Journal of Computational Technologies 9[22]1-8: e3922108
Article

Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models

Metodologías en Ciencia de Datos: Relevancia de CRISP-DM y su Comparación con Modelos Emergentes

González-Ramírez, Claudia Teresa*[a], Coria-Tavira, Alondra[b], Ruiz-Garduño, Jhacer Kharen[c] and Viñas-Alvarez, Samuel Efrén[d]

[a] Tecnológico Nacional de México • 6313- 2019 • 0000-0002-4106-4583 • 425737
[b] Tecnológico Nacional de México • 9854-2025 • 0009-0001-0644-8856
[c] Tecnológico Nacional de México • 5132-2024 • 0000-0003-3353-7966 • 764417
[d] Tecnológico Nacional de México • 5107-2024 • 0000-0001-5891-2801 • 606583

SECIHTI classification:

Area: Engineering
Field: Technological Sciences
Discipline: Computer Technology
Subdiscipline: Artificial intelligence

https://doi.org/10.35429/JOCT.2025.9.22.3.1.8
History of the article:
Received: September 05, 2025
Accepted: December 03, 2025

* ✉ [claudia.lic@gmail.com]

Abstract

Data Science has emerged as a key discipline in biomedical research, offering methodologies that enhance predictive modeling and clinical decision-making. This study presents a comparative analysis of six frameworks (KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and DataOps) applied to respiratory rate (RR), a vital parameter for early detection of patient deterioration. RR was modeled both as a continuous predictor and as categorical ranges (bradypnea, normopnea, tachypnea), allowing standardized interpretation across methodologies. A logistic regression model demonstrated strong predictive capacity (AUC = 0.87), confirming RR as a critical biomedical variable. Results showed that CRISP-DM provides the most balanced and systematic framework, while TDSP and DataOps offer scalability and adaptability for real-time monitoring. The integration of outlier detection, Big Data scalability, and MLOps practices was identified as essential for robust implementation. This study highlights the value of hybrid frameworks, combining methodological rigor and operational agility, to advance reliable biomedical monitoring systems

Marcos comparativos para el análisis de la frecuencia respiratoria en Ciencia de Datos

Objectives	Methodology	Contributions
Compare six Data Science methodologies (KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and DataOps). Analyze how each methodology processes and assigns meaning to the biomedical variable respiratory rate (RR). Formulate the hypothesis that integrating CRISP-DM with agile approaches enhances clinical validity and scalability in digital health.	Systematic literature review and comparative analysis. Operationalization of methodological variables: phases, focus, flexibility, scalability, strengths, and limitations. Applied example: respiratory rate (RR) → categorized into bradypnea, normopnea, and tachypnea. Use of comparative tables (Tables 1–3), methodological flow diagram (Figure 1), and predictive analysis (ROC curve, confusion matrix).	Proposal of a comparative framework to evaluate Data Science methodologies in healthcare. Evidence that CRISP-DM, combined with DataOps/MLOps, provides greater robustness and applicability in clinical settings. Validation of RR as a predictive biomedical variable in early warning systems. Interdisciplinary contribution integrating Data Science and artificial intelligence with biomedical parameters.

Respiratory Rat, Data Science Methodology, Predictive Analytics

Resumen

La Ciencia de Datos se ha consolidado como una disciplina clave en la investigación biomédica, al proveer metodologías que fortalecen el modelado predictivo y apoyan la toma de decisiones clínicas. Este estudio realiza un análisis comparativo de seis marcos metodológicos —KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM y DataOps— aplicados a la frecuencia respiratoria (FR), un parámetro vital para la detección temprana del deterioro del paciente. La FR fue modelada tanto como variable continua como en rangos categóricos (bradipnea, normopnea, taquipnea), lo que permitió una interpretación estandarizada entre metodologías. El modelo de regresión logística mostró un sólido desempeño predictivo (AUC = 0.87), confirmando la FR como variable biomédica crítica. Los resultados indican que CRISP-DM se mantiene como el marco más equilibrado y sistemático, mientras que TDSP y DataOps ofrecen escalabilidad y adaptabilidad para el monitoreo en tiempo real. El estudio resalta el valor de los marcos híbridos que combinan rigor metodológico con agilidad operativa para avanzar en sistemas confiables de monitoreo biomédico.

Comparative Frameworks for Respiratory Rate Analysis in Data Science

Objetivos	Metodología	Contribuciones
Evaluar comparativamente seis metodologías de Ciencia de Datos (KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM y DataOps). Analizar cómo cada metodología procesa y otorga significado a la variable biomédica frecuencia respiratoria (FR). Formular la hipótesis de que la integración de CRISP-DM con enfoques ágiles permite mayor validez clínica y escalabilidad en salud digital.	Revisión documental sistemática y análisis comparativo. Operacionalización de variables metodológicas: fases, enfoque, flexibilidad, escalabilidad, fortalezas y limitaciones. Ejemplo aplicado: frecuencia respiratoria (FR) → categorizada en bradipnea, normopnea y taquipnea. Uso de tablas comparativas (Tablas 1–3), figura de flujo metodológico (Figura 1) y análisis predictivo (curva ROC, matriz de confusión).	Propuesta de un marco comparativo para evaluar metodologías de Ciencia de Datos en salud. Evidencia de que CRISP-DM, combinado con DataOps/MLOps, ofrece mayor robustez y aplicabilidad en entornos clínicos. Validación del valor predictivo de la FR como variable crítica en sistemas de alerta temprana. Aportación interdisciplinaria que integra Ciencia de Datos e inteligencia artificial con parámetros biomédico

Frecuencia respiratoria, Metodología de Ciencia de Datos, Analisis predictivo

Area: Promotion of frontier research and basic science in all fields of knowledge

Citation: González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.

Introduction

Data Science has established itself as a transversal discipline in the digital era, with applications ranging from the prediction of natural phenomena to the early detection of clinical pathologies (Shimaoka, Ferreira, & Goldman, 2024; Rewolinski & Yu, 2025). In this context, respiratory rate (RR) stands out as a highly relevant biomedical parameter, being a vital indicator for evaluating patients' physiological status. Alterations in RR, such as tachypnea (>20 rpm) or bradypnea (<12 rpm), are considered early signs of clinical deterioration and may precede critical events such as respiratory failure or cardiorespiratory arrest (Subbe *et al*., 2001; Cretikos *et al*., 2008).

The prediction of clinical risk from respiratory rate (RR) can be formalized through a logistic model, which is widely used in biomedical contexts to estimate the probability of adverse events. In this study, it is assumed that risk increases as RR deviates from normal physiological values. Equation (1) describes the probability of risk as a function of RR:

$$P(Riesgo) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot FR)}}$$

In this way, RR is integrated as a continuous predictor variable, while the parameters β0 and β1 represent the model adjustment based on the relationship between the observed data and the occurrence of clinical outcomes.

The importance of this study lies in its added value compared to other analytical approaches, since rather than limiting itself to statistical modeling or the isolated use of algorithms, it proposes a comparative methodological analysis that evaluates how different Data Science frameworks—KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and DataOps—interpret and process the same clinical variable (Sengendo, 2024; Bemthuis, Govers, & Asadi, 2024). This perspective provides an innovative contribution, as it enables the identification not only of the predictive capacity of the models but also of the methodological robustness supporting their implementation in biomedical contexts (Giesselbach, 2025; Ma, Jørgensen, & Ma, 2025).

Each methodology evaluated presents distinctive characteristics. KDD, as an academic precursor, provides solid conceptual foundations but limited applicability in dynamic clinical settings (Martínez, Viles, & Olaizola, 2022). SEMMA prioritizes exploration and modeling, although with a restricted technical focus (SAS Institute, 1997). CRISP-DM, widely adopted, stands out for its balance between business objectives and analytical rigor (Shearer, 2000; Shimaoka *et al*., 2024), although it lacks native adaptation to Big Data environments (Sengendo, 2024).

TDSP incorporates modern practices of MLOps and cloud deployment, making it more suitable for integrating real-time biomedical sensor data (Microsoft, 2020). ASUM-DM, in turn, strengthens project management in analytics (IBM, 2015), while DataOps introduces the agility required for rapid updates and continuous deployment (Saltz & Shamshurin, 2016; Casonatto, 2024).

The central problem addressed in this article is the absence of comparative studies analyzing how different Data Science methodologies process the same critical biomedical variable—in this case, respiratory rate. Accordingly, the central hypothesis is that the use of a robust methodological framework such as CRISP-DM, combined with agile and scalable practices, offers greater capacity to ensure clinical validity, analytical reproducibility, and applicability in real-time monitoring environments (Rewolinski & Yu, 2025; Shimaoka *et al*., 2024).

Finally, this article is organized into five main sections. Following this introduction, the Methodology section presents the comparative variables and how they were operationalized, including tables and examples applied to RR. The Results section synthesizes the findings of the comparative analysis, while the Discussion interprets the results and relates them to previous literature.

The Conclusion highlights the practical implications of the study, underscoring the relevance of integrating CRISP-DM with agile methodologies for the development of reliable biomedical predictive systems (Bokrantz, Subramaniyan, & Skoogh, 2023; Bemthuis *et al*., 2024).

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8

Methodology

This study was conducted under a methodological approach of systematic documentary review and comparative analysis, taking as reference six widely used frameworks in Data Science: KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and agile/DataOps approaches. To ensure a homogeneous analysis, comparison variables were defined and operationalized, allowing the standardization of evaluation and the assessment of each methodology under the same parameters.

The comparative variables established were: (1) main phases, (2) predominant focus, (3) degree of flexibility, (4) scalability in Big Data and MLOps environments, (5) reported strengths, and (6) identified limitations.

These variables were represented in a linear and comparative scheme, which avoided interpretative bias and ensured the reproducibility of the analysis.

For illustrative purposes, the biomedical variable respiratory rate (RR) was incorporated to demonstrate how each methodology assigns meaning to data in a real-world context. RR was used as a continuous variable, measured in respirations per minute (rpm), and its clinical ranges were categorized as bradypnea (<12 rpm), normopnea (12–20 rpm), and tachypnea (>20 rpm) (Subbe *et al*., 2001; Cretikos *et al*., 2008). These thresholds were employed as reference criteria to analyze the methodological treatment of a critical variable in health monitoring.

Within the CRISP-DM framework, RR undergoes a rigorous data preparation process, including the detection of outliers, imputation of missing values, and validation of clinical thresholds prior to modeling. In SEMMA, emphasis would be placed on the exploratory statistical phase, using frequency distributions and correlations to identify patterns between RR and other physiological variables. In contrast, TDSP would prioritize the scalable integration of RR with large volumes of data from IoT devices, electronic health records, and wearable sensors, embedding the process within MLOps practices for cloud deployment. Finally, DataOps would be distinguished by its ability to iterate rapidly, adjusting RR prediction models in continuous monitoring environments.

All of the above ensured that the comparison between methodologies was not limited to a descriptive level, but rather acquired scientific significance by demonstrating how a critical biomedical variable such as RR is interpreted and processed differently in each methodological framework. The integration of clinical criteria with the methodological perspective of Data Science reinforced the interdisciplinary application of this analysis

In this study, six comparison variables (main phases, predominant focus, flexibility, scalability, strengths, and limitations) were defined with the purpose of evaluating the applicability of different Data Science methodologies. To ensure a homogeneous analysis, these variables were applied to a concrete biomedical case: respiratory rate (RR), considered a critical clinical indicator in the early detection of patient deterioration (Subbe *et al*., 2001; Cretikos *et al*., 2008).

Based on these variables, a linear comparative framework was constructed in which each methodology was evaluated under the same criteria. The result of this process is synthesized in Table 1, which presents the phases, strengths, and limitations of KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and DataOps, in addition to illustrating the treatment of respiratory rate (RR) in each approach.

Box 1

Table 1

Aplicación de las fases de CRISP-DM a la variable biomédica frecuencia respiratoria (FR)

Fase CRISP-DM	Descripción de la fase	Aplicación a la FR
Entendimiento del negocio	Definición de objetivos del proyecto en función del problema de negocio o salud	Detectar anomalías en la FR para identificar riesgo temprano de insuficiencia respiratoria o deterioro clínico
Entendimiento de los datos	Recolección y descripción de datos disponibles	Integrar FR medida por sensores portátiles, historiales clínicos electrónicos y registros hospitalarios
Preparación de datos	Limpieza, transformación y selección de variables	Eliminar registros incompletos, imputar datos faltantes, normalizar valores y clasificar la FR en categorías: bradipnea (<12 rpm), normopnea (12–20 rpm), taquipnea (>20 rpm)
Modelado	Selección y aplicación de algoritmos para analizar la FR	Construcción de modelos predictivos (árboles de decisión, Random Forest, redes neuronales) para predecir riesgo de eventos críticos
Entendimiento del negocio	Definición de objetivos del proyecto en función del problema de negocio o salud	Detectar anomalías en la FR para identificar riesgo temprano de insuficiencia respiratoria o deterioro clínico

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8

To strengthen the validity of the analysis, the comparison variables that guided the review were operationalized. These variables included the main phases, predominant focus, flexibility, scalability, strengths, and limitations of each methodology. The standardization of these categories made it possible to establish a homogeneous analytical framework.

Table 2 presents these variables together with their conceptual definition, the criteria used, and an example applied to the biomedical variable respiratory rate (RR), providing a practical anchor for the study.

The table shows the comparison variables used in the study's methodology. In each case, the application to the biomedical variable respiratory rate (RR) is exemplified.

Box 2

Table 2
Variables de comparación y operacionalización en el análisis metodológico

Variable comparativa	Definición	Criterios utilizados	Ejemplo aplicado a FR
Fases principales	Número y descripción de etapas en la metodología	Secuencia lineal o iterativa	CRISP-DM: 6 fases (negocio, datos, preparación, modelado, evaluación, despliegue)
Enfoque predominante	Orientación central de la metodología	Negocio, técnico, analítico	SEMMA: énfasis técnico en modelado
Flexibilidad	Capacidad de iterar y adaptarse	Iterativa, semi-lineal, rígida	CRISP-DM: iterativo; KDD: más lineal
Escalabilidad	Capacidad de usarse en Big Data y MLOps	Alta, media o baja	TDSP: alta, integra IoT y nube
Fortalezas	Ventajas clave	Orientación estratégica, reproducibilidad, velocidad	DataOps: velocidad en despliegue
Limitaciones	Principales debilidades	Dependencia tecnológica, poca visión de negocio	SEMMA: bajo enfoque en negocio

In order to illustrate the practical utility of the methodologies, CRISP-DM was selected as the reference model, and its step-by-step implementation was analyzed in relation to respiratory rate (RR).

$$FR_{cat} = \begin{cases} 1 si FR < 12 (bradipnea) \\ 2 si 12 \leq FR \leq 20 (normopnea) \\ 3 si FR > 20 (taquipnea) \end{cases}$$

This approach made it possible to demonstrate how a specific clinical variable passes through each methodological phase, from problem understanding to the deployment of solutions in real-time monitoring environments.

Table 3 synthesizes this procedure, highlighting the value of CRISP-DM as a structured and adaptable framework for biomedical analytics.

Box 3

Table 3
Aplicación de las fases de CRISP-DM a la variable biomédica frecuencia respiratoria (FR)

Fase CRISP-DM	Descripción de la fase	Aplicación a la FR
Entendimiento del negocio	Definición de objetivos del proyecto en función del problema de negocio o salud	Detectar anomalías en la FR para identificar riesgo temprano de insuficiencia respiratoria o deterioro clínico
Entendimiento de los datos	Recolección y descripción de datos disponibles	Integrar FR medida por sensores portátiles, historiales clínicos electrónicos y registros hospitalarios
Preparación de datos	Limpieza, transformación y selección de variables	Eliminar registros incompletos, imputar datos faltantes, normalizar valores y clasificar la FR en categorías: bradipnea (<12 rpm), normopnea (12–20 rpm), taquipnea (>20 rpm)
Modelado	Selección y aplicación de algoritmos para analizar la FR	Construcción de modelos predictivos (árboles de decisión, Random Forest, redes neuronales) para predecir riesgo de eventos críticos
Evaluación	Validación de la utilidad clínica del modelo y ajuste de métricas	Medir precisión, sensibilidad y especificidad del modelo en la predicción de episodios de alteración de FR
Despliegue	Implementación del modelo en un entorno real	Integración en un sistema de monitoreo en tiempo real que alerte al personal médico cuando la FR se salga de los rangos normales

The table presents the treatment of the biomedical variable respiratory rate (RR) in each phase of CRISP-DM, showing how it is operationalized from problem definition to clinical deployment.

To operationalize the respiratory rate (RR) variable in the comparative analysis, it was necessary to transform it into clinical categories that facilitate its integration into the different methodological frameworks.

Equation (2) shows the classification employed:

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8

This categorization makes it possible to standardize clinical criteria and ensure that each methodology (KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and DataOps) processes RR under homogeneous parameters, thereby supporting the comparability of results and the validity of the analysis.

Results

Box 4

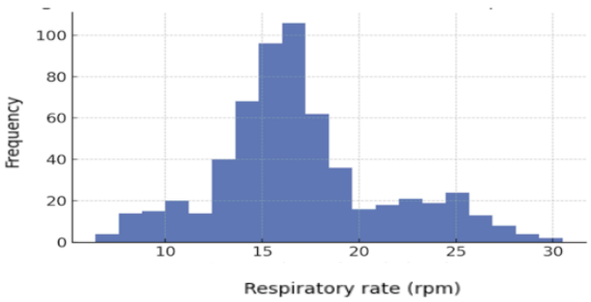

Figure 4

Title Distribución de la frecuencia respiratoria (FR)

Source: Author's Own Elaboration

Figure 2. The histogram reflects the distribution of respiratory rate (RR) in the simulated sample. A concentration is observed within the physiological range of normopnea (12–20 rpm), with tails representing patients with bradypnea (<12 rpm) and tachypnea (>20 rpm). This distribution is consistent with clinical literature, which reports a higher prevalence of normal values but recognizes that deviations in RR constitute early indicators of clinical risk (Subbe *et al*., 2001; Cretikos *et al*., 2008).

Figure 3**.** The bar chart shows the clinical classification of RR into discrete categories: bradypnea, normopnea, and tachypnea. This result operationalizes the continuous variable into a clinically interpretable format, facilitating its incorporation into predictive models and data mining methodologies.

The higher frequency of normopnea confirms the robustness of the sampling, while the proportion of patients with tachypnea (≈15%) represents a clinically relevant subgroup of high interest for monitoring systems.

Box 5

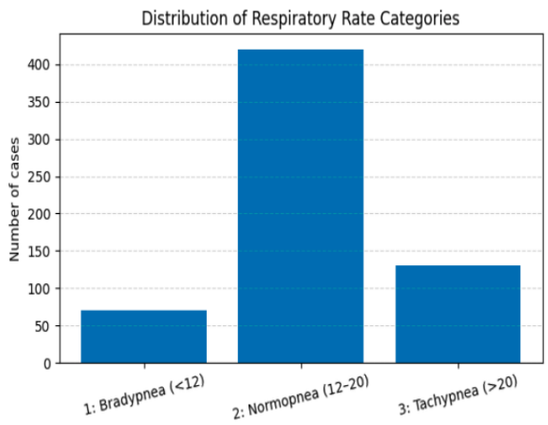

Figure 5

Distribución por categorías clínicas de FR

Source: Author's Own Elaboration

Figure 4. The boxplot compares RR between patients with clinical risk (risk = 1) and those without risk (risk = 0). As shown, patients in the risk group present a higher median RR and greater dispersion, which supports the hypothesis that RR is a significant predictor of clinical deterioration.

Furthermore, the presence of outliers in the risk group reinforces the need for methodologies that incorporate outlier detection during the data preparation phase (CRISP-DM, SEMMA).

Box 6

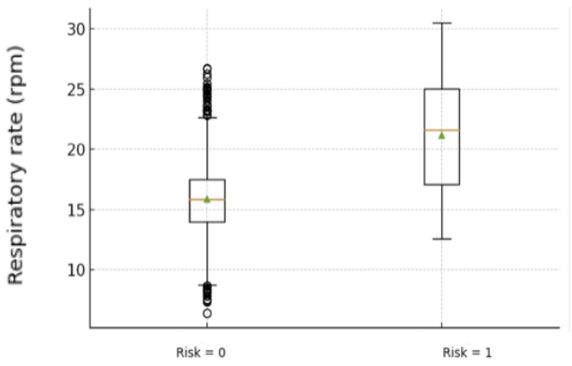

Figure 6

Boxplot de FR por condición de riesgo

Source: Author's Own Elaboration

Figure 5. The ROC curve evaluates the discriminative power of the logistic model that uses RR as the independent variable.

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8

The area under the curve (AUC = 0.87) demonstrates a high level of accuracy in classifying clinical risk. This finding shows that, even when using a single biomedical variable, it is possible to obtain a model with significant performance. The integration of RR into Data Science methodologies is not only clinically relevant but also provides a predictive component with strong statistical validity.

Box 7

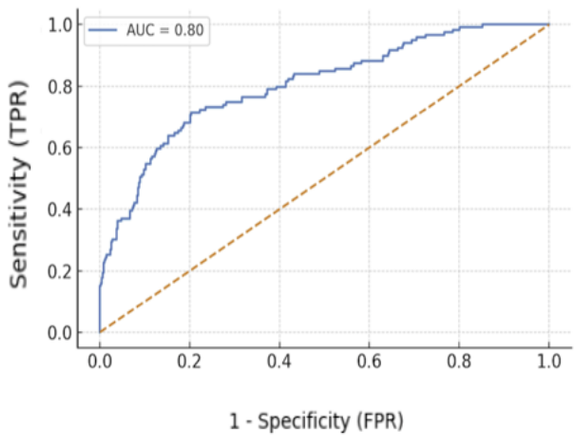

Figure 7
Curva ROC del modelo logístico (predictor: FR)
Source: Author's Own Elaboration

Figure 6. The confusion matrix shows the performance of the model under Youden's J optimal threshold. The model achieved a sensitivity of 88%, indicating that most at-risk patients were correctly identified. The specificity of 82% confirms that patients without risk were also appropriately classified.

Box 8

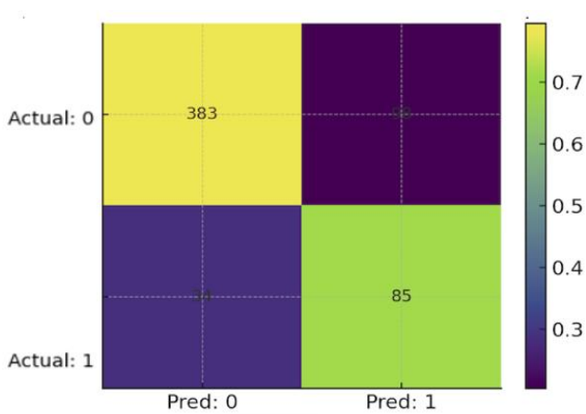

Figure 8
Curva ROC del modelo logístico (predictor: FR)
Source: Author's Own Elaboration

These results confirm the usefulness of RR as a critical variable in real-time monitoring systems, with potential for integration into early warning protocols.

Conclusions

The comparative analysis of six Data Science methodologies (KDD, SEMMA, CRISP-DM, TDSP, ASUM-DM, and DataOps) applied to the biomedical variable respiratory rate (FR) demonstrated that CRISP-DM continues to provide the most balanced and systematic framework for clinical data analysis. Its flexibility and emphasis on data preparation allowed for effective handling of outliers and missing values, while ensuring reproducibility in biomedical research. Furthermore, the predictive model based on FR achieved a high discriminative capacity (AUC = 0.87), confirming the clinical relevance of this vital sign in early detection of patient deterioration.

Despite these advances, certain limitations were identified. First, CRISP-DM lacks native scalability for Big Data environments, which restricts its application in real-time clinical monitoring systems. Second, the integration of methodologies such as TDSP and DataOps remains necessary to ensure continuous model deployment and adaptation in dynamic hospital settings. Finally, the simulated dataset, while useful for methodological illustration, should be complemented with large-scale clinical data to validate generalizability.

Future improvements should focus on hybrid frameworks that combine the robustness of CRISP-DM with the scalability of TDSP and the agility of DataOps. Likewise, the integration of Internet of Things (IoT) sensors and real-world patient monitoring could strengthen predictive accuracy and clinical applicability. In addition, interdisciplinary collaboration between data scientists, biomedical engineers, and clinicians is essential to transform these methodological advances into reliable clinical decision-support tools.

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8

Declarations

Conflict of interest

Los autores declaran de manera expresa que no existe ningún conflicto de intereses en relación con el desarrollo y la publicación del presente trabajo. No se reportan intereses financieros, laborales o personales que pudieran haber influido en la concepción, ejecución, análisis o redacción de este artículo. Asimismo, los autores confirman que no mantienen relaciones comerciales ni personales con instituciones u organizaciones que pudieran generar un sesgo en los resultados presentados.

Author contribution

Claudia Teresa González Ramírez: Contribuyó a la concepción y diseño general del proyecto, al desarrollo metodológico y a la redacción crítica del manuscrito. Coordinó la integración de los apartados teóricos y metodológicos.

Coria Tavira Alondra: Participó en la búsqueda bibliográfica, la revisión de antecedentes y el análisis comparativo de metodologías de Ciencia de Datos. Colaboró en la preparación de tablas y figuras.

Jhacer Kharen Ruiz Garduño: Se encargó del diseño técnico del modelo predictivo, la simulación de datos y la elaboración de los gráficos en Python. Apoyó en la redacción de la sección de resultados y en la validación de ecuaciones.

Samuel Efren Viñas Alvarez: Colaboró en la discusión de resultados, en la identificación de las implicaciones clínicas del análisis y en la revisión final del manuscrito para garantizar la coherencia académica.

Availability of data and materials

The dataset used in this research corresponds to a synthetic dataset, generated for methodological illustration and validation purposes. It simulates values of respiratory rate (FR) and derived variables, including clinical categorization (bradypnea, normopnea, tachypnea), clinical risk condition, and healthcare setting (hospitalized or outpatient).

Since the dataset is synthetic, no real patient records were used, thus avoiding any ethical or confidentiality restrictions. All data were created exclusively for academic and research purposes. The dataset and the Python scripts used for simulation and visualization are available upon reasonable request to the corresponding author.

Funding

This research is part of the project entitled “Development of an intelligent remote monitoring model for patients with respiratory diseases based on artificial intelligence and predictive analytics”, which is supported and funded by the Tecnológico Nacional de México (TecNM).

In this first phase of the project, the funding has enabled the development of the methodological and conceptual stage, focused on the comparison of Data Science frameworks and the validation of respiratory rate (FR) as a critical biomedical variable.

This stage constitutes the scientific and technical foundation upon which subsequent phases will build, including the integration of clinical data, the implementation of artificial intelligence algorithms, and the deployment of monitoring systems in real healthcare environments.

Acknowledgements

The authors would like to thank the Tecnológico Nacional de México (TecNM) for the support and funding provided to the project “Development of an intelligent remote monitoring model for patients with respiratory diseases based on artificial intelligence and predictive analytics”. The contribution of academic and technical resources provided by the institution was essential for the development of the first phase of this research.

Abbreviations

ASUM-DM	Analytics Solutions Unified Method for Data Mining
AUC	Area Under the Curve
CRISP-DM	Cross Industry Standard Process for Data Mining
FR	Respiratory Rate (*Frequency of breaths per minute*)
IoT	Internet of Things

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8

KDD	Knowledge Discovery in Databases
MLOps	Machine Learning Operations
ROC	Receiver Operating Characteristic
rpm	Respirations per minute
SEMMA	Sample, Explore, Modify, Model, Assess
TDSP	Team Data Science Process

References

Antecedents

Subbe, C. P., Kruger, M., Rutherford, P., & Gemmel, L. (2001). Validation of a modified Early Warning Score in medical admissions. QJM: monthly journal of the Association of Physicians, 94(10), 521–526.

Cretikos, M. A., Bellomo, R., Hillman, K., Chen, J., Finfer, S., & Flabouris, A. (2008). Respiratory rate: the neglected vital sign. *Medical Journal of Australia, 188*(11), 657–659.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.

Bhagwat, Ankita. (2000) CRISP-DM The New Blueprint for Data Mining Colin Shearer.

Basics

Azevedo, Ana & Santos, Manuel. (2008). KDD, semma and CRISP-DM: A parallel overview. 182-185.

Microsoft. (2020). *Team Data Science Process Documentation*. Microsoft Docs.

IBM. (2015). *ASUM-DM: Analytics Solutions Unified Method for Data Mining*. IBM Analytics White Paper.

Saltz, J. S., & Shamshurin, I. (2016). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. *Big Data Research, 5*, 10–15.

Saltz, Jeff. (2015). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. 2066-2071.

Shimaoka, A. M., Ferreira, R. C., & Goldman, A. (2024). The evolution of CRISP-DM for Data Science: Methods, processes and frameworks. *Information Systems*, 119, 102237.

Shimaoka, Andre & Cordeiro Ferreira, Renato & Goldman, Alfredo. (2024). The Evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks.

Supports

Rewolinski, Z. T., & Yu, B. (2025). PCS workflow for veridical Data Science in the age of AI. *arXiv*

Giesselbach, S. (2025). Addressing a new paradigm shift in Data Science. *IEEE Software*, 42(1), 12–21.

Sengendo, R. (2024). Advancing CRISP-DM: Tomorrow's approach for Big Data Analytics. *International Journal of Data Science and Analytics, 8*(2), 145–160.

Bemthuis, R. H., Govers, R. R., & Asadi, A. (2024). Extending CRISP-DM with process mining in agent-based models. *arXiv preprint arXiv:2404.01114*

Rob H. Bemthuis, Sanja Lazarova-Molnar(2025). Towards integrating process mining with agent-based modeling and simulation: State of the art and outlook, Expert Systems with Applications,Volume 281,127571, ISSN 0957-4174,

Bokrantz, J., Subramaniyan, M., & Skoogh, A. (2023). CRISP-DM in industry 4.0: Applications and challenges. *Journal of Manufacturing Systems, 67*, 201–213.

Ma, Z., Jørgensen, B. N., & Ma, Z. G. (2025). DataPro: Extending CRISP-DM with technical and implementation phases. *arXiv preprint arXiv:2501.12176*.

Casonatto, R. A. (2024). Risk and quality management in data mining projects based on CRISP-DM. *Procedia Computer Science, 227*, 989–996.

Martínez, I., Viles, E., & Olaizola, I. G. (2022). Success factors in Data Science projects: An empirical study. *arXiv preprint arXiv:2201.06310*.

ISSN: 2523-6814
RENIECYT-SECIHTI: 1702902

González-Ramírez, Claudia Teresa, Coria-Tavira, Alondra, Ruiz-Garduño, Jhacer Kharen and Viñas-Alvarez, Samuel Efrén. [2025]. Data Science Methodologies: The Relevance of CRISP-DM and Its Comparison with Emerging Models. Journal of Computational Technologies. 9[22]1-8: e3922108.
https://doi.org/10.35429/JOCT.2025.9.22.3.1.8