POST-DS: A Methodology to Boost Data Science

Carlos J. Costa
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal
ISEG, Universidade de Lisboa, Lisboa, Portugal
carlos.costa@acm.org

João Tiago Aparicio
Instituto Superior Técnico, Universidade de Lisboa
Lisboa, Portugal
joao.aparicio@tecnico.ulisboa.pt

Abstract — **As the importance of data science is increasing, the number of projects involving data science and machine learning is rising either in quantity or in complexity. It is essential to employ a methodology that may contribute to the improvement of the outputs. In this context, it is crucial to identify possible approaches. And an overview of the evolution of data mining process models and methodologies is given for context. And the analysis showed that the methodologies covered were not complete. So, a new approach is proposed to tackle this problem. POST-DS (Process Organization and Scheduling electing Tools for Data Science) is a process-oriented methodology to assist the management of data science projects. This approach is not supported only in the process but also in the organization scheduling and tool selection.**

Keywords – data science methodology; crisp-dm; data science process; machine learning; data science; data mining

I. INTRODUCTION

Data Science is in the interception of Basic fields: Computer Science and IT (CS), Domain/Business Knowledge (DK), and Mathematics and Statistics (MS) [1]. It includes techniques developed in some traditional fields like artificial intelligence, statistics or machine learning, data science. And so, it is essential to employ a methodology that may contribute to the improvement of the knowledge creation outputs. It is in this context crucial to identify possible approaches. By analyzing the existing methodologies, it is essential to enlarge the scope of knowledge discovery or data mining. Almost all the approaches emphasize the process. The following sections describe KDD (Knowledge discovery in databases), CRISP-DM (CRoss-Industry Standard Process for Data Mining), SEMMA (Sampling, Exploring, Modifying, Modelling, and Assessing), ASUM (Analytics Solutions Unified Method) and TDSP (Team Data Science Process). Then a new approach POST-DS (Process Organization and Scheduling electing Tools for Data Science) is presented. The next section illustrates this methodology usage.

II. LITERATURE REVIEW

A. KDD (Knowledge discovery in databases)

KDD (Knowledge discovery in databases) describes the overall process of discovering useful knowledge from data. Data mining is the usage of specific algorithms for extracting patterns from data. So, data mining refers to an appropriate step in this process. KDD is the non-trivial process of finding valid, new, possibly useful, and ultimately understandable patterns in data. The KDD process is iterative and interactive, involving numerous steps with many decisions made by the analyst. It is essential developing an understanding of the data, creating a target data set, cleaning and preprocessing. Then, several tasks must be performed, like data reduction and projection. The analyst also has to match the goals of the KDD process to a particular data-mining method, exploratory analysis and model and hypothesis selection. An essential task is interpreting mined patterns and using the knowledge directly. [21]

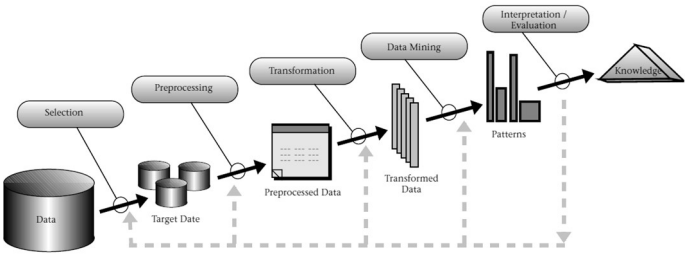

Figure 1. KDD (Knowledge Discovery in Databases) [21].

The first step is developing an initial interpretation in the context of the domain. It consists of the relevant prior knowledge and the goal identification of the KDD process from the customers' perspective.

The second step consists of creating a target data set by selecting a dataset or focusing on a subset of variables.

The third step consists of data cleaning and preprocessing. Basic operations may include removing noise, collecting the necessary information to model or account for noise, deciding on strategies for handling missing values, and accounting for time-sequence information and known changes.

The fourth step involves data reduction and projection. It consists of finding useful features to represent the data. The adequate number of variables can be obtained, with dimensional reduction or transformation methods. Invariant representations for the data may also be identified.

The fifth step consists of matching the goals of the KDD process to a particular data-mining method. Those methods may be summarization, classification, regression or clustering.

The sixth step consists of exploratory analysis and model and hypothesis selection. It involves choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process comprises deciding which models and parameters might be appropriate.

The seventh step consists of data mining. It includes searching for patterns of interest in a particular representational form or a set of such representations. Examples are classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.

The eighth step consists of interpreting mined patterns. In this step, it is possible returning to any of steps 1 through 7 for additional iteration. This step can also include visualization of the extracted patterns and models or display of the data given the obtained models.

The ninth step consists of acting on the discovered knowledge. This step consists of using the knowledge directly, incorporating the knowledge into another system for further action. Alternatively, results may be documenting and reporting to interested parties. This process also contains checking for and resolving potential conflicts with previously believed (or extracted) knowledge. As referred by several authors (e.g. [6]), the primary methodologies are inspired by KDD process as well as CRISP-DM.

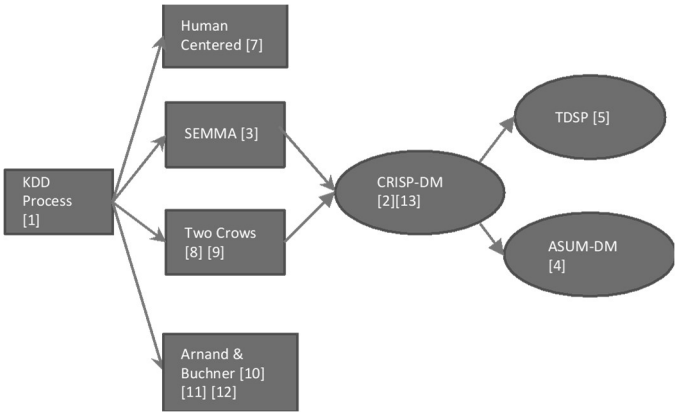

Figure 2. Evolution of data mining process models and methodologies.

B. CRISP-DM (CRoss-Industry Standard Process for Data Mining)

In 1996, four leaders of the nascent data mining market (Daimler-Benz, IntegralSolutions Ltd. (ISL), NCR, and OHRA) created CRISP-DM (CRoss-Industry Standard Process for Data Mining). CRISP-DM is a complete data mining methodology and process model that provides anyone with a comprehensive blueprint for performing a data mining project. The CRISP-DM life cycle has six phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment.[2]

Business understanding is the initial phase. It focuses on understanding the project objectives and requirements from a business perspective. The knowledge is converted into a data mining problem definition. It is also in this phase that the preliminary plan is designed to achieve the objectives. The data understanding phase starts with the initial data collection. Then, it proceeds with activities in order to get familiar with the data: identifying data quality problems; discovering first insights into the data or detecting interesting subsets to form hypotheses for hidden information. The data preparation phase covers all activities until the construction of the final dataset. At the end of this phase, the data that is used by the modelling tool(s). Data preparation tasks are likely to be done multiple times and often not in any fixed order. This task includes table, record and attribute selection as well as transformation and cleaning of data for modelling tools. In the modelling phase, various modelling techniques are selected and applied. And their parameters are calibrated to optimal values. Usually, there are several techniques for the same data mining problem type. Various techniques have specific requirements on the form of data. Thus, moving back to the data preparation phase is often necessary. In the evaluation stage, the built project has a model (or models) that appears to have high quality. Before continuing to the final deployment of the model, it is essential to more systematically assess the model and analyze the steps executed to build the model to be sure it appropriately reaches the business objectives. A crucial objective is to determine if there is some critical business issue that has not been sufficiently considered. A choice on the use of data mining results should be achieved at the end of this phase. Creation of the model is not typically the end of the project. Yet if the aim of the model is to increase knowledge of the data, the knowledge gained need to be organized and displayed in a way that the client can use it. However, conditional to the requirements, the deployment phase can be simple or complex. It may be just generating a report or an implementing of repeatable data mining process across the enterprise. Often, it is the client or another entity, not the data analyst, who carries out the deployment steps. However, even if the analyst does not carry out the deployment effort, the customer needs to understand upfront what actions need to be carried out in order actually to make use of the created models.

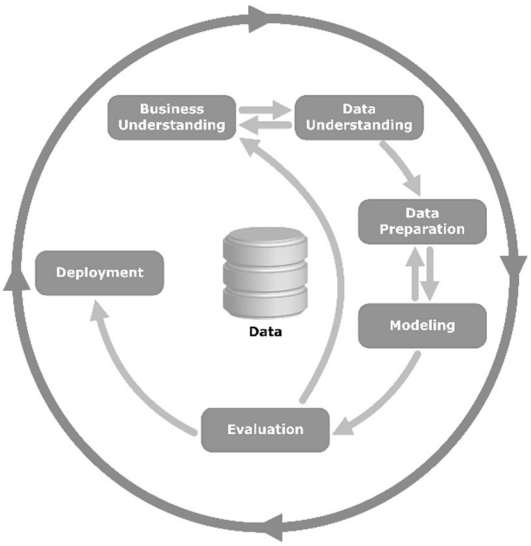

Figure 3. CRISP-DM (CRoss-Industry Standard Process for Data Mining) [2].

Business Understanding focuses on identifying the project objectives and requirements from a business perspective. Then this knowledge is converted into a data mining problem definition and a preliminary plan.

Data understanding begins with an initial data collection. Then, it proceeds with actions to get familiar with the data: identifying data quality problems, discovering first insights into

the data, or to detecting subsets to form hypotheses for non-trivial information.

The data preparation phase includes all activities to construct the final dataset from the initial raw data.

Then, modelling techniques are chosen and applied. As may be analyzed in several examples, e.g. [22], the selection of techniques is an iterative process. Since some techniques have specific requirements regarding the form of the data, there can be a loop-back here to data preparation.

Evaluation is the next phase. Based on any loss functions selected, these models need to be tested to ensure they generalize against unseen data. All critical business issues must also be considered. The result is the selection of the winner model.

Deployment means adopting a code representation of the model into a system. It may be used to score or categorize new unseen data. It must correspond further information in the solution of the original business problem. On the other hand, using presentation techniques is critical in this phase [15][25].

TABLE 1. CRISP-DM (CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING)

Business Understanding	**Data Understanding**	**Data Preparation**	**Modeling**	**Evaluation**	**Deployment**
Determine Business Objectives	Collect Initial Data	Select Data	Select Modeling Techniques	Evaluate Results	Plan Deployment
Assess the Situation	Describe Data	Clean Data	Generate Test Design	Review Process	Plan Monitoring and Maintenance
Determine Data Mining Goals	Explore Data	Construct Data	Build Model	Determine Next Steps	Produce Final Report
Produce Project Plan	Verify Data Quality	Integrate Data	Assess Model		Review Project
		Format Data			

C. *SEMMA (sampling, exploring, modifying, modelling, and assessing)*

The SAS Institute proposed a process of sampling, exploring, modifying, modelling, and assessing large volumes of data to discover previously unknown patterns. This process is called SEMMA, which is the acronym of sampling, exploring, modifying, modelling, and assessing. And it can be applied to business advantage.

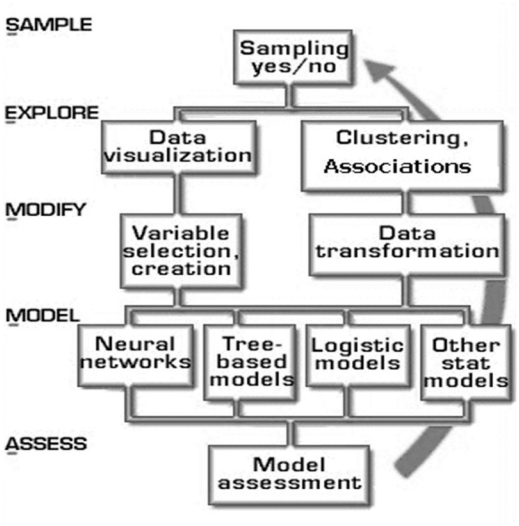

Figure 4. SEMMA [3].

The SEMMA data mining process is appropriate for a variety of industries. It also provides methodologies for such diverse business problems as customer retention and attrition, household, risk analysis, fraud detection, database marketing, market segmentation, affinity analysis, bankruptcy prediction, customer satisfaction, and portfolio analysis [3]

D. ASUM (Analytics Solutions Unified Method)

Analytics Solutions Unified Method (ASUM) is an iterative IBM SPSS Process to implement a Data Mining/Predictive Analytics project. It is based on an extended and refined CRISP-DM methodology. ASUM-DM has five phases: (1) Analyze, (2) Design, (3) Configure & Build, (4) Deploy, and (5) Operate & Optimize. Nevertheless, the first three steps of ASUM (Analyze, Design, and Configure and Build) may be combined because data mining/predictive analytics projects are iterative by nature. It is also possible to add an optional Project Management Process. [4]

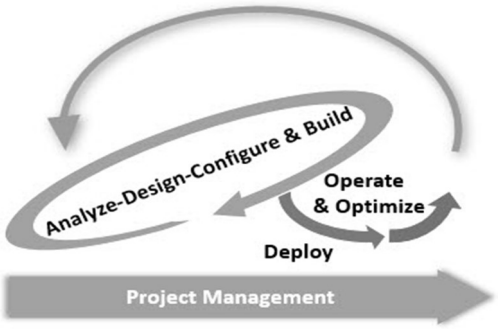

Figure 5. Team Data Science Process (TDSP) [5].

E. *TDSP (Team Data Science Process)*

In order to deliver predictive analytics solutions and intelligent applications efficiently, Microsoft proposed the Team Data Science Process (TDSP). It is an agile, iterative data science methodology. TDSP also suggests how team roles work best together. The purpose is to improve team collaboration and learning. [5]

TDSP includes best practices from Microsoft and other industry players to help toward successful implementation of data science projects. The purpose is helping companies realize the benefits of their analytics program.

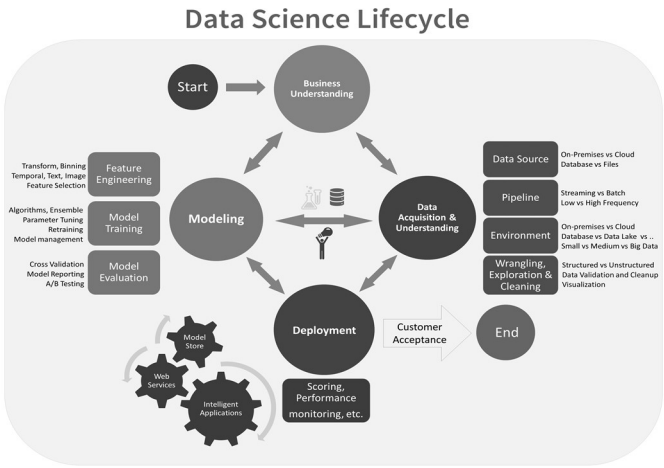

Figure 6. Team Data Science Process (TDSP) [5].

TDSP includes a suggested life-cycle that may be used to structure your data science projects. The life-cycle outlines the steps that projects typically follow, from beginning to end, when they are performed. It is possible to adopt another data science life-cycle, such as the CRISP-DM, KDD or organization's own custom process.

F. Scheduling, Organizations and Tools

In the approaches presented, the emphasis is in the process. Like any project, times are critical. And monitorization must be performed to conclude if the project went successfully. [25] But it is also essential to identify the possible roles involved, and how do those roles may be organized. In this context, using techniques like RACI (Responsible, Accountable, Consulted, Informed) may contribute to improving project organization. [17]

The way how information is analyzed may also help to in the selection of the best way to success. This is why a criterion selection of tools should be followed. Many approaches may be used to select the most appropriate tools [19]. The following figure presents a flowchart developed by a group of researchers from MIT [14] representing main decisions involved in the selection of the tools. This chart is at a high level, but it is essential in the decision processes of choosing Machine Learning tools. On the other hand, other techniques may be used in different phases. For example, in the first phase, it is vital to align corporate strategy and mission with the project mission. Some authors even suggest the use of data knowledge to select appropriate models in the context of the managerial decision. [20]

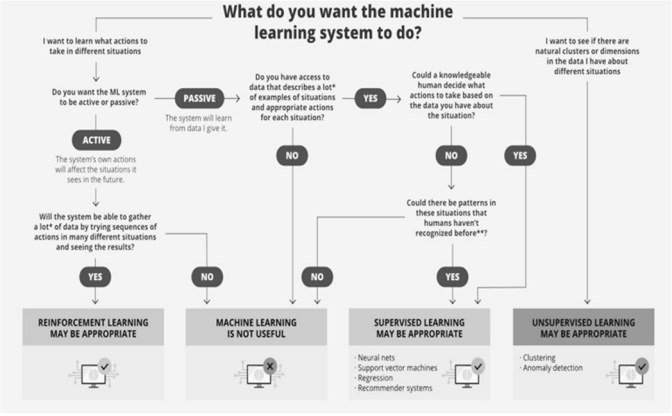

Figure 7. What do you want the machine learning system to do? [14].

Several techniques may also be used in order to select the most appropriate charts in order to create the best interfaces [15][25]

Summarizing, the approaches presented for data mining, machine learning and data science may be interrelated. CRISP-DM is one of the most used and also the one that inspired many of the major approaches. Nevertheless, other features may be added to this approach.

III. PROPOSING A DATA SCIENCE MODEL

Following the literature review, it is possible to identify process, organization, scheduling and tools as the four main blocks to tackle a problem of data science. The following figure represents the proposed model graphically.

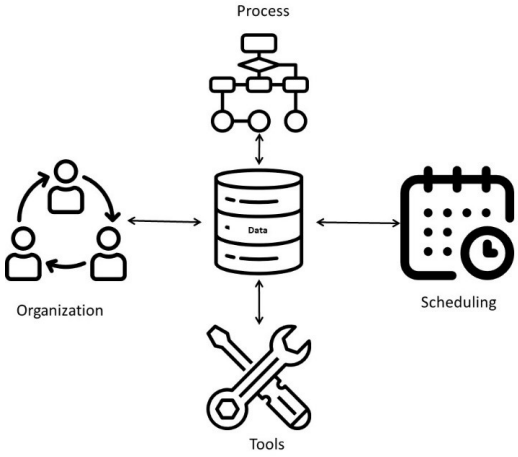

Figure 8. POST-DS Methodology

The POST-DS (Process Organization and Scheduling electing Tools for Data Science) describes the sequence of activities performed in a data science project. The list of the possible task is very well documented in CRISP-DM as it was described previously.

In project management, scheduling is an essential element. Each activity must be performed at a specific time. Often, if a project does not conclude in the expected, it is entirely useless.

It is normal to have different roles in a data science project. The data engineer, data scientist, business analyst or computer

engineer are just possible roles in a data science project. Several people may perform these roles. It is essential to identify what are the roles in a specific project. And how they participate in each phase or task of a project. Using a responsibility matrix, RACI is a possible solution. All those building blocks interact to extract knowledge from data.

A template may be used to support the implementation of this methodology. This template is represented in figure 9.

Activities	Roles	Harmonogram	Tools
(e.g CRISP-DM)	(e.g. RACI)	(e.g. GANTT)	

Figure 9. POST-DS Template.

IV. USING THE MODEL

To assess the feasibility of the proposed approach, it was implemented in the context of a real situation. The purpose was to identify the main characteristics that contributed to the improvement of Instagram presence. In the first phase, a business analyst defined the business objectives, assessed the situation, determined the data science goals, defining the scope of the project, time and budget. Then, data engineer was responsible for data collection and quality assurance of this data. Data preparation and modeling was the responsibility of the data scientist. The business analyst evaluated the models with the support of the data scientist. Designer under the supervision of the business analyst deployed the model, incorporating its results in a website. Figure 10 shows a preliminary version of the implementation of the proposed approach. Some of the activities changed as well as the timings. And some of the roles also regrouped according to the skills of the persons involved in the project. The usage of acronyms is the following for roles: BA – Business Analyst, DE – Data Engineer, DS – Data Scientist, WD-Web Designer; And for RACI: A-Accountable, I-Informed, R-Responsible.

The management of the data science project, in this case, gains from the assistance of such a methodology. As long as it allows integrating the specified components. Because it allowed the adjustment of expectations, clarifying the scope of the project, costs and time. It also makes clear the tasks assigned to each person.

	Activities	BA	DE	DS	WD	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	Tools
1	**Business Understanding**																			
1.1.	Determine Business Objectives	A/R																		meeting
1.2.	Assess the Situation	A/R																		meeting
1.3.	Determine Data Science Goals	A/R																		meeting
1.4.	Produce Project Plan	A/R	R	R																WBS, GANTT
2	**Data Understanding**																			
2.1.	Collect Initial Data	C	A/R	I																open data, scraping,
2.2.	Describe Data	C	A/R	I																Jupyter/python
2.3.	Explore Data	C	A/R	I																Jupyter/python
2.4.	Verify Data Quality	C	C	A/R																Jupyter/python
3	**Data Preparation**			A/R																
3.1.	Select Data	I		A/R																Meeting
3.2.	Clean Data	I		A/R																Jupyter/python
3.3.	Construct Data	I		A/R																Jupyter/python
3.4.	Integrate Data	I		A/R																Jupyter/python
3.4.	Format Data	I		A/R																Jupyter/python
4	**Modelling**																			
4.1.	Select Modeling Techniques	I		A/R																MIT flowchart
4.2.	Generate Test Design	I		A/R																Jupyter/python
4.3.	Build Model	I		A/R																Jupyter/python
4.4.	Assess Model	I		A/R																Jupyter/python
5	**Evaluation**																			
5.1.	Evaluate Results	A/R		R																Jupyter/python
5.2.	Review Process	A/R																		meeting
5.3.	Determine Next Steps	A/R																		meeting
6	**Deployment**																			
6.1.	Plan Deployment	A		R	R															Wordpress or Flask
6.2.	Plan Monitoring and Maintenance	A																		meeting
6.3.	Produce Final Report	A/R	R	R	R															meeting
6.4.	Review Project	A/R		R																meeting

Figure 10. POST-DS Example usage.

I. CONCLUSIONS

An overview of the evolution of data mining process models and methodologies are studied. Given that the methodologies analyzed were not complete, a new approach (POST-DS) was proposed. This approach allows for the identification of processes, organization, scheduling and tools. This approach is inspired particularly in the Cross-Industry Standard Process for Data Mining, but it intends to give additional guidelines. This methodology was applied in a specific data science project. The application made it possible to conclude that this POST-DS can contribute to a better alignment of overall project management.

ACKNOWLEDGMENT

We gratefully acknowledge financial support from FCT-Fundação para a Ciencia e Tecnologia (Portugal), national funding through research grants FCT UIDB/04466/2020 and UIDP/04466/2020.

REFERENCES

[1] S. Aparicio, J. Aparício and C Carlos. Data Science and AI: Trends Analysis. 1-6. 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), 2019 doi 10.23919/CISTI.2019.8760820

[2] C. Shearer. The CRISP-DM Model: the New Blueprint for Data Mining, Journal of Data Warehousing, Volume 5, Number 4, page. 13-22. 2000

[3] SAS Institute Inc. SAS® Enterprise Miner™ 14.3: Reference Help. Cary, NC: SAS Institute Inc. 2017

[4] IBM Analytics solutions unified method-Implementations with Agile principles. 2016. Retrieved Dez, 15, 2019 ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf

[5] G. Ericson, Z. Martens, K Sharkey, CJ Gronlund, Team Data Science Process Documentation. Retrieved January 12, 2020, from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/

[6] G. Mariscal, O. Marban, & C. Fernandez, A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2). P. 137-166. 2010.

[7] R. J. Brachman, T. Anand, The process of knowledge discovery in databases. Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, 37–57. 1996.

[8] Two Crows Corporation Introduction to Data Mining and Knowledge Discovery, 2nd edition. Two Crows Corporation. 1998. ISBN 892095-00-0.

[9] Two Crows Corporation Introduction to Data Mining and Knowledge Discovery, 3rd edition. Two Crows Corporation. 1999. ISBN 1-892095-02-5.

[10] S. Anand, & A Buchner Decision Support Using Data Mining. Financial Times Management, 184, 1998 .

[11] S. Anand, A. Patrick, J. Hughes, D. Bell, A data mining methodology for cross sales. Knowledge-based System Journal 10(7), . 449–461. 1998.

[12] A. G., Buchner, M. D Mulvenna, S. S. Anand, J. G Hughes. An Internet-enabled Knowledge Discovery Process, p 13–27. 1999, citeseer.ist.psu.edu/290505.html

[13] P. Chapman et al., "CRISP-DM 1.0: Step-by-step data mining guide," SPSS inc, vol. 9, p. 13, 2000.

[14] MIT What do you want the machine learning system to do? Intractive infographic, 2019 site: https://s3-us-west-2.amazonaws.com/getsmartergraphics/Courses/MIT+ML/M2/MIT+ML+M2+interactive+infographic.html

[15] C. J. Costa and M. Aparício "Supporting the decision on dashboard design charts". In Proceedings of 254th The IIER International Conference 2019 (pp. 10-15).

[16] K. Adamiecki, "Harmonograf". Przegląd Organizacji, 1931. (translation to english)

[17] Ajit Tewari, Shubha Mishra, Shadab Siddiqui, Priyanka Upadhyay, "Performance measurement at the requirement phase of software development life cycle", Computing for Sustainable Global Development (INDIACom) 2015 2nd International Conference on, pp. 1090-1094, 2015.

[18] J. Cabanis-Brewin, J. S. Pennypacker, "Aligning projects to corporate strategy - Strategic Performance." Paper presented at PMI® Global Congress 2006—North America, Seattle, WA. Newtown Square, PA: Project Management Institute.

[19] S. Bibi and I. Stamelos, "Selecting the appropriate machine learning techniques for the prediction of software development costs," in IFIP International Conference on Artificial Intelligence Applications and Innovations, 2006, pp. 533–540.

[20] S. Banerjee and A. Basu, "A knowledge based framework for selecting management science models," in Twenty-Third Annual Hawaii International Conference on System Sciences, 1990, vol. 3, pp. 484–493 vol.3, doi: 10.1109/HICSS.1990.205381.

[21] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, no. 3, pp. 37–37, 1996.

[22] N. Fernandes, S. Moro, C. Costa and M. Aparício, M "Factors influencing charter flight departure delay". Research in Transportation Business & Management, 100413. https://doi.org/10.1016/j.rtbm.2019.100413

[23] C. J. Costa and M. Aparicio, "Analysis of e-learning processes," in Proceedings of the 2011 Workshop on Open Source and Design of Communication, New York, NY, USA, 2011, pp. 37–40, doi: 10.1145/2016716.2016726.

[24] H. Fayol, Administration industrielle et générale, Dunod. 1916,

[25] M. Aparicio and C. J. Costa. "Data visualization." Communication design quarterly review 3.1 (2015): 7-11 .https://dl.acm.org/doi/pdf/10.1145/2721882.2721883

.