Analyze, Sense, Preprocess, Predict, Implement, and Deploy (ASPPID): An Incremental Methodology based on Data Analytics for Cost-Efficiently Monitoring the Industry 4.0

Jesus Para[a], Javier Del Ser[b,c,d,*], Antonio J. Nebro[e], Urko Zurutuza[a], and Francisco Herrera[f]

[a]*Mondragon University, 20500 Arrasate-Mondragon, Spain.*
[b]*TECNALIA, 48160 Derio, Spain.*
[c]*University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain.*
[d]*Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Spain.*
[e]*University of Málaga, 29071 Málaga, Spain.*
[f]*University of Granada, 18071 Granada, Spain.*

Abstract

Industry 4.0 is revolutionizing decision making processes within the manufacturing industry. Among the technological portfolio enabling this revolution, the late literature has been rich in what regards to the potential of data analytics for improving the production cycle at different stages, from resource provisioning to planning, delivery and storage. However, such a promising role of data analytics has been so far explored without a proper, quantitative inspection of the cost-improvement trade-off, nor has the process of acquiring sensors and extracting valuable information from their captured data formalized in a chain of methodological steps. This paper introduces the Analyze, Sense, Preprocess, Predict, Implement and Deploy (ASPPID) methodology, an iterative decisional workflow that spans from the acquisition of sensing equipment to the quantitative assessment of their contribution to enhance the production step under focus. This methodology aims at helping improvement teams to make informed decisions about which parts of the process needs to be sensed and how to exploit this information towards a verifiable improvement of the production cycle. The implementation of this methodology is exemplified in a real use case focused on the automotive in-

*Corresponding author: javier.delser@tecnalia.com (Prof. Javier Del Ser). TECNALIA. P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain. Tl: +34 946 430 50. Fax: +34 901 760 009.

dustry, where the detection of defects in an annealing process can be modeled as a classification problem with a highly imbalanced dataset. The results obtained after several iterations of proposed ASPPID methodology show that the scrap ratio is reduced by sensing the correct part of the process at reduced investment costs, thus highlighting the crucial role of data science in the business chain of manufacturing plants.

Keywords: Industry 4.0, Methodological Data Analytics, Process Monitoring, Cost Efficiency, Imbalanced Learning

1. Introduction

Since its first use during the Hannover Fair in 2014 [1], the so-called Industry 4.0 paradigm has drifted the interest of the community towards technologies aimed at providing intelligence to industrial process. Industry 4.0 is widely conceived as the fourth industrial revolution following the introduction of steam engines, chain production and process automation. As such, this concept refers to the design and development of *intelligent* factories [2], where traditional manufacturing processes are evolved by collecting large amounts of operational data coming from sensors and other information sources. This massive deployment of sensing equipment throughout the plant empowers multi-criteria decision making based on data analytics, with a target on increasing the efficiency of the manufacturing plants at different levels, from stocking to production and delivery. Not in vain this challenging paradigm has been recently identified as one of the main catalysts of the worldwide economical recovery after the last recession [3].

In order to remain competitive and increase the value chain of their productive processes, manufacturing industries have included the adaptation to this new productive model within their technology roadmap, embracing the inclusion of new technologies and competences within their processes and staff that had been traditionally linked to other sectors. The change has been particularly disruptive in regards to the monitoring of the whole production and delivery chain, with new technological enablers such as Internet of Things (IoT [4]), cybersecurity, augmented reality and collaborative robotics, among many others. However, there is common consensus around the crucial role taken by data analytics (either by itself or as a constituent, core part of Big Data) when providing the intelligence sought for the process at hand. Examples abound in the literature underpinning and evincing the gains yielded by the application of predictive models and optimization algorithms in different industrial problems, such as predictive maintenance or quality assessment (see e.g. [5, 6, 7] and references therein).

Unfortunately, the ever-growing need for integrating sensors and computing technologies for data collection, storage and analysis undergoes several issues in practice. Although manufacturing industries allocate big investments for installing advanced sensing equipment in their processes and machinery, purchases are often biased by the potential of the equipment itself to capture and store as many diverse parameters of the monitored asset as possible, disregarding their inherent utility and contribution for subsequent data modeling stages. While new positions appear within the staff of the majority industries (such as experts in robotics, data architects and data scientists [8]), it is often the case that these roles do not get involved within the purchase decision cycle, thus bounding their contribution to a *best-effort* over data captured by such devices once they have been purchased. This lack of involvement of data-based roles in managerial matters implies significant over-expenditures and a high degree of uncertainty with regards to the potential of data analytics for the industry [9, 10].

This noted fact becomes even more concerning when jointly assessed with the traditional client-provider relational model dominating most of the purchasing models within the small/mid-sized industries. Cost implications of a certain purchase must be evaluated not only from the purely economical perspective (namely, the rated balance between investment and depreciation of the purchased equipment, along with extra expenditures for ancillary tools and maintenance), but also with regards to the *value* of the captured data, the late referring to the contribution of the information captured by the purchased sensor to the data modeling task for which it was acquired. Such a contribution, when quantified in terms of predictive performance, can be translated to economical terms so as to reflect the competitive worth of the investment. However, studies dealing with the cost efficiency of data analysis within the industrial ecosystem are surprisingly scarce, despite the intuitively central role that this technology should play in manufacturing companies towards their convergence to the Industry 4.0 paradigm [11].

In this paper we postulate that all mechanisms and decision processes aimed at the provision of data-based intelligence to the manufacturing cycle should gravitate on the value of the data, quantification that can be only reliably undertaken by staff close to the inference of knowledge from the captured information. As will be exposed in the next subsection, most methodologies used to date in industrial decisional flows aim at improving the production efficiency of the plant by addressing technical aspects that regard data as a byproduct of the process that could be exploited if available, rather than considering this asset as a central matter for decision. In the next section we provide an overview of such methodologies, which helps contextualizing the contribution of this work.

1.1. Related Work

In order to face an ever-growing competitiveness in their respective market, industries have been traditionally forced to be constantly improving their processes. Several methodologies can be adopted for this purpose, among which SPC (Statistical Process Control), TQM (Total Quality Management), LEAN and 6SIGMA can be regarded as the most widely adopted ones:

- SPC [12] was the first methodology to embrace statistical control techniques in the early XX century, used years later by the United States in the World War II to improve the weapons manufacturing process. SPC relies on the use of control graphs where important variables are monitored, to evidence changes in the normal behavior of production processes, allowing operators to act and correct the drift of the processes. More recently insightful reviews have been contributed to the community, such as [13], where the application and experiences of SPC in the food industry is surveyed over 30 years. Most of these reports conclude that the general motivation beneath the adoption of SPC is to reduce defects production, but that some barriers withstand, such as the resistance to change of the personnel involved and the lack of sufficient statistical knowledge. The study in [14] focuses on quality improvement and cost effectiveness. Benefits of the application of SPC include troubleshooting and diagnosis in shorter time frames at reduced production costs. However, SPC undergoes practical issues in its implementation, as operators must be trained beforehand and new documentation must be produced, requiring an workload excess for its preparation.

- TQM [15] is a management strategy aimed at creating quality awareness in all organizational processes. TQM has been widely used in manufacturing, education, government and service industries. It is referred to as *total* because it is concerned with the organization of the company as a whole, including its staff. TQM comprises three general paradigms: 1) management, with specific compounding steps and goals such as planning, organization, control, and leadership; 2) total, standing for the coverage of the methodology all throughout the organization; and 3) quality, denoting the quantitativeness of the methodology by means of usual indicator definitions and all derivatives thereafter. The community has reported several success histories in regards to the particularization of TQM to several industrial sectors; to give two examples, in [16] the TQM methodology is applied to a hotel in Greece, identifying as a result influential TQM factors such as the quality practices of the top management board, strategic quality planning, employee quality management, customer focus and employee knowledge and education. Likewise, [17] studies, through TQM, the impact of Organizational Culture in small and medium Indian automotion com-

panies. Principal factors were discovered to be openness, confrontation, trust, authenticity, proaction, autonomy, collaboration and experimentation.

- LEAN [18] originates directly from the Toyota production system developed between 1950 and 1980. After producing a large batch of cylinder heads and assembling them in the corresponding engines, Toyota's production engineers finally found out that the required power was not achieved because the produced cylinder heads were defective. The enormous costs caused by this problem taught Toyoda one of the lessons that have been maintained in the company's culture: the acquisition and verification of the quality of the components at each stage of the production process, before moving on to the next stage, conforming the so-called Jidoka concept. This was the first motivation for creating a new methodology to eliminate waste, improve quality and reduce production times and costs, based on a set of target operational areas: overproduction, dwell time, transport, excessive procedures, inventory, movements and defects.

- 6SIGMA [19] is defined as a business process that allows companies to improve their final results. Originally created by Motorola in the 80's and consolidated by General Electric, 6SIGMA is a methodology that uses rigorously measured and analyzed data to identify causes of a problem and ways to eliminate them, generating greater customer satisfaction and substantial financial savings. It is supported by statistical and analytical tools and proposes the development of dynamic work groups, working with data in its search for the root cause of the problem studied. This is accomplished by a set of 5 steps: Define, Measure, Analyze, Improve and Control. Recent experiences with 6SIGMA abound: Shokri *et al* in [20] reported a good practice of the application of 6SIGMA to reduce defects in an airbag manufacturing plant, obtaining a scrap reduction on the first year of 0.63%. Similarly, [21] applied 6SIGMA to reduce the defect in the manufacture of glass containers for a cosmetics company, using to this end tools such as Pareto, Process Map and FMEA (Failure Mode Effects Analysis).

The methodologies examined above are nowadays applied to a vast number of diverse industries; however, the advent and proliferation of industrial sensors capable to record high volumes of data spurred the derivation of new methodologies based on the analysis of collected information. Some examples can be found in recent works such as [22], where real-time data is exploited to increase the efficiency of a production process; [23], which focuses on the use of data mining techniques to real industry cases; or [24], which proposes a manufacturing data mining methodology for industrial environments, which spans from the analysis of apparel industries manufacturing unit to the implementation and evaluation of the mining model.

1.2. Contribution

The motivation to introduce a new methodology is that all previous methodological approaches bring into focus data that are registered and then analyzed in a sequential manner, regarding the sensor providing such data as an legacy equipment for the data modeling and exploitation phase. The acquisition of sensors is usually performed under a *max-specs-min-cost* criterion, without taking into consideration the valuable knowledge that data analysis can shed over the suitability of one sensor or another for the predictive task in question. Sensor costs can be dramatically reduced not only by including roles related to data science within the industry personnel, but also by dilating their influence to early stages of the decision workflow, specially those where the selection of sensing equipment is discussed. There resides indeed the novelty of the Analyze, Sense, Preprocess, Predict, Implement, and Deploy (ASPPID) methodology proposed in this paper, which can be broken down in several novel aspects as summarized below:

1. A formal mathematical statement of the addressed cost-efficiency problem, which evinces the interplay between cost and benefit in data-intensive industrial processes, and motivates the iterative nature of the proposed methodology.

2. A data-centered decision workflow, which postulates the data scientist as a member of the core decision team in all decisions made around data-sensitive industrial processes, and a thorough description of the steps involved at each stage of the workflow.

3. An empirical validation of the proposed methodology when adopted to reduce the number of defects in a real aluminum injection plant from the Basque Country (Northern Spain). The lack of previous references in the literature dealing with this brand new injection technology makes the scenario specially suitable for ASPPID, given the high uncertainty around which sensor(s) to use towards decreasing the scrap ratio and the complexity of learning from the highly imbalanced dataset registered by the sensor(s). Evidences will be given on the valuable contribution of the data scientist to the ASPPID methodology.

ASSPID methodogoly has been implemented on a real case study at Edertek, the R&D center of Fagor Ederlan, a tier1 automotive manufacturer, to predict and reduce defective parts in their aluminum injection experimental line, reducing the costs of sensor the process and getting value from data analysis of it, with the particularity that the data response, quality part, is imbalance, with an imbalance rate of 7.69.

So, this paper is structured as follows: Section 2 formulates the optimization problem that models the decision problem tackled by ASPPID, whereas Section

3 and subsections therein details the fundamentals of the proposed methodology. The aforementioned real use case is presented in Section 4 along with a discussion on the obtained results, and finally Section 5 concludes the paper and outlines future research lines springing from this work.

2. Problem Formulation and Rationale

For a better understanding of the motivation for our proposed methodology we herein formulate the decision workflow in which it is framed as an optimization problem. This problem is driven by the balance between the costs derived from the acquisition and installation of the sensing equipment, the collection and analysis of the captured data and the benefits derived from the results of the analysis.

To formally pose this problem, we first mathematically model the cost incurred by acquiring and installing a sensor equipment s from a portfolio of S possible choices $\mathcal{S} \doteq \{s_1, \ldots, s_S\}$ aimed at capturing data from a machine towards a certain predictive modeling task (e.g. energy consumption characterization and minimization, predictive maintenance or defect detection). Let the net price of the sensor(s) be denoted as $C_{sens}(s)$ (in monetary units), to which installation, calibration (both fixed quantities, $C_{cal}(s) + C_{inst}(s)$, including in-machine adjustments and workforce) and maintenance (recurrently imputed cost $C_{mnt}(s)$ with periodicity $T_{mnt}(s)$ [time units]) should be added for a fully operational sensing deployment. Once installed, data must be retrieved from the machinery and stored for its further analysis (either in the Manufacturing Execution Server – MES – or the Cloud). Such a collection should span a minimum time frame $T_{col}^{min}(s)$ so as to provide predictive models with enough data substrate for characterizing the pattern between the collected information and the variable to be predicted (following the previous examples, defect or consumed energy). This minimum time duration depends not only on the sampling specifications of the deployed sensors, but also on the stationarity of the data and the timing requirements of the predictive modeling task under consideration.

In any case, additional costs arise during the data collection period, such as the storage of data $C_{stor}(s, t)$ (which depends on both the sensor – through their data sampling specifications – and the time t for which it is utilized). Depending on the requirements of the use case at hand it might be the case that the machine at hand would be kept aside from the production line during the data capturing period for experimentation purposes, causing further costs due to e.g. cease of production and expenditures in feedstock and raw materials used to build samples for experimentation. All such possible costs will be denoted as $C_{other}(s, t)$, which not only depends on the sensor and time span of the experimentation themselves, but also on the net contribution $R(s, t)$ (in monetary units per unit of time) of the halted

machine to the production of the plant. For instance, should a manufacturing chain in a parallel production chain be completely stopped and left inoperative during the trial, $R(s,t)$ would account for the aggregate difference between the net income of all products that would have been processed by the machine during $T_{col}^{min}(s)$ and the cost of raw material not acquired/consumed during the trial[1].

Once enough data have been retrieved from the sensed machinery, the working flow proceeds by performing different data mining steps aimed at building a model based on the collected data. As has been previously argued this model construction phase aims at achieving a productive gain that can be translated to economical terms. For instance, a predictive model aimed at detecting defective products in a manufacturing machine could be evaluated in terms of its predictive score (e.g. ratio of true positives or recall). In this case, the recall of the developed predictive model could be combined with the scrap ratio of the machine under analysis and the cost of scrap samples to yield a economical benefit $B(s,t)$ [monetary units] associated to the model ergo the acquisition of sensor s and measured over a certain time horizon t after the trial has been completed. The development of the model itself plus the prior preprocessing of the captured data would add an additional time gap $T_{model}(s)$ to the experimentation phase.

Before delving into further derivations, it is of paramount relevance for the scope of this study to note that the estimated economical benefit $B(s,t)$ cannot be quantified beforehand, as the predictive value for the information collected by sensor s is uncertain until captured and fed to the model for its development. This uncertainty should have profound practical implications when deciding which sensor to acquire, as well as the sales and purchase scheme. At this point we mathematically pose two different purchase options:

A. Traditional sales and purchase scheme, by which the sensor supplier receives a monetary quantity $C_{sens}(s)$ for the selected sensor, plus additional payments $C_{cal}(s)$, $C_{inst}(s)$ and $C_{mnt}(s)$ should it also offer maintenance, installation and calibration services.

B. Sensor rental scheme, in which a time-dependent cost $C_{rent}(s,t) \ll C_{sens}(s)$ is negotiated and agreed between manufacturer and supplier, which is billed and paid during the total time of the trial, namely $T_{xp}(s) \doteq T_{col}^{min}(s) + T_{model}(s)$.

Intuitively, a rental scheme would incur a penalty cost $C_{rent}(s, T_{xp}(s))$, which adds to the final cost of the trial if the predictive value of the data captured by

[1]This computation implicitly assumes that the sensed machine is not productive at all during the trial. Otherwise a mixed cost model should be instead adopted so as to reflect in $R(s,t)$ the interplay between material savings and productivity losses.

the sensor is good enough to make a difference in regards to the balance between overall costs and the benefit $B(s,t)$ of the developed model. Based on the above notation, the overall cost of the trial with sensor s when a traditional sales and purchase scheme (A in the above list) is adopted can be expressed as

$$C_A(s) = C_{sens}(s) + C_{cal}(s) + C_{inst}(s) + C_{mnt}(s) \left\lfloor \frac{T_{xp}(s)}{T_{mnt}(s)} \right\rfloor \quad (1)$$

$$+ C_{stor}(s, T_{xp}(s)) + C_{other}(s, T_{xp}(s)) + R(s, T_{xp}(s)), \quad (2)$$

where (1) refers all costs derived to the deployment of the sensor equipment, and (2) collects all expenditures and production losses (if any) during the data collection and model construction and assessment. In case the sales and purchase model is implemented on a rental basis (correspondingly, B in the list) during the time frame $T_{xp}(s)$ of the trial, the overall cost is given by

$$\begin{aligned} &C_B(s, B_{th}(t)) = \\ &C_{rent}(s, T_{xp}(s)) + C_{stor}(s, T_{xp}(s)) + C_{other}(s, T_{xp}(s)) + R(s, T_{xp}(s)) \\ &+ \mathbb{I}\left(B(s,t) \geq B_{th}(t)\right) \left[C_{sens}(s) + C_{cal}(s) + C_{inst}(s) + C_{mnt}(s)\right], \end{aligned} \quad (3)$$

where $\mathbb{I}(\cdot)$ is an auxiliary binary function taking value 1 if its argument is true and 0 otherwise. It can be noted that $C_A(s) \leq C_B(s)$ whenever the final decision is to acquire the sensor and exploit its information within the production line, decision that is made if the obtained benefit $B(s,t)$ from the data collection and modeling phases using sensor s exceeds an expected minimum threshold $B_{th}(t)$ [monetary units per unit of time] agreed by the team before starting the trial. We compile the above formulae as $C(s, m, B_{th}(t)) \in \{C_A(s), C_B(s)\}$, which denotes the cost of the trial with sensor s depending on its purchase mode $m \in \{A, B\}$.

As argued before, the unpredictability of the value of $B(s,t)$ for any t before acquiring sensor s imposes that it cannot be adopted as a metric for selecting a sensor from $\mathcal{S}$ before its acquisition. Therefore, a prior analysis of the available choices for sensing is crucial, yet not always possible, specially in avant-garde manufacturing processes as the one later tackled in this manuscript. The lack of previously reported good sensing practices poses an unavoidable risk when selecting the sensor to be tested.

Based on the above rationale, methodologies as the one proposed in this paper should aim at maximizing the return of investment when a series of trials – each testing a sensor $s \in \mathcal{S}$ – is performed without any a priori knowledge nor estimation of the value of the captured data by each of the choices. Such trials are assumed to occur sequentially in time, such that $s_k \in \mathcal{S}$ denotes the sensor selected in trial k. At every trial k a two decisions must be made by the team in

regards to 1) the acquisition model m_k of selected sensor s_k (A or B, if available upon agreement with the supplier); and 2) a target value $B_{th}(t)$ of the estimated economical benefit of the developed model based on the retrieved data by the purchased sensors over the series of trials. It should be clear that if sensor s_k is selected at trial k, purchased ($m_k = A$) and deployed on the machine/process at hand, it might occur that the captured data and the modeling phase do not achieve the aforementioned minimum value, i.e. $B(s_k, t) < B_{th}(t)\ \forall t$. However, since sensor s_k was purchased, sensors selected at subsequent trials may benefit from the information captured by s_k, potentially – yet uncertainly – furnishing a better model performance than if sensor s_{k+1} was utilized in isolation. On the contrary, benefit yielded by the model constructed from s_k could undergo a collinearity between variables collected at different trials, leading to a lack of predictive relevance of the information provided by the new sensor. This implicitly unveils that the choice of sensors is order-sensitive.

By introducing a slight abuse in notation, the accumulated benefits after K successive trials can be computed as $B(\mathbf{s}_K, \mathbf{m}_K, t)$, where $\mathbf{s}_K = [s_1, s_2, \ldots, s_K]$ and $\mathbf{m}_K = [m_1, m_2, \ldots, m_K]$ (with $m_k \in \{A, B\}$) denote the selected sensors at trials $\{1, \ldots, K\}$ and their purchase mode, respectively. A similar redefinition of the overall costs of the trials can be made as

$$C(\mathbf{s}_K, \mathbf{m}_K, B_{th}(t)) = \sum_{k=1}^{K} C(s_k, m_k, B_{th}(t)), \tag{4}$$

which must be kept below an upper bound C_{max} that represents the budget allocated by the manufacturing plant to perform trials. A purchase and trial methodology should pursue the optimal number K^* and sequence $\mathbf{s}_K^*$ of selected sensors and their purchasing policy $\mathbf{m}_K^*$ so that the sought benefit from the developed model $B_{th}(t)$ is achieved at a minimum overall cost. Mathematically,

$$[K^*, \mathbf{s}_K^*, \mathbf{m}_K^*] = \underset{K, \mathbf{s}_K, \mathbf{m}_K}{\arg\min}\, C(\mathbf{s}_K, \mathbf{m}_K, B_{th}(t)), \tag{5}$$

$$\text{subject to} \quad C(\mathbf{s}_k, \mathbf{m}_k, B_{th}(t)) \leq C_{max},\ \forall k \in \{1, \ldots, K\}, \tag{6}$$

$$B(\mathbf{s}_K, \mathbf{m}_K, t) \geq B_{th}(t), \tag{7}$$

$$m_k \in \{A, B\} \cap \mathcal{M}_{sup}(s), \tag{8}$$

$$s_k \neq s_{k'} \text{ if } k \neq k'\ \forall k, k' \in \{1, \ldots, K\}, \tag{9}$$

where (6) denotes that the overall cost is constrained, (8) indicates that the choice of purchase mode should also take into account the selling policy $\mathcal{M}_{sup}(s) \in \{A, \{A, B\}, B\}$ of the sensor supplier; and (7) imposes that the target benefit should be achieved after the K trials. The process to implement a hypothetic case

comprising $K = 3$ trials is exemplified in Figure 1, where it is important to notice the impacts of the decisions made after every trial in the cost and the accumulated benefit along the process.

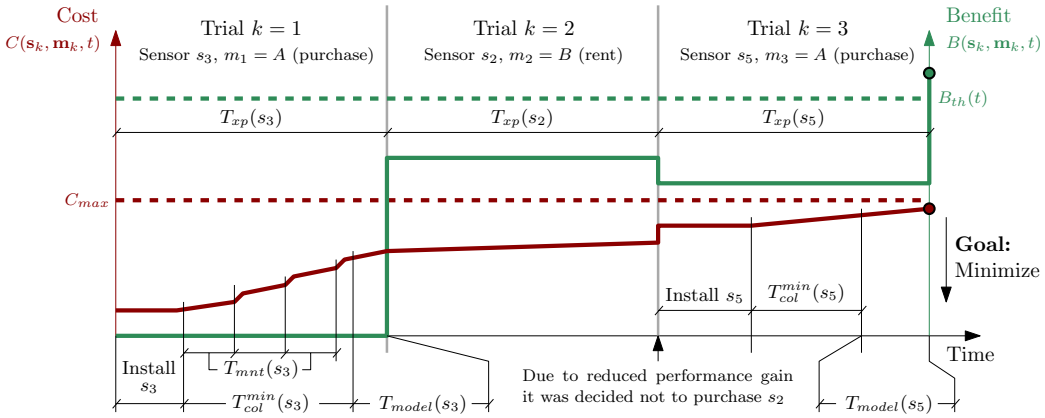


Figure 1: Example of $K = 3$ sensor trials driven by a methodology proposing $\mathbf{s}_K = [s_3, s_2, s_5]$ and $\mathbf{s}_K = [A, B, A]$ for a given predictive modeling task (e.g. defect detection). A double vertical axis has been used to reflect benefits and costs. Due to the uncertainty of the team over the utility of s_2 for the task at hand it was decided to rent it for the duration of the trial, saving costs after its lack of predictive relevance was confirmed.

It should be clear that decisions made on m_k along the trials should be made not only depending on the purchasing options allowed by the supplier, but also on the prior knowledge (if any) about the potential of the sensor to provide value for achieving the goal $B_{th}(t)$. Should this a priori estimation be reliable enough, purchase mode A would incur less costs than B, but in practice this entails a risk of very difficult clearance. The ASPPID methodology proposed in this work is aimed at lowering this risk by including the data scientist all through the decision loop, not only in the data modeling stage. The adoption of our proposed methodology not only reduces the aforementioned risk, but also establishes realistically the expected benefits from the data model by matching a business-related goal $B_{th}(t)$ to data-based insights within an iterative process.

3. Proposed ASPPID Methodology

In order to face the problem posed in the previous section, a methodology is proposed to make the entire decision process gravitate on cost and data-based performance of the considered sensing equipment. The methodology is detailed in what follows.

3.1. General Description

ASPPID is an iterative methodology based on data analysis that pursues a cost-efficient improvement of a certain productive process within manufacturing plants by involving a data scientist from the very beginning of the methodology, i.e. from the moment when economical decisions are made on the acquisition model of sensor equipments to the moment at which the developed model is deployed on the production line. Traditionally, data scientists are part of the improvement team when sensors are already installed on the machines, being in charge of exploiting the information collected by them. In this way, data scientists are restricted to mine the information that the installed sensors are recording. In practice this strategy leads to very frequent errors: already installed sensor(s) may not be recording relevant variables and parameters for the targeted mission of the model, which ultimately jeopardizes the discovery of sufficiently explanatory models to achieve the imposed benefit goal $B_{th}(t)$. As anticipated before, the reason for this mismatch between the information and the performance of the sensor is that in data science, the predictive value of data not only has to do with its cardinality, but also with its relevance: it is often the case when staff responsible for acquiring industrial equipment hinge their decision on the power of sensors and price, sinking in excess of information, incorporating redundant sensors or sensors that register variables with no influence on the predictive model, leading to additional costs.

Figure 2: Schematic diagram illustrating the iterative workflow of the ASPPID methodology.

In ASPPID the data scientist plays a central role within the organization, participating not only in the development of advanced models for data analytics, but

also in the choice and validation of acquired sensors, after in-depth analysis of previous practices reported in the literature and taking into account the cost of the sensors $C(s, m, B_{th}(s))$.

3.2. ASPPID Steps

Specifically ASPPID can be split into 6 fundamental steps: 5 comprising an iterative cycle aimed at fulfilling the project goal(s); and a final step performed only once after the target has been met and the developed model validated. Such steps are depicted in Figures 2 and described in the following subsections:

3.2.1. Analyze

Fueled by the momentum gained by the Industry 4.0 paradigm, it is usual to find organizations machinery with a myriad of sensors capable of recording huge amounts of data captured from different stages of the productive process. In practice this information is almost never exploited and, if used anytime, much of the captured information is highly redundant or does not add any value to the production efficiency of the plant. This is a real problem for organizations: they have large collections of data, but they do not know how to exploit them nor have they analyzed the cost overrun as the result of purchasing decisions.

ASPPID faces this problem by defining a methodology particularly suitable for monitoring industrial machines and processes with sensors. An in-depth analysis of the target problem or task that the organization aims to solve should be the first step before sensing. If this analysis is not properly done or should it steer towards mistaken goals, the entire methodology could give rise to considerably higher costs, hence this step is crucial in our proposed methodology. In this regard, the entire set of decisions to be made has to be driven by a concern, by something to improve quantitatively. This target must be established by the plant/business management board, but has to be achievable, measurable and verifiable. Once the goal has been set, a so-called *improvement* team must be arranged to carry out all the trials needed to meet the target.

3.2.1.1. Contract. ASPPID dictates that trials cannot be started unless the improvement team agrees with the terms and goals imposed by the management board. This agreement is formalized in a *contract*, which reflects and certifies the confidence and support of the management of the plant or business to which the improvement project is destined, feeling it as its own. This contract compels both parts to fulfill the responsibilities posed in its clauses: on one hand, the management board will use it as tool to trace and verify the results obtained by the improvement team within the timing plan; on the other hand, the improvement team will be allocated to an internal project having agreed in their terms, which

is a guarantee of involvement, commitment and identification with its goals, far beyond the usual implications of a project (e.g. work labor, company support for investment). The evolution of all the improvement teams on the plant with respect to their signed contracts will be analyzed at the end of the working year to detect common causes of deviation, unveil management problems in the organization and potential needs of prevailing improvement teams. The contract should have *at least* the following elements:

- Title: it must be descriptive, concise and easy to understand for both client (in ASPPID this role is taken by the management board) and improvement team.
- Description of the problem for which the improvement team is hired. This part must be filled in by the customer and reviewed by the improvement team.
- Objective of the project $B_{th}(t)$, which must be described in detail and quantified in monetary terms as an expected target benefit of the model to be developed within the project. It is of utmost importance to specify the units in which the benefit is measured, either as a rate over time (as in the problem statement in Section 2) or as an alternative measure, e.g. benefit per product (such as monetary units per produced unit). In all cases this goal will be used to decide the success/failure of the project and its closure. It is important to remark that this objective can be reviewed along the iterative ASPPID workflow, yet any change must be reviewed and agreed by both parties.
- Maximum costs of the project C_{max}, which establishes an upper bound on the economic costs associated to the implementation of ASPPID including workforce, sensor purchases, ancillary equipment and external services. Since costs can be revised between trials and for transparency and traceability it is advisable to break down the maximum cost of the trials into the aforementioned concepts.
- Team: the definition of the improvement team should not only determine their members, but also their hours of dedication. Team leadership will be shared by the project manager and the leading data scientist. In constant communication to each other, the project manager is responsible for managing the infrastructure part, purchasing equipment and contacting sensor suppliers, whereas the leading data scientist will be responsible for data analysis. The rest of the improvement team must be composed by personnel aware of the process from a theoretical/technical point of view, as well as by maintenance engineering personnel and/or data architects. The role of them is crucial to determined which parts of the process they suspect to be, under their experience, most promisingly originating the problem to be solved. Data architects and/or maintenance engineers will ensure that sensors are installed and recording data properly.

- Timing plan, including estimated duration of the planned trials $T_{xp}(s)\ \forall s \in \mathcal{S}$, intermediate milestones and reserved work hours for each member of the improvement team (see Figure 3 for an exemplifying schedule of the ASPPID phases for the shown in Figure 1). Creating realistic timing plans helps assessing the progress of the project, and keeps the team motivated. It is important that the customer knows the real status of the project, so presentations will be performed periodically (at an agreed period) and once every trial k is over. When detected, any punctual deviation from the original planning must be thoroughly documented and recorded, and countermeasures and corrective actions must be proposed to the client in duly time, i.e. in the presentation held closest to the detection of the deviation. The client and the project leader will evaluate the cause of the deviation, resolving it (if possible).

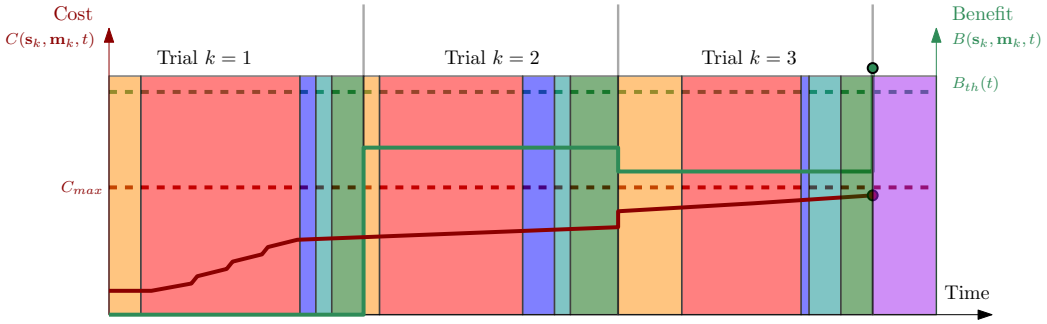


Figure 3: Indicative duration of each phase of ASPPID in the example of Figure 1.

After the contract has been completed, it must be signed by the parts: the customer (plant/business directors) and all team members. A copy of the signed contract must be sent to all managers of the team members, so that they have a record of the hours that their collaborators will devote to the ASPPID project. Once the project has been signed, a first team meeting must be held, where team leaders explain in detail to the rest of the team why the team has been formed, and the role that all the members will play on the team. The objectives, timing and dedication must be clear to all attendants.

3.2.1.2. Brainstorming. Once this step has been taken, the team can begin to approach the project. The first step is to do a brainstorming exercise to determine which causes may be affecting the problem. A brainstorming exercise is a group creativity technique, where all participants must freely contribute ideas about the causes that affect a given problem. It is very important for this dynamic to be effective, that all participants feel free from expressing the ideas that they think can have an influence. This is why team leaders must communicate the rules of

participation to the rest of the team. No opinion can be underestimated nor any person held back. The ranks of the organization must be forgotten. If the presence of a responsible can hinder the correct development of the dynamics, then this exercise should be carried out at different times, isolating the conflicting parts.

To efficiently implement this brainstorming session, ASPPID embraces the so-called Ishikawa diagram [25] – also known as the fish-bone diagram because of its shape – which consists of an axis on which the problem to be tackled resides. From this axis comes the *thorns*, grouping together the main aspects related to the problem: equipment, process, people, materials, environment and management, as depicted in Figure 4. The objective of the Ishikawa procedure is that the team considers all possible causes that may affect the problem at hand, not only those in which each individual member probably focuses biased by his/her role in the plant. When all ideas are contributed, those that are believed to possess the highest influence in the problem are selected. Hypotheses will be established with them. Other variables deemed irrelevant / less relevant are initially discarded, but can be taken again into account in forthcoming trials: due to the uncertainty of the selection of sensors and variables, what seemed to be the root cause and a relevant variable for the task/problem at hand does not often meet the expectations. Therefore, new hypotheses must be again generated by bringing back the Ishikawa methodology along with all previously discarded alternatives.

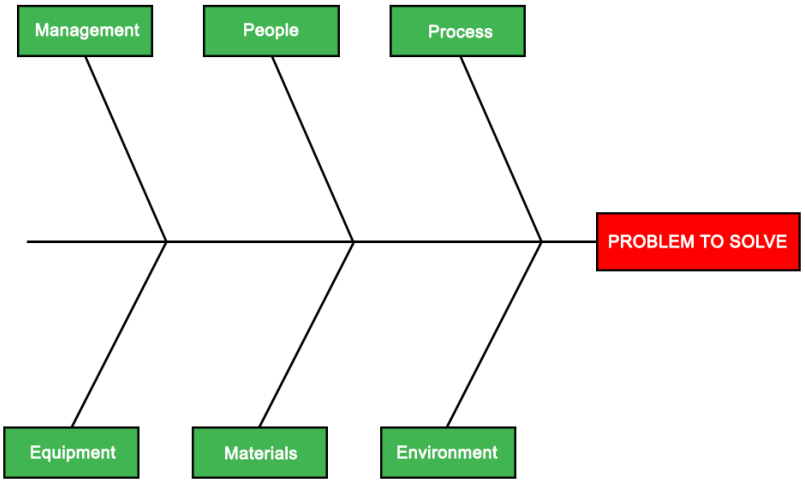


Figure 4: Ishikawa diagram utilized within the *analyze* phase of ASPPID.

At this point we highlight that the participation of the data scientist from the very beginning of the process (including the establishment and agreement of the project's goals) is unprecedented, and sheds certainty on the reachability of the

benefits expected by the rest of the team on the model to be developed. To this end the data scientist can contribute to the brainstorming session with literature on the data-based problem to be solved, drafty estimations based on reported good practices of similar setups, and an explanation of the practical implications of the scores by which the performance of the model to be developed. As a result the brainstorming session should produce a calendar of K trials where to test different monitored variables, and a qualitative yet informed indication of the level of confidence with respect to their utility for the addressed problem.

3.2.2. Sense

Once the brainstorming exercise has been completed, it is time to test the set of hypothesized variables at every trial. To do this, the project manager and the leading data scientist must inspect whether the variables are already being recorded within the plant, and if so, whether the sensor capturing such information is reliable, properly calibrated, in optimum operating conditions [26, 27] and capturing enough data for the needs of the subsequent modeling phase. To verify the latter, the sensor supplier must carry out tests and a calibration plan to avoiding operational errors, repeatability and bias problems.

If the information is not being recorded, the team must gauge the acquisition of sensing equipment aimed at capturing the required variables. At this point ASPPID underpins the importance of performing this acquisition in a cost-efficient way, since purchasing sensors based only on their specifications can be counterproductive due to the unpredictability of the contribution of its captured data to the problem at hand. When the leading data scientist declares the utility of s_k at trial k as extremely uncertain (due to the lack of previous history with the sensor or the problem under analysis), the cost derivation in Section 2 suggests that renting the sensing equipment should be preferred over a traditional purchase, if possible (i.e. if the sensor supplier allows for this model in the acquisition of s_k, i.e. if $B \in \mathcal{M}_{supp}(s_k)$).

Sensors can be a major part of the cost of the project, so acquiring them only after validating if they are estimated to be truly effective is critical to reduce the overall cost $C(\mathbf{s}_K, \mathbf{m}_K, B_{th}(t))$ of the project. As described in Section 2, whenever there is a risk of lack of predictive power of the captured data, sensor renting is more cost-efficient than purchasing it from the very beginning of the trial. This is the reason why ASPPID imposes that before purchasing a sensor it must be proven reliably that its captured data can be expected to yield a quantitative gain towards achieving the project's goal. This a priori estimation must be done by the leading data scientist (and his/her team, if any) based on a prior literature study, interviews with plant personnel, and the study of reports with similar cases.

Once the process/machine has been sensed, it is time to collect and analyze

information of the process. To do this, the team must establish the minimum time period $T_{col}^{min}(s_k)$ for the sensor s_k to collect representative data of all the variability and casuistry of the monitored parameters of the current trial. This estimation can be performed by exploring several aspects of the scenario under analysis, such as the elements that cause the variability of the response variable to be predicted by the model, the variability pattern (e.g. seasonality) of the magnitudes from which data is retrieved, and other statistical features of the data alike.

3.2.3. Preprocess

Once data have been captured and stored, ASPPID proceeds by preprocessing it towards the construction of a model aimed at achieving the project's goal. This is an essential step in the data analysis process, as the extraction of valuable information from data not only depends on the models themselves, but also on a complete data preprocessing process including the following steps [28]:

- Cleansing consists of the removal of data observations that significantly differ from the rest of observations, prone to misfit the subsequent model.
- Data scaling, by which each recorded variable is scaled prior to the use of models that are sensitive to the range of values taken by their input variables (e.g. neural networks). A survey of scaling techniques can be found in [29].
- Imputation, by which data samples with missing values are removed or filled depending on their degree of incompleteness [30, 31].
- Noise identification, by which all instances that result in a loss of predictive model accuracy are removed and not used for model building [32].
- Data transformation, which combines, projects and/or summarizes the recorded data in order to make samples fulfill a priori assumptions of the consequent predictive model (e.g. linearity, normality).

Although the relevance of data preprocessing has been widely acknowledged by the research community, ASPPID reflects explicitly the need for undertaking this step within its cycle due to its high share of the modeling costs $T_{model}(s_k)$. When agreeing on the schedule and costs of the project, the management team must be made aware of the specific data preprocessing needs of the considered sensor portfolio and be warned of eventual over-costs due to this step. To minimize the potential budgetary impact of data preprocessing an informed data quality estimation report should be prepared by the team, if possible before contract negotiation and agreement.

3.2.4. Predict

After data preprocessing, the next step is to generate models that permit to achieve the benefit goal $B_{th}(t)$ established in the contract agreement and, ultimately, to enhance the performance of the monitored machine/process within the manufacturing plant. To this end the retrieved, already preprocessed dataset is used as the starting point of a data analysis phase. As per the majority of practical use cases foreseen for ASPPID, this phase will aim at the development of a *predictive* model capable of explaining the pattern between a set of monitored parameters and a target variable, the latter impacting on $B_{th}(t)$ either directly or indirectly (e.g. consumed power or defective output).

All in all, an exploratory data analysis must be first done [33], comprising techniques such as descriptive statistics, dimensionality reduction and clustering, among others. Once the exploratory analysis is done, the part of the improvement team in charge for data analysis and modeling will be in a better position to construct the planned data model(s); in this regard, the nature of the variable(s) to be predicted and the problem at hand will require different procedures for constructing the model. Intuitively, in industrial environments productive scenarios embracing ASPPID for improving their performance are based on sensors, whose collected information is used to train and build a predictive model for classification or regression, depending on the kind of target variables to be predicted. In industrial environments this family of scenarios often requires the adoption of techniques for class imbalance, feature selection/construction and model validation, since it is likely that the collected data lacks of enough positive examples for properly modeling the pattern between the monitored parameters and the target variable(s). This is the reason why filter, wrapper or embedded feature selection methods become essential in the *predict* phase of ASPPID, along with oversampling, under-sampling or cost-sensitive techniques for class balancing.

Methodologically speaking, during this phase the improvement team must bear in time the *interpretability* of the model towards easing its deployment should it perform satisfactorily with respect to the imposed target benefit $B_{th}(t)$. To this end, ASPPID fosters the consideration of rule mining techniques so as to infer interpretable rule sets from the developed models that can be implemented straightforward by the personnel of the plant upon its approval by the management team.

3.2.5. Implement

The *implement* phase is aimed at verifying that models generated on the previous phase are valid on production and therefore, can be deployed if they meet the imposed performance goal and eventually approved. Depending on the scenario and problem in hands, the implementation can imply 1) that the developed model is tested with new unseen data to assess its estimated generalization capabilities

(as in e.g. a predictive model forecasting the power consumption of a certain machine); 2) and/or that the parameters configuring the monitored machine/process are varied to those optimized based on the insights provided by the data modeling phase. In both cases this phase must prove that models generated in the previous phase perform as expected during real production, so that an empirical guarantee of the improvement yielded by the developed model can be granted to the management board.

To this end, an implementation period must be determined. It should be long enough to allow checking the impact of all process components with a potential influence on the target variable (e.g. changes of tools, molds, etc), but short enough for the equipment tension to be maintained and costs reduced. It is necessary to record the start and end timestamps of this phase so as to compute the benefit $B(\mathbf{s}_k, \mathbf{m}_k, t)$ from the savings obtained during this time frame. Once this period is over, conclusions from the implementation phase must be fully documented, along with an estimation of the economical impact of the tested model, its comparison to the established goal $B_{th}(t)$ and, if deemed necessary, informed recommendations from the improvement team (specially, from the data scientist(s)) on an update of the goals and targets of the ASPPID project. This reported information will be gathered within a deliverable for the client's perusal, who will be called for a review meeting so as to close the k-th loop of the ASPPID project.

3.2.6. Deploy

In the aforementioned meeting a decision on the fulfillment of the project's objective is to be made in light of the reported information and conclusions from the current ASPPID iteration and the evolution of costs incurred within the project. Should the target $B_{th}(t)$ has been met by the developed model, a cost study of the extra cost required to deploy the model definitely within the production line is elaborated, including the purchase of sensors if they have been rented during previous trials. If the estimated overall costs for deployment are contained below C_{max}, the project is declared as successful. Otherwise, both the team and the management team must agree on either a relaxation of the cost constraint or a reversal of the ASPPID cycle to a previous trial, which is decided depending on the suitability of each alternative for the plant. For instance, if the gap between $B(\mathbf{s}_k, \mathbf{m}_k, t)$ and $B_{th}(t)$ is high, the management board could decide to accommodate extra funding for deploying the model; conversely, expenditures beyond C_{max} should be leveraged by assessing small benefit over $B_{th}(t)$ with the investment required to deploy the model.

A platform to keep documents compiling the knowledge acquired during the project is advisable, since many of the obtained data-based insights can be extrapolated to other scenarios dealing with products with equal or similar character-

istics. In these documents the best possible detail should be given as to identify the key conditions and aspects that make the obtained conclusions generalizable to other use cases. The responsible for recording this information is the project manager. Once the information is saved in the platform, a project closing meeting should be organized to crosscheck that the produced documentation is complete. It is also important that the rest of the organization visualizes its success, so a good practice is to install visible panels inside the plant with the most synoptic description possible of the steps taken along the ASPPID cycle.

4. Real Case Study of the Proposed Methodology

ASPPID was originally conceived and put to practice in the manufacturing industrial case of the Basque Country (northern Spain). This section elaborates on the details of its implementation, placing an emphasis on how the process evolved towards getting a cost-efficient operational gain in the production of the manufacturing line under study.

4.1. The Scenario: Edertek (Fagor Ederlan)

In particular the project was run over Edertek, a corporate research and development center of Fagor Ederlan. Fagor Ederlan is a supplier for the automotive industry specialized in chassis and power train applications in different technologies and materials. Nowadays Fagor Ederlan has production plants in Spain, Mexico, China and Slovakia, a worldwide presence for which the company is particularly in pursuing technological advances for production efficiency – through its center Edertek – that can be extrapolated and exploited in several albeit related scenarios. For confidentiality reasons, sensitive information will not be disclosed in the conclusions extracted from the application of ASPPID to this real use case, yet we consider this information of minimal relevance to illustrate how ASPPID helped this company.

The case of use aims at the prediction of defective parts in an aluminum injection line reproduced experimentally in Edertek. From a data analysis perspective, this predictive task can be modeled as a highly-imbalanced binary classification problem where the model to be developed classifies products as ACCEPTED (i.e. compliant with quality requirements) or REJECTED. Although the literature is rich in tackling this particular problem instance by means of predictive models, the aluminum injection is a brand new manufacturing technology for which no previous references are known. This particularity motivated further the participation of the leading data scientist of the company all over the decision workflow, proposing deep changes in the *modus operandi* that prevailed in the company before this use case was started. Those changes gave rise to the definition of an

early iterative procedure to relocate data analysis at the core of the project, which matured towards the herein proposed ASPPID methodology.

4.2. Data mining Methodology and Metrics used on implementation

The objective of the ASPPID project was agreed to be the design and deployment of a process robust enough to manufacture parts with a sufficiently high quality rate for this first experimental phase.

As described on previous subsection, this project has been run on the experimental line of Edertek, a R&D center, to get a predictive analysis on the quality parts, and reduce the costs of sensor. This kind of process has usually a high imblance rate. In this case study, the IR is 7.68.

To deal with imbalance domain it is necessary to use imbalance data-mining techniques to make imbalance class distribution has no influence on model generation. To do this, over and under-sampling techniques has been tested: minority class oversampling, Synthetic Minority Over-sampling Technique (SMOTE, [36]), SMOTE with Tomek links removal [37], the borderline variant of SMOTE [38], Adaptive Synthetic Sampling (ADASYN, [39]), Edited Nearest Neighbours (ENN, [40]) and One-Sided Selection (OSS, [41]).

The most common metrics to validate classification problems is Acc (10). This metric is not appropriate on imbalanced domains, since it is not capable to distinguish between the number of correctly classified instances on the majority and minority class [34].

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

On Edertek all the manufactured parts pass an internal quality control, so metrics used were F1 score (13) and principally, Recall (12). Precision (11) is bypassed because of the fact that all parts pass a quality control, so the focus of the model is to be capable to create good models for majority class (ACCEPTED parts) and minority one (REJECTED parts).

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{(1+\beta)Recall * Preccision}{\beta^2 Recall + Precission} \tag{13}$$

AUC 14 was also calculated to demonstrate is not a good metric, as is influenced by the majority class.

$$AUC = \frac{1 + TPrate - FPrate}{2} \tag{14}$$

Following the ASPPID guidelines for data analysis in Subsection (3.2.4) a Random Forest (RF, [43]) classifier with 500 trees was used as the inner predictive model for choosing among class imbalance techniques, and to discover the features that had more influence on the response. Once the most influent variables are selected, a C4.5 Decision Tree model was utilized as the interpretable predictive model due to its simplicity and straightforward readability in terms of deployable rule sets.

Both, the over & undersampling thecniques + Random Forest and C4.5 tree, were runned in a 5×5-fold cross validation to avoid bias on modelling.

For costs controlling on sensors, the mathematical approach detailed on Section 2 is used.

The way on that the proccess was sensed and data analyzed are described on the next subsection, deploying ASPPID methodology.

4.3. Description of ASPPID Applied to the Use Case

In this case $K = 2$ iterations of ASPPID sufficed to achieve the imposed goal:

4.3.1. First Iteration ($k = 1$)

(1.a) Analyze: Once the target was defined quantitatively (translated from monetary units to a maximum Non-Quality Rate, NQR, equal to 8% that should be achieved by the project to meet the target) and after a budget was provisioned for the project, a multidisciplinary improvement team was arranged, with a project manager and a leading data scientist as its coordinators. The rest of the team was selected based on the technical knowledge of the process to be monitored, and a timing was established to develop the project. Once the contract was created and signed, a first team meeting was launched aimed at developing an Ishikawa exercise to conjecture which parameters may have an influence on the scrap ration of this new injection technology. To do this, a blackboard was used and the data scientist, leading the session,

explained the rules and coordinated the brainstorming through an Ishikawa diagram. As a result, a selection and prioritization of hypothesized factors were agreed by the team. For this first iteration, the following factors were selected:

- Furnace number that processes each injected part.
- Cycle time, i.e. time during which the part is processed in the furnace.
- Cavity number of the part within the mould.
- Mould number within the stock of moulds in the plant.
- X_t^1 [kept confidential], a time series representing a magnitude recorded during the furnace processing.

(1.b) Sense: once identified, the above factors must be sensed from the monitored process. To do this and given the high uncertainty with this injection technology, one of the principles of the ASPPID methodology, to spend the minimum amount of money on sensors, was adopted, avoiding to purchase equipment that would increase the cost of the project without any guaranteed reliability of their contribution to the project's goal, as described in Section 2. In this first iteration the team got advantage of sensing equipment that had previously been installed in another process within the plan and that was out of use by the time when the trial was made. Should this equipment be eventually relevant for the project, new sensing units would be purchased and installed in the monitored process. A period $T_{col}^{min}(s_1)$ equal to 2 months was decided in order to balance the trade-off between loses due to the non-productivity of the monitored process and the collection of enough data to capture the underlying variability of the process. A detailed cost breakdown was done to jointly evaluate all decisions impacting on the budget of the project (use of provisional sensors, loses during the *sense* period, etc.).

(1.c) Preprocess: once collected (at a rate of 1 sample per second in the case of X_t^1), data were preprocessed using one-hot encoding for the categorical variables (furnace, cavity and mould numbers), whereas the cycle time was converted to its Z-score. As for the time-dependent variable X_t^1, a series of interviews with empirical findings from the personnel allowed extracting features of potential importance for defect detection from the time series recorded for each part, as shown[2] in Figure 5. Once extracted, the feature values were each normalized to their Z-score. The preprocessing stage

[2]Plotted curves must be considered just for illustrative purposes, since for the sake of confidentiality the originally processed time series cannot be explicitly shown in the Figure

yielded a total of 19 binary and real-valued predictors.

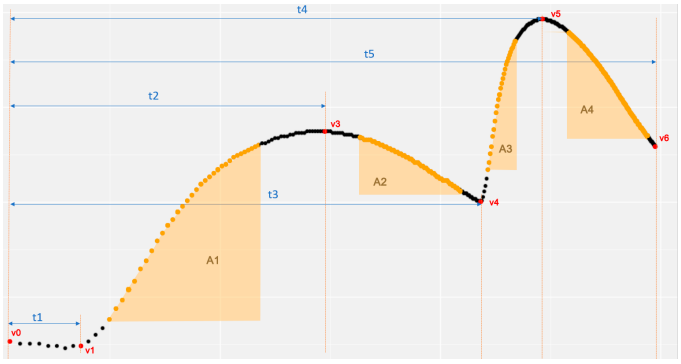
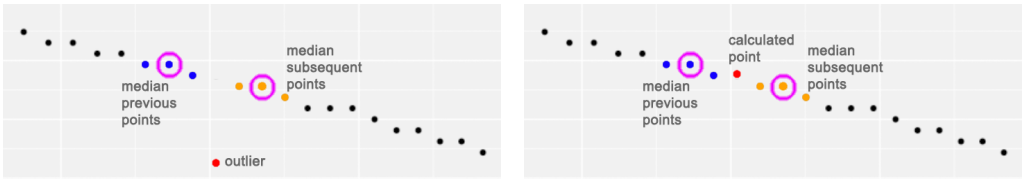


Figure 5: Features extracted from the time series X_t^1 captured by the selected sensor at trial $k = 1$: extreme points $\{v_0, \ldots, v_6\}$, areas $\{A_1, \ldots, A_4\}$ and delays $\{t_1, \ldots, t_5\}$, resulting in 16 features.

No imputation of lost data was needed in this first cycle, whereas outliers were only found in X_t^1 due to reading errors undergone by the installed sensor. A statistical correction was performed over identified atypical samples,substituting by the average of the median of the 3 previous and subsequent values those points whose difference with respect to the previous and posterior point value is greater than a certain percentage. (see Figures 6.a and 6.b).



(a) (b)

Figure 6: (a) Computation of medians after an outlier ● has been detected; (b) Replacement of the atypical sample using the average value of the computed medians.

(1.d) Predict:

Results in terms of precision, recall, F1 score and Area Under the Curve (AUC, [42]) obtained by each scheme in a 5×5-fold cross-validation for under and over-sampling + RF are shown in Table 1. Given the poor results obtained in this benchmark, a final check was performed by the team to determine whether the number of examples in the dataset was enough for

Table 1: Average model scores of the first iteration of the ASPPID methodology.

Model	Recall	Precision	F1	AUC
OVERSAMPLING + RF	0.2325	0.6427	0.3404	0.8745
SMOTE + RF	0.2988	0.4527	0.3480	0.8214
BORDERLINESMOTE + RF	0.2325	0.4510	0.3051	0.8633
SMOTE-TOMEKLINKS + RF	0.313	0.4346	0.3504	0.8105
ADASYN + RF	0.2331	0.5020	0.3160	0.8666
ENN + RF	0.3199	0.4201	0.3463	0.8032
OSS + RF	0.2164	0.6584	0.3249	0.8837

predicting the target variable. This study was done by observing the evolution of the models' performance as the size of the dataset is progressively reduced. Only shown for ADASYN + RF (Figure 7) for brevity, results obtained for all methods lead to the same conclusion: the amount of data retrieved in this first trial was sufficient, since additional data produced by more samples (by e.g. extending $T_{col}^{min}(s_1)$) would not improve significantly the performance of the model.

As mentioned on the previous subsection 4.2, we can confirm that AUC is not a good measure when imbalance class is present. It can be seen that despite the model is not well classifying the positive class, high values of AUC are given. This is because of the weight that majority class has on model generation.

(1.e) Implement: the unsatisfactory results obtained with this first sensor were not convincing to invest on a *implement* phase to validate them in production.

Based on the data-based insights drawn from this first study, the improvement team proposed a new iteration of the ASPPID methodology to the management board of the plant, where to test a new set of sensors within the monitored production process. The first conclusion drawn and agreed by both parties was that sensors used in this first trial were not enough to meet the established target value $B_{th}(t)$, yet the use of sensors already installed in another machine instead of purchasing them had reduced dramatically – more than 90% – the planned costs for this first iteration. This fact buttressed the consideration of cost efficiency as a key aspect in subsequent iterations of ASPPID, and steered thereafter the negotiations with sensor suppliers.

4.3.2. Second Iteration ($k = 2$)

This second iteration, and in general, all iterations within an ASPPID implementation, starts always from the Ishikawa exercise done in previous iterations:

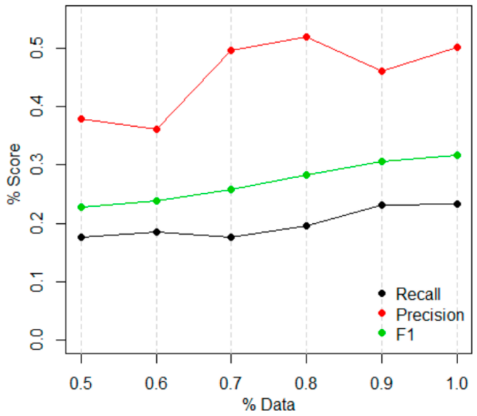


Figure 7: Average model scores versus amount of data for the ADASYN + RF approach.

(2.a) Analyze: the improvement team gathered again and generated new hypotheses with the variables that were initially discarded. Consequently it was decided and agreed to sense new magnitudes and elements along the production line, withstanding the uncertainty of their predictive relevance with the incorporation of new members (upon approval of the management team) with knowledge in Physics and Thermodynamics. Since sensors in the first ASPPID iteration yielded no costs for the project, they were again used in this second trial. New sensors needed to capture the newly decided set of parameters were rented instead of purchased after negotiation with the supplier, who finally admitted a change in their relationship model with the company (i.e. the supplier accepted to include mode B in its sales portfolio $\mathcal{M}_{sup}(s_2)$). At this point it is important to prescribe that in general, variables and sensors capturing them used in early trials should not be discarded from the ASPPID cycle unless the associated sensing equipment is expensive enough to potentially cause a budgetary problem along the project, even if by discarding them the team takes the risk of losing valuable predictive interactions between variables.

(2.b) Sense: Once rented sensors were installed and fully operational, 3 new time series were monitored: X_t^2, X_t^3 and X_t^4 (masked for confidentiality), collected over a period of $T_{col}^{min}(\mathbf{s}_2) = 2$ months at 1 sample per second.

(2.c) Preprocess: repeating the feature extraction procedure made on X_t^1 with the 3 additional time series monitored by the new sensors, a dataset composed of 13522 parts, each characterized by 35 features, was produced, with an non-quality sample rate – namely, imbalance ratio – of 13.44%. Standardization and cleansing were implemented by using the same techniques as in X_t^1 on

X_t^2, X_t^3 and X_t^4.

(2.d) Predict: an identical data analysis to that of the first iteration was done with this newly produced dataset: techniques for handling class imbalance and an inner RF classifier using 5×5-fold cross-validation were used when building the model, rendering the scores listed in Table 2.

From all the generated models, the combination with the highest average F1 and recall values was selected (SMOTE + RF), since all parts undergo a quality control after being processed, thus precision is less relevant in the choice of the model. The study of the predictive importance for each variable (evinced by the RF model) suggested the discrimination of two feature groups comprising more than 50% of the predictive importance upon which to construct 2 different decision tree classifiers: one only with the first variable importance group (Group 1), and the second one with the first and the second group (Group 1+2). The reason for proceeding in this way hinges on the eventual implementation and deployment of the model: operating and fine tuning fewer variables within the production line is more cost-efficient and easily adoptable by the plant personnel.

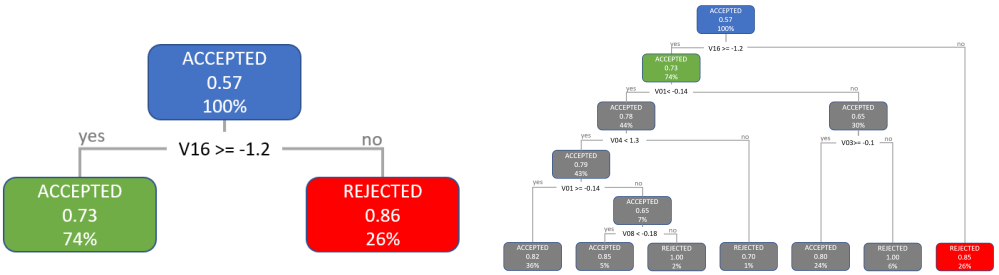
Table 2: Average model scores of the second iteration of the ASPPID methodology.

Model	**Recall**	**Precision**	**F1**	**AUC**
OVERSAMPLING + RF	0.4604	0.7604	0.5733	0.7585
SMOTE + RF	0.5989	0.5728	0.5854	0.6658
BORDERLINESMOTE + RF	0.5160	0.6757	0.5850	0.7227
SMOTE-TOMEKLINKS + RF	0.6070	0.5595	0.5821	0.6593
ADASYN + RF	0.5324	0.6604	0.5893	0.7125
ENN + RF	0.6243	0.5345	0.5756	0.6455
OSS + RF	0.4971	0.7155	0.5864	0.7361

Unlike Random Forest, classification trees can be sensitive to collinearity [43], so the first step is to verify the multi-linearity of the variables involved in Group 1, namely, V09, V10, V15 and V16. In this regard, a pairwise correlation matrix revealed a high linear dependency between the selected variables. Based on this insight, tree models were constructed over different data subsets based only on variable v16, whose decision threshold was proven to be quite stable with respect to different training samples (by comparing among trees trained over different subsets of the available data). If both groups of features were used, variables v14, v1, v8, v4 and v3 composing Group 1+2 would be included in the dataset. A similar correlation

analysis to the one performed previously evinced that collinearity was only present between the variables of the first group. Therefore, all features from the second group were fed to the decision tree model, along with feature v16 from the first group.

Figures 8.a and 8.b show the most frequent classification tree resulting from creating the explanatory model over 5 folds of the balanced dataset with the variables of groups 1 and 1+2, respectively. Each node shows the predominant predicted class (ACCEPTED/REJECTED), the percentage of observations and the occurrence of the predominant predicted within such observations at every node. It can be noted that v16 is capable, by itself, of separating the two classes. As for tree built using group 1+2, the number of variables to be controlled is greater, yielding a too complex set of rules to be implemented in production. Again, variable v16 is evinced to split both classes perfectly.



(a) (b)

Figure 8: Most frequent decision trees inferred from the collected information with different feature subsets (a) Group 1; (b) Group 1+2.

Results corresponding to the application of the above trees are shown in Table 3. Despite the NQR is lower in the tree trained over groups 1 and 2, the rule set extracted from the tree grown by the model was definitely considered to be overly complex to be implemented in the production line. Even though both models enhanced their NQR scores with respect to the benchmark in trial $k = 1$ (supporting their potential to produce a maximum non-quality rate equal to 8% in the implement phase), the Pareto trade-off between implementability of the rules extracted from the model and their performance gain in terms of recall was reported by the improvement team to the engineering manager. Driven by its simplicity, the manager opted for the simplest model based on thresholding variable v16.

(2.e) Implement: in this phase the rule extracted from the explanatory decision tree model selected after the *predict* phase was implemented and run in production for $T_{imp}(\mathbf{s}_2) = 3$ months. Specifically the plant personnel ensured that the conditions dictated by the selected tree model to yield less defective parts (since the simplest model was selected, it reduced to V16 < -1.2) were fulfilled by the machinery. This new production directive produced 23966 parts, with 1672 REJECTED and 22294 ACCEPTED parts, i. e. a non-quality rate of 6.98%, achieving the target for the ASPPID project within $K = 2$ iterations.

Table 3: NQR (mean$\pm$ confidence interval) of rules extracted from feature groups.

		SMOTE + Decision Tree	
Features	**Original data**	**Min NQR branch**	**Max NQR branch**
Group 1+2 (*predict*)	$13.34 \pm 0.47\%$	$7.05 \pm 0.49\%$	$59.67 \pm 2.04\%$
Group 1 (*predict*)	$13.34 \pm 0.47\%$	$7.37 \pm 0.52\%$	$57.76 \pm 2.21\%$
Group 1 (*implement*)	–	6.98%	–

Once verified by the improvement team and presented to the management team, the rented second sensor was purchased, increasing only 8.62% the overall cost of the trial with respect to the case where sensor s_2 was purchase from its beginning. However, as shown on Figure 9 the overall cost of sensing equipment spent during the ASPPID project, including the first and second iteration, was reduced down to 43.11% with respect to a traditional sales and purchase scheme. These cost savings were acknowledged by the management team to a right, informed choice of the sensing equipment, and a close match between the performance metrics of the developed predictive model and the requirements of the use case, which blended together ease of deployment and compliance with the imposed performance target.

4.3.3. Deploy

Once the results of the *implement* phase were verified, the project manager updated the FMEA information in the knowledge base of the company [44]. Although the case study was done on an experimental process within Edertek, the deployment of the model on a real production line will be straightforward thanks to the insertion of this simple yet efficient rule in the FMEA database: a rule extracted by virtue of the proposed ASPPID methodology.

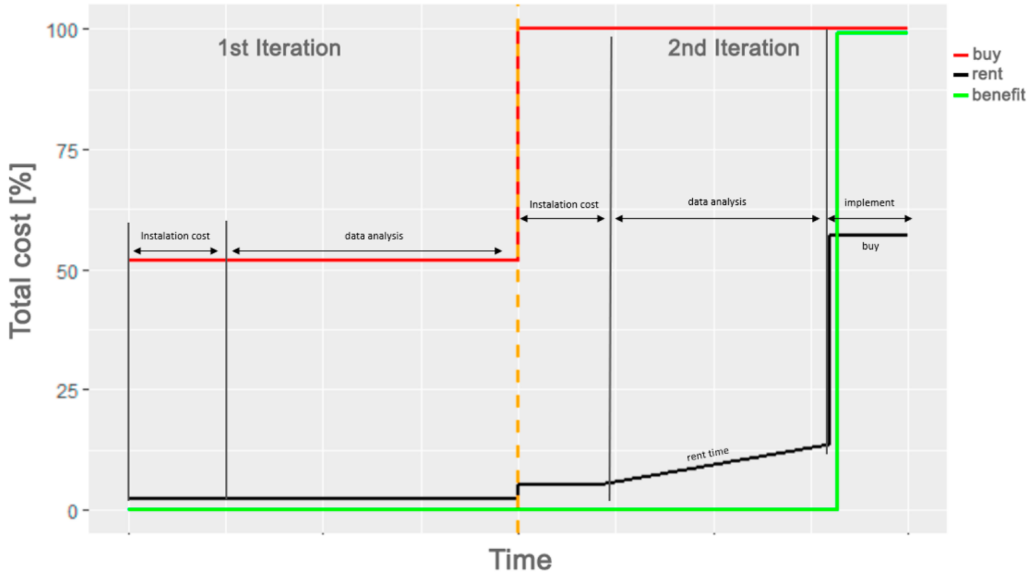


Figure 9: Graphical comparison between costs derived from renting (—) and buying (—) strategies in the acquisition of sensors within the ASPPID methodology. The plot also includes the progression of the benefit (—) along the SPPID cycle.

5. Concluding Remarks and Open Research Lines

This paper has presented ASPPID, a novel iterative methodology for efficiently selecting sensing equipment in industrial processes and machinery, essentially capitalizing on a predominant role of the data scientist along the entire decision workflow, not only within the stages where the captured data are exploited. The proposed methodology comprises steps where all decisions are made towards minimizing the costs and fulfilling the target benefit of the overall project.

For this to be realized, ASPPID proposes to arrange a team where the leading data scientist bridges the traditional gap between the purchase of sensors and its expected utility with respect to the objectives of the project. Furthermore, if allowed by the supplier ASPPID recommends exploring new forms of temporal sensor provisioning and installation, such that the extra renting costs are compensated by a lower risk taken by the team, specially when the project is held with data generated in application scenarios with scarce or null good practices reported by the community. The adoption of this methodology is specially suitable when there is no a priori information about the potential of the data captured by the sensor to bias for the project at hand; in this case a deep involvement of the data scientist in the selection of the sensing equipment to be tested along the trials is deemed crucial not to oversize the sensing requirements of the plant.

A use case held in a real manufacturing industry was used to validate the proposed methodology. The application of SPPID gave rise to a reduction of the sensor costs over 43% with respect to the case where all sensors were acquired from the beginning, while meeting a scrap ratio (NQR) less than the target value imposed at the beginning of ASPPID.

Future research will be devoted towards the inclusion and hybridization of existing production models for data mining (e.g. CRISP-DM) and agile software development (corr. SCRUM or Lean Software, among others) within the ASPPID methodology, placing an special emphasis on how to incorporate multiple conflicting decision criteria to the decision making process. Also new real applications of ASPPID to other industrial scenarios will be performed, such as predictive maintenance of remote machinery and the detection of non-technical losses in Smart Grids, among others.

Acknowledgments

The real case could not be possible without the participation of Fagor Ederlan S.Coop. and its R&D center, Edertek. This work has been supported in part by the ELKARTEK (ref. KK-2016/00096, BID3ABI project) and EMAITEK programs of the Basque Government. Antonio J. Nebro is supported by Grants TIN2014-58304 (Ministerio de Ciencia e Innovación, Spain), and P11-TIC-7529 and P12-TIC-1519 (Plan Andaluz I+D+I, Spain).

Bibliography

[1] Bosch, Industry 4.0 – Germany takes first steps toward the next industrial revolution, http://blog.bosch-si.com/industry-4-0-germany-takes-first-steps-toward-the-next-industrial-revolution/ (accessed 28 September 2017)

[2] Wang, S., Wan, J., Li, D., Zhang, C. 2016. Implementing smart factory of industrie 4.0: an outlook. International Journal of Distributed Sensor Networks 12(1): 3159805.

[3] Latham, S. F., Braun, M. R. 2008. The performance implications of financial slack during economic recession and recovery: observations from the software industry (2001-2003). Journal of Managerial Issues 20(1): 30-50.

[4] Hwang, K., Chen, M. 2017. Big-Data Analytics for Cloud, IoT and Cognitive Computing. John Wiley & Sons.

[5] Browne, W., Yao, L., Postlethwaite, I., Lowes, S., Mar, M. 2006. Knowledge-elicitation and data-mining: Fusing human and industrial plant information. Engineering Applications of Artificial Intelligence 19(3): 345-359.

[6] Camacho, E. F., Bordons, C. 2012. Model predictive control in the process industry. Springer Science & Business Media.

[7] Zhang, Z., He, X., Kusiak, A. 2015. Data-driven minimization of pump operating and maintenance cost. Engineering Applications of Artificial Intelligence 40: 37-46.

[8] Pfohl, H. C., Yahsi, B., Kurnaz, T. 2017. Concept and Diffusion-Factors of Industry 4.0 in the Supply Chain. Dynamics in Logistics, 381-390.

[9] Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., Akter, S. 2017. Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, 70: 308-317.

[10] Chehbi-Gamoura, S., Derrouiche, R. 2017. Big Valuable Data in Supply Chain: Deep Analysis of Current Trends and Coming Potential. In Working Conference on Virtual Enterprises 230-241. Springer, Cham.

[11] Lasi, H., Fettke, P., Kemper, H. G., Feld, T., Hoffmann, M. 2014. Industry 4.0. Business & Information Systems Engineering 6(4): 239-242.

[12] Best, M., Neuhauser, D. 2006. Walter A Shewhart, 1924, and the Hawthorne factory. Quality and Safety in Health Care 15.2 (2006): 142-143.

[13] Lim, S. A. H., et al. 2017. A systematic review of statistical process control implementation in the food manufacturing industry. Total Quality Management & Business Excellence 28(1-2): 176-189.

[14] Madanhire, I., Mbohwa, C. 2016. Application of Statistical Process Control (SPC) in Manufacturing Industry in a Developing Country. Procedia CIRP 40: 580-583.

[15] Deming, W. E. 1981. Improvement of quality and productivity through action by management. Global Business and Organizational Excellence 1(1): 12-22.

[16] Bouranta, N., et al. 2017. Identifying the critical determinants of TQM and their impact on company performance: evidence from the hotel industry of Greece. The TQM Journal 29(1): 147-166.

[17] Sinha, N., et al. 2016. Mapping the linkage between organizational culture and TQM: the case of Indian auto component industry. Benchmarking: An International Journal 23(1): 208-235.

[18] Becker, R. M. 1998. Lean manufacturing and the Toyota production system. Encyclopedia of World Biography.

[19] Harry, M. J. 1998. Six Sigma: a breakthrough strategy for profitability. Quality Progress 31(5): 60.

[20] Shokri, A., Bradley, G., Nabhani, F. 2016. Reducing the scrap rate in an electronic manufacturing SME through Lean Six Sigma methodology.

[21] Delgado López, E. 2016. Propuesta de un plan para la reducción de la merma utilizando la metodología six sigma en una planta de productos plásticos. Doctoral dissertation (in Spanish), Pontificia Universidad Católica del Perú, Escuela de Posgrado.

[22] Winter, D., et al. 2017. Using real-time data for increasing the efficiency of the automated fibre placement process. International Journal of Vehicle Structures & Systems 9.1: 11.

[23] Braha, D., ed. 2013. Data mining for design and manufacturing: methods and applications. Vol. 3. Springer Science & Business Media.

[24] Rahim, M. S., Rahman, M. Chowdhury, A. E. 2017. Mining Industrial Engineered Data of Apparel Industry: A Proposed Methodology. International Journal of Computer Applications 161(7): 0975-8887.

[25] Enarsson, L. 1998. Evaluation of suppliers: how to consider the environment. International Journal of Physical Distribution & Logistics Management 28(1): 5-17.

[26] Burdick, R. K., Borror, C. M., Montgomery, D. C. 2003. A review of methods for measurement systems capability analysis. Journal of Quality Technology 35(4): 342.

[27] Dietrich, E. 2012. Capability of measurement processes based on ISO/FDIS 22514-7 and VDA 5. XX IMEKO World Congress Metrology for Green Growth, Republic of Korea, 9-14.

[28] Chmielewski, M. R., Grzymala-Busse, J. W. 1996. Global discretization of continuous attributes as preprocessing for machine learning. International journal of approximate reasoning 15(4): 319-331.

[29] Shalabi, L. A., Shaaban, Z. 2006. Normalization as a preprocessing engine for data mining and the approach of preference matrix. International Conference on Dependability of Computer Systems, Poland, 207-214.

[30] Santos, M. S., et al. 2017. Influence of data distribution in missing data imputation. Conference on Artificial Intelligence in Medicine in Europe. Springer, Cham, 285-294.

[31] Little, R. J. A., Rubin, D. B. 2014. Statistical analysis with missing data. John Wiley & Sons.

[32] Gamberger, D., Lavrac, N., Dzeroski, S. 2000. Noise detection and elimination in data preprocessing: experiments in medical domains. Applied Artificial Intelligence 14(2): 205-223.

[33] Velleman, P. F., Hoaglin, D. C. 1981. Applications, basics, and computing of exploratory data analysis. Duxbury Press.

[34] Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. "Handling imbalanced datasets: A review." GESTS International Transactions on Computer Science and Engineering 30.1 (2006): 25-36.

[35] He, H., Garcia, E. A. 2009. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering 21(9): 1263-1284.

[36] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16: 321-357.

[37] Tomek, I. 1976. Two modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics 6: 769-772.

[38] Han, H., Wang, W. Y., Mao, B. H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Advances in intelligent computing, 878-887.

[39] He, H., Garcia, E. A., Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks, 1322-1328.

[40] Wilson, D. L. 1972. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics 2(3): 408-421.

[41] Kubat, M., Matwin, S. 1997. Addressing the curse of imbalanced training sets: one-sided selection. International Conference on Machine Learning, 97.

[42] López, V., et al. 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250: 113-141.

[43] Breiman, L. 2001. Random forests. Machine learning 45(1): 5-32.

[44] McDermott, R., Mikulak, R. J., Beauregard, M. 1996. The basics of FMEA. SteinerBooks.