TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Dmitri Poljakov 192460IABM

Application of Machine Learning Methods to Industrial Equipment Fault Detection

Master's thesis

Supervisor: Olga Dunajeva
PhD

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Dmitri Poljakov 192460IABM

Masinõppe meetodite rakendamine tööstusseadmete rikete tuvastamiseks

Magistritöö

Juhendaja: Olga Dunajeva
PhD

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Dmitri Poljakov

10.04.2021

Abstract

The topic of the master's thesis is "Application of Machine Learning Methods to Industrial Equipment Fault Detection". The thesis aims to find methods based on machine learning techniques for predicting the occurrence of malfunctions on the equipment during operation. This topic was formulated on the personal initiative of the author. The data for the work was collected at Enefit Power AS, which is part of the Eesti Energia group. This research proposes a methodology for diagnostics and fault detection based on machine learning techniques such as linear regression, logistic regression, Random Forest, etc. Additionally, the research presents the anomaly detection method developed initially in R, and then technologically prepared for implementation at the enterprise production process.

As a result of the work, it was found that the PPR and linear regression algorithms perform best based on the sample data. PPR model was 94.7% on the test data. The linear regression algorithm after additional training showed the model descriptive ability on the test data by 90.79%. Then a linear regression algorithm was integrated into the company's production process to detect equipment anomalies.

Keywords: industrial equipment, malfunction detection, machine learning models, R, anomaly detection, master's thesis.

This thesis is written in English and is 64 pages long, including 8 chapters, 67 figures and 7 tables.

Annotatsioon

Masinõppe meetodite rakendamine tööstusseadmete rikete tuvastamiseks

Magistritöö teema on "Masinõppe meetodite rakendamine tööstusseadmete rikete tuvastamiseks". Lõputöö eesmärk on leida meetodid seadmete talitlushäirete esinemise prognoosimiseks töötamise ajal. Tootmisprotsesside mitmekesisus tähendab paljude seadmete kasutamist, mis nõuavad õigeaegset hooldust. Mis omakorda põhineb tänapäevaste diagnostikameetodite kasutamisel rikete tuvastamisel. See teema sõnastati autori algatusel. Andmed töö jaoks koguti Eesti Energia kontserni ettevõttes Enefit Power AS. Selles uurimistöös pakutakse välja rikete tuvastamise metoodika, mis põhineb masinõppe tehnikatel nagu lineaarne regressioon, logistiline regressioon, juhuslik mets jne. Lisaks tutvustatakse uuringus anomaaliate tuvastamise meetodit, mis on algselt välja töötatud R-is ja seejärel tehnoloogiliselt ette valmistatud ettevõtte tootmisprotsessis rakendamiseks. Nende uuringute abil lahendatavate praktiliste probleemide ring on väga lai. Diagnostika, tõrke tuvastamine ei ole täielik loetelu nendest valdkondadest, kus selle töö tulemusi saab kasutada.

Samuti autor uuris oma töös rikete tuvastamise kahte erinevat meetodit, võrdles neid omavahel tõhususe ja kasutusmugavuse osas ning kirjeldas meetodite eeliseid ja puuduseid. Töö tulemusena leiti, et PPR ja lineaarse regressiooni algoritmid toimivad kõige paremini. PPR algoritmi mudeli kirjeldusvõime testandmetel oli 94.7%. Lineaarse regressiooni algoritm pärast mudeli täiendavat õppimist näitas mudeli kirjeldusvõime testandmetel 90,79%. Lineaarse regressiooni muudel oli integreeritud ettevõtte tootmisprotsessi seadmete rikete tuvastamiseks.

Olga Dunajeva osales aktiivselt Tallinna Tehnikaülikooli Virumaa Kolledži tööjuhina. Tema akadeemilised teadmised andsid olulise panuse töösse.

Märksõnad: tööstusseandmed, rikete tuvastamine, masinõppemudelid, R, anomaaliate tuvastamine, magistritöö.

See lõputöö on kirjutatud inglise keeles ja on 64 lehekülge pikk, sealhulgas 8 peatükki, 67 joonist ja 7 tabelit.

List of abbreviations and terms

APCS	Automatic process control system
BIT	Business Info Technology
BIT	Business Info Technology
IO	Artificial intelligence
LR	Linear regression
LNG	Liquefied natural gas
ML	Machine learning
MRO	Maintenance and repair
RMSE	Root mean square error
R^2	Determination coefficient
RF	Random forest, ML algorithm
TG	Turbine generator

Table of contents

1 Introduction

The topic of this master's thesis - Application of Machine Learning Methods to Industrial Equipment Fault Detections. The thesis is based on data from the production process of the Enefit Power AS enterprise, which is part of the Eesti Energia group, but the results of the research can also be applied to the production processes of other enterprises.

The Eesti Energia group is the largest energy company in Estonia. The main activity of the company is the production of electricity, as well as liquid fuels such as shale oil and gasoline. The variety of production processes implies the use of a wide range of equipment that requires timely maintenance. Which, in turn, should be based on modern diagnostic methods. The research main idea was formed from the analysis of the real situation at the company, which the operating personnel faced during the equipment operation. The situation occurred at the Estonian Power Plant in 2017 during the turbine No. 8 operation. During work, the operational staff could not identify the malfunction occurrence on the equipment in time. That caused difficulties in the equipment operation and led to unforeseen costs for the enterprise. This situation clearly shows the need to predict the malfunctions occurrence on the equipment. Necessary to find and investigate a method that will inform the operating personnel about the equipment condition deterioration before the consequences are tangible for the enterprise. This served as the main motivation for writing this work, the relevance of which lies in the search for effective methods for detecting equipment malfunctions at the initial stage of their formation.

Machine learning (ML) is the path to smarter and faster data-driven decision-making when performing predictive maintenance [1]. An excellent example of the implementation of ML technology is Santos, an oil company in Australia, specializing in the exploration and production of oil and gas, the production of petroleum products, and liquefied natural gas (LNG). The company's ML technologies were used to predict equipment failure and increase production. With data from equipment, the company can predict equipment failure with 87% accuracy in 48 hours. Using this time as a buffer, the company can initiate preventive maintenance so that the equipment continues to operate and maintain a constant production level [2].

This master's thesis aims to develop a methodology for detecting malfunctions of industrial equipment based on machine learning methods and techniques for its

implementation using the example of the Enefit Power AS enterprise. The use of machine learning methods to diagnose equipment malfunctions will allow Enefit Power AS to meet modern innovative standards. Nowadays these methods are not yet applied in the production process at the Eesti Energia group. Therefore, during the research, it was important to form a basis for introducing new methods into company daily life.

To achieve this goal, the author set the following tasks:

- Collect historical data on the production process for a sufficiently long period.
- Study machine learning methods that are appropriate for the data and the chosen topic, create fault prediction models, and compare their performance.
- Develop a methodology for detecting faults and technology for its implementation on the example of the enterprise Enefit Power AS.

During this research, the author used such machine learning methods as linear and logistic regression, Random Forest (RF), neural networks, MARS, PPR. Based on the R programming language, the author has created fault prediction models, evaluated their efficiency, as well as the implementation possibility in the production process to Enefit Power AS. Based on this research, the author proposes a technique for detecting malfunctions and anomalies in the operation of industrial equipment, and technology for its implementation in the production process of an enterprise. The data was collected from the Enefit Power AS enterprise info servers. The practical part was implemented into Honeywell enterprise management system.

This master's thesis includes several chapters. In Chapter 2 the other authors related works in this direction are indicated. Chapter 3 describes the process of collecting data and choosing the right machine learning methods for the planned task. Further, the work describes two independent research methods. Chapter 4 describes a method for detecting faults based on cases/precedents that have occurred on the equipment in previous operations. Chapter 5 discusses an alternative method for detecting equipment anomalies, which is well applicable to mechanisms where there is no precedents history. Chapter 6 discusses the prospects and opportunities for further project development. Chapter 7 summarizes the obtained results and compares the two methods for detecting equipment malfunctions, as well as describes the benefits of incorporating the technology into plant daily life and the business case. Chapter 8 summarizes the results of the entire project.

2 Related works

The fault prediction topic on industrial equipment is popular, therefore, similar articles can be found in the public domain. The article *"Прогнозирование отказов оборудования в условия малого количества поломок"* published in the journal "*Вестник Череповецкого государственного университета*“, describes methods of creating models for predicting equipment failures based on the Random Forest ML algorithm. The article presents the main stages of model development and configuration. The model includes several sub-models that predict equipment failure using real and predictable sensor examples. A graph of the difference between actual and predicted signal values for the next period is used to identify failures and deviations. The model is trained on regular data, and the model is tuned on past failure data [3].

The TerraLink website offers an already developed special solution that can significantly reduce the number of downtime and malfunctions incidents, as well as reduce possible operating costs. The solution is a software product that collects, stores, and processes real-time data, which is transferred to a prediction model, where the equipment failure probability is determined based on machine learning algorithms [4]. As a result, the user receives a full-fledged tool in the form of a dashboard that allows controlling technological processes and predict possible failures. A solution based on a mathematical model that allows identifying failures and equipment stops in advance.

The article by Viktor Maltsev "Predictive analytics for the effective use of equipment" presents application examples of the ML methods to improve the enterprise asset management efficiency. Shows the benefits that companies have received after introducing the technology into their daily life.

These advantages are:

- improvement of exploration, production,
- reduction of unplanned downtime,
- optimization of equipment repairs,
- cost reduction,
- determination of the optimal operating conditions for the equipment,
- development of long-term plans for capital and current repairs [2].

The article uses many graphs, from which it is seen that predicting technologies have changed the activities of the enterprise towards a more economical distribution of existing assets. For some companies, it was possible to achieve a result when the prediction of failures occurred 48 hours before their actual occurrence. This is a very good indicator because during this time it is possible to manage preventive measures and prevent negative developments [2].

Anton Krudinov's presentation "Using data from sensors to predict the technical condition of equipment" paid much attention to strategies for carrying out maintenance and repairs of equipment.

The presentation shows the following strategies:

- work to failure,
- scheduled maintenance in time,
- planned maintenance according to operating time,
- maintenance as per condition,
- predictive maintenance [5].

The differences of the listed strategies from each other are shown. It also presents tasks and approaches to predicting the technical condition of the equipment. Approaches to predicting the technical condition of the equipment are divided in the article into two classes: machine learning and engineering calculation. Methods for detecting anomalies on equipment during operation and predicting failures are used for the machine learning class. For the class of engineering analysis, it is proposed to use Multiphysics simulation models [5].

The review of articles helped the author to broaden horizons in the field of industrial equipment diagnostics and to pick out some useful ideas for this work. The basic idea is that fault detection can be done in several ways. The first method is based on fault detection by precedents. The second method is based on detecting anomalies on the equipment. Each of their methods has its advantages and disadvantages. In this thesis, both techniques have been applied. The research involved data collected at the enterprise. Besides, the review of articles helped to better organize the research structure following the topic of the project.

3 Research strategy formation

This research is divided into several subtasks that form the overall work structure. Each subtask is responsible for a specific stage in the project. An important aspect is that the result of a particular sub-task serves as a starting point for the next task. In general, the work takes on a strict sequence of steps, which is easy to navigate.

For successful research work, it is necessary to divide the activities into certain steps. The thesis is conditionally divided into two stages. The first stage is preparatory. This stage includes the information collection, data collection and preparation in the form suitable for further research, suitable ML algorithms and their performance metrics selection.

Below are the steps to prepare for the research:

- collection information about existing data storage systems at the enterprise,
- collection information about the equipment,
- software selection,
- data collection and preprocessing,
- ML algorithms selection,
- performance metrics selection for evaluating ML algorithms [6].

The preparatory research tasks stages are presented in Figure 1.

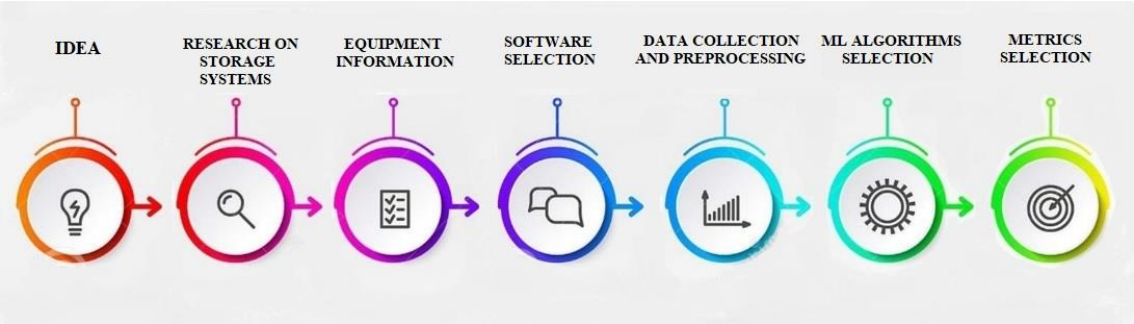

Figure 1. Preparatory stages of research tasks.

The second stage is associated with the research methodology itself. Previously collected data should be splitted into train and test samples. On this stage, the researcher creates ML models, which were identified in the preparatory phase. Each created model must be validated on a test sample. The algorithms performance is evaluated using selected metrics. Only after the model has proven its performance on the test dataset, it will be possible to start implementation it into the production process. At this stage begins the

model applied use phase. Control data is passed through the model to obtain valuable information for the enterprise.

This chapter contains subtasks that are universal and should be performed regardless of the goal and further research progress. This will be done in the thesis next chapters, devoted to methods of detecting faults and anomalies on equipment.

3.1 Collection information about existing storage systems at the enterprise

At large enterprises, processes are divided into production segments, which cannot be controlled without modern automation and computer technology. An automated process control system (APCs) designed to control the enterprise [7].

As a rule, each control system includes information servers. The tasks of these servers are to collect and store process historical data with a certain discreteness. Information servers have functionality that allows providing necessary data about the process for the specified period. Each server has a different user interface. Therefore, the researcher task is to familiarize with the server's functionality, as well as with the information that is available to the user. For this thesis, the Power and Oil Plants information server's functionality was studied in detail.

3.2 Collection information about the equipment

Many rotating mechanisms are involved at the enterprise production processes. These can be turbine units, pumps, blowers, gas blowers, fans, etc. The most interesting for research are powerful rotating mehanisms, especially if these mechanisms are presented in the process in a single quantity. This equipment failure leads to a complete stop of the entire production process. Major equipment breakdown can lead to large financial losses for the enterprise. The subtask of the researcher is to identify the critical equipment involved in the process. For this purpose, the equipment is classified according to its importance. After classification should be paid attention to the equipment that is at the top of the criticality list.

In this research, turbine units were identified as critical plant equipment. They are complex installations consisting of several units interacting with each other. Therefore,

the focus of further research is on this enterprise's equipment. At other industrial enterprises, the key equipment may be alternative units, which are selected based on the specifics of a particular production process.

3.3 Software selection

Different software has different preferences for the data structure. In this research, the following two software products were used to create ML models: Weka and RStudio.

Weka (short for Waikato Environment for Analysis Knowledge) is a modern platform for applied ML. This is free software, the advantage of which is that various ML models are already built into this product functionality. A distinctive feature of this platform is many tuning parameters for precise operating algorithm adjustment when constructing models, as well as when using them to generate predictive values [8].

RStudio is free software, based on the R programming language, and has a wide range of functionality for the researcher [9]. With R and RStudio data can be processed, analysed, and visualized. The product has a huge number of additional packages that can be used for various purposes in different research areas. The difference between these products lies in the ability of R and RStudio to organize research in the form of scripts executed in the form of program code.

3.4 Data collection and preprocessing

For this research, information obtained from different servers was presented in the form of tabular data stored in * .csv format.

By no means always, the initial data received from the corporate repository have a clear structure. Besides, the raw data is often distorted and unreliable: it may contain values outside the acceptable ranges (noise), outliers, and missing values. Therefore, the task of preliminary preparation of the initial data often arises. Data preprocessing techniques usually refer to adding, removing, or transforming values.

Data cleaning consists of identifying and removing errors and inconsistencies in the data to improve the sample quality. Invalid values may appear in the data because of the any sensors malfunction. Such values introduce errors in the research, can lead to inadequate statistics and incorrect conclusions and therefore should be removed from the dataset. The

second reason for deleting data is equipment transients' modes such as starts, stops, and equipment checks, in which the equipment is unstable. Therefore, it is advisable to remove data from the samples referring to the equipment operation in transient modes.

Missing or unreliable data are not always removed, sometimes replaced (imputed) with the mean or median. The choice of a particular method depends on the data amount and the appropriateness of their use in the research [10]. During the data cleansing phase, the researcher also converts data types, aggregate attributes, fills in missing values, and gets rid of noise and outliers [10].

Data transformations to reduce the effects of data skew, or outliers, can lead to significant model performance gains [11]. Variable conversion refers to data normalization or transformation. Normalization allows data to be scaled to a single range for further use in various machine learning models. In practice, the following normalizing attributes methods are most common:

- Minimax - linear data transformation in the range [0...1], where the minimum and maximum scaled values correspond to 0 and 1, respectively,
- Z-scaling data based on mean and standard deviation: Divide the difference between the variable and the mean by the standard deviation [12].

Some models perform better with normalized data and give better predictive results [12].

3.5 Machine learning algorithms selection

ML model is the result obtained by training a ML algorithm using data [6]. The machine learning algorithms used in this research are described in subsections 3.5.1–3.5.5.

3.5.1 Logistic regression

The logistic regression model is designed to solve the problem of predicting the value of the continuous dependent variable Y based on the values of independent variables (predictors) - real X1, ... Xn, provided that this dependent variable can take values in the interval from 0 to 1. Possible use the logistic regression and for solving problems with a binary response when the dependent variable can take only two values 0 (the event did not happen) and 1 (the event happened) [11]. Based on the values of the predictors, the probability of accepting one or another value of the dependent variable Y is calculated [13].

Formula (1) represents the Logistic Regression equation.

$$Y = exp(b0 + b1 * x1 + \ldots + bn * xn)/[1 + exp(b0 + b1 * x1 + \ldots + bn * xn)] \quad (1)$$

The main task for the logistic regression model is classification, which in this thesis is used in the method of fault detecting by precedents, where the data is classified according to the presence or absence of equipment malfunctions.

3.5.2 Linear regression

Linear regression is a regression model used to express the linear dependence of the dependent variable Y on the independent variables X1, ... Xn [11]:

Formula (2) represents the linear regression equation.

$$Y = \mathrm{b}0 + b1x1 + b2x2 + b3x3 + \cdots + bnxn \quad (2)$$

The coefficients of the Linear Regression equation are selected to minimize the sum of squared deviations (SSE) between the observed and predicted values [11].

Formula (3) represents the SSE equation.

$$SSE = \sum_{i=1}^{n} (yi - \hat{y}i)^2 \quad (3)$$

Where yi – are observed values, and $\hat{y}i$ - are model predicted values for sample.

Linear Regression is used in this investigation for both use case fault detection and equipment anomaly detection methods.

3.5.3 Random Forest

Random Forest is an ensemble ML algorithm, where results from several decision trees are combined. Decision trees form a class of learning algorithms that recursively partition the dataset into smaller more pure subsets in order to solve a classification or regression problem. To measure the purity of obtained subsets the information entropy measure is used for classification and SSE for regression problem. Random Forest uses bootstrap samples with replacement to build multiple de-correlated trees that then will be averaged in case of regression problem. In case of classification problem majority votes in the terminal nodes will be used for making a prediction. Random Forest also uses random subsets of predictors for each split and as a result has significant improvements in prediction accuracy comparing to a single tree [14].

3.5.4 Neural networks

Algorithms based on neural networks is a simplified program based on the principles of the human brain. A neural network is built from many neurons, each of which is connected to the rest through synapses. Each of the neurons receives information, processes it, and then transfers the result to the next neuron. The information received by the neuron has a certain weight, which is set through the synapse settings. The more significant the input information has, the more a certain neuron is involved in making the final decision. The neural networks learning process occurs by changing the weights of the connections joining the neurons. Neural networks have several layers [15]. The complexity of a neural network depends on its number (Figure 2).

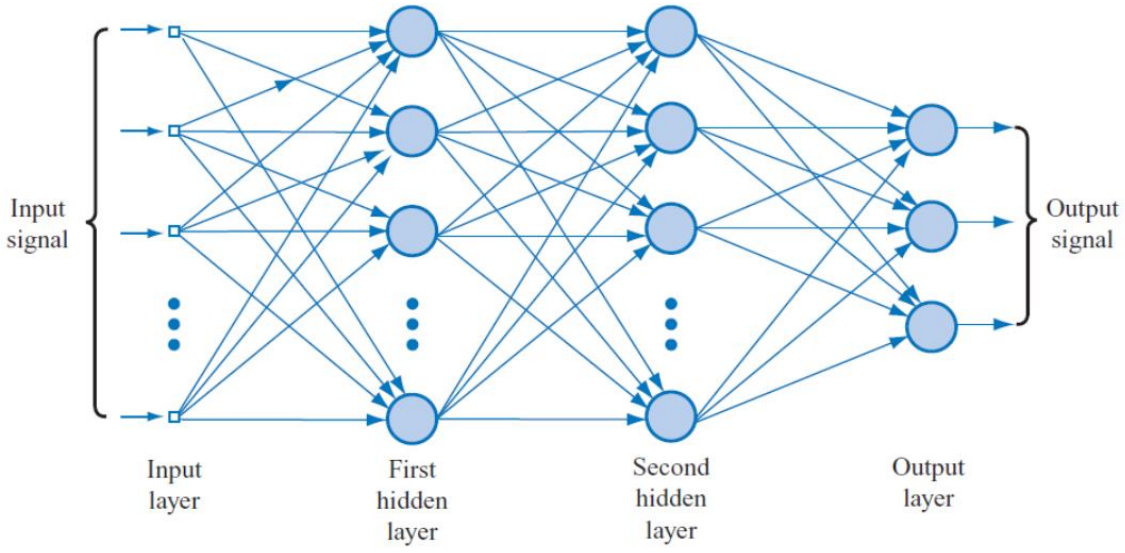

Figure 2. Neural networks [16].

Each neuron has several input channels and only one output channel (Figure 3) [17].

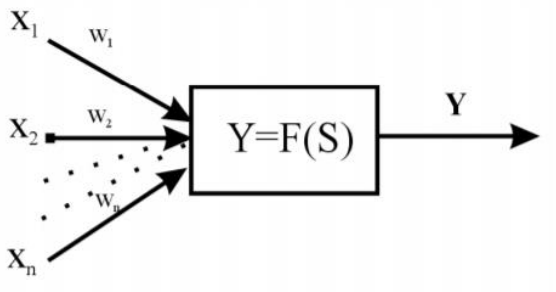

Figure 3. Neuron circuit [17].

The function F (S) is used to activate the neuron. The sum of all outcomes of the signals X1, X2,…, Xn and the weights of these signals w1, w2,…, wn is fed to the input of the

function. The neuron calculates the output Y signal. The most used are linear and sigmoidal activation functions [17].

Complex neural networks perform successfully with tasks that other ML algorithms cannot carry out [15].

3.5.5 MARS and PPR models

Multivariate adaptive regression splines (MARS). MARS is a flexible regression modelling of large data sizes that looks for interactions and nonlinear relationships that help maximize prediction accuracy (Figure 4).

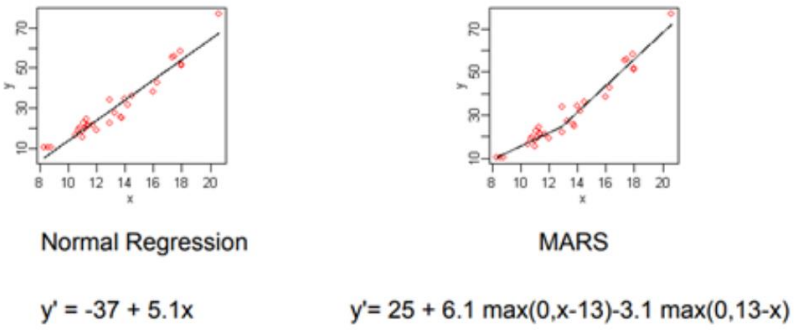

Figure 4. Difference of MARS model from Linear Regression [18].

MARS is generalization of linear regression, which builds the relationship between the dependent and independent variables using so-called basis functions of the form $(x-t)_+$ and $(t-x)_+$, where "+" means positive part, so that $(x-t)_+ = x-t$, if $x > t$ or 0 otherwise. As the value t for each predictor Xj each its observed value xij may be considered [18].

Projection pursuit regression (PPR). PPR adapts additive models in the sense that it first projects a matrix of these explanatory variables in the optimal direction and then applies to smooth functions to those explanatory variables.

Assume $X^T = (X1, X2, \dots, Xp)$ is a vector with p variables. Y is the corresponding response variable. ωm, m = 1,2,…, M is parameter vector with p elements.

Formula (4) represents the Projection Pursuit Regression equation.

$$f(X) = \sum_{m=1}^{M} gm\,(\omega \frac{T}{m} X) \qquad (4)$$

The new feature $Vm = \omega \frac{T}{m} X$ is a linear combination of input variables X. The additive model is based on the new features. Here ωm is a unit vector, and the new feature Vm is actually the projection of X on ωm. It projects the p-dimensional independent variable space onto the new M-dimensional feature space. This is similar to the principal component analysis except that the principal component is orthogonal projection, but it is not necessarily orthogonal here [19].

3.6 Performance metrics selection for evaluating ML algorithms

Performance metrics are the ML model quality indicators. There are many different metrics, the choice of which depends on the research objectives (Figure 5).

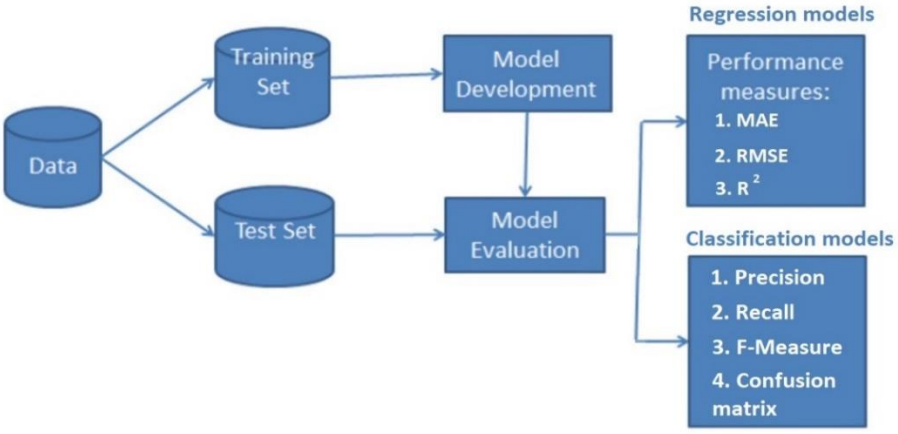

Figure 5. Metrics for evaluating model accuracy.

For classification tasks, there are performance metrics such as Confusion matrix, Accuracy, Precision, Recall, Sensitivity, F-Measure, etc. [20].

To evaluate a binary classification model, the results of which are marked as positive and negative, the Confusion Matrix is often used, which contains 4 cells:

- True positive (TP), objects that have been classified as positive and actually positive (belonging to this class),
- True negative (TN) objects that have been classified as negative and actually negative (do not belong to this class),
- False positive (FP) objects that have been classified as positive but actually negative,
- False-Negative (FN) objects that have been classified as negative but actually positive [20].

Based on the confusion matrix are calculated classification algorithms additional metrics.

Formulas (5) represent metrics equations [20].

$$Precision = \frac{TP}{(TP + FP)} \qquad Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Recall = \frac{TP}{(TP + FN)} \qquad F - Measure = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Most of the tasks in this research are performed using regression models. To evaluate the performance of such models, the following most common metrics were selected:

- RMSE - Root mean square error
- R^2 – Determination coefficient

The root-mean-square-error equations represents the formula (6).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{n}} \tag{6}$$

Where yi – actual values, and $\hat{y}i$ - model predicted results for the sample, n is the total number of errors [21].

The R^2 (r-square) metric indicates the predictive accuracy of regression models measuring the proportion of variance for a dependent variable that is explained by predictors. In the statistical literature, this measure is called the coefficient of determination. A very common definition of this metric is the model r-square R^2 [11].

The coefficient of determination R^2 equations represents the formula (7).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{\sum_{i=1}^{n}(yi-\bar{y}i)^2} \tag{7}$$

Where $\sum_{i=1}^{n}(yi - \hat{y}i)^2$ - sum of squares of regression residuals; yi, $\hat{y}i$ - actual and projected values of the explained variable [11].

$\sum_{i=1}^{n}(yi - \bar{y}i)^2$ - total sum of squares of regression residuals.

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} yi$ - the average of the actual values.

The metric R^2 values can range from 0 (unsuitable match) to 1 (perfect match). The closer the coefficient value is to 1, the better the model will fit the data [11].

4 Case-based fault detection method

The idea behind the case-based fault detecting method is the data used to build ML models is labeled according to the presence of problems on the equipment [22]. The data is classified manually. Data collected when there is a problem with the equipment is classified as True (1). Data collected in the absence of a fault is classified as False (0). The objective of the method is to process the incoming training sample to create logistic regression and linear regression models that will be able to distribute data across classes. In the present investigation, the results of both models were analysed to conclude which model is best suited for the fault detection method. To evaluate the model performance, several test samples were collected. For the logistic regression model, the following metrics were evaluated: precision, recall, F-measure, confusion matrix. For the linear regression model, such performance metrics as RMSE and R^2 were evaluated. Specific minimum requirements for the model's quality metrics were set individually when analysing the obtained results and coordinating them with the production processes in which these models were planned to be used. Only after that, it will be possible to think about using the obtained models with control data in the enterprise production process. The measurements for an arbitrarily chosen period of equipment operation can be used as control data sets. The control samples contain only incoming measurements, without preliminary classification. The model, based on the experience gained during the training stage, predicts the values Yi. In our case, for logistic regression, the predicted variables values will be numbers from 0 to 1. For linear regression, the values may differ up or down from 0 and 1. Values of 0 in the predicted variables Yi will mean no problems on the equipment. Values of 1 in predicted Yi variables will indicate problems on the equipment. These predictions will be the main goal of the case-based fault detection method.

Information about intermediate values between 0 and 1 in the predicted variable Y for regression models will be very valuable. This can be interpreted as intermediate changes in the equipment state. Intermediate values of the predicted variables can be used to generate warning signals to maintenance personnel about deteriorating equipment conditions. So, it becomes possible to predict the situation until the moment when the problem becomes obvious and inform the maintenance personnel about malfunction in advance.

This chapter provides information about the research objects and the instruments used in the investigation with a focus on the case-based fault detecting method. Structurally, the research in this chapter is broken down into the following steps:

- Datasets creation
- Models formation
- Checking models performance
- Testing models in real processes

4.1 Datasets creation

The data for the investigation were collected from the unit No. 8 control system information server of the Estonian Power Plant. The control system is based on METSO DNA equipment. The METSO DNA system includes many programs that provide information in the form of graphs, tables, diagrams, etc. A distinctive feature of the METSO DNA system is the integration of the information server functionality into Microsoft Excel. With the help of an additionally installed component in Excel, it becomes possible to request information of interest directly from the info server and receive an Excel spreadsheet filled with the data.

Several samples were collected for the research. All samples have the same structure but differ only in the class being classified. The classified class can be either True (1) or False (0), depending on the selected turbine operational period. This classification was done manually based on information from turbine operation before and after overhaul. The attribute Y is the target variable in predictive models. All datasets have been tested for outliers - unreliable sensor measurements. Subsequently, such values at which the equipment worked on transient processes were removed from the data.

The samples were collected for the period:

- from 26/11/2015 + one week, 12 variables and 10003 rows,
- from 08/01/2016 + one week, 12 variables and 10080 rows,
- from 06/07/2017 + one week, 12 variables and 10080 rows,
- from 19/07/2017 + one week, 12 variables and 8641 rows,
- from 23/12/2017 + one week, 12 variables and 10080 rows,
- from 20/01/2019 + one week, 12 variables and 10080 rows.

The sampling period is chosen equal to 10 seconds. That is every 10 seconds, the presented data stores up-to-date information about the sensor measurements status. The data for these periods were saved in CSV format.

The samples contain the sensor measurements as:

- turbine speed,
- control signal to control valve 1,
- control signal to control valve 2,
- control signal to control valve 3,
- feedback signal from valve 1,
- feedback signal from valve 2,
- feedback signal from valve 3,
- generalized output signal of the turbine controller,
- electrical load of the turbine,
- turbine power regulator mismatch,
- turbine speed regulator mismatch.

The last column in the data table is the Y classifier - an indicator of whether the data is True (1, problem) or False (0, no problem). To did this, an additional column was created in the dataset. This column was being filled with data based on known information about the period when the turbine is operated with the fault and when the fault has been eliminated (Figure 6).

K	L	M	N
Power	Power regulator mismatch	Speed controller mismatch	Mechanical problem
			Y
181.07	-0.02	-3.17	True
181.15	-0.02	-3.31	True
180.95	-0.02	-3.44	True
181.20	-0.02	-3.59	True
181.62	-0.02	-3.74	True
181.07	-0.02	-3.89	True
181.38	-0.02	-4.06	True

Figure 6. Creating a column with a fault class.

The last step in forming the samples was removing the column with the timestamp and time creation of each line since ML algorithms should not predict data based on date and time values.

The training dataset was created based on data for the turbine operation period from 19/07/17 + one week and from 23/12/17 + one week. Data from 19/07/17 classified as «True B2 (1)». This means that the equipment was in a worn-out condition before the

overhaul. At this time, turbine malfunctions were noticed. The sensor measurements in the presence of mechanical problems on the equipment were included in the dataset. Data from 23/12/17 were classified as «False B (0)». This means that the equipment was in good condition after a major overhaul. Figure 7 shows the algorithm for generating classes for the available data.

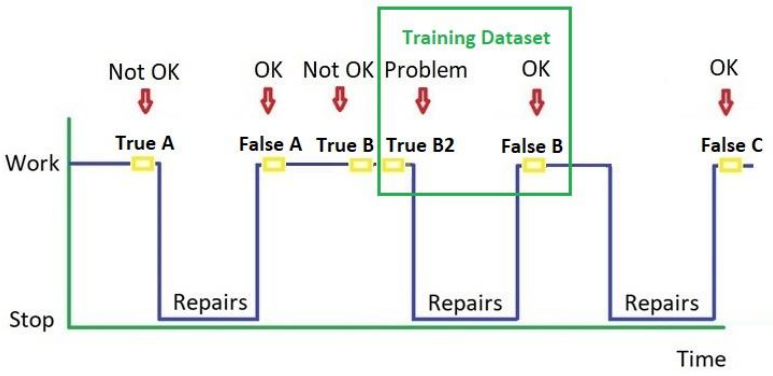

Figure 7. Algorithm for the formation of data classes.

Two samples with different classes were merged into one training set with 18721 rows (Figure 8).

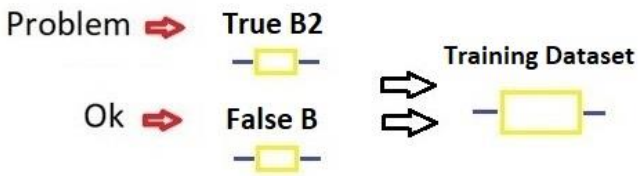

Figure 8. Combining datasets with different classes.

The remaining data are test samples. They are necessary to evaluate and compare predictive models.

The test datasets from 08/01/2016 and from 20/01/2019 are turbine operation data samples after the repair. During the repair, mechanical wear-out on the equipment was eliminated, and the turbine was brought back to its original state. In the Figure 4.2 these datasets are represented by the «False A (0)» and «False C (0)» classes. The difference lies in the time between the next turbine major overhaul. The data were obtained at the beginning of the periods of turbine operation, after each of the repair cycles. This suggests that there is no mechanical wear-out on the equipment.

The test datasets from 26/11/2015 and from 06/07/2017 represent the samples that were collected at the end of the turbine operation cycle, before the next major overhaul. The variation lies in the time difference between the next turbine major overhaul. During these

periods, the turbine worked for a rather large amount of time, and the equipment was subjected to prolonged mechanical stress during operation. This suggests the presence of mechanical worn-out on the equipment. In the Figure 4.2 these datasets are represented by the classes "True A (1)" and "True B (1)".

In this thesis linear regression and logistic regression models were fitted on the training set and evaluated on the test sets. Chapter 4.6 generalized conclusions, characterizing the case-based fault detection method on the equipment.

4.2 Preliminary data analysis

Figure 9 represents the distribution of the turbine electrical load for the training dataset.

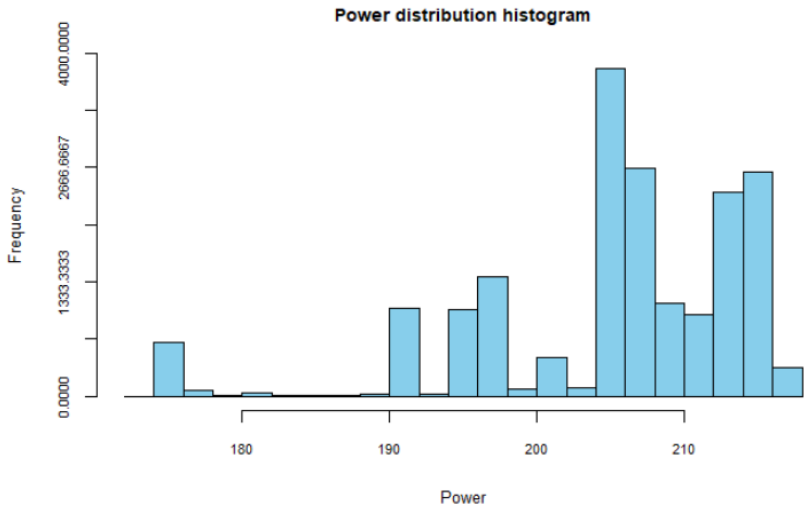

Figure 9. Load distribution histogram for the training dataset.

The load distribution lies in a wide range of different turbine operating modes. The correlation matrix shows the presence of strongly related variables (Figure 10). Control signals to valves and feedback measurements from these valves cause such a dependence.

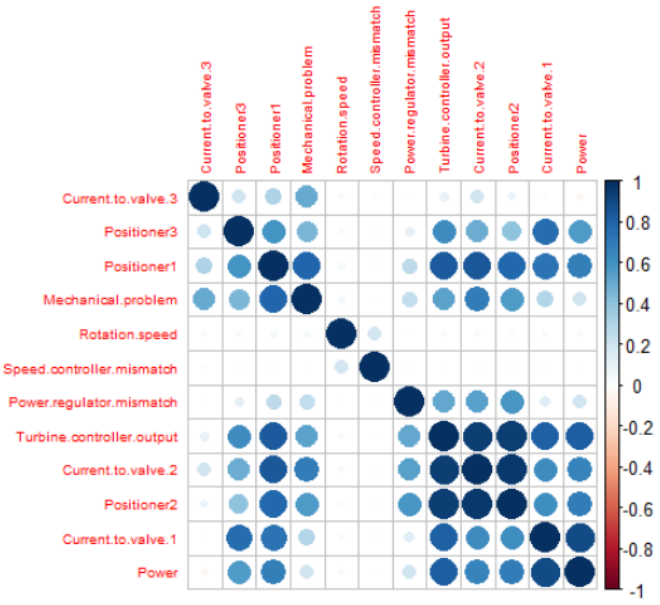

Figure 10. Correlation matrix of the training sample.

Also, it can be noted that the turbine speed and speed controller mismatch did not correlate with other variables. This is because when the turbine is synchronized with the electrical power system, the turbine speed remains constant over the entire operating load range and depends only on the frequency in the electrical network. The speed controller only works in transient modes until the generator is connected to the electrical network. Therefore, the speed controller mismatch was not correlated with other data.

Below is a comparison of the turbine load, turbine controller output, and power regulator mismatch distributions concerning the presence or absence of equipment problems in the training set (Figure 11).

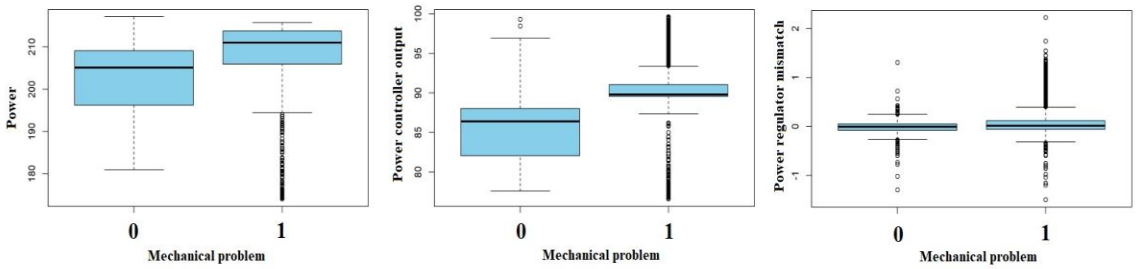

Figure 11. Comparison of distributions regarding the presence or absence of problems on the equipment.

Visual analysis of the distributions shows a slight increase in the amplitude of the turbine regulator mismatch in the presence of a problem on the equipment.

Below is a graph of the actual turbine load for the training sample and a visual separation border of different classes in the data using a vertical line (Figure 12).

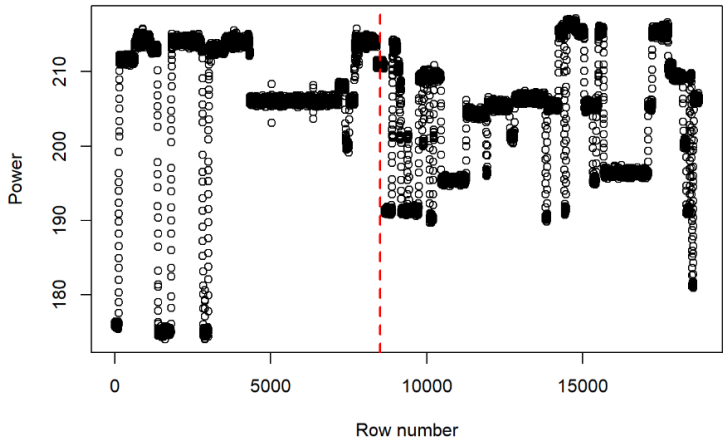

Figure 12. Turbine load in the training sample.

On the left side of the graph are the loads in the presence of a defect/wear-out on the turbine before overhaul. On the right side - loads in the absence of defects after overhaul. It is visually seen that the operation modes in various classes differ. The operation mode in the True (1) class is more stable, but there are several strong changes in the load towards decreasing. This class is dominated by the turbine operating modes closer to the rated load (215 MW). The operation mode in the False (0) class shows a large number, but less noticeable changes in the load over the period participating in the sample. This class is dominated by turbine operating modes in a wide range of operating loads (from 190 to 215 MW).

4.3 Formation of ML models

4.3.1 Linear regression

The R programming language with the RStudio development environment was used to create ML models. The list of available models is very extensive. However, at this stage, the author is limited to the creation of linear and logistic regression models. This is because these models are easy to integrate into the enterprise management system.

A linear regression model was created based on the training sample. The model task is to predict a class value that indicates the presence or absence of a problem on the equipment. The final model was selected with the Akaike Information Criteria (AIC), that iteratively removed the least significant components from the linear regression model [23].

The formula (8) presents the linear regression model equation.

$Y = -44.53 + [Rotation\ speed] * 0.004793 - [Current\ to\ valve1] * 0.03096 + [Current\ to\ valve\ 2] * 0.03566 + [Current\ to\ valve\ 3] * 0.0377 + [Positioner\ 1] * 0.03835 + [Positioner\ 2] * 0.004589 + [Positioner\ 3] * 0.1745 - [Turbine\ controller\ output] * 0.07026 - [Power] * 0.01355 + [Power\ regulator\ mismatch] * 0.04785$ (8)

The RMSE for the training dataset was 0.13. The model's r-square R^2 was 92.6%.

The most significant attributes were positioner # 3 feedback, turbine controller output, and power regulator mismatch. The feedback signal from positioner # 3 turned out to be the most influential attribute because this valve # 3 is the main regulating element that maintains a stable turbine load. Control valves # 1 and # 2 are generally fully open in most operating modes. Until a certain period, they do not participate in load regulation. The purpose of the turbine controller is to maintain a specific, target load. The appearance of gaps in the regulating elements will lead to an increase in the turbine controller output signal oscillation and the power regulator mismatch. This is because the controller will try to compensate for the gaps to keep the valve in a certain position. These arguments explain why these features had the greatest impact on the model predicted values.

4.3.2 Logistic regression

A logistic regression model was created from the training sample. The model task is to predict a class value, which indicates the presence or absence of a problem on the equipment. The WEKA software was used to build the model. The reason was the possibility of comparing two software products, and the accumulating of certain experience to work on both platforms. The logistic regression model metrics for the training dataset are shown in Table 1.

Table 1. Logistic regression model metrics for the training dataset.

Correctly classified values (Accuracy)	99.98%
Incorrectly classified values	0.01%
Precision	1
Recall	1
F-Measure	1
Confusion matrix	a b <-- classified as 10079 1 | a = False 1 8640 | b = True

4.4 Validation on Test Data

4.4.1 Test sample from 20/01/2019 + one week

In the test data for the period from 20/01/2019 + one week, the turbine control system state was classified as False (0), without wear. So, the expected model prediction is 0.

A comparison of the predicted and actual values and the linear regression model residuals are presented below (Figure 13). The red line on the actual vs predicted graph marks the expected result of the classification. The linear regression model RMSE = 0.65.

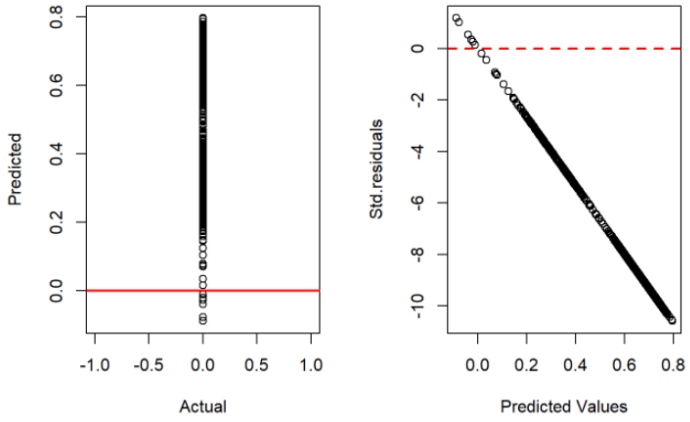

Figure 13. Predicted values distribution and model residuals.

Below are graphs of the predicted class values and the turbine actual load (Figure 14).

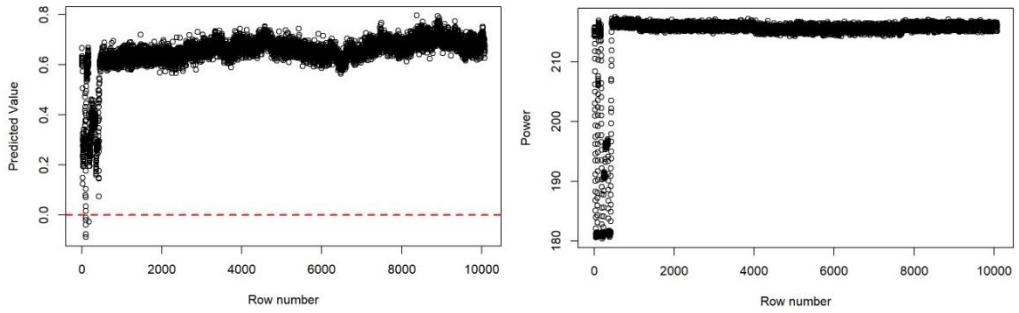

Figure 14. Graphs of the classifier predicted values and the turbine actual load.

In the predicted class values, there are significant amplitude fluctuations associated with a change in the turbine load downward. During turbine stable operation at rated load, the amplitude fluctuations decreased significantly within the range from 0.55 to 0.7. The average of the predicted class values was 0.64.

The predicted values for the logistic regression model are shown in Figure 15.

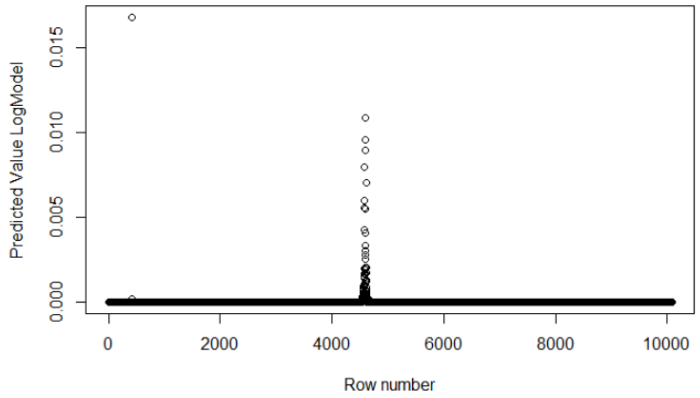

Figure 15. Predicted values distribution result for the logistic regression model.

The logistic regression model performed better with the classification task on the test sample for the period from 20/01/1919 year + one week. The predicted class values are fully consistent with the manual classification in the sample. Logistic regression model metrics on the test sample are shown in Table 2.

Table 2. Logistic regression model metrics on the test sample.

Correctly classified values (Accuracy)	100.00%
Incorrectly classified values	0.00%
Precision	1
Recall	1
F-Measure	1
Confusion matrix	a b <-- classified as 10080 0 | a = False 0 0 | b = True

From the above can be concluded that the mean of the predicted values for the linear regression model did not match the labelled class in the test data, but at the same time, for the logistic regression model, the predicted values completely matched with the expected.

4.4.2 Test sample from 06/07/17 + one week

In the test data starting from 06/07/2017 + one week, the state of the turbine control system was classified as worn-out True (1). So, the expected model prediction is 1.

A comparison of the predicted and actual values and the linear regression model residuals are presented below (Figure 16). The red line on the actual vs predicted graph marks the expected result of the classification. The model RMSE = 0.14.

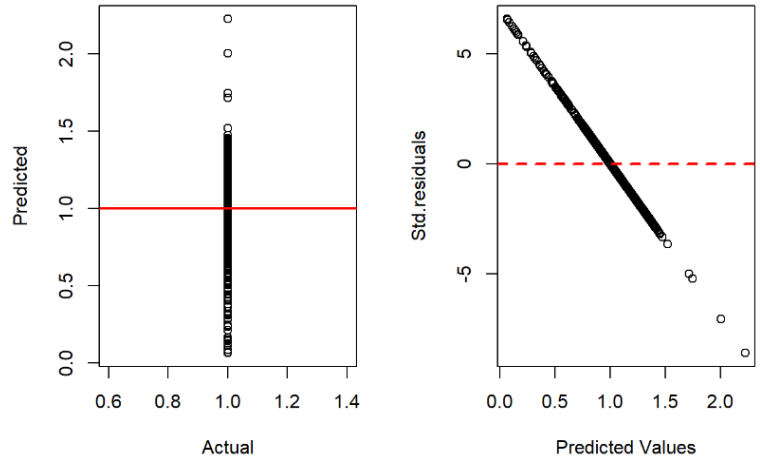

Figure 16. Predicted values distribution and model residuals.

Below are graphs of the predicted class values and the turbine actual load (Figure 17). The average of the predicted class values was 0.96.

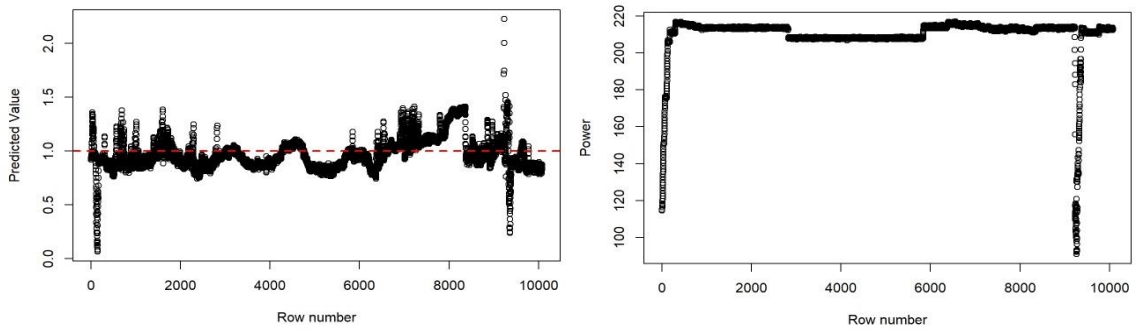

Figure 17. Graphs of the classifier predicted value and the turbine actual load.

The predicted values for the logistic regression model are shown in Figure 18.

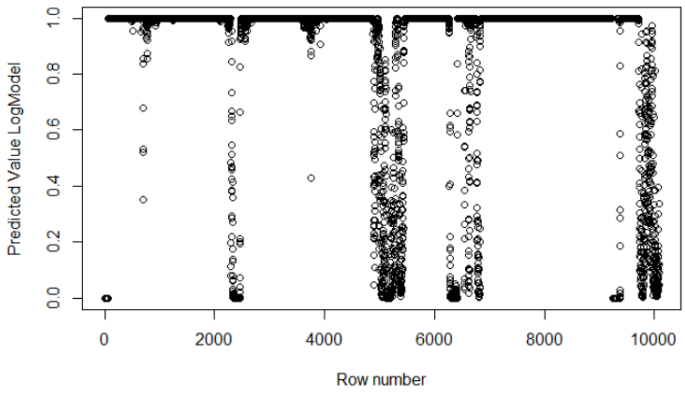

Figure 18. Predicted values distribution result for the logistic regression model.

The predicted class values generally matched manual classification in the sample. However, these values are not evenly distributed across the entire data range. There was a fairly large number of values in which the classifier readings were reversed. The logistic regression model performance has deteriorated (Table 3).

Table 3. Logistic regression model metrics on the test sample.

Correctly classified values (Accuracy)	88.71%
Incorrectly classified values	11.28%
Precision	1
Recall	0.88
F-Measure	0.94
Confusion matrix	a b <-- classified as 0 0 | a = False 1138 8942| b = True

The linear regression model performed better with the classification task for the test sample starting on 06/07/17 + one week. For the linear regression model, the predicted values show amplitude jumps from 0.8 to 1.4. There are also strong jumps associated with changes in the turbine load. But the readings are not reversed, as was found in the logistic regression model. The worn-out state of the control system was confirmed by the predictions of both models.

4.4.3 Test sample from 08/01/16 + one week

In the test data from 08/01/2016 + one week, the turbine control system state was classified as not worn-out False (0). So, the expected model prediction is 0.

Below are graphs of the predicted class values and the turbine actual load (Figure 19).

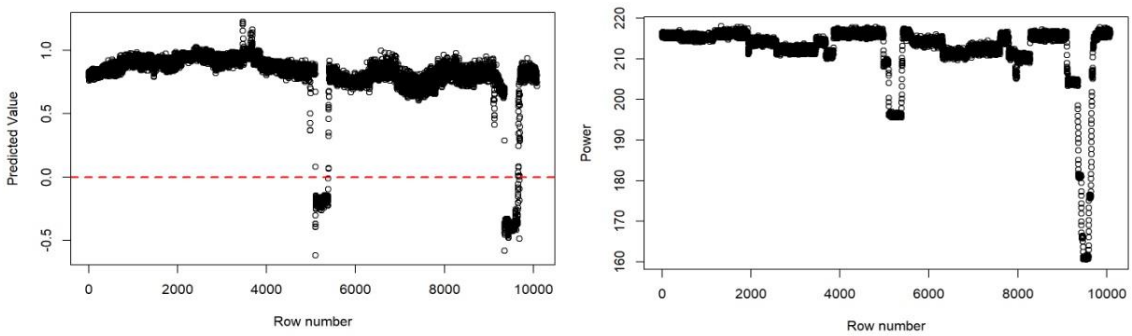

Figure 19. Graphs of the classifier predicted value and the turbine actual load.

The model RMSE = 0.82. The average of the predicted class values was 0.77.

The predicted values for the logistic regression model are shown in Figure 20.

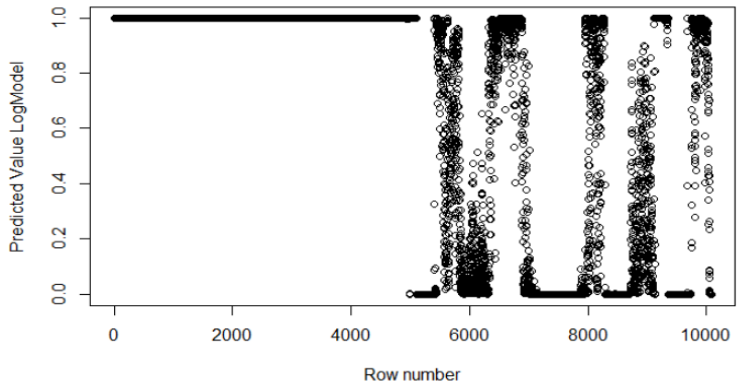

Figure 20. Predicted values distribution result for logistic regression model.

The predicted class values did not match the manual classification in the sample. There was a fairly large number of values in which the classifier readings were reversed. Logistic regression model metrics on the test sample are shown in Table 4.

Table 4. Logistic regression model metrics on the test sample.

Correctly classified values (Accuracy)	32.31%
Incorrectly classified values	67.68%
Precision	1
Recall	0.32
F-Measure	0.48
Confusion matrix	a b <-- classified as 3257 6823 | a = False 0 0 | b = True

On a sample starting from 08/01/16 + one week, both models showed results that did not correspond to the expected ones. It can be assumed that this is not due to the ML models poor performance, but to an error in the manual data classification in the test sample. This may indicate that during the repair work on the turbine at the end of 2015, the repair was not carried out in full volume. The averaged predicted value for the linear regression model (0.77) indicates that the control system mechanical wear-out at that time was 77%. This is even though the turbine has just come out of an overhaul. It should be noted here that this experiment tested the model created on the 2017 training data against the 2016 test data. In reverse time sequence.

This observation may indicate a lack of this method for predicting equipment failures. In this method, the correct data classification for the construction of predictive models plays an important role. An error in data classification can significantly affect the performance of the entire method. Therefore, before classifying the data, it is necessary to collect additional information about the amount of overhaul in each case. Often, access to such information is limited. This can lead the researcher to classify the data based only on assumptions, as was done in this case. But further results analysis showed that this assumption does not always correspond to reality. The dataset was classified incorrectly.

4.4.4 Test sample from 26/11/15 + one week.

In the test data for the period from 26/11/2015 + one week, the turbine control system condition was classified as a worn-out True (1). So, the expected model prediction is 1.

Below are graphs of the predicted class values and the turbine actual load (Figure 21). The model RMSE = 2.0. The average of the predicted class values was 0.73.

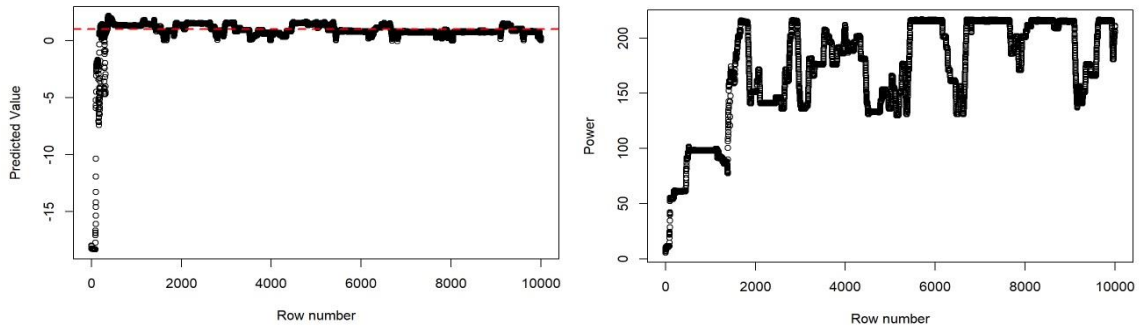

Figure 21. Graphs of the classifier predicted value and the turbine actual load.

The model RMSE was very large. This is because the test sample contains data for the turbine operation transition period, i.e., when the turbine was started up. This shown in the actual load graph. There were no similar transient modes in the training sample, so the model is greatly mistaken for such operation modes. This indicates that it is necessary to approach to the test sample preparation more carefully and remove turbine transient modes operation from it. The predicted values for the logistic regression model are shown in Figure 22.

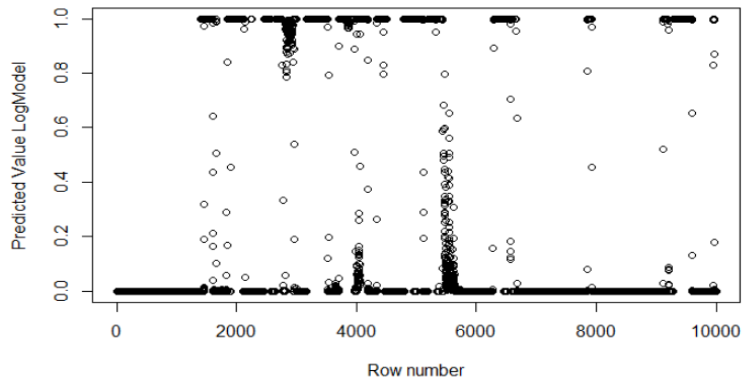

Figure 22. Predicted values distribution result for the logistic regression model.

The predicted class values did not match to manual classification in the sample. There was a fairly large number of values in which the classifier readings were reversed. Logistic regression model metrics on the test sample are shown in Table 5.

Table 5. Logistic regression model metrics on the test sample.

Correctly classified values (Accuracy)	31.10%
Incorrectly classified values	68.89%
Precision	1
Recall	0.3
F-Measure	0.47
Confusion matrix	a b <-- classified as 0 0 | a = False 6892 3111| b = True

On the sample starting from 26/11/15 + one week, the logistic regression model showed results that did not correspond to the expected ones. Most of the values were misclassified. In turn, evaluating the linear regression model results, the average predicted value equal to 0.73 coincides with the expected result. Data classification in the sample has been performed correctly. The regulation system wear-out was 73% on average.

At this stage, a lot of comparative tests between logistic and linear regression models have been carried out. In most cases, the best results were achieved using linear regression models. Therefore, in the next experiments only linear regression models will be used.

4.4.5 Test sample from 19/07/17 + one week

In the test data for the period from 19/07/2017 + one week, the turbine control system condition was classified as worn-out True (1). So, the expected model prediction is 1. Note this period takes part in the training dataset formation.

Below are graphs of the predicted class values and the turbine actual load (Figure 23). The model RMSE = 0.08. The average of the predicted class values was 0.96.

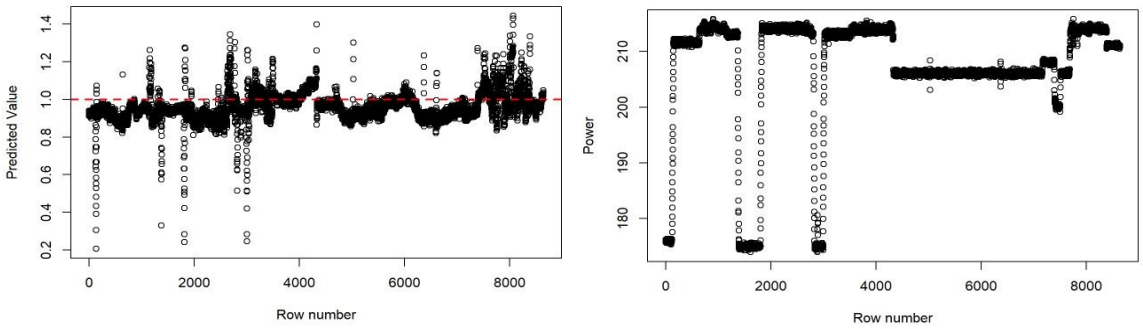

Figure 23. Graphs of the classifier predicted value and the turbine actual load.

The model RMSE is very low. Because this period takes part in the training dataset formation. The predicted values correspond to the labelled values in the dataset. There were jumps in the amplitudes of the predicted values from 0.8 to 1.2. Some of the predictions looked like “outliers”. They were more noticeable at a lower turbine load.

4.5 Additional training of the linear regression model

During the research, the author has repeatedly come across the fact that the changes in turbine load strongly affects the formation of the predicted value for the test samples. The impression is that the model begins to predict not the presence of defects on the equipment, but a change in the turbine operating modes. However, changing the turbine operating mode is a typical operation dictated by the power system requirement to produce electricity at a certain point in time, that has nothing to do with the presence or absence of defects on the equipment. To get rid of this pattern, an attempt was made to additionally train the model by introducing new data into the training set. This action will expand the range of possible equipment operation modes in the training dataset, thereby the model will learn to distinguish a greater number of operation modes and respond less to them when forming predictive values. The data from the previous training sample were merged with additional data starting from 06/07/2017 + one week. These data were

classified as True (1) – wear-out on the turbine was present. The addition of new data to the training sample for this period was chosen due to the observation, that this sample had the smallest classification error in the previous test results. A new linear regression model was built using the new training sample.

The new model RMSE was 0.12. The model r-square was 93%. The previous model had RMSE of 0.13 and an r-square of 92%. Hence the model performance was improved. The new model was validated on a test dataset for the period from 20/01/2019 + one week, which were already used to validate the previous model. In this sample, the turbine control system state has been classified as wear-free False (0). This time, the RMSE was not calculated, because it was already known that the data classification in this test dataset was done incorrectly, and the model error will be large in any case.

Below are the comparative graphs of the predicted values for the new and for the previous model (Figure 24). The left graph shows the predicted values of the new model. The right graph shows the predicted values of the previous model for comparison. For both models, the same test dataset from 20/01/2019 + one week was used.

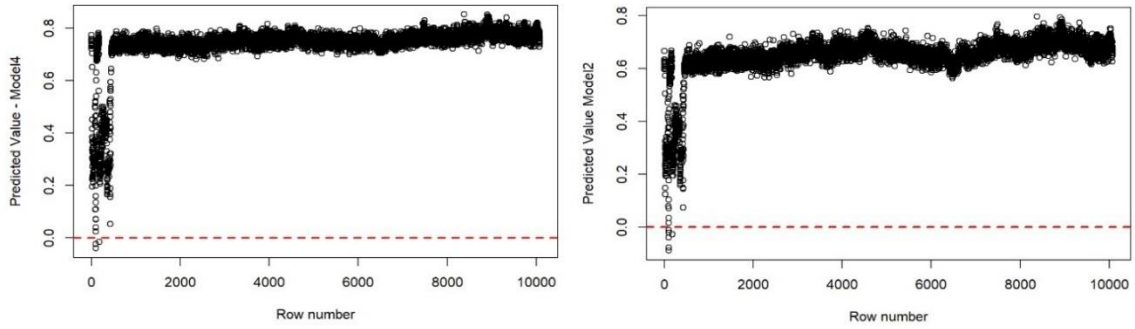

Figure 24. Comparative graphs of predicted values for the new and previous models.

The graphs comparison shows that it was not possible to get rid of fluctuations in the predicted values associated with a change in load. But at the same time, the range of predicted values became narrower at a stable turbine load. This confirms the assumption that the introduction of new, verified data into the training dataset increases the model predictive accuracy. The average predicted value increased from 0.65 to 0.74. The completed model also confirmed that the regulation system state after overhaul for the period 20/01/19 remained worn-out.

4.6 Generalized conclusion by the case-based fault detection method

The case-based method discussed in this chapter needs to be supplemented with additional research. For this method, it is necessary to collect information about cases/faults on the equipment and include the data over the fault's lifetime in the training dataset. The more accurately the precedent data are collected, then more accurate model will be obtained. For the case-based fault detection method, the best results were achieved using linear regression models. But it should be noted that for the test dataset from 20/01/2019 + one week, the logistic regression model showed the best predicting results. Table 6 summarizes the linear regression models results for training and test datasets.

Table 6. Model results for training and test datasets.

Linear regression model	**Dataset type used**	**RMSE**	**R^2**	**Manual Classifier**	**Predicted classifier**
Basic model	Training set from 19/07/17 + 23/12/17	0.13	92.6%	-	-
Basic model	Test set from 20/01/2019	0.65	-	0	0.64
Basic model	Test set from 06/07/2017	0.14	-	1	0.96
Basic model	Test set from 08/01/2016	0.82	-	0	0.77
Basic model	Test set from 26/11/2015	2.0	-	1	0.73
Basic model	Test set from 19/07/2017	0.08	-	1	0.96
Additionally trained model	Training set from 19/07/17 + 23/12/17 + 06/07/17	0.12	93%	-	-
Additionally trained model	Test set from 20/01/2019	-	-	0	0.74

This method of predicting equipment malfunctions can be used in an applied sense, and this method takes place in existence. However, there are some disadvantages. The most significant disadvantage of this method is the incorrect manual data classification during creating datasets, which can affect the model performance. When classifying the data, the researcher relied only on the fact that the turbine was out of overhaul, and based on this information, he classified the data as FALSE (0). Obtained result analysis shows that this is not always true. The obtained predicted values show that the performed maintenance does not guarantee the restoration of control system to the original state, not worn-out. Most likely, in some years, repairs are not carried out fully and the control system components are not affected during repairs. Only this can explain the fact why the

predicted value immediately after the repair was on average about 0.6 (dataset from 20/01/2019) or about 0.8 (dataset from 20/01/2019), and did not approach zero, as the researcher expects. This observation greatly reduces the confidence in this fault prediction method. Therefore, this fault detection method is suitable only for those enterprises where it is guaranteed to obtain accurate information about the amount of repair work performed on the equipment.

The second disadvantage was that the data for the True (1) class was collected from a single-use case that happened on the equipment on 09/07/2017. It was at this time when a malfunction on the equipment exposed itself from a technological point of view. In other cases, the data was classified as True (1) only based on the one-year operation cycle before the major overhaul. As such, the operating personnel did not have any equipment complaints. In the data manual classification, the presence of wear-out on the equipment was made only based on the assumption about the previous long period operation. There is no documentary evidence of the wear-out presence on the equipment. Also, the next possible malfunction on the equipment can develop in a completely different scenario, about which the model does not know anything and, accordingly, it will not be able to predict them. A possible way out of this situation is to create a generalized sample, which will combine different equipment operation periods. But because of the reasons indicated above, there is a possibility that, due to misclassified classes, the generalized sample will be unreliable and will not give good results when building a model.

The third disadvantage is the predicted value jumps, which are associated with a change in the turbine load. But changing the load, in terms of power generation technology, is a standard operation. The predicted value responds to load changes, which can also lead to result misinterpretation. In some cases, a change in the turbine load is perceived by the model as a malfunction/wear-out. The predicted value becomes higher than 1. Thus, this method requires additional research and improvements before it can be used in an applied sense at the enterprise. This conclusion gives reason to consider another alternative fault detection method based on detecting anomalies on equipment, which ideology differs significantly from the precedent-based method. Here the model learns to predict the good equipment condition. Any significant deviation from the healthy condition can be perceived as an anomaly in the equipment. The recorded anomaly will already provide a basis for more detailed research on what causes of this equipment behaviour. The method for detecting anomalies on equipment is discussed in the next chapter of this work.

5 Method for detection anomalies on equipment

Anomalies in control systems involves the following aspects (Figure 25)

- anomalies based on the absolute measurements of one sensor,
- anomalies based on the absolute measurements of the joint sensor's behavior,
- atypical patterns in the sensor measurements,
- global patterns in the joint behavior of the several sensor measurements [5].

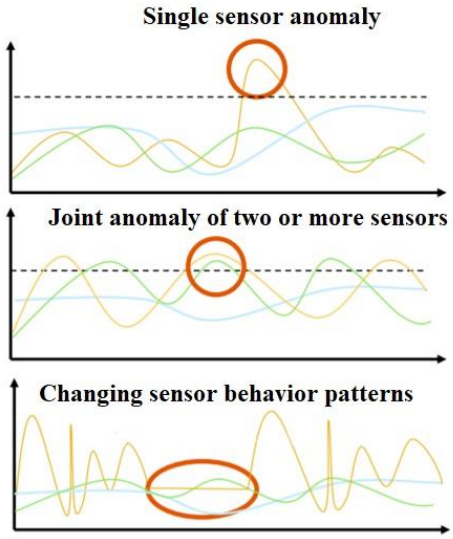

Figure 25. Anomalies in the behaviour of sensor measurements [5].

The idea behind the method for searching for anomalies on equipment is to create such a ML model that will predict the good condition of the equipment. Moreover, any deviation of the predicted value from the normal will be perceived as a malfunction. The schematic algorithm of the anomaly detection method is shown in Figure 26.

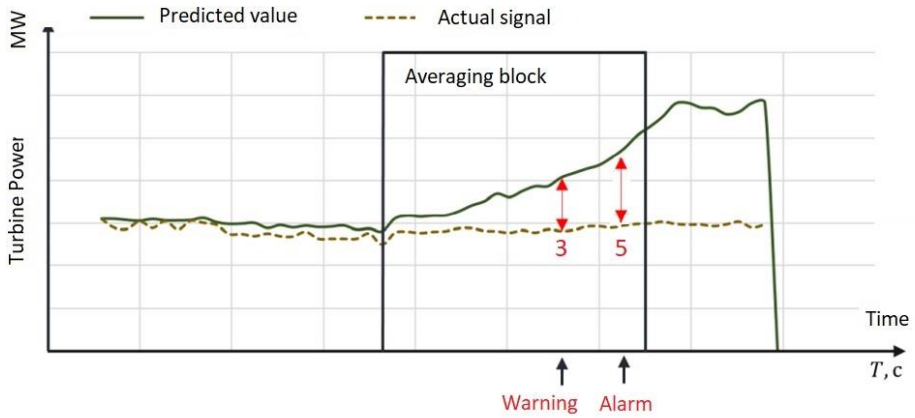

Figure 26. Algorithm of the anomaly search method [3].

The oil plant Enefit 280 turbine was chosen as the object of the research. The oil plant turbine has a rated power of 30 MW. Figure 27 shows the appearance of the operator display in the Honeywell plant control system, as well as the measurements that were involved in the research.

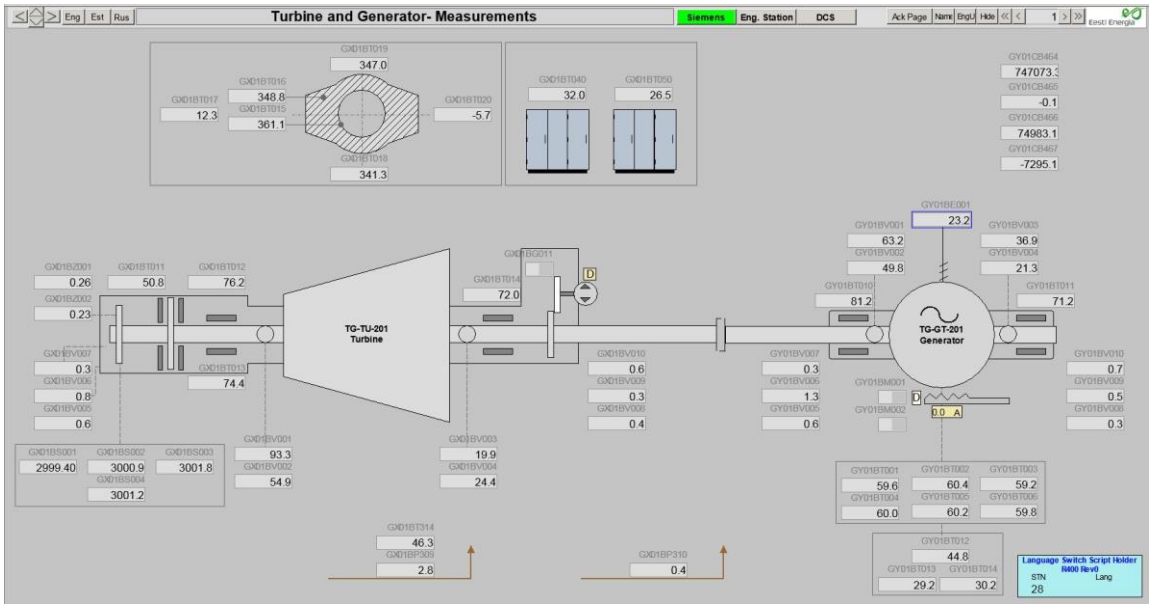

Figure 27. Turbine and generator measurements.

Each measurement in the control system has a unique KKS code. The KKS code has a strict formation structure. Therefore, it determines the belonging of any sensors to one of the functional types of measurements, such as temperature, vibration, axial displacement, pressure, flow rate, etc.

The mechanical sensor measurements of the turbine and generator were used to create the models. Complete list of measurement points can be found via the link in the Appendix 2 (Turbine and generator measurement points Enefit 280).

During this research, such predictive models as linear and nonlinear regression, Random Forest, Neural Networks, MARS and PPR were created.

All models predict the turbine load value Y (GX01BV001) based on the mechanical measurements (bearings temperature, rotor and generator vibration, lubrication oil temperature, cooling air temperature, etc.).

5.1 Data collecting and quality control

The data was collected from the oil plant information server. The data server is part of Honeywell-based control system. To work with data in the Honeywell system, a specialized client program Uniformance Studio is provided. The program functionality allows to request information about the process and display it in the form of tables and graphs. The functionality of the Uniformance Studio program was used to obtain the oil plant turbine operation data.

To create training and test datasets were collected sensor measurements characterizing the turbine unit operation for the period from 24/11/2020 to 21/12/2020. The data were split between the training and test datasets. For the training dataset, were used the turbine operation data for 3 weeks from 24/11/2020 to 16/12/2020 (Figure 28). The training dataset consists of 46 variables and 21142 rows.

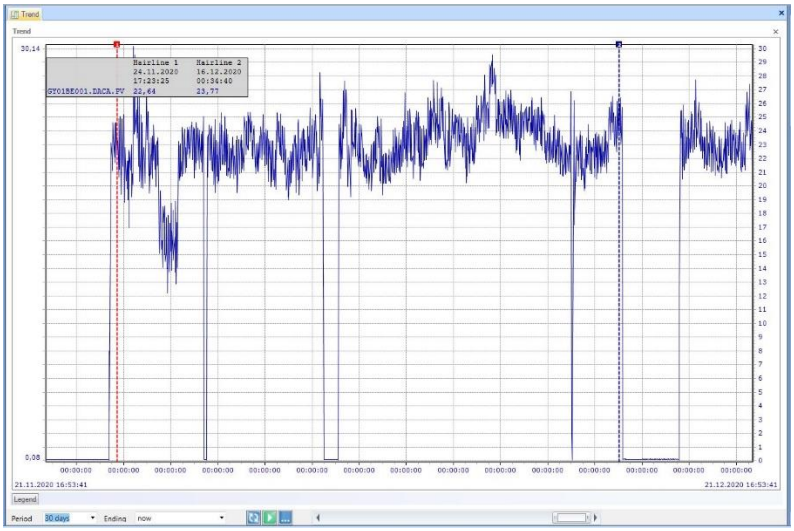

Figure 28. Period for training data.

For the test dataset required for testing and validating the model, was used the turbine operation data for 4 days from 18/12/2020 to 21/12/2020 (Figure 29). The test dataset consists of 46 variables and 2883 rows.

The datasets were uploaded to RStudio for further research. In RStudio, the sample was checked for consistency and the absence of missing values. All attributes in the datasets are in numeric data format.

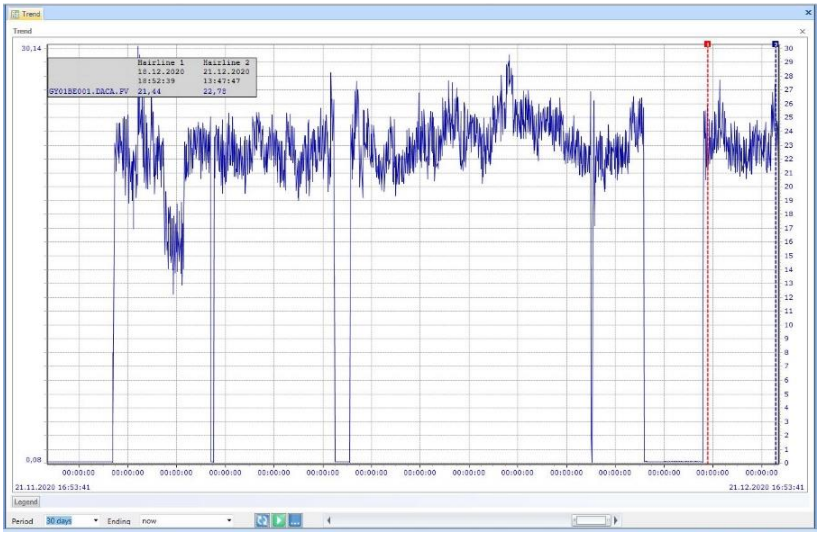

Figure 29. Test data period.

5.2 Preliminary data analysis

The turbine load distribution in the training dataset is presented in Figure 30. The histogram in Figure 30 shows the main turbine load lying in the range from 13 to 30 MW. The average turbine load during the training period was 21.5 MW. This range is the nominal operating area for this turbine. Oil plant turbine does not work at loads below 13 MW. The histogram also shows several zero load values associated with several periods of turbine shutdown. These data were deliberately left in the training dataset to "teach" model to predict the equipment stopped state as well. Otherwise, every time the turbine is stopped, there would be sharp jumps in the prediction values.

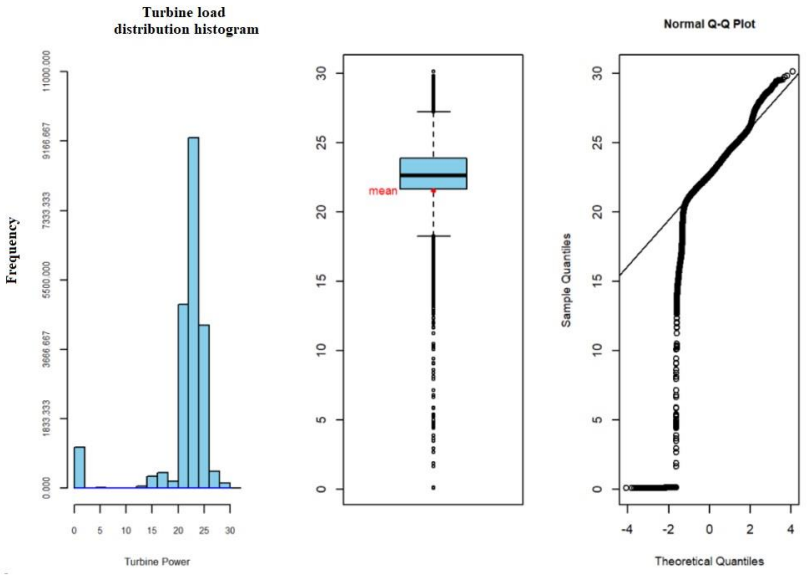

Figure 30. Turbine load distribution histogram.

Figure 31 shows the attributes correlation coefficients concerning the turbine active load Y (GY01BE001).

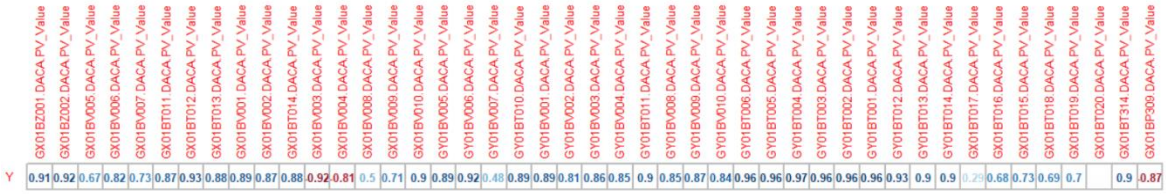

Figure 31. Attributes correlation coefficients concerning the turbine active load.

Most of the attributes are strongly correlated with Y, except for GX01BT020_Value and GX01BT017_Value.

The hierarchical clustered correlation matrix in Figure 32 shows the presence of groups of closely related attributes, which indicates a multicollinearity problem [24].

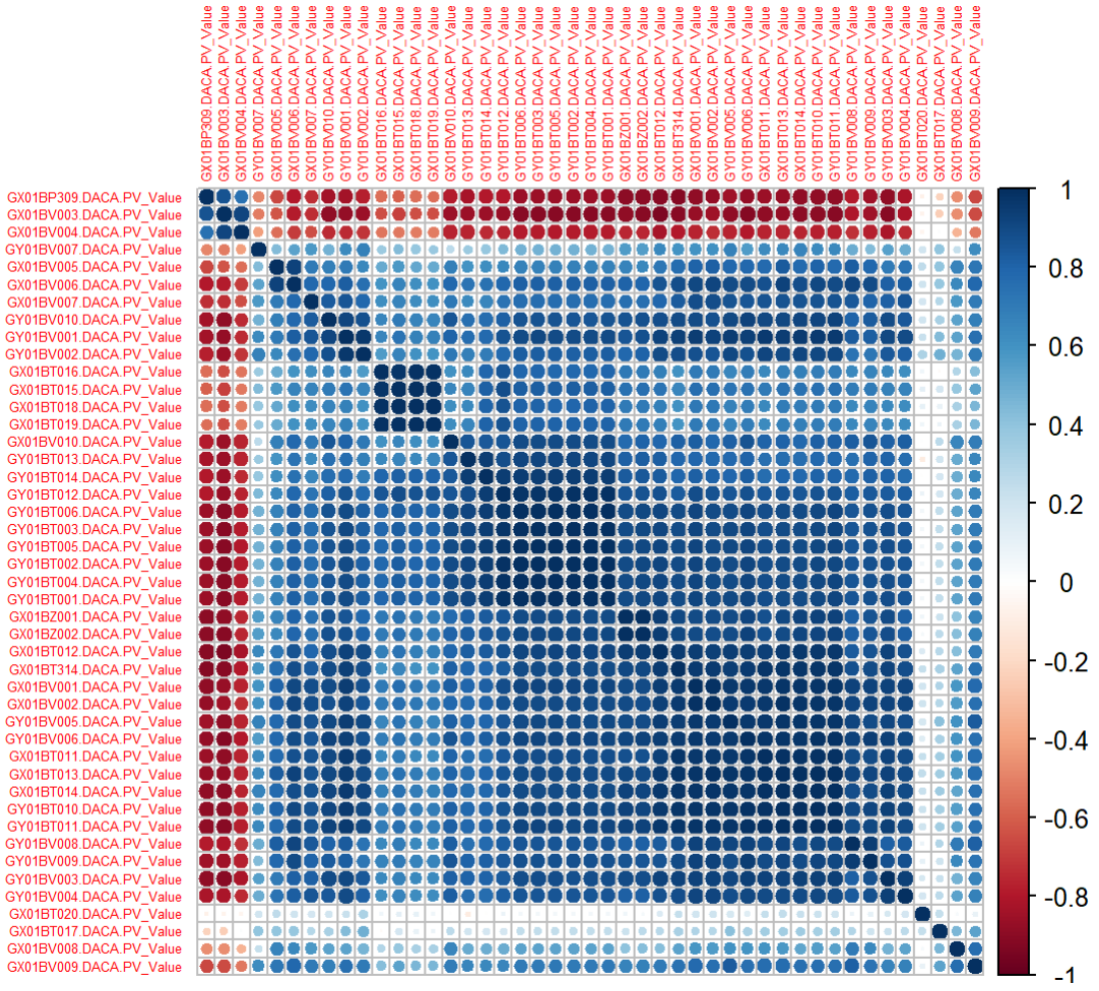

Figure 32. Attribute correlation matrix.

It is also notable that two features GX01BT020_Value and GX01BT017_Value do not correlate with others and with each other.

Some typical features' distributions are presented in Figure 33. All plots can be found via the link in Appendix 2.

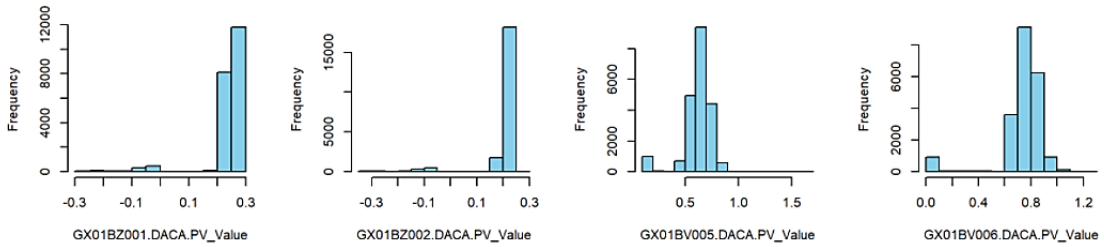

Figure 33. Feature distribution histograms.

The analysis of the features' histograms makes it possible to conclude that the features' distributions are asymmetric, vary within different limits, which means that features require transformations.

Figure 34 explains the relationship of attributes with the turbine current load. Both linear and non-linear relationships are present, with many zero values that reflect the stopped state of the equipment. All plots can be found via the link in Appendix 2.

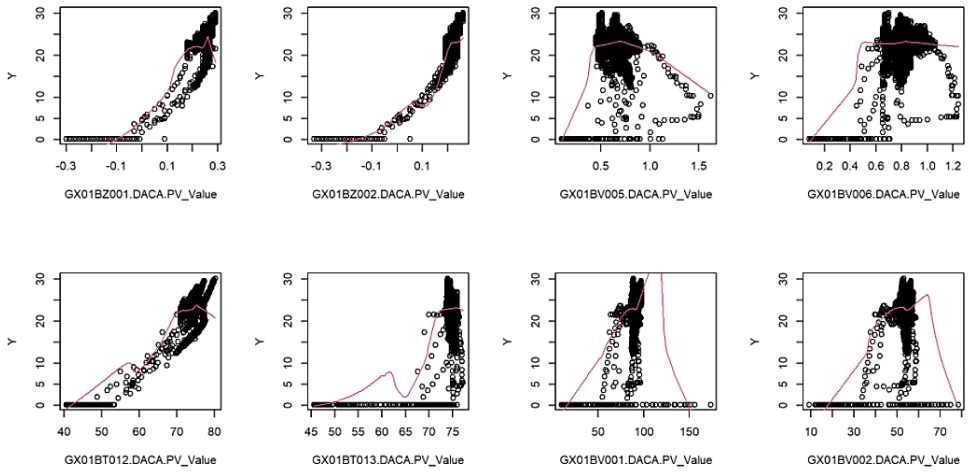

Figure 34. Graphs of the attribute's relationships with the turbine current load.

5.3 Creating a linear regression model

A linear regression model was created based on the training dataset. The model task is to predict the turbine load based on the remaining measurements. The final model was selected with the Akaike Information Criteria (AIC), that iteratively removed the least significant variables from the regression model. The obtained linear regression models,

before and after the AIC variable selection, can be found via the link in Appendix 2 (Linear Regression model Enefit 280).

The model RMSE for the training dataset was 0.59. The model r-square R^2 was 98.8%.

Figure 35 presents the linear regression model's residuals plots, which indicate the presence of outliers.

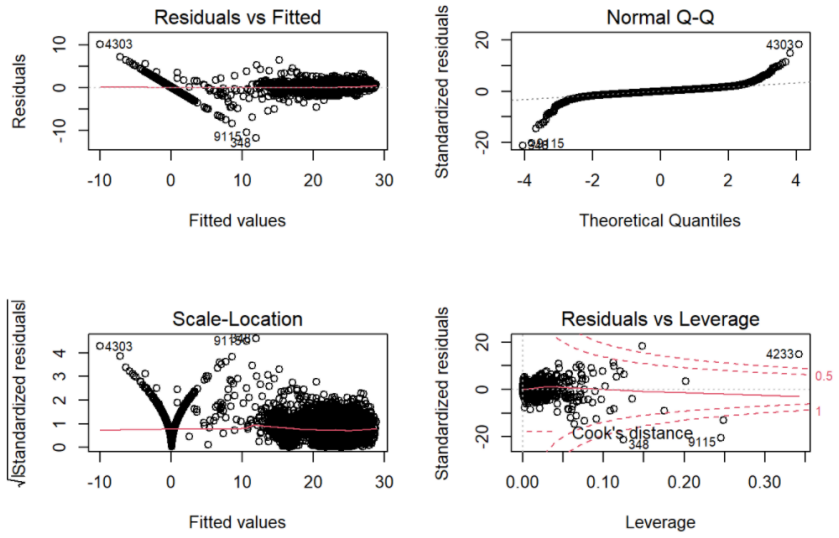

Figure 35. Linear regression model residuals distribution diagrams.

Checking for multicollinearity by calculating the inflation variance factors (VIF) also showed the presence of unacceptably high VIF values, which can lead to inaccurate model predictions [11]. Several measurements of the generator winding's temperature turned out to be the most problematic. There is a technological explanation for the issue of multicollinearity for these points. The generator winding's temperature has a linear dependence on the active load.

A graph of the predicted values versus the actual values of the turbine load for the train dataset is presented in Figure 36. The graph shows that the model is poor at predicting values at zero loads, that is, the operating mode when the turbine was stopped. Variations in the predicted values were found in the range from -8 to +10 MW, which can be explained by the fact that the training dataset provides data for the turbine shutdown period in a limited amount. To improve the model quality at zero workloads, it is necessary to add additional data to the training dataset corresponding to the turbine stopped operating state.

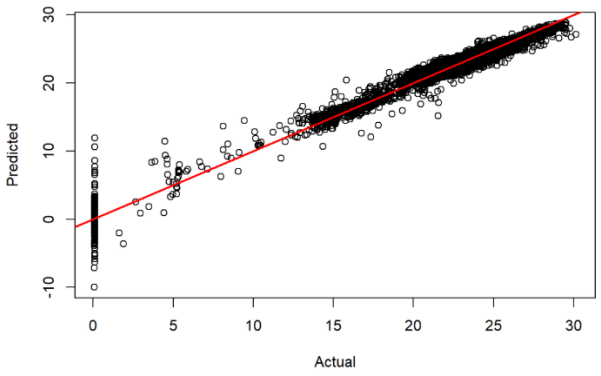

Figure 36. Predicted values relative to the turbine load actual values for training dataset.

5.4 Model validation on the test set

The test set for model validation is based on turbine mechanical sensors measurements for the period from 18/12/2020 to 21/12/2020. The turbine load distribution histogram over the test period shows that the values are in the range from 20 to 26 MW (Figure 37). The average load for the test period was 22.9 MW.

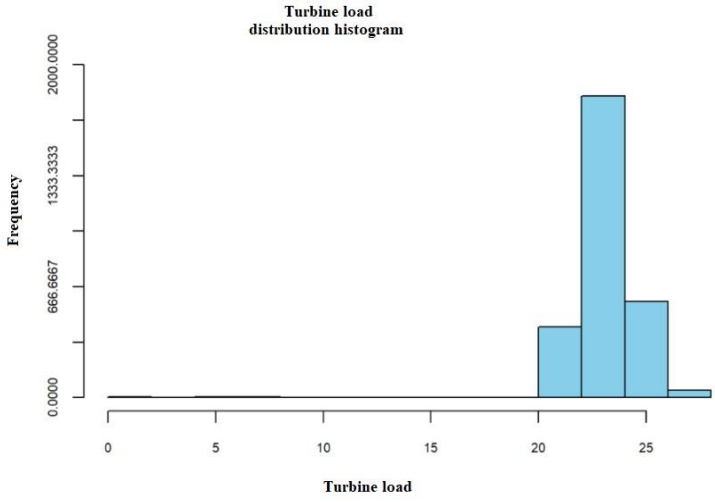

Figure 37. Turbine load distribution histogram for the test period.

A comparison of the predicted and actual values of the turbine load and the model's residuals for the test dataset are presented in Figure 38. The large model errors are noted when the turbine load was below 13 MW. This observation is explained by the fact that according to the production technology, the turbine does not work in such modes. Consequently, these operation modes are not represented in the training dataset, therefore, the model has not learned to predict values for these modes. At loads above 13 MW, the model predicting quality improves, and the errors become smaller.

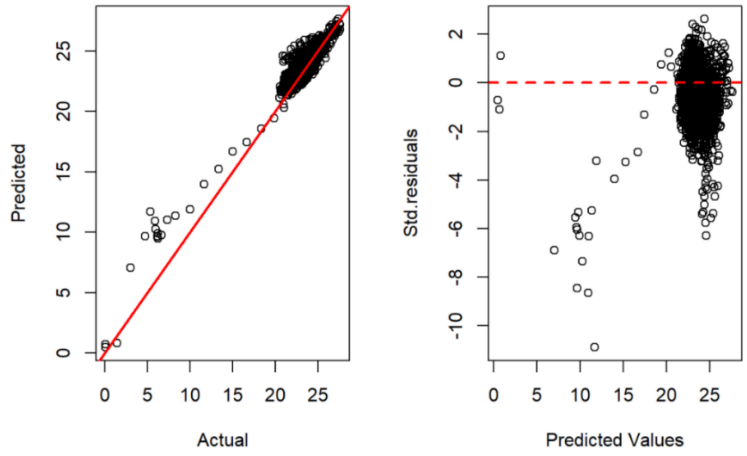

Figure 38. Predicted values distribution and model errors for the test dataset.

On the test data model's R^2 = 89.72% and RMSE = 0.64.

Figure 39 shows the predicted (left graph) and the actual turbine load values (right graph) for the test dataset. From these graphs it can be concluded that at rated turbine loads, the model predicts values with high accuracy. No jumps in the predicted values were found.

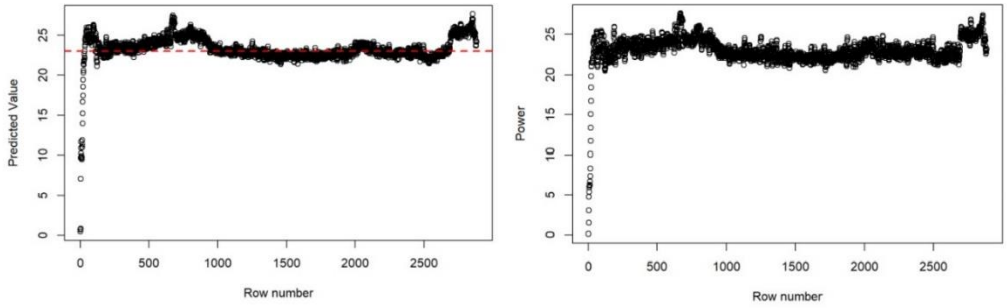

Figure 39. Predicted (left graph) and actual (right graph) turbine load distribution.

5.5 Methods for improving the model quality

An initial analysis of the linear regression model performance showed that some disadvantages need to be eliminated. These disadvantages are the lack of accuracy in predicting zero values and operating modes below 13 MW. Additionally, was found model attributes high multicollinearity. To correct these disadvantages, numerous attempts have been made to improve the model quality.

5.5.1 Data normalization

Attribute's normalization allows to bring all variables to one unified scale. Data normalization was performed by using the Z-scaling method, which is recommended for

most cases [12]. After normalization, the data has a mean of 0 and a standard deviation of 1. After transforming the variables, a new linear regression model was fitted to the training dataset. The model was tested on the test dataset, which has also been previously normalized. For the normalized test dataset RMSE was 5.86. Thus, the attributes normalization led to an increase RMSE for the predicted values. This finding was unexpected because normalization usually increases the model quality.

5.5.2 Data transformation

The aim of data transformation is to obtain a symmetric distribution of the target variable Y [25]. In this case, all the turbine load Y values were raised to the third power. In Figure 40 on the left side presented the Y distribution before the transformation. It is seen that the data is shifted to the left. The Y distribution after the transformation is seen in Figure 40 on the right side. After the transformation Y has an approximately normal distribution.

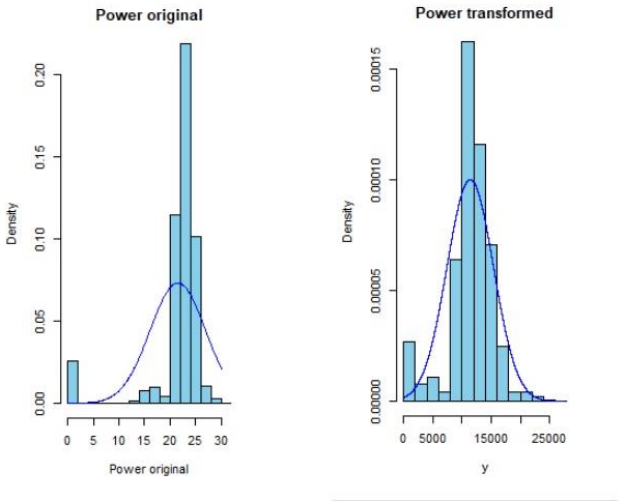

Figure 40. Attribute Y distribution before and after transformation.

A new linear model was created based on the transformed training dataset and tested on the test dataset, which was previously transformed in the same way. After obtaining the predicted values, the back-transformation was applied to return values to the original scale. After the back-transformation RMSE for the test data set was 0.91. This metric turned out to be worse than on the data without transformation.

5.5.3 Using the cross-validation function

The essence of this method is that the training dataset is split into several parts. One of these parts is used to test the model, and the rest are used to train the model. Then the parts are interchanged in such a way that each of them will consistently participate both

in training the model and in testing it. The results of each cycle are averaged together for a final score. Using the cross-validation with 5 layers and 10 repetitions, the linear regression model was validated based on the training dataset. The model r-square R^2 remained at 89.64% and RMSE was 0.64. After the cross-validation, the model accuracy remains at the same level, which indicates that there is no overfitting of the model.

After applying these methods, the model quality metrics either deteriorated or remained at the same level. Thus, it was decided to use the original model in further research, without any transformations.

5.5.4 Additional linear regression model training

During the research, it was repeatedly noticed that the linear regression model did not perform with the prediction of turbine load zero values. To eliminate this disadvantage, additional data were collected for the period of turbine downtime. Data was collected for the period from 17/01/2021 to 22/01/2021. Additional data has been added to the original training dataset. Based on the combined dataset, a new linear regression model was created. Figure 41 shows the predicted values for the training dataset.

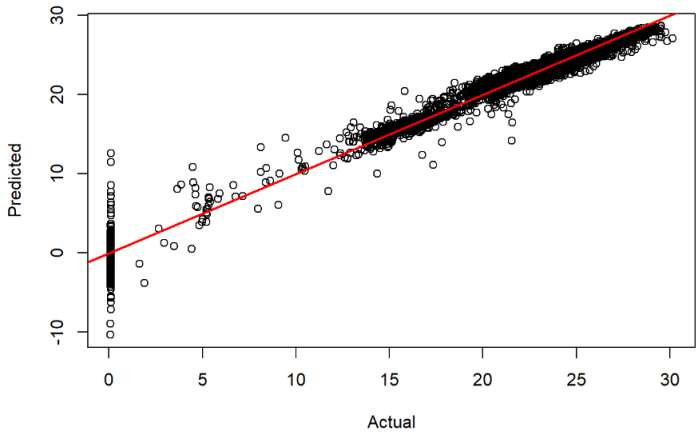

Figure 41. Predicted values distribution for training dataset

Even though the training sample contains data for 4 days of turbine downtime, visual graph inspection showed no improvement in predicting zero values. Then the model was validated on test data. The model r-square increased to 90.79%. The RMSE decreased to 0.57. Thus, the new data introduction into the training dataset improved the linear model quality.

5.6 Process equipment fault simulation

In this chapter, the focus has been on analyzing the predicted value behavior when some equipment faults occur. It is necessary to determine how much the predicted value will deviate from the initial one in the event of any malfunctions on the equipment occurs. Several experiments were performed, the essence of which was to activate the malfunctions occurrence on equipment by simulating some attributes in the test data. In the test samples, some data was manually modified to further analyze the predicted value behavior on the changed data. Several small sub-datasets were taken from the total test dataset, each of which contained 100 observations. The timing of these sub-datasets was randomly selected. In each of these sub-datasets, the data was unchanged in the first 50 observations, and manual changes were made to the data from 50 to 100 rows.

For the first experiment, the temperature sensor GX01BT012 (Rear axial bearing temperature) was simulated (increased) from 76 to 86 degrees. This attribute was selected based on the largest correlation with the turbine load (0.93). Based on this sub-dataset, the turbine load predicted values ware obtained. The left graph in Figure 42 shows the bearing temperature rise from 76 to 86 degrees in the last 50 observations. The right graph in Figure 42 shows how the model predicted values reacted to the temperature change by increasing the load from 22.7 MW to 26.7 MW.

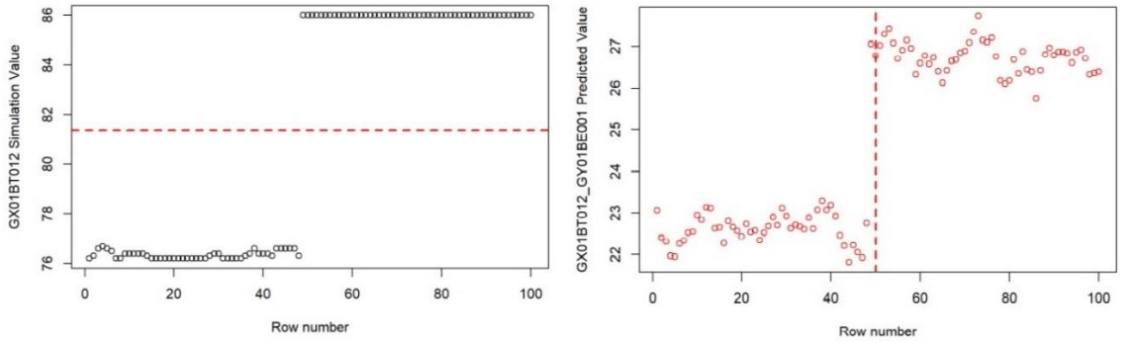

Figure 42. Bearing temperature sensor GX01BT012 increase simulation.

The turbine load average predicted value for the last 50 observations has increased by 4.36 MW. Thus, simulating a 10-degree temperature rise in one of the bearings led to a significant increase in the predicted turbine load.

Next an increase in the vibration sensor values GY01BV007 (Full vibration of the generator - in front of the axial) from 0.3 to 2 mm/s step was simulated. This time, the attribute was selected based on the smallest correlation with the turbine load (0.48). This

simulation result is shown in Figure 43. The left graph shows the process of increasing the generator vibration sensor values from 0.3 to 2 mm/s in the last 50 observations. The right graph shows that the model predicted values responded to the vibration level change by increasing the load from 23.2 MW to 25.3 MW. Thus, simulating an increase in generator vibration sensor values by 1.7 mm/s led to an increase in the predicted turbine load by 2.67 MW.

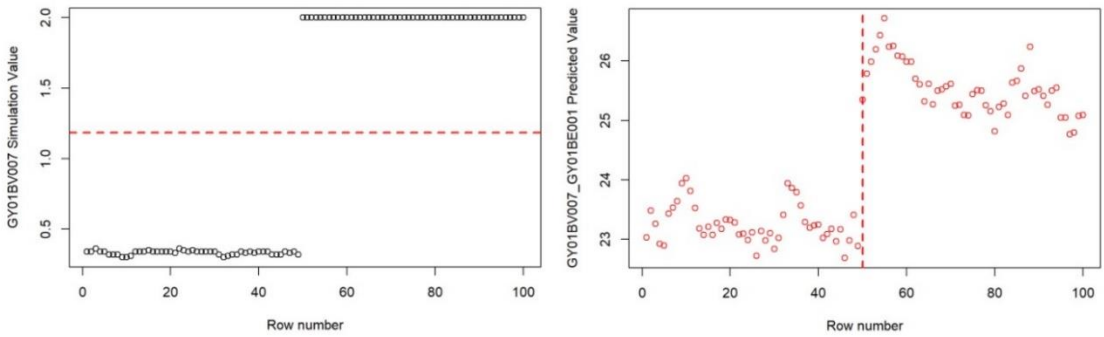

Figure 43. Generator vibration sensor GY01BV007 increase simulation.

The last experiment was to simulate both an increase in the temperature sensor GX01BT012 values and an increase in the vibration level GY01BV007 sensor values using the simulations from the previous experiments. The purpose of this experiment was to test the hypothesis of the cumulative effect of two or more attributes on the model's predicted values. It is assumed that the combined effect of two or more attributes will have a cumulative effect on the turbine load predicted values. This simulation result is shown in Figure 44.

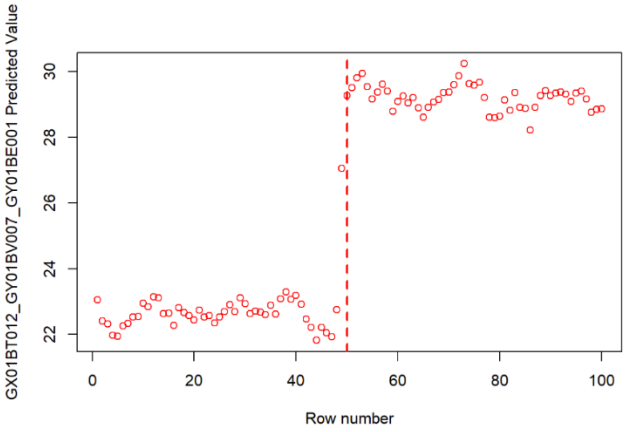

Figure 44. Temperature bearings GX01BT012 and generator vibration GY01BV007 sensors increase simulation.

The increase in bearing temperature joined with increased vibration levels, increased the predicted load from 22.8 MW to 29.2 MW. Thus, simulating an increase in the bearing

temperature by 10 degrees with a simultaneous increase in the generator vibration level by 1.7 mm/s led to an increase in the predicted turbine load by 6.85 MW. Hence, changing the values of two or more attributes leads to a joint, stronger effect on the model predicted values. It can be expected further that a greater number of measurement deviations from normal values will lead to an even greater effect on the predicted load values. This observation is valuable for applying the model to the enterprise management system.

5.7 Alternative machine learning models

Several attempts have been made to improve the original linear regression model, but no significant improvement has been achieved. Therefore, alternative models using other ML algorithms were created based on the same data sets. All created models and related plots can be found via the link in Appendix 2.

5.7.1 Random Forest model

To create a Random Forest model, the initial number of regression trees was set to 400. Figure 45 shows that the model error after 200 trees did not decrease. Therefore, the number of regression trees was reduced to 200 to optimize computing performance.

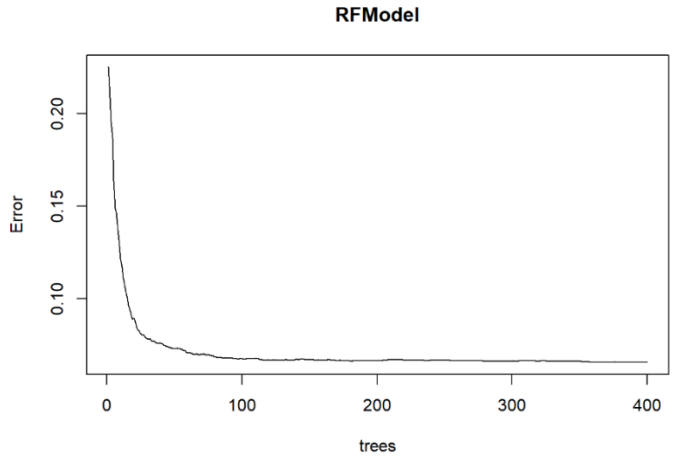

Figure 45. Model errors concerning the number of "decision trees".

A comparison of the predicted and actual values and the Random Forests model's residuals for the test dataset are presented in Figure 46.

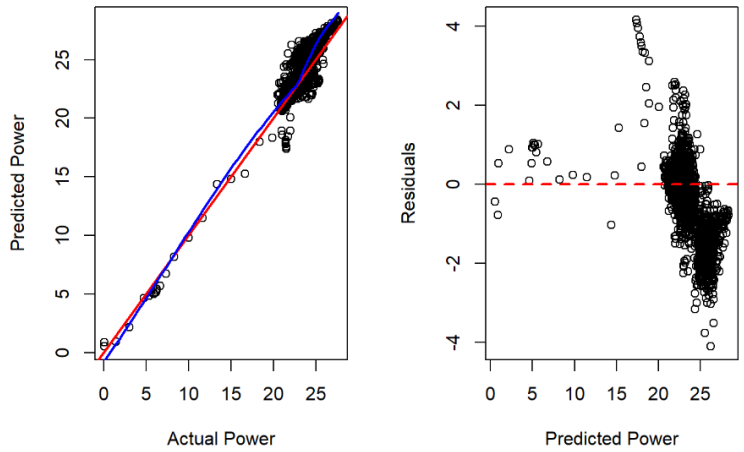

Figure 46. Random Forest model predictions and residuals for the test dataset.

Several attempts were made to improve the Random Forest model by changing the parameters. The best results achieved for the model RMSE and r-square were 0.85 and 86% respectively. The Random Forest model performs worse on the test dataset compared to the linear regression model.

5.7.2 Neural Network

A comparison of the Neural Network model predicted and actual values and the Neural Network model residuals for the test dataset are presented in Figure 47.

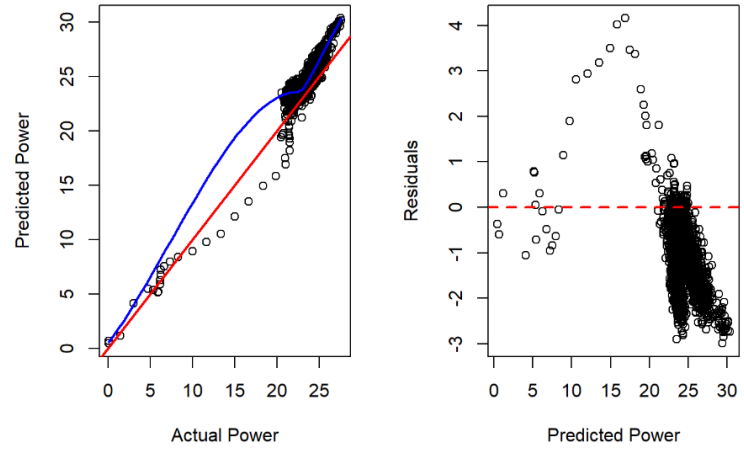

Figure 47. Neural Network model predictions and residuals.

The RMSE of the Neural Network model was 1.24. The model r-square R^2 was 87%. The performance of the Neural Network model turned out to be worse than that of the linear regression and Random Forest models. The high RMSE of the Neural Network model

does not allow it to be successfully used in the method for detecting anomalies on equipment.

5.7.3 Nonlinear regression model

The nonlinear model differs from the linear model by the presence of additional polynomials of different degrees based on some of the original predictors. Component residual plots can be used to decide, which polynomial attributes and of what degree should be add to the model (Figure 48). All plots can be found in the link in Appendix 2.

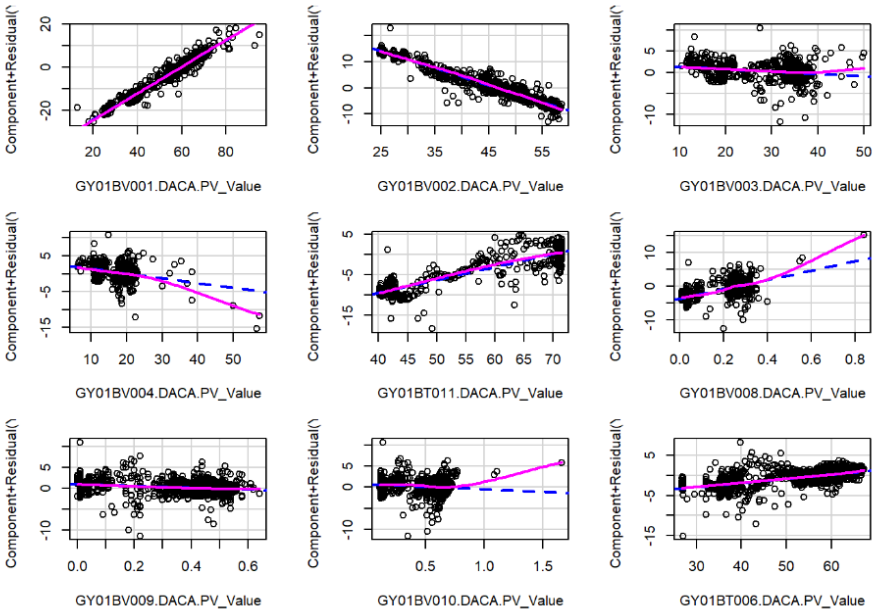

Figure 48. Attributes residuals distribution.

A comparison of the nonlinear model predicted and actual values and the model residuals for the test dataset are presented in Figure 49.

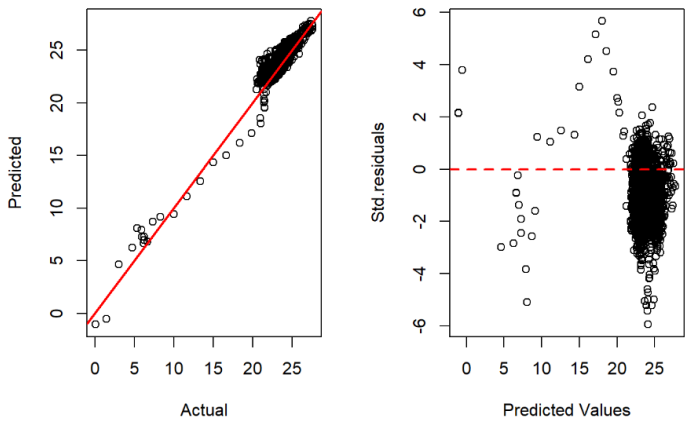

Figure 49. Nonlinear regression model predictions and residuals.

The RMSE for the nonlinear regression model was 0.72. The model r-square R^2 was 91%. The linear regression model, which has been further trained by adding new data into the training dataset has the lowest RMSE so far.

5.7.4 MARS and PPR models

Additionally, two machine learning models, MARS and PPR, were tested. The RMSE of the MARS model on the test data was 0.64 and model r-square R^2 was 89.65%. The result of the MARS model turned out to be worse than its predecessors.

The PPR model RMSE error on the test data was 0.52 and model r-square R^2 was 94.7%. This is the best metric of all the previously reviewed models. The visualization of the PPR model on the training data shows the best results in predicting zero values (Figure 50). It shows a strong correlation between the predicted and actual load values.

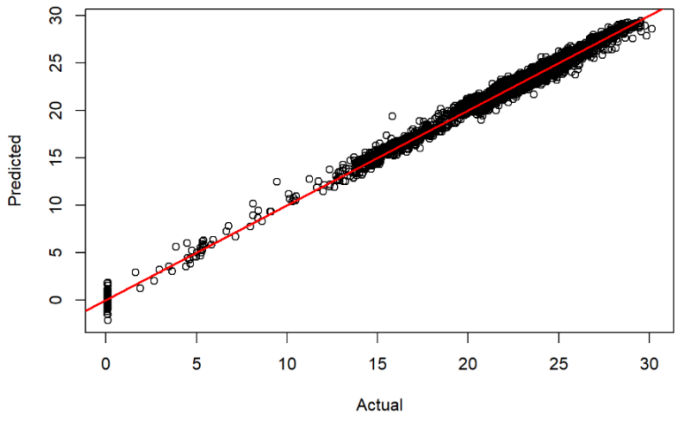

Figure 50. PPR model visualization on training data.

For the test date the same conclusions can be made (Figure 51).

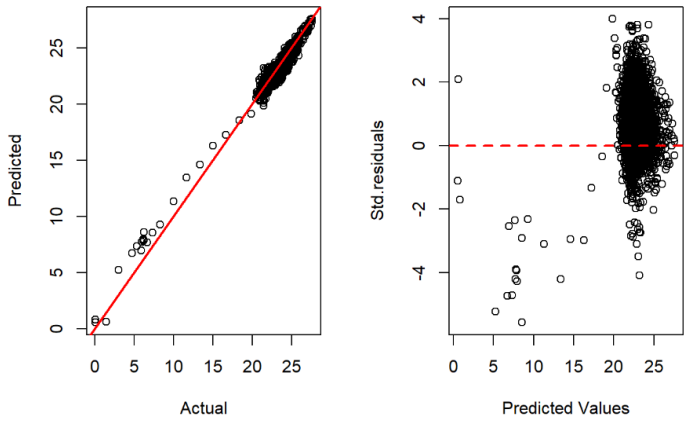

Figure 51. PPR model visualization on test data.

This is the best result among all the previously reviewed models.

5.8 Linear regression model integration into the enterprise management system

Previous research has identified the ML model leading in terms of high r-square and low RMSE. This model became the PPR. But in the applied sense, the PPR model is difficult to implement into an enterprise management system. The obtained models' comparison shows that the linear regression model is slightly inferior to the PPR model in terms of predicting quality. However, at the same time, the linear regression model has a significant advantage over other models considered in the research. This advantage lies in the easy integration of the linear regression model into the enterprise management system. The linear regression R model final formula for the research data is shown in Figure 52.

Y (GY01BE001_Value) = GX01BZ001_Value + GX01BZ002_Value + GX01BV005_Value + GX01BV007_Value + GX01BT011 _Value + GX01BT012_Value + GX01BT013_Value + GX01BV001_Value + GX01BV002_Value + GX01BT014_Value + GX01BV003_Value + GX01BV004_Value + GX01BV008_Value + GX01BV009_Value + GX01BV010_Value + GY01BV005_Value + GY01BV006_Value + GY01BV007_Value + GY01BT010_Value + GY01BV001_Value + GY01BV002_Value + GY01BV003_Value + GY01BV004_Value + GY01BT011_Value + GY01BV008_Value + GY01BV009_Value + GY01BV010_Value + GY01BT005_Value + GY01BT004_Value + GY01BT003_Value + GY01BT002_Value + GY01BT001_Value + GY01BT012_Value + GY01BT013_Value + GY01BT014_Value + GX01BT016_Value + GX01BT015_Value + GX01BT018_Value + GX01BT019_Value + GX01BT314_Value + GX01BP309_Value

Figure 52. Linear regression model formula.

Using above formula, the predicted value of Y (GY01BE001) is calculated based on all remaining variables in the formula. A certain coefficient is assigned to each attribute/variable, which determines the influence of this attribute on the final calculation result. These coefficients were presented in the chapter devoted to the linear regression model creation. Thus, using this formula with the coefficients applied to each attribute, the linear regression model was integrated into the enterprise management system. At the oil plant, the industrial automated system Honeywell is used to control the process. For integration into the enterprise automation system, was selected an additionally trained linear regression model after the removing insignificant components. This allows to optimize the number of attributes involved in the calculations and reduces the load on the Honeywell controllers calculating capability. AUXCALC mathematical modules were used for calculations in the Honeywell engineer software environment. This module has restrictions on the number of inputs, so the general formula was divided into several fragments. Measurements from turbine and generator sensors were connected to each

module following the regression formula. The calculating result of each module was transferred to the next module. Thus, the entire linear regression formula was reproduced in the Honeywell control system program logic. Figure 53 shows a calculations fragment in the first logical unit AUXCALCA1.

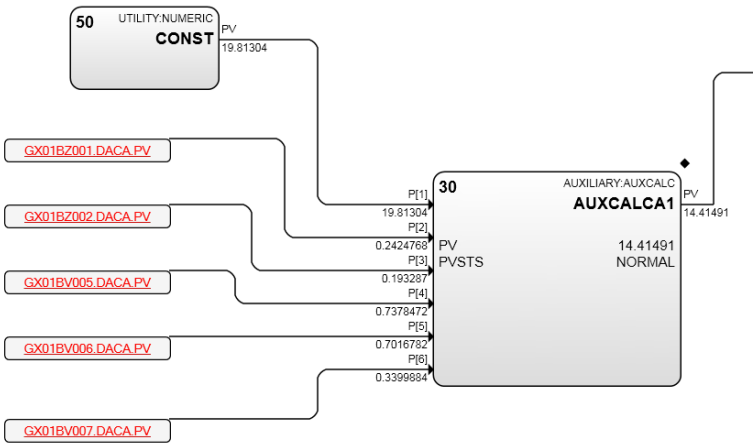

Figure 53. Calculations fragment in the first logical unit AUXCALCA1.

The module configuration parameters indicate the formula by which the AUXCALCA1 module calculates. The formula for the AUXCALCA1 module is as follows:

P[1] + P[2]*(12.170114) + P[3]*(-11.290135) + P[4]*(-3.028353) + P[5]*(-0.745362) + P[6]*(-10.027987)

The first calculation result after AUXCALCA1 module is passed to the second module AUXCALCA2 (Figure 54).

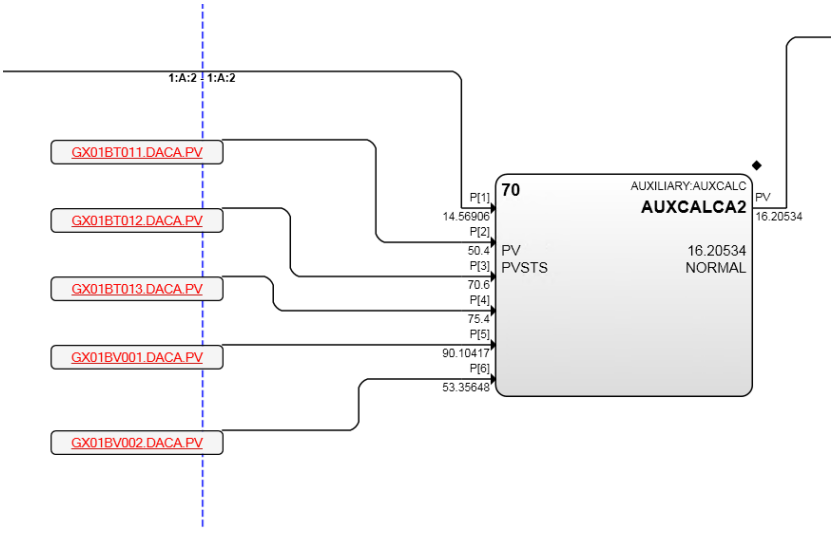

Figure 54. Calculations fragment in the second logical unit AUXCALCA2.

The formula for the AUXCALCA2 module is as follows:

P[1] + P[2]*(-1.200780) + P[3]*(0.437879) + P[4]*(0.558494) + P[5]*(-0.009829) + P[6]*(-0.187109)

The calculation in the rest of the mathematical modules AUXCALC similarly takes place, until all attributes are involved in the calculations. In each of the math modules, the formula contains the coefficients corresponding to the connected attributes. The final value is calculated at the output of the last mathematical module, which is the turbine load predicted value. The program's general view for calculating the predicted value is shown in Figure 55.

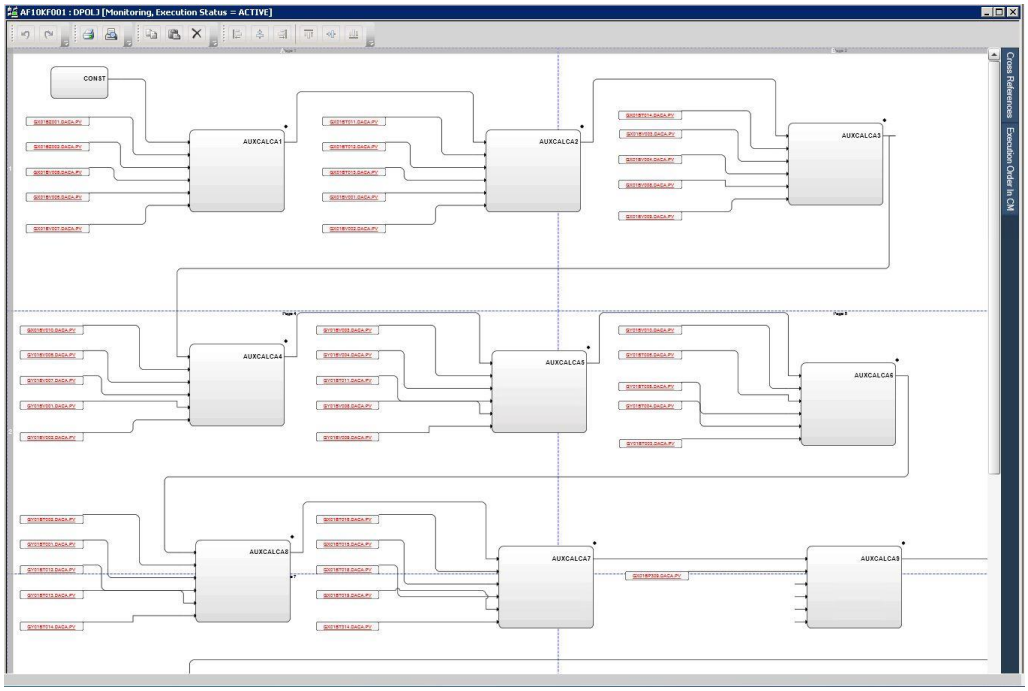

Figure 55. Program general view for calculating the predicted value.

Additionally, the calculated output value was averaged using the ROLLAVGA module (Figure 56).

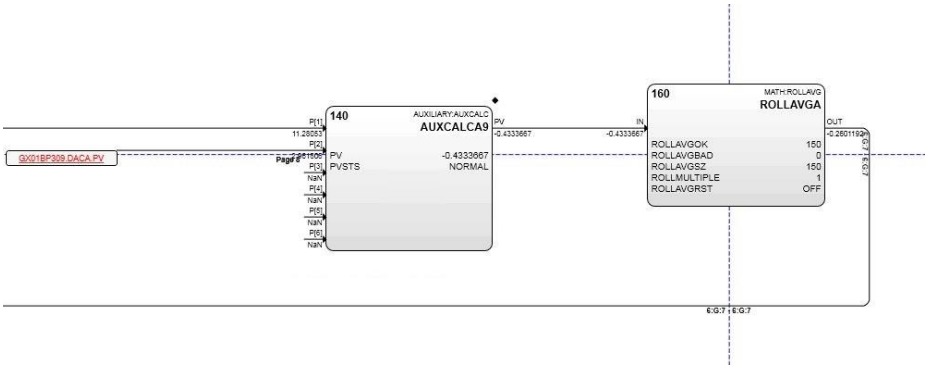

Figure 56. ROLLAVGA module for averaging the predicted value.

ROLLAVGA averages the output to prevent sudden jumps in the predicted values during turbine transients' modes such as a sudden stop or a load rise. The averaging time was adjusted by changing the module setting parameters. For this research, the averaging time was chosen equal to 5 minutes. After averaging, the calculated predicted value was connected to the DACA module input. The DACA module is designed to create an interface for displaying measurements on a certain operator screen. Also, this module formed the organization of recording measurements in the historical database. Recording values in the historical database allows displaying measurements in the form of a graph for a certain period. After the DACA module, were generated emergency messages to operating personnel about the deterioration of the situation at the turbine. For this operation, the predicted turbine load value was compared with the actual load in the AUXCALC10 module (Figure 57).

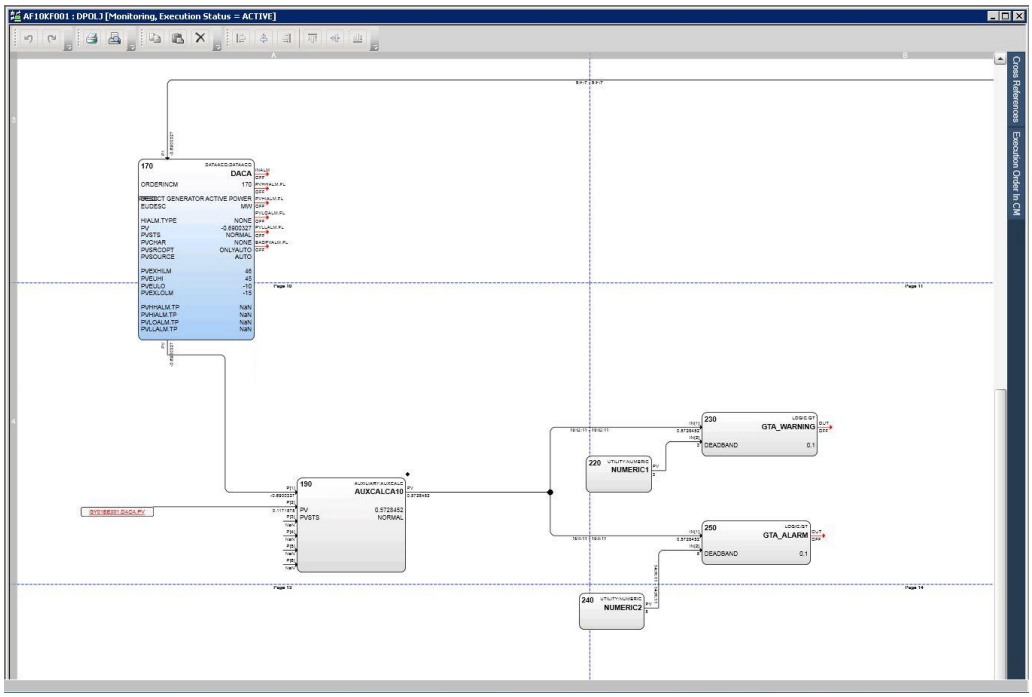

Figure 57. Emergency messages formation to operating personnel.

At the output of the AUXCALCA10 module, was formed absolute deviation between predicted and actual values. After the AUXCALCA10 module, two comparison modules were organized. They are designed to generate a warning and alarm signals for the enterprise operating personnel. The warning limit was chosen to be 3 MW. The alarm limit was chosen to be 6 MW. The signal boundaries were selected based on the analysis of predicted value behavior during the fault's simulation.

After the calculation, the predicted values of the turbine load were recorded on the enterprise information server. Below are the comparative graphs of the actual (blue graph) and predicted (green graph) turbine load (Figure 58).

Figure 58. Comparative graphs of actual and predicted turbine load.

According to the graph, it can be determined that the average deviation of the predicted load value from the actual value was fixed at 1.5 - 2 MW. This deviation is acceptable and did not indicate any negative processes. Below is a more detailed graph of the time when the turbine was stopped. The graph shows that with a certain delay, the predicted value begins to follow the actual turbine load. Additionally, fluctuations in the predicted values at zero loads were noted (Figure 59).

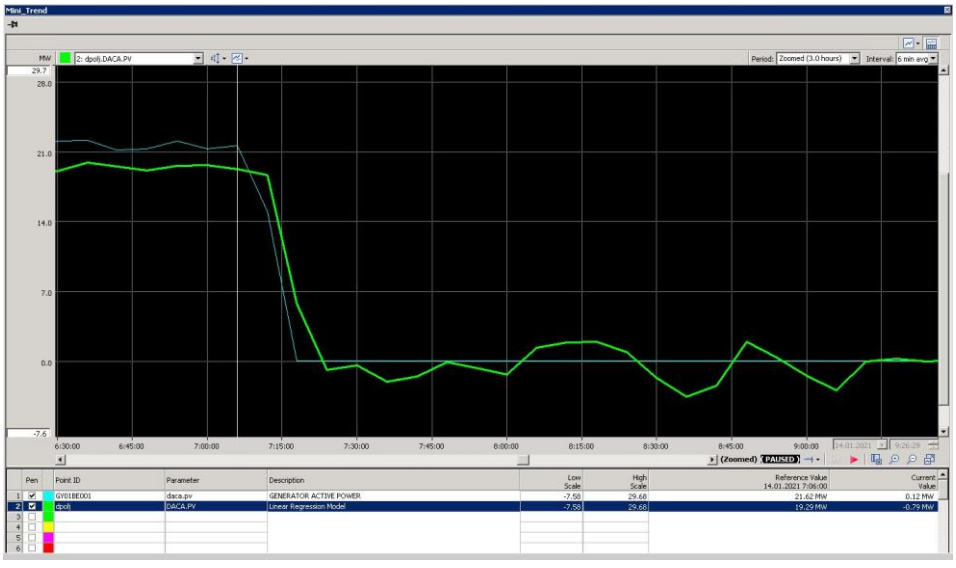

Figure 59. Turbine shutdown moment.

5.9. Conclusions regarding the method of detecting for anomalies on equipment

The statistical performance of created ML models is summarized in Table 7. The model with the best results is marked in yellow. The model that has been integrated into the enterprise management system is highlighted in green.

Table 7. Generalized model results.

Model	**Dataset type used**	**RMSE**	R^2
Basic linear regression model	Test dataset from 18/12/2020 to 21/12/2020	0.64	89.72%
Basic linear regression model	Test dataset from 18/12/2020 to 21/12/2020. Data normalization	5.86	-
Basic linear regression model	Test dataset from 18/12/2020 to 21/12/2020 Data transformation	0.91	-
Basic linear regression model after the cross-validation function	Test dataset from 18/12/2020 to 21/12/2020	0.64	89.64%
Additionally, trained linear regression model.	Test dataset from 18/12/2020 to 21/12/2020	0.57	90.79%
Model Random Forest	Test dataset from 18/12/2020 to 21/12/2020	0.85	86%
Neural Network	Test dataset from 18/12/2020 to 21/12/2020	1.24	87%
Nonlinear regression model	Test dataset from 18/12/2020 to 21/12/2020	0.72	91%
Model MARS	Test dataset from 18/12/2020 to 21/12/2020	0.64	89.65%
Model PPR	Test dataset from 18/12/2020 to 21/12/2020	0.52	94.7%

During the research, several attempts were made to improve the linear model:

- By normalizing data. This experiment showed that the prediction error increases.
- By raising the actual load values to the third power. This experiment showed that the prediction error increases.
- By validating the existing model using the cross-validation function. This experiment showed that the prediction error changes insignificantly, which indicates that there is no overfitting of the model.
- By additional training of the linear regression model by introducing new data into the training set. This experiment improved the linear regression model performance.

The following results were obtained for simulating equipment faults:

- Individually, a change in temperature sensor values led to a difference in the predicted and actual load values equal to 4.36 MW.

- Individually, changing only the vibration bearing sensor values led to a difference in the predicted and actual load values equal to 2.67 MW.
- Anomalies simulation in two variables, in particular an increase in the bearing temperature by 10 degrees and an increase in generator vibration by 1.7 units, led to a significant increase in the predicted turbine load. The average difference between the predicted turbine load and the actual load for the simulated period was 6.85 MW.

The hypothesis of the cumulative effect of two or more attributes was confirmed. Changes to the values of two or more attributes have a stronger combined effect on the model predicted values. It can be expected further that a greater number of measurement deviations from normal will lead to an even greater effect on the predicted load values. This observation was used to apply the model in an enterprise management system.

The Random Forest model showed the best results on the training data, but on the test data the model quality significantly deteriorated, and as a result, the model quality metrics were inferior to linear regression model.

The performance of the Neural Network model turned out to be worse than that of the linear regression and Random Forest models. The Neural Network model high RMSE not allow it to be successfully used in the method of detecting for anomalies on the equipment.

The linear regression model has been integrated (programmed) into Honeywell's automation plant management system. The linear regression model was chosen for integration because of its rather easy execution in the programming environment. This was described in detail in chapter 5.8. Honeywell's system calculates the turbine predicted load in real-time. By calculating the difference between the predicted load and the actual load values, the deviation was obtained in the form of a numerical value. The average deviation of the predicted load value from the actual value was fixed at 1.5 - 2 MW. This deviation is acceptable and does not indicate any negative processes. A greater difference between the values will lead to the generation of warning signals to the operating personnel about the occurrence of an anomaly on the equipment.

6 Work development direction

The material in this chapter is devoted to further project development. This material is not implemented yet, but it can be considered as a theoretical starting point. As the result of the model fitting process, described in the section 4.5, the following linear regression model equation was obtained for the case-based fault detection method: $Y = -44.53 + [Rotation\ speed] * 0.004793 - [Current\ to\ valve1] * 0.03096 + [Current\ to\ valve\ 2] * 0.03566 + [Current\ to\ valve\ 3] * 0.0377 + [Positioner\ 1] * 0.03835 + [Positioner\ 2] * 0.004589 + [Positioner\ 3] * 0.1745 - [Turbine\ controller\ output] * 0.07026 - [Power] * 0.01355 + [Power\ regulator\ mismatch] * 0.04785$

For the equipment anomaly detection method, a similar formula has been implemented in Honeywell's enterprise management system and has proven to work. Therefore, the formula derived from the case-based fault detection method can be similarly implemented in the process control system (DCS). Thus, the system will calculate the predicted value of Y in a range [0, 1] in real-time.

To create alarms about equipment wear-out, it is necessary to divide the possible range of allowable Y values into classes, for example:

- $Y \in [0, 0.4)$ - the problem does not exist,
- $Y \in [0.4, 0.7)$ - the problem exists, not critical,
- $Y \in [0.7, 1]$ - the problem exists, it is critical.

Now, it is possible to create a logical module in the control system, the task of which will be to calculate the predicted value of Y in real-time, compare with the specified limits and generate final information messages to the operating personnel (Figure 60).

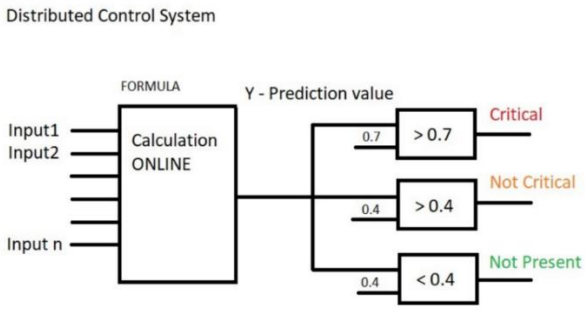

Figure 60. Logic for alarms creation.

At this stage of the project, this section is a theoretical hypothesis that will need to be confirmed or refuted. To confirm the hypothesis, it is necessary to organize the recording of the predicted values of Y in the historical database for further observation. Over time, when the equipment is in operation, occur natural wear-out of mechanical elements. If the hypothesis is correct, then the predicted values of Y should increase smoothly towards the 1 at the end of the equipment operation cycle (before the repair). After a major overhaul, the predicted values Y should lie around 0. The observation results will give grounds to draw certain conclusions about the efficiency of this method.

The case-based fault detection method used Estonian Power Plant turbine No. 8 operation data. Working with the data, it was possible to draw an important conclusion that each time after repairs, the behavior of the equipment changes in the process. ML model classified repaired state of with a large difference for the test datasets, which were taken immediately after the repair, but in different years. This means, that the method of detecting the faulty state of equipment using ML algorithms requires further development. To minimize errors in the case-based fault detection method, it is necessary to create a training dataset in a more complex way. It is necessary to collect data several days before and after each cycle of equipment operation. This means collecting data for a week of equipment operation before and after the next major overhaul. Below is the algorithm for creating a training dataset for the model (Figure 61).

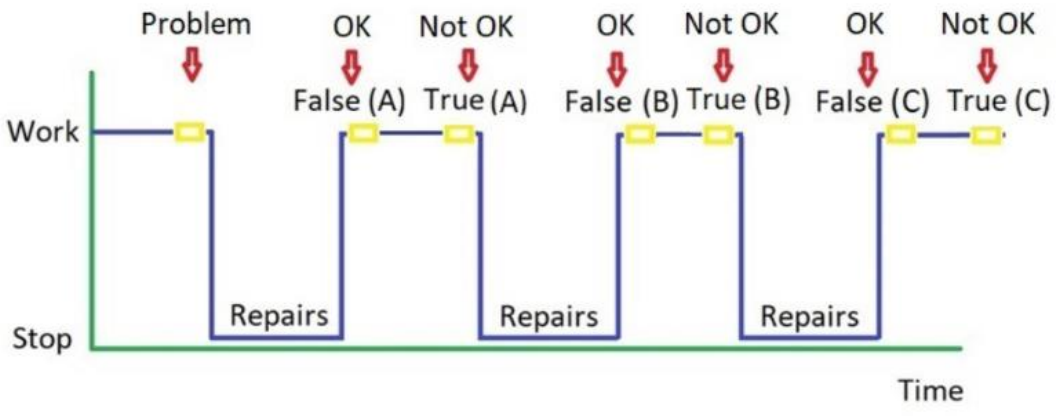

Figure 61. Algorithm for creating a training sample for the model.

The obtained data samples must be combined into one generalized dataset, which will be used as training dataset for various ML models (Figure 62).

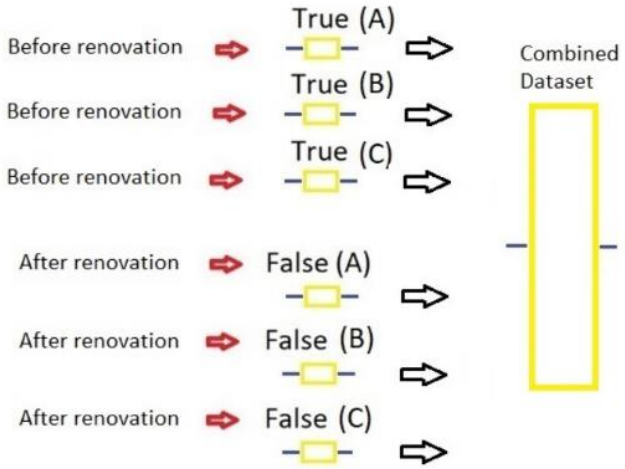

Figure 62. Algorithm for combining data into one generalized dataset.

The expected result of this action is that machine learning algorithms will be able to learn to perceive the healthy state of the equipment as one class, regardless of in which overhaul cycle the sample was collected. The deteriorated condition of the equipment, accordingly, should be perceived as a different class. Significant differences must form between the different classes. In this case, the presented fault detection method will give an expected result.

In this investigation, there is no information about overhauls volume at turbine No. 8. The research result showed that the predicted values after some overhauls did not match with the expected result. The reason for the obtained observations can be explained by the assumption that in some years the overhaul was not carried out fully. Therefore, the model, immediately after the repair, predicts the worn-out equipment condition. At this project stage, these assumptions can neither be confirmed nor refuted. Therefore, the above material remains a theoretical starting point for further research. The key point for the precedent-based fault detection is the need to classify the samples based on information from repair services about the overhauls volume carried out in each cycle of equipment operation. Only in this case it will be possible to create a proper training dataset, and successfully use it in the fault detection method.

7 Conclusions

In this thesis two methods of industrial equipment fault detection were studied: the fault detection by precedents and the detecting anomalies on the equipment. Based on the research results, can conclude that both fault detection methods can be successfully used to generate a correct and accurate message about equipment failures. Experimental results show that machine learning algorithms can recognize changes in the state of equipment during its operation.

However, the precedent-based fault detection method revealed some disadvantages that must be eliminated before using it in a real process. It is necessary to classify the samples more carefully based on information from the repair services. Only then will be possible to minimize errors in data classification and this will make it possible to successfully use the fault detection method on equipment. Before introducing precedents-based fault detection method into the technological process, are required additional tests at the enterprise equipment. The investigation showed that after the new data introduction, the predictive model becomes stronger and more accurate. This suggests that prediction results will get even better over time with adding new, additional data to the model. Finally, this research provided insight into the ongoing changes in equipment after overhaul. This work can serve as a basis for further investigation for finding a more exact method for identification of equipment faults.

The main aim of this work was to develop a methodology for detecting malfunctions of industrial equipment based on ML methods and technologies for its implementation at the enterprise. The ML model developed as the result of the research of the method of detecting anomalies has been successfully integrated into the enterprise management system. The malfunction development will provoke a change in the several sensors response close to the faulty node. Due to the cumulative effect, the several attributes influence on the model predicted value would be significant. Already in the early stages, a warning signal would be generated to the operating personnel about the equipment abnormal behavior. This point has been created two alarm levels in the Honeywell management system. For the warning signal, the deviation limit was chosen equal to 3 MW. For the alarm, the deviation limit was chosen to be 6 MW. Such a deviation will signal that anomalous appearances are beginning to occur on the equipment. Not in all cases, the deviation will be associated with a malfunction. These can be any changes in

the equipment operating modes. A situation may arise when some operation mode is not included in the training dataset. In this case, for the predicting model, this operation mode will be considered as a deviation. It may be necessary to retrain the linear regression model with the new operation mode involvement in the training dataset. When the operating mode did not change, and the deviation of the actual load from the predicted value continues to increase, it is necessary to examine the sensor's measurements more closely. There is a high probability that mechanical faults begin to develop on the equipment. The implemented method of detecting anomalies makes it possible to notice negative changes at the earliest stages of their formation. This will make it possible to take appropriate steps to prevent further negative developments. The preventive steps are taken at the initial stage, ultimately, will give an economic effect during the operation of the equipment.

7.1 Expected benefit analysis and business case

During operation, the equipment is exposed to a large number of negative factors. Such factors are mechanical loads, which contribute to a change in the material strength characteristics and a change in the initial geometric dimensions. The technical condition of the equipment is deteriorating. After overcoming a certain limit, equipment failure occurs, that is, the inability to perform its functions. To restore operability, it is necessary to carry out maintenance and repair (MRO) of the equipment. The negative effects on the equipment are accidental. Therefore, equipment failure can occur at any time in the operation cycle. The following are the well-known strategies applicable to equipment maintenance and repair (Figure 63).

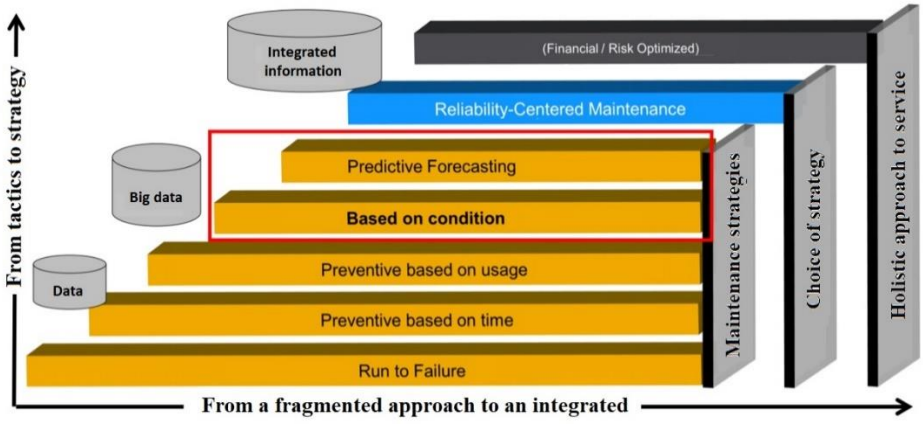

Figure 63. Maintenance and repair strategies [5].

Currently, the company applies three strategies applicable to maintenance and repair. The first strategy is called Work to Failure. This strategy is the simplest and does not require tracking the equipment's technical parameters. This strategy applies to ancillary equipment, the failure of which, in general, is not capable of stopping the production process of the entire plant. But in some situations, under an unfavorable set of circumstances, such a strategy can lead to long downtime. Therefore, if possible, this strategy should be avoided when organizing maintenance. The second and third strategies that are used in the enterprise are called "Preventive maintenance". The essence of these strategies is that equipment repairs are planned. The differences lie in the method of calculating the mechanisms operating time. The timing calculation is based on the equipment operation calendar cycles. The time-based strategy is based on the machine's actual operating time calculation. For this, some counters are provided in the control systems, which are activated at the moment when receiving a signal about the mechanism's active state.

All the above strategies have significant disadvantages. The equipment operation to failure provokes a chain negative reaction of one faulty unit to the system neighboring elements. The failure development stages are shown in Figure 64.

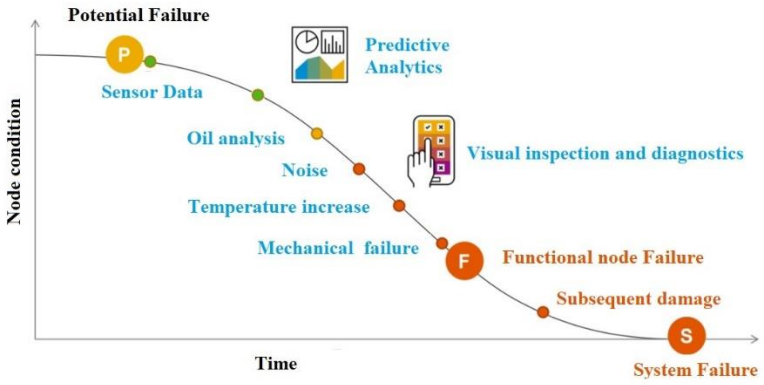

Figure 64. Failure development stages [5].

Before finally falling, some faulty equipment units affect nearby elements for a long time. Thus, subjecting them to increased wear-out. Ultimately, when the complete failure moment of the initial node comes, the system neighboring elements have already become unusable. In this case, it is necessary to repair a lot of nodes, which, in economic terms, is significantly more expensive. The method for determining a malfunction at the initial stage of its development will allow the equipment to be disconnected from work in time.

In this case, only the primary unit needs to be repaired, and the remaining elements would remain in good technical condition.

Scheduled maintenance has the disadvantage that by the time/date of repair, the resource of some equipment has not yet been exhausted (Figure 65).

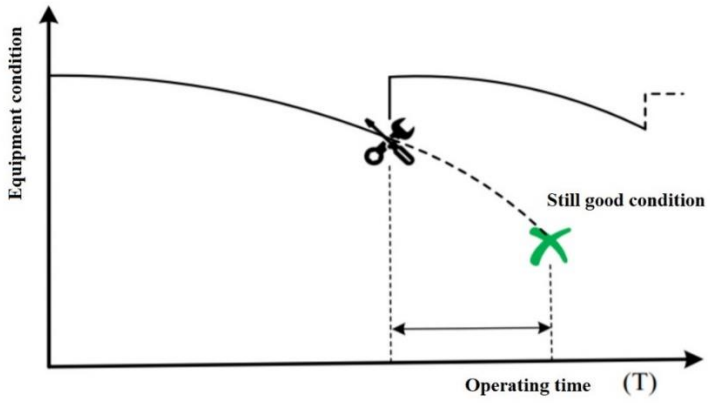

Figure 65. Changing the state of equipment [26].

This leads to ineffective use of the enterprise's repair facilities. Therefore, when organizing maintenance, it is necessary to use other strategies based on more complex algorithms.

The use of ML models allows to implementation strategy called “Repair as per condition” in the enterprise's daily life. The essence of this strategy is to apply a predictive approach to equipment repair and maintenance, which makes it possible to determine in advance possible equipment failures (Figure 66).

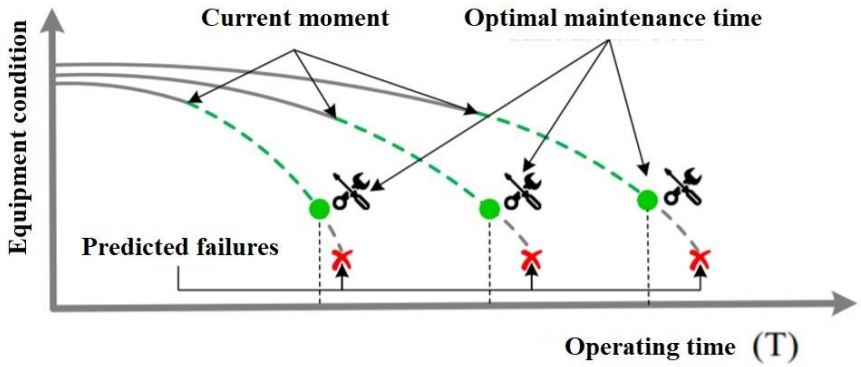

Figure 66. Organization of predictive maintenance strategy [26].

The equipment is operated until the moment of the predicted failure. The alarm threshold is selected individually for each type of equipment. The main criterion for organizing a

threshold value is to analyses the influence of a node on neighboring elements. The main task is to maintain the equipment in operation until the moment when a change in the state of one node begins to negatively affect the entire system. Based on the model's predicted values, possible schedule equipment repairs at the right time (Figure 67).

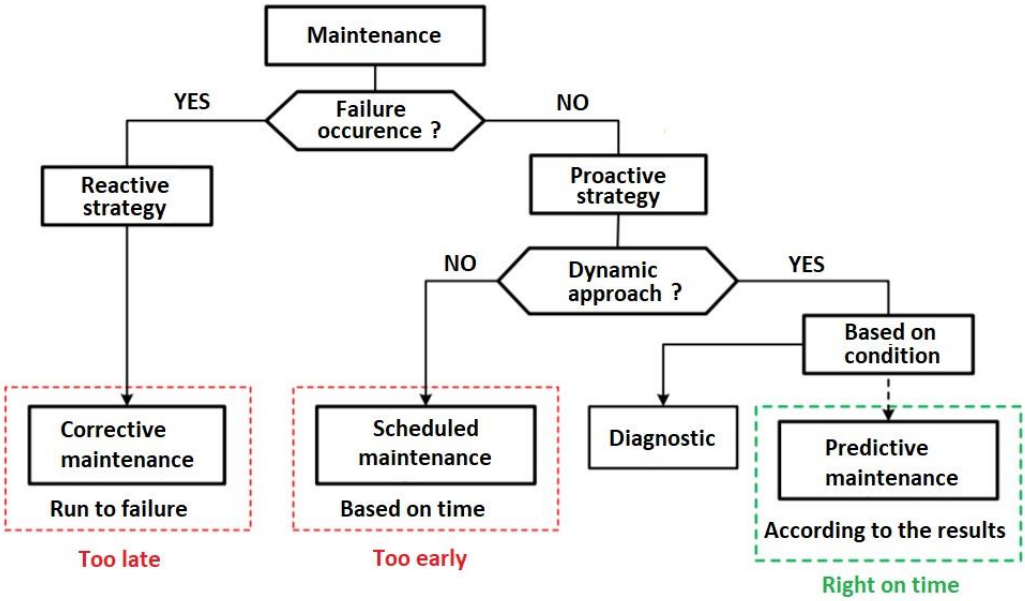

Figure 67. Difference between maintenance strategies [26].

As a result, predicting failures will reduce the time during which the equipment was in repair, thereby increasing the operational reliability of the entire enterprise. The expected benefit from the technology introduced into the process is achieved by reducing the number of unplanned equipment downtime. Besides, the technology allows for improved equipment repair planning. This will reduce the waiting time for spare parts, ultimately minimizing plant downtime. The combination of the above arguments suggests that the introduction of fault detection methods in the enterprise technological process will lead to a decrease in total costs, which is one of the most important economic goals of any enterprise.

7.2 Discussion of the results

The research showed many observations, the answers to which are presented below.

It was found that for the anomaly detection method, many models performed poorly at predictions below 10-13 MW. This observation can be explained as follows, that the training dataset does not have operation modes at such a load. Therefore, the model "did

not learn" to predict such modes. According to the turbine operation technology, modes at such low loads are not used. The operating personnel start the turbine into operation and immediately loads the active power over 13MW. Which explains the insufficient predicting accuracy at low loads.

The second observation was the lack of accuracy in predicting zero loads. This is because there are not enough intervals with such a load in the training dataset. An attempt was made to improve the model by adding additional data into the training dataset for the operational time when the turbine was stopped. This made it possible to improve the model quality and more accurately predict values at zero loads.

The third observation was noted that there were many zero values in some of the data. This is due to the turbine vibration sensor measurements, at the turbine stopped state, which are normal operating measurements. Therefore, there is no need to take any additional action with this observation. In the process of adding additional data for the period of turbine downtime, such zero measurements were added even more, since there is no vibration on the stopped turbine. The above answers revealed the reasons why the models are not behaved enough accurately in some turbine operating modes.

The next discussion was related to the PPR model result. This model has shown the best results in terms of model performance metrics. But this model integration into the enterprise system is a rather difficult task. Of course, models can be found that perform better than linear regression. However, the linear regression model can be programmed in any control system. In this research, the applied meaning of the model was critical. Since the aim of the thesis was to develop a model that will successfully work in the enterprise management system and provide operating personnel with information about the malfunction occurrence.

The last discussion was related to the issue of developing the model in real-time. Indeed, during the turbine operation, new data is constantly being received. A question was raised about the possibility of implementing additional model training with new data in real-time. The expected result of this action would be an increased predictive accuracy of the model. But this functionality is not provided in the Honeywell programming environment. It is possible to perform the additional model training with some other statistical software and the prepared model integrate into the Honeywell system. But, as with other issues in this discussion chapter, this can be viewed as a space for the further development of the project.

8 Summary

The master’s thesis focuses on the application of machine learning methods to detect faults on industrial equipment. For industrial enterprises, the topic of this work is relevant. All key components of production processes depend on reliable, uninterrupted equipment operation.

The goal of the master's thesis was to develop a methodology for detecting malfunctions on industrial equipment based on machine learning methods and the technology of its implementation into the enterprise management system.

In this thesis two methods of industrial equipment fault detection were studied: the fault detection by precedents and the detecting anomalies on the equipment. Based on the data collected from the Enefit Power AS enterprise info servers a number of ML models were fitted to predict the occurrence of malfunctions on the equipment during operation. Programming language R and WEKA software were used for ML models building.

For the case-based fault detection method linear and logistic regression models were created to predict problems on the equipment. For the linear regression model, the RMSE on the training data was 0.13. The model r-square was 92.6%. The logistic regression model correctly classified 99.98% of the training data. On the test samples, was found a large scatter in the model's quality metrics. Comparative metrics analysis showed that for some datasets, the logistic regression model performed better with the classification task. Moreover, on other test datasets, the linear regression model achieved the best performance. As a result, the final choice was made in favor of the linear regression model. The results of this method were summarized at the end of the Chapter 4. The discovered disadvantages and ways to eliminate them were described in detail.

For the method of detecting anomalies on equipment a variety of ML models have been created, such as linear and nonlinear regression, neural networks, Random Forest, MARS model, PPR model. For the linear regression model, many attempts have been made to improve model quality. Summarized results are shown in Table 7. The best performance for the test sample was achieved with the PPR model, the RMSE on the test data was 0.52, r-square was 94.7%. However, due to its simplicity the linear regression model was chosen for integration in the enterprise management system. The linear regression model quality metrics are very slightly inferior to the quality metrics of the PPR model. For the

final linear regression model, the RMSE on the test data was 0.57, r-square was 90.79%. The applied value in this work means the model integration in the Honeywell automated control system to calculate the turbine load predicted values in real-time. The predicted value historical graphs attached to the project prove that the calculations in real-time are fully consistent with the expected results. The graphs visually show that the average deviation of the predicted load value from the actual value was found equal to 1.5 - 2 MW during normal turbine operation. Such deviations are acceptable and do not indicate the development of any negative processes. However, large deviations (3 MW for warning, 6 MW for alarm) will trigger an alarm on the equipment, signaling the development of an anomaly. To validate this assumption, the research performed an equipment fault simulation. To simulate faults, abnormal manual changes were made to the original data. After introducing anomalous values for some measurements into the test samples, was analyzed the model calculated response. The bearing temperature increase simulation by 10 degrees with a simultaneous increase in the generator vibration by 1.7 mm/s led to an increase in the turbine load predicted value by 6.85 MW from the initial one. It can be expected that more measurement deviations from normal values will lead to an even greater impact on the predicted value. Malfunction simulation has proven that these anomalies will not go unnoticed by operating personnel if they occur on the equipment in real life.

Separately, the project noted that the developed methods are innovative for Enefit Power AS. The uniqueness of the work lies in the fact that the author was able to share and distribute the accumulated world experience in this area to the Enefit Power AS enterprise. This research proposes a methodology for diagnostics and fault detecting based on machine learning techniques. The technique is technologically prepared for implementation in the enterprise production process. Uninterrupted equipment operation at the enterprise is of great importance during the production of electricity and liquid fuels. Therefore, the developed methods that will reduce equipment downtime are very valuable for the enterprise.

It should be noted separately that the found methods are universal and applicable to any interested enterprise mechanisms. The information is presented in the project in such a way that it was possible to use this work as a methodological guide for integrating the considered methods in any industrial enterprise.

References

1. Richard Irwin. «Машинное обучение в корне изменит обеспечение надежности».

[WWW] https://sapr.ru/article/25921 (12.03.2021).

2. Мальцев В. «Прогнозная аналитика для эффективного использования оборудования. 2016».

[WWW] https://filearchive.cnews.ru/files/reviews/2016_03_29/2_Maltsev.pdf (17.12.2020).

3. Шаханов Н. И. «Прогнозирование отказов оборудования в условиях малого количества поломок». Publisher: Cherepovets State University (Cherepovets) ISSN: 1994-0637.

[WWW] https://elibrary.ru/item.asp?id=27348490 (11.12.2020).

4. Применение машинного обучения для прогнозирования сбоев оборудования.

[WWW] https://terralink.kz/articles/resheniya-dlya-proizvodstvennoy-sfery/primenenie-mashinnogo-obucheniya-dlya-prognozirovaniya-sboev-oborudovaniya/ (16.12.2020).

5. Антон Курудинов. «Использование данных с датчиков для прогнозирования технического состояния оборудования».

[WWW] https://2019.sapnow.ru/uploads/presentations/forum/03_IoT_SAP_Kurudinov.pdf (17.12.2020).

6. Обработка данных и машинное обучение.

[WWW] https://www.ibm.com/ru-ru/analytics/machine-learning (20.03.2021)

7. В. А. Втюрин. “Автоматизированные системы управления технологическими процессами». //Санкт-Петербургская государственная лесотехническая академия имени С. М. Кирова.

[WWW] https://spbftu.ru/wp-content/uploads/2017/03/asu2.pdf (20.03.2021)

8. Что такое Weka Machine Learning Workbench.

[WWW] https://www.machinelearningmastery.ru/what-is-the-weka-machine-learning-workbench/ (20.03.2021)

9. Полное руководство по изучению R.

[WWW] https://www.machinelearningmastery.ru/a-complete-guide-to-learn-r-29e691c61d1/ (20.03.2021)

10. Зачем нужна очистка данных для Data Mining: 10 главных проблем подготовки датасета и способы их решения.

[WWW] https://www.bigdataschool.ru/blog/%d0%be%d1%87%d0%b8%d1%81%d1%82%d0%ba%d0%b0-%d0%b4%d0%b0%d0%bd%d0%bd%d1%8b%d1%85-data-preparation.html (20.03.2021)

11. M. Kuhn and K. Johnson, «Applied Predictive Modeling»,

DOI 10.1007/978-1-4614-6849-3 3, © Springer Science+Business Media New York 2013

12. Нормализация на практике – методы и средства Data Preparation.

[WWW] https://www.bigdataschool.ru/blog/-data-preparation.html (20.03.2021)

13. Fred C. Pampel, «Logistic regression: A Primer». Series Number 07-132

Leyland F. Pitt, Derek Bromfield, Deon Nel – 1992

14. Kaitlin Kirasich, Trace Smith, Bivin Sadler. «Random Forest vs logistic regression: Binary Classification for Heterogeneous Datasets».

[WWW] https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview (20.03.2021)

15. C. Aldrich and L. Auret, «Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods, Advances in Computer Vision and Pattern Recognition», DOI 10.1007/978-1-4471-5185-2 1, © Springer-Verlag London 2013

16. Essam Seddik. « Fault detection and identification of vehicle starters and alternators using machine learnings techniques».

17. Ф.М. Гафаров, А.Ф. Галимянов. «Искусственные нейронные сети и их приложения». //Издательство Казанского университета».

[WWW] https://kpfu.ru/staff_files/F1493580427/NejronGafGal.pdf (20.03.2021)

18. Алгоритмы ML: один SD (σ) – регрессия.

[WWW] https://www.machinelearningmastery.ru/ml-algorithms-one-sd-%CF%83-regression-47b01d8d51f9/ (20.03.2021)

19. Projection Pursuit Regression.

[WWW] https://scientistcafe.com/ids/projection-pursuit-regression.html (20.03.2021)

20. П.В. Дудченко. «Метрики оценки классификаторов в задачах медицинской диагностики». //Томский политехнический университет.

[WWW] https://core.ac.uk/download/pdf/196226627.pdf (20.03.2021)

21. Простые методы оценки параметров моделей.

[WWW] https://forecasting.svetunkov.ru/etextbook/forecasting_toolbox/estimation-simple-methods/ (20.03.2021)

22. Шаханов Н.И. «Прогнозирование отказов оборудования на основе алгоритмов машинного обучения» \ Н.И. Шаханов, В.М. Осколков, И.А. Варфоломеев, О.В. Юдина \\ Вопросы образования и науки.

23. Информационный критерий Акаике (Akaike's information criterion).

[WWW] https://wiki.loginom.ru/articles/aic.html (20.03.2021)

24. Выдумкин Платон. «Материалы по курсу эконометрика-1».

25. Киран Пиви. «Преобразование данных».

[WWW] https://books.irrp.org.ua/data-design/preobrazovaniya-dannyh/ (20.03.2021)

26. Сай Ван Квонг. «Модели и методы проактивной поддержки принятия решений при управлении техническим состоянием оборудования».

[WWW] http://vstu.ru/upload/iblock/a08/a08aa98012251eb386ce39535d506bf1.pdf (27.03.2021)

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Dmitri Poljakov

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Application of Machine Learning Methods to Industrial Equipment Fault Detection", supervised by Olga Dunajeva
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

10.04.2021

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2

Link to the repository containing the RStudio analytical part code, tables and files with datasets that took part in the project.

https://gitlab.cs.ttu.ee/dmitri.poljakov/application-of-machine-learning-methods-to-industrial-equipment-fault-detection

Case-based fault detection

1. RStudio analytical part (code, plots, tables, models)
   1.1. Case-based fault detection method
2. Datasets
   2.1. RStudio datasets
   2.2. WEKA datasets

Anomaly detection

3. RStudio analytical part (code, plots, tables, models)
   3.1. Anomaly detection method part1
   3.2. Anomaly detection method part2
   3.3. Anomaly detection method part3
4. RStudio datasets
5. Tables
   5.1. Turbine and generator measurement points Enefit 280
   5.2. Linear regression model Enefit 280