Research Article

Received: 20 March 2024 / Accepted: 28 June 2024 / Published: 02 July 2024

Pattern Identification Using Fuzzy Cluster Analysis and Latent Class Analysis: A Case Study in Perú

Jorge Chue Gallardo

César Higinio Menacho Chiock

Jesús Walter Salinas Flores

Iván Dennys Soto Rodríguez

Raphael Félix Valencia Chacón

Rino Nicanor Sotomayor Ruiz

Fernando Rene Rosas Villena

Frida Rosa Coaquira Nina

Academic Department of Statistics and Informatics,
Universidad Nacional Agraria La Molina,
Av. la Molina s/n,
La Molina,
Perú

DOI: https://doi.org/10.36941/ajis-2024-0111

Abstract

The Demographic and Family Health Survey (ENDES) conducted by the National Institute of Statistics and Informatics (INEI) in Peru provides data on fertility and health. The ENDES 2020 report, based on 35,847 surveyed households, undergoes descriptive statistical analysis with the aim of identifying patterns to enhance social conditions. Techniques such as Fuzzy C-Means and Latent Classes, previously applied in various contexts, are employed. Correlation analysis using the R polycor package highlights significant relationships, leading to the exclusion of certain numeric variables in fuzzy clustering due to strong correlations. Random sampling is applied to address the data volume. Three clusters are determined through kmeans clustering, silhouette, Elbow, and Clara methods, assessing their fuzziness with the Dunn's Fuzziness Coefficient. Pattern identification reveals significant differences in family relationships, gender, education, and health insurance among the clusters. The widespread lack of health insurance, particularly ESSALUD/IPSS, stands out as a common issue. Fuzzy clustering and latent class analysis techniques provide groupings with variations in sizes and compositions.

Keywords: *Pattern identification, fuzzy cluster analysis, latent class analysis, Dunn's fuzziness coefficient*

1. Introduction

The Demographic and Family Health Survey (ENDES), conducted annually by the National Institute of Statistics and Informatics (INEI, 2018), aims to provide "updated information and conduct analysis of change, trends, and determinants of fertility, mortality, and health in developing countries." The last application of this survey was from January to December 2020, with results consolidated into 13 data modules with their respective tables and a common identifier. These modules are grouped into two questionnaires: household and individual. The number of interviewed households was 35,847. The ENDES 2020 report presented by INEI (2021) primarily uses descriptive statistical techniques to understand the behavior of the variables under study. This survey is used to monitor various goals related to indicators of the Articulated Nutrition Program, Maternal Neonatal Health, Reduction of Crimes and Offenses affecting citizen security mentioned in the Ministry of Economy and Finance (MEF) of the Republic of Peru (MEF, 2019).

In the aforementioned context, the problem addressed in this research is the identification of patterns in the ENDES 2020 data to deepen its analysis and, therefore, collaborate with the monitoring and follow-up carried out by the MEF and the Ministry of Development and Social Inclusion (MIDIS) of the Republic of Peru. Additionally, identifying groups will allow establishing the possible need to allocate differentiated public budgets by groups to improve their education, health, and employment conditions. There are various techniques for pattern identification such as statistical techniques, structural techniques, matching template, approximation through neural networks, fuzzy model, latent class analysis, and hybrid models (Asht & Dass, 2012). In this research, two techniques were used: the fuzzy c-means clustering algorithm by (Dave & Sen, 2002) chosen for ENDES 2020 because the data are in 13 tables and are of the relational type with a common identifier, and latent class algorithm (LCA) because it determines individual profiles using observable discrete and continuous variables. The LCA technique does not require assumptions of normal distribution, linearity, homogeneity of variances, among others used in classical statistical techniques (Reyna & Brussino, 2011); similarly, (Ondé Pérez & Alvarado Izquierdo, 2019) indicate that LCA allows identifying a categorical latent variable using indicators that define the groups. The Fuzzy C-Means algorithm was used by (Saeipour, Sarbakhsh, Salemi, & Aghdam, 2023) to identify factors of higher risk associated with pedestrian traffic behavior patterns through a survey in Urmia, Iran; (Drouhot, 2021) applied it to study the levels of religiosity among young Muslims in France; (Sardareh, Aghabozorgi, & Dutt, 2015) used the algorithm to determine the effectiveness of reflective dialogue in improving students' problem-solving learning. Some applications of latent class analysis include identifying factors determining unsafe behavior of construction workers by (Deng, Cai, Xie, & Pan, 2023) and determining patterns of motorcycle accidents and driver errors in the provinces of Iran (Tavakoli Kashani, Besharati, & Mohamadian, 2017).

2. Materials and Methods

2.1 Data preparation process

This process began with data collection, where data was gathered from The Household Questionnaire and Individual Questionnaire of the ENDES 2020 (INEI, 2021). Then, data cleaning was performed to address missing values and remove duplicates. The data did not need to be transformed through normalization, standardization, and encoding of categorical variables. Data reduction techniques, such as feature selection and sampling, were applied to condense the dataset while retaining its integrity. Below is a description of the process used.

The Household Questionnaire and Individual Questionnaire of the ENDES 2020 (INEI, 2021) were administered from January to December 2020. The sample is two-stage, balanced probabilistic, stratified, and independent. The sample size was 37,390 households, of which: 15,098 correspond to departmental capitals and the 43 districts of Metropolitan Lima, 9,490 households from other urban sectors, and 12,802 households in rural areas (INEI, 2021). The number of interviewed households was 35,847, with the following characteristics: 35,430 fully interviewed individuals were women aged 12 to 49 years. The analysis focuses on women aged 15 to 49 years, but in some indicators, women aged 12 to 14 years have been considered. In the ENDES 2020, three questionnaires were applied: one to the household and its members, the second to all eligible women, i.e., aged 12 to 49 years, and the third called the Health Questionnaire, which was applied to a person aged 15 years and older. Due to the 2020 pandemic, telephone interviews were implemented, and only face-to-face interviews were conducted considering the biosecurity measures established by the Ministry of Health. During the months of mandatory social isolation, a strategy for recovering medical tests was implemented from July to September 2020. In the two-stage sampling, the sample frames were: first stage, the XI National Population and VI Housing Censuses of 2007, the 2012-2013 SISFOH Update, and the XII National Population and VII Housing Censuses of 2017 (CPV 2017), and second stage, the cartographic update and the registration of buildings and homes carried out prior to the interviews. The questionnaires contain questions on the following topics: Household Characteristics, Housing Characteristics, Basic Data of Women of reproductive age/fertile age (MEF), Birth History - Method Knowledge Table, Pregnancy Childbirth Postpartum and Lactation, Immunization and Health, Marriage - Fertility - Spouse and Woman, AIDS Knowledge and Condom Use, Maternal Mortality - Domestic Violence, Weight and Height - Anemia, Child Discipline, Health Survey, and Social Programs.

The variables are found in the household questionnaire and the observations in the databases: RECH1 and RECH4. The detailed description of the variables is found in (INEI, 2014). The RECH1 database corresponds to variables related to personal characteristics of individuals living in the household such as: relationship of kinship, gender, age, years of education, attendance at an educational institution, number of children under 5 years old to measure weight/height and hemoglobin, number of years of education, among others. The RECH4 database consists of variables related to ESSALUD/IPSS health insurance, military and comprehensive health insurance, insurance company, and level of education.

After performing data cleaning, we applied the following criteria to select variables: removing variables with constant values, missing data, and inconsistent data, the number of variables was reduced to 31 (4 for identification and 27 for responses), and the number of complete records to 139,147. Additionally, 33,399 records were identified as multivariate outliers with the HDoutliers library of R. These multivariate outliers constituted 24% of the total amount of data, leading to the decision to include them in the fuzzy cluster analysis. In this context, we proceed to perform fuzzy cluster analysis and latent class analysis.

3. Results

Correlations between variables were calculated using the hetcor function of the polycor library in R (R Core Team, 2021), considering binary, continuous, and discrete variables. Figure 1 presents a heatmap of correlation values. It was found that variables HV122 and HV124 are highly and directly correlated with a correlation coefficient of 0.97344. Variable HV108 is directly correlated with variables HV109 and HV106 with values of 0.98054 and 0.96254, respectively; variables HV128 and HV126 are directly correlated with a value of 0.971422. Additionally, variables HV121 and HV110 have a perfect direct correlation equal to 1.0000. These high correlations between the variables indicate a strong presence of redundancy in the data, which prevents having a simpler model with lower computational cost. For this reason, we make the decision not to consider the numeric variables HV106, HV109, HV110, HV124, and HV126 in the fuzzy clustering, as well as variable QH25A indicating the nationality of the interviewee. Under these considerations, the analysis ultimately included 22 numeric variables. These variables are: HV101, HV102, HV103, HV104, HV105, HV108, HV117, HV120, HV121, HV122, HV125, HV128, SH11A, SH11B, SH11C, SH11D, SH11E, SH11Y, SH11Z, SH15N, SH15Y, and SH15G.

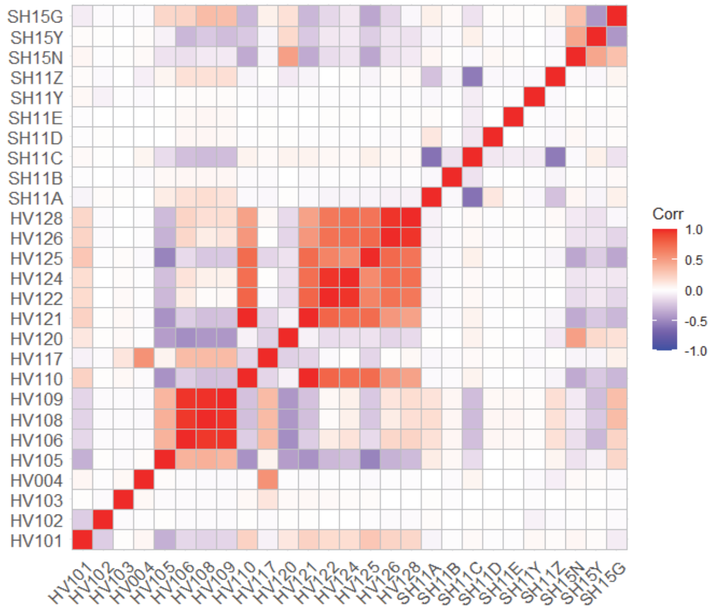

Figure 1. Graph of the correlation matrix of ENDES-2020 data considering 27 variables and 139147 records

The decision to employ sampling techniques in this study arose since in an initial run of the fcm (fuzzy k-means) function developed by (Bezdek, Ehrlich, & Full, 1984) and programmed in R, data processing had to be terminated after approximately 14 hours. This occurred because R programs (R Core Team, 2021) run in computer memory. This characteristic poses a problem when working with large volumes of data and performing complex calculations. Simple random sampling is utilized in this research because it is the sampling technique commonly used by data scientists, as indicated by (Rojas, Kery, Rosenthal, & Dey, 2017).

As the population size is N=139,147, the sample size formula for a proportion of a finite population with 99% confidence and a maximum error of 1% was applied, resulting in a value of n=14,864 records (Scheaffer, Mendenhall, & Ott, 1987). This strategy of selecting a representative sample reduces costs, time, and labor required for analysis (Acharya, Prakash, Saxena, & Nigam, 2013).

To determine the optimal number of clusters in the random sample of size n=14,864, k-means clustering, silhouette, Elbow, and Clara techniques were employed (Martínez, 2022). In Table 1, it is observed that 3 out of 4 techniques indicated that 3 clusters were appropriate for the random sample. This led to the decision to group the random sample into 3 clusters.

Table 1: Cluster number values for the ENDES 2020

Technique for calculating the number of clusters	Kmeans clustering	Silhouette	Elbow	Clara
Number of clusters	3	2	3	3

The results obtained from the application of the fcm function by (Bezdek, Ehrlich, & Full, 1984) indicated that clusters 1, 2, and 3 were composed of 5599, 3182, and 6083 records, respectively. The distance between the clusters is presented in Table 2. Note that cluster 1 is close to clusters 2 and 3, but clusters 2 and 3 are distant from each other.

Table 2. Distance between clusters

	Clúster 1	Clúster 2
Clúster 2	775.9097	
Clúster 3	636.2550	2677.8206

The graph of the clusters is presented in Figure 2. Note that cluster 1 (red color) overlaps with clusters 2 and 3, while clusters 2 and 3 exhibit relative separation. There is no completely exclusive separation observed among the three clusters. To quantify the fuzziness or lack of clarity in cluster grouping, the Dunn Fuzziness Coefficient (DFC) was used, as indicated by (NCSS, 2021), (Dunn, 1973). In Table 3, DFC values for 2, 3, and 4 clusters calculated with the ppclust library and the fcm function of R are presented (R Core Team, 2021). The table results suggest that using 4 clusters is not appropriate for the random sample of ENDES-2020 because it has a lower DFC compared to 2 and 3 clusters. These results confirm the decision to use 3 clusters for grouping ENDES-2020 data for two reasons: first, the normalized DFC values for 2 and 3 clusters have a relatively small difference of 0.01973, and second, because 3 out of the 4 algorithms used in cluster calculation indicated that 3 clusters were appropriate as presented in Table 3.

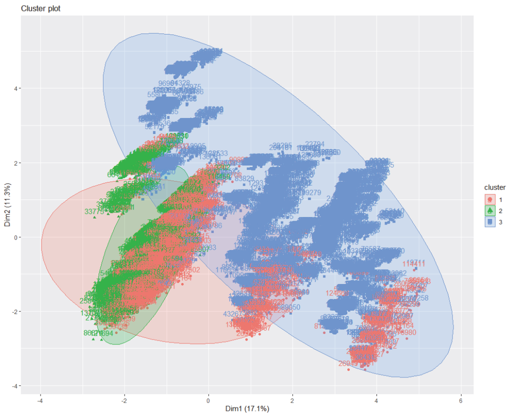

Figure 2. Clusters of ENDES 2020 with a random sample of n=14864

Table 3. Values of Dunn Fuzziness Coefficients

Number of clusters	Dunn Fuzziness Coefficient	Normalized Dunn Fuzziness Coefficient
2	0.8044131	0.6088262
3	0.7260606	0.5890910
4	0.63097	0.50796

For the pattern identification in clusters 1, 2, and 3, descriptive statistical techniques were utilized to construct Figures 4 and 5, presented below. Figure 4 displays the graph of binary variables corresponding to cluster 1. It is notable that the individuals in cluster 1 predominantly live and sleep in the surveyed household.

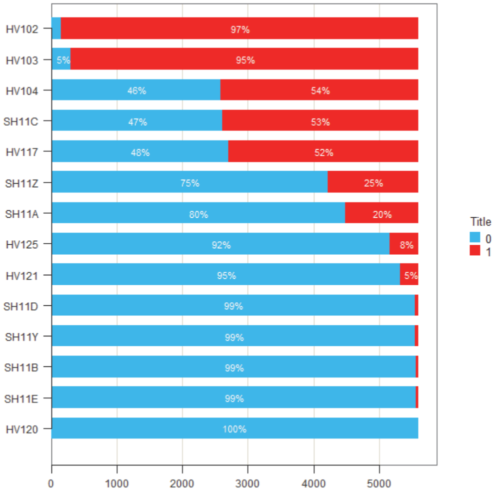

Figure 4. Percentage distribution of binary variables in cluster 1.

The graph of binary variables corresponding to cluster 2 is presented in Figure 5 (a). It is observed in this graph that almost 100% of the respondents do not have any type of health insurance, which is also the case in cluster 1. Figure 5 (b) shows the graph of binary variables corresponding to cluster 3. Note that in cluster 3, similar to clusters 1 and 2, almost 100% of the individuals do not have any type of health insurance. A notable difference among the 3 clusters is that only in cluster 3 are eligible children found for the recording of their height/weight and hemoglobin. This finding should be highlighted to take appropriate measures in the health sector.

Table 4 presents only the results of variables that significantly identify clusters 1, 2, and 3 obtained with the fuzzy c-means clustering algorithm.

Table 4. Comparative summary of the results obtained for clusters 1, 2, and 3 with the fuzzy c-means clustering algorithm

			5599 records	3182 records	6083 records
HV101	Relationship with the head of the family	What is the relationship of (NAME) to the head of the household?	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
		01. Boss.	33.06	57.10	0.10
		02. Wife/Husband.	28.84	29.79	0.33
		03. Son/Daughter.	27.27	2.36	76.60
		04. Son-in-law/Daughter-in-law.	4.52	0.50	0.18
		05. Grandson/Granddaughter.	1.07	----	14.51
		06. Others	5.24	10.25	8.28
HV104	Gender of the head of the household	Is (NAME) a man or a woman?			
		1 HOUR	46.22	50.6	51.22
		2.M	53.78	49.4	48.78
HV105	Age of the head of the household	Average	31.7 years	59.24 years	8.35 years
		Standard deviation	7.43 years	10.22 years	5.45 years
		Median	32 years	57 years	8 years
		Mode	38	48 years	3 years
HV108	Education	Number of years of study			
		less than 6	7.6	42.71	71.20
		6-9 years	18.98	14.11	18.72
		10 years	2.77	1.57	4.19
		11 years	32.98	20.65	4.67
		More than 11 years	37.67	20.97	1.22

HV117	Female Interview Eligibility	ELIGIBILITY	Not eligible 48.29Eligible 51.71	Not eligible 91.01Eligible 8.99	Not eligible 91.80Eligible 8.20
HV120	Children's Eligibility for Height/Weight and Hemoglobin	ELIGIBILITY	Not eligible 100%	Not eligible 100%	Eligible 77.43 Not eligible 22.57
HV121	The member attended school during the current school year	Are you currently attending school or college (NAME)?	Yes 95.07	Yes 100	Yes 54.79
			Not 4.93		No 45.21
HV122	Educational level during the current school year	What level and year or grade is (NAME) currently attending or enrolled although not attending?			
		0. Initial or preschool	95.07	100	57.24
		1. Primary	---	---	23.85
		2. Secondary	0.14	---	17.27
		3. Non-university higher education	4.79	---	1.64
HV125	The member attended school during the previous school year	In the past year (NAME), were you enrolled in school or college (a college or university)?			
		1. Yes	7.68		31.40
		2. No	92.32	100	68.60
HV128	Education in a single year - previous school year	Number of years of study attended or enrolled last year	0 years 92.32	0 years 100%	0 years 47.38
			From 7 to 10 years 0.16		From 1 to 5 years 26.71
			From 11 to 15 years 7.14		From 6 to 10 years 21.78
			More than 15 years 0.38		More than 10 years 4.13
SH11A	Health insurance: ESSALUD/IPSS	Are you affiliated or enrolled in: ESSALUD, Comprehensive Health Insurance or any other health insurance?			
		Yes	19.90	28.16	18.80
		No	80.10	71.84	81.20
SH11C	Health insurance: comprehensive	Are you affiliated or enrolled in: ESSALUD, Comprehensive Health Insurance or any other health insurance?			
		Yes	53.37	50.16	67.76
		No	46.63	49.84	32.24
		Not affiliated	75.26	80.8	87.75
SH15N	The highest level of education	What was the highest level and year or grade of education that (NAME) passed?			
		0. Initial or preschool	0.07	0.72	20.02
		1. Primary	15.56	37.87	32.01
		2. Secondary	45.83	32.00	22.62
		3. Non-university higher education	18.82	9.65	0.51
		4. Higher university	17.90	9.99	0.71
		5. Postgraduate	0.95	1.32	---
		8. He doesn't know	0.88	8.45	24.13
SH15Y	Years of education at level	What was the highest level and year or grade of education that (NAME) passed? (Year)	Less than 5 39.59	Less than 5 36.58	Less than 5 39.19
			5 43.45	5 43.24	5 4.67
			6 0.52	6 0.85	6 ---
			7 5.56	7 10.88	7 32.01
			8 0.86	8 8.45	8 24.13
SH15G	Education degree in level	What was the highest level and year or grade of education that (NAME) passed? (Degree)	Less than 6 6.66	Less than 6 7.04	Less than 6 27.05
			6 8.89	6 3.83	6 4.96
			8 84.44	8 89.13	8 67.99

In the application of the flexmix algorithm by (Grün & Lesisch, 2007) to find the clusters, different variables from Table 1 were used as response variables. It was found that with variable HV101, the prior probabilities of belonging to the three different clusters were 0.0513, 0.8353, 0.1134 with sizes 759, 12800, 1305 for clusters 1, 2, 3 respectively. In percentage terms, the sizes found with LCA are: 5.1%, 86.12%, and 8.77%, while with the fuzzy algorithm, the cluster sizes were 37.67%, 21.41%, and 40.92%. These results differ significantly from those found with fuzzy c-means clustering. The disadvantage of LCA is the a priori identification of a response variable and the inclusion of a model to perform the clustering.

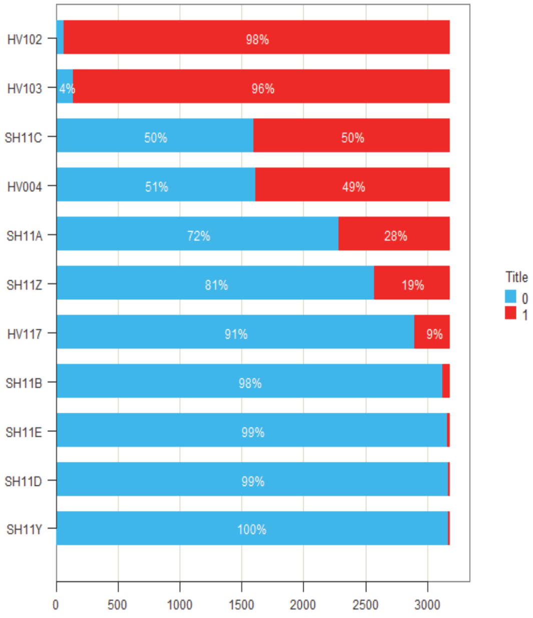

Figure 5(a): Percentage distribution of binary variables for cluster 2

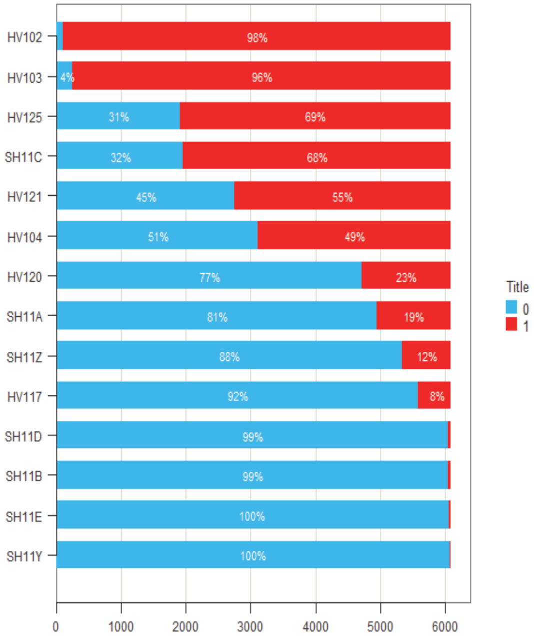

Figure 5 (b): Percentage distribution of binary variables for cluster 3.

4. Discussion

The number of clusters to group the random sample of 14864 records from the ENDES 2020 was determined to be 3. This value was obtained by the techniques of Kmeans clustering, Elbow, and Clara, and confirmed by the Normalized Dunn Fuzziness Coefficients for 2 and 3 clusters (Table 4), which are close to each other (differing by 0.01973). For the case of 3 clusters, the normalized Dunn Fuzziness Coefficient is 0.589091, indicating an acceptable clarity among the clusters (see Figure 2).

The descriptive analysis of the 22 variables common to the three clusters allows us to conclude that variables HV101, HV105, HV108, HV117, HV120, HV121, HV122, HV125, HV128, SH11C, SH11Z, SH15N, SH15Y, and SH15G are the variables that allow the clusters to be identified.

The results obtained with the fuzzy c-means algorithm indicate that the values of some of these variables, mentioned in the previous paragraph, for clusters 1 and 2 differ significantly from cluster 3. Table 3 indicates that cluster 3 is closer to cluster 1 and further away from cluster 2.

In Table 4, it can be observed that cluster 1 is characterized by families whose relationship with the head of the household is distributed approximately uniformly among the head, wife/husband, and child. The head of the household is a woman, with a median age of 32 years. Approximately 70.65% have at least 11 years of education. The majority attend school, but do not have any health insurance, which obliges them to rely on SIS.

Cluster 2 is mostly made up of household heads (57.10%), followed by wives/husbands (29.79%). Approximately 98.02% reside where the survey was conducted. The gender of the head of the household is roughly equal for men (50.6%) and women (49.4%). The average age (59.27 years) and the median (57 years) are quite close. The typical age of household heads is 48 years. The distribution of the ages of the household heads is slightly right-skewed because the Pearson skewness coefficient is 0.6575. Approximately 42.71% have less than 6 years of education, while 41.62% have at least 11 years of education. All respondents attended school during the current school year. Approximately 71.84% are not affiliated with ESSALUD/IPSS, and 98.15% do not have military health insurance. About 49.84% are not affiliated with SIS, and 99.37% do not have private health insurance. The highest education levels are primary (37.87%) and secondary (32%). The highest education level achieved by 89.13% of the respondents was grade 8.

In cluster 3, 76.60% of the respondents are children of the head of the household. About 98.39% are habitual residents of the surveyed dwelling, and 95.80% slept there the previous night. The percentages of male and female respondents are 51.22% and 48.78%, respectively. The average age of the respondents is 8.35 years, and 71.20% have less than 6 years of education. About 77.43% were eligible to study height/weight and hemoglobin. Only 54.79% attend school. Approximately 57.24% are in preschool. About 68.6% were enrolled in school the previous year, and 47.38% had 0 years of education the previous year. About 81.20% are not affiliated with ESSALUD/IPSS or any other health insurance. About 99.28% do not have military health insurance. Approximately 67.26% are affiliated or enrolled in SIS. About 99.28% are not covered by any insurance company, and 99.70% do not know about their affiliation or enrollment in health insurance. About 32.01% studied primary education, and 22.62% studied secondary education. Years of education at levels 7 and 8 were 32.01% and 24.13%, respectively. Grade 8 was the highest attained by 67.99% of the respondents. On the other hand, the results obtained with LCA show that in group 1, the respondent has no relationship with the head of the family; in group 2, children of household heads predominate; and in group 3, the majority are grandchildren.

Some observed limitations in this research are bias in the random sample analyzed that may not be representative of the population, inaccurate answers that could distort the behavior of the variables, the determination of the optimal number of clusters, and the presence of high correlation between the variables. To address these limitations, we can improve data collection methods, change the number of clusters or using alternative clustering algorithms.

The identified patterns reveal distinct differences in the variables among the three clusters analyzed. These differences have critical policy implications, particularly regarding health insurance coverage, access to healthcare services and educational and preventive health initiatives. We propose the following public policies and recommendations for each cluster.

Health Insurance Expansion:

- Cluster 1: Focus on young, educated female heads of households, many of whom lack health insurance and rely on SIS. Introduce subsidies and tailored insurance plans to increase coverage.
- Cluster 2: Target older adults with limited education, many of whom are uninsured. Expand public insurance programs and create affordable options for this demographic.
- Cluster 3: Primarily children with low education levels and high reliance on SIS. Improve health insurance enrollment through school-based programs and partnerships with pediatric providers.

Improved Healthcare Access:

- Clusters 1 and 2: Enhance healthcare facilities in underserved areas, provide transportation subsidies, and incentivize providers to serve these communities.
- Cluster 3: Ensure children have regular access to healthcare through expanded school health programs and parental outreach.

Educational and Preventive Health Initiatives:

- Cluster 1: Implement preventive health education programs through community centers and digital platforms.
- Cluster 2: Offer accessible health education programs emphasizing the importance of insurance and preventive care through workshops and local media.
- Cluster 3: Educate parents about the importance of school attendance and its link to health services, and integrate health education into school curriculums.

These targeted interventions aim to reduce disparities, improve health outcomes, and enhance social equity.

5. Conclusion

1. For the analyzed data from the ENDES 2020 and considering the fuzzy c-means clustering algorithm, it is concluded that:
   a. The quantity of 3 clusters allowed for achieving an acceptable Dunn Fuzziness Index of 0.589091.
   b. Cluster 1 consists of families led by a woman around 32 years old, with 70.65% of respondents having at least secondary education. The majority of children attend school but do not have any health insurance, which requires them to rely on SIS.
   c. In cluster 2, approximately half of the families have male household heads and the other half have female heads. The average age of household heads is 59.27 years, with a median of 57 years. Education levels in this cluster are at least 11 years for 41.62% of people, which is 29% lower than in cluster 1.
   d. In cluster 3, most respondents were children of the household head (76.6%), followed by grandchildren (14.51%). This is a clear difference compared to clusters 1 and 2, where the majority of respondents were household heads who answered the survey.
   e. Regarding the gender of the head of the household, the majority percentage is similar in clusters 2 and 3, which is not the case with cluster 1, where the majority of household heads are women. In terms of education, 71.20% of respondents in cluster 3 have less than 6 years of education, 18.75% have 6 to 9 years, and 10.08% have at least 10 years of education. This distribution of education percentages allows us to conclude that most respondents in cluster 3 only have primary education. The non-eligibility of women for the interview has an overwhelming majority of 91.8%.
   f. Unlike clusters 1 and 2, in cluster 3, 77.43% were eligible children for studying height/weight and hemoglobin. This is another finding that should be taken into account for establishing

future health and food support policies for cluster 3.

2. One aspect that stands out in all three clusters is the absence of health insurance provided by ESSALUD/IPSS in most respondents. This lack of health insurance affects quality of life, personal and family development, access to the latest technologies and treatments, protection against medical expenses, coverage of medicines, among other benefits.
3. Regarding the techniques of fuzzy cluster analysis and latent class analysis, it is concluded that both provided groups of different sizes.

References

Acharya, A., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: Why and How of it? *INDIAN JOURNAL OF MEDICAL SPECIALITIES*, 330-333.

Asht, S., & Dass, R. (2012). Pattern Recognition Techniques: A Review. *International Journal of Computer Science and Telecommunications*, 25-29.

Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM. *Computers & Geosciences, 10*(2-3), 191-203.

Dave, R., & Sen, S. (2002). Robust fuzzy clustering of relational data. *IEEE*, 713-727.

Deng, S., Cai, Y., Xie, L., & Pan, Y. (2023). Group management model for construction workers' unsafe behavior based on cognitive process model. *Engineering, construction and architectural management,, 30*(7), 2928-2946.

Drouhot, L. G. (2021). Cracks in the Melting Pot? Religiosity & Assimilation Among the Diverse Muslim Population in France. *American Journal of Sociology, 126*(4), 795-851.

Dunn, J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*(3), 32-57. doi:10.1080/01969727308546046

Grün, B., & Lesisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis, 51*(11), 5247-5252.

INEI. (2014). *Instituto Nacional de Estadística e Informática*. Obtenido de http://webinei.inei.gob.pe/anda_inei/index.php/catalog/306/variable/V2463

INEI. (2018). *¿ QUE ES ENDES ?* Obtenido de https://proyectos.inei.gob.pe/endes/queesendes.asp

INEI. (2021). Obtenido de Instituto Nacional de Estadística e Informática - Presupuesto: http://www.transparencia.gob.pe/reportes_directos/pte_transparencia_info_finan.aspx?id_entidad=4&id_tema=19&ver=D#.YR5fV07omUk

INEI. (Mayo de 2021). *Encuesta Demográfica y de Salud Familiar ENDES 2020.* Obtenido de https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1795/

Martínez, C. (25 de Marzo de 2022). *Número óptimo de clusters R*. Obtenido de Clustering: https://rensimple.wordpress.com/tag/numero-optimo-de-clusters-r/

MEF. (2019). *MINISTERIO DE ECONOMÍA Y FINANZAS-DIRECCIÓN GENERAL DE PRESUPUESTO PÚBLICO.* Obtenido de Informe de Programación Multianual Presupuestaria 2020-2022: https://www.mef.gob.pe/contenidos/presu_publ/pres_multi/Presupuesto_Multianual_2020_2022.pdf

NCSS. (2021). *Fuzzy Clustering.* Obtenido de NCSS Stastistical Software: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Fuzzy_Clustering.pdf

Ondé Pérez, D., & Alvarado Izquierdo , J. (2019). Análisis de clases latentes como Técnica de Identificación de Tipologías. *International Journal of Developmental and Educational Psychology*, 251-260. Obtenido de https://dehesa.unex.es/bitstream/10662/14065/1/0214-9877_2019_1_5_251.pdf

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Obtenido de https://www.r-project.org/

Reyna, C., & Brussino, S. (2011). Revisión de los fundamentos del análisis de clases latentes y ejemplo de aplicación en el área de las adicciones. *Trastornos Adictivos, 13*(1), 11-19.

Rojas, J., Kery, M., Rosenthal, S., & Dey, A. (2017). Sampling techniques to improve big data exploration. *2017 IEEE 7th symposium on large data analysis and visualization (LDAV)*, 26-35.

Saeipour, P., Sarbakhsh, P., Salemi, S., & Aghdam, F. (2023). A Fuzzy Clustering Approach to Identify Pedestrians' Traffic Behavior Patterns. *Journal of Research in Health Sciences, 23*(3), p. e00592.

Sardareh, S., Aghabozorgi, S., & Dutt, A. (2015). Applying clustering approach to analyze reflective dialogues and students' problem solving ability. *Indian Journal of Science and Technology, 8*(11), 70657.

Scheaffer, R., Mendenhall, W., & Ott, L. (1987). *Elementos de Muestreo.* México: Iberoamérica.

Tavakoli Kashani, A., Besharati, M., & Mohamadian, A. (2017). Analyzing Motorcycle Crash Pattern and Riders' Fault Status at a National Level: A Case Study from Iran. *International Journal of Transportation Engineering,* 5(1), 87-101.