Bol. Soc. Paran. Mat.
©SPM –ISSN-2175-1188 ON LINE
SPM: www.spm.uem.br/bspm

(3s.) **v. 2024 (42)** : 1–10.
ISSN-0037-8712 IN PRESS
doi:10.5269/bspm.62231

Comparison Between Classification Algorithms Using a Multiple Correspondence Analysis with Python

Labdaoui Ahlam and Mekki Soundes

ABSTRACT: To analyse the data of a survey questionnaire we apply multiple correspondence analysis "MCA" as a method to help us convert data into a scatter plot, but it is difficult to study it and get good results, so we need to make a classification to facilitate the study. Among the most usable classification methods, hierarchical ascending classification and k-means. To compare them, we conducted a questionnaire on distance studies during the Corona crisis, which included the opinions of 304 university professors from most universities in Algeria; in our application, we used the Python programming language.

Key Words: Survey, MCA, classifications, Python.

Contents

1. Introduction

This increasingly complex world poses problems for us in the systems of aid and production, which the human being may not be able to solve. Encouraged humans to develop statistics, analyze data, and extract statistical information, through the graphical presentations provided that highlight the difficult relationships that should be understood through direct data analysis. We can divide the analysis of the data into two groups:

Factorial methods: They aim to reduce the number of variables while keeping the information as useful as possible, the importance of reading the data by highlighting the relationships between the variables, there are many ways in which: PCA, MCA, FCA....

2010 *Mathematics Subject Classification:* 62-07.
Submitted January 25, 2022. Published October 09, 2022

Classification method: k-means, hierarchical, etc. Our article consists of two parts in the theoretical part; we have highlighted the analysis of multiple matches plus a small overview of the two classification methods; k-means and hierarchy. In the practical part, we described the application we made on a dataset in Python and the results obtained [1].

2. Statistical Methodology

2.1. Definition of Multiple Correspondence Analysis

Abbreviated as MCA, is an extension of the correspondence factor analysis to summarize and visualize a data table containing more than two categorical variables. It can also be considered as a generalization of the main component analysis when the variables to be analysed are categorical rather than quantitative.
The MCA is generally used to analyze survey or survey data. The MCA starts from a complete disjunctive table (Burt's table) which presents the individuals online and in column all the modalities of the qualitative variables selected. The intersection boxes have a value of 1 if the individual meets the column criterion and 0 if not. As in PCA, the first two axes provide a generally important part of the information contained in the initial table (the horizontal axis being, by convention, the most significant). The proximity of the points indicates, a priori, their associations. The arrangement of the modalities of each variable in relation to the other aids to give a meaning to each axis (which is not always obvious, at the mere observation of the graph) [2].

2.2. Data and Ratings

The Multiple Correspondence Analysis (MCA) allows to study a population of I individuals described by J qualitative variables. A qualitative (or nominal) variable is an application of the set I of Individuals in a finite set on which no structure is considered: for example a set of three colors (blue, white, red). The elements of this Set are called modalities of the variable and it is said for example that a blue individual has the blue modality. The most common application of the MCA is the processing of all survey responses. Each question is a variable whose terms are the proposed answers (among which each respondent must make a unique choice). We begin by reviewing different ways of digitally transcribing all of this data.

2.3. Complete disjunctive table

Another way to present these same data is to build a Complete Disjunctive Table (CDT). In this table, the rows represent the individuals and the Columns represent the modalities of the variables: at the intersection of row i and column k, we find xi k which is 1 or 0 depending on whether the individual i has the modality k or not. The origin of the terminology "Complete Disjunctive Table" is the following: the set of values xi k of the same individual, for the modalities of the same variable, has the value 1 once (complete) and only once (Disjunctive). The columns in this table are numeric functions defined on all individuals called modality indicators [3].

2.4. Scatter plot analysis

Each individual in the scatter plot of NI individuals is represented by the modalities it possesses. As the margin is constant, the transformation into line profiles does not change the data. Thus the NI scatter plot belongs to a hyperactive cube noted HI of stops $\frac{1}{J}$, since the profile of a line is either 0 or$\frac{1}{J}$. An individual i is an Rk point which has to coordinate on the k axis the value $\frac{x_i k}{J}$ with an identical weight for each individual (because the margin is constant) of $\frac{1}{I}$. The GI barycenter of the NI scatter plot has to coordinate $\frac{I_k}{I_j}$j on the k axis. The resemblance between two individuals is defined by the modalities of each individual. If the two individuals have the same overall modalities, then they are similar. The distance that characterizes this resemblance between two individuals i and l is defined by:

$$d^2(i,l) = \sum_{k\in K} \frac{IJ}{I_k}\left(\frac{x_i k}{J} - \frac{x_i l}{J}\right)^2 = \frac{1}{J}\sum_{k\in K} \frac{I}{I_k}(x_i k - x l k)^2 \tag{2.1}$$

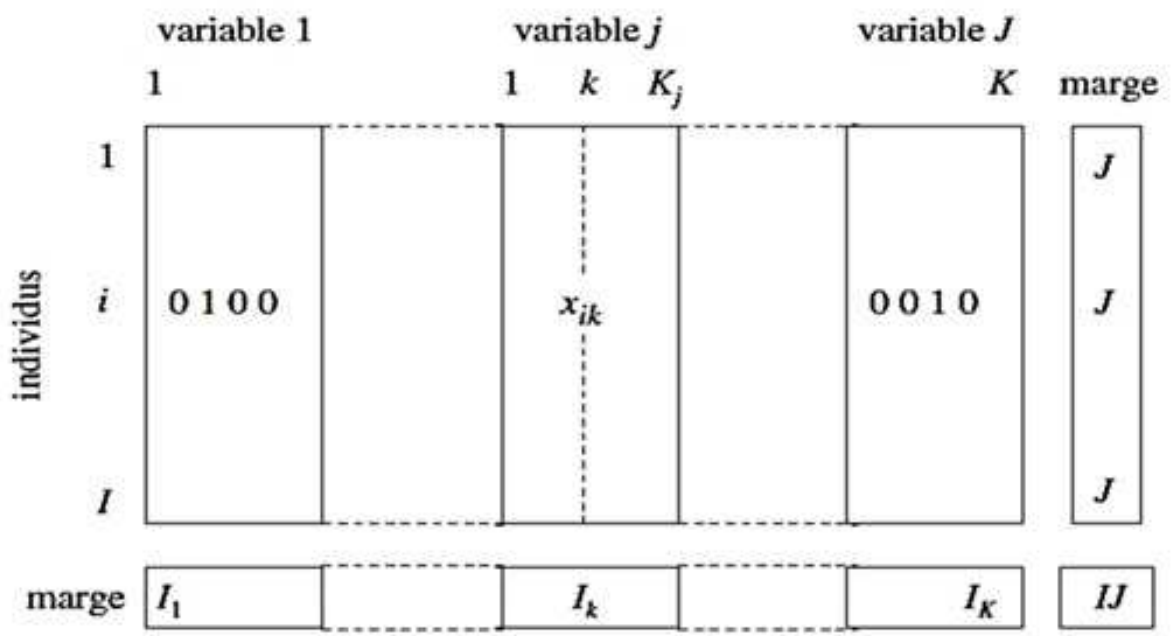

Figure 1: Data table in complete disjunctive form

$K_j = number\ and\ set\ of\ modalities\ of\ variable\ j$

$K = \sum_{j=1}^{j=J} K,\ j = number\ and\ set\ of\ all\ variable\ modalities\ combined.$

$x_i k = 1\ if\ the\ individual\ i\ has\ modality\ k\ and\ 0\ otherwise$

$\sum_{k=1}^{k=K} x_i, k = 1\ for\ all\ (i, j)$

$\sum_{k=1}^{k=K} x_i,\ k = J\ for\ all\ i$

$\sum_{i=1}^{i=I} x_i,\ k = I_k\ for\ all\ k$

$\sum_{k=1}^{k=K} I_k = I\ for\ all\ j$

2.5. Classification algorithms

In addition to the methods for determining the main axes, classification methods are the second part of the geometric analysis of the data. Like all methods geometric data analysis, they aim to summarize the initial data; to do this, they produce homogeneous classes of objects so that objects of the same class resemble each other as much as possible and objects belonging to different classes stand out as possible. In other words, we look for classes, which, on the one hand, each form a coherent whole (compactness) and, on the other, are distinct from each other (reparability). [4]

2.5.1. k-means. The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a data set. There are many types of clustering methods, but k-means is one of the oldest and most affordable. These features make the implementation of k-means in Python reasonably simple, even for novice programmers and data scientists.
The objective of the method is to partition the data into K groups and the value of K is set. The algorithm is relatively simple and it can be shown that at each stage of its execution, the value of W(C) is decreased.

1. We choose the number of K groups we want to obtain.
2. The n observations are randomly partitioned into K groups.
3. The coordinates of the centroids (the vector-mean) are calculated for each of the K groups, i.e. $\mu_k = \frac{1}{N_k} \sum_{i:c(i)=k} X_i, \quad k = 1, ...k$ ou N_k est le nombre d' observation dans le groupe k.
4. The distance between each observation and each of the K vector-means is calculated
5. Each of the n observations is assigned to the group with the closest mean vector.
6. Repeat steps 3-5 until no observations are reassigned to a new group.

2.5.2. Ascending Hierarchical Classification. The HAC makes it possible to build an entire hierarchy of objects in ascending order. We begin by considering each individual as a class and try to merge two or more appropriate classes (depending on the similarity) to form a new class. The process is iterated until all individuals are in the same class. This classification generates a tree that can be cut at different levels to obtain a larger or smaller number of classes.
The hierarchical ascending classification algorithm is very simple. It is due to Lance and William (1967)
Initialization: Construction of the distance table, regardless of the formula used to build it because the HAC algorithm is independent of the metric used. Thus, between each couple of point (x, y) of M, we have a value d(x,y). The initial score is the finest ρ_0 of M.
Grouping: Browse the distance table to determine the nearest element torque
(x^*, y^*) $d(x^*, y^*) \leq min_{x,y \in M} d(x, y)$

We combine the two elements in the same class $A = x^* \cup y^*$ the other classes remain unchanged. We get a new score π_i less than the previous one. [5]

3. Application

3.1. Problematic

Before the Covid-19 outbreak, we had no idea about distance education and had never studied remotely at our university, but after this crisis and continue the study, the intervention of the Ministry of Higher Education and Scientific Research imposed distance education on Algerian universities, and we chose the subject of distance learning because we wanted to deal with a subject in which we live as students. As we developed a questionnaire, asking questions about distance learning during the Covid-19 period. In this article, we chose to question only the professors from most of the Wilaya of Algeria, so we sent more than 1000 professors of the Algerian University via their e-mail, we found the interest of the professors and we received 304 replies, which is sufficient to study the statistics and disseminate them to all professors of the Algerian University. This topic was treated with MCA and two different methods of HAC and k-means classification that were used to find the best of them in order to select it as the best classifier that could

be used in our statistical studies. Of course, statistical study needs a program, and we did not use R or MATLAB, but instead we used Python because it was not used at all in our university. We wanted a new program where we have to know it and learn it for the first time in this job. The version uses is python 3.9, the distribution used is anaconda with the following packages: numpy, pandas, matplotlib, fanalysis. [6], [7].

3.2. Questionnaire Coding

1. What is your field of instruction? → DE
2. How old are you? → Age
3. What state do you work in?(Wilalya) → W-dt
4. Do you have prior experience with distance learning platforms prior to the Covid-19 pandemic? → EXP
5. What method did you use most to communicate with students online? → METH
6. In your opinion, will the wave system reduce the negative impact of this pandemic on university education? → S VAG
7. How did you give your courses to students? → COUR
8. What are the most important difficulties you have encountered when preparing and inserting your online courses? → DIFF
9. Have you asked your students to prepare and send home assignments remotely? →DM
10. What percentage of students sent it? → DM %
11. How do you assess the workload in the preparation and presentation of online courses in relation to the classroom? → Chrg-dT
12. How were the results of your students in your test compared to previous years? → RESLT
13. What is your opinion on distance learning in general as a teaching programmer? → OPN
14. If your answer is "Good program", what are your suggestions for improving this technique in New Algeria? → OPN+
15. If your answer is "Bad program", why? → OPN

3.3. Multiple correspondence analysis

3.3.1. Choice of factor plan. 2.

The elbow criterion applied to the eigenvalues allowed us to choose the first two factorial axes to carry out our analysis. Indeed, we observe on the histogram of eigenvalues a jump. Here, we have 5.25

3.

3.3.2. Graphical representation of modalities and individuals.

The modalities. According to this scatter plot, we notice that the modalities "DIFF_Di -Cr, Mqcnt, Abs-Com", “W-dT_Bsk”, “W-dT_khn”, “chrg-dT_S-Ch”, “RSLT_Pir”, “METH_PF, D-APP, YT”, “DIFF_Di -Cr” are well projected into our factorial plan i.e. they are good representatives. At the same time, these modalities contributed to the construction of this plan.

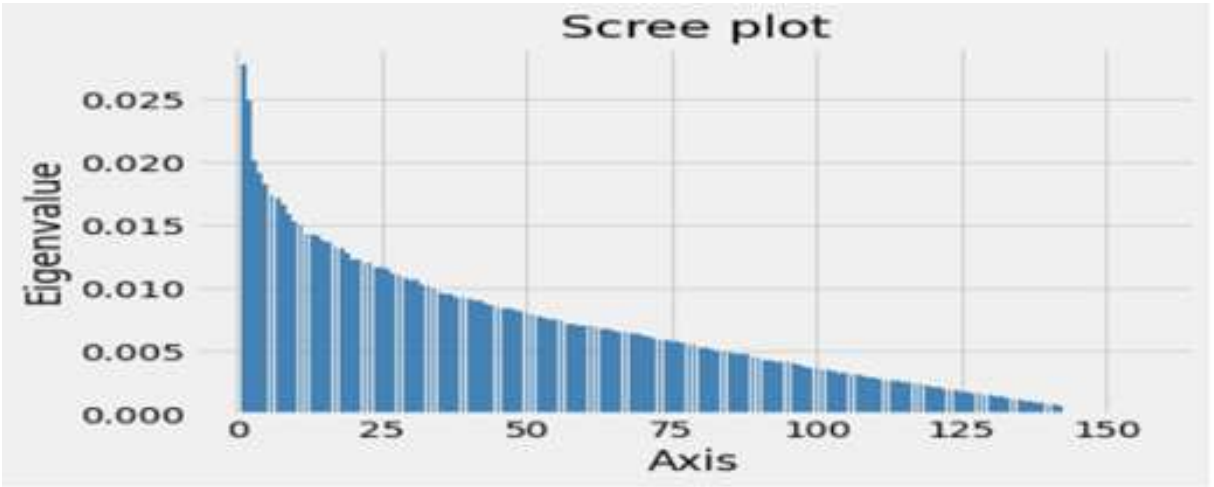

Figure 2: The eigenvalues

	Val.P	%	Cumul %
1	2.770805e-02	2.770805e+00	2.770805
2	2.483482e-02	2.483482e+00	5.254287
3	2.005438e-02	2.005438e+00	7.259724
4	1.916723e-02	1.916723e+00	9.176447
5	1.823446e-02	1.823446e+00	10.999893
..	...	...	...
153	5.699503e-33	5.699503e-31	100.000000
154	5.643840e-33	5.643840e-31	100.000000
155	5.138212e-33	5.138212e-31	100.000000
156	4.871448e-33	4.871448e-31	100.000000
157	4.564159e-33	4.564159e-31	100.000000

[157 rows x 3 columns]

Figure 3: Values of some eigenvalues

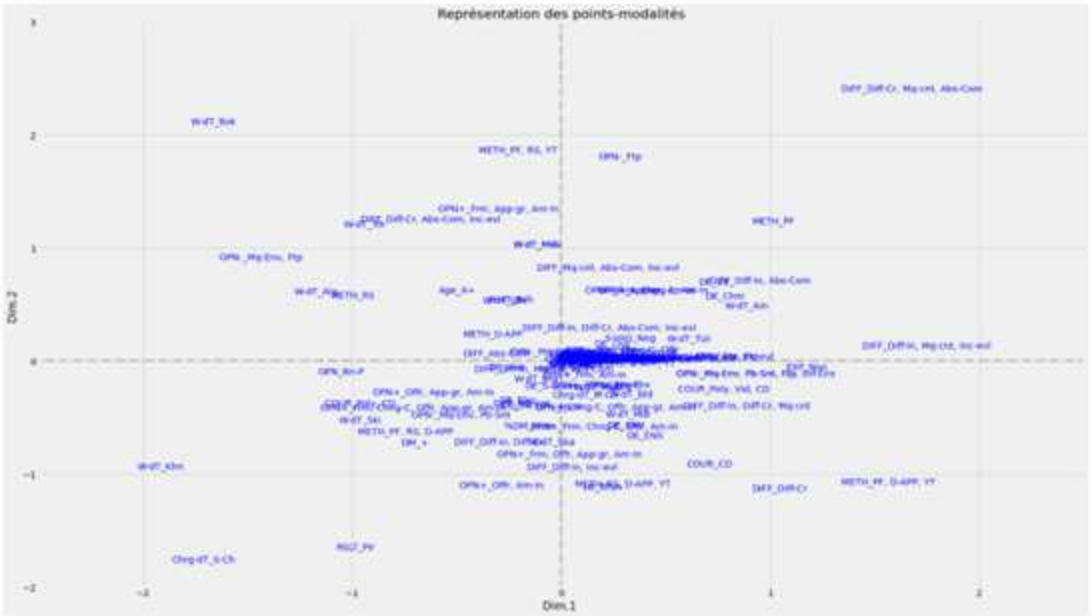

Figure 4: Representation of modalities

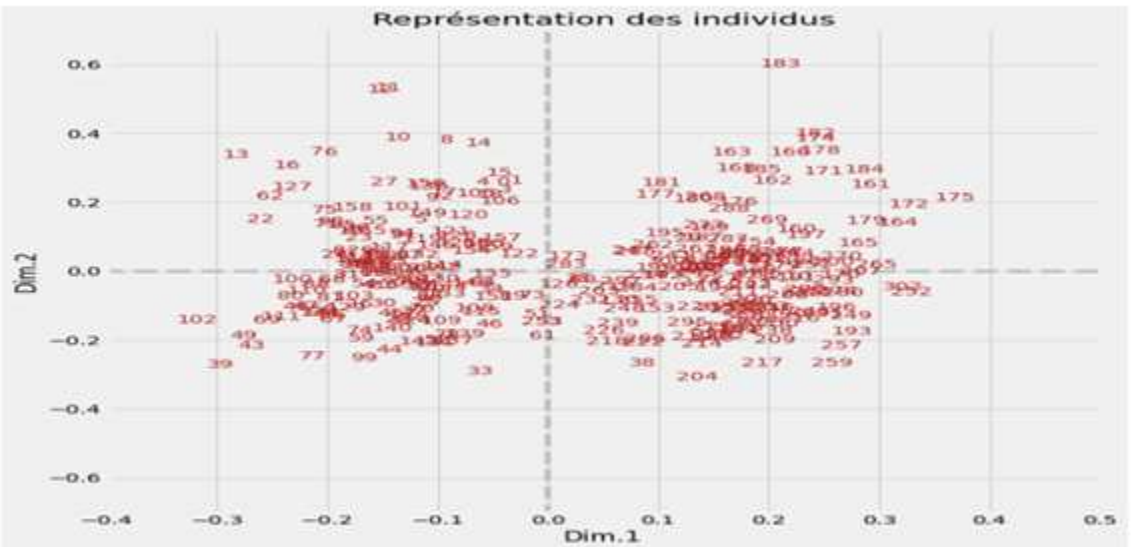

Figure 5: Representation of Individuals

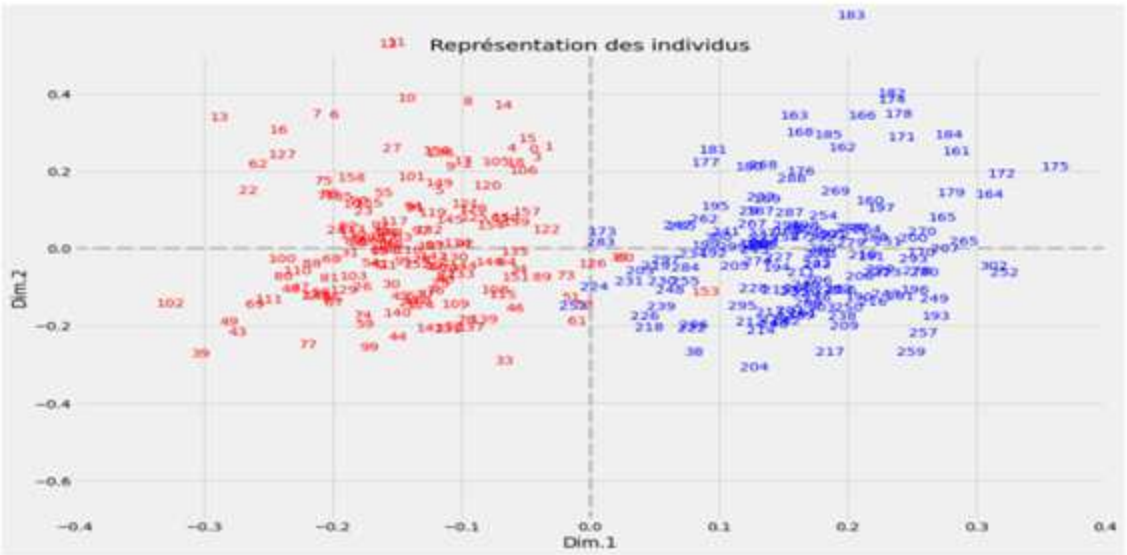

Figure 6: Classification of individuals by k-mean method

The individuals. At the first sight of the scatter plot, we notice the separation of individuals into two groups, separated by axis 1. The group on the left contains the professors were over 45 years old, who said that the results of the students were poor; the workload to prepare and deliver the courses was similar to that of classroom education. The right group contains the teachers have a lack of experience in distance education and several difficulties encountered when preparing and inserting online courses. Teacher number 11 and teacher number 12 work in Biskra, they see distance learning as a bad program because of lack of interactive study environment for students and the technical weakness of the country. Teacher number 183 has no prior experience, and according to him, distance education is a bad program just because of the country's technical weakness.

3.4. Data Classification

3.4.1. K-means. 6.

Group in red. This group presents the oldest teachers, and among them who work in Algiers, Skikda, Khenchela, Biskra, and who have experience in distance education, this is why they see that this teaching technique is good and the workload in the preparation and presentation of online courses compared to the classroom is similar. But they see that it is recommended to train students and teachers on this new technique because the results of their students during exams were catastrophic, although the method they most used to communicate with online students is by social networks and in their consideration going back to their lock experience in this technique.

Group in blue. This group presents the least aged professors, and among them who work in Tizi Ouzou, Blida, Ain Defla, and who have no experience in distance learning, that's why they see that the workload in the preparation and presentation of online courses compared to the classroom is more loaded. The method they had most used to communicate with students online is the platform and unfortunately, they encountered difficulties in the preparation and insertion of their online courses: Lack of direct communication with students, Internet difficulties, inability to evaluate students at a distance, difficulties in delivering course content and lack of control over the technology. Moreover, their opinion on distance education in general as a teaching program is poor because of the technical weakness of the country.

3.4.2. Hierarchical ascending classification. It appears that, the classification obtained with HAC differs from the one we did previously

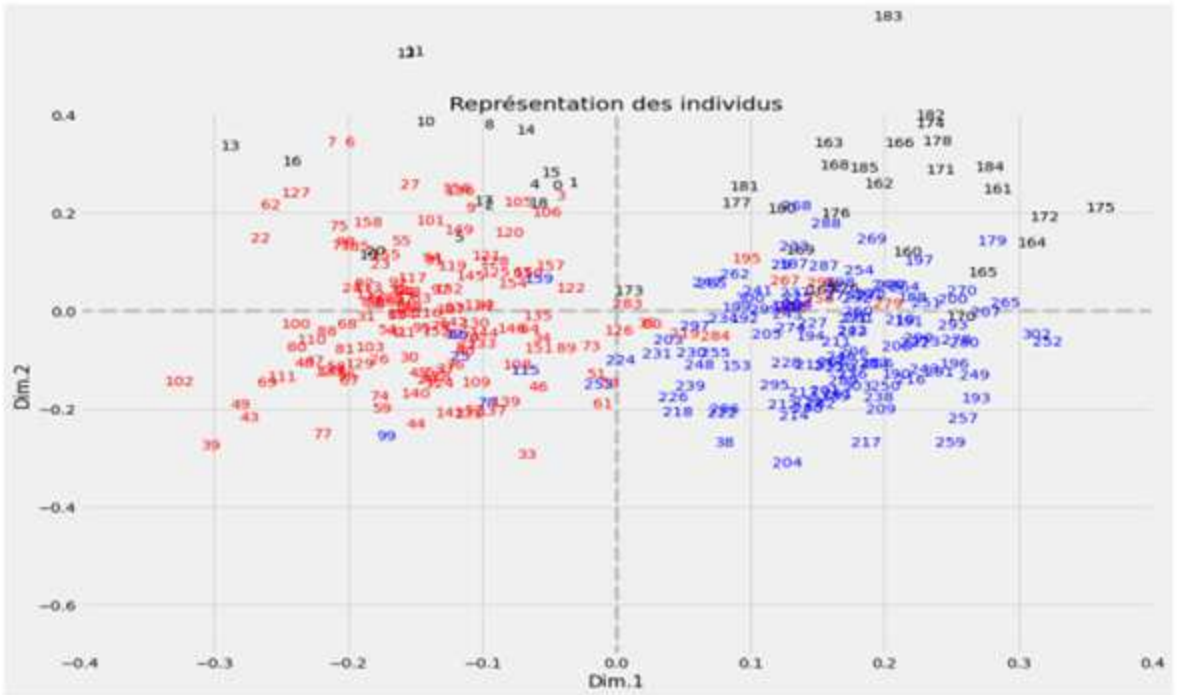

Figure 7: Hierarchical ascending classification of individuals

Group in blue. They are teachers who have no experience in this technique and among them work in Media and Belida. They prefer to give their courses in a CD and the method they had most used to communicate with students online is the platform, YouTube and other applications (Zoom, Google classroom ...). Moreover, the most important difficulties they had encountered when preparing and inserting their courses online is the lack of direct communication with students, the inability to evaluate students at a distance, lack of control over the technology.

Group in red. Most of the teachers here are workers in Khenchela and Skikda. They give their courses as handouts and CDs. They see that this program (distance learning) is good because they do not encounter difficulties in preparing and inserting their courses online and the workload was similar.

Group in black. Most of the teachers here are workers in Biskra, Mila and Tiaret. They see that this program (distance learning) is bad because they encounter difficulties in preparing and inserting their courses online including lack of direct communication with students, Internet difficulties, inability to evaluate students at a distance, difficulties in delivering course content, lack of control over technology. For the method that have preferred to teach its courses on YouTube and social networks. In addition, the main reason why they say it is a bad program is that our country suffers from technical weakness.

4. Discussion

Through our study, we noticed that the teachers' opinions about distance learning were somewhat similar; the students' results were similar to the results of previous years or lower, and even these were better because of the easier exams. One of the most important difficulties that teachers face in distance education is the weakness of internet, and the lack of means, especially for students, which was an obstacle to communicate with them, and would not have prevented the ability to evaluate them. In addition to all this, they encountered difficulties in preparing and presenting the courses, especially for those who have no previous experience on this teaching technique. Teachers' opinions on distance education are still somewhat positive, but several improvements must be made from improving the quality of internet and

providing internet offers for teachers and students and it is better to provide free applications on Google Play Store containing all courses for each specialty, and very necessary to train students and teachers on this new technique. Nevertheless, the success of this mode of teaching requires the availability of adequate technological equipment and a high-speed internet connection. Nevertheless, these conditions are not met in our country; these shortcomings were among the most important reasons that led the rest of the professors to have a negative opinion about distance learning.

5. Conclusion

The goal of our work is to create a framework for comparing statistical classifications on a real data set (survey results: remote study) using Python, where automatically identifying similar data sets in a large data set is an important part of data mining. Automatic classification seeks to group data into clusters so that the data are more similar to each other within the same group than between groups. Since the concepts of similarity and grouping can be explained in several ways and among the existing methods, we have highlighted HAC and K-means to find out the most important differences between them and which is the best:

k-means. It is considered the most widely used in big data analysis and increasing the number of k-groups increases its efficiency in classification, easy to understand and use, it collects similar results and displays the results quickly.

Hierarchical ascending classification. This classification is suitable only for small mode. It provides richer information about the similarity structure of data. It is easy to extract several sections with different levels of "resolution".
According to our application of these two classification methods on our data that used, the k-means classification gave satisfactory results compared to the HAC results, but not always because perhaps with other data, we can see that the HAC gives clearer results. In short, we say that for each specific data its own classification.

References

1. Richmond, B., *Introduction to Data Analysis*, Handbook-ERIC 2006.
2. Maheshwari, A., *Data Analytics*, 2014.
3. Green acre, M., Blass's, J. *Multiple Correspondence Analysis and Related Methods.*, CRC Press 2006.
4. Jacob Kegan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, Cambridge, 2007.
5. McQueen, J. *Some Methods for Classification and Analysis of Multivariate Observations* Dan's Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, pages 281-297, 1967.
6. Python Data Analytics, *Data Analysis and Science Using, Pandas, matplotlib, and the Python Programming Language* 2012.
7. Dawson, M. *Python Programming for the Absolute Beginner,*3rd Edition, 2019.

Ahlem Labdaoui and Mekki Soundes,
Laboratory M.A.M,
Constantine1 University,
Algeria.
E-mail address: ahlem_stat@live.fr, soundesmekki@gmail.com