Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 79 (2019) 403–408

www.elsevier.com/locate/procedia

12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 18-20 July 2018, Gulf of Naples, Italy

DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model

Steffen Huber[a], Hajo Wiemer[a,*], Dorothea Schneider[a], Steffen Ihlenfeldt[a]

[a] *Technische Universität Dresden, Institute for Machine Tools and Control Engineering, 01062 Dresden, Germany*

* Corresponding author. Tel.: +49-351-463-32004; fax: +49-351-463-37073. *E-mail address:* hajo.wiemer@tu-dresden.de

Abstract

The value of data analytics is fundamental in cyber-physical production systems for tasks like optimization and predictive maintenance. The de facto standard for conducting data analytics in industrial applications is the CRISP-DM methodology. However, CRISP-DM does not specify a data acquisition phase within production scenarios. With this work, we present DMME as an extension to the CRISP-DM methodology specifically tailored for engineering applications. It provides a communication and planning foundation for data analytics within the production domain. We show the feasibility of our methodology for engineering applications within a case study in the field of work piece detection.

Keywords: Data mining; Machine learning; Manufacturing data management; Data driven process optimisation

1. Introduction

During the recent years the availability, multitude and potential of useful manufacturing-related data increased massively due to the emerging technologies in the areas of sensor integration and sensor connectivity, which have evolved alongside the ever-growing technical possibilities and the progression of Industry 4.0, Industrial Internet of Things (IIoT) and Cyber-physical Production Systems (CPPS) [1]. These technologies increase the monitoring and networking capabilities along production processes, enabling the massive collection of real-time data about process resources [2].

However, this increasing output in manufacturing-related data is often not utilized to its full extend, e.g. for condition monitoring of individual machine components, or process optimization. Engineers setting up machines and putting production processes into operation are motivated to prepare their facilities for Industry 4.0 applications and equip them with capabilities to provide internal as well as external sensor data. However, while putting the process into operation, obliged to deadlines and various constraints, practitioners tend to rely on traditional procedures and experience-based knowledge. At this stage, data analytics implementation is laid off to some point in the future when further process optimization needs to be obtained. This implies the need to bring in potential experts in the field of data science. Although, to implement useful data analytics procedures requires not only data mining expertise, but also the related domain knowledge. In many cases, the cross-disciplinary cooperation of various domain experts such as data scientists, process and control engineers is required to conduct successful data mining projects [6].

As smaller companies usually cannot afford this kind of service personnel and furthermore qualified experts are rare and high in demand, this expectation tends to impede the implementation process instead of serving it. Furthermore, the existing set-up may already be disadvantageous for the respective data analytics task, when not taken into account adequately from the beginning. In the computer science and business domain the use of data mining methods already is a widely established concept underpinned by sophisticated algorithms and comprehensive concepts. Many procedure models and attempts to standardize the data mining process have been undertaken during the recent years. One especially widespread approach is the Cross-industry standard process for data mining (CRISP-DM) [3]. Although widely accepted in industrial data mining as a kind of de-facto standard, CRISP-DM and similar process models do not cover the domain specific difficulties in obtaining and processing data and furthermore the

2212-8271 © 2019 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering.
10.1016/j.procir.2019.02.106

knowledge involved in the engineering context [4] [5]. Therefore, the technical understanding of the data analysis objects, e.g. machines and production processes, as well as the technical data acquisition by means of machine controls and external sensors, are not yet part of existing data mining methodologies. With this work, we introduce DMME – Data Mining Methodology for Engineering Applications – as an extension to the established CRISP-DM standard. DMME provides a holistic view onto data analytics within the production domain that supports the satisfaction of necessary preconditions for the successful implementation of data driven process and machine analysis. In addition, DMME is developed to be accessible for engineers implementing it to support startup operations as well as optimize production or maintenance processes throughout the value chain.

The paper is structured as follows: Section 2 introduces the CRISP-DM standard and provides a motivating example application for data mining in the production domain. Based on the example, the gap between the methodological requirements and the CRISP-DM standard is illustrated. Section 3 defines our proposed solution DMME as an extension to the CRISP-DM standard. In Section 4, we show the feasibility and transferability of DMME within a work piece detection scenario. Section 5 presents future work and concludes.

Nomenclature

DM	Data Mining
CRISP-DM	Cross-industry standard process for DM
KD	Knowledge Discovery
ML	Machine Learning
IIoT	Industrial Internet of Things
CPPS	Cyber-physical Production Systems
DoE	Design of Experiements

2. State-of-the-Art and Requirements

Data mining methodologies were introduced to provide a more holistic view to the knowledge discovery process, beyond the application of statistical or machine learning algorithms. Their goal is to define a generic workflow, starting with the determination of relevant questions, leading to the targeted processing of the raw and often unstructured data and ultimately aiming at the discovery of new knowledge. Even though model building algorithms are important building blocks of any knowledge discovery process, their selection and application only consume a fraction of the total time involved. In contrast, the acquisition, manual cleaning and preparation of data often requires a considerable time effort.

2.1. Applying CRISP-DM to Manufacturing

The cross-industry standard process for data mining (CRISP-DM) is a framework for translating business problems into data mining tasks and carrying out data mining projects independent of both the application area and the used technology [3]. It is a widely adopted industry-oriented implementation of the generic Knowledge Discovery (KD) process, as described in [7].

Figure 1 shows the six phases of the CRISP-DM process model and their interactions. Any data mining project starts with the project's goal definition that is included in the first phase **"Business Understanding"**. In a manufacturing scenario, a common motivating business goal is to maximize the uptime and efficiency of machines by using predictive analytics. This goal is then transformed into a specific data mining problem, e.g. by identifying the relevant machine components and specifying the physical effects of machine component deterioration. During the **"Data Understanding"** phase hypotheses for hidden information regarding the data mining project goal are formed based on experience and qualified assumptions. In case of a predictive maintenance scenario with the aim to detect the deterioration of bearings, a valid concept would be to look for new frequency patterns in data streams from a motion sensor. In the **"Data Preparation"** phase the engineer collects the relevant data and prepares it for the actual data mining task. This includes the preprocessing, e.g. data reduction and filtering, as well as feature generation with respect to the data mining project goal. Considering the deterioration of bearings, it is a feasible approach to use a Fast-Fourier Transformation (FFT) to eliminate the time component from the signal input and to provide frequency patterns with a fixed length. In addition, the data sets can be labeled according to the engineers' expert domain knowledge, such that different stages and fault types of bearing deterioration are represented in the data set. In the **"Modeling"** phase a data mining workflow is constructed to find the desired parameter settings for the selected algorithms and to execute the data mining task on the preprocessed data. In the predictive maintenance scenario, depending on the size of the training data set, obvious options would be to choose a k-Nearest-Neighbor (kNN) or a neural network-based classification algorithm. The result is a trained classification model including the different stages of deterioration and types auf bearing faults. Within the subsequent **"Evaluation"** phase the trained model is tested against real data sets within a production scenario and the data mining results are assessed according to the underlying business objectives. For this purpose, test data sets are generated following the steps developed in the "Data Preparation" and "Modeling" phases

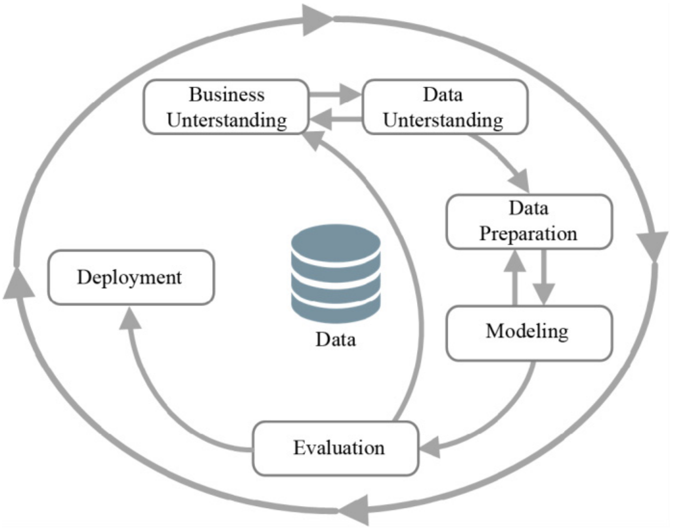


Fig. 1: The CRISP-DM process [3].

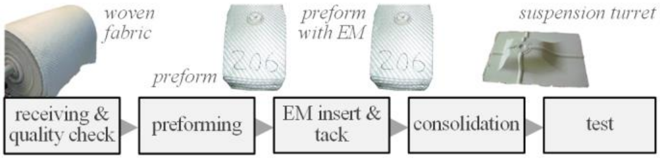
excluding the labeling step. After successful evaluation of the trained model, it is deployed into production in the **“Deployment”** phase. However, the deployment also requires a stable set-up for data acquisition including a sensor and data processing infrastructure.

2.2. Characteristics of Data Analytics in Engineering

To define the typical characteristics of data analytics tasks in engineering we analysed a multitude of research projects carried out during the recent years. In the following we illustrate the synthesized findings referencing to one representative project [8]. Where possible, the workflow is aligned with the steps of the CRISP-DM methodology. The example shows a typical application of data-driven methods to develop novel manufacturing technologies [9]. The project targeted the development of methods for the design and use of novel “textile-reinforced composite components for function-integrating multi- material design in complex lightweight applications”, cf. figure 2. A key subject of the works were technological procedures for efficiently producing textile-reinforced composite components.

A subproject was dedicated to the development of data-driven modelling and simulation of process chains for a targeted adjustment of pre-defined characteristics as well as assuring a reproducible manufacturing of thermoplastic textile composite components. The focus lay on the production of a glass fiber thermoplastic composite component suspension turret with an integrated electronic module. In the project, the IoT-Dashboard Detact® was used, which provides the necessary functions from the connection of the data sources and the analysis of the data to the visualization of the results. The subproject goals were the determination of stable process parameters and getting to investigate the process-immanent interactions. From the data mining perspective, these objectives were achieved within the following work packages:

- **Business Understanding**: The project objective was the development of a manufacturing chain for the composite component suspension turret. The first goal was to ensure quality. The second goal was to achieve short cycle times. The process chain includes the operations: receiving of textile semi-finished products and their quality check, preforming, assembly with electronic modules, consolidation and testing. For this purpose, the dimensions, surface quality and mechanical properties of the component were tested, as well as the functionality of the electronic modules.



Fig, 2: Production process of textile-reinforced composite components.

- Next, the existing influences on the quality parameters and cycle time were analyzed. For this purpose, the *Ishikawa* method was used to systematically derive the influences on the materials, the used production equipment and the environment for each individual operation.
- In the next step, the methods for measuring and logging as well as for collecting the data for all relevant influencing parameters were developed. Some parameters were available in the machine controls, so the interface had to be adjusted to connect to the machine controls. Other parameters required the use of additional sensors or entire measuring devices. In addition, there were also manually added notes necessary that had to be entered into a weblog of the Detact® software.
- After the conceptual design of the parameter acquisition, the sensors and measuring devices as well as the interfaces were installed and tested at each workstation. Afterwards, all parameter channels, i.e. data sources, were integrated into the data management system Detact®.
- Subsequently experiments were conducted to enable the assessment of stable process parameters and establish process understanding. The investigations were planned with methods of the Design of Experiments (DoE), so that on the one hand the complexity could be mastered and on the other hand the effort remained limited. DoE is particularly useful for investigations on process chains with large parameter sets and complex interactions. In addition, to apply DoE methods supports the creation of a particularly valuable data set, thus preparing the following data analysis steps and serving the acquisition of useful results during this phase [10].
- **Data Understanding** and **Data Preparation**: Then, the production tests were carried out and the manufactured components were tested producing the required data sets. The investigations were comprehensively recorded directly in Detact®. This created an analyzable database that was now available for the required analyses.
- **Modeling**: The analysis work package finally carried out the different data analyses. For example, the following questions were of interest: What are the influences on component quality? Which is the optimal range of process parameters? How does the process affect the yield of the electronic modules? How do the electronic modules influence the mechanical properties of the component?
- **Evaluation** and **Deployment**: The final work package focused on evaluating the results in a small series of experiments. Finally, the technology expertise and the data-driven methods were published.

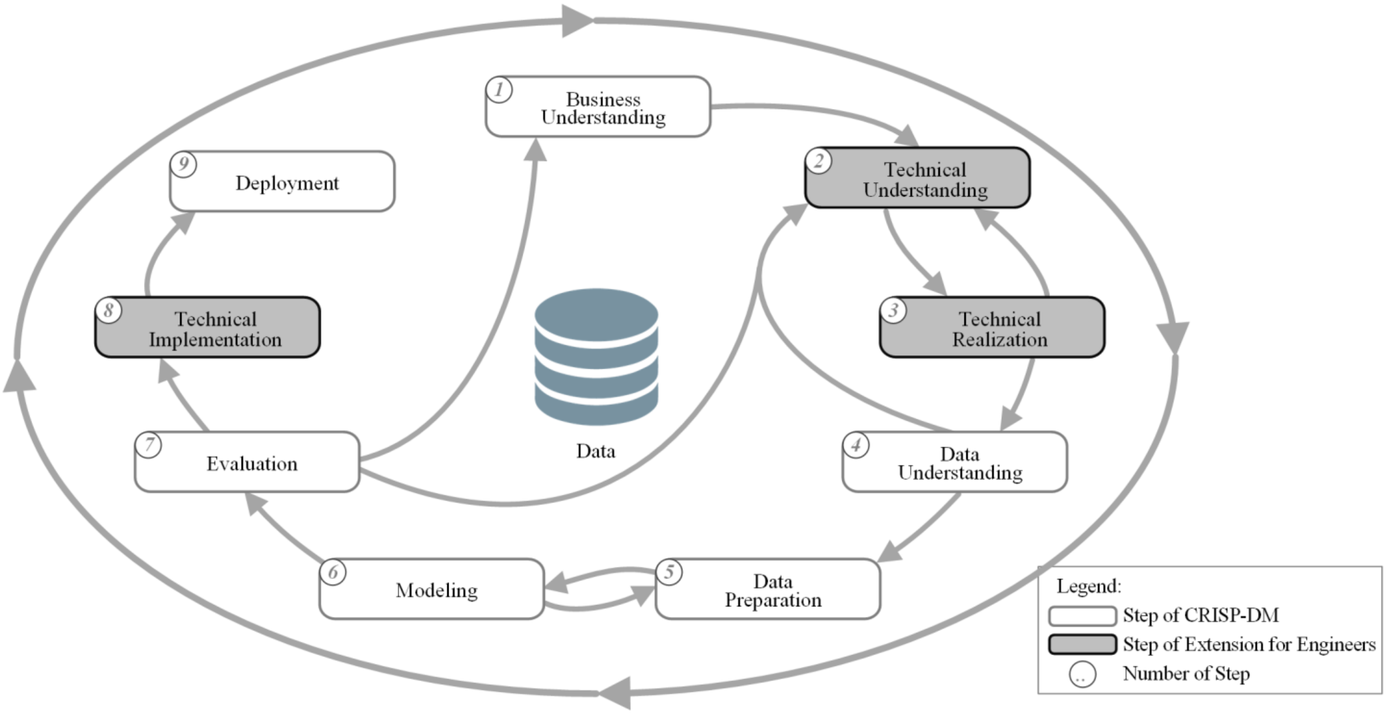


Fig. 3: The DMME process.

2.3. Requirements for a DM-Methodology in Engineering

The given example shows that in the field of engineering several essential tasks have to be executed to identify and acquire the relevant data to enable any subsequent data analytics tasks. This implies that the procedure for data mining in engineering requires additional steps exceeding the cross-domain CRISP-DM methodology. Specifically, the example entails the following requirements for the extension of CRISP-DM:

1. Consideration of the **technical implications** of the business goals. Referring to the example, this corresponds to the application of the *Ishikawa* method to derive parameter influences and the selection of the relevant physical parameters for subsequent measurements (Section 2.2 - Step 2).
2. **Technical development** of a data acquisition method for each of the relevant physical parameters by means of available or additional sensors, interfaces and software. In the example, this corresponds to the concept of logging machine control data as well as the requirements for additional sensors (Section 2.2 - Step 3).
3. **Technical implementation** of the data acquisition methods on site. This corresponds to the actual implementation and installation of the required hardware and software infrastructure (Section 2.2 - Step 4).
4. **Experiment planning** to acquire the desired data sets based on the selected parameters. With regard to the example, this corresponds to the application of DoE for structured and valuable experiment results (Section 2.2 - Step 5).

All of the above requirements originate from the technological context specific to engineering applications. In contrast to existing data mining methodologies like CRISP-DM or KDD, the foundation for analytics, i.e. the data itself, has to be generated first with respect to the specific business goal and to the underlying technological and physical relations. As there are usually production processes or other technical context restrictions involved, the data acquisition has to be considered as an individual step inside the data analysis process.

3. Holistic DM Methodology for Engineering Applications

To cover the requirements defined in Section 2.3 and to provide means for communication and planning in DM projects within the engineering domain, we developed DMME as an extension to the CRISP-DM standard. The goal was to minimize the time exposure for future DM projects by providing a systematic process for the development and documentation of all the required steps allowing for the partial reuse of valuable know-how.

We developed DMME by analyzing the requirements based on several research projects, which each included some kind of data analytics tasks. These comprise production process optimization, condition monitoring of machine components as well as predictive maintenance.

As a result, we provide three novel steps as an extension to the CRISP-DM model as depicted in Figure 3. These are the **"Technical Understanding"**, **"Technical Realization"** and **"Technical Implementation"**, which will be discussed in detail in the following:

3.1. Technical Understanding

The goal of this task is to transform the business goals into measurable technical goals, gather existing expertise of the related physical and process effects as well as to develop an experiment plan. This includes the following steps:

- Analyzation of the **system structure**, process and related parameters.
- Definition of the **technical goals** and target variables.
- Deduction of the technical **analysis tasks**.
- Gathering and documentation of existing know-how about related **physical effects** and **basic conditions**.
- Definition of **relevant physical parameters** and effects for subsequent measurements.
- Definition of **concepts for measuring** each of the relevant physical parameters.
- Develop an **experiment plan** for conducting the measurements.

3.2. Technical Realization

The goal of this task is to test and select the measuring concepts and to conduct the experiment plan. As a result, data is generated that includes all the relevant information and features for the subsequent data analytics task to reach the defined business goals. This includes the following steps:

- Develop a **technical test setup** by selecting the best-fit data acquisition methods from the measurement concepts.
- Conduct the **experiment plan**.
- **Document all steps** of the data acquisition process including technical limitations, possible sources of errors as well as the level of data quality.

3.3. Technical Implementation

The goal of this task is to enable the evaluated model to be provided with run-time data during production. Therefore, the data acquisition method from the **"Technical Realization"** phase of DMME is transformed into a run-time capable infrastructure. This includes the following tasks:

- Check all used sensor technology for their capability to **stream data over a long period**, i.e. provide the required power supply, cope with connection loss and provide self-monitoring capabilities, e.g. by redundant sensors.
- Develop or select a **software infrastructure** to gather, preprocess and analyze the machine control and sensor data streams over a long period.

4. Case Study: Work Piece Detection on a Profiled Rail Guide Carriage using Vibration Analysis

In the following, we present a case study as a proof-of-concept and to show the transferability of the DMME process. The goal was to implement a robust work piece detection, i.e. object detection, on a profiled rail guide carriage based on vibration analysis. For this purpose, we used a Bosch Rexroth profiled rail guide and an IndraControl XM22 machine control. We describe the application of DMME as a case study with emphasis on the newly introduced tasks.

4.1. Business Understanding

The business goal was to detect the type of work piece that was transported on a profiled rail guide. A robust work piece classification could be useful to adapt production process parameters of a machine tool according to the work piece.

4.2. Technical Understanding

The system structure was rather simple, consisting of the machine control, an inverter, the electric drive and the profiled rail guide with a carriage system. The technical goals that followed from the business goal are the creation and detection of vibration signals at the carriage system. To decide if the vibration analysis could be used to detect different types of objects, we conducted a literature review in the domain of condition monitoring of machine tools, where the analysis of vibration modes is a common method. Consequently, we confirmed to develop a method based on vibration signals. Due to the domain knowledge in control programming, we decided to generate a low-frequency vibration signal within the electric drive with a control program. For measuring the generated vibration signal, we chose an acceleration sensor located directly at the carriage system. Our experiment plan for a proof-of-concept validation included several test-runs and measurements for different generated vibration signals ranging from 80 Hz to 500 Hz. These boundaries resulted from the basic conditions of the control program and the electric drive. Furthermore, we planned to classify five different types of work pieces for each of the selected frequencies.

4.3. Technical Realization

For the technical test setup, we implemented the control program that provided a TCP/IP interface for setting the frequency to generate and trigger test runs of the profiled rail guide. Furthermore, we selected a Bosch XDK IoT prototyping platform for measuring the acceleration signals at the carriage system, as it provides wireless connectivity capabilities, e.g. Wi-Fi and Bluetooth LE, as well as a battery pack, a microcontroller, and amongst others, an acceleration sensor with a sampling frequency of 1 kHz. We programmed the firmware for the XDK microcontroller to publish the acceleration sensor data sampled at 1 kHz for all three available axes via the MQTT protocol to a local MQTT broker. Since it provides an easy to manage build-up of a software prototype within a short time span, we developed a data acquisition workflow with Node-Red. The Node-Red workflow was setting a user-defined frequency and started a test run in the machine control via the TCP/IP interface. During the test run, the published acceleration data of the Bosch XDK was gathered via a MQTT Client and written to a CSV-File. With this setup, we executed the experiment plan and generated labeled test data for subsequent data analysis.

4.4. Data Understanding

Based on the generated data sets from the experiment plan, we consulted with a data scientist experienced in the analysis of acceleration, i.e. vibration, data. Based on this input, we aimed at selecting classifying features, e.g. frequency components, within the data by eliminating the time component, e.g. by employing a Fast-Fourier Transformation (FFT).

4.5. Data Preparation

We then transformed the acceleration data sets into fixed length data vectors by employing a FFT within a MATLAB script, while preserving the class labeling information.

4.6. Modeling

For the modeling, we tested different classification algorithms in the MATLAB Statistics and Machine Learning Toolbox. In result, we chose the kNN (k-Nearest-Neighbor) classification algorithm as it could handle the small size of the training data set very well. In addition, it is able to add additional training vectors and classes during run-time. Then, we conducted a Leave-one-out cross validation (LOOCV) on the training data set. The results ranged between 90 % – 100 % classification accuracy for the different frequency ranges.

4.7. Evaluation

We evaluated the model against new types of work pieces and again achieved a high classification accuracy of over 90 %.

4.8. Technical Implementation

To deploy the developed method for the robust classification of work pieces, we developed a software architecture able to cope with the connectivity, data streaming as well as data processing and visualization of results. As a basic condition, the software had to run on-site within an industrial testbed. We choose a Raspberry Pi Model 3B in an industrial case to host the necessary software components. These include an open source MQTT broker (Mosquitto), a Java server backend for communication with the machine control and the MQTT broker and for data processing as well as an Angular 5 Frontend for visualizing the classification results.

4.9. Deployment

All of the system components were then successfully deployed to an industrial testbed.

4.10. Discussion

The example illustrates the benefit of the DMME as the **Technical Understanding**, **Technical Realization** and the **Technical Implementation** are essential steps within many data mining projects in the engineering domain. For now, DMME provides a simple workflow and checklist to enable a better structure, communication and documentation of these crucial tasks. It is important to note, that especially the **Technical Understanding** and **Technical Realization** phases require a close cooperation of personnel with highly specialized expertise, such as process and construction engineers, control programmers, mathematicians, data scientists, software engineers, quality managers and so on.

5. Conclusion and Future Work

Besides the technical possibilities and challenges of Industry 4.0 solutions, we also see the cooperation of highly specialized personnel as a prerequisite for the successful development of the fourth industrial revolution. With DMME, we provide a first solution to guide and document the cooperation process within data mining projects in the engineering domain. For future work, we aim to develop the DMME further to support a more fine-grained level of subtasks as well as additional views including the data, data quality, IT-system, engineering and business perspective as well as their interconnections, representing the cooperation of the involved personnel. In addition, we plan to develop a software assistant to manage these perspectives and guide the different user groups through the related workflows.

Acknowledgements

This research has received funding under the grant number 100302264 (“Smart Data Services für Produktionssysteme” project) by the European Social Fund (ESF) and the German Federal State of Saxony.

References

[1] Xu LD, He W, Li S. Internet of things in industries: A survey. IEEE Transactions on Industrial Informatics. 2014; vol. 10, no. 4, pp. 2233–2243.

[2] Huber S, Seiger R, Kuhnert A, Theodorou V, Schlegel T. Goal-based semantic queries for dynamic processes in the internet of things. Int. Journal of Semantic Computing. 2016; vol. 10, no. 2, p. 269.

[3] Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. 2000; pp. 29–39.

[4] Montero J. Introduction to data mining and its applications to manufacturing. Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. 2008; vol. 3.

[5] Azevedo AIRL, Santos MF. KDD, SEMMA CRISP-DM: A parallel overview. IADS-DM. 2008.

[6] Choudhary AK, Harding JA, Popplewell K. Knowledge discovery for moderating collaborative projects. 2006 4th IEEE International Conference on Industrial Informatics. 2006; pp. 519–524.

[7] Brachman RJ, Anand T. The process of knowledge discovery in databases: A first sketch. AAAI Press. 1994; pp. 1–11.

[8] Modler N, Hufenbach W, et al. Novel Hybrid Yarn Textile Thermoplastic Composites for Function-Integrating Multi-Material Lightweight Design. Advanced Engineering Materials. 2016; vol. 18, no. 3., pp 361-368.

[9] Gröger C, Niedermann F, Mitschang B. Data mining-driven manufacturing process optimization. Proceedings of the world congress on engineering. 2012; vol. 3, pp. 4–6.

[10] Wiemer H, Schwarzenberger M, Dietz G, Juhrisch M, Ihlenfeldt S. A Holistic and DoE-based Approach to Developing and Putting into Operation. Complex Manufacturing Process Chains of Composite Components, Procedia CIRP. 2017; vol. 66, pp 147-152.