Base de Datos y Minería de Datos

# 16285
# Integración de métricas de calidad de datos en procesos de minería de datos: Validación metodológica aplicada al caso INCUCAI


AUTORES:

**Roxana Martínez** roxana.martinez@uai.edu.ar
**Santiago Acha** santiago.acha@uai.edu.ar
**Agustín Luján** AgustinEzequiel.LujanBidondo@alumnos.uai.edu.ar
**Ornella Mansilla** ornellalujan.mansilla@alumnos.uai.edu.ar
**Nicolas Censi** nicolasmartin.censi@alumnos.uai.edu.ar


Ingeniería en Sistemas Informáticos

## LÍNEAS DE INVESTIGACIÓN:
Ingeniería de Software


## PALABRAS CLAVE:
Calidad de Datos; Minería de Datos; ISO/IEC 25012; Ciencia de Datos; Modelos Predictivos; Gobernanza de Datos.


## CONTEXTO:

El presente trabajo se enfoca en la línea de investigación en Calidad y Ciencia de Datos desarrollada en el ámbito del Centro de Altos Estudios en Tecnología Informática (CAETI) de la Facultad de Tecnología Informática de la Universidad Abierta Interamericana (UAI). Esta línea se articula con el proyecto denominado “Investigación y desarrollo de software para la validación de la calidad de datos abiertos e identificación de patrones para predicciones”, orientado al diseño de enfoques y herramientas que integren evaluación de calidad y análisis predictivo en distintos dominios de aplicación.
En este contexto, el Laboratorio de Calidad y Ciencia de Datos desarrolla actualmente una línea orientada a la formalización de un modelo metodológico que incorpora dimensiones de calidad de datos dentro de procesos de minería de datos, promoviendo prácticas reproducibles, medibles y alineadas con estándares internacionales. El trabajo que aquí se presenta consolida dicha línea, proponiendo un enfoque estructurado para la integración sistemática de métricas de calidad en el ciclo de vida de proyectos de análisis de datos, fortaleciendo así la confiabilidad y gobernanza de los resultados obtenidos.

## LÍNEAS DE INVESTIGACIÓN DESARROLLO:

El presente trabajo se enmarca en la línea de investigación en Ingeniería de Software del Centro de Altos Estudios en Tecnología Informática (CAETI), con foco en Calidad de Datos y Ciencia de Datos aplicada a contextos gubernamentales y de gestión pública. En particular, se articula con el desarrollo de enfoques metodológicos y algoritmos orientados a la validación, análisis y explotación de datos públicos abiertos en el ámbito de la salud.

En esta instancia del proyecto, los ejes de investigación y desarrollo se estructuran de la siguiente manera:

- *Desarrollo de algoritmos de validación y evaluación dimensional de calidad*, orientados a detectar inconsistencias, valores faltantes, anomalías estructurales y problemas de integridad en datasets públicos abiertos, con especial atención a datos vinculados a procesos de donación y trasplante de órganos y tejidos.
- *Diseño e implementación de métricas específicas para el dominio sanitario*, incluyendo dimensiones como completitud, consistencia semántica, integridad geoespacial y coherencia temporal, integradas dentro de un modelo metodológico formal.
- *Integración de mecanismos de calidad en pipelines de minería de datos*, permitiendo analizar el impacto de las métricas de calidad sobre el desempeño de modelos predictivos desarrollados mediante técnicas de machine learning.
- *Análisis y validación de modelos predictivos en contextos gubernamentales*, evaluando la estabilidad, robustez y capacidad de generalización frente a distintas condiciones de calidad de datos.
- *Desarrollo de prototipos y funcionalidades para herramientas de validación de datasets públicos*, incorporando criterios de trazabilidad, reproducibilidad y gobernanza de datos.
- *Elaboración de lineamientos y buenas prácticas* para la aplicación de técnicas de aprendizaje automático en procesos de validación y análisis de datos abiertos en salud, promoviendo un enfoque data-centric orientado a la calidad.

## FORMACIÓN DE RECURSOS

El equipo se encuentra conformado por una docente investigadora, Doctora en Ciencias Informáticas, quien dirige el proyecto; un docente auxiliar de la carrera Ingeniería en Sistemas Informáticos; y estudiantes de grado y posgrado que participan activamente en las distintas etapas metodológicas y experimentales del trabajo. En relación directa con la línea de investigación presentada, actualmente se desarrollan 2 tesis doctorales y 2 tesis de maestría vinculadas a la integración de métricas de calidad de datos y su impacto en modelos predictivos aplicados a datos abiertos. Por otro lado, participan estudiantes colaboradores de la carrera de Ingeniería en Sistemas Informáticos de la UAI de distintas sedes, quienes intervienen en tareas de implementación de métricas, validación experimental y desarrollo de prototipos, fortaleciendo su formación en metodologías de investigación, gobernanza de datos y aplicaciones de aprendizaje automático en contextos gubernamentales.
La articulación entre investigación y formación de recursos humanos contribuye a consolidar una línea académica sostenida en el tiempo, promoviendo la generación de capacidades en calidad de datos, minería de datos y Ciencia de Datos aplicada al sector público.

## RESULTADOS OBTENIDOS/ESPERADOS:

***Enfoque del proyecto:***
En el contexto del Gobierno Abierto y la disponibilidad de datos públicos en salud, el presente trabajo aborda el desafío de integrar formalmente métricas de calidad de datos dentro de procesos de minería de datos aplicados a información sanitaria. A diferencia de enfoques centrados exclusivamente en la extracción o explotación analítica, la propuesta incorpora una capa metodológica explícita de evaluación y validación de calidad previa y transversal al modelado predictivo. El caso de estudio se basa en datos derivados del sistema SINTRA (Sistema Nacional de Información de Procuración y Trasplante de la República Argentina) [14], administrado por el Instituto Nacional Central Único Coordinador de Ablación e Implante (INCUCAI) [15], vinculados a procesos de donación y trasplante de órganos y tejidos en la República Argentina. Sobre este dominio, se implementa el modelo metodológico propuesto, incorporando dimensiones de calidad alineadas con estándares internacionales y evaluando su impacto en el desempeño de modelos predictivos.

***Objetivos principales:*** Integrar métricas formales de calidad de datos dentro del ciclo de vida de procesos de minería de datos, validando metodológicamente su impacto en el desempeño, estabilidad y confiabilidad de modelos predictivos.

***Objetivos específicos:***

- Definir y operacionalizar dimensiones de calidad de datos alineadas con estándares internacionales, adaptadas al contexto de datos abiertos en salud.
- Diseñar e implementar algoritmos de evaluación y validación dimensional sobre datasets derivados del sistema SINTRA.
- Integrar los resultados de la evaluación de calidad dentro del pipeline de minería de datos, estableciendo vínculos explícitos entre calidad y desempeño predictivo.
- Comparar el comportamiento de modelos predictivos antes y después del proceso de curación y validación de datos.
- Analizar la trazabilidad y reproducibilidad de los resultados obtenidos bajo un enfoque data-centric.

## REFERENCIAS:


[1] Martínez, R., et al. (2025). Towards a scalable open data platform in public health: Evolving the INCUCAI case for organ donation. En Actas de la 51ª Conferencia Latinoamericana de Informática (CLEI 2025). Valparaíso, Chile.
[2] Kumar, S., Datta, S., Singh, V., Singh, S. K., & Sharma, R. (2024). Opportunities and challenges in data-centric AI. IEEE Access, 12, 33173-33189.
[3] Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric artificial intelligence: A survey. ACM Computing Surveys, 57(5), 1-42.
[4] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021, May). “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
[5] Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., ... & Harmouch, H. (2022). The effects of data quality on machine learning performance. arXiv preprint arXiv:2207.14529.
[6] Gentyala, R. (2025). Ethical Artifacts: Engineering Verifiable Audit Trails for Human-in-the-Loop Decisions in ML Data Pipelines. Journal of Scientific and Engineering Research, 12(10), 240-251.
[7] Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. ACM computing surveys (CSUR), 54(5), 1-39.
[8] ISO 25012 (2008). Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. Disponible en: https://www.iso.org/standard/35736.html
[9] Polyzotis, N., Zinkevich, M., Roy, S., Breck, E., & Whang, S. (2019). Data validation for machine learning. Proceedings of machine learning and systems, 1, 334-347.
[10] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9).
[11] Hosseinzadeh, E., Afkanpour, M., Momeni, M., & Tabesh, H. (2025). Data quality assessment in healthcare, dimensions, methods and tools: a systematic review. BMC Medical Informatics and Decision Making, 25(1), 296.
[12] Mashoufi, M., Ayatollahi, H., Khorasani-Zavareh, D., & Boni, T. T. A. (2023). Data quality in health care: main concepts and assessment methodologies. Methods of Information in Medicine, 62(01/02), 005-018.
[13] Martínez, R. et al. (2021). Metrics proposal to measure the quality of governmental datasets. IEEE Latin America Transactions, 20(2), 301-308.
[14] INCUCAI - SINTRA (2026). “SINTRA - El Sistema Nacional de Información de Procuración y Trasplante de la República Argentina”. Disponible en: https://sintra.incucai.gov.ar/
[15] Argentina.gob.ar - INCUCAI (2026). “INCUCAI”. Disponible en: https://www.argentina.gob.ar/salud/incucai