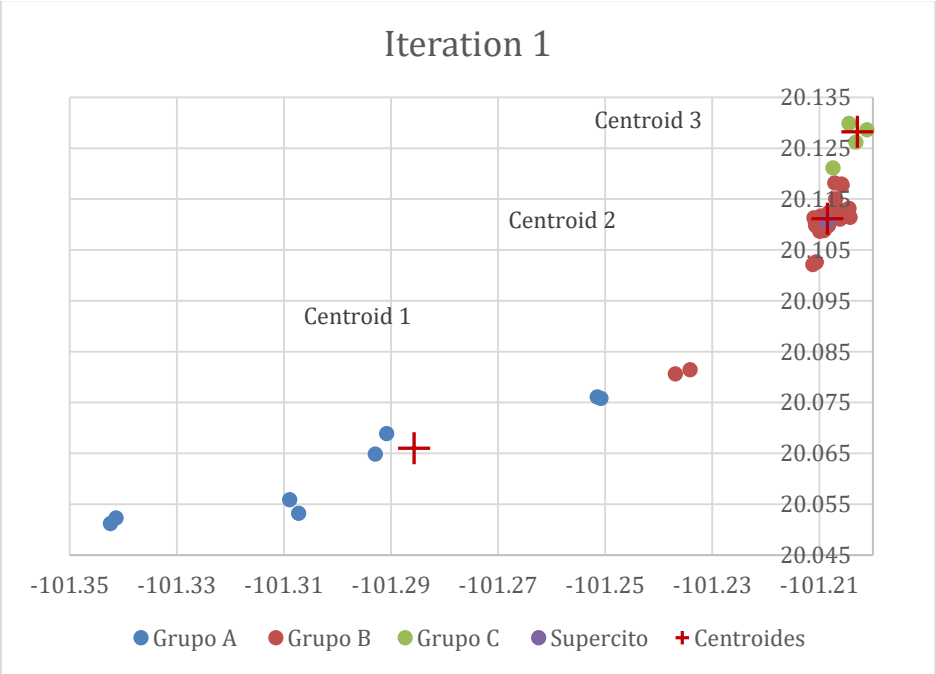

Fig. 4. Iteration 1 of the K-means algorithm on a two-dimensional axis.

Iteration 2. Next, iteration number two of the K-means algorithm with three centroids is presented. In the table, it can be seen in the validation part that there are still some zeros, which indicates that the assignment is not the same as the previous one and a new data reassignment has to be done again.

Table 3. Iteration 2 for the K-means algorithm with 3 centroids.

Iteration 2						
Centroid 1		Centroid 2		Centroid 3		Validation
-101.29822	20.0622613	-101.209937	20.1092921	-101.204038	20.12643	
0.066908645	0	0.03690166	1	0.05414015	0	1
0.064022694	0	0.03935922	1	0.05638279	0	1
0.049347829	1	0.05281626	0	0.06890874	0	1
0.048759158	1	0.05317125	0	0.06916658	0	1
0.009929748	1	0.09040763	0	0.1041251	0	1
0.005898693	1	0.09413478	0	0.10813363	0	1
0.012785408	1	0.11231273	0	0.12654402	0	1
0.01246002	1	0.11247486	0	0.12640556	0	1
0.044238398	1	0.14321546	0	0.15602315	0	1
0.045507276	1	0.14460857	0	0.15746165	0	1
0.106050798	0	0.00607717	1	0.01505123	0	1
0.104215167	0	0.00419424	1	0.01554017	0	1
0.102654977	0	0.0028484	1	0.0154764	0	1
0.102202414	0	0.002377	1	0.0158935	0	1
0.103006205	0	0.00348722	1	0.01471794	0	1
0.106304344	0	0.00631232	1	0.01358766	0	1
0.105062507	0	0.00521006	1	0.01372357	0	1
0.103244176	0	0.0037015	1	0.01453762	0	1
0.106696738	0	0.00671262	1	0.01330579	0	1
0.105440196	0	0.00648244	1	0.011754	0	1
0.095688416	0	0.00732352	1	0.02538439	0	1
0.096454827	0	0.00671586	1	0.02471012	0	1
0.11762256	0	0.02123839	0	0.00366014	1	1
0.115590017	0	0.02127764	0	0.00344479	1	1
0.114537465	0	0.01818405	0	0.00090562	1	1
0.107818814	0	0.00938839	0	0.00890427	1	0
0.108180437	0	0.0120691	0	0.00632344	1	1
0.099859534	0	0.00135071	1	0.01737788	0	1
0.099587936	0	0.00089069	1	0.01794449	0	1
0.099799181	0	0.00027971	1	0.01818633	0	1
0.099758752	0	0.00064299	1	0.01872251	0	1
0.100411605	0	0.00052745	1	0.01810061	0	1
0.100458586	0	0.00093818	1	0.01840456	0	1
0.100302283	0	0.0006769	1	0.01750573	0	1
0.100102986	0	0.00026779	1	0.01789602	0	1
0.100671915	0	0.0009967	1	0.01820414	0	1
0.100402762	0	0.00051783	1	0.01810359	0	1
0.101897588	0	0.00224051	1	0.01591684	0	1
0.101995338	0	0.00242736	1	0.01571657	0	1
0.101883015	0	0.00209546	1	0.01610922	0	1

Iteration 2						
Centroid 1		Centroid 2		Centroid 3		Validation
0.101926251	0	0.0021965	1	0.01598668	0	1
0.10100102	0	0.001877	1	0.01633195	0	1
0.101343511	0	0.00234946	1	0.01585069	0	1
0.100068521	0	0.00225958	1	0.01665606	0	1
0.100550405	0	0.00181776	1	0.01659083	0	1
0.104177582	0	0.00450383	1	0.01396277	0	1
0.106885848	0	0.00929817	0	0.00883168	1	0
0.1078687	0	0.00955995	0	0.00869468	1	0
0.101734374	0	0.001744	1	0.01718677	0	1
0.096469535	0	0.00674278	1	0.02473109	0	1

In the figure 5 shows how the clients, as well as the centroids, have moved from the place they previously had to take a position closer to the clients.

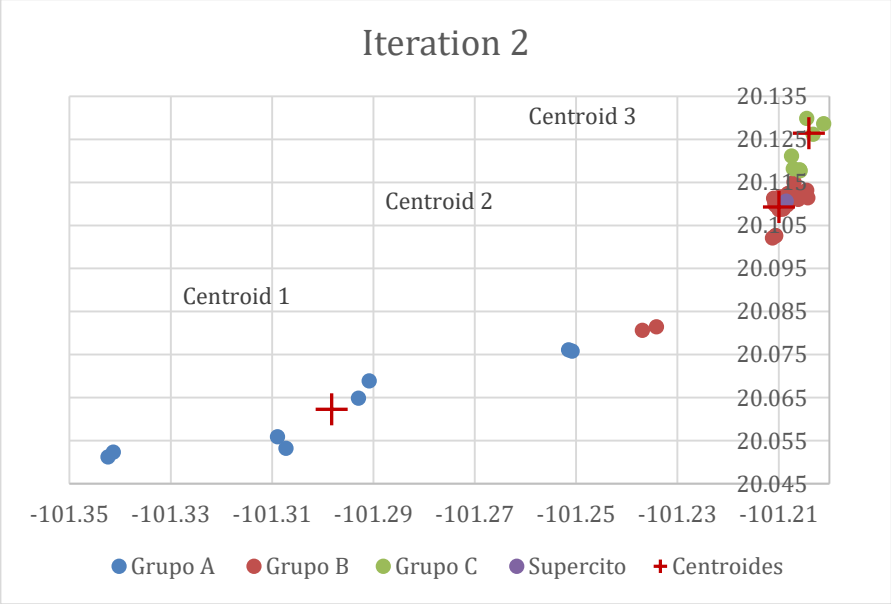

Fig. 5. Iteration 2 of the K-means algorithm on a two-dimensional axis.

Iteration 3 convergence of the algorithm: Table 4 shows the convergence of the K-means algorithm with 3 centroids because, as previously mentioned, with the validation, what is sought is that all the data coincide in comparison with the previous iteration. In this case, our validation for the 50 clients is 1, which means that the assignment made by the algorithm in this iteration is correct and the same as the previous one.

Table 4. Convergence of the K-means algorithm with 3 centroids.

Iteration 3						
Centroid 1		Centroid 2		Centroid 3		Validation
-101.29822	20.0622613	-101.210256	20.1085525	-101.204974	20.1227829	
0.066908645	0	0.03613518	1	0.0506024	0	1
0.064022694	0	0.03860317	1	0.05289015	0	1
0.049347829	1	0.05210229	0	0.06561707	0	1
0.048759158	1	0.05246209	0	0.06589445	0	1
0.009929748	1	0.08979367	0	0.1013594	0	1
0.005898693	1	0.09350623	0	0.10531488	0	1
0.012785408	1	0.11166863	0	0.12369337	0	1
0.01246002	1	0.11184454	0	0.12361756	0	1

Iteration 3						
Centroid 1		Centroid 2		Centroid 3		Validation
0.044238398	1	0.14263014	0	0.15349071	0	1
0.045507276	1	0.14402098	0	0.15492053	0	1
0.106050798	0	0.00665598	1	0.01142712	0	1
0.104215167	0	0.00482087	1	0.01180857	0	1
0.102654977	0	0.00363042	1	0.01171164	0	1
0.102202414	0	0.00316287	1	0.01213048	0	1
0.103006205	0	0.00428778	1	0.01095543	0	1
0.106304344	0	0.00700615	1	0.00992586	0	1
0.105062507	0	0.00595786	1	0.00998234	0	1
0.103244176	0	0.00449946	1	0.01077354	0	1
0.106696738	0	0.00740881	1	0.00966819	0	1
0.105440196	0	0.00728561	1	0.00798848	0	1
0.095688416	0	0.00654253	1	0.02162157	0	1
0.096454827	0	0.0059531	1	0.0209453	0	1
0.11762256	0	0.02204333	0	0.00699477	1	1
0.115590017	0	0.02207456	0	0.00708656	1	1
0.114537465	0	0.01898895	0	0.00383773	1	1
0.107818814	0	0.01019233	0	0.0051501	1	1
0.108180437	0	0.01285948	0	0.00298526	1	1
0.099859534	0	0.00189623	1	0.0136617	0	1
0.099587936	0	0.00128861	1	0.01421819	0	1
0.099799181	0	0.00077352	1	0.01444226	0	1
0.099758752	0	0.00036579	1	0.01496763	0	1
0.100411605	0	0.00102048	1	0.01434127	0	1
0.100458586	0	0.00110626	1	0.01464131	0	1
0.100302283	0	0.00143936	1	0.01376203	0	1
0.100102986	0	0.00103531	1	0.01414886	0	1
0.100671915	0	0.00129548	1	0.01444014	0	1
0.100402762	0	0.00101239	1	0.01434438	0	1
0.101897588	0	0.00304354	1	0.01215954	0	1
0.101995338	0	0.00323199	1	0.01196061	0	1
0.101883015	0	0.00289061	1	0.01234902	0	1
0.101926251	0	0.00299575	1	0.01222746	0	1
0.10100102	0	0.00264588	1	0.01259726	0	1
0.101343511	0	0.00312556	1	0.01211768	0	1
0.100068521	0	0.00283639	1	0.01296802	0	1
0.100550405	0	0.00251907	1	0.01287044	0	1
0.104177582	0	0.00528481	1	0.01019928	0	1
0.106885848	0	0.01009979	0	0.00510144	1	1
0.1078687	0	0.01036453	0	0.00493585	1	1
0.101734374	0	0.00233763	1	0.01342127	0	1
0.096469535	0	0.00598135	1	0.02096615	0	1

In the figure you can see the definitive position of the 3 assigned centroids, as well as the groups of clients which will no longer move, since at this point the algorithm has already assigned each client with the closest centroid.

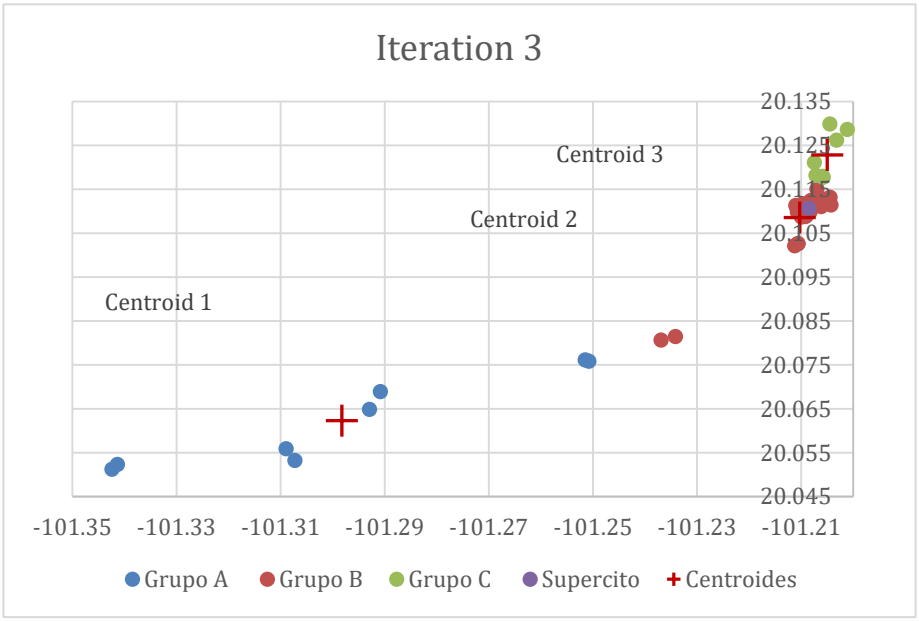

Fig. 6. Convergence of the K-means algorithm with 3 centroids on a two-dimensional axis

Route tracing and comparison with google maps. Figure 7 shows the tracing of the optimal routes that the company's delivery man can follow. In this graph there are 3 figures, one in black which represents the first route as well as the group A of our clients, later we have a figure in blue which represents the second route in which all the clients of group B are found, finally we have the route in green which represents the grouping of clients in group C and the route established for these clients. Finally, the three routes mentioned above are shown on a Google maps map. In figure number 8, it can be seen how the lines and locations of graph 5 coincide with those of the map, having a very good resemblance. This tells us that the algorithm with 3 centroids works optimally since when making the comparison between the graph and the map the results are identical.

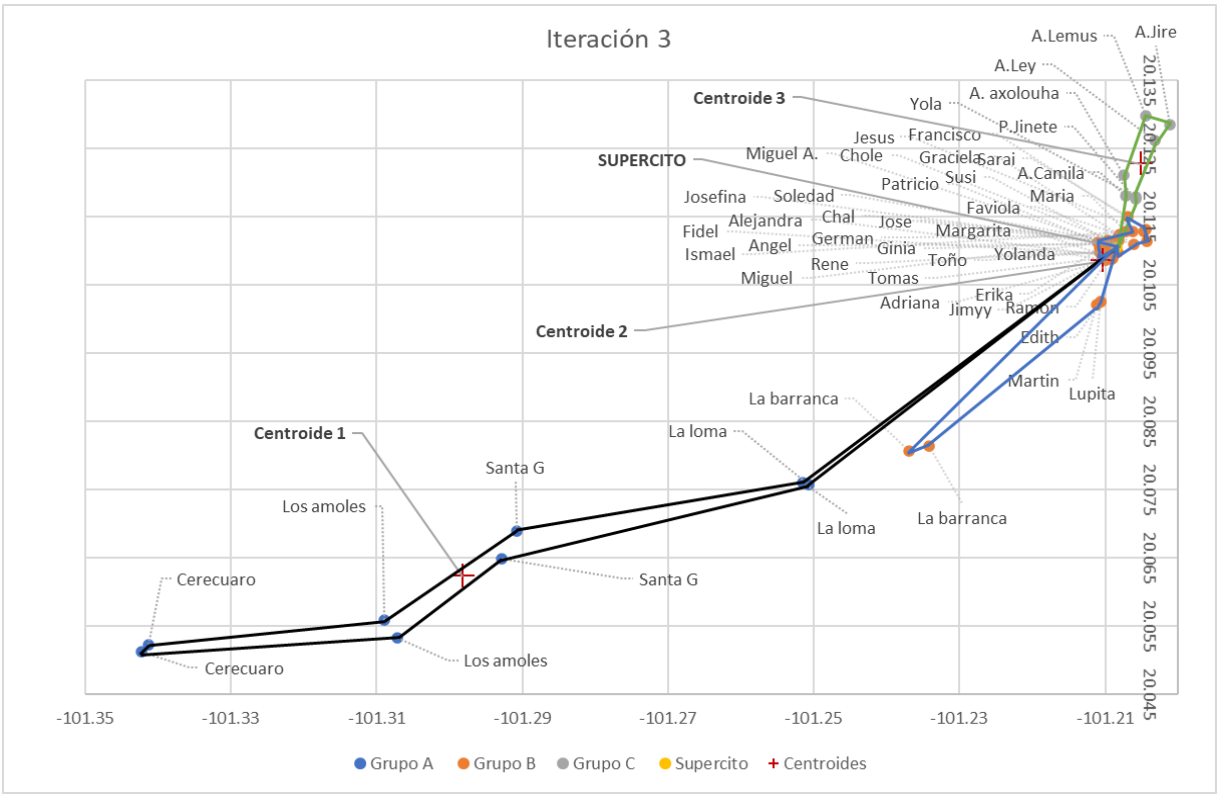

Fig. 7. plotting paths on a two-dimensional axis

Fig. 8. route tracing on Google maps

K-means algorithm with 4,5,6 and 7 centroids.

The process carried out above shows the development of the K-means algorithm with 3 centroids, this process is applied to the development of the K-means algorithm with 4, 5, 6 and 7 centroids. In addition to this, the graphs are presented and a comparison with Google maps, both in the graphs obtained and, in the maps, the routes obtained for each of the developed algorithms are shown, each figure drawn represents a route. On the other hand, regarding the comparison with Google maps in each of the cases, it can be seen that both the maps and the graphs are identical.

K-medias algorithm whit 4 centroids.

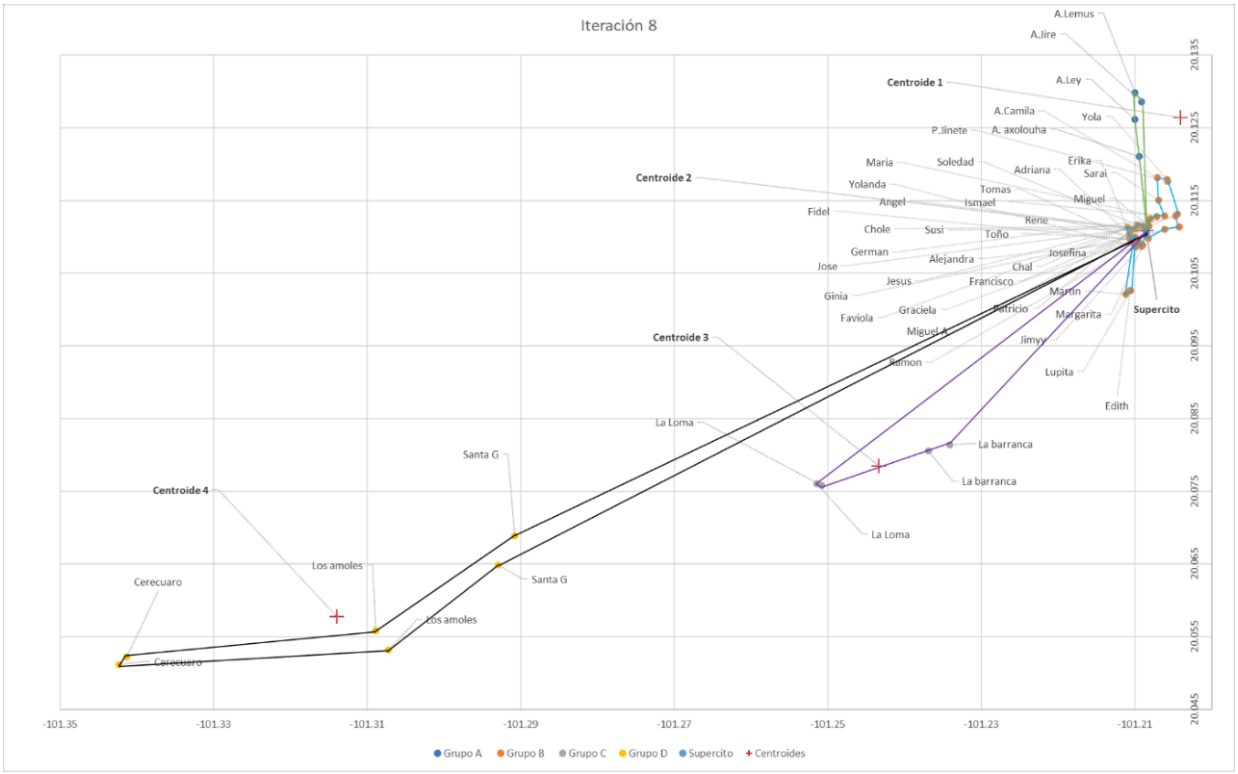

Fig. 9. plotting paths on a two-dimensional axis

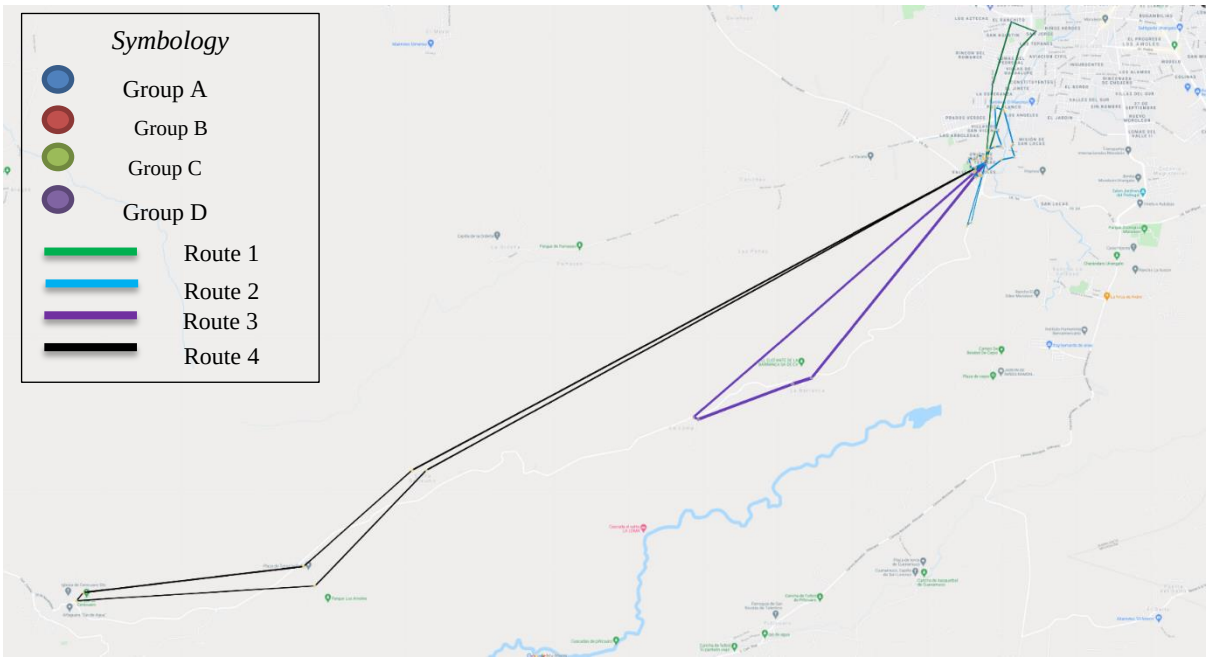

Fig. 10. Plotting routes for the K-means algorithm with 4 centroids

K-medias algorithm whit 5 centroids.

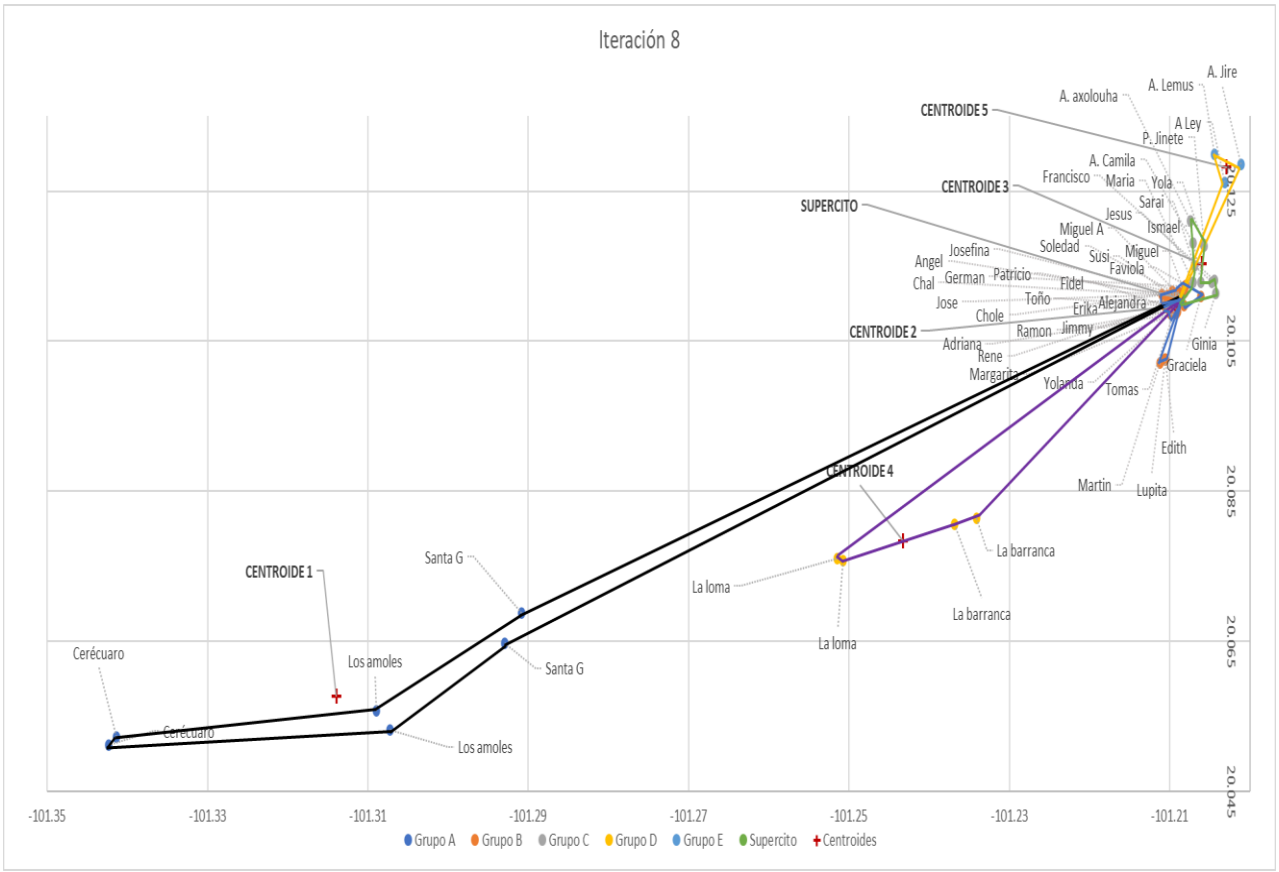

Fig. 11. plotting paths on a two-dimensional axis

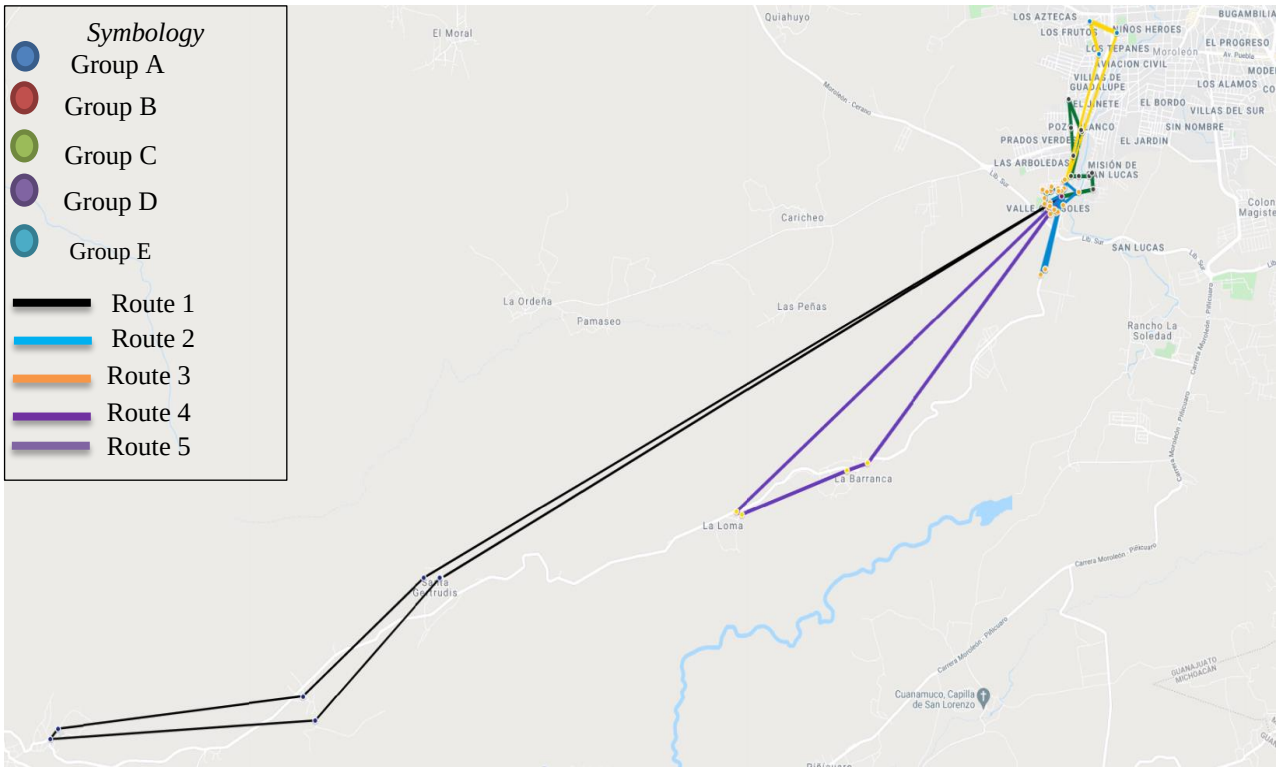

Fig. 11. Plotting routes for the K-means algorithm with 5 centroids

K-medias algorithm whit 6 centroids.

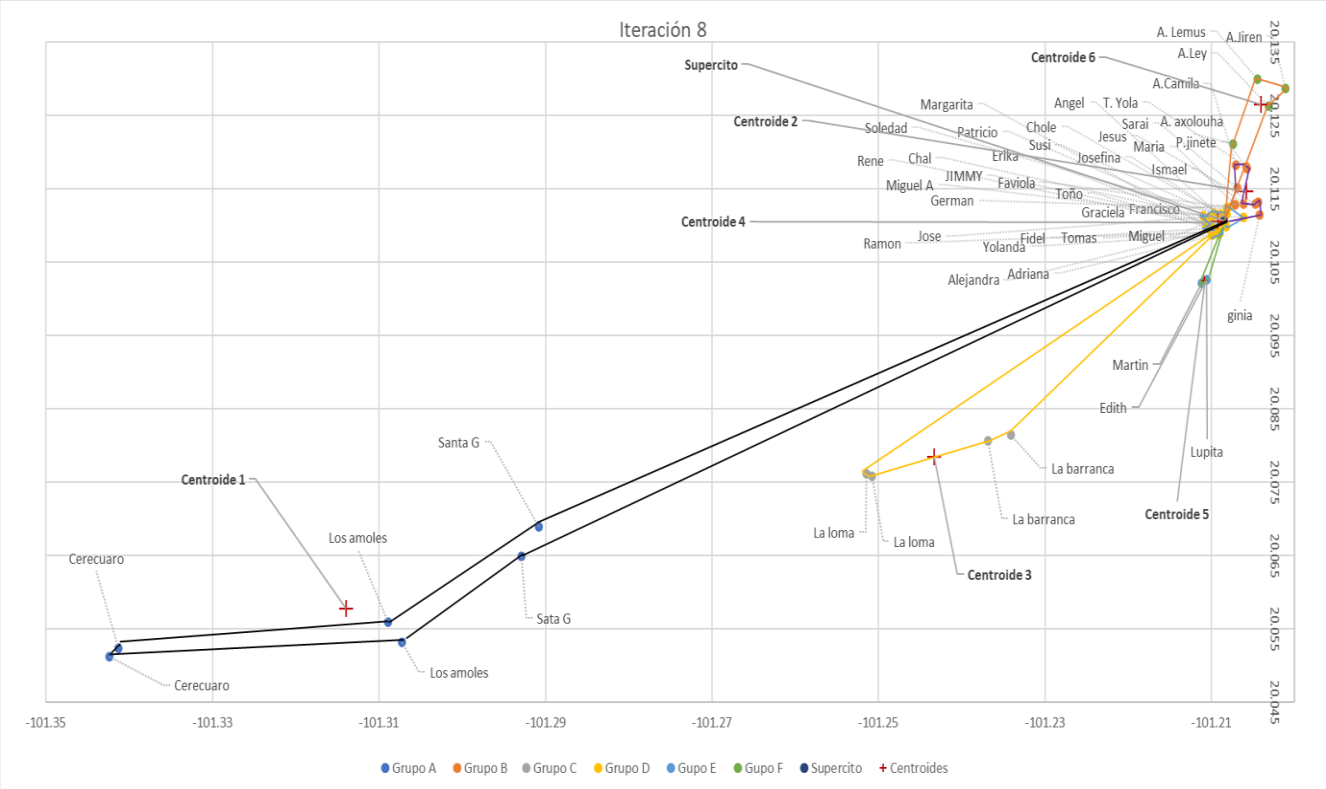

Fig. 12. Plotting paths on a two-dimensional axis.

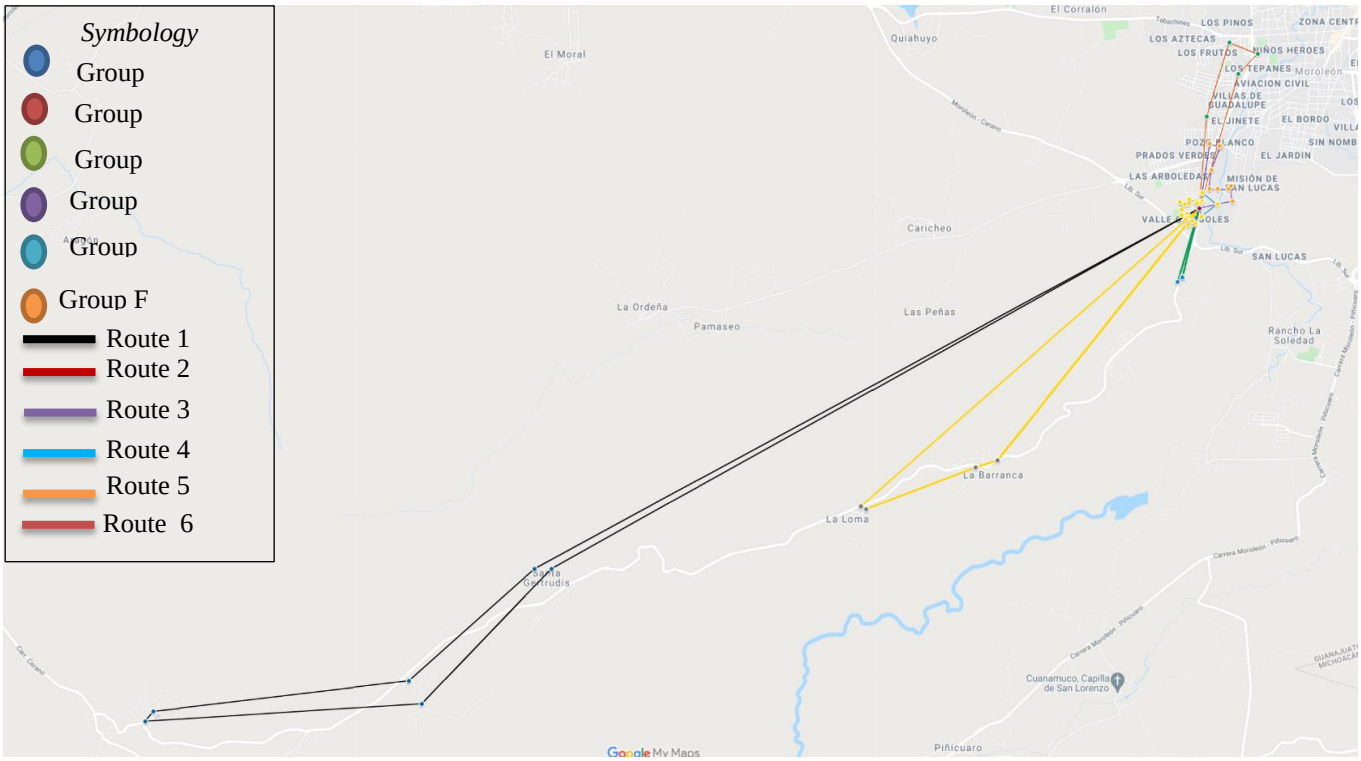

Fig. 11. Plotting routes for the K-means algorithm with 6 centroids.

K-medias algorithm whit 7 centroids.

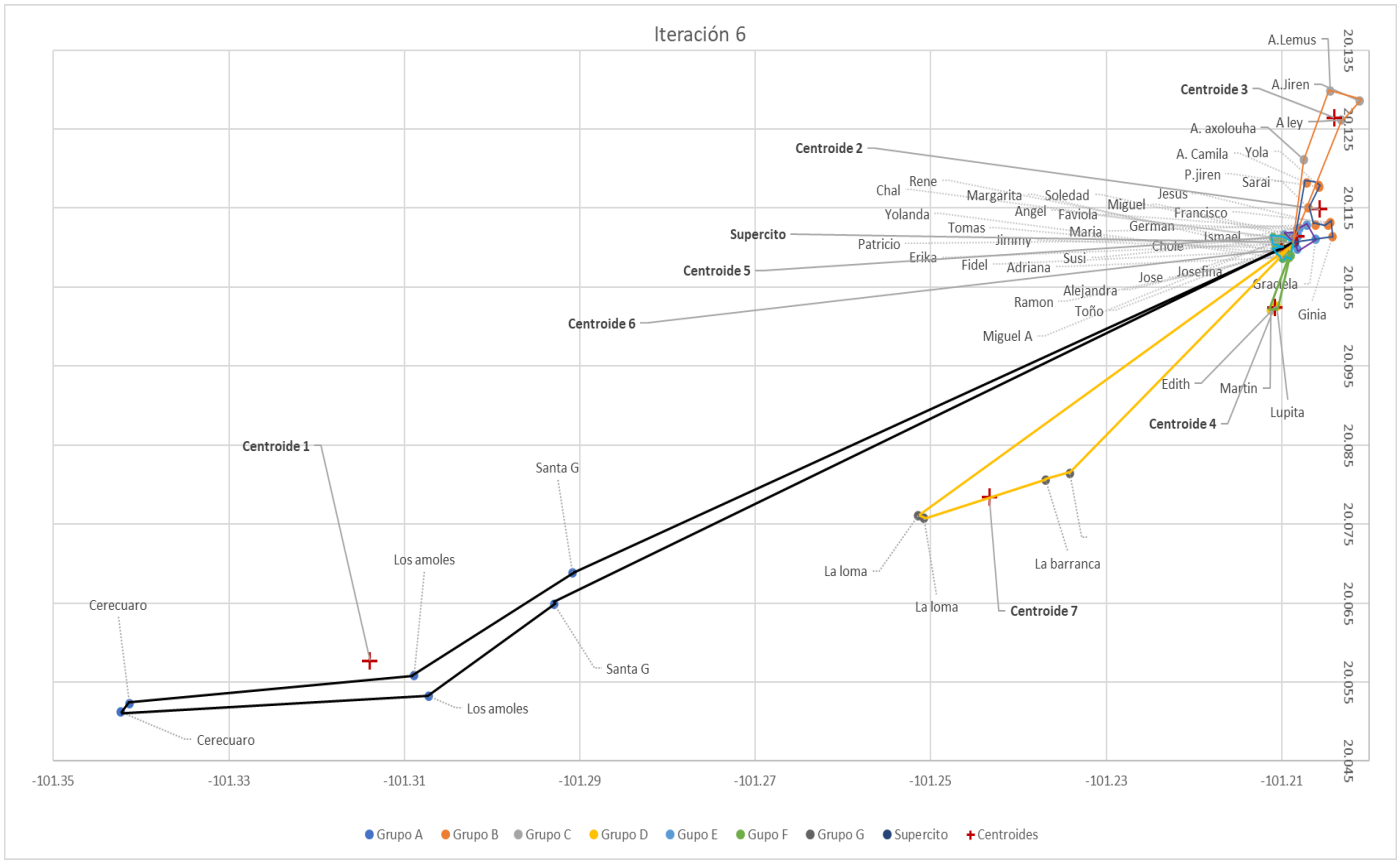

Fig. 12. plotting paths on a two-dimensional axis.

Fig. 13. Plotting routes for the K-means algorithm with 7 centroids.

Elbow method.

In this investigation the Elbow method was used to select an optimal number of clusters for the K-means algorithm adjusting the model with a range of values for K, for this case K has a value of 3,4,5,6 and 7 , in the following graph it can be seen that for this investigation the K-means algorithm works correctly with K= 4.

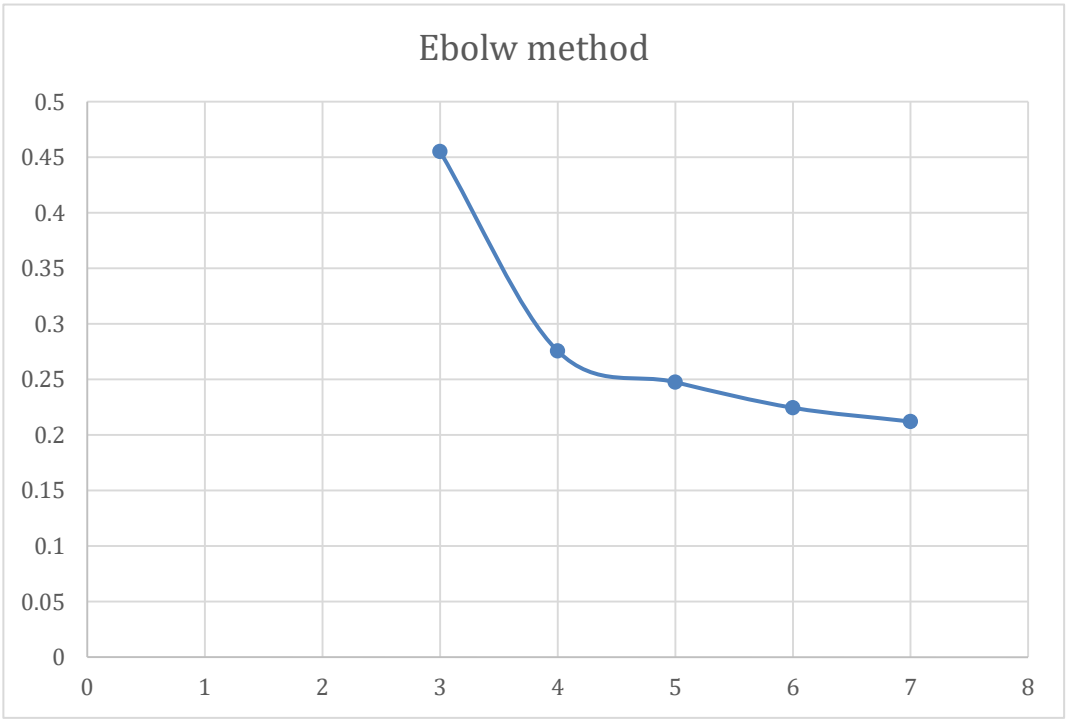

Fig. 13. Elbow method.

CONCLUSIONS

Today it is important that small and medium-sized companies use tools that help them generate a competitive advantage, which is why the K-means algorithm was developed for this research, as well as the Elbow method.

The K-means algorithm is an algorithm that allows discovering grouping in a data set, this algorithm has as its objective to generate a participation in a set of n observations with K groups and each group is represented by a centroid while the Elbow method allows us It helps to make the decision to determine with how many centroids the algorithm works better for the above and based on the results of the K-means algorithm and the Elbow method, this investigation was concluded and through the Elbow method it can be conclude that for this case the optimal result is when K = 4 because, as we could see in the Elbow method, the inflection occurs at 4, this being a favorable result for the company because it only has a delivery man. Thanks to the implementation of these two tools, the dealer will have 4 established routes to deliver their orders, generating savings for the company in terms of fuel.

REFERENCES

[1] Pérez, r., Beltrán, r. (2020, diciembre 12). Situación actual de las mipymes y su adaptación durante la pandemia . Ava Cient , vol. Xi, pp. 81-88.

[2] Instituto nacional de estadística y geografía. (2018). Encuesta nacional sobre productividad y competitividad de las mipymes.

[3] Rojas, j., campos, j. (2016, septiembre). Situación del sector retail y del subramo supermercados e hipermercados de la bolsa mexicana de valores en México 2016. Economía actual, vol. 3 , pp. 34-39.

[4] Pérez, j., et al. . (2007). Mejora al algoritmo de agrupamiento k-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer . Laio2, vol. 2 , pp. 1-7.

[5] Granda Cárdenas, a. A. (2019). Comparativa de extractores de características para clasificación de rostros. 70 hojas. Quito : EPN.

[6] González, y., Galpert, d., molina, r. & Aguero, g. (2020, octubre 20). Integración de rasgos y aprendizaje semi-supervisado para la clasificación funcional de enzimas utilizando k-medias de Spark. Revista Cubana de Ciencias Informáticas , vol. 14, pp. 134-161.

[7] Tapia, p., Meneses, v. (2018, agosto 06). An empirical comparison of em and k-means algorithms for binning metagenomics datasets. Revista chilena de Ingeniería , vol. 26, pp. 20-27

[8] Arancibia, m., Cabero, j., & Marín, v. (2020, Junio). Creencias sobre la enseñanza y uso de las tecnologías de la información y la comunicación (tic) en docentes de educación superior. Formación universitaria, vol. 13 , pp. 89-100.

[9] Gutiérrez, i., Gutiérrez, d., juan, j., rodríguez, l., rico, r., & Sánchez, m. (2020). Aplicación del algoritmo k-means para el análisis de resultados de la prueba planea 2017. Research in Computing Science, vol.149, pp. 407-419.

[10] M. R. S. Mohd, s. H. Herman and z. Sharif, "application of k-means clustering in hot spot detection for thermal infrared images," 2017 IEEE symposium on computer applications & industrial electronics (iscaie), 2017, pp. 107-110, doi: 10.1109/iscaie.2017.8074959.

[11] Sánchez, r. (2021). Clasificación no supervisada de imágenes médicas y minería de datos. Algoritmo s3 vs. K-medias. Revista cubana de investigaciones biomédicas, vol.40, pp. 1-19.

[12] Pham, t., lobos, g., Vidal, c., . (2018, agosto 25). Innovación en minería de datos para el tratamiento de imágenes: agrupamiento k-media para conjuntos de datos de forma alargada y su aplicación en la agroindustria. Información tecnológica, vol.30, pp. 135-142.

[13] Anam, s., Fitriah, z. (2021, febrero 23). Early blight disease segmentation on tomato plant using k-means algorithm with swarm Intelligence-based algorithm. International journal of mathematics and computer sciencie, vol. 16, pp. 1217- 1228

[14] Zhang, l., Deng, s., li, s. (2017). Analysis of power consumer behavior base on the complementation of k-means and dbs. can. IEE, 3, pp. 1-5.

[15] López, d., Fernández, a. (2018, mayo 30). Aplicación en los medios de prensa de un agrupamiento k-means (clustering k-means). Revista chilena de Economía y Sociedad, vol. 12, pp. 27-48.

[16] Carrillo, a., Garatejo, o., pineda, w. (2017, junio). Análisis multivariado de datos funcionales aplicado a curvas de encefalogramas. Comunicaciones en Estadística, vol. 10, pp. 129- 144

[17] Vintimilla, c., Astudillo, f., Severeyn, e., encalada, l., Wong, s. (2017). Agrupamiento de k-medias para estimación de insulino-resistencia en adultos mayores de cuenca. Tic. Ec, 2, pp. 32-39.

[18] Liu, x., he, q., Wang, y., Cong, q. (2021). Geological identification based on k-means cluster of data tree of shield tunneling parameters. Engineering Letters, vol. 29, pp. 432-437.

[19] Neethu, b., Jayanthy, s., Judeson, j. (2019). Greenhouse monitoring and controlling using modified k means clustering algorithm. IEE, 85, pp. 456- 462.

[20] Zhou, r., Zhou, r., Zhu, z. (2021, September 09). K-means clustering algorithm-based detection of carotid atherosclerotic plaque using Contrast-

Enhanced ultrasound images. Hindawi Scientific Programming, vol. 2021, pp.1-7.

[21] X. Min and r. Lin, "k-means algorithm: fraud detection based on signaling data," 2018 IEEE world congress on services (services), 2018, pp. 21-22, doi: 10.1109/services.2018.00024.

[22] González, h., Delbianco, f. (2019). Periodización de procesos económicos mediante k-means: aplicación para el caso del Mercosur. Asociacion Argentina de Economía política, 15, pp. 1-25.

[23] Tang, c., Zhang, h., Liu, s., Zhu, g., Sun., Wu, y., Gan, y. (2021, septiembre 14). Research on the setting of Australian mountain fire emergency center based on k-means algorithm. Hindawi mathematical problems in engineering, 2021, pp. 1-15.

[24] Hossain, s., Afroge, a., zaman, s. (2019). Rfm based market segmentation approach using advanced k-means and agglomerative clustering: a comparative study. IEE, 7, pp. 1-4.

[25] Deng, y., Gao, q. (2020). A study one commerce customer segmentation management based on improved k means algorithm. Information systems and e-business management (2020), vol. 18, pp. 497-510.

[26] Sinaga, k., Yang, m. (2020). Unsupervised k-means clustering algorithm. IEE access, 8, pp. 80716-80727.

[27] Umargono, e., Endro, j., Gunawan, v. (2019). K-means clustering optimization using the elbow method and early centroid determination based-on mean and median. Conrist, 3, pp. 234-240

[28] Cui, m. (2020). Introduction to the k-means clustering algorithm based on the elbow method. Accounting, auditing, and finance, 1, pp.5-8.

[29] Humaira, h., Rasydah, r. (2018). Determining the appropriate cluster number using elbow method for k-means algorithm. WMA, 2, pp. 24-25.

[30] Jeon, j., choi, j., Byun, h. (2017). Implementation of elbow method to improve the gases classification performance based on the rbfn-nsg algorithm. Journal of sensor science and technology, vol. 25, pp. 431-434.

[31] Syakur, m., Khotimah, b., Rochman, e., & Satoto, b. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. Materials science and engineering, 336, pp. 1-6.

[32] Yuan, c., Yang, h. (2019, Junio 18). Research on k-value selection method of k-means clustering algorithm. Multidisciplinary scientific journal, vol. 2, pp. 226- 235.

[33] Aslam, a., Qamar, u., Ayesha, r., Saqib, p. (2020). Improving k-mean method by finding initial centroid points. IEE, 16, pp. 624-627.