MIDA - Method for Industrial Data Analysis based on CRISP-DM

Mateus Mendes*, Torres Farinha
Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, 3030-199 Coimbra
RCM2+ Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, 3030-199 Coimbra
E-mail: mmendes@isec.pt, tfarinha@isec.pt
* Corresponding Author

February 11, 2026

Abstract

As modern computers became increasingly more popular and larger amounts of digital data were available, different methodologies were proposed to extract information from data. CRISP-DM methodology quickly spread and is currently one of the most popular approaches used for data analysis. However, it has some shortcomings, such as beeing too general or business-centered. Different authors have proposed variations, more suitable to specific fields, in order to overcome those limitations. The present paper reviews CRISP-DM, some variations and similar methodologies, and proposes a Methodology for Industrial Data Analysis (MIDA) — a methodology conceived and improved over time, based on previous experience in industrial engineering processes. MIDA consists of eight steps and partially overlaps with CRISP-DM. It has been successfully applied in several previous projects.

Keywords: MIDA, CRISP-DM, Data Analysis, Data Science

1 Introduction

CRoss-Industry Standard Process for Data Mining (CRISP-DM) is one of the most popular methodologies used in data mining and analysis processes. The first references to the methodology date back from 1996. It was developed as part of European CORDIS grants 24959/ 25959 (last checked 2026-01-31), from 1997-07-01 to 1998-12-31 (dates on ISO format YYYY-MM-DD). CRISP-DM was officially presented in 1999. The official document is publicly available online (last checked 2026-01-10) [1].

CRISP-DM is a 6-step methodology. The steps are also referred in the literature as phases. It was conceived to be used in projects where the aim is to extract information from data, and produce models which can be deployed to solve industrial problems. Due to its simple yet powerful structure, CRISP-DM became quickly popular among data scientists, analysts, engineers and other professionals and researchers involved in data analysis projects.

CRISP-DM is a general framework, designed to be applied across different industries. Its level of abstraction is simultanously its power and its main drawback. Is its power because the methodology can and has been successfully used in projects of countless different areas. As long as the steps are correctly followed, the chances of success of the project are very high. But is also its main drawback because it looses specificity requiried for particular projects and fields. In some cases this shortcoming is so important that different authors have proposed variations of the methodology, more adequate to comply with the requirements of particular areas. A comprehensive number of variations are reviewed in Section 3.4.

Physical Asset Management (PAM) is one key area for large industries and institutions, important for optimizing resources and production. Its main goal is to minimize costs of purchasing and ownership of the assets, while maximizing their production and availability. This is a very specific field, involving management, financial data, as well as equipment monitoring and maintenance. Decision-making is increasingly more data-driven, as in many other areas of industry. Part of the data used are financial, management and business related. Most of the times those data can be modeled as time series with very low sampling rate. Other data are normally sensory inputs that come from the assets themselves, reporting their condition, usage and other functioning aspects. Those data are normally time series of variable length and sampling rates.

Whilst the original CRISP-DM methodology can and has been used in PAM projects, its abstract and general structure do cause significant embarrassments in practice, namely when it is necessary to perform all the process necessary for data-driven decision, from requirement analysis to model deployment and maintenance.

This paper proposes a Methodology for Industrial Data Analysis (MIDA), a variation of CRISP-DM that aims to include the most frequent steps necessary for the data collection and analysis explicitly in the project, along with other minor variations that make it more suitable for asset management projects. Nonetheless, the methodology is general enough to be useful in other fields too.

2 Original CRISP-DM

2.1 The 6-step methodology

Figure 1 shows the six steps of the original CRISP-DM methodology. They are:

1. Business understanding — Typically seen as the step where the problem is studied, the objectives of the project are defined and a tentative strategy is outlined.

2. Data understanding — This is the step where data is collected, if not already available, and explored in order to get the first insights both on the data quality and the information it contains.

3. Data preparation — Data are cleaned, transformed and prepared for the modeling phase.

4. Modeling — This is often the step that adds more value to the project, even though sometimes it only lasts a small fraction of the project dura-

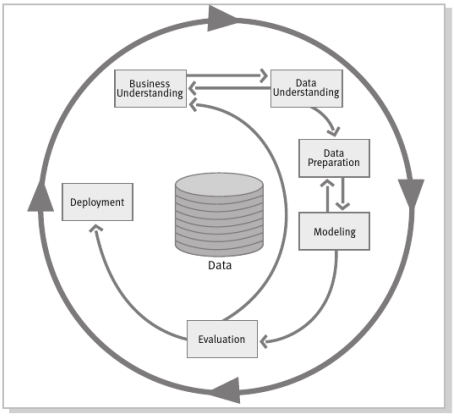

Figure 1: Diagram of the steps of the original CRISP-DM methodology, as proposed in [1].

tion. During this step, statistical, as well as classical or modern machine learning models are tested and trained to model the data to get deeper insights or create models that could be deployed in practice.

5. Evaluation — Sometimes also referred to as "Results," during this step there is an assessement of the quality of the model and how it suits the objectives defined in the Business Understanding step.

6. Deployment — This is the last step, when the models are deployed to production. The outputs of the models can be reports, dashboards, or other tools useful for decision-making.

2.2 Limitations of the original methodology

Even though the CRISP-DM methodology has been widely sucessful, it has quite a few shortcomings that hinder its application in different fields. Numerous gaps have been identified by different authors, as reviewed in Section 3. A quick summary of the main limitations is presented below.

2.2.1 Strong focus on the business aspects

The first step is known as "Business Understanding." It is where goals are defined, and that is done from a business perspective. In many projects the business perspective is not necessarily relevant during this first stage, and stressing this aspect may even cause a loss of focus of the project. For example, in a predictive maintenance project, the focus is usually on reducing equipment downtime and increasing availability. The business perspective is implicit.

2.2.2 Planning step is implicit

No step of the traditional CRISP-DM methodology explicitly refers to the project planning, it is a task of the broader "Business Understanding" step. However, planning is crucial in a project of predictive maintenance or data analysis for data-driven physical asset management, just as in other fields.

2.2.3 Data collection and management are implicit

The original CRISP-DM methodology is also oblivious to the data collection steps in the six main steps. Data collection appears as a task of the broader "Data Understanding" step. Many authors point out that a more explicit reference in the methodology can underline and evidence the importance of this step in the whole project.

Additionally, nowadays the amounts of data generated can easily grow exponentially, posing important challenges to storage and processing systems. Hence, data governance and life cycle management strategies must be defined during the project.

2.2.4 Evaluation step

Naming the analysis of the results simply "Evaluation" can be confuse. In fact, this step is normally when the results are obtained, analysed, compared to the state of the art and the requirements. Then there is a decision whether to get back to one of the previous steps or to proceed ahead to the deployment and/or conclusion of the project. The evaluation is also focused on the results of the models and not a general balance of the project.

2.2.5 Legal and regulatory issues are not addressed

CRISP-DM, as well as other similar methodologies, were conceived in a time when there was still little concern about legal issues when accessing and using the data. In the meantime, however, the regulatory framework evolved in Europe and other continents. In the European Union, Directive 95/46/EC of the European Parliament and of the Council of 1995-10-24 focuses on the protection of individuals with regard to the processing of personal data and on the free movement of such data. This Directive was later replaced by Regulation (EU) 2016/679, popularly known as General Data Protection Regulation (GDPR). This regulation aims to protect personal data and must be taken into account when data is harvested, collected, stored and used for data analysis. More recently, the AI Act, Regulation (EU) 2024/1689, poses rules and constraints on the processing of massive amounts of data using artificial intelligence methods. Those regulations and subsequent local and regional laws and rules must be taken into account in modern data analysis projects.

3 Literature Review

As soon as technology evolved and electronic data was available, after initial research on different knowledge discovery methodologies, some methodologies were proposed. The most notable ones are reviewed below.

3.1 CRISP-DM and Alternative Methodologies

Knowledge Discovery in Databases (KDD) methodology was formulated as early as 1989 [2], even though the landmark paper usually cited as clarifying the steps of the methodology was published in 1996 by Fayyad *et al.* [3]. Compared to

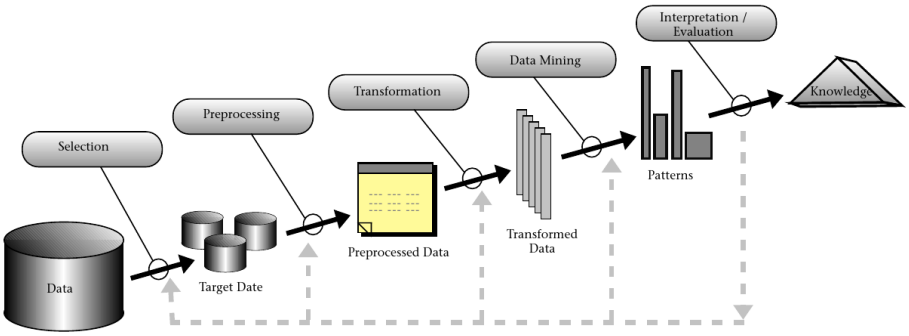

Figure 2: Diagram of the KDD methodology, as proposed in [3]

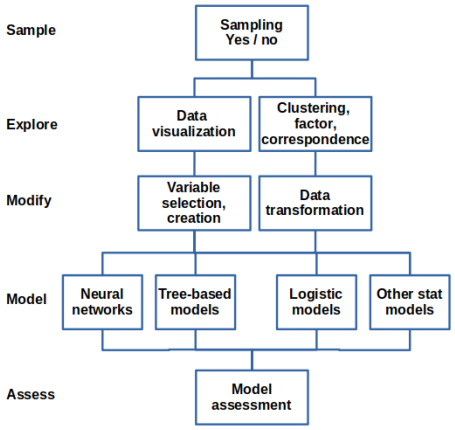

Figure 3: Diagram of the SEMMA methodology, as proposed in [4]

CRISP-DM, KDD is oblivious to the business-side of the project and the deployment issues. Nonetheless, it is often cited as one important landmark in the history of the methodologies and to this day it is still one important alternative to CRISP-DM. Figure 2 illustrates the steps of the KDD methodology.

Sample, Explore, Modify, Model, Assess (SEMMA) was proposed as early as 1997 [4]. Compared to CRISP-DM, this methodology ignores the business and the deployment steps, just as KDD does. Figure 3 illustrates the SEMMA methodology. It is arguably one of the three most popular methodologies, along with KDD and CRISP-DM.

Following on the footstaps of KDD and SEMMA, CRISP-DM was originally proposed in a 1999 document [1] and quickly became the preferred methodology among key players in academia and industry. Wirth and Hipp present one of the first case studies of the successful application of the methodology [5]. Azevedo and Santos compare KDD, SEMMA and CRISP-DM methodologies, highlighting the differences and similarities between the three approaches [6].

Another methodology that gained some popularity was OSEMN — Obtain, Scrub, Explore, Model, iNterpret. According to online records (last checked 2026-01-31), it first appeared in 2010, but has received much less attention from the academic community, when compared to the previously mentioned three

proposals. Shameti and Cico [7] compare OSEMN to CRISP-DM and conclude for the superiority of the latter. Nonetheless, OSEMN has been sucessfully in different data science projects. Dineva and Atasanova [8] used OSEMN in a data science project to process IoT data from beehives. Kumari *et al.* [9] also report the use of the methodology in a sentiment analysis project.

In 2021, Martínez-Plumed *et al.* [10] affirm that CRISP-DM is still the *de facto* standard methodology for data mining and data science projects. This is observed, even though the methods and algorithms evolved over time. Additionally, the term "data science" slowly replaced the old designation of "data mining," more prevalent when CRISP-DM was proposed, for information and knowledge extraction projects. The authors also affirm that when data science projects become more exploratory, a more flexible model is preferred. In the same year, a study by Shröer *et al.* reviews the application of CRISP-DM in data science projects and comes to similar conclusions [11].

In 2024, Shimaoka *et al.* [12] published another survey where they affirm that 82% of the teams do not use any process model — however, CRISP-DM is still the preferred method due to its flexibility, robustness and adaptability to different domains. The authors also compare 16 variations of the methodology and analyze their differences and similarities.

3.2 Comparison of CRISP-DM and alternatives

Zavaleta-Sánchez *et al.* [13] compare the use of KDD and CRISP-DM in a data mining project for phishing identification. They consider that CRISP-DM is more detailed, but KDD is more intuitive and can be followed more naturally, without awareness of being following a formal methodology. The conclusions are understandable, considering that KDD steps are very specific, even though some common steps are missing (*e.g.*, requirement analysis, planning objectives, exploratory data analysis). CRISP-DM entails those tasks, but under a very high level formulation, where problem understanding, requirement analysis and planning fall into the step of "Business Understanding" and exploratory data analysis falls into the step of "Data Understanding."

Palacios *et al.* [14] compare the use of SEMMA and CRISP-DM on a project to construct a repository for studies of land use and cover change. They conclude that SEMMA is more adequate for projects when specific tools are used, while CRISP-DM provides a more general framework. The successful application of the latter, however, may require adherence to tasks/substeps of the main six steps.

Rosander [15] compares the use of CRISP-DM, SEMMA and KDD for signal processing projects. In the particular case, the signals were from radar warning systems. The conclusion was that CRISP-DM was the best methodology, because of its ease of implementation in an agile Scrum project.

3.3 Applications

CRISP-DM has been used in different areas, including industrial, engineering projects [16], teaching [17, 18], paediatrics [19], agriculture [20].

Bosnjak *et al.* [21] report on the use of the CRISP-DM methodology to analyse data from small and medium enterprises. They conclude that many of

the difficulties of the process could have been overcame if the data collection process had been supervised by the data analysists.

Kannengiesser and Gero [22] compare CRISP-DM to models of design in different engineering fields and service design, using a Function-Behaviour-Structure framework. They conclude CRISP-DM is the most solution-oriented methodology, rather than problem-oriented. Therefore, CRISP-DM could be improved by introducing in the methodology some more attention to the problem.

Saltz [23] analyses the use of CRISP-DM in data science projects and concludes that, while it is a robust and powerful method, it misses some key aspects of the project life cycle. Hence, the suggestion is to combine CRISP-DM with other popular project management approaches, such as Scrum.

Plotnikova *et al.* [24] use CRISP-DM in a financial data analysis project. They analyse the strengths and weaknesses of the methodology and identify a total of 18 shortcomings, grouped into three main categories: i) interdependencies between the six steps; ii) requirement analysis and validation; and iii) universality of some of the steps.

3.4 Variations

Because of its flexibility and robustness, CRISP-DM became widely popular. Because of its limitations, some authors have proposed improvements or adaptations of the method for particular fields. Shimaoka *et al.* [12] present a thorough review of the data science methodologies, as well as other agile project management methodologies (namely Scrup and Kanban).

3.4.1 DMME

Huber *et al.* [25] propose Data Mining for Engineering Applications (DMME), as an extension of CRISP-DM to engineering applications. The authors also identify that CRISP-DM does not explicitely foresee a data acquisition phase within production scenarios. According to the authors, DMME provides a communication and planning foundation for data analytics. The key differences of DMME when compared to CRISP-DM are: i) There are two steps added between Business Understanding and Data Understanding. Those steps are "Technical Understanding" and "Technical Realization;" and ii) There is one more step between Evaluation and Deployment. That step is "Technical Implementation." Those new steps may mitigate some of the shortcomings of the original methodology, but it is evident that their names are still quite general.

Cazacu and Titan [26] redefine some tasks of the CRISP-DM methodology for better fit in data science projects. Namely, the business understanding and data understanding are the steps that suffer more changes. This is in line with observations from other authors, who also point out the first steps of CRISP-DM are the ones that exhibit more gaps.

Venter *et al.* [27] propose a variation of CRISP-DM for evidende mining in forense data. The variation is called CRISP-EM, where EM stands for "Evidence Mining." The key differences to the original methodology are: i) Business Understanding is called "Case Understanding;" ii) Modeling is called "Evidence Modeling;" iii) "Evaluation" is also called "Evidence Extraction;" and iv) there is "Evidence Reporting" in place of Deployment.

Nagashima and Kato [28] propose APREP-DM, as a framework for automatic pre-processing of sensory data based on CRISP-DM. The authors evaluate KDD, SEMMA and CRISP-DM, and propose a variation of the latter where data preparation includes specific preparation and cleaning steps, performed automatically.

Ayele [29] proposes CRISP-IM, an adaptation of CRISP-DM for idea mining. The original method was modified, so that instead of Business Understanding there is technology need assessment. Data Understanding is called "Data collection and understanding," Modeling is renamed "Modeling for Idea Extraction," Evaluation is called "Evaluation and Idea Extraction," and Deployment is replaced by "Reporting Inovative Ideas."

Bokrantz *et al.* [30] focus on the application of AI tools in manufacturing and propose an enhanced version of CRISP-DM, with an additional step of "Operations and Maintenance." This new step is placed after Deployment, and therefore stands between the end of a project (or iteration) and the Business Understanding step at the beginning of a new project (or iteration).

Kristoffersen *et al.* [31] use the CRISP-DM methodology in the context of circular economy data analysis. The method is enhanced with an additional phase of "Data Validation" and integrates the concept of "analytic profiles." The Data Validation step is introduced after the typical Data Preparation. The Analytic Profiles are used between Business Understanding and Data Understanding, in order to improve communication of knowledge/insights.

Acuña-Cid *et al.* [32] propose the integration of Network Analysis methods with CRISP-DM methodology. The resulting method is called CRISP-NET. The steps are the same of the original method. However, the subtasks of each step are strictly aligned with the network analysis methodoly, aiming at better integration of the relationships within the systems.

Catley *et al.* [33] propose CRISP-TDM, a variation of the original method that aims at incorporating a temporal dimension into the process, specifically tailored for the case of medical data. Steps 1, 2, 4 and 6 of CRISP-DM are enriched with tasks such as clinical reporting.

Asamoah and Sharda [34] propose CRISP-eSNeP — Cross Industry Standard Process for Electronic Social Network Platforms. This variation is specially tailored to process social networks' big data. The steps are very different from the original method: Cluster Development, Data Acquisition, Data Cleaning, Data Formatting, Data Validation, Data Analysis, and Deployment.

Ramos *et al.* [35] propose CRISP-EDM, a variation adapted to projects of data analysis in the educational domain. The steps are the same, even though renamed and with the tasks slightly adapted to the educational domain.

Niaksu [36] proposes an extension of CRISP-DM for the medical field. The modified method is called CRISP-MED-DM. In this proposal, the six steps of the original methodology are maintained, only the tasks are adapted to the particularities of the medical field.

Shafer *et al.* [37] propose QM-CRISP-DM, a variation that aims to adapt the methodology for quality management. The steps of the original CRISP-DM methodology are maintained, only the tasks are redefined to integrate quality management tools.

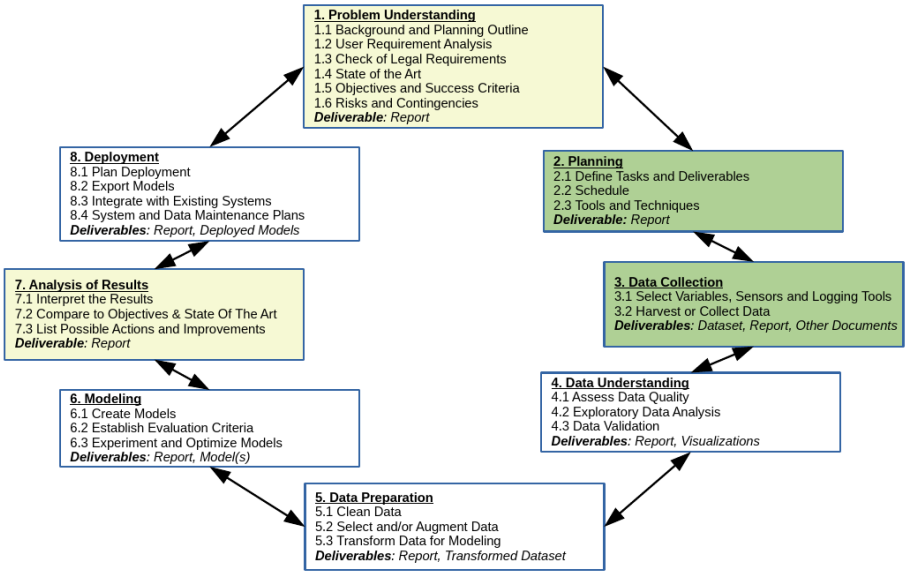

Figure 4: Diagram of the proposed methodology. Steps different from the original CRISP-DM are highlighted in color: yellow steps are significantly changed, green are new

4 Methodology proposed

Figure 4 gives an overview of the proposed method. It shows the eight steps of the methodology and also the tasks of each step. Hence, it aims at being a reference guide of the MIDA methodology.

MIDA consists of eight steps, with a variable number of tasks. Some of the steps coincide with those of CRISP-DM, others are variations or proposed as new. As in other methodologies, the steps and tasks are optional and the methodology is agile, flexible and iterative.

The steps and tasks are described in more detail below.

1. **Problem Understanding** — Inspired by the Business Understanding step of CRISP-DM, this is the first step of the MIDA methodology. Its name, however, aims to give the method a more problem-oriented focus, which is possibly more adequate in engineering and industrial settings. The tasks in this step are described below.

   1.1 Background and Planning Outline — Most industrial projects start with an idea or need that is identified. The first task should be a quick check of the background need or idea that led to the project, in order to get an overview of the needs and feasibility of the project. Initial planning of the project should be done.

   1.2 User Requirement Analysis — If the first background check is successful, the project should proceed to the second task, which is a more detailed analysis of the requirements. Adequate formal methodologies, such as use cases, can and must be used if needed.

1.3 Check of Legal Requirements — Legal implications should be checked in this step. Those include, for example, restrictions on the access to the data, imposed by copyright protections or regulations such as GDPR or the AI Act.

1.4 State of the Art — Once the user requirements are understood and the legal aspects are verified, a survey of the state of the art is recommended. This is important to verify the latest solutions available, from the technical and scientific points of view. This is fundamental to help establish realistic objectives and lay the ground for the Planning step.

1.5 Objectives and Success Criteria — Once the previous tasks are completed, clear objectives and success criteria should be defined.

1.6 Risks and Contingencies — A risk analysis and solutions to mitigate those risks should also be done, before proceeding to the next steps.

Deliverable: Report — The output of this step should be a report, detailing the main findings of the Problem Understanding step, along with the conclusions, and proposing the roadmap for the next steps.

2. **Planning** — After Problem Understanding, detailed planning should be performed. Adjustments and refinements to the initial planning should be performed as needed.

2.1 Define Tasks and Deliverables — The work should be divided into adequate tasks. Deliverables should be defined.

2.2 Schedule — Tasks should be scheduled and milestones defined. Adequate tools, such as Gantt diagrams, should be used as needed.

2.3 Tools and Techniques — The methods and strategy to be used to achieve the goals should be defined. Materials necessary to the project should be listed and chosen. This includes data analysis software and other materials, if needed, for data collection, storage or deployment.

Deliverable: Report — The output of this step should also be a report, which includes the planning as well as list of methods and tools necessary for the successful completion of the project.

3. **Data Collection** — If the data necessary for the project is not immediately available, it should be collected in this step. The MIDA tasks to perform during this step are as follows.

3.1 Select Variables, Sensors and Logging Tools — Data collection could be achieved through different methods, depending on the type of problem. Sensors and support equipment, such as data loggers, may be necessary.

3.2 Harvest or Collect Data — The necessary, or possible, amounts of data, should be harvested or collected in order to build the dataset necessary for the following steps.

Deliverables: Dataset, Report, Documentation — The output of this step is a dataset, along with a report detailing the collection process, variable specifications and other details relevant for the analysis process. Other materials can be included, such as equipment manuals, datasheets or other technical documents.

4. **Data Understanding** — Once the dataset is available, the next step is its exploration, using adequate tools. In this step the first conclusions about the data and the processes in their origin are made. The tasks in this step are as follows.

    4.1 Assess Data Quality — The first task is assessment of the quality of the data. Missing values, discrepant values and artifacts, should be identified.

    4.2 Exploratory Data Analysis — This task involves normally descriptive statistics, different visualizations and manipulations that lead to a better understanding of the equipment and/or processes that generated the data.

    4.3 Data Validation — From the other tasks, conclusions can be taken about which parts of the dataset are relevant for the remainder of the project, which ones may need transformation, as well as those that must be discarded.

    Deliverables: Report, Visualizations — The output of this step should be a report with the statistical results, visualizations and other descriptions and conclusions about the quality of the data and the nature of the underlying processes.

5. **Data Preparation** — Once the data and the underlying processes are understood, data should be prepared for the modeling step. The MIDA tasks are the following.

    5.1 Clean Data — Invalid data, such as discrepant samples, or irrelevant data, such as categories with insufficient data, should be removed from further processing. Variables can also be eliminated in this task based on different criteria.

    5.2 Select and/or Augment Data — The important variables, or data chunks, should be selected for further steps. Data augmentation techniques could be applied, if needed, for balancing of skewed datasets.

    5.3 Transform Data for Modeling — Data should be transformed as needed for the models to be developed in the next step. This could involve normalization, standardization, application of sliding windows, filters and other transformations. Encoding of qualitative variables is also performed in this step.

    Deliverables: Report, Transformed Dataset — The output of this step should be a dataset prepared for the modeling step, along with an explanatory report.

6. **Modeling** — This step is the one where normally more value is added to the data analysis process. However, it is often one of the shortest ones. The MIDA tasks are described below.

    6.1 Create Models — Most models are nowadays classical machine learning or deep learning models. Supervised and unsupervised machine learning models are often used. Statistical, numerical and other approaches can also be used in many processes.

    6.2 Establish Evaluation Criteria — The models should be evaluated with adequate performance metrics. The metrics should be chosen according to the type of variables and models applied, as well as the objectives and aceptance criteria defined in the first step.

    6.3 Experiment and Optimize Models — Models should be optimized according to the results of the evaluation criteria.

    Deliverables: Report, Model(s) — The output of this step should include the optimized models as well as an explanatory report.

7. **Analysis of Results** — This step includes thorough analysis of the best results, physical interpretation, scope, advantages and limitations of the approach. The MIDA tasks should be as follows.

    7.1 Interpret the Results — The results must be interpreted to uncover the physical, processual and other possible implications. This often requires deep understanding of the tools, equipment and processes involved.

    7.2 Compare Results to Objectives & State Of The Art — The results must be compared to the objectives of the project established in the first step, as well as the state of the art, for better framing of their quality, as well as advantages and limitations of the approach followed.

    7.3 List possible actions and improvements — After interpretation of the results in a broad context, lines of action should be defined, namely aiming at the deployment step if the results are acceptable, or one of the previous steps of the methodology.

    Deliverable: Report — The output of this step is normally a report. This report can contain the conclusion of the project if there is no deployment step.

8. **Deployment** — Last step of the project. The implementation depends not only on the objectives of the project but also on the quality of the results obtained. The MIDA tasks are listed below.

    8.1 **Plan Deployment** — Deployment of the models and/or results in production often requires specific and careful planning. In industrial settings, for example, it can impact the production lines, pose safety risks or economic risks.

8.2 Export Models — It is desirable that models are deployed in standard, or at least platform-independent, formats. This often requires the use of formats and platforms such as ONNX [38] for machine learning models.

8.3 Integrate with Existing Systems — Proper deployment of the new models often requires installation of software, as well as creation of pipelines for data inputs and channeling of model outputs.

8.4 System and Data Maintenance Plans — Models deployed in industrial settings, for example, require adequate monitoring and maintenance, to guarantee good working conditions and minimize downtime or chances of misbehaviour. Hence, careful planning of the monitoring and maintenance tasks required to maintain the system working well should be outlined before completion of the project. Additionally, the data generated may require life cycle planning and governance strategies, which should be outlined for future maintenance and use of the system.

Deliverables: Report, Deployed Models — The output of this step should be a technical report on the deployment environment, including monitoring and maintenance plans.

5 Discussion

The MIDA methodology has been successfully applied in previous projects led by the authors' team.

Rodrigues *et al.* [39] used the methodology in a project where the goal was to predict bus motor oil condition using artificial neural networks. The data were collected from two different bus companies. The models were not deployed. The datasets in both companies consisted of tens of results of laboratory analysis.

Cruz *et al.* [40] also followed the methodology in a project where the goal was to automatically detect problems in heat-sealed bottles in a chemical industry. In this case, the methodology was followed, from Problem Understanding to the successful deployment of the models in the shop floor.

Malta *et al.* [41] also followed all steps of the methodology in a forest engineering project, where the goal was to determine the wood volume of *pinus pinaster* pine trees. The methodology was followed also from Problem Understanding until model was also deployed in a web-based simulator.

Costa *et al.* [42] applied the full methodology to determine the working states of a plastic injection machine. The model was also deployed in a large plastic industry. Contrary to the previously mentioned projects, where the data collection process was also designed during the project, in this case, the dataset consisted of several years of records already available. Nonetheless, the methodology was successfully applied from Problem Understanding to Deployment.

Silva *et al.* [43] also used the MIDA methodology in a project where the aim was to improve automatic help to encoding of Electronic Health Records in ICD-10 taxonomy. This project did not aim at developing a model ready for deployment. However, it was particularly important because of the sensitivity of the data used. Access to EHR requires particular attention to the ethical and

legal implications of accessing, storing and processing such data, and was one of the main motivators to include task 1.3 in the methodology.

During the projects cited above in this section, the weaknesses revealed by the original CRISP-DM, its variations and competing methodologies lead to the refinement of MIDA over time. In the end, the methodology proposed revealed to be needed and sufficient to the successful completion of all the projects. The methodology is also being used in other ongoing projects, whose results will be published soon.

6 Conclusion

CRISP-DM remains one of the most widely adopted approaches for data-driven projects. However, its high-level structure, lack of explicit planning and data collection stages, and absence of considerations for modern regulatory constraints, reduce its effectiveness in modern real-world industrial environments.

To overcome those shortcomings, we proposed MIDA — Method for Industrial Data Analysis, an eight-step methodology designed and refined through practical applications in engineering and industrial projects. MIDA builds upon the foundations of CRISP-DM but incorporates explicit stages for planning, data collection, legal analysis, and a more detailed and domain-oriented understanding of the problem. Its structure aims to provide a clearer roadmap for projects where multidisciplinary knowledge, regulatory compliance, and data heterogeneity are the norm.

MIDA has already been successfully applied in different real-world projects, such as the analysis of operation states in injection-molding equipment, estimation of forest biomass using deep-learning approaches and ecoding of EHR using ICD-10. These applications demonstrate the methodology's flexibility and its suitability for different scenarios.

The advantages of MIDA include its explicit emphasis on planning, legal requirements, and structured data collection, as well as its balance between technical rigor and practical applicability. Nevertheless, MIDAmay still need adaptation when used in domains radically different from those for which it was originated.

Future work includes validating the methodology across additional industrial sectors, refining specific tasks for particular domains if needed. Ongoing projects developed by the authors and collaborators will provide further feedback to improve MIDA and reinforce its role as a practical, domain-oriented alternative to the classical CRISP-DM framework.

Author contributions

Conceptualization: Mateus Mendes; Methodology: Mateus Mendes and Torres Farinha; Formal analysis and investigation: Mateus Mendes and Torres Farinha; Writing - original draft preparation: Mateus Mendes; Writing - review and editing: Torres Farinha.

Ethics declarations

Conflict of interest

The authors declare that they have no known competing financial or personal relationships with other people or organizations that could inappropriately influence the work reported in the paper. There is no professional or other personal interest of any nature or kind in any product or company that could be

influencing the positions presented in, or the review of, the manuscript.

References

[1] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0*. IBM, 1999 online (last checked 2026-01-31).

[2] Gregory Piatetsky-Shapiro. Knowledge discovery in real databases: A report on the IJCAI-89 workshop. *AI magazine*, 11(4):68–68, 1990 online (last checked 2026-01-31).

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996 online (last checked 2026-01-31).

[4] SAS Institute. Data mining and the case for sampling. Technical report, SAS Institute, 2003 online (last checked 2026-01-31).

[5] Rüdiger Wirth and Jochen Hipp. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the* 4^{th} *international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000 online (last checked 2026-01-31).

[6] Ana Azevedo and Manuel Santos. KDD, SEMMA and CRISP-DM: a parallel overview. In *IADIS European Conf. Data Mining*, volume 8, pages 182–185, 2008 online (last checked 2026-01-31).

[7] Ketjona Shameti and Betim Cico. Comparison of methodological approaches: CRISP-DM vs OSEMN methodology using linear regression and statistical analysis. In *Balkan Conference in Informatics*, pages 47–60. Springer, 2024 online (last checked 2026-01-31).

[8] Kristina Dineva and Tatiana Atanasova. OSEMN process for working over data acquired by iot devices mounted in beehives. *Curr. Trends Nat. Sci*, 7(13):47–53, 2018 online (last checked 2026-01-31).

[9] Kajal Kumari, Mahima Bhardwaj, and Swati Sharma. OSEMN approach for real time data analysis. *International Journal of Engineering and Management Research*, 10(2), 2020 online (last checked 2026-01-31).

[10] Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach. CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061, 2021 online (last checked 2026-01-31).

[11] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. A systematic literature review on applying CRISP-DM process model. *Procedia computer science*, 181:526–534, 2021 online (last checked 2026-01-31).

[12] André Massahiro Shimaoka, Renato Cordeiro Ferreira, and Alfredo Goldman. The evolution of CRISP-DM for data science: Methods, processes and frameworks. *SBC Computing Reviews*, 4(1):28–43, Oct. 2024 online (last checked 2026-01-31).

[13] Esteban Zavaleta-Sánchez, Gabriel Domínguez-Sánchez, Cecilia-Irene Loeza-Mejía, and Eddy Sánchez-DelaCruz. Comparative study of KDD and CRISP-DM methodologies for phishing identification. In *International Congress on Information and Communication Technology*, pages 317–330. Springer, 2024 online (last checked 2026-01-31).

[14] Herman Jair Gómez Palacios, Robinson Andrés Jiménez Toledo, Giovanni Albeiro Hernández Pantoja, and Álvaro Alexander Martínez Navarro. 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*, 2(3):598–604, 2017 online (last checked 2026-01-31).

[15] Antonia Rosander. Evaluating frameworks for implementing machine learning in signal processing. Master's thesis, Kth Skolan för Electroteknik Och Datavetenskap, Sweden, 2018 online (last checked 2026-01-31).

[16] Nils Doede, Paulina Merkel, Mareile Kriwall, Malte Stonis, and Bernd-Arno Behrens. Implementation of an intelligent process monitoring system for screw presses using the CRISP-DM standard. *Production Engineering*, 19(1):77–88, 2025 (online (last checked 2026-01-31)).

[17] Sanjiv Jaggia, Alison Kelly, Kevin Lertwachara, and Leida Chen. Applying the CRISP-DM framework for teaching business analytics. *Decision Sciences Journal of Innovative Education*, 18(4):612–634, 2020 (online (last checked 2026-01-31)).

[18] Jairo Acosta Solano, Diana Janeth Lancheros Cuesta, Samir F. Umaña Ibáñez, and Jairo R. Coronado-Hernández. Predictive models assessment based on CRISP-DM methodology for students performance in colombia - saber 11 test. *Procedia Computer Science*, 198:512–517, 2022 (online (last checked 2026-01-31)). 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare.

[19] Ayi Purbasari, Fedri Ruluwedrata Rinawan, Arief Zulianto, Ari Indra Susanti, and Hendra Komara. CRISP-DM for data quality improvement to support machine learning of stunting prediction in infants and toddlers. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6, 2021 (online (last checked 2026-01-31)).

[20] Lendy Rahmadi, Hadiyanto, Ridwan Sanjaya, and Arif Prambayun. Crop prediction using machine learning with CRISP-DM approach. In Abhishek Swaroop, Zdzislaw Polkowski, Sérgio Duarte Correia, and Bal Virdee, editors, *Proceedings of Data Analytics and Management*, pages 399–421, Singapore, 2023 (online (last checked 2026-01-31)). Springer Nature Singapore.

[21] Z. Bosnjak, O. Grljevic, and S. Bosnjak. CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. In *2009 5th International Symposium on Applied Computational Intelligence and Informatics*, pages 509–514, 2009 (online (last checked 2026-01-31)).

[22] Udo Kannengiesser and John S. Gero. Modelling the design of models: An example using CRISP-DM. *Proceedings of the Design Society*, 3:2705–2714, 2023 (online (last checked 2026-01-31)).

[23] Jeffrey S. Saltz. CRISP-DM for data science: Strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2337–2344, 2021 (online (last checked 2026-01-31)).

[24] Veronika Plotnikova, Marlon Dumas, and Fredrik P. Milani. Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. *Data & Knowledge Engineering*, 139:102013, 2022 (online (last checked 2026-01-31)).

[25] Steffen Huber, Hajo Wiemer, Dorothea Schneider, and Steffen Ihlenfeldt. DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model. *Procedia Cirp*, 79:403–408, 2019 (online (last checked 2026-01-31).

[26] Mihaela Cazacu and Emilia Titan. Adapting CRISP-DM for social sciences. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 11(2Sup1):99–106, 2021 (online (last checked 2026-01-31)).

[27] Jacobus Venter, Alta de Waal, and Cornelius Willers. Specializing CRISP-DM for evidence mining. In *IFIP International Conference on Digital Forensics*, pages 303–315. Springer, 2007 (online (last checked 2026-01-31)).

[28] Hiroko Nagashima and Yuka Kato. APREP-DM: a framework for automating the pre-processing of a sensor data analysis based on CRISP-DM. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 555–560, 2019 (online (last checked 2026-01-31)).

[29] Workneh Yilma Ayele. Adapting CRISP-DM for idea mining: a data mining process for generating ideas using a textual dataset. *International Journal of Advanced Computer Sciences and Applications*, 11(6):20–32, 2020 (online(last checked 2026-01-31)).

[30] Jon Bokrantz, Mukund Subramaniyan, and Anders Skoogh. Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM. *Production Planning & Control*, 35(16):2234–2254, 2024 (online(last checked 2026-01-31)).

[31] Eivind Kristoffersen, Oluseun Omotola Aremu, Fenna Blomsma, Patrick Mikalef, and Jingyue Li. Exploring the relationship between data science and circular economy: An enhanced CRISP-DM process model. In Ilias O. Pappas, Patrick Mikalef, Yogesh K. Dwivedi, Letizia Jaccheri, John Krogstie, and Matti Mäntymäki, editors, *Digital Transformation for a Sustainable Society in the 21st Century*, pages 177–189, Cham, 2019 (online (last checked 2026-01-31)). Springer International Publishing.

[32] Héctor Alejandro Acuña-Cid, Eduardo Ahumada-Tello, Óscar Omar Ovalle-Osuna, Richard Evans, Julia Elena Hernández-Ríos, and Miriam Alondra Zambrano-Soto. CRISP-NET: Integration of the CRISP-DM model with network analysis. *Machine Learning and Knowledge Extraction*, 7(3), 2025 (online(last checked 2026-01-31)).

[33] Christina Catley, Kathy Smith, Carolyn McGregor, and Mark Tracy. Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. In *2009 22nd IEEE International Symposium on Computer-Based Medical Systems*, pages 1–5, 2009 (online(last checked 2026-01-31)).

[34] Daniel Asamoah and Ramesh Sharda. Adapting CRISP-DM process for social network analytics: Application to healthcare. 2015 (online(last checked 2026-01-31)).

[35] Jorge Luis Cavalcanti Ramos, Rodrigo Lins Rodrigues, João Carlos Sedraz Silva, and Pamella Letícia Silva de Oliveira. CRISP-EDM: uma proposta de adaptação do modelo CRISP-DM para mineração de dados educacionais. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1092–1101. SBC, 2020 (online(last checked 2026-01-31)).

[36] Olegas Niaksu. CRISP data mining methodology extension for medical domain. *Baltic Journal of Modern Computing*, 3(2):92, 2015 (online(last checked 2026-01-31)).

[37] Franziska Schäfer, Christian Zeiselmair, Jonas Becker, and Heiner Otten. Synthesizing CRISP-DM and quality management: A data mining approach for production processes. In *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, pages 190–195, 2018 (online (last checked 2026-01-31)).

[38] Purvish Jajal, Wenxin Jiang, Arav Tewari, Erik Kocinare, Joseph Woo, Anusha Sarraf, Yung-Hsiang Lu, George K Thiruvathukal, and James C Davis. Analysis of failures and risks in deep learning model converters: A case study in the onnx ecosystem. *arXiv preprint arXiv:2303.17708*, 2023 (online (last checked 2026-01-31)).

[39] João Rodrigues, Ines Costa, J Torres Farinha, Mateus Mendes, and Luis Margalho. Predicting motor oil condition using artificial neural networks and principal component analysis. *Eksploatacja i Niezawodność*, 22(3), 2020 (online (last checked 2026-01-31)).

[40] Samuel Cruz, António Paulino, Joao Duraes, and Mateus Mendes. Real-time quality control of heat sealed bottles using thermal images and artificial neural network. *Journal of Imaging*, 7(2):24, 2021 (online (last checked 2026-01-31)).

[41] Ana Malta, José Lopes, Raúl Salas-González, Beatriz Fidalgo, Torres Farinha, and Mateus Mendes. Pinus pinaster diameter, height, and volume estimation using mask-RCNN. *Sustainability*, 15(24):16814, 2023 (online (last checked 2026-01-31)).

[42] João Costa, Rui Silva, Gonçalo Martins, Jorge Barreiros, and Mateus Mendes. Analysis of the state and fault detection of a plastic injection machine—a machine learning-based approach. *Algorithms*, 18(8):521, 2025 (online (last checked 2026-01-31)).

[43] Hugo Silva, Vítor Duque, Mário Macedo, and Mateus Mendes. Aiding ICD-10 encoding of clinical health records using improved text cosine similarity and PLM-ICD. *Algorithms*, 17(4):144, 2024 (online (last checked 2026-01-31)).