Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 3, June 2025, pp. 2171~2187
ISSN: 2302-9285, DOI: 10.11591/eei.v14i3.8381

Public health challenges in the Cuzco region: a decade of anemia in vulnerable populations applying data mining

Inoc Rubio Paucar[1], Laberiano Andrade-Arenas[2]
[1]Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú
[2]Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú

Article Info

Article history:

Received Feb 24, 2024
Revised Nov 30, 2024
Accepted Dec 25, 2024

Keywords:

Anemia
Data mining
Information technologies
Knowledge discovery
Rapid miner studio

ABSTRACT

The objective of the research is to carry out an exhaustive analysis of anemia in the province of Cusco using the Rapid Miner Studio tool that allows an analysis of the number of most concurrent cases in each district of the province of Cusco. Different sources of information were consulted to take as a reference the impact of the disease in different parts of the world. Likewise, information was introduced about how information technologies manifest positive responses in certain diseases around the world. The knowledge discovery in databases (KDD) methodology was used, which consists of several phases proposed in the project, such as data selection, data preprocessing, data mining and evaluation of results. Consequently, this research will help to recognize the most abundant cases in the districts of the province of Cusco. The results obtained were that 348 confirmed cases of anemia occurred in the district of Espinar, being the most affected district. Finally, it was concluded that in different provinces, not only in Cusco, there is a high prevalence of the disease due to factors associated with its treatment.

Corresponding Author:

Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades
University Avenue 5175, Lima, Peru
Email: landrade@uch.edu.pe

1. INTRODUCTION

One of the most relevant problems in the province of Cusco located in Peru is anemia. There is a high excessive prevalence of anemia cases exclusively in children who have a weakened immune system due to a lack of poor nutrition by their parents. Economic problems are a triggering problem for this type of disease, combined with the lack of constant information from the health authorities that has led to the increasing number of cases in recent years.

It is a disease in which high concentrations of red blood cells (RBCs) called hemoglobin are below normal. This reduces the predisposition to transport blood to the tissues accompanied by oxygen in body parts and reduces the body's ability to do so. On the other hand, one of the causes presented is the deficient intake of iron either in food for human consumption or dietary supplements based on iron and other substances to combat anemia [1]. It is stated that the amount of anemia in pregnant women reaches almost 42% and 30% of non-pregnant women reflect risk situations associated with the disease. On the African continent a high mortality of mothers who contract the symptoms of anemia is preserved referring to other factors that are related. Research conducted in Ethiopia for the year 2016 reported that 24% of women of reproductive age contracted the disease especially women in rural areas are the most affected with 25% and 17% of established cases [2]. According to other research, anemia has several causes for the conception of the disease. There are health programs promoted by the state government of each country, such as the iron supplementation program for adolescents with this disease. On the other hand, the proposed study refers to the evaluation of a cross-sectional study conducted with a sample of 399 high school girls in the northern, southern, eastern, western and central regions of Iran [3]. Another point to highlight is malnutrition in women in a reproductive state and children under preschool age. Consequently, in any part of the world, there is a percentage of low-income families, which is the main problem faced by several countries worldwide [4]. In Colombia, mothers of families present anemia in both pregnant and non-pregnant women, taking into account that low-income people are the most prone to contracting this disease. The main objective of this research was to develop the danger that exists about the disease focused on mothers and children who attend programs called healthy child control. The conclusion was carried out according to the symptoms presented by the patients such as diarrhea, fever or respiratory symptoms due to lack of food nutrition and nutritional supplements [5].

Hemoglobin concentration maintains a physiological concept for the evaluation of the disease anemia. This leads to it being one of the most common and dangerous diseases worldwide in developing countries like India according to the World Health Organization (WHO). Consequently, research was conducted on artificial intelligence (AI) for the detection of the disease, which by conventional practices can observe the pallor of the nail to determine if the patient has the disease. This disease requires constant monitoring due to the number of symptoms that manifest the disease that in many cases can be felt physically or seen visually [6]. This disease requires constant monitoring due to the number of symptoms that manifest the disease that in many cases can be felt physically or seen visually. The detection of anemia can be detected according to various characteristics such as rapid laboratory tests called hemograms. Therefore, the implementation of a prediction model based on four concepts such as beta thalassemia trait (BTT), iron deficiency anemia, hemoglobin E (HbE), and combined anemias despite the presence of multiple RBC indices. Using the extreme learning machine (ELM) algorithm to determine whether a patient has the disease [7]. However, specialized treatment to combat anemia has high costs based on medical tests and other characteristics that are not available in some countries. On the other hand, waste leads to biological risks and pollutants for the environment. To avoid this type of problem that affects not only patients but also the environment, an AI-oriented chatbot was developed to detect anemia. The software is created using segmentation models of regions of interest and classification models that allow the detection of anemia cases outside the normal ones. For this purpose, data were collected from 160 patients with anemia and 140 without the disease. A data training phase is performed with cloud services using REAN chatbot services [8].

The combination of new technologies with medicine has provided a great opportunity to provide hope for treatment or cure of the disease. AI has a high impact on the prediction of various factors that cause people to contract anemia. The goal is to find solutions to this disease that is growing in different population districts that go hand in hand with certain economic and social factors.

2. LITERATURE REVIEW

2.1. Theoretical basis

2.1.1. Data mining

Data mining techniques encompass a broad array of methodologies, ranging from traditional statistical analysis to cutting-edge machine learning algorithms. These techniques are applied across various domains such as business, healthcare, and finance, to extract valuable insights from massive datasets [9]. Supervised learning algorithms like Khilari are utilized to classify data into predefined categories, while unsupervised learning techniques like clustering help identify hidden patterns or structures within the data without predefined labels [10]. Additionally, association rule mining uncovers relationships and dependencies among variables, while anomaly detection flags unusual or suspicious data points. The synergy of these diverse techniques empowers organizations to make data-driven decisions, optimize processes, and gain a competitive edge in today's data-centric landscape.

Applied data mining techniques

In data mining, different techniques are applied for data analysis through the application of algorithms in each technique applied to different fields within the social field [11]. Therefore, algorithms described in different groups are applied according to their use.

- Association rules: this technique performs a series of relationships called Itemset which is a collection of one or more items. For this, there is a relationship between both Itemset based on rules where there is an antecedent and a consequent.
- Classification: classification algorithms are computer approaches for categorizing or classifying instances or items. These algorithms discover patterns in labelled training data and then use those patterns to categorize new data. A classification algorithm's fundamental objective is to construct a model that can generalize underlying patterns in data and, as a result, assign new examples to the associated classes.

- Clustering: for the clusters, the grouping of specific data according to their similarities is performed. With this, similarity patterns are found so that each collection is distinct between clusters. But the objects are similar from each cluster in some way.
- Sequence and trajectory analysis: it is intended to look for patterns in a set of events or values that flow to subsequent ones. It recognizes several variants in the data that occur at regular intervals or in the ebb and flow of data points over a long time.

2.1.2. Anemia

Anemia is oriented to the reduction in hemoglobin concentration by its threshold in blood. In this sense, the risk problems predisposed to the disease are specified in severe blood loss, for example during childbirth in pregnant women. The symptoms attributed to the disease to factors such as fatigue and weakness among others of severe cases. Consequently, it generates a risk factor for premature birth in newborns with low birth weight and postpartum depression [12]. It is important to consider that one of the problems for the development of the disease is the lack of iron. Excessive intake of iron-inhibiting foods or insufficient intake of bioavailable iron is due to the lack of micronutrients such as vitamin B12, vitamin A, or genetic disorders [13].

- Anemia in pregnant women

Conception in women, a pregnancy has several risks associated with it if it is not constantly monitored. One of the problems is anemia which results in dangers such as infant mortality, low birth weight, premature birth, and irrecoverable neurobehavioral and cognitive deficits that affect the newborn in their life over time [14]. Research estimates that 115,000 maternal deaths are caused by anemia each year. 46% of pregnant women in Nepal are anemic. Family involvement and counseling of pregnant women can increase adherence to iron and folic acid medication tablets unveiling an integrated strategy to prevent anemia. However, these interventions are often less accessible to marginalized women [15]. On the other hand, other research showed that 14 weeks postpartum had a high impact of recommending supplementation to pregnant women with serum ferritin below 20 g/L in early pregnancy, as well as exploring which factors were related to changes in iron status according to different iron indicators [16]. In pregnant women, anemia is one of the threatening public health problems in both mothers and children. For this reason, a study was carried out to investigate the administration of iron supplements during pregnancy. For this research, data were collected from women between 15 and 49 years of age, obtaining a total of 14,564 women in this age range [17].

- Anemia in children and adolescents

In the African continent, according to the data, there is a high infant morbidity and mortality rate [18]. There is even a prevalence of factors associated with other diseases related to this disease. Research estimates hypothyroidism at 6% in children and adolescents [19]. On the other hand, anemia can be associated with hereditary factors which is a challenging concept in areas specifically with limited use of genetic studies. Therefore, it is difficult to draw an accurate diagnostic approach to undiagnosed anemia which is a challenge for specialists in the field [20]. Decreased hemoglobin in people with sickle cell disease (SCD) is associated with lower oxygen saturation (SpO_2) and increased risk of stroke, both of which are associated with lower intelligence quotient (IQ) scores. Therefore, hemoglobin and SpO_2 may be increased in people with SCD [21]. In another context, anemia also refers to acute lymphoblastic leukemia which is prescribed as a blood cancer due to the affection of the RBCs. The cause that generates this disease is unknown. However, the predisposition to the disease of infantile acute leukemia presents 80% of leukemia in children and 20% in adults [22]. In that sense, anemia can bring serious consequences with the progressive growth of the infant affecting the educational and healthy perspective of the child's life. These risk factors guarantee a non-anemic study in children with growth retardation studies that are analyzed as serious consequences below [23].

3. METHOD

3.1. Definition of the knowledge discovery in databases method

The knowledge discovery in databases (KDD) methodology aims to perform a structured approach on a large amount of data. It aims to discover important information by following the steps of the method to subsequently make decisions about what is to be predicted [24]. Data mining includes several steps such as data processing, feature engineering, model training, model evaluation, result visualization, and model interpretation, among others [25]. For this purpose, engineering has characteristics that must be known to perform data processing integrating data mining such as knowledge domain, correlation analysis, and elimination among other important factors.

Figure 1 depicts the several steps that comprise the KDD technique. Throughout this technique, each stage of the research's development is described in depth. As a result, each of these steps greatly contributes to the enhancement of the quality of information stored internally in a database. This organized technique promotes a systematic and efficient approach to information management by facilitating the extraction, processing, and interpretation of useful knowledge from data sets.

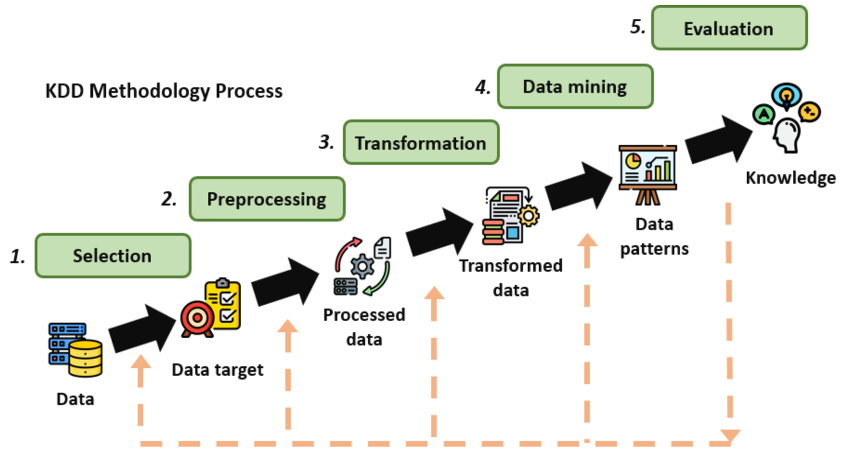

Figure 1. KDD methodology

3.2. Definition of the knowledge discovery in databases method process

The KDD methodology is a structured process that addresses the transformation of insignificant data into meaningful information and ultimately into useful knowledge. The complete KDD framework consists of several stages that are followed in this discovery journey. The success of the process depends on each step, from data selection and preprocessing to the evaluation of the results. This process begins with an understanding of the domain and a precise definition of the business objectives, which guide all subsequent stages. This introductory paragraph will discuss the different stages of the KDD methodology, highlighting its role in extracting valuable knowledge from large data sets.

3.2.1. Data selection process

At this stage, data are selected for analysis. This involves the identification and extraction of data from a variety of sources, including databases, files, and records. The information obtained is translated as a set of data selected from reliable sources of information. For this purpose, we intend to perform an analysis in terms of people contracting anemia between the years 2010 onwards. This information is made up of the following data. Table 1 shows the data selected for the model using the decision tree algorithm in the Rapid Miner Studio tool. In this database, we present the important records that will allow us to corroborate whether the number of people who contract anemia in the Department of Cusco is the same as the number of people who contract anemia in the Department of Cusco.

Table 1. Anemia database

Department	Province	Distrito	Case	Total	Cod_Ess	Eess	Age	Year
Cusco	Acomayo	Acomayo	1	1	2317	C.s. Acomayo	6	2010
Cusco	Acomayo	Acomayo	5	6	2317	C.s. Acomayo	6	2013
Cusco	Acomayo	Acomayo	3	24	2317	C.s. Acomayo	7	2010
Cusco	Acomayo	Acomayo	7	10	2317	C.s. Acomayo	7	2013
Cusco	Acomayo	Acomayo	2	5	2317	C.s. Acomayo	9	2014
Cusco	Acomayo	Acomayo	1	1	2317	C.s. Acomayo	11	2012
Cusco	Acomayo	Acomayo	3	3	2317	C.s. Acomayo	13	2012
Cusco	Acomayo	Acomayo	0	9	2317	C.s. Acomayo	15	2012
Cusco	Acomayo	Acomayo	0	5	2317	C.s. Acomayo	17	2011
Cusco	Acomayo	Acomayo	0	6	2317	C.s. Acomayo	17	2014
Cusco	Acomayo	Acomayo	0	5	2317	C.s. Acomayo	21	2014
Cusco	Acomayo	Acomayo	3	5	2317	C.s. Acomayo	24	2013
Cusco	Acomayo	Acomayo	2	2	2317	C.s. Acomayo	25	2011
Cusco	Acomayo	Acomayo	8	9	2317	C.s. Acomayo	28	2013
Cusco	Acomayo	Acomayo	0	3	2317	C.s. Acomayo	34	2012
Cusco	Acomayo	Acomayo	1	1	2317	C.s. Acomayo	35	2013

– Filtering by conditions

The use of (1) entails selecting all records, which results in a variable reaching or exceeding a predefined limit. This procedure is required to perform the appropriate calculations in the context of current research or analysis. The formula provides a formal framework for evaluating the link between the records and the variable under consideration, allowing the identification of patterns, trends, or values that fit the defined criteria. These calculations' correctness and relevance promote informed decision-making and contribute to the creation of conclusions based on the information collected from the selected records. Formulation (conceptual):

$$Selected\ data = \{x \mid condition\ is\ met\ for\ x\} \quad (1)$$

– Filtering by ranges

This method of data selection within a designated interval, outlined by (2), plays a pivotal role in the management and analysis of datasets. By delineating a subset of data using explicit boundaries, researchers are equipped with a precise approach to identify and manipulate information pertinent to a specific investigation. This targeted approach not only enhances the efficiency of data analysis but also ensures that resources are allocated effectively, optimizing the extraction of meaningful insights for informed decision-making. Moreover, the systematic nature of this method facilitates reproducibility and fosters transparency in research practices, contributing to the overall reliability and validity of findings. Formulation (conceptual):

$$Selected\ data = \{x \mid a \leq x\ \leq b\} \quad (2)$$

– Frequency filtering

In this scenario, data selection hinges on the frequency of occurrence of specific values, a process governed by (3). By applying this criterion, researchers can filter the dataset to encompass only those categories with a noteworthy frequency, thereby honing in on the most relevant and prevalent data points. This targeted filtration method ensures that the analysis focuses on the most influential and informative aspects of the dataset, minimizing noise, and maximizing the utility of extracted insights. Moreover, by prioritizing categories with substantial frequency, researchers can uncover patterns, trends, and relationships that hold greater significance within the dataset, facilitating more robust, and actionable conclusions. Formulation (conceptual):

$$Selected\ data = \{x \mid frequency\ of x\ is\ above\ a\ certain\ threshold\} \quad (3)$$

– Outlier data removal (Outliers)

Outlier elimination is a crucial step aimed at enhancing the robustness and accuracy of data analysis by identifying and excluding anomalies that may distort the results. Statistical techniques like the interquartile range (IQR) offer a systematic approach to detect and remove outliers, ensuring that the analysis is based on reliable and representative data points. In (4) delineates specific criteria tailored to the contextual nuances of the problem, providing a flexible framework to address outliers effectively. By employing such methods, researchers can mitigate the influence of aberrant observations, thereby fostering more dependable, and meaningful analytical outcomes. Formulation (conceptual):

$$Selected\ data = \{x \mid x\ is\ not\ an\ outlier\} \quad (4)$$

Indeed, statistical measures like the IQR serve as valuable tools in outlier detection, allowing researchers to pinpoint data points that deviate significantly from the bulk of the dataset. By calculating the IQR, which represents the range between the first and third quartiles of the data distribution, outliers can be identified as observations lying beyond a certain multiple of the IQR away from the quartiles. Subsequently, data points falling outside this defined threshold can be deemed outliers and excluded from further analysis.

– Random sample

Random sampling, a fundamental technique in statistical analysis, involves the selection of data points from a larger population entirely at random. This approach ensures that each member of the population has an equal chance of being included in the sample, thereby minimizing bias and promoting the creation of representative subsets. In (5) encapsulates the principles of random sampling, providing a systematic framework to generate samples that accurately reflect the characteristics of the population. Formulation (conceptual):

$$Selected\ data = random\ sample\ n \quad (5)$$

You can use random sampling methods to select a representative sample of your data.

– Selection of relevant variables

Variable selection is a critical preprocessing step aimed at identifying and retaining the most informative features or columns within a dataset for analysis or model development. By carefully selecting relevant characteristics, researchers can streamline the analytical process, reduce the computational burden, and enhance the interpretability and performance of predictive models. In (6) encapsulates the essence of variable selection, providing a structured approach to identify and prioritize features based on their contribution to the desired outcome or target variable. Formulation (conceptual):

$$Selected\ data = \{x_1, x_2, \ldots\ x_n\} \tag{6}$$

Variable selection enables the prioritization of relevant factors for analysis, as depicted in Figure 2. This process involves extracting data from a comma separated values (CSV) file and employing operators to isolate pertinent fields for constructing a model related to anemia cases in Cusco. By focusing solely on variables crucial to the analysis, statistical calculations can be tailored to specific parameters extracted from the database, facilitating the identification of anemia cases. This targeted approach ensures that resources are allocated efficiently, optimizing the accuracy and efficacy of the analytical process.

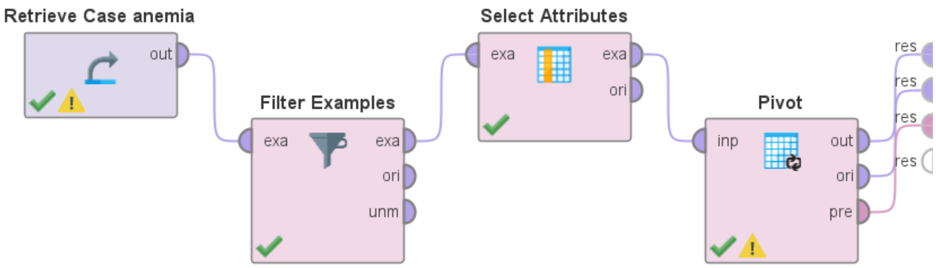

Figure 2. Data mining process

3.2.2. Data preprocessing

To carry out an effective analysis, some data is essential to a comprehensive cleaning and transformation process. It involves addressing aspects such as the identification and elimination of outliers, the imputation of missing data, the normalization of variables, and the selection of relevant characteristics. These steps are crucial to ensuring that the data is free from inconsistencies and optimally prepared for subsequent phases of the method. Meticulous attention to data quality at this stage lays the foundation for a more robust and reliable analysis in the later stages of the selected method.

– Standardization

Elimination of outliers is a pivotal step in data preprocessing, necessitating the identification and exclusion of observations that deviate substantially from the majority of the dataset. These outliers, due to their extreme values or unusual characteristics, have the potential to skew statistical analyses and lead to erroneous conclusions. In (7) encapsulates the method for detecting and eliminating outliers, providing a systematic framework to enhance the robustness and reliability of subsequent analyses. By employing statistical techniques such as the calculation of z-scores or the IQR, researchers can effectively identify outliers and mitigate their impact on the analysis. Formula of z-score (standardization):

$$z = \frac{(x - \mu)}{\sigma} \tag{7}$$

where x is the original value, μ is the mean, and σ is the standard deviation.

– Imputation of data

The data imputation, represented by (8), involves the systematic addition of estimated values to a dataset to fill in missing or incomplete entries. This process is crucial for ensuring the completeness and integrity of the dataset, particularly in scenarios where missing values could compromise the validity of subsequent analyses or models. By leveraging statistical techniques such as mean imputation, regression imputation, or machine learning algorithms, researchers can infer plausible values for missing data points based on the information available in the dataset. Mean to impute missing values:

$$x = \frac{\sum_{i=1}^{n} x_i}{n} \tag{8}$$

In an ordered dataset, the center value is called the median. The imputation of the median is the process of substituting the mean value of the variable for any missing value. Compared to the imputation of the mean, this approach is less vulnerable to extreme values, or outliers, and is more resistant against them as mentioned in (9).

\text{Median} = \frac{x_{{\frac{n+1}{2}}\text{ if n is odd,o }\frac}{\frac{n}{2} + \frac{n}{2} + 1}{2} \text{ if n is even} (9)

– Selection of features

The feature selection process, as illustrated in (10), encompasses the strategic identification and inclusion of variables that wield significant influence on the analysis while disregarding those that contribute minimal value or introduce noise. This critical step in data preprocessing is essential for streamlining analyses, enhancing model performance, and fostering interpretability. Through techniques such as filter methods, wrapper methods, or embedded methods, researchers can systematically evaluate the relevance of each variable to the target outcome and retain only those that contribute meaningfully to the predictive power of the model. Correlation coefficient (p):

$$p(X, Y) = \frac{cov(X,Y)}{\sigma_X \, \sigma_Y} \tag{10}$$

The significance of features, elucidated through model selection methods like decision trees' Gini importance score, is paramount in guiding the construction of effective predictive models. This metric quantifies the contribution of each feature to the predictive power of the model, thereby enabling researchers to prioritize variables that exert the most substantial influence on the target outcome. By leveraging decision trees' Gini importance score, researchers can discern which features play pivotal roles in driving predictive accuracy and refine their models by focusing on these influential factors.

– Elimination of outliers

Identifying and excluding atypical values, as demonstrated in (11), is crucial to safeguarding the integrity and reliability of analytical conclusions. These outliers, characterized by their significant deviation from the majority of the dataset, possess the potential to distort statistical analyses and lead to erroneous interpretations. By systematically detecting and removing such outliers, researchers can mitigate their influence on the analysis, ensuring that conclusions are based on a more representative and accurate portrayal of the data distribution. Techniques such as IQR, z-scores, or clustering algorithms enable researchers to effectively pinpoint and exclude observations that exhibit aberrant behavior. Interquartile range (IQR):

$$IQR = Q3 - Q1 \tag{11}$$

In range $Q1 - 1.5 * IQR \; to \; Q3 + 1.5 * IQR$, where Q1 represents the first quartile and Q3 denotes the third quartile, the defined range serves as a critical threshold for identifying atypical values within a dataset. Observations falling outside this range are deemed atypical or outliers, as they deviate significantly from the central tendency of the data distribution. By delineating this range based on quartile values, researchers can systematically flag observations that lie beyond the expected variability of the dataset, thereby facilitating the detection and exclusion of aberrant data points.

At this pivotal juncture of the process, Figure 3 serves as an invaluable visual aid, offering a comprehensive overview of the essential components within the Rapid Miner Studio tool interface. This dynamic environment equips researchers with a suite of indispensable tools tailored to execute fundamental tasks crucial for data preprocessing and analysis. From data filtering to field selection and addressing integrity-related issues, Rapid Miner Studio provides a user-friendly platform to streamline workflows and expedite analytical endeavors. By leveraging its robust functionalities, researchers can navigate through complex datasets with ease, ensuring the extraction of meaningful insights while upholding data integrity and quality standards.

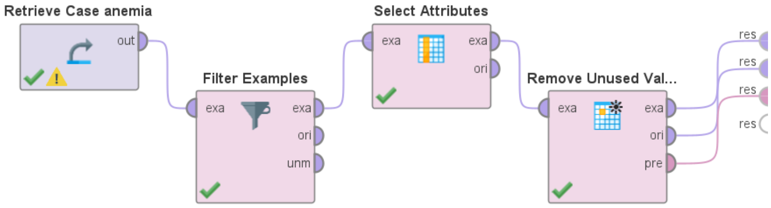

Figure 3. Data preprocessing stage

3.2.3. Data transformation process

A number of changes are made to the data during the data transformation step in order to make it more suitable for analysis. In order to improve the information available for more complex analyses, this process includes steps like variable normalization to guarantee a consistent scale, data format conversion for more effective representation, and the development of new characteristics derived from preexisting ones. To put it simply, the goal of this stage is to modify the data in a way that will make it easier to apply data mining algorithms and extract meaningful patterns.

– Binning (Agrupación)

The technique of grouping, sometimes referred to as discretization or binning, divides a continuous variable's range of values into intervals known as bins or containers. To make analysis and interpretation easier, this method aims to make continuous data representation more straightforward. The width of a container can be calculated using an adaptive formula that adapts automatically to the data distribution or based on predetermined criteria.

– Logarithmic transformation

The logarithmic transformation is employed to enhance the suitability of data for certain statistical models and stabilize variance. When the data show signs of bias or deviate from a normal distribution, this method is quite helpful. Taking the logarithm of every value in the data set is necessary to apply the logarithmic transformation. As a result, extreme numbers typically have less of an effect, and the logarithmic scale can highlight patterns in parts of the data that would otherwise be hard to see.

– Reducing dimensionality

Reducing dimensionality principal component analysis (PCA) is a key step in simplifying complexity and enhancing computer efficiency in analysis because it entails reducing the number of variables in a data collection. There are many ways to complete this work, but two popular ones are choosing attributes and applying PCA. Whereas feature selection selects an ideal subset of variables, keeping pertinent data for analysis, PCA aims to describe the original data in terms of a smaller collection of non-correlated core components.

– Discretization

The process of discretization includes giving continuous variables discrete values in order to make data administration and analysis easier. Rounding down with the soil function $(|x|)$ is a typical example. This method is useful for classifying continuous data into ranges or categories, which is advantageous for statistical analysis and some machine learning techniques.

Figure 4 shows the key Rapid Miner Studio operators involved in the data transformation process. During this stage, new data sets are generated derived from the information previously obtained in previous stages. The fundamental purpose of this transformation is to create relevant and enriched information that will be crucial for the subsequent phase. In this next phase, the information will be classified by implementing the decision tree algorithm. This strategic approach to data generation and optimization aims to optimally prepare information for the classification process, thus helping to improve the effectiveness of the predictive model.

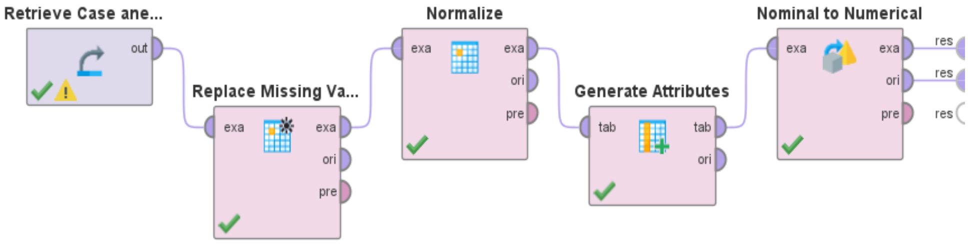

Figure 4. Data transformation process

3.2.4. Data mining process

The data mining process involves the application of various techniques aimed at discovering significant patterns related to the topic raised. This encompasses the use of machine learning algorithms, which play a crucial role in anticipating and modeling these patterns within data analysis. In the context of the research in question, the specific application of the decision tree algorithm will be chosen, which belongs to the group of classification algorithms. This approach aligns directly with the proposed objectives, as decision trees are particularly effective for data classification, providing a visual and logical representation of the decisions made during the analysis. The choice of this algorithm is based on its ability to offer clear interpretations and facilitate the understanding of the determining factors in decision-making within the analyzed data set. This data mining process, with its specific focus on the decision tree algorithm, will reveal underlying patterns and generate valuable insights for the subject of study.

– Decision tree algorithm

A decision tree is a nonparametric supervised learning algorithm that can be used for classification and regression tasks. It consists of a root node, branches, internal nodes, and leaf nodes, and has a hierarchical tree structure [26]. There are no incoming branches at the root node of a decision tree. Decision nodes, or internal nodes, are fed by outgoing branches that originate from the root node. Both kinds of nodes carry out assessments to create homogenous subsets, which are denoted by leaf nodes or terminal nodes, depending on the attributes that are accessible. Every conceivable result in the data collection is represented by the leaf nodes [27]. In this section, a figure is presented that illustrates the structure of the decision tree algorithm that will be used in the prediction model associated with the issue raised. This tree starts with a root node representing the first decision based on a crucial feature. As one moves along the branches of the tree, further subdivisions occur into internal nodes, each representing additional decisions based on different features. This process of expansion continues until the final nodes, or leaves, of the tree are reached, where the final predictions of the model are located. The hierarchical structure of the decision tree allows for a visual and logical representation of how decisions are made in the prediction process. Each node, whether root, internal, or leaf, encapsulates important information about the characteristics and results of the model, facilitating the interpretation of the patterns identified during data mining as specified in Figure 5.

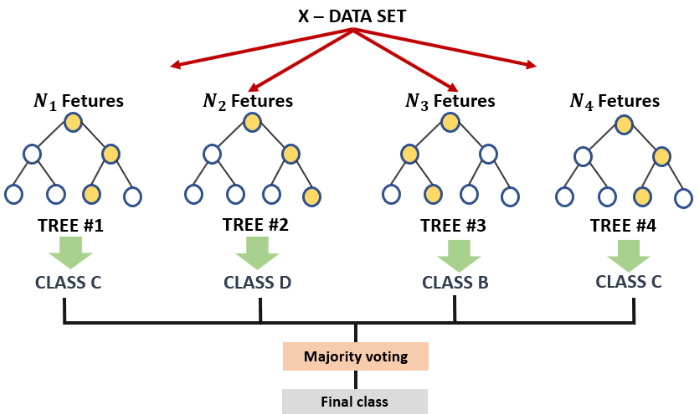

Figure 5. Structure of the decision tree algorithm

– Concept of entropy

Entropy is a metric used in machine learning, particularly in decision tree methods, to quantify the impurity of a data collection. Entropy is a measure of the degree of uncertainty in a class distribution within an example collection. Entropy is used in decision tree design to help determine how to partition a data collection into more homogeneous groupings. In a decision tree, a node is deemed pure if every example within it comes from the same class; if it contains a mix of classes, it is deemed more impure (greater entropy). In the context of machine learning, the entropy formula is modified as mentioned in (12):

$$H(S) = -\sum_{i=1}^{C} Pi * log_2(p_i) \tag{12}$$

where S is a data set, c is the number of classes in S, and p_i is the number of examples in S that belong to class i.

– Information gain

When evaluating the value of dividing a data set into smaller subsets for the purpose of building decision trees, information gain is a key idea. The fundamental idea is to quantify the extent to which splitting the data according to a particular feature reduces uncertainty (also known as entropy) in the classification of the data. The entropy before and after partitioning is compared to gain information. Higher information gain suggests feature-based partitioning. A successfully decreases the degree of ambiguity as mentioned in (13):

$$Gain\ (S, A) = Entropy(S) - \sum v\epsilon Values(A)\ \frac{|S_v|}{S} * Entropy(S_v) \tag{13}$$

where: S is the original data set, A is the characteristic by which the data set is to be divided, *Values(A)* are the possibilities of the characteristic A, $|S_v|$ is the number of examples in S where characteristic A has value v, and $Entropy(S)$ is the entropy of the dataset S, which measures its impurity or uncertainty.

– Gini impurity

Gini impurity is a metric used in conjunction with decision trees and other classification techniques to quantify the homogeneity of a data set. This metric assesses the probability that a randomly selected piece of data will be misclassified relative to the classes that make up that set. In essence, it provides a measure of how mixed or impure the classes within a data set are and is used to guide decision-making in classification processes. The general formula for the Gini impurity *Gini(S)* as mentioned in (14) in a data set *S* are *c* classes is:

$$Gini(S) = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{14}$$

p_i is the portion of examples in the set *S* that belongs to the class *i*.

The values of the Gini impurity range from 0 to 1. A value of 0 denotes a pure data collection, meaning that every sample is part of a single class. Maximum impurity, or a homogeneous distribution of the data set across all classes, is indicated by a value of 1. The smooth transition from data mining to cross-validation underscores the importance of careful data preparation prior to model training. This comprehensive approach strengthens the robustness of the model and its ability to make accurate predictions in real-world situations, as verified in Figure 6.

Figure 7 shows the results obtained after training the model, highlighting the effectiveness and accuracy achieved. The inclusion of the "apply model" and "performance" operators not only facilitates the practical deployment of the decision tree algorithm but also provides a quantitative assessment of its performance. This comprehensive approach supports informed decision-making and continuous improvement of the model based on the results observed during the training phase.

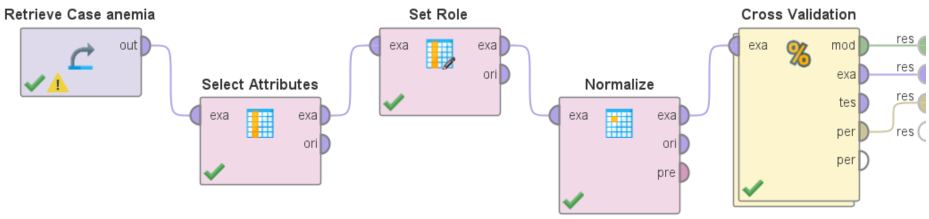

Figure 6. Performing cross validation

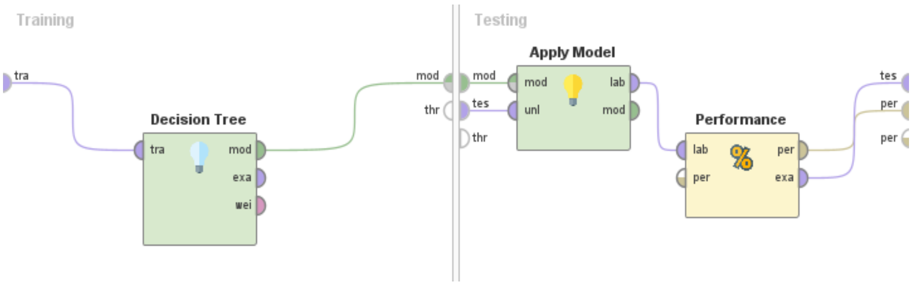

Figure 7. Applying the decision tree algorithm

4. RESULTS
4.1. Evaluation of result

Once the data mining process is complete, it is crucial to evaluate the quality and usefulness of the discovered patterns and models. This assessment encompasses several key aspects, including cross-validation, comparison with relevant metrics and a thorough review of model performance. This analysis can be instrumental in understanding the spatial distribution of case incidence, identifying possible geographic patterns, and prioritizing areas that require special attention. In addition, data visualization through Figure 8 facilitates effective communication of information to different audiences, thus contributing to informed and evidence-based decision making to address the health situation in the province of Cusco.

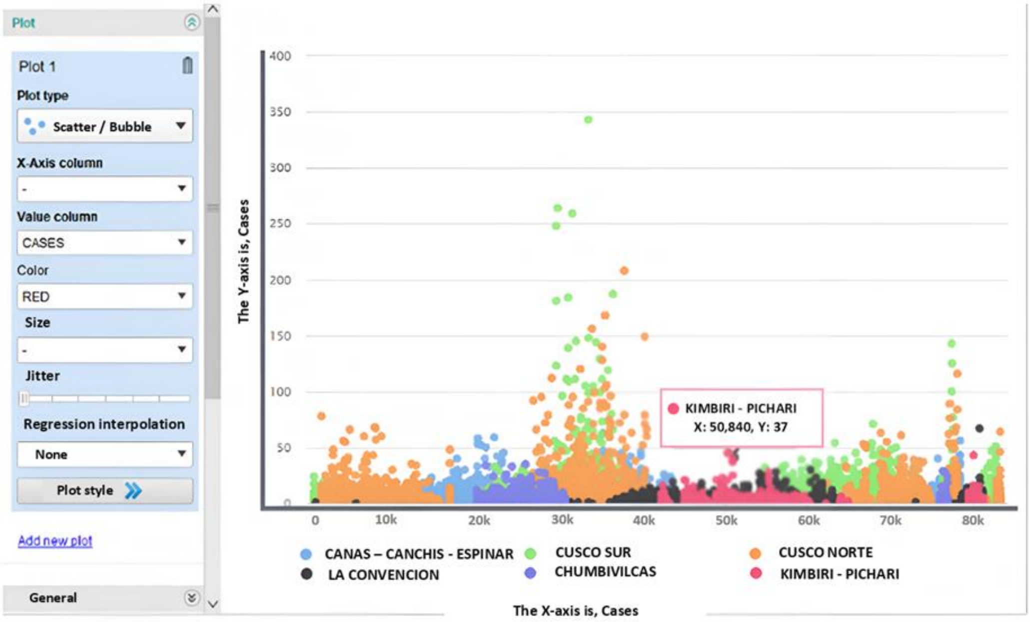

Figure 8. Results of model statistics

To establish the model with a confidence threshold of 0.25, an algorithm was applied that identified that the most prominent districts in terms of cases were Canas, Canchis, Espinar, La Convención, Cusco Sur, Chumbivilcas, Cusco Norte, Limbiri, and Pichari, these being the most impacted by the disease. This analysis provides valuable insight into the geographic areas that require special attention, allowing precise targeting of resources and efforts to effectively address the situation. The graphical representation in Figure 9 not only highlights the magnitude of the impact in San Jeronimo and Echarate, but also facilitates the visual interpretation of the data for a more effective understanding. This visualization can serve as a basis for specific intervention strategies and preventive measures tailored to the particular needs of these districts, thus strengthening the response to the epidemiological situation.

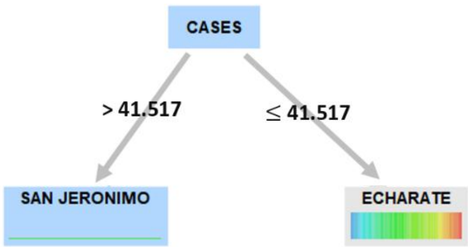

Figure 9. Results of the two districts with the most cases of anemia

Table 2 presents in detail the results obtained after the normalization process. Each cell of the table contains crucial information, including the resulting minimum and maximum impact values as well as the deviation in the previously empty fields. This normalization approach is essential to homogenize data magnitudes and facilitate meaningful comparisons between different variables or regions.

Table 2. Statistical data standardization

Name	Type	Filter attributes
Year	Real	Min: -2.136 Max: 1.403 Deviation: 1.000
Ubigeo	Polynominal	Least: 29 Most: 80902 (5675)
Total	Real	Min: -0.389 Max: 46.230 Deviation: 1.000
Red	Polynominal	Least: 2848 Most: Cusco Norte (25219)
Province	Polynominal	Least: Null (12) Most: La convención (20770)
Normal	Real	Min: -0.517 Max: 33.894 Deviation: 1.000
Microred	Polynominal	Least: Sin Microred (30) Most: Santo tomas (5052)
EESS	Polynominal	Least: PAMPA (1) Most: C.S TTIO (667)
Distrito	Polynominal	Least: Null (29) Most: Echarate (5675)
Departament	Polynominal	Least: Cusco (83412) Most: Cusco (83412)
Cut_date	Real	Min: 0 Max: 0 Deviation: 0
Cod_EEss	Polynominal	Least: 27011 (13) Most: Null (26788)
Cases	Real	Min: -0.340 Max: 62.629
Age	Real	Min: -1.740 Max: 2.112

The graphical representation in Figure 10, where each bar corresponds to a district and its height reflects the magnitude of recorded cases, provides a snapshot and effective view of the geographical distribution of the situation. This visual approach simplifies the rapid identification of the most affected districts, providing an intuitive understanding of the areas that may need more focused attention.

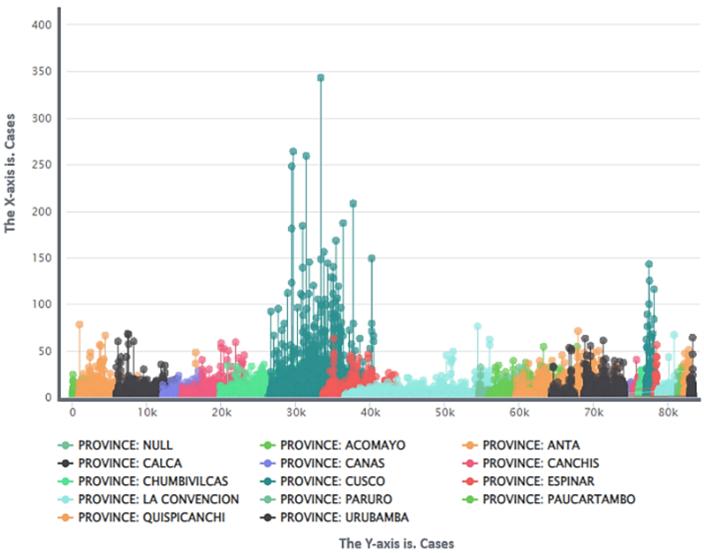

Figure 10. Comparison of cases in the districts of the province of Cusco

In the province of Cusco, the problem of anemia is significantly manifested in several districts. Among them, the district of Espinar stands out as one of the most affected, showing the complexity of the situation through the X-axis of the graph. This representation shows 33,626 cases that, due to various factors described on the X-axis, present worrying indicators related to anemia. These factors may include a predisposition to the disease, susceptibility, or the presence of associated symptoms. Figure 11 depicts a Rapid Miner Studio representation of the correlation matrix application. The relevant attributes were chosen in this section to generate a heat map with these notions. In that sense, using this theory entails creating a heat map that allows for the comparison of the variables used.

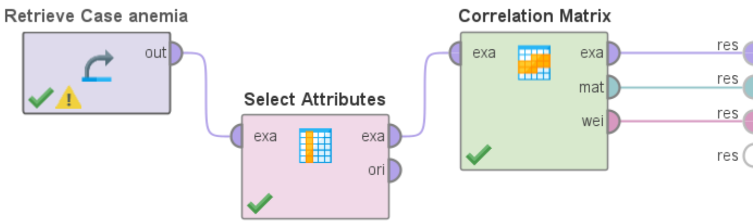

Figure 11. Correlation model

In another instance, Figure 12 shows the heat map itself according to the processes established in previous stages with the Rapid Miner Studio tool. The correlation matrix, also known as attribute weighting in the context of Rapid Miner Studio, is a valuable tool for understanding the relationships between different variables in a data set. This matrix shows how variables are interrelated and provides information on the strength and direction of those relationships as shown in Table 3.

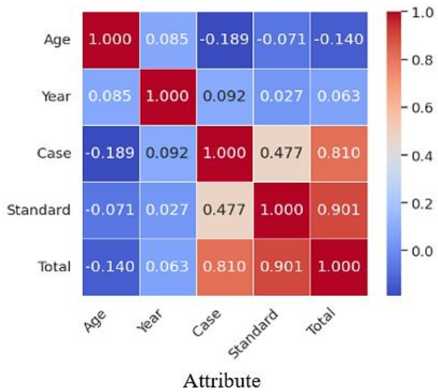

Figure 12. Heat map

Table 3. Statistical data standardization

Attribute weight (correlation matrix)	
Age	0.988
Year	1
Case	0.772
Standard	0.742
Total	0.630
Cut-off date	0

4.2. Evaluation of result

This section compares data mining algorithms in order to assess their usefulness in approaching data mining solutions. To accomplish this, it was chosen to employ a Cartesian plane, which is seen to be the most distinguishable from one to the other. Figure 13 compares models like Naive Bayes, decision trees, and rule induction. The most noticeable algorithm with 1.0 is rule induction, while the rest have lower results, which helps us understand some of the algorithms that may be developed with the tool.

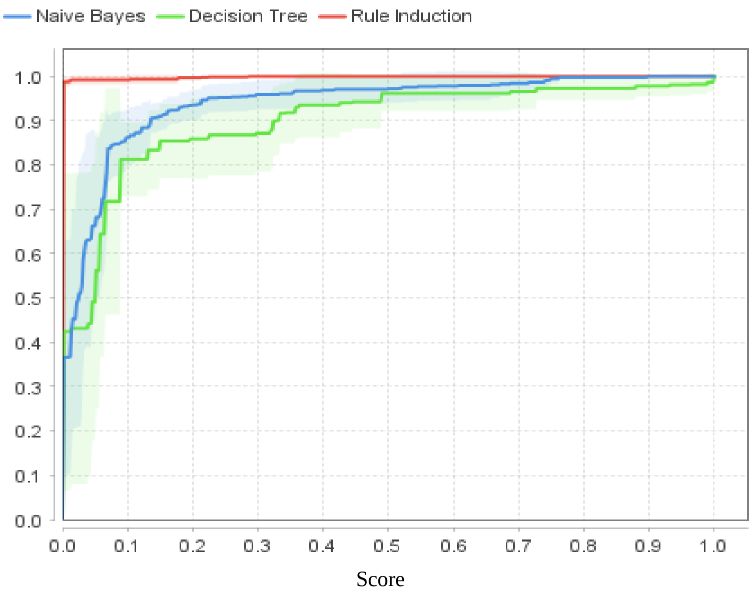

Figure 13. Model comparison

4.3. Comparison of methodologies

SEMMA and CRISP-DM picked the KDD methodology because it was better suited to the proposed data mining project. The KDD methodology is unique in that it covers all aspects of the data mining process, from data selection and preparation to model validation. Furthermore, KDD has been shown to be particularly effective in uncovering useful patterns and insights in massive data sets, which was critical to our research. As shown in Table 4, KDD surpassed SEMMA and CRISP-DM in terms of variety and capacity to meet the specific issues of our data mining project more efficiently and thoroughly.

Table 4. Comparison of methodologies

Comparison attributes	KDD methodology	CRIPS-DM methodology	SEMMA methodology
Structure and sequence	Its structure is not as rigid as other methodologies which consist of: data selection, data cleaning, transformation, data mining and evaluation and application of knowledge.	This method consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment.	SEMMA'S method has five stages: sampling, exploration, modification, modeling and evaluation.
Business orientation	Recognizes the importance of the commercial objectives of the business, seeking to generate knowledge by using it to obtain competitive advantages.	Realizes the understanding of business objectives from the outset and ensures that the results are actionable and valuable for decision making.	It comprises the performance of information analysis taking into account the business purpose and the application of results.
Flexibility	It has a broad and less structured approach that provides a general framework for knowledge discovery.	It is flexible to adapt to a wide variety of domains and projects and can be scalable for business use.	It is flexibly adaptable to different projects, although it follows a predefined sequence of stages.
Interaction	It promotes the iterative process. But, on the contrary, it does not have a defined structure like the SEMMA or Crisp-DM method.	It comprises the implementation of an iterative process reviewing results. In which it is adapted to constantly evolving projects.	It has an iterative process in stages as needed. Adjustments can be made throughout the process.

5. DISCUSSION

The ability to make predictions to obtain meaningful results by identifying significant patterns in a large amount of data is known as data mining. Consequently, a variety of approaches are used to manipulate information according to the needs of the people who use it [9]. A variety of data mining techniques are used to perform these activities using mathematical algorithms, such as machine learning classification, such as Khilari, among [10]. The decrease in hemoglobin concentration below its blood threshold is known as anemia.

The main risk factors for the disease include significant blood losses, such as those occurring during childbirth in pregnant women. Symptoms of the disease are associated with factors such as fatigue and weakness, among other serious symptoms. Thus, it increases the risk of premature delivery in low-birth-weight babies and postpartum depression [12]. Iron deficiency is one of the factors contributing to the development of the disease. Lack of micronutrients such as folate, vitamin B12, vitamin A, or genetic disorders can cause excessive intake of iron-inhibiting foods or insufficient intake of bioavailable iron [13].

6. CONCLUSION

As a conclusion and future work, a series of investigations have been carried out according to the topic raised about people with anemia in the regions of Cusco. In our country anemia is one of the most serious problems, especially in the provinces where the lack of resources is very common. The poor nutrition and poverty is one of the indicators that many people do not have the optimal nutrition that people of different ages need. Another problem is the lack of health centers close to the surrounding districts, which means that many people do not have medical services to treat their illnesses. On the other hand, the few health centers do not have the necessary equipment for the treatment of the disease and therefore the presence of anemia is increasing over the years. The research proposes to analyze the cases of anemia by applying data mining to obtain results through the decision tree algorithm. To perform this task, the KDD methodology was used to perform the corresponding procedures according to the data obtained, allowing to classify the cases presented during the next years from 2010 onwards. The approach of these procedures is given with the data training process that serves to obtain the model according to the fields selected in previous stages. Finally, the results are shown as statistical data for the provinces of Cusco with the most cases presented during the last few years. Finally, the suggested recommendation is to apply technologies related to data mining as expert systems to solve the cases of people with anemia. Because many of them do not have the necessary resources for treatment of the disease or in extreme cases there are no health centers nearby. Therefore, systems based on other technologies would provide a great help in the prevention of other common diseases.

FUNDING INFORMATION

The research described was funded with support from the University of Sciences and Humanities.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Inoc Rubio Paucar		✓	✓	✓	✓			✓	✓		✓		✓	
Laberiano Andrade-Arenas	✓			✓		✓	✓			✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability was obtained from Kaggle open data in this study. The website is https://www.kaggle.com/models.

REFERENCES

[1] S. K. Singh, H. Lhungdim, C. Shekhar, L. K. Dwivedi, S. Pedgaonkar, and K. S. James, "Key drivers of reversal of trend in childhood anaemia in India: evidence from Indian demographic and health surveys, 2016-21," *BMC Public Health*, vol. 23, no. 1, p. 1574, Dec. 2023, doi: 10.1186/S12889-023-16398-W.

[2] M. Yalew *et al.*, "Individual and contextual-level factors associated with iron-folic acid supplement intake during pregnancy in Ethiopia: a multi-level analysis," *BMC Pregnancy Childbirth*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/S12884-023-05593-7.

[3] A. Jafari, Z. Hosseini, H. Tehrani, and A. Alami, "Evaluation of the barriers and facilitators of iron supplementation program among adolescent females," *Clinical Nutrition ESPEN*, vol. 56, pp. 36–42, Aug. 2023, doi: 10.1016/J.CLNESP.2023.04.024.

[4] J. M. Were *et al.*, "The double burden of malnutrition among women of reproductive age and preschool children in low- and middle-income countries: a scoping review and thematic analysis of literature," *Nutrition*, vol. 111, p. 112053, Jul. 2023, doi: 10.1016/J.NUT.2023.112053.

[5] L. Del Castillo *et al.*, "Prevalence and risk factors of anemia in the mother-child population from a region of the Colombian Caribbean," *BMC Public Health*, vol. 23, no. 1, p. 1533, Dec. 2023, doi: 10.1186/S12889-023-16475-0.

[6] S. Das *et al.*, "Smartphone-based non-invasive haemoglobin level estimation by analyzing nail pallor," *Biomedical Signal Processing and Control*, vol. 85, p. 104959, Aug. 2023, doi: 10.1016/J.BSPC.2023.104959.

[7] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, "A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia," *Healthcare (Switzerland)*, vol. 11, no. 5, Mar. 2023, doi: 10.3390/HEALTHCARE11050697.

[8] Pallavi, B. Basumatary, R. Shukla, R. Kumar, B. Das, and A. K. Sahani, "A deep learning-based system for detecting anemia from eye conjunctiva images taken from a smartphone," *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, 2023, doi: 10.1080/02564602.2023.2242318.

[9] P. Venkata, V. Pandya, and A. V. Sant, "Data mining model based differential microgrid fault classification using SVM considering voltage and current distortions," *Journal of Operation and Automation in Power Engineering*, vol. 11, no. 3, pp. 162–172, Oct. 2023, doi: 10.22098/JOAPE.2023.10185.1722.

[10] J. C. Macuácua, J. A. S. Centeno, and C. Amisse, "Data mining approach for dry bean seeds classification," *Smart Agricultural Technology*, vol. 5, Oct. 2023, doi: 10.1016/J.ATECH.2023.100240.

[11] D. Zhang, "Research on dance art teaching system based on data mining and machine learning," *Computer-Aided Design and Applications*, vol. 21, no. S2, pp. 54–68, 2024, doi: 10.14733/CADAPS.2024.S2.54-68.

[12] J. Yourkavitch *et al.*, "A rapid landscape review of postpartum anaemia measurement: challenges and opportunities," *BMC Public Health*, vol. 23, no. 1, p. 1454, Dec. 2023, doi: 10.1186/S12889-023-16383-3.

[13] S. O. Sama, G. S. Taiwe, R. N. Teh, G. E. Njume, S. N. Chiamo, and I. U. N. Sumbele, "Anaemia, iron deficiency and inflammation prevalence in children in the Mount Cameroon area and the contribution of inflammatory cytokines on hemoglobin and ferritin concentrations: a cross sectional study," *BMC Nutrition*, vol. 9, no. 1, Jul. 2023, doi: 10.1186/S40795-023-00748-3.

[14] S. R. Nadhiroh, F. Micheala, and S. T. E. Hui, "The association between maternal anemia and stunting in children aged 0-60 months: a systematic literature review," *Nutrition*, p. 112094, Nov. 2023, doi: 10.1016/J.NUT.2023.112094.

[15] S. Bhattarai *et al.*, "Contextual factors affecting the implementation of an anemia focused virtual counseling intervention for pregnant women in plains Nepal: a mixed methods process evaluation," *BMC Public Health*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/S12889-023-16195-5.

[16] J. Ghosal *et al.*, "To what extent classic socio-economic determinants explain trends of anaemia in tribal and non-tribal women of reproductive age in India? Findings from four National Family Heath Surveys (1998–2021)," *BMC Public Health*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/S12889-023-15838-X.

[17] F. Efendi, I. Israfil, K. Ramadhan, L. McKenna, A. Z. Alem, and H. Malini, "Factors associated with receiving iron supplements during pregnancy among women in Indonesia," *Electronic Journal of General Medicine*, vol. 20, no. 5, pp. 1-7, May 2023, doi: 10.29333/EJGM/13266.

[18] B. T. Zewude and L. K. Debusho, "Multilevel proportional odds modeling of anaemia prevalence among under five years old children in Ethiopia," *BMC Public Health*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/S12889-023-15420-5.

[19] G. Kaudha *et al.*, "Prevalence and factors associated with hypothyroidism in children with sickle cell anemia aged 6 months − 17 years attending the sickle cell clinic, Mulago Hospital, Uganda; a cross-sectional study," *BMC Endocrine Disorders*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/S12902-023-01317-2.

[20] N. H. Aly *et al.*, "A stepwise diagnostic approach for undiagnosed anemia in children: A model for low-middle income country," *Blood Cells, Molecules, and Diseases*, vol. 103, p. 102779, Nov. 2023, doi: 10.1016/J.BCMD.2023.102779.

[21] J. P. MacEwan, A. A. King, A. Nguyen, A. Mubayi, I. Agodoa, and K. Smith-Whitley, "Cognition and education benefits of increased hemoglobin and blood oxygenation in children with sickle cell disease," *PLoS One*, vol. 18, no. 8, Aug. 2023, doi: 10.1371/JOURNAL.PONE.0289642.

[22] S. A. M. Saghir, "A new insight updates in diagnosis and management of acute lymphoblastic leukemia, cytogenetics, immunophenotyping, and proteomic profile," *Electronic Journal of General Medicine*, vol. 20, no. 5, Jun. 2023, doi: 10.29333/EJGM/13386.

[23] Z. W. Htay, T. Swe, T. S. S. Hninn, M. T. Myar, and K. M. Wai, "Factors associated with syndemic anemia and stunting among children in Myanmar: A cross-sectional study from a positive deviance approach," *Archives de Pediatrie*, Aug. 2023, doi: 10.1016/J.ARCPED.2023.03.010.

[24] B. Molina-Coronado, U. Mori, A. Mendiburu, and J. Miguel-Alonso, "Survey of network intrusion detection methods from the perspective of the knowledge discovery in databases process," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2451–2479, Dec. 2020, doi: 10.1109/TNSM.2020.3016246.

[25] C. Zhang, J. Lu, and Y. Zhao, "Generative pre-trained transformers (GPT)-based automated data mining for building energy management: Advantages, limitations and the future," *Energy and Built Environment*, Feb. 2023, doi: 10.1016/J.ENBENV.2023.06.005.

[26] S. Vishwakarma and B. Chilwal, "Detection and classification of leaf blast disease using decision tree algorithm in rice crop," in *Mathematics and Computer Science*, vol. 2, wiley, 2024, pp. 49–58, doi: 10.1002/9781119896715.ch4.

[27] H. Yan, Y. Ma, and Z. Ru, "Computer-aided medical analysis of smart university student's psychological fitness data using fuzzy clustering and decision tree algorithms," *Computer Aided Design Applications*, vol. 21, no. S9, pp. 277–291, Jan. 2024, doi: 10.14733/cadaps.2024.S9.277-291.

BIOGRAPHIES OF AUTHORS

Inoc Rubio Paucar is bachelor of the systems engineering and computer science career at the Norbert Wiener University. He like to carry out research on software development applying the RUP methodology. He also apply the SCRUM methodology. He has done his pre-professional practices in national and international companies. He is going to specialize in the field of software development with programming languages such as Java, PHP, and JavaScript. He can be contacted by email: Enoc.Rubio06@hotmail.com.

Dr. Laberiano Andrade-Arenas is doctor in systems and computer engineering. master in systems engineering. Graduated from the master's degree in University Teaching. Graduated from the master's degree in accreditation and evaluation of educational quality (UNE) Systems Engineer. ITILV3 Fundamentals International Course (Zonngo-Peru/IMLAD-Mexico). Scrum fundamentals certified, research professor with publications in SCOPUS indexed journals. He has extensive experience in the University Chair in face-to-face and blended classes at different undergraduate and postgraduate universities in Lima. He can be contacted at email: landrade@uch.edu.pe and landradearenas@gmail.com.