RESEARCH REPORT ON CRISP - DM

JANUARY 1, 2023

BY
Mayur Korde

Abstract

CRISP-DM is the de-facto standard and an industry-independent process model for applying data mining projects. Twenty years after its release in 2000, we would like to provide a systematic literature review of recent studies published in IEEE, ScienceDirect and ACM about data mining use cases applying CRISP-DM. We give an overview of the research focus, current methodologies, best practices and possible gaps in conducting the six phases of CRISP-DM. The main findings are that CRISP-DM is still a de- factor standard in data mining, but there are challenges since the most studies do not foresee a deployment phase. The contribution of our paper is to identify best practices and process phases in which data mining analysts can be better supported. Further contribution is a template for structuring and releasing CRISP-DM studies.

1. Introduction

Processing huge amounts of data to support decision processes is becoming increasing attention in corporate IT strategies [1,2]. Data Science is the discipline to gain valuable insights from data by mathematical and analytical models and applications. Data science projects can profit from project management and process methodologies. Such methodologies work as success factors [3,4]. However, following a project methodology strictly could be a challenge for data science teams. Process models like CRISP-DM [5] and the Analytics Life Cycle [6] should help and can be enhanced by agile approaches as well [7]. In current data science projects, process models are not well established, although process models exists for several years. As Saltz (2018) shows, only 18% of data science teams follow an explicit process model [3]. Deployment in data science has also become considerable attention in the last years. Due to the own phase of deployment in CRISP-DM we would like to focus on answering how this phase is conducted through the process models in research.

There are several studies about project and process methodologies in data mining or data science projects. However, a recent systematic literature review of the overall process of CRISP-DM is still missing. The studies of Saltz (2016) and Cato (2016) show the importance of process methodologies in big data projects in general [4,8]. Saltz (2017) has compared process methodologies such CRISP-DM and agile approaches by interviewing data science project teams [9]. We focus in this paper on each phase of CRISP-DM only and exclude other process methodologies explicitly because CRISP-DM is still a popular process in practice and in research. Therefore, we will find out benefits and challenges how CRISP-DM is conducted in research studies. This paper contributes to the current state of research by identifying best practices and process phases, in which analysts can be more supported.

This systematic literature review is structured as follows. The following section is about the CRISP-DM process model. The second section describes the research questions and methodology. The third section presents the results. The fourth section shows the interpretation and the limitations of the results.

2. Background to CRISP-DM

CRISP-DM is an industry-independent process model for data mining. It consists of six iterative phases from business understanding to deployment (see Table 1). Table 1 describes the main idea, tasks and output of these phases shortly, based on the user guide of CRISP-DM [5,10].

Table 1: CRISP-DM process model descriptions

Phase	Short description
Business Understanding	The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase. First the data mining type should be explained (e. g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created.
Data understanding	Collecting data from data sources, exploring and describing it and checking the data quality are essential tasks in this phase. To make it more concrete, the user guide describe the data description task with using statistical analysis and determining attributes and their collations.
Data preparation	Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Dependent on the used model (defined in the first phase) derived attributes have to be constructed. For all these steps different methods are possible and are model dependent.
Modeling	The data modelling phase consists of selecting the modeling technique, building the test case and the model. All data mining techniques can be used. In general, the choice is depending on the business problem and the data. More important is, how to explain the choice. For building the model, specific parameters have to be set. For assessing the model it is appropriate to evaluate the model against evaluation criteria and select the best ones.
Evaluation	In the evaluation phase the results are checked against the defined business objectives. Therefore, the results have to be interpreted and further actions have to be defined. Another point is, that the process should be reviewed in general.
Deployment	The deployment phase is described generally in the user guide. It could be a final report or a software component. The user guide describes that the deployment phase consists of planning the deployment, monitoring and maintenance.

3. Method: Systematic literature review

The purpose of a systematic literature review is to identify, evaluate and interpret relevant research papers with regard to specific research questions [11]. Papers about CRISP-DM use cases can facilitate valuable insights in how to conduct CRISP-DM in research and how to structure those studies. In this systematic literature review we follow the instructions and recommendations of Kitchenham (2007): defining the goal, research questions, a search strategy, selecting the papers based on predefined inclusion and exclusion criteria. We defined the following research questions:

- RQ1: Which reasons are mentioned in the papers for using CRISP-DM? CRISP-DM is often noun as the most popular process model in data mining. Therefore, we would like to find evidence in the selected papers.

- RQ2: In which domains is CRISP-DM applied? CRISP-DM is an industry-independent process model. With this research question we would like to quantify the domain interests on conducting CRISPDM and identify domainvenues where related paper about CRISP-DM are conducted.

- RQ3: How are each of the six phases of CRISP-DM be conducted? CRISP-DM was developed from industrial practitioners since there was a need for a standard process model for conducting data mining. However, there are researchers in data mining discipline who follow CRISP-DM as well. Therefore, this research question deals with the concrete phases. Furthermore, it is important how the findings can be transferred back to industry. Answering RQ3, we have developed an initial classification framework. We have adjusted this framework continuously while reading and compared it also with the original user guidelines. Following Kitchenham (2007), we also developed and used a data collection form including all the questions needed to answer the research questions.

We searched in the relevant scientific databases ScienceDirect, IEEE and ACM Digital Library to obtain studies from journals or conferences. Thus, these sources provide a sufficiently broad range of literature in the field of information systems. We have chosen to limit the results regarding the publication year in order to obtain recent best practices from the last completed years between 2017 and 2019. Thus, we can extend the study of [8], as it focuses on the years 2014 to 2016. We applied the following search strategy in the beginning of 2020. The search terms focus on different notations of CRISP-DM. Due to CRISP-DM includes the deployment phase, narrowing the search terms explicitly to deployment focus is not necessary.

("CRISP-DM" OR "CRISP DM" OR "CRoss Industry Standard Process for Data Mining" OR "CRISPDM" OR "CRISP reference model")

We have defined the inclusion criteria that papers must describe use cases and be written in English. Exclusion criteria are enhancements of CRISP-DM, use cases outside of data mining, missing abstracts, grey literature and only pre-proofed studies. Due to better comparability, the exclusion of enhanced CRISP-DM models is mandatory. Totally, we have reviewed 24 studies from three completed years that should be a suitable number for answering the research questions (see Figure 1 and Figure 2). As Figure 2 shows, CRISP-DM is still used by several researchers twenty yearsafter its release.

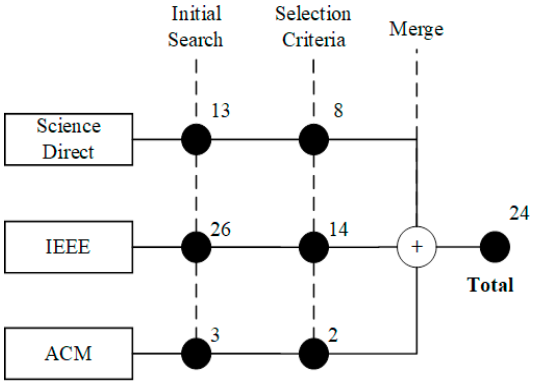

Figure 1: Search strategy for answering research question.

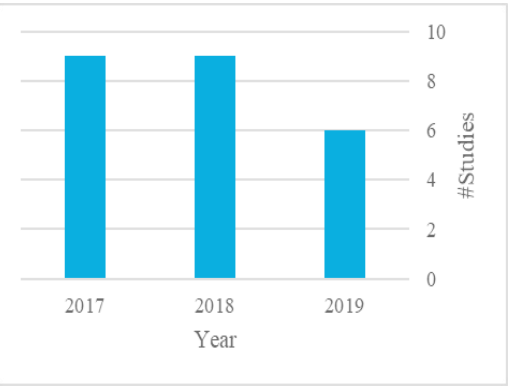

Figure 2: Count of the selected papers until search date in beginning of 2019 for answering research questions RQ1 to RQ3

4. Results for RQ1 and RQ2: Reasons for CRISP-DM and applied domains

Figure 3 shows the results of the first research question. Most of the authors defines CRISP-DM as the de-facto standard for applying a process model in data mining projects. Eight authors do not give any reasons why they have chosen CRISP-DM. Two authors have compared different process models before deciding for CRISP-DM (e. g. with SEMMA in [12]). Other authors have described CRISP-DM as an easy and structured, reliable, commonly used and industry-independent process model.

Figure 4 lists different domains in which the studies could be categorized. The classes of the domains have been inferred from the title and abstract of the papers. Most use cases are found in the health and education domain. For example, CRISP-DM is used in cases like cancer diagnostics in combination with classification models [13]. Four papers follow with engineering use cases. One example is data mining in lithium battery production to classify quality of battery dependent on several key factors [14]. Without a significant frequency, CRISP-DM is also used in many further domains (see Figure 4).

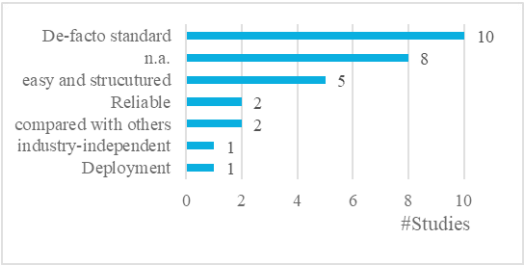

Figure 3: CRISP-DM reasons.

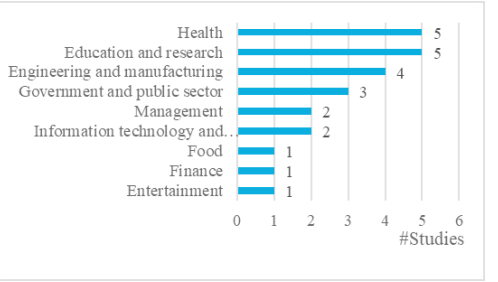

Figure 4: Applied Domains.

5. Results for RQ3: Methods of each phase

The third research question answers how each of the six phases of CRISP-DM is implemented (see Figure 5). Identified methods serves as a class to which the phases can be classified. This helps to identify best practices or new and innovative methods for implementing and conducting CRISP-DM phases.

5.1 Business Understanding and Data understanding

The business understanding phase is described very differently, since CRISP-DM is used in different domains. All of them have a textual description of the business goals and why data mining is useful for the specific use cases. Solely there are differences regarding the structure of the studies since eleven studies do not have a dedicated section for this phase (e. g. the description can be found in the introduction section). Furthermore, the concrete data mining goal is specified in a structured manner in eight studies, e. g. like "[...] In terms of data mining, it is a classification with the binary target attribute." [13]. Furthermore, three studies use a self-developed case study to clarify business goals and to harvest data from their case study itself (e. g. case study battery production in [14]).

The authors of the selected studies use different approaches to explain the data they use in their research. As the CRISP-DM user guide recommends as well, most of the authors entitle the concrete data sources and explain from where the data has been collected during the data understanding phase. The most important method of data understanding is the descriptive statistic. Further methods are data visualization, showing example data rows or interviewing experts. This helps authors and readers to understand the data.

Furthermore, multiple data sources have been used in three papers, e. g. for gaining data completeness: "The data retrieved from the ART database did not include all relevant variables. Hence, we collected data manually from the patient folders." [15] . Four papers also mention the technology they use for this phase and also for the whole CRISP-DM process (Xcyt [13], MySQL Workbench [16], Matlab [17], Weka [18]).

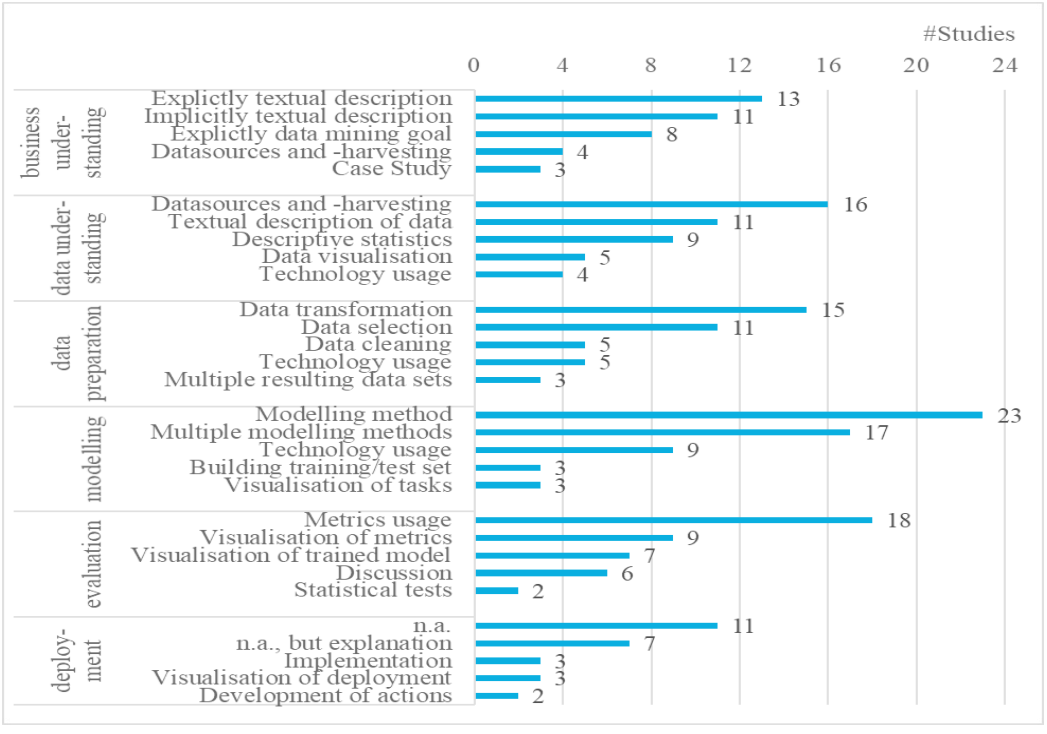

Figure 5: Approaches of each phase of CRISP-DM.

5.2 Data Preparation, Modeling and Evaluation

The typical data preparation tasks such as data selection, transformation and cleaning are also used in the reviewed papers. The data transformation task is described most frequently, because this is one of the most complex tasks in data preparation. Data selection and data cleaning tasks follow. Authors of five studies noun the technologies explicitly (RapidMiner [14], R [19], MS Access and WEKA [15], Microsoft SQL Server and RapidMiner [20]). Three studies also prepare their data with different methods as it results in multiple data sets.

Except of one conceptual paper [21], the other papers explain the models for the specific use cases. Training, testing and analyzing more than one model and algorithm is are useful to compare results of different models during evaluation. Authors of nine studies entitle the used technology in this phase (WEKA [21–25], Spark [26], Orange Canvas and RapidMiner [16]).

In the evaluation phase, most of the papers use metrics to evaluate the quality of trained models. Furthermore, metrics are also visualized to illustrate the results, e. g. in a confusion matrix [27]. The visualization of the trained models means, that e. g. a decision tree are shown to explain and evaluate the models (e. g. in [27,28]). Furthermore, a discussion and statistical tests can strengthen the results. Two papers do not conduct the evaluation phase as the paper [29] and [30] are more conceptually and not technically.

5.3 Deployment

17 studies do not perform a deployment phase, although this is part of the process model. Seven studies give an explanation why they do not deploy the models, for example in [21]: "In this particular case, the number of false positives must be minimized as much as possible, even though the obtained results were satisfactory, before implementing a decision-support system [...]". Accordingly, bad performing models should not be deployed. In [31], the reason for the lack of deployment is that the model will be deployed in the future.

Only three papers implement new architectures or the models in an existing system. Two studies describe the deployment phase less technically and develops concrete actions based on the results of the data mining models.

6. Interpretation and limitations of the results

6.1 Understandable process model and potentials for adoption

The objective of this systematic literature review is to identify the research focus, best practices and new methods for applying CRISP-DM phases. The results show that several data mining projects since 2017 base on CRISP-DM. It is seen as the de-facto standard in data mining projects, as the results of RQ1 show. This finding coincides also with references in other research disciplines close to data mining, like in big data research. CRISP-DM is still a suitable process model [32].

The recommendations from the CRISP-DM user guide have mostly been used in the phases from business understanding to evaluation. There are differences in structure and whether or how specific tasks have been described. For example, not all studies have described all data preparation tasks like data selection, transformation and cleaning. Depending on the specific data mining goals, depending on the data, and depending on the domain, different modelling methods have been evaluated. During the evaluation phase, metrics have been defined and used to compare or to analyze the trained models.

For presentation of the data mining results, the structure can follow the CRISP-DM process phases. Table 2 shows how such structure could look like and could be used as a template. It contains in the rows each phase of the CRISP- DM process model and in the second column the methods and approaches in descent order as its mentioned in the research studies. It is worth to mention that the template is an addition to the user guide of CRISP-DM [10] and is derived from papers of this systematic literature review.

Table 2: Template for presenting and structuring papers following CRISP-DM.

One section per CRISP-DM phase	Methods and approaches as subsections
1. Business understanding	1.1 Textual description in own section
	1.2 Defining the data mining goal explicitly
2. Data understanding	2.1 Mention of the data source and harvesting process
	2.2. Structural description (data model, example data)
	2.3 Descriptive statistics obligatory
3. Data preparation	3.1 Describing input and output data
	3.2 Methods and approaches (transformation, selection, cleaning)
4. Modeling	4.1 Mention of the modeling approach
	4.2 At least the used technology should be mentioned here
	4.3 Building test and training sets
5. Evaluation	5.1 Defining metrics
	5.2 Visualization of model and metrics
6. Deployment	6.1 If deployment in the scope, the implementations should be described

6.2 Enhancements of CRISP-DM

As we have already mentioned in the introduction, CRISP-DM is one of many process methodologies. Furthermore, there are also enhancements of CRISP-DM due to some drawbacks. For example, the APREP-DM (a Framework for Automating the Pre-Processing of a Sensor Data Analysis) enhances the CRISP-DM to specific pre-processing tasks for sensor analysis since the CRISP-DM does not describe sensor data explicitly [33]. The authors of the Lean Design Thinking Methodology for Machine Learning and Modern Data Projects (LDTM) assumes, that CRISP-DM is limited in managing the requirements of current technologies like machine learning algorithms. Therefore, they integrate design thinking approaches in CRISP-DM. LDTM is domain-independent but focuses on new technologies [34].

Such enhancements could be focused and evaluated in further research. However, there is not a significant number of studies using the enhancements until now. The focus on CRISP-DM for this systematic literature review was therefore a suitable way for researching process methodologies in data mining.

6.3 Technology heterogeneity

CRISP-DM is an organizational process model and not restricted to any technology. Therefore, several technologies can support the process. Figure 5 shows that the used technologies have been mentioned in several studies. For storage, used technologies are MS Access, MySQL or Microsoft SQL Server. For data preparation, used technologies are Xcyt, Matlab RapidMiner, WEKA, and R. For modelling, WEKA is mainly used in data mining projects. We can assume that WEKA is a popular tool in data mining since it is a free software licensed under the GNU General Public License. Furthermore, it is still maintained (last version was released in the end of 2019) and is developed by a research institution, the university of Waikato.

6.4 Challenges for deployment

It is worth noting that the deployment phase is missing in the majority of the studies. This is an interesting insight because CRISP-DM defines a deployment phase explicitly. The deployment phase is understood by one study as the development of concrete actions based on the trained model. This is an innovative understanding of the deployment phase due to the user guide defines the deployment phase more technically. Technical solutions have been developed by three studies (see Figure 5). In all other cases, it is not comprehensible why the deployment phase is missing. Therefore, we infer two implications and possible explanations. First, the focus in the reviewed papers is on building and evaluating models only, but not on deployment. Second, there are missing guidelines in how to conduct deployment in data mining projects.

This could indicate a lack of possibilities for integrating models and data into a productive environment. Therefore, CRISP-DM does not adequately cover the deployment of the analysis in the productive environment, where model performance must be continuously monitored and controlled. Therefore, CRISP-DM does not cover the whole project lifecycle including machine learning approaches. However, there are existing approaches of deployment of data mining models, but they are not integrated in CRISP-DM yet. Furthermore, modern approaches also deals with the automation and extraction of services automatically. These services could be deployed, so the deployment phase is integrated directly [35]. Such approaches could reduce time-to-science or time-to-research.

6.5 Limitations of the systematic literature review

This systematic literature review focuses on research studies only. Therefore, the results could be limited for potential industry adoption. Furthermore, only the data that is published could be reviewed. That means, that potential further tasks in data mining that are not made public available in the studies could not be reviewed. This paper can identify the tasks that are described in the publications. We cannot infer that other studies have not implicitly performed the mentioned tasks or that the original data are already well-prepared. However, it is crucial to describe all preparatory data mining tasks because the results will be reliable and traceable. The short time period of three years could also be a limitation as the number of papers are insufficient. However, more recent papers could strengthen the results.

7. Conclusion

This paper explores CRISP-DM phases in recent studies. CRISP-DM is a de-facto standard process model in data mining projects. This systematic literature review is used to give an overview of how CRISP-DM is used in recent studies and to find research focus, best practices and innovative methods. A total of 24 studies were selected as relevant after conducting the search strategy and inclusion and exclusion criteria. We have analyzed reasons for the selection of CRISP-DM, domains, each phase (business understanding, data understanding, data preparation, modelling, evaluation and deployment). Subsequently, inductive categories were formed from the marked text passages in order to analyze them descriptively and to develop a classification framework. In a further investigation, potential limitations and enhancements of CRISP-DM should be reviewed and potential industrial adoption should be researched. Also, the overview of used technologies could be practical to support scientists for technology selection. One result is that the deployment phase is mostly missing. Future research should explore suitable ways to integrate models into a productive environment.

References

[1] Krcmar, Helmut. 2015. *Informationsmanagement*. 6th ed. Wiesbaden: Springer Gabler.

[2] Laudon, Kenneth C., Jane Price Laudon, and Detlef Schoder. 2010. *Wirtschaftsinformatik. Eine Einführung*. 2nd ed. München: Pearson Deutschland. (IT).

[3] Saltz, Jeff, Nicholas Hotz, David Wild, and Kyle Stirling. "Exploring Project Management Methodologies Used Within Data Science Teams Orleans, LA, USA, August 16-18, 2018." *24th Americas Conference on Information Systems, AMCIS 2018, New Orleans, LA, USA, August16-18, 2018*: Association for Information Systems.

[4] Cato, Patrick. 2016. Einflüsse auf den Implementierungserfolg von Big Data Systemen. Ergebnisse einer inhalts- und kausalanalytischen Untersuchung. Dissertation. Friedrich-Alexander-Universität Erlangen-Nürnberg.

[5] Wirth, Rüdiger and Jochen Hipp. 2000. "CRISP-DM: Towards a Standard Process Model for Data Mining." *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (4), pp. 29–39.

[6] Long, Carol and Kelly Talbot. eds. 2015. *Data science & big data analytics. Discovering, analyzing, visualizing and presenting data*.Indianapolis, IN: John Wiley & Sons.

[7] Grady, Nancy W., Jason A. Payne, and Huntley Parker. 2017. "Agile Big Data Analytics. AnalyticsOps for Data Science." *IEEE International Conference on Big Data (BIGDATA)* (17), 2331–2339.

[8] Saltz, Jeffrey S. and Ivan Shamshurin. 2016. "Big data team process methodologies: A literature review and the identification of key factors for a project's success." *IEEE International Conference on Big Data (Big Data)*, pp. 2872–2879.