A smart data holistic approach for context-aware data analytics (AETHER-UA)

Ana Lavalle[1], Alejandro Maté[1], Juan Trujillo[1], Miguel A. Teruel[1] and Alexander Sánchez Díaz[1]

[1]*Lucentia Research Group - Department of Software and Computing Systems, University of Alicante, Carretera de San Vicent del Raspeig, s/n, 03690, San Vicent del Raspeig, Spain*

Abstract

A smart data holistic approach for context-aware data analytics (AETHER-UA) is one of the four subprojects, developed in the University of Alicante, being part of the whole project AETHER. This project is being developed by four partners: (i) University of Malaga - Coordinator; (ii) University of Alicante, (iii) University of Castilla La-Mancha, and (iv) University of Seville. The project is funded by the Ministry of Science and Innovation. The main goal of this project is to advance towards a knowledge-based framework integrating novel solutions for data, process and business analytics. The research activities for designing and developing Aether will mainly focus on three main challenges: the characterization of the datasets, the improvement and automation of the algorithms, and the generation of mechanisms to enhance model explainability and interpretation of the results. The project is highly related to data processing, integration, analysis and modeling. More concretely, within the AETHER-UA project, several proposals are being developed for the modeling of user's requirements for Machine Learning applications, the developing of a framework based on Model Driven Development (MDD) for eXplanable Artificial Intelligence and several approaches for the data bias analysis.

Keywords

Smart Data, Data Integration, Data Bias, Conceptual Modeling, User's requirements, Modeling of Machine Learning Applications

1. Project description

This project will develop Aether, a data platform oriented as an extension of BIGOWL [1], an ontology to support knowledge management in Big Data analytics, by including new dimensions such as data quality, and cybersecurity, plus application domains like process analytics and business analytics. Thus, the target metadata framework will include aspects that have never been integrated into a single solution. This novel framework will open new research challenges in taking advantage of it in data analytics: how the metadata can be discovered, linked to the different analysis, and exploited.

The first challenge is the characterization of the input datasets. Knowing properties from

ER2023: Companion Proceedings of the 42nd International Conference on Conceptual Modeling: ER Forum, 7th SCME, Project Exhibitions, Posters and Demos, and Doctoral Consortium, November 06-09, 2023, Lisbon, Portugal

alavalle@dlsi.ua.es (A. Lavalle); amate@dlsi.ua.es (A. Maté); jtrujillo@dlsi.ua.es (J. Trujillo); materuel@dlsi.ua.es (M. A. Teruel); alexander.sanchez@ua.es (A. S. Díaz)

0000-0002-8399-4666 (A. Lavalle); 0000-0001-7770-3693 (A. Maté); 0000-0003-0139-6724 (J. Trujillo); 0000-0003-0102-6918 (M. A. Teruel); 0000-0002-6236-9095 (A. S. Díaz)

the datasets such as semantics, topology, or statistical features, is critical for data optimization, process and business analytics. These properties can be used to automatically select the best algorithms depending on the problem and the input data, and fine-tune their configuration to get high-quality results.

However, the characterization of the data at the beginning of the analysis processes will discover properties that may not be valid as the data are transformed or on the inferred knowledge. So, the second challenge will be to research how these properties are transmitted throughout the analysis workflows or how new properties can be derived, and how they can be used to: facilitate or automate the selection and configuration of the algorithms needed in the analysis of the process; measure the actionability of the results based on values of authority and provenance or reliability and confidence; derive an explanation of why the result is the one we obtain and be able to make it intelligible. Thus, we aim to move from the use of analysis algorithms such as black boxes to facilitate the understanding and interpretability of the results by analysts.

All of these steps forward cannot be isolated from some fundamental aspects: (1) data tend to be managed through workflows that can be represented by using business processes, which implies new challenges; (2) the quality of the data is crucial for later analysis; (3) not only the data are relevant for analysis, cybersecurity aspects must also be included, and (4) the advance after the application of the various techniques must be measured by using KPIs.

All these research activities will be applied in real contexts by developing pilots in industrial contexts in different application domains taking advantage of existing collaborations with other research organizations or companies in a multidisciplinary approach: precision agriculture, personalized medicine, bioinformatics, Smart Cities, logistics, industry 4.0 and remote sensing.

This project is being developed by four partners: (i) University of Malaga - Coordinator; (ii) University of Alicante, (iii) University of Castilla La-Mancha, and (iv) University of Seville. The project is funded by the Ministry of Science and Innovation and has a duration of four years.

2. Project information

- **Project name:** A smart data holistic approach for context-aware data analytics (Aether)
- **Duration:** From Jan, 1st, 2021 until Dec, 31st, 2024
- **Partners:** In the following, we classify the participants grouped by the corresponding partner participating in the project
  - **Aether-UMA: KHAOS Research group - University of Málaga (UMA).** Khaos main background is related to semantics technologies.
  - **Aether-UCLM: Alarcos and GSyA research groups - University of Castilla-La Mancha (UCLM).** The Alarcos Research Group main background is data quality, the quality of databases, data warehouses, data models, and data itself. The GSyA research group's main background is related to software development security and security management and governance.
  - **Aether-UA: Lucentia Research group - University of Alicante (UA).** The Lucentia Group has extensive experience in the field of Business Intelligence (BI), Data Analytics, Big Data, and Conceptual modeling of BI applications.

  - **Aether-US: IDEA Research group - University of Sevilla (US).** The IDEA Group is devoted to research on issues related to business process models and diagnosis of systems based on data analysis.

- **Funding agency:** Ministry of Science and Innovation
- **Website:** https://aether.es/

3. Project goals

Figure 1 summarizes the project goals. The main goal is to advance towards a knowledge-based framework integrating novel solutions for data, process and business analytics. This goal will be approached by extending BIGOWL to cope with relevant dimensions as the core of Aether framework (General Goal 1). The research activities for designing and developing Aether will mainly focus on three main challenges: the characterization of the datasets (General Goal 2), the improvement and automation of the algorithms (General Goal 3), and the generation of mechanisms to enhance model explainability and interpretation of the results (General Goal 4). The project results will be applied to relevant scenarios (General Goal 5). These five general goals can be divided into more specific goals with their responsable researchers (due to space constraints we cannot provide the list of all the researchers):

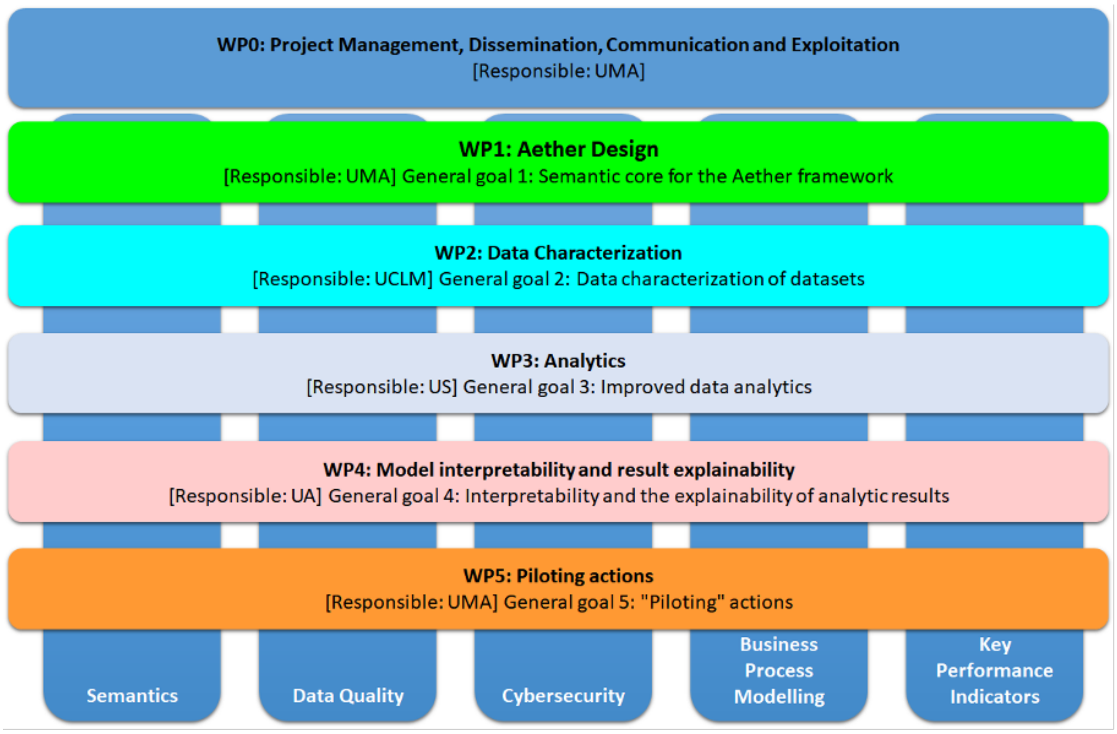

WP0: Project Management, Dissemination, Communication and Exploitation
[Responsible: UMA]
WP1: Aether Design
[Responsible: UMA] General goal 1: Semantic core for the Aether framework
WP2: Data Characterization
[Responsible: UCLM] General goal 2: Data characterization of datasets
WP3: Analytics
[Responsible: US] General goal 3: Improved data analytics
WP4: Model interpretability and result explainability
[Responsible: UA] General goal 4: Interpretability and the explainability of analytic results
WP5: Piloting actions
[Responsible: UMA] General goal 5: "Piloting" actions
Semantics
Data Quality
Cybersecurity
Business Process Modelling
Key Performance Indicators

Figure 1: Project goals

- **General goal 1 (responsible UMA):** To design the semantic core for the Aether framework to deal with relevant aspects of data, process and business analytics.
- **General goal 2 (Responsible UCLM):** To investigate data characterization mechanisms for the analysis of datasets (not only Big Data) to obtain semantic (domain), topological, statistical (data biases, kind of distribution, data type, descriptive statistics, categories) and other properties (e.g., quality, security, provenance).
- **General goal 3: (Responsible US)** To design and develop new solutions for improved data analytics, process analytics and business analytics, taking advantage of the analysis of the datasets and processes.
- **General goal 4 (Responsible UA):** To facilitate the interpretability and the explainability of analytic results, with particular focus on aligning Machine Learning techniques with business objectives. This will lead to new insights and the identification of novel, relevant analysis.
  - **Specific goal 4.1 (I. Navas):** To study the use of knowledge graphs and ontologies to enhance the explicability of the models (of the algorithms) used in the analysis of data, to improve the actionability of results maintaining the privacy rules about data determined.
  - **Specific goal 4.2 (J.C. Trujillo):** To define a methodology to derive indicators and visualizations from the metadata generated by the various investigations and consolidated in BIGOWL, taking into account user goals, semantics and machine learning algorithms outputs.
  - **Specific goal 4.3 (A.J. Varela):** Reasoning about the security vulnerabilities and vulnerable configurations to improve the interpretability and usability based on the results over the datasets and the analysis of their security requirements, features, configurations, and the information of the system.
  - **Specific goal 4.4 (J.A. Cruz):** To define quantitative (synthetic) indicators to measure the interpretability of the results of data analysis and machine learning algorithms. These indicators will allow us to warn users about underlying aspects that may affect their interpretation.
  - **Specific goal 4.5 (M.T. Gómez):** To study the alignment between the applied techniques in the data analytics of business processes, and the obtained information, using the knowledge graphs and ontologies to facilitate the interpretability and usability of the results (of the algorithms) of the data analysis even before the algorithms are applied to ascertain the complexity and the usability of the used techniques.
  - **Specific goal 4.6 (A. Maté):** To develop reasoning techniques that use the semantics of the analytic workflow and KPI definitions to facilitate the identification of potential causes for underperforming objectives as well as their implications for the organization.
- **General goal 5 (Responsible UMA):** To develop ”Piloting” actions, through the development of pilot experiences with companies and organizations in the application of the results of the project.

4. Methodology and Work Plan

4.1. Methodology

Our work methodology builds on the reference framework provided by the Enterprise Unified Process (Enterprise Unified Process. Several authors. Enterprise Unified Process Publishing, 2020, ISBN: 978-1867420781), which proposes an iterative, incremental life cycle that is very appropriate for research projects. It has been adapting it for the last decade, so it is very solid nowadays. As a summary, the main phases into which the project was divided and then the disciplines that must be performed are as follows: (i) Inception, (ii) Preparation, (iii) Construction, (iv) Transition, and (v) Recapping.

- Inception: the goal is to identify the working hypotheses and the objectives, verify that they are realistic and aligned with current trends, ensure that they are well aligned with the existing research programs, and check that they will have positive scientific, technical, social and economic impacts. It is a good idea to use brainstorming meetings in this phase, so that researchers may contribute with their ideas. This phase has been developed at the beginning of the project proposal preparation.
- Preparation: the goal is to work on the general goal and the motivation of the project plus a summary of the state of the art that supports them. It is a good idea to use a brainstorming meeting so that every researcher can contribute to his or her ideas. After that, the focus must shift towards devising the work packages, their specific goals, the dissemination, transfer, and contingency plans. It is a good idea to work in small groups in brainstorming meetings so that every researcher can contribute ideas. This phase has been developed during the project proposal preparation. Still, it will be revisited in the first month of the project development to update it according to any new insight discovered.
- Construction: this phase focuses on defining the goals within each work package and dividing these work packages into several tasks. It is a good idea to work in small groups that organise a kick-off meeting per goal in which they agree on the tasks to be done, the workflow, the intermediate milestones, and the key performance indicators. Then, the researchers work individually or very small groups and meet periodically in plenary workshops in which they present their results, discuss new ideas, and identify deviations and corrective actions.
- Transition: the goal is to transfer the research results to the end-users and the industry. We will pay special attention to transferring them to the companies that support the proposal using seminars and workshops. We will also pay special attention to complementary projects in which we can apply our results in the context of their industrial projects.
- Recapping: the goal is to recap on how the project has gone on to identify strong and weak points that may help work better in future projects.

Our software development methodology combines Kanban and iterative prototyping. Kanban is an agile methodology based on keeping a board to visualise the tasks to carry out (backlog), in progress, and finish. Thus, the team members can see all the tasks in their context. When some tasks are completed, the next one is taken from the backlog. The idea behind using prototypes is to develop incomplete versions of the software project to get a working system quickly. So, it can be easily tested, and the users can provide feedback early. Once a prototype is ready, a new one is planned to cover new features, and so on. Prototyping can be combined with continuous integration, so that software is produced in short cycles to ensure that it can be ready to be released at any time by frequently applying the steps of testing, building, and releasing. The project will be managed with the help of OpenProject (https://www.openproject.org/), which was specifically designed to help with project planning and scheduling, to define user stories and tasks, to assign them to developers, and to report on timings and costs.

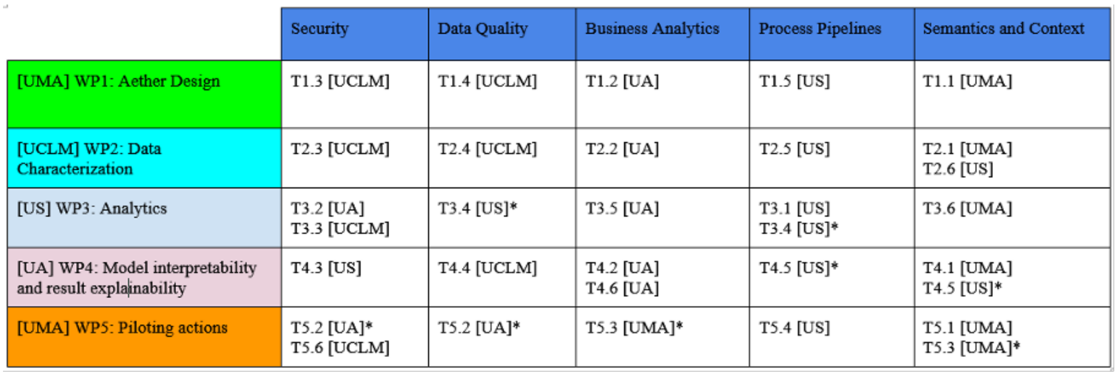
The source code of the software projects will be monitored by the Travis (https://www.travis-ci.org) continuous integration system. The applications will be packed into containers using Docker (https://www.docker.com), which has proven to help deploy the results very easily.

5. Work plan

The table shown in Figure 2 summarises the work plan and how it is related to interdisciplinary dimensions of the project (* means that the task involves several dimensions). In the following Section 5.1, we describe the main work Package developed by the University of Alicante pointing out the main role of conceptual modeling in different areas.

5.1. WP4: Model interpretability and result explainability (WP Leader: A. Maté)

The goal of this WP is to facilitate more confident and accurate decisions by improving the interpretability and explainability of analytic workflows and outputs. The aim is to provide decision-makers with additional pieces of evidence and insights on the rationale behind analytical workflows such as data subsets that lead to the activation of Machine Learning & Artificial Intelligence rules, the semantic relationship between the data being analyzed and other data



	Security	Data Quality	Business Analytics	Process Pipelines	Semantics and Context
[UMA] WP1: Aether Design	T1.3 [UCLM]	T1.4 [UCLM]	T1.2 [UA]	T1.5 [US]	T1.1 [UMA]
[UCLM] WP2: Data Characterization	T2.3 [UCLM]	T2.4 [UCLM]	T2.2 [UA]	T2.5 [US]	T2.1 [UMA] T2.6 [US]
[US] WP3: Analytics	T3.2 [UA] T3.3 [UCLM]	T3.4 [US]*	T3.5 [UA]	T3.1 [US] T3.4 [US]*	T3.6 [UMA]
[UA] WP4: Model interpretability and result explainability	T4.3 [US]	T4.4 [UCLM]	T4.2 [UA] T4.6 [UA]	T4.5 [US]*	T4.1 [UMA] T4.5 [US]*
[UMA] WP5: Piloting actions	T5.2 [UA]* T5.6 [UCLM]	T5.2 [UA]*	T5.3 [UMA]*	T5.4 [US]	T5.1 [UMA] T5.3 [UMA]*

Figure 2: Work plan

sets stored by the organization, or the connection between analytic outputs and business goals and their implications.

- **Task 4.1. (UMA, UA, US, UCLM):** Knowledge graph based explainability (9 Months). This task will analyse the use of knowledge graphs to enhance the understanding of the datasets and the models used to transform them through data analytics. The semantics defined at the different levels of knowledge will enable the exploration of these knowledge graphs to extract relevant knowledge. Data privacy will be preserved when exploring external knowledge graphs.
  - **Deliverable 4.1** [REPORT] Analysis of knowledge graphs to improve model explainability (M39).
- **Task 4.2. (UA, UMA):** Semantically enhanced dashboards and scorecards (9 months). This task aims at developing a methodology for deriving dashboards and scorecards exploiting the semantic information in the Aether framework. To this aim, previous works of the UA will be combined with the expertise on semantics from UMA to create a methodology that takes into account not only user goals and the data at hand, but also data semantics to create richer visualizations and dashboards.
  - **Deliverable 4.2** [REPORT] Methodology for the derivation of semantically enhanced dashboards and scorecards (M39).
- **Task 4.3. (US, UCLM):** Testing of Vulnerabilities (9 months). Automatic creation of security tests from exploiting databases based on the extracted vulnerabilities from relevant databases to create models that describe the possible vulnerabilities and evolution of the systems. These models will be used to verify and diagnose the running configuration further to define tests according to their potential vulnerabilities.
  - **Deliverable 4.3** [REPORT, SOFTWARE] Report the automatic extraction of the models from databases, and software to verify and diagnose configurations, and security tests (M41).
- **Task 4.4. (UCLM, UMA, UA):** Data Analytics Results' Interpretability (12 months). The interpretation of the results of the Machine Learning models is vital because (1) it helps to correct unwanted biases, (2) it facilitates the detection of noise that alters the prediction and (3) it allows to confirm that the variables that the model has selected are significant in the framework of the problem. However, when faced with different possible models, there is currently no mechanism for measuring the level of interpretability of the model. Therefore, in this task, a series of experiments will be carried out to identify possible metrics that will allow the user to define the level of interpretability of the algorithms.
  - **Deliverable 4.4** [REPORT] Experimental results in measuring interpretability of data analytics results in the project (M44).
- **Task 4.5. (US, UMA, UA):** Alignment among processes and discovery techniques (10 months). How the best configuration of the process mining techniques can be ascertained before the techniques are applied. To fulfil this task, it is necessary to characterise the processes to discover the relation between the parameters of the techniques and the processes where they are applied.

  - **Deliverable 4.5** [REPORT] Methodology to facilitate the alignment of the most suitable algorithms and setting according to data sets in process discovery (M44)

- **Task 4.6. (UA, UMA):** Business reasoning framework (10 months). This work package aims at developing a novel methodology for reasoning on the current performance of the organization by making the system aware of existing KPIs and analytic workflows (ML outputs and data transformations) to provide as much insights as possible on the root cause of underperforming business objectives.
  - **Deliverable 4.6** [REPORT] Business reasoning framework description (M46).

6. Project relevance

Aether project aims to enhance the whole data analytics process, seamlessly augmenting datasets and algorithms with semantic and domain knowledge, while also improving their security, quality and actionability. The expected result is a new way to approach data analytics, enabling advanced reasoning and applications with less effort, augmenting the data catalogue incorporated in the business decisions. In turn, this change has profound implications for academia, society, and industry. Numerous organizations, such as hospitals, universities or industries, can harness this new approach to easily integrate new datasets into their analysis while having a clearer view of their implications for their objectives.

7. Current state

As it can be summarized, the project is highly focused on "Data". In a more particular way, we should point out that conceptual modeling has a relevant role in many Work Packages and transversal tasks. One of the main novel research lines under development in this project is the modeling of user's requirements for Machine Learning Applications. Currently, we are working on the extension of the iStar [2] modeling approach in order to be able to gather the main ML requirements.

On the other hand, data bias is one of the most relevant issues to be tackled when working with external data and/or developing ML models. In this way, we have developed several works where conceptual modeling and visualization play relevant roles in finding the main data bias in a set of data [3, 4, 5, 6]. These works can be applied to small and also Big Data sources [7, 8]. On the other hand, one of the current research lines will be focused on the modeling of ML models in order to model the most relevant data to train the ML models [9, 10, 11, 12, 13].

We also apply different Process Mining techniques for the discovery of defacto process models, conformance analysis and extensions by combining AI models. We use algorithms such as Fuzzy Miner to manage levels of abstraction in the models by tackling both correlation and relevance variables, and the Trace Alignment algorithm, for clustering based on Agglomerative Hierarchical Clustering (AHC) to detect patterns in process traces [14].

References

[1] C. Barba-González, J. García-Nieto, M. del Mar Roldán-García, I. Navas-Delgado, A. J. Nebro, J. F. Aldana-Montes, Bigowl: Knowledge centered big data analytics, Expert Systems with Applications 115 (2019) 543–556.

[2] J. M. Barrera, A. Reina-Reina, A. Lavalle, A. Maté, J. Trujillo, An extension of istar for machine learning requirements by following the prise methodology, Available at SSRN 4358075 (2023).

[3] A. Lavalle, A. Maté, J. Trujillo, J. García-Carrasco, Law modeling for fairness requirements elicitation in artificial intelligence systems, in: Proceedings of the 41st International Conference on Conceptual Modeling, ER 2022, Springer, 2022, pp. 423–432.

[4] M. del Mar Roldán-García, J. García-Nieto, A. Maté, J. Trujillo, J. F. Aldana-Montes, Ontology-driven approach for kpi meta-modelling, selection and reasoning, International Journal of Information Management 58 (2021) 102018.

[5] A. Lavalle, A. Maté, J. Trujillo, J. García, A methodology based on rebalancing techniques to measure and improve fairness in artificial intelligence algorithms, in: Proceedings of the 24nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data DOLAP@EDBT/ICDT 2022, volume 3130 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 81–85.

[6] A. Lavalle, A. Mate, J. Trujillo, M. A. Teruel, A data analytics methodology to visually analyze the impact of bias and rebalancing, IEEE Access 11 (2023) 56691–56702.

[7] A. Maté, J. Peral, J. Trujillo, C. Blanco, D. García-Saiz, E. Fernández-Medina, Improving security in nosql document databases through model-driven modernization, Knowledge and Information Systems 63 (2021) 2209–2230.

[8] C. Blanco, D. García-Saiz, D. G. Rosado, A. Santos-Olmo, J. Peral, A. Maté, J. Trujillo, E. Fernández-Medina, Security policies by design in nosql document databases, Journal of Information Security and Applications 65 (2022) 103120.

[9] S. García-Ponsoda, J. García-Carrasco, M. A. Teruel, A. Maté, J. Trujillo, Feature engineering of eeg applied to mental disorders: a systematic mapping study, Applied Intelligence (2023) 1–41.

[10] J. García-Carrasco, A. Maté, J. Trujillo, A data-driven methodology for guiding the selection of preprocessing techniques in a machine learning pipeline, in: International Conference on Advanced Information Systems Engineering, Springer, 2023, pp. 34–42.

[11] R. Tardío, A. Maté, J. Trujillo, Beyond tpc-ds, a benchmark for big data olap systems (bdolap-bench), Future Generation Computer Systems 132 (2022) 136–151.

[12] J. M. Barrera, A. Reina, A. Mate, J. C. Trujillo, Fault detection and diagnosis for industrial processes based on clustering and autoencoders: a case of gas turbines, International Journal of Machine Learning and Cybernetics 13 (2022) 3113–3129.

[13] A. Reina Reina, J. M. Barrera, B. Valdivieso, M.-E. Gas, A. Maté, J. C. Trujillo, Machine learning model from a spanish cohort for prediction of sars-cov-2 mortality risk and critical patients, Scientific Reports 12 (2022) 5723.

[14] J.-F. Rodríguez-Quintero, A. Sánchez, L. Iriarte Navarro, A. Maté, M. Marco Such, J. Trujillo, Fraud audit based on visual analysis: A process mining approach, 2021-05-21.