Salud, Ciencia y Tecnología. 2026; 6:2446
doi: 10.56294/saludcyt20262446

ORIGINAL

Method with machine learning to carry out feasibility in data mining projects

Método con aprendizaje automático para realizar factibilidad en proyectos de minería de datos

Juan Camilo Giraldo Mejia[1], Fabio Alberto Vargas Agudelo[2], Jorge Guadalupe Mendoza León[2]

[1]Tecnológico de Antioquia – Institución Universitaria. Antioquia, Medellín, Colombia.
[2]Instituto Tecnológico de Sonora- ITSON. Estado de Sonora, México.

Cite as: Giraldo Mejia JC, Vargas Agudelo FA, Mendoza León JG. Method with machine learning to carry out feasibility in data mining projects. Salud, Ciencia y Tecnología. 2026; 6:2446. https://doi.org/10.56294/saludcyt20262446

Submitted: 31-07-2025 **Revised:** 03-10-2025 **Accepted:** 16-11-2025 **Published:** 01-01-2026

Editor: Prof. Dr. William Castillo-González

Corresponding Author: Juan Camilo Giraldo Mejia

ABSTRACT

Currently, there are Data Mining techniques aimed at increasing the accuracy of the information and the agility in the analysis. These are applied in the productive sector to characterize behaviors, based on the discovery of knowledge and, in this way, base decision-making in real and dynamic situations. Artificial intelligence (AI) drives research methods and data mining techniques for knowledge acquisition. For its use, the life cycle of data mining projects is followed, which involves stages of extraction, cleaning, preparation and transformation, modeling, and data evaluation. However, it is important to consider a feasibility study for data mining projects, with the objective of positively impacting organizations, by minimizing costly errors and guaranteeing an efficient distribution of resources, as well as the decision on the continuity of a project. This article presents a Machine Learning method to carry out feasibility in data mining projects, seeking to impact organizations by minimizing costly errors and guaranteeing an efficient distribution of resources.

Keywords: Data Mining; Machine Learning; Feasibility in Projects; Knowledge Discovery.

RESUMEN

En la actualidad, existen técnicas de Minería de Datos destinadas a incrementar la exactitud de la información y la agilidad en el análisis. Estas se aplican en el sector productivo para caracterizar comportamientos, apoyándose en el descubrimiento de conocimiento y, de esta manera, fundamentar la toma de decisiones en situaciones reales y dinámicas. La inteligencia artificial (IA) impulsa métodos de investigación y técnicas de minería de datos para la adquisición de conocimiento. Para su uso, se sigue el ciclo de vida de los proyectos de minería de datos, el cual involucra etapas de extracción, limpieza, preparación y transformación, modelado y evaluación de datos. Sin embargo, es importante considerar un estudio de factibilidad para los proyectos de minería de datos, con el objetivo de impactar positivamente en las organizaciones, mediante la minimización de errores costosos y la garantía de una distribución eficiente de recursos, así como de la decisión sobre la continuidad de un proyecto. Este articulo presenta un método de Aprendizaje Automático para realizar factibilidad en proyectos de minería de datos, buscando impactar en las organizaciones minimizando los errores costosos y garantizando una distribución eficiente de los recursos.

Palabras clave: Minería de Datos; Aprendizaje Automático; Factibilidad en Proyectos; Descubrimiento de Conocimiento.

INTRODUCTION

The purpose of data mining is to extract hidden patterns from large volumes of data and derive relevant conclusions that support organizational decision-making. In turn, artificial intelligence drives research methods and data mining techniques to acquire knowledge.[1]

Currently, data mining techniques are designed to improve the accuracy of information and the agility of analysis.[2] These are applied in the productive sector for behavior analysis, supported by knowledge discovery, to inform decision-making in real, dynamic situations. They enable organizations to obtain valuable insights from large amounts of data across various sources, improving operational and competitive effectiveness. Among the most common techniques are classification, regression, clustering, and associative rule extraction,[3] which enable organizations to anticipate trends, detect errors, make informed decisions, and improve profitability. In addition, they make it easier to identify hidden data, draw valuable conclusions, and guide strategic decisions in the business and scientific contexts.[4] Data mining also integrates statistics, artificial intelligence, and databases, and obtains information from various sources. Its application is broad, spanning sectors such as education, industry, finance, technology, markets, and health, and contributing to behavior prediction, resource optimization, and risk prevention.[5] To implement these techniques, the life cycle for data mining projects is followed, which includes extraction, cleaning, preparation and transformation, modeling, and data evaluation. The data preparation and cleaning phases, as well as the selection of methods and techniques, are essential for responding to specific query requirements from extensive data collections.[6] However, it is necessary to conduct a feasibility study for this type of project, as it enables decisions on its continuation, thereby minimizing costly errors and ensuring efficient resource allocation.[7] It should be noted that methodologies for data mining projects do not always adequately manage implementation elements, as technological, organizational, and data aspects need to be specified more effectively.[8] Data mining helps extract hidden patterns from large datasets, but it poses challenges for project management and administration. A data mining project cannot be successful without structure and purpose, which involves conducting a technical, economic, and operational feasibility study. It is essential to analyze financial, technical, operational, legal, and programming elements to propose a method that supports decision analysts in estimating and assessing the feasibility of data mining projects, as well as in identifying potential risks or challenges.[9,10]

In companies, technological tools play an essential role in meeting project objectives, which require technical resources such as information systems that are crucial for executing activities, as well as modeling and analyzing data in an agile manner. Physical resources are also necessary, including equipment with processing capacity and servers capable of supporting large amounts of information. For organizations to conduct feasibility studies on mining and data centers, it is necessary to determine the business process, identify weaknesses and reliable sources, innovate and adopt new practices, establish a strategic team, incorporate modularity, and define decision-making criteria. Operational feasibility determines whether a project can be sustained over time and be successful in the organization. To do this, it is necessary to involve a team trained to operate and analyze data using the various technological tools implemented.[10]

Likewise, any data mining project requires a financial investment, which is why the initial and operating costs, expected economic benefits, and process optimization must be validated, as the project is not feasible if the expenses exceed the return on investment. This implies having complete, sufficient, and precise information, which, in turn, requires a cleansing process to minimize, or preferably eliminate, unnecessary data. It is recommended to improve processes and use specific techniques when establishing the particular requirements of the data mining project, taking into account the size of the data, the scope and complexity of the project, as well as the available resources, to facilitate the design, execution, and monitoring of workflows.[11] Data mining brings significant benefits to organizations through process improvement; however, it also poses challenges related to data quality, system scalability, and ethical issues. Existing methodologies for data mining projects are often applied without significant changes or extensions, including the addition of new phases or stages, to address technological and organizational needs. Therefore, it is suggested that they be optimized by incorporating the necessary elements and resources at the technological, operational, managerial, and data levels.[11]

METHOD

Step 1—Review of characteristics and stages of methodologies for data mining projects

The characteristics of three of the most widely used methodologies for developing data mining projects were reviewed. This was done to understand the dynamics of each in its different phases or stages.

SEMMA Methodology

It is one of the most widely used methodologies in data mining projects. It is structured around five stages: sampling, which involves selecting a sample from a specific source. The second stage, known as exploration, allows us to understand the structure of the data obtained previously. Subsequently, in stage three, the data is

prepared and transformed. Once the data has been prepared, it moves to the modeling stage, where algorithms are selected and applied to generate predictive or descriptive models. The stage of evaluating the functionality and usefulness of data mining models.[12,13] This methodology is also applied in a variety of contexts.[14,15]

CRISP-DM Methodology

The CRISP-DM methodology arose from the need to establish a standard process model.[16,17] It consists of six stages. In stage one, understanding the business, the organization's purposes, and the scope of the process indicators are established. In stage two, data is obtained from different sources. In stage three, data preparation, data cleaning, and transformation are performed. In stage four, mining techniques are applied to generate models that explore the articulation between data. In stage five, the models obtained are evaluated based on their performance. The last stage is called deployment; the models generated are used to benefit the organization by incorporating them into its dynamics.

Catalyst Methodology

It was validated by the scientific and industrial communities for the development of data mining projects, having been created around 2000, at the same time as CRISP-DM.[18,19] Six stages support it: understanding the business, during which the project's scope is specified. Stage two is understanding the data, during which the data sources are identified. In stage three, the data is prepared, and activities are carried out to ensure its quality. In phase four, modeling is carried out using appropriate tools and techniques to achieve the data mining objectives. Stage five corresponds to the evaluation of the models obtained. Stage six is called implementation, in which reports are generated.[18,19,20]

Step 2- Analysis of characteristics and activities of the methodologies

The three methodologies reviewed are widely used to develop data mining projects, each with specific phases that guide best practices for generating predictive or descriptive models that respond to organizational needs. However, these methodologies do not include activities related to feasibility or viability analyses of data mining projects. They do not indicate activities that encourage learning about and establishing the technical resources necessary to carry out the project, nor do they address aspects related to the project's economic and operational feasibility.

Table 1. Comparison of activities carried out with the reviewed methodologies

Methodology	Activities				
	Understanding the organization and the process	Knowledge and preparation of data	Modeling	Evaluation and Deployment	Technical, economic, and operational feasibility
SEMMA	Not applicable	Performed	Performed	Performed	Not performed
CRISP-DM	Performed	Performed	Performed	Performed	Not performed
CATALYST	It is performed	Performed	Performed	Performed	Not performed

The methodologies are clear and detailed in specifying the scope of the projects, based on the selection and description of the organizational processes involved in the data mining analysis. In addition, the study's purposes and the process indicators defining the project's scope are established.

Step 3- Specification of the phases for the proposed method

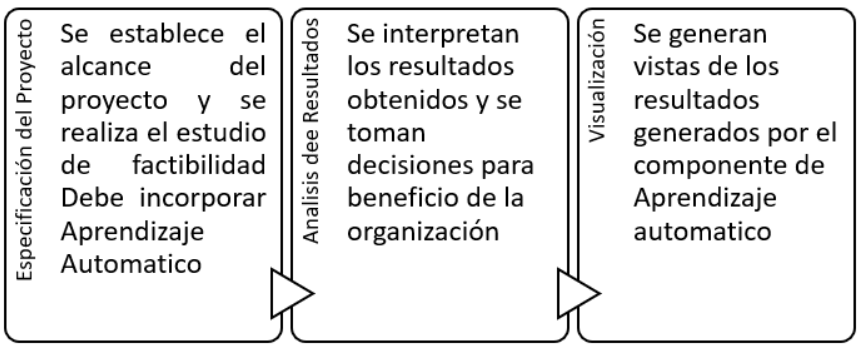

Figure 1. Phases of the method for data mining project feasibility

The method must be able to be articulated with the phases of any of the methodologies used for data mining projects, such as CRISP-DM, SEMMA, Catalyst, KDD, and Kimball, among others. The method is structured into three phases. The first is called project specification, the second phase is results analysis, and the third is visualization.

Step 4- Specification of method components

In the case of the CRISP-DM and SEMMA methodologies, each method is linked to the phases of business identification or business knowledge. If the methodology does not have a phase or component in which the business is contextualized, the proposed method is integrated as a new stage or phase of the methodology, being phase one or the initial phase of the method, as shown in figure 2.

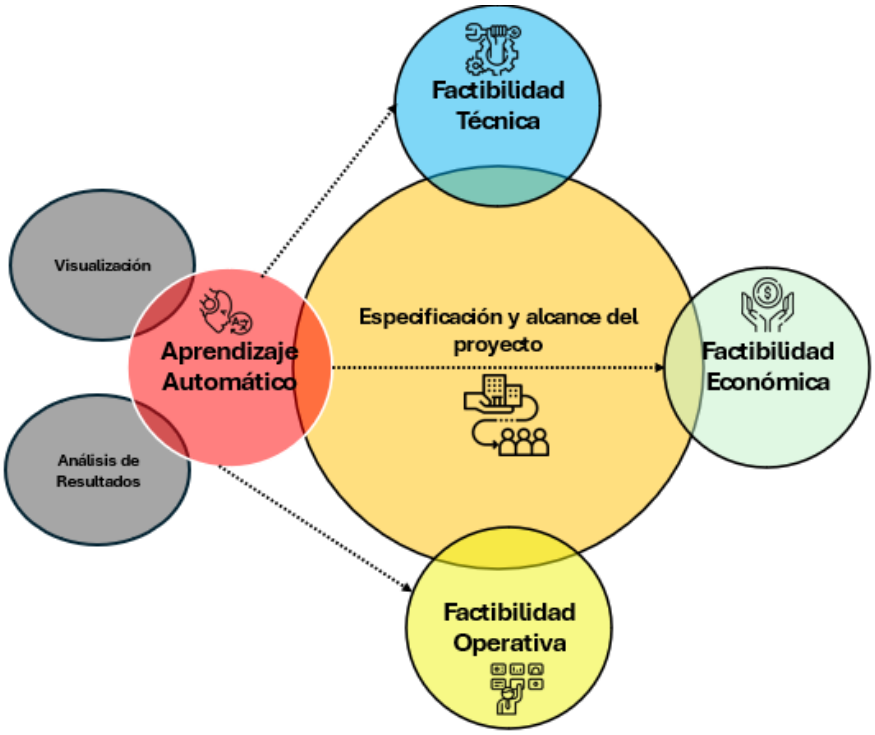

Figure 2. Machine learning method - Feasibility in data mining projects

Phase 1 – Project specification and scope

In this stage, the process to be intervened with Data Mining is identified and described in detail, knowing the activities that comprise it and the data flow. In addition, activities are carried out to determine the technical, economic, and operational feasibility of the project.

Technical feasibility

The amount of data and the data sources (types) are specified, as are the algorithms to be used to generate mining models, as well as the necessary hardware and software infrastructure. Data cleaning is performed. Likewise, the type of algorithm and its complexity are identified based on the characteristics of the data mining project.

To facilitate the dynamics of this phase, machine learning is used to leverage data from previously completed projects. This information is stored in a knowledge base.

Machine learning: the model collects data from completed data mining projects. With this information, feasibility results can be predicted using trained machine learning techniques. The model will indicate the probability that the project is technically feasible.

Economic Feasibility

The total cost of the project is determined by accounting for hardware, software, personnel, return on investment, and initial budget resources.

Machine learning: the model predicts the project's return on investment and final cost. To do this, a regression model such as Linear Regression can be used, or a decision tree-based model such as a Housing Price Prediction model, when the problem is that a real estate company wants to predict home sale prices based on structural and location characteristics. In addition, the inference provides a projection of the return on investment and indicates whether the project is economically viable with a certain level of probability.

Operational Feasibility

The work team is established, and the required human resources and the time needed for project implementation, start-up, and monitoring are determined.

Machine learning: the model predicts the time required for implementation, start-up, and monitoring of the project, as well as the number of human resources needed, based on its specific characteristics. As in the technical model, a classification or regression model can be used, depending on the dependent and independent variables defined for training. As a result, the model indicates the project's operational feasibility with a certain level of probability.

Phase II - Analysis of Results for Decision Making

In accordance with the previously established requirements for the data mining project, the predictions obtained in each feasibility activity are interpreted and used for decision-making.

Activities

Data Collection: data is extracted from various sources, specifically from models generated by the machine learning component in the three feasibility phases. The data is in comma-separated values (CSV) format and serves as the input for review and analysis. Visualization tools facilitate the interpretation of results.

Phase III - Visualization

A graphical view of the data generated by the machine learning component is obtained, and information related to the project's scope is loaded, enabling more accurate decisions about the organizational process's needs.

RESULTS

Phase I - Project Specification and Scope

Specification of the sales management process: the process corresponds to sales management and concerns the strategies and best practices the organization applies to manage its customers.

Purpose of the analysis: the organization's sales have been affected by customer churn. The organization planned a data analysis to determine which customers are at risk of churning. This will allow the organization to establish a set of retention strategies.

A predictive data analysis model was implemented to determine, based on specific demographic characteristics, which customers are highly likely to churn.

Indicators for data analysis

Indicators were established for customer behavior regarding sales, with respect to satisfaction levels, tastes, inclinations, and product preferences.

Technical Feasibility

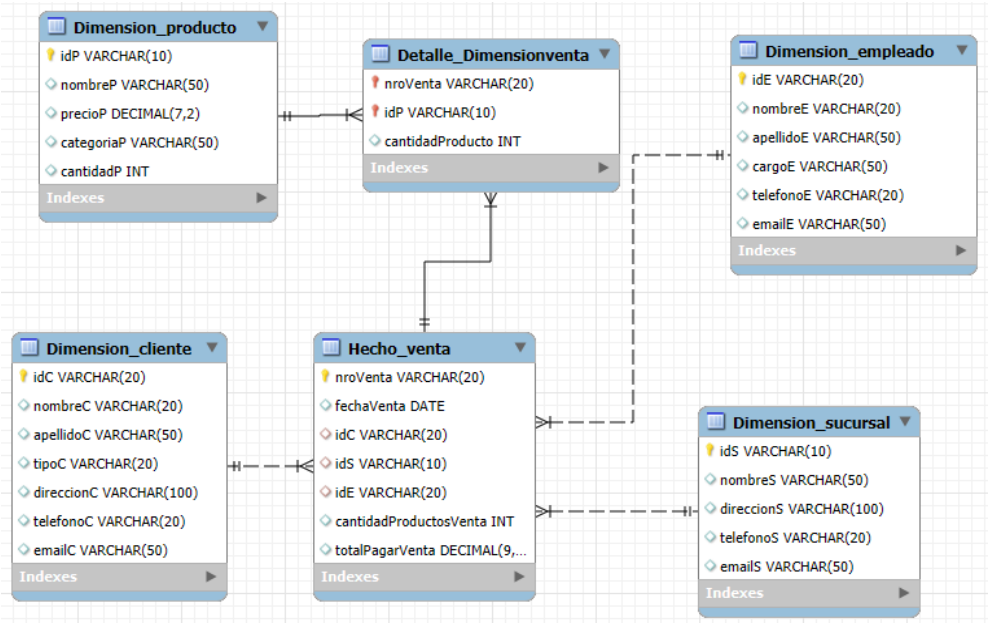

Figure 3. Data Warehouse – customer sales events

Data source: data was identified in CSV format, with approximately 2000 records, structured to include sales characteristics, including customer data, products, prices, product types, categories, branches, employees, and invoiced values. The data was taken from transactional databases of sales, PQRS, and customer management.

Data Storage: the data was consolidated into a data warehouse comprising dimension tables and a fact table. Dimensions (product, branch, employee, customers, and preferences), and a fact table that records the movement of historical events generated from customer sales transactions.

Knowledge Base Technical Feasibility

Identification of the technical project, project execution date, project end date, project purpose, data source, proposed machine learning model, estimated cost of technical feasibility, estimated time, technical feasibility result indicators.

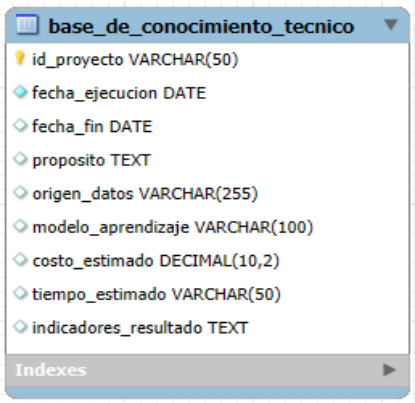

Figure 4. Knowledge Base – Technical Feasibility

Machine learning: a model was generated using Python (Pandas libraries), training a Neural Network classification algorithm. The source used for training is the technical feasibility knowledge base. The inference presented an 82 % probability of technical feasibility.

Economic Feasibility

The project duration of three months is taken into account, as is the cost of the software and hardware required for implementing machine learning models and supporting the knowledge bases and applications.

Software cost: open source licenses for MySQL, Power BI, and Python.

Human Resource Cost: Data Scientist, Data Analyst. Thirty-six million pesos (36 000 000).

Hardware cost: cloud licenses and a dedicated server for the database and applications. Forty million pesos (40 000 000).

Total project cost: seventy-six million pesos (76 000 000).

Economic knowledge base

This section contains historical records for other projects, including calculations of personnel, hardware, and software costs, total project costs, and return on investment percentages. All of this takes into account the project's duration and the organization's available budget.

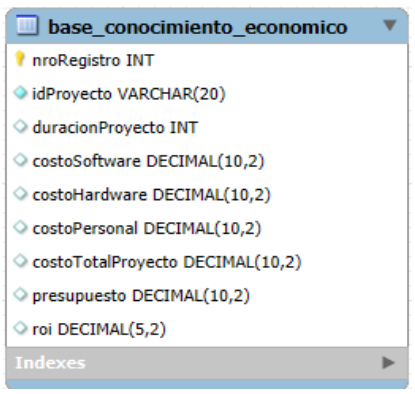

Figure 5. Knowledge Base – Economic Feasibility

Machine learning for economic estimation: a model was generated using Python (Pandas libraries), training

a Decision Tree algorithm. The source used for training is the financial feasibility knowledge base. The model's conclusion indicates that the project is economically viable, as it has a high return on investment and the total cost is within the organizational budget.

Operational Feasibility

Operational knowledge base: this repository contains historical events related to the operational feasibility determined for other projects, based on the duration, the total human resources involved in the project, and the role and scope that each of them has to contribute to development, deployment, and administration.

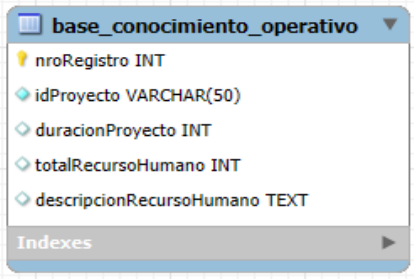

Figure 6. Knowledge Base – Operational Feasibility

Machine learning for operational estimation: a classification algorithm was used to generate a model that determined the project is operationally viable, taking into account resource requirements, estimated time, and effort. It has a 79 % probability, based on other established and executed projects. The model was trained with events from the operational knowledge base.

Phase II – Analysis of Results for Decision Making

Project feasibility: the project is feasible, and its development and deployment are approved. The budget and required personnel are available, and it is within the estimated time frame.

Phase III – Visualization

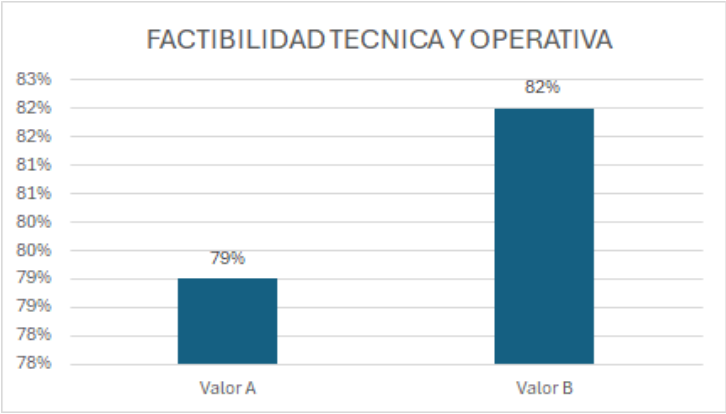

Figure 7. Technical and Operational Feasibility Behavior

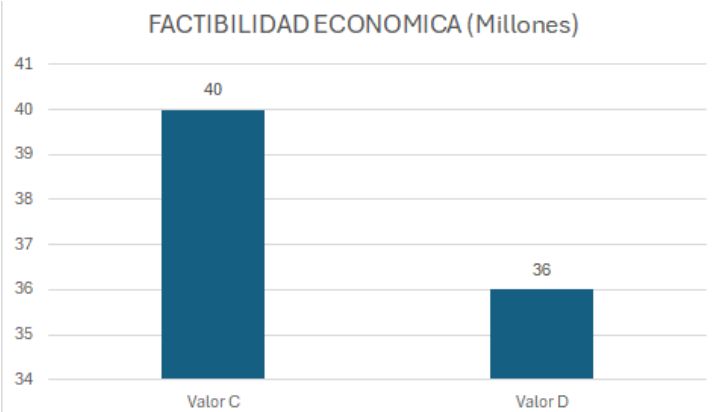

Figure 8. Economic Feasibility Performance

DISCUSSION

The implementation of the method during the methodological phases of data mining contributes to reducing cost overruns, improving project quality, and efficiently distributing the human resources required for the various activities. Additionally, it allows confirmation of the project's profitability after implementation and monitoring, thereby justifying the investment and ensuring fulfillment of the strategic objectives established during the planning phase.

CONCLUSIONS

Incorporating this method into the phases of data mining methodologies enables controlling cost overruns, improving project quality, and optimizing the distribution of human resources across different activities. It also allows confirming the project's profitability once it has been implemented and monitored, thereby justifying the investment and ensuring achievement of the strategic objectives established during planning.

The feasibility of a data mining project requires rigorous and balanced analysis. Adopting a comprehensive approach that addresses technical, economic, and operational dimensions increases the likelihood of success for organizations.

This proposal seeks to build confidence among those responsible for such initiatives, ensuring that they are not only technologically functional but also profitable, ethical, and genuinely transformative for the organization.

BIBLIOGRAPHIC REFERENCES

1. Funes A, Dasso A. Methods and Techniques of Data Mining. En: Encyclopedia of Information Science and Technology, Fifth Edition. IGI Global; 2021. p. 749-767. Doi: 10.4018/978-1-7998-3479-3.ch045.

2. Zhu S. Analysis of the severity of vehicle-bicycle crashes with data mining techniques. Journal of Safety Research (J Saf Res). 2021. P. 76:218-227. Doi: $10.1016/j.jsr.2020.11.011$

3. P R, Nayak PP, Poojary P, Talekar PP. Data Mining: Concepts, Techniques, and Applications. Int J Adv Res Sci Commun Technol. 2024;120–128. doi:10.48175/ijarsct-2282.

4. Begum DPI, Banu DN. Data mining techniques. 2024. p. 188–206. Doi:10.58532/v3bfit2p6ch2.

5. Xie B, Zhang F. Design and Implementation of Data Mining in Information Management System. En: 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES). IEEE; 2022. p. 1-5. Doi: 10.1109/ICKES58461.2022.9995817

6. Thakur A. A systematic review on data mining methods and applications. Int J Adv Res Comput Sci. 2022;13(2):28–31. doi:10.26483/ijarcs. v13i2.6807. Doi: 10.26483/ijarcs.v13i2.6807

7. Kumar M. FEASIBILITY STUDY & INPUT/ OUTPUT FORM DESIGN. GRF BOOKS; 2023. doi: 10.52458/9788196869434.2023.eb.grf.ch-12.

8. Plotnikova V, Dumas M, Milani F. Adaptations of data mining methodologies: a systematic literature review. PeerJ. 2020;6. doi:10.7717/PEERJ-CS.267.

9. Vasiliev AA, Goryachev AV. Models and Methods of Data Mining in Project Management. En: 2022 International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS). IEEE; 2022. p. 52-56. Doi: 10.1109/ElConRus54750.2022.9755501.

10. McLeod S. Feasibility studies for novel and complex projects: Principles synthesised through an integrative review. 2021; 2:100022. Doi: 10.1016/j.plas.2021.100022.

11. Renggli C, Rimanic L, Kolar L, Wu W, Zhang C. Automatic Feasibility Study via Data Quality Analysis for ML: A Case-Study on Label Noise. 2023. Doi: 10.1109/ICDE55515.2023.00024.

12. Gómez Palacios HJ, Jiménez Toledo RA, Hernández Pantoja GA, Martínez Navarro ÁA. A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. Adv Sci Technol Eng Syst J. 2017;2(3):598–604. Doi: 10.25046/aj020376.

13. Sangacha-Tapia L, González-Cañizalez Y, Rivas-Herrera J. Optimización de Criterios de Búsqueda avanzada

para Nuevas Tendencias en la Académica mediante Machine Learning. Rev Cient Zambos. 2025;4(2):197–211. Doi:10.69484/rcz/v4/n2/114.

14. Alika A, Mirza H, Andri, Ferdiansyah A. Classification Of South Sumatra Songket Woven Fabric Motifs Using Deep Learning. Data J Inf Syst Manag. 2024;2(2):24–35. Doi:10.61978/data. v2i1.

15. Castro Coria EG, Ruiz Flores López P. Guía metodológica para el uso de minería de datos en la Plataforma Nacional de Transparencia. Estud Derecho Inf. 2024;(17). Doi:10.22201/iij.25940082e.2024.17.18782.

16. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS; 2000.

17. Espinosa-Zúñiga JJ. Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ing Investig Tecnol. 2020; XXI (1):1–17. Doi:10.22201/fi.25940732e.2020.21n1.008.

18. Eckert KB. Modelo basado en la Toma Decisiones con Criterios Múltiples para la elección de Metodologías de Data Science. Universidad Nacional de Misiones; 2019.

19. Moine JM, Gordillo SE, Haedo AS. Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. En: XVII Congreso Argentino de Ciencias de la Computación. 2011.

20. Rodríguez Montequín MT, Álvarez Cabal JV, Mesa Fernández JM, González Valdés A. Metodologías para la realización de proyectos de Data Mining. En: VII Congreso Internacional de Ingeniería de Proyectos. 2003. p. 257-265.

FINANCING

None.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Fabio Alberto Vargas Agudelo, Jorge Mendoza.
Data collection: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Formal analysis: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Research: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Fabio Alberto Vargas Agudelo, Jorge Mendoza.
Methodology: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Project management: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Resources: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Supervision: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Validation: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Writing – original draft: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.
Drafting – review and editing: Fabio Alberto Vargas, Juan Camilo Giraldo Mejía, Jorge Mendoza.