Laura Bebek, BSc

Optimizing the Overall Equipment Efficiency in Thermal Processes Through Data Analysis

Master's Thesis

to achieve the university degree of

Master of Science

Master's degree programme: Softwaredevelopment and Management

submitted to

Graz University of Technology

Supervisor

Dipl-Ing.Dr.techn. Roman Kern

Institute for Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Graz, March 2019

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

Thermal processes in the manufacturing industry involve highly optimized equipment for production. In order to run the process the equipment has to be maintained, replaced and adjusted in their settings regularly. This requires a certain amount of effort, concerning the economic and timely aspects.

The goal of this thesis is to purpose an approach for further improvement of the equipment efficiency, based on data-driven methods. Initially historic product and process data had been collected, mapped and pre-processed. In order to train selected machine learning algorithms features had been engineered and extracted. To ensure the state of the equipment can be represented through the available data, several models have been trained and evaluated. The presented heuristic approach dealt with the quality of the collected data and included a predictive maintenance model. This model was further analyzed to identify the influencing parameters on the lifespan of the equipment. Besides the prediction of maintenance actions, a proposal to optimize the utilization of the equipment is presented.

The data-driven methods applied in this thesis revealed the potential for future improvements in processes and according parameters.

Zusammenfassung

Thermische Prozesse in der Fertigungsindustrie erfordern hochoptimierte Anlagen für die Produktion. Um einen möglichst reibungslosen Ablauf der Prozesse zu ermöglichen, muss das Equipment regelmäßig gewartet, getauscht und eingestellt werden. Die wirtschaftliche und zeitliche Komponente des Aufwands sind signifikant.

Ziel dieser Arbeit war es, einen Ansatz zur weiteren Verbesserung der Maschineneffizienz auf der Grundlage datengetriebener Methoden zu entwickeln. Initial wurden historische Produkt- und Prozessdaten beschafft, kombiniert und vorverarbeitet. Um die ausgewählten Algorithmen für maschinelles Lernen zu trainieren, wurden beschreibende Attribute, sogenannte Features entwickelt und extrahiert. Um sicherzustellen, dass der Zustand des Equipments anhand der verfügbaren Daten dargestellt werden kann, wurden mehrere Modelle trainiert und deren Performance evaluiert. Der vorgestellte heuristische Ansatz befasste sich mit der Qualität der gesammelten Daten und beinhaltete ein Modell für die vorhersagende Wartung. Die Parameter mit dem meisten Einfluss auf das trainierte Modell wurden identifiziert, um die Einflüsse der Einstellungen auf die Lebensdauer des Equipments zu ermitteln. Neben der Vorhersage von Wartungsmaßnahmen wurde ein Vorschlag zur Optimierung der Einstellungen des Equipments entwickelt.

Die datengetriebenen Methoden, die in dieser Arbeit angewandt wurden, zeigten ein Potential für zukünftige Verbesserungen der Prozesse und den dazugehörigen Parameter auf.

Acknowledgment

First of all I want to thank Markus Puff, Günther Herold and his team from TDK electronics (formerly Epcos) for giving me the opportunity to accomplish the practical part of this thesis and for providing the data set. Further I want to thank the team of thermal process experts from TDK which supported me with their know-how and made the work very enjoyable.

A special thanks goes to my supervisor Roman Kern, for his constant support during the whole thesis with his expertise and for the patient guidance.

Further i want to thank my family and friends for being there for me whenever i needed them. At last but not least i want to thank my ”fellow students”, which became more like friends after our long journey, for mental encouragement and keeping my motivation high. (”Du muasst es hoid schon a bissl wolln a”).

Contents

1. Introduction 1

2. Related Work 3
2.1. Background 3
2.1.1. Industry 4.0 3
2.1.2. Big Data 4
2.1.3. Machine Learning 6
2.1.4. Logistic Regression 7
2.1.5. Support Vector Machine 9
2.1.6. Random Forest 11
2.1.7. Thermal Process 13
2.2. State of the Art 13

3. Requirements towards a Data Analysis Process 21
3.1. Data Lake 21
3.1.1. Structure and Requirements 24
3.1.2. Implementation 26
3.2. Data Analysis Models 26
3.2.1. KDD Process Model 27
3.2.2. Crisp-DM Process Model 29
3.2.3. Comparison of Selected Models 32

4. Overview and Evaluation of Data Analysis Tools 35
4.1. Selection of the Tools 35
4.2. Evaluation 36

5. Data Set and Process 45
5.1. Production Process 45
5.2. Data Set 47

5.3. Data Quality 50

6. Data Pre-Processing and Feature Extraction 55
6.1. Pre-Process 55
6.2. Sensor Data 56
6.3. Maintenance Records 58
6.4. Product Data 61
6.5. Extracted Features 61
6.6. Feature Selection 66

7. Method 67
7.1. Selection of Method 67
7.1.1. Setting Target Value 68
7.1.2. Training Method 70
7.1.3. Evaluation Method 71
7.2. Approach 74
7.2.1. Predict Missing Data 74
7.2.2. Outlier Detection 75
7.2.3. Predictive Maintenance 79
7.2.4. Parameter Importance 80

8. Evaluation 83
8.1. Results 83
8.2. Discussion 89

9. Conclusion 93

A. Features 99
A.1. List of All Extracted Features 99
A.2. Graphic Representation of Feature: Integral of actuating variable 104
A.3. Graphic Representation Extracted Maintenance Actions 105
A.4. Feature Importance according to MDA 115
A.5. Temperature Integrals and Ceiling - Timeline 116

B. Additional Graphics 127
B.1. Example of Data Quality 127
B.2. Example of Predicted Missing Data 129
B.3. Product Quality Measurements Compared in the Kilns 131

B.4. Used Software and Packages 135

Bibliography **137**

List of Figures

2.1. The Vs of Big Data 5
2.2. Overview of Machine Learning methods 7
2.3. Logistic Regression 9
2.4. Support Vector Machine 10
2.5. Random Forest Simplification 12
2.6. Usescases for Big Data Analytics in Manufacturing 15

3.1. Data Lake 23
3.2. KDD process model 27
3.3. Crisp-DM process model 30

5.1. Production Process 46
5.2. Rotary kiln for the sintering process 47
5.3. Sensor values from one run in the sintering kiln 49

6.1. Procedure of pre-processing and feature extraction 56
6.2. Temperature curve of one run in the kiln 59
6.3. Representation of Trapezoidal Rule 62
6.4. Example of extracted features 64
6.5. Example of extracted set top temperature 65

7.1. Ceiling within the kiln with the top and inside heating zones 69
7.2. k-fold Cross Validation 70
7.3. Basic Confusion Matrix 72
7.4. Heuristic Approach 74
7.5. Example of predict missing data 76
7.6. Graphical presentation of the calculated mad values 77
7.7. Semantic context of the detected outliers 78
7.8. Semantic context of the detected outliers 79
7.9. Runs until ceiling exchange versus duration of the ceiling . . 81

8.1. Confusion matrix of trained SVM 84
8.2. Confusion matrix of trained LG 85
8.3. Confusion matrix of trained RF 85
8.4. Top and inside heating zones with according lifespan of the ceiling 88

Acronyms

API	Application Programming Interface
CRISP	Cross Industry Standard Process
CRM	Customer Relationship Management
DL	Data Lake
DM	Data Mining
ERP	Enterprise Resource Planning
GPL	General Public License
GUI	Graphical User Interface
IDE	Integrated Developer Environment
KDD	Knowledge Discovery in Databases
KNN	k-nearest Neighbor
KPI	Key Performance Indicator
LR	Logistic Regression
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MDA	Mean Deviation Accuracy
ML	Machine Learning
MSE	Mean Squared Error
OEE	Overall Equipment Efficiency
OIE	Overall Input Efficiency
PMO	Production Management and Optimization
RD	Research and Development
RDBMS	Relational Database Management System
RF	Random Forest
SVM	Support Vector Machine

1. Introduction

On the way to "Industry 4.0" machines, sensors, several devices and applications produce huge amounts of data on a daily basis. Those data-sets have to be stored, processed and analyzed to gain information out of it and make the manufacturing ‚smart '.

In the electronic manufacturing industry, especially those involving thermal processes, equipment efficiency is of significant interest. Optimizing the equipment is not only impacting the economic factor by decreasing the costs but also the environmental influence is gaining more importance. In thermal manufacturing the processes and the equipment are already highly optimized. Further optimization should be reached with a Big Data analysis. By analyzing historical process data new information and patterns for deeper insight should be extracted. The approach of this thesis is to conclude a proposal to optimize the utilization of the equipment in order to reach an improvement, which would not have been reached without data-driven methods. The term Overall Equipment Efficiency is a standard to measure the equipment productivity defined by Semiconductor Equipment and Materials International (SEMI)[1]. The measurement is calculated based on Availability, Performance and Quality. Whereas Availability describes the time the equipment is available to production per total time, in percentage. The performance is measured by operational efficiency and rate efficiency, which refers to the efficiency of the up-time of the machine and the units produced. Further the quality is measured by calculating the ratio of accepted produced units regarding quality per total produced units. The theoretical part should describe the procedure of a data analysis in general, give an overview of the basic terms and describe the set-up and

[1] http://ams.semi.org/ebusiness/standards/SEMIStandardDetail.aspx?ProductID=211&DownloadID=3473

requirements of a data analysis process. Further available platforms, open source as well as proprietary ones, for those analysis should be evaluated.

The second part, the practical one, consists of realizing the described process model from the first part. The data should be extracted from several platforms in the production system, processed, cleaned and analyzed. Based on the process data, models for several algorithms will be trained and compared. This should lead to an approach to optimize the Overall Equipment Efficiency (OEE) by extracting and using new information concerning the specific process. This part is done in collaboration with a manufacturer from the electronic industry. It requires a close and frequent interaction with the experts of the process to gain the specific information regarding the needed domain knowledge.

The central research issue of this thesis is defined as follows:

- Is it possible to represent the state of the equipment as it is based on historical process, product and maintenance data?
- Is the quality of the collected data sufficient?
- Does the state of the equipment have an impact on the energy consumption?
- How to discover hidden knowledge from historical production data to generate an optimizing strategy?
- Does the insight of the data lead to a proposal for optimizing the overall equipment efficiency?

As a first step the separated data sources are collected and linked to each other. As the data is not collected for an analysis purpose, the data quality has to be reviewed. The data-set is pre-processed and provides the basis for several machine learning algorithms to be trained. After comparing and evaluating the methods one of them is chosen for optimization and should lead to new insights concerning the process and the equipment.

2. Related Work

2.1. Background

In this chapter some fundamental definitions and explanations are given for an enhanced understanding of the data analysis in this thesis. Additionally the sate of the art section gives an overview of the current research of the following three sections: areas of application of data analysis in manufacturing in general, improving manufacturing through data analysis and in particular improving thermal process equipment.

2.1.1. Industry 4.0

The manufacturing sector is currently on its way towards Industry 4.0 as described by Schwab [40]. Where several devices, machines and platforms are connected in a physical and logical way. The structure would change from a hierarchical system with separated operating units to a interconnected collaborating production network. Within this system the machines and devices could communicate via sensors and event-based logs on their own and therefore be independent from human operators. This would lead to a smart factory, where machines and work pieces could be localized within the manufacturing and their conditions could be monitored. The processes are transparent and visualized in real time. Along the way to Industry 4.0 there are still a lot of challenges to overcome as defined by Kagermann, Wahlster and Helbig [16], such as standardization, legal restrictions, security aspects and as well the role of the human working within this environment.

2.1.2. Big Data

The term Big Data has been described with the initial 3 V's by Laney [20], where the V's stand for Volume, Velocity and Variety. Volume describes the huge amount of data involved and increasing exponentially. Velocity characterizes the high frequency the data is generated and has to be proceeded. The term Variety addresses the different types of data originating from various sources. That is why the data can further be separated into structured, semi-structured and unstructured. According to Wamba et al. [48] there are two additional attributes when it comes to the definition of Big Data which are Value and Veracity. The extent of Value generated from insights of the data and adds economic benefits to the business. Veracity described the trustworthiness of the sources and the quality of the data sets. Those described 5 V's are located in the center of the figure 2.1 related to further V's. Over time several V's had been defined in the context of Big Data, such as by Kirk [18].

Interacting with the core consisting of Volume, Variety and Velocity the terms coming up within the context of Big Data are as follows: Variability in terms of data describes how wide spread the data is. The Viability is the ability of the underlying system to process new data sets and changing data. To interact and explore the data or the new insights gained out of it, some kind of Visualization is needed. To understand the connections of the data towards applications, users or other data the Vitality has been discussed in the Big Data context. Viscosity is described by Shafer [41] as the difficulty to work with the data. Data Volatility should not have any impacts toward the stability of the system where data is processed. The Validity of the data and also for further consumption should be ensured. The Vocabulary concerning Big Data varies depending on the domain where it is used, therefore a common understanding should be created. The Venue refers to the system where the data analysis is performed, depending on the nature of the data. The Vagueness of the found insights depend on the interpretation of the meaning.

Dealing with Big Data faces a variety of challenges, the two biggest challenges has been defined by Katal, Wazid and Goudar [17]. First of all the design of a system that can store and process those huge amounts of data

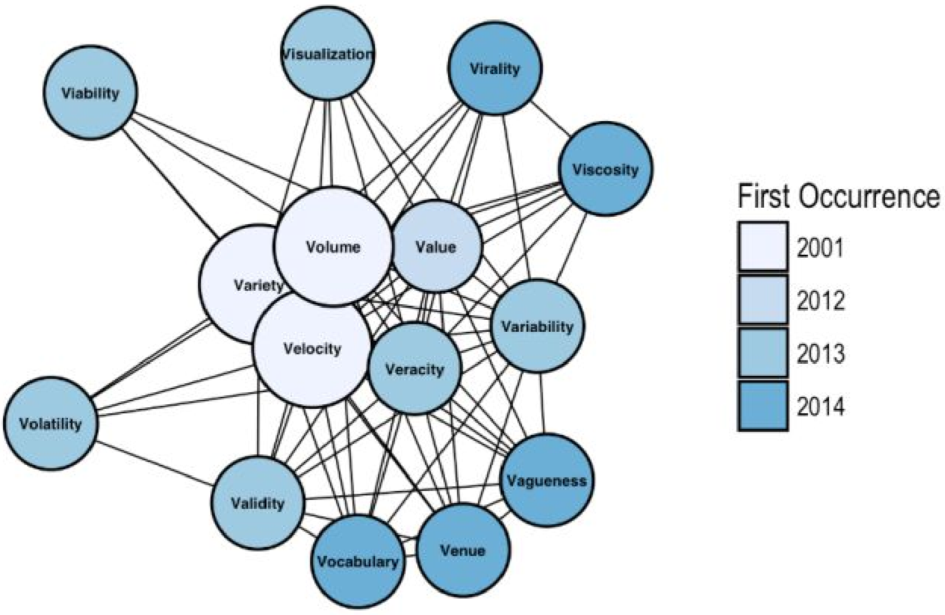


Figure 2.1.: The most relevant V's and its relations defining Big Data and Data Science according to [41]. Whereas the most dominant terms are Volume, Variety and Velocity. At the same time those three terms correspond to the initial explanation in 2011.

in an effective and efficient way. The second challenge is to select the most essential data out of the huge data collection and further generate additional value for the business through data analysis. Big Data in manufacturing has various sources in different formats, such as sensors within machines, ambient sensors, controllers in production system, log files, process requirements, etc.

2.1.3. Machine Learning

The application of Machine Learning (ML) methods can solve a given problem based on sample data and past experiences. In order to train a ML model, the step by step instructions to solve the problem are not known. Therefore the lack of knowledge is ‚learned ‘from the data, as described by Alpaydin [1]. ML algorithms can be trained to predict the output variable (Y) based on the input data (X) which represents a set of features. The basic ML techniques can be separated into supervised and unsupervised learning as shown in figure 2.2 [26]. The main difference between the two learning methods is that the data set for supervised learning is labeled with the output data (Y) and the unsupervised is not.

Therefore unsupervised learning is used to find unknown structures and relations in data sets. For example the data points can be clustered to explore categories. In supervised learning models can be trained with the input and output data in order to classify new unseen data. The goal is to train the algorithm, in order to find the function and according parameters to describe the given data Y = f(X). The goal is to predict the label for input data with the lowest possible error rate. In other words to find the model which approximates the process with high predictive accuracy. ML approaches should be able to adapt to changing environment by learning from the data. That is why ML is considered to be an application of artificial intelligence. In the practical part of this thesis some of the ML models are trained, therefore a selection of three supervised algorithms are chosen for comparison. The applied algorithms are explained in the following sections.

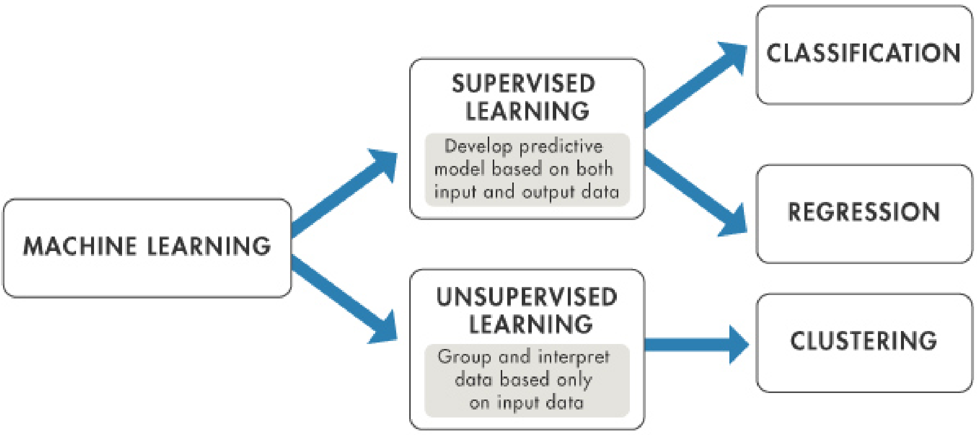

MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Figure 2.2.: Overview of the ML techniques according to [26] classified into supervised and unsupervised learning. The supervised learning consisting of Classification and Regression is trained with already labeled data to train the model. Whereas unsupervised learning can be used to cluster the data set into different groups.

2.1.4. Logistic Regression

The Logistic Regression (LR) approach is a supervised learning method, which is suited for binary classification problems. Basically the input data is weighted differently within a logistic function to predict the output. [1] [14] Each of the input values x has an coefficient value to weight the features, this coefficients are learned from the data. As shown in 2.1b the input data is combined in a linear way to represent the output data, where β corresponds to the coefficients of the according input value. The underlying logistic function, also known as Sigmoid function is shown in 2.1a. In the Sigmoid function the a is substituted by the linear expression of the input data. The logistic function limits the value to the range between 0 and 1. The hypothesis representing the model, which predicts the probability, that the given input data belongs to the default class is shown in 2.1c.

$$\sigma(a) = \frac{1}{1+e^{-(a)}} \tag{2.1a}$$

$$Z = \beta_0 + \beta_1 \cdot x \tag{2.1b}$$

$$p(x) = \sigma(Z) = \frac{1}{1-e^{-(\beta_0+\beta_1 \cdot x)}} \tag{2.1c}$$

$$\frac{p(x)}{1–p(x)} = e^{\beta_0+\beta_1 \cdot x} \tag{2.1d}$$

The prediction is based on the ratio of the probability that the input belongs to default class 1 divided by the probability that the input data does not belong to default class 1. The described ratio 2.1d is called odds. The graphical representation of the logistic regression is shown in figure 2.3 where the green line corresponds to the hypothesis representing the trained model. In this example the classes represented by the blue and red dots are strictly separable [24].

In order to optimize the model the optimal coefficients have to be chosen. Optimal coefficients are representing the model with minimum error. Therefore the cost function shown in 2.3 is minimized via gradient descent by finding the global minimum.

$$Cost(h\Theta(x), y) = \begin{cases} -log(h\Theta(x)) & \text{if } y = 0 \\ log(1-h\Theta(x)) & \text{if } y = 1 \end{cases} \tag{2.2}$$

$$J(\Theta) = -\frac{1}{n}\sum[y^{(i)}log(h\Theta(x(i))) + (1-y^{(i)})log(1-h\Theta(x(i)))] \tag{2.3}$$

There is a possibility that LR is not restricted to binary data. With the so called Multinomial LR more than 2 classes can be predicted.

$$P(Y_i = y|X_i) = \frac{e^{\beta_i \cdot X_i \cdot y}}{1+e^{\beta_i \cdot X_i}} \tag{2.4a}$$

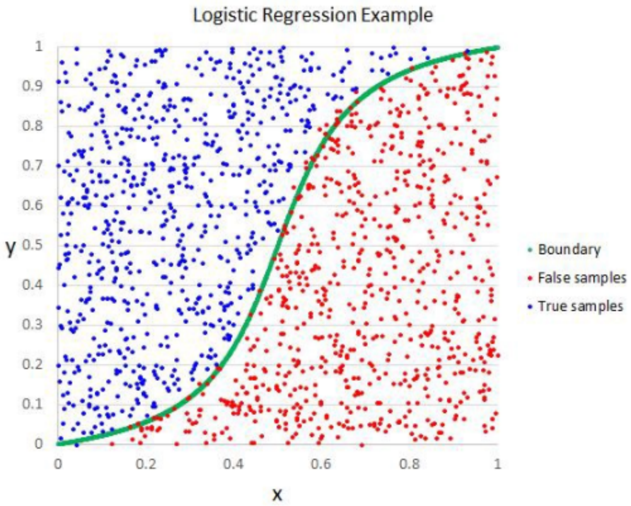


Figure 2.3.: Example of separating the input data points with the LR taken from [24]. The input data is classified in terms of the highest probability of output value. In this example the two classes, False and True Examples, are strictly separate-able by the decision boundary.

In this method the basic functionality is the same, except the probabilities for the classes are calculated and compared to all the other classes. The representation of the model still contains a linear combination of the input data. The probability that the input data X_i belongs to class y is calculated through 2.4a. This calculation has to be done for each of the classes. Further the data sample i has to be assigned to the class corresponding to the highest probability value.

2.1.5. Support Vector Machine

The Support Vector Machine (SVM) is an algorithm allocated to supervised learning techniques and can be used for Classification and Regression. The goal of the algorithm is to find a hyperplane which separates the data points in the most sufficient way [30] [1]. A sufficient way would be to chose the hyperplane, which fits best not only for the training data but also for the new unseen data. Therefore the goal is to find the hyperplane with the maximum

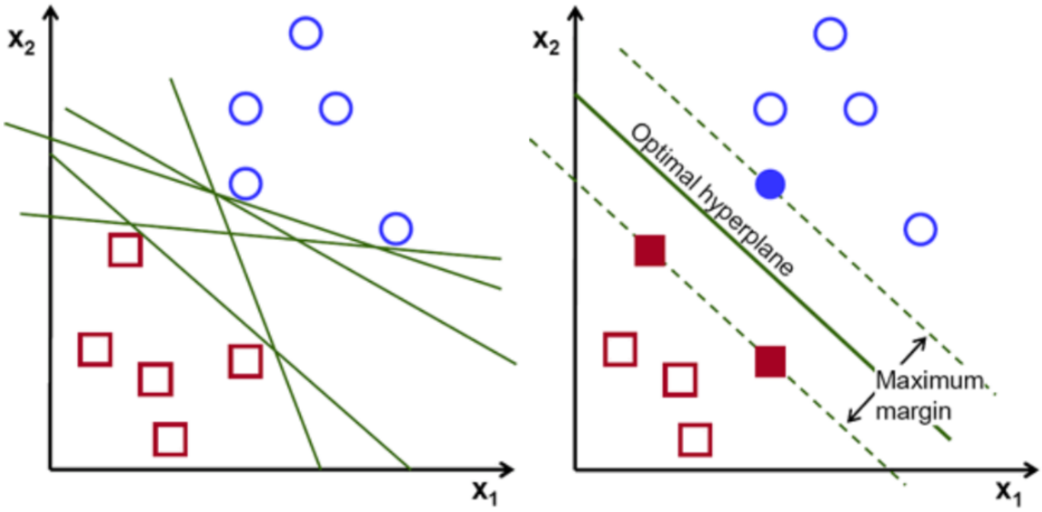


Figure 2.4.: Graphical representation of SVM algorithm [45]. There is a variety of hyperplanes to separate the data points. The optimal hyperplane is chosen by maximizing the distance to the nearest data points. This is called the separation margin.

distance to the closest data points. The distance between hyperplane and the nearest data points is called separation margin. A graphical explanation is shown in figure 2.4 [45], where the optimal hyperplane with the according maximum margin is presented with the green line. The margin is described through the support vectors, they describe the separating function.

The definition of the sample data is defined in 2.5a, where x_i refers to the data point and y_i to the according classification. To classify new data 2.5b is defined. The normal vector w represents the hyperplane and b the bias. In order to get the optimal hyperplane, one has to find the maximum separation margin to data point with the minimum distance. Therefore the square norm 2.5c has to be minimized by adjusting b and w, on condition defined in 2.5d.

Definition: The geometric margin of a hyperplane w with respect to a dataset D is the shortest distance from a training point x_i to the hyperplane defined by w.

$$\{(\mathbf{x}_i, y_i) | i = 1, \ldots, m; y_i \in \{-1, 1\}\} \tag{2.5a}$$
$$y_i = \operatorname{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \tag{2.5b}$$
$$\frac{1}{2}\|\mathbf{w}\|_2^2 \tag{2.5c}$$
$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \tag{2.5d}$$

In the shown formulas the classes are restricted to two $(-1, 1)$ for simplification. SVM model could be trained for more classes, therefore the same procedure as described is executed for each class. This is done by separating the actual class from all others, one-vs-all. Depending on the data set the classification could be done by a hyperplane in n-dimensional space. The function representing the hyperplane can be influenced by choosing different kernels. This could be a linear function, therefore the underlying classification problem has to be linear. For other kernels the above mentioned formulas have to be extended.

2.1.6. Random Forest

The Random Forest (RF) learning approach belongs to the supervised ML techniques. It has been designed by Breiman [6] and can be used for Classification as well as for Regression. The underlying functionality as explained by Peterek et al. [33] is a decision tree, where each sample is classified according to the given features. Initially the features are split up randomly. Each feature range is divided at a specific value based on the training data. A tree is trained by optimizing the split of each feature, this is reached through minimizing the squared error. The values should be split in a way, that the deviation from the average is minimized.

The word forest indicates that the algorithm uses more than one decision tree. Within this method several decision trees are trained. Each of them classify each sample, the final classification is the class with the most number of votes or in case of regression the averaged value. The idea of the algorithm

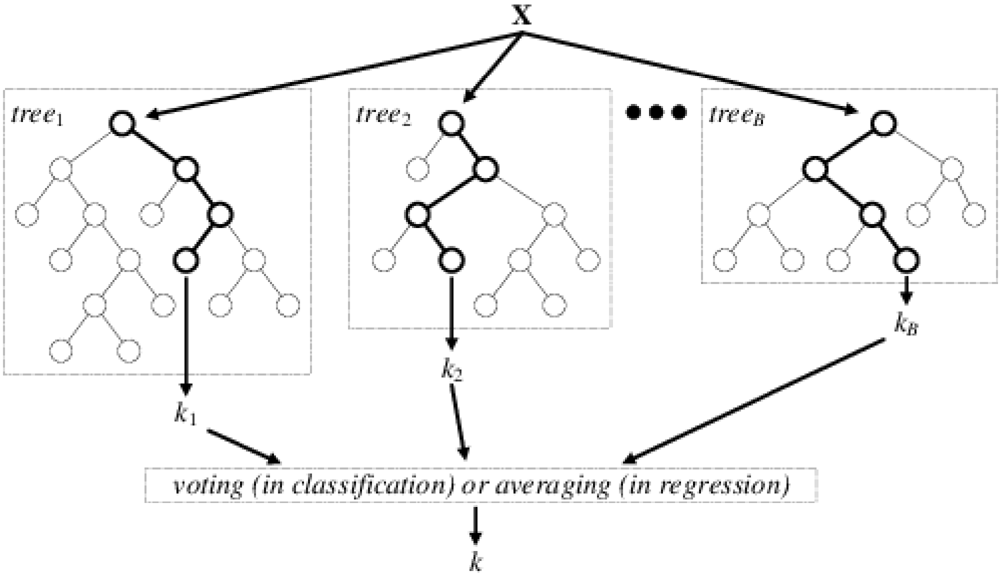


Figure 2.5.: The Figure shows the simplified functionality of the Random Forest method *Random Forest* [38]. The sample with the features is the input to the model where each of the trees in the forest classifies the sample. The final classification is done by counting the votes of each sample. As this method can used for regression as well, the classified values are averaged for the final result.

is based on the wisdom of the crowd, which applies the assumption that a crowd is wiser compared to an individual. Each tree forming the forest representing an individual could classify the sample independent from the other trees. By taking the aggregated result of all classifications of the forest the error is expected to be smaller.

A simplified functionality of RF is shown in figure 2.5 [38]. The trees are build independent, therefore each path is different, as the features are chosen randomly.

A main advantage within this method is the time the model is trained and evaluated. Because of the separated trees forming the forest the execution of the algorithm can be parallel. It is robust when it comes to outliers, due to the separated trees. Further the training data does not have to be normalized.

2.1.7. Thermal Process

In manufacturing steps sometimes it is required to change materials through chemical and physical reactions. The change in structure of the material can be reached through a temperature-controlled process. The thermal process consisting of highly complex mechanisms depending on multivariate process parameters, such as heating in different zones, gas flow, material load. Most commonly a thermal process part of a production process and takes place in a special furnace or kiln. The basis data set described in 5.2 for the data analysis process in this thesis is generated in a thermal process. The specific thermal process takes place in a rotary kiln , where parts for the electronic industry are produced. A more detailed description of the underlying thermal process will be described in 5.1.

2.2. State of the Art

As Auschitzky, Hammer and Rajagopaul [3] revealed how Big Data analysis can improve manufacturing by enhancing services, processes, maintenance and products. Isolated data sets are connected to obtain deeper knowledge of the processes. Several examples were described where the dependency of the data is investigated, the inter-linkages where pointed out and the complexity had been analyzed. The core techniques for gaining further understanding and optimization of the manufacturing processes. Improvements could be reached through the following four aspects: (1) exploratory data analysis: visualizing the data and finding initial patterns with basic statistic approaches, (2) using correlation analysis to understand the dependency within the data and forming hypothesis, followed by (3) significance testing to investigate the hypothesis and training artificial neural networks to determine the optimal parameters. Even in companies with already high optimized manufacturing procedures, data analysis could lead to further achievements.

Over the past decade had been different approaches to use the insights of Big Data analysis to improve manufacturing processes in the electronic industry. In the paper of Lv et al. [25] applications of Data Mining (DM) in the electronic industry had been reviewed. Over 75% of the articles applied Big Data analysis to production and control management, 17% dealt with process design and less than 8% analyzed topics regarding sale, service and recycling. For recognizing a pattern out of these areas the most used had been: prediction, classification and clustering functions. Whereas in Production Management and Optimization (PMO) most commonly a hybrid analysis consisting of statistical analysis and knowledge discovery analysis had been used.

Research regarding machine-learning applications in manufacturing had been reviewed by Pham and Afify [34]. The paper indicates that there is no universally applicable method. The methods are independent of the domain and can be very useful for discovering unknown pattern, when properly used. To choose a technique which is suitable few conditions have be met: a solid understanding of the requirements and of the problem. Further (a) the problem has to be of sufficient degree of complexity (b) and can be formulated as input for machine learning application. The training data has to be (c) in a adequate format and (d) include representative samples. The data set should have the ability to be (e) cleansed in an effective way. Methods for (f) learning and evaluation should be chosen under consideration of the application. Concrete achievements could be accomplished through several applications, Nagorny et al. [29] made a Review of use cases regarding Big Data analysis in smart manufacturing. An overview can be found in figure 2.6. They figured out several major applications within the production process:

- enables the identification of patterns and additional knowledge of influencing factors
- monitoring and observations of defined patterns, could help to detect anomalies or defects
- possibility to predict trends for optimizing the runtime, avoid failures etc.
- allows the diagnosis and causality-finding of failures
- increase the visibility of production systems
- could provide decision support
- enables optimization of Key Performance Indicator (KPI)

The wide variety of applications shows that there is a huge potential for

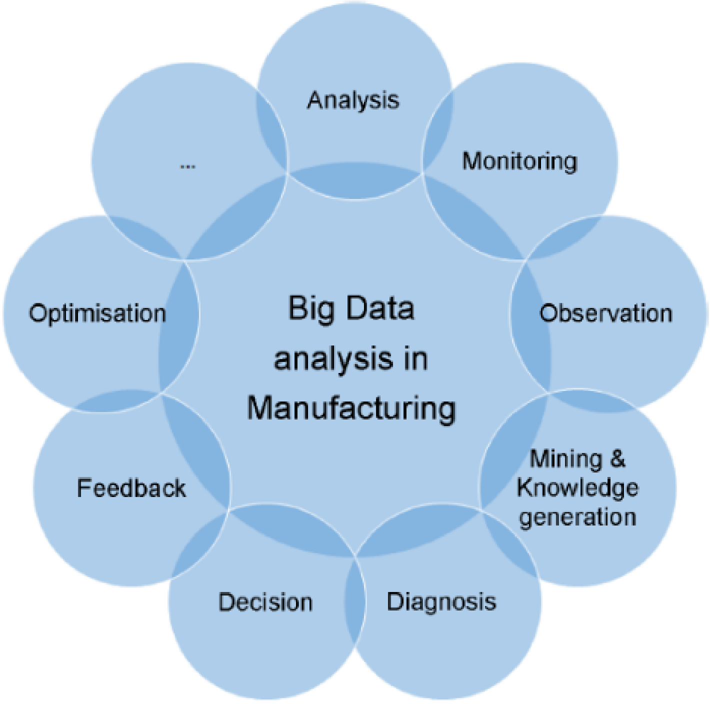


Figure 2.6.: The figure shows the uses cases for Big Data analytics in manufacturing according to Nagorny et al. [29]. There is a wide variety of applications within the production industry with enormous potential to exploit.

exploitation, especially in production processes.

On the way to efficient applications of those technologies are still a lot of challenges to overcome. One challenge is to build an sufficient platform to collect and process the necessary data for further analytics. Zhang et al. [51] designed a framework to represent the overall life-cycle of the product in the manufacturing processes. Based on this framework they analyzed data from sensors, mobile devices, machines, radio-frequency identification (RFID) tags and other devices by clustering, association, classification and prediction techniques. The proposed model including an early fault warning method for intelligent maintenance, energy efficiency and process analysis for improve production and optimizing the warehouse by zero inventory spare parts brought the company a competitive advantage. Based on those insights new equipment for reusing the heat and energy waste had been installed.

According to Sheu [43] the true OEE should also consider the input of the equipment, which is called Overall Input Efficiency (OIE). The input includes the manpower, spare parts, raw material etc. everything needed to operate the equipment. The researchers try optimize the total equipment efficiency by minimizing the required resources and keeping the produced goods at a constant level. This would cut the costs to operate the machine. Due to high temperatures needed within the thermal processes, they are one of the most energy-intense manufacturing sectors. Resulting from the high consumption of energy, it is considered as a significant cost and input factor in the production process. Barletta et al. [5] added the energy consumption to measure the OEE. The researchers revealed the potential energy losses to reach an optimization instead of focusing on the planned production time like the traditional OEE. One way to optimize the OEE is to reach an improvement the energy efficiency within the equipment. Zhang et al. [49] reviewed several approaches for optimization of the energy consumption and proposed a data analytic method for the overall production process, including maintenance and manufacturing process data. Those insights lead to immense improvements in maintenance, service and production which further decreased the material and energy consumption.

Among imaging the overall process there had been several researchers focusing on optimization of only one part of the production process. As thermal processes in manufacturing are one of the most energy-intense production steps there are a lot of researchers in this field focusing on energy efficiency. Therefore another way to optimize the OEE is to improve the energy efficiency. In order to increase the energy efficiency based on the data of a single machining process a model for predicting and simulation the energy consumption was presented by Liu, Xie and Liu [23]. Because of different energy pattern in the production, the energy consumption was split into periods and further classified into three different states: idle, start-up and cutting. For each of those states a model of the energy consumption was generated. Depending on the production process the energy consumption can be simulated by adding the energy prediction for the corresponding state. Further the models were evaluated and optimized, in order that the prediction is also suitable for different speed levels. Optimizing the resource usage within the generated models lead to energy savings. Another approach was presented by Zhao et al. [52]. The authors also differentiated between various states and further distinguished between fixed and variable energy, whereas fixed energy is needed to switch the machine on and keep it in a ready state and variable consumption is needed in the production state. Based on those data collections they trained a Neural Network predicting the energy needed in each step the machine needs to accomplish. With this network they simulated different settings within the machine and detected causal relationships. By predicting the energy consumption for various parameters, a deeper understanding about the processes and its products has been achieved. The prediction also supports the operators of the machine to choose the optimal parameters regarding energy. Further adaptation in the machining cycle, acceleration-deceleration approach, could cut the energy needed by shorten the standby energy consumption. Machines within the idle state imply a big potential for improvement in energy efficiency, according to Vikhorev, Greenough and Brown [47] they are accountable for 20-30% of the energy losses. In order to use this knowledge and improve the energy performance of several machines O'Driscoll, Kelly and O'Donnell [31] introduced the intelligent sensor. Therefore features had been extracted from energy data stream to describe and separate the individual loads. Further the operational state within the machine had been classified with the k-nearest Neighbor (KNN) algorithm into idle, running and cutting state. The improvement regarding energy efficiency could be achieved by understanding and adapting the inefficient states. Additionally an optimization of the overall energy consumption could be designed by adapting the time and acceleration ratio.

The study of Zhang et al. [50] analyzed the research intersecting energy consumption, manufacturing and Big Data in the context of energy-intense production. They found out that, each of them was highly researched independently, whereas the intersection of all three topics has mostly been analyzed in a theoretical way. A framework for energy Big Data had been developed to represent the energy consumption of the whole production process. The framework includes four sections: the first one is called energy Big Data perception and acquisition where the data from several sensors and devices is collected. Followed by energy data big storage and preprocessing which proposed a sequence to clean the data and transform it in a adequate format. The energy Big Data mining and energy-efficient decision-making layer including different data mining techniques such as regression, classification, clustering. The top level consist of application services regarding energy Big Data where the energy is monitored and different savings are proposed.

In detail there had been researchers focusing on the core equipment needed in thermal process. Zhou and Chai [53] proposed a pattern-based hybrid intelligent control for rotary kiln process. Within this research different temperatures are clustered with fuzzy logic, the clusters resulted in the different burning zones within the kiln. As any change in the parameters got a relatively long lag, it is a complex process. Depending on the temperature and coal feeding they defined two states Normal and Abnormal and the according rules which have to be applied to control the parameters of the kiln. By applying this model within an rotary kiln in an aluminum plant the production capacity, the operating rate and the operational life span of the equipment has been increased.

The technical report of Steck-Winter and Unger [44] outlines the thermal processing plants in smart factories. The authors try to determine the present state of the thermal processing and which steps will be necessary in future. Besides the required digitalization of the overall production process, also the condition of the machines and the parts within the machines have to be represented in digital way. Therefore a so called digital twin of each machine, including the condition of the parts, the maintenance logs, production capacities etc. will be mandatory in order to perform data-driven optimization. Predictive maintenance could help to reduce inspections and optimize the maintenance plan to keep the failure rate low and therefore the availability of the machines high. Within these thermal process plants predictive maintenance is not yet applicable due to missing data records. The systematic condition acquisition of several spare parts is challenging because of their specific and diverging applications and is therefore not yet sufficient. As a result the knowledge regarding the lifespan of the parts is not yet exhausted.

As presented there is a variety of research concerning optimization of equipment and energy within manufacturing, including several data-driven and machine learning approaches. Although already implemented machine-learning applications have shown that there is a big potential to improve, support and develop manufacturing processes, they are not implemented at scale yet. Still there are a lot of challenges to face, such as underlying infrastructure or how to collect and process the data in an efficient way.

There is a broader field in the research when it comes to energy optimization based on Big Data. There are various models trying to predict the energy consumption and further optimize the parameters to reach an overall optimization of the process.

In manufacturing in general there is a collection of attempts for data driven improvements, but not particularly including thermal processes. The multivariate process is very complex and the traditional improvements were not based on data driven approaches. The acquisition of the representing data is challenging, therefore the optimization of equipment is not yet researched sufficient.

3. Requirements towards a Data Analysis Process

In order to perform a data analysis in an efficient way, some requirements have to be met. The retrieval and access to various data sources can be laboriously and long-winded. Due to different underlying systems and storage methods, the data records vary in their format. The data mostly is stored in separated locations and has no linkage. Which requires additional effort to map the data records to each other. To avoid these efforts the basic requirements towards a Data Lake (DL) are mentioned below. Further to ensure a good practice for the data analysis process, the sequence of common process models are explained. This chapter handles the requirements towards data analysis process, consisting of the requirements of the underlying structure and the process model itself.

3.1. Data Lake

Accessing and collecting data in a traditional IT Infrastructure can be very time-consuming and limited. Within the research of this thesis the term DL arose several times in the context of underlying infrastructure as a basis for data analysis.

The term DL was initially used by Dixon [9] as an approach to handle the challenges arising from Big Data as described in 2.1.2. A DL should therefore be drafted to handle the high amount of data, the speed of the data and the heterogeneous structure. The idea is to create a repository for storing raw data in their native format, such as structured, unstructured and semi-structured and process it whenever it is needed. Based on the raw data stored in the DL several data analysis on different aspects could be performed, therefore the data has to be pre-processed and transformed. An improvement compared to traditional architectures is that the so called data silos could be connected for exploring relations and correlations to gain further knowledge. The principle of a DL structure within an enterprise will be explained in the following section.

The lake could be described as a data management platform where data should be available and accessible for everyone in an organization. The data flowing into the DL originates from a variety of sources producing different kinds of data. Possible input types are shown in 3.1 originating from relational database management systems (RDBMS) which are present in nearly every enterprise such as customer relationship management (CRM), enterprise resource planning (ERP) and further business integrated systems. Logs could also be input streams for the DL, originating from several mobile devices and machines. Another input towards the DL could be sensors or various timeline data sources. The last shown input type consisting of files such as emails, reports, guidelines, tables and similar data. These data can be separated into structured (RDBMS), unstructured (Files) and semi-structured (Logs, Sensors). Another segregation within the data could be dividing repetitive and non-repetitive data. Depending on those categorizations they have to be processed differently. Miloslavskaya and Tolstoy [28] described the way the data is stored as raw, which means they are stored as they were proceeded from various inputs. To find the desired data it has to be search-able, this could be done by labeling the data or adding metadata. As described by Inmon [15] not creating an understandable context to the data could lead to a 'one way' lake where the data is only stored and not consumed. The meta-data should consist of additional information, such as regarding time, place, amount and purpose of the data generation. The meta-data is needed to put the data in context and perform the analytic. The architecture within the lake is flat where each data set get a unique identifier.

As we do not know the questions arising in the future or deriving from data explorations, the possible data operations should be highly flexible. Thus the analytic applications should be designed dynamic, which means they should not be pre-build functions.

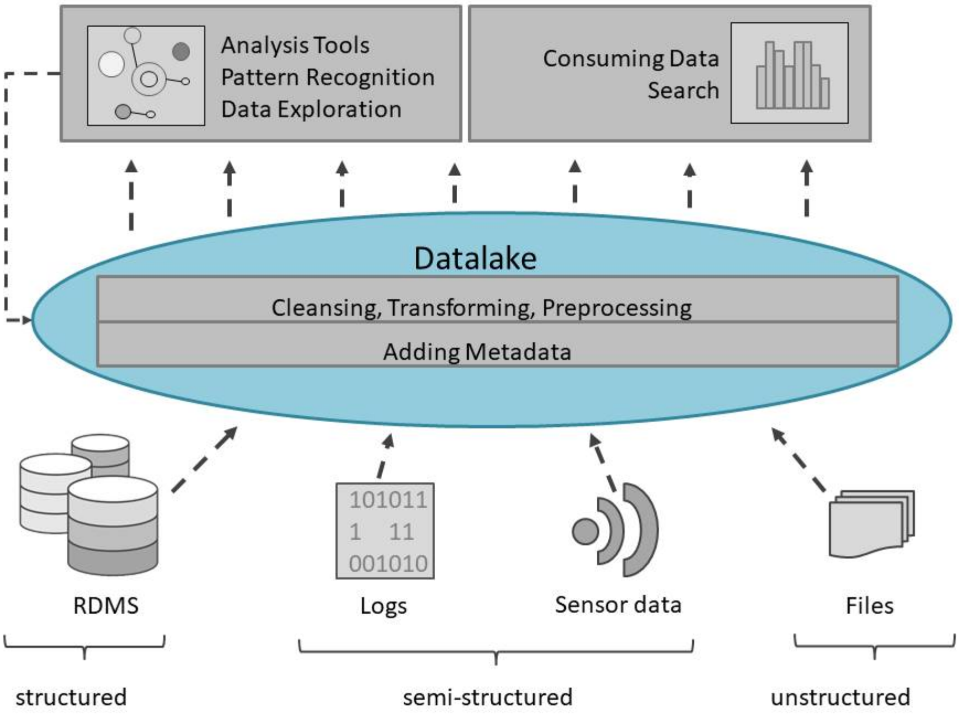


Figure 3.1.: Data Lake design showing of the input of various data sources from traditional RDBMS, logs could originate from several mobile devices, another input is data from sensors or other timeline dependent sources and several files such as emails, reports, guidelines. Whereas the data could be separated into structured, semi-structured and unstructured types. Proceeded in the DL, metadata is added to the raw data. Data-sets are processed, cleaned and transformed on demand and could be accessed for analysis or information purpose.

3.1.1. Structure and Requirements

The requirements towards a DL structure had been defined by Miloslavskaya and Tolstoy [28] and Fang [10].

- **Scalability**
  Handling a growing amount of data also the architecture has to be scalable. The storing, versioning, searching, indexing and archiving features should have the capability to scale as your data. Therefore the format of the data should also support an the growing amount. Further data processing should be scalable by supporting of the needed data formats and processing techniques.
- **Strategy for archiving data**
  When huge amount of data is stored into the lake on a daily basis, there has to be a strategy to archive the data. A possible approach would be identify those parts of data which are no longer used and move them to cheaper store. It could still be accessed, if needed. The data should not be deleted at any time if possible.
- **Searchability**
  The raw data stored in the lake should be accessible and therefore needs to be searchable. To find the data in the lake whenever it is needed, metadata has to added to the files. The data should be set into context to support text-based search queries. Information about who, where, what, when, etc. to to find the according data sets. An Indexing scheme would create the opportunity to identify the files uniquely.
- **Cardinality**
  The relations within the individual data sources have to be represented in a way. At least in a mathematical way and additionally in a visual presentation. The user should be given the ability to explore the interrelations between the different data records, to avoid the data silo representation.
- **Trackability**
  After the raw data is stored in the lake, all the operations should be replicable and therefore be documented. Also information from the metadata (where, when, who) could be helpful here, further applied changes should be within the records. To ensure the operations performed on the data can be tracked and no gained information is lost.

Additionally implementing analysis multiple times can be prevented in this way.

- **Interfaces**
  The Data Lake should be Integrated into the It-Infrastructure for accessing and writing the data. For this purpose defined interfaces should be designed.
- **Shared-access model**
  A shared-access model should ensue that data can be accessed in a central point to multiple formats of data. Additionally it should no require any extracting or transformation of the data for the users, but it should support some form of in-memory computing.
- **Device independency**
  The accessibility to the Data Lake requires to be independent from the device, it should be possible with mobile as well as fixed devices. It should therefore not be bound to a specific operating system or architecture.
- **Agile analytics**
  The platform should be designed to allow a variety of data analytics. Analytical approaches can reach from low to high in complexity and range. The design should be as open as possible for the questions arising in the future.
- **Quality**
  The data quality should be assured at every step in the DL, independent whether the data is structured, unstructured or semi-structured. The quality checks should be separated from ördinaryẅork-flows.
- **Efficieny**
  Due to the high amount of data stored, the processes should be optimized in their efficiency. This could be reached by compression or aggregation of the data.
- **Data is never moved**
  The data should not be moved for any reasons. Instead the analytic process should be based on the data.

3.1.2. Implementation

The implementation of a Data Lake should be an agile approach rather than an ”all in one integration” as described by Hagstroem et al. [13]. Best practices had been achieved by companies which focused on what they want to accomplish within data analysis rather than focusing on the technology factors first. After identifying some use cases, those were implemented in a few pilot projects on different platforms and further evaluated. Within small phases of roll-outs challenges regarding the implementation can be identified in an early stage. Further the future users could give feedback on the system and it could be adapted or redesigned. Fang [10] also recommends to ẗhink big and start smallẗo figure out beforehand where and how to start. Further there should be a team including business and It experts as well as data scientists, otherwise the value from the data would not be extracted.

The underlying architecture mostly is a distributed processing to ensure the large data could be handled in an adequate amount of time. In this thesis the data set for further analysis is collected within an IT-platform from a manufacturing company. As described in the process of collecting the data was very time consuming and not straight forward. As there was no central point of data access and the lack of context it required additional information from process experts.

3.2. Data Analysis Models

The way from a data collection leading to further insights is called data analysis process, also known as data mining process. There are several models to structure the process of a data analysis. In this chapter the two most common ones will be described and compared afterwards. The data mining (DM) process has been defined by Fayyad, Piatetsky-Shapiro and Smyth [11] as follows: *’The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.’* Whereas non-trivial defines the data mining phase, where there are no predefined quantities. Instead there should be extracted complex functions and patterns out of the data sets. The word process implies that there are several steps

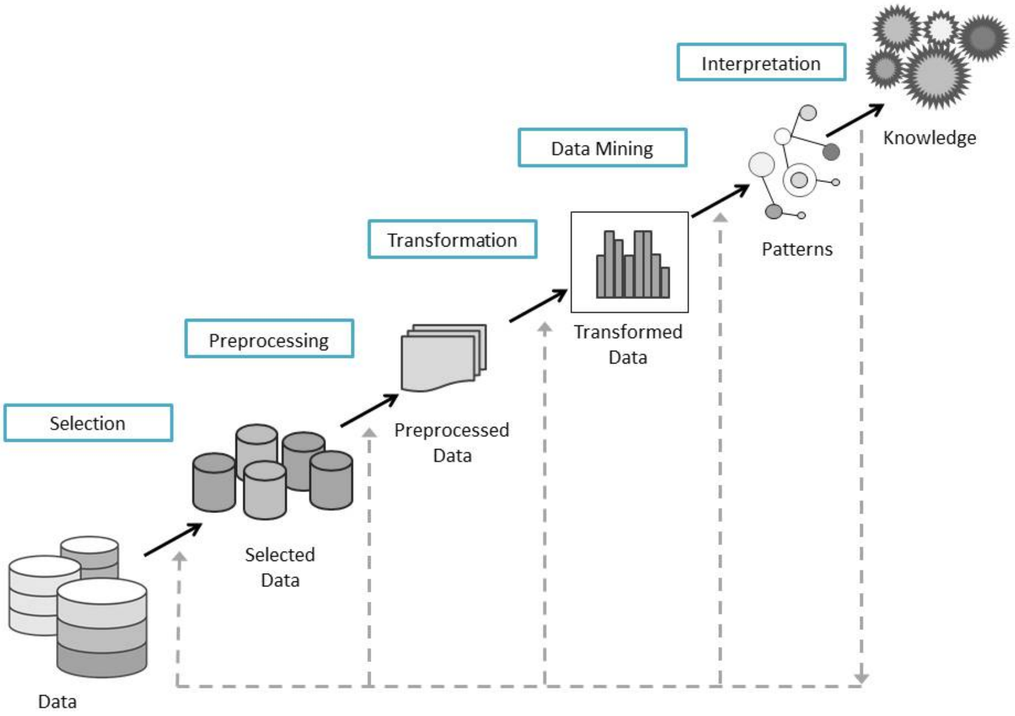


Figure 3.2.: KDD process model shows the main phases and its relations in the data mining process according to Fayyad, Piatetsky-Shapiro and Smyth [12]

that need to be passed, these steps depending on the chosen model and will be described in this chapter.

3.2.1. KDD Process Model

The Knowledge Discovery in Databases (KDD) process model is an iterative and interactive model, starting at the raw data collection leading to further knowledge as shown in Figure 3.2. The process is described by 5 phases consisting of 9 steps and has been defined in 1996 by Fayyad, Piatetsky-Shapiro and Smyth [12].

- **Developing and understanding of the application domain**
  The first step is about understanding the application domain, which requires prior knowledge of the business. Further the application goals should be defined.

- **Creating a target data set**
  Based on the defined goals the relevant data should be extracted from the overall data collection. As the whole data collection would be too voluminous for an analysis, it has to be delimited by focusing only on a subset out of the data.

- **Data cleansing and preprocessing**
  The data has to be cleaned and preprocessed to increase the quality and make analysis more significant. Within this step outliers and noise are removed, as well as consideration how to handle missing data take place.

- **Data transformation**
  The algorithms require different input formats of the data, therefore the data extracted from several platforms have to be transformed into the adequate format.

- **Choosing the function of data mining**
  The chosen data mining function could be for example based on summarization, classification, clustering. The model should be be made upon the already defined goals.

- **Choosing the data mining algorithm**
  The data mining algorithm consists of decisions based on choosing the methods (usually more than one) with the adequate parameters for detecting correlations and patterns within the data.

- **Data mining**
  Detecting correlations and patterns in the data for deeper understanding or discover further questioning in the data mining process. This should be done within the chosen algorithms, forms such as classification rules or trees, regression, clustering, sequence modeling,

dependency, and line analysis.

- **Interpretation**
  The extracted pattern have to be understood and interpreted, mostly within visualization. Information should be filtered, not all of them is relevant to the discovered processes. Further they have to be presented in an understandable way for the future users. From this stage of the model previous steps could be reentered and the data mining process could be redesigned.

- **Using discovered knowledge**
  To gain a benefit of the discovered knowledge, it has to be integrated in the corporation. This could be providing it understandably to the interesting areas or taking actions towards integrating the newly discovered knowledge.

3.2.2. Crisp-DM Process Model

Cross-Industry Standard Process for Data Mining (Crisp-DM) has been developed in 1996 within a project funded by the European Union from Chapman et al. [7]. The model can be applied independent from the industry sector. As shown in 3.3 the model is divided into six phases, whereas the order of those phases can vary. The sequence of entering the process steps is cycling and can lead on both directions, backward as well as forward.

- **Business Understanding**
  In this initial step the business background, the objectives as well as the success criteria should be investigated. Based on the understating of the current business situation the goal for the data analysis should be defined.

- **Data Understanding**
  The phase should include getting the data collection and exploring it with different methods, for example visualizing. This would lead to first insights, identifying quality issues and getting familiar with

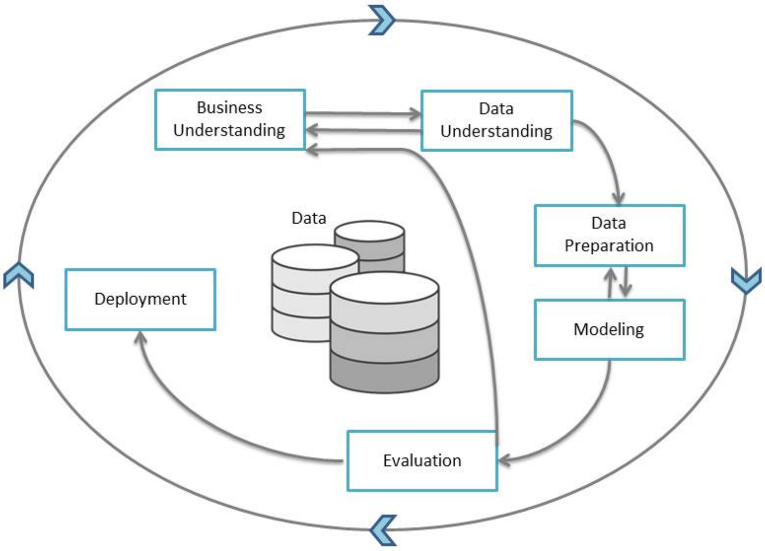


Figure 3.3.: Crisp-DM according to Chapman et al. [7] process model shows the six phases and its relations in the data mining process

the data. Clarifying which data is available and whether it is even useful. The data understanding is closely connected to the business understanding, as to define a data mining goal, at least some basic knowledge of the data is required.

- **Data Preparation**
  Preparation of the data includes choosing the adequate data sets, as well as filtering and cleansing. Depending on the appliance of the data sometimes merging and deriving attributes is necessary for further steps. The needed preparation depends on model, which will be chosen in the next phase. That is the reason why data preparation and modeling states are not independent and therefore alternating.

- **Modeling**
  When it comes to modeling the data adequate models, algorithms and techniques have to be chosen. Within the implementation of the models, the parameters have to be adapted until the results are sufficient. For example the error measure is low enough.

- **Evaluation**
  To ensure that the quality of the process executed is high enough, it has to be evaluated. The evaluation and interpretation of the results could lead to new insights and therefore deeper business understanding. This could lead to new questions and demand further data understanding , therefore the circle would be reentered. At the end the results for the deployment should be selected.

- **Deployment**
  The gained insights should be deployed and presented to the stakeholders. Depending on the use case this could demand the knowledge of a domain expert or no specific background at all. That is why it is important to know in advance what the created model will be used for. There is a need of visualization and organization of the results for the future users, in a traceable and understandable way.

3.2.3. Comparison of Selected Models

In Table 3.1 the phases of KDD and Crisp-DM are compared. As you can see the Data Understanding phase of Crisp-DM is equivalent to Creating a target data set and data cleaning and pre-processing in the KDD model. There are three phases: Choosing the function of data mining, Choosing the data mining algorithm and the data mining stage can be compared to the Modeling stage in the Crisp-DM model. The other stages have corresponding steps in the other models, for example Data preparation can be identified with Data transformation.

Phases of DM - processes	
KDD	**Crisp-DM**
Developing and understanding of the application domain	Business Understanding
Creating a Target Data Set	Data Understanding
Data Cleaning and Pre-processing	
Data Transformation	Data Preparation
Choosing the function of data mining	Modeling
Choosing the data mining algorithm	
Data mining	
Interpretation	Evaluation
Using discovered knowledge	Deployment

Table 3.1.: Comparison of the phases of the data mining process models KDD and Crisp-DM

According to Azevedo and Santos [4] the Crisp-DM model is an implementation of the KDD model, with the purpose for practical use in the industry.

Shafique and Qaiser [42] has made a comparison of the two models and came to the conclusion that the KDD model is more accurate and structured compared to the Crisp-DM model. That is why the KDD model is mostly used by researches and data experts. Due to this comparison the applied process model will be the Crisp-DM as it is better suited within the practical approach. Whereas CRISP-DM model is more sufficient within the industrial sector. The guidance is more adequate for implementing the data mining process in practice.

4. Overview and Evaluation of Data Analysis Tools

There is a variety of tools available for the data analysis process. In this chapter a selection of tools will be compared on several aspects. Depending on the purpose within the data analysis process different criteria have to be met. Therefore the tools will be separated into three stages: scripting tools, visual programming and visualization.

4.1. Selection of the Tools

As not every tool is suitable for every use case, the selected tools are separated into different levels as shown in 4.1. The first category contains scripting tools, where the data process is represented in program code which is executed in the command line. There are several Integrated Developer Environment (IDE)s which support the data analysis process. The second stage of tools is the graphical programming which do not require programming skills but an understanding of the data analysis process. The process is modeled through nodes which represent one step executed on the data set, those nodes have an input and output and can be linked through connections. There are a variety of functionality a node can have, such as reading the data set, filter, merge, applying an algorithm, visualize certain data etc. The third and last category is visualization of the data which offers different techniques to explore the data set in a graphical way. Therefore plots, information graphics, charts, tables etc. could be used.

The selection of tools which will be compared in this chapter is based on the list of mostly used data processing tools according to Nagorny et al.

[29]. As the list contains mostly graphical programming tools, only a part of it: Orange, Weka and Knime had been chosen for evaluation. Further as a programming language R was listed and according to the poll of used data mining tools Piatetsky [35]. Python got the highest share within the used tools, therefore it was added to the tool selection. While researching R and Python, almost every time it was in context with Matlab, consequently this was also added to the selection of tools. 4.1 show that the selection for scripting environments consists of Python, R and Matlab. As an additional category a visualization tool had been added to complete the selection. The tool of choice for visualization is Tableau as it also got a very high rank within the mentioned poll Piatetsky [35]. The following list of tools is chosen for comparison and will be used in the practical part 7 within the data analysis process.

4.2. Evaluation

The tools will be evaluated according to different criteria. One point of comparison is the license needed, whether the software is open source or proprietary. Open source software is listed in the table with the GNU General Public License (GPL) [1]. Further basic skills required to use the tools are evaluated. The scalability, transparency and adaptability will be rated with a number reaching from 1-5, whereas 1 corresponds to very high and 5 corresponds to very low. Further the tools will be reviewed whether each of it supports machine learning and or deep learning in any form. The evaluation containing all criteria is shown in 4.2 and 4.3.

The most obvious difference between the programming frameworks is the license. As Python and R are open source software and licensed under GNU GPL, Matlab is only available with a commercial license. Each of the tools are not only limited to data analysis and have a variety of useful functionalities. These functionalities can be accessed through additional libraries or packages. Ozgur et al. [32] compared those three environments regarding their pros and cons, as well as their suitability in teaching environments. There is a variety of comparing reports of these three tools such as by *R vs*

[1] https://www.gnu.org/licenses/gpl-3.0.en.html

Selection of Tools		
Level	Tool	Description
Scripting	Python	multipurpose programming language, which provides a huge set of helpful libraries for the data mining process such as pandas, numpy, scikit-learn and matplotlib for visualization
	R	is a programming language for statistical computing, besides it also provides machine learning packages for further data analysis and visualization techniques
	Matlab	provides numerical computing, optimized for matrices, also funtions and data can be plotted, further there are also several toolboxes which provide machine learning functions
Graphical Programming	Orange	a workflow can easily be modelled with the nodes representing different functions, data pre-processing and machine-learning can be implemented very fast
	Weka	data-mining tool where the process can be represented through nodes and provides a large set of funtionalities which can be aplied to the data set
	Knime	modelling the data analysis process through nodes which represent one functionality executed on the data set, those nodes have an input and output and can be linked through connections
Visualization	Tableau	easy to use visualization tool, which provides interactive filtering, selecting, highlighting... on the data

Table 4.1.: Overview of the selected to tools for evaluation separated into the three categories: Scripting, Graphical Programming and Visualization

Python vs MATLAB vs Octave [37]. Matlab is very fast concerning execution time when it comes to vectorized operations, but it is the slowest out of all three regarding iterative loops. R is faster than Matlab but still slower than Python. In the subject of scale-ability all three tools are very scalable. Regarding R and Matlab they are limited in scale when using not only vectorized operations, might be the case when pre-processing the data. All three of the listed programming frameworks in Table 4.2 are very adaptable as they offer a massive set of libraries or packages. As they are not only limited to data analysis they provide possibilities to manipulate data and therefore a changing data set, as well as the feasibility to adapt to changing circumstances. Whereas python and R provide a wider range of libraries for adaption. Although some libraries are restricted to a particular version of Python. Therefore some of the libraries and packages with specific versions do not work together at all. Python further allows to create a stand-alone application, independent from the underlying system. R got limited capabilities in creating stand-alone applications, therefore an additional framework is needed. In Matlab this is even harder to accomplish.

All three tools are very transparent when working with them. Considering the libraries have to be used and therefore the parameters and functions have to be known by the user. The required skill-set for Python consisting of programming skills and further have a basic understanding of the libraries, how to find, install and use them. In order to use R you have to know R scripting, which packages to use and how to use them. The same applies to Matlab, you should be familiar with programming in Matlab and depending on what the user tries to accomplish have a solid mathematical skill set. As a basis the user has to have some fundamental understanding of the data set, no matter which of the tools will be used.

The data set mostly has to be cleansed, transformed, filtered etc. before analyzing it. All three scripting frameworks got a vast amount of libraries providing helpful functions to pre-process the data. In Python R and Matlab, Machine learning and Deep Learning functionalities can be accessed through several packages, which need to be installed additionally.

In Table 4.3 the selected data analysis tools are Orange, Weka, Knime and Tableau, whereas the last one is limited to visualization. Orange, Weka and Knime are the selected graphical programming tools licensed under GNU GPL, which provide a Graphical User Interface (GUI) to model the workflow

	Python	R	Matlab
License	GNU	GNU	commercial
Scale	1	2	2
Adapt	1	2	3
Transp.	1	1	1
Skills	programming skills, data understanding, knowledge of libraries	programming skills, data understanding, familiar with packages	mathematical skills, data understanding
Visualization	3	3	1
data set	raw format, pre-processing possible	pre-processing possible	manipulation of data for common formats supported
Addition	huge set of libraries (pandas, numpy, scikit-learn) not limited to data analysis	huge set of packages for statistics and machine learning	ideal for data in matrix format
machine learning	yes	yes	yes
deep learning	yes	yes	yes

Table 4.2.: Evaluation of the selected scripting tools Python, R and Matlab. The criteria of comparison is License (Open Source or proprietary), Scalability, Adaptability and Transparency are rated from 1-5, whereas 1 corresponds to very high and 5 to very low.

	Orange	Weka	Knime	Tableau
License	GNU	GNU	GNU	commercial
Scale	3	3	2	1
Adapt	2	2	2	1
Transp.	2	3	2	2
Skills	knowledge of data analysis process	knowledge of data analysis process	knowledge of data analysis process	understanding the data set
data set	supports common formats (csv, xlsx, sql tables)	supports only particular format for import ->data preparation necessary	supports common formats (csv, xlsx, sql tables)	all data has to be present (no feature generation)
Visualization	2	2	2	1
Addition	supports python scripts	external interfaces available: R, Python, Knime, Matlab	supports python and R scripts as input	excellent for exploratory data analysis, Extensions available
machine learning	yes	yes	yes	no
deep learning	no	no	yes	no

Table 4.3.: Evaluation of the selected graphical programming and visualization tools Python, R and Matlab. The criteria of comparison is License (Open Source or proprietary), Scalability, Adaptability and Transparency are rated from 1-5, whereas 1 corresponds to very high and 5 to very low.

of the data mining process 4.1. Basically all three tools are providing a way to visualize the data flow, where specific nodes representing operations on the data[39]. The order of the process steps executed on the data set can be adapted and changed very fast. Which is why those tools perform very well within an initial comparison of various algorithms. In Weka this is calledK̈nowledge Flow,̈ it further provides the Experimenter and Explorer functions which provide an environment to perform the exploratory data analysis.

The scalability of Orange and Weka is limited, due to runtime. When adapting the data set, the reload requires a certain amount of time depending on the size. When it comes to Adaptability all three tools got a very high rank, because the nodes can be rearranged very quickly and in an easy way. Adaptions concerning parameters within the algorithms are limited to the possibilities provided by each tool, whereas further adaptions could be implemented through several scripts such as python and R. Orange for example allows a python script to be executed on the data set, represented by a node. Whereas Knime offers this possibility for both, R and Python scripts. Further data generated with Knime could be exported for visualization purpose into a report from Tableau. Weka offers an Application Programming Interface (API) which could be accessed within an Python or an R script. Also within the Weka GUI several packages can be installed to support R and Python scripts. When running an external script within those tools, there is a loss concerning execution time. The Transparency of the different tools is quite good, as the user has to select and adapt the parameters in nodes the process is plausible. When executing the nodes Orange and Knime both show the user the progress of each node. Further each node could be started independently, when there is a sufficient input available.

The skill-set to use the tools do not require the user to have programming skills. It is necessary to understand the software tool, which is very intuitive within all three of them. Further the user has to have a solid understanding of the data set itself and what he wants to accomplish and how. The basic functionality of the algorithms would be necessary to adapt the parameters and model the process. When it comes to the data set, Orange, Knime and Weka support the common formats as input. The supported formats within Orange, Weka and Knime are xlsx, csv, txt and several database formats (such as postgreSql or mssql) which require according packages to

be installed. Knime offers additionally to the common formats supported by Orange, integration for a variety of other data formats. Although the csv-reader and the xls packge of Weka offers limited possibilities and the data might require some pre-processing.
Orange, Weka and Knime offer a variety of built-in visualization tools and plots. The data set as well as the results and various outcomes of the analysis can be visualized by applying the corresponding nodes to the data set. Tableau offers interactive and intuitive visualization options for the data set. As input it supports most of the common data formats such as csv, xls etc. Further particular parameters from the input data set can be visualized per drag and drop. Additionally it offers the functionality of highlighting the data points with colors, shapes etc. depending on parameters. Out of these visualizations recurring reports can be generated.
Machine learning is supported in all of the three tools Orange, Weka and Knime. For using classification and regression techniques no additional packages have to be installed. Deep Learning algorithms are not supported within Weka and Orange, but they could be realized in a more complex way, such as implemented via external libraries implemented through Python. In Knime Deep Learning techniques are ready to use within the basic installation.

Depending on the quality and format of data set the pre-processing process requires certain a amounts of steps. Python an R got a variety of libraries and are therefore not limited in their functionality regarding data manipulation. Matlab would be a good fit, when it comes to visualizing the data set in an easy and understandable way. Orange and Weka are a good choice for an initial training of various machine learning algorithms to compare the outcomes of the different algorithms in a very efficient way. Knime offers a lot more built-in functions compared to Orange and Weka, further it comes with an intuitive GUI. Knime could be very useful in various aspects, not only pre-processing of the data , but also training and using machine learning models. All of the compared tools offer visualization techniques and in many cases this would already fulfill the requirements. In case it does not, there is the visualization tool Tableau which visualizes the data set in a interactive way and give the user a great opportunity to explore the data. Most of the tools offer several extensions to integrate within the other tools. For example data generated with Knime could be exported into a report from Tableau. Weka offers an API which could be accessed within an Python script.

There is no multipurpose tool which fit all the needs. In order to map the overall DM process, there might be more than one tool necessary. Depending on use case, data set and the desired outcome of the data analysis process adequate tools has to be chosen. To cover the whole data analysis process reaching from mining the data to visualizing the results, an individually chosen mix of tools will be sufficient.

5. Data Set and Process

In this thesis an the overall equipment efficiency is pursued through a data driven analysis. The underlying data set is generated within thermal processes, more precisely within the electronic industry. The fundamental data set for the data analysis consisting of historical process and product data collected on several platforms. It includes historical data from the machinery, the process and products, as well as maintenance records and quality data. In this chapter the data set is briefly described. Further the environment and the thermal production process where the data is generated, are explained.

5.1. Production Process

The basic data set is generated within a thermal process in the technical ceramic production, where the end products are highly optimized ceramics for technical applications. The ceramic material is widely used, because of its high heat and wear resistance. Further qualities of the material are thermal conductivity and electrical insulation. The applications within the electronic industry includes components, such as sensors, capacitors, actuators and similar parts. The data analysis will focus on the production data of piezo-ceramic components, which could be described as electromechanical transducer. Mechanical energy is dissipated into electrical energy and vice versa.[1]

The production process of multilayer piezo actuators consists of various steps. A simplified outline of the process according to *Production process technical ceramics* [36] is shown in 5.1. The manufacturing of the components

[1] https://de.wikipedia.org/wiki/Piezoelement(Accessed on: 2018-11-02)

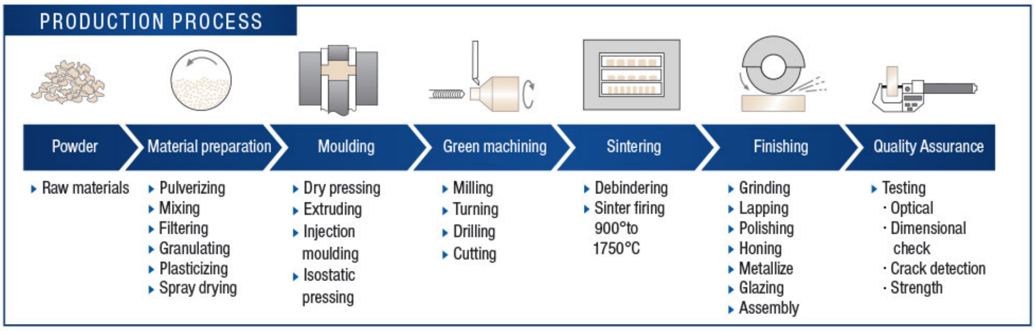


Figure 5.1.: Simplified outline of the production process for technical ceramic components as stated in *Production process technical ceramics* [36]. The manufacturing consists of various steps, reaching from mixing the material to final inspection.

starts at mixing the raw materials as well as the binder for the tape casting of the ceramic foil. The mixture is pressed into desired form, in the case of multilayer pirezo actuators, the foil is generated. Followed by printing the inner electrodes and stacking the foil-layers. In the first thermal process the components are debindered, which means the binder is burnt out at about 500 degree Celsius. Afterwards the ceramic is sintered, which is the firing process with temperatures over 1000 degree Celsius. As a last step in the manufacturing process the contact pins are soldered to the components and they are tested and taped for delivery.

One of the most essential steps regarding product quality within the process is sintering. Further in this step the availability of the sensor data was very high compared to the other steps. In the sintering process the produced parts pass through a controlled heating process in the kiln. In this step the binder within the ceramic-mixture is removed through burn-out. Therefore the fired components will be smaller afterwards and acquired more density which give them their typical properties such as heat and wear resistance. Within the sintering process the temperature exceeds 1000 degree Celsius and has to be held for a certain amount of time. At the same time the temperature acceleration has to be controlled, this means it should not increase or decrease too fast. Due to the high temperatures needed, this manufacturing segment is of intense in energy demand.

The described sintering process takes place in a rotary sintering kiln shown

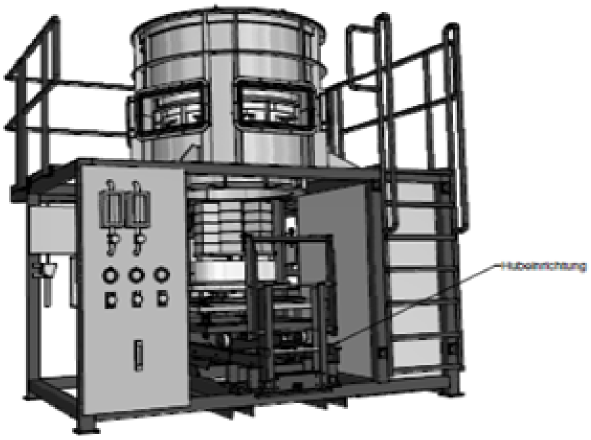


Figure 5.2.: Rotary kiln for the sintering process, generating the sensor data for the data analysis. In this kiln the piezo-electric components are fired and gain their well known properties such as heat and wear resistance.

in 5.2. Within this kiln, where the sensor data originates there are 3 heating - zones, at the top, bottom and one in between. Corresponding to these heating-zone there are also 3 control loops actuating the heating units. The temperature within the sintering kiln is essential to the quality of the products and is defined as a constant value in each run. The defined overall temperature in the kiln is regulated through 3 set temperatures in the according heating zones. The sintering program stays always the same to ensure the product quality does not change. An overview and a more detailed description of the program can be found in the next chapter 6.2. This process and its parameters are already highly optimized and had been adapted and researched for more than 20 years.

5.2. Data Set

The data generated within the sintering process is chosen for analysis, because of the high impact of the condition in the kilns on the end product. The data set described in this section serves as basis for analyzing the process, to gain new insight which would lead to approach for optimize the efficiency within the kilns.

As a first step the data had been retrieved from several platforms within the company specific IT-Landscape. Following data sources had been selected for the final data set:

- sensor data
- product quality data
- maintenance records
- stock disposal history
- process parameters

A detailed description of the data set can be found in 5.1. Each channel of the sensor data had been exported from a industry-specific software solution where all available sensor data has been collected. The quality data including geometry and electrical measurements each had been retrieved from an Oracle Databases. The maintenance action data had been acquired from the central maintenance management system. Several process regulations and further unstructured information had been collected from the process and production engineers. The process of the data acquisition was very time consuming and not straight forward. There was no central point of data access available, instead the records had to be accessed in different ways and on different platforms. This took certain amount of time and access rights for the platforms were necessary. The lack of context required additional information from process experts to collect and understand the data.

The collected data set includes sensor data from 20 sintering kilns, maintenance data, quality measurements geometric and electric as well as several process information. The data listed in table 5.1 is available in the time range of over 2 and a half years.

The sensor data is available in comma separated format, consisting of a time-stamp and the according value. A visualization of the values from the sensor channels is schown in 5.3. The frequency of the data stored reaches from 30 seconds to 5 Minutes intervals, in the time range from 01.01.2016 to 09.09.2018. The temperature values are available from three different locations for top, bottom and inside of the kiln. The values are measured in degree Celsius and available for target and actual values each.

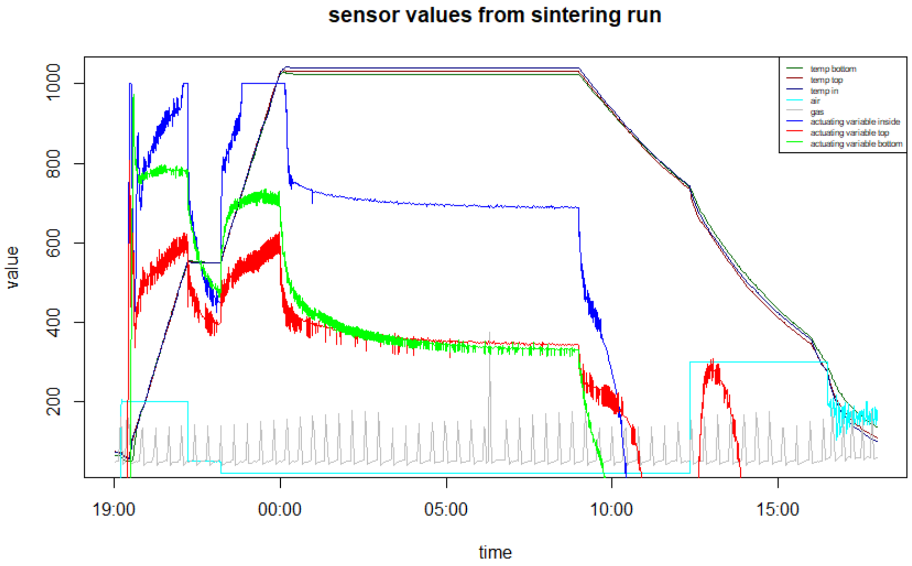


Figure 5.3.: The plot represents one run in the kiln. It contains the available sensor values: the temperatures for top inside and bottom zones, the air and gas flow and the three actuating values which regulate the heaters. The air, gas and actuating values are scaled by the factor 10 due to better illustration.

The sensor values representing the air and gas quantity flowing through the kiln in liters per minute, also target and set values. The actuating variables are available for each heating element, representing a percentage of needed capacity in the control loop. The machine state is representing the operation state of the kiln as described in 6.1. The product quality data consists of geometric and electrical measurements. The geometric as well as the electrical ones are tied to a unique lot number. A so called lot or job consisting of a variable number of parts. The quality measure data for each lot is available statistically aggregated, consisting of average, minimum and maximum value of the charge.

The geometric quality data of the product is measured after the sintering process. Each entry also contains the according time stamp when the measurement is taken. Within the geometric measurements of the parts the width, length, shrinkage, weight, curvature etc. are stored for the transparency of the process.

The electric quality measures consisting of the high and low level signals and the internal resistance of the components. For each of the measurements specification limits are available to ensure the constant quality of the products. When exceeding these limits the parts are rejected for further processing.
The maintenance records consisting of entries about the actions taken on several parts of the kiln. The actions could be visual inspection, which happens on a regularly basis and where the state of each part should be evaluated regarding exchange or repair actions. Another entry could be exchange of a specific part, such as heater. Lining and renewal of the ceiling are further possible maintenance actions. Due to insufficient quality within the maintenance data as described in 5.3, additional data source is needed. Therefore the disposal in the warehouse stock is exported and added to the data set. The collected warehouse data is limited to the specific spare parts from the rotary kilns used in the sintering process. One entry within the stock export includes the name of the kiln, the date when the part is consumed, the description, the amount of part and an additional text.

5.3. Data Quality

The described data set was collected over the past years for tracking purpose. Therefore it was not of importance how the data was stored, but that it was stored. As a result the data was not stored initially for improving the processes, that is why the quality of the data requires certain amount of pre-processing before using the data set for further analysis. Regarding the sensor data the frequency of the timestamps and according values has a high variance. Which varies from 2 seconds between entries to 5 minutes. A value is only stored, when the value changes compared to the last entry but at least every 5 minutes. Independent of that condition the frequency also changes within the running state of the kiln where the data changes continuously. As a result the sensor data has to be re-sampled and interpolated, which would lead to inaccuracies within the data. Within the maintenance logs several challenges appeared. Most of the maintenance actions taken are based on visual inspections, except when there is a defect recognizable in any of the sensor data. First of all the maintenance actions were not clearly defined, due to free text entries from the operator. Therefore the wording is not explicit and leaves room for interpretation. Some entries could not even be categorized whether the action had been taken or not. For example one repetitive entry was 'Exchange of the spare part according to actual wear, further comments needed', where no further comments exist.

Misspelling, abbreviations and different wording for several parts and actions are most challenging, examples are shown in B.1. These circumstances led to a variety of words for the same meaning. Even after figuring out which spare part is affected, the entries were not explicit about the action. The entries also vary in the way the actions are called, examples are listed in the appendix. Summarized the records of maintenance actions are challenging in various aspects. In the survey of Laranjeiro, Soydemir and Bernardino [21] the following criteria to measure the quality within the data are pointed out and examples from the data set are added for better understanding:

- **Accessibility**
  The accessibility of the data is aggravated due to duplicated entries, which refer to the same taken actions. The records containing also ambiguous data, which can be interpreted in different ways. For example the entry including the description of the part and how it is affected, but not whether it has been exchanged or repaired or just recorded. Another problematic case are the abbreviations of different words, see the appendix for detailed examples B.1. Further challenges are different word orderings, special character use etc.

- **Accuracy**
  The point in time when the action has been taken is not accurate, due to more than one entry for the same procedure. This leads to no precise information concerning the timely manner of certain actions. The description of the activity is not clearly defined and leads to subjective interpretations. The entries are not standardized. Further some entries contain misspelled data.

- **Completeness**
  The entries of the records are inconclusive, several actions are based

on the condition of other elements in the kiln. As an example '... the action is taken based on the condition of the element x... the actual extent of repair has to be added'. But no further comments were added. So no information is present whether the action has been taken or not. Further the amount of exchanged parts is not always available. Some maintenance actions are not recorded at all. For example when a heater is exchanged in the kiln. When an entry exists, details regarding which heating zone in the kiln is affected, are missing.

- **Consistency**
  The data is not consistent concerning the question: What kind of action was executed? Is there a record of every maintenance process taken? The way a maintenance action is not unique, therefore the entries are not consistent. The wordings have different orders and representation forms, such as the units of heaters.

- **Currency**
  Outdated temporal data, this could occur when a kiln is closed for production due to needed maintenance and the clearance afterwards is not within the records.

The described challenges made a automated extraction impossible without manual preparation and further information. Therefore an additional data source is added to the data set, the disposal in the warehouse stock. As the entries are standardized, the quality is of high standard. Based on the described records of data, an automated extraction of the maintenance actions taken is not possible. Before further proceeding with the extractions, manual preparation and pre-processing 6.1 6.3 is needed.

Data set		
type	description	size
sensor data	temperature (top, bottom, inside) air quantity actuating variable (top, bottom, inside) machine state gas quantity	about 300 MB per machine (5.99 GB in total)
product quality data	removal bottom and top delta stack length stack width sintered (wing 1-4) bending (wing 1-3, 2-4) total length loss of mass stack weight stack density length shrink internal resistance (RIS) high level signal (GS) low level signal (KS) polarity optical end control	14.4 MB

Data set		
maintenance data	date maintenance action total count operating hours additional description	618 KB
stock disposal	date description of the part machine amount textual description of action	102 KB
process parameter	part specifications load amount offset for each kiln	2.68 MB

Table 5.1.: Description of the parameters of the data set originating from the sintering step within the thermal process available for 20 kilns for a time period of more than 2.5 years.

6. Data Pre-Processing and Feature Extraction

The data set described in the previous chapter has to be linked, cleansed and transformed for further analysis. The data is available in various formats, therefore it has to be pre-processed in order to train machine learning algorithms. As a first step the features are extracted from the cleansed sensor data. Further the available data sources are mapped to each other, so the relations can be explored. The details of the procedure consisting of pre-processing and feature extraction are explained within this chapter.

6.1. Pre-Process

In order to train a supervised machine learning algorithm the data has to be processed into a format which fits for this purpose. An overview of the data sources and their operations is shown in Figure 6.1. The raw data consisting of sensor data, maintenance records and product data has to be cleaned from invalid data. Further the runs representing an operating cycle within a kiln are extracted from the sensor data, based on the state sensor. For a more precise description of the runs they are separated into segments, which represent the heating and stable phase within the sintering process. The described runs and the respective segments are shown in 6.2. The features are calculated from the various sensor channels for the extracted segments. After cleansing and preparing the maintenance records the relevant information about the actions taken are extracted and transformed in a suitable form. The maintenance features further have to be mapped to the runs, the process is briefly described in 6.3. Another part of the raw data is the product specific quality data set consisting of lot number and

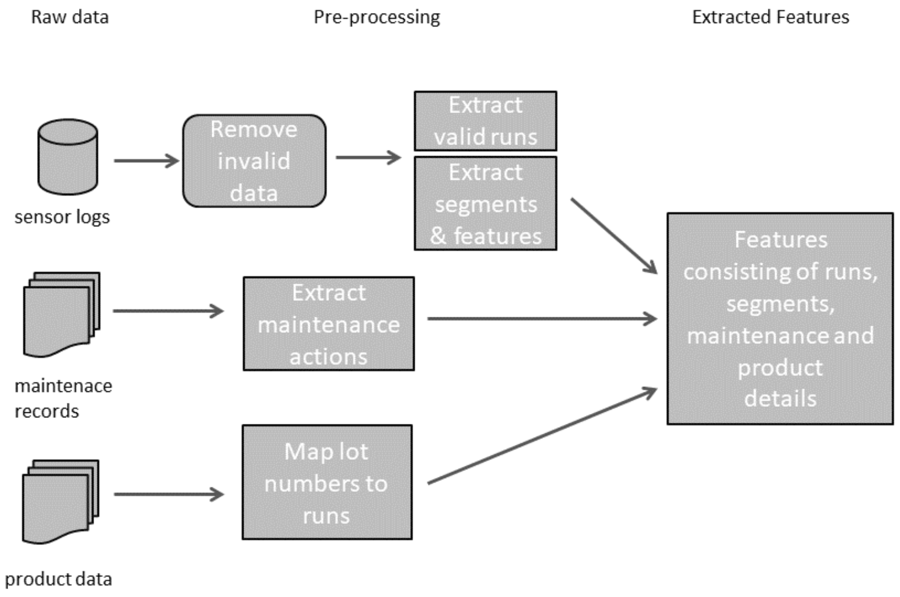


Figure 6.1.: The procedure of pre-processing and feature extraction based on raw data set consisting of sensor data, maintenance records and product data. First the invalid values from sensor logs are removed and the runs are extracted. Further the runs are separated into segments and according values from the sensor channels are aggregated and several features are calculated.

corresponding measures of the products. Whereas each lot number has to be assigned to one run as described in 6.4.

6.2. Sensor Data

Initially all sensor values are validated. Each sensor channel has a range of correct values, due to defects some values exceed the limits and will be removed. For example there are several invalid target temperatures within the sensor data, which are present when a malfunctioning occurs. The defined temperature limit should not exceed 1400 degree Celsius. There are aberrant temperatures present, ranging from 1400 degrees up to 1800 degrees Celsius and had been removed from the data set.

In each rotary kiln are so called runs which represent an operating cycle. A run has a load amount and reaching the time span when the kiln is switched on until the sintering program is terminated. Each of the sintering kilns got a sensor channel representing the state of the kiln. Based on this state the runs had been extracted and validated. The state is represented by the according bits set, which are listed in the table 6.1 below. The kiln is switched off when the state is set to 0, this is only the case when the kiln is closed due to maintenance. Regularly the kiln is in state 256, which represents the idle state. When a program is running the state sensors is set to 512. After the program terminates the state is set to 256 again, except when unforeseen circumstances occur. This could be a manufacturing incident, caused by too high temperatures, too high gas fluid or any other defects. This could be detected through invalid sensor values, then the program is ended manually by the operator. Then the state is set to 1024 and afterwards set to 256 again. The state 8074 is not defined.

Due to extraordinary activities in a kiln such as to maintenance or Research and Development (RD) purposes there are several runs which are not valid or do not represent the production process. A valid run is defined as follows: last at least 20 hours, the maximum upper time limit of 30 hours should not be exceeded and the run should have a valid lot number representing the load.

A run is defined through a start and an end point and within one kiln the runs are numbered. To assure the run is valid and match the sintering curve shown in 6.2 several steps had been defined. According to these states a run can be identified by meeting following conditions: the state of a kiln has to switch from state 256 to state 512 or 2048. Within this run the segments can be extracted based on target temperature sensor.

Further a run is divided into 6 segments (s1, s2, s3, s4, s5 and s6). The Segment s1 is a heating segment where the temperature is raised from environment temperature to 550 degree Celsius. After reaching the 550 degree Celsius limit the temperature stays constant for an hour which

States of the kln		
Value	Name	Description
0	Inactive	Kiln is switched off.
256	Reset	Kiln is switched on and in idle state.
512	Run	Kiln is in active state and sintering programm is currently running.
1024	Halt	The programm has been aborted due to unforeseen circumstances.
2048	Ende	The programm has been manually ended.
8704	not defined	Further possible state which is undefined.

Table 6.1.: Overview of the possible states within a kiln. Depending on the bit set a value represents the operating state.

represents s2. In segment s3 the temperature within the kiln is raised to operating temperature over 1000 degree Celsius. The stable segment s4 at top temperature has to last 9 hours to guarantee the quality of the material as defined in process specifications, whereas the other segments can vary in their duration.

As mentioned in 5.3 there are missing values within the data stream. The missing values are up-sampled and linearly interpolated in order to achieve the same frequency. This step leads to inaccuracies.

6.3. Maintenance Records

The maintenance actions from the records including the exchange of several parts within the kiln. The most invasive operation is the lining of the kiln, which takes several months. Within this time span the kiln is closed for operation. Another action which takes certain amount of time

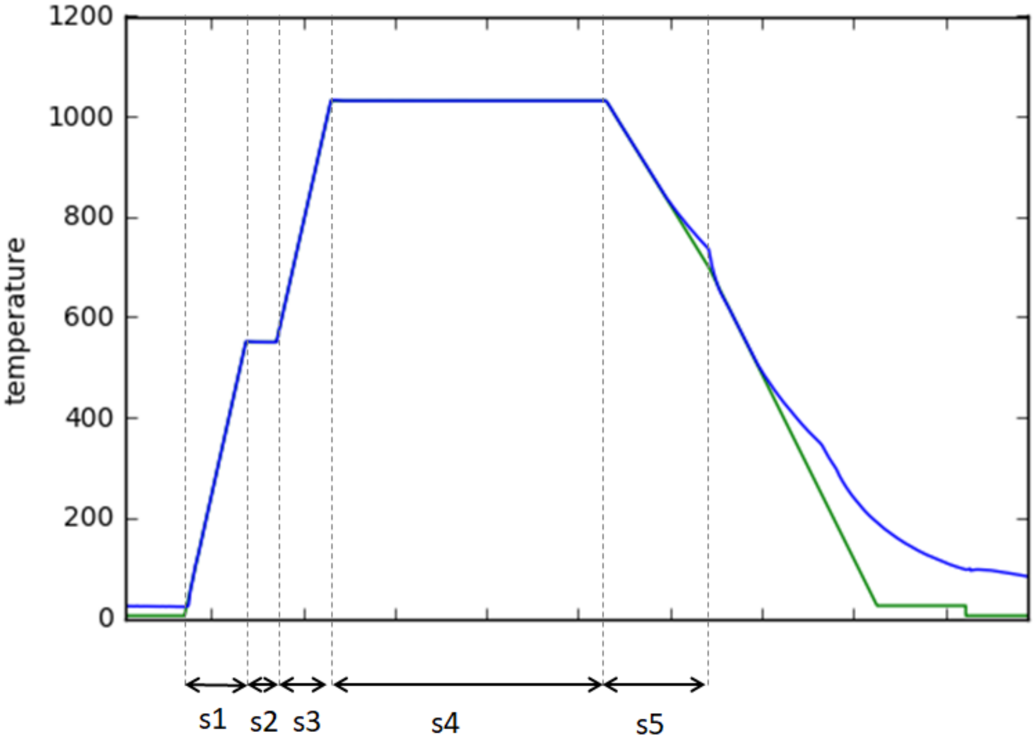


Figure 6.2.: The segmentation of the curve of the temperature representing one run in a sintering kiln. Segment s1 starting at environment temperature raising to 550 degrees Celsius. In segment s2 the temperature is maintained constantly at 550 degree Celsius for an hour. In the next segment s3 the temperature within the kiln is raised to the operating temperature 1035 degree Celsius. The stable segment s4 at top temperature lasts 9 hours. In s5 the temperature is cooled down to 700 degrees Celsius.

is the exchange of the ceiling. Afterwards the target temperature in the 3 heating zones have to be re-calibrated, this is done during a so called 3-point-measurement. Therefore the actual temperatures in three points are measured in the kiln, the chosen points are located where the material is placed during production. Further the heating elements in the three zones can be exchanged separately. In each zone a different amount of heating elements are incorporated. The bottom zone includes 12 heaters, the top heating zone consists of 6 heaters and the inside zone got only 1 heater. To describe the maintenance action accordingly, not only the exchange but also the amount of exchanged elements is extracted. On the heaters rise plugs due to chemical reactions in the kilns. These plugs are removed regularly and therefore described in the extracted features. Additionally each zone has a thermocouple for temperature measurement, which can be replaced. This element is referenced as S1 (inside), S2 (top) and S3 (bottom).

As the maintenance records are not standardized, the quality of the data is not sufficient 5.3 to extract the actions taken fully automatic. Therefore the files have to be prepared to extract the features. First of all a Regular Expression [1] for matching all possible explanations and combinations of words is created. This still implies insecurities whether particular actions are taken or not. To increase the certainty of the extracted entries they will be validated with the trained model as explained in 7.2 Due to insufficient records regarding the heater exchange activities, further data is needed. Additional records from the warehouse operations containing information about how many heaters are exchanged are mapped to the data set.
To represent the data in an adequate format for machine learning a counter of runs is added as a feature. The counter for each part representing how long the specific part has been in the kiln. As an example at run x the ceiling is in the kiln for 30 runs and the lining has not been exchanged for 70 runs. Every time a part of the equipment has been replaced the according counter is set to 0.

[1] https://en.wikipedia.org/wiki/Regular_expression

6.4. Product Data

Charges from the production can be uniquely identified through the so called lot number. This unique identifier is listed within the quality data set to assign the available measurements and the according time-stamps.
For the validation purpose of the runs not only in a timely manner, but also in the load aspect those lot numbers are dedicated to the runs. The load of one run could consist of several lot numbers. A lot number is already assigned to one kiln, but not to the according run. The assignment within one kiln is done through the following matches:

- t is the time of demounting at the end of the current sintering process
- r is the current run number (unique within one kiln)
- r_{start} representing the start time of the run r
- r_{end} representing the end time of the run r
- t_{const} is defined as the time a kiln could be unloaded before r_{end}
- the lot_number belongs to one run if t is within the timerange $r_{end} - t_{const}, (r_{end} + 1) - t_{const}$

With other words, when the load is demounted during two runs within one kiln, then the lot is assigned to this run. As the kiln could be unloaded before the actual end of the run, a time-constant of 1.5 hours is subtracted from the range of the end of each run.

6.5. Extracted Features

Following features had been extracted to describe the runs of the sintering process. Each feature calculated for the overall time range of one run and for each of the segments. Further separated into the three heating zones, representing the top zone, bottom zone and inside zone of the kiln:

- **Integral of the actuating variable**
  The temperatures in the three heating zones are regulated by the controller in each zone. These controllers got a sensor representing the actuator reaching from 0 to 100 %. As there is no power measure and

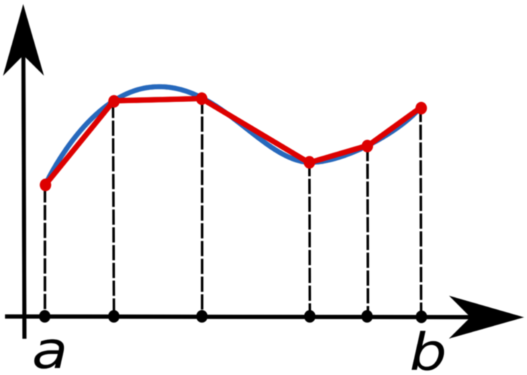


Figure 6.3.: Graphical illustration of trapezoidal technique [46]. The approximation of the integral from a to b is done with the trapezoidal rule. Therefore the area within two sample points is approximated with a trapezoid and the area is calculated. For the overall area from a to b all trapezoids are summed up.

to represent the energy consumption, the integral from the actuating variable is calculated. The trapezoidal rule is used to approximate the integral from the samples:

$$A = \int_a^b f(x) \cdot \mathrm{d}x \approx (b-a) \cdot \frac{F(a)+F(b)}{2} \tag{6.1}$$

As demonstrated in 6.3 the area between two sample values is approximated with a trapezoid and the area is calculated [2]. For the overall area the areas of the trapezoids have to be summed up. The feature is calculated for bottom, inside and top zone in the kiln and also for segment s1-s6 each.

- **Weighted integrals of the heating zones**
 The calculated integrals are summed to represent the overall actuator value needed. As the three actuator variables of the heating zones do not refer to the same amount of power the integrals are weighted.

$$A_{weighted} = c_{top} \cdot A_{top} + c_{inside} \cdot A_{inside} + c_{bottom} \cdot A_{bottom} \tag{6.2}$$

[2] https://en.wikibooks.org/wiki/Introduction_to_Numerical_Methods/Integration

The weight is calculated based on the amount of heating elements in each zone. In total 19 heaters are inside the kiln, with the following distribution: the top zone inherits 6 heaters, inside is only one heater and the bottom zone includes 12 heaters. This leads to the following weights:

$$c_{top} = \frac{6}{19} \approx 0.315 \quad c_{inside} = \frac{1}{19} \approx 0.053 \quad c_{bottom} = \frac{12}{19} \approx 0.632 \quad (6.3)$$

- **Minimum, Maximum and Arithmetic Mean**
  To describe the various sensor values for each of the segments within the runs the minimum, maximum and arithmetic mean are calculated as follows:

  $$a_{avg} = \frac{1}{n} \cdot \sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \ldots + a_n}{n} \quad (6.4)$$

  Those three describing statistics are extracted for each of the sensor values for the overall run and for each of the segments.

- **Average gradient**
  To represent the slope of the sensor values in each zone the gradient is calculated between each of the sample points. Further the arithmetic mean is calculated over all gradients.

  $$m_{avg} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \quad (6.5)$$

  The gradient is calculated for the target and set temperatures and the actuating variables.

- **Ratios**
  To relate the different measures to each other, the following ratios are calculated as additional features.
  - proportion of the integral per degree
  - heating segment integral (s2-s3) per stable segment integral (s4)
  - difference of temperatures in different zones, e.g. difference of top to inside divided by difference of inside to bottom $diff\_oben\_diff\_unten$

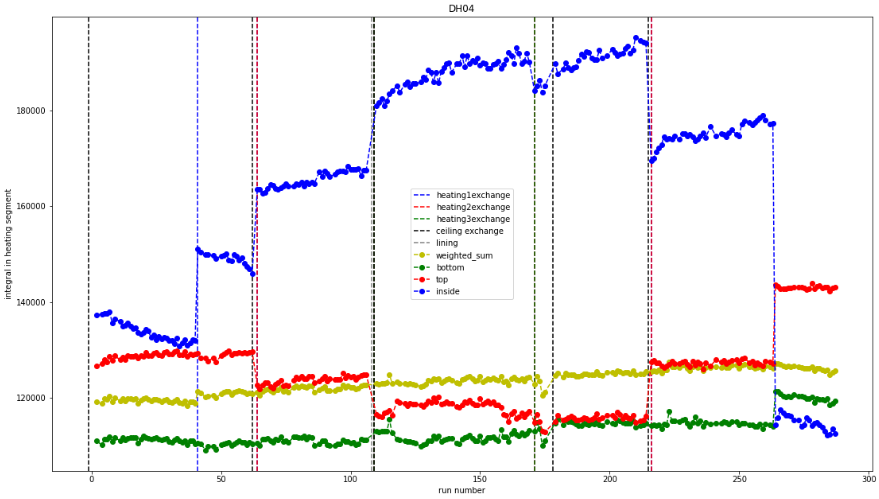


Figure 6.4.: Example plot from kiln ‚DH04 ‘of extracted features including integrals of actuating variables from top, inside and bottom heating zones. The graphic also contains the maintenance actions extracted such as ceiling exchange, lining, heater exchange etc.

- **Difference**
  For the purpose of describing the change between the runs a difference of the mean values is calculated. The value from previous run is subtracted from the value of the actual run.

The graphic representation of extracted features for example of the kiln ‚DH04 ‘6.4. The plot contains the integrals of the actuating variables for the three heating zones and the extracted maintenance actions. Further plots can be found in the appendix B.3.
The overall amount of extracted features had been 220. An overview of the detailed features can be found in the appendix A.1.

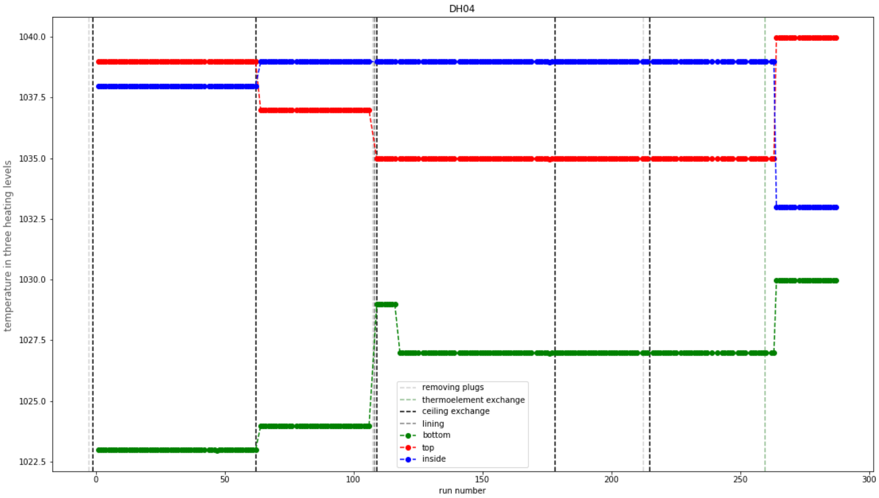


Figure 6.5.: Example of extracted top temperature in each run from kiln ,DH04 ʻ. The target top temperature is set in each of the three heating zones top (red), inside(blue) and bottom (green). The shown temperature is set during the stable segment s4.

6.6. Feature Selection

In order to train the models not all of the extracted features are needed. Based on the huge amount of data the process of training would take longer than necessary. Not all of the attributes contain useful information, some of them do not have variance and are therefore redundant. The model accuracy could be improved by selecting a subset of features.

A way to measure the importance of a feature for the trained model is the Mean Deviation Accuracy (MDA), described by Archer and Kimes [2]. Therefore the feature is „removed "or permuted and the decrease in accuracy is measured. The more important the variable is for classification the higher the decrease of the accuracy. As the features contain correlated values, this does not mean when removing a feature with a high MDA the model will get worse by this amount. When removing this feature, instead any other correlated value might be used.
Although the MDA value is a good indicator, the selection of the features included the consideration of their meaning. As not every feature is representative depending on the use-case respectively the target value that should be predicted. For example the feature *removing_plugs* counted in runs, had always the highest MDA measure. This action is mostly carried out immediately prior to the exchange of the ceiling, therefore this feature has been excluded. In the first segment the heating integral, representing the used power, varies according to the temperature measured at the start of the run. In some runs the temperature is still higher than the environment temperature, because if the residual heat from the run before.

The heater is turned off manually at any point in segment 5 which is not represented in any of the available data. The actuating variable is still active, but no power is consumed from this point in time. Therefore also segment 5 as well as 6 will be removed from the feature-set. During training the features will be scored according to their importance within the models. The final selection of the features is shown in A.4.

The amount of the extracted runs representing the data described in 5.2 consists of 4950. Depending on the purpose of the chosen method, not all of the samples are suitable. Further details are described in 7.2.4.

7. Method

In this chapter the initial selection of the method and comparison of the algorithms is described. This includes the details of training the models and calculating the measurements for the evaluation. Based on the gained insight during the exploratory analysis of the available data, an heuristic approach is designed to optimize the equipment.

7.1. Selection of Method

Initially the three machine learning algorithms, described in 2.1, are trained and compared to select the best fit for the data set and the purpose. Further this comparison is used to ensure that the extracted features and available data is sufficient to represent the state of the equipment and can be used to train the machine learning methods. This initial training is done with the graphical programming tools listed in 4.1. The selected ML techniques LR, SVM and RF are used to classify the target value explained below 7.1.1. As these initial comparison only is done to chose an appropriate algorithm and the used tool, Orange, only supports classification and no classification techniques, the target value has to be transformed from continuous values to classes. To do so the target value is discretized as explained in 7.1.1. In order to compare the algorithms, based each of the techniques a model is trained and evaluated. The model with the highest accuracy is chosen for further improvement and is used for implementation of the approach in described in 7.2.

7.1.1. Setting Target Value

The selected algorithms belong to the supervised learning techniques. Therefore a target variable has to be chosen for training the model. During exploratory data analysis and feature extraction a range of variables had been considered. The prediction should cover the exchange of the ceiling, as there is a wide variety for the durability of a ceiling within the kilns. This part is also of significant interest for the process experts, as they have experimented with several materials and settings within the last decades to find the causalities for the varying duration. The figure 7.1 shows the ceiling including the heating zones top and inside. According to the maintenance records a ceiling could last from 20 to 140. As shown in figure 7.8 the average lasting of a ceiling reaches from around 40 runs per ceiling up to more than 100 runs.

Further the effort to exchange the ceiling is big, not only economical but also in a timely manner. For the exchange procedure the kiln is closed for production for at least 4 weeks where the maintenance action is taken. Afterwards the kiln has to be re-calibrated by measuring and setting the temperature within the heating zones. This means there is an interest to improve the lasting of the ceiling, this would lead also to an improvement of the OEE by raising the availability of the machines. Further the mentioned input factors included in OIE could be decreased by demanding less spare parts.

For the initial training this numeric value was discretized, therefore the ceiling age was divided into 5 intervals representing the 5 classes as shown in figure 8.3. The segmentation in 5 intervals seems to be reasonable, as the classes would represent the age of the ceiling in an adequate way. Reaching from the class representing a new ceiling (< 26 runs) up to a very old one (≥ 105 runs), whereas the time between is separated into 3 additional classes ([26-52], [52-79], [79-105]).

Not all of the extracted runs with the according features are suitable for the training set, due to lack of information. The first respectively the last interval in each kiln do not contain information when the ceiling has been exchanged respectively will be exchanged, see the explanation in 7.2.4. Therefore these samples have to be removed, this reduces the samples from 4950 to 3951.

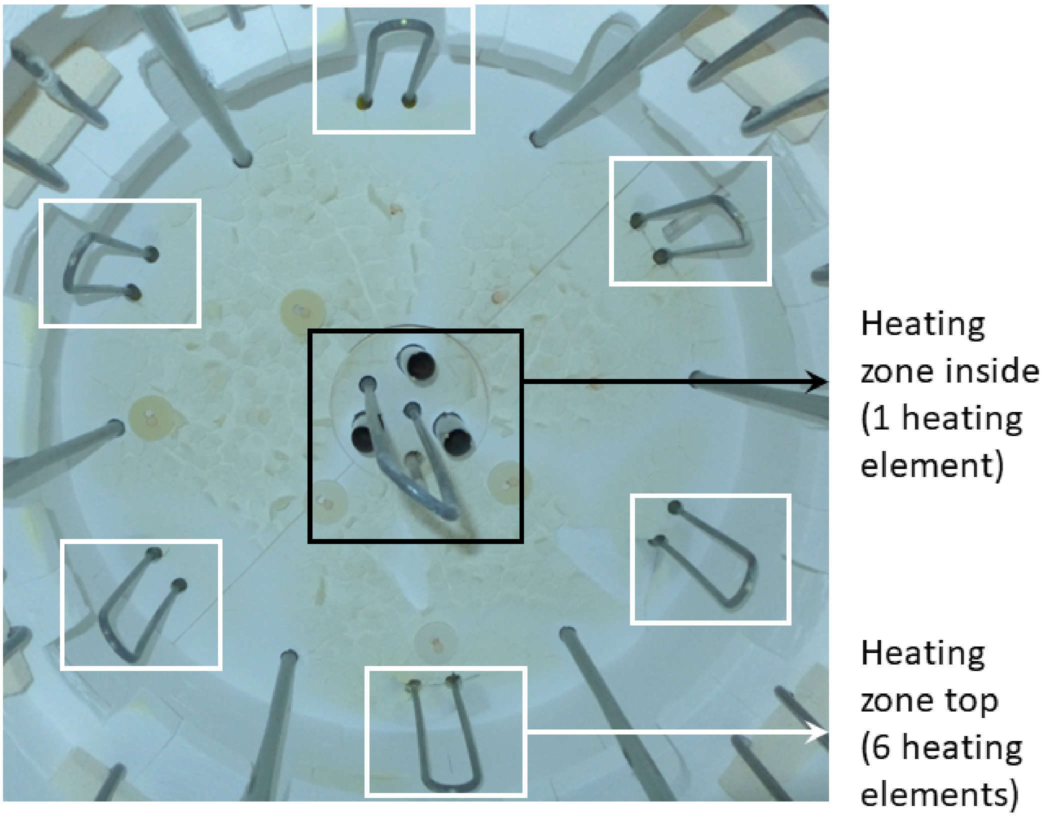

Heating zone inside (1 heating element)
Heating zone top (6 heating elements)

Figure 7.1.: On the picture the ceiling within a kiln with the top and inside heating zones is shown. Whereas the top heating zone consisting of 6 heater, they are marked white and the inside one consisting of only 1 heater is marked black. The black mark covers an area of about 20 x 20 cm. Additionally the wear and tear of ceiling can be recognized by the change in color yellow areas and the small cracks.

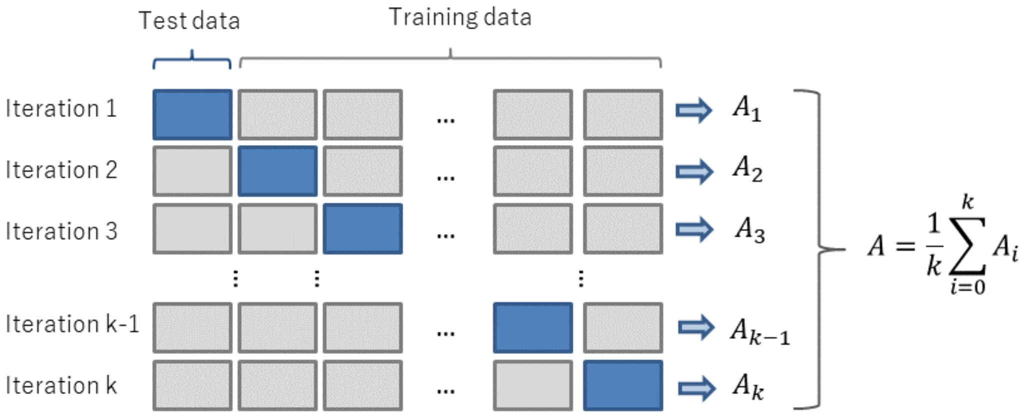


Figure 7.2.: The cross validation method is shown in the graphic. Initially the data set is is randomly split into k-folds, where k has to be chosen accordingly. One fold is chosen to be the test set, whereas the residual k-1 folds represent the training set. The model is trained with k-1 folds, tested with the last fold and discarded. These steps are done for each iteration. The final evaluation is the average over all trained models in each of the k iterations.

7.1.2. Training Method

The chosen training and evaluation method is the so called k-fold cross validation, described by Kohavi et al. [19]. In this method the data set initially is shuffled randomly and split into k folds. Where k-1 fold representing the training set and the test set consist of the last fold as shown in figure 7.2. The model is trained with the training set and evaluated on the one fold representing the test set. After evaluation the model is discarded. These steps are done for each of the k iterations. Choosing k should be good balance between performance and computational cost, in literature a k of 10 is very common [19]. When choosing a k of 10, the concrete numbers for this thesis are as follows: the training set consisting of 3555 records and the test set including 396 samples in each iteration.

The overall performance of the model such as overall accuracy is calculated through averaging the accuracy rates from each trained model.

7.1.3. Evaluation Method

In order to compare the performance of the trained models, the confusion matrix is generated. The confusion matrix is shown in figure 7.3 [8] further explanations can be found in [1]. It basically measures the amount of the accurate and inaccurate predicted classes compared to the actual values.

The confusion matrix is divided into the following four sections:

- True Positives (TP)
  Count of the data samples which actually belong to the class Positive (P) and are correctly classified as P.

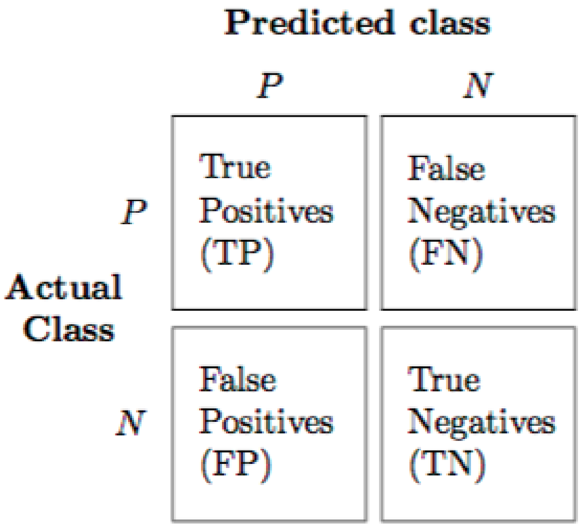
- True Negatives (TN)
  Count of the data samples which actually belong to the class Negative (N) and are correctly classified as N.

- False Positives (FP)
  Count of the data samples which actually belong to the class Positive (P) and are classified as N, therefore the prediction is incorrect.

- False Negatives (FN)
  Count of the data samples which are predicted as belonging to the class P, but actually belong to the class Negative (N). For samples within this category, the classification is incorrect.

True (T) and False (F) corresponds to the prediction, whether the classification is accurate or not. Positive (P) and Negative (N) corresponding to the classes. This example is only to demonstrate the basic functionality of a confusion matrix. The classes could be replaced by any other classes of choice and are not limited in their number.

Further the measurement do not have to consist of counting the data samples, but can be a percentage. Therefore the amount of the data samples in each category (TP, TN, FP, FN) have to be divided by the total amount of the samples classified. Based on the separation in the confusion matrix there are several measurements for a classifier. First the accuracy is measured by



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 7.3.: The figure shows the structure of a basic confusion matrix [8]. It represents a method to measure the performance of a trained classification model. The actual and predicted data are separated into the classes Positive (P) and Negative (N). The actual Positive class, which are predicted as P are called True Positives (TP). Whereas the actual Positive class, which is wrongly predicted as N is called False Negative (FN). True (T) and False (F) corresponds to the prediction, whether the classification is accurate or not. Further P and N could easily be replaced by any other classes.

adding all correct classified samples and divided by all samples classified. As listed in 7.1 the precision of the model is calculated by dividing the True Positive samples by all positive samples. This measure is also known as False Positive Rate. The recall is defined by setting the True Positive Samples in relation to True Positive and False Negatives. Therefore it is also called the True Positive rate or sensitivity. As precision and recall are not independent there is a further measurement which combines those two dimensions. In order to improve the model only the F1 measurement has to be improved, as it represents the harmonic mean.

$$Accuracy = \frac{TP + TN}{Total} \tag{7.1a}$$

$$Precision = \frac{TP}{TP + FP} \tag{7.1b}$$

$$Recall = \frac{TP}{TP + FN} \tag{7.1c}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7.1d}$$

$$\tag{7.1e}$$

The listed measurements are suitable for a classifier, but not for regression techniques. To measure the performance of regression models further metrics are needed. To measure how well the model fits to the data R^2 7.2 is defined, it is also called coefficient of determination. The value can be within [0,1], where 1 corresponds to a model which fits the data exactly. When the model fits the data worse than expected, it also can have negative values [1]. The Mean Absolute Error (MAE) calculates the mean deviation of the predicted values from the actual values.

$$MAE = \frac{1}{n} \sum |Yact_i - Ypred_i| \tag{7.2a}$$

$$MSE = \frac{1}{n} \sum (Yact_i - Ypred_i)^2 \tag{7.2b}$$

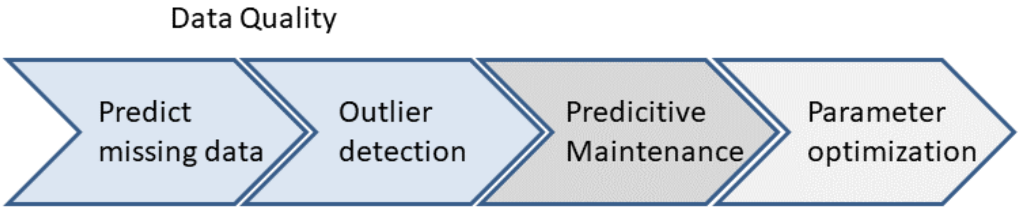


Figure 7.4.: The heuristic approach to optimize the overall equipment efficiency consisting of four steps. Whereas the first two steps dealing with the data quality, first the missing and false maintenance data is predicted and corrected and as second step the outliers are detected. Based on the corrected data set a predictive maintenance model is build. Further the parameters which are influencing the duration of the equipment lifetime are identified and an optimization of those parameters is proposed.

7.2. Approach

Based on the insights and outcome of the exploratory data analysis an heuristic approach has been designed to represent the process for data driven optimization. The approach is based on the pre-processed data set described in 5.2.

The process shown in figure 7.4 consisting of four steps, whereas the first two dealing with the data quality. As the quality of the data, particularly the records of the maintenance actions are not sufficient as presented in 5.3, the maintenance records have to be verified beforehand. Where missing entries or additional unwarranted entries are identified as false and removed from the data set. As a second step dealing with the data quality the outliers will be detected and removed for further analysis. Based on the pre-processed data a predictive maintenance model will be trained and evaluated.

7.2.1. Predict Missing Data

Within the data analysis process the model trained with RF as explained in 7.1. Based on this model the overall data set is classified, for each sample the runs since exchange is predicted. When the calculated difference between actual and predicted variable *runs_since_exchange* is deviating from the average, it is marked from the algorithm. The marked samples have to be manually approved, this has been done with the process experts. In a further step the missing entries are inserted and the wrong entries are removed.

To visualize the process step, an example from kiln DH26 is shown in figure 7.5. Where the first graphic shows the records as they were and the second one shows the revised data set. Further examples can be found in the appendix B.2. After applying the process step there is still uncertainty in the data samples. Not all entries can verified within this theses. For improving the quality, several validation steps could be designed for the data to run through already during acquisition.

7.2.2. Outlier Detection

The detection of irregular runs, which do not represent the routine production process, Median Absolute Deviation (MAD) is used. As it is has been pointed out by Leys et al. [22] the MAD is better suitable to detect outliers compared to other statistical methods, such as standard deviation. MAD is calculated with the following formula:

$$MAD = median(\left| X_i - \widetilde{X} \right|) \tag{7.3}$$

The basic idea is to calculate the absolute deviations of the median value and to find the median of that values. An acceptance threshold is set to a constant value, which could vary depending on the strictness. According to Miller [27] the value 2 corresponds to poorly conservative, 2.5 to moderately conservative and 3 to very conservative acceptance criteria. In this example the threshold is set to 3, as the outliers should be detected very cautious. The according mathematically expression is as follows:

$$\frac{x_i - \widetilde{X}}{MAD} < |\pm 3| \tag{7.4}$$

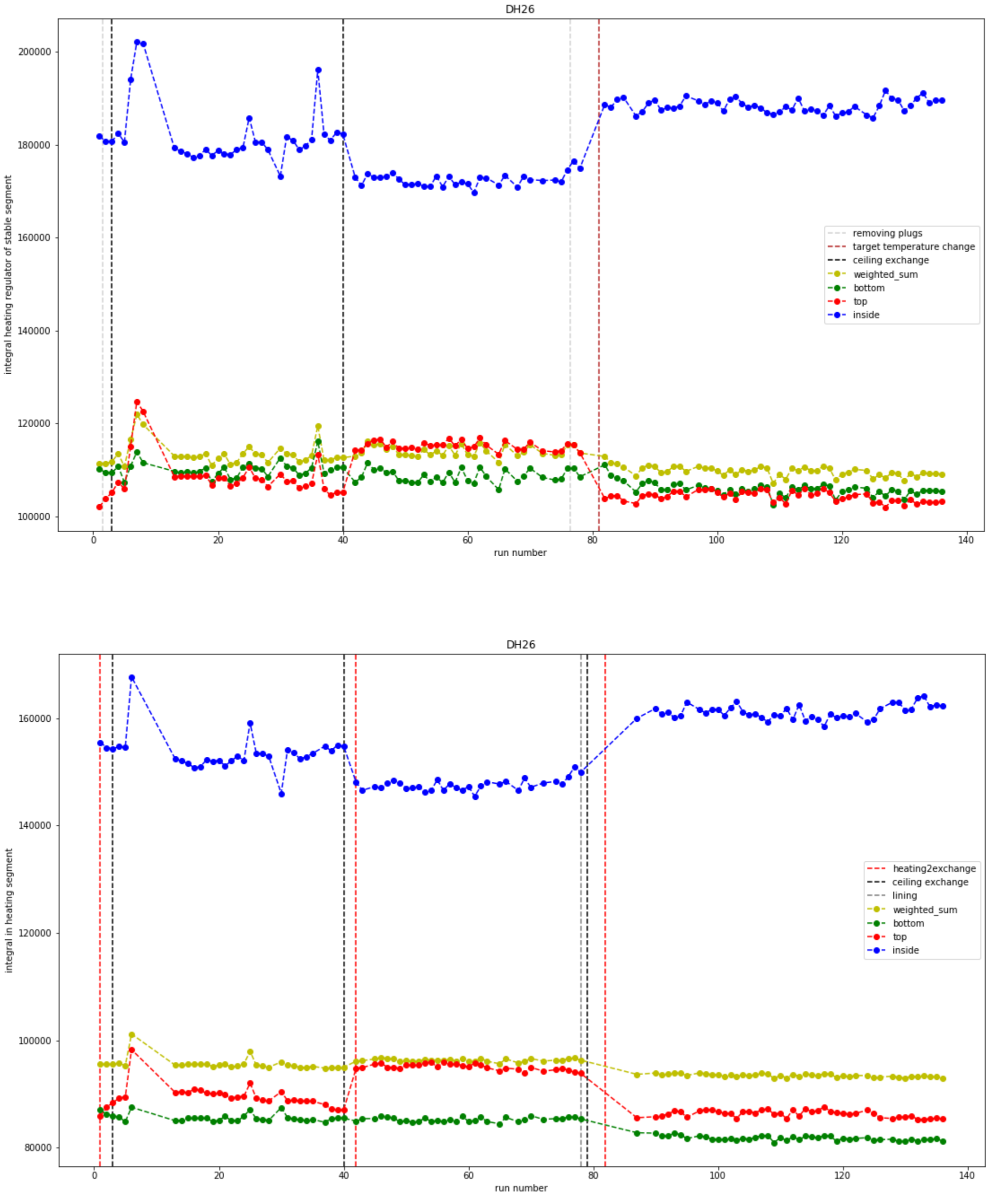


Figure 7.5.: The graphic shows an example for kiln DH26, where a ceiling exchange was not in the records but predicted by the algorithm. As it is shown at around run 80 the ceiling is predicted to be exchanged. In this case this seems legit, due to the lining of the kiln.

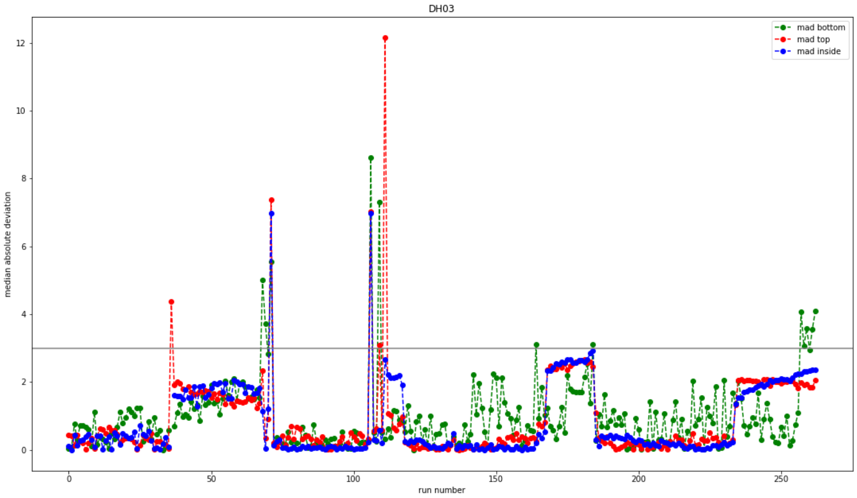


Figure 7.6.: The graphical representation of the mad values for kiln ‚DH03 ‘. The mad is calculated for the integrals which represent the power consumption in the top, inside and bottom heating zones of the kiln. The constant threshold is chosen very conservative represented by the value 3. Each mad exceeding the threshold is marked as an outlier.

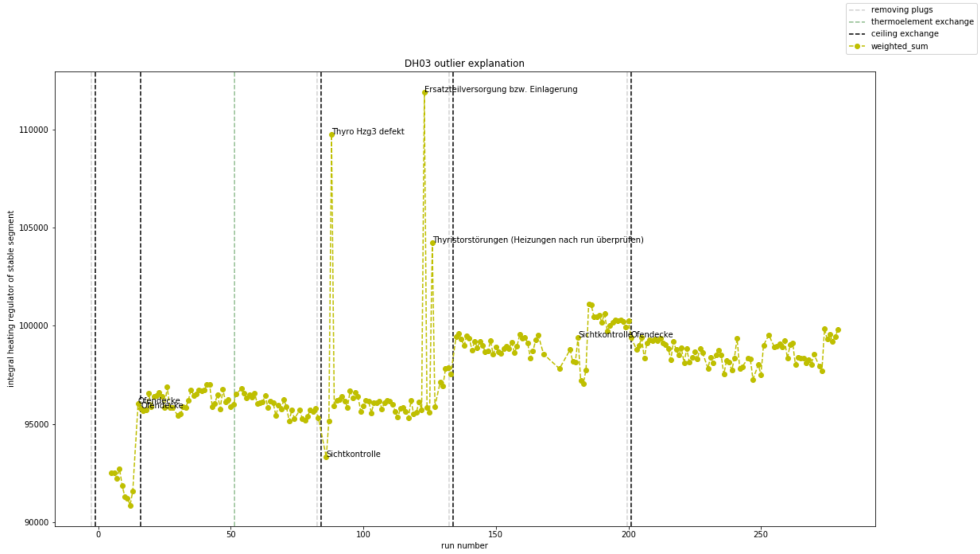


Figure 7.7.: The plot contains the detected outliers from one kiln (DH03) set in a semantic context. Therefore the anomalies are mapped to the according records. The time frame ranging from the end of the last valid run and the start of the next run are chosen to select the incidents of the maintenance entries. Those entries are plotted to the according point in the curve. When no textual description is present, the data point is marked with an *.

As shown in 7.6 the calculated MAD from the data the values which exceed the chosen limit 3 are marked as an outlier.

To set the found anomalies in a semantic context, the data points are mapped to the according maintenance records. All incidents recorded between the end of the last valid run and the start of the next valid run are mapped to the outlier. The figure 7.7 shows an example from one kiln and the according outliers. When a data point is detected as outlier, but there were no records available the according point in the curve is marked with a *. To provide better overview the semantic meaning of outliers are plotted on the weighted integrals, containing the integrals of the three heating zones 6.5.

The MAD is calculated for each setting interval in the kilns. Because changing the settings could also lead to an abrupt change of the the actuating variables. An example is shown in figure 6.4. Therefore no sliding window

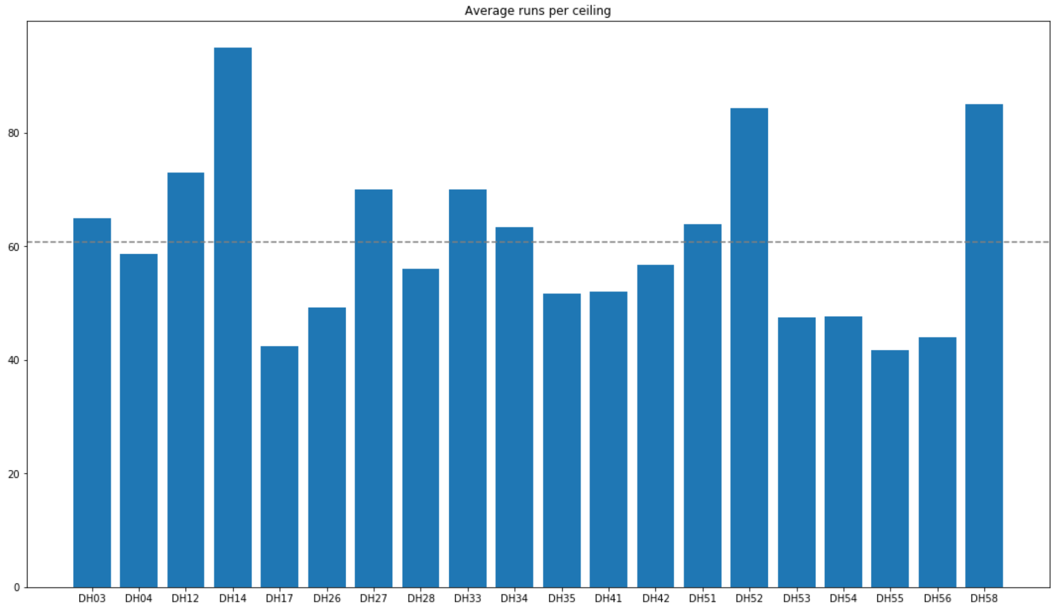


Figure 7.8.: The plot shows the average lifespan of a ceiling represented in runs for each of the inspected 20 kilns. There is a variation reaching from 20 runs to 140 runs, whereas the overall average is marked at about 60 runs per ceiling.

is chosen, this would lead to mark outliers around the jumps of the values, which would be wrong. This procedure is performed 2 times before in a further step the samples are removed from the data set in order to train the models.

7.2.3. Predictive Maintenance

As the trained algorithms have a moderate accuracy the next step within the approach is to build a predictive maintenance model. The prediction should cover the exchange of the ceiling, because there is a wide variety for the durability of a ceiling within the kilns. The plot 7.8 is a visualization of the extracted feature, it shows the mean value of the life-cycle of a ceiling counted in runs.

The actual state is that the ceiling is visually inspected every 10 runs in each kiln. Therefore a person from the maintenance has to go to the kiln and make a prediction whether or not the ceiling will last for another 10 runs. An predictive maintenance model is build to predict the duration of ceiling, by predicting how many runs the ceiling will last. The prediction of this model could be used to send a trigger to the maintenance software. Based on this a message could be generated to visually inspect the state of the ceiling before the exchange is predicted, prior to a chosen threshold. This would decrease the amount of visual inspections needed. The evaluation of the model is described in the next chapter 8.1.

7.2.4. Parameter Importance

To discover the influencing factors of the duration of the ceiling, the model will be retrained. Instead of predicting when the ceiling is exchanged, the model should predict how long the ceiling will last until it has to be exchanged. Therefore the counter of runs since ceiling exchange is rearranged to total runs the ceiling will last. For better understanding the two different values have been plotted in figure 7.9.

As not all the records in the data set contain the information of how long the lifespan of the ceiling is, those records have to be removed for this training purpose. The intervals in the beginning and or at the end of the data series, as an example in figure 6.4 the last exchange at the ceiling is about at run 220. The runs after this exchange does not contain the data about the lasting of the ceiling, therefore it is removed. Based on this model the most influencing parameters for the duration of the ceiling can be identified. Further the outcome can be discussed with the process experts. The parameters could be adapted in the data set to see the effect on the classification of the model. Additionally an overall description of statistics for each kiln is generated, for the overview for the process experts for future usage.

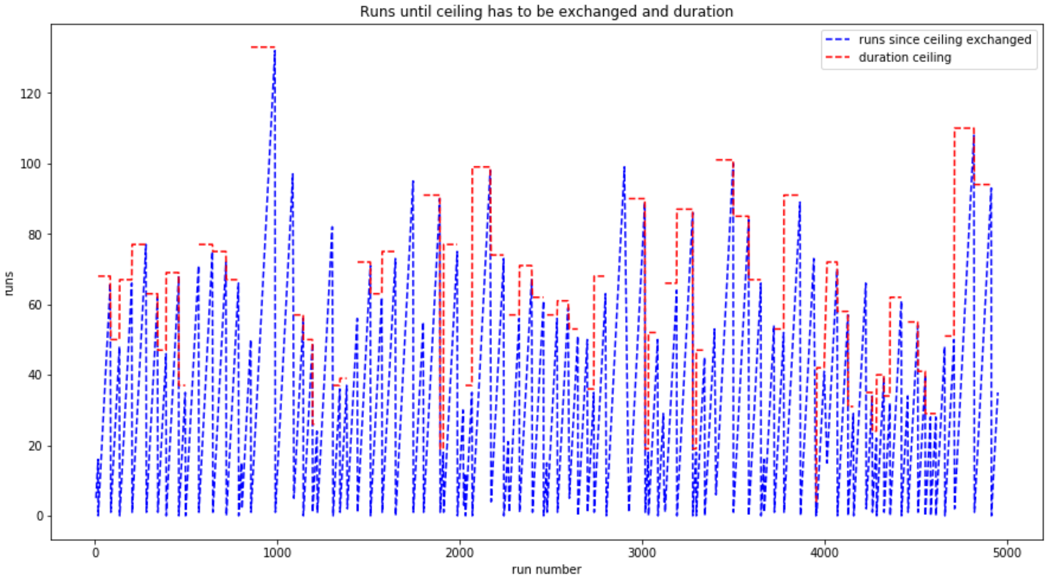


Figure 7.9.: The graphic shows the counter until the ceiling has to be exchanged compared to the duration of the ceiling in the kiln. Overview of all 4900 runs, where some intervals do not have a duration counter, which is represented by the red line. This is because in the beginning of the data points there is no information how long the ceiling is already in the kiln. At the end of the data there is a similar problem, no information is given of how long the ceiling will last in the kiln.

8. Evaluation

The results of the described approach in the last chapter will be evaluated this chapter. To asses the performance of the trained models the measured indicators will be presented. Therefore the results of the various measurements are presented and discussed and interpreted afterwards.

8.1. Results

The initial training of the models to select the method as explained in 7.1 led to the results listed in table 8.1. The table contains the prediction accuracy averaged over all classes for each method. The definition of the used measurements in this section can be found in 7.1.3.

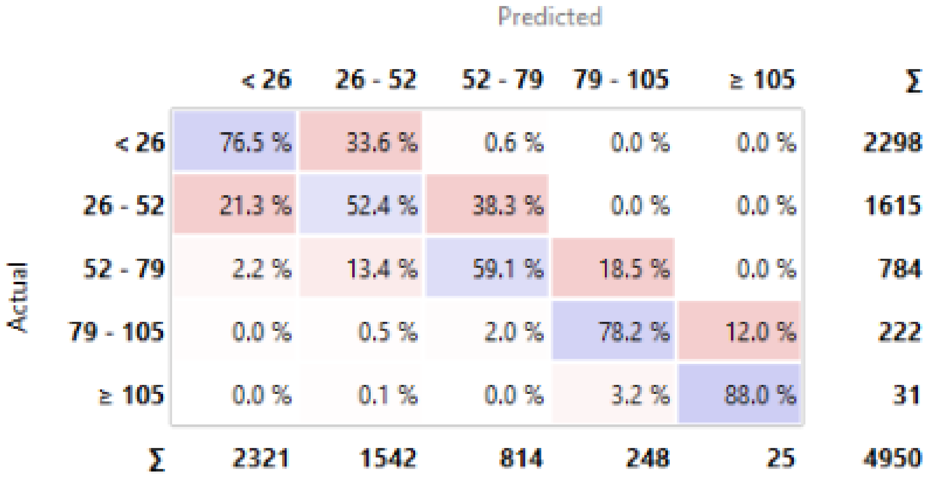
All three algorithms, LR, SVM and RF, had been evaluated with cross validation. Where the chosen k was 10, as explained in 7.1.2. The initial training of the models was done with Orange 4.1 and the training parameters were set to default. The classification accuracies, in table 8.1 referenced as CA, of the models had been very distinct. The accuracy of the LR model was about 0.56, which was quite similar to the SVM technique with the linear kernel, where the accuracy was about 0.60. The SVM with the non linear kernel Radial Basis Function (RBF) had a slightly better performance, the accuracy was 66%. Whereas the initial trained RF model was 0.95 accurate. The classes can be seen in the confusion matrix 8.3, they represent the discretization of the target variable into 5 intervals as explained in 7.1.1.

Based on the results of this first training approach, the choice was to implement the described process 7.2 based on the RF algorithm. First step in order to have a more accurate model and be able to make more precise

Scores

Method	CA	F1	Precision	Recall
Logistic Regression	0.568	0.514	0.517	0.568
SVM (linear kernel)	0.603	0.611	0.639	0.603
SVM (RBF kernel)	0.663	0.661	0.660	0.663
Random Forest	0.959	0.959	0.959	0.959

Table 8.1.: Initial comparison of the selected machine learning techniques. The graphic contains the prediction accuracies of the different trained models.

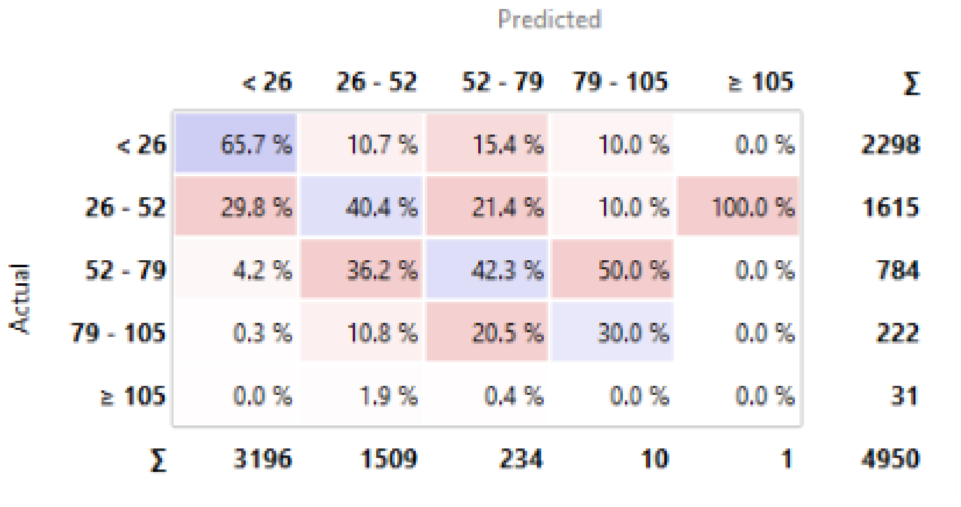


Predicted

Actual

	< 26	26 - 52	52 - 79	79 - 105	≥ 105	∑
< 26	76.5 %	33.6 %	0.6 %	0.0 %	0.0 %	2298
26 - 52	21.3 %	52.4 %	38.3 %	0.0 %	0.0 %	1615
52 - 79	2.2 %	13.4 %	59.1 %	18.5 %	0.0 %	784
79 - 105	0.0 %	0.5 %	2.0 %	78.2 %	12.0 %	222
≥ 105	0.0 %	0.1 %	0.0 %	3.2 %	88.0 %	31
∑	2321	1542	814	248	25	4950

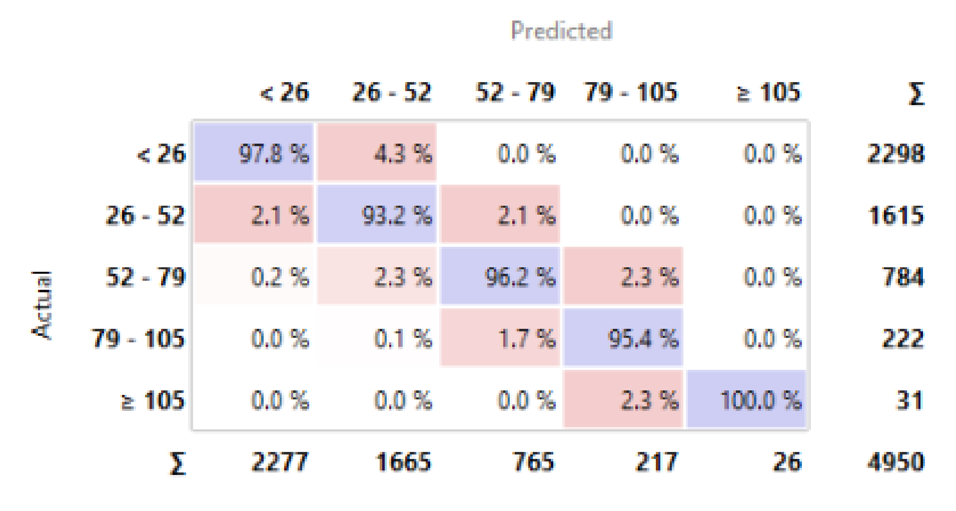
Figure 8.1.: The confusion matrix of the trained SVM showing the result of each target class. The overall Classification Accuracy is 66.3%.



Predicted

Actual	< 26	26 - 52	52 - 79	79 - 105	≥ 105	Σ
< 26	65.7 %	10.7 %	15.4 %	10.0 %	0.0 %	2298
26 - 52	29.8 %	40.4 %	21.4 %	10.0 %	100.0 %	1615
52 - 79	4.2 %	36.2 %	42.3 %	50.0 %	0.0 %	784
79 - 105	0.3 %	10.8 %	20.5 %	30.0 %	0.0 %	222
≥ 105	0.0 %	1.9 %	0.4 %	0.0 %	0.0 %	31
Σ	3196	1509	234	10	1	4950

Figure 8.2.: The confusion matrix of the trained LR showing the result of each target class. The overall Classification Accuracy is 56.8%.



Predicted

Actual	< 26	26 - 52	52 - 79	79 - 105	≥ 105	Σ
< 26	97.8 %	4.3 %	0.0 %	0.0 %	0.0 %	2298
26 - 52	2.1 %	93.2 %	2.1 %	0.0 %	0.0 %	1615
52 - 79	0.2 %	2.3 %	96.2 %	2.3 %	0.0 %	784
79 - 105	0.0 %	0.1 %	1.7 %	95.4 %	0.0 %	222
≥ 105	0.0 %	0.0 %	0.0 %	2.3 %	100.0 %	31
Σ	2277	1665	765	217	26	4950

Figure 8.3.: The confusion matrix of the trained RF showing the result of each target class. The overall Classification Accuracy is 95.9%.

trees	features	R^2	MAE	MSE
50	5	0.917	4.198	47.155
50	10	0.934	3.358	37.702
50	15	0.938	3.036	35.310
100	5	0.922	4.087	44.874
100	10	0936	3.263	36.757
100	15	0.939	2.979	34.831
1000	15	0.940	2.904	34.201

Table 8.2.: Overview of the results trained with varying training parameter. Where trees corresponding to the trees in the forest, features to the maximum features in each tree.

predictions, the RF was used for regression. Instead of building classification model, the exact amount of runs since ceiling exchange is predicted.

The model has been trained with varying parameters, such as the amount of trees in the forest($n_estimators$ = 50, 100, 1000) and the limitation of features ($max_features$ = 5, 10, 15) used in each tree. The model with 100 trees was a good balance between the needed time for training the model and accuracy. The increase of the performance for 1000 trees was significantly low, for comparison the Mean Squared Error (MSE) and MAE are shown in 8.2.

Depending on the parameters of the training algorithm, the MAE of the model is between 2.9 and 4.1 referencing the average error of the regression. The value corresponds to runs in the kiln. The MSE reaching from 34 to 47. The described results in table 8.2 were measured in models trained with the selected subset of features. Compared to the initial training with all available features, where the MAE was 7.716 and the MSE was 108.146, the result has significantly improved within the model trained with the selected features. So has the time needed for the model to be trained.

Some of the last intervals in the kilns had been removed due to lack of information. See runs approximately 210 - 300 in 6.4 as example. These records do not contain data when the ceiling will be exchanged the next time, so they can not be used as training or test set. About 400 of these runs can be classified with the maintenance records and were used as an additional validation set. As some of the kilns are closed for ceiling exchange within the time frame of available data, these extra information had been extracted manually for evaluation purpose. Based on these unseen data, as it has not been used for training or test set, the model for predictive maintenance has been evaluated. The target of the model was to predict the runs until the ceiling has to be exchanged. The evaluation on the described validation set led to an MAE of 5.438 and an MSE of 33.42762. The MAE increased from less than 3 to more than 5, but the MSE decreased for about 1.4. Which indicates that the variation of the prediction error in the validation set is smaller than in the test set. As the validation set only consisted of 5 intervals, this result is has to be seen carefully.

The significance according to MDA 6.6 of the features had been plotted after training of the several models. There has been differences due to the random chosen features, but the most influencing factors in each of the models had always been a selected group of features. An example result can be seen in the appendix A.4.

According to the MDA the age of the heaters as well as the lining of the ceiling were always important rated features. This could be summarized by the overall condition of the kiln seems to influence the lifespan of the ceiling. The target temperatures had been identified as the other influencing factors. Not only the absolute temperatures set, but the differences of the temperatures to the other zones. The inspected kilns are the same type of construction, the only vary in their settings. The most significant difference in the settings is the target temperature in each heating zone. The program temperature should always be the same and constant, but the overall temperature in the kiln can be reached through various temperatures in each level. As the are measured and set from the maintenance personnel and there are no restrictions and specification except for the overall temperature. Based on this insights further analysis had been done, the statistic for each interval had been extracted.

To illustrate this the average top temperature from top and inside had been

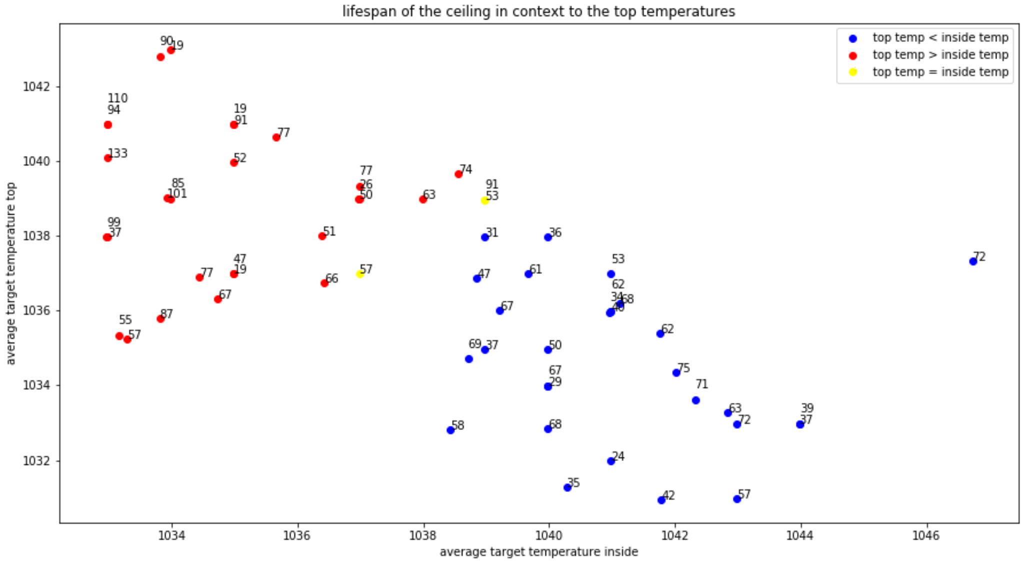


Figure 8.4.: Visualization of the intervals of the ceilings with the according average target temperatures from the top and inside heating zones. Each dot corresponds to one interval of the ceiling and the lifespan is plotted next to them with according lifespan of the ceiling. The data is separated by colors of the corresponding maximal temperatures, red is top heating zone and blue is inside heating zone.

plotted representing the intervals 8.4. The intervals representing a point in the graphic are colored with red if the maximum temperature from the top heating zone is higher than the temperature in the inside heating zone. Next to the dots representing an interval, the corresponding lifespan of the ceiling is plotted. The intervals related to the highest lifespans (> 90 runs) had two aspects in common: the maximal target temperature has always been in the top heating zone and the temperature set in the inside heating zone has always been less than 1040 degree Celsius. The the ceiling intervals represented by the red dots lasted in average 70.19 runs, whereas the intervals plotted as blue dots had an average ceiling lifespan of 53.93 runs.

8.2. Discussion

The comparison of the initial training results shows a significant difference between the RF, the SVM and LR techniques. The variety of these results may derive from the fact that the data is not linearly separable. Whereas SVM and LR both trying to find a suitable hyperplane dividing the different classes, RF classifies the data set according to splits in their features. As one of the major features to describe the actual state is the integral of each heating zone and these integrals move within different ranges depending on the actual settings of the kiln. When considering the integrals and their behavior, the observation shows that the change of the integrals with growing age of the ceiling is not consistent. Depending on the settings of the kiln there is a different behavior of the features over time. Further visualizations of these properties are shown in the appendix A.2. The integrals in each zone depending and reacting on each other. For example the setting of the maximum target temperature effects the control behavior. The heating zone with the maximum target temperature is increasing over time, whereas the according other zone is decreasing. An exemplified illustration of this behavior is shown in 6.4 and 6.5, where the temperatures zones are switched several times. When top and inside heating zone are set to the same target value the control behavior seems to be random. As the integrals were calculated based on percentage points which refer to undocumented states they are not in the same range. These undocumented states could be actual electrical voltage, temperature offset, etc. which is individually set in each kiln depending on the actual point in time. Summarized the integrals vary in their initial values depending on the settings and in their alteration type and their alteration rate. This fact makes it really hard to compare the kiln to each other, as any of the kilns got different settings and resulting control behavior. The circumstance that the performance of SVM, even when trained with the non linear kernel, compared to RF is significantly worse may derive from the fact that the behavior of each kiln cannot be described with one function. An improvement in the results of the trained SVM model could be reached through calculating the normalized integrals per setting interval and kiln. Further improvement could be achieved by acquiring not referencing values available in percentage but absolute values, especially when it comes to any kind of power measures.

The performance of the predictive maintenance model based on the RF algorithm seem to be sufficient for triggering the message for the maintenance system as described in 7.2.3. As the MAE is around 3 and 5 in the very limited validation set, the message sent for visual inspection could be chosen with an according offset. For example when the prediction of the ceiling to be exchanged reaches 10, the message should already be sent to the experts to decide how long the ceiling will last. The predictive maintenance model in this case has only a supportive function, the final decision when to change the ceiling would still be on the expert but could help to reduce the visual inspections. In order to run this scenario fully automated, the data quality has to be taken care of in the first place.

The ranking of the features from the trained model indicates that the settings of the target temperatures in the three heating zones has an impact on the lifespan of the ceiling. Besides the overall condition of the ceiling and the particular parts, most influencing factors seem to be the temperatures in the two heating zones right below the ceiling. These zones are shown in 7.1, referenced as top and inside heating zones. By increasing the period of the ceiling in a kiln, the OEE could be increased as well. Based on the available data not only the absolute temperature is of importance to the life-cyle of the ceiling, but primarily the difference within the zones. When the top heating zone has the maximum set temperature the lifespan is more likely to last longer 8.4. By only applying this restriction, top heating zone has to have the maximum set temperature, the potential lifespan improvement seems to be valuable. The red intervals shown in 8.4 might be improved by more than 30% from an average of 53 runs to 70 runs. As the actual average lasting is about 60 runs, as shown in 7.8, this would lead to an overall increase of the average ceiling lifespan of 16%.

Further improvement of the OEE seems to be feasible through the adjustment of the differences between the three heating zones and their absolute values. In order to attempt extending the lifespan of the ceiling a model could be trained for prediction of the optimal settings for each heating zone. Additionally the predictive maintenance seem to have potential to be extended to further parts and actions.

9. Conclusion

The described data analysis in thermal processes was based on the data set consisting of sensor data from the equipment, product quality data described through measurements, maintenance records extended through stock disposal and a diverse set of process parameters.

The data set had been collected in the company specific distributed infrastructure. Therefore the data had to be accessed and retrieved, from several platforms. Further the records were cleansed, mapped and pre-processed in order to train the machine learning algorithms. To accomplish these steps various tools were necessary, within this thesis a selection of tools had been compared. The selected tools covering different areas such as scripting languages, graphical programming and visualization techniques. The comparison was based on attributes such as license, transparency, scalability and needed skill set for the particular software. The all-in-one tool which fits all the needs as there are so many differentiations in the data set, in the use-case, in the environment, etc. does not exist. There are multipurpose tools which cover a variety of functions and applications, but it should always be selected for the precise use-case.

Within the particular production environment this kind of analysis has not been done before. This fact led to the focus whether it is possible at all to train machine learning algorithms with the available data.

Is it possible to represent the state of the equipment as it is based on historical process, product and maintenance data?
The analysis has shown that the state of the investigated equipment, the rotary kilns, can be represented definitely with the available data sources. By choosing the appropriate data, various ML algorithms has been trained to represent and predict the actual state of particular parts of the equipment. The detail of the representation depend on the amount and availability of

the data, with further data sources, such as more sensor data, more detailed failure description, further setting records etc. the machine learning models could be extended to various predictions.

Is the quality of the collected data sufficient?
Within this thesis the data quality was challenging. Due to the lack of standardization of the records it was highly time consuming to extract the features. Further the missing data made it really hard to describe the actual happenings. That is the reason for the first step in the presented approach to predict the missing respectively wrong actions taken. The sensor data required up-sampling and interpolation due to strong varying frequency of the data points. Whereas this step led to inaccuracies, but could be done automatically and does not require manual adjustments. It further required some significant effort to to map the data to other available data sources. Anyway with the right amount of preparation the quality of the data seems to be sufficient to train a model. For further analysis, and in order to generate a fully automated pipeline for data preparation, an improvement of the quality is of significance.

Does the state of the equipment have an impact on the energy consumption?
In order to relate the state of the equipment to the energy consumption, any kind of energy measurements would be required. As the available actuating variable represented in percentage refers to undocumented settings, they are not suited for comparison. They are not absolute values and could therefore only be used to describe the change within one setting interval. Based on the available data and the calculated integrals representing the power consumption, it seems that the age of the equipment has an impact on the energy needed.

How to discover hidden knowledge from historical production data to generate an optimizing strategy?
When training different machine learning models the most important features, according to their influence on the accuracy of the model, had been identified. In order to assure the plausibility of those features, they had been discussed with several domain experts. It seems that the target temperatures in the different heating zones, as well as their differences to each other, have an impact on the lifespan of the investigated equipment. This discovered

factor could be used to optimize the life-cyle of the equipment by restricting the settings accordingly.

Does the insight of the data lead to a proposal for optimizing the overall equipment efficiency?
By extending the lifespan of the inspected equipment through optimization of the production parameter, the OEE could be improved as well. Based on the knowledge that the state of the equipment can be represented with according features there is huge potential for optimization to exploit. The trained model is suitable for predictive maintenance and could be used to trigger a message in the maintenance software. This message would inform the maintenance unit to visually inspect the according kiln. The scenario could help to decrease the amount of visual inspections. The predictive maintenance is not limited to one part of the equipment. In this thesis the ceiling exchange was particularly described. Further the prediction could be extended to the break of a heater, the lining, the exchange of different thermal elements etc. Besides predictive maintenance also the settings for the target setting measurement (referenced as 3P measurement), where the temperature levels are set, might be predicted. This could decrease the effort to adjust the equipment substantially and would further decrease the resources required without influencing the productivity efficiency. The impact on the quality of the product is very unlikely, as long as it is assured that the overall temperature equals the defined value.

There is not only one appropriate way to model a data analysis process, there are numerous ways. The data analysis in this thesis had been designed to fit the particular needs within the production environment and the generated data. The way to design the process was not straight forward as there are a variety of principles, methods and strategies to chose from. Initial premise might turn out as incorrect or were adapted during the process, in order to adjust to these circumstances one has to be resilient.

Appendix

Appendix A.

Features

A.1. List of All Extracted Features

1 index

2 run

3 duration

4 duration_h

5 from

6 furnace

7 s1_deviation_temp_innen

8 s1_deviation_temp_oben

9 s1_deviation_temp_unten

10 s1_duration

11 s1_gas_quantity

12 s1_integral_hzg_innen

13 s1_integral_hzg_oben

14 s1_integral_hzg_unten

15 s1_luft_ist_mean

111 s4_temp_ist_unten_mean

112 s4_temp_ist_unten_min

113 s4_temp_soll_innen_grad_mean

114 s4_temp_soll_innen_mean

115 s4_temp_soll_oben_grad_mean

116 s4_temp_soll_oben_mean

117 s4_temp_soll_unten_grad_mean

118 s4_temp_soll_unten_mean

119 s5_deviation_temp_innen

120 s5_deviation_temp_oben

121 s5_deviation_temp_unten

122 s5_duration

123 s5_gas_quantity

124 s5_integral_hzg_innen

125 s5_integral_hzg_oben

16	s1_luft_soll_mean
17	s1_temp_ist_innen_grad_mean
18	s1_temp_ist_innen_max
19	s1_temp_ist_innen_mean
20	s1_temp_ist_innen_min
21	s1_temp_ist_oben_grad_mean
22	s1_temp_ist_oben_max
23	s1_temp_ist_oben_mean
24	s1_temp_ist_oben_min
25	s1_temp_ist_unten_grad_mean
26	s1_temp_ist_unten_max
27	s1_temp_ist_unten_mean
28	s1_temp_ist_unten_min
29	s1_temp_soll_innen_grad_mean
30	s1_temp_soll_innen_mean
31	s1_temp_soll_oben_grad_mean
32	s1_temp_soll_oben_mean
33	s1_temp_soll_unten_grad_mean
34	s1_temp_soll_unten_mean
35	s2_deviation_temp_innen
36	s2_deviation_temp_oben
37	s2_deviation_temp_unten
38	s2_duration
39	s2_gas_quantity
126	s5_integral_hzg_unten
127	s5_luft_ist_mean
128	s5_luft_soll_mean
129	s5_temp_ist_innen_grad_mean
130	s5_temp_ist_innen_max
131	s5_temp_ist_innen_mean
132	s5_temp_ist_innen_min
133	s5_temp_ist_oben_grad_mean
134	s5_temp_ist_oben_max
135	s5_temp_ist_oben_mean
136	s5_temp_ist_oben_min
137	s5_temp_ist_unten_grad_mean
138	s5_temp_ist_unten_max
139	s5_temp_ist_unten_mean
140	s5_temp_ist_unten_min
141	s5_temp_soll_innen_grad_mean
142	s5_temp_soll_innen_mean
143	s5_temp_soll_oben_grad_mean
144	s5_temp_soll_oben_mean
145	s5_temp_soll_unten_grad_mean
146	s5_temp_soll_unten_mean
147	diff_diff
148	s1_diff_grad_oben
149	s1_diff_grad_unten

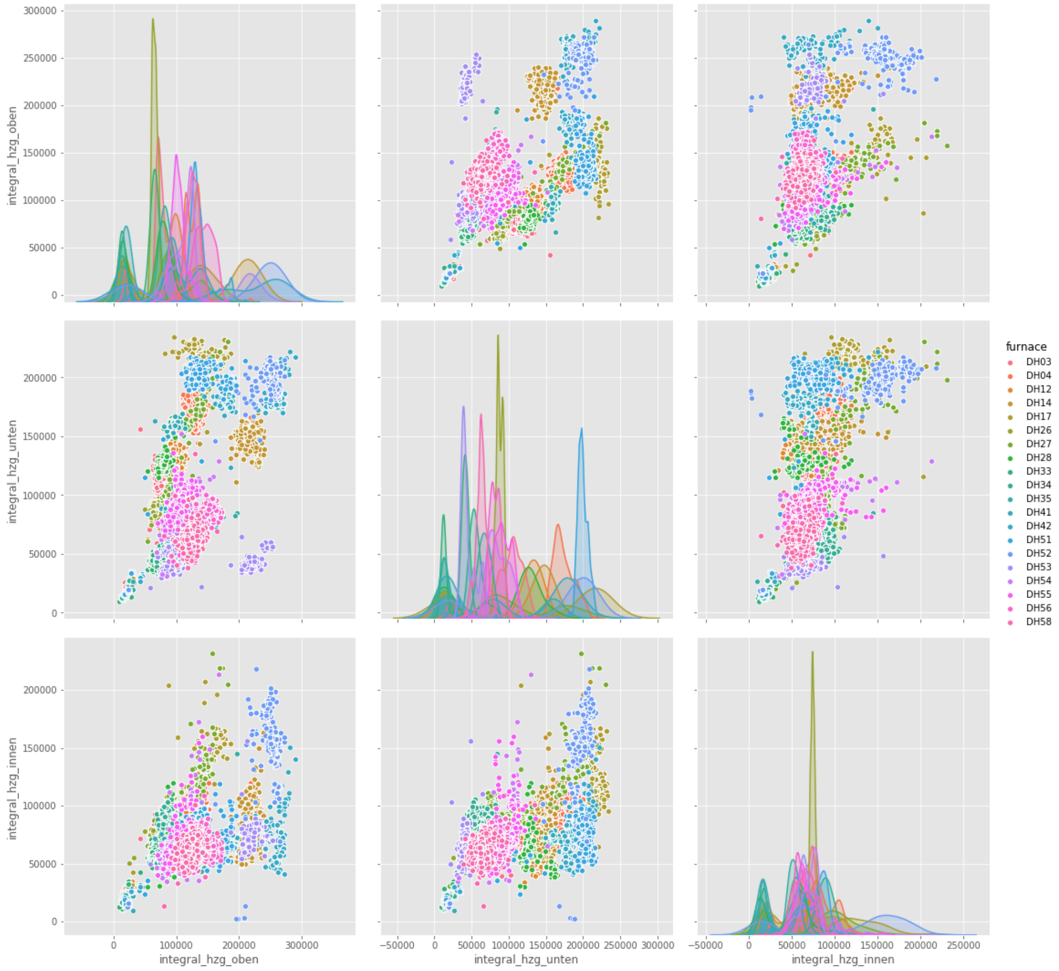
40	s2_integral_hzg_innen
41	s2_integral_hzg_oben
42	s2_integral_hzg_unten
43	s2_luft_ist_mean
44	s2_luft_soll_mean
45	s2_temp_ist_innen_grad_mean
46	s2_temp_ist_innen_max
47	s2_temp_ist_innen_mean
48	s2_temp_ist_innen_min
49	s2_temp_ist_oben_grad_mean
50	s2_temp_ist_oben_max
51	s2_temp_ist_oben_mean
52	s2_temp_ist_oben_min
53	s2_temp_ist_unten_grad_mean
54	s2_temp_ist_unten_max
55	s2_temp_ist_unten_mean
56	s2_temp_ist_unten_min
57	s2_temp_soll_innen_grad_mean
58	s2_temp_soll_innen_mean
59	s2_temp_soll_oben_grad_mean
60	s2_temp_soll_oben_mean
61	s2_temp_soll_unten_grad_mean
62	s2_temp_soll_unten_mean
63	s3_deviation_temp_innen
150	s1_diff_grad_innen
151	s3_diff_grad_oben
152	s3_diff_grad_unten
153	s3_diff_grad_innen
154	integral_hzg_oben
155	integral_hzg_unten
156	integral_hzg_innen
157	diff_integral_hzg_oben
158	diff_integral_hzg_unten
159	diff_integral_hzg_innen
160	deviation_temp_oben
161	deviation_temp_unten
162	deviation_temp_innen
163	temp_oben_mean
164	temp_unten_mean
165	temp_innen_mean
166	diff_temp_oben_innen
167	diff_temp_oben_unten
168	diff_temp_innen_unten
169	diff_oben_diff_unten
170	diff_innen_diff_unten
171	diff_temp_avg_oben_unten
172	s2_s3_integral_per_degree_oben
173	s2_s3_integral_per_degree_unten

64 s3_deviation_temp_oben
65 s3_deviation_temp_unten
66 s3_duration
67 s3_gas_quantity
68 s3_integral_hzg_innen
69 s3_integral_hzg_oben
70 s3_integral_hzg_unten
71 s3_luft_ist_mean
72 s3_luft_soll_mean
73 s3_temp_ist_innen_grad_mean
74 s3_temp_ist_innen_max
75 s3_temp_ist_innen_mean
76 s3_temp_ist_innen_min
77 s3_temp_ist_oben_grad_mean
78 s3_temp_ist_oben_max
79 s3_temp_ist_oben_mean
80 s3_temp_ist_oben_min
81 s3_temp_ist_unten_grad_mean
82 s3_temp_ist_unten_max
83 s3_temp_ist_unten_mean
84 s3_temp_ist_unten_min
85 s3_temp_soll_innen_grad_mean
86 s3_temp_soll_innen_mean
87 s3_temp_soll_oben_grad_mean
174 s2_s3_integral_per_degree_innen
175 s2_s3_integral_per_degree_oben_grad
176 s2_s3_integral_per_degree_unten_grad
177 s2_s3_integral_per_degree_innen_grad
178 s4_integral_per_degree_oben
179 s4_integral_per_degree_unten
180 s4_integral_per_degree_innen
181 s4_integral_per_degree_oben_grad
182 s4_integral_per_degree_unten_grad
183 s4_integral_per_degree_innen_grad
184 sum_intergal_per_degree_oben
185 sum_intergal_per_degree_unten
186 sum_intergal_per_degree_innen
187 weighted_sum_integral
188 weighted_sum_integral_per_degree
189 s2_s3_weighted_sum_integral
190 s4_weighted_sum_integral
191 sum_intergal_per_degree_oben_grad
192 sum_intergal_per_degree_unten_grad
193 sum_intergal_per_degree_innen_grad
194 s2_s3_weighted_sum_integral_grad
195 s4_weighted_sum_integral_grad
196 heat_stable_innen
197 heat_stable_oben

88	s3_temp_soll_oben_mean
89	s3_temp_soll_unten_grad_mean
90	s3_temp_soll_unten_mean
91	s4_deviation_temp_innen
92	s4_deviation_temp_oben
93	s4_deviation_temp_unten
94	s4_duration
95	s4_gas_quantity
96	s4_integral_hzg_innen
97	s4_integral_hzg_oben
98	s4_integral_hzg_unten
99	s4_luft_ist_mean
100	s4_luft_soll_mean
101	s4_temp_ist_innen_grad_mean
102	s4_temp_ist_innen_max
103	s4_temp_ist_innen_mean
104	s4_temp_ist_innen_min
105	s4_temp_ist_oben_grad_mean
106	s4_temp_ist_oben_max
107	s4_temp_ist_oben_mean
108	s4_temp_ist_oben_min
109	s4_temp_ist_unten_grad_mean
110	s4_temp_ist_unten_max
198	heat_stable_unten
199	valid
200	runs_since_ofendecke
201	runs_since_ausmauerung
202	pfropfen
203	bias
204	S1
205	S3
206	S2
207	temp_change
208	hzg_1
209	hzg_2
210	hzg_3
211	hzg_2_menge
212	hzg_1_menge
213	hzg_3_menge
214	temp_change_oben
215	temp_change_unten
216	temp_change_innen
217	runs_since_ofendecke_dis
218	outlier
219	description_outlier
220	top_temp

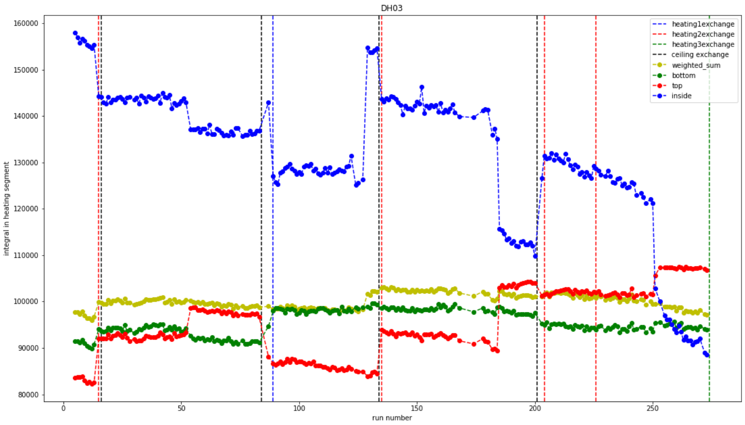
A.2. Graphic Representation of Feature: Integral of actuating variable

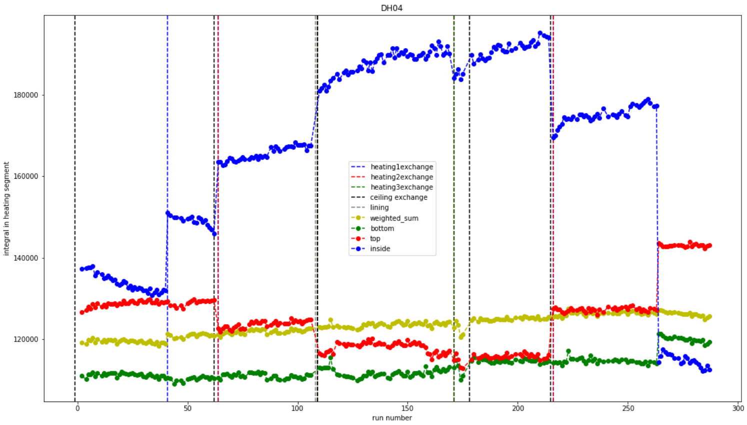
Each of the kilns got their individual heating profile. Depending on the settings and actual state of the kiln, they have a strong variety. In the following plot the three heating integrals and their relations towards the others are shown.

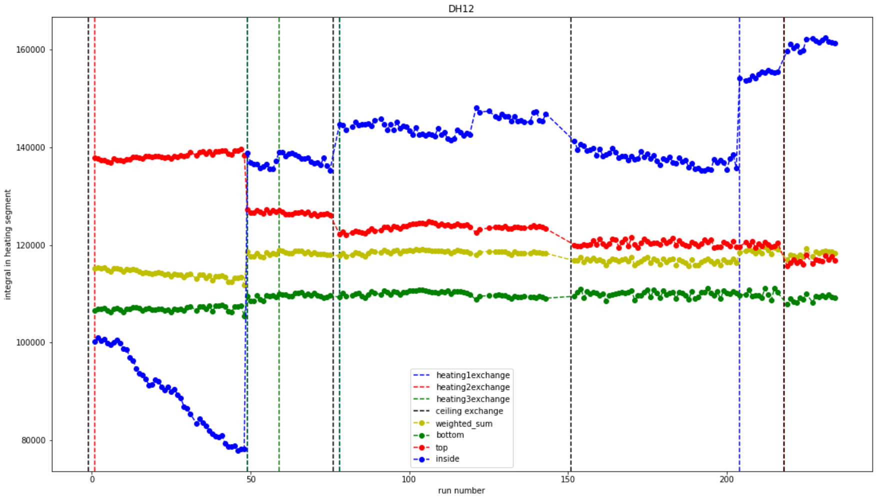


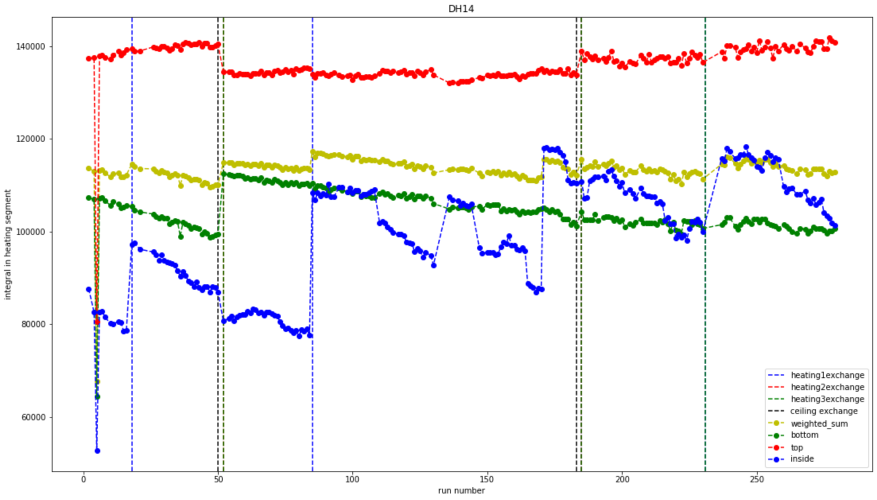
A.3. Graphic Representation Extracted Maintenance Actions

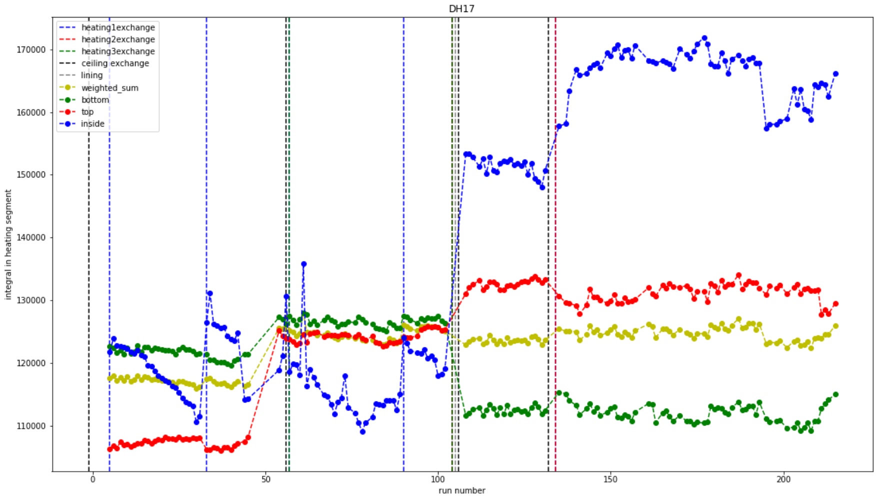
The following graphics show the integrals and their behavior. It can be observed that the change of the integrals with growing age of the ceiling is not consistent. Depending on the settings of the kiln there is a different behavior of the features over time.

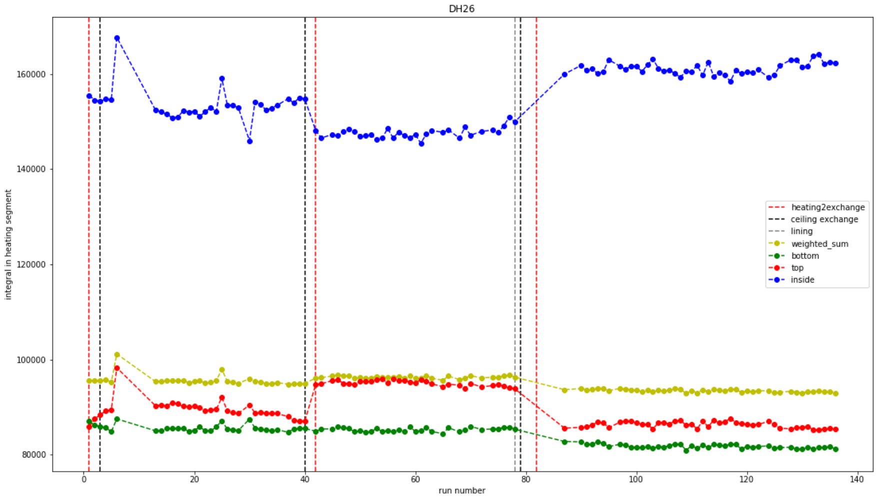


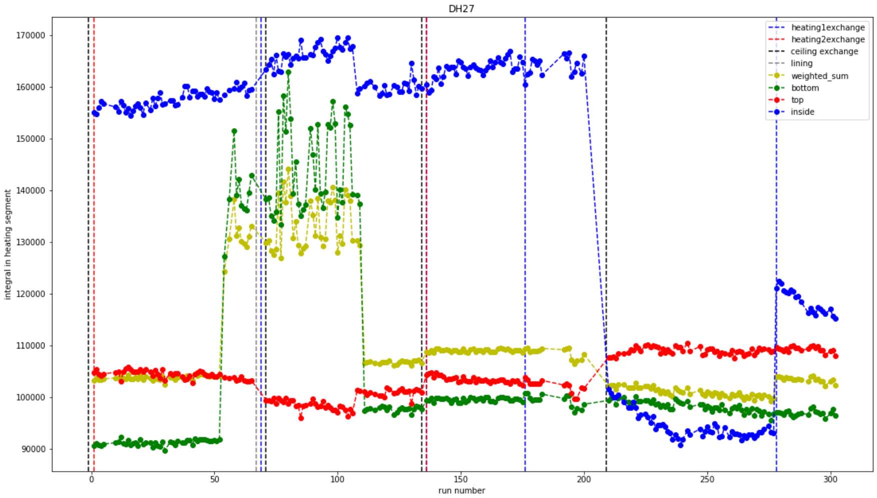


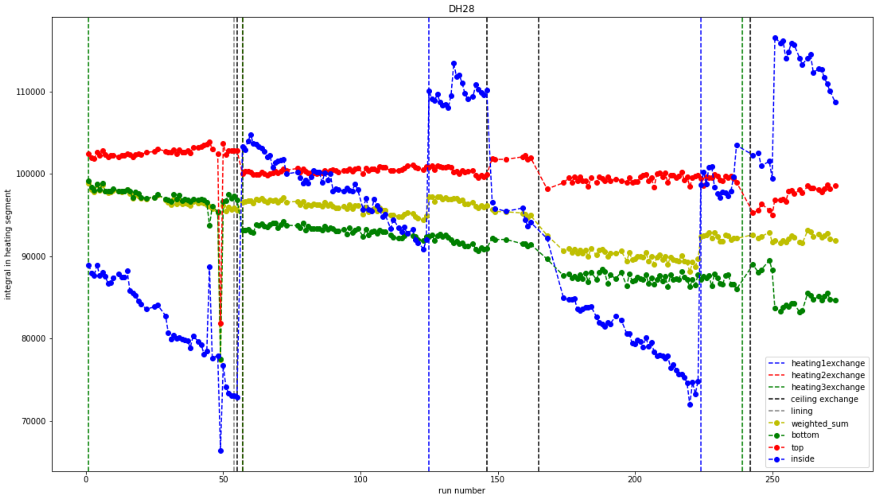

DH12
integral in heating segment
160000
140000
120000
100000
80000
0
50
100
150
200
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside

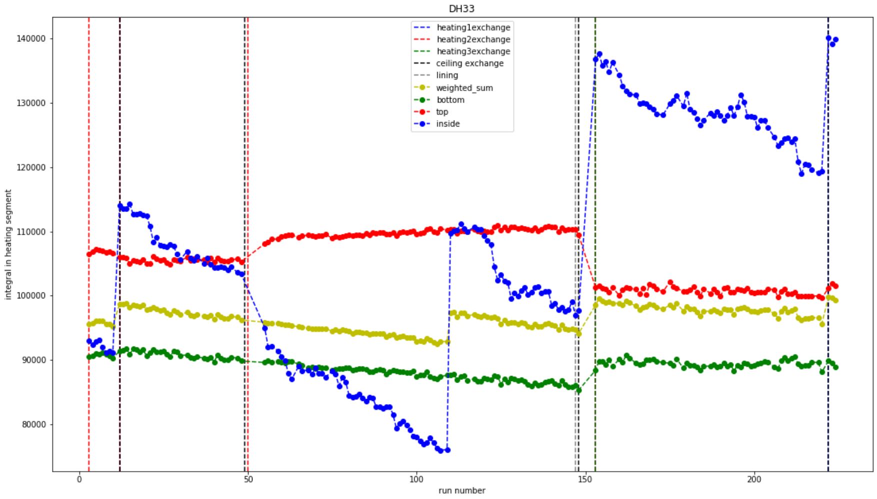

DH14
integral in heating segment
140000
120000
100000
80000
60000
0
50
100
150
200
250
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside

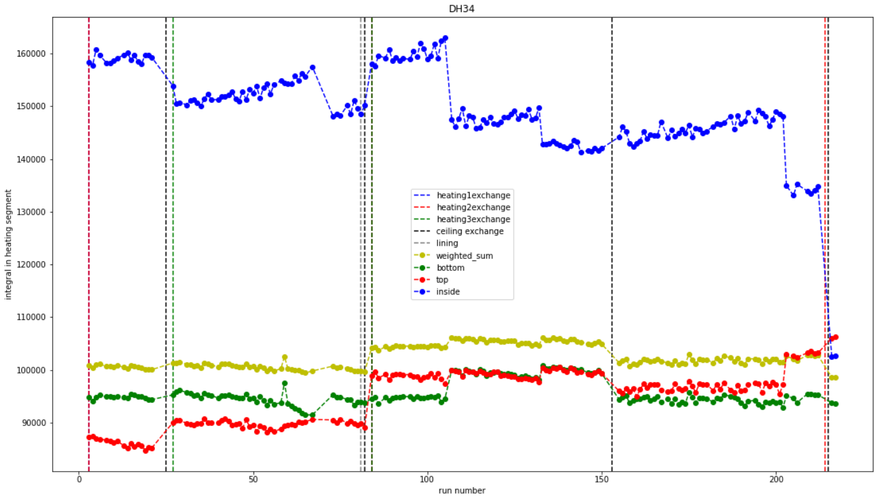

DH17
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside
integral in heating segment
170000
160000
150000
140000
130000
120000
110000
0
50
100
150
200
run number

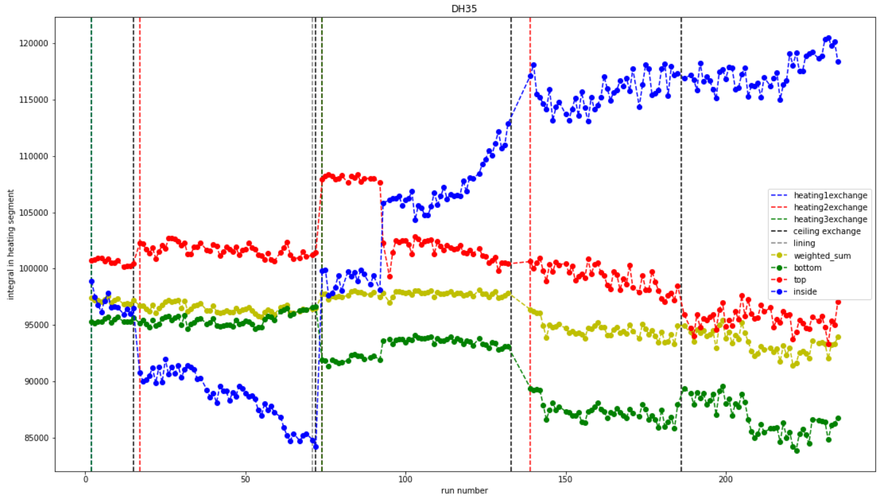

DH26
heating2exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside
integral in heating segment
160000
140000
120000
100000
80000
0
20
40
60
80
100
120
140
run number

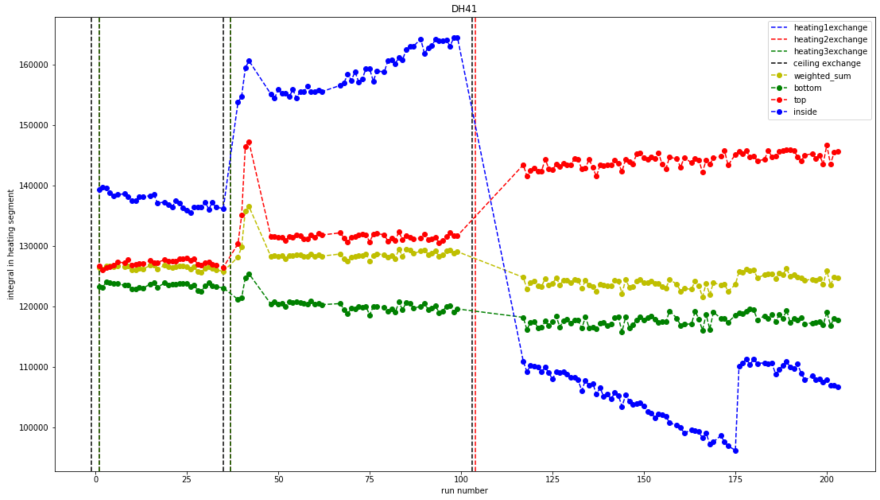

DH27
integral in heating segment
170000
160000
150000
140000
130000
120000
110000
100000
90000
0
50
100
150
200
250
300
run number
heating1exchange
heating2exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

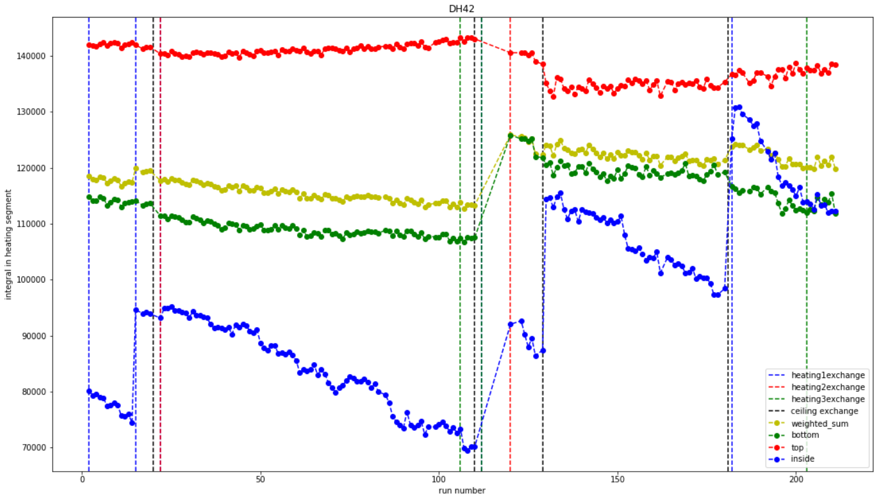

DH28
integral in heating segment
110000
100000
90000
80000
70000
0
50
100
150
200
250
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

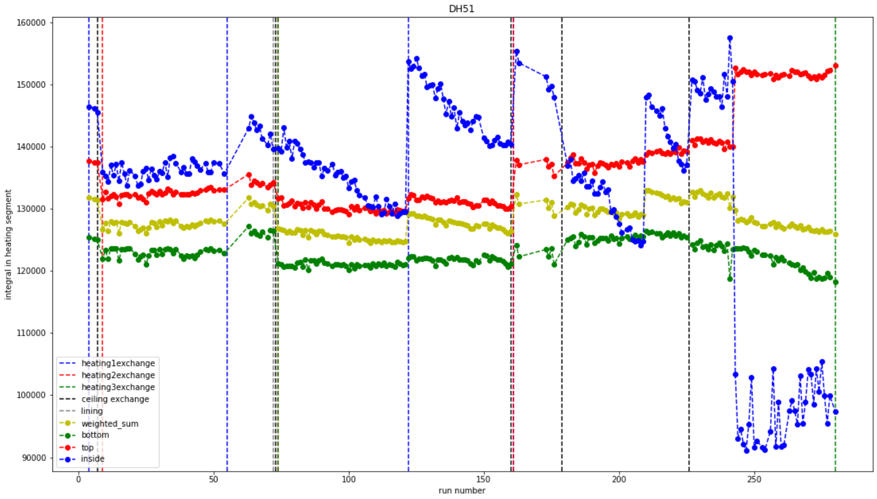

DH33
integral in heating segment
140000
130000
120000
110000
100000
90000
80000
0
50
100
150
200
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

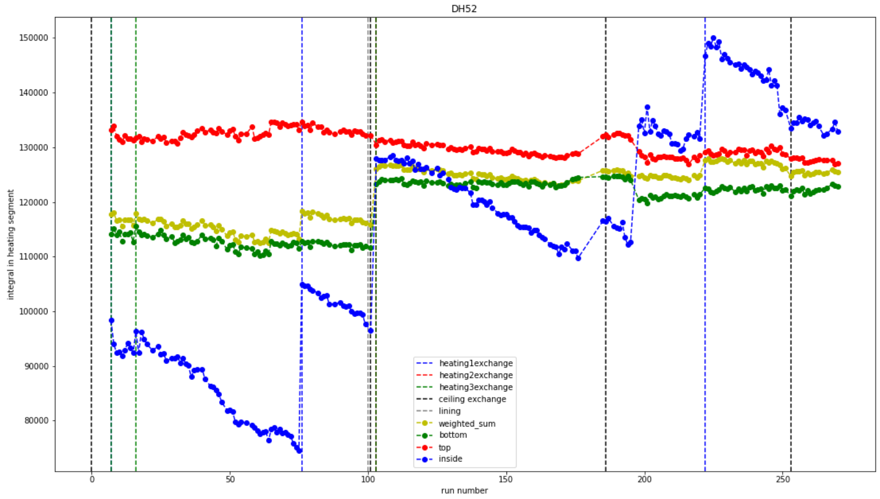

DH34
integral in heating segment
160000
150000
140000
130000
120000
110000
100000
90000
0
50
100
150
200
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

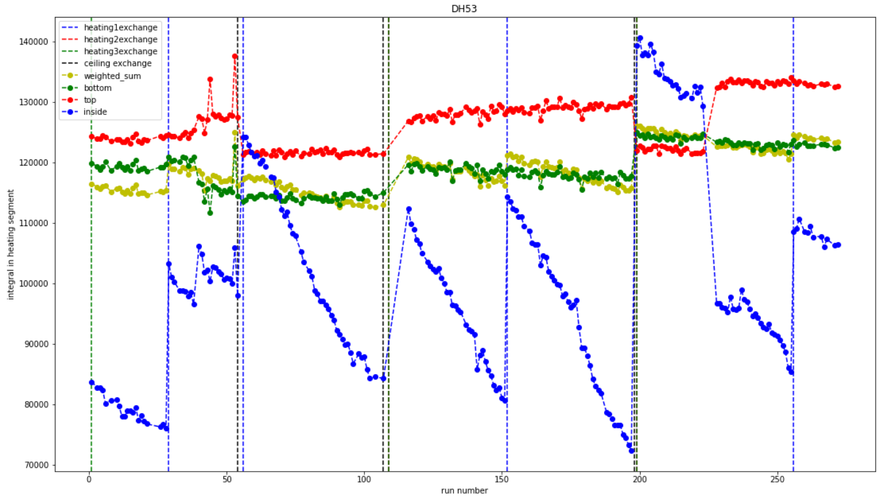

DH35
integral in heating segment
120000
115000
110000
105000
100000
95000
90000
85000
0
50
100
150
200
run number
heatinglexchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

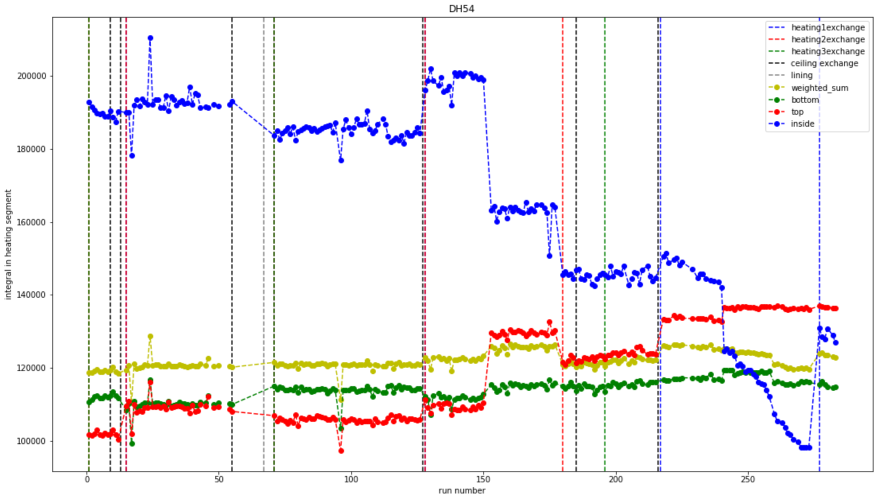

DH41
integral in heating segment
160000
150000
140000
130000
120000
110000
100000
0
25
50
75
100
125
150
175
200
run number
heatinglexchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside

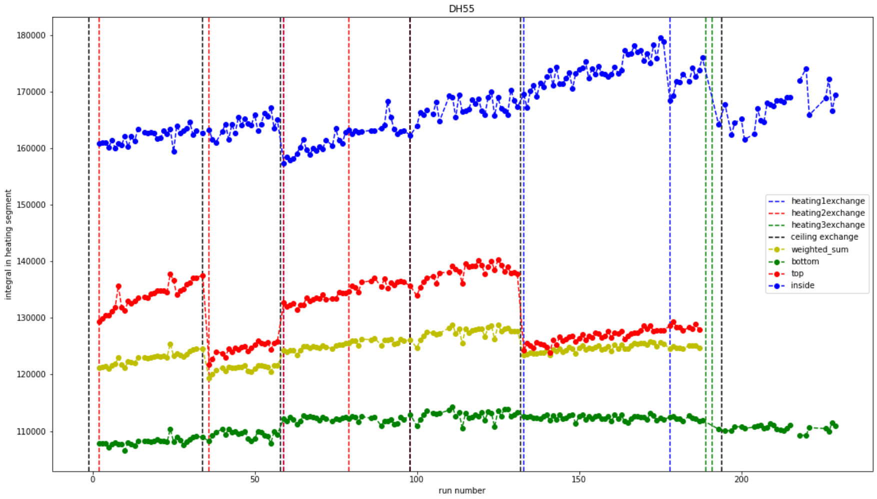

DH42
integral in heating segment
140000
130000
120000
110000
100000
90000
80000
70000
0
50
100
150
200
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside

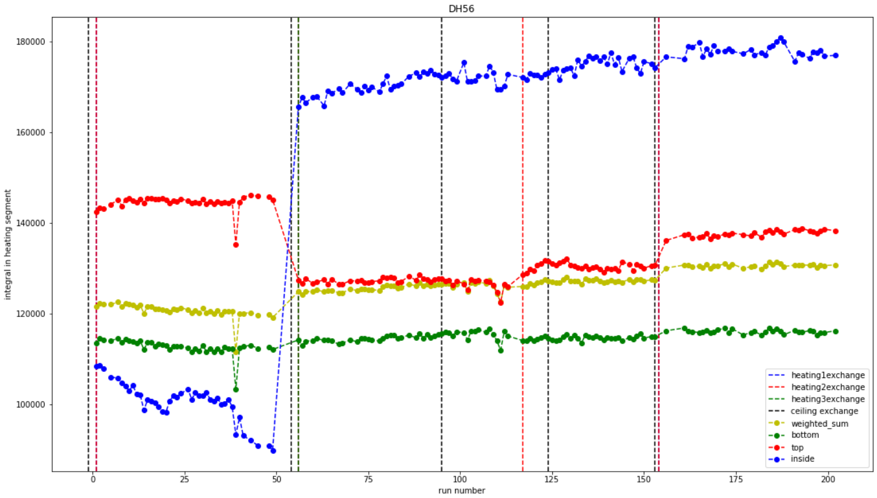

DH51
integral in heating segment
160000
150000
140000
130000
120000
110000
100000
90000
0
50
100
150
200
250
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

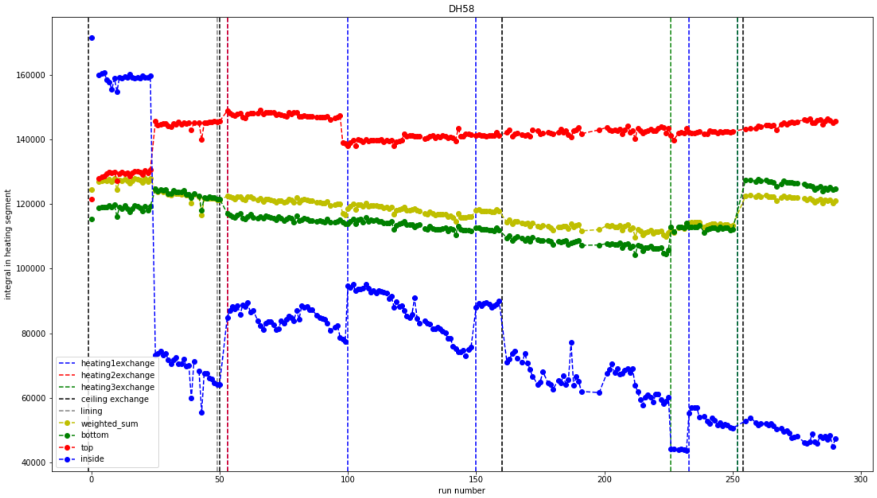

DH52
integral in heating segment
run number
150000
140000
130000
120000
110000
100000
90000
80000
0
50
100
150
200
250
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside


DH53
integral in heating segment
run number
140000
130000
120000
110000
100000
90000
80000
70000
0
50
100
150
200
250
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside


DH54
integral in heating segment
run number
200000
180000
160000
140000
120000
100000
0
50
100
150
200
250
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside


DH55
integral in heating segment
run number
180000
170000
160000
150000
140000
130000
120000
110000
0
50
100
150
200
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside


DH56
integral in heating segment
180000
160000
140000
120000
100000
0
25
50
75
100
125
150
175
200
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
weighted_sum
bottom
top
inside


DH58
integral in heating segment
160000
140000
120000
100000
80000
60000
40000
0
50
100
150
200
250
300
run number
heating1exchange
heating2exchange
heating3exchange
ceiling exchange
lining
weighted_sum
bottom
top
inside

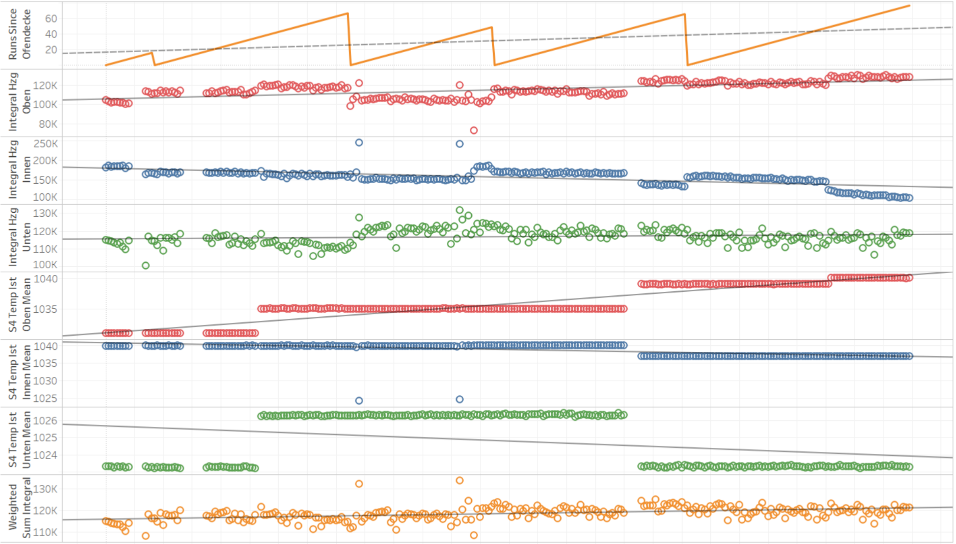
A.4. Feature Importance according to MDA

Example for the top rated features according to MDA. The age of the heaters as well as the lining of the ceiling had been rated as important. Further the difference of the temperatures in the three heating is rated high.

```
(0.8016 'hzg_2')                        (0.0937 'runs_since_ausmauerung')
(0.0665 'hzg_3')                        (0.0664 'run')
(0.0636 'hzg_1')                        (0.0515 'diff_diff')
(0.0511 'diff_innen_diff_unten')        (0.0495 'heat_stable_innen')
(0.0281 's4_integral_hzg_oben')         (0.0238 'diff_temp_innen_unten')
(0.0231 's4_integral_hzg_innen')        (0.0185 's2_integral_hzg_oben')
(0.0135 's4_integral_hzg_unten')        (0.0122 's4_temp_soll_innen_mean')
(0.0114 's2_integral_hzg_innen')        (0.0109 'furnace_')
(0.0108 's3_integral_hzg_unten')        (0.0104 'heat_stable_unten')
(0.0101 's4_temp_ist_oben_mean')        (0.0095 'diff_temp_oben_innen')
(0.0092 's4_temp_soll_oben_mean')       (0.0088 's3_integral_hzg_oben')
(0.0079 'diff_temp_avg_oben_unten')     (0.0078 'S1')
(0.0076 'heat_stable_oben')             (0.007 's1_integral_hzg_innen')
(0.0064 'diff_oben_diff_unten')         (0.006 's3_integral_hzg_innen')
(0.0057 'diff_temp_oben_unten')         (0.0046 's2_integral_hzg_unten')
(0.0027 's1_integral_hzg_oben')         (0.0026 's1_integral_hzg_unten')
(0.0026 'S2')                           (0.0018 'duration_h')
(0.0014 'S3')                           (0.0006 's1_temp_ist_oben_min')
(0.0001 'top_temp')
```

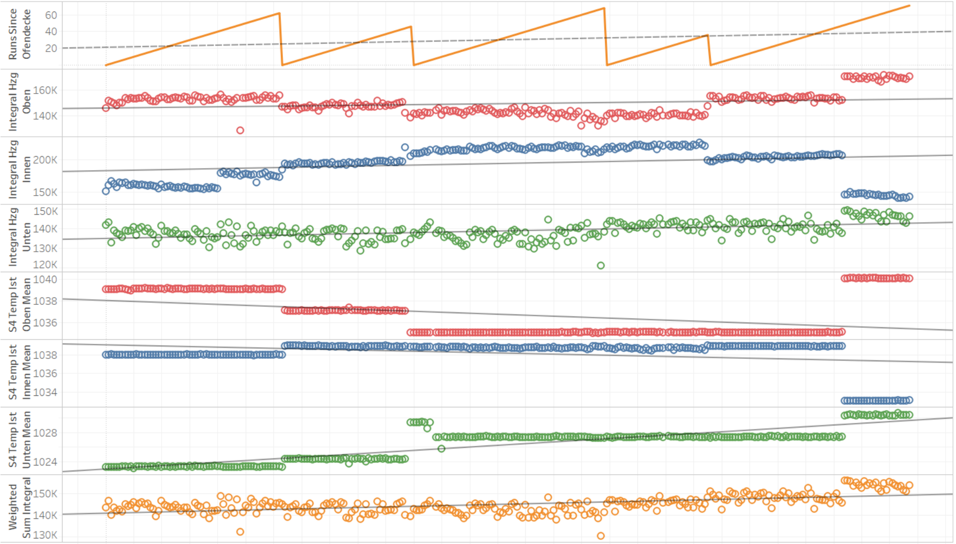
A.5. Temperature Integrals and Ceiling - Timeline

DH03 Temperature Integral Ceiling



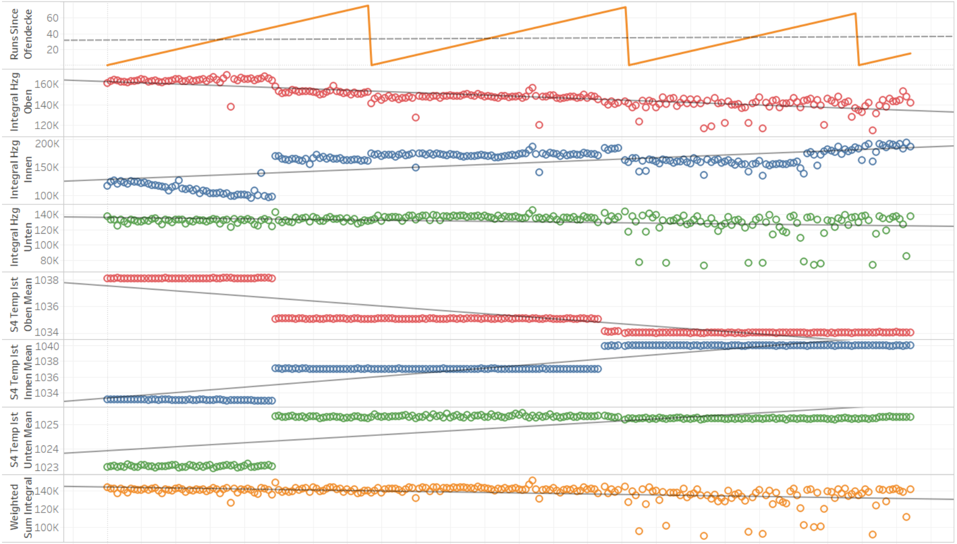
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH03. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), which keeps 4,855 members.

DH04 Temperature Integral Ceiling



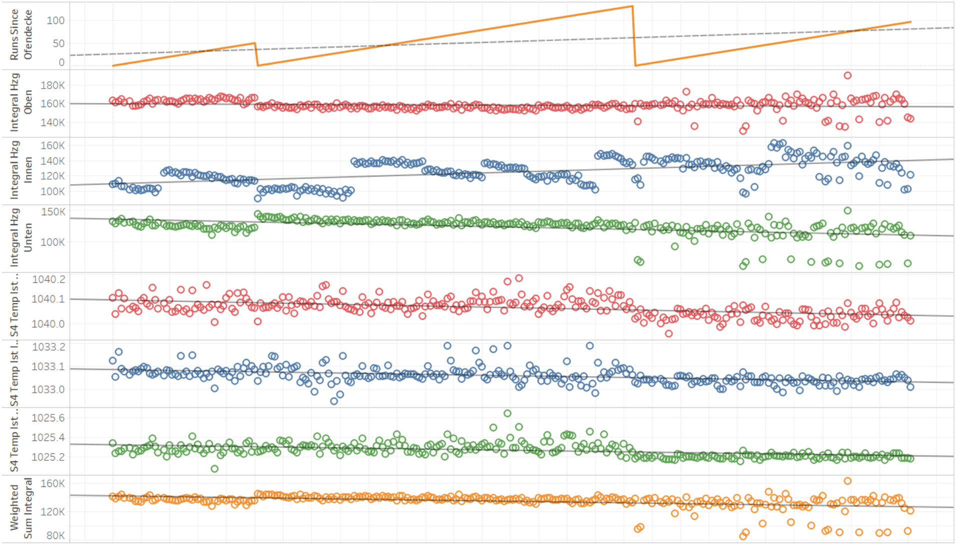
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH04. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), which keeps 4,855 members.

DH12 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH12. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), which keeps 4,855 members.

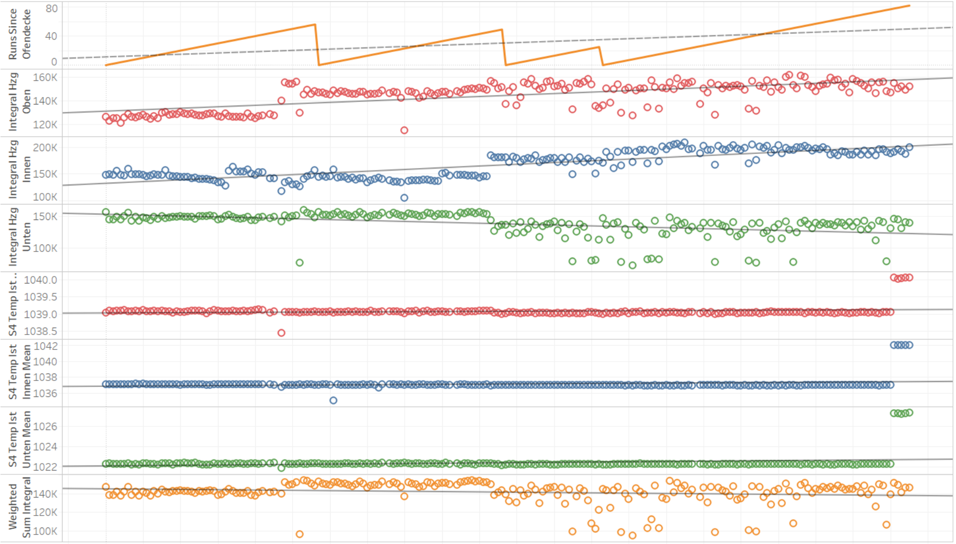
DH14 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH14. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run) and Exclusions (Run,S4 Temp Ist Unten Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members.

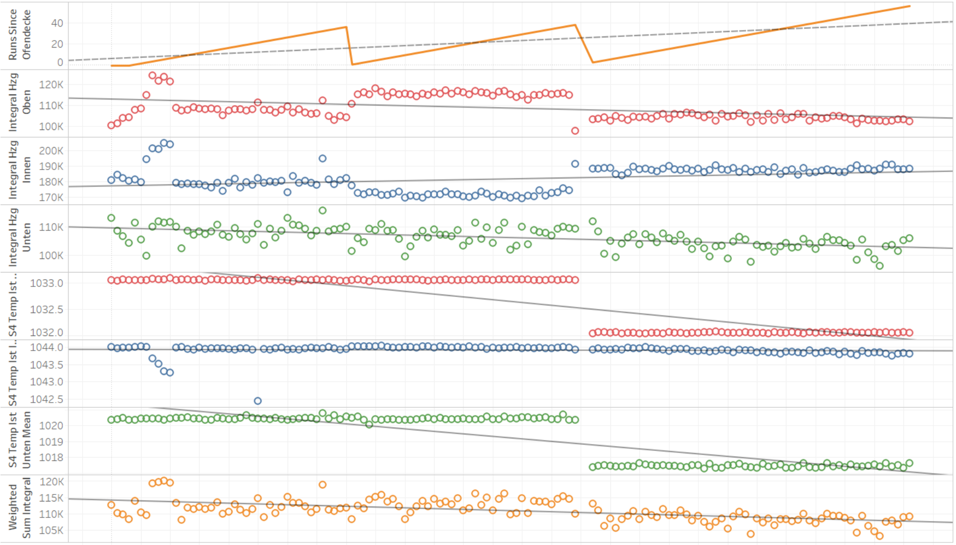
Appendix A. Features

DH17 Temperature Integral Ceiling



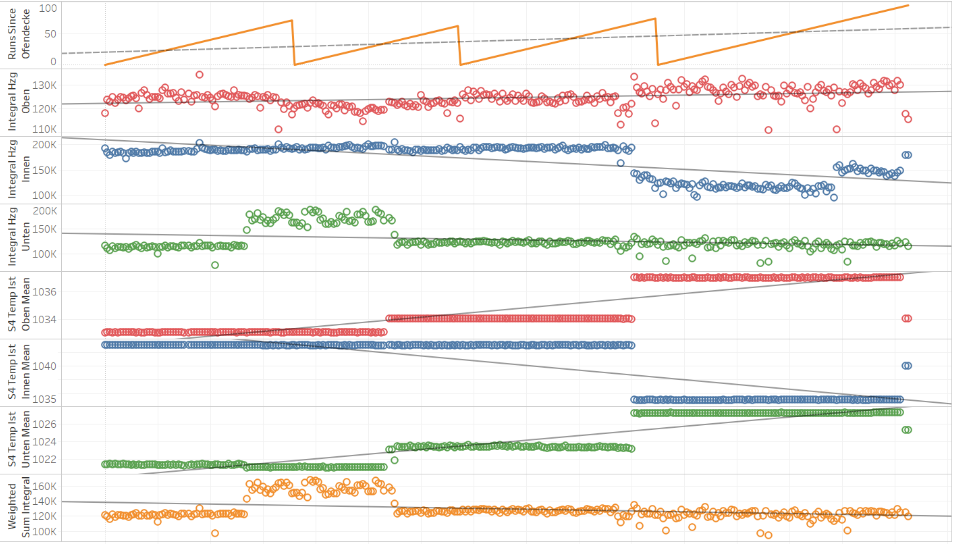
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH17. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH26 Temperature Integral Ceiling



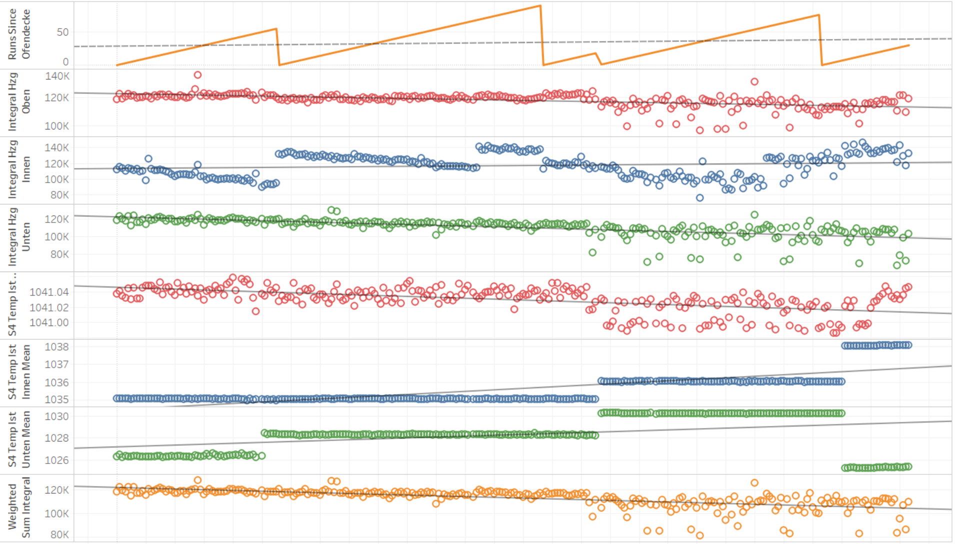
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH26. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH27 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH27. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

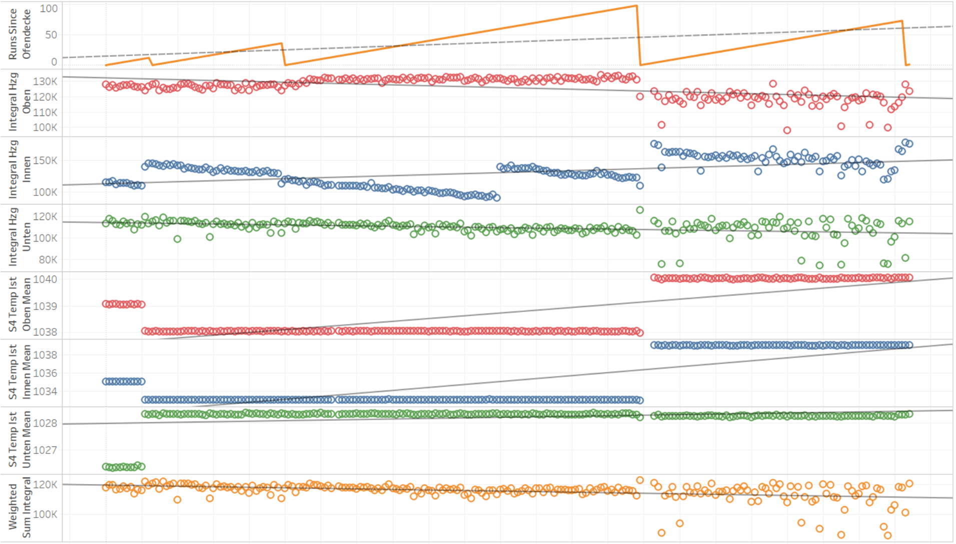
DH28 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH28. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

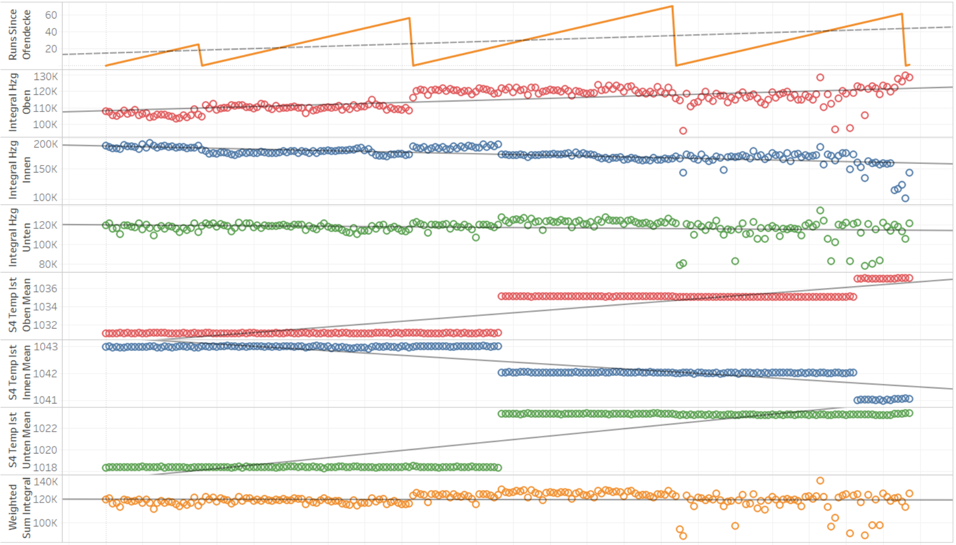
Appendix A. Features

DH33 Temperature Integral Ceiling



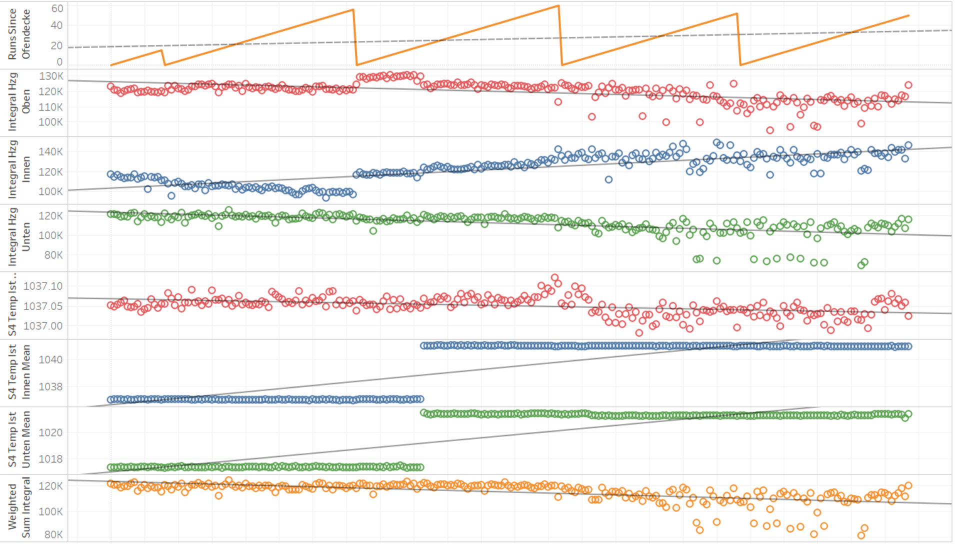
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH33. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH34 Temperature Integral Ceiling



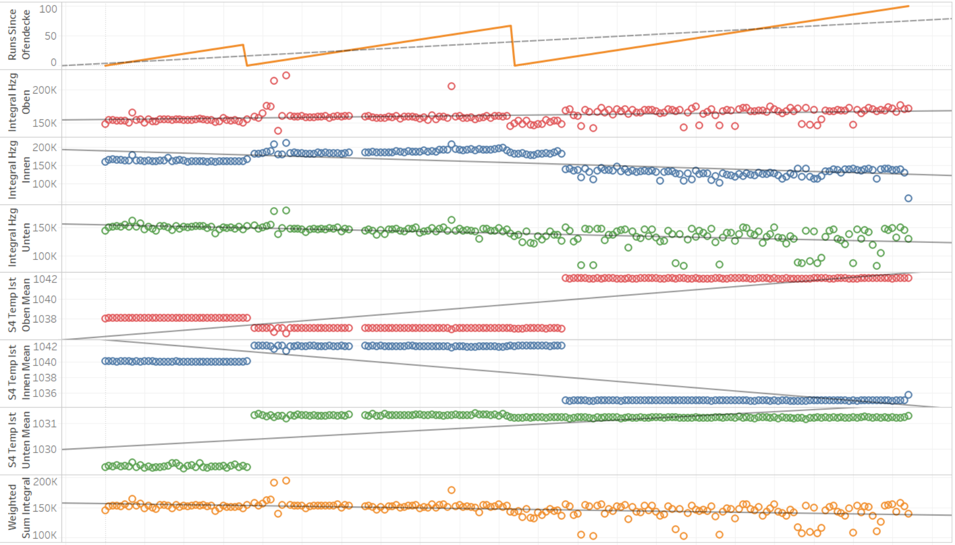
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH34. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH35 Temperature Integral Ceiling



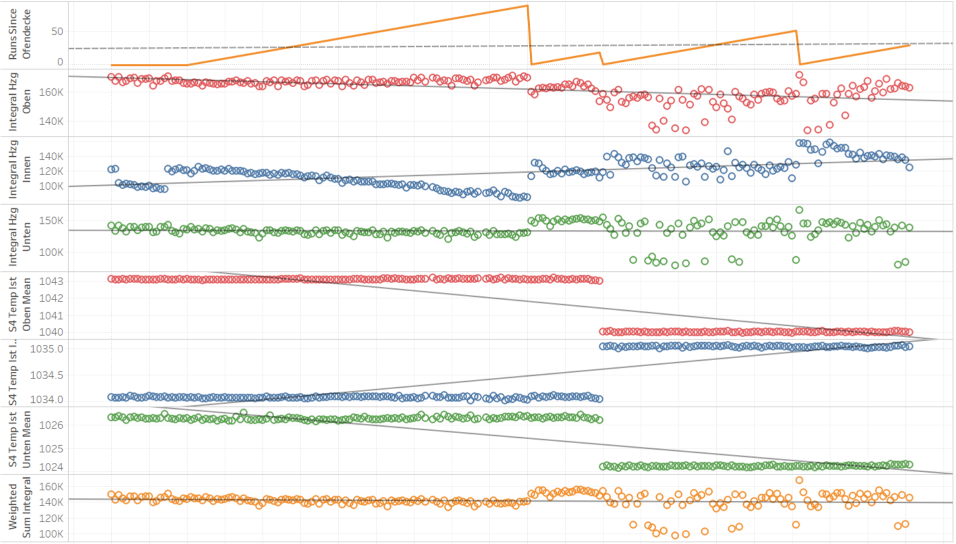
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH35. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH41 Temperature Integral Ceiling



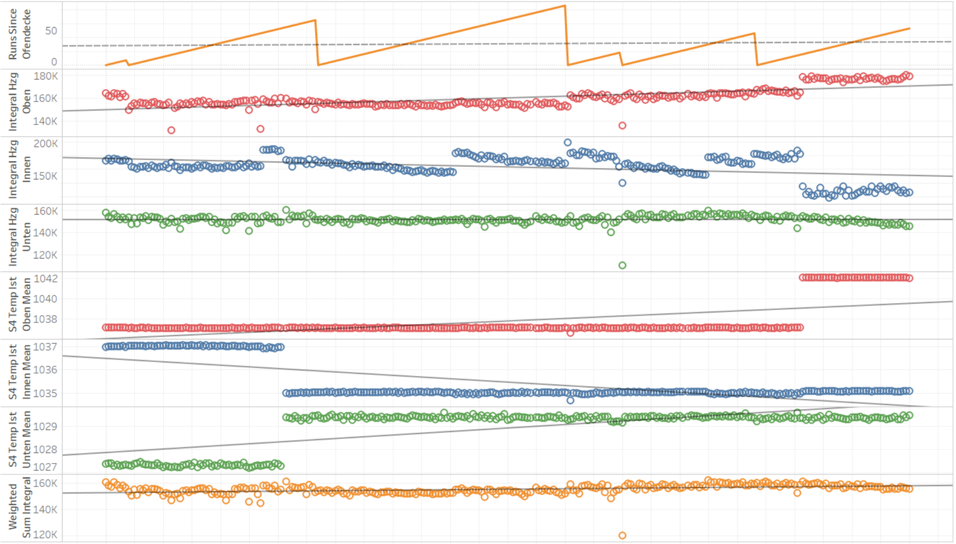
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH41. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH42 Temperature Integral Ceiling



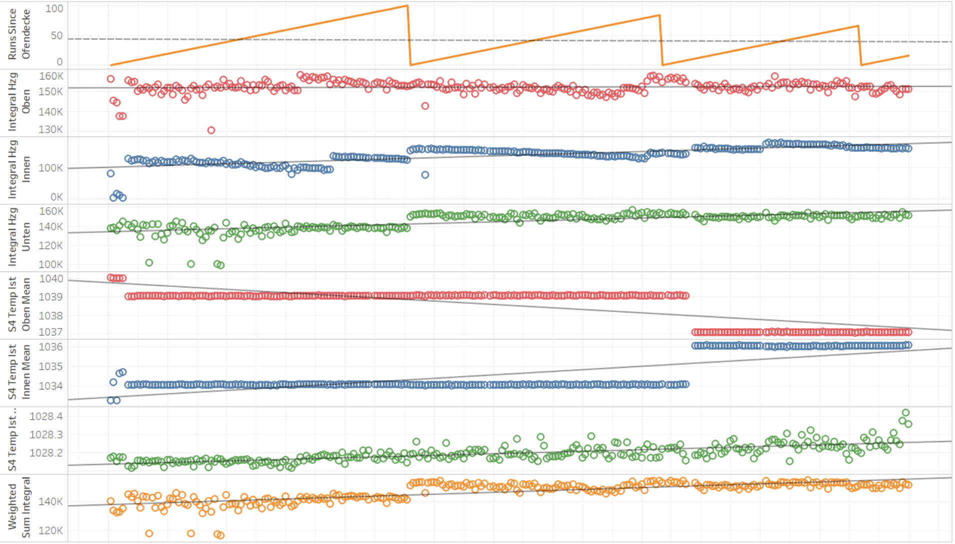
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH42. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH51 Temperature Integral Ceiling



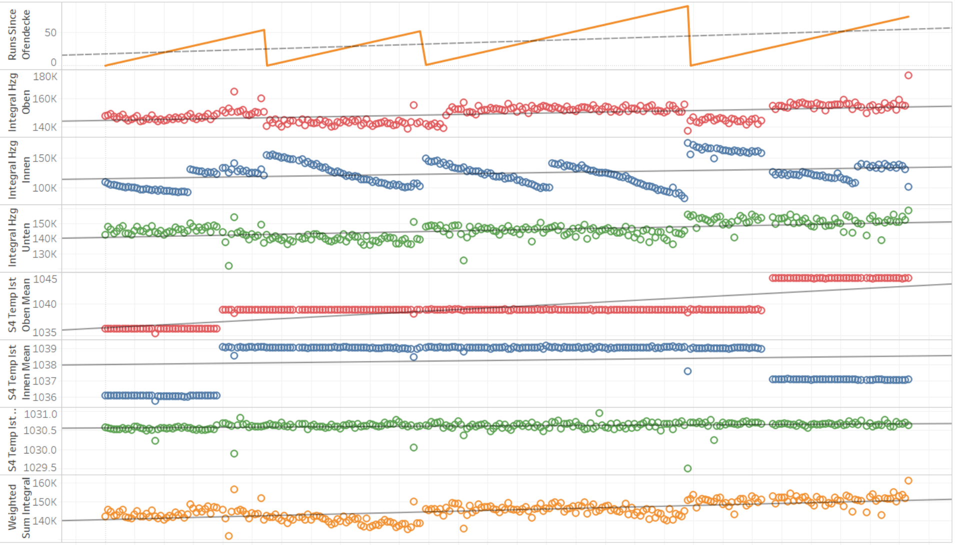
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH51. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,857 members.

DH52 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH52. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,856 members.

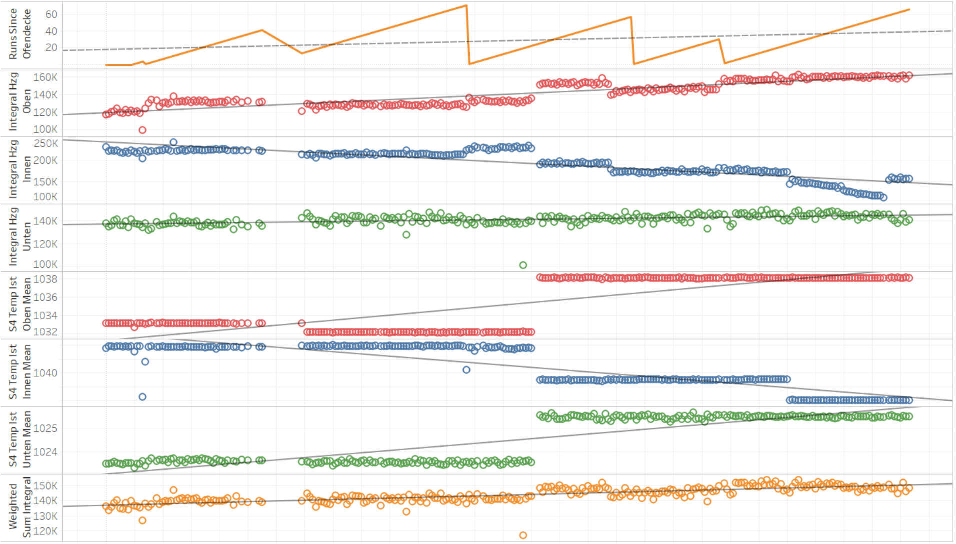
DH53 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH53. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,856 members.

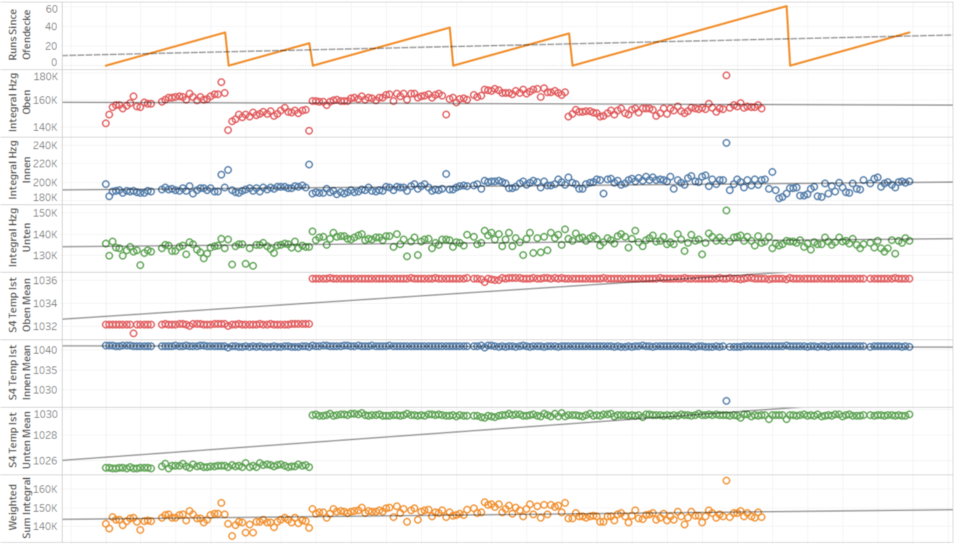
Appendix A. Features

DH54 Temperature Integral Ceiling



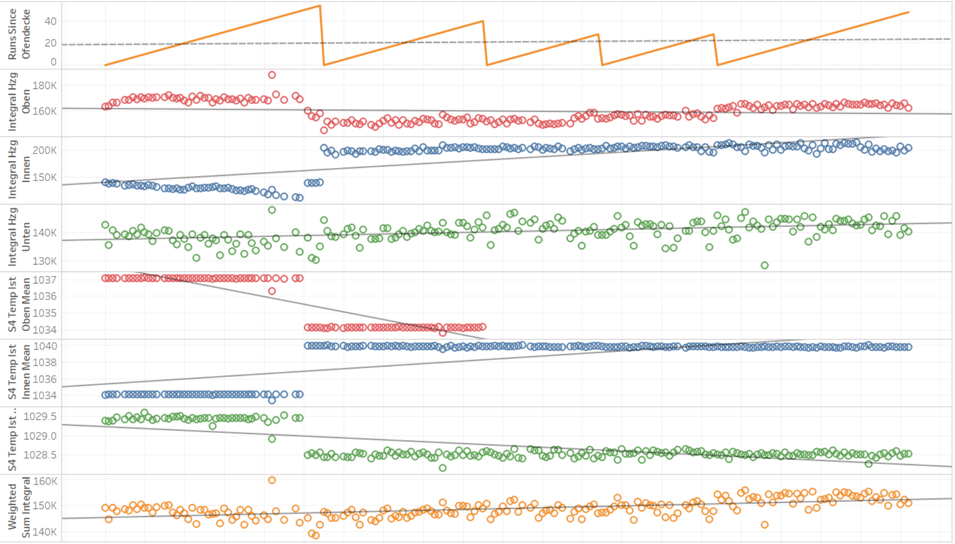
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH54. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,856 members.

DH55 Temperature Integral Ceiling



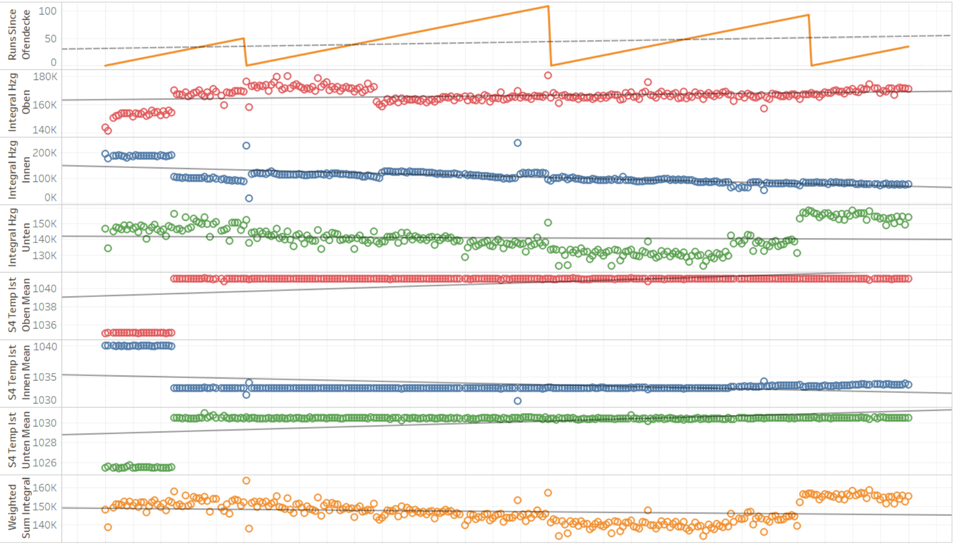
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH55. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,856 members.

DH56 Temperature Integral Ceiling



Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH56. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,857 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,856 members.

DH58 Temperature Integral Ceiling



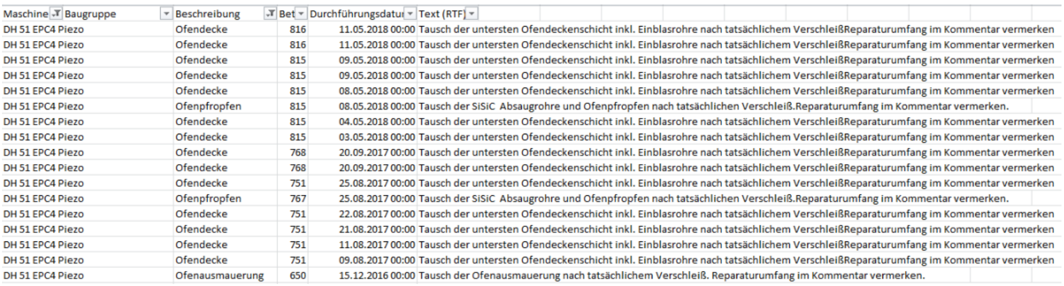
Run vs. Runs Since Ofendecke, Integral Hzg Oben, Integral Hzg Innen, Integral Hzg Unten, S4 Temp Ist Oben Mean, S4 Temp Ist Innen Mean, S4 Temp Ist Unten Mean and Weighted Sum Integral. The data is filtered on Furnace, Exclusions (Run,Sum Intergal Per Degree Unten Grad) and Exclusions (Run,S5 Temp Soll Oben Grad Mean). The Furnace filter keeps DH58. The Exclusions (Run,Sum Intergal Per Degree Unten Grad) filter keeps 4,843 members. The Exclusions (Run,S5 Temp Soll Oben Grad Mean) filter keeps 4,845 members. The view is filtered on Exclusions (Integral Hzg Unten,Run), Exclusions (Run,S4 Temp Ist Unten Mean), Exclusions (Run,S4 Temp Ist Oben Mean) and Exclusions (Run,S4 Temp Ist Innen Mean). The Exclusions (Integral Hzg Unten,Run) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Unten Mean) filter keeps 4,855 members. The Exclusions (Run,S4 Temp Ist Oben Mean) filter keeps 4,856 members. The Exclusions (Run,S4 Temp Ist Innen Mean) filter keeps 4,856 members.

Appendix B.

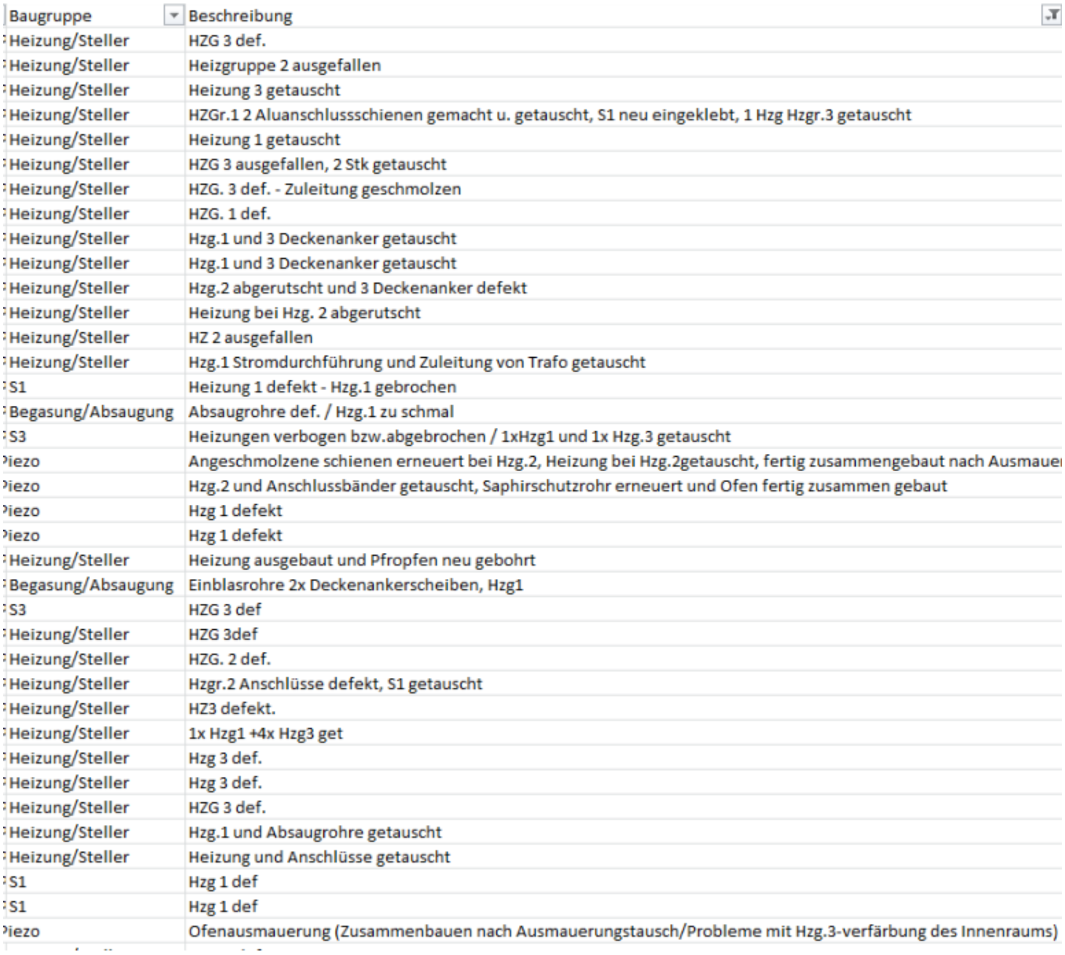
Additional Graphics

B.1. Example of Data Quality

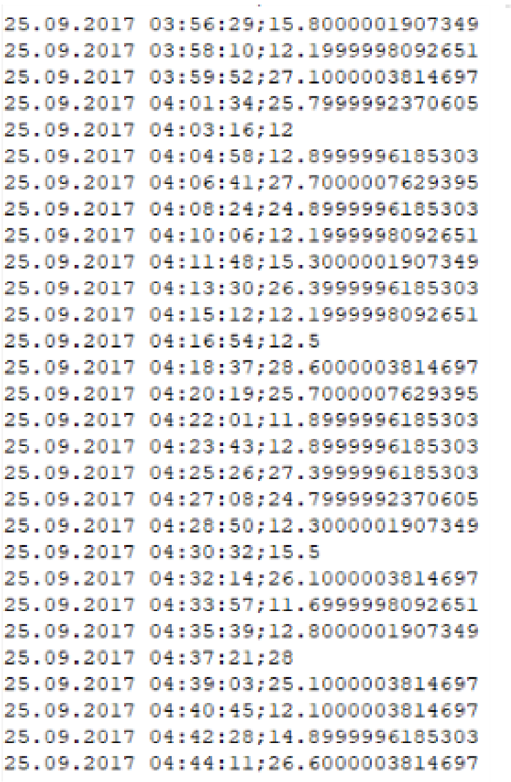
Her some examples from the maintenance records are listed. Particularly the difference in wording for exchange of the heater. Further some entries of the sensor values are shown to illustrate the varying frequency.



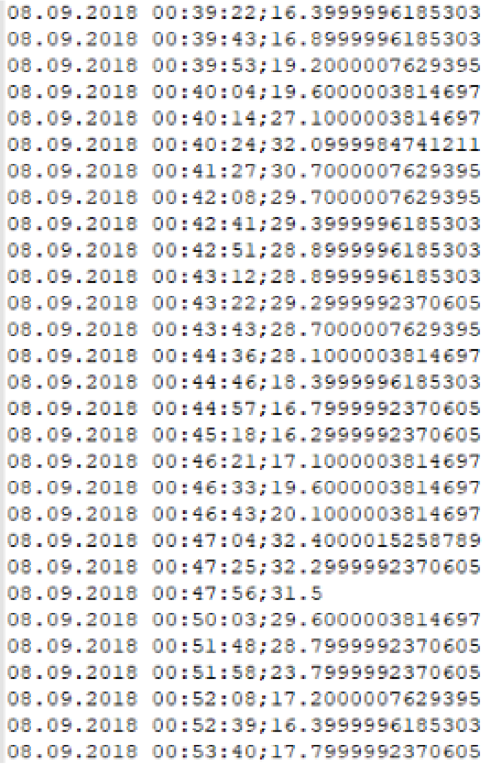
Maschine	Baugruppe	Beschreibung	Bet	Durchführungsdatu	Text (RTF)
DH 51 EPC4 Piezo		Ofendecke	816	11.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	816	11.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	815	09.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	815	09.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	815	08.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofenpfropfen	815	08.05.2018 00:00	Tausch der SiSiC Absaugrohre und Ofenpfropfen nach tatsächlichen Verschleiß.Reparaturumfang im Kommentar vermerken.
DH 51 EPC4 Piezo		Ofendecke	815	04.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	815	03.05.2018 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	768	20.09.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	768	20.09.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	751	25.08.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofenpfropfen	767	25.08.2017 00:00	Tausch der SiSiC Absaugrohre und Ofenpfropfen nach tatsächlichen Verschleiß.Reparaturumfang im Kommentar vermerken.
DH 51 EPC4 Piezo		Ofendecke	751	22.08.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	751	21.08.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	751	11.08.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofendecke	751	09.08.2017 00:00	Tausch der untersten Ofendeckenschicht inkl. Einblasrohre nach tatsächlichem VerschleißReparaturumfang im Kommentar vermerken
DH 51 EPC4 Piezo		Ofenausmauerung	650	15.12.2016 00:00	Tausch der Ofenausmauerung nach tatsächlichem Verschleiß. Reparaturumfang im Kommentar vermerken.



Baugruppe	Beschreibung
Heizung/Steller	HZG 3 def.
Heizung/Steller	Heizgruppe 2 ausgefallen
Heizung/Steller	Heizung 3 getauscht
Heizung/Steller	HZGr.1 2 Aluanschlussschienen gemacht u. getauscht, S1 neu eingeklebt, 1 Hzg Hzgr.3 getauscht
Heizung/Steller	Heizung 1 getauscht
Heizung/Steller	HZG 3 ausgefallen, 2 Stk getauscht
Heizung/Steller	HZG. 3 def. - Zuleitung geschmolzen
Heizung/Steller	HZG. 1 def.
Heizung/Steller	Hzg.1 und 3 Deckenanker getauscht
Heizung/Steller	Hzg.1 und 3 Deckenanker getauscht
Heizung/Steller	Hzg.2 abgerutscht und 3 Deckenanker defekt
Heizung/Steller	Heizung bei Hzg. 2 abgerutscht
Heizung/Steller	HZ 2 ausgefallen
Heizung/Steller	Hzg.1 Stromdurchführung und Zuleitung von Trafo getauscht
S1	Heizung 1 defekt - Hzg.1 gebrochen
Begasung/Absaugung	Absaugrohre def. / Hzg.1 zu schmal
S3	Heizungen verbogen bzw.abgebrochen / 1xHzg1 und 1x Hzg.3 getauscht
iezo	Angeschmolzene schienen erneuert bei Hzg.2, Heizung bei Hzg.2getauscht, fertig zusammengebaut nach Ausmaue
iezo	Hzg.2 und Anschlussbänder getauscht, Saphirschutzrohr erneuert und Ofen fertig zusammen gebaut
iezo	Hzg 1 defekt
iezo	Hzg 1 defekt
Heizung/Steller	Heizung ausgebaut und Pfropfen neu gebohrt
Begasung/Absaugung	Einblasrohre 2x Deckenankerscheiben, Hzg1
S3	HZG 3 def
Heizung/Steller	HZG 3def
Heizung/Steller	HZG. 2 def.
Heizung/Steller	Hzgr.2 Anschlüsse defekt, S1 getauscht
Heizung/Steller	HZ3 defekt.
Heizung/Steller	1x Hzg1 +4x Hzg3 get
Heizung/Steller	Hzg 3 def.
Heizung/Steller	Hzg 3 def.
Heizung/Steller	HZG 3 def.
Heizung/Steller	Hzg.1 und Absaugrohre getauscht
Heizung/Steller	Heizung und Anschlüsse getauscht
S1	Hzg 1 def
S1	Hzg 1 def
iezo	Ofenausmauerung (Zusammenbauen nach Ausmauerungstausch/Probleme mit Hzg.3-verfärbung des Innenraums)



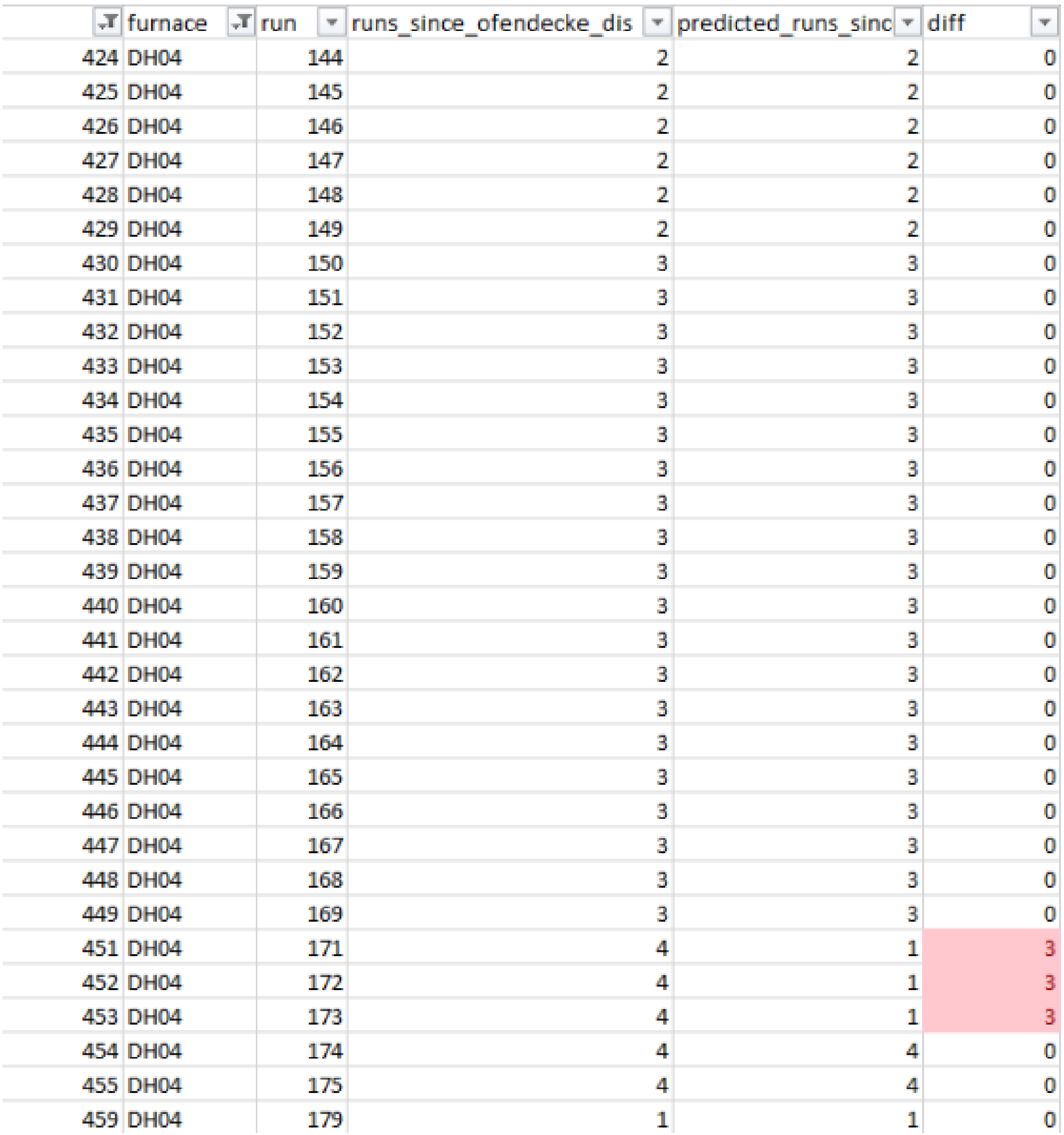
```
25.09.2017 03:56:29;15.8000001907349
25.09.2017 03:58:10;12.1999998092651
25.09.2017 03:59:52;27.1000003814697
25.09.2017 04:01:34;25.7999992370605
25.09.2017 04:03:16;12
25.09.2017 04:04:58;12.8999996185303
25.09.2017 04:06:41;27.7000007629395
25.09.2017 04:08:24;24.8999996185303
25.09.2017 04:10:06;12.1999998092651
25.09.2017 04:11:48;15.3000001907349
25.09.2017 04:13:30;26.3999996185303
25.09.2017 04:15:12;12.1999998092651
25.09.2017 04:16:54;12.5
25.09.2017 04:18:37;28.6000003814697
25.09.2017 04:20:19;25.7000007629395
25.09.2017 04:22:01;11.8999996185303
25.09.2017 04:23:43;12.8999996185303
25.09.2017 04:25:26;27.3999996185303
25.09.2017 04:27:08;24.7999992370605
25.09.2017 04:28:50;12.3000001907349
25.09.2017 04:30:32;15.5
25.09.2017 04:32:14;26.1000003814697
25.09.2017 04:33:57;11.6999998092651
25.09.2017 04:35:39;12.8000001907349
25.09.2017 04:37:21;28
25.09.2017 04:39:03;25.1000003814697
25.09.2017 04:40:45;12.1000003814697
25.09.2017 04:42:28;14.8999996185303
25.09.2017 04:44:11;26.6000003814697
```



```
08.09.2018 00:39:22;16.3999996185303
08.09.2018 00:39:43;16.8999996185303
08.09.2018 00:39:53;19.2000007629395
08.09.2018 00:40:04;19.6000003814697
08.09.2018 00:40:14;27.1000003814697
08.09.2018 00:40:24;32.0999984741211
08.09.2018 00:41:27;30.7000007629395
08.09.2018 00:42:08;29.7000007629395
08.09.2018 00:42:41;29.3999996185303
08.09.2018 00:42:51;28.8999996185303
08.09.2018 00:43:12;28.8999996185303
08.09.2018 00:43:22;29.2999992370605
08.09.2018 00:43:43;28.7000007629395
08.09.2018 00:44:36;28.1000003814697
08.09.2018 00:44:46;18.3999996185303
08.09.2018 00:44:57;16.7999992370605
08.09.2018 00:45:18;16.2999992370605
08.09.2018 00:46:21;17.1000003814697
08.09.2018 00:46:33;19.6000003814697
08.09.2018 00:46:43;20.1000003814697
08.09.2018 00:47:04;32.4000015258789
08.09.2018 00:47:25;32.2999992370605
08.09.2018 00:47:56;31.5
08.09.2018 00:50:03;29.6000003814697
08.09.2018 00:51:48;28.7999992370605
08.09.2018 00:51:58;23.7999992370605
08.09.2018 00:52:08;17.2000007629395
08.09.2018 00:52:39;16.3999996185303
08.09.2018 00:53:40;17.7999992370605
```

B.2. Example of Predicted Missing Data

In the following graphic the misclassified data for the ceiling exchange are highlighted.



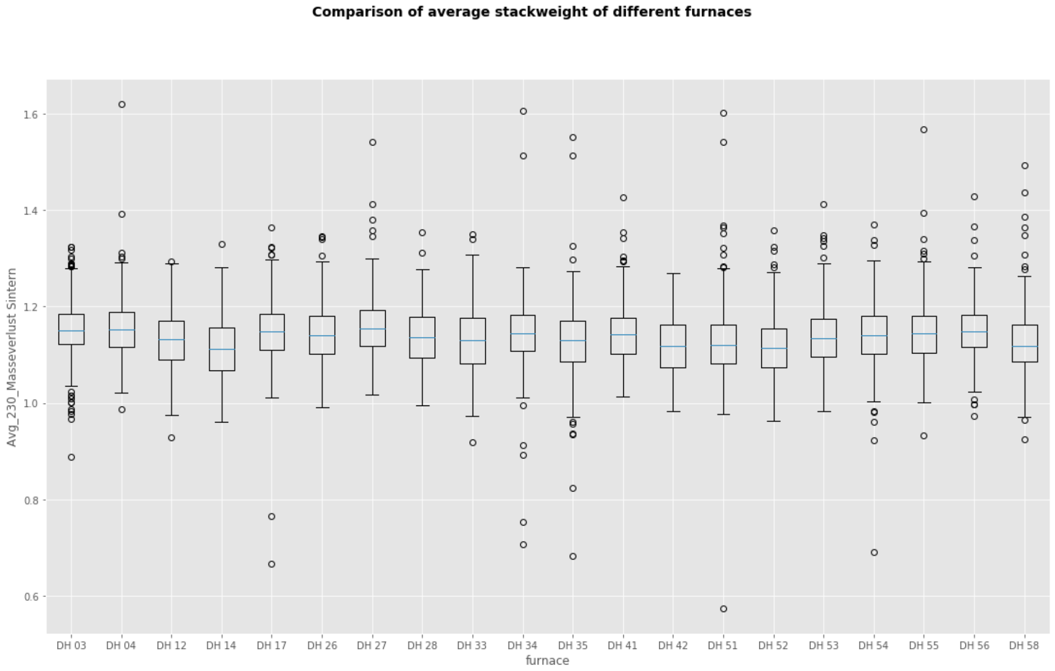
	furnace	run	runs_since_ofendecke_dis	predicted_runs_sinc	diff
424	DH04	144	2	2	0
425	DH04	145	2	2	0
426	DH04	146	2	2	0
427	DH04	147	2	2	0
428	DH04	148	2	2	0
429	DH04	149	2	2	0
430	DH04	150	3	3	0
431	DH04	151	3	3	0
432	DH04	152	3	3	0
433	DH04	153	3	3	0
434	DH04	154	3	3	0
435	DH04	155	3	3	0
436	DH04	156	3	3	0
437	DH04	157	3	3	0
438	DH04	158	3	3	0
439	DH04	159	3	3	0
440	DH04	160	3	3	0
441	DH04	161	3	3	0
442	DH04	162	3	3	0
443	DH04	163	3	3	0
444	DH04	164	3	3	0
445	DH04	165	3	3	0
446	DH04	166	3	3	0
447	DH04	167	3	3	0
448	DH04	168	3	3	0
449	DH04	169	3	3	0
451	DH04	171	4	1	3
452	DH04	172	4	1	3
453	DH04	173	4	1	3
454	DH04	174	4	4	0
455	DH04	175	4	4	0
459	DH04	179	1	1	0

B.3. Product Quality Measurements Compared in the Kilns

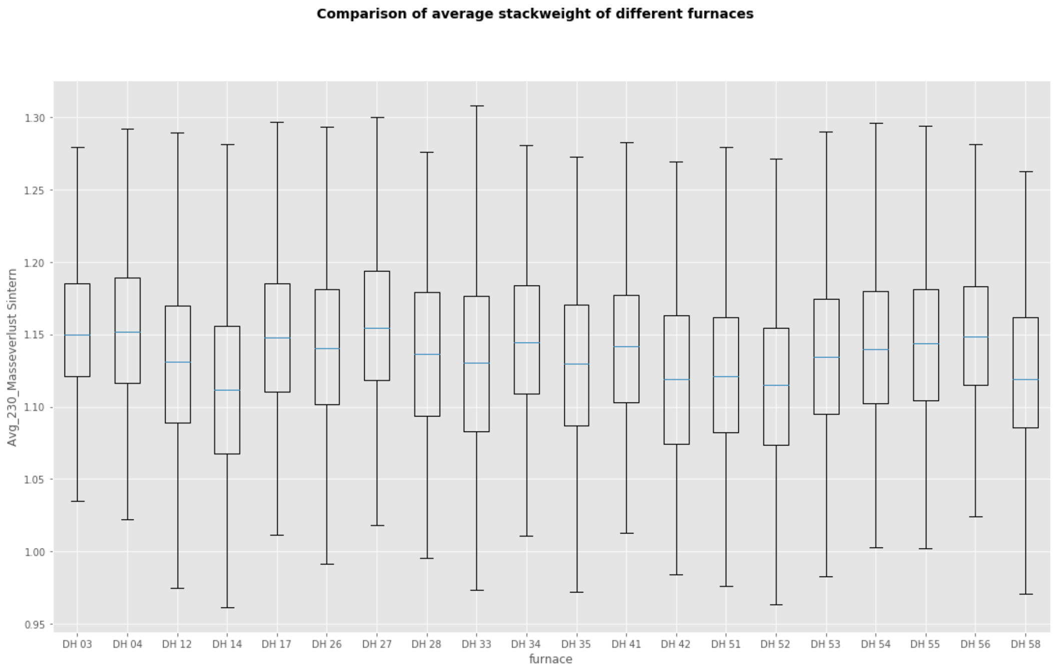
As the plots show there are variations within the measured quality data. For better visualization the outliers have been removed from the plots, compare the first and the second plot.

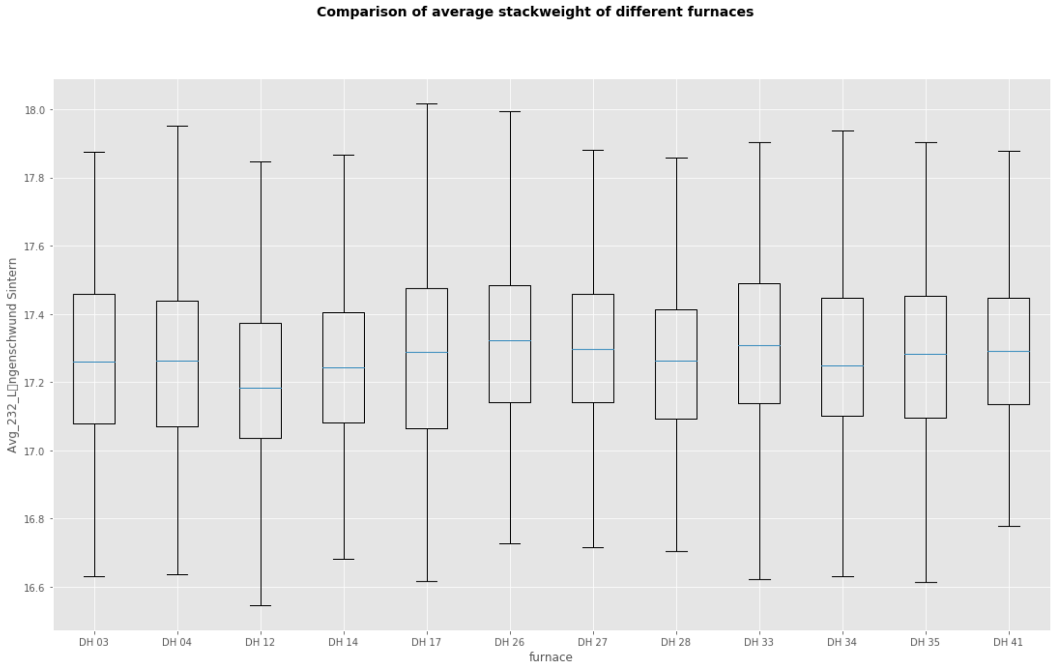
For further investigations concerning the causalities of the variations, additional data would be necessary. As the quality results strongly depend on the material mix.

Further several quality measures are plotted in respect of the age of the ceiling within the kilns. The following graphics show the measures from curvature and length shrinkage. The longer the ceiling is in the kiln the higher the value of the product quality measure.

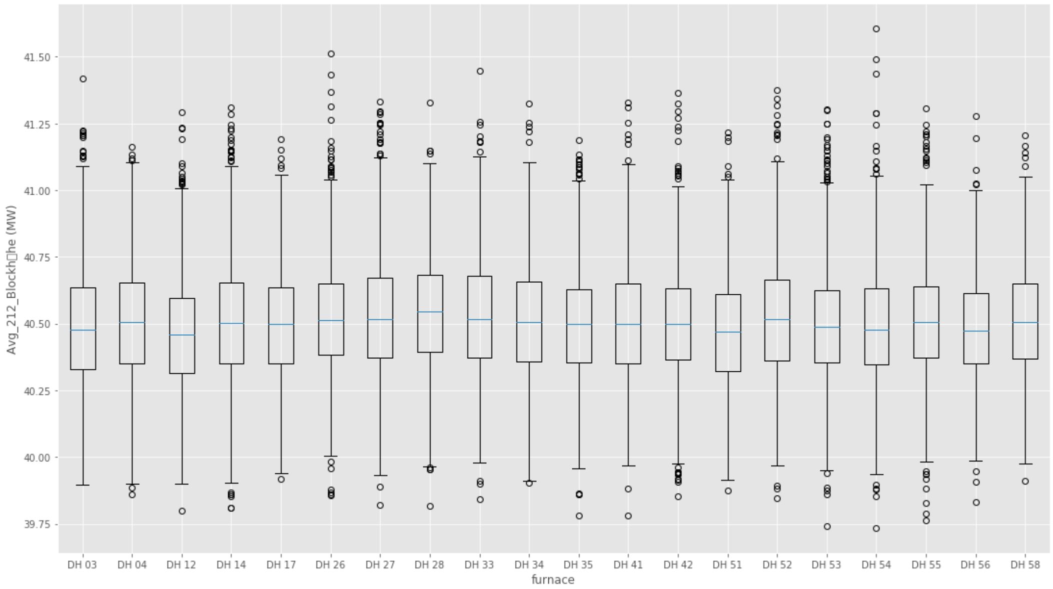


Appendix B. Additional Graphics

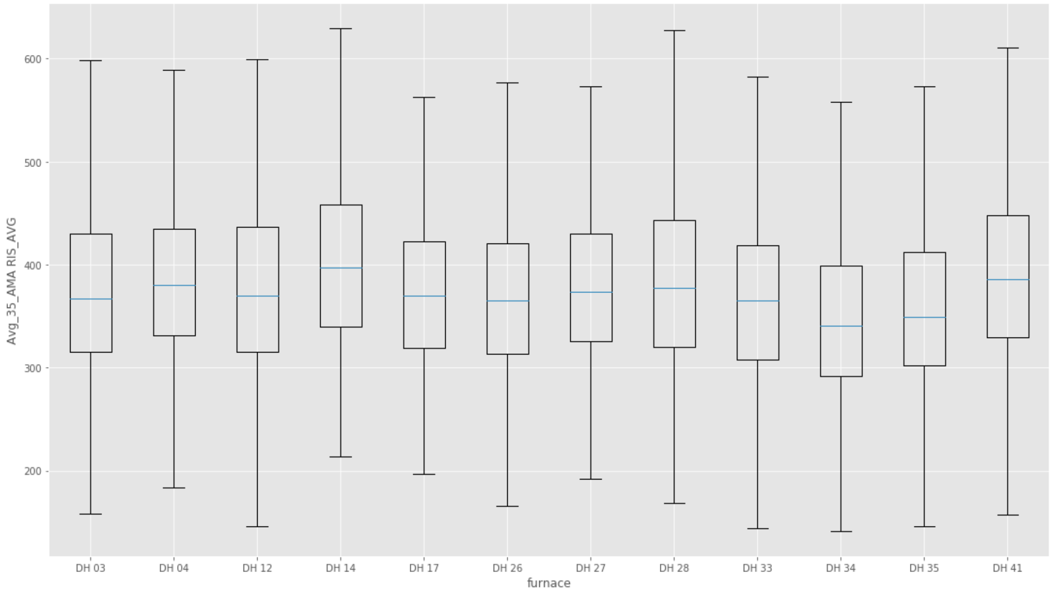


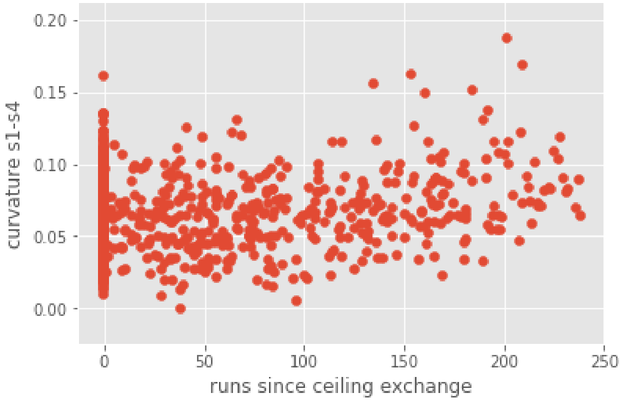


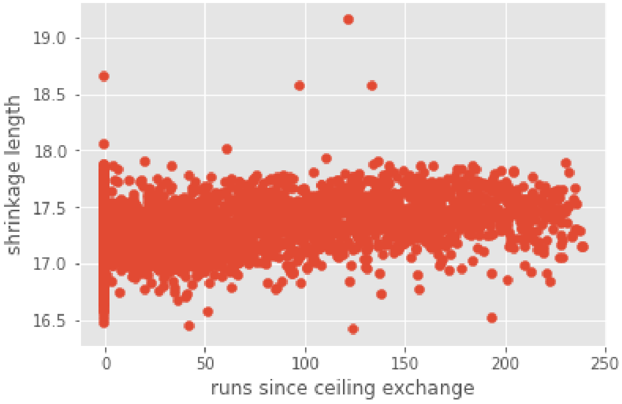
Comparison of average stackweight of different furnaces

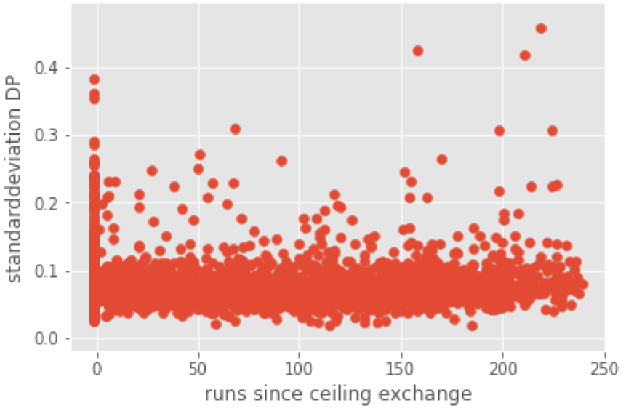


Comparison of average stackweight of different furnaces




curvature s1-s4
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
runs since ceiling exchange


shrinkage length
19.0
18.5
18.0
17.5
17.0
16.5
0
50
100
150
200
250
runs since ceiling exchange


standarddeviation DP
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
250
runs since ceiling exchange

B.4. Used Software and Packages

Software / Tool	Version	Packages
R	3.5.1	dplyr 0.8.0.1 tidyr 0.8.1 stringr 1.3.1 ggplot2 3.1.1
Python	2.7	matplotlib 2.2.3 scikit-learn 0.20.1 numpy 1.15.4 pandas 0.23.4 scipy 1.1.0 statsmodels 0.90.0 seaborn 0.9.0
Matlab	9.2	Test license
Knime	3.7.1	
Weka	3.8.3	
Orange	3.19.0	
Tableau Desktop	2018.3.0	Student license

Bibliography

[1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, Dec. 2009, p. 537. ISBN: 978-0262012430 (cit. on pp. 6, 7, 9, 71, 73).

[2] Kellie J. Archer and Ryan V. Kimes. 'Empirical characterization of random forest variable importance measures'. In: *Computational Statistics and Data Analysis* 52.4 (2008), pp. 2249–2260. ISSN: 0167-9473. DOI: `https://doi.org/10.1016/j.csda.2007.08.015`. URL: `http://www.sciencedirect.com/science/article/pii/S0167947307003076` (cit. on p. 66).

[3] Eric Auschitzky, Markus Hammer and Agesan Rajagopaul. 'How big data can improve manufacturing'. In: (July 2014), pp. 1–4. URL: `https://www.mckinsey.com/business-functions/operations/our-insights/how-big-data-can-improve-manufacturing` (cit. on p. 13).

[4] Ana Isabel Rojão Lourenço Azevedo and Manuel Filipe Santos. 'KDD, SEMMA and CRISP-DM: a parallel overview'. In: *IADS-DM* (2008). DOI: `http://hdl.handle.net/10400.22/136`. URL: `http://recipp.ipp.pt/handle/10400.22/136` (cit. on p. 32).

[5] I. Barletta et al. 'Assessing a proposal for an energy-based Overall Equipment Effectiveness indicator through Discrete Event Simulation'. In: (Dec. 2014), pp. 1096–1107. ISSN: 0891-7736. DOI: `10.1109/WSC.2014.7019968` (cit. on p. 16).

[6] Leo Breiman. 'Random Forests'. In: *Mach. Learn.* 45.1 (Oct. 2001), pp. 5–32. ISSN: 0885-6125. DOI: `10.1023/A:1010933404324`. URL: `https://doi.org/10.1023/A:1010933404324` (cit. on p. 11).

[7] Pete Chapman et al. *CRISP-DM 1.0 Step-by-step data mining guide*. 1999. URL: `https://www.the-modeling-agency.com/crisp-dm.pdf` (visited on 08/08/2018) (cit. on pp. 29, 30).

[8] *Confusion Matrix*. URL: http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/ (visited on 02/03/2019) (cit. on pp. 71, 72).

[9] James Dixon. *Pentaho, Hadoop and Data Lakes*. 2010. URL: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (visited on 08/08/2018) (cit. on p. 21).

[10] H. Fang. 'Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem'. In: *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, June 2015, pp. 820–824. DOI: 10.1109/CYBER.2015.7288049. URL: https://ieeexplore.ieee.org/abstract/document/7288049/ (cit. on pp. 24, 26).

[11] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. 'From Data Mining to Knowledge Discovery in Databases'. In: *AI Magazine* 17 No 3 (1996), pp. 234–246. DOI: https://doi.org/10.1609/aimag.v17i3.1230. URL: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230 (cit. on p. 26).

[12] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. 'The KDD process for extracting useful knowledge from volumes of data'. In: *Communications of the ACM* 39.11 (1996), pp. 27–34. DOI: https://doi.org/10.1145/240455.240464. URL: https://dl.acm.org/citation.cfm?id=240464 (cit. on p. 27).

[13] Mikael Hagstroem et al. *A smarter way to jump into data lakes*. Aug. 2017. URL: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/a-smarter-way-to-jump-into-data-lakes (visited on 09/05/2018) (cit. on p. 26).

[14] Joseph M. Hilbe. *Logistic Regression Models*. CRC Press, May 2009, p. 656 (cit. on p. 7).

[15] Bill Inmon. *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications, 2016, p. 166. URL: https://books.google.at/books?id=gCfdCwAAQBAJ (cit. on p. 22).

[16] Henning Kagermann, Wolfgang Wahlster and Johannes Helbig. 'Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0'. In: *Bundesministerium für Bildung und Forschung, Promotorengruppe Kommunikation der Forschungsunion Wirtschaft – Wissenschaft* (2013). DOI: https://doi.org/10.1016/j.ijpe.2014.12.031. URL: https://www.bmbf.de/files/Umsetzungsempfehlungen_Industrie4_0.pdf (cit. on p. 3).

[17] A. Katal, M. Wazid and R. H. Goudar. 'Big data: Issues, challenges, tools and Good practices'. In: *2013 Sixth International Conference on Contemporary Computing (IC3)*. Aug. 2013, pp. 404–409. DOI: 10.1109/IC3.2013.6612229 (cit. on p. 4).

[18] Borne Kirk. *Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's*. URL: https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/ (visited on 08/09/2018) (cit. on p. 4).

[19] Ron Kohavi et al. 'A study of cross-validation and bootstrap for accuracy estimation and model selection'. In: 14.2 (1995), pp. 1137–1145 (cit. on p. 70).

[20] Douglas Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. META Group, Feb. 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (cit. on p. 4).

[21] N. Laranjeiro, S. N. Soydemir and J. Bernardino. 'A Survey on Data Quality: Classifying Poor Data'. In: (Nov. 2015), pp. 179–188. DOI: 10.1109/PRDC.2015.41 (cit. on p. 51).

[22] Christophe Leys et al. 'Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median'. In: *Journal of Cleaner Production* 49 (2013), pp. 764–766. DOI: https://doi.org/10.1016/j.jesp.2013.03.013. URL: https://www.sciencedirect.com/science/article/pii/S0022103113000668 (cit. on p. 75).

[23] Fei Liu, Jun Xie and Shuang Liu. 'A method for predicting the energy consumption of the main driving system of a machine tool in a machining process'. In: *Journal of Cleaner Production* 105 (2015). Decision-support models and tools for helping to make real progress to more sustainable societies, pp. 171–177. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2014.09.058. URL: http://www.sciencedirect.com/science/article/pii/S0959652614009883 (cit. on p. 16).

[24] *Logistic Regression*. URL: https://cdn-images-1.medium.com/max/800/1*eDeJCcodhj72njIo0x5j0A.jpeg (visited on 21/03/2019) (cit. on pp. 8, 9).

[25] Shengping Lv et al. 'A Review of Data Mining with Big Data towards Its Applications in the Electronics Industry'. In: *Applied Sciences* 8 (2018). ISSN: 2076-3417. DOI: 10.3390/app8040582. URL: http://www.mdpi.com/2076-3417/8/4/582 (cit. on p. 13).

[26] *Machine Learning techniques*. URL: https://de.mathworks.com/discovery/machine-learning.html (visited on 18/12/2018) (cit. on pp. 6, 7).

[27] J. Miller. 'Reaction time analysis with outlier exclusion: Bias varies with sample size'. In: *The Quarterly Journal of Experimental Psychology* 43.4 (1991), pp. 907–912. DOI: 10.1080/14640749108400962 (cit. on p. 75).

[28] Natalia Miloslavskaya and Alexander Tolstoy. 'Big Data, Fast Data and Data Lake Concepts'. In: *Procedia Computer Science* 88 (2016). 7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016, held July 16 to July 19, 2016 in New York City, NY, USA, pp. 300–305. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2016.07.439. URL: http://www.sciencedirect.com/science/article/pii/S1877050916316957 (cit. on pp. 22, 24).

[29] Kevin Nagorny et al. 'Big Data Analysis in Smart Manufacturing: A Review'. In: *Int. J. Communications, Network and System Sciences* 10 (2017), pp. 31–58. ISSN: 1913-3723. DOI: https://doi.org/10.4236/ijcns.2017.103003. URL: https://file.scirp.org/Html/1-9702186_75656.htm (cit. on pp. 14, 15, 35).

[30] Andrew Ng. *Support Vector Machines*. URL: http://cs229.stanford.edu/notes/cs229-notes3.pdf (cit. on p. 9).

[31] Eoin O'Driscoll, Kevin Kelly and Garret E. O'Donnell. 'Intelligent energy based status identification as a platform for improvement of machine tool efficiency and effectiveness'. In: *Journal of Cleaner Production* 105 (2015). Decision-support models and tools for helping to make real progress to more sustainable societies, pp. 184–195. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2015.01.058. URL: http://www.sciencedirect.com/science/article/pii/S0959652615000621 (cit. on p. 17).

[32] Ceyhun Ozgur et al. '"MatLab vs. Python vs. R"'. In: (Apr. 2016), pp. 355–372 (cit. on p. 36).

[33] T. Peterek et al. 'Performance evaluation of Random Forest regression model in tracking Parkinson's disease progress'. In: (Dec. 2013), pp. 83–87. DOI: 10.1109/HIS.2013.6920459 (cit. on p. 11).

[34] DT Pham and AA Afify. 'Machine-learning techniques and their applications in manufacturing'. In: *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 219.5 (2005), pp. 395–412. DOI: https://doi.org/10.1243/095440505X32274 (cit. on p. 14).

[35] Gregory Piatetsky. *Python eats away at R: Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis*. May 2018. URL: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html (cit. on p. 36).

[36] *Production process technical ceramics*. URL: https://www.vogt-ceramic.de/production.html (visited on 03/05/2019) (cit. on pp. 45, 46).

[37] *R vs Python vs MATLAB vs Octave*. Nov. 2018. URL: https://towardsdatascience.com/r-vs-python-vs-matlab-vs-octave-c28cd059aa69 (cit. on p. 36).

[38] *Random Forest*. URL: https://www.researchgate.net/profile/Evaldas_Vaiciukynas/publication/301638643/figure/fig1/AS:355471899807744@1461762513154/Architecture-of-the-random-forest-model.png (visited on 22/03/2019) (cit. on p. 12).

[39] Kalpana Rangra and K. L. Bansal. 'Comparative Study of Data Mining Tools'. In: 04 (June Kalpana Rangra). ISSN: 2277 128X (cit. on p. 41).

[40] Klaus Schwab. *The Fourth Industrial Revolution*. Penguin Books Limited, 2017, p. 192. ISBN: 9780241980538. URL: https://books.google.at/books?id=OetrDQAAQBAJ (cit. on p. 3).

[41] Tom Shafer. *The 42 V's of Big Data and Data Science*. URL: https://www.elderresearch.com/blog/42-v-of-big-data (visited on 09/06/2018) (cit. on pp. 4, 5).

[42] Umair Shafique and Haseeb Qaiser. 'A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)'. In: *International Journal of Innovation and Scientific Research* 12.1 (2014), pp. 217–222. URL: https://pdfs.semanticscholar.org/bd13/74cd5f7ee859da6fdf32ea4e646969e9
pdf (cit. on p. 33).

[43] D. D. Sheu. 'Overall Input Efficiency and Total Equipment Efficiency'. In: *IEEE Transactions on Semiconductor Manufacturing* 19.4 (Nov. 2006), pp. 496–501. ISSN: 0894-6507. DOI: 10.1109/TSM.2006.884718 (cit. on p. 16).

[44] Hartmut Steck-Winter and Günther Unger. 'Thermoprozessanlagen in der smarten Fabrik'. In: *gwi - gaswärme international - Ausgabe 02 2017* (2017), pp. 53–61. ISSN: 0020-9384. URL: http://www.bibliothek.uni-regensburg.de/ezeit/?2271953 (cit. on p. 18).

[45] *Support Vector Machine*. URL: https://cdn-images-1.medium.com/max/660/0*9jEWNXTAao7phK-5.png (visited on 07/03/2019) (cit. on p. 10).

[46] *TRapezoidal Rule*. URL: https://en.wikipedia.org/wiki/File:Composite_trapezoidal_rule_illustration.png#filelinks (visited on 03/12/2018) (cit. on p. 62).

[47] Konstantin Vikhorev, Richard Greenough and Neil Brown. 'An advanced energy management framework to promote energy awareness'. In: *Journal of Cleaner Production* 43 (2013), pp. 103–112. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2012.12.012. URL: http://www.sciencedirect.com/science/article/pii/S0959652612006580 (cit. on p. 17).

[48] Samuel Fosso Wamba et al. 'How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study'. In: *International Journal of Production Economics* 165 (2015), pp. 234–246. ISSN: 0925-5273. DOI: https://doi.org/10.1016/j.ijpe.2014.12.031. URL: http://www.sciencedirect.com/science/article/pii/S0925527314004253 (cit. on p. 4).

[49] Yingfeng Zhang et al. 'A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products'. In: *Journal of Cleaner Production* 142 (2017). Special Volume on Improving natural resource management and human health to ensure sustainable societal development based upon insights gained from working within 'Big Data Environments', pp. 626–641. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2016.07.123. URL: http://www.sciencedirect.com/science/article/pii/S0959652616310198 (cit. on p. 16).

[50] Yingfeng Zhang et al. 'A big data driven analytical framework for energy-intensive manufacturing industries'. In: *Journal of Cleaner Production* 197 (2018), pp. 57–72. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2018.06.170. URL: http://www.sciencedirect.com/science/article/pii/S0959652618318201 (cit. on p. 17).

[51] Yingfeng Zhang et al. 'A framework for Big Data driven product lifecycle management'. In: *Journal of Cleaner Production* 159 (2017), pp. 229–240. ISSN: "0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2017.04.172. URL: http://www.sciencedirect.com/science/article/pii/S0959652617309150 (cit. on p. 15).

[52] G.Y. Zhao et al. 'Energy consumption in machining: Classification, prediction, and reduction strategy'. In: *Elsevier* 133 (2017), pp. 142–157. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2017.05.110. URL: http://www.sciencedirect.com/science/article/pii/S0360544217308666 (cit. on p. 17).

[53] Xiaojie Zhou and Tianyou Chai. 'Pattern-based hybrid intelligent control for rotary kiln process'. In: *2007 IEEE International Conference on Control Applications*. IEEE. 2007, pp. 31–35. DOI: https://doi.org/10.1109/CCA.2007.4389201 (cit. on p. 18).