Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 7 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i07.47647

PAPER

Proposed Feature Selection Technique for Pattern Detection in Patients with Pneumonia Records

Jesus Orlando Gil Jauregui[1], Angel Gerardo Carmen Cruzatti[1], Miguel Angel Cano Lengua[2](✉), Hugo Villaverde Medrano[3]

[1]Universidad Nacional Mayor de San Marcos, Lima, Perú

[2]Universidad Tecnológica del Peru, Lima, Perú

[3]Universidad Cayetano Heredia, Lima, Peru

mcanol@unmsm.edu.pe

ABSTRACT

Pneumonia in Peru is a very serious problem. Its impact in recent years has been aggravated due to the Covid-19 pandemic, generating an increase in infections and deaths without distinguishing the age range, which placed this country on the mortality list due to the pandemic. That is why this research seeks the causes of this problem and evaluates what patterns were detected between the years 2019–2022 in patients with pneumonia in Peru from data set from the Comprehensive Health Insurance (SIS). The data presented values related to age, gender, medication and other significant values to understand the disease. The results of the research were achieved by using the PCA technique where the dimensionality of the data was reduced from 28 to 4 main features (Patient's year of health care, Age, BMI, Department). Finally, with this processed data set, the K-Means algorithm was used, where it was determined that patients in the 60 to 85 years range are the most affected by J189 pneumonia. In addition, an environmental pattern was found in J189 pneumonia. J128, resulting in a focus on patients on the Peruvian coast in places like Lima or La Libertad.

KEYWORDS

clustering, data mining, scientific data, patterns, pneumonia

1 INTRODUCTION

Currently, there is a significant growth in the information generated or produced by various sectors, such as health, due to the digitalization of our environment. In this regard, [1] mention that changes are evident in different economic sectors, including the health sector. Due to these changes, it is important to know how to adapt to them, and to perform accurate analysis of all the information generated.

Concerning health in Peru, many problems affect the quality of life of the population. One of the most worrying aspects is the lack of access to quality health services in many rural and urban areas of the country, as mentioned by [2] where, with respect to health workers, "almost 50% (102,777) of them were concentrated in the

Gil Jauregui, J.O., Carmen Cruzatti, A.G., Cano Lengua, M.A., Medrano, H.V. (2024). Proposed Feature Selection Technique for Pattern Detection in Patients with Pneumonia Records. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(7), pp. 69–89. https://doi.org/10.3991/ijoe.v20i07.47647

Article submitted 2023-12-28. Revision uploaded 2024-02-14. Final acceptance 2024-02-14.

city of Lima." In addition, the lack of resources and trained personnel in some health facilities also contribute to the inefficiency of the services offered. On the other hand, unequal access to health care is also a significant problem, as many people living in poverty cannot afford efficient medical treatment for their health.

A relevant point in these health concerns is the lack of investment in research and development in the health sector, which prevents significant advances in the treatment of prevalent diseases in the country, such as diabetes and hypertension. In addition, the covid-19 has led to an increase in the number of pneumonia cases, placing Peru in high mortality rankings. For this reason, Peru's healthcare system is going through a difficult time and faces challenges in terms of accessibility, quality and investment in research that must be addressed as soon as possible.

Pneumonia, according to [3], is "a common respiratory infection and warrants careful consideration of antibiotic initiation and choice." This is divided into three main groups as mentioned by [4]. "Pneumonias are classified into community-acquired pneumonia (CAP) and nosocomial or hospital-acquired pneumonia (HAP)," and this classification comes from the place where the disease was acquired.

Based on this context, it is necessary to be able to find the most significant features associated with pneumonia and its types. A constant problem when choosing features is to properly select the algorithm, which is why [5] highlights the need to improve prognostic prediction of patients with pneumonia by using advanced artificial intelligence (AI) methods instead of the Support Vector Machines (SVM) algorithm that has been used in previous research. In addition, the authors suggest the importance of improving the selection of more significant features from the presented data set to improve prediction accuracy.

Similarly, [6] indicate that it is important to identify and use new variables, relevant features or patterns related to COVID-19 pneumonia in order to improve and facilitate classification models (such as those implemented by the authors) for the diagnosis and identification of this disease. And, thus contribute to the competent authorities in their decision-making process related to health policies. On the other hand, [7] indicate that the provision of more relevant features to the diagnosis of childhood pneumonia should be taken into account in the data set. Knowing that children also have other types of respiratory diseases, even with similar symptoms, the learning algorithms must know how to differentiate pneumonia from these diseases. For this, the identification and disposition of several features with sufficient relevance associated with the disease is important.

The growth of information needs to be controlled in order to treat it efficiently and to achieve an improvement in people's health. This is why the authors [8] state, "Use big data to perform proactive measures and ensure timely stakeholder engagement to prevent disease occurrence in the case of noncommunicable diseases. Ultimately, that will reduce the cost of management and the burden of chronic disease." The improvements are both at the economic level and in the quality of life so that the countries that apply this knowledge achieve efficiency in these fields.

Big data analysis is commonly used through data mining, which provides multiple contributions in various areas of knowledge, the health sector being one of those that benefited the most by this type of process. An effective application of data mining in health could make health decisions of any entity mor effective, which in turn would allow savings in costs and time spent in the search for better proposals. This is evidenced by the application of artificial intelligence algorithm to detect respiratory diseases in patients [9].

The application of an unsupervised machine learning model fits well with the objective of the present research as it has been previously used in healthcare for various purposes. [10], in their research, applied unsupervised machine learning to databases of electronic medical records for the discovery of latent diseases and to separate patients based on subgroups with common diseases. The authors highlight the usefulness of unsupervised learning for the identification of disease patterns and the clustering of patients with common diseases using a number of significant features.

Within unsupervised learning, we find the k-means clustering algorithm also often has various applications in the health sector. In research done by [11], the k-means algorithm is used to cluster a group of university students based on their mental health. The data were obtained from an online survey, which solicited information about students' mental and emotional health. In addition, the authors point out that this algorithm can help in the identification of patterns in health, in this case, mental health.

Finally, our research has the main objective of implementing the k-means algorithm using data mining to obtain patterns in patients with pneumonia in Peru, through a data set of people diagnosed with some type of pneumonia during the period 2020–2022. This data set comes directly from affiliated patients of the SIS (Comprehensive Health System) and different fields such as age, sex, glucose levels, among other values relevant to the research will be taken into account.

2 LITERATURE REVIEW

A total of 22 scientific articles have been selected, which were extracted from different journals and conferences indexed in repositories such as IEEE, Web of Science, Scopus, Sciencedirect, Springer. And we included some articles from non-indexed journals, but they were selected for their contribution to our research.

2.1 Algorithms for pattern detection in pneumonia

Research [12] and [13] relied on correlation methods and decision tree algorithm respectively to identify the main variables to be entered into the algorithms and thus were used when detecting the patterns needed in their research.

If we look at issues related to the origin of the data set, we have authors like (Lai et al., 2019); [5]; [15]; [13]; [16] who used as data source public data sets like NHRID. On the other hand, we have the Lazio Hospital Information System [17], which got its data from private centers like the Marshfield Clinical Health System. In summary, we have a majority use of public data for health research.

About the algorithms used [12]; [14]; [5]; [15] and [13] the most used were classification algorithms such as the SVM, Multi Layer Perceptron, Bayesian Boosting algorithm. Furthermore, in the research [14] and [5], the AUC was the main metric to identify the classification capability of the algorithm created. In the first case, it had a value in the range of 0,7518–0,7601 and the other had an AUC for each algorithm created of 0,7727 and 0,7758 to make way for future research in the same context to improve that figure.

2.2 Patterns or risk factors associated with pneumonia

The detection of patterns or risk factors related to pneumonia has been the objective of several investigations with application cases in different countries and varied populations. In recent research and because of the pandemic context, the search for risk factors for pneumonia has been closely linked to COVID-19. For example, [18] investigated to identify independent risk factors for COVID-19 pneumonia in Mexican adults who have already been vaccinated. On the other hand, [19] have a similar objective in their research, which is to analyze patterns and risk factors associated with the development of COVID-19 pneumonia. But in this case, they apply the analysis to data on children and adolescents in Mexico.

There are also different cases and application scenarios external to COVID-19, focused on the objective of identifying patterns or risk factors related to pneumonia. In relation to childhood pneumonia, there is research that seeks to identify risk factors associated with pneumonia among children who were admitted to hospitals or other medical centers because of this disease. For example, [20] conducted an investigation aimed at analyzing the prevalence and patterns of pneumonia among children admitted to the University of Port Harcourt Teaching Hospital in Nigeria. Similar research was conducted by [21], who sought to identify the factors of pneumonia among children who visited Adama Hospital Medical College in Ethiopia. On the other hand, there is the research of [22], which sought to determine the incidence and risk factors for community-acquired severe pneumonia in children hospitalized at the C.R. Gardi Hospital in India. In the investigation of [23], which deals with progressive community-acquired pneumonia, they sought to identify the associated risk factors in children hospitalized in various medical centers in Taiwan.

In addition to cases of children hospitalized for pneumonia, there are also other scenarios and application cases for identifying patterns or risk factors associated with pneumonia. Among them, there is the research conducted by [24], in which they sought to identify risk factors associated with the high incidence of childhood pneumonia in a rural area of the Bojonegoro region in Indonesia. Another case is that of [25], who investigate to estimate the sensitivity of pneumonia diagnosis and to investigate its determinants or patterns in common among Malawian children.

To search for and identify patterns or risk factors associated with pneumonia, different data sets or data samples have been used in research related to this disease, which allowed obtaining relevant information and contributing to knowledge in this field. In relation to the above investigations, there are differences in the features of their data sets or sample for their respective investigations. For example, for COVID-19 pneumonia, [18] used 1607 records of adults over 20 years of age with confirmed disease and onset between March and July 2021. Such a number contrasts with that used by [19], who use 215,656 records of patients under 18 years of age diagnosed with COVID-19, updated to May 23, 2020. These were extracted from the database of the General Directorate of Epidemiology of the Ministry of Health of Mexico.

Excluding the cases of COVID-19 pneumonia, there is the research of [26], who also used a much higher amount than [18], but a lower amount than [19], being almost half of the latter. They used a total of 115,036 hospital admission data records from 2012, 2013 and 2014 in the Netherlands, which were obtained from the Dutch Hospital Data (DHD) data source. On the other hand, there is the research of [27], who used a much smaller number of records than the 3 mentioned, with a total of 667 records of children between 2 and 16 years of age, and who have been diagnosed with pneumonia. Medical records extracted from the University Malaya Medical Center between 1 January 2012 and 31 December 2014 were used for this research.

Another research with an even smaller amount of data than [27] has been carried out by [20], which analyzed 286 physical medical records of children aged 1 to 17 years admitted to the pediatric emergency room within the period from January 2017 to December 2018. The authors indicate that the original number of children admitted in this period was 2169 (1089 males and 1080 females), of whom 312 were admitted for pneumonia. However, 26 of the records could not be recovered, leaving the 286 already mentioned.

Considering questionnaires and interviews as data collection techniques, on the one hand, there is the research done by [25], where the sample data used were obtained from the Service Provision Assessment (SPA) in the period 2013–2014, which consisted of a census that was carried out by the Malawi Demographic and Health Survey Program. The survey included an audit of facility resources, surveys of clinical practices, and observations of clinical care for children under the age of 5. In total, 3136 clinical visits were collected from children between 2–59 months of age. On the other hand, there is the research of [22], who, in this case, detail how they carried out the data collection process in the period July 2015 and June 2016. They indicate that the chosen participants were children from 2 to 59 months of age, admitted to pediatrics or the pediatric intensive care unit of C.R. Gardi Hospital. It has a calculated sample size of 151 children with severe pneumonia. Regarding the data collection process, the authors mention that the oxygen saturation level (SaO2) was taken on admission to the hospital, using a pulse oximeter. Interviews were conducted with the mother or caregiver of the children who met the inclusion criteria, which consisted of answering questions from a questionnaire related to the signs and symptoms of pneumonia on admission.

On the other hand, there are also investigations or studies of the type case control, where they divide the studied population into two groups: case and control. For example, [24] halved their population of 176 children into 88 children for each group. The group of cases was children between 10–59 months of age who were diagnosed with pneumonia, while the control group was children of the same age who were not diagnosed with pneumonia. Their analysis was made in the period from January 1, 2018, to January 31, 2019. With a similar population and control groups, there is the research of [21], with 248 children between 2–59 months of age who visited the outpatient department of Adama Hospital Medical College in Ethiopia, between January 1 and March 15, 2021. The group of cases were also children diagnosed with pneumonia, and the control, those who were not diagnosed with pneumonia in that period. In this case, to obtain the data, a structured questionnaire was designed that included questions associated with independent variables (sociodemographic variables), environmental factors, nutrition and vaccination status and disease history.

Based on the above-mentioned investigations, different risk factors associated with the revised investigations could be identified. That may be present in the general public, but in some cases, may be more harmful to children. Summarizing some of the most common ones we have:

Common symptoms include the following:

- Smoking.
- Diabetes.
- Obesity.
- Malnutrition.
- Diarrhea.

Considering previous and current health conditions, we have:

- Previous upper respiratory diseases.
- Incomplete immunization in the case of COVID-19 pneumonia.
- Immunosuppression.
- Low level of hemoglobin and white blood cells.
- Vitamin A deficiency.

Other factors to consider include:

- High incidence of pneumonia in regions of high population density.
- Vulnerability to disease for older adults, such as complications of COVID–19.

In addition, in cases of childhood pneumonia, there were some characteristic factors such as:

- Low educational level of parents.
- High percentage of children raised in families in extreme poverty.
- Large family.
- Non-exclusive breastfeeding.
- Low birth weight.
- Premature birth.

3 METHOD

The CRISP-DM methodology was used in the development of the research, which allows for an order and an adequate procedure to carry out the respective analysis of the data and to obtain the expected results for the present work.

3.1 Business understanding

Pneumonia is a health problem that negatively affects the general population in our country. And that is why a data mining model was developed for pattern detection in patients with pneumonia in Peru. In this context, it was considered to use medical records of patients affiliated to the Comprehensive Health System (SIS) who have suffered from some type of pneumonia and have been treated in different health institutions in our country.

3.2 Data understanding

The data set used in this research was obtained by the SIS (Comprehensive Health System) through a request made on its transparency portal. In total, 591317 records of patients diagnosed with pneumonia were obtained between 2019 and 2022. In total, 37 fields were provided, including age, department, gender, medications, etc. Furthermore, these data are based on the features of the Single Care Form prepared by the Ministry of Health.

Data structure. The data set awarded by the SIS is an Excel document with different tables and fields related to patients with pneumonia and with a cut-off date of April 26, 2023. This Excel document is divided into 5 tables containing data for the years 2019–2022: 1 table of health care, and 4 tables (1 for each year) related to certain medications in that period.

In Table 1 you can see the variables with their respective descriptions related to the patient, similarly in Table 2 in relation to the prescribed medication.

Table 1. Description of collected table of health care

Variable	Description	Variable	Description
DIAGNOSTIC CODE	ICD-10 code of the type of pneumonia diagnosed	RENIPRESS NAME	IPRESS Name where the health care was performed
DIAGNOSTIC NAME	Type of pneumonia diagnosed	NIVELEESS	Health facility level
TYPE OF DIAGNOSIS	It can be presumptive, definitive or repeated	CATEGORY	Sub-level to which a medical establishment belongs
DATE OF BIRTH	Date of birth of the patient	TYPE_CARE	Type of health care provided to the patient
SEX	Sex of the patient	DESTINY_CARE	It is the patient’s final situation after health care
SERVICE	Health service code	BP	Blood pressure in mmHg
SERVICEDESC	Description of the health service	HEIGHT	Height in centimeters (cm)
AGE	Age of patient	WEIGHT	Weight in kilograms (kg)
DEPARTMENT	Department where health care was registered	BMI	Body mass index
PROVINCE	Province where health care was registered	PNEUMOCOCCUS	Indicates whether the patient is immunized against pneumococcus
DISTRICT	District where health care was registered	INFLUENZA	Indicates whether the patient is immunized against influenza
RENIPRESS COD	IPRESS code where the health care was performed		

Table 2. Description of collected table of medications

Variable	Description
MEDICATION CODE	Medication code prescribed to the patient
MEDICATION NAME	Description of the medication prescribed to the patient
QUANTITY	Number of medications prescribed to the patient

3.3 Model design

The proposed model for detecting patterns in pneumonia begins with data entry, which represents the described records of patients diagnosed with pneumonia in Peru. It then continues with the data preparation stage, intending to condition the data for the implementation of the K-means algorithm. Finally, there is the results phase, where the clusters produced by the algorithm are shown. And how the associated pneumonia patterns are represented, as can be seen in Figure 1.

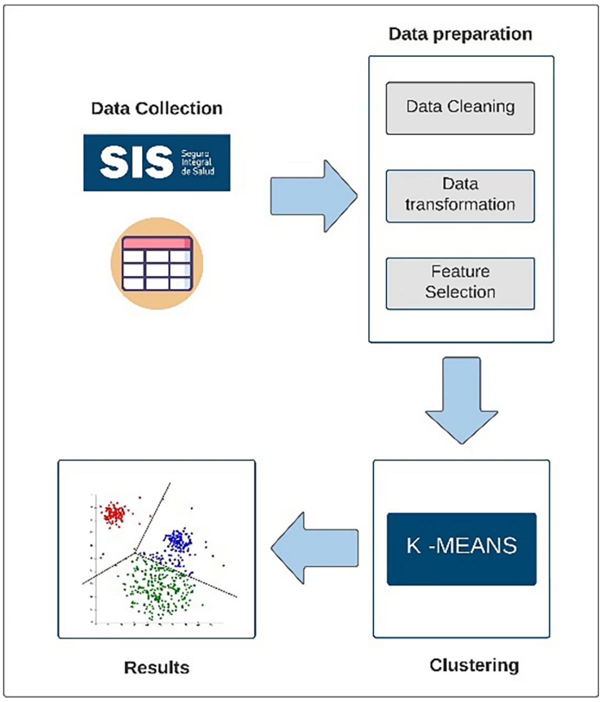

Fig. 1. Model for pattern detection in pneumonia

Data preparation. In this phase, there is a set of activities aimed at constructing final data. It is acceptable that the task of data preparation or imputation can be performed more than once.

a) **Data selection.** Data associated with 'id' such as 'MEDICATIONID' (column n), 'HEALTHCAREID' (column n), etc. will be removed from the data set to be processed because they have no greater relevance to the objective of the project. We will only keep the 'ATENDEDID' (column n) for control of the number of health care per person.

b) **Data cleansing.** To deal with missing data, the elimination of values was used, since [28] states that if the data sample is very large and the MCAR (Missing Completely at Random) criterion is met, then elimination by list/row is the most recommended solution.

c) **Data transformation.** The numerical transformation (Gutiérrez et al., 2016) provides a conversion of the categories by integers without repetitions, making it the most indicated over One – Hot because the latter only provides 0 or 1 as numbers to be replaced per category.

   Because our research uses k-means, which is a machine learning algorithm that uses numerical variables, a normalization of the categorical data is applied. In our dataset as we have 19 fields of this type, the conversion to numerical data was applied. In addition, such conversion was useful for performing main feature analysis, since the data had to be numerical.

   Furthermore, the principal component analysis (PCA) technique was used. We chose this methodology based on the research of [29] which indicates that the PCA technique is the best when selecting the main features in Big Data, with better results than SVM, Random Forest and Naive Bayes since it retains a large proportion of fields unlike others.

d) **Feature selection.** A key point to consider for the objectives of this study is the selection of main features in patients with pneumonia. For this purpose, the Principal Component Analysis (PCA) technique was used.

We used this methodology based on research from [29], which indicates that the PCA technique is the best when selecting the main features in Big Data, with better results than SVM, Random Forest and Naive Bayes because it retains a large proportion of fields unlike these. This technique defines the principal components as linear combinations of the original variables, and they are ordered according to the amount of variance they capture.

Below are the steps commonly used to determine the principal components.

- Standardize.
  Equation (1) can be used to subtract the variables with their mean and then divide it by their standard deviation.

$$\frac{X_{j,i} - \overline{X_j}}{sd(X_j)},\ i = 1,\ldots,n,\ j = 1,\ldots,p \tag{1}$$

- Calculate principal components
  Equation (2) shows that the principal components to be found are defined by multiplying the variables by their weight (variance), *m* indicates the number of components.

$$Y_j = \alpha_{1j}X_1 + \alpha_{2j}X_2 + \cdots + \alpha_{pj}X_p,\ j = 1,\ldots,m \tag{2}$$

The coefficient vectors in the principal components are one.

$$\alpha^{t}_{j_j}\alpha_1 = 1 \tag{3}$$

The *n* principal components found were established in the function to be maximized with the weights of each variable in our data set on patients with pneumonia.

Function to be maximized:

$$\sum_{i=1}^{p}\alpha_{i1}^{2} = 1 \tag{4}$$

- Lagrange multiplier
  Since maximizing the summation is sought, the Lagrange is applied with the following equation (5).

$$F(\alpha_1,\lambda) = \alpha_1^{t}\sum\alpha_1 - \lambda(\alpha_1^{t}\alpha_1 - 1) \tag{5}$$

Once the weights per variable were located to maximize the function, the main variables were determined with which the patterns per patient were determined using K-means in the next step.

Modelling. For the implementation of the pattern detection model for pneumonia in Peru, it was decided to use the K-means clustering algorithm because of its ability to efficiently segment the data into groups based on the minimization of variance within each group.

a) **K-means algorithm.** For this research, the data of patients diagnosed with pneumonia and their respective features or attributes (after having gone through the preparation phase) were analyzed. These allowed us to calculate the Euclidean distance to the centroids of the initially defined clusters, so that each patient or individual is assigned the closest cluster. Complementing the above, [30] and [31] refer to the steps to be followed when applying the k-means algorithm to a data set. They can be visualized in the following flowchart (see Figure 2):

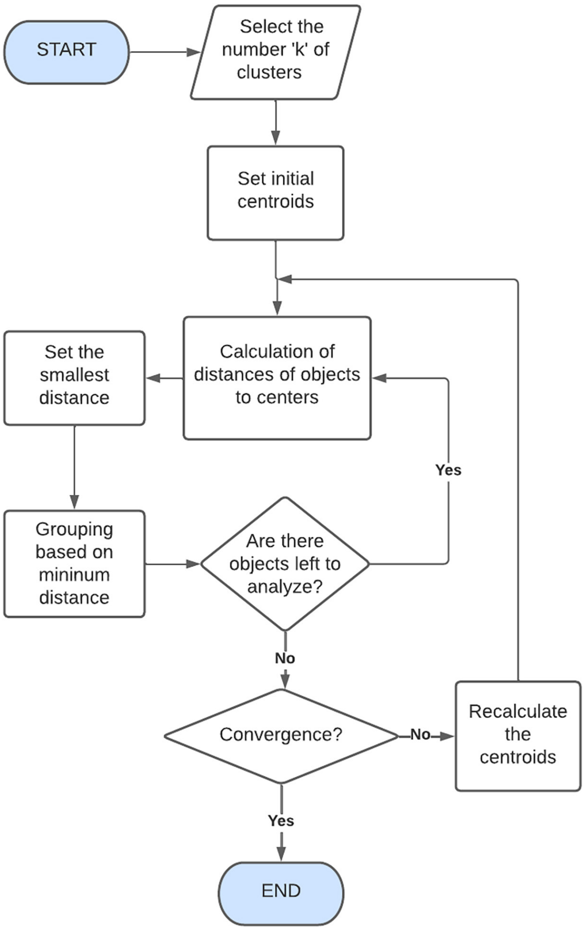
START
Select the number 'k' of clusters
Set initial centroids
Calculation of distances of objects to centers
Set the smallest distance
Grouping based on mininum distance
Are there objects left to analyze?
Yes
No
Convergence?
No
Recalculate the centroids
Yes
END

Fig. 2. K-means algorithm flowchart

Focusing on the data of patients diagnosed with pneumonia, we have the following:

- Given the data set of size *'N'* represented by Patients *'P'* = {*P*1, *P*2, *P*3, …., P_n}. It contains a number *'M'* of fields or features *'X'* = {*X*1, *X*2, *X*3, …., X_m}. Where *'M'* is the number of fields that were kept after the data preparation and cleansing phase. In addition, the features X_i, refer to described in the above. Such as: Diagnosis, date of birth, sex, age, department, blood pressure, height, weight etc.

  After defining the data set, the number *'k'* of clusters is established, and the patients P_i to be chosen as the initial centroids *'C'* = {*C*1, *C*2, … C_k} for the algorithm are selected from the data set.
- Subsequently, the Euclidean distance between the patients P_i to the centroids C_i is calculated, taking the fields or features X_i as reference points. This calculation can be represented by equation (6) described by [32]:

$$d(Z_p, C_j) = \sqrt{\sum_{k=1}^{N_d} (Z_{pk} - C_{jk})^2} \tag{6}$$

Where:

- N_d refers to the number of remaining fields or columns in the final patient data set.
- Z_p refers to the features of patients P_i.
- C_j refers to the features of patients defined as centroids C_i.

– After calculating the distance of each patient P_i to the centroids C_i corresponding to each cluster, the cluster with which the least respective distance was obtained is assigned.

– Subsequently, the algorithm completion criterion is evaluated. For this criterion, the value of Inertia is commonly used, seeking to minimize it in each iteration until reaching some set value as convergence. As another alternative to the completion criterion, a fixed number of iterations performed by the algorithm can be set.

[33] represents the calculation of inertia by equation (7).

$$SS_{Wi} = \sum_t \left\| x_t - \mu_i \right\|^2 \quad \forall\ i \in (1, K) \tag{7}$$

- t refers to the number of patients P_i.
- x_t refers to patient fields.
- μ_i refers to defined patient fields centroids C_i.

– Finally, the centroids of the initially defined clusters are reassigned. Then, steps b), c) and d) are repeated until the completion criterion of the established algorithm is met.

For centroid reassignment, for each cluster the average of the values of the features X_i of patients it contains must be calculated. [32] represents this by equation (8):

$$m_j = \left(\left(\Sigma_{\forall Z_p inc_j} Z_p\right) / n_j\right) \tag{8}$$

Where:

- Z_p refers to the fields or features of patients P_i grouped in the cluster C_i.
- n_j is the number of patients for cluster C_i.

b) Techniques for choosing the number “k”. For the implementation of the k-means algorithm in the proposed model, it will be necessary to take into account the existing techniques for choosing the optimal number of clusters. On this subject, [34] points out that it is important to set an appropriate value for the number of clusters k, and that an incorrect choice of this value can cause the algorithm to give poor results. Since some clusters could be merged or split so as to couple to the set amount.

[33] mentions commonly used techniques for choosing the number of clusters. Such as:

– *Optimization of inertia (elbow method*). This method evaluates the Inertia value from the minimum and maximum number of possible clusters, and thus obtains the optimal value according to the input data. This technique is evaluated graphically, and the number of clusters where the inertia value ceases to vary significantly will be chosen.

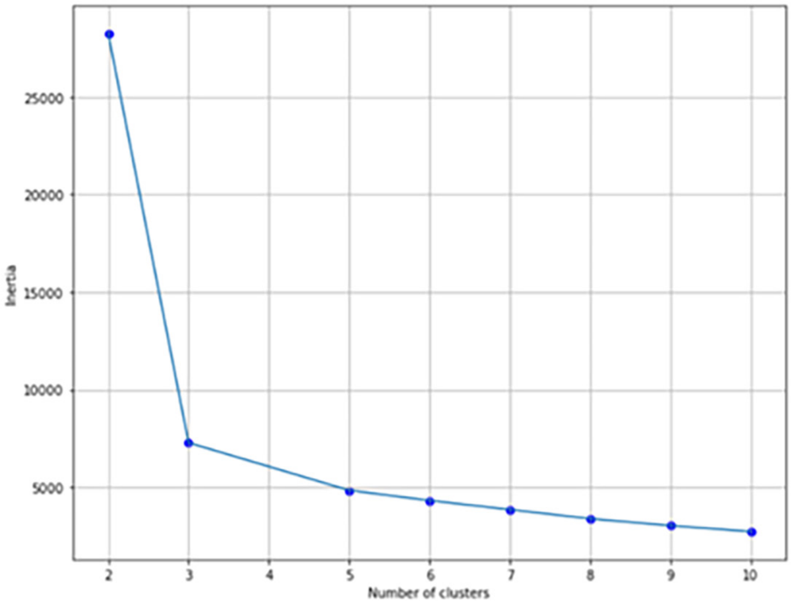

Fig. 3. Elbow method. Adapted from machine learning algorithms (p. 189), by G. Bonaccorso, Packt Publishing

Figure 3 shows how the inertia value no longer varies significantly when the number of clusters is 5. So, in this specific case, it would be an optimal amount to initialize the algorithm.

– *Silhouette coefficient.* According to the author, the coefficient or score of the silhouette is based on the principle: "maximum internal cohesion and maximum group separation." What is sought is that the grouping of the data set produces subdivisions that are correctly separated from each other. That is, the distance between two patients in the same cluster should be less than the distance between two patients in different clusters. The author defines the silhouette coefficient calculation using equation (9) (taking into account a patient X_i from the data set):

$$S(\overline{X_{\iota}}) = \frac{b(\overline{X_{\iota}}) - a(\overline{X_{\iota}})}{max\{a(\overline{X_{\iota}}), b(\overline{X_{\iota}})\}} \tag{9}$$

The coefficient value varies in the range of [–1, 1] , where:

- $a(\underline{x}_i)$ is the mean distance between patient X_i and other patients in the same cluster.
- $b(\underline{x}_i)$ is the mean distance between patient X_i and patients in the nearest cluster.

In addition, the values are interpreted as follows:

- If the value is close to 1, it means that the coefficient reaches its best possible value. And $b(\underline{x}_i) > a(\underline{x}_i)$

- If the value is close to 0, it means that the difference between the distances from the patient to the patients in the same cluster, and from the patient to the patients in the nearest cluster, is almost zero. That is, there is cluster overlap.
- If the value is close to –1, it means that the patient was assigned the wrong cluster, therefore, another 'k' number of clusters must be established. Since $a(\underline{x}_i) > b(\underline{x}_i)$

Model evaluation. This task is based on the interpretation according to each one, our criteria and knowledge determined our mastery over the use of the model and being able to explain the results.

For the implementation of the pattern detection model for pneumonia in Peru, it was decided to use the K-means clustering algorithm. This allowed grouping data from patients with pneumonia to obtain common patterns.

3.4 Deployment

The model deployment is applied based on data set collected for Peruvian patients diagnosed with some type of pneumonia obtained from SIS. In this way, and using the clusters detected by the model, the most common patterns associated with pneumonia were obtained in this real data set. In Figure 4, you can see the development of the process.

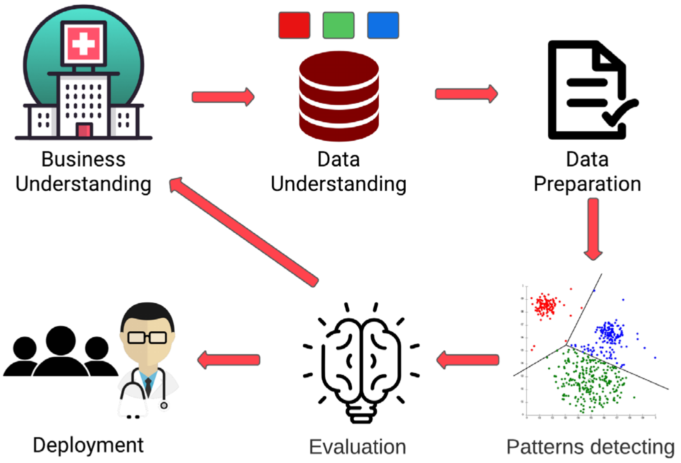

Fig. 4. CRISP – DM Process

4 RESULTS AND DISCUSSION

4.1 Results

In the data preprocessing phase, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data set from 28 original fields to 4, which are finally the main features (see Table 3). This significant reduction in the complexity of the data set was achieved by applying the "elbow rule" technique in PCA, while retaining most of the data variance, simplicity and interpretability of the model. With respect to the records, 11184 rows were finally obtained for the data set finally processed (see Table 4).

Table 3. Preprocessing techniques

Technique	Description	Affected Variable(s)
MCAR	This technique eliminates null values due to the fact that no pattern is found in the data with no value	BP, height, weight, BMI
Numerical Transformation	Categorical variables are converted to numerical values for subsequent use of PCA	birth_date, sex, service, servicedesc, department, province, district, category, type_care, destiny_care, pneumococcus, influenza
PCA	Selection of main features	Final variables: • year_care • age • bmi • department

Table 4. Number of records in stages

Stage	# Records
Initial	591317
MCAR	20887
Outliers & PCA	11184

Subsequently, we advanced to the application phase of the K-Means clustering algorithm. In this stage, K-Means was configured with a k value of 13 clusters for data clustering. This is an optimum number that arose after an exhaustive analysis of the inertia value, applying the techniques: elbow method and silhouette coefficient. Values were tested between a range of 1 to 20 clusters, where the number k = 13 obtained by the elbow method was chosen over the number k = 3 obtained by the silhouette method, since the inertia value is lower.

First, an analysis of the types of diagnosis (diagnosis code) was performed, based on the number of clusters, where it can be evidenced that Pneumonia, unspecified organism (J189), was the most common diagnosis within our data set used (see Figure 5), distributed evenly across the rest of the clusters.

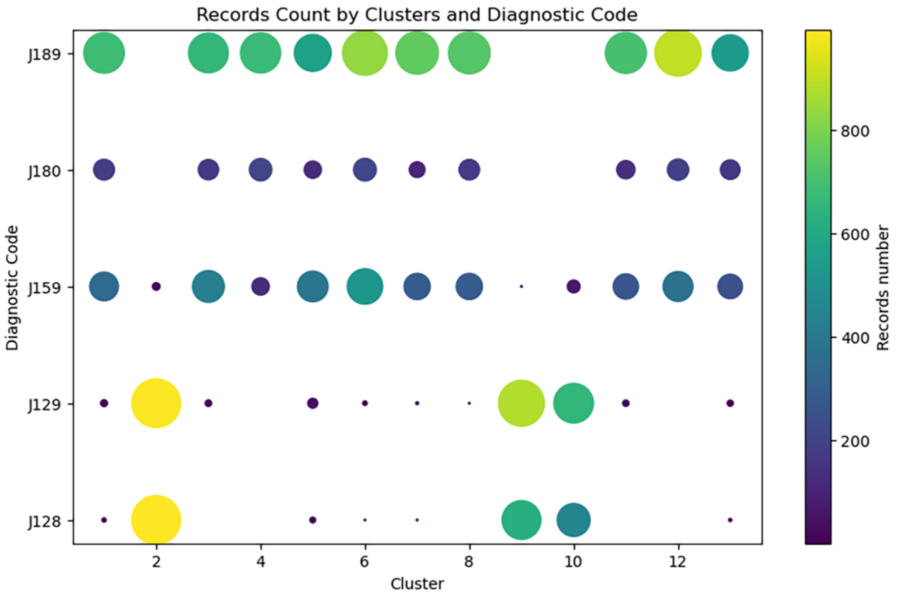

Fig. 5. Cluster vs Diagnostic code

Regarding the analysis of patients diagnosed taking into account the year of health care, cluster plots were evaluated for each of the five diagnostic types mentioned above. For the types: Bacterial pneumonia (J159) and bronchopneumonia (J180), no notable differences were found between the year of health care.

Regarding the type of pneumonia, unspecified organism (J189). It can be evidenced that the largest number of cases occurred in the years 2019–2021 (see Figure 6).

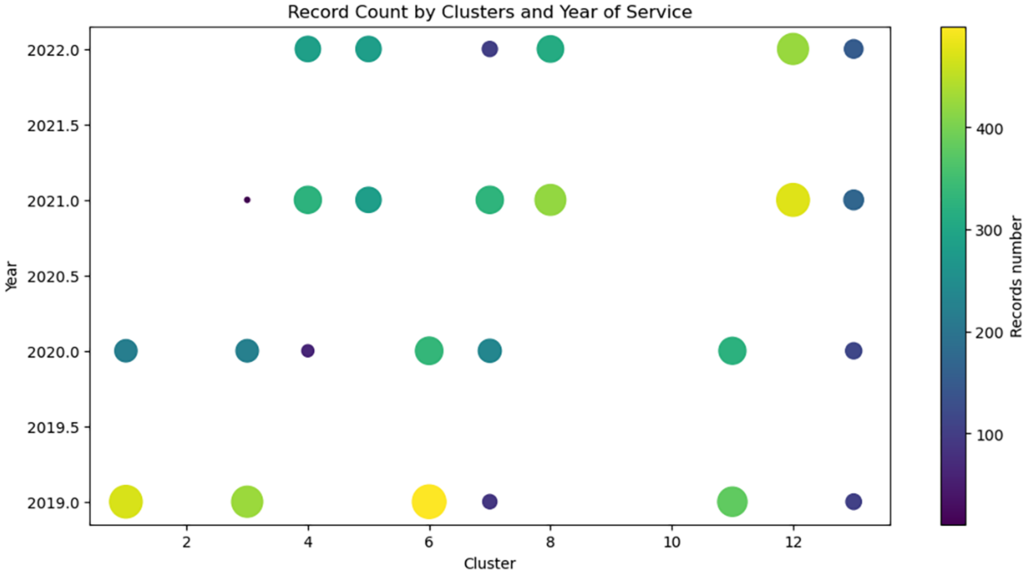

Fig. 6. Number of clusters vs Year of care (Pneumonia type J189)

On the other hand, for the types Viral pneumonia, unspecified (J129) and Other viral pneumonia (J128). It was evident that the highest number of cases occurred in 2021 (see Figures 7 and 8), which was the second year of the coronavirus outbreak in Peru.

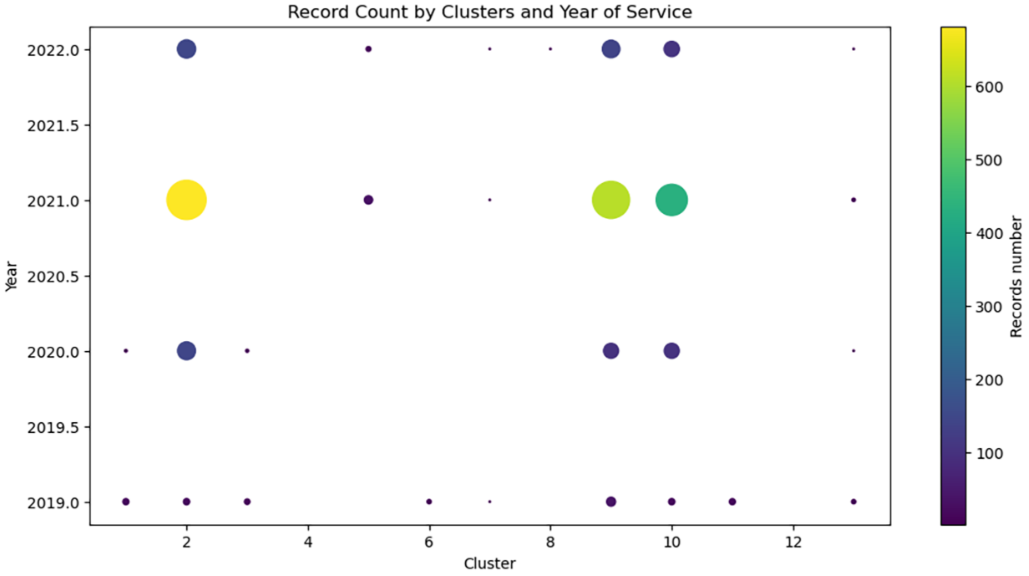

Fig. 7. Number of clusters vs Year of health care (Pneumonia type J129)

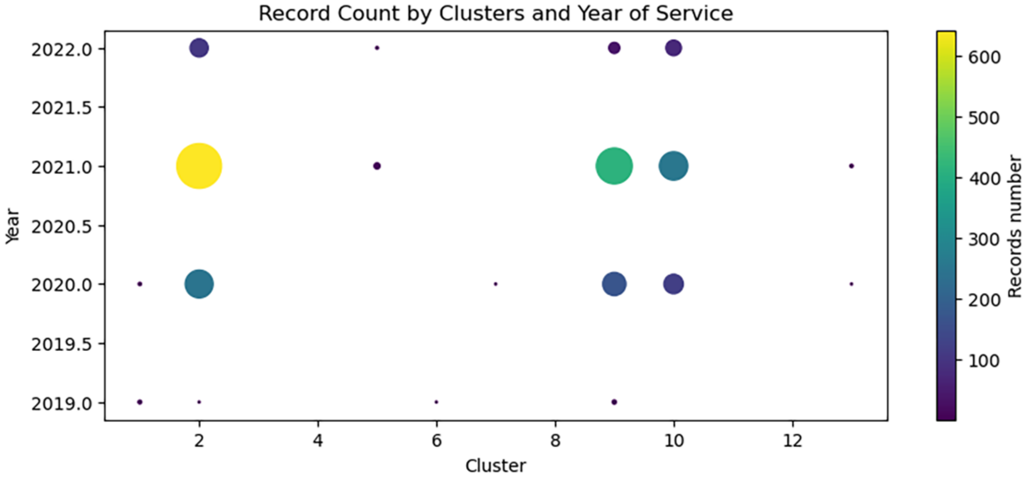

Fig. 8. Number of clusters vs Year of health care (J128)

The analysis of J189 pneumonia, which has showed the highest number of cases, reveals a predominance in elderly patients, specifically between 60 and 85 years of age (see Figure 9). This trend in age is similar to the findings of the research carried out by [13], which indicated that the patients most vulnerable to pneumonia were between 74 and 88 years old. Therefore, it is concluded that pneumonia largely affects elderly people.

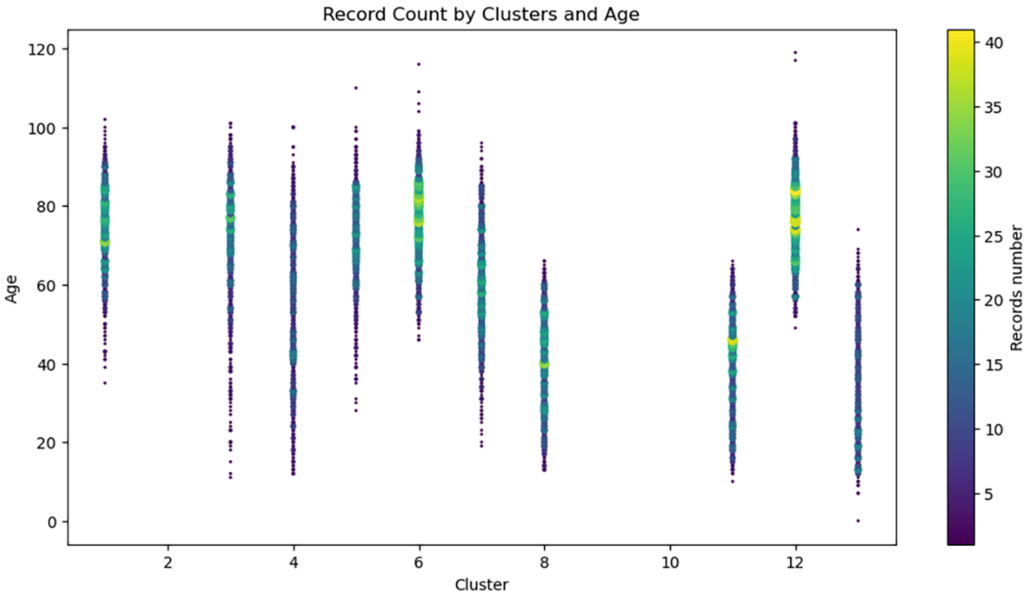

Fig. 9. Number of clusters vs Age of patients (Pneumonia type J189)

In the analysis of patients with J129 pneumonia (see Figure 10), it was observed that a BMI range between 22 and 30 was associated with a high degree of obesity, suggesting possible distinctive health characteristics and risks in this specific group. This agrees with what was identified by [19] where obesity is the predominant comorbidity in more than 82% of their patient sample. These findings highlight the significant presence of obesity as a key condition in patients with pneumonia.

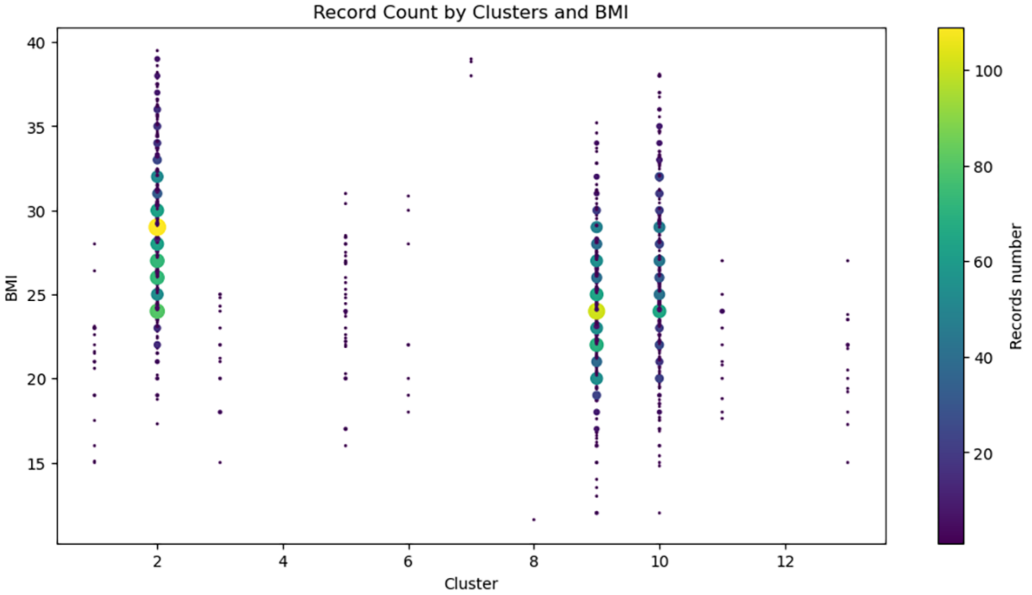

Fig. 10. Number of clusters vs BMI of patients (Pneumonia type J129)

In the analysis of patients diagnosed with J128 pneumonia (see Figure 11), a high number of patients diagnosed in the departments of Metropolitan Lima, La Libertad and Piura, which are regions with a high population density in our country, can be seen.

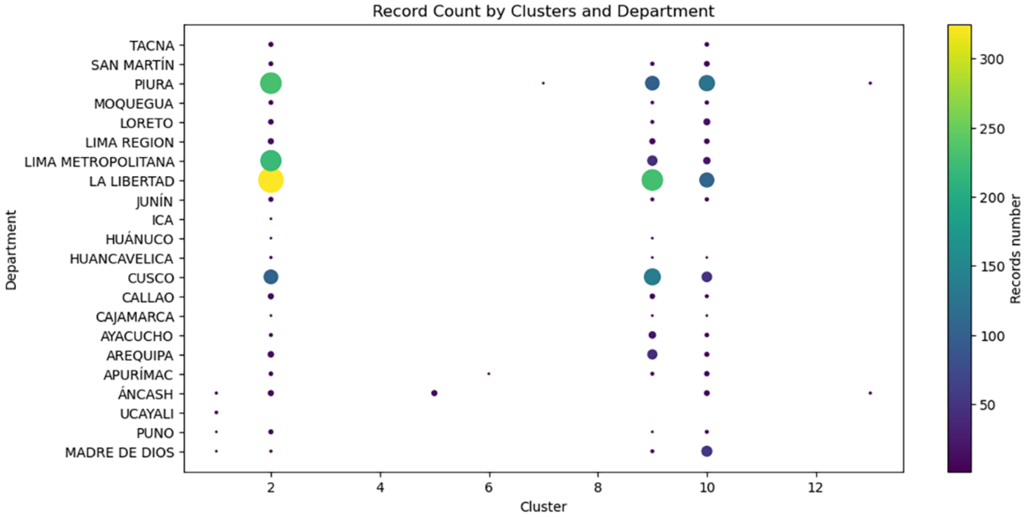

Fig. 11. Number of clusters vs. Department of patient residence. (Pneumonia type J128)

4.2 Discussion

It can be seen that J189 pneumonia predominated in elderly patients (60–85 years of age) is important from a clinical perspective. This could indicate that certain types of pneumonia are more associated with specific age groups, which could influence prevention and treatment strategies. This trend in age resembles the findings of research conducted by [13], which indicated that patients with the highest vulnerability to pneumonia were aged 74–88 years.

On the other hand, the research findings differ from what was obtained by [19] and [20] for childhood patients. Since, in this age range, the highest prevalence of the disease was concentrated between 1–4 years, being more frequent for children under 1 year. This is contradicted by the results (see Figure 9), since no significant differences were observed between these ages.

Therefore, it can be mentioned based on the data set used, that pneumonia largely affects elderly people.

Furthermore, the importance of body mass index (BMI) in patient segmentation is a key observation. The prevalence of BMI in the 22–30 range in patients with J129 pneumonia suggests that this group of patients may have distinctive features in terms of health and risk factors.

This is consistent with what was identified by [19] where obesity is the predominant comorbidity in more than 82% of its patient sample. These findings highlight the significant presence of obesity as a key condition in patients with pneumonia.

Pneumonia type J128 shows a predominance in the coast, especially in the departments of Metropolitan Lima and La Libertad. This finding suggests that there are possible geographic, environmental or exposure factors that may influence the distribution of the disease. It is important to explore these factors further in future research to better understand the underlying causes of this variability.

As mentioned, the trend of detected cases in relation to the department, is similar to that indicated by [26], who indicate that regions of high incidence of pneumonia are characterized by a high population density. This can be seen in the high number of patients diagnosed in the departments of Metropolitan Lima, La Libertad and Piura, which are regions with a high population density in our country.

There are other works related to the prediction of different diseases which use other algorithms related to machine learning [35–37].

5 CONCLUSION

The application of the data mining process to this research contributed as a framework for the treatment of data before, during and after the recognition of Patterns by applying K-means. Therefore, the utility of data mining and CRISP-DM methodology in Big Data analysis related to the health field can be highlighted. The use of the K-Means algorithm made it possible to use various measures to evaluate how clustering algorithms work. Through the implemented visual representations, it was possible to observe how the data of patients with pneumonia behave, which allows informed decisions to develop preventive programs by health managers in the future.

This research highlights the relevance of models that detect behavioral patterns in patients with pneumonia, providing a valuable contribution to guide the care of cases of this disease in Peru. A relevant future consideration involves the incorporation of geographical and environmental data due to Peru's territorial diversity and high levels of pollution in the capital, according to multiple studies. In addition, the analysis of socioeconomic level, an aspect not addressed in this research, represents a relevant field that we suggest exploring in future studies complementary to ours.

6 REFERENCES

[1] E. Oviedo and A. Fernández, "Tecnologías de la información y la comunicación en el sector salud: Oportunidades y desafíos para reducir inequidades en América Latina y el Caribe, 165," *Serie Políticas Sociales,* 2010. http://hdl.handle.net/11362/6169

[2] F. Inga-Berrospi and C. Arosquipa Rodríguez, "Avances en el desarrollo de los recursos humanos en salud en el Perú y su importancia en la calidad de atención," *Rev Peru Med Exp Salud Publica*, vol. 36, no. 2, pp. 312–318, 2019. https://doi.org/10.17843/rpmesp.2019.362.4493

[3] S. N. Grief and J. K. Loza, "Guidelines for the evaluation and treatment of pneumonia," *Primary Care – Clinics in Office Practice*, W. B. Saunders, vol. 45, no. 3. pp. 485–503, 2018. https://doi.org/10.1016/j.pop.2018.04.001

[4] A. A. Vásquez Gaibor, S. C. Reinoso Tapia, M. N. Lliguichuzca Calle, and J. V. Cedeño Caballero, "Neumonía asociada a ventilación mecánica," *Recimundo*, vol. 3, no. 3, pp. 1118–1139, 2019. https://doi.org/10.26820/recimundo/3.(3).septiembre.2019.1118-1139

[5] J. C. Hsu, F. H. Wu, H. H. Lin, D. J. Lee, Y. F. Chen, and C. S. Lin, "AI models for predicting readmission of pneumonia patients within 30 days after discharge," *Electronics (Switzerland)*, vol. 11, no. 5, p. 673, 2022. https://doi.org/10.3390/electronics11050673

[6] I. Quesado, J. Duarte, A. Silva, M. Manuel, and C. Quintas, "Data mining models for automatic problem identification in intensive medicine," *Procedia Computer Science*, Elsevier B. V., vol. 210, pp. 218–223, 2022. https://doi.org/10.1016/j.procs.2022.10.140

[7] E. Naydenova, A. Tsanas, S. Howie, C. Casals-Pascual, and M. De Vos, "The power of data mining in diagnosis of childhood pneumonia," *J R Soc Interface*, vol. 13, p. 20160266, 2016. https://doi.org/10.1098/rsif.2016.0266

[8] K. E. Alhajaj and I. A. Moonesar, "The power of big data mining to improve the health care system in the United Arab Emirates," *J Big Data*, vol. 10, no. 1, 2023. https://doi.org/10.1186/s40537-022-00681-5

[9] D. Perna, and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proceedings – IEEE Symposium on Computer-Based Medical Systems, Institute of Electrical and Electronics Engineers Inc.*, 2019, pp. 50–55. https://doi.org/10.1109/CBMS.2019.00020

[10] Y. Wang *et al.*, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *J Biomed Inform*, vol. 102, 2020. https://doi.org/10.1016/j.jbi.2019.103364

[11] Y. Liu, "Analysis and prediction of college students' mental health based on k-means clustering algorithm," *Applied Mathematics and Nonlinear Sciences*, vol. 7, no. 1, pp. 501–512, 2022. https://doi.org/10.2478/amns.2021.1.00099

[12] M. Pulgar-Sánchez *et al.*, "Biomarkers of severe COVID-19 pneumonia on admission using data-mining powered by common laboratory blood tests-datasets," *Comput Biol Med*, vol. 136, p. 103364, 2021. https://doi.org/10.1016/j.compbiomed.2021.104738

[13] N. Khajehali and S. Alizadeh, "Extract critical factors affecting the length of hospital stay of pneumonia patient by data mining (Case study: An Iranian hospital)," *Artif Intell Med*, vol. 83, pp. 2–13, 2017. https://doi.org/10.1016/j.artmed.2017.06.010

[14] H. J. Lai, P. C. Chan, H. H. Lin, Y. F. Chen, C. S. Lin, and J. C. Hsu, "A web-based decision support system for predicting readmission of pneumonia patients after discharge," in *Proceedings – 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018, Institute of Electrical and Electronics Engineers Inc.,* 2019, pp. 2305–2310. https://doi.org/10.1109/SMC.2018.00396

[15] S. Cascini *et al.,* "Pneumonia burden in elderly patients: A classification algorithm using administrative data," *BMC Infect Dis*, vol. 13, no. 1, 2013. https://doi.org/10.1186/1471-2334-13-559

[16] P. A. Kache *et al.,* "Geospatial cluster analyses of pneumonia-associated hospitalisations among adults in New York City, 2010-2014," *Epidemiology and Infection*, vol. 147, Cambridge University Press, 2019. https://doi.org/10.1017/S0950268818003060

[17] H. Hegde *et al.,* "Identifying pneumonia subtypes from electronic health records using rule-based algorithms," *Methods Inf Med*, vol. 61, nos. 1–2, pp. 29–37, 2022. https://doi.org/10.1055/a-1801-2718

[18] E. Murillo-Zamora, R. A. Sánchez-Piña, X. Trujillo, M. Huerta, M. Ríos-Silva, and O. Mendoza-Cano, "Independent risk factors of COVID-19 pneumonia in vaccinated Mexican adults," *International Journal of Infectious Diseases*, vol. 118, pp. 244–246, 2022. https://doi.org/10.1016/j.ijid.2022.02.003

[19] M. Moreno-Noguez, R. Rivas-Ruiz, I. A. Roy-García, D. O. Pacheco-Rosas, S. Moreno-Espinosa, and A. A. Flores-Pulido, "Risk factors associated with SARS-CoV-2 pneumonia in the pediatric population," *Bol Med Hosp Infant Mex*, vol. 78, no. 4, pp. 251–258, 2021. https://doi.org/10.24875/BMHIM.20000263

[20] N. Gabriel-Job and U. S. Azubogu, "Prevalence and pattern of pneumonia among children admitted into university of Port Harcourt teaching hospital: A two year review," *Int J Trop Dis Health*, pp. 1–6, 2020. https://doi.org/10.9734/ijtdh/2019/v40i230225

[21] T.-A. Abebaw, W. K. Aregay, and M. T. Ashami, "Risk factors for childhood pneumonia at Adama Hospital Medical College, Adama, Ethiopia: A case-control study," *Pneumonia*, vol. 14, no. 1, 2022. https://doi.org/10.1186/s41479-022-00102-4

[22] S. K. Kasundriya, M. Dhaneria, A. Mathur, and A. Pathak, "Incidence and risk factors for severe pneumonia in children hospitalized with pneumonia in Ujjain, India," *Int J Environ Res Public Health*, vol. 17, no. 13, pp. 1–15, 2020. https://doi.org/10.3390/ijerph17134637

[23] C. Y. Huang *et al.*, "Risk factors of progressive community-acquired pneumonia in hospitalized children: A prospective study," *Journal of Microbiology, Immunology and Infection*, vol. 48, no. 1, pp. 36–42, 2015. https://doi.org/10.1016/j.jmii.2013.06.009

[24] V. N. Sutriana, M. N. Sitaresmi, and A. Wahab, "Risk factors for childhood pneumonia: A case-control study in a high prevalence area in Indonesia," *Clin Exp Pediatr*, vol. 64, no. 11, pp. 588–595, 2021. https://doi.org/10.3345/cep.2020.00339

[25] O. T. Uwemedimo *et al.*, "Distribution and determinants of pneumonia diagnosis using Integrated Management of Childhood Illness guidelines: A nationally representative study in Malawi," *BMJ Glob Health*, vol. 3, no. 3, p. e000506corr1, 2018. https://doi.org/10.1136/bmjgh-2017-000506corr1

[26] E. Benincà, M. Van Boven, T. Hagenaars, and W. Van Der Hoek, "Space-time analysis of pneumonia hospitalisations in the Netherlands," *PLoS One*, vol. 12, no. 7, p. e0180797, 2017. https://doi.org/10.1371/journal.pone.0180797

[27] J. M. Ooi, K. P. Eg, K. Chinna, A. M. Nathan, J. A. de Bruyne, and S. Thavagnanam, "Predictive risk factors for complicated pneumonia in Malaysian children," *J Paediatr Child Health*, vol. 55, no. 4, pp. 406–410, 2019. https://doi.org/10.1111/jpc.14213

[28] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013. https://doi.org/10.4097/kjae.2013.64.5.402

[29] G. T. Reddy *et al.*, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020. https://doi.org/10.1109/ACCESS.2020.2980942

[30] A. M. Fahim, A. M. Salem, F. A. Torkey, and M. A. Ramadan, "Efficient enhanced k-means clustering algorithm," *J Zhejiang Univ Sci*, vol. 7, no. 10, pp. 1626–1633, 2006. https://doi.org/10.1631/jzus.2006.A1626

[31] S. Abadi *et al.*, "Application model of k-means clustering: Insights into promotion strategy of vocational high school," *International Journal of Engineering and Technology (UAE)*, vol. 7(2.27), Special Issue 27, p. 182, 2018. https://doi.org/10.14419/ijet.v7i2.11491

[32] M. Kushwaha, H. Yadav, and C. Agrawal, "A review on enhancement to standard k-means clustering," in *Lecture Notes in Networks and Systems*, 2020, vol. 100. https://doi.org/10.1007/978-981-15-2071-6_26

[33] G. Bonaccorso, Machine Learning Algorithms – Second Edition, Packt Publishing, 2018.

[34] A. Fahim, "K and starting means for k-means algorithm," *J Comput Sci*, vol. 55, p. 101445, 2021. https://doi.org/10.1016/j.jocs.2021.101445

[35] S. O. Akinola, Q.-G. Wang, P. Olukanmi, and T. Marwala, "Predicción temprana del brote del virus de la viruela del mono mediante aprendizaje automático," *Transacciones IETI sobre análisis y pronóstico de datos (iTDAF)*, vol. 1, no. 2, pp. 14–29, 2023. https://doi.org/10.3991/itdaf.v1i2.40175

[36] R. Al-ahmadi, H. Al-ghamdi, and L. Hsairi, "Classification of diabetic retinopathy by deep learning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 20, no. 1, pp. 74–88, 2024. https://doi.org/10.3991/ijoe.v20i01.45247

[37] X. Wei, N. Gao, M. Xu, M. Rezaeitalesh, N. Mu, Z. Lyu, … X. Chang, "Predicción de la positividad del β-amiloide en la enfermedad de Alzheimer: análisis completo de la selección de características y modelos de aprendizaje automático para una identificación precisa," *Revista internacional de ingeniería biomédica y en línea (iJOE)*, vol. 19, no. 17, pp. 98–114, 2023. https://doi.org/10.3991/ijoe.v19i17.44173

7 AUTHORS

Jesus Orlando Gil Jauregui is a software engineering undergraduate student at Universidad Nacional Mayor de San Marcos (UNMSM), Peru. His research interests include data science, machine learning, big data, data mining. He can be contacted at email: ogiljauregui@gmail.com

Angel Gerardo Carmen Cruzatti is a bachelor's student in software engineering at the National University of San Marcos (UNMSM), Peru. His research interests include machine learning, big data, data mining, and software development. He can be contacted at email: angelcarmen0302@gmail.com

Miguel Angel Cano Lengua is a professor at the Technological University of Peru (UTP) and the National University of San Marcos (UNMSM), has a degree in Mathematics, a PhD Engineering of Systems and Computer Science from the Universidad Nacional Mayor de San Marcos, a Master's in Systems Engineering from the National University of Callao (UNAC). He works on continuous optimization, artificial intelligence algorithms, conical programming, numerical methods, methodology, and software design. He can be contacted at email: mcanol@unmsm.edu.pe

Hugo Villaverde Medrano PhD in Educational Administration and Master's in Administration and Computer and Systems Engineering. He works as a Business Intelligence Professional at the Comprehensive Health Insurance (SIS) of Peru. Professor at the Technological University of Peru (UTP) and others. Specialist in Oracle Database Management System and SQL Server.