SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA

Ing. Jela Abasová

Offprint of Dissertation Thesis

Big Data – knowledge discovery from heterogenous data storages – implementation of best practices

for being awarded the academic degree Doctor ("philosophiae doctor", abbrev. as "PhD.")

in PhD. study programme: Process Automation and Informatization

in study branch: Cybernetics

form of study: full-time

Place and date: Trnava, 17.08.2021

Dissertation Thesis was elaborated at: Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava

Submitted by: Ing. Jela Abasová
Jána Bottu 25
917 24 Trnava

Supervisor: Prof. Ing. Pavol Tanuška, PhD.
Jána Bottu 25
917 24 Trnava

Reviewers: .....
.....
.....

.....
.....
.....

Offprint was sent on: .....

Defence of Dissertation Thesis will be on...... **at** .....**o´clock**.

at .....

.....
Rector of STU or Faculty Dean
if PhD. study programme is executed on the Faculty
(name, surname and degrees)

ABSTRACT

ABASOVÁ, Jela: *Big Data – knowledge discovery from heterogenous data storages – implementation of best practices*. [Dissertation thesis]– Slovak University of Technology in Bratislava. Faculty of Materials Science and Technology in Trnava; Institute of Applied Informatics, Automation and Mechatronics – Supervisor: prof. Ing. Pavol Tanuška, PhD.– Trnava: MTF STU, 2021, pp. 193.

Abstract: CRISP-DM (Cross-industry standard process for data mining) methodology was developed as an intuitive tool for data scientists to help them with applying Big Data methods in a complex technological environment of Industry 4.0. The review of numerous recent papers and studies uncovered that most of papers focus either on application of existing methods in case studies, summarizing existing knowledge, or developing new methods for a certain kind of problem. Although all these types of research are productive and required, I identified lack of complex best practices for a specific field. Therefore, my goal is to propose best practices for the data analysis in production industry. The foundation of the proposal is based on three main points: the CRISP-DM methodology as the theoretical framework; the literature overview as an expression of current needs and interests in the field of data analysis; and case studies of projects I was directly involved in as a source of real-world experience. The results are presented as lists of the most common problems for selected phases ('Data Preparation' and 'Modelling'), proposal of possible solutions, and diagrams for these phases. These recommendations can help other data scientists avoid certain problems or choose the best way to approach them.

Keywords: Big Data; Industry 4.0; CRISP-DM; predictive maintenance; production industry

ABASOVÁ, Jela: *Big Data – získavanie znalostí z heterogénnych dátových úložísk – implementácia osvedčených postupov*. [Dizertačná práca]– Slovenská technická univerzita v Bratislave. Materiálovotechnologická fakulta so sídlom v Trnave; Ústav aplikovanej informatiky, automatizácie a mechatroniky – Školiteľ: prof. Ing. Pavol Tanuška, PhD.– Trnava: MTF STU, 2021, 193 s.

Metodológia CRISP-DM („Cross-industry standard process for data mining“) bola vyvinutá ako intuitívny nástroj pre dátových vedcov, aby im pomohla aplikovať metódy Big Data v komplexnom technologickom prostredí Industry 4.0. Prehľad početných súčasných vedeckých článkov a štúdií odhalil, že sa väčšina z nich sústredí buď na aplikáciu existujúcich metód v prípadových štúdiách, sumarizuje existujúce znalosti, alebo vyvíja nové metódy pre určitý typ problému. Hoci všetky tieto typy výskumu sú zmysluplné a žiadané, identifikovala som nedostatok komplexných osvedčených postupov pre špecifickú oblasť. Preto je mojím cieľom navrhnúť osvedčené postupy pre dátovú analýzu v produkčnom priemysle. Môj návrh je postavený na troch hlavných bodoch: na metodológii CRISP-DM čoby teoretickom rámci; na prehľade literatúry, ktorý vyjadruje súčasné potreby a záujmy na poli dátovej analýzy; a prípadové štúdie projektov, do ktorých som sa zapojila, ako zdroj skúseností z praxe. Výsledky sú prezentované ako zoznam najčastejších problémov pre jednotlivé fázy CRISP-DM, návrhy možných riešení, a procesné diagramy postupu riešenia. Odporúčania, ktoré poskytujem, môžu pomôcť iným dátovým vedcom vyvarovať sa určitých problémov alebo si vybrať vhodný prístup k nim.

Kľúčové slová: Big Data; Industry 4.0; CRISP-DM; prediktívna údržba; produkčný priemysel

CONTENT

1 INTRODUCTION

This age is the age of connection: connecting fields of knowledge, which have been developing independently for a long time, but also connecting ideas, cultures, philosophies, and peoples. Globalization is closely linked to the digitalization of the world. More and more actions take place in the virtual reality instead of the physical sphere, the material world is enriched by the extended reality. The progress influences the society as well as individuals, all the inhabitants of the developed parts of the world, and the whole planet.

There is an increasing development in the field of data analysis. The new data sets are bigger and more complex than those recorded in previous periods. They are often unstructured or semi-structured, ideally enriched by metadata. These new sets are frequently measured in terabytes or even petabytes. There is an increasing demand for executing the data in real or near-real time, or for predictive maintenance, which helps to prevent failures, thus significantly decreasing the time and finances needed to fix the problems which have already occurred. The methods for operating data, used in previous periods, are ceasing to be effective. They are not flexible enough and they do not suffice the rising demands anymore.

The traditional statistical methods for operating data are therefore being substituted by more complex and more practical methods collectively known as Big Data, developed specifically for working with huge data volumes. The phases of data processing are described by CRISP-DM methodology, which is generally accepted for its intuitive nature, simpleness, and logicality. This methodology serves as a firm base for data analysists and data scientists.

The analysis, investigation of error sources and searching for hidden patterns is the simpler the better we know the process we are working with, the data sources we draw from, and the overall context. The data scientists cooperate with the experts from the examined field and consult the project's next steps with them. It is advantageous for the data scientists to follow the progress of data analysis, its current fields of interests, and new methods and solutions, as well.

2 LITERATURE OVERVIEW

I performed a systematic literature overview on three digital libraries: ScienceDirect, IEEE Xplore and MDPI. I chose the aforementioned ones because they provide many outputs concerning the topics of interest and I have very good previous experience with these databases. I searched for the following keywords: 'Industry 4.0', 'Big Data', 'CRISP-DM', and 'predictive maintenance'. I focused mostly on recent studies (year 2021) with a small number of older studies used as a comparison of the changes in trends of the topic. I decided to prefer current studies, because the implementation of Big Data into praxis is an issue of recent 3 – 4 years, and problems connected with the implementation have become more evident during that time.

The literature overview was based on 43 papers and studies. A summary of the important aspects of the reviewed studies is presented, as follows (a study can belong to more than one category).

Year of publication:

- 2021: 28 studies;
- 2019-2020: 7 studies;
- 2018 and older: 8 studies.

Main topic:

- Big Data: 20 studies;
- Industry 4.0: 14 studies;
- Predictive maintenance: 8 studies;
- CRISP-DM: 5 studies.

Most recurring additional topics:

- Quality;
- Sustainability;
- Smart factory;
- Internet of things;
- Management.

Type of work:

- Case study: 14 studies;
- Literature review: 14 studies;
- Method(s) forming/comparing: 11 studies;

- Improving/extending methodology: 4 studies;
- Theoretical work: 4 studies.

Most of the authors deal with practical problems and applications of existing methods and methodologies, or with compiling and comparing previous works in the field. Part of these two types of researchers also apply their practical/theoretical knowledge to construct new methods, and significantly less of them focuses on improving the existing methodologies or describing complex best practices. Purely theoretical works are also rare.

3 GOALS

As the literature overview suggested, most of the research focus either on solving a certain problem, summarizing knowledge of the field, or developing a specific solution. All these types of research are productive and required. But I identified lack of complex best practices based on real-world problems. Out of 43 examined projects, only 4 dealt with this topic. Therefore, I decided to fill in this gap and to come up with my own best practices for the field of production industry.

I base the proposal on three main points: the CRISP-DM methodology as the theoretical framework; the literature overview as an expression of current needs and interests in the field of data analysis; and case studies as a source of real-world experience.

The proposal will be presented in a form of Process diagrams of BPMN (business process modeling notation), lists of most common errors, possible solutions to them, and other recommendations.

To achieve this goal, the following steps need to be taken:

1. Literature overview:

When performing the literature overview, I divided my area of interest into four partially overlapping topics: Industry 4.0, Big Data, CRISP-DM, and predictive maintenance. All of them are closely connected to the main topic of this research. Overwhelming majority of the review papers and studies focus on production industry and related fields, but briefly deal also with other fields where Big Data and CRISP-DM methodologies are currently being used, aimed on the up-to-date trends.

2. Case studies:

There are seven real-world projects from the autor's praxis used as a practical base for the paper, and they will be described in detail. All the projects deal with problems from real companies and represent important factors in forming the best practices.

3. Solutions described on case studies for the phases of CRISP-DM:

At least two projects (case studies) together with the practical solution described in details will be chosen for every phase of CRISP-DM. Every project will be used at least once in the whole process, and some will have several instances. There, the practical problems will be demonstrated in order to extract the knowledge later to make it more general.

4. Best practices for each of the selected phases:

The solutions of all the projects for a phase will be combined and generalized to form best practices for the phase. This proposal will include a list of problems typical for the phase, recommendations, and diagrams.

5. Best practices for the whole process:

The overall best practices will be formed out of combination of best practices of the phases. Apart from typical problems, their solutions, and diagrams, it will also include applicability of the best practices, as well as industry production specifics description and possible transferability to other areas.

4 MATERIALS AND METHODS

I was using several software tools when performing data analysis. Most often I used the following ones:

- RapidMiner – for preprocessing during 'Data Preparation' phase, for various analyses during 'Data Understanding', 'Modelling' and 'Evaluation' phases; [79]
- Python – universal tool helpful in data acquisition during 'Data Understanding' phase, numerous preprocessing and analysis tasks during 'Data Preparation', 'Modelling' and 'Evaluation' phases, and visualization of results during 'Deployment' phase. [80]

- Elasticsearch – for visualizing data during ‘Deployment’ phase and for operating extremely large datasets (unmanageable by RapidMiner) during ‘Data Understanding’ and ‘Modelling’ phases. [81]

I assembled the documentation from real-world projects from my praxis. I have chosen current projects, thematically similar (from the production industry), but different in their nature (goals of analysis, data acquisition methods, data structure) to serve as a basis for experience extraction and best practices generalization.

I was directly involved in all of the projects I describe in this section as either a data analyst or a data scientist. I participated in the projects during my PhD. studies, including the time I spent on Erasmus+.

The projects introduced below serve as case studies for demonstrating one or more phases of CRISP-DM methodology. Companies’ names (places of project implementation) and other distinguishing details were omitted to keep the confidentiality contract. Also, because of data sensitivity, most of the data were anonymized.

Note that I use plural (we/us/our etc.) when describing the progress in case studies, as the projects were a team work (even though I focus mostly on parts which were my responsibility). I return to singular (I/me/my etc.) when drawing conclusions.

4.1 Screwing heads

The goal of the project was to predict production failures on automated screwing heads, therefore lowering failure rate on the screwing mechanism. Failures were connected to violation of established boundaries of key parameters: adhesive pressure, axial torque, and screwing depth. There were other available parameters of the process: quality and chemical constitution of material, its thickness, position of screw, etc.

Data sources:

- Screwing heads failures;
- Failures of the connecting technology attached to the screwing heads;
- Measure points;
- Product moves.

The last data source (‘product moves’, containing timestamps for various points of the product occurrence during the production process) was originally not meant to be used at all. However, it had to be included early after the start of ‘Data Preparation’ phase, since there

was no identifier in the data common for all the sources and, furthermore, they were not time-synchronized, making joining of the sources impossible. The ‘product moves’ data source contained enough information to connect it with all the three other sources.

Specific problems arose due to multilingual nature of the company. Attribute names were often listed as a mixture of two or more languages, and the same applied to notes written manually by workers. This made some words illegible when using specific characters (such as Slovak letters with diacritics), particularly when transferring from one data source to another. Another problem connected to the issue was mixed usage of decimal point and decimal comma. In some cases, also dates were affected by multilingualism: date formats ‘yyyy-MM-dd’ and ‘yyyy-dd-MM’ were used inter-changeably and on some occasions, some of the records were accepted in incorrect format, while others were identified as text by analytical tools. Another minor (but still important) issue was the numerical values being recorded together with their units (such as [Nm] in screwing moment). If this was omitted during the preprocessing phase, the attribute was often misclassified as textual during data modelling phase, while it was clearly a numerical attribute with wrong formatting.

After successful preprocessing, in the ‘Modelling’ phase the appearance of various error types during individual weeks was analyzed on six-month period sample. We realized that the major problem of the screwing process was the screwing technique, so after that we focused mostly on ‘Errors of the connecting technology’ data source, which contains more detailed description of the attributes evaluated during the process. Further analysis was applied on ten robots with the biggest number of errors.

Part of results of this project was published as a conference paper ‘Proposal of data pre-processing for purpose of analysis in accordance with the concept Industry 4.0’ in Springer and presented on Advances in Intelligent Systems and Computing conference, 2019.

4.2 Financial data

The goal of the project was investigating correlations between various departments of the company in order to optimize financial flows. Important part of the task was to prepare financial data from ERP (Enterprise Resource Planning) level to a form suitable for standard business reports. That included the set-up of regular monthly exports, data preprocessing, and construction of correlation matrices, as to tabularly display relations between company

departments. The project was dealing with a lot of sensitive data, therefore data anonymization was an important part of the task.

Data sources:

- Financial data from various accounts and departments, collected every month during two-years period.

In the ‘Business Understanding’ phase we defined requirements for export from required part of database. Exports were provided in ‘Data Understanding’ phase by the financial department employees. In the first attempt, data were exported via script. The plan was to merge several tables at once. The exported file contained large matrix of data, with too complex structure, and column separator was not integrated correctly. After that, several files were exported, with one account per file, and with names of departments used as attributes (column names) as a result of the second attempt, resulting in a very high number of attributes (more than 700) and making further processing of data quite laborious and impractical. Finally, the requirements for export were redefined one more time. New set of exports was performed, containing multiple files with one department per file. This structure was validated by both us and the management and used in the next phases of the process.

There were three different ways in which accounts can be identified: ID, name, or ID+name. The redundancy of the data set represented the biggest problem (alongside the general issues such as empty rows). Furthermore, while the (anonymized) IDs were consistent, the names of the same account differed in spelling (with or without diacritics) and letter case. These issues were addressed in the ‘Data Preparation’ phase.

Once prepared, the data were used for construction of various correlation matrices in the ‘Modelling’ phase, as to uncover and explain relations between finances of different departments, accounts, for different employees, time periods and sources. In the ‘Deployment’ phase, the results were presented to the management of the company and after incorporating their suggestions, regular exports were set up.

Part of the results was published in conference paper ‘Proposal of effective preprocessing techniques of financial data’ in INES 2018.

4.3. Paintshop

The goal of the project was to integrate data sources in painting process in order to find correlations between painting attributes and their subsequent optimalisation. Painting

process is one of the most complex processes within the company, consisting of several subprocesses, and connected with other processes via inputs and outputs. Inconsistency and insufficient documentation evocate need for describing and visualizing relations between individual components.

Data sources:

- Paint thickness;
- Paint colors;
- Paint structure;
- Paint usage;
- Paint audit;
- Process values;
- Alarms;
- Product moves;
- Technical information;
- Meteorological data.

The company management was interested in integration of the sources concerning ‘Paintshop’. This would be impossible without deeper understanding of sources and the overall process. Every part of the ‘Paintshop’ process was supervised by a different person or people, and just a few of them were in mutual contact, causing data fragmentation. After internal discussion during ‘Business Understanding’ phase we came up with the following steps that were to be taken before moving on to the integration:

First, we had to accumulate all the accessible information about the painting process as a whole and use it to create scheme of the process, as to mark all the subprocesses, their respective data sources, and relations between them within the process flow. Second, we had to make a list of all responsible persons, contact them and gain as much documentation and information about the sources as possible. Third, we had to describe each of the sources and decide about their usability within the integration.

The data were heterogenous in both quantity and availability. Only four of the sources offered regular exports. Most data could be obtained only via one-time exports by a third person, who could provide us with the required data only after we first obtained permission from the management of the company. This prolonged the data-obtaining process and rapidly diminished possibilities of real-time monitoring and analysis. Each data source had to undergo data preparation phase individually. Two of them may serve as examples here.

The ‘paint thickness’ data source contained information about paint thickness, measured at several points of the car body. Its structure proved as extremely problematic for the further analysis, as the key attributes were situated in the first column, what precluded analyzing multiple car bodies at a time. To handle this, we designed a script via Microsoft VBA in Microsoft Excel environment, providing following tasks: transposing key attributes from rows to columns, creating new columns and names of attributes, computing min, max, and average from the measured values, recording the values into new attributes, and merging all the Excel sheets.

Meteorological data were obtained from a small meteorological station placed in the company, monitoring several weather attributes – meteodata – such as temperature, humidity, and wind speed. The problem with the data source was inconsistency in frequency of recording. While in 2018 (the year when we were working on the project) the data were recorded every 5 minutes, the historical data (2015 – 2017) were much less frequent, being recorded only once an hour. Therefore, we proposed a process in RapidMiner, having three basic steps: generating new time units attributes (year, month, day, hour) from the more-detailed (newer) data, aggregating newer data by all the time units and averaging their values, and merging the two sets.

Compatible time interval represented the main criterion for data integration on the selected sources and possibility to connect them all in real time to support predic-tive maintenance, and therefore improve quality of the painting process was set out as the ultimate objective.

Part of the results was published as a conference paper ‘Proposal of data preparation model for Big Data analytics in painting process’ in ELECTRO 2020 proceedings and presented on the conference.

4.4 Formal concept analysis

Goal of the project was to design an approach for joining formal concepts [83], calculated from separated parts of a formal context. There were two variants to be considered: a formal context divided into two relatively equally big parts, and a formal context divided into two disproportional parts (where the second, smaller part is being added iteratively). We disposed with a code for computation of a formal concept from a formal context (ideal due

to given criteria). Unlike other projects, this one was not bound to a process in a company, it was more abstract, with possible application in big area of fields.

Data sources:

- Matrices comprising formal contexts;
- Script for building formal concepts. [82]

There are a few possible ways how to address the problem. One of our inputs is, by definition, a set of pertinent concepts. The other can either be a raw formal context, or we can use the original algorithm to turn it into another set of pertinent concepts.

That means: We can use either one set of pertinent concepts plus one formal context, or two sets of pertinent concepts as our inputs. Depending on the inputs we proposed two Approaches: 'Merge coverage + context' and 'Merge coverages'. (*Coverage* is here used as another name for the set of pertinent concepts.)

Purpose of the algorithm, 'Merge coverage + context', is to merge the initial set of pertinent coverage of formal concepts (F_1) with newly added objects, i.e. the new formal context (K_2). The prerequisite of using this approach is that the formal context K_1 from which the coverage F_1 was computed is remarkably bigger than the formal context K_2.

Purpose of the second algorithm, 'MergeCoverages', is to merge two sets of coverages of pertinent concepts F_1 and F_2. The algorithm is applicable mostly in cases when the two coverages are of almost same size.

Having two formal contexts which need to be merged, we can use either of the algorithms (along with the original algorithm, the one for building pertinent concepts) to obtain desired results. But due to the different natures of the two Approaches, it is strictly recommended to use the former for merging two uneven contexts and the latter for merging two contexts of similar size. In this sense, they are not fully replaceable by each other.

I worked on the project during my stay in Erasmus+ in Tallinn, Estonia on Tallinn University of Technology under supervising of Professor Sadok Ben Yahia.

4.5 Gluing-machines

The goal of the project was to uncover causes of production failures in gluing process. We disposed with process values with clearly marked failure states, but there was neither documentation about process and its parameters nor an error log, and it proved impossible to contact a responsible person. Therefore, a big part of analysis consisted of identifying

cycles and sequences within the process and categorization of failures. Only that allowed to form theories about causes of failures.

Data sources:

- Process values from one gluing-machine recorded during half a year period.

The original goal, as defined by the company management, was rather vague: They were dissatisfied with the number of errors cumulating within the process, and wished the error rate to be decreased. Along with that, they wanted the cause(s) of errors to be discovered, so that failures can be prevented in the future.

After obtaining the data, followed by internal discussions over them, we decided that this goal must be extended. With documentation inaccessible, we knew we were going to spend significant amount of time figuring out and describing the inner functioning of the process. Therefore, we included this important step into the business goal.

Before we started looking for anomalies in the data, we needed to know what normal process flow looks like. Some attributes in the set were extremely disperse (represented by coefficient of variation > 0.3). Normal distribution was scarce. That implied we should look for more than one pattern (or set of desired values) within the data. When we applied a heatmap for the data sorted by timestamp, we were able to see some repeating sequences. They were visible in most of the real and average values but seemed not to affect the last values. There were at least six different sequences with a different frequency.

After ‘Modelling’ phase of this project, it proved to be impossible to use data mining methods in the contemplated way; there was clearly no connection between available attributes and the errors, as the attributes did not degrade over time, and their values directly and non-directly preceding production failures were not unusual in any way. But even though the contemplated way of uncovering patterns was not usable, we noticed a couple of interesting features in the data while performing analysis. We discussed this with the management, and together defined new goal for the project. We agreed we would assemble rules for cases possibly leading to failures.

In ‘Evaluation’ phase, after investigating the conditions by which the errors happen, mostly based on categorical attributes (which values occur in categorical attributes during failures, and to which cycles they usually belong), on consecutiveness (what sequences of cycles usually lead to a failure) and on time criteria (how often the failures happen, how much do they tend to cumulate) we were able to form a hypothesis. We hypothesized that an algorithm responsible for switching cycles/tools/locations on a product is not working correctly. As one of the categorical attributes kept the last value of one of the previous cycles,

it seemed possible the algorithm gets stuck on the last point of that cycle without resetting to the required position, leading to inevitable collapse. Such theory was unproveable without further insight into how the algorithm works (and, in fact, whether such algorithm existed, as there was a fair possibility the process is operated manually).

After that, in the 'Deployment' phase, monitoring was set up. All the numerical attributes were displayed as process flows, grouped based on their nature. Categorical attributes were visualized tabularly, divided by cycles. Errors were displayed in consideration of the facts exposed during the analysis.

4.5 Welding-machines

The goal of the project was to design a monitoring system with anomaly detection for welding process in real time. Similar to the previous project, no documentation was accessible, and because of that, initial identification of patterns within data was necessary. Moreover, there was no information about production failures (it was impossible to clearly distinguish error state from non-error state), leading to the need of drafting the identification procedure of (possible) failures.

Data sources:

- Process values from six welding-machines.

Similar to the previous project, no documentation to the process or data set was available, as well as no information on normal or error states. Furthermore, the error states were not even marked in the set. Therefore, it was necessary for us not only to look for cycles and sequences typical for the process, but also estimate those parts where the cycles were broken or had incorrect values. That was a purpose of 'Data Understanding' phase, what was a dominant phase for this project.

We repeated the approach from the previous project and depicted the data via a heatmap. This time, however, we were not able to use the real values of the attributes in every moment, so we used values aggregated by second instead, in order to visualize all the attributes at once. This led to decreasing the accuracy of the data, and the patterns we were able to see were not absolutely reliable. We needed to add one more factor to improve reliability of the findings.

Hence, we investigated the 'Timestamp' attribute. The differences between two consequencing rows (in seconds, which were the smallest time units after the aggregation)

were of two categories: they were either under 5 seconds, or above 30 seconds. We assumed the former were two rows of the same cycle, while the latter were two rows of different cycles, with the > 30 s gap/pause between them.

Understanding so we were able to divide the process flow into individual cycles, and to identify these cycles, to which we assigned names from the end of the alphabet, i. e. X, Y, Z, etc. We could not be sure, however, whether they each belonged to a different procedure, or whether they were modifications of the same procedures, which had developed over time. There were 7 different cycles in terms of process values (visualized via heatmap), but no more than three of them were present at a time during the observed period taking almost 4 months. When we compared their lengths, we concluded that there may be three cycles in the process, updating as the time goes, one of them significantly major in frequency, and two minor ones.

When marking possible error states, we divided the estimated process failures into three categories: The first type were blocks with incorrect values, which significantly differed from the values typical for any of the cycles. As these values might alternatively represent other, yet undefined, cycles, we decided we would call a cycle those blocks which represent at least 5 % of all the blocks, otherwise we would consider it an anomaly and a potential error. The second type were blocks which copy parts of some of the cycles in term of process values, but which were significantly shorter than the respective cycle. This might mean the process had either finished too early or began too late and therefore could not be completed successfully. It is reasonable to consider these blocks to be failures within the process. The third type were blocks consisting of a single value. They are indeed an anomaly, though it stays unclear now what caused them: they may belong to the previous or following cycle, being recorded at the wrong time, or they may be accidental recordings of the process values of the machine.

In the ‘Deployment’ phase, we set up online monitoring with anomaly detection similarly to the previous project.

4.6 Permissible variations

The goal of the project was to design and implement a system for prediction of the final state (error/non-error), based on input parameters of a production process in real time. Input and

output numerical values of two parameters served as a training set. The data mining part could be approached as both classification and regression problem.

Data sources:

- Input and output values of two parameters of a production process.

The data set consisted of measurements from several points on the surface of the products, recorded for 2 months. Each point had been saved in an individual .csv file and contained app. 19,000 records. Out of the points, 26 contained enough measurements for both numerical attributes on input and output to performs further analysis.

Before applying any data mining models, we constructed correlation matrices for each point as to see whether there are any meaningful correlations. There was at least one moderate correlation between inputs and outputs in 18 out of 26 points. We also searched for correlations between points of the groups. As a results, there was at least moderate correlation found in 7 out of 8.

We decided to use three different approaches which we would compare to each other in the end. Two of them dealt with individual points: First one used models built via RapidMiner, the second one was an experimental setting of boundaries via Python. The third approach was similar to the first one, but instead of individual points, groups were used. We chose classification over regression for the task (though both methods would be possible due to the nature of the data) and we used sensitivity as an evaluation criterion.

Though the initial correlation matrix looked more in favor of groups (87.50 % at least moderately correlated vs. 69.23 % for individual points), the results of validation proved different: Approach 3 (data mining methods used specifically for groups) was successfully trained for only 37.50 % of cases, what was less than for any other case, and furthermore, no groups were validated successfully.

This left us with the first two Approaches, from which the first one was the very best (53.85 % validated successfully on both cross and split validation), followed by the second Approach (46.15 % successfully validated). In the last step of ‘Modelling’ phase we combined the two Approaches in an ensemble model with results calculated via voting.

The monitoring was set up in near-real time in the ‘Deployment’ phase. Before visualizing, the data were bulked from Qdb database (‘Data Sources’) into Kibana/Elasticsearch (‘Data Storage Layer’) via a Python parser (‘Data Acquisition Layer’). On frontend, the data from Input and their predicted Output were displayed in form of ‘traffic lights’. The table contained ID of a product, timestamp of recording, point on the surface of

the product, values from two Input attributes, and predicted Output – green for ‘OK’, red for ‘NOK’, intuitively.

Part of the results was published as a workshop paper ‘Classification models for purpose of predictive maintenance in a production process’ and presented on WAIT conference, 2021. That was during my Erasmus+ stay in Budapest, Hungary, which I spent under supervision of Doctor Bálint Kiss.

5 RESULTS

The projects I was working on provided me with two types of useful experience. First, there were aspects in which most of the projects followed the same line. I think about them as typical for production industry. This knowledge is the easiest to generalize and forms a base for the best practices.

Then there were features that differ for certain projects. These specifics impersonate problems (and solutions) that cannot be fully generalized because of the lack of examples (and therefore lack of variability required for generalization), but they are still going to be incorporated to some extent, serving as a basis for future projects closely dealing with problems only briefly touched within the proposal.

5.1 Proposal of best practices for ‘Business Understanding’ phase

Learning from the real-world experience, I was able to assemble a process diagram of top layer of ‘Data Understanding’ phase (Fig. 1). It possesses rather linear structure and is divided into four subphases. ‘Data Understanding’ is the opening phase of CRISP-DM. According to praxis, it usually starts with ‘Defining business goal’: firstly as internal discussion on the side of the company, secondly as transferring requirements to the data science team, thirdly as an internal discussion of said team. It usually contains more than one iteration of redefining the goal, as both parties aim to find a compromise between desired and possible. Follows ‘Defining data mining goal’, if applicable, as the project does not necessarily contain a DM part; and sometimes it would, but defining such goal is impossible without further understanding of the process and sources (and is therefore possible only after ‘Data Understanding’ phase). This step is similar to the previous one, with many internal and external discussions and repeated redefinitions of the goal. The next one is ‘SWOT

analysis', where the requirements are judged in terms of their strengths, weaknesses, opportunities and threats, and after these are documented, they will serve as a base for 'Finalising project plan'. The plan should contain information about required steps, resources, time, people, and also a backup plan (alternatives) for the risky parts of the project.

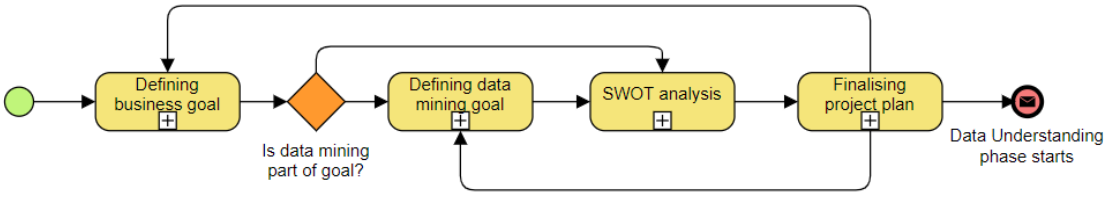

Fig. 1: Conceptual process diagram proposal for 'Business Understanding' phase

Some of the problems had recurring characteristics when dealing with real-world projects. I assembled them and divided them into six categories based on the data quality dimensions.

Accuracy:

- Vague business/data mining goal;
- Sensitivity-specificity tradeoff;
- Risk of misunderstanding goal/process/data;
- Slim margin for error.

Completeness:

- Uncomplete business/data mining goal;
- No contact to responsible person(s);
- Missing documentation.

Consistency:

- Some of the steps/order of steps unpredictable.

Timeliness:

- Plan has to be scattered and rebuilt from scratch.

Validity:

- Business/data mining goal cannot be achieved;
- Process/structure unknown;
- Unknown (number of) sources.

Uniqueness:

- Not enough/too many methods/ways to grasp the task;
- Not enough/too many sources/parameters;
- Unprecedented task.

Proposal for ‘Business Understanding’ phase, composed due to real-world experience, can be seen in a detailed process diagram (Fig. 2).

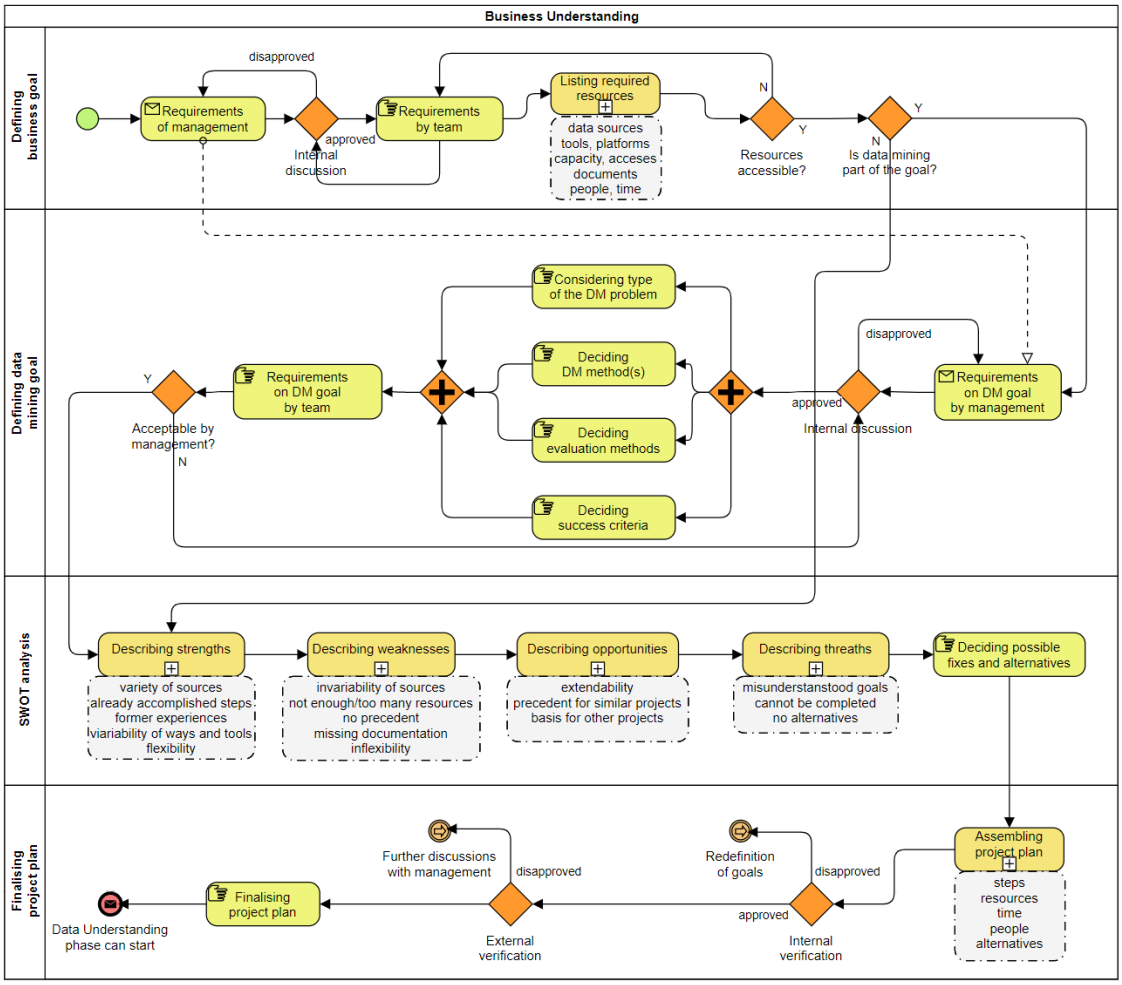

Fig. 2: Detailed process diagram proposal for 'Business Understanding' phase

5.2 Proposal of best practices for ‘Data Understanding’ phase

Experience from all the aforementioned projects allowed me to assemble most common problems and their solution for the process. I constructed a conceptual process diagram of the ‘Data Understanding’ phase (Fig. 3), consisting of six steps, subset of which is usually performed when acquiring and inspecting the possible data sources. The process follows after successful completion of ‘Business Understanding’ phase and is based on the business and data mining aims established there. In praxis, ‘Data Understanding’ starts with ‘Constructing data requirements’, criteria and specifications the data sources should meet. This is first discussed internally (within the data scientist’s team), then verified with customer (management of the company). Then, ‘Listing of available data sources’ is

performed according to mutual agreement. This may lead to communication with various departments of the company. The accessibility of the sources is checked and the first exports are set-up. Result of this step is a list of accessible sources; eventually, this may lead to redefinition of data requirements. The process continues with ‘Describing data source’ – each source separately. This includes constructing descriptive statistics, basic diagrams and distribution of data. Obtaining documentation is a crucial part; if it does not exist or is not accessible, there is a need to perform ‘Describing normal process state’ and/or ‘Marking possible anomalies’ (in case errors are not marked within the process). The phase is then closed with ‘Verifying data sources’, a final step which approves the acquired sources, when documentation about all the sources is assembled. After that, ‘Data Preparation’ phase may start.

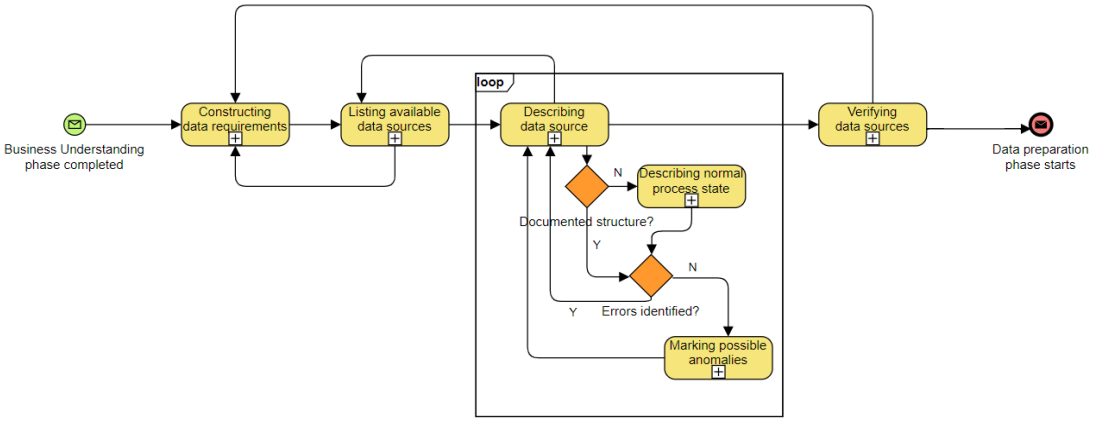

Fig. 3: Conceptual process diagram proposal for 'Data Understanding' phase

As before, I classified the most common problems into data quality dimensions. Following problems occurred repeatedly when working on the projects:

Accuracy:

- Unclearly specified requirements.

Completeness:

- Missing process documentation;
- Not enough sources;
- No documentation to a source;
- Limited/permitted permanent access to a source;
- Missing documentation about data (unknown sequences within process);
- Missing identifier of OK/NOK states.

Consistency:

- Process changing over time (without records about these changes).

Validity:

- Requirements unreal, given available sources;
- Responsible person(s) not available;
- Sensitive data.

With that information I was able to draft an expanded process diagram for ‘Data Understanding’ phase (Fig. 4).

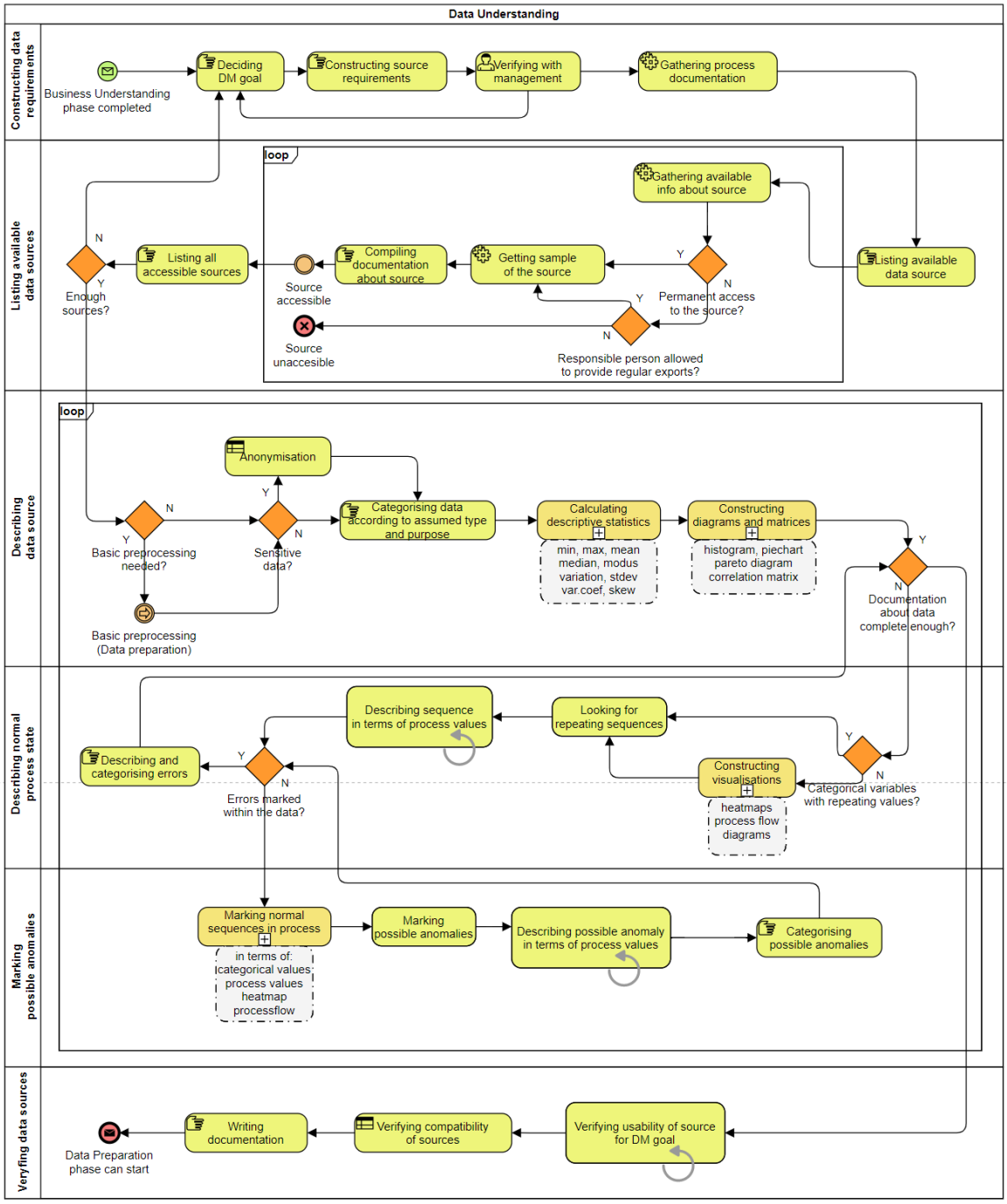

Fig. 4: Detailed process diagram proposal for 'Data Understanding' phase

5.3 Proposal of best practices for ‘Data Preparation’ phase

After documenting ‘Data Preparation’ phase for various real-world projects, I assembled the most common problems and their solutions. In the top process diagram (Fig. 5), there is a proposal for the top level of this phase. The structure is linear, with two main parts, which can iterate, if necessary, until the desired result is achieved. The process follows after ‘Data Understanding’ phase. It starts with ‘Adjusting structure’, either working with given exports, or returning to ‘Data Understanding’ phase for a while, to design a new export structure, close enough to the required structure (if possible and necessary). The next step is ‘Integration of data sources’ via common identifier (like ID), timestamp, composed keys, etc. If two sources cannot be integrated in a meaningful way, return to ‘Data Understanding’, or even ‘Business Understanding’ phase may be necessary, to reconsider either available data sources or business/data mining goal, respectively. Successful integration is followed by adjustments inside the source. The following order proved to be the most advantageous one: First, ‘Adjusting attributes (columns)’ is performed. The desired attributes are selected, new ones are calculated, and supplementary ones removed. ‘Adjusting records (rows)’ follows, as to keep only the records happening in the desired time periods (e. g. omitting records from holidays and weekends or keeping just records from specific working shifts). This also includes dealing with empty and semi-empty rows. The latter may be either removed or reconstructed. The last step is ‘Adjusting values’, in which various issues are addressed, concerning data types and formats. In multilingual companies, language issues were found to be very important and may take time to be solved (and unearthed, in the first place). After that, ‘Data Preparation’ phase is completed, and ‘Modelling’ phase can start.

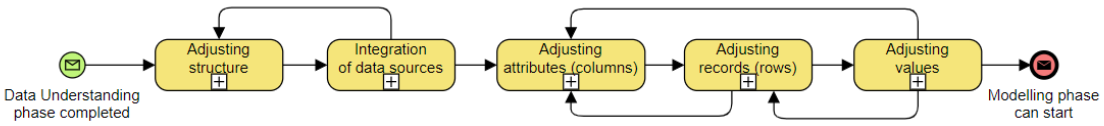

Fig. 5: Conceptual process diagram proposal for ‘Data Preparation’ phase

I listed the most common problems that occurred to us during the ‘Data Preparation’ phase, and classified these problems into groups, based on their affiliation to one of the six data quality dimensions.

Accuracy:

- Incorrectly guessed types;
- Incorrect formats of numbers and dates.

Completeness:

- Missing rows;
- Rows with but empty values & rows with mostly empty values.

Consistency:

- Structure incompatible with other sources;
- Missing common unambiguous identifier;
- Two time periods have completely different structure;
- One period recorded much less frequently;
- Naming inconsistencies;
- Spelling inconsistencies (aka case, diacritics);
- Inconsistent usage of decimal point/comma;
- Different time formats;
- Names/text written in mix of languages.

Timelines:

- Only one parameter recorded at a time;
- Sources not time-synchronised.

Validity:

- Structure not suitable for software tools;
- Structure too complex (e. g. consisting of one large matrix);
- Big number of attributes;
- Attributes in rows instead of columns;
- Composed (non-atomic) attributes;
- Categorical attributes not usable in data mining methods;
- Records from non-working days misrepresent results;
- Characters not manageable by the process;
- Words illegible where using specific characters;
- More versions of spelling.

Uniqueness:

- Redundancy due to usage of multiple sources.

I composed a process diagram of ‘Data Preparation’ (Fig. 6) phase, based on the aforementioned knowledge.

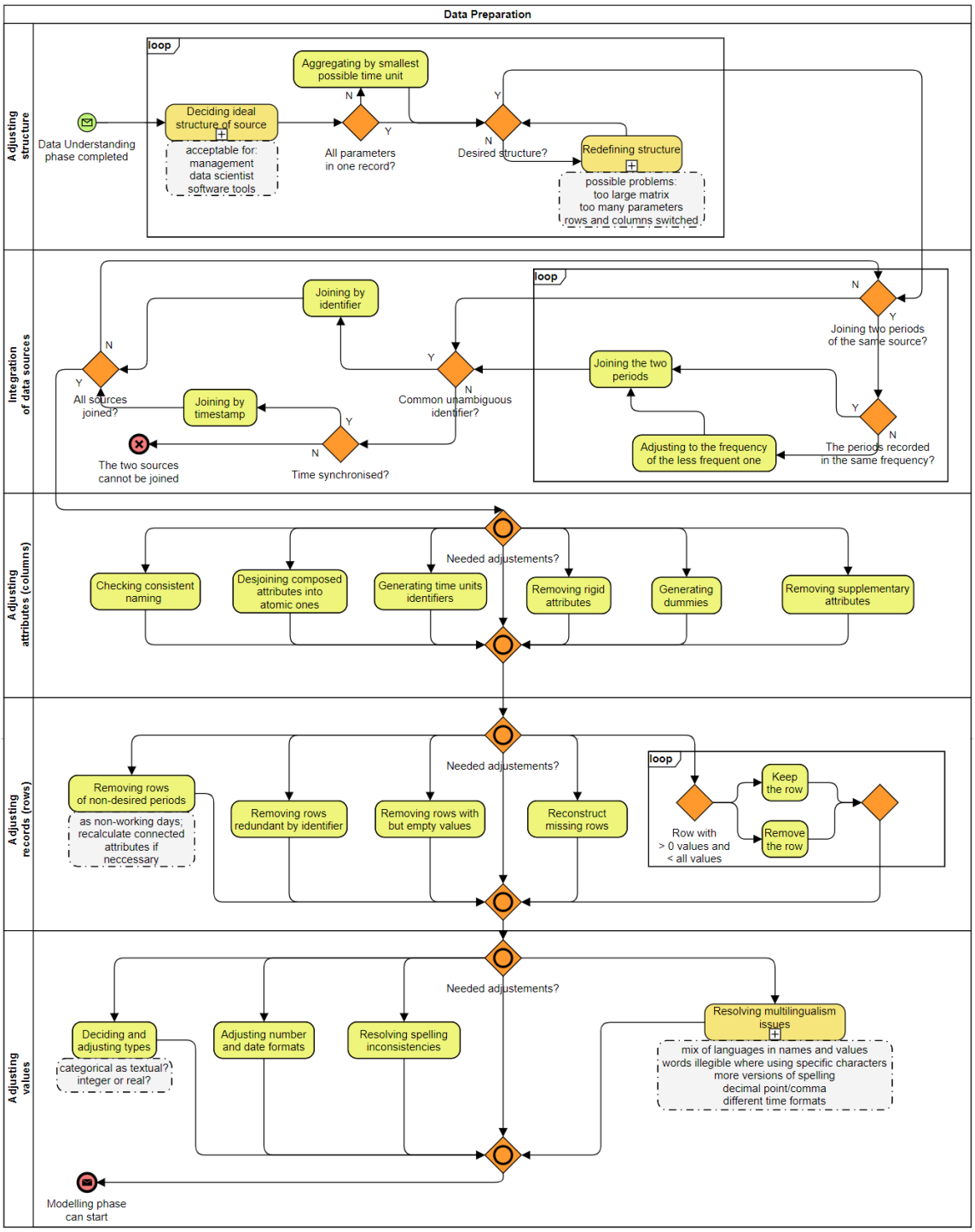

Fig. 6: Detailed process diagram proposal for 'Data Preparation' phase

5.4 Proposal of best practices for 'Modelling' phase

I assembled a conceptual process diagram (Fig. 7) based on the information from the real-world projects. The diagram has five main parts. After the data were preprocessed during

‘Data Preparation’ phase, they undergo ‘Initial Analysis’. Note that an analysis was already performed during the ‘Data Understanding’ phase – but that was a general analysis, whereas this one is already focused on a clearly defined goal. Its purpose is to uncover features important for the data mining goal (like future input and output attributes). With this knowledge, ‘Possible solutions’ are discussed. Usually there are many ways how to strive for the goal, and these should be listed and drafted in this step, without actually implementing them. Yet, in this step all possibilities should be investigated, focusing on their advantages, disadvantages, demandingness, how much of the original goal they may cover, etc. After that, the most useful solutions are picked. (From this step on, I am going to call them approaches.) Their number depends on how many of them met the desired criteria, and also how much time and resources the team disposes with. ‘Building an Approach’ is a step that follows, usually in loop for every chosen approach. In this step, the chosen ideas are implemented. It can be performed via available data mining methods or as a script built from scratch. The model is applied to a dataset, validated, and verified. When this is finished, ‘Comparing Approaches’ takes place. All the viable approaches are selected. One or more of them may be chosen for future evaluation and deployment. If none of the approaches is good enough, combining them may be considered, as to improve their characteristics. In case all the approaches are completely unusable, return to the ‘Business Understanding’ phase and redefining data mining and/or business goal to something that is both achievable and acceptable by management is necessary. On the other hand, if there is at least one viable solution, the ‘Final Approach’ step follows. If there are more solutions to be combined, they are combined either using one (or more) already existing ensemble methods or by scripting own methods. Finally, the final Approach is finished and documented. Then the ‘Evaluation’ phase may start.

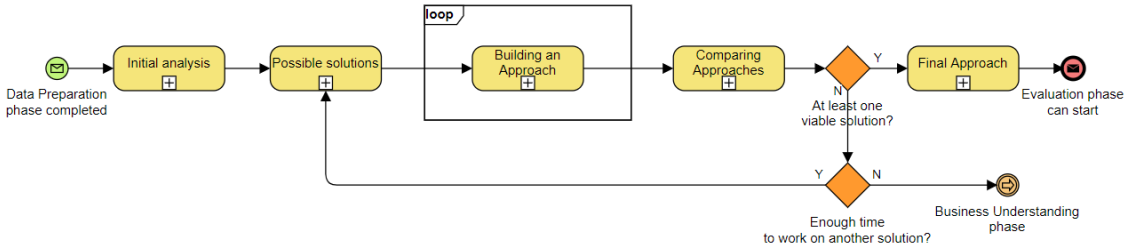

Fig. 7: Conceptual process diagram proposal for ‘Modelling’ phase

Similarly to the previous phases, I assembled all the problems occurring during the work on the case studies and made a list of those which occurred at least in two of them, using them as a base to form a proposal for the phase. For ‘Modelling’ phase, there is a big number

of possible tasks – there are various types of data mining as classification, regression and clustering, and aim of the project can extend even beyond those types. Therefore I did not focus on every type and method, I rather tried to form a general concept on finding the right approach and tools to fit the desired goal.

The following problems (again, categorized into groups based on data quality dimensions) were repeating most frequently during this phase:

Accuracy:

- No meaningful correlations found;
- Results are far worse than expected.

Completeness:

- No meaningful solution available;
- The DM/business goal is unfulfilled;
- No approach is viable;
- An approach is viable, but for different purposes.

Consistency:

- Initial analysis does not confirm the management's hypothesis;
- The future results of a solution are doubtful;
- Results do not confirm the team's hypothesis;
- The approaches are difficult to combine/cannot be combined;
- Combining approaches rips them of certain characteristics.

Timeliness:

- An approach/combined approaches take too much time/computational power.

Validity:

- The data could not be put in a form ideal for the desired analysis;
- Desired analysis is not possible;
- No solution leading to the desired goal;
- No build-in methods for the purpose we need;
- Results are leading somewhere else than expected.

Uniqueness:

- Too many solutions available;
- Too many ways how to work on an Approach;
- Too many viable Approaches.

Based on all the acquired information I was able to form a proposal for a detailed process diagram of 'Modelling' phase (Fig. 8).

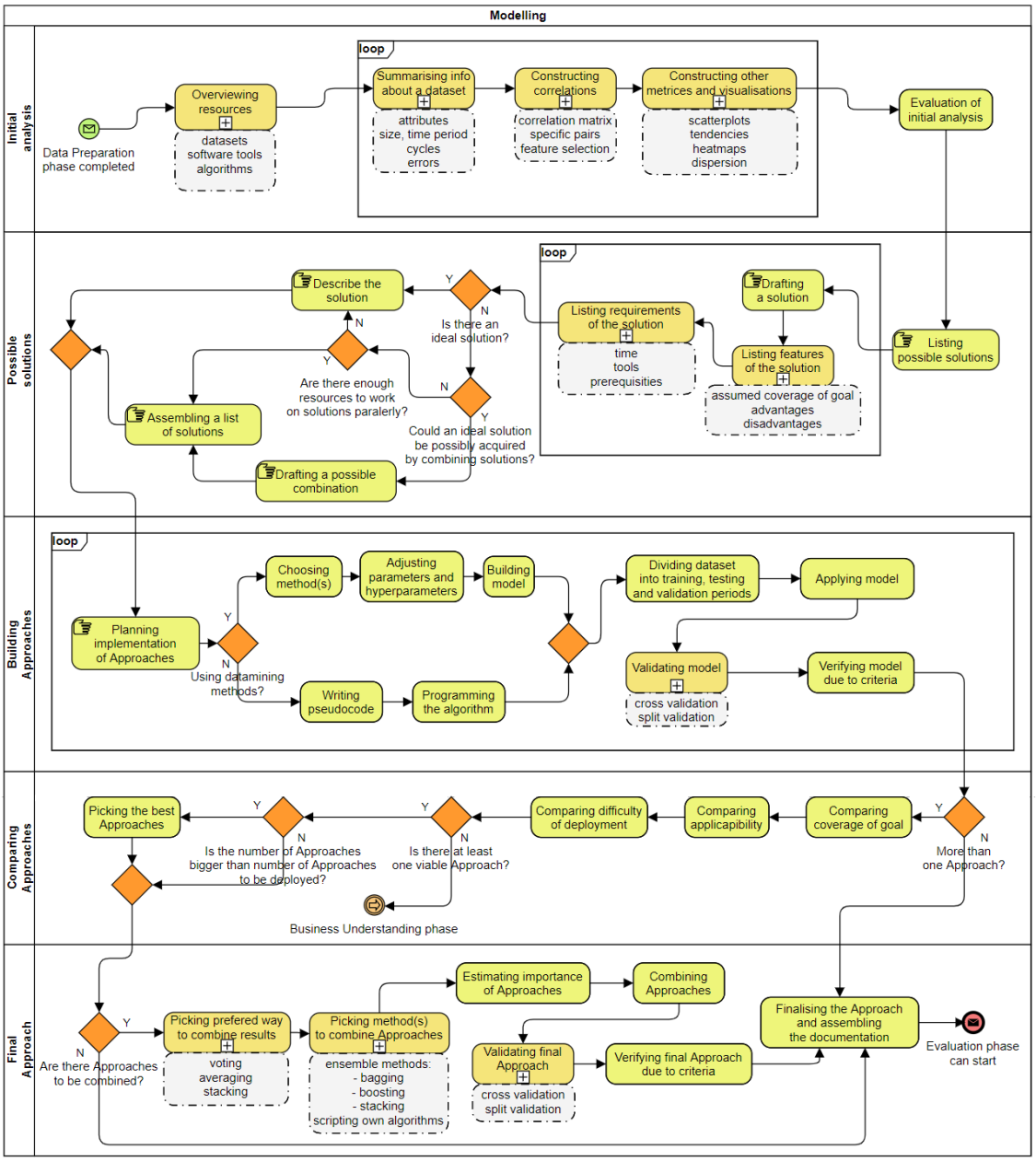

Fig. 8: Detailed diagram proposal for 'Modelling' phase

5.5 Proposal of best practices for 'Evaluation' phase

The experience I extracted from the case studies were used in building a conceptual process diagram of 'Evaluation' phase, as can be seen in the picture (Fig. 9). It consists of four parts. Their occurrence in a particular project depends on successfulness of the previous analysis (especially the 'Modelling' phase). If the analysis provided us with meaningful results, these will be validated and (if validated successfully) improved during this phase. 'Validation of solution' is, ideally, performed on new time periods, and can be extended by trying the same

model on similar data (of other parts of product or process). The ‘Solution improvements’ may include tuning hyperparameters in data mining models, making algorithms more effective by changing the code, combining partial solutions in different ways, etc. Aside from the original data mining and business goal, other interesting facts may have been uncovered during the analysis. ‘Other results’ are collected and visualised in various ways, e.g. rules, recommendations, schemes, tables, charts. The last step of the phase is ‘Goals evaluation’, which focuses on evaluating the completion of business and data mining goal up to this point and lays basis for the ‘Deployment phase’.

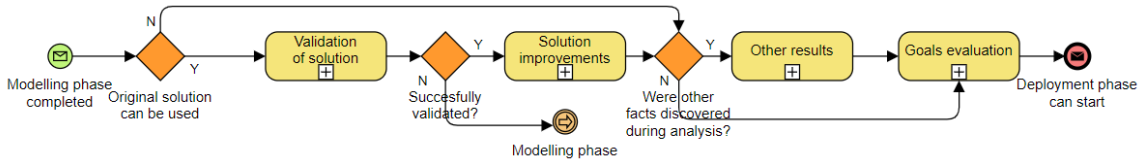

Fig. 9: Conceptual process diagram proposal for 'Evaluation' phase

During this phase I encountered some of the problems and challenges several times. Here they are categorised into groups of data quality dimensions:

Accuracy:

- Specificity-sensitivity trade-off;
- The improved solution is not as good as the improved partial solutions.

Completeness:

- Data mining goal unfulfilled;
- Business goal unfulfilled.

Consistency:

- A model trained on original data does not provide good results;
- The process changes often;
- A model cannot be used for other, similar parts of the product/process;
- The improved solution is unstable.

Timeliness:

- Training periods too big/small.

Validity:

- Hyperparameters picked improperly;
- Values of hyperparameters picked improperly;
- Difficult to verbalise/display other results;
- Other results seem not to have any practical usage;
- Analysis took a different course.

The proposal of a detailed process diagram for ‘Evaluation’ phase can be seen in the figure (Fig. 10).

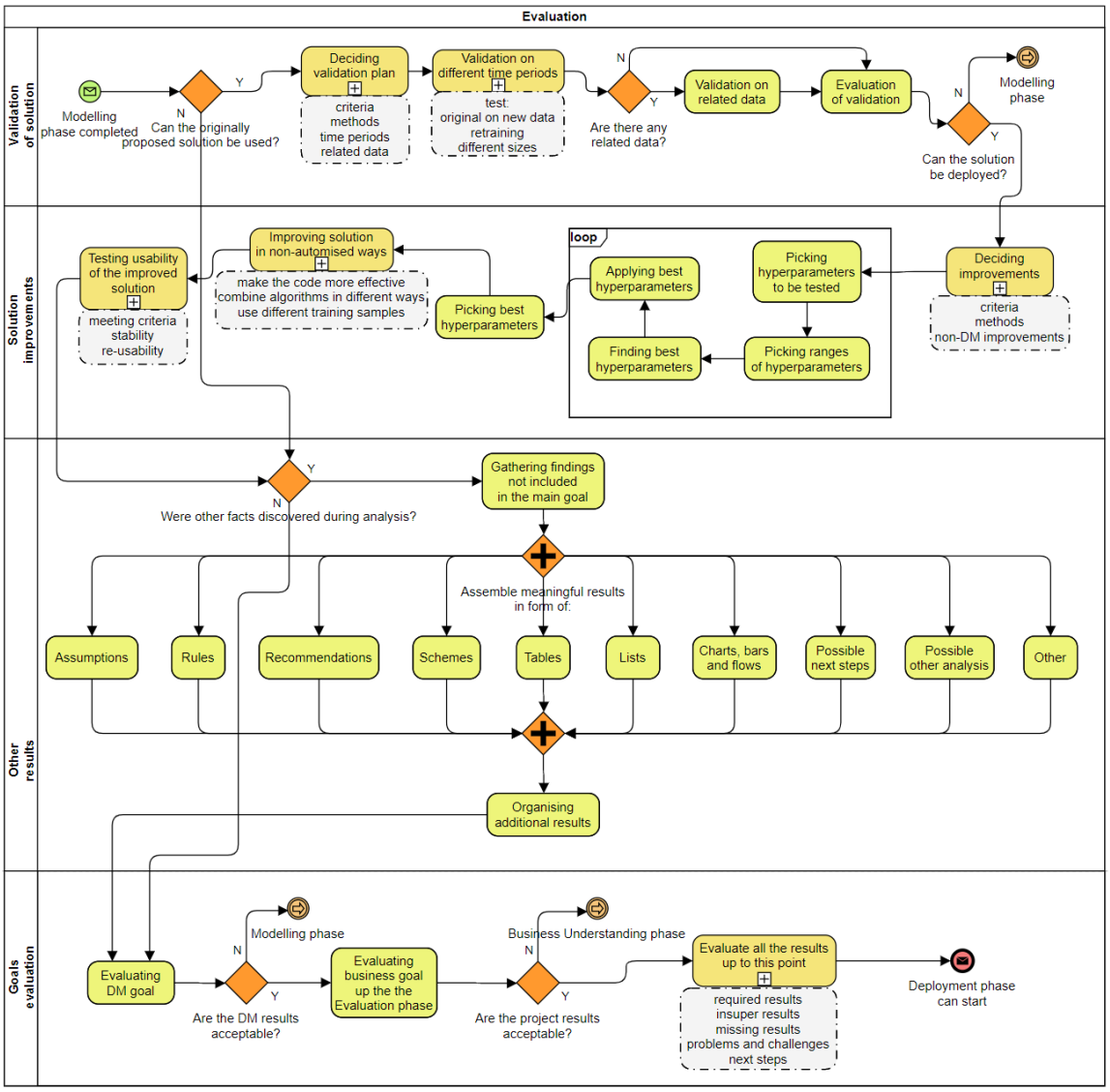

Fig. 10: Detailed process diagram proposal for 'Evaluation' phase

5.6 Proposal of best practices for ‘Deployment’ phase

After completion of the aforementioned projects, I was able to assemble a conceptual process diagram, as can be seen in the picture below (Fig. 11). It is divided into five phases, three of which are optional: all of them can be performed, or none of them, or anything in between, depending on the nature of the project. ‘Deployment’ is the final (sixth) phase of CRISP-DM. Similarly to the other phases, it starts with deciding a ‘Deployment plan’, which includes visions and proposals of the team and management, selection of visualisation platform and sources to be used for visualisation, then bulking these sources into the

platform. ‘Monitoring’ follows, with designing visualisation of the process in form of process flows, tables, pie charts, bar charts, metrics or other, and their combination. ‘Monitoring’ itself passively represents the process by focusing on specific features, including errors, in real time (of course with option for storaging), but does not provide any sort of opinion or advice as what parts of the process may be abnormal. That is, in fact, task of the next, ‘Anomaly detection’, step. It includes marking (possible) abnormalities, based on the previous patterns (normal process states), calculation of severity of these abnormalities, and forecasting future states. The next step, ‘Error prediction’, is similar but more specific, with predictions focused on process failures. The last step is ‘Delivering results’. There, results of ‘Deployment’ and all the previous phases are validated according to business goal, and final documentation is assembled. The results (application, prediction model, findings, documentation, etc.) are delivered to the management. The knowledge gained during working on the project can be used as a basis for similar projects, published in scientific papers, and serve for building team’s know-how.

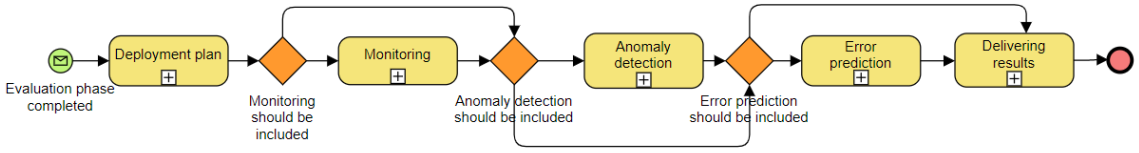

Fig. 11: Conceptual process diagram proposal for 'Deployment' phase

Problems which reappeared several times, categorised into data quality dimensions, are the following:

Accuracy:

- ambivalent requirements for deployment

Completeness:

- Not enough data for a certain visualisation.

Timeliness:

- All attributes cannot be monitored at once due to unique data structure.

Validity:

- Need for scaling;
- Expected exceptions;
- Skews from criteria.

Uniqueness:

- Too many combinations leading to big number of visualisations;
- Too many cycles = too many patterns.

In the figure below (Fig. 12) can be seen proposal for ‘Deployment’ phase, based on real-world experience.

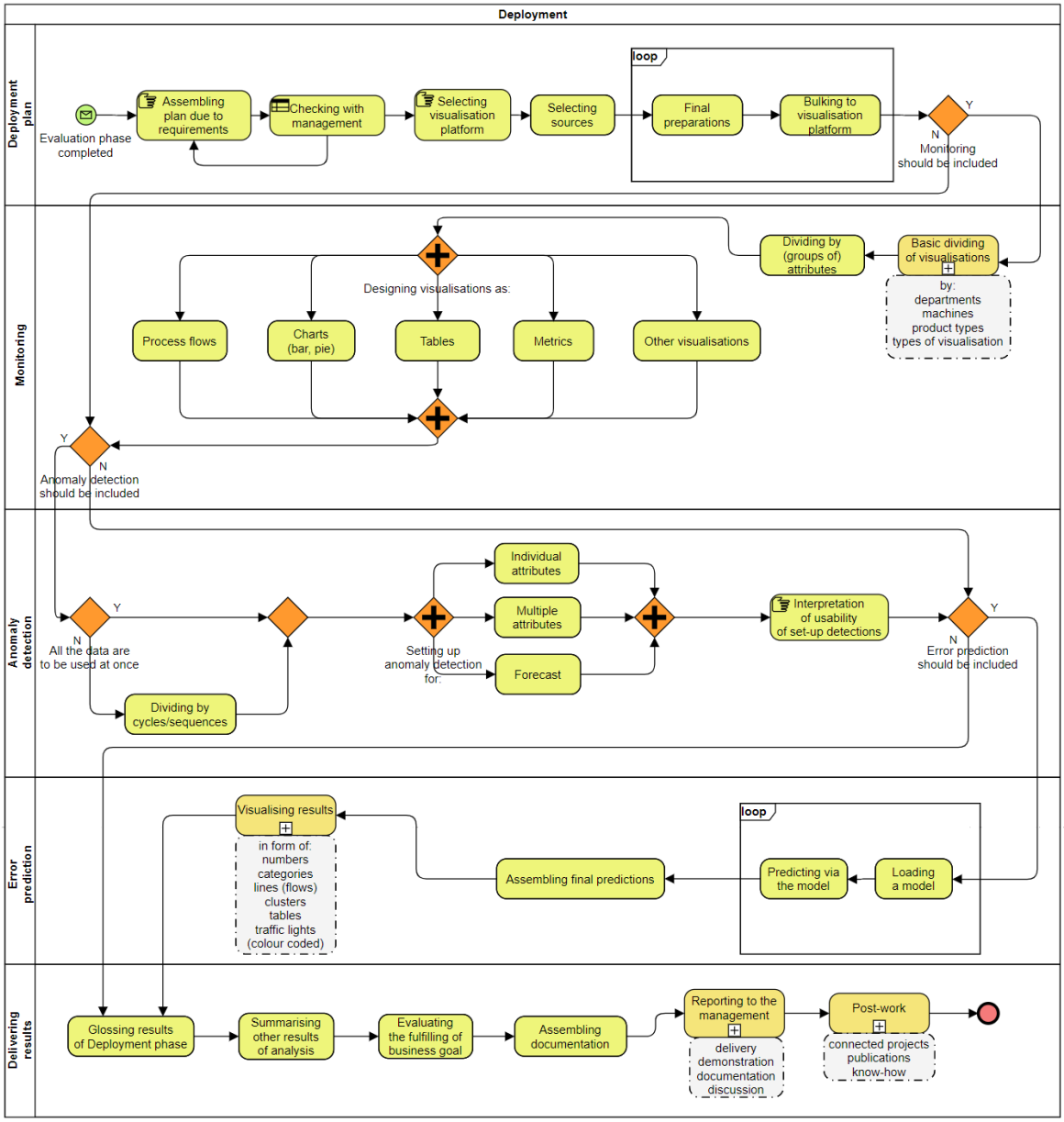

Fig. 12: Detailed process diagram proposal for 'Deployment' phase

5.7 Integration of Phases

I made two versions for every diagram for every phase: a conceptual one (featuring just subphases of every phase) and an advanced one (with elaborated subphases). By joining all members of the former category, we get a new diagram composed of conceptual diagrams (Fig. 13).

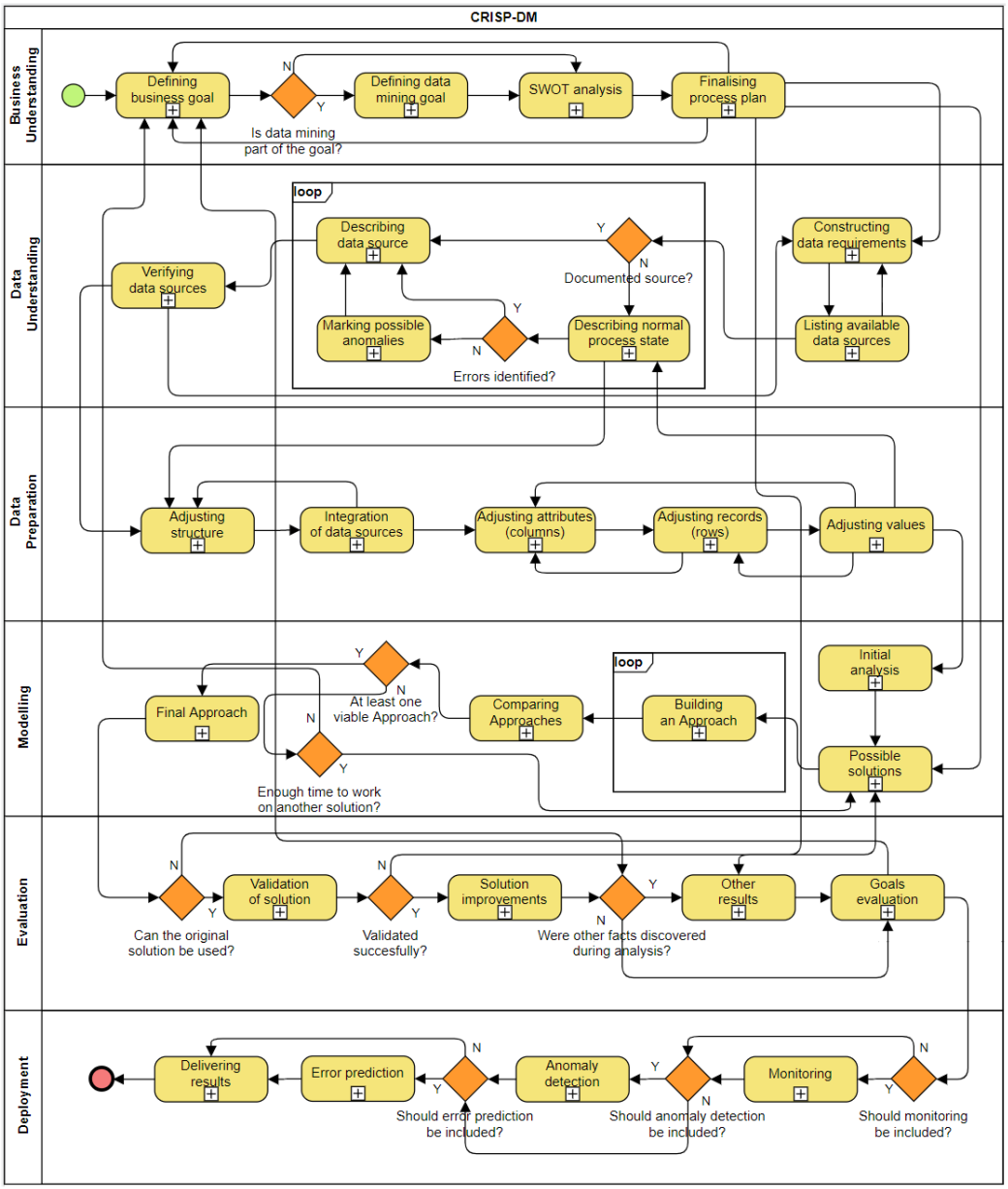

Fig. 13: Detailed process diagram proposal for integration of phases

As one can see from Fig. 13, not only the inner structure is different from typical CRISP-DM, but also the sequence has been enriched by new connections between the phases. The original methodology already includes returning to some of the previous phases, but the praxis shows the phases are much more connected among each other, in both directions, e.g. if the data mining goal was not fulfilled in the 'Modelling' phase, the return to the 'Business Understanding' phase may be necessary to define a new goal according to the remaining resources and possibilities. After that we may usually skip the next two or three phases and execute either 'Modelling' or 'Evaluation' phase, based on the nature of the new desired results.

The connections between phases are addressed in the conceptual diagram composed from the basic phases (Fig. 14).

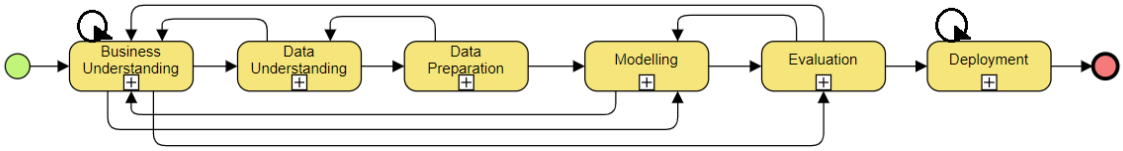

Fig. 14: Conceptual diagram proposal for the combination of phases

One of the main contributions of my research is implementation of new approaches, obtained via analysis of several real-world projects, into the lifecycle of solving Big Data projects based on CRISP-DM methodology. Application of these approaches can help with avoiding/early eliminating problems, which usually occurs during work on Big Data projects. Furthermore, the best practices I proposed contribute to the more effective problem-solving, with several prearranged solutions to be used when necessary.

That way, the best practices can help with sparing company's time and money, as well as time and sources of the data scientist's team.

6 DISCUSSION

The presented best practices were based on experiences of our team. I oriented the proposals on problem-solving of a specific phase and designed process diagrams for such purpose. The best practices emerge from CRISP-DM methodology as a theoretical base. From the methodology I took division into phases and their sequence. I got the content and inner structure of the phases by tracking the projects' documentation.

As an example we can take 'Data Preparation' phase. The CRISP-DM methodology consists of six subphases: data cleaning, data integration, data transformation, data reduction, data discretization, and feature engineering. I include these subphases in one form or another, but I tracked the typical course of the projects to reorder tasks due to their purpose (rather than to their nature), as this proved to be the most suitable for most of the projects I have carried out. That lead to forming five subphases: 'Adjusting structure', 'Integration of data sources', 'Adjusting attributes (columns)', 'Adjusting records (rows)', and 'Adjusting values'. This order represents a logical sequence of events, starting with rough changes of sources' structure as to prepare them for integration – followed by the actual integration – then changes of internal structure take place, first addressing the most important units of a

data set (attributes), then the individual instances (records), and finally focuses on minor (yet important) issues like format, spelling and language inconsistencies (i.e., values issues). I performed similarly for all the other.

6.1 Most Common Problems

Even though most errors which served as a base for the best practices were phase-specific, there were a couple of them which affected various phases and arose during several projects. Once more, I classified the most prominent and the most often occurring errors into groups based on data quality dimensions.

Accuracy:

- Results do not confirm the hypothesis;
- The business/data mining goal not fulfilled.

Completeness:

- Vaguely defined/uncomplete business/data mining goal;
- Missing documentation;
- Permitted/limited access to a source;
- Lack of suitable data sources;
- Not enough data for certain visualizations.

Consistency:

- Process changing over time;
- Missing common unambiguous identifier;
- Structure changing over time;
- Naming, spelling and language inconsistencies;
- Validation on a different period/part unsuccessful.

Timeliness:

- All attributes cannot be monitored at once.

Validity:

- Sensitive data;
- Structure not suitable for software tools;
- Desired analysis not possible;
- Analysis took a different course.

Uniqueness:

- Too many combinations leading to a big number of visualizations.

Many of the problems come from inadequately defined requirements and a lack of documentation. The former problem can be usually helped by enough communication (with company management and within team). Specifically important is to know beforehand which sources will be available and what information can be extracted from them. The latter problem is something the data scientist cannot influence; however, early knowledge about the state of the documentation can be very helpful in estimating the time needed for the project. Also, it can lead to incorporating the building of documentation into the project's goals. In any case, knowing as much as possible about the project before starting it is a key to finishing it successfully and on time.

We can find parallels to these problems in many of the papers I worked with in the literature overview. Zhao et al. (2018) remark that introducing industrial Big Data is a big challenge for the company – that is the same struggle I observed many times, e.g. in incompatibility of data sources. Inconsistencies of data are also addressed by Keskar et al. (2021). [15,21]

Problems with choosing the right data to be collected is also described by Sun et al. (2019), what leads to lack of suitable data sources, which is also one of the problems I identified. Lacam et al. (2021) focuses on importance of generating Smart Data by enriching Big Data – and I discuss the general lack of documentation. Solid documentation is a base for such enrichment. [33,20]

Juez-Gil et al. (2021) came to a conclusion that more complex methods may prove less useful, similarly to one of the possible problems I identified for 'Modelling' phase. Küfner et al. (2021) describe problems with working close to real time, what is also reflected in my experience with not being able to access certain data sources in real-time or monitor all attributes at once. [22,7]

Human factor is another aspect which I addressed often in the projects (in issues like missing documentation, vaguely defined business goal or denied access to a database). Importance of involving employees into Industry 4.0 was accentuated by Gallo et al. (2021) and Shin et al. (2021). Schuh et al. (2019) also focus on extensive internal knowledge in a company. [9,38,46]

As I stated before, summarizing best practices or extending existing methodologies is rather unusual topic among the studies on key terms of this thesis. But still I was able to find similarities with my work. Schröer et al. (2021) also build their best practices on CRISP-

DM – they focus on a single phase, ‘Deployment’. Huber et al. (2019) do something similar for ‘Data acquisition’ subphase. [30,31]

That proves the problems I was discussing are not limited to a small group of projects. But even though they are often present, they are but scarcely summarized and discussed in other studies. I collected and systematized them and used them as a basis for a new set of best practices. Along with listing the problems, I also discussed them in detail, providing possible solutions. The best practices include process diagrams as a visual representation of generalised project solutions. They can be helpful when choosing a path of approaching a project, as they often provide more than one possible solution and allow to skip certain parts of the process.

By getting familiar with challenges I (repeatedly) encountered, other data scientist can prepare in advance when working on similar projects.

6.2 Specifics of Production Industry

I used six projects as case studies for this paper. All of them were dealing with real processes in companies. And since I based the research on problems found in production, also my proposal was focused at this area.

One of the most important features of production process is, obviously, its practical orientation. Most tasks focus either on improving the production, diminishing errors, or on more general predictive maintenance. Getting knowledge just for knowledge is not demanded, every result should have its appliance in praxis.

A customer, and not the data scientist’s team, is the one to decide. Therefore, a more logical solution (from the data scientist’s point of view) may be dropped in favor of the one picked by the client (whose reasons may be rooted in company policy, long-term goal, etc.). Here, the team’s ability to negotiate and to honestly present pros and cons of every solution and to provide the customer with enough information proved to be crucial for a successful communication with the client. It is very useful when there is not only management of the company present to the meetings but also experts from praxis.

The level of deployment of Industry 4.0 in a company can make a huge difference in difficulty of acquiring and processing data. Ideally, all the processes are time-synchronized and storing the data is unified for the whole company. In my best practices I address both this ideal case and the opposite (sadly, frequent) case. But still, most of the data are recorded

and often even stored, what decreases the need of the team to set up their own measuring. It may be necessary only in the cases where the project explicitly states the need for measuring either completely different features of the process, or already measured features in an unprecedented way.

Another important thing about the production process data is that most of them are artificial in their nature, produced by machines programmed by humans, or connected to the scheduled processes (meteorological data captured to uncover influences of weather on the production process represent an exception from the rule). The artificialness of the data means we are often able to guess the possible structure of the process flow, which will in many cases consist of repeating cycle or cycles (their number guessable due to how many different product types and tasks are operated on the machine). Most of the attributes will have either ideal values or boundaries set-up, allowing us to mark those parts of the process where these values are incorrect.

7 CONCLUSION

Even though the best practices were meant to be applied for the production industry projects, I was also thinking about the possible transmission of it to other fields from the very beginning.

After stating the specifics of the original field, it is clear that there are parts of the proposal which cannot be intactly used for other fields, but there are definitely some that could be used, either fully or with just minor adjustments. By the other fields are meant any other areas prone to data analysis, like biology, medicine, social sciences, marketing, etc.

The parts of the best practices which would not need major changes are ‘Business Understanding’, ‘Data Preparation’, ‘Modelling’ and ‘Evaluation’.

‘Business Understanding’, despite its name, can be understood as Project Understanding or even Defining Goals. Similarly, if we substitute management by terms like institution, university or even client, the proposal is still viable. In cases where the team decides about the project independently, the sub-steps connected to the management may be omitted.

‘Data Preparation’ deals with preprocessing the raw data into the desired format; the origins and meaning of the data are not important in this context. The process can be applied basically to any other data.

‘Modelling’, even though it can mean very different things for different fields, is generalized to the extent where only the path is defined (like the decision between already existing DM methods and designing own solution), but the proposal is general enough to be usable for projects with completely different nature.

‘Evaluation’ validates the model(s) designed in the previous phase and summarizes other discoveries made during the analysis. Again, the proposal describes mostly the sequence of steps, and is therefore applicable in other areas as well.

Then there are two phases which cannot be so easily applied out of the production process: ‘Data Understanding’ and ‘Deployment’.

‘Data Understanding’ in industry production works with records of a process designed by humans and performed by machines; the conditions are controlled, tasks predefined and purpose clear, the process is divided into clearly distinguishable cycles and sequences. My proposal of getting to know previously unknown process builds on these features. Choosing the most suitable representation in e.g. behavioral biology can be far more complex and therefore need a very different treatment, according to specifics of the field. Another issue is data acquisition: in industry production we can often work with data recorded by a company and stored in the company’s data warehouses; the situation can be very different for other fields – e.g., sociological research may demand proposal not only for the data storage but also their gathering via manually filled answer sheets.

‘Deployment’ can mean fundamentally different things for other fields. While in production industry it usually means deploying the solution (e.g., set-up of predictive maintenance), followed by a delivery to the company’s management, in more scientific fields, the main goal may be publishing a paper/book, in ecology the focus can lie on acquiring reasoning strong enough to persuade government to strengthen nature preservation, etc. Therefore, the phase should be reshaped for any new field and type of project.

BIBLIOGRAPHY

If it is not stated otherwise, the figures and tables are my own.

1. The 4 industrial revolutions. Available online: https://www.sentryo.net/the-4-industrial-revolutions/ (accessed on 16.05.2019)
2. SPENDLA, L.; KEBISEK, M.; TANUSKA, P.; HRCKA, L. Concept of predictive maintenance of production systems in accordance with Industry 4.0. In IEEE 15th International Symposium on Applied Machine Intelligence and Informatics 2017, Volume 1, pp. 26-28
3. SHARMA, A.K.; BHANDARI, R.; PINCA-BRETOTEAN, C.; SHARMA, C.; DHAKAD, S.K.; MATHUR, A. A study of trends and industrial prospects of Industry 4.0. In Materials Today: Proceedings 2021
4. DI BONA, G.; CESAROTTI, V.; ARCESE, G.; GALLO, T.; Implementation of Industry 4.0 technology: New opportunities and challenges for maintenance strategy. In Procedia Computer Science 2021, Volume 180
5. BELTRAMI, M.; ORZES, G.; SARKIS, J.; SARTOR, M. Industry 4.0 and sustainability: Towards conceptualization and theory. In Journal of Cleaner Production 2021
6. GHOBAKHLOO, M.; FATHI, M.; IRANMANESH, M.; MAROUFKHANI, P.; MORALES, M.E. Industry 4.0 ten years on: A bibliometric and systematic review of concepts, sustainability value drivers, and success determinants. In Journal of Cleaner Production 2021, Volume 302
7. KÜFNER, T.; SCHÖNIG, S.; JASINSKI, R.; ERMER, A. Vertical data continuity with lean edge analytics for industry 4.0 production. In Computers in Industry 2021, Volume 125
8. DURÁN, V.J.R.; BERGES, I.; ILLARRAMENDI, A. Towards the implementation of Industry 4.0: A methodology-based approach oriented to the customer life cycle. In Computers in Industry 2021, Volume 126
9. GALLO, T.; SANTOLAMAZZA, A. Industry 4.0 and human factor: How is technology changing the role of the maintenance operator?. In Procedia Computer Science 2021, Volume 180
10. BEAL, V.: Big Data. Available online: https://www.webopedia.com/TERM/B/big_data.html (accessed on 16.05.2019)
11. The Definition of Big Data. Available online: https://www.oracle.com/big-data/guide/what-is-big-data.html (accessed on 16.05.2021)
12. AZEEM, M.; HALEEM, A.; BAHL, S.; JAVAID, M.; SUMAN, R.; NANDAN, D. Big data applications to take up major challenges across manufacturing industries: A brief review. In *Materials Today: Proceedings* **2021**
13. WANG, J.; XU, C.; ZHANG, J.; ZHONG, R. Big data analytics for intelligent manufacturing systems: A review. In *Journal of Manufacturing Systems* **2021**
14. BAO, J.; YANG, Z.; ZENG, W.; SHI, X. Exploring the spatial impacts of human activities on urban traffic crashes using multi-source big data. In *Journal of Transport Geography*. **2021**, *Volume 94*
15. ZHAO, H.; HOU, J. Design concerns for industrial big data system in the smart factory domain: From product lifecycle view. In *Proceedings of the IEEE International Conference on Engineering of Complex Computer Systems, ICECCS* **2018**, *Volume 2018*, pp. 217-220
16. MUJEEB, S.; JAVAID, Data Analytics for Price Forecasting in Smart Grids: A Survey. In *Proceedings of the 21st International Multi Topic Conference* **2018**
17. ZHANG, Y.; XU, S.; ZHANG, L.; YANG, M.; Big data and human resource management research: An integrative review and new directions for future research. In *Journal of Business Research*. **2021**, *Volume 133*, pp. 34-50.
18. BHATNAGAR, N. 2020. Role of Robotic Process Automation in Pharmaceutical Industries. In: Advances in Intelligent Systems and Computing. 2020, Volume 921, s. 497-504.
19. LIANG, Y.; ZHENG, X.; ZENG, D. 2019. A survey on big data-driven digital phenotyping of mental health. In: Information Fusion. 2019, Volume 52, s. 290-307.
20. LACAM, J.; SALVETAT, D. Big data and Smart data: two interdependent and synergistic digital policies within a virtuous data exploitation loop. In *The Journal of High Technology Management Research* **2021**, *Volume 32, Issue 1*
21. KESKAR, V.; YADAV, J.; KUMAR, A. Perspective of anomaly detection in big data for data quality improvement. In *Materials Today: Proceedings* **2021**
22. JUEZ-GIL, M.; ARNAIZ-GONZÁLEZ, Á.; RODRÍGUEZ, J. J.; GARCÍA-OSORIO C. Experimental evaluation of ensemble classifiers for imbalance in Big Data. In *Applied Soft Computing* **2021,** *Volume 108*
23. CHANG, V. An ethical framework for big data and smart cities. In *Technological Forecasting and Social Change* **2021**, *Volume 165*
24. HUANG, W.; LI, T.; LIU, J.; XIE, P.; DU, S.; TENG F. An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. In *Information Fusion* **2021**, *Volume 75*, pp. 28-40

25. KEBISEK, M.; SPENDLA, L.; TANUSKA, P. Analysis of temperature impact on production process with focus on data integration and transformation. In *Software engineering trends and techniques in intelligent systems : proceedings of the 6th Computer Science on-line conference 2017 (CSOC 2017)* **2017**, *Volume 3*, pp. 317-325

26. What is the CRISP-DM methodology?. Available online: https://www.sv-europe.com/crisp-dm-methodology/ (accessible on 16.05.2021)

27. SALEH, S.J.; ALI, S.Q.; ZEKI, A.M. Random Forest vs. SVM vs. KNN in classifying Smartphone and Smartwatch sensor data using CRISP-DM. In *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)* **2020**, pp. 1-4

28. SCHÄFER, F.; ZEISELMAIR, C.; BECKER, J.; OTTEN, H. Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes. In *IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)* **2018**, pp. 190-195

29. CATLEY, C.; SMITH, K.; MCGREGOR, C.; TRACY, M. Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. In *22nd IEEE International Symposium on Computer-Based Medical Systems* **2009**, pp. 1-5

30. SCHRÖER, C.; KRUSE, F.; MARX GÓMEZ, J. A Systematic Literature Review on Applying CRISP-DM Process Model. In *Procedia Computer Science* **2021**, *Volume 181*, pp. 526-534

31. HUBER, S.; WIEMER, H.; SCHNEIDER, D.; IHLENFELDT, S. DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. In *Procedia CIRP* **2019**, *Volume 79*, pp. 403-408

32. ZONTA, T.; DA COSTA, C.A.; DA ROSA RIGHI, R.; DE LIMA, M.J.; DA TRINDADE, E.S.; LI, G.P. Predictive maintenance in the Industry 4.0: A systematic literature review. In *Computers & Industrial Engineering* **2020**, *Volume 150*

33. SUN, Y.; XU, Z.; ZHANG, T. On-board predictive maintenance with machine learning. In *SAE Technical Papers* **2019**, *Volume 2019-April*

34. APILETTI, D.; BARBERIS, C.; CERQUILETTI, T. ISTEP, an integrated self-tuning engine for predictive maintenance in industry 4.0. In *16th IEEE International Symposium on Parallel and Distributed Processing with Applications* **2018**

35. DALZOCHIO, J.; KUNST, R.; PIGNATON, E.; BINOTTO, A.; SANYAL, S.; FAVILLA, J.; BARBOSA, J. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. In *Computers in Industry* **2020**, *Volume 123*

36. DE PATER, I.; MITICI, M. Predictive maintenance for multi-component systems of repairables with Remaining-Useful-Life prognostics and a limited stock of spare components. In *Reliability Engineering & System Safety* **2021**, *Volume 214*

37. AYVAZ, S.; ALPAY, K. Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time. In *Expert Systems with Applications* **2021**, *Volume 173*

38. SHIN, W.; HAN, J.; RHEE, W. AI-assistance for predictive maintenance of renewable energy systems. In *Energy* **2021**, *Volume 221*

39. NIKIFOROVA, A. Smarter Open Government Data for Society 5.0: Are Your Open Data Smart Enough? *Sensors* **2021**, *21*, 5204. https://doi.org/10.3390/s21155204

40. JAMWAL, A.; AGRAWAL, R.; SHARMA, M.; GIALLANZA, A. Industry 4.0 Technologies for Manufacturing Sustainability: A Systematic Review and Future Research Directions. Appl. Sci. 2021, 11, 5725. https://doi.org/10.3390/app11125725

41. FLORESCU, A.; BARABAS, S.A. Modeling and Simulation of a Flexible Manufacturing System—A Basic Component of Industry 4.0. Appl. Sci. 2020, 10, 8300. https://doi.org/10.3390/app10228300

42. JIANG, J.-R.; KAO, J.-B.; LI, Y.-L. Semi-Supervised Time Series Anomaly Detection Based on Statistics and Deep Learning. Appl. Sci. 2021, 11, 6698. https://doi.org/10.3390/app11156698

43. MACH-KRÓL, M.; HADASIK, B. On a Certain Research Gap in Big Data Mining for Customer Insights. Appl. Sci. 2021, 11, 6993. https://doi.org/10.3390/app11156993

44. JELLASON, N.P.; ROBINSON, E.J.Z.; OGBAGA, C.C. Agriculture 4.0: Is Sub-Saharan Africa Ready? Appl. Sci. 2021, 11, 5750. https://doi.org/10.3390/app11125750

45. GHASEMAGHAEI, M.; CALIC, G.. Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decision Support Systems 2019, pp. 38-49

46. SCHUH, G.; REBENTISH, E.; RIESENER, M.; IPERS, T.; TÖNNES, C.; JANK, M. Data quality program management for digital shadows of products. Procedia CIRP. 2019, Volume 86, pp. 43-48

47. SAKIB, N.; WUEST, T. Challenges and Opportunities of Condition-based Predictive Maintenance: A Review. Procedia CIRP. 2018, Volume 78, pp. 267-272

48. RAGUSEO, E.; Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. International Journal of Information Management. 2018, Volume 38, Issue 1, pp. 187-195

49. KOVACS, O.; The dark corners of industry 4.0 – Grounding economic governance 2.0. Technology in Society. 2018, Volume 55, pp. 140-145
50. HARPER, D.: data (n.). Available online: https://www.etymonline.com/word/data (accessed on 19.05.2019)
51. Cambridge University Press. Available online: https://dictionary.cambridge.org/dictionary/english/data (accessed on 19.05.2019)
52. Data Types: Structured vs. Unstructured Data. Available online: https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/ (accessed on 19.05.2021)
53. KING, T.: Structured Data vs. Unstructured Data; What's the Difference?. Available online: https://solutionsreview.com/data-management/key-differences-between-structured-and-unstructured-data/ (accessed on 19.05.2019)
54. KW model for knowledge management and data value extraction. Available online: https://www.i-scoop.eu/big-data-action-value-context/dikw-model/ (accessed on 30.10.2019)
55. What is the DIKW Pyramid?. Available online: https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/ (accessed on 31.10.2019)
56. BRAHMACHARY, A.: DIKW Model: Explaining the DIKW Pyramid or DIKW Hierarchy. Available online: https://www.certguidance.com/explaining-dikw-hierarchy/ (accessed on 01.11.2019)
57. BIG DATA. Available online: https://sk.adastragrp.com/biznis-riesenia/big-data (accessed on 31.10.2019)
58. Big Data. Available online: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html (accessed on 28.09.2020)
59. Artificial Intelligence (AI). Available online: https://www.techopedia.com/definition/190/artificial-intelligence-ai (accessed on 28.09.2020)
60. COPELAND, J.: Artificial intelligence. Available online: https://www.britannica.com/technology/artificial-intelligence (accessed on 23.12.2020)
61. What Is Machine Learning?. Available online: https://www.mathworks.com/discovery/machine-learning.html (accessed on 16.05.2021)
62. NG, Andrew. 2015. *Machine Learning.* [s. l.] : Stanford University, 2015.
63. FAGGELLA, D.: What is Machine Learning?. Available online: https://emerj.com/ai-glossary-terms/what-is-machine-learning/ (03.02.2021)
64. What is Machine Learning? A definition. Available online: https://www.expertsystem.com/machine-learning-definition/ (accessed on 15.03.2021)
65. Machine Learning. Available online: https://www.sas.com/en_us/insights/analytics/machine-learning.html (accessed on 03.01.2021)
66. Deep Learning. Available online: https://www.mathworks.com/discovery/deep-learning.html (09.05.2021)
67. ŠPENDLA, Lukáš - HRČKA, Lukáš - TANUŠKA, Pavol. Proposal of knowledge discovery platform for big data processing in manufacturing. In *Mathematics and Computers in Science and Industry : MCSI 2015 : Sliema, Malta, August 17-19, 2015*. 1st ed.
68. ZVADA. [s. a.] *Proces objavovania znalostí z databáz.* Košice : Technická univerzita Košice
69. ŠIMONČIČOVÁ, Veronika - HRČKA, Lukáš - ŠPENDLA, Lukáš - TANUŠKA, Pavol - VAŽAN, Pavel. Pattern recognition for predictive analysis in automotive industry. In *Cybernetics and mathematics applications in intelligent systems : proceedings of the 6th Computer science on-line conference 2017. 26.4. - 29.4. 2017, Praha, ČR. Vol. 2*.
70. ALTON, L.: The 7 Most Important Data Mining Techniques. www.datasciencecentral.com, 2017, [cit. 16. máj 2019]. Dostupné na webovskej stránke (world wide web): https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques
71. PATRIZIO, A.: Top 15 Data Mining Techniques for Business Success. www.datamation.com, 2019, [cit. 16. máj 2019]. Dostupné na webovskej stránke (world wide web): https://www.datamation.com/big-data/data-mining-techniques.html
72. KOŠŤÁL, Michal. 2007. *Dolovanie dát a jeho techniky a využitie.* Bratislava : Slovenská technická univerzita, 2007.
73. BROWN, M.: Data mining techniques. developer.ibm.com, 2012, [cit. 16. máj 2019]. Dostupné na webovskej stránke (world wide web): https://developer.ibm.com/articles/ba-data-mining-techniques/
74. TANUŠKA, Pavol - ŠPENDLA, Lukáš - KEBÍSEK, Michal - VAŽAN, Pavel - HRČKA, Lukáš. Data integration and transformation proposal for big data analyses in automotive industry. In *INES 2017*
75. What is the CRISP-DM methodology?. Available online: https://www.sv-europe.com/crisp-dm-methodology/ (accessed on 03.02.2021)
76. CRISP - DM. Available online: https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf (accessed on 13.12.2020)

77. VORHIES, W.: What is the CRISP-DM methodology?. Available online: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome (accessed on 09.08.2020)
78. The CRISP-DM Methodology. Available online: http://www.pamanalytics.com/downloads/The%20CRISP-DM%20Methodology.pdf (accessed on 28.02.2021)
79. RapidMiner Studio. Available online: https://rapidminer.com/products/studio/ (accessed on 22.09.2020)
80. KUHLMAN, D. A Python Book: Beginning Python, Advanced Python, and Python Exercises. Platypus Global Media, 2011.
81. Elasticsearch. Available online. https://www.elastic.co/elasticsearch/ (accessed on 06.04.2020)
82. MOUAKHER, A.; YAHIA, S. B. QualityCover : Efficient binary relation coverage guided by induce d knowle dge quality. In: *Information Sciences* **2016**, pp. 355-356
83. Formal concept analysis. Available online: https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume24/cimiano05a-html/node3.html (accessed on 15.04.2021)
84. SHUNG, K. P. Accuracy, Precision, Recall or F1? Available online: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 (accessed on 19.05.2020)
85. Machine Learning: GridSearchCV & RandomizedSearchCV Papa Moriba Kouate. Available online: https://towardsdatascience.com/machine-learning-gridsearchcv-randomizedsearchcv-d36b89231b10 (accessed on 31.10.2020)
86. Explaining the 68-95-99.7 rule for a Normal Distribution. Available online: https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2 (accessed on 03.02.2020)

LIST OF PUBLICATIONS

Proposal of effective preprocessing techniques of financial data

ABASOVÁ, J. -- JÁNOŠÍK, J. -- ŠIMONČIČOVÁ, V. -- TANUŠKA, P. Proposal of effective preprocessing techniques of financial data. In *INES 2018.* IEEE, 2018: 2018, p. 000293--000298. ISBN 978-1-5386-1121-0.

Preparing influence analysis of meteoparameters on production process

ABASOVÁ, J. -- TANUŠKA, P. Preparing influence analysis of meteoparameters on production process. In SILHAVY, R. *Software Engineering and Algorithms in Intelligent Systems.* Cham: Springer, 2019, p. 109--120. ISBN 978-3-319-91185-4.

Proposal of data pre-processing for purpose of analysis in accordance with the concept Industry 4.0

GRIGELOVÁ, V. -- ABASOVÁ, J. -- TANUŠKA, P. Proposal of data pre-processing for purpose of analysis in accordance with the concept Industry 4.0. In ŠILHAVÝ, R. *Artificial Intelligence Methods in Intelligent Algorithms.* Cham: Springer Verlag, 2019, p. 324--331. ISBN 978-3-030-19809-1.

Proposal of data preparation model for Big Data analytics in painting process

ABASOVÁ, J. -- GRIGELOVÁ, V. -- TANUŠKA, P. Proposal of data preparation model for Big Data analytics in painting process. In HOCKICKO, P. -- DUBOVAN, J. *ELEKTRO 2020.* Danvers, USA: IEEE, 2020, p. 1--6. ISBN 978-1-7281-7542-3.

Classification models for purpose of predictive maintenance in a production process

ABASOVÁ, J. -- RYDZI, Š. -- ZAHRADNÍKOVÁ, T. -- TANUŠKA, P. Classification models for purpose of predictive maintenance in a production process. In KISS, B. -- SZIRMAY-KALOS, L. *Proceedings of the Workshop on the Advances of Information Technology 2021*. p. 145--149. ISBN 978-963-421-844-9.

Big Data – Knowledge discovery in production industry data storages – implementation of best practices

ABASOVA, J. – TANUSKA, P. – RYDZI, S. Big Data – Knowledge discovery in production industry data storages – implementation of best practices.

The article has been accepted for publication in MDPI journal *Applied Sciences* and is now in proceeding to be published.