Journal of Information Systems and Technology Management – Jistem USP
Vol. 19, 2022, e202219018
ISSN online: 1807-1775
DOI: 10.4301/S1807-1775202219018 Original Article

EVALUATION OF A PROCESS FOR THE EXPERIMENTAL DEVELOPMENT OF DATA MINING, AI AND DATA SCIENCE APPLICATIONS ALIGNED WITH THE STRATEGIC PLANNING

Methanias Colaço Júnior[1,2], https://orcid.org/0000-0002-4811-1477
Rodrigo Fontes Cruz[2], https://orcid.org/0000-0001-6783-9263
Luciano Vieira de Araújo[1], https://orcid.org/0000-0002-9687-5367
Ana Carla Bliacheriene[1], https://orcid.org/0000-0002-5705-3950
Fátima de L. S. Nunes[1], https://orcid.org/0000-0003-0040-0752

[1]Universidade de São Paulo, São Paulo, SP, Brazil

[2]Universidade Federal de Sergipe, Aracajú, SE, Brazil

ABSTRACT:

The Big Data phenomenon has imposed maturity on companies regarding the exploration of their data, as a prerogative to obtain valuable insights into their clients and the power of analysis to guide decision-making processes. Therefore, a general approach that describes how to extract knowledge for the execution of the business strategy needs to be established. The purpose of this research paper is to introduce and evaluate the implementation of a process for the experimental development of Data Mining (DM), AI and Data Science applications aligned with the strategic planning. A case study with the proposed process was conducted in a federal educational institution. The results generated evidence showing that it is possible to integrate a strategic alignment approach, an experimental method, and a methodology for the development of DM applications. Data Mining (DM) and Data Science (DS) applications also present the risks of other Information Systems, and the adoption of strategy-driven and scientific method processes are critical success factors. Moreover, it was possible to conclude that the application of the scientific method was facilitated, besides being an important tool to ensure the quality, reproducibility and transparency of intelligent applications. In conclusion, the process needs to be mapped to foment and guide the strategic alignment.

Keywords: Big Data, Strategic Alignment, Experimentation, Small Data, Reproducibility

Manuscript first received: 2022-05-18. Manuscript accepted: 2022-09-20

Address for correspondence:
Methanias Colaço Júnior, Universidade de São Paulo – EACH – USP, São Paulo, SP, Brazil;
Universidade Federal de Sergipe Aracajú, SE,Brazil. E-mail: mjrse@hotmail.com
Rodrigo Fontes Cruz, Universidade Federal de Sergipe, Sergipe, Brazil.
E-mail: rodrifontes@gmail.com
Luciano Vieira de Araújo, Universidade de São Paulo – EACH – USP, São Paulo, SP, Brazil.
E-mail: lvaraujo@usp.br
Ana Carla Bliacheriene, Universidade de São Paulo – EACH – USP, São Paulo, SP, Brazil.
E-mail: acb@usp.br
Fátima de L. S. Nunes,Universidade de São Paulo – EACH – USP, São Paulo, SP, Brazil.
E-mail: fatima.nunes@usp.br

CC BY

INTRODUCTION

In competitive markets, each detail is important at the time of defining the profit and loss of business organizations, in which making the most accurate decisions has become fundamental for their survival (Botelho & Filho, 2014). In this context, data appear as an important source of obtaining competitive advantage (Kubina; Varmus & Kubinova, 2015), which is based on the knowledge gained from the analysis of such data, and has improved areas like Business Intelligence (BI), Data Mining (DM), and Data Science, being the latter a more recent jargon that tries to incorporate the two first ones, and scientific validation to applications as well.

However, before any attempt to perform the extraction of this useful knowledge can be made, a general approach that describes how to extract knowledge needs to be established (Kurgan & Musilek, 2006). Thus, several models of Data Mining processes were proposed by researchers and professionals. The examples include Fayyad, et al. (1996), Cabena et al. (1998), Cios et al. (2000), CRISP-DM (2003), Berry & Linoff (1997), Sharma, Osei-Bryson & Kasper (2012) and Ławrynowicz & Potoniec (2014).

A part of these methods was proposed in a different context than the one we live in today, in which data analysis is very different from the one previously practiced (Sedkaoui, 2018). CRISP-DM for example, which is considered the most widely used methodology according to many opinion surveys, was originated in the second half of the 1990´s, thus being about two decades old (Sharma, Osei-Bryson & Kasper, 2012; Schäfer et al., 2018; Martínez-Plumed et al., 2019). Therefore, although these changes haven´t occurred fast, and new technologies have been proposed to accommodate some of the changes, the essence of the methodologies does not fully encompass the diversity of Data Science projects (Martínez-Plumed et al., 2019), and does not explicitly guide the strategic alignment. As examples of these technologies, IBM introduced ASUM-DM (IBM, 2005), and SAS introduced SEMMA (SAS, 2005).

The necessity for data scientific analysis comes from the Big Data phenomenon. Big Data can be conceptualized as a Business Intelligence approach characterized by 9 Vs: Volume, Velocity, Variety, Viscosity, Virality, Visualization, Veracity, Validity and Value. In other words, the term entitles a set of technologies to explore the big *volume* of data produced on a yearly basis, in a structured and unstructured manner, by extracting *valuable* patterns for the decision-making process.

From the viewpoint of Experimentation, it is important to analyze the threats to the *Validity* of the results, by considering graphs and *Visualizations* that are balanced, simple and inclusive, and that trigger a decision with only a quick look. Besides, the possibilities of forwarding and easily publishing the results on social medias and software to create presentations, it means, *Virality*, is fundamental to a good strategy. Finally, the messages must be clear and objective, just like the refrain of a hit song; they stick to the mind and call to action, *Viscosity*.

For this context, in a literature review performed by Cruz; Colaço Júnior & Gois (2022), it was evidenced that there was not a heterogeneous approach that could comprehend the development of Business Intelligence and Data Mining applications aimed for the Strategy, and/or that could predict the Experimental evaluation. In other words, the exposition of methodological evidence is limited to the scarcity of publications, the ones that do not exist or that are limited by unpublished empiricism.

As a limited exception, for example, we can list the process published by Colaço Júnior et al. (2019), which despite the advance in guiding the strategic alignment, it does not comprehend Data Mining and AI solutions yet, and does not predict an experimental evaluation of the knowledge models, which are important complementary and integral parts of this research study.

CC BY

As previously mentioned, this prediction is important, because besides its important strategic matter, standardizing the intelligence projects for the use of an experimental approach is an alternative to meeting the specific presuppositions of Data Science (Bock et al., 2018; Costa et al., 2016; Santos et al., 2018), considering that the application of a rigorous scientific method coadunate the attempt to make data analysis a science, with principles that reduce the threats to the validity of the generated knowledge. In the literature, some projects such as the ones performed by Kohavi et al. (2013) and Costa et al. (2015) already used experimentation as a way to select the solutions with the highest value return to the business organization.

Therefore, the purpose of this research work is to introduce and evaluate the implementation of a process for the experimental development of Data Mining (DM) and Data Science applications aligned with the strategic planning.

For this, a case study of the application of the process was conducted in a federal educational institution, which presented initial results to the awareness of the intelligence team about the need for the transparency of the strategic goals and the creation of applications to reach them. Besides, the qualitative evaluation regarding the facilitated adoption of the scientific method was positive, working as one more support to ensure the quality of the intelligent applications. In relation to the strategic alignment per se, fomentation, emphasis and communication benefit from the mapping of processes that transform the search for strategic goals into explicit requirements.

The used methodological procedures, the related works, the conceptual basis for the understanding of this research work, the process to be evaluated, the evaluation of the proposed process, and the final conclusions will be described in the next sections of this article.

METHODOLOGY

A quasi-systematic literature review, published in (Cruz; Colaço Júnior & Gois, 2022), identified the lack of methods or processes integrated with the development of DM applications that could present any heterogeneous approach to guiding the strategic alignment and experimentation, predicting clear support to the strategic goals, and an experimental phase in the validation of the results. Consequently, in view of the identified gap, a process for the strategic alignment and for the Experimental development of Data Mining applications was developed.

Lastly, a case study with the intelligence area of a federal educational institution was planned and conducted, in order to evaluate the proposed process. According to Yin (2015), a case study is an in-depth empirical investigation of some contemporary phenomenon within its real-life context, especially when the boundaries between the phenomenon and the context are not clearly evident. It refers to a detailed analysis of a specific case, supposing that the knowledge of this phenomenon through a meticulous study of a single case is possible.

In section 6, the case study and its methodology are **detailed in a self-contained** manner.

RELATED WORKS

Scientific research works with the same object of research as the one in this article were not found, considering the use of any approach to guide the strategic alignment towards the experimental development of DM applications, which increases the importance of the results presented here.

CC BY

However, some works mention the importance of this alignment or propose some methodology for the efficient experimental development of this type of application.

In a survey conducted in Brazil (Lima et al., 2017), it was verified that 67,50% of the companies do not use an experimental methodology for the development of BI, which contributes to the lack of success of the projects. Associated with this result, 72,00% of the companies do not use a strategic alignment methodology. The lack of a methodology aligned with the strategy of the company demonstrates that the administrators may have been making decisions based on information that is irrelevant to the institution or that is unaligned with the business strategy.

In this case, it is worth highlighting that despite the fact that the literature and practice conceptually separate the Data Mining and BI areas, there is a strong convergence and integration between them, because the "I", or the Intelligence in BI, can only be achieved by applying Data Mining techniques. This indicates that the lack of alignment methodologies must also reach the Data Mining projects, since there are BI projects without Data Mining, Data Mining without BI, and in the best scenario, a complete BI, which uses analytical data integration (BI), statistics and artificial intelligence (AI), i.e, Data Mining (Data Base + Statistics + AI).

Sharma, Osei-Bryson & Kasper (2012) address the several limitations identified in the existing Data Mining process models in their work, and suggest to solve them through the proposal of a new improved model named Integrated Knowledge Discovery and Data Mining Process model (IKDDM), which presents an integrated view of the KDDM (Knowledge Discovery and Data Mining) process and provides explicit support for the execution of each one of the tasks described in the model. The efficacy and efficiency provided by the IKDDM model against the CRISP-DM, a leader model in the KDDM process, were evaluated as well. The results of the statistical tests indicated that the IKDDM model surpasses the CRISP-DM model in terms of efficiency and efficacy. The IKDDM model also surpassed the CRISP-DM model in terms of the quality of the process model itself.

Kohavi et al. (2013) present the Bing Experimentation System, a system that tries to guide the development of products and enables the business organization to evaluate the ROI (Return Over Investment) of the projects through experimentation. The system enables the simultaneous execution of more than 200 experiments, exposing about 100 million clients to billions of Bing variables, which include the implementation of new ideas and variations of the existing ones.

Cheng et al. (2009) present an ontology-based approach for BI applications, specifically in Statistical Analysis and Data Mining, implementing the approach in a Financial Knowledge Management System. The knowledge resulting from each experiment, which consists of data sequences, model, parameters and reports, is stored, shared, disseminated, and thus, being useful for the decision-making process.

Finally, Colaço Júnior et al. (2019) present a process that merges the Goal/Question/Metric + Strategies approach with an agile development methodology for Business Intelligence applications, proposed by himself, aiming to ensure the strategic alignment. The proposed process was evaluated through a case study, in a Latin American multinational retail company, in which it was evidenced that it is possible to integrate the adopted strategic alignment approach with a methodology for the development of BI applications. With the good initial evidence, the researchers of this article evolved the process and predicted the use of experimentation for the validation of intelligent models that can be created with AI techniques, from a BI application that is ready or not.

CC BY

CONCEPTUAL BASIS

Necessary concepts for the understanding of this research work are presented in this section.

GQM+STRATEGIES

GQM+Strategies expanded the GQM model, the latter renowned as a systematic approach that integrates the business goals, adapting them to the software process models, products and quality-related perspectives, based on the specific needs of the project (Basili et al., 2007). GQM stands for the definition of **G**oals, **Q**uestions to be answered to achieve the goals and **M**etrics used to answer the questions.

According to Basili et al. 2010, GQM+Strategies is an approach to align the business organizations through measurement. It enables a business organization to consistently integrate the strategic alignment with its goals in different units, to make decisions based on identified metrics, to communicate goals and organizational strategies, as well as to monitor the achievement of the goals and the success/failure of the defined strategies.

The main result of this approach is a strategic measurement program that enables data-based decisions (Basili et al., 2010). In order to attach a software development technology to the strategic alignment, GQM+Strategies has two main components (Basili et al., 2010):

- Grid – It documents the strategic goals the business organization wishes to focus on, its justification for attaching the goals to different organizational units, and a measurement method to evaluate and interpret the data to be measured for the decision-making process.
- Process – It defines how to create the model, the implementation of its strategies, the gathering and analysis of data, besides how to initiate the improvement actions within the process.

In summary, there are six repetitive phases, plus an initiation one. The six phases are organized as a continuous improvement cycle and are based on the Quality Improvement Paradigm (QIP) proposed by Victor Basili et al. (2014). The phases are grouped into 3 macro stages, each one containing 2 specific phases. The 3 macro stages are:

- Develop – It develops the hierarchical model (grid), which aligns the goals, strategies and measurement data;
- Implement – It executes the strategies and measurements defined in the previous process, and thus verifies the execution of the goals and the efficacy of the strategies;
- Learn – It involves the knowledge from the stages that were conducted, through data analysis, to improve the process of generating new goals and strategies.

Such elements assist with the data gathering to be conducted in the implementation of the methodology, the flow and relation between them. They are defined by 6 phases: Initiation; Characterization of Environment; Definition of Goals, Strategies and Measurements; Implementation of Model; Planning of the Implementation of Model; Execution of Planning; Analysis of Results and Package of Improvements (Basili et al., 2014).

CC BY

EXPERIMENTAL DEVELOPMENT PROCESS OF DATA MINING AND DATA SCIENCE APPLICATIONS ALIGNED WITH THE STRATEGIC PLANNING

This section presents a process for the development of strategy-driven and experimentally evaluated Data Mining applications, as an integral part of the BI process. Initially, we will introduce in the next subsection another process that also aims for the strategic alignment, but of BI strategies without explicitly considering Data Mining.

GQM+STRATEGIES AND AN AGILE METHODOLOGY FOR THE REQUIREMENTS ELICITATION OF BI PROJECTS

The process proposed by Colaço Jr et al. (2019) is also an adaptation of the GQM+Strategies approach, by adding a new development methodology for BI applications, without the extension for the detailing of features that use Data Mining and Artificial Intelligence.

After executing the preliminary stages of the GQM+Strategies, the Intelligence Area is involved in phase 2 to execute the stage: Develop BI. This stage is divided in 6 (six) activities: 1.0 – Define the BI goal; 2.0 – Specify metric-based indicators; 3.0 – Define detail levels and visualization dimensions of the indicators; 4.0 – Write User Stories; 5.0 – Implement Prototype and 6.0 – Validate Strategic Goals.

In this regard, the purpose of this research work is to extend the strategy-driven BI methodology proposed by Colaço Jr et al. (2019) to contemplate experimentally evaluated Data Mining and AI applications. Figure 1 presents the macro proposal as a unified process, and the process itself (Develop DM) will be discussed in the next section.

EXPERIMENTAL DEVELOPMENT OF DATA MINING APPLICATIONS ALIGNED WITH THE STRATEGIC PLANNING

The proposed process, *Develop DM,* consists of 6 activities (Figure 2): (i) Develop the DM Process Goal; (ii) Prepare Data; (iii) Design Model; (iv) Evaluate the Process Experimentally; (v) Validate Strategic Goals and (vi) Implement DM.

All the activities that compose each stage of the process will be described next, and the Goal, Input, Subactivities and Results (Outputs) of each activity will be detailed as well.

DEVELOP THE GOAL OF THE DM PROCESS

The focus of this activity is to comprehend the goals and requirements of the business, converting the knowledge into a data mining problem. The goal, inputs, subactivities and results are shown in Table 1.

In order to define the mining goal for a business organization, the basic motivation needs to be determined, and the logic that leads to the goal definition needs to be written.

INPUTS

The GQM+Strategies grid is a GQM+Strategies approaching element already described in the conceptual basis of this research work. In this element, the strategic goals that the business organization wishes to focus on, its justifications for goal attachment, as well as a measuring method to evaluate and interpret the data to be measured for the decision-making process are documented.

CC BY

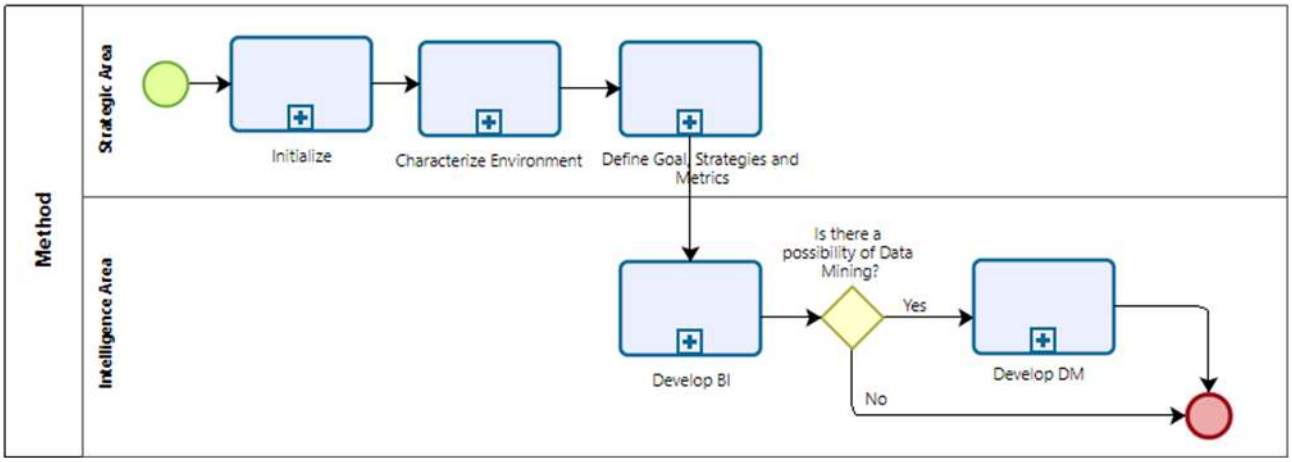

Figure 1. The Proposed Macro Process

Source. Author

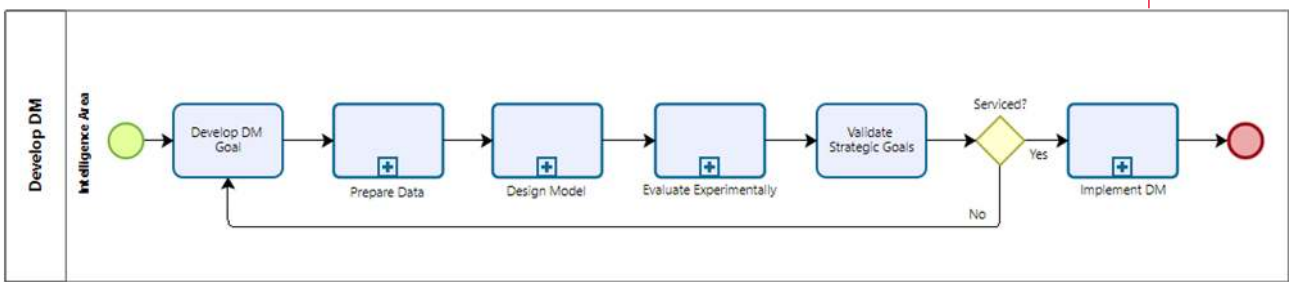

Figure 2. Activities of the Process Develop DM

Source. Author

Table 1. Activity Description: Definine Data Mining Goal

Activity: Define Data Mining Goal	
Goal	Defining the overall mining goal, attached to the strategic planning.
Inputs	The *GQM+Strategies* grid or element of another methodology, and Priority goal of the client.
Subactivities	Identify the organizational goals; Select the strategic goal in the grid; Select questions and metrics regarding the strategic goal; Write the overall increment goal based on the GQM questions; Review and Adjust.
Results (Outputs)	Preliminary scope and selected grid.

In this phase, the project goals and requirements are identified from the business perspective, and then these goals are converted into a data mining project. It is worth highlighting that the customer can use another methodology or another framework, like COBIT (**C**ontrol **Ob**jectives for **I**nformation and Related **T**echnologies) (Svatá, 2019), and can list the strategic goals and priorities in a different manner.

SUBACTIVITIES

The subactivities refer to all the necessary effort to execute the macro activity named: Define Data Mining Goal. Five subactivities were defined for the activity:

1. ***Identify the organizational goals*:** In order to achieve this subactivity, it is necessary to have the GQM+Strategies grid and the main goal of the Customer available. All the strategic goals defined by the business organization must be contained in them or in the artifacts of another methodology. This list is necessary so it is possible to identify which strategic goal will be selected.

CC BY

2. ***Select the strategic goal in the grid*:** Promising goals related to viability, benefit and cost are considered in this subactivity. Concentrating on the goals that have the highest success impact on the business is recommended. The selection process of the goals is quite interactive, demanding the participation of several unities of the business organization.
3. ***Select questions and metrics regarding the strategic goal*:** Once the questions and metrics in their strategic level are already defined and identified, then it is necessary to select the questions and metrics that will assist to define the scope, limit it, and constitute reasoning for the goals and their strategies selected in the GQM+Strategies grid. For example, for the goal of a superior level business, the questions and metrics usually refer to the restrictions and external opportunities and will be related to the vision and mission of the business organization. The restrictions and external opportunities include aspects like competitive products, commercial strategies with suppliers, and market tendencies. The restrictions and internal opportunities include aspects like competence level of the team, satisfaction of the internal client, technological advancements and existing infrastructure.
4. ***Write the overall increment goal, based on the GQM questions*:** It is necessary to describe the goals that are intended to reach through data mining, indicating the new data that will be visualized through their patterns and/or predictions. It is worth highlighting that all the previous stages, ideally, were already defined for a BI project Data Mart. In other words, what will happen to Data Mining in this case is the addition of new features that can describe the data in new manners through the discovery of patterns, and/or that can execute useful predictions to the strategy. BI User Stories may be improved or new ones may be created, by also considering the use of dynamic medias (multimedia) associated with the project, which can contain recordings of interviews and reports of clients about the desired features, such as those suggested in (Colaço Júnior et al., 2017) and in (Santos et al., 2020). Therefore, in this subactivity, the **delivery object** is the overall data mining goal, as well as the predictions and new visualization ways that will be available, based on the GQM questions (or strategic goals), **in order to** generally describe the visual and predictive insights of the current increment of the mining project, **from the viewpoint of** the Specific Client or the Specific Area, **in the context** of the business organization, influenced by the selected goal and the current increment of the mining project.
5. **Review and Adjust:** After all the previous subactivities are concluded, it is recommended to analyze and discuss the preliminary scope and the selected grid in a group meeting. It is also recommended that all the people in the organizational areas that are affected by the defined goal take part in the meeting. During it, the intelligence area explains all the process that resulted in the definition of the chosen goal. The participants must then verify if the connection between the goals and the strategies are logical and pertinent to the real world. Any issues brought up in the meeting are immediately discussed and the solution to each one can be: Planned, Immediately Addressed or Discarded in Session.

RESULTS

As an expected output of the subprocess **Define Data Mining Goal**, we have the "*Preliminary Scope and Selected Grid*" described in this section. In Table 2, we provided a model to be used, which is an adaptation of the GQM model.

CC BY

Table 2. Output Model of the Subprocess: Define Data Mining Goal

GOAL				
<Describing the strategic goal that the project aims to meet. It corresponds to an anticipated future state that a business organization aims to reach. Answer the question: "What must be reached?>				
FROM THE VIEWPOINT OF THE BUSINESS				
GOAL	**PURPOSE**	**QUALITY FOCUS**	**VIEWPOINT**	**CONTEXT**
< Object of Analysis >	*< Purpose of the Project >*	*< What are the possible metrics to measure the desired goal, according to the members of the project? >*	*< Stakeholders, people who will influence the output of the analysis >*	*< Target Audience >*
FROM THE VIEWPOINT OF MINING				
GOAL	**PURPOSE**	**QUALITY FOCUS**	**VIEWPOINT**	**CONTEXT**
< Object of Analysis >	< Purpose of the mining >	*<Predictability that is intended to reach, efficacy, efficiency, etc. >*	*< Stakeholders, people who will influence the output of the analysis >*	*< Context and suppositions >*
QUALITY FOCUS (QUESTIONS AND METRICS)			**VARIATION FACTORS**	
< What are the possible metrics to measure the object of interest, according to the members of the project? >			*< What contextual factors will influence the metrics, according to the expectation of a member of the project? Important information for the comprehension of the baseline hypotheses can be provided. >*	
BASELINE HYPOTHESES			**IMPACT ON BASELINE HYPOTHESES**	
< What is the current knowledge of the members of the project regarding the metrics? Can it be available from the actual data of previous projects? Or can it represent any type of opinion from an expert, in other words, suppositions about what can be true? >			*< How can the variation factors influence the actual measurements? What type of dependence between the metrics and the influence factors are assumed? What other data, which are necessary to interpret the model and the metrics, does it provide information on? >*	
MODEL INTERPRETATION				
< Interpretation of the described goals. >				
OVERALL GOAL WITH PRELIMINARY VIZUALIZATION SCOPE				
< Description of the overall goal with scope delimitation and expected final result. For the formalization of the goal, the use of the GQM model below is recommended. It is also recommended to add an outline of the expected result, in a manner to enable an easy understanding of the problem. For this reason, an initial prototype of the mining process outputs and User Stories can be used, according to the example shown in the case study. > *A summary of the goal can be described like this:* *ANALYZING* ***<Object of Study>,*** *IN ORDER TO* ***<Goal>,*** *REGARDING THE* ***<Approach>,*** *FROM THE VIEWPOINT OF* ***<Stakeholders>,*** *IN THE CONTEXT OF* ***<Context>.***				

PREPARE DATA

This activity focuses on the preparation of the Dataset that will be used in mining. The majority of the data used in a mining process is originally gathered and preserved for other purposes, and need some refining, integration, cleaning or transformation, before they are ready for a knowledge model training. The goal, inputs, subactivities and results of this activity are presented in Table 3.

INPUTS

The input of this subprocess is the document described in item **Results on page 8**.

Table 3. Activity Description: Prepare Data

Activity: Prepare Data	
Goal	Building the Dataset(s)
Inputs	Preliminary scope and selected grid.
Subactivities	Pre-select the Data; Supervise the Database; Balance the Database; Normalize the Database.
Results (Outputs)	Optimized Dataset for the mining process.

CC BY

SUBACTIVITIES

Subactivities refer to all the necessary effort for the execution of the macro activity named: *Prepare Data*. For this activity, 4 subactivities were defined:

1. ***Pre-select Data*:** A decision about the data used for the analysis must be made. The criteria include relevance to the data mining goals, quality and technical restrictions, such as limits to data volume or types of data.
2. ***Supervise the Base*:** In a supervised classification task, the supervision is performed when it is intended to find a function that is capable of predicting unknown labels from a previously defined labeled dataset.
3. ***Balance the Base*:** When the dataset presents an inequality among the samples from its different classes, the application of balancing techniques becomes necessary, in a manner to have a homogeneous dataset. If the accurate metric is used, the balancing is the sine qua non condition.

The sampling- and/or resampling-based methods are among the main balancing techniques. They consist in modifying the structure of the unbalanced dataset, in a manner to leave it with equivalent amounts of samples for the present classes, be it through subtraction (undersampling) or addition (oversampling) of new samples (Maione, 2020).

4. ***Normalize Base*:** The normalization of data consists in putting attributes in the same range of values. It is an operation that adjusts the value range of each attribute in a manner that the values are in small intervals, such as from -1 to 1, or from 0 to 1. Such adjustment is necessary to avoid that some attributes, for showing a value range higher than others, tendentiously influence specific data pattern methods. The variables can be normalized according to amplitude or according to distribution. There are some data normalization methods: linear normalization (or Max-min normalization), standard deviation normalization (or Z-score), decimal scaling normalization, normalization through the addition of elements and normalization through the maximum value of elements (Mín/Max) (Goldschmidt & Passos, 2005).

In the image below (Figure 3), we can observe how such subactivities are disposed in the subprocess or in the macro activity *Prepare Data*. In this case, there is an inclusive gateway, that is, all 3 paths can be activated.

RESULTS

As an expected output of the process Prepare Data, we have the optimized Dataset that will be used in mining. In Table 4, we provided a model to be used. In the first column of the table we have the attribute, in the second one its measurement unit, in the third column its domain, that represents the possible value set of the attribute, and lastly, a brief description of the attribute.

DESIGN MODEL

The most appropriated Data Mining techniques are selected and applied here, based on the goals identified in the first phase. The goal, inputs, subactivities and results of the activity are shown in Table 5.

CC BY

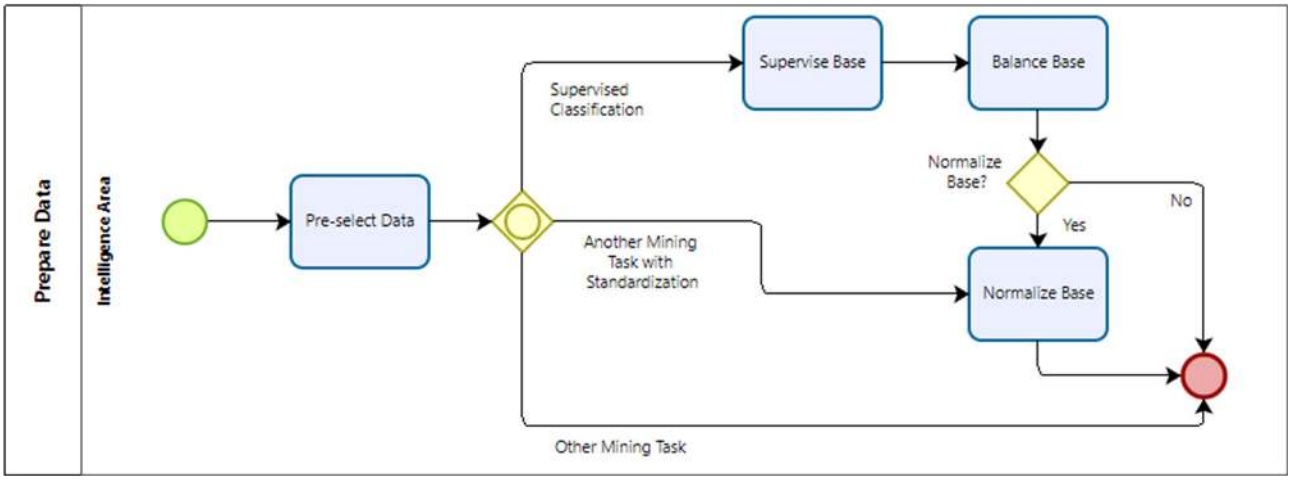

Figure 3. Activities of the Process Prepare Data
Source. Author

Table 4. Output Model of the Subprocess: Prepare Data

Attribute	Unit	Final Domain	Description
Attribute 1	Unit of Attribute 1	Domain of Attribute 1	Description of Attribute 1
Attribute 2	Unit of Attribute 2	Domain of Attribute 2	Description of Attribute 2
Attribute *n*	Unit of Attribute *n*	Domain of Attribute *n*	Description of Attribute *n*

Table 5. Activity Description: Design Model

Activity: Design Model	
Goal	Building Data Mining Model.
Inputs	Optimized Dataset.
Subactivities	Select Attributes; Select Algorithms; Transform Data; Define Parameters.
Results (Outputs)	Selected Algorithms and Attributes; Necessary Parameterization and Data Transformation.

INPUTS

The input of the subprocess is the document described in item **Results on page 10**.

SUBACTIVITIES

Subactivities refer to all the necessary effort for the execution of the macro activity named: *Design Model*. For this activity, 4 subactivities were defined:

1. ***Select Attributes*:** The purpose of selecting attributes is the elimination of redundant and uninformative attributes, as well as the creation of new ones. The elimination of these attributes can bring benefits, such as facilitating the understanding and visualization of data, as well as reducing the computing cost of the applied algorithm.

An exhausting search for the best possible subsets of attributes is usually inevitable from the viewpoint of computing, due to the number of possible subsets. Thus, the problem of selecting attributes can be approached via heuristic search methods. Some of the most used algorithms are: Random forest (Ma & Fan, 2017), Greedy hill Climbing, Best first and Genetic Algorithms (Covões et al., 2010).

CC BY

2. ***Select Algorithms*:** It consists in the selection of the algorithms that will constitute the data mining model. The types of involved variables, the techniques available in the tools that will be used, and the business goals gathered in the first stage of this research work must be considered.
3. ***Transform Data*:** The main goal of this phase is transforming the data representation in order to overcome any existing limitations in the algorithms that will be used for the extraction of patterns. Generally, the decision about which transformations are necessary depends on the algorithm that will be used in the data mining phase. Specific tools can be applied only to datasets with normal attributes, while other algorithms are able to infer and discover patterns related only to numeric variables, for example.
4. ***Define Parameters*:** There is always a high number of parameters that can be adjusted in any mining algorithm. The parameters and their chosen values must be listed, along with the justification for choosing the configurations of the parameters.

In Figure 4, we can observe the flow of such activities in the macro activity *Design Model*.

RESULTS

As an expected output of the subprocess **Design Model**, we have all the prospective algorithms for composing the mining model, which will be evaluated, and one or only some will be selected for use. Moreover, the selected attributes, the parameterization and data transformation are necessary. In Table 6, we provided a pattern to be used in the documentation of this model. In the first column of the table we have the algorithms selected for mining, in the second column the Dataset attributes used by the algorithms, and subsequently, the value of the parameters the algorithms were executed with.

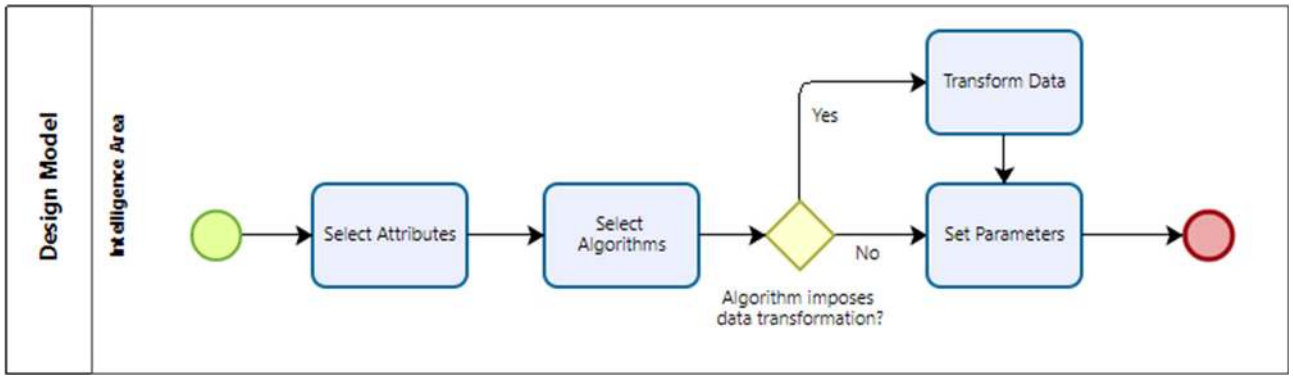

Figure 4. Activities of the Process Design Model
Source. Author

Table 6. Output Model of the Subprocess: Project Model

Algorithm	Attribute	Parameters	Transformation
Algorithm 1	Attribute 1, Attribute 2, Attribute n	P1: Value, P2: Value, Pn: Value	In case of transformation, fill in with the transformed attributes and the applied formula.
Algorithm 2	Attribute 1, Attribute 2, Attribute n	P1: Value, P2: Value, Pn: Value	
Algorithm n	Attribute 1, Attribute 2, Attribute n	P1: Value, P2: Value, Pn: Value	

CC BY

EVALUATE EXPERIMENTALLY

This stage can evaluate the degree to which the algorithms and applications meet the business goals, and it tries to determine if there are any business motives for which any algorithms or objects of study are defective. It is worth highlighting that this subprocess can be used to evaluate and validate any software components or any elements related to reaching business goals. Therefore, the use of an experimental approach is suggested, with which the scientific classic method is applied to validate if all the goals determined in the first stage of the process were achieved. The goal, inputs, subactivities and results of this activity are shown in Table 7.

INPUTS

The input of this subprocess is the document described in items **Results on page 10** and **Results on page 12**.

SUBACTIVITIES

Subactivities refer to all the necessary effort for the execution of the macro activity named: *Evaluate Experimentally*. For this activity, 5 subactivities were defined:

1. ***Define Experiment Goal*:** In this stage we defined the goal of the experiment. In other words, what we want to learn, what the object of analysis is, what the aspects of interest are, what the purpose of the study is, from which viewpoint and in which context the study will be conducted.
2. **Plan Experiment:** This stage corresponds to the planning of the experiment, with the following phases:
   a. ***Select the Context*:** The overall context, including the locality or business organization, along with the whole target audience (the population, not the sample), the people we want the results to be applied to.
   b. ***Enunciate the Research Questions*:** A research question is the declaration of a specific inquiry that the researcher needs to answer in order to address the research problem. The research question or questions guide the types of data to be gathered and the type of study to be developed.
   c. ***Define Dependent Variables:*** They refer to all the variables that are wished to be studied in an experimentation process, it means, they are the variables in which the researchers wish to observe the effect of the changes applied to the independent variables. They are also named study variables and are directly derived from the established hypotheses.
   d. ***Define Independent Variables:*** They refer to all the variables of an experimentation process that are manipulated and controlled, for example, a development method, the experience of people, and the environment. They are also named input variables.
   e. ***Formulate Theoretical Hypotheses:*** The purpose of the study must be translated into formal hypotheses. A hypothesis is a conjecture, a provisory answer that, according to certain criteria, will be Rejected or Non-rejected. A hypothesis must be formulated in a manner in which the occurrence of the observed phenomenon is attributed to chance (H0: null hypothesis). Immediately afterwards, another hypothesis that can be an alternative to the first one must be formulated, if it is demonstrated that chance cannot be held accountable by the observed phenomenon (H1: alternate hypothesis).

CC BY

Table 7. Activity Description: Evaluate Experimentally

Activity: Evaluate Experimentally	
Goal	Evaluating the prospective algorithms that will compose the Data Mining model to be built.
Inputs	Dataset with the selected attributes; Selected and Parameterized Algorithms; Transformations.
Subactivities	Define the Experiment Goal; Plan Experiment; Operate Experiment; Analyze and Interpret Data; Describe Threats to the Validity.
Results (Outputs)	Results and Detailing of the Experimental Process; Selected Data Mining Model (with one or more selected algorithms).

In this moment, during the formulation of hypotheses, theoretical variables defined through abstract concepts must be used. In the data validation stage, the operationalization of such variables will be formalized, it means, the representation of a theoretical variable via an operational variable is necessary to infer predictions from the hypotheses, aiming to avoid a possible misunderstanding between the operationalized hypothesis and the theory, or the generalization that this stage intends to test (Kluger & Tikochinsky, 2001).

For example, in a study on depression, which is a phenomenon related to other abstract variables such as rumination and anxiety, measuring these variables directly is a difficult task, once the operationalization of the hypotheses demands objectiveness. In other words, the researcher will have to recur to concrete things, such as the hormone levels of the person or psychological questionnaires that generate objective scores.

Therefore, after the definition of the hypotheses and the theoretical variables, one or more operational variables that appropriately represent the concept that will be evaluated must be chosen.

In data validation, after the normality test, the mathematically defined operational hypothesis might be formalized, once it will be already defined. For example, if a mean (parametric test) or median (nonparametric test) will be really used.

f. ***Select Participants and/or Objects:*** In face of the impossibility to evaluate the whole universe, in the majority of the cases it is necessary to define a part of the universe (sample) that we can evaluate. This sample must be the most representative one. It is ideal to calculate the size of the sample, or minimally use a sample for an infinite population.

g. ***Design Experiment:*** The experimental project is a complete plan to evaluate the experimental variables against the control variables, by following and mitigating the influence of the state variables. It involves the objects, measures, instructions, techniques, experimental format, and treatments. When there is not any possibility to evaluate two or more experimental groups with consolidated products, processes or methods to be evaluated, there must be a control group minimally materialized with an A/B test.

h. ***Define Instrumentation:*** The used tools and environment will be described here.

3. ***Operate Experiment:*** After planning the experiment, we go to the stage in which the experiment is run. This stage is divided in:

a. ***Prepare Experiment:*** Before performing the real experiment, the preparation may include a **pilot study** to confirm the experimental scenario, to help organize the experimental factors (for example, experiences of the participants) or to inoculate the participants. The pilot study will be a **homomorphism** of the whole activity of the *Operation* and is **strongly recommended**. If problems are detected, it may be necessary to return to the activity Plan Experiment, as can be seen in Figure 5.

CC BY

b. ***Execute Experiment:*** It consists of the following stages:

i. ***Describe Environment:*** Description of the environment in which the experiment was conducted;

ii. ***Run Experiment:*** The execution of the experiment per se;

iii. ***Gather Data:*** The gathering of the data of the experiment.

c. ***Define and Perform Data Validation:*** It consists in the definition of statistical methods that will be used to validate the data gathered by the experiment, as well as its execution. Data analysis may include a combination of quantitative and qualitative methods. The preliminary data selection, probably by using graphs and histograms, generally precedes the formal data analysis. The data analysis process usually requires the investigation of any adjacent suppositions (distributives, for example) before the application of statistical models and tests. In other words, after the confirmation of the normality and homoscedasticity of the data, the operational hypothesis may be formalized and tested, in mathematical terms.

4. ***Analyze and Interpret Data:*** It consists in the analysis and interpretation of the data, by also describing the statistical results that were used as the basis for the conclusions.

5. ***Describe Threats to the Validity:*** It is a description of everything that may threaten the result of the experiment. The threats can be (Colaço Júnior, 2018):

a. Conclusion: It is related to the ability of reaching a correct conclusion about the relations between the treatment and the result. Example: Choosing the statistical test;

b. Construction: Aspects related to the project and human factors. Example: Researcher develops a project based on what he/she expects;

c. Internal: It defines if the relation between the treatment and the result is casual, without the influence of a factor that may not have been measured. Example: A participant gives the other one some information;

d. External: Conditions that limit the ability to generalize. Consequences from other threats. Example: Selection of unrepresentative participants, atypical time and/or place.

In Figure 5, we can observe how these subactivities are disposed in the macro activity *Evaluate Experimentally*.

In Figure 6, we have the flow of the subprocess *Plan Experiment*, and in Figure 7 the flow of the subprocess *Operate Experiment*. These subprocesses have a simple sequential path, however, they facilitate the documentation of the experiment for future replications.

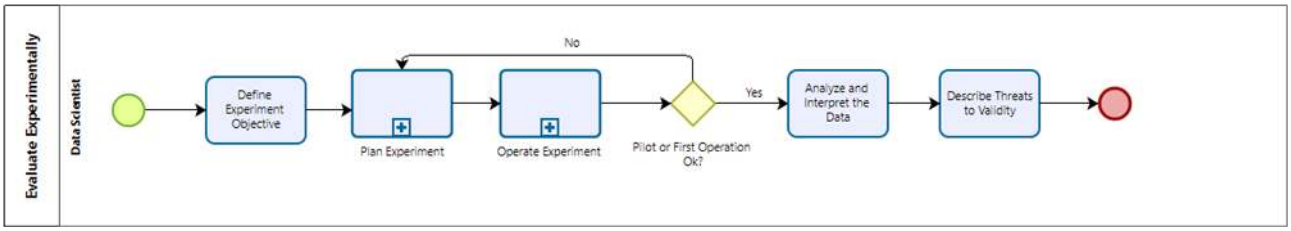

Figure 5. Activities of the Process Evaluate Experimentally

Source. Author

CC BY

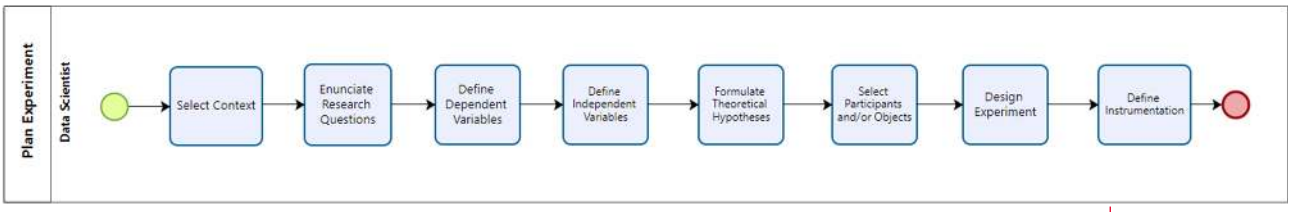

Figure 6. Activities of the Process Plan Experiment
Source. Author

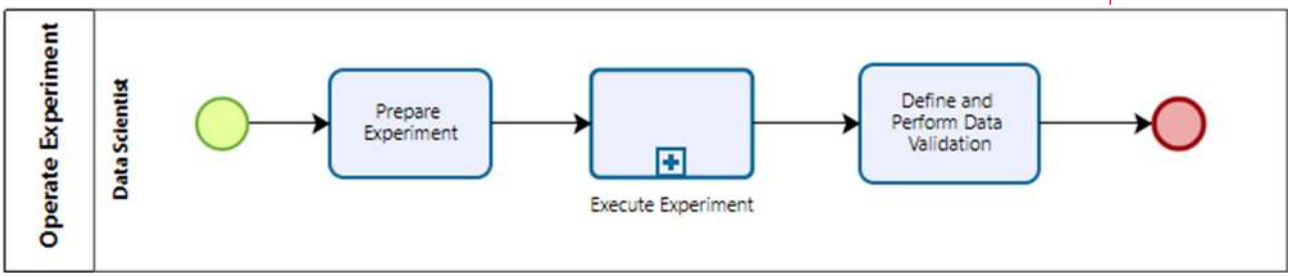

Figure 7. Activities of the Process Operate Experiment
Source. Author

RESULTS

As an expected output of the subprocess **Evaluate Experimentally**, we have the definition, the execution and the result of the whole experimental project, as well as the selected mining model.

VALIDATE STRATEGIC GOALS

The strategic goal of DM is validated in this activity. As the main goal, validations and acceptance are gathered by the client, for the implementation of the created mining model. The goal, inputs, subactivities and results of this activity are shown in Table 8.

INPUTS

The input of this subprocess is the document described in items **Results on page 8** and **Results on page 16**.

SUBACTIVITIES

Subactivities refer to all the necessary effort for the execution of the macro activity named: *Validate Strategic Goals*. For this activity, the following subactivities were defined:

1. ***Validate Strategic Goals:*** The validation of the strategic goals is the time to externalize everything that was gathered and developed with the technical teams and the business area, in order to validate, confirm and make the commitment to implement what was defined for DM. Nonconformity can still be found and the process can be restarted for validation and adjustments. In this stage, it is also important to revalidate the application prototype and visualizations to be implemented.
2. ***Review and Adjust:*** It is what suggests the subactivity of the same name in the macro activity *Develop the Goal of the DM Process*.

CC BY

Table 8. Activity Description: Validate Strategic Goals

Activity: Validate Strategic Goals	
Goal	Formalization and acceptance of the strategic goals for the implementation of DM.
Inputs	Preliminary scope and selected grid; Selected Data Mining Model; Description and/or recording, with new or incremented User Stories (optional); Visualization Prototype.
Subactivities	Validate Strategic Goals; Review and Adjust.
Results (Outputs)	Acceptance Document or Non-conformity List.

RESULTS

As an expected output of the subprocess **Validate Strategic Goals**, there is a nonconformity list or an acceptance document that will authorize the continuity of the process "Implement DM". In Table 9, we provided a pattern to be used in the documentation of this process.

IMPLEMENT DM

Finally, once the strategic goals were validated and approved, the data mining model must be implemented according to the defined prototype. The goal, inputs, subactivities and results of this activity are shown in Table 10.

INPUTS

The input of this subprocess is the document described in items **Results on page 8** and **Results on page 16**.

SUBACTIVITIES

Subactivities refer to all the necessary effort for the execution of the macro activity named: *Implement DM*. For this activity, the following subactivities were defined:

Table 9. Validation of the Strategic Goals

Validation Checklist for the Implementation of DM		
Evaluating Team		**Date**
Intelligence Area and Management		-
Set Goals and Prototypes		**Validation**
Strategic Goal		<Accepted/Refused>
Mining Goal		< Accepted/Refused >
Prototype		< Accepted/Refused >
Conclusion		
This document formalizes the acceptance delivery by considering it in conformity with the defined requirements and acceptance criteria, as well as by considering the validation of all produced documents.		
Participant	Signature	Date
Intelligence Area		-
Management (Business Area)		-

Table 10. Activity Description: Implement DM

Activity: Implement DM	
Goal	Implementation of the Selected Mining Model.
Inputs	Selected Data Mining Model. Visualization Prototype.
Subactivities	Select the Most Effective Algorithm(s); Train them with all the available Database; Implement Application by using the Algorithm(s).
Results (Outputs)	Implemented Data Mining Model.

CC BY

1. ***Select the Most Effective Algorithm(s):*** As a result of the experimental evaluation, we have the Selected Data Mining Model, with one or more selected algorithms.
2. ***Train them with all the Available Database:*** The whole database is used for training the selected algorithms.
3. ***Implement Application by using the Algorithm(s):*** The application is developed based on the defined and approved prototype.

In Figure 8 we have the flow of the subprocess *Implement DM*.

RESULTS

As an expected output of the subprocess **Implement DM**, we have the implementation of the selected data mining model according to the defined prototype.

CASE STUDY

A case study was planned and conducted in a federal educational institution (FEI) to evaluate the proposed process. Even though the context Big Data Volume was not totally explored independently of the size of the used database, the experimental evaluation is a necessary stage, and in some phase of the analysis, sampling must be used for a better performance, within a more specific business interest concept. In other words, the use of Small Data to search the best performance, to extract the best of Big Data or to use more recent data, avoiding that Artificial Intelligence models are not broken by events like the Covid-19 pandemic. The case study is presented and detailed in the following sections.

CASE STUDY STAGES AND GUIDELINES

- The main stages for the achievement of the case study were:
- Definition of the project: Defining the goals of the case study;
- Planning: Plans, instruments, and project of the case study;
- Operation of the Case Study: Defining the preparation and execution.
- Consolidation and publishing of results.

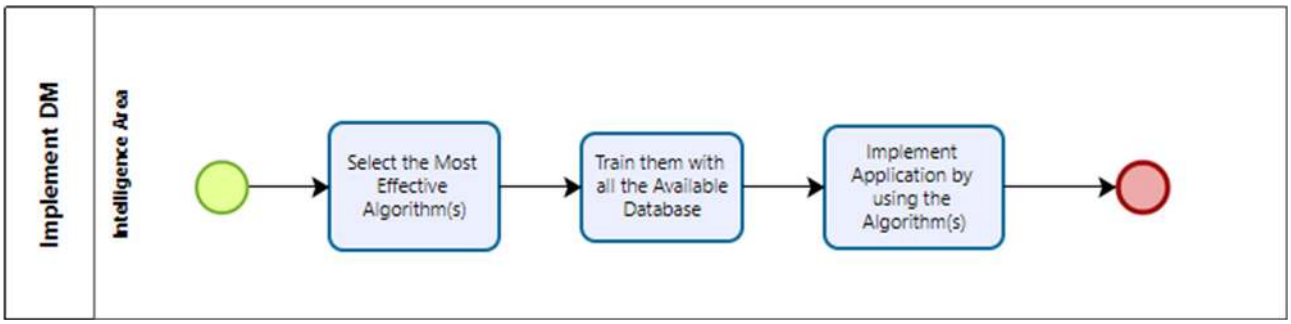

Figure 8. Activities of the Process Implement DM
Source. Author

CC BY

DEFINITION OF THE GOAL

The purpose of this study was to evaluate the process described **on pages 6 to 18**. Therefore, the research questions that need to be answered are these ones: 1st) Can a strategy-driven BI process be extended for the development of experimentally evaluated *Data Mining* and *Data Science* applications? 2nd) Will a process that encapsulates experimentation guide the exercise of Data Science?

Finally, the elaboration of an investigable supposition within the purpose of this work becomes necessary. The supposition in question: A strategy-driven BI methodology can be extended for the contemplation and development of experimentally evaluated *Data Mining* and *Data Science* applications.

PLANNING

The project was conducted in November and December of 2021, with the participation of the members of "5A Program", an academic intelligence project of the referred institution. During the application, the process guided the development of a Data Mining project, performing the 6 activities previously described.

A meeting for the presentation of the project with all the involved team was planned, as well as weekly follow-up meetings regarding the use of the process. As a communication plan, there was an agreement on the use of mechanisms provided by the company, such as email, chat and video conference.

SELECTION OF THE PARTICIPANTS

The selection was made through convenience and quota, through the access of the researchers into the institution, which has a rare well-defined intelligence team with specific characteristics of the population that the project aims to reach. The amounts, functions and professional experience time of the participants were: one Manager with up to 3 years of experience in the professional area, specifically; two Data Analysts, one with up to 3 years and the other with between 5 and 10 years; and six scholarship holders, with less than 1 year of experience in the area.

CASE STUDY INSTRUMENTATION

Process presented **on pages 6 to 18**, which had its execution evaluated through a questionnaire available at https://forms.gle/CkRso2fVD7JfM7QM9.

OPERATION

The preparation and execution of the study in question are described in this subsection.

PREPARATION

After the training in the process, the team acknowledged the necessity of the strategic goals of the business organization, and gathered them, selecting a prioritized one for the application of the process, which could be supported by intelligent applications. The gathering happened along with the strategic team, the one which agreed on a Data Mining project to support the mitigation of school dropout.

CC BY

EXECUTION

For each activity, the produced artifacts that would be used as requirements for the following activities were documented. Lastly, the support for the respective selected strategic goal was validated. The detailing of the execution will be presented in the next section.

RESULTS

The produced documents resulting from the activities were analyzed, so that the research questions could be answered. They will be presented in the following sections.

DEVELOPING THE GOAL OF THE DM PROCESS

In order to develop the Mining project to support the achievement of the selected business goal, the goal of the Data Mining process was developed. The document was produced from the template defined in section **Results on page 8**, according to Table 11.

Table 11. Preliminary Scope of the DM Goal

GOAL				
Reducing school dropout by identifying which factors may contribute to the academic unsuccess of the students.				
FROM THE VIEWPOINT OF BUSINESS				
OBJECT	**GOAL**	**QUALITY FOCUS**	**VIEWPOINT**	**CONTEXT**
Students of the Graduation Level of all the campuses of the Federal Institution.	Evaluating which characteristics of the students can compose the academic unsuccess factors.	Reducing the academic unsuccess of the students in the school dropout issue by 5% by the end of the year.	Rector, Pro-rector, Student and Professor Coordinators, Office Managers and Principals.	Education Area
FROM THE VIEWPOINT OF MINING				
OBJECT	**GOAL**	**QUALITY FOCUS**	**VIEWPOINT**	**CONTEXT**
Data Mining Algorithms	Evaluating and Predicting	Reaching a school dropout prediction with accuracy of 90% or more	Office Managers, students, data analysis professionals and data scientists	Students of the graduation level of all the campuses of the FEI.
QUALITY FOCUS (QUESTIONS AND METRICS)			**VARIATION FACTORS**	
PE-G-Q1 – What is the school dropout rate in a school year? *Sdr (A):* ***School dropout rate in year A.*** *ALG-G-Q1 – What are the accuracies of the main machine learning algorithms for the school dropout prediction task?* *Acur (n): Algorithm n accuracy*			–	
BASELINE HYPOTHESES			**IMPACT ON BASELINE HYPOTHESES**	
Sdr(2020) = 8,22% *Acur(Decision Tree) = 70%*			–	
MODEL INTERPRETATION				
PE-G-Q1 = Sdr(2021) / Sdr(2020) <= 0.95 <Reduction of 5% in school dropout from 2021 in relation to the dropout of 2020> *ALG-G-Q1 = Acur (n) >= 90%*				
OVERALL GOAL WITH PRELIMINARY VISUALIZATION SCOPE				
ANALYZING data mining algorithms, **IN ORDER TO** predict and evaluate, **REGARDING THE EFFICACY** of the school dropout prediction, **FROM THE VIEWPOINT OF THE** Office Managers, Students, Data Analysts and Data Scientists, **IN THE CONTEXT** of the students who are active and enrolled in the graduation level of a Federal Institution. As the mining result, the general dropout probability will be presented, as well as the individual probability, identifying the active and enrolled students who tend to drop out, enabling the adoption of measures that can change such scenario. In the image below (Figure 9), we have the initial prototype of the mining process outputs.				

CC BY

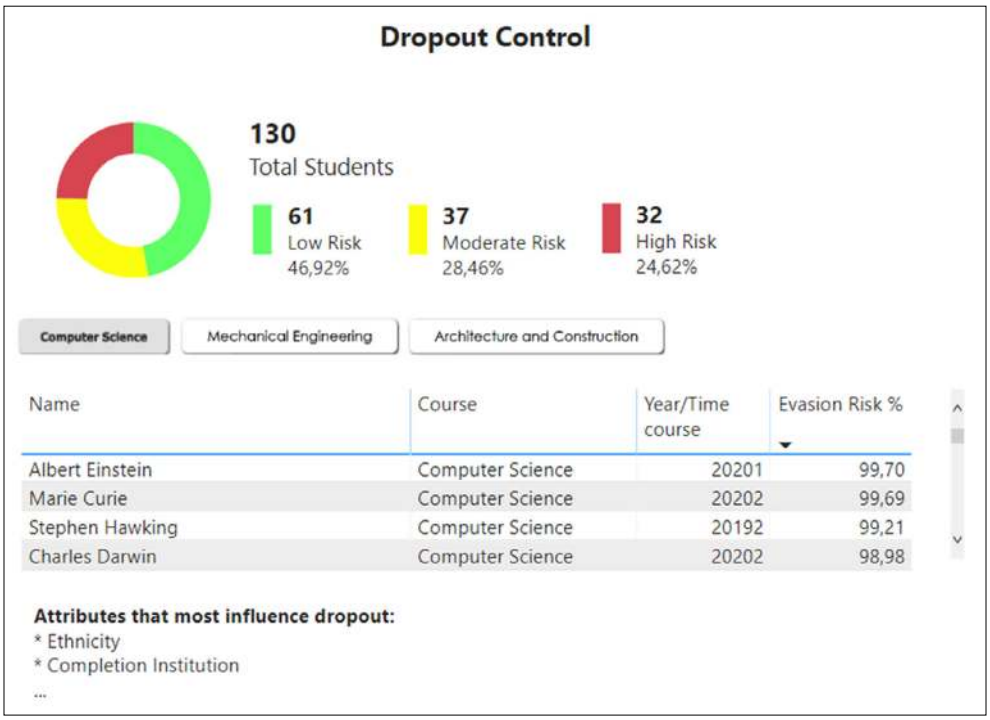

Figure 9. Prototype
Source. Author

PREPARING DATA

In the activity Prepare Data, the following Dataset (Table 12) was selected:

Table 12. Dataset

Attribute	Description
Sexo	Gender of the student
Idade	Age of the student
tipo_instituicao_conclusao	Type of educational institution where the student concluded high school
Raca	Ethnicity
est_civil	Marital status
qtd_trac	Amount of subjects interrupted
reab_matricula	It indicates whether the student reinstated the course
qtd_ap_med_p	Average amount of approved subjects by semester
qtd_ap_1p	Amount of approved subjects in the first semester
qtd_rep_med_p	Average amount of failed subjects by semester
qtd_rep_1p	Amount of failed subjects in the first semester
qtd_per_cur	Amount of semesters attended by the student
Cra	Academic Performance Coefficient
qtd_disciplinas_concluidas	Total amount of subjects student passed
qtd_disciplinas	Total amount of subjects student enrolled in
media_geral	Overall performance rate of the student in the course
media_faltas	Overall absence rate of the student in the course
Cotista	It indicates whether the student was admitted through quota system

CC BY

DESIGNING MODEL

The classification algorithms provided by the Pycaret library were selected. Pycaret is a Python open source machine learning library, which was selected due to its compatibility with Google Colaboratory, a collaborative environment used in the education organization (Gaian & Hotti, 2021). The algorithms are: Ada Boost Classifier, Decision Tree Classifier, Extra Trees Classifier, Gradient Boosting Classifier, K Neighbors Classifier, Light Gradient Boosting Machine, Linear Discriminant Analysis, Logistic Regression, Naive Bayes, Quadratic Discriminant Analysis, Random Forest Classifier, Ridge Classifier and SVM - Linear Kernel.

The following procedures were conducted for data transformation: The attribute 'tipo_instituicao_conclusao' had its null values filled with the value 'OUTRA' (OTHER); The attribute 'est_civil' had its null values filled with the value 'Other'; The attributes 'sexo', 'tipo_instituicao_conclusao', 'raca' e 'est_civil' had their values altered to numbers, instead of texts; Only the records that have status in: 'CANCELED', 'ACTIVE', 'CONCLUDED', 'EXPELLED', 'ACTIVE – LAST STAGE OF UNDERGRADUATION', 'ACTIVE – UNDERGRADUATE WAITING FOR DIPLOMA' were filtered; The text status were changed to binary status, as it follows: status: 'CANCELED', 'EXPELLED' were replaced by 1 (target – dropout); status: 'ACTIVE', 'CONCLUDED', 'ACTIVE - LAST STAGE OF UNDERGRADUATION', 'ACTIVE - UNDERGRADUATE WAITING FOR DIPLOMA' were replaced by 0 (control – students who did not drop out).

The algorithms parameters were the Pycaret defaults.

EVALUATE EXPERIMENTALLY

An experiment to evaluate the best algorithm focused on school dropout, in terms of efficacy, was conducted. The main algorithms may be used to form a prediction metamodel that will assist with the dropout management. Therefore, the experiment described in Table 13 was conducted.

The results were used to answer the research question. The algorithms Light Gradient Boosting Machine and Gradient Boosting Classifier (GBC) had quite similar and close averages for the metrics, once Light is a customization by Microsoft. However, it had execution time inferior to GBC, which was followed, in terms of efficacy, by Random Forest Classifier, with averages that were also similar to the ones of the first ranked algorithms. The remaining presented algorithms did not reach the established minimal goal of 90%.

Besides the ranking of the algorithm Light, sufficiently conclusive statistical evidence was necessary to define the best algorithm among the ones that reached the goal. Therefore, a significance level of 0,05 was established. When applying the Shapiro-Wilk Test for the normality analysis of the distribution of data, we obtained the p-values presented in Table 14, in which it is possible to observe values above the adopted significance level, concluding that the distribution of data for the metrics Accuracy and F1-Measure are normal.

As only three algorithms remained and the data are normal, we opted for comparing the algorithms in pairs, through the Paired T Test. Moreover, also because of the normality, the operationalization of the hypothesis could already be conducted with the use of the average. Therefore, the following hypotheses were formalized (for each evaluated metric and for each pair of algorithms):

CC BY

Table 13. Experimental Evaluation of the Mining Model

DEFINITION OF THE EXPERIMENT GOAL
Goal Definition
It was already described in Table 11.
EXPERIMENT PLANNING
Context Selection
It was an "in vitro" experiment, because the data were taken from the real environment, so they could be transformed and then used in a controlled environment. The data of the students in all Undergraduate courses were considered, by incorporating the freshmen between 2003, the year the first undergraduate courses started, and 2020. The data gathering considered personal, academic and social attributes.
Research Questions
In the context of school dropout, from the selected algorithms, which ones presented the best indicators in terms of efficacy?; Does the accuracy exceed the established goal of 90%?;
Dependent Variables
Classifications, from which can be derived: Accuracy, Sensitiveness (Recall), Precision and F1-Measure.
Independent Variables
The dataset described in Table 12 and the algorithms previously listed.
Formulation of Theoretical Hypotheses
• H_0: The algorithms ($_{1,2..n}$) have the same efficacy. • H_1: The algorithms ($_{1,2..n}$) have different efficacy.
Selection of Participants and Objects
All the undergraduate students of the Institution were selected, totalizing 10,949 students, of which 6,961 (63,57%) were part of the class **target** (Dropouts) and 3.988 (36,42%) were part of the class **control** (Non-dropouts). One of the metrics used was accuracy, which requires the balancing of the classes. Thus, it was necessary to conduct the balancing process, the one that considered the highest amount of records presented in each class, 3,988, being the final total a value superior to a sample for infinite population, according to the literature foundation (Pinto, 2015). Considering the population of the Institute, the final sample of 7,976 students has margin of error of 0,57%, for a reliability of 95%.
Experiment Project
For the evaluation of the model, we used the 10-Fold Cross-Validation approach, in which the data are divided into 10 parts, keeping their proportions. Therefore, 10 tests were conducted, separating a part of the data to be tested later. Besides, it will be possible to obtain annual, semi-annual, bi-monthly, quarterly, monthly, and half-monthly results
Instrumentation
For the data mining process, we used the Python pycaret library, which is a high-level open source machine learning library, whose purpose is to maximize the comparison and usability performance of the Scikit-Learn library. For the execution of the Python code, we used the Google Colab cloud environment, which is destined to the creation and execution of codes in Python, directly from a browser, without the necessity to install any software into local machines. The data used for analysis came from SIGAA, the academic system used in the institution, which uses the PostgreSQL as DBMS (Database Management System).
EXPERIMENT OPERATION
Preparation
It consisted in the preparation of the dataset that was used in mining. The preparation occurred by following the subprocess *Prepare Data* of the proposed process.
Execution
It consisted in the execution of the classificatory process in the data, planned in the experiment Project, for each selected mining algorithm, by using the other independent variables.
Data Validation
Three types of statistical tests were used to assist with the analysis, interpretation and validation: the Shapiro-Wilk Test, the Levene Test and the Paired T Test. The Shapiro-Wilk Test was used for the Normality test, and the Levene Test for the homoscedasticity test. Once the normality assumption was fulfilled and the homoscedasticity was not, the Paired T Test was used to test the hypotheses.
RESULTS
Data Analysis and Interpretation
After the execution of the algorithms, the results of the classifications, which will be presented in Figure 10, were obtained by using the 10-Cross-Validation approach.
Threats to the Validity
Threats to the internal validity: The current academic system has been present in the institution since 2017, and has inherited the basis of the previous academic system with several inconsistent information, mainly by the middle of 2007. This threat was mitigated by conducting the cleaning process of the data, reducing the probability of the use of older incorrect information.

CC BY

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9070	0.9693	0.8912	0.9207	0.9056	0.8141	0.8147	0.163
gbc	Gradient Boosting Classifier	0.9045	0.9670	0.8891	0.9179	0.9031	0.8091	0.8098	0.919
rf	Random Forest Classifier	0.9022	0.9654	0.8933	0.9101	0.9014	0.8044	0.8050	0.866
ada	Ada Boost Classifier	0.8841	0.9564	0.8790	0.8883	0.8835	0.7682	0.7685	0.325
et	Extra Trees Classifier	0.8807	0.9522	0.8797	0.8820	0.8807	0.7614	0.7617	0.823
dt	Decision Tree Classifier	0.8714	0.8718	0.8733	0.8706	0.8717	0.7428	0.7431	0.057
knn	K Neighbors Classifier	0.8494	0.9154	0.8665	0.8383	0.8520	0.6987	0.6994	0.157
lr	Logistic Regression	0.8110	0.9005	0.8175	0.8076	0.8123	0.6221	0.6225	0.911
ridge	Ridge Classifier	0.8105	0.0000	0.8264	0.8016	0.8135	0.6210	0.6219	0.034
lda	Linear Discriminant Analysis	0.8105	0.8992	0.8257	0.8021	0.8134	0.6210	0.6218	0.068
nb	Naive Bayes	0.7625	0.8535	0.6804	0.8186	0.7397	0.5251	0.5360	0.032
svm	SVM - Linear Kernel	0.6896	0.0000	0.7741	0.7410	0.7003	0.3791	0.4400	0.110
qda	Quadratic Discriminant Analysis	0.5157	0.5160	0.1682	0.5710	0.2444	0.0319	0.0496	0.044

Figure 10. Comparative of the Algorithm Metrics
Source. Author

Table 14. Result of the Shapiro-Wilk Test for the analysis of data normality

Algorithm	Accuracy	F1-Measure
Light Gradient Boosting Classifier	0.0856	0.6809
Gradient Boosting Classifier	0.2881	0.1274
Random Forest Classifier	0.1525	0.7854

- H_0: The algorithms ($_{1,2}$) have the same averages for the metric.

$\mu_1(\text{metric}) = \mu_2(\text{metric})$;

- H_1: The algorithms ($_{1,2}$) have different averages for the metric.

$\mu_1(\text{metric}) \neq \mu_2(\text{metric})$.

By applying the Paired T Test for each pair of algorithms, the following p-values were obtained (Table 15):

Once the p-values presented values higher than the established significance level, the null Hypotheses (H_0) could not be rejected. In other words, the differences between the averages of the algorithms were not high enough to be statistically significant, indicating that the 3 algorithms were equivalent, regarding the compared measures. Considering the most important metrics for the problem in question, it was possible to build a system with a model that used three winner algorithms, by deciding the classification through the most recurrent result. The creation of this metamodel will generate a new hypothesis to be tested.

Table 15. Result of the Paired T Test

Algorithm 1 against Algorithm 2	Accuracy	F1-Measure
Light Gradient Boosting Classifier/Gradient Boosting Classifier	0.4951	0.5045
Light Gradient Boosting Classifier/Random Forest Classifier	0.08812	0.1038
Gradient Boosting Classifier/Random Forest Classifier	0.142	0.1603

CC BY

VALIDATING STRATEGIC GOALS

A validation checklist for the implementation of DM was applied in the activity *Validate Strategic Goals* (Table 16).

IMPLEMENT DM

The implementation activity was not followed by this study.

EVALUATION OF THE PROCESS

In order to interpret the presented results, a qualitative evaluation was proposed. According to Demo (2012), the qualitative evaluation of a research study tries to preserve and seek information in the real world. The qualitative information can obtain reliability in the correct execution of the procedures.

EVALUATION METHOD

The evaluation was applied to the expert team of the institution, with diverse experiences in the area, by using the questionnaire described in **Case Study Instrumentation**, which contains a set of qualitative questions about the process proposed here. The whole process described in this research work, as well as the produced artifacts, were provided and experienced by each appraiser. At the end of the project, the questionnaire was applied under supervision.

ANALYSIS OF THE EVALUATIONS

The studied institution did not use any formal methodology to develop its Data Mining projects, confirming the data verified by Lima et al. (2017), and the quasi-systematic literature review conducted in the methodology of this research work (Cruz; Colaço Júnior & Gois, 2022).

All the interviewees mentioned the Strategic Alignment of the Data Mining projects as something important. Furthermore, three informed that the experimentation can also be used to guide the development of several products and can enable the institution to evaluate the ROI (Return Over Investment) of the software projects.

Table 16. Validation of the Strategic Goals

Validation Checklist for the Implementation of DM		
Evaluating Team		**Date**
Intelligence Area and Management		01/18/2022
Set Goals and Prototype		**Validation**
Reducing the academic unsucess of the students in the school dropout issue by 5% by the end of the year		Accepted
Reaching a school dropout prediction with accuracy of 90% or more		Accepted
Prototype		Accepted
Conclusion		
This document formalizes the delivery acceptance by considering it in conformity with the defined requirements and acceptance criteria, as well as by considering the validation of all produced documents.		
Participant	**Signature**	**Date**
Intelligence Area	<Supressed>	01/18/2022
Management (Business Area)	< Supressed >	01/18/2022

CC BY

About the possibility to attach the Data Mining project to the strategic planning of the company, there was full agreement in relation to the support given by the proposed process. Regarding the experimentation efficacy in the evaluation of the algorithms that will compose the data mining, 70% of the interviewees answered that they fully agree, and 30% that they agree.

When asked if the proposed process assists the professional to follow the experimental model without neglecting any phases, 30% answered that they fully agree and 70% answered that they agree. The summary of the evaluation can be seen in Figure 11.

Therefore, the initial results brought evidence that it is possible to integrate the strategic alignment, scientific method and a development methodology of DM applications, fomenting the experimentation as a feedback element for the Strategic Planning.

THREATS TO THE CASE STUDY INTERNAL VALIDITY

The case study may be positively influenced by the proponents of the process, which tend to defend their products and hide problems. This fact may cause a phenomenon studied by psychology named Demand Characterization, which considers that an experimental artifact may be an interpretation of the participants regarding the purpose of the experiment, leading to inconsistent behavior changes, in order to adapt to such interpretation (Orne, 1962). To mitigate this factor, it can be said that at least two different approaches were used: The More The Merrier and Unobtrusive Manipulations and Measures (Orne, 1962). Respectively, in the first one, the process was applied by the employees of the company who were not involved in the research, in order to mitigate bias. The second one guided us not to inform which factors and metrics would be used during the conclusion, so the participants did not have any clues about the hypothesis of the research. Lastly, an interview was conducted with the participants, aiming to evaluate the results qualitatively.

CONCLUSION AND FUTURE WORKS

The main contribution of this research work was the proposal and evaluation of an approach to guide the strategic alignment and the experimental development of *Data Mining* and *Data Science* applications. The work was consolidated by the achievement of a case study in a federal educational institution, in order to help consolidate the proposed process.

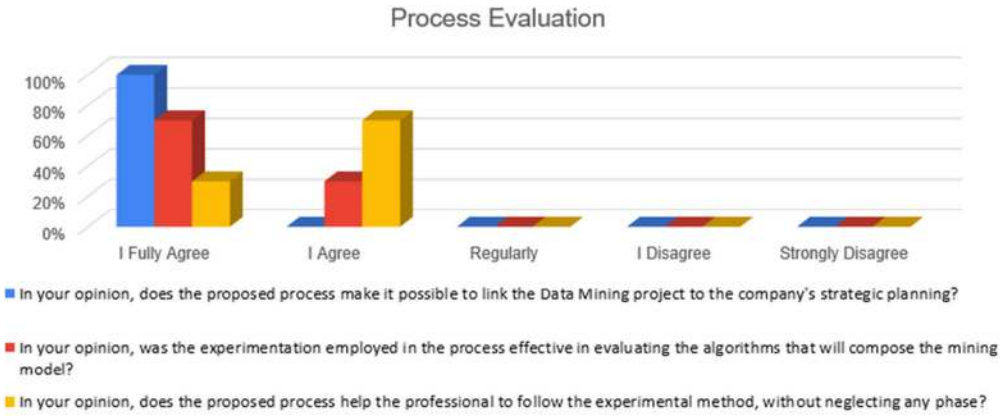

Figure 11. Evaluation of the Proposed Process
Source. Author

CC BY

The combination of the GQM+Strategies approach with the experiment was successful, once it was possible to guide and align the development of the DM application with the strategic planning of the business organization, by using the scientific method and metrics that support the implementation of this project as basis. Regarding the strategic alignment per se, fomentation, emphasis and communication benefit from the mapping of processes that transform the search for strategic goals into explicit requirements, and the systematic approach of the scientific method is also facilitated.

It is important to highlight that the processes of knowledge discovery may have their goals altered, once the business processes supported by this technological subprocess may not be functioning well, according to their performance indicators. In other words, the goals to be reached by the DM applications may be directly technical, such as reaching a minimal time to execute a search, and demand for the accuracy of the information. However, these goals must also reflect the range of the business goals, such as the acquirement of new clients or dropout reduction. If the intelligent application meets its goals but the supported business process does not, the level where the problem is must be identified (software and/or business) and a planning with an improvement package and new goals must be put into practice. For example, if the system indicates that the Ethnicity of the student is a preponderant dropout factor, what actions should be taken from such viewpoint?

As future works, it is necessary to apply the process in other companies, with different sizes and complexities, evaluating the adherence to the process by any business organizations. Moreover, as the subprocess "Evaluating Experimentally" can be used to evaluate and validate any software components or any elements related to the range of the business goals, it can be replicated and evaluated in other areas, adding the optional activity of creating an experimental package, which shares the tools and data used in the experiment. This subprocess is in line with the proposals made by Sculley, Snoek, Rahimi & Wiltschko (2018), Google AI researchers, systematizing an experimental process for AI applications, whose activities titles can even serve as sections standardized of articles on AI experimentation.

Lastly, rigorous specification of experiments increases transparency and helps to mitigate the reproducibility problem in the fields of Data Mining and AI, in which researchers are unable to replicate each other's results because of inconsistent experimentation and publishing practices.

REFERENCES

Basili, V. R. (1996, March). The role of experimentation in software engineering: past, current, and future. In *Proceedings of IEEE 18th International Conference on Software Engineering* (pp. 442-449). IEEE.

Basili, V., Heidrich, J., Lindvall, M., Munch, J., Regardie, M., & Trendowicz, A. (2007, September). GQM+Strategies – Aligning Business Strategies with Software Measurement. In *First international symposium on empirical software engineering and measurement (ESEM 2007)* (pp. 488-490). IEEE.

Basili, V. R., Lindvall, M., Regardie, M., Seaman, C., Heidrich, J., Münch, J., ... & Trendowicz, A. (2010). Linking software development and business strategy through measurement. *Computer*, *43*(4), 57-65.

V. Basili, A. Trendowicz, M. Kowalczyk, J. Heidrich, C. Seaman, J. Münch, D. Rombach. (2014). *Aligning Organizations Through Measurement: The GQM+Strategies Approach*. *Springer Publishing Company, Incorporated.*

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Bock, C., Gumbsch, T., Moor, M., Rieck, B., Roqueiro, D., & Borgwardt, K. (2018). Association mapping in biomedical time series via statistically significant shapelet mining. *Bioinformatics*, *34*(13), i438-i446.

CC BY

Bosch-Sijtsema, P., & Bosch, J. (2015). User involvement throughout the innovation process in high-tech industries. *Journal of Product Innovation Management*, *32*(5), 793-807.

Botelho[1], F. R., & Filho, E. R. (2014). Conceituando o termo business intelligence: origem e principais objetivos. *Sistemas, Cibernética e Informática*, vol. 11, n.º 11, pp. 55–60.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Cheng, H., Lu, Y. C., & Sheu, C. (2009). An ontology-based business intelligence application in a financial knowledge management system. *Expert Systems with Applications*, *36*(2), 3614-3622.

Cios, K. J., Teresinska, A., Konieczna, S., Potocka, J., & Sharma, S. (2000). Diagnosing myocardial perfusion from PECT bull's-eye maps-A knowledge discovery approach. *IEEE Engineering in Medicine and Biology Magazine*, *19*(4), 17-25.

Clancy, T. (1995). The standish group report. *Chaos report*.

Colaço Júnior, Methanias, de Fátima Menezes, M., Corumba, D., Mendonça, M., & Santos, B. S. (2015). Do software engineers have preferred representational systems?. *Journal of Research and Practice in Information Technology*, *47*(1), 23-46.

Colaço Júnior, M. (2018). *Vocabulário e Definição de Estudos Experimentais* [Material da Disciplina de Engenharia de Software Experimental]. Mestrado em Ciência da Computação, Universidade Federal de Sergipe, São Cristóvão, Sergipe.

Colaço Júnior, M. ; Cruz, R. F. ; Lima, A. S. (2019). Proposal and Evaluation of a Strategy-Driven Business Intelligence Applications Development Process; Proposta e Avaliação de um Processo para o Desenvolvimento de Aplicações de Business Intelligence Dirigido à Estratégia. In: *International Conference on Information Systems and Technology Management*, 2019, São Paulo. ContecSI. DOI: 10.5748/16contecsi/kmg-6129.

Côrte-Real, N., Oliveira, T., & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. *Journal of Business Research*, *70*, 379-390.

Costa, J. K. G., Santos, I. P. O., Nascimento, A. V. R., & Júnior, M. C. (2015, May). Experimentation at Industrial Setting to Improve the Effectiveness of the ETL Procedures Implementation in a Business Intelligence Environment. In *SBSI* (pp. 459-466).

Costa, J. K., Santos, I. P., junior, M. C., & Nascimento, A. V. (2016, May). An Experiment in an Industrial Business Intelligence environment to improve data loads maintenance. In *Proceedings of the XII Brazilian Symposium on Information Systems on Brazilian Symposium on Information Systems: Information Systems in the Cloud Computing Era-Volume 1* (pp. 534-541).

Covões, T. F. (2010). *Seleção de atributos via agrupamento* (Doctoral dissertation, Universidade de São Paulo).

CRISP-DM. (2003). *Cross Industry Standard Process for Data Mining 1.0: Step by Step Data Mining Guide.* [Online] 20 de Junho de 2019. http://www.crisp-dm.org/.

Cruz, R. F.; Colaço Júnior, Methanias; Gois, V. M. (2022). How experimental and strategic are Business Intelligence (BI) and Data Mining applications?; Quão experimentais e estratégicas são as aplicações de Business Intelligence (BI) e Data Mining? *Revista Ibero-Americana de Estratégia*. DOI: 10.5585/riae.v21i1.17689.

Demo, P. A. e Silva, R. (2012). *Pesquisa e Informação Qualitativa* 5. ed. São Paulo.

Dittrich, Y., Nørbjerg, J., Tell, P., & Bendix, L. (2018, May). Researching cooperation and communication in continuous software engineering. In *2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)* (pp. 87-90). IEEE.

Endres, A., & Rombach, H. D. (2003). *A handbook of software and systems engineering: Empirical observations, laws, and theories*. Pearson Education.

Fagerholm, F., Guinea, A. S., Mäenpää, H., & Münch, J. (2017). The RIGHT model for continuous experimentation. *Journal of Systems and Software*, *123*, 292-305.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

Gain, U., & Hotti, V. (2021, February). Low-code AutoML-augmented Data Pipeline–A Review and Experiments. In *Journal of Physics: Conference Series* (Vol. 1828, No. 1, p. 012015). IOP Publishing.

Goldratt, E. M., & Cox, J. (2016). *The goal: a process of ongoing improvement*. Routledge.

Goldschmidt, R., & Passos, E. (2005). *Data mining: um guia prático*. Gulf Professional Publishing.

Hohnhold, H., O'Brien, D., & Tang, D. (2015, August). Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1849-1858).

IBM. (2005). *Analytics solutions unified method*. ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf.

Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the" theoretical" null hypothesis: the rise, fall, and resurrection of commonsense hypotheses in psychology. *Psychological bulletin*, *127*(3), 408.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, *18*(1), 140-181.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013, August). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1168-1176).

Kohavi, R., Deng, A., Longbotham, R., & Xu, Y. (2014, August). Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1857-1866).

Kohavi, R., & Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. *Encyclopedia of machine learning and data mining*, *7*(8), 922-929.

Kubina, M., Varmus, M., & Kubinova, I. (2015). Use of big data for competitive advantage of company. *Procedia Economics and Finance*, *26*, 561-565.

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review*, *21*(1), 1-24.

Ławrynowicz, A., & Potoniec, J. (2014). Pattern based feature construction in semantic data mining. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *10*(1), 27-65.

Lima, Adriano; Colaço Júnior, Methanias; Nascimento, Andre Vinicius RP. (2017). Um Survey com Empresas Brasileiras acerca da Utilização de Business Intelligence (BI) e um diagnóstico sobre a infraestrutura e metodologias associadas. *Conference: Ibero-American Conference on Software Engineering (CIBSE) - Experimental Software Engineering Track (ESELAW)*.

CC BY

Ma, L., & Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC bioinformatics*, *18*(1), 1-18.

Maione, C. (2020). *Balanceamento de dados com base em oversampling em dados transformados*. 2020. 135 f. Tese (Doutorado em Ciência da Computação em Rede) - Universidade Federal de Goiás, Goiânia.

Mandić, V., Basili, V., Harjumaa, L., Oivo, M., & Markkula, J. (2010, September). Utilizing GQM+ Strategies for business value analysis: An approach for evaluating business goals. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 1-10).

Martin, R. C. (2002). *Agile software development: principles, patterns, and practices*. Prentice Hall.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., ... & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

Olsson, H. H., & Bosch, J. (2014). The HYPEX model: from opinions to data-driven software development. In *Continuous software engineering* (pp. 155-164). Springer, Cham.

Orne, M. T. (1962). *Sobre a psicologia social da experiência psicológica: Com referência particular para exigir características e suas implicações*.

Pinto, P. (2015). *Introdução à Análise Estatística-Vol 2* (Vol. 2). Sílabas & Desafios.

Rodríguez, P., Haghighatkhah, A., Lwakatare, L. E., Teppola, S., Suomalainen, T., Eskeli, J., ... & Oivo, M. (2017). Continuous deployment of software intensive products and services: A systematic mapping study. *Journal of Systems and Software*, *123*, 263-291.

Roy, R. K. (2001). *Design of experiments using the Taguchi approach: 16 steps to product and process improvement*. John Wiley & Sons.

Santos, A. C. M., Colaço Junior, Methanias, & de Carvalho Andrade, E. (2020). Multimedia resources as a support for requirements engineering and software maintenance. In: *Journal of Software: Evolution and Process*. DOI: 10.1002/smr.2327.

Santos, B. S., Junior, M. C., & de Souza, J. G. (2018, June). An Experimental Evaluation of the NeuroMessenger: A Collaborative Tool to Improve the Empathy of Text Interactions. In *2018 IEEE Symposium on Computers and Communications (ISCC)* (pp. 00573-00579). IEEE.

SAS. (2005). *Semma data mining methodology*. http://www.sas.com/technologies/analytics/datamining/miner/semma.html.

Schäfer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018, November). 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)* (pp. 190-195). IEEE.

Sedkaoui, S. (2018). Statistical and Computational Needs for Big Data Challenges. In *Big Data Analytics in HIV/AIDS Research* (pp. 21-53). IGI Global.

Sjøberg, D. I., Hannay, J. E., Hansen, O., Kampenes, V. B., Karahasanovic, A., Liborg, N. K., & Rekdal, A. C. (2005). A survey of controlled experiments in software engineering. *IEEE transactions on software engineering*, *31*(9), 733-753.

Sharma, S., Osei-Bryson, K. M., & Kasper, G. M. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. In *Expert Systems with Applications*, *39*(13), 11335-11348.

Singh, B. (2016). *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. *Bangalore* Vol. 11, Ed. 2

CC BY

Sculley, D., Snoek, J., Rahimi, A., Wiltschko, A. (2018). Winner's curse? On pace, progress, and empirical rigor. *In Proceedings of the 6th International Conference on Learning Representations, Workshop Track*.

Svatá, V. (2019). COBIT 2019: Should We Care? *9th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 329-332.

Vasconcelos, N., Júnior, M. C., Almeida, T., & da Silva, V. M. (2019). Comparative Analysis of Data Mining Algorithms Applied to the Context of School Dropout. In *FedCSIS (Communication Papers)* (pp. 3-10).

Yin, R. (2015). *Estudo de Caso*: *Planejamento e Métodos*. 5. ed., Bookman.

CC BY