

A study of Tobacco use and mortality by data mining

Laberiano Andrade-Arenas¹, Inoc Rubio Paucar², Cesar Yactayo-Arias³

¹Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú

²Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú

³Departamento de Estudios Generales, Universidad Continental, Lima, Perú

Article Info

Article history:

Received Feb 27, 2024

Revised Jul 11, 2024

Accepted Jul 17, 2024

Keywords:

A priori

Data mining

Knowledge discovery in data

Rules of association

Tobacco

ABSTRACT

The use of data mining to address the issue of people who consume tobacco and other harmful substances for their health has led to a significant dependence among smokers, which over time causes illnesses that may result in the addict's death. As a result, the research's goal is to apply a data mining study whose findings showed that the confidence intervals are less than 0.355. However, the lift and conviction in the last three rules are also lower, making it unlikely that these rules will be followed. On the other hand, the knowledge discovery in data bases method was used. It consists of the following stages: data selection, preparation, data mining, and evaluation and interpretation of the results. To that end, comparisons of agile data mining methodologies like crisp-dm, knowledge discovery in data, and Semma are also done. As a result, using specific criteria, dimensions are segmented to allow for the differentiation of these methodologies. As a result, a comparison graph of models such as naive Bayes, decision trees, and rule induction is used. To sum up, it can be said that the rules of association apply to men, the number of admissions, and the cancers that can be brought on by smoking. Also, the percentage of male patients admitted with cancers that can be brought on by smoking Last but not least, the number of admissions and cancers that can be brought on by smoking

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Cesar Yactayo-Arias

Departamento de Estudios Generales, Universidad Continental

Lima, Perú

Email: cyactayo@continental.edu.pe

1. INTRODUCTION

People who are dependent on this substance have a progressive problem as a result of using tobacco. Because nicotine is their active component, smokers frequently suffer from a variety of diseases due to their consumption or abstinence from nicotine. According to a study conducted by the World Health Organization (WHO), 22.3% of the world's population is expected to use tobacco in 2020. To be precise, among men, 36.7% are addicted to tobacco, and among women, 7.8% are addicted to this substance [1].

One of the leading causes of death in recent years is tobacco. This implies that certain diseases manifest themselves in accordance with excessive nicotine consumption [2]. Tobacco consumption has the ability to cause diseases that result in death, though. But the truth is that countries that have used methods to curb this disease have helped to achieve positive results [3]. The younger generation is being increasingly exposed to the consumption of harmful substances like tobacco [4]. The use of tobacco during adolescence acts as a precursor to the adaptation of other brain-harming substances in the future [5]. Long-term effects result in a decline in the person's cognitive abilities and changes to their level of agitation [6]. Studies were conducted on the following symptoms, such as psychological issues, conduct issues, and neurological issues, regarding nicotinic intake and the use of exogenous progesterone for treating the disease [7]. Cancer-related

diseases are caused by more than just excessive tobacco use. Patients with cancer are also impacted by a temporary tobacco dependence. Therefore, the link between smoking tobacco and cancer affects the patient psychologically, leading to undesirable reactions in the person's behavior [8].

Smoking-related illnesses have played a role in the development of diseases, including lung cancer. osteoarthritis (OA), osteoporosis (OP), and rheumatoid arthritis (AR) have all been linked to the development of cadmium (CD), which is found in tobacco through humus [9]. On the other hand, this infection causes a serious increase in CD levels in the blood, which causes oxidative stress and results in a loss of cholesterol [10]. The rise in smokers has led to the development of serious health issues like lung cancer. Early disease detection offers the potential to control the condition and save many people's lives. Certain treatments prescribed by oncology specialists have been developed thanks to technologies like computerized radiology [11]. The removal of lung nodules and other surgical procedures have prevented the growth of cancerous cells brought on by tobacco use while establishing health precautions for their treatment [12].

In one study, data mining with classification algorithms was used with a sample of 22,000 students between the ages of 14 and 19. There were signs there that mentioned teenagers smoking tobacco [13]. On the other hand, a prototype that predicts the high risk of tobacco smokers compared to a group of non-smokers was used. Using automatic learning algorithms, 318 people were subjected to this prediction according to its symptoms [14]. The use of specialized systems managed with the knowledge of a data base known as San Factor implemented for the prevention of quitting smoking has helped decision-makers adopt specific medical treatments [15]. However, giving up smoking requires a significant psychological effort for someone addicted to another substance, especially if they have been diagnosed with cancer. To achieve this goal, the use of deep learning models enables the processing of large amounts of data, finding crucial characteristics like the cessation of smoking habits [16]. Another study's use of algorithms like naive Bayes and Random Forest implemented using the WEKA tool to analyze the abandonment of smokers resulted in the provision of 20 distinct attributes to predict the number of people who will give up this habit [17].

The research is aimed at contributing new knowledge or perspectives to the specific field of study, addressing gaps in the existing literature, solving practical or complex problems, improving the understanding of specific phenomena. The data mining aims to make decisions in accordance with some predictions by employing algorithms that allow for societally acceptable outcomes. As a result, the investigation's goal is to conduct a study on people who are addicted to tobacco use using data mining.

2. LITERATURE REVIEW

This section proposes two very interesting segments. The first focuses on research projects dealing with the topic of tobacco consumption while utilizing specific technologies like data mining. Similarly, the second point is related to the research that is done on the raised topic.

2.1. Theoretical basis

Data mining is used to find patterns in behavior among a large amount of data by extracting them as variables and applying specific algorithms developed for each activity [18]. On the one hand, a clear example is the automatic learning application, which in some studies aims to identify and compile crucial characteristics of a particular subject using mathematical algorithms [19]. To that purpose, the application of rule-based analysis takes into account how the data analysis algorithms work. However, making decisions is based on the outcomes of accurate predictions made with technologies that provide confidence in their outcomes [20]. The algorithms used in data mining include classification, grouping, and association rules-based algorithms, among others, to predict specific problems in various social spheres. Each algorithm has unique characteristics depending on how it is used in specific artificial intelligence (AI)-related areas [21]. In reference to the discussion of data mining, many tools allow for the creation of models that are suitable for each unique situation. Organizations that manage large amounts of data are able to make decisions in a more specific manner thanks to the application of certain algorithms [22].

2.1.1. Classification algorithms

The classification algorithms in data mining create models that attempt to pre-classify data related to decision trees or neural networks. For this reason, the system must carry out a learning process by analyzing data through training. These trained algorithms enable the determination of specific sets of prescribed parameters through classification rules [23].

2.1.2. Clustering

The clustering topic allows for the acquisition of untagged grouping data in order to carry out the construction of groupings of data known as clusters. In this sense, each cluster is a consolidated graph made

up of a group of items or data that, upon analysis, are shown to be comparable to one another. But at the same time, they obtain elements that are different from other items and data sets [24].

2.1.3. Neural networks

In neural network algorithms, it is understood as a collection of interconnected entry and exit units. The neural network learns to adjust certain weights to predict the class tuple labels during model learning. For this reason, neural networks can extract meaning from complex or printed data used to extract patterns and identify crucial trends for decision-making [25].

2.1.4. Decision trees

A decision tree algorithm is a non-parametric learning algorithm that is employed for both classification and regression tasks. It has a jerry-rigged tree structure made up of a central node, branches, internal nodes, and hoary nodes. A decision-making tree has an open node without entwined branches. The internal nodes, or decision nodes, are fed by the radial branches of the primary node. Both types of nodes do evaluations to create homogeneous subsets. To put it another way, depending on the available characteristics, they are indicated by nodes hoja or nodes terminals. All potential results from the collection of data are shown as hoary nodes [26].

2.2. Related work

Various issues have been brought on by the addictive smoking of cigars and their constituent parts. In this sense, those who overindulge become dependent on these substances because they like feeling relaxed and fighting the psychological stress associated with smoking [27]. As a result, it was proposed in research to classify smokers using the technique of data mining, categorizing them as light or heavy smokers, and applying the knowledge discovery in data methodology (KDD) to certain processes. On the other hand, in a different study, data mining will be used to predict the market demand for cigars and the individuals who purchase this product [28]. One of the key contributors to the early diagnosis of cardiovascular diseases is excessive alcohol and tobacco use. This is why the use of data mining involves the detection of factors linked to some symptoms, like high blood sugar levels, high cholesterol, and high artery pressure. In this context, medical data are analyzed using data mining techniques based on automatic learning [29]. Finally, the use of these mathematical algorithms enables the acquisition of favorable results in the person's behavior to counteract excessive tobacco consumption. Making decisions on meditation and the likelihood of quitting smoking using data mining tools will be possible thanks to the development of models for hybrid predictive analysis [30].

3. METHOD

3.1. Definition of the KDD methodology

The KDD methodology is a process that extracts knowledge based on a large volume of data samples. Because of this, iterative, practical methods are used whenever it is necessary to obtain this information. Its main quality is adaptability to the completed project [31]. In other words, it allows returning to earlier phases that were developed for updating later phases. Several steps make up the KDD methodology's data analysis process. To that end, it is necessary to extract patrons, which are rules or functions that enable decision-making as shown in Figure 1. The KDD is justified in the study of tobacco use and mortality due to its ability to address the complexity of data, identify hidden patterns, and significant relationships in large datasets. This rigorous methodological approach ensures validation and reproducibility of results, critical aspects in public health research and risk factors such as tobacco use. Furthermore, its previous applicability in medical research supports its utility in extracting useful insights that can inform health policies and prevention strategies.

3.2. Development of the KDD methodology

In this section, we delve into the development of the KDD methodology, leveraging the powerful capabilities of RapidMiner Studio. By harnessing this robust tool and leveraging the data gathered from the specified topic, we apply a set of predetermined criteria to guide our analytical process. With RapidMiner Studio's intuitive interface and comprehensive suite of data mining functionalities, we embark on a systematic journey through the various stages of the KDD process. From data selection and preprocessing to modeling and evaluation, each step is meticulously executed to uncover hidden patterns, trends, and insights within the dataset.

3.2.1. Data selection

A database of data about people who are tobacco addicts was obtained during the data selection process. Which provides the diagnoses as a consequence of medical examinations performed by subject-matter experts. The data cleaning process is then carried out after making sure that there are no blank spaces or noise in the records. A question mark is displayed next to extraneous data in the records, as shown in Table 1. This could cause issues when it comes time to make the prediction. For this reason, data cleaning is done to reduce noise.

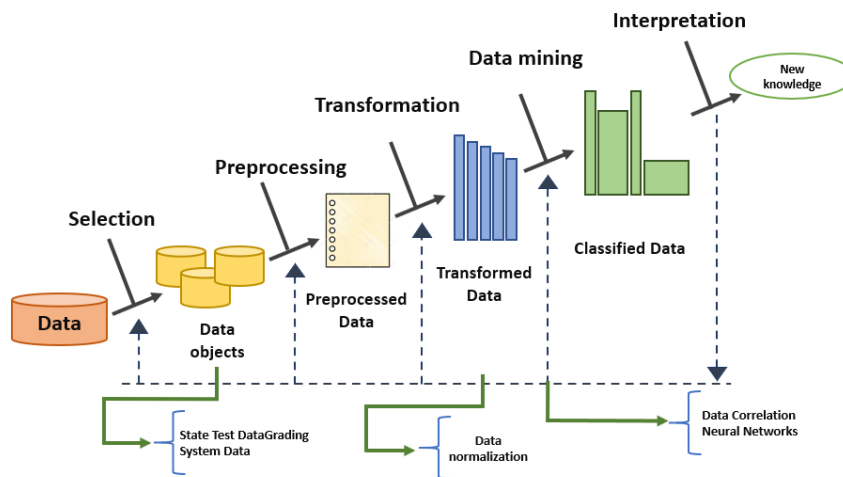


Figure 1. KDD methodology

Table 1. Database with noise

Year	ICD10 Code	ICD10diagnosis	Diagnosis	Metric	Sex
20/14/15	I70	Atherosclerosis	Circulatory diseases which can be caused by smoking	Attributable number	Male
20/14/15	K25-K27	Stomach/Duodenal Ulcer	Digestive diseases which can be caused by smoking	Attributable number	Male
20/14/15	K50	Crohn’s disease	Digestive diseases which can be caused by smoking	Attributable number	Male
20/14/15	K05	Periodontal disease/Periodontitis	Digestive diseases which can be caused by smoking	Attributable number	Male
20/14/15	H25	Age related Cataract 45+	Other diseases which can be caused by smoking	Attributable number	Male
20/14/15	S72.0-S72.2	Hip Fracture 55+	Other diseases which can be caused by smoking	Attributable number	Male
20/14/15	O. O3	Spontaneous Abortion	Other diseases which can be caused by smoking	Attributable number	Male

One of the main advantages of using RapidMiner Studio is its intuitive and easy-to-use interface, which allows users to perform complex data analysis without prior programming experience. Figure 2 shows an example of this data selection process, where different components are used to process the data and select the relevant fields for analysis. The first stage in the data selection process usually involves importing data sets from various sources, such as local files, databases or even cloud services. Once the data is loaded into RapidMiner Studio, users can use a variety of components to preprocess the data, such as data cleansing, normalization and variable transformation.

a. Retrieve admissions

This operator is used to load data into RapidMiner Studio. It is crucial that there is no import limit on the number of records. This operator ensures that all data is processed without restrictions. In this way, the adequate number of records is guaranteed.

b. Filter examples

Enables the selection of the most pertinent data for the analysis. In other words, choose the most crucial fields to prepare it for analysis. This ensures that the data is ready for accurate and efficient processing. In this way the filter is made to work with the required fields.

c. Select attributes

This operator extracts useless data by generating new characteristics from the existing data. For this reason, certain subsets of attributes or characteristics from a collection of data are chosen for further analysis. This process enhances the quality of data used in the analysis.

d. Detect outlier

Depending on the distance between his k closest neighbors. This operator discovers exceptional values in the collection of examples provided. The parameters can be used to determine the variables n and k .

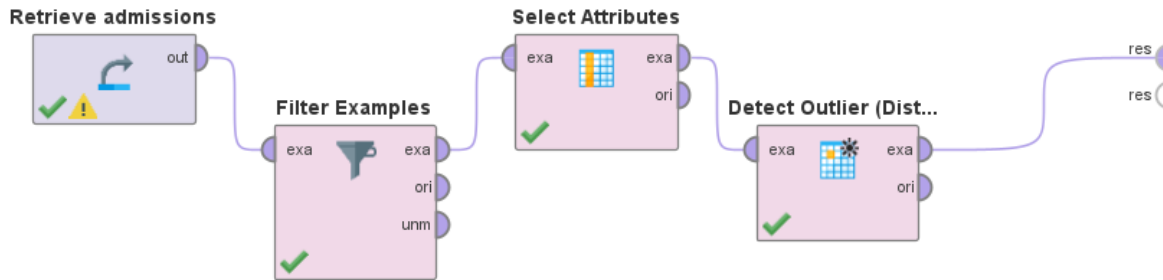


Figure 2. Data selection process

3.2.2. Data preprocessing

Most frequently, during this process, the data imported for analysis contains negative characteristics, including missing data, excess characters, and empty fields. As a result, data cleanup is done in this phase, fixing any errors that were discovered. Additionally, pertinent characteristics and variables are obtained to make it easier to mine the data for the analysis.

a. Detection of missing values

Representing the values that are missing from a columns or proposed attributes. The values that are lacking are filled in as a result. One technique for carrying out this task is to replace values with the median, average, or predetermined value, as shown in (1).

$$\% V. faltantes = \frac{N.V. faltantes}{N.T. de registros} \times 100 \% \quad (1)$$

V is represents missing values (data), $N.V$ is represents the number of missing numbers in the model, $N.T$ is represents the number of records in the database.

b. Detection of missing values

Interquartile range (IQR) as shown in (2). The third quartile of this formula is $Q3$, while the first quartile is $Q1$. Values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are considered exceptions and may be eliminated or treated in a specific manner.

$$IQR = Q3 - Q1 \quad (2)$$

c. For the standardization of data

Min-max scaling: Determine the numerical attribute scaling in a certain range; a clear example could be 0 and 1, as seen in (3). This helps in normalizing the data for better analysis. In this sense, the data can be normalized correctly.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

d. Z-Score standardization

To standardize numerical attributes with a mean of 0 and a standard deviation of 1, as shown in (4). Where μ is the mean and σ is the standard deviation. In that sense, the operation is performed as visualized in (4). Consequently, Figure 3 shows a graph as far as data processing is concerned. That is, in this process, we perform data cleaning, normalization, data duplication, and other functions.

$$Z = \frac{X - \mu}{\sigma} \quad (4)$$

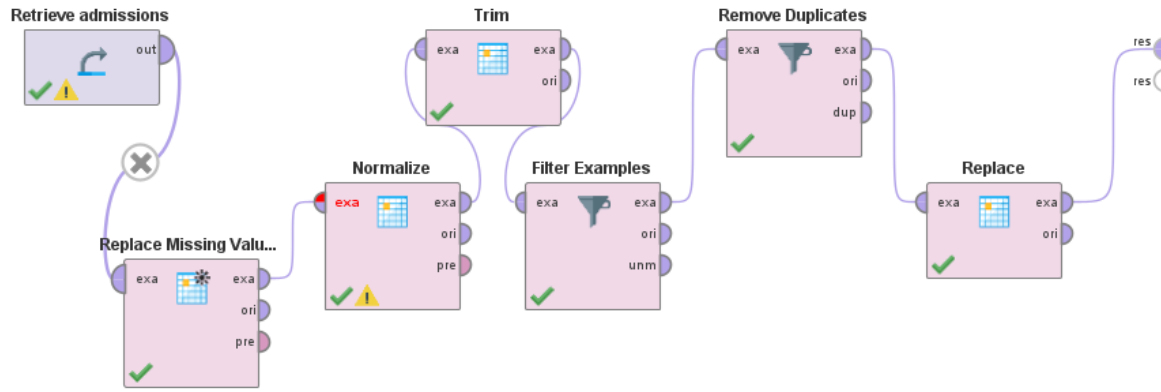


Figure 3. Data processing

3.2.3. Data transformation

At this point, the data previously processed in the previous steps is transformed into a more accurate representation of the desired investigation according to the objectives established in the research. This includes adding new features to the established model to decrease the dimension of the data. In this context, the concepts of the data transformation step will be applied and explained in detail by mathematical formulas in each process.

a. Normalization

The minimum and maximum normalization formula to adjust the values of a variable X to a specific range, for example, from 0 to 1 as specified in (5). This technique ensures that all values fall within the desired range. In this way normalization can be performed optimally.

$$X_{norm} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (5)$$

The Z-score, also known as a standard score, is a statistical measure that measures the number of standard deviations at a specific data point that are above or below the mean of a data set. It is a useful tool in statistics and data analysis to assess the relative position of a data point in a distribution. The Z-score normalization formula for standardizing on the variable X as visualized in (6).

$$X_{std} = \frac{X - \text{mean}(X)}{\text{std}(X)} \quad (6)$$

b. Imputation of missing data

– Imputation based on mean

Mean-based imputation is a missing data handling technique that includes replacing missing values with the associated variable's arithmetic mean. When dealing with missing data in a data collection. This strategy is widely employed to maintain statistical consistency, as described in (7).

$$X_{imputed} = \text{mean}(X) \quad (7)$$

– Median-based imputation

Median-based imputation is a missing data management strategy that replaces missing values with the variable's median. This strategy, like median-based imputation, is employed when there is missing data and aims to maintain statistical consistency in the data set, as shown in (8). This approach helps to preserve the central tendency of the data.

$$X_{imputed} = \text{median}(X) \quad (8)$$

c. Dimensionality reduction

– Formula for principal components analysis (PCA)

As indicated in form (9), it is a statistical technique used to turn a set of correlated variables into a set of uncorrelated variables known as principal components. This method simplifies the data by reducing its

dimensionality while retaining most of its original information. Principal component analysis is widely used in various fields such as data compression, feature extraction, and data visualization.

$$PCA = \text{eigenvectors} \times \text{original data} \quad (9)$$

The operators utilized in the data transformation process are depicted in Figure 4. This prepares the data for the following procedure, which consolidates the model presented in the objectives. These steps are crucial for ensuring accurate and meaningful results in the analysis.

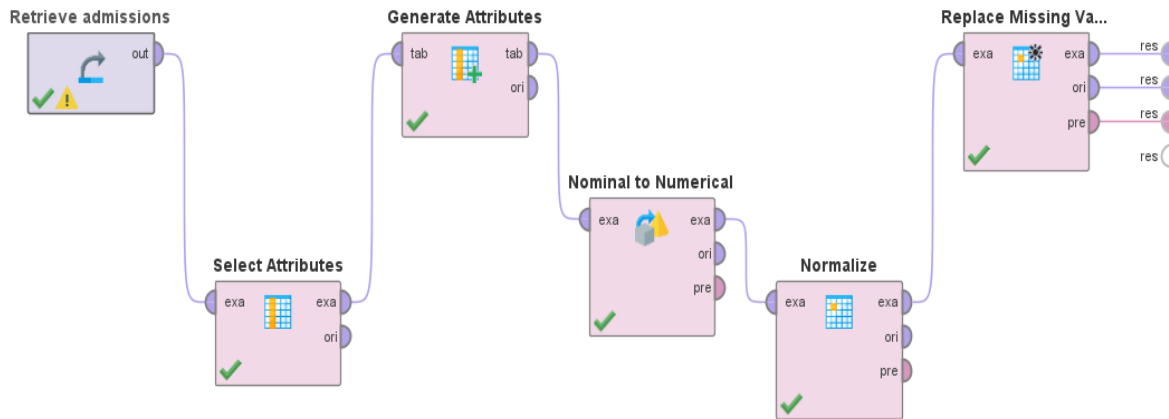


Figure 4. Data transformation process

3.2.4. Data mining

The representation of the data miner is the phase of the KDD methodology that is most important in the overall process. This process establishes some criteria, such as the identification of significant patrons that will provide representations depending on the types of models. As a result, choosing a data mining algorithm also involves choosing the methods to find patterns in the data as well as the best models and parameters depending on the kind of data (categorical or numerical).

a. Apriori algorithm

For the processing of combinations of frequently occurring elements, prior knowledge algorithms are used. As a result, create a collection of data known as an item, for which a database is used to provide support. Enables the effective discovery of a set of frequently occurring sets of terms that serve as the foundation for associative rules [32].

– Itemset

Itemsets play a fundamental role in data mining and association rule learning, particularly in the context of transactional databases. In such databases, each transaction typically represents a set of items purchased or observed together. An itemset is essentially a collection of one or more items that frequently occur together within these transactions.

– Support

The concept of support is central to frequent itemset mining and association rule learning. It quantifies the frequency with which a particular itemset appears in a database relative to the total number of transactions in the database. In other words, support measures the proportion of transactions that contain a specific combination of items.

– Frequent itemset

In association rule mining, a frequent itemset is a set of items (or items) that occur together frequently in a dataset, surpassing a predefined threshold known as the support threshold or minimum support. The support of an itemset is the proportion of transactions in the dataset that contain that particular itemset. The support threshold, or minimum support, acts as a criterion for determining which itemsets are considered frequent. Itemsets with a support value equal to or higher than the support threshold is deemed frequent, while those with a support value below the threshold are considered infrequent and typically discarded.

– Rules of association

The use of association rules in data mining makes it possible to analyze databases in search of patterns or co-occurrences. A clear example is the discussion of one article in relation to another. Given the possibility that one item is chosen, given that another item has already been chosen. It is possible to state that a customer is likely to buy both items if he/she buys item A, as visualized in (10).

$$\{item A, item B\} \rightarrow \{item C\} \tag{10}$$

b. Support and trust

In other words, support and confidence are two key parameters in rule-oriented algorithms. Let X be a list. As a result, the support is to perform the following transcountable fracture as mentioned in (11). The number of transactions in a database is represented by the letter N . In the equation, $(X Y)$ refers to the total number of transactions in a database, with X serving as the antecedent and Y serving as the result. As shown in (12), a fraction of transactions must occur in which the itemsets X appear in order for a rule to be considered reliable. It could be seen as the frequency with which a transaction involving item X also includes item Y .

$$support(X \Rightarrow Y) = support(X \cup Y) = \frac{cont(X \cup Y)}{N} \tag{11}$$

$$conf(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \tag{12}$$

c. Lift proposition

The lift is carried out to acquire the observed frequency through the application of a rule that is carried out with the anticipated frequency by chance (if the rule is not actually there), as seen in (13). On the other hand, the lift helps us identify instances of increased sales probability, for instance, in accordance with the rules followed. A graph illustrating the use of association rules while implementing the previous algorithm is shown in Figure 5. In this sense, all the concepts mentioned in the earlier-established definition, such as the calculation of support, confidence, and lift, are applicable.

$$Lift = conf \frac{(X \rightarrow Y)}{s(Y)} \tag{13}$$

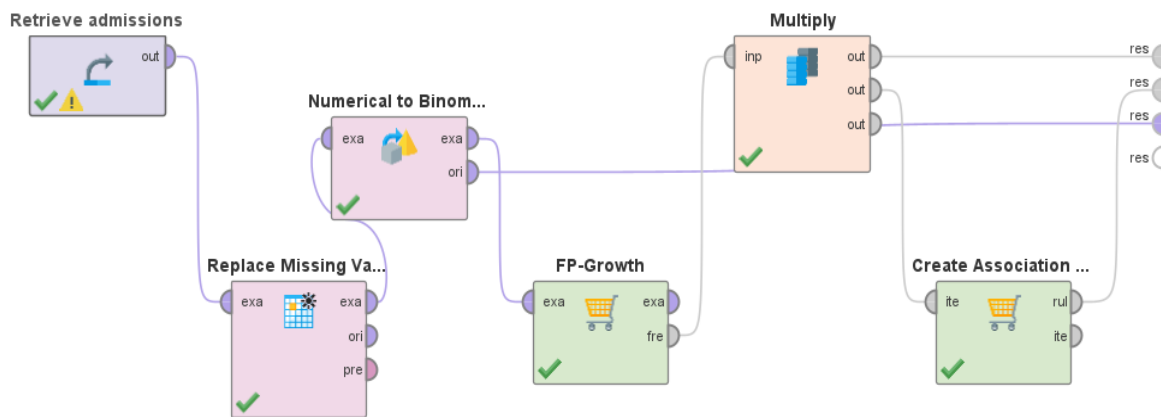


Figure 5. Application of the a priori algorithm

d. Numerical to binomial

This operator enables us to store Boolean variables (true or false). Realize the transformation of characteristics or attributes that originally included numeric values into a binary or binomial variable. This transformation simplifies the data representation for certain analyses.

– FP-growth

We are able to filter using the minimal amount of support required and locate frequently used itemsets. The frequently used itemsets are filtered with a minimum level of support. This filtering process helps in identifying meaningful patterns in the data.

– Multiply

To duplicate an operator output. Imagine the often-occurring itemsets performing operations involving the multiplication of numerical attributes with the help of this operator. This duplication process aids in creating multiple instances of the same output for further analysis.

– Create association rules

Permits seeing the frequent item sets' association rules. separating relationships using a minimal level of trust. This helps in identifying meaningful associations with higher confidence levels. In this way, it can be done more efficiently.

4. RESULTS

4.1. Evaluation of results

After completing the process of evaluating the algorithm's application beforehand and implementing the procedures related to this algorithm using the tool RapidMiner Studio. The process of producing the project's results will be initiated. The established items in each field are grouped together into itemsets, allowing the fusion of the antecedent variables and the consequent variables. For this reason, the support criteria listed in Table 2 for each item are applied.

Table 2. Data on items and their support

Size	Support	Item 1	Item 2	Item 3
1	0.667	Male		
1	0.508	Number of admissions	Cancer which can be caused by smoking	
1	0.492	Attribute number	Female	
1	0.349	Cancer which can be caused by smoking		
1	0.333	Cancer which can be caused by smoking	Female	
1	0.190	Circulatory diseases which can be caused by smoking		
2	0.339	Female		
2	0.328	Male		
2	0.233	Male	Number of admissions	
2	0.127	Male	Attribute number	
2	0.175	Male	Cancer which can be caused by smoking	
2	0.169	Male	Circulatory diseases which can be caused by smoking	
2	0.175	Male	Number of admissions	Cancers that can be caused by smoking
2	0.164	Male	Attribute number	Cancers that can be caused by smoking
2	0.116	Number of admissions		
3	0.116	Number of admissions	Cancers that can be caused by smoking	
3	0.116	Number of admissions	Female	

The rules of association, crucial for understanding the interplay between various factors, are generated with stringent parameters. Employing a minimum support of 0.5 and a lift evaluation criterion of 0.1 ensures that only robust associations are considered, filtering out noise and spurious correlations. This rigorous approach, exemplified in Table 3, lays the foundation for meaningful analysis and interpretation. In Table 4, the association rules derived from the model shed light on the intricate relationships within the analyzed database. Specifically, these rules illuminate the connection between sex and smoking-related diseases, providing insights into potential risk factors and underlying patterns. Each rule is accompanied by a confidence measure, offering a quantitative assessment of the reliability of the association.

Table 3. Rules of association

No.	Premises	Conclusion
5	Countable number	Cancers in men that can be induced by smoking
15	Countable number	Cancers that are caused by smoking
14	Female	Cancers that are caused by smoking
13	Male	Cancers that are caused by smoking
16	Attributable number, male	Cancers that are caused by smoking
12	Number of admissions, male	Cancers that are caused by smoking
4	Admissions figures	Cancers in men that can be induced by smoking
11	Admissions figures	Cancers that are caused by smoking

Table 4. Association rules obtaining confidence

Identified association rules	
[Male] --> [Number of admissions, Cancers which can be caused by smoking] (confidence: 0.175)	
[Male] --> [Attributable number, Cancers which can be caused by smoking] (confidence: 0.175)	
[Male] --> [Circulatory diseases which can be caused by smoking] (confidence: 0.190)	
[Number of admissions] --> [Male, Cancers which can be caused by smoking] (confidence: 0.229)	
[Attributable number] --> [Male, Cancers which can be caused by smoking] (confidence: 0.237)	
[Number of admissions] --> [Female] (confidence: 0.333)	
[Attributable number] --> [Female] (confidence: 0.333)	
[Cancers which can be caused by smoking] --> [Female] (confidence: 0.333)	
[Cancers which can be caused by smoking] --> [Male, Number of admissions] (confidence: 0.333)	
[Cancers which can be caused by smoking] --> [Male, Attributable number] (confidence: 0.333)	
[Number of admissions] --> [Cancers which can be caused by smoking] (confidence: 0.344)	
[Male, Number of admissions] --> [Cancers which can be caused by smoking] (confidence: 0.344)	
[Male] --> [Cancers which can be caused by smoking] (confidence: 0.349)	
[Female] --> [Cancers which can be caused by smoking] (confidence: 0.349)	
[Attributable number] --> [Cancers which can be caused by smoking] (confidence: 0.355)	
[Male, Attributable number] --> [Cancers which can be caused by smoking] (confidence: 0.355)	
[Male] --> [Attributable number] (confidence: 0.492)	
[Female] --> [Attributable number] (confidence: 0.492)	
[Cancers which can be caused by smoking] --> [Number of admissions] (confidence: 0.500)	
[Cancers which can be caused by smoking] --> [Attributable number] (confidence: 0.500)	
[Male, Cancers which can be caused by smoking] --> [Number of admissions] (confidence: 0.500)	
[Male, Cancers which can be caused by smoking] --> [Attributable number] (confidence: 0.500)	
[Male] --> [Number of admissions] (confidence: 0.508)	
[Female] --> [Number of admissions] (confidence: 0.508)	
[Number of admissions] --> [Male] (confidence: 0.667)	
[Attributable number] --> [Male] (confidence: 0.667)	
[Cancers which can be caused by smoking] --> [Male] (confidence: 0.667)	
[Circulatory diseases which can be caused by smoking] --> [Male] (confidence: 0.667)	
[Number of admissions, Cancers which can be caused by smoking] --> [Male] (confidence: 0.667)	
[Attributable number, Cancers which can be caused by smoking] --> [Male] (confidence: 0.667)	

Important outcomes such as confidence support or lift, as listed in Table 5, are highlighted to fulfill the specified criteria. These criteria serve as benchmarks for evaluating the strength of associations between different elements. The confidence criteria, depicted in Figure 6, are structured in a scheme aligning with their association rules and organized elements. Figure 7 presents the association rules alongside the outcomes of applying the lift criterion, enabling the selection of rules exceeding a value of 1.

Table 5. Result on the rules of association

Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
0.116	0.237	0.748	-0.868	0.002	1.016	1.005
0.175	0.355	0.747	-0.810	0.003	1.016	1.009
0.116	0.349	0.837	-0.550	0	1	1
0.233	0.349	0.740	-1.101	0	1	1
0.116	0.355	0.841	-0.540	0.002	1.016	1.009
0.116	0.344	0.834	-0.561	-0.002	0.984	0.992
0.116	0.229	0.740	-0.899	-0.002	0.984	0.995
0.175	0.344	0.779	-0.841	-0.003	0.984	0.992

This systematic approach aids in identifying significant patterns and relationships within the dataset, facilitating informed decision-making and strategy development. Moreover, by examining these metrics such as confidence support and lift, researchers can gain insights into the reliability and significance of the identified associations. These metrics play a crucial role in determining the strength and importance of relationships between variables, guiding subsequent analysis and interpretation. The structured organization of criteria and association rules streamlines the process of identifying meaningful patterns and correlations, ultimately contributing to a deeper understanding of the underlying data dynamics. Through meticulous analysis and adherence to established criteria, researchers can make informed decisions, optimize processes, and uncover valuable insights that drive progress and innovation in their respective fields.



Figure 6. Confidence criterion tree

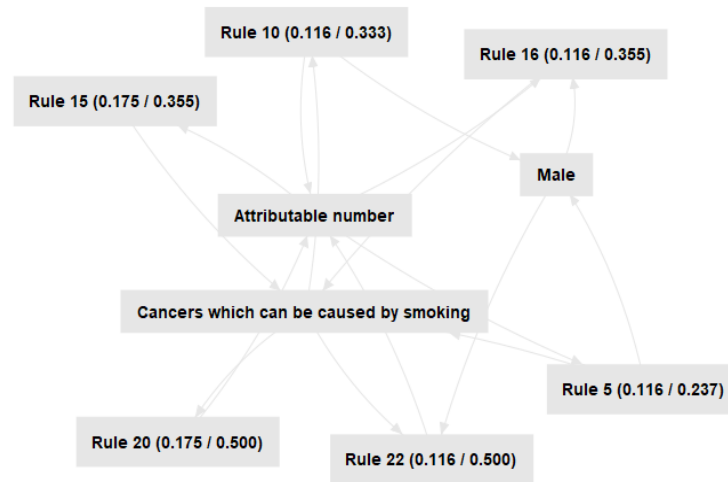


Figure 7. Tree criteria based on the lift

4.2. Comparison of methodologies

A comparison of three significant methodologies, including KDD, Crips-dm, and Semma, is done in order to choose one for the proposed project. Finally, each methodology has specific procedures that must be followed in order to do the proper analysis, like we did in earlier processes. The comparison of the most common methods used in the implementation of data mining is shown in Table 6. To that end, the differences between the methodologies are described via attribute comparison.

4.3. Model comparison

In this section, various models and algorithms are compared. These models and algorithms are used in a classification system that enables them to produce precise results in line with the proposed research. A comparison of models like naive Bayes, decision trees, and rule induction is shown in Figure 8. The algorithm with the best results under 1.0 is the induction of rules, while the other algorithms have worse results.

Table 6. Comparison of methodologies

Comparison of attributes	Methodology KDD	Methodology CRIPS-DM	Methodology SEMMA
Structure and sequence	Other methodologies include data selection, cleaning, transformation, and extraction, as well as evaluation and application of knowledge, although their organization is less precise [33].	This methodology consists of six steps: understanding the business, understanding the data, preparing the data, modeling, assessing, and implementing [34].	The five steps of Semma's methodology are demonstration, exploration, modification, modeling, and evaluation [35].
Business orientation	Recognizes the significance of the company's commercial goals and strives to learn how to gain a competitive advantage.	Understands the commercial goals from the outset and ensures that the results are useful and processable for decision-making.	Perform an information analysis taking into account the company's goals and how the results are used.
Flexibility	Using a broader and less structured focus, it provides a general framework for the discovery of knowledge.	It is adaptable to a variety of situations and projects and can be expanded for commercial use.	Although it follows a predetermined sequence of steps, it can be tailored to various projects.
Interaction	Calls for repetition. However, it lacks a clear structure, much as the Semma or Crisp-DM methodologies.	Learn how to use an iterative process for outcome review. Projects that are always evolving are adjusted to.	It is a procedure that can be carried out in stages as necessary. Adjustments may be made throughout the process.

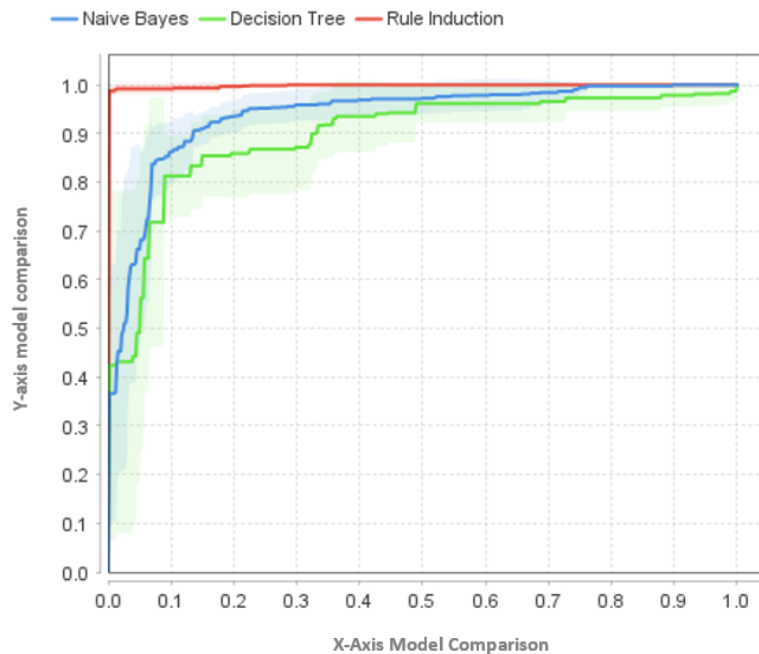


Figure 8. Model comparison

5. DISCUSSION

Data mining uses specific algorithms that are implemented for each activity to extract patterns of behavior from large amounts of data and identify them as variables [18]. On the one hand, the application for automatic learning makes use of mathematical algorithms to identify and compile key characteristics of a certain subject [18]. To do this, the use of rule-based analysis includes an explanation of how the data analysis algorithms work.

However, making decisions is based on the outcomes of accurate predictions made with reliable technology [20]. Due to the substances, they contain, cigarillos have an addictive component that has led to a

number of issues. As a result, those who abuse these substances become dependent on them because they allow them to unwind and relieve the psychological stress associated with smoking [27]. As a result, it is proposed in the studies to classify them as light or heavy smokers and apply certain processes using data mining techniques [28]. One of the most important factors in the early detection of cardiovascular diseases is excessive alcohol and tobacco use. As a result, data mining calls for looking for factors connected to symptoms like high blood pressure, high levels of glucose in the blood, or high cholesterol. In this context, techniques for data mining based on automatic learning are used to analyze medical data [30].

The study aims to classify smokers and detect cardiovascular disease risk factors through data mining techniques. Its importance lies in early detection and potential intervention strategies. Unanswered questions include the accuracy of predicting health outcomes solely based on data patterns and the effectiveness of intervention strategies identified through data mining. Future research could explore refining predictive models and evaluating the impact of interventions on health outcomes

6. CONCLUSION

To conclude, the work focuses on tobacco consumption or smoking for people addicted to these substances, which are generally considered a remedy to overcome stress and relaxation problems. The pleasure of smoking brings certain problems in the future to the addicted person, which makes diseases manifest themselves more easily. The application of data mining allows for accurate results about the rules created with the tool RapidMiner Studio. According to the rules created based on the application of the a priori algorithm it is necessary to consider certain criteria for their evaluation such as the application of concepts like a collection of itemsets, the number of times an itemset appears in a database called frequent itemsets support, the calculation of confidence among other factors. In this sense, it was proposed to use a minimum support of 0.5, which represents 50%, with the lift criterion of 0.1, which is equivalent to 10%, which makes the rule valid and obtains the appropriate results for the model. The combination of studies on tobacco consumption and data mining represents a powerful tool for better understanding tobacco consumption patterns, identifying risk factors and intervention effectiveness, and designing more effective strategies for tobacco prevention and cessation. By utilizing data mining techniques, it is possible to analyze large datasets related to smoking, such as consumption patterns, demographic profiles, associated risk factors, and intervention outcomes. This enables researchers and healthcare professionals to make more informed and personalized decisions, as well as develop more effective health policies and programs to address the issue of smoking and its impacts on public health. To conclude, it is recommended for future work to use these data mining models on large amounts of data, such as big data. For this purpose, it is necessary to propose the appropriate model for certain investigations. You can also create programmable algorithms with a programming language such as Python and also use application programming interfaces (APIs) that handle graphical interfaces that perform these tasks.




REFERENCES

- [1] M. A. Parker, R. Cruz-Cano, J. M. Streck, E. Ballis, and A. H. Weinberger, "Incidence of opioid misuse by cigarette smoking status in the United States," *Addictive Behaviors*, vol. 147, Dec. 2023, doi: 10.1016/j.addbeh.2023.107837.
- [2] A. Prabhakaran, V. Restocchi, and B. D. Goddard, "Improving tobacco social contagion models using agent-based simulations on networks," *Applied Network Science*, vol. 8, no. 1, Aug. 2023, doi: 10.1007/s41109-023-00580-5.
- [3] R. G. Salloum *et al.*, "An effectiveness-implementation hybrid trial of phone-based tobacco cessation interventions in the Lebanese primary healthcare system: protocol for project PHOENICS," *Implementation Science Communications*, vol. 4, no. 1, Jun. 2023, doi: 10.1186/s43058-023-00456-w.
- [4] S. Pratik, D. S. K. Nayak, R. Prasath, and T. Swarnkar, "Prediction of smoking addiction among youths using elastic net and KNN: a machine learning approach," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13119, Springer International Publishing, 2022, pp. 199–209.
- [5] A. Copeland, T. Stafford, and M. Field, "Recovery from nicotine addiction: a diffusion model decomposition of value-based decision-making in current smokers and ex-smokers," *Nicotine and Tobacco Research*, vol. 25, no. 7, pp. 1269–1276, Mar. 2023, doi: 10.1093/ntr/ntad040.
- [6] E. K. O'Brien, M. Roditis, A. Persoskie, and K. A. Margolis, "Youths' perceptions of nicotine harm and associations with product use," *Nicotine and Tobacco Research*, vol. 25, no. 7, pp. 1302–1309, Mar. 2023, doi: 10.1093/ntr/ntad028.
- [7] A. M. Novick *et al.*, "Progesterone increases nicotine withdrawal and anxiety in male but not female smokers during brief abstinence," *Nicotine and Tobacco Research*, vol. 24, no. 12, pp. 1898–1905, Jun. 2022, doi: 10.1093/ntr/ntac146.
- [8] T. Jose, D. R. Schroeder, and D. O. Warner, "Changes in cigarette smoking behavior in cancer survivors during diagnosis and treatment," *Nicotine and Tobacco Research*, vol. 24, no. 10, pp. 1581–1588, Mar. 2022, doi: 10.1093/ntr/ntac072.
- [9] X. Liu, M. Xu, W. Jia, Y. Duan, J. Ma, and W. Tai, "PU.1 negatively regulates tumorigenesis in non-small-cell lung cancer," *Medical Oncology*, vol. 40, no. 2, Jan. 2023, doi: 10.1007/s12032-023-01946-6.
- [10] J. Fernández-Torres *et al.*, "Impact of cadmium mediated by tobacco use in musculoskeletal diseases," *Biological Trace Element Research*, vol. 200, no. 5, pp. 2008–2015, Jul. 2022, doi: 10.1007/s12011-021-02814-y.
- [11] J. Djekic Malbasa *et al.*, "Decade of lung cancer in Serbia: tobacco abuse and gender differences," *European Review for Medical and Pharmacological Sciences*, vol. 27, no. 7, pp. 3105–3116, 2023, doi: 10.26355/eurrev_202304_31945.




- [12] D. Piloni *et al.*, "Smoking habit and respiratory function predict patients' outcome after surgery for lung cancer, irrespective of histotype and disease stage," *Journal of Clinical Medicine*, vol. 12, no. 4, Feb. 2023, doi: 10.3390/jcm12041561.
- [13] M. Nuñez, M. M. Narváez-Ríos, P. A. Quezada-Sarmiento, and L. X. Suárez-Morales, "Prediction techniques in tobacco consumption by adolescents," *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, vol. 2021, pp. 155–163, 2021.
- [14] S. M. Yang *et al.*, "Performance of automated oral cancer screening algorithm in tobacco users vs. non-tobacco users," *Applied Sciences (Switzerland)*, vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13053370.
- [15] S. C. Hanganu and L. C. Hanganu, "Knowledge-based expert system applied in oral health towards tobacco use prevention and smoking cessation," *Annals of DAAAM and Proceedings of the International DAAAM Symposium*, pp. 615–616, 2009.
- [16] A. Karlsson *et al.*, "Impact of deep learning-determined smoking status on mortality of cancer patients: never too late to quit," *ESMO Open*, vol. 6, no. 3, Jun. 2021, doi: 10.1016/j.esmoop.2021.100175.
- [17] K. Rijhwani, V. R. Mohanty, Y. B. Aswini, V. Singh, and S. Hashmi, "Applicability of data mining and predictive analysis for tobacco cessation: An exploratory study," *Frontiers in Dentistry*, vol. 17, Nov. 2020, doi: 10.18502/fid.v17i24.4624.
- [18] G. Feng and M. Fan, "Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization," *Expert Systems with Applications*, vol. 237, 2024, doi: 10.1016/j.eswa.2023.121555.
- [19] D. Zhang, "Research on dance art teaching system based on data mining and machine learning," *Computer-Aided Design and Applications*, vol. 21, no. 2, pp. 54–68, Jul. 2024, doi: 10.14733/cadaps.2024.s2.54-68.
- [20] J. Liu, "The impact of data mining on management and digital marketing in the age of big data," *Computer-Aided Design and Applications*, pp. 229–247, Aug. 2023, doi: 10.14733/cadaps.2024.s3.229-247.
- [21] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance," *Trends in Neuroscience and Education*, vol. 33, 2023, doi: 10.1016/j.tine.2023.100214.
- [22] A. Gehlot and N. Misra, "Retracted: An IoT based smart healthcare medical system using deep learning algorithm," in *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Oct. 2022, pp. 1–6, doi: 10.1109/MysuruCon55714.2022.9972370.
- [23] S. Mohapatra *et al.*, "A stacking classifiers model for detecting heart irregularities and predicting cardiovascular disease," *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2022.100133.
- [24] T. Li, Z. Wan, and J. Guo, "A new nonmonotone spectral projected gradient algorithm for box-constrained optimization problems in $m \times n$ real matrix space with application in image clustering," *Journal of Computational and Applied Mathematics*, vol. 438, Mar. 2024, doi: 10.1016/j.cam.2023.115563.
- [25] M. H. T. Najaran, "An evolutionary ensemble convolutional neural network for fault diagnosis problem," *Expert Systems with Applications*, vol. 233, Dec. 2023, doi: 10.1016/j.eswa.2023.120678.
- [26] K. M. Kim, J. H. Kim, H. S. Rhee, and B. Y. Youn, "Development of a prediction model for the depression level of the elderly in low-income households: using decision trees, logistic regression, neural networks, and random forest," *Scientific Reports*, vol. 13, no. 1, Jul. 2023, doi: 10.1038/s41598-023-38742-1.
- [27] M. Hua, S. Sadah, V. Hristidis, and P. Talbot, "Health effects associated with electronic cigarette use: Automated mining of online forums," *Journal of Medical Internet Research*, vol. 22, no. 1, Jan. 2020, doi: 10.2196/15684.
- [28] T. Wei, F. Huang, X. Zhang, and S. Zuo, "Research and application of data mining based on artificial intelligence in cigarette delivery," in *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Jun. 2020, pp. 184–189, doi: 10.1109/ICBAIE49996.2020.00046.
- [29] H. Singh, T. Gupta, and J. Sidhu, "Prediction of heart disease using machine learning techniques," in *Proceedings of the IEEE International Conference Image Information Processing*, Nov. 2021, vol. 2021-Novem, pp. 164–169, doi: 10.1109/ICHIP53038.2021.9702625.
- [30] S. Thammaboosadee and K. Yuttanawa, "A multi-stage predictive model for smoking cessation: Success and choices of medication approaches," *International Journal of Electronic Healthcare*, vol. 11, no. 3, pp. 239–255, 2021, doi: 10.1504/ijeh.2021.117125.
- [31] A. Dekhtyar and J. H. Hayes, "Automating requirements traceability: two decades of learning from KDD," in *2018 1st International Workshop on Learning from other Disciplines for Requirements Engineering (D4RE)*, Aug. 2018, pp. 12–15, doi: 10.1109/D4RE.2018.00009.
- [32] Q. Hao, W. J. Choi, and J. Meng, "A data mining-based analysis of cognitive intervention for college students' sports health using Apriori algorithm," *Soft Computing*, vol. 27, no. 21, pp. 16353–16371, Sep. 2023, doi: 10.1007/s00500-023-09163-z.
- [33] 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, pp. 201–208, Mar. 2021, doi: 10.33166/AETiC.2021.05.025.
- [34] 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.
- [35] 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 LNCS, Springer International Publishing, 2020, pp. 222–234.

BIOGRAPHIES OF AUTHORS






Laberiano Andrade-Arenas    doctor in systems and computer engineering. master 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 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.



Cesar Yactayo-Arias    master's study in applied mathematics at the Universidad Nacional Mayor de San Marcos. Since 2016 he has been teaching mathematics subjects at the Universidad de Ciencias y Humanidades and the Universidad Continental. Currently, he is the author and co-author of several peer-reviewed articles in high-impact journals. He can be contacted at e-mail: yactayocesar@gmail.com.