How to Discover Knowledge for Improving Availability in the Manufacturing Domain?

Fabian Utz
Hochschule Darmstadt
University of
Applied Sciences, Germany
fabianutz@gmail.com

Christian Neumann
Freudenberg IT
GmbH & Co. KG
Germany
christian.neumann@freudenberg-it.com

Omid Tafreschi
Hochschule Darmstadt
University of
Applied Sciences, Germany
omid.tafreschi@h-da.de

Abstract

This paper presents a specific process model for Knowledge Discovery in Databases (KDD) projects aiming at availability improvement in manufacturing. For this purpose, Overall Equipment Efficiency (OEE) is analyzed and used, since it is an approved approach to monitor and improve the degree of availability in manufacturing. To define the specific process model, we use the generic CRISP-DM reference model and conduct a mapping for availability improvement. We prove the applicability of our model in the context of a specific KDD project in a large enterprise in the manufacturing industry.

1. Introduction

Manufacturing companies strive for the detection and reduction of loss in their productions to improve their efficiency and competitiveness. One approved approach is to calculate the Overall Equipment Effectiveness (OEE). OEE is a Key Performance Indicator (KPI) that represents the degree of effectiveness of a production and compasses the three factors performance, quality and availability [1], [2].

Information Technology enables new possibilities to improve OEE. Technical developments and the increasing availability of manufacturing data resulting from automation and digitization not only facilitate an easier, automatized and more accurate calculation but also allow for a proactive improvement of OEE [3], [4]. Particularly, the field of Knowledge Discovery in Databases (KDD) gained relevance in using manufacturing data for knowledge generation, e.g., predicting the failure of a machine and increasing availability by timely maintenance.

Cross Industry Standard Process for Data Mining (CRISP-DM) is the de facto standard to realize such KDD projects in industry [5]. It is characterized by its general applicability in many domains. However, this generality leads to costly mappings for specific domains.

According to [6], a mapping of CRISP-DM to specific domains is required and can be done either for a single project, i.e., mapping for the present, or for a set of projects, i.e., mapping for the future. In this paper, we propose a mapping for the future in the manufacturing domain. In other words we provide the answer to following question: How to discover knowledge for improving availability in the manufacturing domain?

For this purpose, the paper delivers a specific process model for KDD projects in the manufacturing domain using CRISP-DM. The proposed process model provides a bridge to the gap in-between the generic CRISP-DM and its specific deployments in the manufacturing domain. In other words, the proposed model contains a set of specific, efficient, and reusable processes required for conducting KDD projects in the mentioned domain. To assess its efficiency and reusability, we apply the process model to a specific KDD project and analyze the efforts and results.

The paper is organized as follows. After the introduction, Section 2 provides an overview of KDD by depicting the reference model CRISP-DM and discussing KDD applications in the manufacturing domain. Section 3 presents a specific process model for availability improvement based on CRISP-DM by outlining the modeling approach and describing each phase in detail. In order to assess the applicability of the proposed process model, it is applied to a KDD project in manufacturing domain in Section 4. Section 5 discusses the related work and Section 6 provides the conclusion and future work.

2. KDD

As first defined by Fayyad et al., KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [7]. The KDD process is highly iterative and interactive and contains different phases depending on the reference model used. In this section we provide an overview of KDD by depicting

URI: http://hdl.handle.net/10125/50441
ISBN: 978-0-9981331-1-9

the approved reference model CRISP-DM. We then discuss OEE as KPI and classify KDD applications of the manufacturing domain into the three OEE factors describing their overarching goal.

2.1. CRISP-DM

Chapman et al. define CRISP-DM as generic process model for KDD projects in industry [6]. Today, CRISP-DM is the de facto standard and utilized by the majority of KDD experts [5]. Figure 1 provides an overview of the CRISP-DM process which contains six phases. Each phase includes generic tasks, which require specific mappings for different domains. A mapping for the manufacturing domain is presented in Section 3.

The CRISP-DM process is iterative with non-strictly defined loops between phases [8]. This is due to the variety of different outcomes of each phase that determine the next step. The overall cyclical nature of a KDD project results from gained experiences that can be applied in future projects. Below, the six phases are described generically [6]:

The initial Business Understanding phase focuses on understanding the project objectives, defining a data mining problem and a preliminary plan to solve the problem.

In the phase Data Understanding a big picture of available and useful data is created. Data quality problems are identified, first insights into the data are discovered and interesting subsets are detected to form hypotheses regarding hidden information.

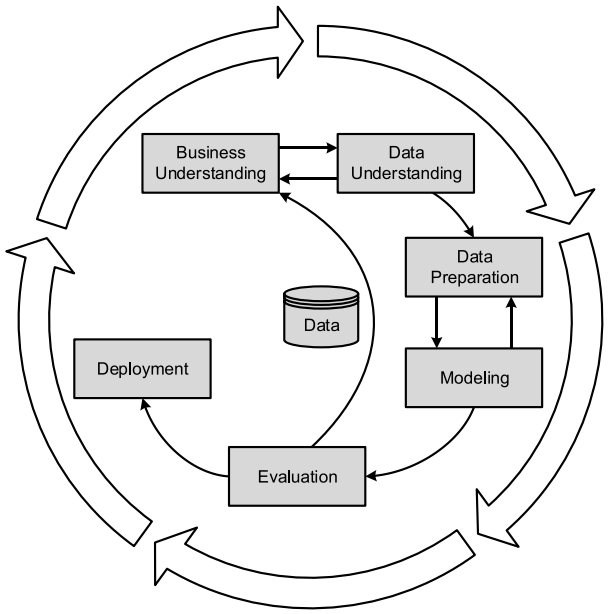

Figure 1. CRISP-DM Lifecycle [6]

The Data Preparation phase covers all activities needed to construct the final dataset from initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection, as well as transformation and cleaning of data for modeling tools.

In the phase Modeling, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the Data Preparation phase is often necessary.

Before the Evaluation phase, a model was built that appears to have high quality from a data analysis perspective. To assure that the model properly achieves the project objectives, it is important to thoroughly evaluate it and review the steps executed during creation, before proceeding to the final deployment of the model.

Creation of a model is generally not the end of the project, because the solution has to be deployed. Even if the purpose of the model is to increase knowledge based on data, the knowledge gained will need to be organized and presented in a way that it can be used. This often involves applying live models within an organization's decision making processes. These activities are done in the phase Deployment.

2.2. KDD in Manufacturing

Automation and digitization have found their ways into manufacturing. As stated in the 2016 Global Industry 4.0 Survey from PricewaterhouseCoopers (PwC), digitization in manufacturing companies has reached 33 % in 2016 and is expected to be doubled by 2020 [9]. With these technological achievements high volumes of structured and unstructured data is created, which can be used to generate knowledge and improve OEE in the long term. To be successful in this manner, two prerequisites have to be met. First, the effective questions from a manufacturing perspective to improve OEE have to be defined. Second, relevant data to answer these questions has to be identified. Therefore, we analyze the calculation of OEE determined as proposed by [10]:

$$OEE = Availability \cdot Performance \cdot Quality \quad (1)$$

Further details for the calculation are provided by [1] and [11]. Availability is the percentage of scheduled time that a machine is available to operate and is reduced by machine failures, unplanned maintenance, and set up. Performance is the speed at which the machine runs as a percentage of its designed speed and is reduced by microstops and reduced speed. Quality represents the good units produced as a percentage of the total units produced. Waste and rework, i.e., products that do not or did not meet defined specifications, reduce quality.

There is a set of concepts which are related to OEE. We classify them to the three OEE factors:

- Availability improvement
  - Predictive Maintenance [12], [13], [14]
  - Prognostics and Health Management [13], [15]
- Performance improvement
  - Predictive Production Planning [16]
  - Predictive Manufacturing Control [17]
- Quality improvement
  - Predictive Quality Control [18]
  - Control Chart Pattern Recognition [19], [20]

These concepts provide theoretical foundations and require additional specifications for direct applicability. For this purpose, we propose a specified process model focusing on improving availability, since the costs for keeping a high availability represent 15 – 60 % of total costs in manufacturing [21]. Additionally, availability can be quantified in an accurate and reasonable manner. In focusing on one OEE factor, i.e., availability, we can define a specific, applicable and reusable mapping of CRISP-DM for the manufacturing domain.

3. Mapping of CRISP-DM for Increasing Availability

In this section we provide a specific KDD process model for availability improvement based on CRISP-DM. The basic idea is outlined and followed by a detailed description of each of the six phases including their specific tasks.

3.1. Basic Idea

Chapman et al. recommend a deductive approach for mapping CRISP-DM [6]. However, an efficient mapping requires consideration of specific requirements of a domain. Therefore, we conducted several KDD projects in the manufacturing domain and studied specific models from other domains as proposed by [8], [22]. This approach was inductive and complements the recommended deductive approach. As a result, we derive a holistic mapping of CRISP-DM including a specific modeling of tasks. For this purpose, following options exist:

1) an existing task is used without changes;
2) an existing task is omitted due to missing relevance;
3) an existing task is specified to support availability improvement;
4) a new task is added to a phase.

We use Business Process Model and Notation 2.0 (BPMN 2.0) to visualize the process models, since it is the de facto standard for business process modeling and prevalently used in research and industry [23], [24]. Figure 2 depicts the generic CRISP-DM process model.

XOR operators are used to model feedback loops between phases and tasks which are necessary to enable an iterative approach. It must be noted that the BPMN 2.0 models provide an overview of each phase. The conditions of jumping from one phase or task into another are very complex and versatile. In sake of clarity, we refrain from detailing these conditions in all process models.

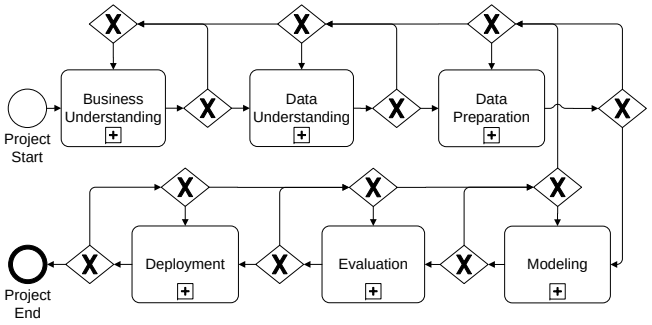

Figure 2. Process Model

The model depicted by Figure 2 has six collapsed sub-processes. We explain them by providing an expanded view for each sub-process in the following.

3.2. Business Understanding

Figure 3 presents the tasks required in the Business Understanding phase, i.e., Describe Production Environment, Define Project Goal, Define Data Mining Goal, Perform Project Management.

Highly automated machines and complex manufacturing processes are key features of today's productions. A basic understanding of these processes and machines as well as sensors and the maintenance strategy is a prerequisite for applying KDD methods to improve availability. For this purpose, we define the new task Describe Production Environment.

The task Define Project Goal answers the question, what the outcome of the project is and thus should be related to availability improvement. Quantifying the goal in terms of availability improvement helps evaluating it at the end of the project.

To break down the project goal and define the answer for the question of how availability can be improved, a detailed data mining goal is defined in task Define Data Mining Goal. Basically, two options exist to answer the question. We can find the root cause for failures and thus

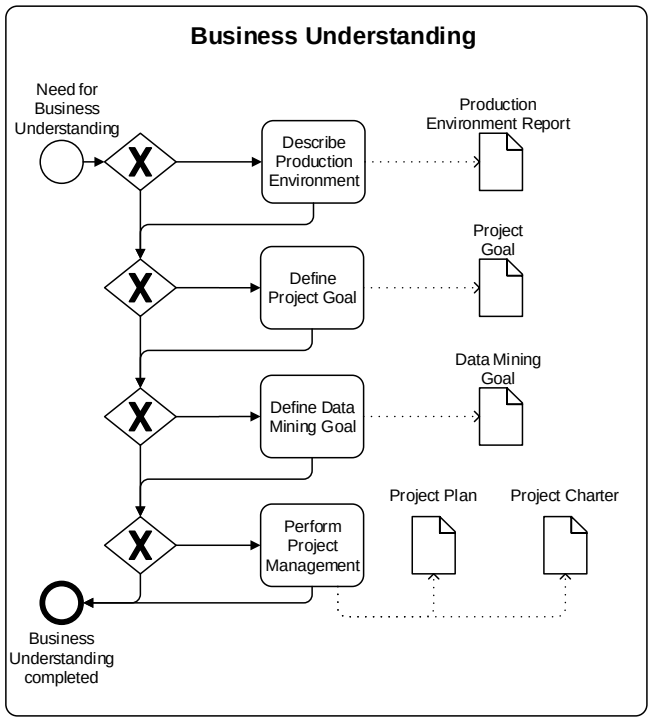

Figure 3. Phase Business Understanding

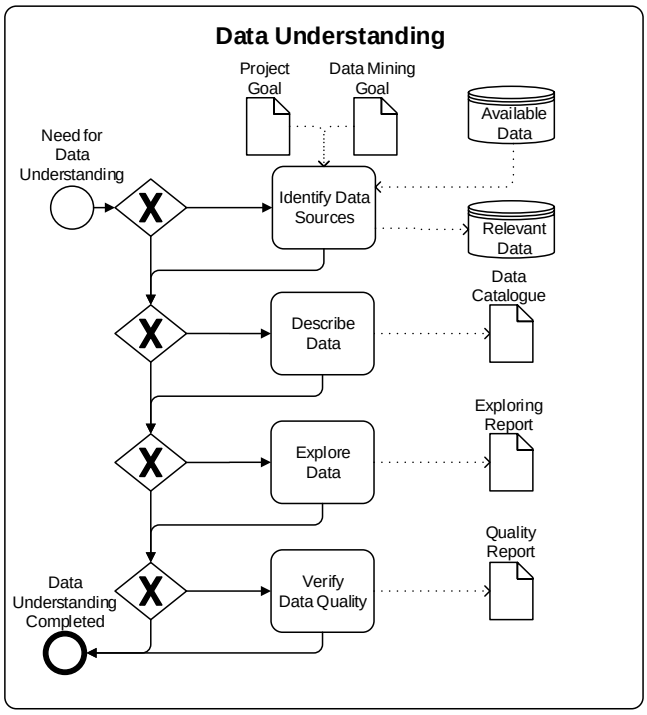

Figure 4. Phase Data Understanding

enable eliminating them or we can predict failures to prevent unplanned maintenance. Additionally, the used data and success metrics are defined, serving a specification for the model produced in the Modeling phase.

We specify the task Perform Project Management, in which a project plan and a project charter are created and activities such as staffing, risk assessment, project planning or calculation are performed [6].

3.3. Data Understanding

In the Data Understanding phase a big picture of available and relevant data sources in the manufacturing environment is created. To do so, four tasks are performed: Identify Data Sources, Describe Data, Explore Data, Verify Data Quality. These tasks, depicted in Figure 4, are directly transferred from CRISP-DM and specified in their content.

To increase availability with KDD, two types of data are necessary and to be identified in task Identify Data Sources:

1) historical data about machine failures, which are often recorded in MES (Manufacturing Execution Systems) and
2) data that indicate these failures or describe the condition of machines, such as sensor data.

With this data collected, it is possible to either find the root cause of failures or predict failures in the future resulting in less failures and unplanned maintenance and thus increasing availability.

The task Describe Data serves a concise overview and first understanding of the identified data sources and collected data. Especially describing sensor data with measurement units and value ranges create a first understanding.

In exploring the data in task Explore Data with visualization, clustering, correlation and other methods, the data structure and strong relations between machine failures and indicator variables are identified.

The quality of the discovered knowledge is strongly related to data quality and thus a high data quality is important. To assess data quality in task Verify Data Quality, we use the four criteria as proposed by [25], i.e., accuracy, timeliness, consistency, and completeness. Emphasis has to be put on completeness due to the occurrence of sensor data and WSN (Wireless Sensor Networks) in manufacturing and thus issues in data communication and dropped readings [26].

3.4. Data Preparation

Data Preparation is usually the most time intensive phase in KDD. Particularly, in machine-related projects, where sensor data and thus time series data are frequently used, it is highly challenging to prepare data for modeling. To facilitate data preparation, we define a certain format the

prepared data has to meet, which is depicted in Table 1. In order to bring the data into this target format, four tasks have to be conducted which we adopt from CRISP-DM and present in Figure 5: Select Data, Clean Data, Construct Features and Create Feature Set.

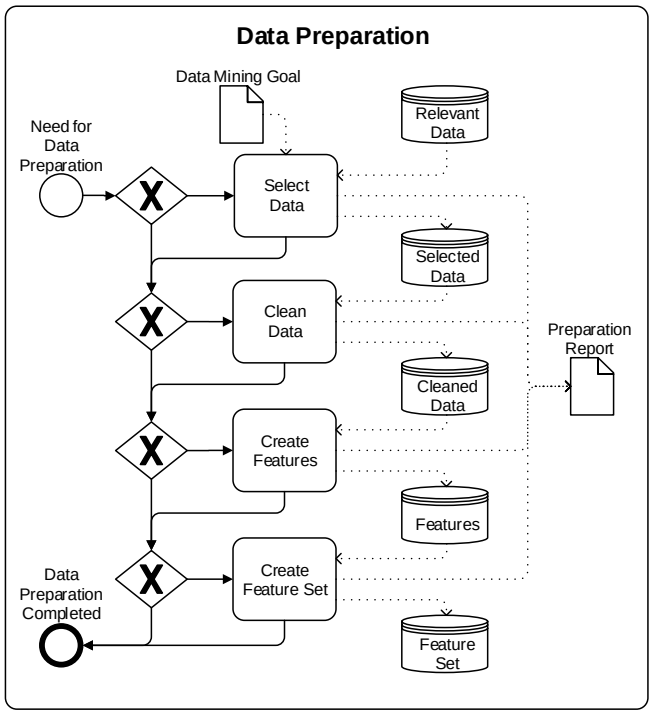

Figure 5. Phase Data Preparation

Selecting the relevant variables of the data sources in task Select Data is crucial to the quality of the model produced in the Modeling phase. As stated in Section 3.3, data about machine failures are as important as data, that describe the condition of machines and indicate those failures. To select the best indicators and machine condition variables, several techniques [27], [28], e.g., Pearson, Kendall and Spearman Correlation, Fisher Score, and Chi-Square-Test, can be applied in different situations.

In task Clean Data, the identified data quality issues are handled. Challenging are missing sensor values and the occurrence of sensor noise. To eliminate noise and reconstruct missing values, regression models and probabilistic models as well as outlier detection methods are applied [26].

To build a high quality model for predicting failures, so-called features are needed. Those features are derived from existing variables such as sensor values, machine information and component replacement records. Such a derived feature is the age of a component calculated based on its last replacement. The goal is to produce the best possible representation of a machine and its components at particular points in time. This allows an algorithm to better identify emerging patterns in the data. Tumbling Window, Sliding Window and Hopping Window are methods for data stream management [29] and can create features for predicting machine failures.

Table 1. Target Format for the Feature Set

Time Stamp	Input Variables				Target Variable
$\mathbf{t}$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_n$	$\mathbf{y}$
t_1	$x_{1,1}$	$x_{2,1}$	...	$x_{n,1}$	y_1
t_2	$x_{1,2}$	$x_{2,2}$	...	$x_{n,2}$	y_2
...	...	...	...	...	...
t_m	$x_{1,m}$	$x_{2,m}$	...	$x_{n,m}$	y_m

For most modeling techniques, one consistent table called feature set is used for training (see Table 1). All selected variables and constructed features are merged considering time stamps and other dependencies. The construction of a target variable, which can contain information whether the machine had a failure at a particular point in time, is done separately. With a predictive model we then try to predict this target variable for a new set of features.

3.5. Modeling

Modeling consists of four tasks depicted in Figure 6. Selecting the suitable modeling technique in task Select Modeling Technique and choosing appropriate metrics for evaluating the model's quality in task Assess Model can be specified in our context.

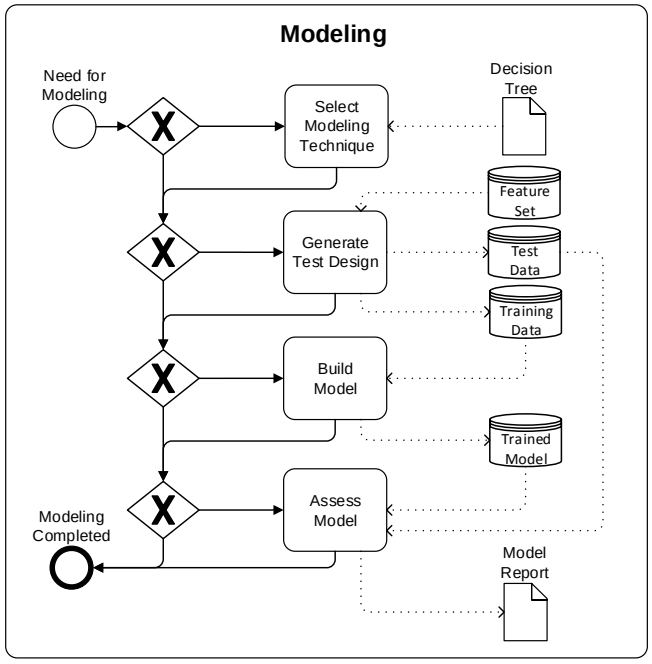

Figure 6. Phase Modeling

There are mainly four modeling techniques that can be used to either find the root cause of machine failures or predicting them. We studied various data mining packages, tools and functions compiled in [30] and data mining techniques presented in [31], [32], [33] in order to define a process of how to select the best technique in our specific context. Figure 7 shows the resulting decision tree, that helps selecting the modeling technique suitable with a defined data mining goal related to availability improvement.

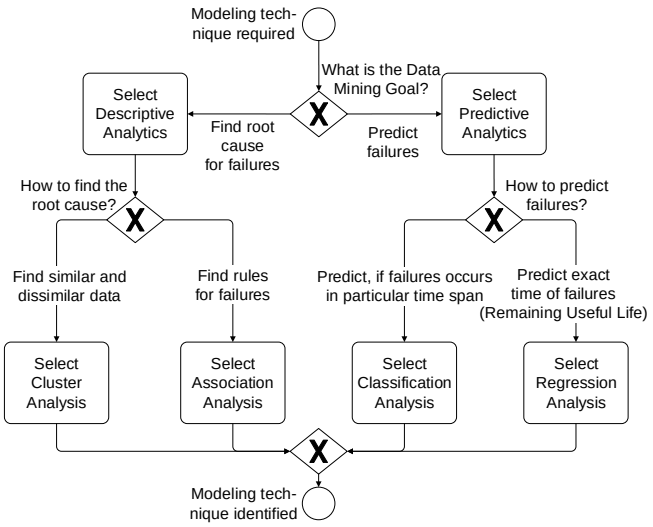

Figure 7. Decision Tree for Selecting a Modeling Technique

Descriptive analytics including clustering and association analysis are techniques primarily used to find the root cause of machine failures. Predictive analytics including classification and regression analysis are techniques used to predict machine failures.

Generating a test design in task Generate Test Design is the activity of splitting the feature set into training and test data. To build a model in task Build Model, the training data is used.

To assess the model in task Assess Model, it is applied to the test data to verify if its outcomes, e.g., predictions, are correct. Different metrics can be applied to assess model quality. It is important to pick the right metrics dependent on the problem to be solved. When predicting a machine failure with predictive analytics techniques, a wrong prediction of a machine failure is producing costs resulting from unneeded maintenance or needed maintenance not executed. On the other hand, a correct prediction can save costs resulting from timely maintenance – a failure that would've been occurred does not occur. To assess a model predicting machine failures with classification analysis, the metric recall in combination with precision should be used. In regression analysis, the weighted sum of prediction error is to be minimized, in order to get a good quality model [34].

3.6. Evaluation

In agile project management methods, the Evaluation phase can be compared with the retrospective, in traditional project management with the lessons learnt [35], [36]. In four tasks, depicted in Figure 8, results are evaluated, the project is reviewed, next steps are determined and a project report is produced.

We do not specify the tasks in this phase, since they remain generic in every domain. However, we transfer the task Produce Project Report from the Deployment phase to the Evaluation phase, since evaluating the project and writing down the findings in parallel saves time to create the project report. Apart from that, deployment is often conducted separately and thus all the results should be well documented.

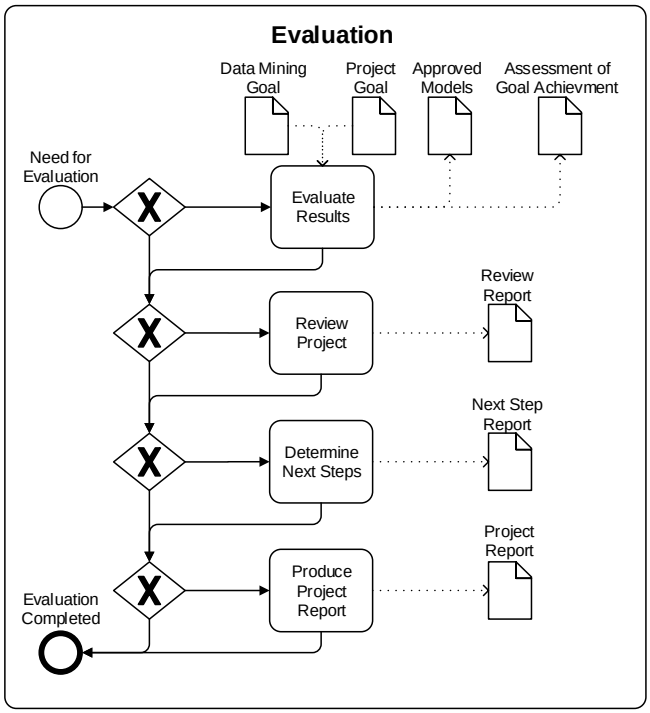

Figure 8. Phase Evaluation

3.7. Deployment

In CRISP-DM, the deployment of a model is only planned and not conducted. Hence, the results and outputs gained in the KDD process consisting of built models and findings are not operationalized. In order to ensure applicability, we concretize activities to be conducted in different steps of deployment particularly in respect of availability improvement. The four specified tasks of the Modeling phase are presented in Figure 9.

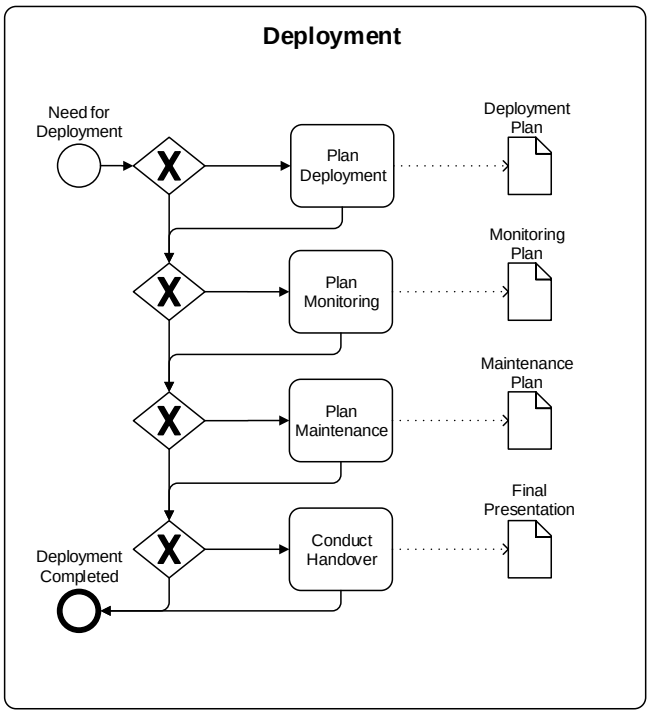

Figure 9. Phase Deployment

In the task Plan Deployment we define how the approved models and necessary production data are provisioned and how the results are presented, e.g., visualized.

The model's quality has to be monitored since production data change over time. For this, we define the task Plan Monitoring. As a result of data changes, the model creates bad predictions and needs to be maintained, i.e., retrained with new data. For this, we define the task Plan Maintenance. There are three options to trigger the maintenance of a model:

1) occurring events in manufacturing of which we assume that patterns in data have changed, e.g., changed machine settings;
2) periodical retraining, e.g., every week;
3) deviation from a defined threshold of model quality metrics, e.g., recall is 5 % under threshold.

The considerations and decisions made in the tasks Plan Deployment, Plan Monitoring and Plan Maintenance have to be documented in the corresponding reports shown in Figure 9. To assure, that built models, data, findings and documents are available and handed over to the deployment team properly, a final project meeting with a presentation is performed in the new task Conduct Handover at the end of a project.

4. Validation

To assess the applicability of the proposed process model, we applied it to a KDD project in the manufacturing domain. The project aimed at generating knowledge to improve availability. More specifically, the project aimed at building a model for predicting failures of a carding machine to reduce unplanned maintenance. We applied the aforementioned six process models as follows.

Business Understanding.
The manufacturing company produces nonwovens which are used for indoor filters and other products. These materials are produced in a continuous production process on a plant with several machines and process steps. One of these process steps is carding in which the loose fibers are first aligned to form a nonwoven. This machine consists of 20 engines powering different rollers and belts. A schema of the carding machine is shown in Figure 10.

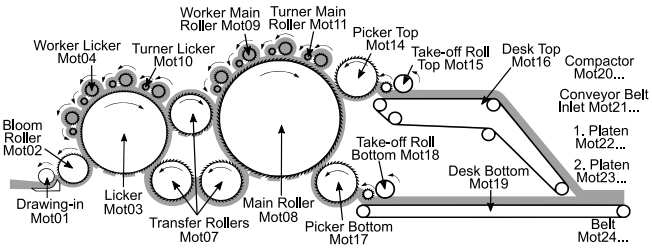

Figure 10. Schema Carding Machine

For each of the 20 motors, three sensor values are measured:

- motor speed V in revolutions per minute rpm, measured every second;
- power consumption I in ampere A, measured every second;
- temperature T in degree Celsius °C, measured every minute.

The company's strategic objective is to improve the availability of the carding machine from under 90 % to 98 % in one year. Goal of the project was to support this strategic objective in predicting failures of the carding machine to reduce unplanned maintenance and thus improve availability.

Data Understanding.
Data source for the sensor data was a supervisory control and data acquisition system (SCADA). Elicitation of failure data is done via a MES. Table 2 shows an excerpt of the sensor values extracted. The Pen Number is referencing a certain sensor in the machine. Table 3 contains information about the sensors which we need to identify whether the

generated sensor values represent motor speed (V), power consumption (I) or temperature (T). Table 2 and Table 3 are linked to each other, e.g., the first record in Table 2 contains the measured speed of motor one, i.e., Mot01. In Table 4, data about occurred machine failures is shown containing a time stamp, the reason for failure, the affected equipment and the breakdown duration in minutes.

Table 2. Excerpt Sensor Values

Pen Number	Date	Value
1	14.01.2017 13:30:01	23,943863
1	14.01.2017 13:30:02	23,712383
⋮	⋮	⋮
2	14.01.2017 13:30:01	43,427383
2	14.01.2017 13:30:02	40,477303
⋮	⋮	⋮

Table 3. Excerpt Sensor Description

Pen Number	Pen Name	Minimum	Maximum
1	V_Mot01	0	188,58
2	I_Mot01	0	45,22
3	T_Mot01	0	36
⋮	⋮	⋮	⋮

Table 4. Excerpt Failure Data

Date	Reason	Equipment	Duration
12.01.2017 21:31	product change		51
13.01.2017 17:19	electrical disturbance	winder	120
14.01.2017 07:05	mechanical disturbance	carding	102
⋮	⋮	⋮	⋮

If we analyze the sensor data just before failures, power consumption (I) stands out. While motor speed (V) and temperature (T) do not change before a machine failure, power consumption values show anomalies a few minutes before failures occur as shown in Figure 11. This failure is the third record in Table 4.

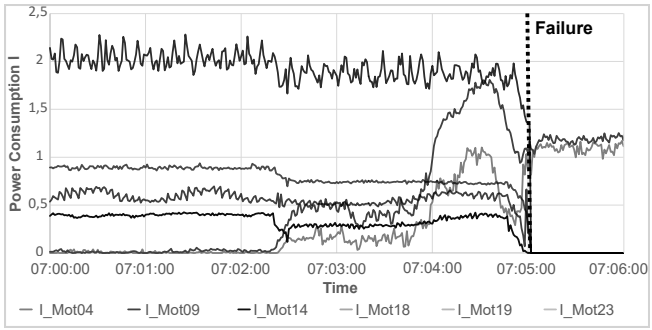

Figure 11. Power Consumption Anomalies before a Machine Failure

The quality of the sensor data extracted from the SCADA was high, except from Motor 20 sensors, which produced no data. However, the time stamps of machine failures extracted from the MES are not correct. Operators type them in manually which results in incorrect time stamps to be adjusted manually.

Data Preparation.

As motor speed and temperature values do not indicate a machine failure, we discard them and select power consumption as indicator variables. To create the feature set with respect to the target format, equidistant time stamps for every second are created. Then the values of the 19 power consumption sensors are matched to these time stamps in the feature set. The target variable is a boolean value indicating anomalies linked to potential failures. Table 5 is an excerpt of the constructed feature set. This data is used during the next phase to train the algorithms and assess their quality with regard to anomaly detection.

Table 5. Excerpt Feature Set

Date	I_Mot01	I_Mot02	...	Failure
14/01/2017 06:55:00	2,202	6,216	...	0
14/01/2017 06:55:01	2,202	5,120	...	0
⋮	⋮	⋮	⋮	⋮
14/01/2017 07:09:05	0	0	...	1

Modeling.

We used Microsoft Azure Machine Learning Studio for modeling. Since we want to detect anomalies before a failure occurs and thus predict machine failures in a particular time span, we built classification models, i.e., anomaly detection models. The tool provides two different algorithms to train such models, i.e., One-Class Support Vector Machine (SVM) and Principal Component Analysis (PCA) Anomaly Detection. We trained both, compared them and chose the one with higher quality. Since we cannot measure recall or precision of these particular algorithms in the tool, the model which detects a failure earlier and more distinctly is of better quality. The predictions, i.e., anomaly score as metric for failure probability, of both models are depicted in Figure 12. The higher the anomaly score, the more likely an anomaly is detected.

Comparing the output of the models, they both detect the failure which occurred at 07:05:00 three minutes in advance around 07:02:00 at the same time. However, SVM is more distinctly. In standard machine operation it calculates an anomaly score of just above 0, three minutes before the failure occurs it scores over 0.8. PCA scores 0.5 even in standard machine operation, which is not acceptable for failure prediction. Hence, we choose SVM.

Evaluation.

The outcome of the project is reviewed with all involved parties to assess, whether the project goal is reached, i.e.,

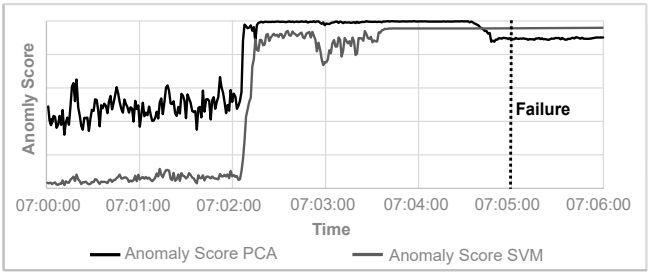

Figure 12. Model Comparison for Prediction of Failure

the availability can be improved with implementing the predictive model into production. We came to the conclusion that, if a failure can be detected three minutes before it occurs, the consequences, i.e., breakdown time and unplanned maintenance efforts, can be reduced. In this case, the jamming of the carding machine can be avoided. As a result, the effort for cleaning can be reduced. Hence, the availability can be improved and the project goal is reached.

A project report is created summarizing the six phases. Next steps are determined as follows. Validation of the model on other machines data; deployment of the model with live data streaming from the machine; collection of more failure and sensor data in order to further improve models.

Deployment.
For planning and testing deployment of the model, a web service is created in the cloud where the model is stored. Via an Application Programming Interface (API) and a R-script, the model can be fed with sets of sensor data and returns anomaly scores accordingly. The model can be retrained periodically with data from standard machine operation to learn new patterns and remain of high quality. The project is handed over with a final presentation.

5. Related Work

KDD process models have been defined and specified for certain domains. [22] proposes a specified process model based on a deductive and an inductive approach. However, it focuses on text mining in contrast to our model considering the manufacturing domain. The methodology in [8] is comparable to our work. It maps CRISP-DM to define a specific process model for the medical domain. However, such a mapping does not exist for the manufacturing domain. [37] proposes a data mining approach for analyzing semiconductor data to enhance a KPI called overall usage effectiveness (OUE). In contrast to our approach, enhancing OUE is more specific and can only be used for wafer fabrication, whereas OEE is accepted cross-industrial. Moreover, it aims at yield enhancing instead of availability improvement and thus contributes to performance improvement. [38] discusses the nature and implications of data mining techniques in manufacturing. Though, it only focuses on the content of the subject and does not provide the process of KDD. [39] postulates the new approach Knowledge Discovery and Analysis in Manufacturing (KDAM). It sensitizes the topic, reviews methods for KDD in manufacturing and discusses related applications, but comparable to [38], is not focusing on the process of KDD. [40] reviews data mining in manufacturing on the kind of knowledge and proposes a 10 step process. However, it is still very generic and a rough guideline without a detailed layer describing specific tasks. A similar approach is given in [41], which lacks detailed descriptions.

6. Conclusion and Outlook

Low availability constitutes a big part of manufacturing costs. KDD can be used to generate knowledge for improving availability. For this purpose, we presented a specific KDD process model. We used CRISP-DM as reference, studied specific models from other domains and considered various conducted projects. The process model contains six phases with specific, efficient and reusable tasks traversed in several iterations throughout a project. To assess the efficiency of the process model, we applied it to a KDD project aiming at improving availability by predicting failures of a carding machine. This objective has been achieved and since the process model provides a reusable approach, the required efforts conducting the project have been reduced.

Based on the results of this paper, we intend to validate the defined process model in further KDD projects and complement specific models for quality and performance improvement in order to provide a holistic KDD-based approach for improving OEE.

7. References

[1] C. May and A. Koch, "Overall Equipment Effectiveness (OEE): Werkzeug zur Produktivitätssteigerung," *Zeitschrift der Unternehmensberatung (ZUb)*, no. 6, pp. 245–250, 2008.

[2] S. J. Blöchl, C. T. Stemplinger, and H. Winkler, "Overall Equipment Effectiveness gezielt verbessern," *ZWF Zeitschrift fuer Wirtschaftlichen Fabrikbetrieb*, vol. 109, no. 11, pp. 830–834, 2014.

[3] P. Lade, R. Ghosh, and S. Srinivasan, "Manufacturing Analytics and Industrial Internet of Things," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 74–79, May 2017.

[4] M. P. Gallaher, Z. T. Oliver, K. T. Rieth, and A. C. OConnor, "Economic Analysis of Technology Infrastructure Needs for Advanced Manufacturing Smart Manufacturing," NIST, Tech. Rep., 2016.

[5] KDnuggets, "Crisp-dm, still the top methodology for analytics, data mining, or data science projects," 2014. [Online]. Available: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0: Step-by-step data mining guide," 2000. [Online]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf

[7] U. Fayyad, G. Piatsky-Shapiro, and P. Smyth, Eds., *Advances in Knowledge Discovery and Data Mining*, 5th ed. Menlo Park: AAAI Press, 1996.

[8] O. Niaksu, "CRISP Data Mining Methodology Extension for Medical Domain," *Baltic Journal of Modern Computing*, vol. 3, no. 2, pp. 92–109, 2015.

[9] P. (PwC), "Industry 4.0: Building the digital enterprise." [Online]. Available: https://www.pwc.com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016.pdf

[10] S. Nakajima, *TPM tenkai*. Japan Institute of Plant Maintenance (JIPM), 1982.

[11] D. H. Stamatis, *The OEE Primer: Understanding Overall Equipment Effectiveness, Reliability, and Maintainability*. New York: Productivity Press, 2010.

[12] K.-S. Wang, Z. Li, J. Braaten, and Q. Yu, "Interpretation and compensation of backlash error data in machine centers for intelligent predictive maintenance using ANNs," *Advances in Manufacturing*, vol. 3, no. 2, pp. 97–104, 2015.

[13] J. Lee, E. Lapira, S. Yang, and A. Kao, "Predictive Manufacturing System - Trends of Next-Generation Production Systems," *IFAC Proceedings Volumes*, vol. 46, no. 7, pp. 150–156, 2013.

[14] C. Park, D. Moon, N. Do, and S. M. Bae, "A predictive maintenance approach based on real-time internal parameter monitoring," *The International Journal of Advanced Manufacturing Technology*, vol. 85, no. 1-4, pp. 623–632, 2015.

[15] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications," *Mechanical Systems and Signal Processing*, vol. 42, no. 1-2, pp. 314–334, 2014.

[16] G. Figueira, M. Furlan, and B. Almada-Lobo, "Predictive production planning in an integrated pulp and paper mill," *IFAC Proceedings Volumes*, vol. 46, no. 9, pp. 371–376, 2013.

[17] J. Krumeich, S. Jacobi, D. Werth, and P. Loos, "Big Data Analytics for Predictive Manufacturing Control - A Case Study from Process Industry," in *BigData Congress 2014*, P. P. S. Chen and H. Jain, Eds. Los Alamitos, California and Washington and Tokyo: Conference Publishing Services, IEEE Computer Society, 2014, pp. 530–537.

[18] J. B. O. Meré, A. G. Marcos, F. A. Elías, and C. M. Fernández, "Advanced predictive quality control strategy involving different facilities," *The International Journal of Advanced Manufacturing Technology*, vol. 67, no. 5-8, pp. 1245–1256, July 2013.

[19] N. A. Rahman and I. Masood, "Methodology For Designing A Control Chart Pattern Recognizer In Monitoring Metal Stamping Operations," *ARPN Jounal of Engineering and Applied Sciences*, vol. 11, no. 10, pp. 6439–6441, 2016.

[20] W.-A. Yang and W. Zhou, "Autoregressive coefficient-invariant control chart pattern recognition in autocorrelated manufacturing processes using neural network ensemble," *Journal of Intelligent Manufacturing*, vol. 26, no. 6, pp. 1161–1180, 2015.

[21] Y. Bahei-El-Din, *Advanced Technologies for Sustainable Systems: Selected Contributions from the International Conference on Sustainable Vital Technologies in Engineering and Informatics, BUE ACE1 2016, 7-9 November 2016, Cairo, Egypt*. Cham: Springer International Publishing, 2016.

[22] A. Schieber and A. Hilbert, *Entwicklung eines generischen Vorgehensmodells für Text Mining*, ser. Dresdner Beiträge zur Wirtschaftsinformatik. Technische Universität Dresden, Fakultät Wirtschaftswissenschaften, 2014.

[23] T. Allweyer, *BPMN 2.0: Introduction to the Standard for Business Process Modeling*. BoD–Books on Demand, 2016.

[24] P. Harmon and C. Wolf, "State of Business Process Management 2016," BPTrends, Tech. Rep., 2016.

[25] B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics*, vol. 154, pp. 72–80, 2014.

[26] C. C. Aggarwal, *Managing and Mining Sensor Data*. Boston: Springer, 2013.

[27] P. P. Eckstein, *Statistik für Wirtschaftswissenschaftler: Eine realdatenbasierte Einführung mit SPSS*, 4th ed. Wiesbaden: Springer Fachmedien Wiesbaden, 2014.

[28] R. Barga, V. Fontama, and W. H. Tok, *Predictive Analytics with Microsoft Azure Machine Learning*, 2nd ed. Berkeley CA: Apress, 2015.

[29] M. Garofalakis, J. Gehrke, and R. Rastogi, Eds., *Data Stream Management: Processing High-Speed Data Streams*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[30] KDnuggets, "50+ Data Science and Machine Learning Cheat Sheets," 2015. [Online]. Available: http://www.kdnuggets.com/2015/07/good-data-science-machine-learning-cheat-sheets.html

[31] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Amsterdam: Elsevier/Morgan Kaufmann, 2012.

[32] O. Maimon, L. Rokach, and T. Kohonen, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Boston: Springer Science+Business Media LLC, 2010.

[33] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge: MIT Press, 2001.

[34] S. Malinowski, B. Chebel-Morello, and N. Zerhouni, "Remaining Useful Life estimation based on discriminating shapelet extraction," *Reliability Engineering & System Safety*, vol. 142, pp. 279–288, 2015.

[35] D. Maximini, *Scrum - Einführung in der Unternehmenspraxis: Von starren Strukturen zu agilen Kulturen*. Berlin and Heidelberg: Springer, 2013.

[36] H. Meyer and H.-J. Reher, *Projektmanagement: Von der Definition über die Projektplanung zum erfolgreichen Abschluss*. Wiesbaden: Springer Gabler, 2016.

[37] C.-F. Chien, A. C. Diaz, and Y.-B. Lan, "A data mining approach for analyzing semiconductor MES and FDC data to enhance overall usage effectiveness (OUE)," *International Journal of Computational Intelligence Systems*, vol. 7, no. sup2, pp. 52–65, 2014.

[38] K. Wang, "Applying Data Mining to Manufacturing: The Nature and Implications," *Journal of Intelligent Manufacturing*, vol. 18, no. 4, pp. 487–495, 2007.

[39] M. Polczynski and A. Kochanski, "Knowledge Discovery and Analysis in Manufacturing: A Next Generation Quality and Reliability Paradigm," *Quality Engineering*, vol. 22, no. 3, pp. 169–181, 2010.

[40] A. K. Choudhary, J. A. Harding, and M. K. Tiwari, "Data mining in manufacturing: A review based on the kind of knowledge," *Journal of Intelligent Manufacturing*, vol. 20, no. 5, pp. 501–521, 2008.

[41] W. Wojcik and K. Gromaszek, "Data Mining Industrial Applications," in *Knowledge-oriented applications in data mining*, K. Funatsu and K. Hasegawa, Eds. Rijeka: In-Tech, 2011.