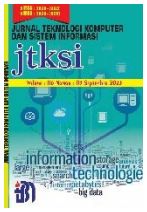

JTKSI (Jurnal Teknologi Komputer dan Sistem Informasi)

JTKSI, Volume 7, Nomor 1, Januari 2024
E ISSN: 2620-3030; P ISSN: 2620-3022, pp.57-65
Accredited SINTA 4 Nomor 200/M/KPT/2020
https://jurnal.ftikomibn.ac.id/index.php/jtksi/index

Received: 18 Desember 2023 Revised: 20 Desember 2023; Accepted: 2 Januari 2024

Venezuelan Insurance Clusters with Unsupervised Machine Learning

Jorge Luis Aquino Olmos[1], Luis David Lara Rodríguez[2], Elizabeth López Meléndez[3]
[1] Data Science Program. Universidad Tecnológica Latinoamericana en Línea (UTEL)
[2]Information Technology Department, Universidad Politécnica de Puebla (UPPue)
[3]Mechatronics Department, Universidad Tecnológica de Huejotzingo (UTH)
e-mail: aquinosuarez@gmail.com[1], luis.lara406@uppuebla.edu.mx[2], elizabeth.lopez@uth.edu.mx[3]

Abstract

Insurance companies play an important role in a healthy economy, as they provide a security service to goods and people. The Venezuelan insurance market has faced great challenges in the last decades in a narrow environment, where knowing the closest competitors is of utmost importance. The public governing body that regulates the comparison between insurers has only made use of the premiums charged as a rating factor, however this governing body makes public an additional range of indicators, for this research we have selected five of these indicators from the last three years and contrast whether the premiums charged represent a unique rating characteristic. Several machine learning methods have been applied for this study, from dimensionality reduction techniques, unsupervised clustering and optimal clustering performance criteria; all of them have allowed us to discover the hidden patterns of the data and present a clustering that reaches 100% accuracy, of the same number of classes as the government entity; which reflect the naturalness of the data as opposed to the arbitrary one given by this same entity. From its derived groupings, it is possible to affirm that the parameter of premiums collected does not represent a determining grouping characteristic, allowing the Venezuelan insurance market to compare itself efficiently with its competitors.

Keywords: Machine Learning, PCA, Mixture of Gaussian, K-means, BIC, AIC, Insures.

I. INTRODUCTION

The purpose of insurance is to compensate for damages that may occur when an event occurs, which is why the insurance industry plays an important role in the economy of countries, due to the relief effect it provides, becoming an ally in economic stability in times of crisis. As an example of this importance, it is pertinent to mention that, in Colombia, by 2021, this industry contributed 3.72% of GDP, a figure that is on average at the regional level [1]. Noting that the insurance industry throughout history has been an ally in the development of countries, no economy that has flourished has done so without the support of this important industry [2].

The development of the insurance sector goes hand in hand with the economic development of any economy [3], which is why it is vital to be able to carry out a classification that leads to the comparison of insurance companies; in practice, this classification is done by companies specialized in credit and risk rating. These agencies evaluate certain aspects related to the financial soundness and capacity of the insurer to meet its obligations to its policyholders, and then give them a rating such as "A+", "AA-", "B", etc., where a higher rating indicates higher solvency and lower risk.

The regulatory framework for the insurance sector in Europe is established through Solvency II, which sets out the guidelines and rules that regulate and serve to supervise the insurance and reinsurance industry, significantly helping the stability of this fundamental pillar of the economy. All these mechanisms and ways of classifying insurance companies help the supervisory bodies to carry out the necessary supervision to which they are obliged. However, this technicality does not penetrate the players in the industry, which is why simpler indicators are used to classify the companies, serving as a sales mechanism and a commercial strategy that reaches future insurance purchasers. The figures used in the study correspond to those generated in several consecutive years by the Superintendency of Insurance SUDEASEG the governing body of the sector in Venezuela [4].

The evolution of technology has come on leaps and bounds, helping companies to have more and more sophisticated and innovative tools at their disposal, as well as to store a large and varied amount of information, which is processed through modern computer systems. Similarly, this technological advance offers them novel analysis techniques, but it

has also created a challenge for them; to be able to make use of this enormous amount of data for business and commercial decision making. In view of this challenge, the aim of this work is to determine whether the way in which the classification of insurance companies categorized by means of the premiums charged has been carried out is statistically correct; by applying a set of common statistical techniques of machine learning, the aim is to find the similarity between the insurance companies operating in the Venezuelan market, taking into account the indices that are regularly published by the sector's regulatory body, i.e. the aim is to create groups of companies that present similar characteristics, and this knowledge can provide advantages, among which we can mention: Identify strengths and weaknesses, detect market opportunities, improve marketing strategy, promote innovation, prepare for changes in the market, learn from the mistakes of the competition. This will allow you to take certain actions that will lead to your business growth. Among the actions that could be implemented are: improving your sales channels by observing how the competition does it, where to establish new offices, observing the market niche of your competitors and thereby increasing your sales, and increasing the network of commercial allies that offer the insurer [5]. Ultimately, having a thorough understanding of the competition is fundamental to survive in highly competitive markets such as the Venezuelan insurance market [6].

Table 1 shows the variables that take part in the statistical analysis. The data that are part of this study correspond to those published by the governing body in the period 2020-2022, being three years with six variables per year, with which we will have a data set made up of 18 variables. With respect to insurers, the information published corresponds to 50 insurers, however, only 42 of these are part of the study, since the other 8 insurers do not present data or are at zero in a large number of years, which will cause problems when evaluating these indicators in the different methods applied in the study.

Table 1. Variables to be analyzed accompanied by their description.

GC: Commissions and acquisition costs Represents the cost of insurance intermediation, arising from the payment of commissions and bonuses to its insurance producers.
GA: Administrative costs Total amount paid to cover the cost of total administrative expenses, including staff costs and overheads incurred in the conduct of its insurance business.
PC: Premiums collected Fee premium paid by the insured to cover the coverage of a specific risk.
SO: Balance of operations It is the positive or negative balance resulting from all the insurance undertaking's operations of the insurance undertaking at the end of the period is obtained by adding together the net technical result and the general the general management result, giving the positive or negative operating balance (profit or loss) of operations (profit or loss) of the company.
SP: Paids Claim Total amount of claims indemnified by the insurance company during a given period, this amount is reflected net of claims salvage.
ST: Total Claims Refers to the sum of claims paid plus reserves for benefits and gross outstanding claims and reserves for benefits and gross outstanding claims.

The Table 2 shows the classification made in the year 2022 by the internal regulatory body (SUDEASEG) of the insurers present in the study, it should be noted that this classification is made with the only parameter of the premiums collected in that year and the number of insurers belonging to its four different groupings, corresponds to a finite number of elements which are: The first three groupings have ten insurers each in order of importance to the mentioned item, the last grouping compiles all other insurers.

The data of the study variables show a high skew (to the right), lower values, which is a common feature in this type of variables. A factor that must be taken into account is represented by one of the monetary measures adopted by the Venezuelan government, which consisted of eliminating zeros from the currency, which creates a distortion when comparing variables from different years, for this reason transformations must be made on the data to avoid this distortion see equation 1, the range must be [0-1].

Table 2. SUDEASEG's classification of Venezuelan insurance companies in 2022

ID	Insurance	Class	ID	Insurance	Class
1	Altamira	2	23	Mercantil	1
2	American	4	24	Mundial	3
3	Andes	2	25	Nuevo Mundo	3
4	Atrio	2	26	Occidental	3
5	Ávila	4	27	Oceánica	1
6	Banesco	1	28	Oriental	4
7	BBVA	3	29	Pirámide	1
8	Bolivariana	4	30	Previsora	2
9	Capital	4	31	Primus	4
10	Caracas	1	32	Proseguros	3
11	Caroni	3	33	Qualitas	2
12	Catatumbo	3	34	Real	2
13	Constitución	1	35	Uniseguros	3
14	Corporativos	4	36	Universal	4
15	Estar	2	37	Universitas	1
16	Fé	3	38	Venezolana	2
17	Hispana	1	39	Venezuela	2
18	Horizonte	2	40	Virgen	4
19	Iberoamericana	4	41	Vitalicia	4
20	Interbank	4	42	Vivir	4
21	Internacional	1	43	Zuma	3
22	Mapfre	1			

$$Z_j = \frac{X_i - \min(X_j)}{\max(X_j) - \min(X_j)} \qquad (1)$$

This normalization allows algorithms that make use of distances in their calculations, which, in case of not

making use of such normalization, the results could be affected by the existence of discrepancy between the magnitudes present in the data, on the other hand, by performing the transformation each variable is given the same importance when they are used in any clustering algorithm.

An important factor in our study is the correlation that may exist between the variables; therefore, a heat map is presented that identifies this relationship in the data; knowing the relationship that may exist between the variables will allow us to make the necessary adjustments using the computer learning methods that have been applied in the search for dimensional reduction of the variables, an important factor in this study.

$$r = \frac{\operatorname{cov}(x,y)}{S_x S_y} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{(n-1)S_x S_y} \qquad (2)$$

We have used the Pearson correlation statistic, defined in Equation 2, which allows us to measure or observe the relationship that exists between each pair of variables. This statistic indicates a null value as a non-existent relationship between the variables being measured, and a value close to unity (positive or negative) expresses a relationship (positive or negative) between the variables being compared. By calculating the correlation between each pair of variables using this statistic, the correlation matrix of the variables can be constructed, as shown in Figure 1.

The results obtained in the matrix clearly reflect a high correlation between the variables. This inconvenience must be dealt with, otherwise it can cause problems in the analysis and interpretation of the results obtained. Some steps that can be taken to minimize the problem are as follows Analyze the nature of the correlation (positive or negative) or opt for a dimensionality reduction analysis.

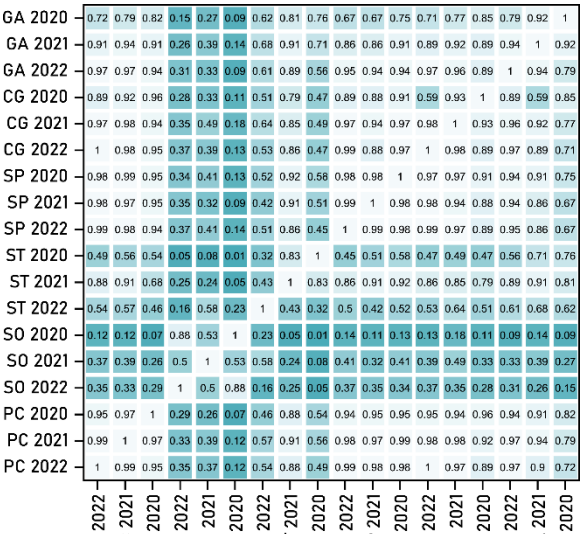

Figure 1. Correlation matrix of study variables

II. RESEARCH METHODS

2.1 Machine Learning

Machine Learning (ML) is one of the most extensive and potential areas of Artificial Intelligence (AI), where one of the objectives of ML is the development of the ability to learn and provide expert recommendations in a narrow domain by relying on adaptive and learning methods or algorithms; the vast majority of these techniques are agglomerated in two large groups: **Supervised and Unsupervised** [7].

The first group of these decide the classification problem, when certain objects are known that make up finite groups and these allow an infinite set of objects to be put together, generally this classification is carried out by an expert. This group is subdivided into linear and non-linear classifiers, being representative of the first subgroup: the Perceptron, Bayesian Classifier, Linear Discriminant Analysis, etc. [8]; and of the second subgroup: Neural Networks, Support Vector Machines, Logistic Regression, Linear Discriminant Analysis and so on [9].

Unsupervised ML methods solve the classification-clustering problem by considering the range of initial indeterminate objects to be clustered with the help of an automatic process based on their properties; the number of clusters can be obtained automatically or initially determined by [10]. From this classification, typical methods are: K-means (K-means), k-medians (k-medians), Fuzzy Clustering (Fuzzy C-Means, Soft K-means), K harmonic means (KHM), among others [11].

Applications of ML in various areas of knowledge are varied, these include physics [12], medical physics [13], robotics [14], mining [15], agriculture [16], biomedicine [17], genetics [18] , medical [19, 20, 21].

2.2 Principal Component Analysis

When studying a phenomenon, it is common to look for as many variables as possible that reflect the nature of the experiment to be analyzed; this generally leads to an increase in the calculation of correlation coefficients and a strong correlation between the variables under study, as well as a difficult visualization between them. The concept of information entropy, which relates to the fact that the more information there is, the greater the variability of the data, should not be overlooked.

The principal component analysis (PCA) technique studies the relationship between *p* correlated variables of an original set, which are transformed into a new set of uncorrelated variables, known as the principal component set [22].

The new variables are linear combinations of the previous ones, with the characteristic of being ordered according to the variability they represent of the data; if the original data are not uncorrelated, we will have a smaller set of *p* variables that represent the highest variability of the data. Algorithm 1 shows the pseudocode of this technique, based on the calculation of eigenvalues and eigenvectors.

Algorithm 1. Principal Component Analysis

Input : Dataset $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$

Dimension d' of the lower dimensional space.

Output : The projection matrix

$\mathbf{W}^* = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{d'})$

1 Center all samples $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$
2 Compute the covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$
3 Perform eigenvalue decomposition on the covariance matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$
4 Take the eigenvectors $\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{d'}$ corresponding to the d' largest eigenvalues

2.3 K-means clustering algorithm

The K-means clustering algorithm was proposed by Mac Queen in 1967 [23], other authors such as Forgy, Lloyd, Hartigan and Wong worked on Mac Queen's algorithm in order to Mac Queen's algorithm, with the aim of improving his proposal. The main use of this algorithm is to obtain qualitative and quantitative information on large multivariate datasets that helps to find only one definitive clustering of the data [24].

The K-means algorithm is the simplest unsupervised algorithm that solves clustering problems, this algorithm is based on defining ***k*** centroids for each cluster, given a certain number of clusters; these centroids must be placed in a cautious manner because different locations may cause a different result, the more distant the centroid is between clusters the better the result will be. Subsequently, each point of the clustering is then taken to associate it with the nearest centroid, in case a point does not exist, the clustering is performed with those values, however, it must be those values, however, the resulting ***k*** new centroids of the previous resulting clusters must be recalculated, with these new centroids we have to assemble the set of with the nearest new centroid, the ***k*** new centroids will change their location step by step until will change their location step by step until there is no longer any change and the centroid does not change its location [25].

$$C_M = \left\| \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x} - \boldsymbol{\mu}_i \right\|_2^2 \qquad (3)$$

K-Means has an optimal iterative nature, whose purpose is to minimize Equation 3, which represents the proximity of an item to be classified and the mean of a cluster, a small value of which indicates a higher intra-cluster similarity; the development of this method is presented in Algorithm 2.

Algorithm 2. k Means Clustering

Input : Dataset $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$

Number of clusters k

Output : Clusters $C = \{C_1, ..., C_k\}$

1 Choose k random , $k \leftarrow \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, .., \boldsymbol{\mu}_k\}$
2 **While** $C_n \neq C$
3 $Ci = \varnothing;\ (1 \le i \le k)$
4 **for** $j = 1, 2, ..., m$
Compute the distance between $\boldsymbol{x}_j$ and media vector
5 $d_{ij} \leftarrow \left\| \boldsymbol{x}_j - \boldsymbol{\mu} \right\|_2$
Cluster Labeling $\boldsymbol{x}_i$
6 $\lambda_j \leftarrow \arg\min d_{ij};\ \ i \in \{1, 2, ..., k\}$
Move $\boldsymbol{x}_i$ its corresponding cluster
7 $C_{\lambda j} \leftarrow C_{\lambda j} \cup \{\boldsymbol{x}_j\}$
8 **end for**
9 **for** $i = 1, 2, ..., k$
Compute the distance between x_j
10 $\boldsymbol{\mu}_i^{'} = \frac{1}{|C_i|}\sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$
11 **if** $\boldsymbol{\mu}_i^{'} \neq \boldsymbol{\mu}_i$ then
12 $\boldsymbol{\mu}_i \leftarrow \boldsymbol{\mu}_i$
13 **else**
14 Leave media vector without change
15 **end if**
16 **end for**

2.4 Mixture of Gaussian Clustering

Unlike k-means, Mixture-of-Gaussian clustering does not use prototype vectors but probabilistic models to represent clustering structures [26]. This technique is a particular type of unsupervised learning algorithm, so called because it assumes that the data points to be clustered are not labelled with the value to be predicted. This method is commonly expressed as a mixture of Gaussians, which can be written as a linear superposition of Gaussians, it's noticed in Equation 4.

$$p(x) = \sum_{i=1}^{k} \alpha_i N(\boldsymbol{x}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \qquad (4)$$

where α_i is called the proportion or weight of the mixture and must satisfy that, $0 \le \alpha_i \le 1$ and $\sum_i \alpha_i = 1$.

Each component represents an individual variable. Each Gaussian is a probability density function that defines the probability of a data value within a given distribution. The model assigns a probability to each cluster, indicating the likelihood that the data point belongs to that class. The Mixture-of-Gaussian clustering algorithm is given in Algorithm 3.

Algorithm 3. Mixture of Gaussian Clustering

Input : Dataset $D=\{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$

Number of clusters k

Output : Clusters $C=\{C_1,...,C_k\}$

1 Choose k random , $k \leftarrow \{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$
2 **repeat**
3 **for** $j=1,2,...,m$
Compute the posterior probabilities $\boldsymbol{x}_j$
4 $\gamma_{ji} = p(z_j = i|\boldsymbol{x}_j)(1 \le i \le k)$
5 **end for**
6 **for** $j=1,2,...,k$
7 Compute de updated mean vector
$$\boldsymbol{\mu}_i^{'} = \frac{\sum_{j=1}^{m}\gamma_{ji}\boldsymbol{x}_j}{\sum_{j=1}^{m}\gamma_{ji}}$$
8 Compute de updated covariance matrix
$$\boldsymbol{\Sigma}_i^{'} = \frac{\sum_{j=1}^{m}\gamma_{ji}(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{'})(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{'})^{\mathrm{T}}}{\sum_{j=1}^{m}\gamma_{ji}}$$
9 Compute the updated mixture coefficients:
$$\alpha_i^{'} = \frac{1}{m}\sum_{j=1}^{m}\gamma_{ij}$$
10 **end for**
11 Update $\{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i | 1 \le i \le k\}$ to $\{\alpha_i^{'}, \boldsymbol{\mu}_i^{'}, \boldsymbol{\Sigma}_i^{'} | 1 \le i \le k\}$
12 **until** The termination condition is met
13 $C_i = \varnothing$; $(1 \le i \le k)$
14 **for** $i=1,2,...,m$
15 Determine the cluster label λ_j of $\boldsymbol{x}_j$
16 Move $\boldsymbol{x}_j$ to the corresponding cluster
$C_{\lambda j} = C_{\lambda j} \cup \{\boldsymbol{x}_j\}$
17 **end for**

Algorithm 3 presents the Gaussian Mixture Clustering model. It starts by initializing the parameters in line 1. Then, in lines 2-12, the parameters are updated iteratively until the maximum number of iterations is reached or the log-likelihood function stops increasing, i.e. its increment is small, giving way to the clustering assignments performed in lines 14-17.

2.5 Agglomerative Hierarchical Clustering

Hierarchical clustering aims to create clustered structures to form a new one or to split an existing one and give rise to underlying ones, being iterative in nature it forms by adopting an agglomerative (bottom-up) or splitting (top-down) strategy [27].

AGNES (AGNES stands for AGglomerative NESting) is a typical hierarchical clustering algorithm that uses the bottom-up strategy. The algorithm first considers each sample of the data set as an initial clustering. Then, at each iteration, two closer clusters are merged as a new cluster, and this process is repeated until the number of clusters reaches the desired value. The key is how to measure the distance between clusters. Since each cluster is a set of data points, we need to define a distance measure over sets. The distance metrics of two clusters C_i and C_j are defined in Equation 5,

$$\begin{aligned} d_{\min}(C_i,C_j) &= \min_{\boldsymbol{x}\in C_i,\boldsymbol{z}\in C_j} dist(\boldsymbol{x},\boldsymbol{z}) \\ d_{\max}(C_i,C_j) &= \max_{\boldsymbol{x}\in C_i,\boldsymbol{z}\in C_j} dist(\boldsymbol{x},\boldsymbol{z}) \\ d_{\text{avg}}(C_i,C_j) &= \frac{1}{|C_i||C_j|}\sum_{\boldsymbol{x}\in C_i}\sum_{\boldsymbol{z}\in C_j} dist(\boldsymbol{x},\boldsymbol{z}) \end{aligned} \quad (5)$$

Algorithm 4. AGglomerative NESting (AGNES)

Input : Dataset $D=\{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$

Cluster distance metric function d

Number of clusters k

Output : Clusters $C=\{C_1,...,C_k\}$

1 **for** $j=1,2,...,m$
2 $C_j = \{\boldsymbol{x}_j\}$
3 **end for**
4 **for** $j=1,2,...,m$
5 **for** $j=i+1,...,m$
6 $\boldsymbol{M}(i,j) = d(C_i,C_j)$
7 $\boldsymbol{M}(j,i) = \boldsymbol{M}(i,j)$
8 **end for**
9 **end for**
10 $q=m$
11 **While** $q>k$
12 Find two clusters C_{i^*} and C_{j^*} with the shortest distance
13 Merge C_{i^*} and C_{j^*}: $C_{i^*} = C_{i^*} \cup C_{j^*}$
14 **for** $j=j^*+1, j^*+2,...,q$
15 $C_j = C_{j-1}$
16 **end for**
17 Delete j^* th row and column of the distance Matrix $\boldsymbol{M}$
18 **for** $j=1,2,...,q-1$
19 $\boldsymbol{M}(i^*,j) = d(C_{i^*},C_j)$
20 $\boldsymbol{M}(j,i^*) = \boldsymbol{M}(i^*,j)$
21 **end for**
22 $q=q-1$
23 **end while**

The minimum distance between two clusters is determined by their nearest samples; the maximum distance is determined by the farthest samples between the clusters; and the average distance is determined by all samples in both clusters.

When the distances between clusters are measured by $d_{\min}$, $d_{\max}$ or d_{avg} the corresponding AGNES algorithms are called single linkage (maximum similarity), complete linkage (minimum similarity) or average linkage. respectively. The pseudocode of AGNES is given in Algorithm 4.

2.6 Optimal Clustering Criteria Akaike and Bayesian Information Criteria

The aim is to minimize the AIC to obtain a network with the best generalization, however, the root mean square error (RMSE) statistics are expected to improve gradually as more parameters are added to the model, the AIC and BIC statistics penalize the model for having more parameters and therefore tend to result in more parsimonious models. Model selection is done by looking for the minimum BIC value. It turns out that the final form of this criterion is quite similar to that of the AIC, but it is noted that the penalty due to the number of model parameters is multiplied by the natural logarithm value of the RMSE. Consequently,

the BIC is more biased than the AIC towards smaller models [28].

The calculation of the criteria is shown in Equation 6.

$$\begin{aligned} \mathrm{AIC} &= n\ln(\mathrm{RMSE}) + 2(p+q) \\ \mathrm{BIC} &= n\ln(\mathrm{RMSE}) + (p+q)\ln n \end{aligned} \qquad (6)$$

Silhouette Criteria

The silhouette coefficient is a metric used to calculate the goodness of a clustering technique; the range is from -1 to 1. When the value of 1 is obtained it means that the clusters are located in the center of the assigned cluster, if it is 0, the distance between clusters is on the border, finally, if the value is minus 1, the clusters are misassigned. This metric is based on cluster geometry [29].

Silhouette function is shown in Equation 7,

$$\begin{aligned} s(i) &= \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, \quad \text{where} \\ a(i) &= \frac{1}{|C_I|-1}\sum_{j\in C_I, i\neq j} d(i,j), \\ b(i) &= \min_{k\neq 1}\frac{1}{|C_k|}\sum_{j\in C_k} d(i,j) \end{aligned} \qquad (7)$$

Calinski-Harabasz Criteria

The Calinski-Harabasz criteria is based in the degree of dispersion between clusters [29], this criterion is shown on equation 8,

$$\begin{aligned} CH(k) &= \frac{B(k)(N-k)}{W(k)(k-1)}, \quad \text{where} \\ B(k) &= \sum_{i=1}^{k} a_i \left\|\overline{x_i}-\overline{x}\right\|^2 \\ W(k) &= \sum_{i=1}^{k}\sum_{C(j)=i}\left\|x_j-\overline{x_k}\right\|^2, \end{aligned} \qquad (8)$$

Where k is the corresponding number of clusters, $B(k)$ is the inter-cluster divergence, $W(k)$ is the intra-cluster divergence and n is the samples. If $B(k)$'s value is large, the degree of dispersion is high between cluster, if $W(k)$ is smaller, a closer relationship between clusters exists. When $CH(k)$ index is higher the clustering effect is better.

III. RESULTS

Below are the performance plots with the normalized data for the selected BIC, AIC, Silhouette and Calinski metrics. These are grouped in pairs and accompanied by the normalized absolute difference plot; the latter helps us to estimate the number of clusters in which the best performance is observed.

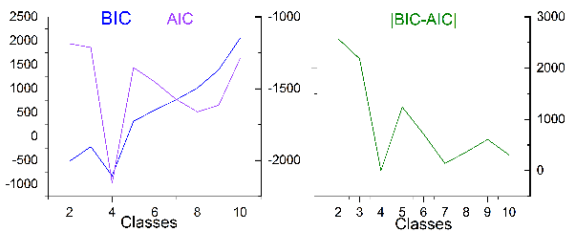

Figure 1. Left: Performance of BIC and AIC metrics. Right: Normalized absolute difference of these metrics.

The graph of the BIC and AIC metrics in Figure 1 and the pair of BIC and Calinski in Figure 2 share the value of 4 clusters as the best for both combinations. In contrast, the combination of BIC and Silhouette shown in Figure 3 has a value of 5 clusters.

The above data on optimal metrics allow us to estimate that between 4 and 5 clusters will be the best values representing the hidden patterns of the quantitative features of the insurers. The 18 indicators of the 43 insurers are reduced to 35% of them, i.e. 6 new normalized characteristics that account for 99.1291% of the variance of the data using the PCA technique.

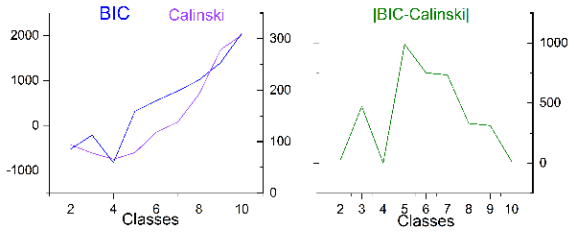

Figure 2. Left: Performance of BIC and Calinski metrics. Right: Normalized absolute difference of these metrics.

These data are clustered in search of these hidden patterns, using the three selected unsupervised classification methods; for both, optimal values of the number of clusters are obtained with the metrics in Figures 1-3.

The clustering obtained by the Mixture of Gaussian method is defined as the ground truth, since this method is an improvement of K-means and tends to perform better than the AGNES method. This ground truth is compared with the clusters obtained by the other classification methods, which allows us to estimate the similarity between them, using two classical metrics for estimating the performance of classification algorithms among the clusters obtained by the selected algorithms.

The first of these is the confusion matrix, which is a mechanism for evaluating the performance of a machine learning algorithm. Where each row of the matrix represents the instances of the expert class, the columns represent the number of predictions of each [30].

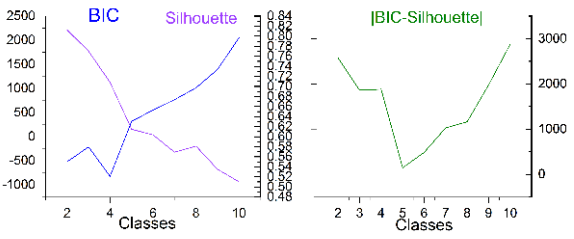

Figure 3. Left: Performance of BIC and Silhouette metrics. Right: Normalized absolute difference of these metrics.

Accuracy is the second metric chosen, where q is used to evaluate the performance of correct predictions relative to the total number of elements, and is defined in Equation 9.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{All Samples}} \quad (9)$$

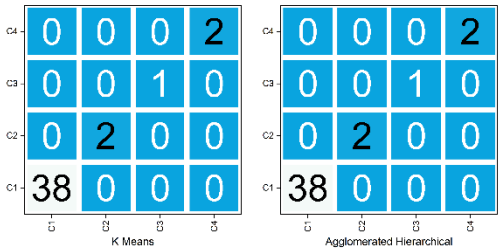

Figure 4. Confusion matrices for 4 clusters. Left: K Means. Right: AGNES.

By their very nature, unsupervised clustering techniques predict labels of dissimilar value, the latter being required for the calculation of accuracy and confusion matrices. This is solved by relabeling the predictions given by the clustering method to be compared and the ground truth one; the labels of the latter are sorted in ascending order, creating blocks of clusters; the indices of the vector of predictions obtained are applied to the vector of predictions given by the method to be compared. The process ends when, for each block, the mode is obtained, this metric being the new label: if the result is multimodal, the value of the mode (least repeated in the other blocks) is selected. If splices persist, they are broken randomly.

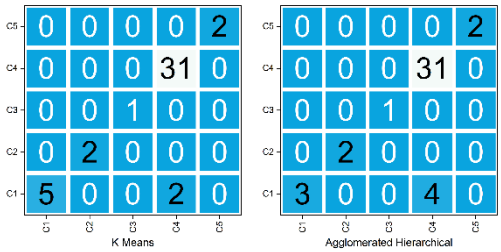

Figure 5. Confusion matrices for 5 clusters. Left: K Means. Right: AGNES.

The accuracy achieved by the two clustering methods with respect to the ground truth by clustering in four classes is 100% for both, this complete self-recognition is observed in both confusion matrices, where two insurers are seen in the second and fourth clustering, one in the third and the rest in the first, the matrices are shown in Figure 4. When grouping the insurers in five classes, an accuracy of 95.34% is obtained for K-means and 90.69% for AGNES, observing in the confusion matrices the discrepancy in the first grouping, these matrices are presented in Figure 5.

The accuracies obtained with four and five clusterings correspond to the number of classes given by the optimal metrics estimated by the tests in Figures 1-3, observing that the clustering in four classes gives equal rankings by the three clustering algorithms chosen for this research; this number of clusters makes sense to the one given by the government institution, but with a non-arbitrary number of insurers per category, but rather by their hidden patterns.

Table 3 shows the clustering predicted by the Mixture-of-Gaussian method for each insurer, identified by its ID from Table 2, for four and five classes.

Table 3. Mixture of Gaussian clusters of Venezuelan insurers with 4 and 5 classes.

ID	4 Classes	5 Classes	ID	4 Classes	5 Classes
1	1	4	23	2	4
2	1	4	24	1	4
3	1	4	25	1	4
4	1	4	26	1	1
5	1	1	27	1	4
6	1	4	28	1	3
7	1	4	29	3	4
8	1	4	30	1	4
9	1	2	31	1	4
10	2	5	32	1	4
11	4	4	33	1	4
12	1	1	34	1	4
13	1	5	35	1	4
14	4	4	36	1	4
15	1	4	37	1	1
16	1	1	38	1	4
17	1	4	39	1	4
18	1	4	40	1	4
19	1	4	41	1	4
20	1	1	42	1	4
21	1	1	43	1	4
22	1	2			

The predictions for 4 and 5 classes do not show a simple relationship between them; it still retains the characteristic that one grouping concentrates the majority of insurers, being 88.37% and 72.09% for four and five classes respectively.

The four-class cluster retains the same predictions by reducing to 10% of its original characteristics, this percentage is achieved with the first pair of principal components, which represent 86.61% of the variability of the data; this reduction helps us to construct its scatter plot, in which the intrinsic nature of its clusters is noticed, this graph is shown in Figure 6.

The analysis of the 43 insurance companies in this same number of groups, which reflect a stratified view of the Venezuelan insurance sector, reveals the diversity and singularities that characterise these companies in the market. The second group includes Caracas C.A. Seguros and Mercantil C.A. Seguros, two of the largest insurers in the country. The fourth group concentrates Caroní C.A Seguros and Corporativos C.A Seguros, two companies with the same level of collected premiums. The third grouping shows a unique pattern with the exclusive presence of Pirámide C.A Seguros, which has shown remarkable growth in the market in recent years. Finally, the largest grouping, with the remaining 38 insurance companies, shows a heterogeneity that makes it difficult to break down and identify similar features.

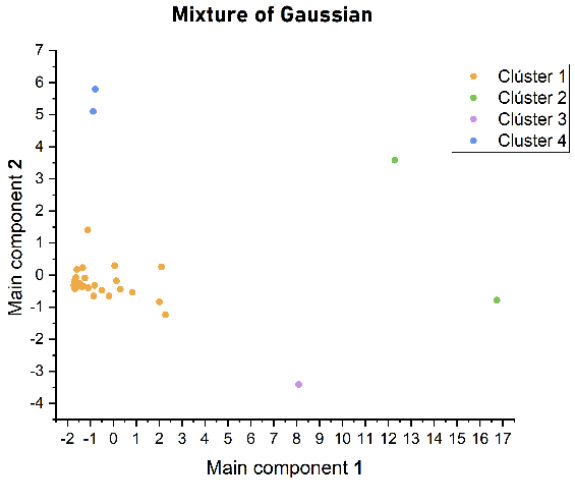

Figure 6. Scatter plot for four classes

IV. CONCLUSIONS

Clustering techniques that do not require prior labelling (unsupervised learning) help to find hidden patterns in the data; the nature of these techniques has been applied in this study, which has allowed us to scrutinize that the groupings of insurers given by the government institution do not obey this nature. In addition to the fact that the use of the premiums collected criterion represents a preponderant factor, it is not sufficient for this purpose on its own.

The data are efficiently grouped with four and five groupings, showing that the first of these two groupings achieve 100% accuracy even when the principal components are reduced to 10%. These predictions support the ranking of insurers with clear and unbiased criteria when choosing an insurer, this will result in a healthy growth of the Venezuelan insurance market.

It is recommended that the study be expanded to include other variables that could be factors for comparison, such as: branches nationwide, number of employees, policies issued, policies cancelled, number of intermediaries and their ranking, etc. Thus, the ranking methods will operate with valuable information in the discrimination of the groupings.

Acknowledgements

Jorge is grateful to UTEL for the academic training he received while studying for his Master's degree in Data Science. Elizabeth is grateful to UTEL. Luis David thanks the National System of Researchers (SNI-CONAHCYT) for the support through the grant number 332238.

REFERENCES

[1] C. Hernández Avendaño, "La industria aseguradora en el PIB de Colombia," *Revista Faselcolda*, vol. 189, pp. 2-10, 2023.

[2] S. A. Guevara and G. B. Ruegeles, "Importancia de la Industria aseguradora en el encuentron y desarrollo económico del país.," pp. 19-25, 1985.

[3] W. Mayorga, "2013. Un año en transacción," *Revista Fasecolda*, pp. 18-23, 2014.

[4] G. B. d. Venezuela, 2022. [Online]. Available: https://www.sudeaseg.gob.ve/.

[5] Á. d. C. Aragón Gallegos, S. I. Cerquín Silva, R. O. Escurra Yactayo and A. L. Roncalla Viena, "Segmentación de clientes para mejorar la experiencia de compra de productos electrónicos en Falabella," Creative Commons, Perú, 2023.

[6] M. E. Porter, Estrategia competitiva: Técnicas para el análisis de los sectores industriales y de la competencia., México: Grupo Editorial Patria, 2015, pp. 68-73.

[7] Z.-H. Zhou, Machine Learning, Australia-China: Springer, 2016.

[8] M. Ravil, "Machine learning methods: An overview," *Computer modelling and new technologies*, vol. 19, pp. 14-29, 2015.

[9] A. Hamza, I. Hamza Awad, N. Sulaiman Mohd, M. Aliyu and B. Abuagla, "Taxonomy of Machine Learning Algorithms to classify real-time Interactive applications," *International Journal of Computer Networks and Wireless Communications 2*, pp. 69-73, 2012.

[10] T. Ayodele, "Types of machine learning algoritms," *New advances in machine learning*, vol. 3, pp. 19-48, 2010.

[11] W. Barbakh, "Review of Clustering Algorithms. In: Non-Standard Parameter Adaptation for Exploratory Data Analysis," in *Studies in Computational Intelligence*, vol. 249, Springer, Berlin, Heidelberg, 2009, pp. 7-28.

[12] F. N. Khan, Q. Fan, C. Lu and A. P. T. Lau, "An optical communication's perspective on machine learning and its applications," *Journal of Lightwave Technology*, vol. 37, no. 2, pp. 493-516, 2019.

[13] E. Harki and Z. S. Rashid, "Analysis of factors that affect radiation dose level during interventional cardiology procedures using logistic regression," *Revista Mexicana de Física*, vol. 69, no. 3, pp. 1-8, 2023.

[14] D. Kim, S.-H. Kim, T. Kim, B. B. Kang, M. Lee, W. Park, S. Ku, D. Kim, J. Kwon, H. Lee, J. Bae, Y.-L. Park, K.-J. Cho and S. Jo, "Review of machine learning methods in soft robotics," *Plos One*, p. 16, 2021.

[15] J. T. Mc. Coy and L. Auret, "Machine learning applications in minerals processing: A review," *Minerals Engineering*, vol. 132, pp. 95-109, 2019.

[16] A. Sharma, A. Jain, P. Gupta and V. Chowdary, "Machine learning applications for precision agriculture: A comprehensive review," *IEEE Access*, vol. 9, pp. 4843-4873, 2020.

[17] L. Patel, T. Shulka, X. Huang, D. W. Ussery and S. Wang, "Machine Learning Methods in Drug Discovery," *Molecules*, vol. 25 (22), no. 5277, 2020.

[18] M. W. Libbrecth and S. W. Nonble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics volume*, vol. 16, pp. 321-332, 2015.

[19] B. Abbasi and D. M. Goldenholz, "Machine learning applications in epilepsy," *Epilepsia*, vol. 60, no. 10, pp. 2037-2047, 2019.

[20] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8-17, 2015.

[21] M. Field, N. Hardcastle, M. Jameson, N. Aherne and L. Holloway, "Machine learning applications in radiation oncology," *Physics and Imaging in Radiation Oncology*, vol. 19, pp. 13-24, 24 06 2021.

[22] T. Kurita, "Principal component analysis (PCA).," *Computer Vision: A Reference Guide*, pp. 1-4, 2020.

[23] J. MacQueen, "Classification and analysis of multivariate observations.," in *5th Berkeley Symp. Math. Statist. Probability.*, Los Angeles California, 1967.

[24] J. Wang and S. Xialong, "An improved K-Means clustering algorithm," in *2011 IEEE 3rd International Conference on Communication Software and Networks*, China, 2011.

[25] J. Wu, "Cluster Analysis and K-means Clustering: An Introduction," in *Advances in K-means Clustering*, Berlin, Heidelberg, Springer, Berlin, Heidelberg, 2012, pp. 1-26.

[26] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Procedia Computer Science*, vol. 171, pp. 158-167, 2020.

[27] A. Alam, M. Muqueem and S. Ahmad, "Comprehensive review on Clustering Techniques and its application on High Dimensional Data," *IJCSNS International Journal of Computer Science and Network Security*, vol. 21, no. 6, pp. 237-244, 2021.

[28] C. Agiakloglou and A. Tsimpanos , "Evaluating the performance of AIC and BIC for selecting spatial econometric models," *Journal of Spatial Econometrics*, pp. 1-35, 2023.

[29] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 5, pp. 1-6, 2019.

[30] R. B. Pereira, A. Plastino, B. Zadrozny and L. H. C. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Information Processing & Management*, vol. 54, no. 3, pp. 359-369, Mayo 2018.

[31] P. Ferrando, U. Lorenzo and J. Muñiz, "Decalogue for the factor analysis of test items," *Psicothema*, pp. 7-17, 2022.

[32] V. Nateski, "An overvies of the supervised machine learning methods," *Horizon*, pp. 51-62, 2017.

[33] W. A. Barbakh, Y. Wu and C. Fyfe, "Review of Clustering Algorithms," in *Studies in Computational Intelligence book series (SCI,volume 249)*, vol. 249, Berlin, Springer, Berlin, Heidelberg., 2009.

[34] P. J. Hardin and J. M. Shumway, "Statistical significance and normalized confusion matrices," *Photogrammetric engineering and remote sensing*, vol. 63, no. 6, pp. 735-739, 1997.