energies

MDPI

Article

A Framework for Big Data Analytical Process and Mapping—BAProM: Description of an Application in an Industrial Environment

Giovanni Gravito de Carvalho Chrysostomo [1], Marco Vinicius Bhering de Aguiar Vallim [1], Leilton Santos da Silva [2], Leandro A. Silva [1,*] and Arnaldo Rabello de Aguiar Vallim Filho [3,*]

1 Postgraduate Program in Electrical Engineering and Computing, Mackenzie Presbyterian University, Rua da Consolação, 896, Prédio 30—Consolação, São Paulo 01302-907, Brazil; giovannigravito@gmail.com (G.G.d.C.C.); vallim.marco@gmail.com (M.V.B.d.A.V.)

2 EMAE—Metropolitan Company of Water & Energy, Avenida Nossa Senhora do Sabará, 5312—Vila Emir, São Paulo 04447-902, Brazil; leilton@emae.com.br

3 Computer Science Department, Mackenzie Presbyterian University, Rua da Consolação, 896, Prédio 31—Consolação, São Paulo 01302-907, Brazil

* Correspondence: leandroaugusto.silva@mackenzie.br (L.A.S.); aavallim@mackenzie.br (A.R.d.A.V.F.)

Received: 30 July 2020; Accepted: 9 November 2020; Published: 18 November 2020

Abstract: This paper presents an application of a framework for Big Data Analytical Process and Mapping—BAProM—consisting of four modules: Process Mapping, Data Management, Data Analysis, and Predictive Modeling. The framework was conceived as a decision support tool for industrial business, encompassing the whole big data analytical process. The first module incorporates in big data analytical a mapping of processes and variables, which is not common in such processes. This is a proposal that proved to be adequate in the practical application that was developed. Next, an analytical "workbench" was implemented for data management and exploratory analysis (Modules 2 and 3) and, finally, in Module 4, the implementation of artificial intelligence algorithm support predictive processes. The modules are adaptable to different types of industry and problems and can be applied independently. The paper presents a real-world application seeking as final objective the implementation of a predictive maintenance decision support tool in a hydroelectric power plant. The process mapping in the plant identified four subsystems and 100 variables. With the support of the analytical workbench, all variables have been properly analyzed. All underwent a cleaning process and many had to be transformed, before being subjected to exploratory analysis. A predictive model, based on a decision tree (DT), was implemented for predictive maintenance of equipment, identifying critical variables that define the imminence of an equipment failure. This DT model was combined with a time series forecasting model, based on artificial neural networks, to project those critical variables for a future time. The real-world application showed the practical feasibility of the framework, particularly the effectiveness of the analytical workbench, for pre-processing and exploratory analysis, as well as the combined predictive model, proving effectiveness by providing information on future events leading to equipment failures.

Keywords: big data process; predictive maintenance; machine learning

1. Introduction

Interest in data-based knowledge applied to decision-making processes has been growing in different industrial segments [1]. The importance of this movement of data-driven decisions is understood, since organizations with better performance have used data analysis five times more than those with low performance [2].

This movement of implementing a so-called KDD—knowledge discovery in databases—environment is relatively new in industrial business, and it is due, on the one hand, to the huge volume of data generated (big data), which is largely the result of the Internet of Things (IoT), where sensors connected to a variety of objects, spread across the planet, have accelerated the big data phenomenon. On the other hand, data availability has sparked interest in using these historical data to support decisions, based on mathematical models and algorithms, mainly those of artificial intelligence (AI), which allow predictions of different types of events, such as the imminence of equipment failure, triggering a preventive maintenance schedule [3].

The combination of concepts, such as big data, IoT and AI, has had a considerable impact on industrial business, defining the main dimension of Industry 4.0, which can be defined as a concept that encompasses automation and information technology, transforming raw materials into value-added products from data-driven sources [3,4].

One of the main areas is AI-based predictive maintenance. In this type of maintenance rather than scheduling operation suspension for maintenance, based on fixed time intervals, the best stopping moment is defined based on AI inference, as a result of an analytical model, calibrated (trained) on the basis of historical data [4–6].

A continuous monitoring of equipment, by AI algorithms, can have an important impact by allowing the reduction of corrective maintenance, which occurs unexpectedly and is strongly undesirable, compromising budgets and industrial production. Advance information that an equipment failure is close allows for proactive and planned actions to mitigate these financial impacts. This is clearly a trade-off between investment in research and development and equipment productivity [7,8].

New companies are already starting operations considering the modern concepts of Industry 4.0, but traditional industries are also entering this new world, seeking to improve their processes by including Industry 4.0 elements.

1.1. Motivation

In this paper, we will deal with one of these cases. It is a real-world case of a hydroelectric power plant, operating since 1926, which despite being an operation within traditional standards, has, over time, been updated to receive monitoring systems based on data collection sensors. The objective now was to go one step further, developing an analytical "workbench" for data exploration and, furthermore, implementing applications of AI algorithms to support a predictive process.

So far, the plant updating process has been developed incrementally, but with little documentation. The mapping of processes and sensors, for example, were not fully updated. Therefore, if new improvements were desired, these mappings should be a must before any new action. Such mappings could provide a clear understanding of the power plant system and subsystems, as well as the types of sensors installed and variable observations collected. With an understanding of this entire universe, the road was open for new developments. As a result of these process mappings, as well as an exploratory data analysis, a favorable environment would be created for the application of AI algorithms to support the implementation of predictive models, and thus achieving a consistent KDD environment.

Therefore, the main motivation that led to the development of this paper was to report in the literature the experience obtained in this research project in which all phases of a big data process were covered and which led to the construction of a framework (BAProM) that can be used in industrial systems of different types.

The description of this framework, accompanied by an implementation in a real-world case, may lead other researchers to develop similar works, and professionals in the field to make better-informed decisions, and therefore, become more secure.

1.2. Research Question

This subsection presents the research question (RQ) that drove all the development of the study described in this paper.

RQ:

What are the phases and their respective internal structures to constitute a consistent framework focused on the big data process, which could be applied in real-world cases of predictive maintenance?

As the question states, its purpose is to define the phases, tasks and techniques that must be employed in each step of a big data process, considering from the identification of relevant processes and variables to be studied to the implementation of prediction models. Such a framework should be suitable for application in predictive maintenance use cases.

1.3. Objectives

Based on the RQ, the objective of this study, therefore, is to address these issues, and it must do so through a framework proposal that has been called BAProM—Big Data Analytical Process and Mapping.

As specific objectives of the study, we must:

(a) Define and test the BAProM framework as a pipeline of four modules: Process Mapping, Data Management, Data Analysis, and Predictive Modeling.
(b) Ensure the modules are adaptable to different types of industry and problems and can be applied independently.
(c) Develop an application of BAProM in the hydroelectric plant (UHB) as a decision support tool for predictive maintenance.
(d) Identify all relevant operational processes in UHB
(e) Identify all variables significantly associated with equipment failures.
(f) Conclude the application with the development of a prediction model of equipment failures
(g) The prediction model must have the ability to identify, by the values of the significant variables, whether an equipment would be close to a failure point or not.
(h) The prediction model must predict the probability of an equipment failure in a future period.

1.4. Implications and Contributions

The importance of studying the big data process is the relevance that the subject has acquired in Industry 4.0, since more and more stakeholders are adopting data-driven decision-making practices.

The implications of data analysis and prediction models, expected products of a big data process, are far beyond Industry 4.0. In fact, its benefits spread across all areas of activity.

In Industry 4.0, in particular, the implications of a framework that could be implemented as systematic procedures in the operation can be huge. Such models would lead to a minimization of corrective maintenance occurrences, in addition to optimizing the periodic maintenance schedule. Productivity can increase, as can profit. As the amounts involved in industry can be significantly high, so would be the benefits of costs savings.

This paper, therefore, can bring an important practical contribution to an important economic sector.

On the other hand, the conceptual and technical implications of the paper can also be significant, since novelties are proposed and validated by a complete implementation in a real-world case.

The mapping of processes and variables is often not present in the big data processes described in the literature, and this paper seeks to draw attention to this fact and show its relevance in the direction the project took.

The development of an analysis and data exploration tool, with the demonstration in the article of its use in different stages of the process, is another contribution of the study that should have implications in the way the projects are developed.

In addition, a combined prediction model, employing a decision tree complemented by an artificial neural network to forecast critical variables for a future period, as will be presented in this paper, is not often seen in the literature.

The article thus acquires some relevance with these contributions and may have positive implications both from a conceptual and practical point of view.

The description of the BAProM framework, as well as the real-world application case, is presented in the paper over five more sections. In Section 2, we give a literature overview of the works related to this research. Section 3 presents the methodology employed in the conception of the framework and shows how it could be implemented. In Section 4 we describe the Case Study developed in the hydroelectric power plant, and Section 5 shows and discusses the results of these practical applications. Finally, Section 6 presents the conclusions and gives directions for future works.

2. Related Works

Every industry, including power generation, wants its equipment to be as efficient as possible, which means operating at full load (or close to it), producing as much as possible and having the equipment for the maximum available time [9]. Therefore, maintenance aims to inspect any equipment to ensure its effectiveness, avoiding unexpected failures [10].

The most common type of maintenance is a periodic one, called preventive maintenance, which consists of stopping the equipment according to a predefined schedule, and performing scheduled services and inspections to check for additional repair needs. Most preventive maintenance stops can prove to be unnecessary, resulting in maintenance expenses and loss due to production stoppage. However, even so, this type of maintenance is sustained by the industry, as it is still the best resource to avoid corrective maintenance [11].

Corrective maintenance comes from a failure in an equipment throughout the industrial process, generating a high financial impact on budgets due, above all, to the immediate need for repair and spare parts, in addition to interrupting the production chain in an unplanned way [12,13].

The best scenario would be one in which the ideal time for maintenance is known in advance. But, this type of discovery is not trivial, as it involves a complex system of variables related to operation, maintenance, production and even the human order of those who are handling the equipment [14].

These questions increase the interest in installing sensors in a variety of equipment, collecting data almost in real time (in the order of seconds) about their mechanical, electrical or operational conditions. Having the data and developing analyses makes it possible to get to models supporting decisions regarding when maintenance should occur and what procedures should be adopted for eventual failures. Decisions, in this case, are supported and based on information extracted from data (Data-Driven approach) [7,8,15].

When a maintenance decision is based on information extracted from data collection, it generates a proactive action. In addition, this paradigm shift between reactive (corrective) to proactive maintenance actions is also seen in the literature by transforming time-based maintenance (TBM) into condition-based maintenance (CBM) [7,8].

Proactive maintenance uses concepts of Internet of Things (IoT), big data (BD) and artificial intelligence (AI). Simply put, for conceptualization purposes, the sensor used in monitoring is associated with the IoT component, the process of collecting and exploratory processing of data to the database is associated with BD, and the training of algorithms for the generation of models to be used for decision-making is addressed to AI.

Literature points towards a new industry revolution. After the mechanical, electrical and automation revolutions that brought mass production, assembly lines and information technology, raising workers' income and making technological competition the core of economic development, the fourth industrial revolution is characterized by a set of technologies, where the operation is modernized with sensors for monitoring, collecting, and storing data and using data-mining techniques, with intelligent algorithms to support decision-making [3,16,17].

The approaches of Industry 4.0 used together are optimistic because they can monitor, diagnose and predict possible failures in addition to indicating the best time for maintenance to occur. The papers focusing on anticipating the best time for maintenance define this approach as predictive maintenance [18–20].

Related work emphasizes the choice of specific algorithms or composite algorithms, in order to seek the best performance in predicting the best time for a maintenance service. Composite algorithms imply, on the one hand, the use of techniques for dimensionality reduction which may occur due to the high number of sensors. These are techniques such as the Principal Components Analysis (PCA) [15–17] or data clustering algorithms, as K- Means [21] or yet, probabilistic models such as the Bayesian Belief Network (BBN) [3]. On the other hand, there is the use of AI algorithms, where the most used in predictive maintenance are Support Vector Machines (SVM) [16,17,22], Artificial Neural Networks (ANN) [18,22], Bayesian Belief Network [3], Random Forest (RF) [22], Partial least squares (PLS) [15], Markov Chain and deterministic methods [23,24]. These mentioned works are discussed in more detail below.

Yin et al. present a survey of studies employing statistical methods for monitoring and detecting failures in large-scale industrial systems. As their main results, database problems stand out, and among them can be highlighted the high number of variables, wrong measurements and missing values. For variable treatment, especially dimensionality reduction, and monitoring to detect flaws, the authors conclude that the best approaches are PCA and regression by PLS. The combination also allows identification of the most significant variables in an equipment failure [15].

Another paper, developed by Jing and Hou used the Tennessee-Eastman Process (TEP) to simulate an industrial chemical environment in order to assess process control, process monitoring and diagnostic methods. As far as diagnosis is concerned, the authors used PCA to reduce the dimensionality of the data and SVM for the diagnostic classification [16].

A survey of articles from 2007 to 2015 using SVM to detect failures in industrial environments is presented in a paper of Yin and Hou. The main conclusion of this research was that the best results were obtained when the SVM was combined with some other dimensionality reduction technique [17].

Lee et al. proposed an analytical framework with Prediction-Health Management (PHM) algorithms aiming to learn how to operate normal equipment and to predict its lifespan. Self-analysis of the equipment is performed using unsupervised algorithms such as the ANN Self-Organizing Maps (SOM), defining normal operating standards. Therefore, when the operation comes to the point of having a certain level of dispersion in relation to its standard behavior, learned by SOM, the algorithm infers that it just started a degradation process [3].

The development of an ANN based on operation data of machining equipment is the content of a paper of Yan et al. The objective of the research was to estimate the remaining life of the most relevant component of that equipment. The work also proposes the need for a standardization of semi-structured and unstructured data from industrial processes, to improve the accuracy of the prediction algorithms. An improvement occurs because heterogeneous data, such as vibration signals from the machine and images of the machine's working environment, can provide important information for the prediction model after being structured and standardized [18].

Gatica et al. propose two approaches to predictive maintenance, named online and offline. The approaches have top-down and bottom-up strategies. In the "top-down" approach, the process begins with understanding the use case, as well as the machines employed. Following from this, a mental model of the process is made, where a hypothesis of how the process impacts data collection, is elaborated. Finally, the hypothesis is tested by analyzing the sensor data. In the 'bottom-up' strategy, the process has the following flow: data collection, exploratory analysis, selection of variables, predictive modeling and results validation based on the experience of the industrial process team [20].

A model to evaluate equipment failure time by collecting data with a vibration sensor was proposed by Sampaio et al. Their objective was to develop a relationship between the vibration levels and the equipment failure time, thus raising a characteristic curve that was learned by three AI algorithms: ANN, RF and DT. The lowest RMSE (Root Mean Square Error) was achieved by ANN [22];

Wang et al. presented a framework named Policy Semi-Markov Decision Process (PSMDP) to find the best time for predictive maintenance, based on the system deteriorating state. The proposal aimed to understand the equipment operating status, so that maintenance would be planned

considering the aspects of production efficiency and maintenance expenses. The work aims to discover when equipment is about to present a failure and consequently establish an action plan for the best maintenance moment [23].

A paper developed by Gao et al. presented a bibliographic review of works dealing with approaches involving fault detection based on signals and methods of deterministic models. The result is a taxonomy of fault diagnosis approaches for deterministic systems, stochastic fault diagnosis methods, discrete and hybrid event diagnostic approaches, and diagnostic techniques for networked and distributed systems [24].

The works presented in this section focus on different aspects of predictive maintenance. Among all the works mentioned here, only Gatica et al., as explained above, thought of the problem in the form of a process [20]. The others focused on the techniques involved and among these, the problem of the data set is noted. The data collected from sensors has problems of outlier, missing values, standardization and dimensionality that were pointed out in full by only [18]. Others were concerned only with dimensionality reduction, which was resolved with the use of PCA. Regarding prediction processes, the SVM algorithm is widely used, but without further discussion of parameterization and the kernel used. In part, the strong use of this algorithm is due to its performance in comparison with other methods. However, most of the applications are in contexts that are not necessarily industrial environments.

A systematic review of Machine-Learning methods applied to Predictive Maintenance on two scientific databases: IEEE Xplore and Science Direct [25], gave an overview of the maintenance types—corrective, preventive and predictive—and tried to show the machine-learning methods being explored and the performance of the techniques. An analysis of the papers between 2009 and 2018 showed that techniques of the most diverse types have been widely used, such as: Decision Tree, RF—Random Forest, k-NN—k-Nearest Neighbors, , SVM—Support Vector Machine, Hierarchical clustering, k-means, Fuzzy C-means, ANN—Artificial Neural Network, LSTM- Long Short-Term Memory Network, ARIMA—Autoregressive Integrated Moving Average, ANOVA—Analysis of Variance, Linear Regression, GLM—Generalized Linear Model, and others.

In another paper, the authors presented a machine-learning approach for detecting drifting behavior—so-called concept drifts—in continuous data streams, as potential indication for defective system behavior and depict initial tests on synthetic data sets. The machine-learning techniques used in the study were LR, RF and Symbolic Regression (SR). They also presented a real-world case study with industrial radial fans and discuss promising results from applying their approach [26].

The literature also presents a predictive maintenance framework based on sensor measurements [27] and a prognostic is developed, oriented towards the requirements of operation planners, which is based on a Long Short-Term Memory network. Its performance is compared with two benchmark maintenance policies: a classical periodic and an ideal case (perfect prognostics information) called the ideal predictive maintenance (IPM). The mean cost rate of the proposed framework was lower than the periodic maintenance policy and close to the ideal case IPM. It is possible to find works yet, with confirmations that big data and IoT play a fundamental role in data-driven applications for Industry 4.0, as is the case of predictive maintenance [28]. The authors in this paper reviewed the strengths and weaknesses of open-source technologies for big data and stream processing and tried to establish their usage in some cases. As a result, they proposed some combinations of such technologies for predictive maintenance in two cases: one in the transportation industry, a railway maintenance, and another in the energy industry, a wind turbine maintenance.

Another study proposed a Weibull proportional hazards model to jointly represent degradation and failure time data. The authors explained that degradation refers to the cumulative change of the performance characteristic of an object over time, as the capacity of batteries of hybrid-electric vehicles, the leveling defects of railway tracks and so on. The proposed strategy was applied to the predictive maintenance of lead-acid batteries and proved to be adequate [29].

This review sought to provide an overview of the main aspects associated with the theme of this research. Thus, works were presented showing the context of the Industry 4.0 environment, involving the maintenance of equipment, the acquisition of data for monitoring, based on sensors, the use of AI algorithms based on ANN for failure prediction, the use of statistical methods for monitoring and fault detection, and other proposed analytical structures. A rich field of opportunities has been presented.

From this picture of opportunities, verified in the literature review, emerged one of those opportunities with the proposal of the big data Analytics Process Mapping framework, BAProM, which is the development of an analytical framework covering the entire big data process and also including a first phase of a detailed mapping of processes and variables, which is not frequently seen in the literature. As stated before, synthetically, the framework consists of four modules: Process Mapping, Data Management, Data Analysis and Predictive Modeling.

Such a framework, including the mapping of processes and variables to a predictive analysis and showing results of an implementation in a real-world case, is a novelty in the literature.

The details of each of these modules are presented in the next section, as well as the reasons for each technique selected to became part of this first version of the framework, which was validated in a real-world case of the Henry Borden hydroelectric plant Section 4.

In addition to its conceptual relevance, the research gains practical importance by being applied to a relevant industrial system in the real world creating mechanisms for monitoring the operation and predicting equipment failures, which could be avoided once they were known in advance.

3. Framework

A classical development of a big data project starts by data collection regarding the important variables of the system under study [1]. However, in some cases, it is not so clear what these variables are, since a comprehensive documentation may not be available. In these cases, an earlier phase of mapping processes and relevant variables to characterize the state of the system is necessary.

The framework proposed in this paper introduces in the big data process phase of mapping processes and variables as an initial fundamental part of the process, which is followed by data management, in which part lies the collection of primary data. After that, there would be a phase of data analysis, with more exploratory characteristics, and in the end there is a predictive modeling.

The entire process was consolidated into four modules, whose details are shown as follows:

Module 1: Mapping Process

- Process mapping to identify critical operation points of the system under study;
- Variables mapping to identify critical operation indicators;
- Analysis of technical reports to define the standard behavior for variables monitored by sensors;
- Interviews with specialists responsible for the operation of the system.

Module 2: Data Management

- Data Acquisition from sensors at critical points of the system, properly mapped;
- Aggregation to dataset any relevant historical data registered in other systems;
- Pre-processing of data involving preparation, transformation of the data into a final format for analysis and selection;
- Consolidation of data on a single database.

Module 3: Data Analysis

- Implementation of computational tools for analysis and visualization of information stored in dataset;
- Development of analyses and new insights about interrelations, correlations and/or operational trends;

Module 4: Predictive Modeling

- Design and Construction of an Incident Predictive Model;
- Validation of the predictive model;
- Application of the predictive model to optimize processes;

Figure 1 illustrates the complete framework, including the techniques and the computational tools applied in each step. Please note that the framework proposed is an extension of the big data process proposed by [1]. Here, the flow of activities incorporates mapping, which therefore becomes an integral part of a big data process. This, in a way, is recommended in the CRISP-DM (Cross Reference Industrial Standard Process for Data Mining) model, which suggests as initial phases the understanding of processes and data [30]. However, this understanding phase is not directly related to a mapping of processes or variables, in CRISP-DM, as it is here in this proposal.

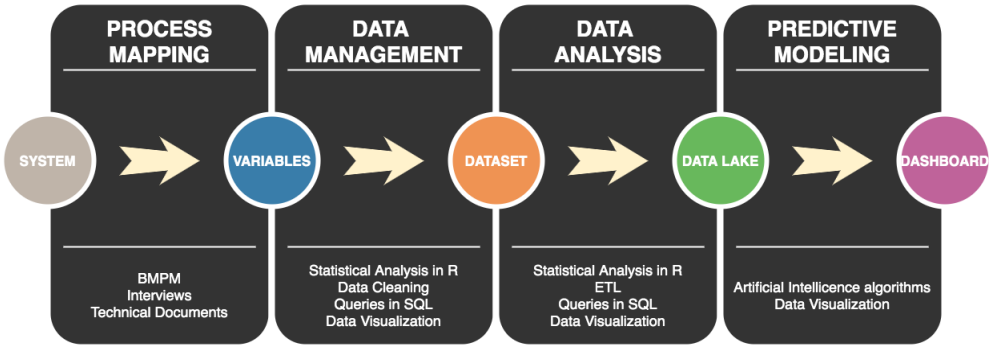

Figure 1. BAProM—Modules and techniques applied by module.

3.1. Process Mapping

The mapping of processes is a fundamental step, since it unveils the set of variables, which are those "keeping the knowledge" of the system under study, often obscured under a surface of a mass of data. This mapping of variables, which follows the mapping of process, opens the access doors to this knowledge. A process mapping can be defined as a modeling technique used to understand in a clear and simple way how a business unit is operating, representing each step of its operation in terms of inputs, outputs, and actions. As a result, a model of the system operation is built, with all its flows, relations, variables and complexities [31,32]. This is a fundamental step in research and development studies, as it provides a broader view of the object of study and makes it possible to improve the basis for decision-making, since at the modeling stage all processes are identified, mapped, understood and validated, which may lead to a process redesign. The characteristics of the processes (flows and/or activities) may be redesigned, aiming optimization and/or adaptation to recurrent needs.

All these concepts were initially applied to business processes, to improve and to automate a process. In fact, process automation by the means of applications is one of the major uses of process modeling [32]. However, by the characterization and validation of a process, it is possible to identify critical points in the system and, therefore, to identify and/or define critical variables, which form the basis for the data collection phase of a big data process. The start point for a consistent data collection is a set of representative variables of the system under study. Therefore, even though the aim here is the study of a big data process, the modeling process to identify this set of representative variables is similar to classical business process modeling. Thus, this paper tries to demonstrate how a tool originally designed for modeling business processes, the well-known software engineering tool BPMN—Business Process Model and Notation—can also be applied to a big data process.

BPMN is the notation of the methodology of business process management, widely used in software engineering for process modeling and validation of the process from the prototype generation of an application. The BPMN was developed by the Business Process Management Initiative (BPMI)

and is currently maintained by the Object Management Group, maintaining the current version of BPMN in 2.0 [31,32].

A proposal [33] to use this process to align the business process with that of the analysis, corroborates the benefits pointed out in other articles [34]. The relevance of this type of application can also be demonstrated in a work which proposes, in an embryonic way, the improvement of a BPMN for better use in an analytical context [35].

The BPMN provides a standard notation, easily understood by all members of the business. Stakeholders include business analysts who create and refine the processes, the technical developers responsible for implementing the processes, and the business managers who monitor and manage the processes. Consequently, BPMN intends to serve as a common language to bridge the communication that often occurs between designing business processes and implementing a process automation. It is a process-modeling notation comprehensible to the process owner (definition); to the participant in the process (use); to system developers (automation); to the business manager (monitoring) and; to decrease the distance between definition and implementation of the defined solution [31,32]. Based on these characteristics, this proposed methodology considers the use of BPMN as an adequate tool for the development of the mapping of processes and variables, the initial stage of the big data process conceived here.

3.2. Data Management

When we talk about data, we are in fact referring to observations of a set of variables which is the fundamental pillar for an analytical description of a system. It represents a synthetic framework of the system knowledge map, and through the variables observations it is possible to penetrate often complex paths existing in the masses of data, obscured by a variety of noises, as random observations, missing value outliers and so on. Data management means collecting and dealing with these observed values of the variables, and assures quality to the data, since it is the base of the entire analytical process of the system. Data quality is essential for a descriptive analysis of the system and an understanding of its behavior, as well for predictive models.

As described at the beginning of this section, data management begins by the acquisition and recording steps, which are strongly dependent on the application domain. This collection step, based on the set of critical variables, is the basis for the next analytical phases. In the case of Industry 4.0, the theme of this work, sensors usually make the acquisition. However, it may also be done by data sources other than sensors, such as photos and/or sounds collected in the operating environment or even by very simple processes such as notes registering operating situations of equipment.

The second step of data management, referred to as extraction, cleaning and annotation, also known as pre-processing, is dedicated to improving data quality. The pre-processing has two fundamental segments: data preparation and dimensionality reduction.

Data preparation means, basically, cleaning, integration and representation or transformation of the data, preparing the data for the analytical phases.

The cleaning involves treatment of data noise, characterized mainly by outliers (points with behavior quite different from the others) and missing values (lack of observations). Due to the diversity of data sources from different databases, noise, inconsistencies or missing values are very common. Even data from a single database is not exempt from such problems, and neither is data collected automatically by sensors, as these are liable to fail [36–38]. The cleaning consists of eliminating noise, correcting inconsistencies and handling missing values. The treatment of noisy data consists of identifying attribute values outside an expected standard (outlier) or other unexpected behaviors. The causes are diverse, such as measurement variations of equipment, human interference or extraordinary events, among others. The solution can be by simply removing the value, if the observation is identified as an anomaly, or by treatment using binning, clustering or other procedures. The elimination of an outlier is the simplest solution, but, before eliminating such a value, it must be considered that an occurrence with a value other than the usual may be the result of a measurement never seen before and

therefore it should be carefully studied rather than being eliminated. An outlier, in fact, may represent an opportunity of a discovery, which might conduct a research to new paths not considered before. Correcting data inconsistency is also a part of cleaning. Inconsistency is the presence of conflicting values in the same attribute, which in many cases may be caused by the integration of different databases. An example would be if each database uses a different scale to measure power. One could use kilowatt and the other megawatt. In integration, the values would be inconsistent. The correction can be done manually, automatically, in some cases, or even considering some kind of normalization (see Data Transformation, ahead in this section). Another cleaning task is to deal with the absence of data, which occurs when one or more attribute values do not exist. There can be several causes, such as failure to fill manually, no knowledge of the attribute, or low importance of the attribute, among others. The problem can be solved simply by removing the attribute or removing the entire sample, if this may cause a problem to other attributes of the same sample. There are other types of solutions with more elaborate techniques, such as to assign the mean, a moving average, or even the minimum or maximum values to those missing values [37]. Data cleaning is an essential step for the analytic stage. After cleaning there is the integration and representation or transformation, as a final data preparation for the analytic stage. These are pre-processing procedures applied to the data to gain efficiency and effectiveness. The integration activity occurs when one has many data sources, and seeks the construction of a single database. Otherwise, it loses importance. The representation in many cases means a data transformation, converting types and/or values of attributes. In some cases it is necessary, for example, transforming a continuous numerical value into a discrete value, or a discrete value for categorical ordinal, categorical nominal for discrete binary and categorical ordinal for discrete. It may be necessary, however, to normalize attributes that present values in broad ranges, in order to make them have the same level of importance in an analytical process. For normalization, the literature presents different methods, such as the z-score that transforms attribute values so that they remain with zero mean and standard deviation equal to one. Another method, considered as standard by many authors, is the min-max method [37]. The pre-processing so far included its first segment, the data preparation, involving cleaning, integration and representation or transformation.

Dimensionality reduction is a second segment of pre-processing. It is associated with data redundancy, which is another problem that must be treated. It occurs when two attributes have a dependency on each other. In this case, they may have the same values, or they may be very similar. It may happen for different reasons, such as lack of information of a database (metadata) that an attribute generates another one, or it may also exist between copies of a database. Typically, redundancy can be identified using correlation analysis, where the Pearson Correlation Coefficient is one of the most frequently used [37]. However, it may also be identified by using techniques such as factor analysis or Principal Components Analysis (PCA). The result of applying these techniques is a selection of records in the database and/or attributes, which will form the final database for the analysis phase. This selected data is a reduced database, without redundancy, but with equivalent analytical capacity.

It should be noted that each project has different needs and it is not always necessary to develop all pre-processing steps described here. Anyway, if all the steps are necessary, the natural sequence would be preparation, involving cleaning, integration and representation or transformation, and dimensionality reduction, in an iterative way and with interactions between the steps, in a feedback process, until the final data quality is effectively guaranteed [38]. A final step may still be necessary at this phase, which would be consolidating the data into a single database.

3.3. Data Analysis

This analytical module, as shown in Figure 1, is basically defined by exploratory analyses of the critical variables, which is essential for a descriptive analysis of the system, building a clear understanding of its behavior. Similar to the variables stored regarding the information about a

system, once this data is properly explored and interpreted, the information obtained will represent an accumulated knowledge regarding the system.

The exploratory analysis is based on the observed values of the variables, and usually, it works with tools as the Structured Query Language (SQL) to create consolidated databases from multiple queries on different data sources. SQL allows data modeling, relating tables created by extracting, transforming and loading data, the so-called ETL process (extract, transform and load), and constructing analytical repositories appropriate for discoveries [39].

Typical examples of this analytical approach are multidimensional data models, supported by data warehouses (DW). A DW is a repository constructed with data extracted from transaction systems, the so-called OLTP (Online Transaction Processing) data, and is exclusively for analysis, and so, it is not constantly updated (non-volatile) [40].

Online analytical processing tools (OLAP) are useful instruments to explore a DW. This kind of tool provides the exploration of different perspectives of a database. Moreover, SQL and other statistical tools may provide aggregate functions to summarize data, generating descriptive statistics measures such as sum, mean, median, standard deviation, minimum and maximum values, counts, etc. As a result, the descriptive statistic provides a clear view of the behavior of the variables under study and furnishes indicators for consolidated reports.

The Interpretation of these indicators must be strongly supported by computational tools integrating statistical analysis with visualization resources, as different types of graphics, dashboards and other instruments. The implementation of such analytical tools is part of this proposal. With an analytical computational tool, the development of analyses and new insights about interrelations, correlations, and/or operational trends of the variables becomes a reality.

3.4. Predictive Modeling

Differently from the approach described in the previous section, the predictive module is based on Data-Mining (DM) techniques. DM is a process of analytically exploring databases for the purpose of making findings that are not obvious, whose outcomes are effective in decision-making processes. DM is a core component of a KDD process [38], and usually involves prediction, clustering, or data association techniques.

The prediction process can be developed based on AI algorithms, which are strongly based on the data, including an auto-adjustment of its internal free parameters, calibrating the model (the so-called "training" of the model) which is performed from data history. This parameter adjustment (the model training) makes the algorithm able to be applied in other datasets, distinct from the one where the training process took place. A training model can, for example, estimate future values of variables, as the probability of an equipment failure. In such an example, the prediction could support the estimation of an optimal period for equipment maintenance [1].

There are different types of algorithms for prediction. One of the classical ones is the Decision Tree (DT), which is a type of AI algorithm whose model, generated after the training process, can be interpreted by humans and machines [41].

In a DT algorithm, the training process is simple and intuitive. In DT, each variable (attributes) is analyzed in its capacity of estimating a class of an object in a dataset. The DT defines a sequence (in a hierarchical tree structure) of attributes to be used to estimate the category (the class) of the object under analysis and depending on this sequence, different results may be generated. Therefore, a metric must be employed to stablish the "best" attributes sequence. One of the most used indicators is the entropy, which is a measure of the uncertainty associated with an attribute. The entropy is computed in terms of separation between classes. The variables are combined, and a measure of the entropy is performed [37]. The final model is a hierarchical structure by variable importance leading to a process of classification of the objects.

In an industrial context, for example, a class may be an equipment failure or not. Based on the values of the attributes of an equipment, the DT algorithm decides if the values of its attributes in a

certain point in time means an imminence of equipment failure or not. An important characteristic of a DT algorithm is its ability to allow interpretation by humans and not only by machines. It provides a reasonable understanding to experts of how the model is making its decisions, what leads stakeholders to trust the model. In this BAProM framework a DT algorithm has been developed to be used in the optimization of the maintenance programming of equipment.

Another important AI algorithm type is the ANN. ANN mimics or simulates the behavior of the human brain. In fact, it is a computational algorithm that implements a mathematical model inspired by the brain structure of intelligent organisms. As a result, it is possible to implement a simplified functioning of the human brain in computers [42]. The human brain receives information, processes and returns a response, and does so through neurons, connected in an immense network, and which communicate with each other, by electrical signals (synapses). An ANN seeks to imitate, in a simplified way, this process, to solve a problem. ANN, therefore, is an artificial network of neurons (nodes) connected. These artificial neurons are connected in layers: an input data layer, intermediate layers (varying from 0 to "n") and an output layer.

ANN is a powerful tool for solving complex problems, and can be used, for example, in classification, clustering, associations and in time series forecasts. In this BAProM framework an ANN was employed to forecast time series of critical variables defined in the DT model. The two models, therefore, worked together to forecast an equipment failure, allowing the operation team to act before the fail takes place.

4. Case Study

This section presents a more detailed description of the case study and the experimental methodology applied in real-world operation. The computational tools used in the experiment are also discussed.

Therefore, the core purpose that has been implemented in this study was mapping operation, facilities, and processes to identify variables that would be relevant for decision-making of maintenance. Then, with those variables identified, it would be possible to start an analytical repository, and, further, training machine-learning algorithms to the prediction.

4.1. Case Description

The case studied in this article is the Henry Borden Power Plant (UHB), located in Cubatão, about 60 km from São Paulo, capital of the state of São Paulo, in Brazil.

Its power generation complex is composed of two high drop-off power plants (720 m) called External and Underground, with 14 groups of generators powered by Pelton turbines, with an installed capacity of 889 MW. Pelton turbines are characterized by blade-shaped fins that are the main cause of maintenance [43].

The External Power Plant is the oldest. It has eight external forced ducts and a conventional powerhouse. The first unit started operations in 1926, the others were installed up to 1950, in a total of eight generator sets, with an installed capacity of 469 MW.

Each generator is powered by two Pelton turbines, which receive water flows from the Rio das Pedras reservoir. These flows arrive at the so-called "Valve House", where they pass through two butterfly valves in penstocks. Then, they descend a slope, reaching their respective turbines, covering a distance of approximately 1500 m.

The Underground Power Plant is composed of six generator sets, installed inside the rocky massif of Serra do Mar, in a 120 m long, 21 m wide and 39 m high cave with an installed capacity of 420 MW.

The first generator set went into operation in 1956. Each generator is triggered by a Pelton turbine driven by four jets of water. The operation of the UHB was developed according to an Integrated System of Generation of Electrical Energy composed of four large interdependent subsystems, interrelated in a continuous way, in the sense of generating electric energy delivered to the Brazilian Interconnected System, distributed cross country.

The general framework in UHB today is management practices combining modern instruments, such as computerized monitoring systems and dashboards for visualizations of different types of indicators with empirical practices based on its team experience. The system in the operation center runs uninterruptedly, allowing information from the entire system constantly. Therefore, appropriate decisions can be made at every moment. However, many of these operating parameters and metrics are established based on empirical practices. A typical example is the timeframe between inspections and preventive maintenance of turbines of the generating system. These parameters, which are determinant for the quantification of operational costs and level of the service of the electric system, should be periodically re-evaluated and, if possible, optimized, in order to find the optimal point of the tradeoff between costs and service levels.

Hydropower plants and generation facilities represent a high level of investments requiring management based on robust processes and standards, to guarantee the adequate return of the investments. Its operation and maintenance must be developed to guarantee the preservation and maximization of the use of this patrimony, within the operational conditions in which it operates.

The primary objective of such facilities is to maximize the availability of the energy generation and use of equipment. This is only achieved with high-level operating standards and procedures to guarantee the facilities productivity and the quality of the services offered.

The operational and maintenance standards of a power plant have their own costs, which may be considerable, given the complexity and size of the operations, inducing managers to search for practices leading to costs minimization.

Therefore, the use of BAProM framework (Section 3), seeking to establish optimized parameters to minimize operational costs and maintenance associated with equipment shutdowns could represent a relevant contribution to the operation of a hydroelectric plant. The BAProM framework was applied to this case, to develop a predictive model to establish the probable occurrence of an incident in Generating Units (GU) that could cause an interruption in the operation and, consequently, the need for corrective maintenance. Based on these predictions, it would be possible to establish optimal periods between maintenance.

In the following section, we present the approach and results involved in the predictive modeling.

4.2. Experimental Methodology

The methodology applied to the case study strictly followed the four modules of the proposed BAProM framework, but throughout the project, the technical team had to face some concrete questions, which only when an experimental methodology is effectively put into practice is it possible to have the real dimension of certain issues. Given the impact of such practical issues on the time dedicated by the team to resolve them, it is understood that they deserve a record and a discussion, as they can occur in many real projects. On the other hand, some steps that might have seemed difficult, in practice, demanded much less time and dedication from the team than one could imagine previously.

Thus, this section presents the four modules of the experimental methodology highlighting and discussing the main practical aspects related to the experience developed during the project. This type of record can be a useful contribution to the definition of the steps of a methodology, and also, to emphasize the attention that a team must dedicate to each step of the application of a methodology. In addition, it can also be an important contribution to scale, in a schedule, the time of each phase in a practical project.

Module 1: Mapping Process

The mapping of UHB operational processes involved the identification and characterization of the operation systems of the power plant, and the main physical variables (electrical, mechanical and electromagnetic) associated with the processes. The main procedure for gathering information to build the mapping focused on a search directly with the power plant staff, since the individuals working on the plant showed to have a consistent knowledge and deep domain of the business to be

modeled. These professionals with relevant experience in management and operations showed to have knowledge not only of the general process, but also of the some important details, allowing a consistent and reliable description of the processes and their associated variables, identifying accurately some points to be mapped and highlighted. This module was one of the major challenges of the project, involving extensive discussion with the power plant operation team, to learn the operational process and the relevant variables and this is a lesson to be learned. Sometimes, it is not a trivial task for the technical data science team to learn the technicality of the business being modeled. In addition to the information extracted from the meetings with the operation personnel, other relevant information on equipment was obtained from documents provided by the company.

Module 2: Data Management

If the previous module represented one of the major challenges of the project, this data management module was the biggest one. First, it was decided that the data collection would be focused on two UHB Generating Units (UG), known as UG4 and UG6, both with the same mechanical, electrical and technical characteristics, and the data would be obtained from a supervisory system database fed by sensors coupled to the plant equipment, which were connected to this database. The problems started to appear when analyzing the collected data. The data has not been properly stored over the years. Most of the data collected was used just as input to dashboards and discarded after use. The fact is that although the hydroelectric plant had a good data collection infrastructure, the team did not have a culture of data analysis, but only used the information for an instant monitoring of the operation. In the team there was no qualification for data analysis, so what happened was that the data was used for monitoring and most of them was then discarded. In fact, there was little historical data to analyze. Therefore, what happened is that the effective data collection had to be started from the beginning of the study. This led to a considerable delay in the project's planned schedule. In addition, there was a great deal of heterogeneity among the collection time periods of the diverse variables. The intervals between two collections could be quite different from one variable to another. There was no standardization of these periods. Moreover, we found many variables without data (missing data) and many problems of noise, including inconsistency and outliers.

Beside these problems, during the project one of the Generating Units, UG4, had a technical problem and had to be deactivated for a long period. Therefore, the data collection had to be focused only on one of the Generating Units of the UHB, UG6, which became, therefore, the object of this study.

Anyway, all problems had to be solved, especially the interval between two consecutive collections, which was adjusted so that all variables were always collected at the same time stamp. New time series of observations of the variables started to be generated. Once these difficulties were overcome, the data was successfully collected and a succession of analyses could be conducted. The final collection period was from May 2017 to January 2018 and they kept being updated continuously.

Module 3: Data Analysis

This module did not present any significant practical problems. The challenge here was technical, related to the development of a computational tool to support the analytical phase, which should be effective, but also be friendly, so that it could also be used by the operation team, non-technical users. Therefore, as soon as the mapping phase ended, an analysis and visualization tool, an analytical workbench, was developed to be used not only in this phase 3, but also, in the previous one, data management, to assist in the characterization of the variables and in the identification of data quality problems. Each critical variable has been filtered by the dashboard tool, providing visualizations of their domains, through statistical diagrams and summaries, presenting time series graphs, boxplots and histograms to identify trends, seasonality, statistical distributions, presence of outliers and missing values, as well, metrics such as mean, median, standard deviation and quantiles. Once the analyses were completed, knowledge about the system increased significantly and the entire database was ready to be subjected to predictive modeling.

Module 4: Predictive Modeling

As the previous one, this fourth module did not present significant practical problems. Once more, the challenge here was technical, since the predictive modeling involves AI algorithms and data modeling, since the data should be properly prepared to input the models.

The modeling was subdivided in two predictive models: the first one was a Decision Tree, where relevant variables associated with equipment failure were identified, as well as, thresholds, indicating the imminence of a failure when a variable reaches that value; the second model was an ANN dedicated to forecast time series of the significant variables, and so, could be possible to foresee in a future period when one of these variables would reach a threshold. Therefore, this was a module in which the tasks went without unexpected occurrences, including the techniques and tools employed in the practical application, which proved to be well adapted to the tasks.

4.3. Computational Tools

This section discusses the computational instruments used in the case study, for the four modules of the proposed framework.

The mapping of operational processes and variables was the fundamental starting point of the methodology, first module of the BAProM framework. In this case, a tool originally designed for modeling business process, the software engineering tool, BPMN, could be applied even though this was a big data process. A graphical tool based on the last existing version of BPMN (v. 2.0) showed to be well suited for this development. The software provided appropriate resources for modeling processes allowing validation of operation rules, flows definition and identification of critical variables. Moreover, these features were essential for validation the mapping with the power plant team.

The next step, module 2 in the proposed methodology, was Data Management, where data collection played a relevant role. Data was provided from different sources, such as the supervisory system database, fed by a set of sensors, an application named Impediment Registry, a by-product of these project, and which records equipment occurrences, as well as some external data. The SQL Server Database Management System was the basis for data storage, in a unified repository.

Another step of this module 2, the pre-processing for data validation, as the data cleaning, had the support of the analysis and visualization tool mentioned earlier, an analytical workbench, developed specifically for this framework, in the R-Shiny, a language library and statistical environment in R. This computational tool was a fundamental support for the characterization of the variable's standards and identification of data quality problems, providing for example, the treatment of missing values and outliers.

In module 3, the complete exploratory analysis, was totally based on this analytical workbench. The Shiny package from R language made it possible to build interfaces for the application with a high level of usability, as well as the processing of basically, all kinds of statistical operations.

For analysis and visualization purposes the tool provided the flexibility of filters, allowing the selection of a generating unit, a specific subsystem and a variable, as for example, selects "UG6", subsystem "Generator" and variable "Stator Armature". The tool provided yet, diverse types of dashboard outputs, graphical and metrics, which represented the core of the data analysis in module 3. Its features included statistical summaries, graphical visualization of time series, boxplot and histograms.

Module 4 was developed basically, through algorithms coded in R. This language provides a variety of functionalities, as statistical and machine-learning functions, allowing the development of most of the data science algorithms, from the simplest ones to the most sophisticated, such as those of artificial intelligence algorithms. Both algorithms employed in module 4, DT and ANN, were developed in R, which proved to be well suited for the job.

5. Practical Application of BAProM: Results and Discussion

The practical application strictly followed the modules of the BAProM framework, which are presented in the following subsections.

5.1. Process Mapping Results and Discussion

The mapping of Processes and Variables, developed on the first module, was fundamental to identify and formalize all the relevant flows and processes of the UHB Integrated Energy Generation System, as well as all the relevant variables. An integrated macro model for the entire system was developed with the purpose of showing a more comprehensive view of all subsystems identified in the process. Therefore, it was possible to verify and analyze the major components of the four stages in their sequential order.

The modeling was developed, in most part through information gathering with the power plant team. Four subsystems were identified, making up the entire UHB energy generation process. These subsystems, which are: Adduction, Turbine, Generator and Transmission, are illustrated in Figure 2.

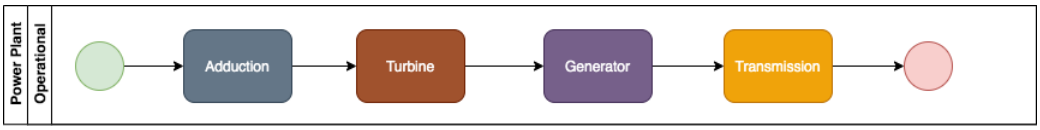

Figure 2. Macro view of the system with 4 subsystems.

These four subsystems make up, at UHB, an Integrated Electricity Generation System, which delivers energy to the Brazilian Interconnected Central System, composed of several power plants spread over the country, which in turn, distributes the energy throughout the country.

A synthetic description of the subsystems interaction, could be, as follows: the Adduction System carries a water flow from a reservoir descending a slope, to reach the turbines, covering a distance of approximately 1500 m (almost 1 mile). The pressure is enough to promote a high-speed rotation of the Turbine (Turbine System).

Each turbine, in turn, generates a rotation of the bearing axis on which it is supported, transmitting energy to the Generator to which it is connected.

The Generator by means of this kinetic energy, creates a magnetic field generating electrical current for the Transmission System. This system increases the voltage and prepares the energy ("packs") leading to transmission lines, integrating the Brazilian Interconnected System, for later distribution.

In this module, which represents the process mapping task of the BAProM approach, process modeling was applied to the entire system, which was a very extensive work, composed of vast documentation. Each of the four subsystems had its own modeled process, as well as the identification of its relevant variables.

For the purpose of illustrating this process, the mapping of one of the subsystems, the Generator, is presented here, with a special emphasis on one of its components, the "Stator Armature". The mapping of the other subsystems and their components was very similar to what is presented here.

The mapping of the Generator subsystem is shown in Figure 3, where the "Stator Armature" appears as its third component.

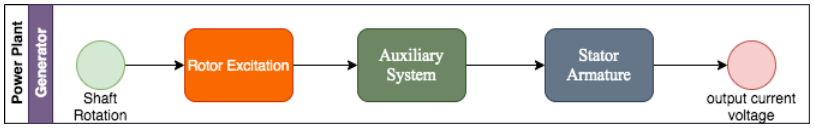

Figure 3. Generator Subsystem.

An integrated model for Stator Armature was developed showing a detailed view of this subsystem component (see Figure 4), and already including some aspects of data management,

as well. This mapping provides a solid basis for applying the other framework modules, in their sequential order.

The mapping provides a broad overview of all critical variables. For the specific case of the Stator Armature, five critical variables were identified: Active Power, Active Energy, Armature Tension, Armature Current and Rotor Groove Temperature. These five variables, now, should be continually monitored by sensors, and their observed values subjected to an ETL process, for future analyses.

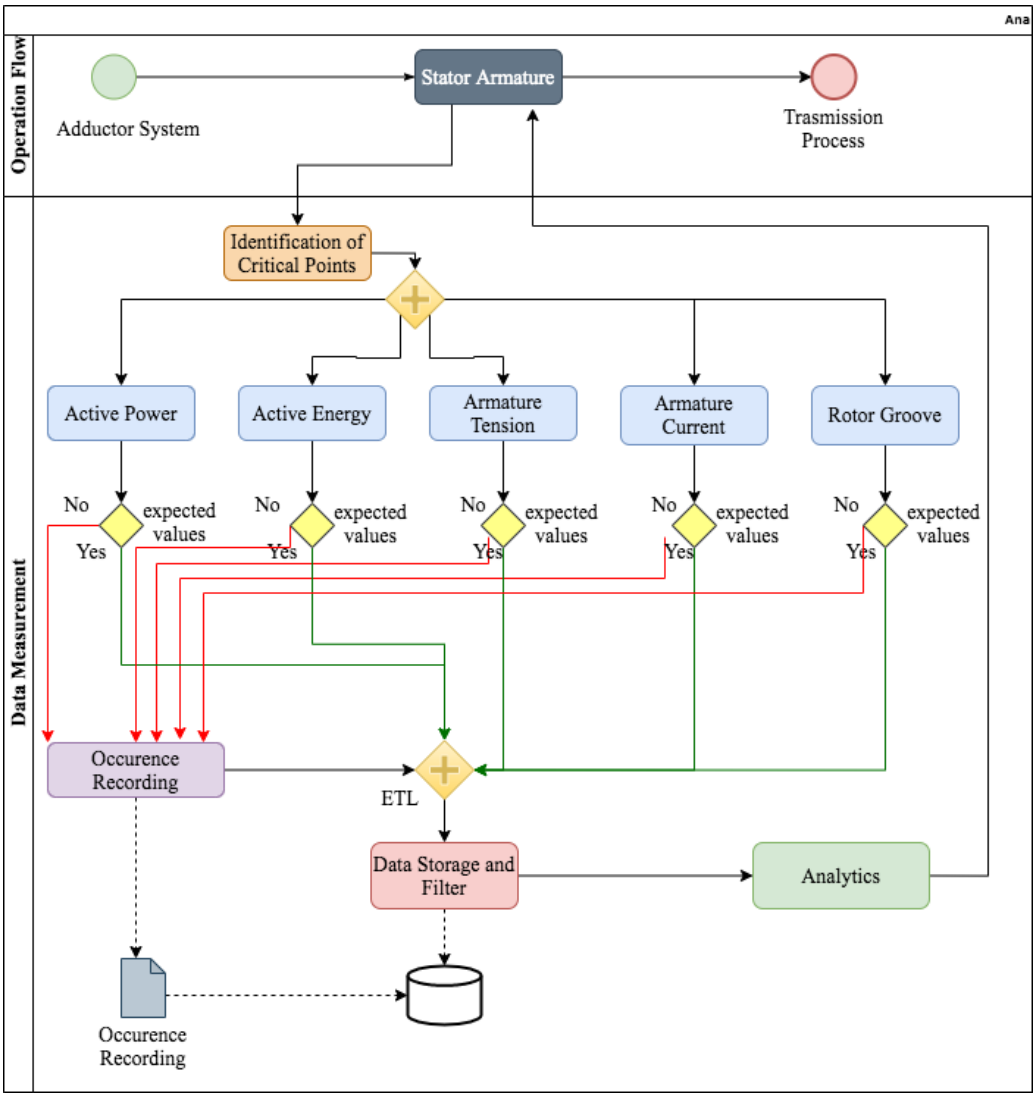

Figure 4. Mapping of the Component "Stator Armature".

The mapping of the complete operation of UHB was an effective practical contribution for the company, since it did not have this type of documentation, comprehensive and detailed, involving its entire operation.

Having completed the mapping, the next step was data management, which is the subject of the next section.

5.2. Data Management Results

The data management began with the data acquisition and recording, and it was favored by the previous phase, which provided an effective road map to the ETL procedure, by just following the flow throughout the mapping. In fact, as can be seen in Figure 4, whenever data is collected, a check is performed to verify that its value is within a specified range; if so, the values are extracted from the

source, transformed into a compatible format and then loaded into the database. As the ETL procedure extracted and transformed the data for storage, simultaneously, was carried out an analysis of data quality, on all kinds of anomalies.

The data quality is essentially the pre-processing step, which was developed in Module 2. In this step, the data was submitted to a rigorous quality analysis process, based on data preparation, which involved cleaning, integration and transformation of the data into a final format for analysis.

The cleaning involved treatment of data noise, characterized mainly by outliers and missing values. This phase had the support of the analytical workbench, developed specifically for this framework, which provided statistical analysis and visualizations, having been a fundamental support for this cleaning task.

The number of variables resulting from the data quality checking for each subsystem (Figure 2) are summarized in Table 1. Please note that in percentage terms, the proportion of variables with outliers represented 57% of the total of variables, while variables with missing values were 43%.

Table 1. Distribution of variables among the operation subsystems and number of anomalies.

System	Number of Variables	Number of Variables with Outliers	Number of Variables with Missing Values
Adduction	12	3	9
Turbine	35	30	5
Generator	37	23	14
Transmission	16	1	15
Total	**100**	**57**	**43**

These are relatively high numbers and, therefore, were brought up for discussion with the UHB technical team, to understand the reasons for such values. Regarding outliers, while in some cases there was just a possibility of a value outside the expected standard, in other situations the observed values in fact corresponded to problems to be treated. As an example, some temperature sensors measured negative values. Since this scenario was impossible in the region of the power plant and the equipment should accompany the operating environment, it was clear that the observed negative temperatures were errors in the data collection, and consequently, those values were discarded. It was found that the errors were due to sensor failures. Another reason for outliers was data collections performed at system startup times. In these cases, peaks occur in certain variables, but they soon stabilize, entering in an equilibrium state. The missing values were also related to sensor problems. In this case, for a period, some sensors were disconnected from the system due to technical causes. The missing values were then treated. In most cases, they were filled with averages for near periods.

Another aspect identified during this phase was that some relevant data, collected in other systems operating at UHB, were not being integrated into the database of the supervisory system. As a combination of different sources can be useful to develop exploratory analyses, as well as robust predictive models, this integration has been implemented. One of the important data sets incorporated was a maintenance database, since the predictive models of this study are focused on maintenance. Thus, for exploratory and predictive analysis, the built repository integrated data from controls and records of equipment maintenance to the data of the critical variables of the system. Therefore, a single database started to store all the relevant variables for the analyses developed in Modules 3 and 4.

Once these problems had been solved, a last question arose, concerned with the time interval between successive data collections. This problem could be better perceived when analyzing the groups to which the variables belonged. The collected variables belong to four different types: electrical, pressure, temperature and speed regulation. This pattern was already part of a tacit knowledge of the UHB's operating team, which was formally defined in module 1 of process mapping.

Regarding these four groups, the periods between successive collections were too long, and there was still, a considerable heterogeneity among periods of collection of the different types of variables. There was no standardization of these periods.

The different collection periods can be seen in Table 2. These turned out to be a serious problem, since the analysis of variables for different timestamps creates basic problems under two aspects: analytical and systemically. Moreover, the value itself of each collection period, was a problem, since it varied from 5 min to 15 min. These were long periods for this type of data collection.

Table 2. Distribution of Variables in Categories and Collection Time Interval per Category.

Type of Variables	Number of Variables	Period of Collection (in Minutes)
Electrical	8	15
Pressure	14	5
Temperature	18	15
Speed Regulator	14	5

The question was analyzed with the UHB technical team and from these discussions came out a resulting standard period for all variables, defined in a fixed time interval of 30 s.

Once all quality problems were resolved, the data started to be regularly collected. By the end of this module, the result was a consistent dataset, without outliers and without missing values and with all variables on the same time scale (timestamp).

As a final comment, it could be highlighted that previously to this study, most of the data collected in UHB was just used as an input for computation and presentation of operation indicators on dashboards in the plant's supervisory system. The data, after use, was then discarded. There were no historical data and, consequently, no analytical treatment of the data. This was changed with this project.

Today, all variables have their historical data kept in the database of the supervisory system, for a period of 6 months. At the same time, there is now a data warehouse, built on a separate data server, where new data is continuously incorporated into the historical series of the variables, which can now be increased for an almost indefinite period. It is a very different scenario. In fact, it would be reasonable to consider that the results of this module 2, mainly the ETL procedure, the exclusive data server and the data warehouse were successful, being, thus, relevant contributions of this work and that could be followed in other applications.

5.3. Data Analysis Results

The data analysis was based on the analytical workbench, illustrated in Figure 5, developed specifically for this framework. This application was fundamental for this analytical module. Before describing the workbench, it must be remembered that it was designed for a Brazilian company and, therefore, for Brazilian users. Thus, all labels and titles in the application were defined in the Portuguese language. The figures presented here in this paper are maintaining the original screens of the tool; however, this aspect should not affect the understanding of the tool, in relation to its functionalities, since a detailed description of each one will be provided.

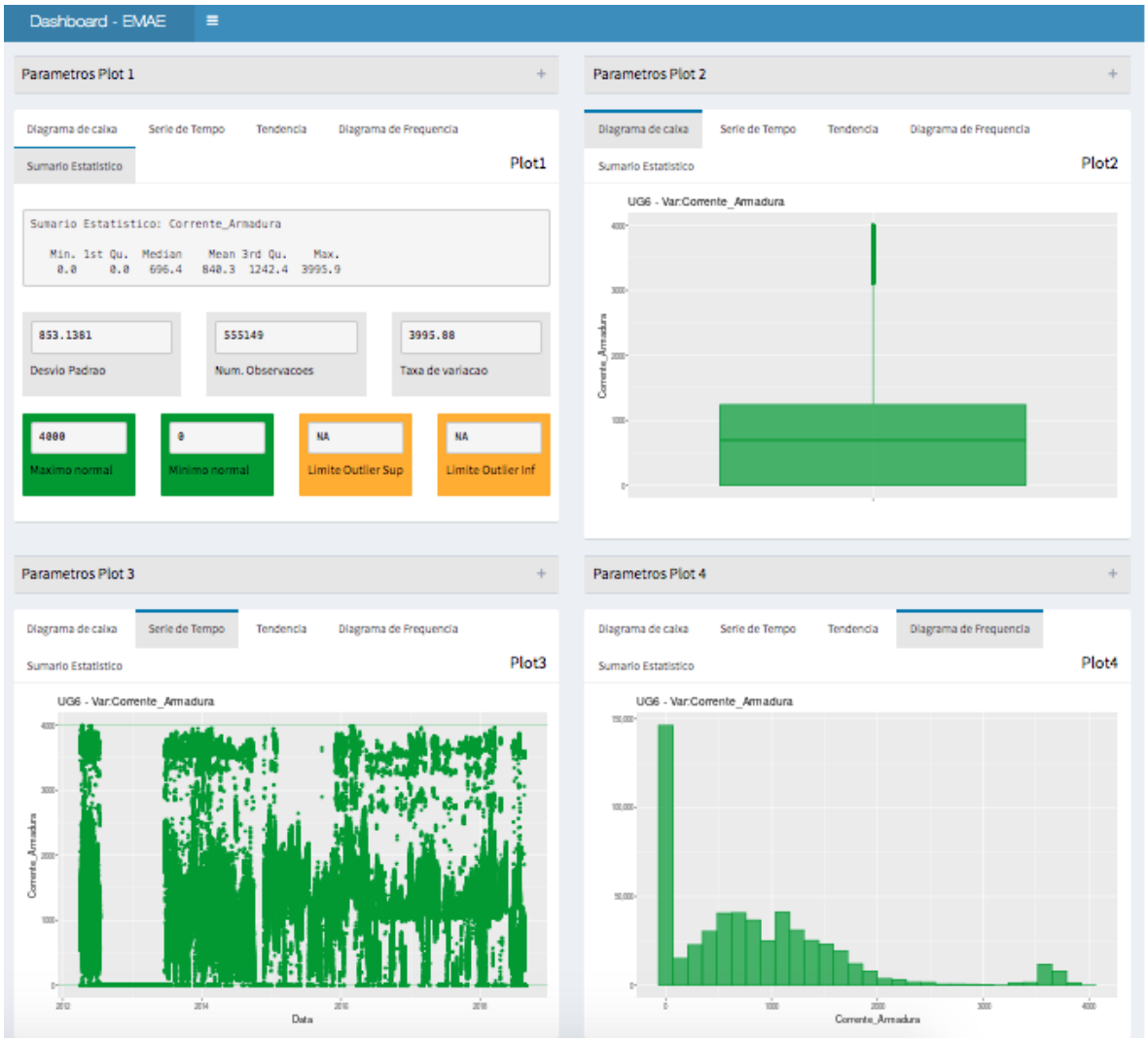

Figure 5. Variable "Armor Current"—Exploratory Analysis in the Analytical Workbench.

The analysis could be developed from many angles. The tool provided the filtering of a generating unit, a specific subsystem and a variable of that subsystem. To select the variables the tool follows the hierarchy, starting at the system, going through its subsystems, then, its components and finally, the variables. At any hierarchical level, an analysis can be defined.

Once the analysis parameters are defined, four types of outputs can be viewed: a statistical summary, a boxplot diagram, a time series plot and a histogram. The user can select all these features to analyze a single variable or one of them for a comparative analysis among variables.

The analysis results are presented by subdividing the screen into quadrants, and in each of the quadrants one of these four types of outputs is presented. Therefore, the output interface is, in fact, a dashboard, combining graphical visualizations and with statistical measures.

As in Section 5.1, component "Stator Armature" will be used here, once more, to demonstrate the tool. An analysis of one of its critical variables, the "Armor Current", will be shown. An exploratory analysis of this variable is represented in the four blocks of Figure 5, in which all possibilities of statistical metrics and graphic analyses can be visualized.

From that Figure 5 one analysis that can be done for the "Armor Current" variable (in Portuguese *VarCorrente_Armadura*), is based on the boxplot (see Plot 2, graphic at superior right in Figure 5), where it can be seen a group of values between 3000 and 4000, considering the scale of the vertical axes, distorting the visualization of a potential outlier. However, from the statistical summary (see Plot 1,

graphic at superior left in Figure 5, the maximum expected value (in Portuguese: Máximo Normal) is 4000, indicating that there were no outliers in these data. A confirmation can be obtained by the maximum observed value of the variable which happens in the situation of high energy generation. The time series diagram and the histogram plot (Plot 3 and Plot 4, respectively from left to right in the bottom of Figure 5) also provide relevant information for analysis. In the case of plot 3, it is possible to identify the period in which the maximum value was reached and the histogram (plot 4) shows the distribution of observed values. In this case, value zero has the largest frequency, which stands the period when the GU was off for maintenance. Another type of analysis can be seen in the Figure 6, which shows the "Rotor Groove Temperature", another variable of the Stator Armature component. A comparison of the behavior of the variable in GU4 and GU6 is presented through the Boxplot and Time Series diagrams. The graphs show very similar behavior, as expected, even though the boxplot shows less dispersion for GU4, the green one. A complement to this comparison was developed based on the frequency histograms (Figure 7) and the same behavior was detected. With histograms, it can be seen more clearly the dispersion of the data.

This type of comparative analysis between the two generating units was developed for all variables, whenever data from both generating units were available.

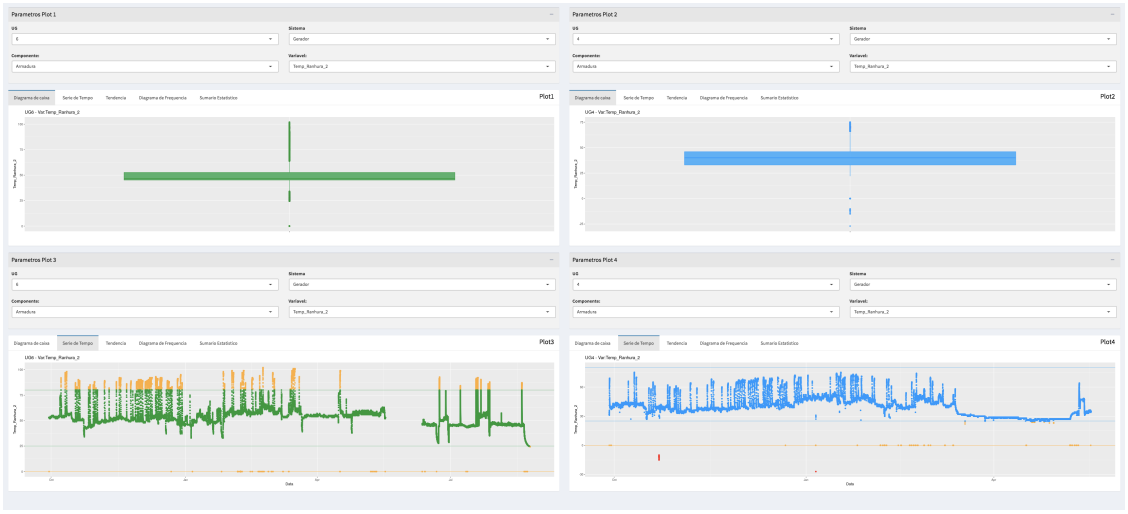

Figure 6. Rotor Groove—Comparison GU4 vs. GU6—Boxplot and Time Series.

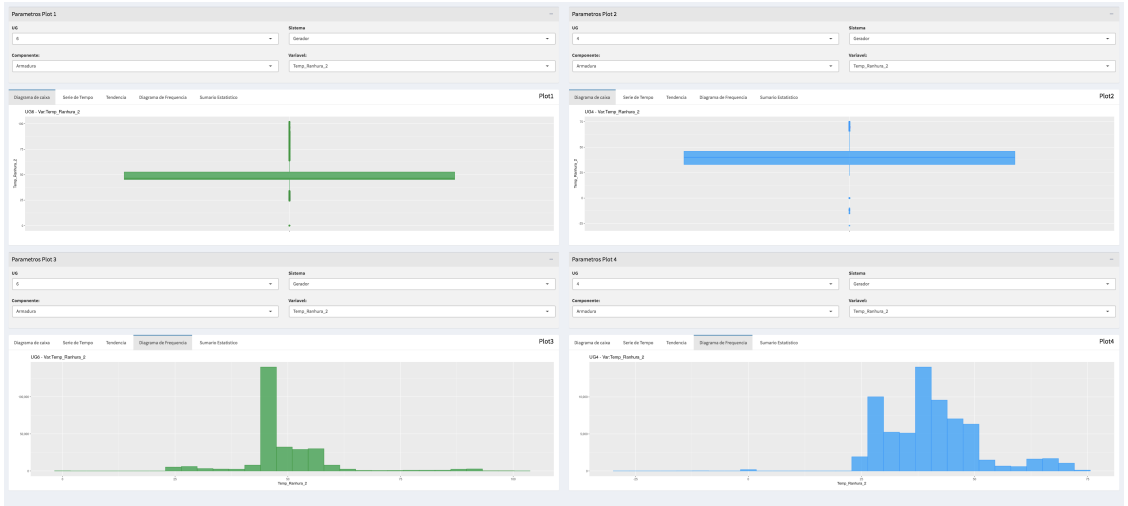

Figure 7. Rotor Groove—Comparison GU4 vs. GU6—Boxplot and Histogram.

A final analysis developed for all variables is illustrated in Figure 8, a time series decomposition, in this case for the Armature current. This is an important analysis, since when we look to the original data, many times we do not see certain behaviors as they are obscured by random effects. The time series decomposition shows three components of the series: the tendency, seasonality, and random effect (remainder). Through these decomposition tendencies, seasonality effects became much clearer, giving the stakeholders important information for the decision. These effects are clear in Figure 8. The figure shows in its upper part the original data. Then it presents the trend and seasonality curves in sequence. In addition, in its lower part it presents the random component (remainder). It is perfectly possible to see in the trend curve that in two moments in time the variable showed a growth trend, which was later reversed. Regarding seasonality, it can be seen that there are reasonably well-defined cycles, in which peaks occur. In addition, these peaks are reflected in the original data curve, as can be seen.

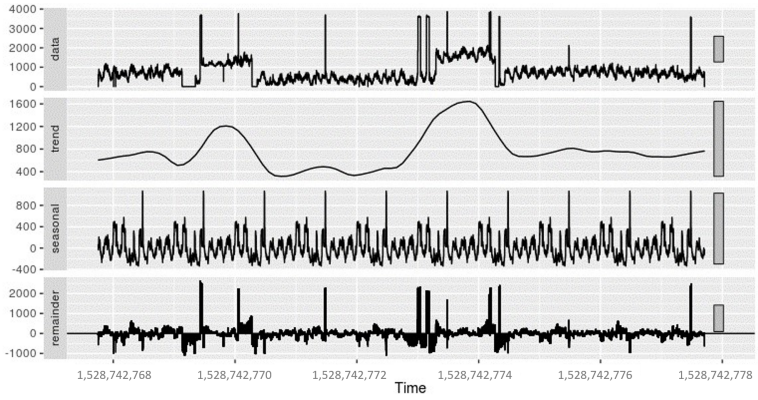

Figure 8. Armature Current—Time Series Decomposition.

Some important results were obtained in this module, and for this, the role of the analytical workbench in the developed analyses must be highlighted, not only in this Module 3 but also in Module 2, as already reported, having contributed significantly to the identification of anomalies associated with critical variables.

As stated earlier, the section showed some examples with focus on the Stator Armature component, but the figures and discussions presented in this section are just an illustration of the analysis developed. In fact, the entire set of variables was submitted to an exploratory analysis, which was a comprehensive and extensive work. Indeed, more than 750 diagrams have been generated, including dashboards of the types presented above, graphs of time series decomposition, and other types of diagrams involving comparisons and correlations among variables.

This extensive analysis provided a reasonably deep knowledge about the behaviors of the variables, individually and as a system. At the end of this module, the technical team was confident that the accumulated knowledge about the system at UHB was robust and that they could move on to the next module to develop predictive modeling, discussed in the next section.

5.4. Predictive Modeling Results

Before describing this last module of the framework is important to point out once more, the general objective of this application, which is to support the predictive maintenance decisions, by a minimization of corrective maintenance occurrence, and so, the objective of this application is to support predictive maintenance decisions, identifying when an equipment is in the imminence of a failure.

The predictive modeling was subdivided in two modules:

A first part of this application is a predictive model to identify the variables significantly associated with equipment failures, and, by the values of these significant variables, whether the equipment is close to a failure point or not.

The second part of the application is a time series forecasting model, developed to predict the values of the significant variables, and, thus, to predict the probability of an equipment failure in a future period.

The literature review presented in Section 2 shows an intense use of algorithms such as Artificial Neural Network, Support Vector Machine and others considered to be black box [41]. This term, black box, is used because the algorithm's decision criterion is not visible to humans. In this research, it was preferred for its first model, to use an AI algorithm more similar to a white box [41], where decision-making based on the algorithm, can be more easily interpreted by experts, and even by laypeople. In this sense, one of the algorithms employed in the predictive modeling, part 1, of BAProM was the Decision Tree (DT) and for the time series forecasting model was employed an ANN model.

The choice in both cases was technical, but in the first model it was also defined to better take advantage of the experience of the UHB team, who could better visualize the results and who, with their experience, with solid knowledge of the causes of maintenance, could better assess the outputs of the model. However, this model, as well as the whole framework, can be used for any industrial system.

Once again, the "Stator Armature", one of the components of the Generator subsystem (Figure 4), was used to illustrate the experiments carried out with DT, which are presented in this subsection. As mentioned before, the Stator Armature has five critical variables: Active Energy, Armature Tension, Armature Current, Active Power and Rotor Groove temperature. More than one DT were built to contemplate different possibilities of variables combinations. The DT presented here shows one of these DT, which included the three first variables:Active Energy, Armature Tension, Armature Current.

For this experiment, data from 2017 to 2018 were used, equivalent to 5000 samples. The data for training the model were classified with two types of labels: OP (operating normal) and CM (corrective maintenance), defined for data collected 15 days before and after the maintenance. The modeling of the DT algorithm follows the well-known cross-validation process, with 5-fold [37]. The resulting DT Model for this experiment is illustrated in Figure 9.

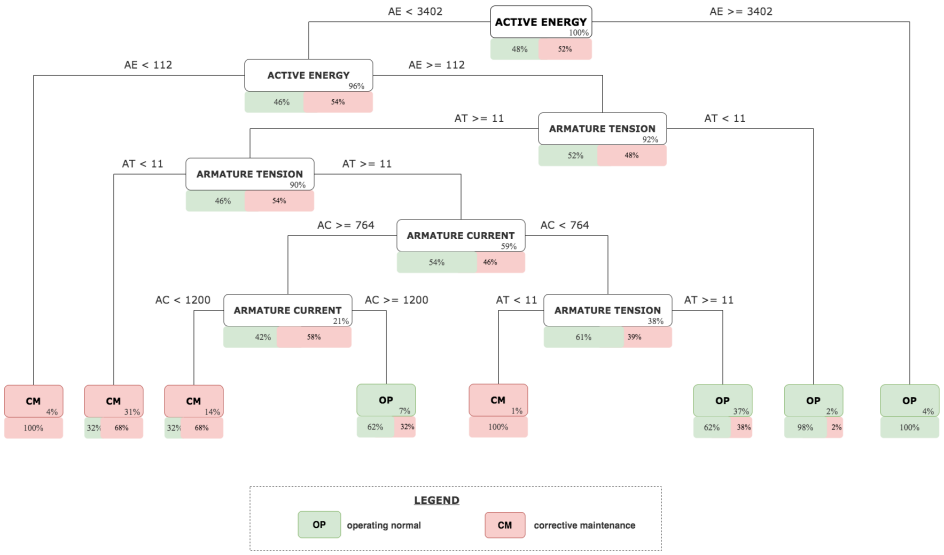

Figure 9. Predictive Decision Tree Model.

Once the decision tree has been generated, from the application of the technique to the data, it is possible to analyze the model output and identify the leaves whose classifications presented the best results in terms of accuracy and quantity, and retrace the path from these leaves to the root node to identify the rules that generated these leaves. With this knowledge of the rules, it will be possible

to establish the degree of importance of each variable, defining those that deserve closer observation. Please note that the model provides the ranges of values and probabilities for each variable, which lead to the conditions of OP or CM. These allows the monitoring of these variables so that when reaching these thresholds, an alarm may be triggered to evaluate the possibility of having a maintenance stop before a failure occurs, generating corrective maintenance. The model furnishes, yet, for each node the percentage of observations of the dataset. In terms of performance, the accuracy of the different DT models developed, ranged from 70% to 96%. In addition, in this specific case, a qualitative analysis of the variables at the different levels of the decision tree was developed by UHB specialists, who agreed with the results, which showed the degrees of importance of the variables in the maintenance decision. The DT model, therefore, showed to be a consistent predictive maintenance tool, supporting the decision-making of scheduling equipment stops.

Complementing the DT model, a second type of predictive model based on ANN was developed to forecast the critical variables that would need to be monitored. Thus, in addition to monitoring the actual value of a variable, one can also identify in a future period, when one of these variables would reach a threshold that could lead to an equipment failure. An MLP—Multilayer Perceptron neural network was employed in this model and the forecasting results for the variable "Active Energy" are presented in Figure 10. In that figure, time is expressed in 30-s intervals, which were the time intervals used in data collection and there is a trend curve projected for the future, also showing the curves of the lower and upper limits, of confidence intervals for the forecasts. The intervals are presented for two confidence levels: 80% level, with a narrower range, and a 95% confidence level.

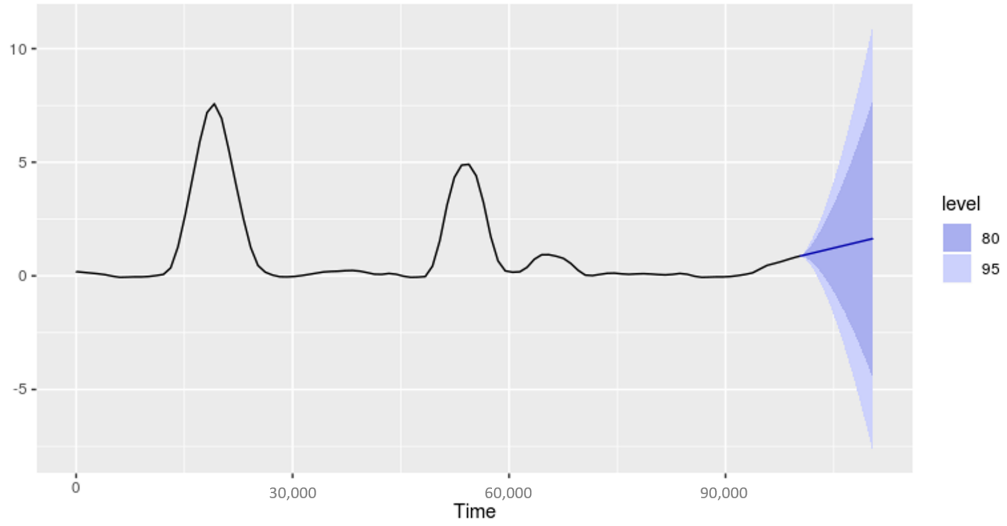

Figure 10. Active Energy Forecasting.

It should be noted that with the two predictive models working together, it can be said that there is a predictive modeling with reasonable robustness, once it can have reference parameters for monitoring the variables in real time, triggering preventive actions every time that a critical variable enters a level of equipment failure; and at the same time, there is an implementation of an effective instrument to project this type of situation for some time in the future, providing even more time, so that the operation teams can prepare and/or prevent such occurrences.

As said before, the research was conducted at UHB in a real-world environment. Therefore, the predictive model described here was tested with real data of UHB's operation. As previously stated, the data used in this study varied from May 2017 to January 2018. Therefore, to validate the model, what was done was to use data from the first months of this period to predict occurrences of failure for the final months of this period. In addition, since the actual data from these forecast months were known, it was possible to compare the predictions made by the model with the actual occurrences. The model was able to identify most of the failures that could have been avoided and to identify

maintenance that could have been reprogrammed. These predictions would result in cost savings and productivity increasing.

6. Conclusions and Future Work

This paper presented an application of a methodological proposal, expressed by the framework BAProM—Big Data Analytical Process and Mapping—which sought to contemplate all the phases of a KDD process, from the mapping of processes and critical variables, going through data management, exploratory analysis and even implementing predictive models. The complete framework has been tested in a real-world application in an industrial environment, making it possible to validate and demonstrate its practical feasibility. This real-world application started with the mapping of the entire operational process of the plant and an ETL procedure. Next, a data analysis tool, an "analytical workbench", was developed and implemented. This workbench has been shown to be suitable for different types of analysis, such as pre-processing or exploratory analysis. The tool has multiple possibilities for graphical analysis and statistical metrics computation, in addition to allowing monitoring of system variables, indicating anomalous behavior. It was used in the pre-processing phase and in exploratory analyses with satisfactory results.

A predictive model was developed, based on decision trees, which allowed the identification of more relevant variable thresholds, indicating the imminence of an equipment failure which consequently allows the programming of a predictive maintenance, avoiding unplanned stops for corrective maintenance. The predictive model made it possible to implement a management process for critical variables. Operators can act before an interruption event occurs. The whole process proved to be effective and efficient, given the feasibility of its implementation in a real-world operation.

In addition, a time series forecasting model for these critical variables, based on ANN, was also designed and implemented, which made the process even more effective, since managers can have information on future times when these variables should reach their thresholds, leading to the need for corrective maintenance. The forecasts provide additional time for teams to act, avoiding unexpected equipment stops.

The main conclusions of the research can be expressed as follows:

(a) The phases and tools proposed in the framework proved to be well suited to an industrial process, allowing it to pass effectively through all stages of a big data process.
(b) The process and variable mapping, the first phase of the framework, is a novelty proposed in the research which proved to be a fundamental step. The knowledge obtained in this phase about the entire operation under study was an essential driver for the following phases, mainly to define the most relevant variables to be analyzed.
(c) The development of a computational tool focused on data exploration was essential to support the pre-processing of the data and also the analytical phase, where the behaviors of the variables and their interrelations are identified. The dashboard developed in the project was fundamental to identify non-standardized behaviors in some variables, as well as to identify reference parameters used to propose patterns for monitoring variables.
(d) A predictive model, based on a decision tree, proved to be well suited to identify, with reasonable accuracy, the critical variables that lead to equipment failures and to predict the limit values of those variables that can cause a failure. An additional advantage of this model is that as a decision tree is a "white box" model, the rules identified by the technique are totally clear and known, which favors an implementation to trigger alerts, whenever a threshold of a critical variable is reached.
(e) Another point to be highlighted is that in order to have a projection of the future, the decision tree model must be complemented by a model for forecasting the critical variables considered in the tree. Thus, it is possible to identify in a future period when one of these variables would reach a threshold, leading to equipment failure. An ANN prediction model for such variables proved to be an effective alternative.

Despite the positive points of this framework, it must be considered that there are some limitations that should be considered in future studies and projects. One of these improvements concerns ETL, which relies on operational personnel to transfer production data to a repository dedicated to analytic. This process could be automated. Another limitation refers to data pre-processing in which part of the work is done by inspecting the variables with the support of the dashboard. Some of these tasks could also be automated. Furthermore, the dashboard could be improved by automatically generating some standard graphics and metrics to all or to a group of variables.

Moreover, regarding future works, it would be important to implement on the dashboard the critical values identified in the predictive decision tree model, so that alarms would be automatically triggered without the need for human monitoring when one of those variables is close to those values.

Another opportunity for future work is the application of this methodology in other industrial systems, including other subsystems of the case studied. Finally, one can also develop a validation of the results obtained through decision trees with other types of predictive models, such as artificial neural networks and support vector machines.

Author Contributions: Conceptualization, L.A.S., A.R.d.A.V.F.; methodology, L.A.S., A.R.d.A.V.F. and G.G.d.C.C.; software, G.G.d.C.C. and M.V.B.d.A.V.; validation, L.A.S., A.R.d.A.V.F. and L.S.d.S.; formal analysis, M.V.B.d.A.V. and G.G.d.C.C.; investigation, M.V.B.d.A.V. and G.G.d.C.C.; resources, L.A.S. and L.S.d.S.; data curation, M.V.B.d.A.V. and G.G.d.C.C.; writing–original draft preparation, G.G.d.C.C., L.A.S. and A.R.d.A.V.F.; writing–review and editing, A.R.d.A.V.F., L.A.S., G.G.d.C.C. and M.V.B.d.A.V.; visualization, G.G.d.C.C. and M.V.B.d.A.V.; supervision, L.A.S.; project administration, A.R.d.A.V.F.; funding acquisition, L.S.d.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is a part of the R&D project "EMAE-ANEEL-P&D 00393-0008/2017", funded by EMAE—-Metropolitan Company of Water & Energy, of the state of São Paulo, Brazil.

Acknowledgments: We thank all the EMAE staff who participated in the R&D project "EMAE—ANEEL-P&D 00393-0008/2017", and all the faculty and student members of the BigMAAp research lab at Mackenzie Presbyterian University.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
2. Lavalle, S.; Lesser, E.; Shockley, R.; Hopkins, M.; Kruschwitz, N. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Manag. Rev.* **2011**, *52*, 21–32.
3. Lee, J.; Kao, H.A.; Yang, S. Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. *Procedia CIRP* **2014**, *16*, 3–8. [CrossRef]
4. Li, Z.; Wang, Y.; Wang, K.S. Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. *Adv. Manuf.* **2017**, *5*, 377–387. [CrossRef]
5. Rigatos, G.; Siano, P. Power transformers' condition monitoring using neural modeling and the local statistical approach to fault diagnosis. *Int. J. Electr. Power Energy Syst.* **2016**, *80*, 150–159. [CrossRef]
6. Márquez, F.P.G.; Tobias, A.M.; Pérez, J.M.P.; Papaelias, M. Condition monitoring of wind turbines: Techniques and methods. *Renew. Energy* **2012**, *46*, 169–178. [CrossRef]
7. Bousdekis, A.; Papageorgiou, N.; Magoutas, B.; Apostolou, D.; Mentzas, G. A Proactive Event-driven Decision Model for Joint Equipment Predictive Maintenance and Spare Parts Inventory Optimization. *Procedia CIRP* **2017**, *59*, 184–189. [CrossRef]
8. Ahmad, R.; Kamaruddin, S. An overview of time-based and condition-based maintenance in industrial application. *Comput. Ind. Eng.* **2012**, *63*, 135–149. [CrossRef]
9. Froger, A.; Gendreau, M.; Mendoza, J.E.; Pinson, É.; Rousseau, L.M. Maintenance scheduling in the electricity industry: A literature review. *Eur. J. Oper. Res.* **2016**, *251*, 695–706. [CrossRef]
10. Stenström, C.; Norrbin, P.; Parida, A.; Kumar, U. Preventive and corrective maintenance – cost comparison and cost–benefit analysis. *Struct. Infrastruct. Eng.* **2015**, *12*, 603–617. [CrossRef]

11. Nasr, A.; Gasmi, S.; Sayadi, M. Estimation of the parameters for a complex repairable system with preventive and corrective maintenance. In Proceedings of the 2013 International Conference on Electrical Engineering and Software Applications, Hammamet, Tunisia, 21–23 March 2013.
12. Arno, R.; Dowling, N.; Schuerger, R. Equipment failure characteristics & RCM for optimizing maintenance cost. In Proceedings of the 2015 IEEE/IAS 51st Industrial & Commercial Power Systems Technical Conference (I&CPS), Calgary, AB, Canada, 5–8 May 2015.
13. Sheut, C.; Krajewski, L.J. A decision model for corrective maintenance management. *Int. J. Prod. Res.* **1994**, *32*, 1365–1382. [CrossRef]
14. Liu, B.; Xu, Z.; Xie, M.; Kuo, W. A value-based preventive maintenance policy for multi-component system with continuously degrading components. *Reliab. Eng. Syst. Saf.* **2014**, *132*, 83–89. [CrossRef]
15. Yin, S.; Ding, S.X.; Xie, X.; Luo, H. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. *IEEE Trans. Ind. Electron.* **2014**, *61*, 6418–6428. [CrossRef]
16. Jing, C.; Hou, J. SVM and PCA based fault classification approaches for complicated industrial process. *Neurocomputing* **2015**, *167*, 636–642. [CrossRef]
17. Yin, Z.; Hou, J. Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes. *Neurocomputing* **2016**, *174*, 643–650. [CrossRef]
18. Yan, J.; Meng, Y.; Lu, L.; Li, L. Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance. *IEEE Access* **2017**, *5*, 23484–23491. [CrossRef]
19. Civerchia, F.; Bocchino, S.; Salvadori, C.; Rossi, E.; Maggiani, L.; Petracca, M. Industrial Internet of Things monitoring solution for advanced predictive maintenance applications. *J. Ind. Inf. Integr.* **2017**, *7*, 4–12. [CrossRef]
20. Gatica, C.P.; Koester, M.; Gaukstern, T.; Berlin, E.; Meyer, M. An industrial analytics approach to predictive maintenance for machinery applications. In Proceedings of the 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA), Berlin, Germany, 6–9 September 2016.
21. Vallim-Filho, A.R.A.; Okido, P.; Silva, L.A.; Vallim, M.V.B.A.; Silva, L.S. Data Dimensionality Reduction based on Variables Clustering. *XI Int. Stat. Congr.* **2019**, *1*, 1–10.
22. Sampaio, G.S.; de Aguiar Vallim Filho, A.R.; da Silva, L.S.; da Silva, L.A. Prediction of Motor Failure Time Using an Artificial Neural Network. *Sensors* **2019**, *19*, 4342. [CrossRef]
23. Wang, N.; Sun, S.; Si, S.; Li, J. Research of predictive maintenance for deteriorating system based on semi-markov process. In Proceedings of the 2009 16th International Conference on Industrial Engineering and Engineering Management, Beijing, China, 21–23 October 2009.
24. Gao, Z.; Cecati, C.; Ding, S.X. A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3757–3767. [CrossRef]
25. Carvalho, T.P.; Soares, F.A.A.M.N.; Vita, R.; da P. Francisco, R.; Basto, J.P.; Alcalá, S.G.S. A systematic literature review of machine learning methods applied to predictive maintenance. *Comput. Ind. Eng.* **2019**, *137*, 106024. [CrossRef]
26. Zenisek, J.; Holzinger, F.; Affenzeller, M. Machine learning based concept drift detection for predictive maintenance. *Comput. Ind. Eng.* **2019**, *137*, 106031. [CrossRef]
27. Nguyen, K.T.; Medjaher, K. A new dynamic predictive maintenance framework using deep learning for failure prognostics. *Reliab. Eng. Syst. Saf.* **2019**, *188*, 251–262. [CrossRef]
28. Sahal, R.; Breslin, J.G.; Ali, M.I. Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case. *J. Manuf. Syst.* **2020**, *54*, 138–151. [CrossRef]
29. Hu, J.; Chen, P. Predictive maintenance of systems subject to hard failure based on proportional hazards model. *Reliab. Eng. Syst. Saf.* **2020**, *196*, 106707. [CrossRef]
30. Moro, S.; Laureano, R.; Cortez, P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In Proceedings of the European Simulation and Modelling Conference-ESM'2011 (EUROSIS-ETI), Guimaraes, Portugal, 24–26 October 2011; pp. 117–121.
31. Recker, J. Opportunities and constraints: The current struggle with BPMN. *Bus. Process. Manag. J.* **2010**, *16*, 181–201. [CrossRef]
32. Völzer, H. An overview of BPMN 2.0 and its potential use. In *International Workshop on Business Process Modeling Notation*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 14–15.

33. Park, G.; Chung, L.; Zhao, L.; Supakkul, S. A goal-oriented big data analytics framework for aligning with business. In Proceedings of the 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), San Francisco, CA, USA, 6–9 April 2017; pp. 31–40.
34. Nalchigar, S.; Yu, E. Conceptual modeling for business analytics: A framework and potential benefits. In Proceedings of the 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, Greece, 24–27 July 2017; Volume 1, pp. 369–378.
35. Abdelsalam, H.M.; Shoaeb, A.R.; Elassal, M.M. Enhancing Decision Model Notation (DMN) for better use in Business Analytics (BA). In *Proceedings of the 10th International Conference on Informatics and Systems*; ACM: New York, NY, USA, 2016; pp. 321–322.
36. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer: Berlin/Heidelberg, Germany, 2015.
37. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011.
38. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **1996**, *17*, 37–37.
39. Turban, E.; Sharda, R.; Aronson, J.E.; King, D. *Business Intelligence: A Managerial Approach*; Pearson Prentice Hall: Corydonê, IN, USA, 2008.
40. Elmasri, R.; Navathe, S. *Fundamentals of Database Systems*; Pearson: London, UK, 2017; Volume 7.
41. Abdallah, I.; Dertimanis, V.; Mylonas, H.; Tatsis, K.; Chatzi, E.; Dervili, N.; Worden, K.; Maguire, E. Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data. In *Safety and Reliability–Safe Societies in a Changing World*; CRC Press: Boca Raton, FL, USA, 2018; pp. 3053–3061.
42. Haykin, S.O. *Neural Networks and Learning Machines*, 3rd ed.; Pearson: London, UK, 2008.
43. Grein, H.; Lorenz, M.; Angehrn, R.; Bezinge, A. Inspection periods for Pelton runners. *Water Power Dam Constr.* **1985**, *37*, 49.