Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Insured Profiles Segmentation using Unsupervised Machine Learning

Ibtissam Abdellaoui [1], Souhila Benbrahim[2] , Assia Cherif[3]

[1]*The School of Higher Commercial Studies (EHEC Algiers), Algeria*

[2]*LASAP, The National Higher School of Statistics and Applied Economics (ENSSEA), Algeria*

[3] *LASAP, The National Higher School of Statistics and Applied Economics (ENSSEA), Algeria*

Corresponding Author: tissamaabdellaoui@gmail.com

ARTICLE INFO

Received: 25 Dec 2024

Revised: 21 Sep 2025

Accepted: 30 Nov 2025

ABSTRACT

The National Social Insurance Fund (CNAS) in Algeria manages a crucial pillar of social security against risks such as illness or disability, covering salaried workers through various benefits. This study applies the K-means algorithm, an unsupervised machine learning method, to the Big Data from the General Directorate of CNAS, which is processed using the scikit-learn package in Python. to segment the insured into eight homogeneous profiles based on age, the frequency of prescribed medication care, and the amounts of reimbursements.

The segments identified by the execution of the algorithm reveal a behavioral diversity in healthcare consumption, providing the CNAS with a basis for targeted actions in prevention, awareness, and optimization of reimbursements.

Keywords: Big Data, Segmentation, Machine Learning, K-means, Behavior, Social security

INTRODUCTION

The emergence of machine learning has profoundly transformed data processing and analysis methods. Among the most commonly used algorithms for segmentation, K-means holds an important place. It is an unsupervised clustering algorithm that allows for the grouping of similar individuals based on given characteristics (Naeem et al., 2023). Appreciated for its ease of use, speed of execution, and ability to handle large volumes of data, K-means is a powerful tool for segmenting heterogeneous populations (Yin et al., 2024).

Our article is situated in this context, focusing on the application of these techniques to the field of social security, particularly the National Social Insurance Fund for Salaried Workers (CNAS). This institution, which plays a central role in the social protection system in Algeria, has

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

a significant volume of data on insured individuals, which opens interesting perspectives for fine and targeted segmentation.

This paper addresses the following issue: How can the CNAS employ segmentation techniques from Machine Learning, namely the K-means algorithm, to identify homogeneous groups of insured individuals and customize its services in a more targeted and pertinent manner?

The selection of CNAS as an area of research is justified by its essential position in the Algerian social framework, as well as the depth and diversity of its databases. Despite the presence of other social security organizations in Algeria, the CNAS distinguishes itself by its nationwide coverage, the variety of insured profiles, and the accessibility of extensive data, which are essential for the implementation of Machine Learning techniques.

The implementation of a Big Data system within the CNAS aims to afford the insurance organization under examination a genuine possibility to augment the worth of its informational assets. This paper attempts to illustrate how the astute utilization of this data might enhance the services offered to insured individuals.

LITERATURE REVIEW

Machine learning, a dynamic sub-branch of artificial intelligence, allows systems to learn from data without explicit programming, extracting regularities, patterns, and predictions to continuously improve. Clustering is a statistical technique aimed at organizing raw data into homogeneous groups called clusters. Each group gathers objects with similar characteristics while being distinct from other groups. The main objective is to determine the optimal number of clusters, which can prove to be complex in practice (Kashwan & Velu, 2013).

The clustering process is iterative and exploratory. It is commonly used in machine learning, pattern recognition, and data mining, particularly for processing unlabeled data(Kashwan & Velu, 2013). This process falls within a broader framework of knowledge discovery from large sets of unstructured data.

According to (Talia & Trunfio, 2012), in the face of the exponential growth of data volumes, manual labeling by humans quickly becomes a costly and difficult task. In this context, automatic labeling, particularly thru clustering, has become an essential step in the analysis processes. Clustering involves grouping unlabeled data into clusters, so that similar elements are gathered into the same cluster, while dissimilar elements are separated into different clusters.

Clustering can be used in two main approaches:

- As an exploratory tool, to reveal hidden structures or patterns within the data;
- As a preprocessing step, when the identified clusters serve to improve the performance of a supervised learning algorithm or another analytical task.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The notion of similarity is at the heart of this technique: the elements of the same cluster are closer to each other than to those belonging to other clusters, according to an adapted measure (Euclidean distance, Jaccard coefficient, etc.).

Clustering is a statistical technique aimed at organizing raw data into homogeneous groups called clusters. Each group brings together objects with similar characteristics while being distinct from other groups. The main objective is to determine the optimal number of clusters, which can be complex in practice (Kashwan & Velu, 2013).

According to (Talia & Trunfio, 2012), in the face of the exponential growth of data volumes, manual labeling by humans quickly becomes a costly and difficult task. In this context, automatic labeling, particularly through clustering, has become an essential step in the analysis processes. Clustering involves grouping unlabeled data into clusters so that similar elements are gathered into the same cluster, while dissimilar elements are separated into different clusters.

METHODOLOGY

1. K-means clustering method

The k-means algorithm is one of the most popular ways to do partial segmentation because it is simple and works well. Hugo Steinhaus (1957) first suggested this method, and Stuart Lloyd made it more popular. It works by dividing a dataset into K clusters by minimizing an objective function called the sum of squared errors, which is also known as intra-cluster variance (Azencott, 2018).

The objective of the k-means algorithm is to minimize the overall intra-cluster variance by assigning each observation to one of the K groups such that:

$$\boldsymbol{Arg\ min} \sum_{k=1}^{k} \sum_{x_{i \in c_k}} \|x_i - c_k\|^2$$

Where c_k represents the centroid of cluster C_k, fined as the average of the observations assigned to it:

$$c_k = \frac{1}{|C_k|} \sum_{x_{i \in C_k}} x_i$$

2. Lloyd's algorithm

The Lloyd algorithm, a standard implementation of K-means, follows an iterative procedure described as follows:

- **Initialization**: randomly choose K observations as initial centroids;
- **Assignment:** assign each observation to the nearest centroid, according to a distance measure (often the Euclidean distance);

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

- **Update:** recalculate the centroids of each cluster;
- **Repetition**: iterate steps 2 and 3 until convergence (when the assignments no longer change or a tolerance criterion is met).

This process is greedy: at each step, it seeks a local improvement of the solution, which guaranties convergence to a local minimum of the criterion. It is thus recommended to restart the algorithm several times with different starting points, keeping the solution with the lowest SSE (Azencott, 2018).

The K-means algorithm is a stochastic algorithm because it starts with random centroids. Consequently, the results may be different each time, and some initializations may lead to partitions that are very stable and far from the best solution. The K-means++ variant uses a method for starting that is meant to make the centroids spread out as much as possible from the start.

The procedure proceeds in the following order:

- A first centroid c_k is randomly chosen from the observations in the dataset D;
- For $K = \underline{2, k}$, the centroid c_k is selected from the remaining observations $D\backslash\{C_l, \ldots . C_{k-1}\}$, according to a probability proportional to the square of the distance separating them from the nearest already selected centroid.

This method doesn't make the algorithm completely deterministic, but it does make it much less likely that you'll get bad partitions. In practice, it gives more stable and efficient results than starting with completely random data (Azencott, 2018).

3. Advantages, complexity, and limitations of the k-means algorithm:

The K-means algorithm has several advantages that explain its widespread adoption in segmentation tasks. It is primarily an easy-to-implement algorithm, based on an intuitive logic of attracting points toward centers. Its computational efficiency is also a major advantage: its algorithmic complexity is on the order of $O(n.p.k.t)$, where:

- n is the number of observations;
- p is the data dimension;
- k is the number of clusters;
- and t is the number of iterations required for convergence.

The procedure is linear in the number of observations since k and t are usually small relative to n. This feature makes it much faster than hierarchical clustering approaches, which cost a lot more. K-means is faster and more scalable since it simply calculates distances to the centroids instead of to all the other points, like hierarchical clustering does. But this algorithm also has a lot of big problems. To start, it uses a greedy strategy, which means it tries to make the answer better at each step. This method doesn't guarantee that the SSE criterion will reach a

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

global minimum. The algorithm could only reach a local minimum, depending on where the centroids start. To address this problem, it is recommended to run the algorithm multiple times and keep the solution that minimizes the intra-cluster variance the most. Moreover, the k-means algorithm is sensitive to outliers. An observation very distant from the others can indeed attract a centroid to it, or even form a cluster by itself, which harms the coherence of the segmentation. This characteristic can nevertheless be diverted for anomaly detection purposes: points alone in their cluster can be considered atypical.

Finally, the obtained clusters are always convex due to the very nature of the algorithm, which generates a Voronoi diagram. This limits the use of k-means to well-separated and spherically shaped data structures, to the detriment of more complex or non-linear structures (Azencott, 2018).

RESULTS AND DISCUSSION

1. Data Collection

The study covers all active social security contributors under the CNAS who have made at least one pharmaceutical care transaction within Algerian national territory from January 1, 2024, to October 31, 2024.

This population constitutes a relevant basis for segmentation analysis, as it reflects real and traceable healthcare consumption behaviors, allowing for the identification of different insured profiles according to their frequency of use, geographical location, or the types of treatments sought.

From the initial database of pharmaceutical operations, a data analysis was initiated. While the original database was structured around transactions, this new version is organized by individuals, with each line corresponding to a unique insured person. This reprocessing allowed for the isolation of a set of **139,401** active social security contributors during the studied period, thus providing a suitable basis for the analysis of individual characteristics.

Each record represents an insured person identified by a unique identifier. The variables retained for each individual are as follows: The unique identifier of the insured (Insured number), Gender, age of the insured by bracket (Age).

2. Overview of the data and statistical evaluation

2.1. Age distribution

The analysis of the age distribution of the insured reveals a high concentration among seniors, particularly those aged 65-69 (10.70%, 14,909 people), followed by those aged 70-74 (10.57%, 14,736) and 60-64 (10.09%, 14,067). The 50-59 age group constitutes approximately 17.35% (25,192 covered individuals), whereas the 30-49 age group comprises almost 20% (30,749 insured individuals). Individuals aged older than 80 constitute 16.50% (22,012), while those

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

under 30 represent merely 5.41% (7,533), indicating a significant underrepresentation of minors (less than 1%).

2.2. Gender distribution

Women constitute 50.7% (70,651 insured), while men represent 49.3% (68,750), indicating a nearly equal balance.

3. Identification of relevant variables

The variables retained, for the period from January 1 to October 31, 2024, are as follows: **Frequency**, corresponding to the sum of the quantities of medications dispensed for each insured, which allows for an assessment of the individual level of pharmaceutical consumption, and **Expense**, which represents the total amount reimbursed by the CNAS to the pharmacy for a given insured.

	Frequency
count	139401.000000
mean	11.141054
std	12.986276
min	1.000000
25%	4.000000
50%	7.000000
75%	14.000000
max	478.000000

dtype: float64

	Expense
count	139401.000000
mean	3501.663285
std	6160.105698
min	40.000000
25%	1014.910000
50%	1910.100000
75%	3762.700000
max	495526.170000

dtype: float64

Fig 01. Descriptive statistics of the Frequency and Expense variable

The mean quantity of medication provided to an insured individual throughout the study period is 11 units, with a standard deviation of 13 units, signifying considerable diversity in pharmaceutical consumption levels. The minimum recorded value is 1 unit, whilst the greatest value attains 478 units, indicating the presence of instances of significantly high consumption that require meticulous scrutiny. Moreover, the analysis of quartiles reveals that 25% of insured patients obtained a maximum of 4 units of medication (first quartile). The median, established at 7 units, indicates that half of the insured utilized 7 units or fewer. Finally, 75% of insured individuals took at most 14 units of medication, which corresponds to the third quartile.

The average reimbursement paid by the CNAS to pharmacies is 3,502 DA per insured person, with a standard deviation of 6,160 DA, which indicates a high variability in the amounts reimbursed. The smallest observed reimbursement is 40 DA, while a median of 1,910.10 DA means that 50% of insured individuals received an amount equal to or less than this value. At the other extreme, the maximum reimbursement reaches 495,526 DA.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

4. Data cleaning

The data cleaning methodology adopted in this study is inspired by the one proposed by (Benbrahim et al., 2022), while being adapted to the specific characteristics of the data used.

4.1. Treating NaN values

In the database provided by the CNAS, we retained the following variables: Age, Frequency (frequency of medications), Expense (cost). It is reported that no missing values were detected using **isnull() function.**

4.2. Standardization of variables

At this stage, the Frequency and Expense variables were standardized through centering and reduction, achieved by removing the mean of each variable and subsequently dividing the result by its standard deviation. This transformation yields variables with a mean of zero and a standard deviation of one, so ensuring their equal contribution within the K-means method. This is accomplished via **StandardScaler** from the **sklearn.preprocessing** package.

4.3. Detection and Treatment of Outliers

The K-means algorithm is particularly susceptible to outliers, as the mean utilized for centroid calculation might be significantly skewed by these atypical data points. To limit this effect, a univariate analysis is first conducted on each variable to identify outliers, which are temporarily excluded during the initial K-means pass, allowing for the definition of more stable and representative cluster boundaries. Upon establishing these divisions, the extreme values are reallocated to the most pertinent group to preserve the information they contribute to the overall interpretation.

The following figures compare the distribution of standardized Frequency variable and standardized Expense variable using a box plot before and after excluding outliers:

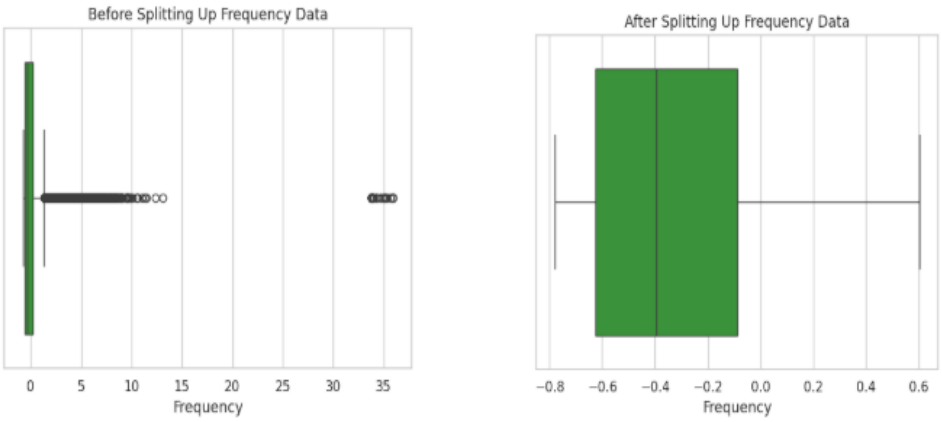

Fig 02. Boxplots of Frequency Data Before and After Splitting Them Up

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

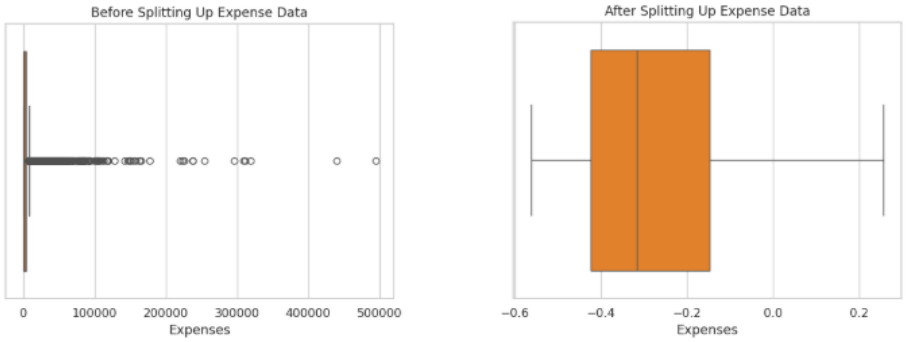

Fig 03. Boxplots of Expense Data Before and After Splitting Them Up

The figures also highlight the presence of a large number of extreme values (outliers) in the data before their temporary exclusion from the analysis set.

5. Number of Clusters

The number of clusters for the K-means algorithm was set to three for all variables, to which a specific cluster dedicated to extreme values is added. This system aims to finely tailor the segmentation to the specific characteristics of CNAS insured individuals and to facilitate the alignment of future audit tools with the different identified segments.

The following figure illustrates the structure of the groups obtained for the Frequency variable, presenting the clusters resulting from the K-means algorithm applied to the previously standardized data using **KMeans** from **sklearn.cluster package**. To facilitate the reading of the results, the values were then converted to their original scale.

	count	mean	std	min	25%	50%	75%	max
Cluster_Frequency								
1	43495.0	2.900632	0.936511	1.0	2.0	3.0	4.0	4.0
2	47672.0	7.057896	1.674006	5.0	6.0	7.0	8.0	10.0
3	26743.0	14.284299	2.533254	11.0	12.0	14.0	16.0	19.0
4	21491.0	32.964590	20.370651	20.0	23.0	28.0	37.0	478.0

Fig 04. The range of the clusters of the variable Frequency

The first group (cluster 1) comprises 43,495 insured whose prescriptions encompass between 1 and 4 units of medications. The second cluster consists of 47,672 covered persons, with prescriptions varying from 5 to 10 units, while the third cluster has 26,743 insured with dosages ranging from 11 to 19. The fourth cluster comprises 21,491 insured persons, distinguished by prescriptions of a minimum of 20 units, potentially reaching up to 478 units.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The table below proposes a correspondence between the clusters obtained by the K-means algorithm and segment labels, namely, Minimal (01), Ordinary (02), Considerable (03), and Excessive (04).

The four (4) segments ranging from minimal to excessive presented in the following figure illustrate the quantitative proportion of each segment:

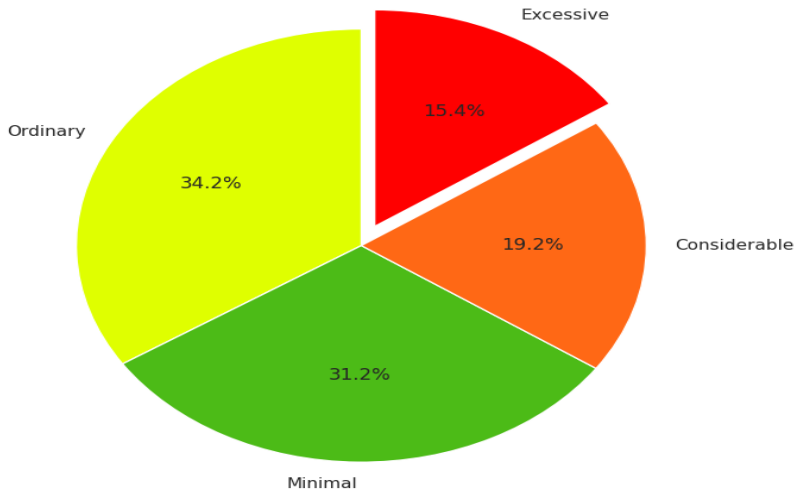

Fig 05. Pie chart of the segments of the Frequency variable

The distribution of insured individuals according to the quantity of prescribed medications shows a predominance of the “Ordinary” (34.2%) and “Low” (31.2%) segments. This means that most people who have insurance get only a small or moderate amount of medicine. In contrast, the “Considerable” (19.2%) and “Excessive” (15.4%) segments concern a smaller share of the population, reflecting less frequent cases of high medication consumption.

The following figure shows the range of clusters associated with the expense variable, obtained after applying the K-means algorithm to the standardized data. For clearer and more realistic interpretation, the results are presented using the actual (non-standardized) values:

	count	mean	std	min	25%	50%	75%	max
Cluster_Expense								
1	58019.0	896.483413	375.513690	40.00	604.425	892.07	1208.815	1557.23
2	37389.0	2211.534093	432.001194	1557.26	1831.030	2157.12	2562.720	3074.34
3	20295.0	3927.669741	569.776266	3074.43	3427.735	3853.17	4390.660	5085.38
4	23698.0	11550.476282	11752.621209	5085.49	6245.640	8348.71	13241.545	495526.17

Fig 06. Range of the clusters for the Expense variable

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The first group (cluster 1) includes 58,019 insured individuals, with reimbursed amounts ranging from 40 to 1,557 DA. The second cluster comprises 37,389 insured individuals, with expenditures ranging from 1,557 to 3,074 DA, while the third cluster contains 20,295 insured individuals, with reimbursements between 3,074 and 5,085 DA. Finally, the fourth cluster brings together 23,698 insured individuals, characterized by reimbursed amounts ranging from 5,086 DA up to a maximum of 495,526 DA.

The table below provides a correspondence between the clusters obtained by the K-means algorithm and segment labels, namely, Inexpensive (01), Reasonable (02), Expensive (03), and Exorbitant (04). The four segments range from inexpensive to exorbitant, as shown in the following figure, which illustrates the quantitative share of each segment:

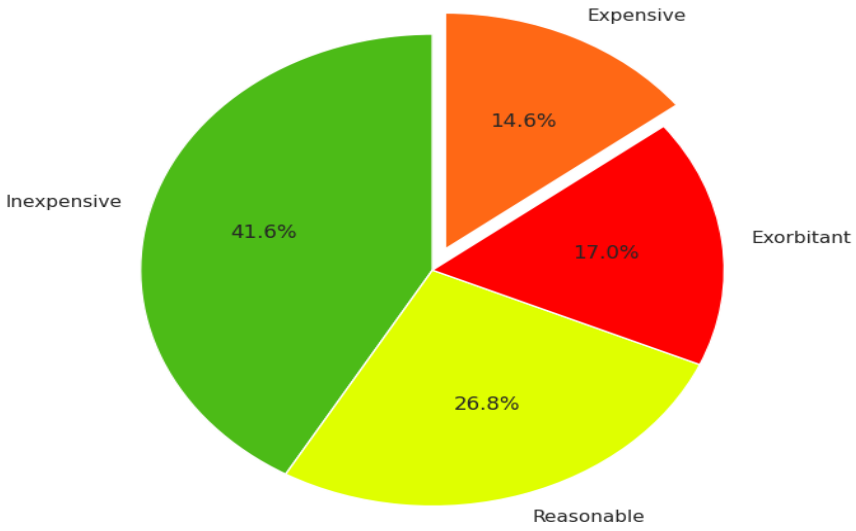

Fig 07. Pie chart of the segments of the Expense variable

The figure shows the distribution of insured individuals across the segments of the Expense variable. The "Inexpensive" segment is the most frequent, with approximately 41.6% of insureds, followed by the "Reasonable" segment, which accounts for 26.8% of individuals. The "Exorbitant" and "Expensive" segments are less represented, with respective shares of 17% and 14.6%. These results indicate that the majority of insureds generate low to moderate expenditures for CNAS, while a minority concentrates high reimbursement amounts, which can reach up to 495,526 DA for a single insured person and place a significant burden on the institution's reimbursement budget.

To perform an overall segmentation of CNAS insureds, three variables are combined. Starting from the age variable, a variable called Cluster_Age was created, dedicated to age group, and broken down into three clusters as follows:

- Cluster 1, "**Young**," groups the age range from 0 to 19 years;
- Cluster 2, "**Adult**," groups the age range 20 to 59 years;

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

- Cluster 3, "**Senior**," groups the ≥ 60 age range.

Furthermore, the variables Cluster_Frequency and Cluster_Expense are utilized.

The subsequent script was employed to combine these three clusters:

```
data["OverallScore"] = (
    data["Cluster_Age"].astype(str) + "-" +
    data["Cluster_Frequency"].astype(str) + "-" +
    data["Cluster_Expense"].astype(str)
)
```

Fig 08. OverallScore script combining the three axes

Tab 01. The obtained segments and their characteristics

Segment	Label	OverallScore	Age group	Frequency	Expense
1	Moderate youth	1-1-1, 1-1-2, 1-2-1, 1-2-2	Youth	Minimal or Ordinairy	Inexpensive or Reasonable
2	Moderate seniors	3-1-1, 3-1-2, 3-2-1, 3-2-2	Seniors		
3	Economical but frequent	1-3-1, 1-4-1, 1-3-2, 1-4-2, 2-3-1, 2-4-1, 2-3-2, 2-4-2, 3-3-1, 3-4-1, 3-3-2, 3-4-2	All ages	Considerable or Excessive	Inexpensive or Reasonable
4	Intense but Occasional	1-1-3, 1-1-4, 1-2-3, 1-2-4, 2-1-3, 2-1-4, 2-2-3, 2-2-4, 3-1-3, 3-1-4, 3-2-3, 3-2-4		Minimal or Ordinairy	Expensive or Exorbitant
5	Young Heavy Consumers	1-3-3, 1-3-4, 1-4-3, 1-4-4	Youth	Considerable or Excessive	Expensive or Exorbitant
6	Heavy-Consuming Seniors	3-3-3, 3-3-4, 3-4-3, 3-4-4	Seniors		
7	Moderately Active Adults	2-1-1, 2-1-2, 2-2-1, 2-2-2	Adults	Minimal or Ordinairy	Inexpensive or Reasonable

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

8	Very Active Adults	2-3-3, 2-3-4, 2-4-3, 2-4-4		Considerable or Excessive	Expensive or Exorbitant

The table is a classification table describing eight segments of insured people, combining age group, level of medicine consumption, and spending.

Age groups used are Youth, Adults, Seniors, plus two "All ages" segments that can concern any age.

Each segment is labelled by behavior, such as Moderate Youth/Seniors, Economical but Frequent, Intense but Occasional, Young Heavy Consumers, Heavy-Consuming Seniors, Moderately Active Adults, and Very Active Adults.

For every segment, the table specifies whether the quantity of consumption is "Minimal or Ordinary" or "Considerable or Excessive", and whether spending is "Inexpensive or Reasonable" or "Expensive or Exorbitant".

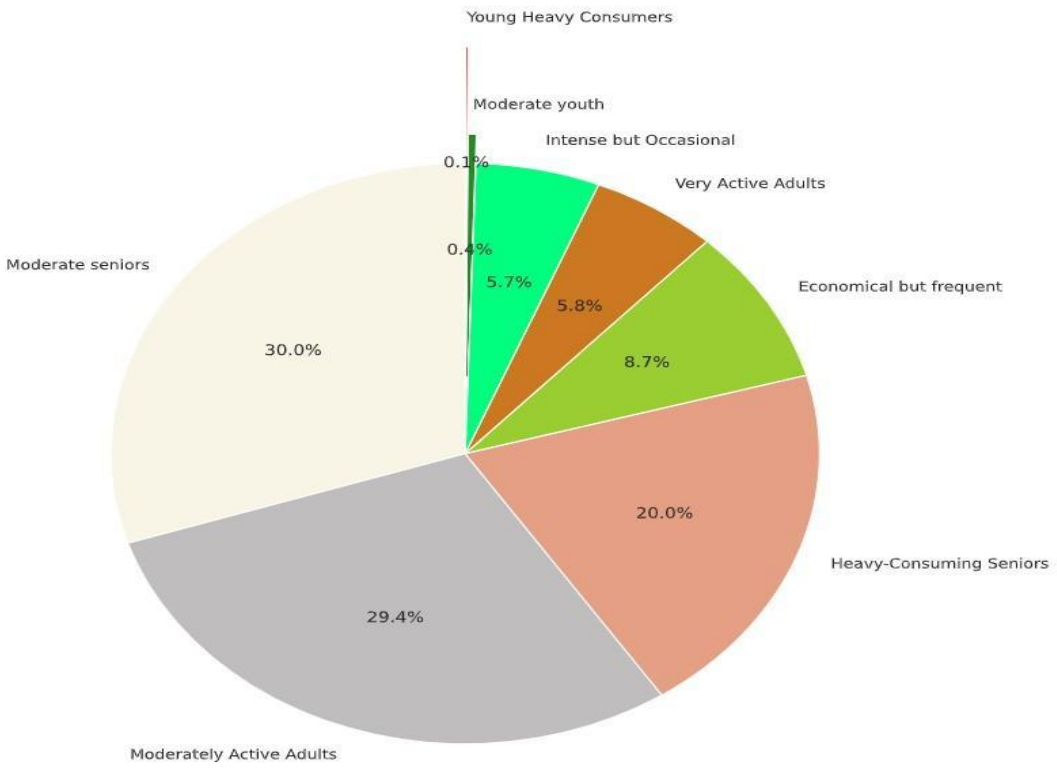

Fig 08. Pie chart of insured profile segments

The analysis of CNAS data identifies eight segments of insured individuals, based on age, medication use, and reimbursed expenditures, which highlight the following patterns:

- **Main segments**

The largest segments are "Moderate Seniors" (29.99%, ≥60 years, low/normal medication use, limited expenditures), "Moderately Active Adults" (25.44%, 30–59 years, similarly low-

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

consumption profile), and “Heavy-User Seniors” (19.99%, ≥60 years, high consumption and high expenditures).

- **Minor segments**

The remaining segments include “Frugal but Frequent” (8.70%, all ages, high consumption but low expenditures), “Intense but Occasional” (5.66%, all ages, low consumption but high expenditures), “Highly Active Adults” (5.31%, 30–59 years, high consumption and high expenditures), “Moderate Youth” (4.31%, <30 years, low consumption), and “Heavy-User Youth” (0.59%, <30 years, …).

CONCLUSION

In conclusion, this research article, based on the K-means algorithm applied to key variables such as the frequency of prescriptions and the amounts reimbursed as expenditures for the institution, has made it possible to identify homogeneous segments of CNAS beneficiaries, highlighting distinct behavioral profiles according to their level of consumption and their impact on expenditures.

This behavioral segmentation is of major strategic interest, as it provides a robust basis for designing targeted actions in prevention, awareness-raising, and reimbursement budget optimization.

By combining rigorous descriptive analysis with a refined clustering approach, CNAS thus equips itself with an operational segmentation tool to adjust its services, strengthen its audit and control mechanisms, and better manage its financial commitments to its beneficiaries.

REFERENCES

[1] Azencott, C.-A. (2018). *Introduction au machine learning*. Dunod Malakoff, France. http://www.cazencott.info/dotclear/public/lectures/2017-06-26-intro-ml.pdf

[2] Benbrahim, S., Chaouche, S. N., & Toumache, R. (2022). Countries’ economic segmentation using k-means clustering for the year 2021. *مجلة الباحث الاقتصادي-Economics Researcher’s Journal-*, *9*(1), 512–528.

[3] Kashwan, K. R., & Velu, C. M. (2013). Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering*, *5*(6), 856.

[4] Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.

[5] Talia, D., & Trunfio, P. (2012). *Service-Oriented Distributed Knowledge Discovery*. Chapman and Hall/CRC. https://doi.org/10.1201/b12990

[6] Yin, H., Aryani, A., Petrie, S., Nambissan, A., Astudillo, A., & Cao, S. (2024). *A Rapid Review of Clustering Algorithms* (No. arXiv:2401.07389). arXiv. https://doi.org/10.48550/arXiv.2401.07389