Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 87–95
DOI: 10.15439/2023F8583
ISSN 2300-5963 ACSIS, Vol. 36

Analysis of the Public Health Service in Bogotá, Colombia: a Study Based on Customer's Complains and Using Unsupervised Learning Algorithms

Sebastian Quinchia-Lobo, Daniela Salazar-González, Daniel Salas-Álvarez,
Rubén Baena-Navarro, Isaac Caicedo-Castro
0000-0002-2491-737X, 0000-0001-9434-2156, 0000-0002-7097-7883
0000-0001-5055-6515, 0000-0002-7567-3774
Universidad de Córdoba, Socrates Research Team, Faculty of Engineering,
Carrera 6 No. 76-103, 230002 Montería, Córdoba, Colombia
Email: {squinchia31, dsalazargonzalez, danielsalas, rbaena, isacaic}@correo.unicordoba.edu.co

***Abstract*—In this study, our aim is to analyze the public health services in the city of Bogota, Colombia. We used unsupervised learning algorithms for clustering requests, complaints, claims, and denunciations issued to Supersalud in 2021. We collected the data from Supersalud's databases. We adopted clustering algorithms such as K-Means, Bisecting K-Means, and Gaussian Mixture, thus, we evaluated the quality of the combination using the silhouette coefficient. The algorithm with the best clustering quality to generate the clusters has been improved. Of the eight clusters, the first two present the highest incidences, with 181 and 249 affiliates affected for every 2,000 in the year 2021. In the first cluster, with 55% support and 100% confidence, a strong association was found between problems related to medical care facilities and restricted access to health services. In addition, in these two clusters RCCD with pathologies such as chronic communicable and non-communicable diseases (respiratory, diabetes, renal, risk factors and cardiovascular) associated with restricted access to health services were found. In conclusion, The unsupervised grouping allowed to analyze the public health services from the perspective of the RCCD, providing valuable information on the experiences of the users and the challenges in the provision of health services in Bogotá, these findings demonstrate the restriction in the access to health services from different perspectives of a deficient state regarding the provision of health services in the city of Bogotá, Colombia.**

***Index Terms*—Health Indicators, Health Services, Health Statistics, Cluster Analysis, Machine Learning.**

I. Introduction

IN Colombia, health services are regulated by the National Policy for the Provision of Health Services as per Law 1753 of 2015, in conjunction with the surveillance and control body for health services stipulated by Law 1122 of 2007, known as the National Superintendency of Health (Supersalud). Requests, Complaints, Claims, and Denunciations (RCCD) serve as pivotal tools for citizens to voice their concerns to public entities, being recognized as one of the most robust indicators of care quality. However, the quality of service provision in the public health system has been compromised due to adverse situations [1] that impact the contribution of Health Service Providers (HSP) to society.

This work was funded by the University of Córdoba in Colombia

Decision-making in public health can be jeopardized due to biases in the due to the lack of analysis, interpretation and erroneous registration of the same [2], a crucial aspect that can obscure judgments about potential outbreaks and pandemics like Covid-19. In Colombia, restricted access to databases poses a challenge to promoting national public health research [3]. Additionally, the Territorial Health Plan of Bogota 2020-2024 [4] proposes measures to enhance the efficiency of health teams and transition towards a systematized and automated health system.

This study is an extension of a previous investigation [5], where we aim to expand this work with a larger dataset by implementing unsupervised clustering algorithms to group the RCCD filed by healthcare users with Supersalud in Bogota in 2021 [6] and analyze the public healthcare services in this city. Through machine learning, we have utilized open-access data and conducted analyses to meet the primary objective. Our findings suggest that a group of RCCD with chronic pathologies has a strong association with restricted access to health services, as do pathologies such as Covid-19 and cancer.

The study is divided into three parts: the first introduces the challenges of the health sector in Colombia and briefly describes the study conducted along with the methods and techniques applied; the second addresses the findings and the achievement of the objective; and finally, a conclusion is provided that summarizes our findings and discusses their implications for future research.

II. Methodology

This study is classified as a quantitative research of descriptive type, adapted from other studies [7], widely explained by Sampieri [8], with an extended data mining approach of the Cross-Standard Process for Data Mining in Industry (CRISP-DM) [9] to analyze RCCD and meet the objective of the study. The phases of the process used are described below:

Topical area: Advanced Artificial Intelligence in Applications

A. Software

For the analysis of information, the Python 3.11.1 programming language, the Apache Spark platform and the PySpark 3.3.2 library were used. These tools allow large amounts of information to be processed in a distributed, parallel, and replicated manner [10], which guarantees reliable and reproducible results. In addition, the Google Colab development environment was used, which makes it easy to work with Jupyter Notebooks online.

B. Data

The dataset for this study is sourced from the open RCCD database interposed to Health Promoting Entities (HPE) with Supersalud in Colombia for the year 2021 [6]. Acquired on February 15, 2023, the data adheres to the *Habeas Data* Law, ensuring obfuscation to protect individual identities. Each RCCD is assumed to represent a unique individual. Our analysis is centered on Bogotá's RCCD, encompassing both the contributory and subsidized regimes. Nationally, 993,349 RCCDs were recorded, with Bogotá being the city with the most RCCD records, accounting for a total of 226,230. Specifically, the contributory and subsidized regimes in Bogotá accounted for 213,375 RCCDs out of 7,927,520 affiliates as of December 2021 [11]. This study narrows its focus to these regimes in Bogotá, as they comprise 94% of the city's RCCDs, split between 169,431 (contributory) and 43,944 (subsidized).

1) Understanding the problem and its data: To comprehend the problem and its data, we integrated the business understanding and data comprehension phases of CRISP-DM. Initial assessments confirmed the integrity of two key features: the municipality code of the entity receiving the RCCD and the affiliation regime of the affected individual. Filtering was applied based on these features, specifically `ENT_COD_MPIO` (municipality code, with 11001 representing Bogotá) and `AFEC_REGAFILIACION` (affiliation regime, focusing on subsidized and contributory). A subsequent deep dive into RCCD features revealed several categorical attributes crucial for understanding sector challenges, including gender, macro-motive, life risk, age range, pathology, and high cost. These attributes were found to be comprehensive, with no data loss observed. Their value domains after applying the filter are distributed as follows:

- **The gender** named in the data set as `AFEC_GENERO` determines the gender of the affected person. With a cardinality of 2, it takes two values: Man and Woman, at the national level (without applying the filter) another value called "Not Applicable" was observed.
- **The life risk** named in the data set as `RIESGO_VIDA` determines if the reason for which the RCCD was filed puts the life of the affected person at risk. With a cardinality of 2, it takes two possible values Yes and No (Without applying the filter, no others were observed).
- **The age range** of the affected person in the data set called `AFEC_EDADR`. With a cardinality of 9, due to the obfuscation of the data, it takes values such as: Between 0 and 5 years, between 6 and 12 years, between 13 and 17 years, between 18 and 24 years, between 25 and 29 years, between 30 and 37 years, between 38 and 49 years, between 50 and 62 years and over 63 years, at the national level (without applying the filter) another value called "Not Applicable" was observed.
- **The macro-motive** named in the data set as `MACROMOTIVO`. With a cardinality of 6, it takes the following categorical values (Without applying the filter, no others were observed): Restriction in access to health services, Deficiency in the effectiveness of health care, User dissatisfaction with the administrative process, Non-recognition of economic benefits, Lack of availability or inappropriate management of human and physical resources for care and Petitions, complaints and claims filed by HPS-HPE, territorial entities and control and surveillance agencies.
- **The high cost** feature named in the data set as `ALTO_COSTO`, this determines if the reason for the RCCD is related to a high cost disease, it presents a cardinality of 22 and takes the following categorical values (Without applying the filter, it does not others were observed): Not Applicable, Peritoneal dialysis, Hemodialysis, Management of patients in intensive care units, Diagnosis and management of the HIV-infected patient, Chemotherapy and radiation therapy for cancer, Medical-surgical management of major burns, Management of major trauma among others.
- **The pathology** feature named in the data set as `PATOLOGIA_1`, contains the disease related to the reason that the RCCD occurred. This feature have a cardinality of 22 and take the following categorical values (Without applying the filter, no others were observed): Chronic non-communicable respiratory diseases, Problems related to health care facilities or other health services, Non-communicable chronic diseases - diabetes, Osteoarticular diseases, Cancer, Chronic non-communicable cardiovascular diseases among others.

2) Data preparation and experimental configuration: At this stage, the CRIPS-DM data preparation phase was adopted and an experimental setup was included. Regarding the preparation of the data, the following considerations were taken into account:

- As can be seen previously, the "high cost" feature, which determines the cause of a high value for which the RCCD is interposed, has a high cardinality, for this reason it was reduced to using two values yes and no, in the case if it does not have any high cost (Not Applicable) its value is no and yes in otherwise.
- The pathology called "problems related to medical care facilities or other health services", despite not being its own pathology, was not excluded because it facilitates understanding of the impact and demonstrates shortcomings in the quality of health care provision services.

For the experimental configuration, combinations of features were created according to the objective of the study for the

subsequent phases (see table I).

We use one-hot coding [12] because the domain of the study features is categorical (see section II-B1). This coding assumes a categorical variable d, which takes values in the set $O = o_1, o_2, \ldots, o_H$, one-hot transforms d into an H-dimensional vector p, such that each dimension h_i comprises a value between zero and one corresponding to its value in the set O (see eq. 1). This process increases the dimensionality of the database, which implies a slightly higher processing cost due to this increase.

$$h_i = \begin{cases} 1 \text{ if } d = o_i \\ 0 \text{ if } d \neq o_i \end{cases} \tag{1}$$

Where h_i is the ith dimension of a categorical variable d, which takes values in the set O. o_i is the ith value taken by the categorical variable d.

Utilizing one-hot, categorical values like "Over 63 Years" in age range are transformed into binary columns. If an RCCD instance has this value, it's marked as 1; otherwise, it's 0. Because we use Apache Spark, we use their optimized approach that customizes one-hot for large data sets that creates a single vector column, capturing all values per feature (see eq. 1).

3) Clustering Algorithms: In this phase, the CRISP-DM modeling phase was adopted, using the previously prepared data set. Accordingly, it is proposed to use the unsupervised grouping technique to group the RCCD according to the study variables. This would allow segmenting the experiences of patients. For example: if a cluster shows that a group of RCCD have restrictions regarding some pathologies, the regulatory entities of the health sector could investigate ways to reduce the restrictions for these patients.

In addition to the above, due to the nature of the data, they do not have a predefined class or category that allows separating the data to predict, classify or group [13], therefore, in this research, we have adopted unsupervised machine learning clustering algorithms, such as K-Means, Bisecting K-Means, Gaussian Mixture [14] from the Apache Spark Pyspark library.

K-Means is an unsupervised clustering algorithm guided by proximity in a search space, which is calculated in closeness to the centroids (central point of a cluster that defines it) [15] until it is minimal, noting that it considers each cluster as convex due to the Euclidean distance determining the closeness to the centroids (see eq. 2) [16]. The implementation of K-Means available in the PySpark library for distributed processing of large volumes of information of Apache Spark in its official documentation [17] mentions that it uses a variant of the original algorithm called K-Means++, in whose particular case according to the documentation is based on the article by Bahmani [18], where a controlled random sample is used (see eq. 3), generating distant centroids. The following definitions have been recovered from the previously mentioned article:

$$d(x, C) = \min_{c \in C} ||x - c||^2 \tag{2}$$

$$C \leftarrow C \cup \{x\}; x \in X \text{ with a probability of } \frac{d^2(x, C)}{\Phi_X(C)} \tag{3}$$

Where X is the set of observations in the dataset, x is a specific observation of X, C is the group of selected centroids, c is a center of the set C, $d(x, C)$ is the Euclidean distance from x to the closest centroid in C, $\Phi_X(C)$ is the objective function to minimize using the Euclidean distance from the points in X to the centroids in C.

Bisecting K-Means is a variant of K-Means as its name indicates, the difference from the original being that it splits the data into subgroups forming a tree as its nodes subdivide until they are indivisible, in such a way that the specified or identified k clusters are generated, using K-Means with its Euclidean distance (see eq. 2). According to the official documentation of the implemented library [19], it is based on the article by Steinbach [20, 21], whose implementation is still used with some updates.

Gaussian Mixture is a probability-based algorithm which allows generating clusters considering that each of them follows a Gaussian distribution. According to the implemented library, the expectation maximization originally described by Dempster [22] is used to generate Gaussian Mixture Models (GMMs) [23]. These models are given by the multivariate probability density function, see eq. 4 in [24].

$$g(x, \Phi) = \sum_{k=1}^{K} P_k f(x, \mu_k, \Sigma_k) \tag{4}$$

4) Evaluation and quality metrics: This phase extends from the evaluation phase of CRISP-DM, in which an analysis was carried out using the silhouette coefficient as a criterion to select the clustering algorithms and determine their optimal configuration in terms of number of clusters and experimental parameters (see table I), where the different configurations were evaluated, for each one a range of groups between two and eight was assigned due to the highest silhouette coefficient being obtained with eight clusters.

Without classes, the evaluation and comparison between these algorithms is challenging, where it is necessary to have alternatives to similarity metrics between clusters, such as the silhouette coefficient, which is a dimensionless measure [25] for determining clustering coherence [26], ensuring optimal quality, with values ranging between -1 and 1, 1 being the best result and -1 the opposite, see eq. 5

$$s(i) = \frac{b(i) - a(i)}{\{a(i), b(i)\}}, \text{ where } -1 \leq s(i) \leq 1 \tag{5}$$

Where i is the ith observation, $s(i)$ is the silhouette coefficient for observation i, $a(i)$ is the average distance between i and all other observations in the same cluster, $b(i)$ are all the observations of the nearest cluster different from i.

5) Analysis tools: To understand the magnitude of the problem, it is necessary to determine the proportional (cumulative) incidence of each cluster [27, 28], taking as parameters the total number of affiliates in December 2021 of the regimes of this investigation extracted from external reports [11], using as a reference the name and code of the municipality (Bogotá, 001) and the number of RCCD for each cluster, in terms of

TABLE I
COMBINATIONS OF FEATURES USED TO PERFORM TESTS ON THE DATASET.

Feature Combination Identifier	Features					
	Gender	macro-motive	Life risk	Age range	Pathology	High cost
1	x		x	x		
2	x	x	x	x		
3	x	x	x		x	
4	x	x	x	x	x	
5	x	x	x		x	x
6	x	x	x	x	x	x
7			x		x	x
8		x	x		x	x
9		x	x			x

every 2,000 affiliates per year (see eq. 6):

$$I_p(k) = \frac{w_k}{T_a} p \tag{6}$$

Let k be the kth cluster, $I_p(k)$ the proportional incidence of k, w_k the number of observations in k, T_a the total number of affiliates in the regimes of study, and p the estimation proportion in terms of affiliates per year.

It is also necessary to use association rules to evaluate the frequency and probability of occurrence of a set of values in each cluster, for which support and confidence are used, which through the following expressions recovered from external sources [29]:

$$\text{Support} = P(T \cap K) \tag{7}$$

$$\text{Confidence} = P(K|T) = \frac{P(T \cap K)}{P(T)} \tag{8}$$

Let T and K be two set of items.

III. RESULTS

A. Clustering Outcomes

After carrying out several experiments, the best experiments were selected based on their silhouette coefficients, in table II where the 10 best results are presented, ordered from highest to lowest, of these 10 best results they have the same combination of features given by macro-motive, life risk and high cost, identified with ID 9 (see table I). The two best results, with eight and seven clusters respectively, were obtained using the K-Means algorithm and presented an acceptable adjustment time.

Figure 1 depicts heat maps using the data from the 189 iterations performed for each algorithm using the number of clusters, features combinations, and their silhouette coefficient. It can be observed that the highest silhouette coefficient is found in combinations 7 and 9 of the clustering algorithms. Additionally, it is highlighted that K-Means achieved a better silhouette coefficient in the conducted tests, while Gaussian Mixture exhibited negative silhouette coefficient and other very low values, indicating its limitations.

Figure 2 shows the processing cost (time it takes each algorithm to process the data) for the three clustering algorithms: K-Means, GaussianMixture and BisectingKMeans. It can be observed that K-Means obtained the lowest processing cost, with an interquartile range (IQR) of 2.5 seconds and a total range of 7.05 seconds, with a median of 8.76 seconds, a mean of 9.2 seconds, and an outlier of 29.57 seconds. Unlike the other two algorithms, whose performance was lower due to having a longer processing time. In GaussianMixture a greater dispersion is observed, with a IQR of 27.23 seconds and a total range of 70.6 seconds, with a median of 19.32 seconds and a mean of 28.38 seconds. This algorithm has 2 outliers of 80.77 and 103.84 seconds. BisectingKMeans has a median of 41.59 seconds and a mean of 38.2 seconds, also an IQR of 15.82 seconds and a total range of 34.11 seconds. This algorithm has no outliers.

B. Identified Clusters

In summary, with the previous findings, the best result was obtained when using K-Means with the combinations of features given by macro-motive, life risk and high cost, identified with ID 9 (see table I) and with 8 clusters compared to the others through the silhouette coefficient; In a second run using these same parameters, eight clusters with a silhouette coefficient of approximately 0.97 were identified. The RCCD instances with the identified clusters were matched to the prepared data to be able to use the other features such as pathology, gender, and age range, which were not present in the combination of features (experimental setup) used for clustering.

For each cluster, the cumulative incidence was calculated in terms of 2,000 affiliates in December 2021 for each cluster (see eq 6), two of them present higher incidences compared to the total corresponding to clusters one and two (see table III), the remaining clusters present low incidences. From Table (see table IV), it can be seen that the amount of RCCD accumulates in some values of the macro-motive feature of the clusters. On the other hand, in most of the clusters, chronic diseases are present.

The first cluster represents 181 out of every 2,000 affiliates in December 2021. This cluster presents a life risk, but not a high cost, with a support $P(X \cap Z) = \{$ Quantity of RCCD in the cluster with chronic communicable diseases and noncommunicable (respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access

TABLE II
ITERATIONS OF THE DIFFERENT COMBINATIONS OF FEATURES, ALGORITHMS AND NUMBER OF CLUSTERS

Algorithm	Number of clusters	Silhouette coefficient	Model fitting time in seconds
KMeans	8	0.96	7.55
KMeans	7	0.95	5.15
GaussianMixture	8	0.93	46.80
BisectingKMeans	8	0.91	47.08
KMeans	6	0.91	7.00
GaussianMixture	7	0.89	19.32
GaussianMixture	6	0.88	36.84
BisectingKMeans	7	0.86	46.10
GaussianMixture	5	0.85	36.32
BisectingKMeans	6	0.83	42.89

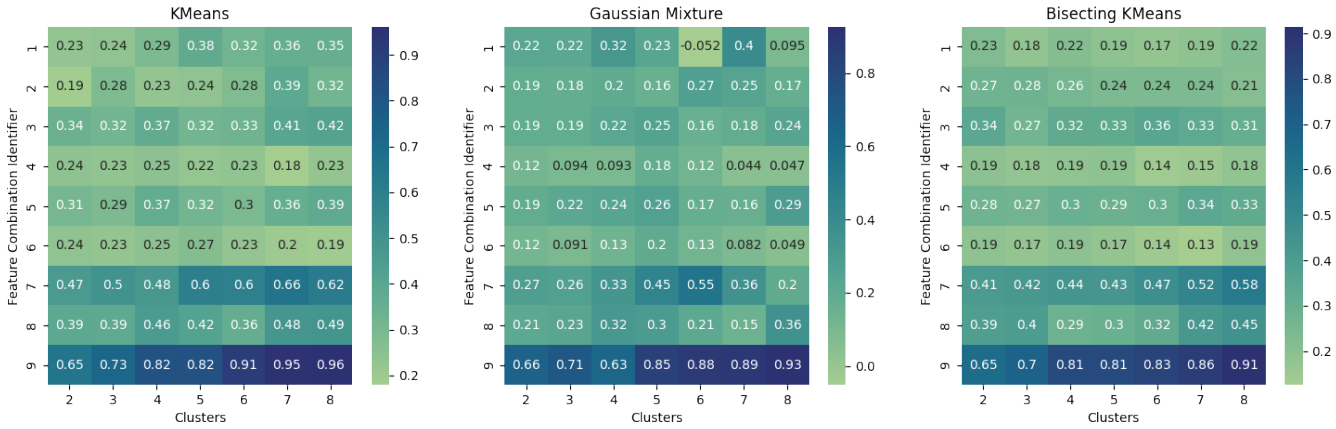

Fig. 1. Heat maps of the clustering algorithms with the combinations of features and the number of clusters.

TABLE III
QUANTITY OF RCCDS RECEIVED FOR EACH VALUE PER FEATURE - PART 1

Feature	Value	Quantity of RCCD per cluster							
		1	2	3	4	5	6	7	8
		size = 71819	size = 98843	size = 6831	size = 7826	size = 8059	size = 15411	size = 3735	size = 851
		$I = 181$	$I = 249$	$I = 17$	$I = 20$	$I = 20$	$I = 39$	$I = 9$	$I = 2$
Range of age	0 - 5	3184	4806	276	196	30	752	153	19
	6 - 12	2562	4068	196	76	31	591	97	5
	13 - 17	2489	2991	151	104	21	411	111	8
	18 - 24	4517	5257	503	241	319	1281	208	43
	25 - 29	5021	6168	581	382	1037	1363	279	55
	30 - 37	8068	10778	948	827	2727	2121	435	90
	38 - 49	11205	15949	1330	1471	2255	2657	570	167
	50 - 62	13262	21064	1236	1994	1174	2847	675	205
	63+	21511	27762	1610	2535	465	3388	1207	259
Gender	Man	30498	38427	2815	4076	3764	6503	1600	432
	Woman	41321	60416	4016	3750	4295	8908	2135	419
High cost	No	71819	98045	6708	0	7898	14923	3735	0
	Yes	0	798	123	7826	161	488	0	851
Risk of life	No	0	98843	6831	0	7847	11849	0	0
	Yes	71819	0	0	7826	212	3562	3735	851

TABLE IV
QUANTITY OF RCCDS RECEIVED FOR EACH VALUE PER FEATURE - PART 2

Feature	Value	Quantity of RCCD per cluster							
		1	2	3	4	5	6	7	8
Macro-motive	Deficiency in the effectiveness of health care	0	0	6831	0	0	0	3735	851
	Lack of availability or inappropriate management of human and physical resources for care	133	119	0	19	1	0	0	0
	User dissatisfaction with the administrative process	0	0	0	0	0	15186	0	0
	Non-recognition of economic benefits	0	0	0	0	8058	0	0	0
	Petitions, complaints and claims filed by HPS-HPE, territorial entities and control and surveillance agencies	336	0	0	86	0	225	0	0
	Restriction in access to health services	71350	98724	0	7721	0	0	0	0
Pathology	Covid-19	13285	12	9	452	43	482	584	48
	Intensive care for any pathology	3	21	2	43	0	10	0	12
	Cancer	9131	441	159	3183	231	640	710	479
	Overall effectiveness of care	15	8	6	2	0	1	3	0
	Chronic communicable disease	37	8	1	1	1	1	0	0
	Vector-borne disease	3	1	1	0	0	0	1	0
	Chronic non-communicable respiratory diseases	6196	1566	170	166	84	500	384	12
	Chronic non-communicable diseases - diabetes	3291	3319	158	134	40	332	131	6
	Chronic non-communicable diseases - renal	954	471	55	844	77	142	47	68
	Chronic non-communicable diseases - risk factors	588	1807	200	22	34	190	37	3
	Chronic non-communicable cardiovascular diseases	9900	13473	667	664	128	1410	431	30
	Orphan diseases	1982	181	28	61	15	95	120	5
	Immune-preventable diseases	6	10	0	1	0	0	0	0
	Neurological diseases	2378	718	31	14	23	152	62	2
	Osteoarticular diseases	4266	14809	757	243	413	1137	161	9
	Great burn	6	3	2	18	1	4	0	4
	Maternal-child	3352	3660	246	119	45	739	255	7
	Not applicable	134	88	12	2	17	38	1	0
	Problems related to medical care facilities or other health services	10209	54709	3932	360	6797	8742	489	20
	Prosthetic hip or knee joint replacements	20	130	4	274	8	19	1	6
	Mental health	5831	1118	168	33	58	476	306	3
	Oral health	60	2158	198	2	24	188	1	1
	Sexual and reproductive health	10	57	4	0	0	8	2	0
	HIV AIDS and other sexually transmitted diseases	162	75	21	1188	20	105	9	136

to health services } / {cluster size} $= 29\%$ and a confidence $(X \rightarrow Z) = P(Z|X)$ = { Quantity of RCCD in the cluster with communicable and non-communicable chronic diseases (respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access to health services} / { Quantity of RCCD in the cluster with chronic communicable and non-communicable diseases (Respiratory, diabetes, renal, risk factors and cardiovascular)} $= 99\%$, this suggests a pattern in the data from this cluster. The pathologies of covid-19 and Cancer are also highlighted, also with the previous macro-motive, which have a support $P(X, Z) =$ { Quantity of RCCD in the cluster with covid-19 and cancer that have the macro-motive of restriction in access to health services } / { cluster size} $= 31\%$ and with a confidence$(X \rightarrow Z) = P(Z|X)$ = { Quantity of RCCD in the cluster with covid-19 and cancer who have the macro-motive for restriction of access to health services} / { Quantity of RCCD in the cluster with covid-19 and cancer } $= 99\%$, this reveals a strong pattern in the data, something that is worrisome since it reveals that in this cluster to these pathologies, access to health services is restricted, which can increase their risk of serious complications or long-term effects by not having timely or

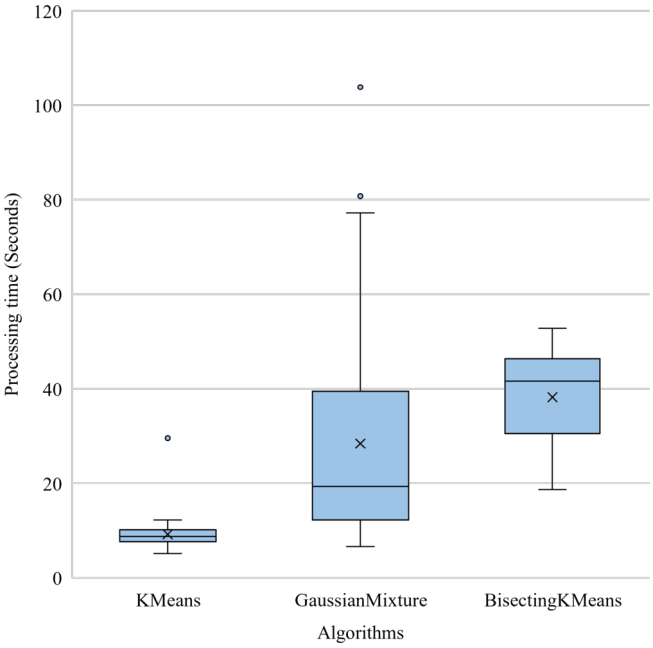

Fig. 2. Boxplot of the processing time of the different algorithms

adequate treatment; In addition, cancer and Covid-19 are two pathologies to highlight because cancer patients usually contract an immunosuppressed state that increases the risk of infections and exposes them to complications [30], both pathologies according to the previous results are restricted to them. access to health services, this causes greater concern, because health should be for everyone and should not be a privilege.

In the second cluster, 61% correspond to women, without a life risk and for the most part does not present a high cost, this represents 249 of every 2,000 affiliates in December 2021. With a support $P(X \cap Z) = \{$ Quantity of RCCD in the cluster with the pathology called problems related to medical care facilities or other health services that have the macro-motive of restriction in access to health services$\}$ / {Cluster size$\}$ $= 55\%$ and a confidence $(X \rightarrow Z) = P(Z|X) = \{$ Quantity of RCCD in the cluster with the pathology called problems related to health care facilities or other health services that have the macro-motive of restriction in access to services of health$\}$ / $\{$ Quantity of RCCD in the cluster with the pathology called problems related to medical care facilities or other health services$\}$ $= 100\%$, this indicates a strong pattern in the data, revealing the shortcomings of the users in the public health care, which are not satisfied with the care and has a strong association with the restriction of health services. It is also highlighted from this cluster that with a support $P(X \cap Z) = \{$ Quantity of RCCD in the cluster with chronic communicable and non-communicable diseases (Respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access to health services $\}$ / {Cluster size$\}$ $= 21\%$ and confidence $(X \rightarrow Z) = P(Z|X) = \{$ Quantity of RCCD in the cluster with chronic communicable and non-communicable diseases (Respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access to health services$\}$ / $\{$ Quantity of RCCD in the cluster with communicable and non-communicable chronic diseases (Respiratory, diabetes, renal, of risk and cardiovascular)$\}$ $= 100\%$, this demonstrates the strong association with this macro-motive, in addition to these chronic pathologies that can affect the quality of life of the affected person if the restriction in access to health services persists, implying a imminent risk to public health, due to the fact that they are not given sufficient attention or the state of medical care services are not suitable, leading to a worsening of their health status, such that it presents a life risk to this population.

IV. Discussion

In our study, we employed unsupervised clustering algorithms, including K-Means, Bisecting K-Means, and Gaussian Mixture, to analyze the Right to Health Care Claims (RCCD) in Bogotá, Colombia's public health services. Our findings revealed that K-Means outperformed the other two algorithms in terms of efficiency, enabling us to segment user experiences and gain a nuanced understanding of the challenges they encounter when accessing health services.

Our analysis underscores that health services for chronic pathologies, predominantly evident in clusters 1 and 2, are notably constrained. Within these clusters, a recurring macro-motive consistently yielded the highest RCCD count. Notably, cluster 1 emphasized issues within medical care facilities. These observations resonate with Lemy's research [30], which identified administrative barriers in accessing health services. Furthermore, the prominence of Covid-19 and Cancer pathologies in the first cluster indicates restricted health service access during the 2021 pandemic. This aligns with other studies [31] that observed a decline in preventive services and an absence of comprehensive vaccination strategies for Covid-19, especially for cancer patients.

Interestingly, while some rules exhibited low support, they demonstrated high confidence, indicating robust relationships between specific variables within data subsets. For instance, a rule associating Covid-19 and cancer pathologies with restricted health service access exhibited 31% support but a staggering 99% confidence. Such patterns can be instrumental in discerning trends within specific data groups. However, it's pivotal to remember that correlation doesn't equate to causation, necessitating further research to ascertain any causal links.

The implications of our findings are profound. They accentuate the imperative to enhance the health system's efficacy and structure, especially in crisis scenarios like pandemics. We advocate for the relevant authorities to devise strategies bolstering the availability and accessibility of public health care services, drawing insights from our study.

However, our research is not without limitations. The scope was confined to RCCD in Bogotá, covering the contributory and subsidized healthcare regimes, the two regimes with the

most RCCD. Our quantitative approach, based on RCCD from Supersalud, might not capture the data's full depth. The study's lens was solely on user perception and satisfaction, overlooking other potential influencers like healthcare personnel, infrastructure, or funding. We experimented with various feature combinations for the clustering algorithms, but other configurations remain unexplored. For instance, we limited our cluster count to eight based on the optimal coefficient. Lastly, due to data obfuscation for legal compliance, individual identification was impossible, leading us to assume each RCCD represents a distinct individual in our analysis.

V. Conclusion

This study underscores the power of sophisticated data analysis techniques in the domain of public health, specifically when working with RCCD. By leveraging mathematical and statistical models for information preprocessing and employing algorithms like K-Means, Bisecting K-Means, and Gaussian Mixture, we achieved a comprehensive analysis of public health services in Bogotá, Colombia. The silhouette coefficient played a pivotal role, ensuring the best clustering quality and preventing the generation of ambiguous or non-informative results. With a strong association identified, having a support of 55% and a confidence of 100%, we found significant issues related to healthcare facilities and restrictions in access to health services.

One of the standout findings was the evident deficiencies in the health system, especially concerning the quality of services. Chronic pathologies, both communicable and non-communicable, were prominently present in clusters 1 and 2, with the highest incidence. These clusters revealed a significant restriction in access to health services, with this restriction being a dominant motive. The macro-motives in these two clusters reflected the challenges in accessing health services, emphasizing the urgency to prioritize and address these issues. Without timely and appropriate care, the risk to patients' lives is substantially heightened, especially given the current restrictions in the public health sector.

Considering these findings, it's recommended to expand the scope of this study, incorporating data from different regions of Colombia to capture a more comprehensive national perspective. Such an approach could unveil crucial insights into the challenges faced by public health services across the country. Moreover, a deeper exploration of the patterns and strong associations identified in this study is essential. Understanding the intricate relationships between various variables and their impact on health service quality can pave the way for more informed administrative decisions, ultimately enhancing patient care in the health sector.

Acknowledgment

We thank Universidad de Córdoba in Colombia for supporting this study and Supersalud for publishing the dataset used in this study. Too thanks to the projects SFCB-01-21 and FI-01-22 of the Universidad de Cordoba. Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. Finally, we thank the anonymous reviewers for their comments that contributed to improve the quality of this article

References

[1] D. Mendieta and G. Rojas, "Corruption the biggest epidemic that colombia suffers," *Revista Opiniao Juridica*, vol. 19, no. 32, pp. 296–315, Sep. 2021. [Online]. Available: https://periodicos.unichristus.edu.br/opiniaojuridica/article/view/3979

[2] S. C. Johnson, M. Cunningham, I. N. Dippenaar *et al.*, "Public health utility of cause of death data: applying empirical algorithms to improve data quality," *BMC Medical Informatics and Decision Making*, vol. 21, no. 175, p. 20, Dec. 2021. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01501-1

[3] J.-S. Franco and D. Vizcaya, "Availability of secondary healthcare data for conducting pharmacoepidemiology studies in Colombia: A systematic review," *Pharmacology Research & Perspectives*, vol. 8, Oct. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/prp2.661

[4] Secretaría Distrital de Salud de Bogotá D.C., "Plan Territorial de Salud para Bogotá D.C. 2020-2024," Secretaría Distrital de Salud de Bogotá D.C., Bogotá D.C., Tech. Rep., 2020. [Online]. Available: https://subredsuroccidente.gov.co/planeacion/DOCUMENTO%20PTS%202020-2024%20%2027042020.pdf

[5] S. Quinchia-Lobo and D. Salazar-González, "Análisis exploratorio de las pqrd del sector salud mediante aprendizaje no supervisado para identificar las principales barreras y oportunidades de mejora en la prestación del servicio en la salud pública del municipio de montería," B.Sc. thesis, Universidad de Córdoba, Monteria, Córdoba, Jul. 2023, supervisors: Salas-Alvarez D. and Baena-Navarro R. [Online]. Available: https://repositorio.unicordoba.edu.co/handle/ucordoba/7408

[6] SUPERSALUD. Base de datos pqrd 2021 - csv — portal de datos abiertos de la sns. [Online]. Available: https://mapas.supersalud.gov.co/arcgisportal/apps/sites/#/datos-abiertos/datasets/3824e636c1b748269364c0e57c680d58/about

[7] M. Hinojosa, I. Derpich, M. Alfaro *et al.*, "Procedimiento de agrupación de estudiantes según riesgo de abandono para mejorar la gestión estudiantil en educación superior," *Texto Livre*, vol. 15, p. 22, Mar. 2022. [Online]. Available: https://periodicos.ufmg.br/index.php/textolivre/article/view/37275

[8] R. Hernández Sampieri, C. Fernández Collado, and P. Baptista Lucio, *Metodología de la investigación*, 5th ed. México, D.F: McGraw-Hill, 2010.

[9] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050921002416

[10] E. Nazari, M. H. Shahriari, and H. Tabesh, "BigData Analysis in Healthcare: Apache Hadoop , Apache spark and Apache Flink," *Frontiers in Health Informatics*, vol. 8, no. 1, pp. 92–101, Jul. 2019. [Online]. Available: http://ijmi.ir/index.php/IJMI/article/view/180

[11] ADRES. (2022) Reporte de afiliados por departamento y municipio. [Online]. Available: https://www.adres.gov.co/eps/bdua/Paginas/reporte-afiliados-por-departamento-y-municipio.aspx

[12] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 4, p. 12, 2016.

[13] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, pp. 301–315, Aug. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417419301472

[14] U. N. Wisesty and T. R. Mengko, "Comparison of dimensionality reduction and clustering methods for sars-cov-2 genome," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2170–2180, 2021. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111115089&doi=10.11591%2fEEI.V10I4.2803&partnerID=40&md5=9f4a6b2b087f1e835402560d8081947c

[15] S. Jian, D. Li, and Y. Yu, "Research on Taxi Operation Characteristics by Improved DBSCAN Density Clustering Algorithm and K-means Clustering Algorithm," *Journal of Physics: Conference Series*, vol. 1952, no. 4, p. 7, Jun. 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1952/4/042103

[16] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9072123/

[17] Apache Software Foundation. Kmeans — pyspark 3.3.2 documentation. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.KMeans.html#pyspark.ml.clustering.KMeans

[18] B. Bahmani, B. Moseley, A. Vattani *et al.*, "Scalable k-means++," *Proceedings of the VLDB Endowment*, vol. 5, no. 7, pp. 622–633, Mar. 2012. [Online]. Available: https://dl.acm.org/doi/10.14778/2180912.2180915

[19] Apache Software Foundation. Bisectingkmeans — pyspark 3.3.2 documentation. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.BisectingKMeans.html#pyspark.ml.clustering.BisectingKMeans

[20] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD Workshop on Text Mining*, 2000.

[21] M. Vichi, C. Cavicchia, and P. J. F. Groenen, "Hierarchical Means Clustering," *Journal of Classification*, vol. 39, no. 3, pp. 553–577, Nov. 2022. [Online]. Available: https://link.springer.com/10.1007/s00357-022-09419-7

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–22, Sep. 1977. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x

[23] Apache Software Foundation. Gaussianmixture — pyspark 3.3.2 documentation. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.GaussianMixture.html#pyspark.ml.clustering.GaussianMixture

[24] K. Aziz, D. Zaidouni, and M. Bellafkih, "Leveraging resource management for efficient performance of Apache Spark," *Journal of Big Data*, vol. 6, p. 23, Dec. 2019. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0240-1

[25] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257

[26] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. sydney, Australia: IEEE, Oct. 2020, pp. 747–748. [Online]. Available: https://ieeexplore.ieee.org/document/9260048/

[27] A. Fajardo-Gutiérrez, "Medición en epidemiología: prevalencia, incidencia, riesgo, medidas de impacto," *Revista Alergia México*, vol. 64, no. 1, pp. 109–120, Feb. 2017. [Online]. Available: http://revistaalergia.mx/ojs/index.php/ram/article/view/252

[28] L. Rychetnik, P. Hawe, E. Waters *et al.*, "A glossary for evidence based public health," *Journal of Epidemiology & Community Health*, vol. 58, pp. 538–545, 2004. [Online]. Available: https://jech.bmj.com/lookup/doi/10.1136/jech.2003.011585

[29] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2014. [Online]. Available: http://doi.wiley.com/10.1002/9781118874059

[30] O. M. Al-Quteimat and A. M. Amer, "The Impact of the COVID-19 Pandemic on Cancer Patients," *American Journal of Clinical Oncology*, vol. 43, no. 6, pp. 452–455, Jun. 2020. [Online]. Available: https://journals.lww.com/10.1097/COC.0000000000000712

[31] L. Bran Piedrahita, A. Valencia Arias, L. Palacios Moya *et al.*, "Barreras de acceso del sistema de salud colombiano en zonas rurales: percepciones de usuarios del régimen subsidiado," *Hacia la Promoción de la Salud*, vol. 25, no. 2, pp. 29–38, Jul. 2020. [Online]. Available: https://revistasojs.ucaldas.edu.co/index.php/hacialapromociondelasalud/article/view/2358