The Review Of
DIABETIC
STUDIES OPEN ACCESS

Application Of Data Mining Techniques For The Analysis Of Work Absenteeism At In The Human Resources Records Of An Academic Institution

Juan Sebastián Salazar-Osorio[1] , Sonia Jaramillo-Valbuena[2] , Jorge Iván Triviño-Arbeláez[3]

[1,2,3]*Universidad del Quindío, Armenia, Quindío, Colombia.*

Submission date: December 2024 Acceptance date: December 2025

Abstract

Objective: Identify trends and patterns in employee absenteeism in the Human Resources Department at the Academic Institution, based on data analysis using the CRISP-DM methodology, in order to provide strategic input that strengthens institutional human talent management.
Method: Quantitative, descriptive, and exploratory design, based on data mining, from CRISP-DM. The population corresponded to a total of 1,406 disability records in the period 2012-2024. Cleaning, preparation, attribute selection, and modeling processes were used, employing statistical and computational techniques.
Results: General illness is the main cause of absences; administrative workers account for the highest proportion of absences; prolonged absences are associated with middle and advanced ages. Furthermore, the incorporation of ensemble methods revealed more stable and accurate predictive patterns across these groups, reinforcing the utility of advanced analytical approaches for understanding the dynamics of absenteeism.
Discussions: While the prevalence of general illness is consistent with what has been documented in the literature, the concentration of absenteeism among administrative staff raises a particular issue in the university context. Organizational and contractual factors influence this phenomenon. The usefulness of data mining as a predictive and interpretive tool for understanding workplace absenteeism is highlighted.
Conclusions: The application of the CRISP-DM methodology made it possible to identify relevant patterns and transform them into strategic inputs to strengthen institutional human talent management. Additionally, the results demonstrate the value of ensemble methods, which outperformed individual classifiers and provided more robust predictive capacity, highlighting their potential as reliable analytical tools for supporting evidence-based decision-making

Keywords: Advanced analytics, Work absenteeism, Data mining, Predictive models, Human talent.
JEL Classification: J28-C55-I12-M54.

Introduction

The phenomenon of absenteeism is recognized as one of the main difficulties in the management of institutions (Calle and Gil, 2025) since it directly affects the continuity of production processes (Angarita et al., 2024), generates overload in the work teams (Mendoza, 2024) and increases the costs associated with social security (Guerrero et al., 2025) and its incidence, according to Béjar et al. (2025), Pozo-Pozo and Espinosa-Tigre (2025) and Rivero et al. (2025). It transcends economic losses, becoming embedded in the relational dynamics of groups, causing tensions, decreasing motivation, and altering the organizational climate. In the Colombian case, this phenomenon has been categorized as a public health problem (Barrera-

Sigua et al., 2023); since, according to Arrieta-Burgos et al. (2025) In 2023, the average was 1.71 cases of absenteeism per worker, an increase of 13% compared to 2022 (1.51); thus, the main cause continues to be general or common illnesses (67% of cases), with an average of 0.91 cases per worker; likewise, work accidents and illnesses represent around 4.8% of cases with a downward trend; permits and licenses represent 29% of cases, with a notable increase in 2023.

This is due to multiple causes, including both medical reasons and social and psychological factors (Grijalba and Riascos, 2025; Mesa-Mesina et al., 2025; Tatamuez-Tarapues et al., 2018) However, despite this consideration, organizations tend to focus on accounting for absences rather than analyzing their underlying causes. Colombian academic university institutions are no exception; although they have human resources management systems that systematically store information related to absences, the use of these records has been limited to accounting and administrative purposes, without generating analytical processes that would allow for a thorough understanding of behavioral patterns or the factors associated with absenteeism. This situation creates a knowledge gap and justifies the problem statement: having information on absences does not guarantee its strategic use if methodologies capable of transforming the data into inputs for decision-making are not employed.

Consequently, the lack of in-depth analysis hinders the identification of risk factors, the understanding of trends over time, and the design of effective institutional policies. Therefore, it becomes necessary to employ data mining techniques to uncover hidden relationships within the available records, transforming dispersed information into actionable knowledge. Thus, the problem statement focuses on the lack of studies using advanced analytical methods to examine absenteeism at an Academic Institution located in the Colombian Coffee region, which limits the institution's capacity to anticipate and manage this phenomenon. In this context, the research question is formulated as follows: What trends and patterns can be identified in absenteeism within the Human Resources Department of the Academic Institution, (based on data analysis using the CRISP-DM methodology), that provide strategic input for strengthening the institution's human talent management?

This question highlights the need to apply approaches that go beyond simple statistical description and move towards building predictive models that promote more efficient human talent management. In response to this question, the general objective was to identify trends and patterns in absenteeism at the Human Resources Department of the Academic Institution, based on data analysis using the CRISP-DM methodology, in order to provide strategic inputs that strengthen the institution's human talent management. From this central purpose, the following specific objectives were derived: a) to characterize the recorded information on absenteeism at the Human Resources Department of the Academic Institution, identifying the relevant variables for analysis; b) to prepare the dataset through cleaning, transformation, and variable selection processes, in order to guarantee its quality and relevance for modeling; c) to apply data mining techniques using the CRISP-DM methodology to build and evaluate analytical models that allow for the recognition of patterns and trends in absenteeism. This methodological approach ensures coherence between the question posed, the procedures used, and the expected results.

Operationally, the following specific variables were analyzed using information available from the Human Resources Department of the Academic Institution: days of absence, diagnosis ID, absence reason code ID, age at date of absence, marital status, date of birth, absences by Gender for the years 2013 to 2023, payroll class, birth municipality ID, and discretized age. The combination of these variables provides a comprehensive view of the phenomenon, allowing for the identification of how personal, institutional, and clinical factors interact in shaping absenteeism. In terms of scope, the research was limited to the analysis of 1,406 records of work-related absences from the period 2012 to 2024 at the Academic Institution. This temporal delimitation allowed for the identification of trends and patterns of behavior over twelve years, although the findings cannot be automatically extrapolated to other organizational contexts. On the other hand, the study's limitations are related to the quality of the available records, the accuracy of the diagnoses,

the veracity of the reports, and the consistency of the information that can influence the results obtained. Furthermore, the research did not include qualitative variables related to employees' perceptions of the organizational climate or their motivation, which would have enriched the understanding of the phenomenon from a more holistic perspective.

Concomitantly, the justification for this research lies in the need to make the most of institutional records by incorporating advanced analytical methodologies that transcend conventional descriptive approaches. Data mining, of course, emerges as an ideal tool for identifying hidden patterns, building predictive models, and generating strategically valuable knowledge for organizational management. In the case of the Academic Institution, the application of these techniques opens the possibility of transforming administrative data into evidence that supports the design of wellness programs, self-care strategies, and policies aimed at reducing absenteeism and its economic and social impacts. Thus, the study, in addition to providing a precise diagnosis, offers input for formulating informed decisions that promote institutional sustainability.

Regarding the background, the literature shows a growing trend towards applying data mining and machine learning techniques to study absenteeism. The studies of Araujo et al. (2019), Bayram & Burgazoglu (2020), Berón et al., (2021), Lawrence et al. (2021), Muñoz (2020), Rodríguez (2025), Skorikov et al. (2020) and Zupančič & Panov (2024) demonstrated the usefulness of algorithms such as neural networks, Naive Bayes, and decision trees in predicting work absences. Random Forest and Gradient Boosted Trees were also implemented to anticipate the occurrence and duration of work-related disabilities, achieving high levels of accuracy. These studies confirm that advanced analytics is an effective alternative to traditional methods of study. In contrast, in Colombia, research on absenteeism has been largely limited to the development of occupational health indicators and descriptive analyses, without systematically incorporating predictive or analytical approaches. Therefore, this study represents a novel contribution by applying data mining in a public higher education institution, demonstrating that it is possible to generate valuable information for human talent management and the formulation of evidence-based policies.

Theoretical Foundation

Absenteeism

Absenteeism, according to Angarita et al. (2024) and Tatamuez-Tarapues et al. (2018), is recognized as one of the most relevant problems in studies of human resource management, occupational health, and labor economics, because it directly connects organizational productivity with the biopsychosocial conditions of workers (Arrieta-Burgos et al., 2025;Skorikov et al. (2020) So, in basic terms, it refers to the employee's failure to attend their post during the corresponding time (Barrera-Sigua et al., 2023) However, this concept transcends physical absence and becomes a complex indicator of organizational dynamics, staff health, and the contractual and social ties that mediate the work/employee relationship (Bayram & Burgazoglu, 2020; Mamani, 2023; Pedraza, 2021; Rodríguez, 2025)

Along these lines, authors such as Béjar et al. (2025), Luna y Brokate (2014), Ortiz et al. (2021), Perez et al. (2024), Pulido et al. (2021), Rivero et al. (2025) and Sánchez (2015) define absenteeism as the worker's absence from the workplace during agreed-upon periods, whether there is justification for it or not. This definition emphasizes the quantitative dimension of the phenomenon, measuring the frequency and duration of absences. However, the specialized literature warns that understanding absenteeism must also consider qualitative aspects, such as its causes, motivations, and repercussions. Araujo et al., 2019; Calle and Gil, 2025; Pozo-Pozo and Espinosa-Tigre, 2025) In this sense, absenteeism is a reflection of working conditions, institutional commitment, and health environments.

In this line of thought, Chiavenato (2006, 2008, 2009a, 2009b) this is understood as the set of unforeseen failures that affect the continuity of processes and require the redistribution of workloads. Furthermore, Arrieta-Burgos et al. (2024), Berón et al., (2021) and Muñoz (2020) broaden the discussion by conceiving

it as a behavior that expresses the degree of the worker's connection with the organization and that, of course, influences both productivity and team cohesion. Therefore, absenteeism is interpreted as a symptom of the psychosocial relationships between employees and entities, making it a deeper phenomenon than mere temporary absence. Guerrero et al., 2025; Mesa-Mesina et al., 2025 interpreted this phenomenon as an indicator of occupational morbidity and quality of life, the measurement of which allows for the evaluation of the effectiveness of prevention and health promotion programs in organizations (Grijalba and Riascos, 2025; Mendoza, 2024).

According to its nature, the literature distinguishes several forms of absenteeism: justified absenteeism relates to medical incapacities or authorized leaves, while unjustified absenteeism refers to absences without formal support (Borda et al, n.d.; Saldarriaga and Martínez, 2008; Tatamuez-Tarapues et al., 2018). Added to this is the notion of presentism or "presenteeism" (Morquera, 2017; Perez et al., 2024), in which the worker remains in their position despite not being in adequate health conditions, which reduces their performance (García and Martínez, 2016). This typology shows, according to Cifuentes et al. (2020) and Garcia et al. (2023) that the problem, more than the physical absence of the worker, lies in the deterioration of work capacity and its consequences for the organizational climate. Furthermore, absenteeism is recognized as a Key Performance Indicator (KPI) that directly impacts productivity, efficiency, and operating costs. (Boada et al., 2005; Duke and Valencia, 2021). Its analysis allows the detection of the interaction between sociodemographic variables (age, gender, marital status), contractual variables (type of contract, position, dependency) and clinical variables (diagnosis, days of incapacity, type of license).

Thus, it is evident that it is a multidimensional and multicausal phenomenon, in which individual, organizational and social factors converge (Angarita et al., 2024; Becerra et al., 2024) In higher education institutions, absenteeism is particularly relevant, as the absence of faculty, administrative staff, or support personnel impacts the continuity of educational processes and the quality of service. It is not simply a matter of economic loss or administrative adjustment; on the contrary, it is a factor that shapes pedagogical practices and institutional planning. Therefore, its analysis requires an approach that considers both the economic, educational, and social repercussions.

Data Mining

Data mining has established itself as one of the most relevant tools in the field of advanced analytics (Berón et al., 2021); given that it responds to the need to transform large volumes of information into useful knowledge for decision-making (Dávila and Sánchez, 2012; Ruiz and Armoa, 2023). Therefore, its purpose goes beyond data processing, to discover structures, trends, and relationships within them that are not evident at first glance (Cedillo et al., 2024; Orozco et al., 2021). According to Calle et al. (2024) and Rodríguez (2013) this process consists of identifying valid, novel, understandable patterns with potential applications, which positions data mining as a methodology that transcends conventional statistical description and is oriented towards knowledge discovery.

Therefore, data mining falls within the paradigm of knowledge discovery in databases (KDD), understood as a sequence of interconnected phases that encompass the selection of information, its cleansing, transformation, analysis and, finally, the interpretation of results (Alujah, 2001; Armero et al., 2023; Marulanda et al., 2017; Marcano and Talavera, 2007). In this line of thought, data mining corresponds to the analytical phase, in which computational algorithms are applied to reveal latent patterns in the information (Cabrera, 2024) thus, its main strength lies in the ability to work with massive volumes of data and detect relationships that could not be identified using traditional statistical techniques (Anchundia, 2025; Cotta, 2025; González, 2025; Mullo et al., 2025).

The techniques used in data mining are generally divided into supervised and unsupervised methods. The former include classification and regression, which allow for predicting values or categories based on predictor variables, while the latter include clustering, association rules and correlation analysis, whose

purpose is to explore information without the need for a predefined objective (Noroña et al., 2025; Risco-Ramos, et al., 2023; Seclen, 2025). These methodologies are complemented by the most recent developments in artificial intelligence and machine learning (Díaz et al., 2025; Morales et al., 2025), such as deep neural networks, assembly algorithms and support vector machines, that expand the scope of data mining towards scenarios of high complexity and greater predictive capacity (Orozco et al., 2021; Ruiz and Armoa, 2023).

In this vein, and to bring order and rigor to this process, standardized methodological frameworks have been proposed, with CRISP-DM (Cross-Industry Standard Process for Data Mining) being the most widely used (Corzo, 2025; Zhang 2021This model comprises six interdependent stages: a) understanding the business; b) understanding the data; c) preparing the data; d) modeling; e) evaluation; f) deployment. Thus, its value lies in offering a framework adaptable to different sectors and problems (Dávila and Sánchez, 2012), which ensures that the results have technical coherence and relevance to the strategic purposes of the institutions (Calle et al., 2024). Thanks to its flexibility, CRISP-DM has established itself as a benchmark in academic research and business applications (Sánchez and Pérez, 2021; Wirth & Hipp, s.f.) thus, the applicability of data mining is evident in various contexts, namely: in the financial sector it is used to detect fraud and calculate risks; in commerce, to segment customers and project consumption patterns; in health, to analyze medical records and identify epidemiological factors; and in human talent management, to anticipate turnover, evaluate performance, and analyze phenomena such as absenteeism. In all these scenarios, data mining makes it possible to understand what has happened and generate predictive and prescriptive models that facilitate more impactful strategic decisions (Rodríguez, 2013).

Finally, it should be noted that data mining is based on the premise that raw data is more than just numerical records; it carries underlying structures and hidden meanings that can be transformed into strategic knowledge once discovered. This knowledge is distinguished by its non-trivial nature that is, it is not limited to what is obvious and by its capacity to guide decisions based on empirical evidence. Consequently, data mining acts as a bridge between the wealth of information and strategic action, enabling organizations to anticipate scenarios, optimize resources, and develop sustainable policies.

Method

Design.

A quantitative, observational, and retrospective design for secondary analysis of administrative data was adopted, framed within an empirical-analytical paradigm and developed using the CRISP-DM methodology for data mining (business and data understanding, preparation, modeling, evaluation, and deployment) as a systematic and reproducible workflow. The project specifically includes data preparation (cleaning, dimensionality reduction) and experimentation with classifiers based on performance metrics, in accordance with the CRISP-DM stages described in the document. The analysis included descriptive statistics (distributions, homologations, and summaries by type of disability, ICD diagnosis, age/discretized age, marital status, and payroll class) and supervised modeling to predict duration (active/inactive) and detect patterns and trends by subgroups. The document reports, for example, the distribution of days of absences (median 7; p75=16.75; maximum 176; n=1,406) and the construction of the class variable; as well as analyses by age and by marital status, among others.

Participants.

The unit of analysis consisted of medical leave records for employees of the Academic Institution, regardless of their employment status (administrative staff; tenured, contract, and adjunct faculty; and employment contracts). These events are recorded and monitored by the Human Resources Department and processed through the EPS (Health Promoting Entities), in accordance with current regulations.

Inclusion criteria: Records from the KACTUS HCM human resource management and payroll system (SQL Server 2022 engine) that corresponded to absences due to illness with a diagnosis coded according

to ICD, within the period 2012–2024 (April), and that had the necessary fields for the analysis (e.g., days of incapacity, type/code of incapacity, diagnosis, sociodemographic and linkage variables).

Exclusion: In accordance with CRISP-DM, during data preparation, incomplete, inconsistent, duplicate records and outliers that would prevent modeling, as well as variables irrelevant to the objectives, were removed.

Sample size: the consolidated base for the study consisted of 1406 records of incapacities in the indicated period.

Instruments.

Primary data source: KACTUS HCM system supported on SQL Server 2022, which stores the different types of disabilities and ICD diagnoses reported by officials.

Key variables include, (among others): days of disability, diagnosis ID, and disability type/code; as well as age/discretized age, gender, marital status, payroll class, and identification and location variables (properly anonymized for analysis). The document presents the dictionary of variables and equivalencies (e.g., disability types and payroll classes).

Measurement instrument: institutional administrative record (disability events) with ICD diagnosis and approved disability typology (General Illness [GI], Maternity/Paternity Leave, Work Accident, Occupational Disease, etc.), which provides validity of health content and institutional coverage of cases.

Validity and reliability: Construct and content validity: supported by ICD standards for diagnoses and institutional homologations of disability types and payroll class; in addition, the selection of variables with SelectKBest reinforces the relevance of the predictor set with respect to the class.

Model reliability and robustness: supported by the use of multiple classifiers, K-fold cross-validation and evaluation with standard metrics (accuracy, recall, precision), plus a 70/30 split for independent testing.

Data quality: addressed during preparation (cleaning, handling of outliers and dimensionality reduction) according to CRISP-DM.

Results

The results obtained allow for a comprehensive characterization of absenteeism at the Academic Institution during the period 2012–2024, based on the analysis of 1,406 absentiism records. The analysis was structured following the phases of the CRISP-DM methodology, which facilitated the refinement and organization of the information, the selection of the most relevant variables, and the identification of descriptive patterns and significant trends in the data. These findings are presented at different levels: initially, the general characteristics of the database and the dependent variable, days of disability, are described; then, the distributions by type of disability, age, sex, marital status, payroll category, and municipality of birth are analyzed; and finally, the results derived from the data preparation and modeling are highlighted. This order of presentation ensures a coherent and progressive reading of the findings, by linking the descriptive aspects with the predictive elements that underpin the purpose of the research.

1. Dataset and Variable Selection: The analysis was performed on a consolidated database of 1,406 disability records reported at the Academic Institution between April 2012 and April 2024. The original dataset includes 26 categorical variables and 8 numerical variables. To guide the modeling, a feature importance filter (SelectKBest) was applied, which identified the 10 most informative variables, in descending order: diagnosis ID; disability code ID; age at disability; marital status; date of birth; gender; payroll class ID; municipality of birth ID; and discretized age. These variables guided the standardization stages and the creation of the target variable.

2. Dependent Variable: number of days absent (descriptive and discretization for the class). The number of days absent variable presents a positively skewed distribution. Descriptive summary: n = 1406; mean = 16.63 days; standard deviation = 27.02; median = 7 days; Q1 = 3; Q3 = 16.75; minimum = 1; maximum = 176. The interquartile range is 13.75; outliers were identified below −17.625 and above 37.375. Following the criteria of the Human Resources Management Department, the variable was discretized into Absence Classification: Low (≤ 7 days) and High (> 7 days), where the majority of events (75%) correspond to absences of 17 days or less.

3. Distribution by type of disability and annual trend: By type of disability, the vast majority of records correspond to IGE (Incapacity for Work): 1,255 out of 1,406 cases (≈ 89.2%). These are followed by maternity leave (≈ 6.7%) and work accidents (≈ 3%). Occupational disease and other types appear in very low proportions, but require monitoring due to their potential budgetary and preventative impact. Regarding the temporal dynamic, there is evidence of a drop in reports in 2020 (n = 48) probably associated with the pandemic and remote work arrangements and a rebound in 2023 (n = 183), suggesting changes in in-person work or a greater reporting culture within the institution.

4. Age and disability: When comparing age at the time of disability between the defined classes, records classified as High (absences >7 days) show a mean of 47 years (SD ≈ 11) and an IQR that places the central 50% between 37 and 56 years. For the Low class, the mean is 45 years (SD ≈ 12) with an IQR between 34 and 54 years. Both groups show wide variability (ranging from young adults to the elderly), although the trend indicates a greater concentration of prolonged absences in middle- to older-aged ranges.

5. Sex, marital status, and age (discrete): Between 2013 and 2023, 1,216 disabilities were recorded in the sub-analysis for that period; of these, approximately 59.2% corresponded to females and approximately 40.8% to males, with IGE (Individualized Geriatric Status) being the predominant category in both sexes. Regarding marital status, the single group accounted for the majority of reports (903 cases, 64.2%), followed by married individuals (298; 21.2%), those in common-law unions (123; 8.7%), divorced individuals (76; 5.4%), and widowed individuals (6 cases). The age (discrete) variable ("Young Adult," "Adult," "Older Adult," "Senior Person") shows that general illnesses predominated among adults and older adults, while maternity leave was concentrated among young adults, a pattern consistent with the demographic and reproductive distribution of the workforce.

6. Payroll Class (type of employment) and municipality of birth: When analyzing payroll class, the largest proportion of absences comes from administrative staff (724 records; 51.8% of the total), followed by contracted workers (≈25.3%), part-time teachers (13.4%), and tenured teachers (9.4%). This pattern suggests that the conditions and workloads of administrative and contract workers influence the frequency of absences. Regarding geographic origin, the municipality of Armenia has the highest number of cases (616 absences), followed by Calarcá (121) and La Tebaida (116), demonstrating a strong local presence among the absent personnel.

7. Preprocessing and variable preparation results: Before modeling, code standardization (diagnosis ID, disability code ID, payroll class ID, birth municipality ID) was performed to facilitate interpretation; discretizations (e.g., discretized age, disability classification), outlier handling, and inconsistencies were eliminated as part of the CRISP-DM preparation phase. The 10 variables selected by SelectKBest guided the pre-training dimensionality reduction.

Figure 1 shows that absenteeism is mainly associated with general illness, whereas work related and occupational causes account for a relatively small proportion of cases.

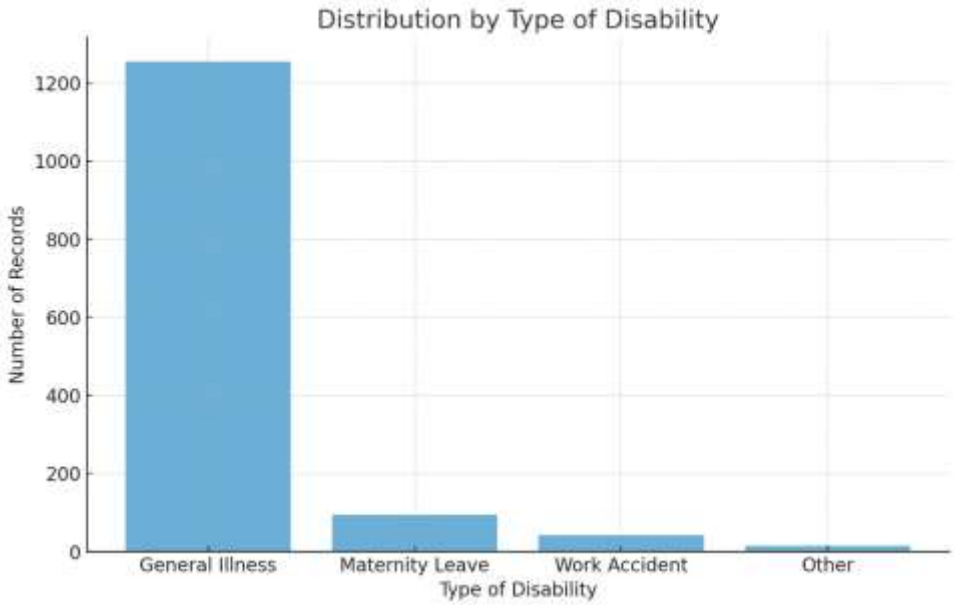

Figure 1. Distribution by Type of Disability among employees of Academic Institution.

Figure 2 indicates that absenteeism varies across payroll classes, with administrative staff showing the highest incidence and tenured faculty the lowest.

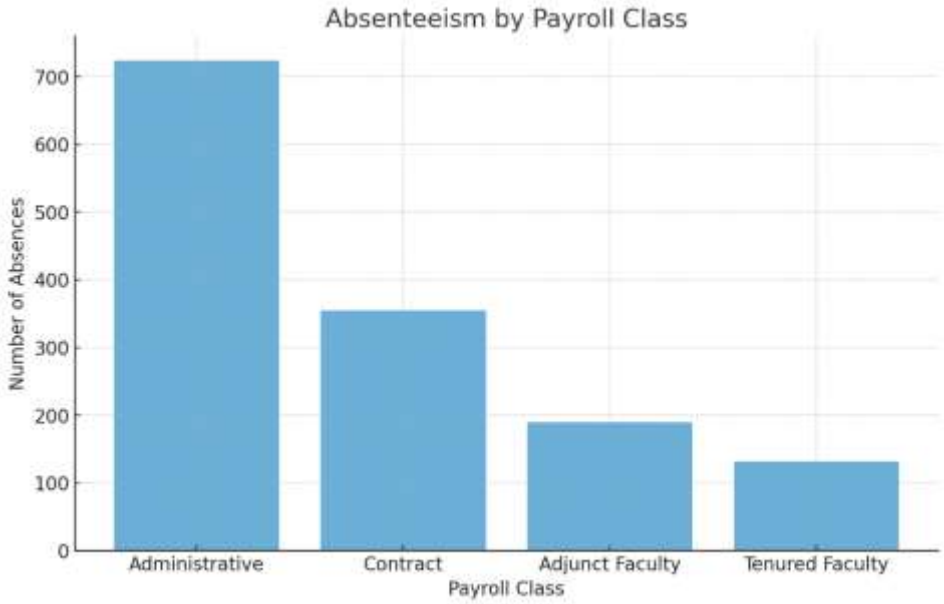

Figure 2. Absenteeism by payroll class in employees of Academic Institution.

To predict the duration of medical incapacity, several supervised classification models were implemented, including decision trees, Naïve Bayes classifiers, support vector machines (SVM), and artificial neural networks. Decision trees classify instances by recursively partitioning the data based on the predictor that maximizes information gain, producing an interpretable hierarchical structure. Naïve Bayes applies Bayes' theorem under the assumption of conditional independence among predictors, allowing efficient probabilistic classification. Support vector machines (SVM) construct an optimal separating hyperplane that maximizes the margin between classes, enhancing generalization in high-dimensional spaces. Artificial neural networks model nonlinear relationships through interconnected layers of neurons trained by weight

optimization. Together, these techniques enable the identification of patterns associated with short and long duration incapacities and support the construction of predictive models aligned with the CRISP-DM methodology.

These classification algorithms were trained and evaluated within the CRISP-DM modeling stage to determine their effectiveness in distinguishing between short and long duration absences. The accuracy results obtained for each model are presented in Table 1, providing evidence of their predictive capacity and identifying the approaches best suited to support institutional decision-making. The accuracy results indicate that the Random Forest model achieved the highest predictive performance (0.758), outperforming the Decision Tree (0.677), Artificial Neural Network (0.554), and Naïve Bayes classifier (0.503). These findings suggest that ensemble-based methods may capture more complex interactions among the predictors compared to single-tree or probabilistic approaches. Overall, the results confirm that supervised classification techniques are capable of distinguishing between short- and long-duration medical incapacities, with the Random Forest model emerging as the most reliable option for supporting institutional decision-making within the CRISP-DM analytical framework.

Table 1. Model Accuracy

Decision Tree	Naïve Bayes	Random Forest	Artificial Neural Network
0,677	0,503	0,758	0,554

Discussions

The study on absenteeism at the Academic Institution reveals that general illness is the main cause of disability (≈ 89%), in line with what has been indicated by Angarita et al. (2024) and Arrieta-Burgos et al. (2025) in their reports on the Colombian context. However, the prevalence level identified in the educational institution is higher than in other sectors, raising doubts about the effectiveness of the promotion and prevention programs implemented in the university setting. This contrast demonstrates that, while some private companies have strengthened their occupational health strategies, gaps persist in the public sector that explain the magnitude of the phenomenon. Another finding that generates debate is the concentration of absenteeism among administrative workers, who account for more than half of the sick leaves (51.8%). This pattern partially coincides with that reported by Barrera-Sigua et al. (2023) and Calle and Gil (2025) in small and medium-sized enterprises, where contractual factors directly influence absenteeism rates. However, this differs from what has been documented by Duke and Valencia (2021) in the health sector where healthcare workers were the most affected.

These contrasts highlight that the causes of absenteeism depend, to a large extent, on specific occupational conditions, which reinforces the claims of Boada et al. (2005) Regarding the relationship between job demands and the frequency of absences, the research indicates that, in the university context, the sedentary nature of the work, the bureaucratic burden, and administrative pace appear to be more significant triggers than teaching duties, directly challenging the institution's internal management. With respect to sociodemographic characteristics, the research shows that prolonged absences are concentrated among middle-aged and older individuals, a finding that aligns with previous studies. Mamani (2023) and Rivero et al. (2025) point to the relationship between aging in the workforce and greater vulnerability to long-term disabilities. However, the prevalence among single people contrasts with what has been suggested by Pulido et al. (2021), who highlight the weight of psychosocial factors as modulators of absenteeism.

This dynamic raises an important question: to what extent do social support networks or structural working conditions influence the duration of absences? Although the data presented here do not allow for a definitive conclusion, they do highlight the value of future studies incorporating qualitative methodologies to further explore workers' perceptions, experiences, and motivations.

Furthermore, the implementation of data mining using the CRISP-DM methodology proved successful, allowing for the identification of the most influential variables and the structuring of the analysis. This approach has been validated byCorzo (2025) and Wirth & Hipp (2024) as a standard in data analytics, and coincides with the proposals of Cabrera (2024) and Díaz et al. (2025) regarding the importance of preprocessing and data quality. Unlike studies such as those of Araujo et al. (2019) and Guerrero et al. (2025) that contrast other studies that prioritize the predictive capacity of models, this research focused on identifying explanatory relationships between clinical, contractual, and sociodemographic variables, providing a more interpretive than strictly technical approach.

Therefore, this decision represents a contribution from the researcher, as it emphasizes that understanding the phenomenon is as important as anticipating it, in contrast to purely quantitative approaches such as those of Skorikov et al. (2020) In addition to this, the results reaffirm that absenteeism is a complex and multidimensional phenomenon in line with Chiavenato's organizational and human talent management theory (2006, 2008, 2009a, 2009b) Furthermore, they strengthen the legitimacy of the use of data mining as a tool for addressing problems in the social sciences, as has been proposed by Alujah (2001) and Marcano and Talavera (2007) by demonstrating its capacity to generate useful knowledge from administrative records. Therefore, the Academic Institution is called upon to implement an absenteeism monitoring and surveillance system based on data analytics, following experiences such as those of Rodríguez (2025) and Perez et al. (2024).

Of course, this system should integrate dynamic dashboards, preventative strategies for administrative staff and middle-aged adults, as well as self-care promotion programs focused on reducing general illnesses. Under this logic, human talent management must evolve from a record-keeping approach to a preventative and strategic model, consistent with Chiavenato's vision (2008, 2009a) on the centrality of human resources in organizational productivity. In this regard, the distinctive contribution of this study lies in demonstrating that absenteeism in the Colombian university context has its own profile, differing from other economic sectors and therefore requiring specific intervention measures. At the same time, it demonstrates the potential of data mining as a management tool in higher education, adding to the arguments ofCedillo et al. (2024) and Anchundia (2025)on the possibilities of these techniques in that sector.

Regarding future projections, it is suggested to advance longitudinal research that allows observation of the evolution of absenteeism over longer periods, complement the analyses with qualitative methods to explore workers' perceptions, implement advanced predictive models that improve the ability to anticipate critical cases and calculate the economic costs of absenteeism, following the line of Béjar et al. (2025) and Grijalba and Riascos (2025). In summary, this discussion allows us to answer the research question: What trends and patterns can be identified in employee absenteeism at the Human Resources Department of the Academic Institution, based on data analysis using the CRISP-DM methodology, that provide strategic input for strengthening the institution's human talent management? This facilitated the achievement of the overall objective and, furthermore, provided strategic input that, if utilized, will allow the institution to design more effective, preventative management policies adapted to its specific context.

Conclusions

This study demonstrates that the Random Forest classifier outperforms Decision Tree, Naive Bayes, and Multilayer Perceptron models in terms of classification accuracy for the considered dataset. The results highlight the suitability of ensemble-based tree models for handling complex, nonlinear relationships in tabular data, particularly in scenarios with limited data availability.

Furthermore, the findings confirm that simpler probabilistic classifiers may fail when strong feature dependencies are present, while neural network models require careful architectural design and sufficient data to achieve competitive performance. Consequently, Random Forest emerges as the most appropriate model for the problem under analysis, providing a favorable balance between predictive accuracy and generalization capability.

The analysis of absenteeism at the Human Resources Department of the Academic Institution identified significant patterns and trends that confirm the multifactorial and multidimensional nature of the phenomenon. The findings highlight that absenteeism stems from the interaction of clinical, sociodemographic, and contractual variables, which together create distinct risk profiles. This comprehensive understanding demonstrates that absenteeism should be interpreted as an indicator of institutional health, not merely as an administrative record, making a substantial contribution to the strategic management of human talent.

Thus, this research demonstrated that data mining is an interpretive resource for understanding underlying relationships that traditional approaches fail to reveal. Consequently, this study reaffirms that data analysis is a strategic tool for university management, offering reliable and actionable information. However, the most significant contribution of this work lies in having transformed available information into strategic inputs for institutional planning. Recognizing that absenteeism is concentrated in particular groups such as administrative staff and middle or older aged employees provides a strong basis for developing differentiated, targeted prevention and support policies. Therefore, these inputs encourage a shift from a reactive model, focused on justifying absences, to a preventive model that prioritizes health, promotes well-being, and fosters the sustainability of human talent as the organization's core resource.

In addition, it is important to conclude that the use of data mining for work force management presents an ethical and organizational challenge. This use must be accompanied by principles of transparency, data protection, and worker participation to avoid excessive control or the stigmatization of high-risk profiles. In this regard, the research urges the Academic Institution to consolidate a culture of responsible analysis that balances the strategic use of information with the safeguarding of individual rights. Finally, the study opens avenues for future research aimed at exploring absenteeism from longitudinal and mixed-methods perspectives, integrating not only administrative records but also workers' perceptions and associated economic costs. This would contribute to strengthening the academic literature and designing replicable models for other higher education contexts.

Future Research Directions

Future research could broaden the analytical framework by incorporating additional contextual and organizational variables that may influence the duration of medical incapacity. Factors such as job demands, occupational exposure, or seasonal health trends could enrich the predictive space and allow the classification models (decision trees, Naïve Bayes classifiers, support vector machines (SVM), and artificial neural networks) to capture subtler patterns associated with absenteeism. Expanding the temporal and structural dimensions of the dataset would also enable longitudinal analyses capable of identifying emerging tendencies and shifts in workforce health profiles.

Building on this expanded analytical foundation, future studies may explore advanced modeling strategies, such as ensemble learning or hybrid approaches, to assess whether combining techniques enhances predictive accuracy. Furthermore, developing automated reporting tools or monitoring dashboards could translate the analytical insights into real-time support for institutional decision-making, strengthening early detection of risk groups and informing preventative interventions. Together, these extensions would consolidate a more comprehensive and data-driven approach to understanding absenteeism in academic environments.

References

1. Angarita, A., Bedoya, S., & Álzate, M. (2024). Absenteeism profile of workers in a Colombian company in the retail sector.Journal of the Spanish Association of Specialists in Occupational Medicine, 33(4), 435-448. Retrieved from: https://scielo.isciii.es/pdf/medtra/v33n4/3020-1160-medtra-33-04-435.pdf

2. Aluja, T. (2001). Data mining, between statistics and artificial intelligence.Question, 25(3), 479-498. Retrieved from:https://dialnet.unirioja.es/servlet/articulo?codigo=2364961
3. Anchundia, E. V. (2025).Analysis of the data mining-based model to determine student dropout factors at the Southern State University of Manabí(Undergraduate thesis). Southern State University of Manabí, Manabí: Ecuador.
4. Araujo, V., Rezende, T., Guimarães, A., Silva, V., Campos, P. (2019). A hybrid approach of intelligent systems to help predict absenteeism in companies. Applied. Sciences, 1(536). DOI: https://doi.org/10.1007/s42452-019-0536-y
5. Armero, A. A., Becerra, P. A., & Mora, E. A. (2023).Proposal for improvement in data mining and service management in the operations division of the company SAYCO(Undergraduate thesis). ECCI University, Bogotá: Colombia.
6. Arrieta-Burgos, E., Sepúlveda, C., Hurtado, I., Restrepo, J., & Jaramillo, T. (2024).Work absenteeism and medical incapacities 2022Bogotá: Center for Social and Labor Studies of the National Association of Businessmen of Colombia ANDI.
7. Arrieta-Burgos, E., Sepúlveda, C., Restrepo, J., & Jaramillo, T. (2025).Work absenteeism and medical incapacities 2023Bogotá: Center for Social and Labor Studies of the National Association of Businessmen of Colombia ANDI.
8. Barrera-Sigua, Y., Payares-Celins, L. J., Estupiñán-Fernández, D. A., Ordoñez-López, S. Y., Malaver-Cardenas, J. A., & Monsalve-Jaramillo, E. (2023). Factors related to absenteeism in a Colombian mining company.Journal of Health Research. University of Boyacá, 10(1), 45-57. DOI: https://doi.org/10.24267/23897325.915
9. Bayram, M., & Burgazoglu, H. (2020). The Relationships Between control measures and absenteeism in the context of internal control. Safety and Health at Work, 11(4), 443-449. DOI: https://doi.org/10.1016/j.shaw.2020.07.007
10. Becerra, J. P., García, Ó. A., Velásquez, E. A., Villareal, I. L., Vásquez-Trespalacios, E. M. (2024). Absenteeism due to medical reasons in virtual and alternating work modalities in workers of a Colombian company.Journal of the Spanish Association of Specialists in Occupational Medicine, 33(1), 74-84. Retrieved from: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S3020-11602024000100008
11. Béjar, V., Madrigal, F., & Madrigal, S. (2025). Factors that influence absenteeism and its economic impact on organizations.LATAM, 6(1), 3022-3032. DOI: https://doi.org/10.56712/latam.v6i1.3555
12. Berón, E., Mejía, D., & Castrillón, Ó. (2021). Main causes of work absenteeism: an application from data mining.Technological information, 32(2), 11-18. DOI: http://dx.doi.org/10.4067/S0718-07642021000200011
13. Boada, J., Vallejo, R. D., Agulló, E. & Mañas, M. Á. (2005). Work absenteeism as a consequence of organizational variables.Psychothema, 17(2), 212-218. Retrieved from: https://www.redalyc.org/pdf/727/Resumenes/Resumen_72717205_1.pdf
14. Borda, M. J., Rolón, E., Díaz-Piraquive, F. N., & González, J. (s.f.). Work absenteeism: impact on productivity and control strategies from corporate health programsRetrieved from: https://repository.urosario.edu.co/server/api/core/bitstreams/6d970520-6b36-4ad0-bcc9-45702bbb9b7e/content
15. Cabrera, J. F. (2024). Improving data quality: a perspective on preprocessing techniques for data mining. (Doctoral thesis). University of Cádiz, Cádiz: Spain.
16. Calle, A., Moreira, I., Quimis, S., & Yoza, R. (2024). Data mining with a focus on measurement in consumer behavior.Science and Development, 27(1), 173-182. Retrieved from: http://revistas.uap.edu.pe/ojs/index.php/CYD/index
17. Calle, P. T., & Gil, H. A. (2025). The Impact of absenteeism in small and medium-sized enterprises in the city of Bogotá.Studies and perspectives, 5(1), 3431-3445. DOI: https://doi.org/10.61384/r.c.a..v5i1.1063

18. Cedillo, J. M., Beltrán, H. M., Saltos, M. I., & Soriano, F. (2024). Exploring data mining in higher education management: challenges and opportunities in the digital age.Reincisol, 3(5), 1367-1385. DOI: https://doi.org/10.59282/reincisol.V3(5)1367-1385
19. Chiavenato, I. (2006).Introduction to the general theory of administration. México, D.F.: McGraw-Hill.
20. Chiavenato, I. (2008). Human talent management. Mexico City: McGraw-Hill.
21. Chiavenato, I. (2009a).Human resource management. The human capital of organizations. México, D.F.: McGraw-Hill.
22. Chiavenato, I. (2009b). COrganizational behavior. The dynamics of success in organizations. México, D.F.: McGraw-Hill.
23. Cifuentes, P. A., Ortega, D., & Rojas, L. (2020). Absenteeism in the construction sector: From self-care to excellence, a path towards quality of life(Undergraduate thesis). CES University, Medellín: Colombia.
24. Cotta, J. G. (2025). Improving the description of open educational resources, using techniques based on Artificial Intelligence, machine learning and data mining. (Master's thesis). Francisco José de Caldas District University, Bogotá: Colombia.
25. Corzo, S. (2025). Setting a standard: CRISP-DM and artificial intelligence.Neuronum, 11(3), 30-32. Retrieved from: https://eduneuro.com/revista/index.php/revistaneuronum/article/view/579
26. Dávila, F., & Sánchez, Y. (2012). Data mining techniques applied to the diagnosis of clinical entities.RCIM, 4(2), 174-183. Retrieved from: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18592012000200007
27. Díaz, M., Caballero, A., Caballero, A., Pérez, R., & Pérez, R. (2025). Techniques to guarantee data quality in an intensive care service software application: the key to reliable data mining.Cuban Journal of Computer Science, 19(1), 20-45. Retrieved from: https://rcci.uci.cu/index.php/RCCI/article/view/13018
28. Duque, L., & Valencia, A. (2021). Impact of absenteeism in health sector organizations in the city of Medellín for the year 2020.CIES, 12(2), 287-301.
29. García, J. C., Cruz, J. A., & Villarreal, H. Z. (2023) Main causes of work absenteeism in nursing staff assigned to the General Hospital of Zone No. 3 Tuxtepec, Oaxaca.Latam, 6(6), 1183-1195 DOI: https://doi.org/10.56712/latam.v4i6.1515
30. García, V. H., & Martínez, R. (2016). Absenteeism and health: a study of its importance in teleworking.Straight, 11(1), 13-25. DOI: https://doi.org/10.17163/ret.n11.2016.01
31. González, A. A. (2025). Zero digital competence: training needs via data mining towards an innovative and disruptive digital training system.Pixel-Bit, 73(4), 1-30. DOI: https://doi.org/10.12795/pixelbit.108664
32. Grijalba, A. J., & Riascos, A. M. (2025). Direct costs of work absenteeism due to medical incapacity in healthcare workers at the Proinsalud Clinic, Pasto-Nariño, 2023(Master's thesis). Mariana University, San Juan de Pasto, Nariño: Colombia.
33. Guerrero, L. M., Moreno, N., & Molina, J. D. (2025). Prediction of work absenteeism due to disability: an approach from generalized linear models.THIS IS IT, 22(44), 1-27. DOI: https://doi.org/10.24050/reia.v22i44.1890
34. Lawrance, N., Petrides, G., & Guerry, M-A. (2021). Predicting employee absenteeism for cost effective interventions. Decision Support Systems, 147, 1-10 DOI: https://doi.org/10.1016/j.dss.2021.113539
35. Luna, R. M., & Brokate, F. J. (2014). Characterization of absenteeism among nursing staff at the University Hospital of the Caribbean in Cartagena during the period from September 2012 to September 2013(Specialization thesis). University of Cartagena, Cartagena: Colombia.
36. Mamani, N. (2023).Absenteeism and its relationship with staff productivity in the district municipality of Kelluyo, Chucuito province and Puno Region, 2023(Undergraduate thesis). National University of the Altiplano, Puno: Peru.
37. Marcano, Y. J., & Talavera, R. (2007). Data mining as support for business decision making.Option, 23(52), 104-118. Retrieved from: https://www.redalyc.org/pdf/310/31005208.pdf

38. Marulanda, C. E., López, M., & Mejía, M. H. (2017). Data mining in knowledge management of SMEs in Colombia.Virtual Journal of the Catholic University of the North, (50), 224-237. Retrieved from: https://www.redalyc.org/pdf/1942/194250865013.pdf
39. Mendoza, C. (2024).Analysis of absenteeism: causes, impact and management strategies in companies(Specialization thesis). National Unified Corporation of Higher Education (CUN), Bogotá: Colombia.
40. Mesa-Mesina, F., Quevedo-León, R., & Espinoza-Téllez, T. (2025). Work absenteeism due to the use of medical leave in four economic sectors in Peru, 2000-2025.Journal of the University of Zulia, 16(47), 58-76. DOI: https://doi.org/10.5281/zenodo.17058338
41. Morales, J., Romero-Torres, J., & Gutiérrez, D. (2025). Use of data mining to evaluate the quality and mobility of university student transport.Edges, 12(20), 87-93. Retrieved from: http://revistaaristas.tij.uabc.mx/index.php/revista_aristas/article/view/406/396
42. Muñoz, S. R. (2020). Design of models for predicting absenteeism and lateness for a human resources consulting firm(Undergraduate thesis). University of Chile, Santiago, Chile: Chile.
43. Morquera, N. (2017).Factors that influence absenteeism and its impact on the organizational climate(Master's thesis). Nueva Granada Military University, Bogotá: Colombia.
44. Mullo, A., Reinoso, J., Chamba, M., & Lozada, C. (2025). Analysis and characterization of power quality using data mining.Energy, (22), 33-45. DOI: https://doi.org/10.37116/revistaenergia.v21.n2.2025.702
45. Noroña, N., Cajas, E., Chamba, M., & Lozada, C. (2025). Transient stability analysis using the concept of inertia and data mining.Energy, (22),1-11 DOI: https://doi.org/10.37116/revistaenergia.v21.n2.2025.700
46. Orozco, W., Villao, A., Orozco, J., & Villarroel, M. (2021). Application of data mining techniques to predict the academic performance of students at the 'Lic. Angélica Villón L.' school.UPSE, 8(2), 68-75. DOI: 10.26423/rctu.v8i2.637.
47. Ortiz, D., Vallejo, D. C., Fajardo, I. B., Mejía, J. F., & Tobar, M. M. (2021). Administrative management of absenteeism in diagnostic support services at the Nariño Departmental University Hospital(Specialization thesis). Pasto, Nariño: Colombia. Catholic University of Manizales.
48. Pedraza, M. A. (2021).Analysis of absenteeism and its impact on productivity at Offibank Y Cia SAS, 2018-2019(Specialization thesis). Francisco José de Caldas District University, Bogotá: Colombia.
49. Pérez, B. J., Medina, L. M., Rolón, G. C., & Contreras E. C. (2024). Absenteeism and presenteeism at work: impact on productivity and control strategy through a health prevention and promotion program.Fesc World, 14(28), 115-134. DOI: https://doi.org/10.61799/2216-0388.1590
50. Pozo-Pozo, D. A., & Espinosa-Tigre, R. M. (2025). Organizational factors and their influence on absenteeism among workers at a health center in Tulcán, Ecuador.MQRInvestigar, 9(1), 1-33. DOI: https://doi.org/10.56048/MQR20225.9.1.2025.e327
51. Pulido, E., Lora, L., & Jiménez, L. K. (2021). Psychosocial factors that influence absenteeism: evaluation of an explanatory model.Interdisciplinary, 38(1), 149-162. DOI: https://doi.org/10.16888/interd.2021.38.1.10
52. Risco-Ramos, R., Pérez-Aguilar, D., Casaverde-Pacherrez, L., Malpica, M., Pérez-Aguilar, J., & Pérez-Aguilar, A. (2023). Use of a business intelligence and data mining framework as computational tools in SMEs: Production forecasting of a hydroelectric power plant as a case study. Laccea, 1-8. Retrieved from: https://laccei.org/LACCEI2023-BuenosAires/papers/Contribution_832_a.pdf
53. Rivero, R. R., Velazco, A. R., Rivera, S. D., Barriga, M. C., & Cueva, C. H. (2025). Association between socio-labor factors and absenteeism in workers of the agro-export sector.University, Science and Technology, 29, 341-348. DOI: https://doi.org/10.47460/uct.v29ispecial.939
54. Rodríguez, D. P. (2025).Design of a data analytics tool for measuring and managing work absenteeism due to medical reasons in the official fire department of Bogotá(Specialization thesis). National Unified Corporation of Higher Education (CUN), Bogotá: Colombia.
55. Rodríguez, J. E. (2013). Development of tools for data mining "UDMINER".Links, 9(1), 21-40. DOI: https://doi.org/10.14483/2322939X.4207

56. Ruiz, V. V., & Armoa, A. (2023). The importance of data mining as a strategic tool in companies.Latin American Science, 7(1), 9267-9276. DOI: https://doi.org/10.37811/cl_rcm.v7i1.5119
57. Saldarriaga, J. F., & Martínez, E. (2008). Factors associated with work absenteeism due to medical reasons in a higher education institution.Journal of the National Faculty of Public Health, 25(1), 32-39. DOI: https://doi.org/10.17533/udea.rfnsp.207
58. Sánchez, D. C. (2015). Work absenteeism: a view from the management of occupational safety and health.Forest Health, 5(1) 43-54. Retrieved from: https://revistas.unbosque.edu.co/index.php/RSB/article/view/182/114
59. Sánchez, M. G., & Pérez, G. Á. (2021). CRISP-DM methodology in the management of Data Mining projects. Case of dermatological diseases.International Conference on Project Management 2021Retrieved from: https://repository.universidadean.edu.co/server/api/core/bitstreams/d5d07653-c175-4bfe-9745-d8d9e8cc16ad/content
60. Seclen, C. A. (2025). Intelligent system based on data mining to predict fertilizer production at Nutrition Vegetable Corporation Figal SAC(Undergraduate thesis). Catholic University Santo Toribio de Mogrovejo, Chiclayo: Peru.
61. Skorikov, M., Hussain, M. A., Khan, M. R., Akbar, M. K., Momen, S., Mohammed, N., & Nashin, T. (2020). Prediction of Absenteeism at Work using Data Mining Techniques. 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 1-6. DOI: 10.1109/ICITR51448.2020.9310913.
62. Tatamuez-Tarapues, R. A., Domínguez, A. M., & Matabanchoy-Tulcán, S. M. (2018). Systematic review: factors associated with absenteeism in Latin American countries.University and Health, 21(1), 100-112. DOI: http://dx.doi.org/10.22267/rus.192101.143
63. Wirth, R., & Hipp, J. (s.f.). CRISP-DM: towards a standard process model for data mining. Recuperado de: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf
64. Zhang, Y. (2021). Sales Forecasting of Promotion Activities Based on the Cross-Industry Standard Process for Data Mining of E-commerce Promotional Information and Support Vector Regression. Journal of Computers, 32(1), 212-225 DOI: doi:10.3966/199115992021023201018
65. Zupančič, P., & Panov, P. (2024). Predicting employee absence from historical absence profiles with machine learning. Applied. Sciences, 14(16), 1-29. DOI: https://doi.org/10.3390/ app14167037