Analogy of the Application of Clustering and K-Means Techniques for the Approximation of Values of Human Development Indicators

José Luis Morales Rocha, Mario Aurelio Coyla Zela, Nakaday Irazema Vargas Torres, Genciana Serruto Medina
Gestión Pública y Desarrollo Social
Universidad Nacional de Moquegua, Moquegua, Perú

***Abstract*—The objective of this study was to apply Clustering and K-Means' techniques to classify the departments of Peru according to their Human Development Index. In this article, the elbow method was used to determine the optimal number of clusters, applying the classification algorithms to group the departments of Peru according to their similarities, in addition to the Principal Component Analysis (PCA) technique for a better display of clusters. After applying the unsupervised algorithms, the results were more relevant in clusters 2 and 4 according to their HDI, made up of the departments of Arequipa, the Constitutional Province of Callao, Ica, Lima, Moquegua and Tacna, where the most notable is the life expectancy at birth, the population with full secondary education, the number of years of education, the average per capita income, and the state's density index. The results obtained by the K-Means algorithm show more cohesive results than the Clustering algorithm.**

Keywords*—Clustering; K-Means; elbow method; cohesion; separation; human development index

I. Introduction

The main reason for developing this research is to establish indicators of similarities and identify which departments have a lower or higher level of HDI for the characteristics analyzed and manage better levels of life expectancy, access to education, income level and index of density of the State, in the different areas that allow an adequate formulation of public policies and prioritize the social agenda that allows better opportunities and degree of progress and equality of citizens.

The research proposes the application of unsupervised Machine Learning algorithms (K-Means and Clustering) to observe the formation of clusters, with their respective indicators, grouping the departments of Peru into four clusters, according to the similarities between them, to measure human development through life expectancy, access to education and income level.

In this research, unsupervised learning algorithms were proposed to group the departments into clusters, according to optimization criteria; being one of the most used the K-Means; this algorithm ranks the indicators into clusters. In [1] K-Means is a partition grouping technique; the data objects are divided into groups that do not overlap. In [2] The clusters allow interaction in external networks in which information flows and facilitates its transfer. For [3] clustering techniques meta-learning tools are useful to analyze the knowledge produced by modern applications. The elbow method is used to determine the optimal number of clusters and a suitable observation, fixing the distances between each cluster.

The most relevant departments according to their HDI are found in cluster 2 and cluster 1 are perceived with the lowest HDI values, so the State must provide public policies focused on the populations of the departments in cluster 1.

The use of the K-Means and Clustering algorithms require the classification of groups with similar characteristics, according to the Human Development Indicators (HDI) with a high incidence due to altitude and State Density Index; therefore, quality information for decision-making in the design of public policies to improve HDI by departments and regions is provided.

The structure of the research article presents the state of the art according to the study variables, theoretical background emphasizing classification techniques, determined the results of the clusters, the discussion and conclusions of the research.

II. Related Work

To achieve high precision in terms of time and space, in [3] considers K-Means to be the best option for large and categorical data. It concludes that the K-Means genetic algorithm is faster than evolutionary algorithms. In [4] two clustering methods: K-Means and hierarchical clumping in air pollution studies were reviewed, with the aim of providing a review of clustering applications, specifically by using hierarchical clustering and k- mean. It was stated that each grouping technique has its own advantages and disadvantages and there is no a "best" method.

According to [5] the performance of classification algorithms is influenced by certain characteristics of the data sets on which they are modeled, such as imbalance in class distribution, class overlap, and lack of density. At the same time, the circumstances of class overlap and lack of density of the minority class in unbalanced data sets are observed.

As [6] artificial intelligence in medicine shows that ultrasonic imaging technologies have a true diagnosis; Two types of neural network algorithms have been proposed in three categories: USCT images of healthy, fractured and osteoporotic bones. Initially, a Convolutional Neural Network classifier system is presented and then an evolutionary neural network with the AmeobaNet model for the USCT images

classification. In [7] emotion recognition through an artificial neural network that detects spoken expressions, proposing a regularized Bayesian artificial neural network model that recognizes emotions through speech. The Berlin database with 1470 samples of emotions: 500 angry, 300 happy, 350 neutral and 320 sad. The performance of the methodology is compared with other avant-garde ones used for the same purpose, the proposed methodology achieved 95% precision in the recognition of emotions, being one of the highest compared to other methodologies used.

In [8] Hybrid approaches to data classification and optimization algorithm increase the precision of data classification. The study performs Moth Flame (MFO) and Fuzzy Min Max Neural Network (FMMNN) optimization applications to classify medical data. In terms of classification, the experiment achieved 97.74% accuracy for liver disorders and 86.95% accuracy for the diabetes data set that is related to the achievement of good human health.

As [9] states, data analysis is used as a tool in different fields, clustering plays an important role in the composition of the data analysis, thus dealing with the segmentation of the data structure in an unknown segment; using the K-Means algorithm. This article explains the applications of clustering methods and the objectives of clustering with big data. It also introduces the clustering technique for identifying data patterns by performing sample data analysis.

In [10] the research examines the CatBoost ranking algorithm on loan approval and staff promotion. This algorithm outperforms other implemented classifiers. Two types of analysis were carried out, in the first one the amount, the type, the income of the applicant and the purpose of the loan that help to predict the approvals of the loan were considered, in the second case the division, the schooling abroad, the geopolitical zones, qualification and working years, which had a high impact on the promotion of personnel. Based on the performance of CatBoost, the algorithm is interesting for a better prediction of loan approvals and staff promotion.

In [11] the K-Nearest Neighbor algorithm is used in multidimensional and outlier data due to its precision. A hybrid K-Nearest Neighbor approach with optimized particle scoring to improve K-Nearest neighbor performance, which is implemented in two stages: it first resolves multidimensional data by selecting the features with the swarm optimization algorithm and the second resolves the presence of outlier's values with the results of stage 1 and applying a new K-Nearest Neighbor technique scored.

In [12] computer diagnosis of tumors is important, as their segmentation is difficult to diagnose. The Fuzzy K Means fast clustering algorithm based on super pixels was used. These images bring a multi-scale morphological gradient reconstruction operation that allows getting segmentation precision. The results reveal that this approach is fast and accurate compared to segmentation algorithms; which provide a high precision of 99.58% and an improved RFN value of 8.34% compared to other methods analyzed. In [13] the logistic regression, K-NN applied to the data set in breast cancer, was found to determine the well based prediction of the data set. Also, with logistic regression an accuracy of 91% was achieved, and the detection was early and accurate. Likewise, it is seen that to reduce and classify heart disease, the support vector machine (SVM) has been adopted, the closest K-NN neighbors and the linear discriminant analysis. It has been shown that the vector machine turned out to be a better classifier with an accuracy of 80.4%.

In [14] the paper uses fine-tuning transfer learning on RNA-Seq gene expression data, classifying 5 types of cancer that affect women. The data comes from the genomic data commons (GDC) portal, with 2166 samples, along with 19,947 common genes. Spearman's correlation was used to narrow down the number of genes, eliminating those that are highly correlated. Gene expression is filtered by selecting values greater than 0.25 in the samples. In the gotten profile, the samples are transformed into 2D images as data, adapting to the convolutional layer of the CNN architecture. We fit four previously trained models on the RNA-Seq gene expression data, namely ResNet50, DenseNet, Xception and VGG16. The Xception architecture shows the highest and most accurate performance 98.6%, recovery 97.8%, and F1 score of 98% in a five-time cross-validation test and training approach.

According to [15] the study proposes a model based on machine learning to predict new infections expected by COVID 19. The model is tested in Egypt and in the 10 highly rated countries in September 2020. The proposed model is implemented based on algorithms supervised machine learning regression. Then compared with one of the more accurate prediction models The Bayesian crest, and the results show the power of the model compared to its counterpart in all the countries studied.

III. Theoretical Background

The rapid development of data collection techniques and new storage technologies [11] have allowed organizations to retain a large amount of data. With the help of machine learning algorithms, the quality of decision-making can be supported thus human error can be avoided. Classification [16] are supervised techniques that categorize unknown data into a specific class or group. In classification, the classes are known in advance.

A. K-Means

In [17] the K-Means technique is a clustering algorithm, machine learning technique. In [18] K-Means is a partition clustering technique; data objects are divided into groups that do not overlap.

The K-Means algorithm [3] groups clusters iteratively. Calculate the distance means, using an initial centroid, with each class that is represented by the centroid, using the distance as a metric and giving the k classes in the data set. In the K-Means algorithm, the mean value of the elements within the group is represented in the center of each group. The K-Means algorithm [19] groups the data into groups, defining a fixed number of groups, assigning data iteratively to the groups formed by adjusting the centers in each group.

The K-Means technique [18] learns the characteristics of a data set and forms partitions with them, these partitions are called clusters, which represent data with similar features. For

numerical data, each group is represented by a centroid, which is the mean of the elements in the group. For categorical variable data, it corresponds to the object that occurs most frequently, which is used as the group prototype.

K-Means uses the squared Euclidean distance as a measure of similarity for cluster membership:

$$d_{sq} = \sum_{i=1}^{D}(x_i - y_i)^2 \quad (1)$$

In (1) x and y are points in a D-dimensional space. The number of clusters k is determined by minimizing the sum of squared errors (SSE), which is given by the sum of the squared error in each data pair and its closest centroid. It is given by (2).

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\left\|x_j - c_j\right\|^2 \quad (2)$$

where c_j is the centroid of the j-th group, and $w_{ij} = 1$ if the data pair x_i is in group *j* and w_i, j = 0, if x_i is not in group *j*.

The K-Means algorithm [20] randomly selects k data points from an original data set to later add them as the center of the initial clustering. First, each piece of data is considered a data point. Then, the Euclidean distance algorithm is used to determine the distance between the data points and the cluster center, the data set is preliminarily clustered according to the distance. Finally, the average distance of the data in each group is calculated, and the center of the group is adjusted, and the final result of the grouping is obtained through multiple iterations.

B. Clustering

Clustering techniques [21] are used in different areas of research, such as data classification, taxonomy, document retrieval, image segmentation and pattern classification. The Clustering algorithm [18] is the technique of grouping elements using a similarity measure. The grouping can be hierarchical or partitioned, exclusive, overlapping or fuzzy, and complete or partial. The Clustering algorithm [3] presents as a result the reduction of the dimensionality of a data set. The goal of a clustering algorithm is to identify the various groups within a data set.

The clustering technique [22] is a method to group data into classes with identical characteristics in which the similarity between classes is maximized or minimized. Grouping is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes.

C. Types of Clustering

Clustering is divided into two types:

1) Hard clustering: each data point is or is not part of a cluster. It means that each element is grouped into one of the k groups.

2) Soft clustering: A probability is assigned to the data point to be in certain clusters instead of placing each data point in a separate cluster, a probability of being in k groups is assigned to each element.

D. Clustering Methodologies

Because the Clustering technique is subjective. Cluster analysis is not an automated activity, but an iterative information discovery process or a multi-objective collaborative optimization that involves trial and error.

E. Validation of the Classification Algorithms

As the goal of clustering is to group similar objects in the same cluster and different objects to be placed in different clusters, [23] internal validation metrics are usually based on two criteria:

1) Cohesion: The element of each cluster must be as close as possible to the other elements of the same cluster.

The Sum of Squared Within (SSW), internal measure to evaluate the Cohesion of the clusters the grouping algorithm generated is:

$$SSW = \sum_{i=1}^{k}\sum_{x\in C_i} dist^2(m_i, x) \quad (3)$$

where *k* is the number of clusters, *x* a point of cluster c_i and m_i the centroid of cluster c_i.

2) Separation: Clusters must be widely separated from each other. There are different approaches to measure this distance among cluster: distance between the closest member, distance between the most distant members, or the distance among centroids.

The Sum of Squared Between (SSB), a measure of separation used to evaluate the inter-cluster distance is given by:

$$SSB = \sum_{j=1}^{k} n_j\, dist^2\left(c_j, \bar{x}\right) \quad (4)$$

where k is the number of clusters, n_j is the number of elements in cluster *j*, c_j is the centroid of cluster *j*, and x is the mean of the data set.

IV. Results

This section shows the results obtained from the application of Unsupervised Machine Learning algorithms (K-Means and Clustering).

A. Indicators used

The following indicators were used in the application of Machine Learning techniques for the classification of the Human Development Index.

- Human development Index.
- Life expectancy at birth.
- Population with full secondary education (18 years).
- Years of education (Population aged 25 and over).
- Family income per capita.
- Altitude.
- State Density Index.

B. The Elbow Method

The criterion used to establish the number of clusters to be used was determined by the elbow method.

The elbow method uses the mean distance of the observations to their centroid. The larger the number of clusters k, the intra-cluster variance decreases more. The smaller the intra-cluster distance the better it is, since it means that the clusters are more compact. The elbow method looks for the value k that satisfies that an increase in k does not substantially improve the mean intra-cluster distance.

According to Fig. 1, 4 clusters were established for the classification of the Human Development Index.

C. K-Means

To graphically illustrate the formation of the clusters and because there are seven indicators and it is not possible to make a graph that represents all these characteristics, a technique called Principal Component Analysis (PCA) was used, which reduces the quantity of variables to be analyzed, in this case to be visualized, creating a smaller quantity of new variables that best represents the original variables.

In Fig. 2, the graph of the HDI classification is shown, by means of the two main components, coloring it according to the cluster to which each department belongs according to its HDI.

In this graph, it is observed that each of the departments are well defined according to their HDI, in components 1 and 2, each of the departments is represented with the points and with the colors the cluster to which they belong.

The clusters are organized according to the following colors:

Cluster 1: Blue

Cluster 2. Green

Cluster 3: Red

Cluster 4: Yellow

According to Fig. 3, it is observed that there is a significant difference between the four clusters, the number 2 presents a higher HDI, followed by cluster 4, meanwhile cluster 3 and cluster 1 present a low HDI. In addition, it can be seen that cluster 3 presents greater dispersion and cluster 4 less dispersion.

Table I shows that the most relevant departments according to their HDI are found in cluster 2, made up of the departments of Arequipa, the constitutional province of Callao, Ica, Lima, Moquegua and Tacna. The positions in favor of these departments are found in almost all their dimensions, making life expectancy at birth more noticeable, in the population with full secondary education, years of education, average per capita income and the state's density index.

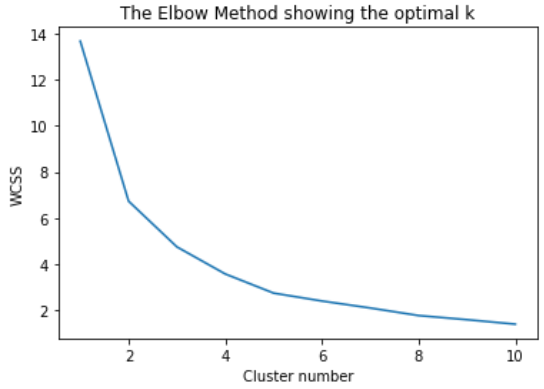

Fig. 1. The Elbow Method.

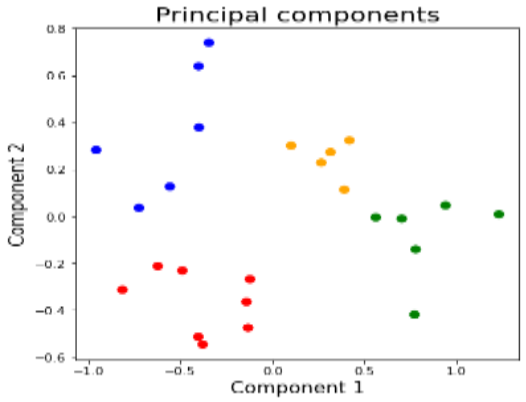

Fig. 2. Principal Components – Kmeans.

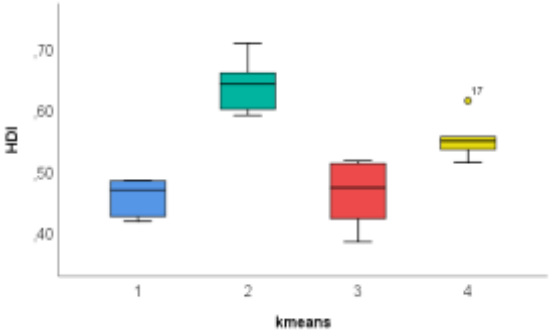

Fig. 3. Box Plot – K-Means.

TABLE I. CLASSIFICATION USING K-MEANS

Cluster	HDI	Life expectancy at birth	Population with full secondary education	Years of education	Family income per capita	Altitude	State Density Index
1	0.457	71.77	50.12	7	728.96	1359.00	0.66
2	0.639	76.85	74.11	10	1184.19	843.50	0.79
3	0.463	73.08	66.37	8	633.27	3383.75	0.72
4	0.552	75.74	64.92	8	937.99	74.40	0.74

In the departments of cluster 1, the lowest HDI values are noted, made up of the departments of Amazonas, Cajamarca, Huánuco, Loreto, San Martin and Ucayali. Although life expectancy at birth is relatively high (approximately 72 years), the figures for the population with full secondary education are approximately 50%, 7 years of education on average, it is appreciated that there is an average per capita family income of S /. 728.96 and a state density index of 0.66.

Cluster 4 is made up of the departments of La Libertad, Lambayeque, Madre de Dios, Piura and Tumbes; cluster 3 made up of the departments of Ancash, Apurímac, Ayacucho, Cusco, Huancavelica, Junín, Pasco and Puno.

D. Clustering

In Fig. 4, the dendrogram for the classification of the departments of Peru according to the HDI is shown.

In Fig. 5, the graph of the classification of the departments of Peru according to their HDI is shown, by means of the Clustering algorithm, through the two main components, coloring it according to the cluster each department belongs based on its HDI.

In this graph it is observed that each of the departments are also well defined according to its HDI.

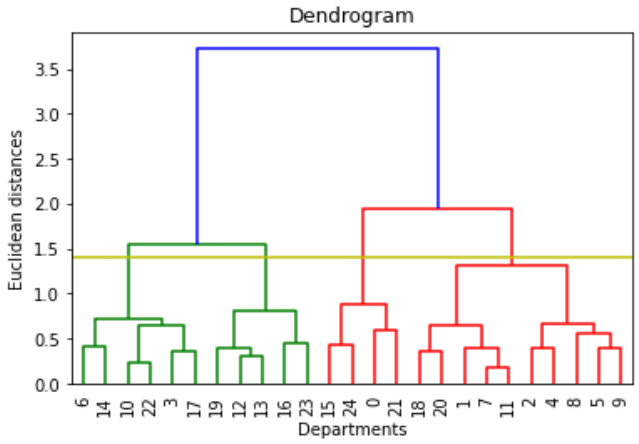

Fig. 4. Dendrogram – Clustering.

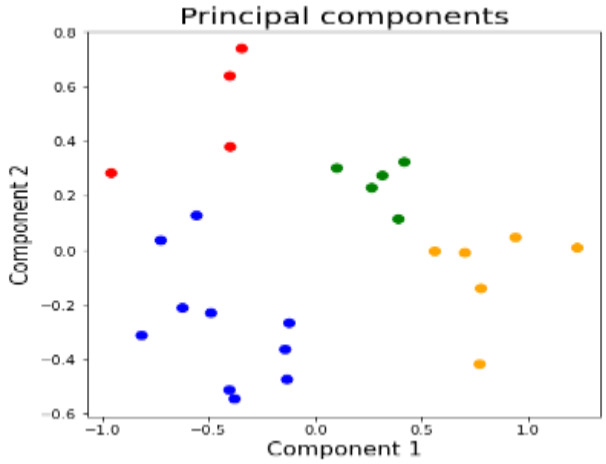

Fig. 5. Principal Components – Clustering.

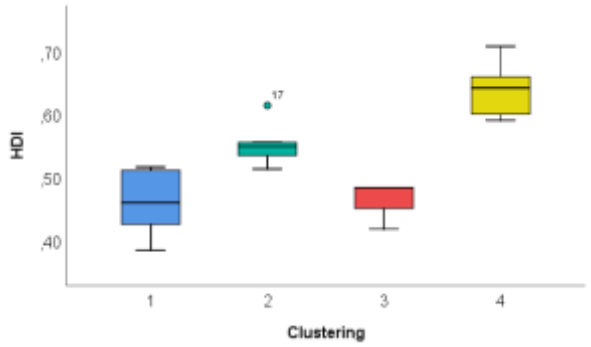

Fig. 6. Box Plot – Clustering.

According to Fig. 6, in the box-and-whisker plot of the Clustering algorithm, it is also observed that there is a significant difference among the four clusters, cluster 4 has a higher HDI, followed by cluster 2, cluster 3 and cluster 1 have a low HDI. In addition, it can be seen that cluster 1 presents greater dispersion and cluster 2 lower dispersion.

Table II displays that the most relevant departments according to their HDI are found in cluster 4, made up of the departments of Arequipa, the constitutional province of Callao, Ica, Lima, Moquegua and Tacna. The positions in favor of these departments are also given in almost all their dimensions, highlighting the life expectancy at birth, the population with full secondary education, the years of education, the average per capita income and the density index of the state.

In cluster 1 departments, made up of Ancash, Apurímac, Ayacucho, Cajamarca, Cusco, Huancavelica, Huánuco, Junín, Pasco and Puno, the lowest HDI values are perceived. Although life expectancy at birth is relatively high (approximately 73 years) and the population with full secondary education (64%), above 7 years of education on average are shown, an average per capita family income of S /. 635.10 and a state density index of 0.71.

The departments of La Libertad, Lambayeque, Madre de Dios, Piura and Tumbes compound cluster 2, and cluster 3 the departments of Amazonas, Loreto, San Martin and Ucayali.

E. Internal Validation Metrics

1) Cohesion: internal measure to evaluate the Cohesion of the clusters of the clustering algorithms:

$SSW_{Clustering} = 1.90$

$SSW_{K\text{-}Means} = 1.74$

The Sum of Squared Within (SSW) of the K-Means algorithm (1.74) shows more cohesive clusters than the Clustering algorithm.

2) Separation: measure of separation used to evaluate the inter-cluster distance.

$SSB_{Clustering} = 4.9856$

$SSB_{K\text{-}Means} = 4.9859$

The Sum of Squared Between (SSB) of the K-Means and Clustering algorithms show similar inter-cluster distances.

TABLE II. CLASSIFICATION USING CLUSTERING

Cluster	HDI	Life expectancy at birth	Population with full secondary education	Years of education	Family income per capita	Altitude	State Density Index
1	0.459	73.05	64.00	7	635.10	3172.20	0.71
2	0.553	75.74	64.92	8	937.99	74.40	0.74
3	0.467	71.19	47.91	8	772.25	875.50	0.66
4	0.640	76.85	74.11	10	1184.19	843.50	0.79

V. DISCUSSION

According to the objective, to apply Clustering and K-Means techniques to classify the departments of Peru according to their Human Development Index, the results exhibit in Table I show that the departments with greater relevance according to their HDI are in cluster 2, positions in favor of these departments arise in almost all their dimensions, making life expectancy at birth more noticeable, in the population with full secondary education, the years of education, the average per capita income and the density index of the state, these results were achieved using the K-Means technique. According to the Clustering technique, Table II shows the results obtained with the most relevant departments according to their HDI in cluster 4, positions in favor of these departments are also given in almost all their dimensions, making the hope of life at birth, population with completed high school, years of education, average per capita income, and state density index more relevant. The results obtained by the K-Means algorithm show more cohesive results than the Clustering algorithm.

Results that when compared with what was found by [3], who determined that the K-Means algorithm shows better results for the classification of big data. The author in [4] claims that K-Means and hierarchical clumping techniques have their own advantages and drawbacks and there is no "best" method. These results can affirm that the K-Means algorithm shows significant results regarding to the Clustering classification algorithms, especially in the cohesion measures.

On the other hand [24] K-Means is a classic prototype-based clustering technique that attempts to group data into K groups specified by the user. In [22], [25] Clustering is a method to group data into classes with identical characteristics, in which intraclass similarity is maximized or minimized.

VI. CONCLUSION

According to the K-Means algorithm, it identifies cluster 2 with the highest HDI because it groups the departments in the Coastal Region with higher population density on average, higher per capita income, strategic geographic location in metropolitan areas and zones of industrial, commercial, agricultural and mining activity that contribute to development, greater employment, health and education, contribution by mining canon and increase in government investment.

The K-Means algorithm accurately determines the grouping by departments in cluster 3, which shows a lower level of HDI doing its classification by similar characteristics of geographical location, belonging to the Sierra Region, which have a relationship and incidence due to higher altitude and lower relative population density, with inequalities and inadequate application of equitable public policies and less development of economic activities in these departments, in which, also, the level of human development of the population decreases.

Through the Clustering classification, the highest level of HDI is confirmed by the same characterization in the grouping of cluster 4, made up of the departments of the Coastal Region and considered as metropolitan cities, with greater mining development and lower HDI than cluster 1, which integrates the departments of the Sierra Region.

In both cases of application of the K-Means (cluster 1) and Clustering (cluster 3) algorithms, demonstrate better effectiveness in the separation of the clusters in a broad way, through the grouping of the departments of Loreto, San Martín and Ucayali that have the lowest HDI level and are located in the Jungle Region, characterized by lower population density and less development of economic activities.

To sum up, the study concludes that the K-Means and Clustering techniques require the classification of groups, cohere and optimize the information for decision-making in the departments under study, in order to be used to manage and achieve better levels of DHI.

VII. FUTURE WORK

Future work will be related to the measurement of Quality-of-Life Indices (ICV) of the adult population and human development indicators at a comparative level among Ibero-American countries.

Plan to analyze the Regional Competitiveness Indices and their relationship with economic and social development and compare the indicators by regions and departments in order to know their evolution and determining factors for changes in position.

In addition, project studies on problems, trends and progress in development policies by measuring management indicators according to results and products structural gaps, which guarantee the application, follow-up and monitoring of public policies with equality and equity criteria in all regions and departments of Peru.

REFERENCES

[1] D. L. Pineda-Ospina, E. G. Rodríguez-Guevara, and D. A. García-Bonilla, "Regional clusters as a strategy to overcome competitive disadvantages," Res. Dev. Innov. Mag., vol. 11, no. 1, pp. 49–62, 2020, doi: 10.19053/20278306.v11.n1.2020.11682.

[2] V. Bhagat, Y. Izad, J. Jayaraj, R. Husain, K. Che Mat, and M. Moe Thwe Aung, "Emotional Maturity Among Medical Students and Its Impact on Their Academic Performance," Tost, vol. 4, no. 1, pp. 48–54, 2017, [Online]. Available: http://transectscience.org/.

[3] M. Faizan, M. F. Zuhairi, S. Ismail, and S. Sultan, "Applications of Clustering Techniques in Data Mining : A Comparative Study," vol. 11, no. 12, pp. 146–153, 2020.

[4] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," Atmos. Pollut. Res., vol. 11, no. 1, pp. 40–56, 2020, doi: 10.1016/j.apr.2019.09.009.

[5] M. R. Ayyagari, "Classification of Imbalanced Datasets using One-Class SVM, k-Nearest Neighbors and CART Algorithm," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 1–5, 2020, doi: 10.14569/ijacsa.2020.0111101.

[6] M. Fradi, M. Afif, and M. Machhout, "Deep Learning based Approach for Bone Diagnosis Classification in Ultrasonic Computed Tomographic Images," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 80–87, 2020, doi: 10.14569/IJACSA.2020.0111210.

[7] M. Iqbal, S. Ali, M. Abid, F. Majeed, and A. Ali, "Artificial Neural Network based Emotion Classification and Recognition from Speech," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 434–444, 2020, doi: 10.14569/ijacsa.2020.0111253.

[8] A. K. Dehariya and P. Shukla, "Medical Data Classification using Fuzzy Main Max Neural Network Preceded by Feature Selection through Moth Flame Optimization," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 655–662, 2020, doi: 10.14569/IJACSA.2020.0111276.

[9] M. Faizan, M. F., S. Ismail, and S. Sultan, "Applications of Clustering Techniques in Data Mining: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 12, pp. 146–153, 2020, doi: 10.14569/ijacsa.2020.0111218.

[10] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," vol. 11, no. 11, 2020.

[11] R. Kadry and O. Ismael, "A New Hybrid KNN Classification Approach based on Particle Swarm Optimization," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 291–296, 2020, doi: 10.14569/IJACSA.2020.0111137.

[12] M. Rela, S. N. Rao, and P. R. Reddy, "Liver Tumor Segmentation using Superpixel based Fast Fuzzy C Means Clustering," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 380–387, 2020, doi: 10.14569/IJACSA.2020.0111149.

[13] T. A. Khan, K. A. Kadir, S. Nasim, M. Alam, Z. Shahid, and M. S. Mazliham, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease classification," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 560–569, 2020, doi: 10.14569/IJACSA.2020.0111170.

[14] F. Alharbi, M. K. Elbashir, M. Mohammed, and M. E. Mustafa, "Fine-Tuning Pre-Trained Convolutional Neural Networks for Women Common Cancer Classification using RNA-Seq Gene Expression," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 676–683, 2020, doi: 10.14569/IJACSA.2020.0111182.

[15] T. Sh. Mazen, "A Novel Machine Learning based Model for COVID-19 Prediction," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 523–531, 2020, doi: 10.14569/IJACSA.2020.0111166.

[16] A. D. Dondekar and B. A. Sonkamble, "Harmonic Mean based Classification of Images using Weighted Nearest Neighbor for Tagging," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 11, pp. 240–244, 2020, doi: 10.14569/IJACSA.2020.0111131.

[17] T. A. Sipkens and S. N. Rogak, "Technical note : Using k -means to identify soot aggregates in transmission electron microscopy images," J. Aerosol Sci., no. September, p. 105699, 2020, doi: 10.1016/j.jaerosci.2020.105699.

[18] E. Patel and D. S. Kushwaha, "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," Procedia Comput. Sci., vol. 171, no. 2019, pp. 158–167, 2020, doi: 10.1016/j.procs.2020.04.017.

[19] J. Morales, N. Vargas, M. Coyla, and J. Huanca, "Classification model of municipal management in local governments of Peru based on K-means clustering algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 568–576, 2020, doi: 10.14569/IJACSA.2020.0110770.

[20] W. Yang, H. Long, L. Ma, and H. Sun, "Research on clustering method based on weighted distance density and k-means," Procedia Comput. Sci., vol. 166, pp. 507–511, 2020, doi: 10.1016/j.procs.2020.02.056.

[21] M. Mateen, J. Wen, M. Hassan, and S. Song, "Text clustering using ensemble clustering technique," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 9, pp. 185–190, 2018, doi: 10.14569/ijacsa.2018.090925.

[22] B. M, J. Tomy, U. A, and P. Jacob, "Clustering Student Data to Characterize Performance Patterns," Int. J. Adv. Comput. Sci. Appl., vol. 1, no. 3, pp. 138–140, 2011, doi: 10.14569/specialissue.2011.010322.

[23] E. León Guzmán, "Metrics for Clustering Validation," 2019, [Online]. Available: http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion13_validacion_Clustering.pdf.

[24] T. Tamer, A. Haydar, and I. Ersan, "Data Distribution Aware Classification Algorithm based on K-Means," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 9, 2017, doi: 10.14569/ijacsa.2017.080946.

[25] M. Khalid, N. Pal, and K. Arora, "Clustering of Image Data Using K-Means and Fuzzy K-Means," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 7, pp. 160–163, 2014, doi: 10.14569/ijacsa.2014.050724.