Data mining for predictive analysis in gynecology: a focus on cervical health

Laberiano Andrade-Arenas[1], Inoc Rubio-Paucar[2], Cesar Yactayo-Arias[3]
[1]Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú
[2]Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú
[3]Departamento de Estudios Generales, Universidad Continental, Lima, Perú

Article Info

Article history:

Received Nov 3, 2023
Revised Jan 24, 2024
Accepted Jan 25, 2024

Keywords:

Crisp-dm
Data mining
Knowledge discovery in databases
Rapid miner studio
Semma

ABSTRACT

Currently, data mining based on the application of detection of important patterns that allow making decisions according to cervical cancer is a problem that affects women from the age of 24 years and older. For this purpose, the Rapid Miner Studio tool was used for data analysis according to age. To perform this analysis, the knowledge discovery in databases (KDD) methodology was used according to the stages that this methodology follows, such as data selection, data preparation, data mining and evaluation and interpretation. On the other hand, the comparison of methodologies such as the standard intersectoral process for data mining (Crips-dm), KDD and sample, explore, modify, model, evaluate (Semma) is shown, which is separated by dimensions and in each dimension both methodologies are compared. In that sense, a graph was created comparing algorithmic models such as naive Bayes, decision tree, and rule induction. It is concluded that the most outstanding result was -1.424 located in cluster 4 in the attribute result date.

Corresponding Author:

Cesar Yactayo-Arias
Departamento de Estudios Generales, Universidad Continental
Avenida Alfredo Mendiola 5210, Lima, Peru
Email: cyactayo@continental.edu.pe

1. INTRODUCTION

According to research done by the World Health Organization (WHO), cases of uterine cancer in women are currently the fourth most common cancer among women worldwide. For this reason, it is predicted that in 2020 there will be around 576,000 confirmed new cases of breast cancer and 67,000 new cases of uterine cancer [1]. According to certain health sciences experts, uterine cancer is one of the cancer-causing abnormalities of malignant tumors that has an effective prevention when compared to other types of cancers like human papillomavirus (VPH) cancer [2]. But they do demonstrate their presence among women everywhere. According to research conducted on the African continent, this anomaly ranks as the fourth most common fatal form of genotoxic cancer. To be more specific, Malawi has a mortality rate of 51,5/100,000, with an age-specific mortality rate [3]. Additionally, other results show that each year, 528,000 cases of uterine cancer in women are diagnosed globally, with 200,000 cases of them dying from the disease [4]. The cells of the *cuello uterino*, the lower portion connecting the uterus with the vagina, are where the *cuello* uterine cancer (CC) kind of cancer develops. The primary cause of CC is papillomavirus infection, along with several other risk factors like early marriage, early sexual relationships, having multiple sexual partners, and smoking. Beginning without symptoms, it later manifests as vaginal bleeding and pain during sexual relations [5]. This type of disease is treated using mechanisms based on the use of specific procedures such as surgery, q-wave therapy, and radiotherapy. The patient's psychological condition, which results in a state of persistent depression and the person's suffering, is the main issue [6].

The many information technologies have a positive impact on various social spheres, particularly in medicine. An indication of this is the data mining application using the random forest algorithm, which provides results on the diagnosis of uterine cancer well before the illness manifests [7]. The application of citopatolytic characteristics aids in the development of novel concepts based on these principles. This procedure was carried out using techniques for image analysis while taking into consideration a sizable amount of data to create an expert system [8]. Additionally, intelligent systems like big data and the internet of things (IoT) have a significant impact on medical specialties. The Google Cloud Platform (GCP) was utilized to conduct data analysis through classification stages for the use of algorithms for categorizing images related to uterine cancer [9]. The use of advanced learning technologies has helped to produce significantly more accurate results for diagnosing uterine cancer [10]. This technology is focused on a collection of cancer-related cells that can be used to predict through learning algorithms whether a patient has cancerous cells or not [11]. The application of computerized technology also makes predictions about uterine cancer based on trained neural networks that are a component of hybrid grouping techniques. These grouping algorithms are used in image analysis to tackle the issue of detecting and identifying practiced images for image recognition [12].

On the other hand, other illnesses, such as the human papillomavirus (VPH), cause symptoms related to uterine cancer. These infections enable the development of cancerous tumors that rupture the basement membrane and release adenosine diphosphate (ADN) [13]. Although it is true, it has been thought that the VPH has taken on a crucial role in the propensity of those with uterine cancer [14]. In other studies, a vaginal bio adhesive film was used to apply the proper dosage in cases of uterine cancer using third-dimensional (3D) technology [15]. Last but not least, the immunological environment plays a critical role in the development of uterine cancer and endocervical adenocarcinoma [16].

The necessity of using information technology has led to significant discoveries over the years. The use of these technologies gives patients new opportunities for life while maintaining control over their illness. Because of this, the goal of this study is to propose a data mining model that complies with an analysis of cases of uterine cancer in women and the age at which this disease is most likely to manifest itself.

2. LITERATURE REVIEW

In this section, the literature review is divided into two parts. The first point focuses on research papers related to data mining. Likewise, the second point will be written according to the works related to the topic raised. This allows you to have a broader overview of your analysis. It also serves as a fundamental basis for research.

2.1. Theoretical basis

To add value and use the data to make decisions, data mining focuses on methods of analysis carried out on computer systems to identify certain key patterns in a large volume of data [17]. Another definition is that data mining may analyze a large volume of data to uncover key informational trends that can be used to make decisions and achieve desired results [18]. In this sense, data mining employs statistical processes that focus on the analysis of data as variables usable in various informational tools. The automatic machine learning algorithm (ML), which is used in these processes, is one of the algorithms used. This enables the categorization of certain data to identify the most important variables and use statistical significance for their interpretation [19]. The science of data mining has applications in a variety of social fields, including electronic learning, intelligent tutoring systems, text mining, and social network mining, among others. Due to this, the primary task of the data mining application is to make accurate predictions using various types of mathematical algorithms [20].

- Rules of association: the decision-tree algorithms are the fundamental idea of association rules. This algorithm's main capability is to create rules to get a single conclusion. To put it another way, they relate several characteristics to one another to create a rule that must be followed [21].
- Classification algorithms: these classification algorithms carry out methods for automatic learning by the categorization of items. It is based on infinite results from processed combinations [22]. This allows for an adequate classification to continue with the investigation.
- Clustering: specific data are grouped into clusters based on their similarities. For each collection to be distinct between clusters, patterns of similarity are found. However, in some sense, the objects in each cluster are similar [23].

2.2. Related word

In studies, the advancement of medical technology has helped to prevent and quickly identify certain criteria, such as uterine cancer [24]. Through the application of technology, data mining techniques are combined to extract crucial information for studying behavior patterns in images using statistical classification methods [25]. On the other hand, new platforms like the BGISEQ-500 connected to the classification of the database are used in quercetin-based pharmaceutical antitumor treatments for uterine cancer to produce results on the genes that are differentially expressed [26]. Certain medical diagnoses related to the illness have been applied in some cases with an emphasis on crucial information extraction algorithms [27]. The value added is the knowledge that physicians contribute that may have a significant impact on the advancement of medicine by bringing awareness to information technology. Therefore, the goal of this study is to use data mining to the combination of image recognition technologies, extracting the most significant characteristics using this analyses [28].

More and more documented information about the symptoms and progression of the disease is being added to the electronic records of information on electronic media (ERH). However, the main issue is the quantity of information shown that is too noisy [29]. As a result, to predict whether a patient has a condition for which early detection is advantageous for appropriate treatment, it was proposed to group by case 1,321 patients who had the condition to obtain information about the characteristics reported using data mining platforms [30]. The use of medications to treat uterine cancer has started to lose some of its effectiveness as certain levels of control over the illness have been lost. Technology integration has taken on the task of understanding the presented anomalies so that the disease cannot further manifest. Additionally, the metastasis in the lumbosacral ganglia (PLNM) is a prognostic indicator and an independent parameter that guides treatment plans for uterine cancer. The role played by long non-coding RNAs (lncRNA) in the process of the tumors' biological functions is becoming more and more clear. This study's goal was to extract lncRNA linked to metastasis in lymph tissue and explore its potential functional pathophysiological mechanisms in lymph tissue metastasis of uterine cancer [31].

3. METHOD

3.1. Information about the database

It was possible to obtain a certain amount of information from a database to carry out the data analysis based on the proposed topic. Which was located in the Peruvian state's National Open Data Platform. The data obtained serve as input for the investigation.

3.1.1. Open data

Certain tools and techniques by established normative requirements that are met by the public administration bodies make up the framework of data governance that Peru has established. To ensure a fundamental level of acceptance for the collection, processing, and storage of data, it must be implemented in its legal, technological, and strategic context. The supreme decree (D.S.) 157-2021-PCM, stipulates that this information may be used by the public sector for academic research or other purposes.

3.1.2. Data set information

This data source contains information on cases of uterine cancer in the year 2023 at health facilities in the Diris Lima Center. To confirm that the data are protected by the entity providing the information, the information about the extracted database is shown in Table 1. Besides, in Figure 1, it shows a graph of the ages and districts with the most cases of women with cervical cancer in the province of Lima.

Table 1. Data set Information

Camp	Value
Editor	Directorate of Integral Health Networks Lima Centro
Date modified	2023-08-17
Launch date	2023-07-19
Identifier	5c9d58e9-34af-4da6-a411-bcb212200c3a
License	License not specified
Public access level	Public

In Y-axis shows the age ranges of the patients inserted in the database. Which will be evaluated according to the range from 0 to 100. Besides, in axis X**,** the districts with the most cases of cervical cancer in the province of Lima are shown. For this purpose, they are identified by different particular colors.

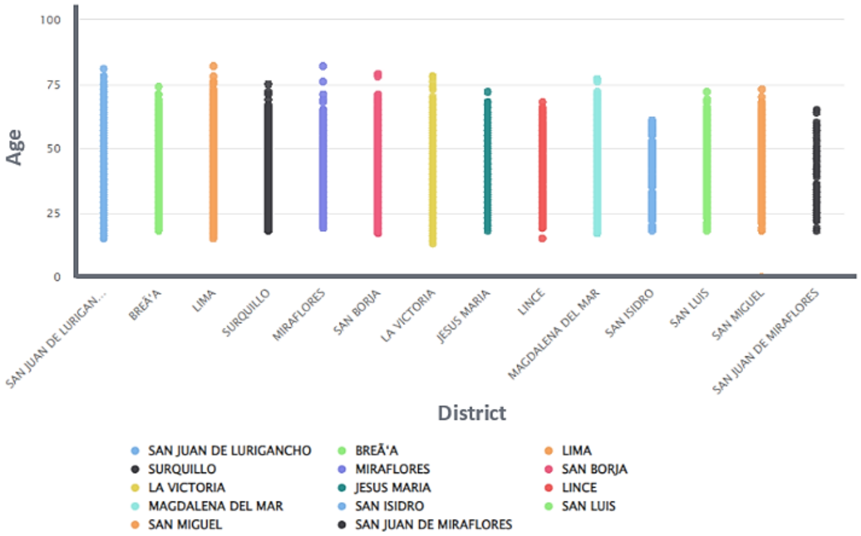


Figure 1. Cases of women with cervical cancer

3.2. Definition of KDD methodology

The KDD methodology is focused on the discovery of knowledge by some defined stages. This refers to an interactive and iterative process. To put it another way, this methodology is adaptable, allowing for easy returns to earlier steps [32]. By applying it to specific tasks in the field of artificial intelligence (AI), the KDD methodology also has a statistical component. The recognition of patrons in a database enables the extraction of valuable information from large quantities of data for the making of decisions that are applied to many fields. The user's use of AI-based techniques must be of utmost importance in this methodology. Because of this, KDD analyzes vast quantities of data, something that a human being is not capable of doing. Figure 2 displays an illustration of the subsequent KDD methodology application stages. It demonstrates how each process interacts to identify key considerations that enable one to make an appropriate decision on the posed challenge.

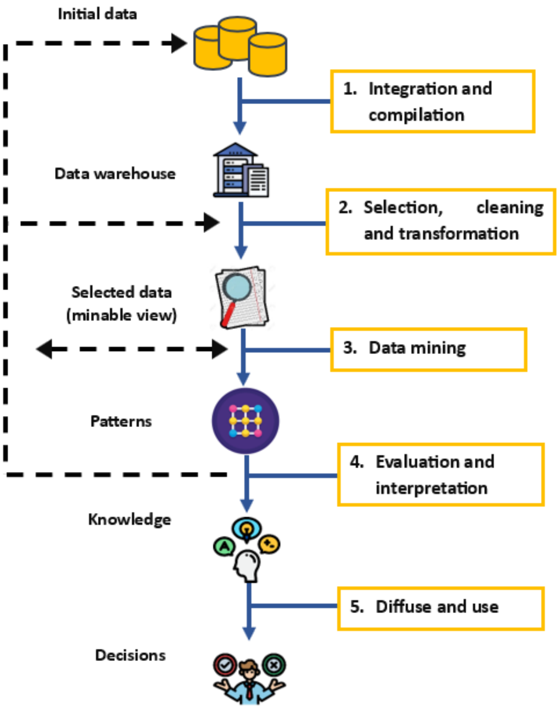

Initial data
1. Integration and compilation
Data warehouse
2. Selection, cleaning and transformation
Selected data (minable view)
3. Data mining
Patterns
4. Evaluation and interpretation
Knowledge
5. Diffuse and use
Decisions

Figure 2. Process steps for KDD

3.3. Development of the KDD methodology

In this section, each step of the KDD methodology is developed using the tool Rapid Miner Studio to produce satisfactory research results. By analyzing each step, it allows you to have a better organization for your orientation in the investigation. In addition, each step will be analyzed in a focused manner.

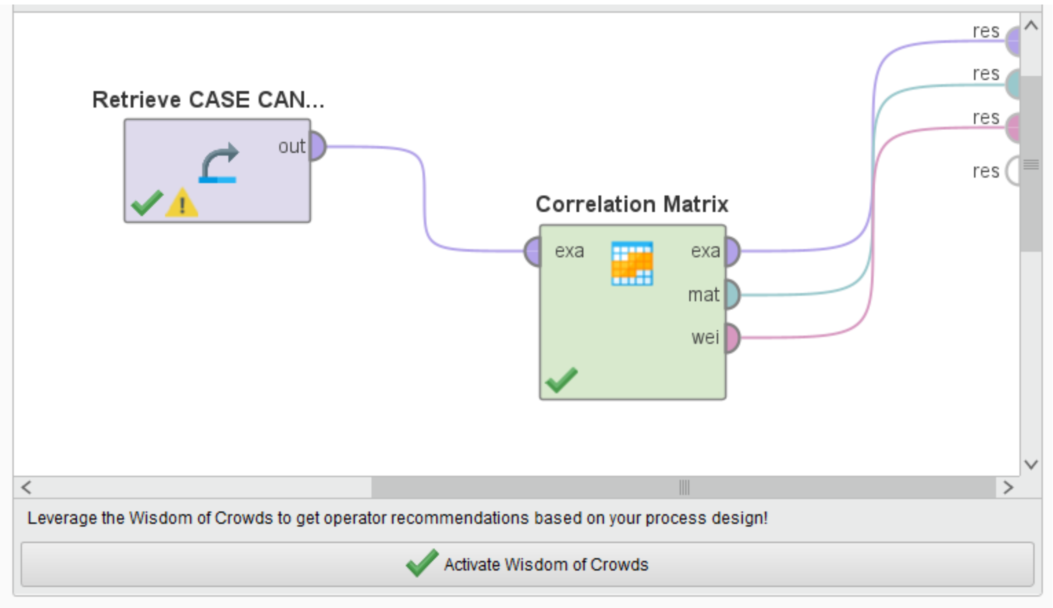
3.3.1. Data selection

In the data selection stage, information is sought from a database of data that is organized according to the proposed topic, in our case, the analysis of women with uterine cancer. For this reason, many data repositories were consulted to choose the necessary information in large quantities. According to the data obtained mentioned in earlier chapters, the purpose of the current article is to explore information on uterine cancer in women from Lima-Per Province. The age field will be used to classify the most concurrent ages at risk for cervical cancer; the district where cervical cancer is prevalent is verified and, the final score, it is determined whether the patient has uterine cancer or not. Which uses two negative or positive options.

A correlation matrix is nothing more than a table of correlation coefficients across several variables. This matrix illustrates how all possible pairs of values in a table can be related to one another, allowing for the retrieval of a large number of data points and the display of significant patterns as shown in (1). On the other hand, Figure 3 shows the graph of what components make up the realization of a correlation matrix in the Rapid Miner Studio tool. Figure 4 shows the results of the variables used comparing their correlation attributes.

$$r = \frac{(n\Sigma XY - \Sigma X\Sigma Y)}{sqrt\left((n\Sigma X^2 - (\Sigma X)\hat{}2)(n\Sigma Y\hat{}2 - (\Sigma Y)\hat{}2)\right)} \quad (1)$$

where r is correlation coefficient, n is number of observations, ΣXY is Sum of product of each pair of corresponding observations of the two variables, ΣX is Sum of the observations of the first variable, ΣY is Sum of the observations of the second variable, $\Sigma X^{\hat{}2}$ is Sum of the squares of the observations of the first variable. $\Sigma Y^{\hat{}2}$ is Sum of the squares of the observations of the second variable.

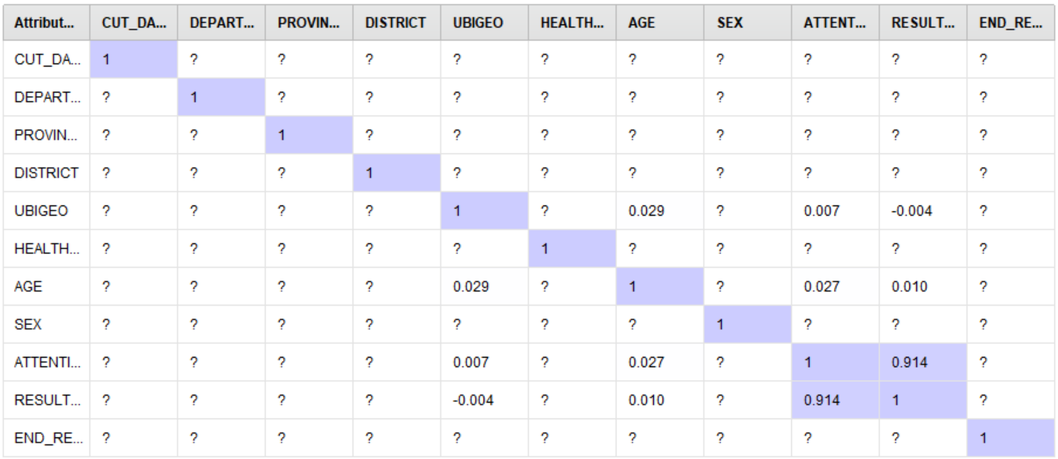


Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 3. Correlation matrix in the Rapid Miner Studio tool

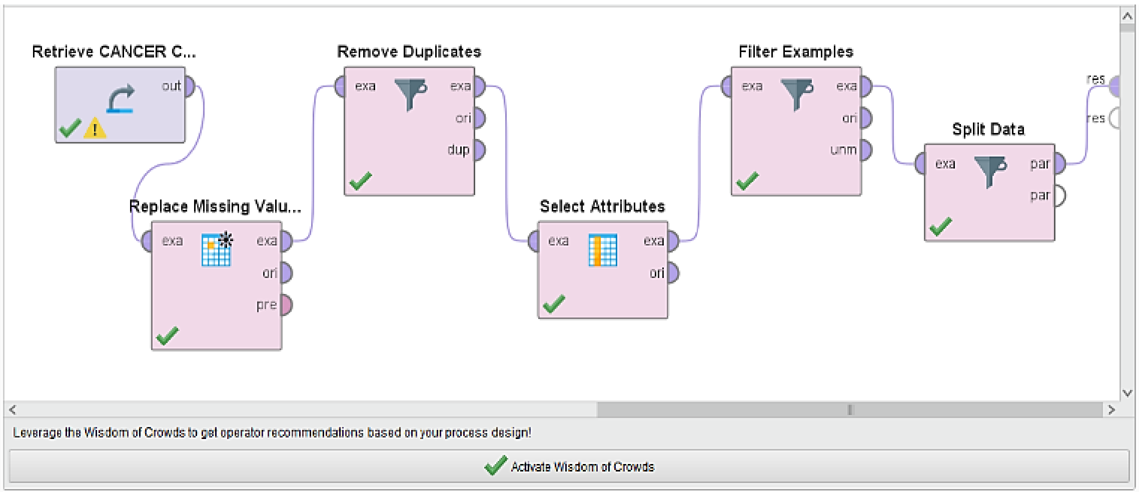
3.3.2. Data pre-processing

The majority of the time during this process, the data imported for analysis contain negative characteristics like missing data, extraneous characters, and empty fields, among others. Due to this, data cleaning is done in this phase, correcting any errors that were discovered. Additionally, pertinent attributes and variables are obtained to make data mining in the analysis easier. Certain components used to carry out an efficient data cleanup and prepare the data for the subsequent process are shown in Figure 5.



Attribut...	CUT_DA...	DEPART...	PROVIN...	DISTRICT	UBIGEO	HEALTH...	AGE	SEX	ATTENT...	RESULT...	END_RE...
CUT_DA...	1	?	?	?	?	?	?	?	?	?	?
DEPART...	?	1	?	?	?	?	?	?	?	?	?
PROVIN...	?	?	1	?	?	?	?	?	?	?	?
DISTRICT	?	?	?	1	?	?	?	?	?	?	?
UBIGEO	?	?	?	?	1	?	0.029	?	0.007	-0.004	?
HEALTH...	?	?	?	?	?	1	?	?	?	?	?
AGE	?	?	?	?	0.029	?	1	?	0.027	0.010	?
SEX	?	?	?	?	?	?	?	1	?	?	?
ATTENTI...	?	?	?	?	0.007	?	0.027	?	1	0.914	?
RESULT...	?	?	?	?	-0.004	?	0.010	?	0.914	1	?
END_RE...	?	?	?	?	?	?	?	?	?	?	1

Figure 4. Correlation of attributes



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 5. Data processing

a. Retrieve cervical cancer

To put data import into action, you need to gather knowledge from several sources of information previously consulted. This involves collecting data from various places and sources for subsequent analysis and application in a specific context. This information gathering is essential for making informed decisions and developing strategies based on the acquired knowledge.

b. Replace missing value

Implementing the elimination of missing values using this operator is crucial for enhancing data analysis. It ensures a cleaner dataset, reducing the risk of errors and inaccurate insights during the analytical process. Proper handling of missing data improves the overall quality and reliability of the results obtained, contributing to more robust and accurate decision-making.

c. Remove duplicates

Eliminating duplicate data from a database is a fundamental step in data management. This process not only prevents conflicts stemming from redundant information but also significantly improves the accuracy and efficiency of data analysis. By ensuring that each data point is unique, you enhance the reliability of the database and avoid any potential issues that may arise from repeated or conflicting records.

d. Select attributes

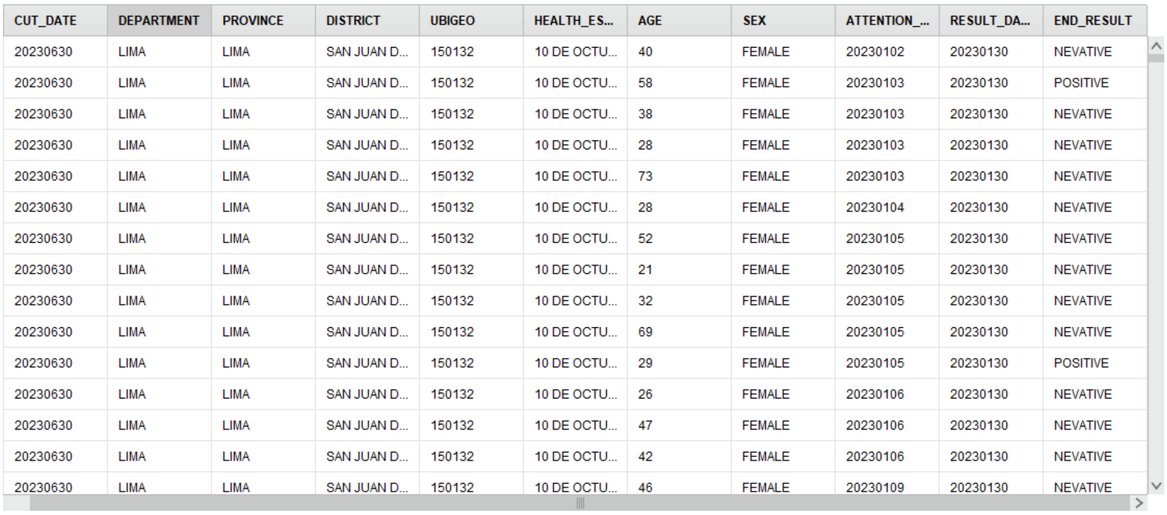
The operator plays a critical role in data analysis by facilitating the extraction of relevant information while generating new features from the existing dataset. This process helps refine and streamline the data, enabling analysts to focus on the most pertinent variables for their specific goals. Identifying and creating these new characteristics enhances the dataset's quality and ultimately contributes to more precise and insightful decision-making processes.

e. Filter examples

The operator serves as a powerful tool for refining datasets by selecting subsets of examples that meet particular criteria, which can involve filtering based on attribute values or other relevant factors. This process is instrumental in enhancing data quality and reducing noise, ensuring that the dataset is tailored to the specific needs of a modeling or analysis task. Additionally, it assists in selecting only the most pertinent attributes, and streamlining the dataset to optimize the performance and accuracy of the model or analysis.

f. Split data

The operation is a vital step in the data preparation process, where the dataset is divided into distinct groups for training and other purposes. This partitioning entails creating subsets with specific relative sizes, which are determined by the specific needs of the analysis or modeling project. It ensures that the data is appropriately allocated for training and testing, allowing for the development and evaluation of models, ultimately leading to more robust and reliable results. The outcome of the data processing is shown in Figure 6, where issues like blank records and additional characters, among others, are resolved.



CUT_DATE	DEPARTMENT	PROVINCE	DISTRICT	UBIGEO	HEALTH_ES...	AGE	SEX	ATTENTION_...	RESULT_DA...	END_RESULT
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	40	FEMALE	20230102	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	58	FEMALE	20230103	20230130	POSITIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	38	FEMALE	20230103	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	28	FEMALE	20230103	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	73	FEMALE	20230103	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	28	FEMALE	20230104	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	52	FEMALE	20230105	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	21	FEMALE	20230105	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	32	FEMALE	20230105	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	69	FEMALE	20230105	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	29	FEMALE	20230105	20230130	POSITIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	26	FEMALE	20230106	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	47	FEMALE	20230106	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	42	FEMALE	20230106	20230130	NEVATIVE
20230630	LIMA	LIMA	SAN JUAN D...	150132	10 DE OCTU...	46	FEMALE	20230109	20230130	NEVATIVE

Figure 6. Data cleansing result

3.3.3. Data mining

The data mining representation is the most important stage in the whole process of the KDD methodology. In this process, certain criteria are established such as the identification of important patterns that will count with representations in dependencies of model types. That is why it is necessary to develop these criteria to select the most appropriate task for the data mining model to be applied; represent the model to be used according to the techniques used in previous stages to represent the knowledge obtained; establish the appropriate algorithm for the development of the model about the selected technique; the K-means algorithm is a kind of unsupervised classifier based on clustering that divides objects into k groups according to their unique characteristics and the clustering is done by reducing the number of groups in each cluster [33]. Which employs quadratic distance in its structure.

The representation of the objects is called d dimensional real vectors $(x_1, x_2, \ldots, x_n)$. The K-means algorithm provides k groups where the sum of distances of the objects within each group $S = \{s_1, s_2 \ldots, s_n\}$ to their centroid is minimized as shown in (2). S belongs to a data set which are elements x_j objects represented by vectors. Each element represents a certain characteristic or attribute. K groups represent clusters with their centroid μ_ι as shown in (3). Components of data cleaning and the use of the k-means algorithm for data grouping in the Rapid Miner Studio tool are shown in Figure 7.

$$\frac{min}{s} E\,(\mu_\iota) = \frac{min}{s} \sum_{i=1}^{n} \sum_{x_j\, \epsilon s_i} ||x_j - \mu_i\,||^2 \qquad (2)$$

$$\frac{\partial E}{\partial \mu_i} = 0 => \mu_i^{(t+1)} = \frac{1}{s_i^{(t)}} \sum_{x_j\, \epsilon s_i^{(t)}} x_j \qquad (3)$$

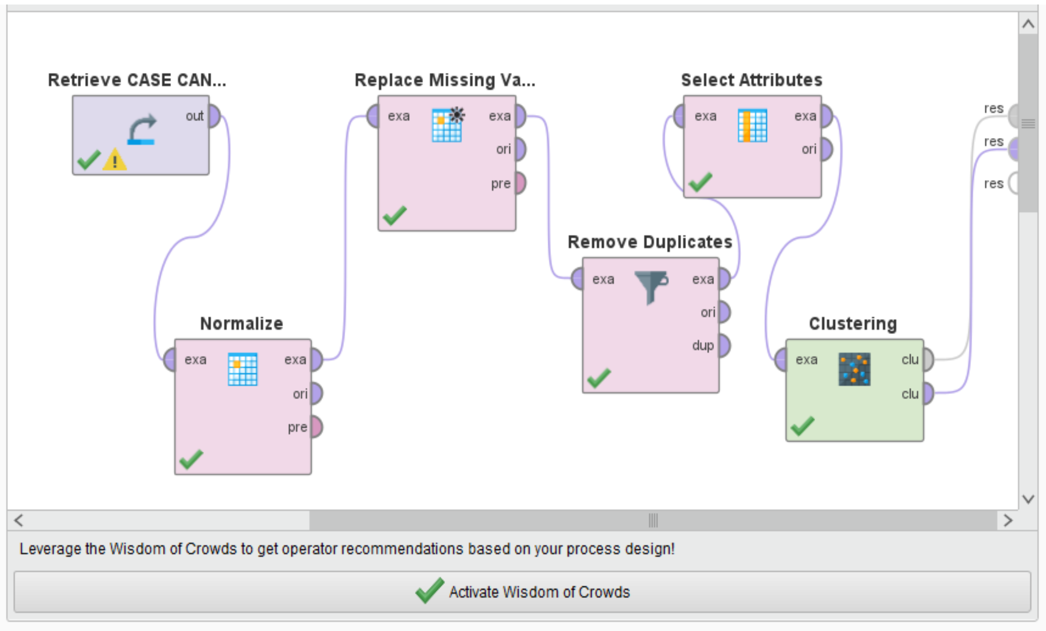


Figure 7. K-means algorithm for data grouping

4. RESULTS

After the data extraction process, it is crucial to assess the value and usefulness of the discovered models and patrons. For this reason, K-means applications are categorized into several groups using the sum of distances according to the characteristics of the attributes set out in the model. In Figure 8, the Y axis represents ages, and the X axis represents clusters. Every cluster is represented by n women, and as a result of the distance between groups, clusters 2 and 1 are the most overrepresented.

The clustering results for each variable chosen in the earlier processes are displayed in Table 2. The attributes used and their corresponding values are shown in Figure 9. in numeric form. Additionally, age-based cluster classifications are checked.

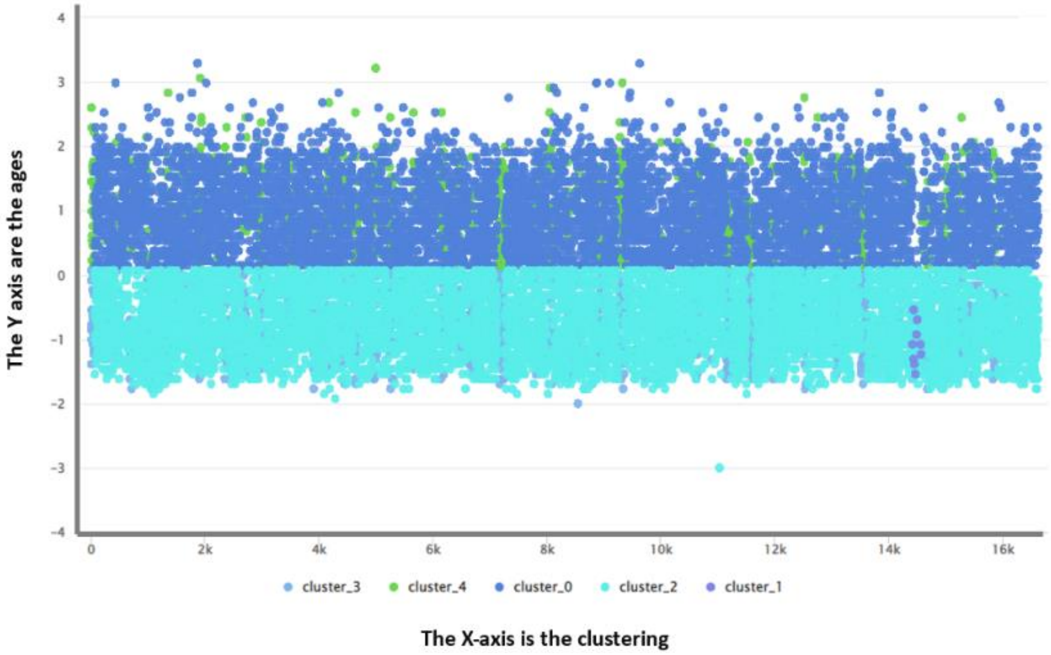
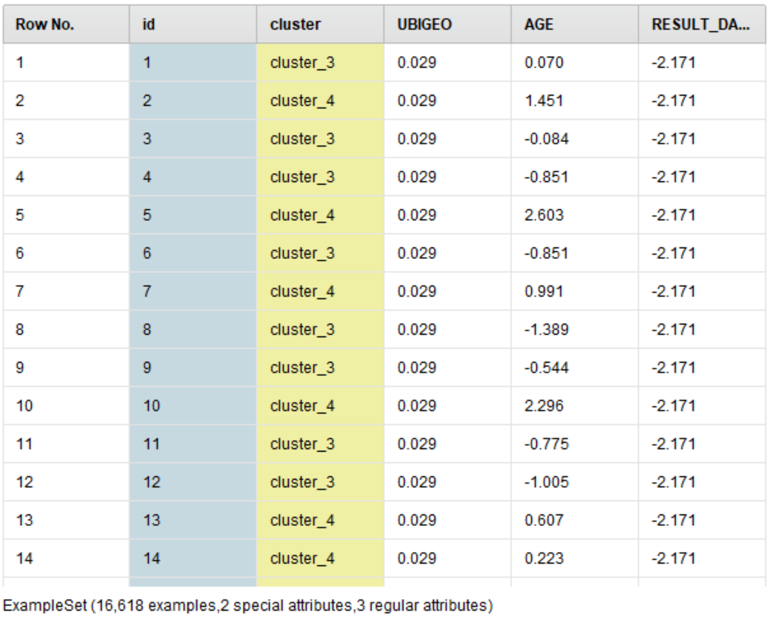


Figure 8. K-means prediction

Table 2. Clustering results

Attribute	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
UBIGEO	0.025	-39.848	0.025	0.025	0.025
AGE	0.988	-1.144	-0.765	-0.750	0.948
RESULT_ DATE	0.279	0.101	0.267	-1.410	-1.424



Row No.	id	cluster	UBIGEO	AGE	RESULT_DA...
1	1	cluster_3	0.029	0.070	-2.171
2	2	cluster_4	0.029	1.451	-2.171
3	3	cluster_3	0.029	-0.084	-2.171
4	4	cluster_3	0.029	-0.851	-2.171
5	5	cluster_4	0.029	2.603	-2.171
6	6	cluster_3	0.029	-0.851	-2.171
7	7	cluster_4	0.029	0.991	-2.171
8	8	cluster_3	0.029	-1.389	-2.171
9	9	cluster_3	0.029	-0.544	-2.171
10	10	cluster_4	0.029	2.296	-2.171
11	11	cluster_3	0.029	-0.775	-2.171
12	12	cluster_3	0.029	-1.005	-2.171
13	13	cluster_4	0.029	0.607	-2.171
14	14	cluster_4	0.029	0.223	-2.171

ExampleSet (16,618 examples,2 special attributes,3 regular attributes)

Figure 9. Clustered clustering database

4.1. Model comparison

This section compares several data mining algorithms in order to assess how effective they are in coming up with data mining solutions. The comparison of models like naive Bayes, decision trees, and rule induction is shown in Figure 10. The which rule induction algorithm outperforms the others with a score of 1, while the other algorithms perform worse.

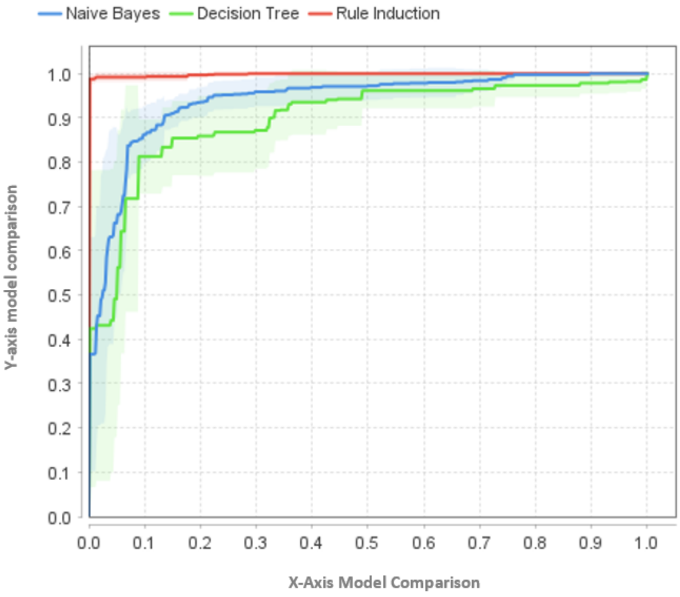


Figure 10. Model comparison

4.2. Comparison of methodologies

This section examines the most prevalent data mining methodologies. To do this, three different methods have been chosen, each with its own unique characteristics that may be compared. A methodology enables the development of procedures that include tasks, steps, and specific organizing methodologies. A comparison of several data mining methodologies is shown in Table 3. In this sense, KDD was chosen as the methodology because it suited the proposed project.

Table 3. Comparison of methodologies

Comparison of attributes	Methodology KDD	Methodology CRIPS-DM	Methodology SEMMA
Structure and sequence	Its structure is less rigid than that of other methodologies, which also include data selection, cleaning, transformation, extraction, and assessment and application of knowledges [34].	The six steps that make up this methodology are understanding the business, understanding data, preparing data, modeling, evaluating, and implementing [35].	Semma's methodology consists of five steps: demonstration, exploration, modification, modeling, and evaluation [36].
Business orientation	Recognizes the significance of the company's commercial goals and seeks to gain knowledge to gain a competitive advantage.	Understands the commercial goals from the outset and ensures that the data is processable and useful for decision-making.	Learn how to do information analysis while taking the company's goals and use of the results into consideration.
Flexibility	Provides a framework for the widespread discovery of knowledge through a broad, less structured focus.	It can be expanded for commercial use and is adaptable to a variety of contexts and projects.	Although it follows a predetermined sequence of steps, it is flexible enough to accommodate various projects.
Interaction	Encourage repetition. However, it lacks a clearly defined structure like the Semma or Crisp-DM methodologies do.	Learn how to use an iterative method for results review. It adapts to projects that are always changing.	It is a process that can be repeated in stages as necessary. Modifications may be made during the process.

5. DISCUSSION

The goal of data mining is to add value and use the data to inform decisions through the discovery of significant patterns in a large amount of data [17]. According to another definition, data mining can also analyze the massive amount of data that is obtained and reveal key informational patterns that are necessary for decision-making and achieving desired results [18]. On the other hand, data mining employs statistical methods that concentrate on data analysis as variables that can be used in a variety of informational tools. These procedures make use of machine learning (ML). This makes it possible to group certain data in order to identify the most crucial factors and then interpret those results using statistical significance [19]. However, the sciences of data mining are used in a variety of social fields, including electronic learning, intelligent tutoring systems, text mining, and social network mining, among others. Due to this, the primary goal of the data mining application is to make accurate predictions using various types of mathematical algorithms [20]. On the other hand, research has shown that advancements in medical technology have made it possible to prevent and identify uterine cancer early [24]. Through the combination of technologies for data mining, it is possible to obtain information that is essential for studying patterns of behavior in images using statistical classification methods [25].

6. CONCLUSION

In conclusion, the prevalence of women with cervical cancer has been increasing in recent years. For this, the application of data mining using the Rapid Miner Studio tool allows obtaining results according to the grouping proposed as clusters to classify them according to the variable age that has been selected. In this sense, several numerical variables were compared such as: *ubigeo*, *age* and *date_result*. The selected attributes show results located in attributes with their corresponding cluster number. The model shows the following: Cluster 0 (6,093 items), Cluster 1 (11 items), Cluster 2 (7,782 items), Cluster 3 (1,525 items) and Cluster 4 (1,207 items) which has a number of items of 16,618. Also, the KDD methodology was applied to perform data mining oriented to each process established in the methodology that fits the research topic. Finally, the comparison of certain described methodologies with their dimensions and characteristics of each one is performed. Finally, the comparison of algorithmic models such as: naive Bayes, decision tree and rule introduction, which is offered in the Rapid Miner Studio tool, was applied. To conclude, it is recommended to extend the research based on data mining by applying the algorithms that exist today for the analysis of large amounts of data. Also, the application of big data involves obtaining large amounts of data obtained in different business processes. For this, the appropriate algorithm should be used by extending the knowledge in machine learning based on supervised learning and unsupervised learning.

ACKNOWLEDGEMENTS

Thank you for the constant support of the Universidad de Ciencias y Humanidades, which through its authorities provides support so that the research materializes.

REFERENCES

[1] A. Znaor *et al.*, "Breast and cervical cancer screening practices in nine countries of Eastern Europe and Central Asia: a population-based survey," *Journal of Cancer Policy*, vol. 38, Dec. 2023, doi: 10.1016/j.jcpo.2023.100436.

[2] B. Zhang *et al.*, "Knowledge, willingness, uptake and barriers of cervical cancer screening services among Chinese adult females: a national cross-sectional survey based on a large e-commerce platform," *BMC Women's Health*, vol. 23, no. 1, Aug. 2023, doi: 10.1186/s12905-023-02554-2.

[3] R. Zou *et al.*, "Effects of metalloprotease ADAMTS12 on cervical cancer cell phenotype and its potential mechanism," *Discover Oncology*, vol. 14, no. 1, Aug. 2023, doi: 10.1007/s12672-023-00776-2.

[4] S. Dadipoor, A. Alavi, Z. Kader, S. Mohseni, H. Eshaghi Sani Kakhaki, and N. Shahabi, "Predictive power of PEN-3 cultural model in cervical cancer screening among women: a cross-sectional study in South of Iran," *BMC Cancer*, vol. 23, no. 1, Aug. 2023, doi: 10.1186/s12885-023-11240-3.

[5] K. Dhakal, P. Wang, J. F. Mboineki, M. A. Getu, and C. Chen, "Assessment of supportive care needs among cervical cancer patients under treatment in Nepal: a cross-sectional study," *BMC Women's Health*, vol. 23, no. 1, Aug. 2023, doi: 10.1186/s12905-023-02484-z.

[6] S. Salta, J. Lobo, B. Magalhães, R. Henrique, and C. Jerónimo, "DNA methylation as a triage marker for colposcopy referral in HPV-based cervical cancer screening: a systematic review and meta-analysis," *Clinical Epigenetics*, vol. 15, no. 1, Aug. 2023, doi: 10.1186/s13148-023-01537-2.

[7] R. Weegar and K. Sundström, "Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations," *PLoS ONE*, vol. 15, Aug. 2020, doi: 10.1371/journal.pone.0237911.

[8] S. W. K. Chan, K. S. Leung, and W. S. F. Wong, "An expert system for the detection of cervical cancer cells using knowledge-based image analyzer," *Artificial Intelligence in Medicine*, vol. 8, no. 1, pp. 67–90, Feb. 1996, doi: 10.1016/0933-3657(95)00021-6.

[9] P. Subashini, T. T. Dhivyaprabha, M. Krishnaveni, and M. B. J. Susan, "Smart intelligent system for cervix cancer image classification using google cloud platform," in *Enabling Technologies for Effective Planning and Management in Sustainable Smart Cities*, Springer International Publishing, 2023, pp. 245–281.

[10] D. Sambyal and A. Sarwar, "Recent developments in cervical cancer diagnosis using deep learning on whole slide images: an Overview of models, techniques, challenges and future directions," *Micron*, vol. 173, Oct. 2023, doi: 10.1016/j.micron.2023.103520.

[11] R. V. Ulia, Suryati, and A. Santoni, "Cytotoxic potential of essential oil isolated from Semambu (*Clibadium surinamese L*) leaves against T47D breast and hela cervical cancer cells," *Molekul*, vol. 18, no. 2, pp. 289–299, Jul. 2023, doi: 10.20884/1.jm.2023.18.2.7816.

[12] L. Mishra, R. Dasgupta, Y. S. Chowdhury, S. Nanda, and S. Nanda, "Cervical cancer detection using hybrid pooling-based convolutional neural network approach," *Indian Journal of Gynecologic Oncology*, vol. 21, no. 2, Apr. 2023, doi: 10.1007/s40944-023-00712-w.

[13] S. J. Bryan *et al.*, "Circulating HPV DNA as a biomarker for pre-invasive and early invasive cervical cancer: a feasibility study," *Cancers*, vol. 15, no. 9, May 2023, doi: 10.3390/cancers15092590.

[14] R. Montero-Macías *et al.*, "TRANSLACOL project: Nodal human papillomavirus tumoral DNA detection by ddPCR for survival prediction in early cervical cancer patients without pelvic lymph node invasion," *Journal of Clinical Virology*, vol. 161, Apr. 2023, doi: 10.1016/j.jcv.2023.105418.

[15] A. Almotairy *et al.*, "Disulfiram 3D printed film produced via hot-melt extrusion techniques as a potential anticervical cancer candidate," *International Journal of Pharmaceutics*, vol. 635, Mar. 2023, doi: 10.1016/j.ijpharm.2023.122709.

[16] Q. Gan *et al.*, "Prognostic value and immune infiltration o of HPV-related genes in the immune microenvironment of cervical squamous cell carcinoma and Endocervical Adenocarcinoma," *Cancers*, vol. 15, no. 5, Feb. 2023, doi: 10.3390/cancers15051419.

[17] 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*, vol. 5, no. 1, pp. 143–169, Feb. 2024, doi: 10.1016/j.enbenv.2023.06.005.

[18] K. Kaygisiz *et al.*, "Data-mining unveils structure–property–activity correlation of viral infectivity enhancing self-assembling peptides," *Nature Communications*, vol. 14, no. 1, Aug. 2023, doi: 10.1038/s41467-023-40663-6.

[19] K. R. Flores, L. V. F. M. de Carvalho, B. J. Reading, A. Fahrenholz, P. R. Ferket, and J. L. Grimes, "Machine learning and data mining methodology to predict nominal and numeric performance body weight values using large white male turkey datasets," *Journal of Applied Poultry Research*, vol. 32, no. 4, Dec. 2023, doi: 10.1016/j.japr.2023.100366.

[20] W. E. Brown and E. Moreno-Centeno, "A data mining transmission switching heuristic for post-contingency AC power flow violation reduction in real-world, large-scale systems," *Computers and Operations Research*, vol. 160, Dec. 2023, doi: 10.1016/j.cor.2023.106391.

[21] J. Lyu, Z. Cao, and E. Song, "A data association algorithm for the robust confidence ellipsoid filter," *Signal Processing*, vol. 213, Dec. 2023, doi: 10.1016/j.sigpro.2023.109201.

[22] S. Hong, K. Kim, and S. H. Lee, "A hybrid jamming detection algorithm for wireless communications: simultaneous classification of known attacks and detection of unknown attacks," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1769–1773, Jul. 2023, doi: 10.1109/LCOMM.2023.3275694.

[23] Y. Zhang, X. Li, L. Wang, S. Fan, L. Zhu, and S. Jiang, "An autocorrelation incremental fuzzy clustering framework based on dynamic conditional scoring model," *Information Sciences*, vol. 648, Nov. 2023, doi: 10.1016/j.ins.2023.119567.

[24] L. Zhang *et al.*, "Intelligent diagnosis of cervical cancer based on data mining algorithm," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–9, Nov. 2021, doi: 10.1155/2021/7690902.

[25] N. Mei, H. Chen, N. Zhao, Y. Yi, and C. Li, "A comprehensive pan-cancer analysis of RBM8A based on data mining," *Journal of Oncology*, vol. 2021, pp. 1–16, Jul. 2021, doi: 10.1155/2021/9983354.

[26] Y. Li, J. Kou, T. Wu, P. Zheng, and X. Chao, "Screening of therapeutic candidate genes of quercetin for cervical cancer and analysis of their regulatory network," *OncoTargets and Therapy*, vol. 14, pp. 857–866, Feb. 2021, doi: 10.2147/OTT.S287633.

[27] S. D. Annapurna *et al.*, "Identification of differentially expressed genes in cervical cancer patients by comparative transcriptome analysis," *BioMed Research International*, vol. 2021, pp. 1–13, Mar. 2021, doi: 10.1155/2021/8810074.

[28] V. Garcia-Rios, M. Marres-Salhuana, F. Sierra-Liñan, and M. Cabanillas-Carbonell, "Predictive machine learning applying cross industry standard process for data mining for the diagnosis of diabetes mellitus type 2," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, pp. 1713–1726, Dec. 2023, doi: 10.11591/ijai.v12.i4.pp1713-1726.

[29] H. Luo *et al.*, "Genome-wide somatic copy number alteration analysis and database construction for cervical cancer," *Molecular Genetics and Genomics*, vol. 295, no. 3, pp. 765–773, Jan. 2020, doi: 10.1007/s00438-019-01636-x.

[30] R. V. Angadi, J. A. Mangai, V. J. Manohar, S. B. Daram, and P. V. Rao, "An ensemble based data mining model for contingency analysis of power system under STLO," *International Journal of Applied Power Engineering (IJAPE)*, vol. 12, no. 4, pp. 349–358, Dec. 2023, doi: 10.11591/ijape.v12.i4.pp349-358.

[31] C. Shang *et al.*, "Characterization of long non-coding RNA expression profiles in lymph node metastasis of early-stage cervical cancer," *Oncology Reports*, vol. 35, no. 6, pp. 3185–3197, Mar. 2016, doi: 10.3892/or.2016.4715.

[32] L. Aguagallo, F. Salazar-Fierro, J. García-Santillán, M. Posso-Yépez, P. Landeta-López, and I. García-Santillán, "Analysis of student performance applying data mining techniques in a virtual learning environment," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 11, pp. 175–195, Jun. 2023, doi: 10.3991/ijet.v18i11.37309.

[33] A. A. Abdulnassar and L. R. Nair, "Performance analysis of K-means with modified initial centroid selection algorithms and developed Kmeans9+ model," *Measurement: Sensors*, vol. 25, Feb. 2023, doi: 10.1016/j.measen.2023.100666.

[34] A. H. Azizan *et al.*, "A machine learning approach for improving the performance of network intrusion detection systems," *Annals of Emerging Technologies in Computing*, vol. 5, no. 5, pp. 201–208, Mar. 2021, doi: 10.33166/AETiC.2021.05.025.

[35] J. Bokrantz, M. Subramaniyan, and A. Skoogh, "Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM," *Production Planning and Control*, pp. 1–21, Jul. 2023, doi: 10.1080/09537287.2023.2234882.

[36] S. López-Torres *et al.*, "IoT monitoring of water consumption for irrigation systems using SEMMA methodology," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11886, Springer International Publishing, 2020, pp. 222–234.

BIOGRAPHIES OF AUTHORS

Laberiano Andrade-Arenas doctor in systems and computer engineering. Master's in systems engineering. Graduated with a master's degree in University Teaching. Graduated with a master's degree in accreditation and evaluation of educational quality. Systems engineer scrum fundamentals certified, a research professor with publications in Scopus-indexed journals. He can be contacted at email: landrade@uch.edu.pe.

Inoc Rubio-Paucar bachelor's in systems and computer engineering. He has a background in database management and computer system design, with a focus on artificial intelligence applications, machine learning, and data science. His research interests are in the area of computer science. He can be contacted at email: Enoc.Rubio06@hotmail.com.

César Yactayo-Arias obtained a bachelor's degree in administration from Universidad Inca Garcilazo de la Vega and a master's degree in education from Universidad Nacional de Educación Enrique Guzmán y Valle, he is a doctoral candidate in administration at Universidad Nacional Federico Villarreal. Since 2016 he has been teaching administration and mathematics subjects at the Universidad de Ciencias y Humanidades and since 2021 at the Universidad Continental. Currently, he also works as an administrator of educational services at the higher level, he is the author and co-author of several refereed articles in journals, and his research focuses on TIC applications to education, as well as management using computer science and the internet. He can be contacted at email: yactayocesar@gmail.com.