

Industry 4.0 Application on Diagnosis Prediction of Construction Machinery: A New Model Approach

Semra Erpolat Taşabat^{1,2,*}, Tayfun Özçay³, Salih Sertbaş³, Esra Akca³

¹Department of Statistics, Faculty of Science and Literature, Mimar Sinan Fine Art University, Turkey

²Social Sciences Institute, Bahçeşehir University, Turkey

³Borusan CAT, Turkey

Received February 26, 2020; Revised June 13, 2020; Accepted June 19, 2020

Copyright©2020 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract The transition from large data stacks obtained as a result of rapid development in computer technologies to meaningful information is only possible with data mining and statistics. In this study, a model has been developed to provide early fault detection and vehicle maintenance needs by using instant data obtained from Caterpillar Inc. construction vehicles. With the Early Warning System, primarily, the selected sensor data coming from the satellite related to the vehicles is used to predict the failure possibility of the vehicles in a certain time ahead remotely by using the methods of machine learning and using the internet of things and cloud technology. Then, prediction data are integrated into decision-making mechanisms in business processes. Finally, the information acquired by using data visualization technologies is made available for being reported and made traceable through summary data. The location of data mining on machine learning is illustrated by the necessary algorithms. In order to determine the correct fault in accordance with the data obtained from the sensors of the machines the gradient boosting, logistic regression and C5.0 algorithm is used. From the results obtained, the gradient boosting algorithm produced the best training results for all categories, while for the test data, the gradient boosting algorithm produced the best results for the categories C1000 and C3000, and logit regression for the C3030, C5070 and C5459 categories. The focus of the personalized product mentioned by Industry 4.0, the system developed in this study, can be easily adapted to the operation of different machines.

Keywords Machine Failure, Machine Learning, Gradient Boosting, Logistic Regression, C5.0 Algorithm

With the developing technology, the production and use of computerized machines and vehicles have gathered momentum and consequently the amount of information obtained has increased gradually. The conversion of continuously increasing and streaming data stacks into logical and useful information becomes more and more important. Various methods and algorithms have been developed for this purpose. This series of methods and algorithms, which are called data mining in the most general terms, have been combined with statistics and transformed into methods of more comprehensible and logical solutions.

Industry 4.0 is emerging as a new vision with developing technology. Industry 4.0 was first used in 2011 at the Hannover Messe in Germany. Its aim is to achieve successful results in cooperation by enabling devices and machines to communicate with each other and with humans. Sustainable production, real-time data monitoring, system intervention before a failure occurs, high efficiency, reduction in production costs and product launch time, and flexibility in production have made Industry 4.0 advantageous (What is Industry. <http://www.armolis.com/akademi/endüstri-40/> [1 December 2019].)

The development of the industry is briefly summarized in Figure 1 (Recommendations for implementing the strategic initiative industry 4.0. 2013. <http://alvarestech.com/temp/tcn/CyberPhysicalSystems-Industrial4-0.pdf/> [4 October 2019].) As can be seen from the figure, the development of the industry started in 1784 with the development of the production sector based on steam engines. Then, the development gained momentum in 1870, with the development of electrical energy-based factories and production lines, and in 1969 with electronics and automation-based production, and finally reached its highest level in 2011 with cyber-physical systems (Industry 4.0).

1. Introduction

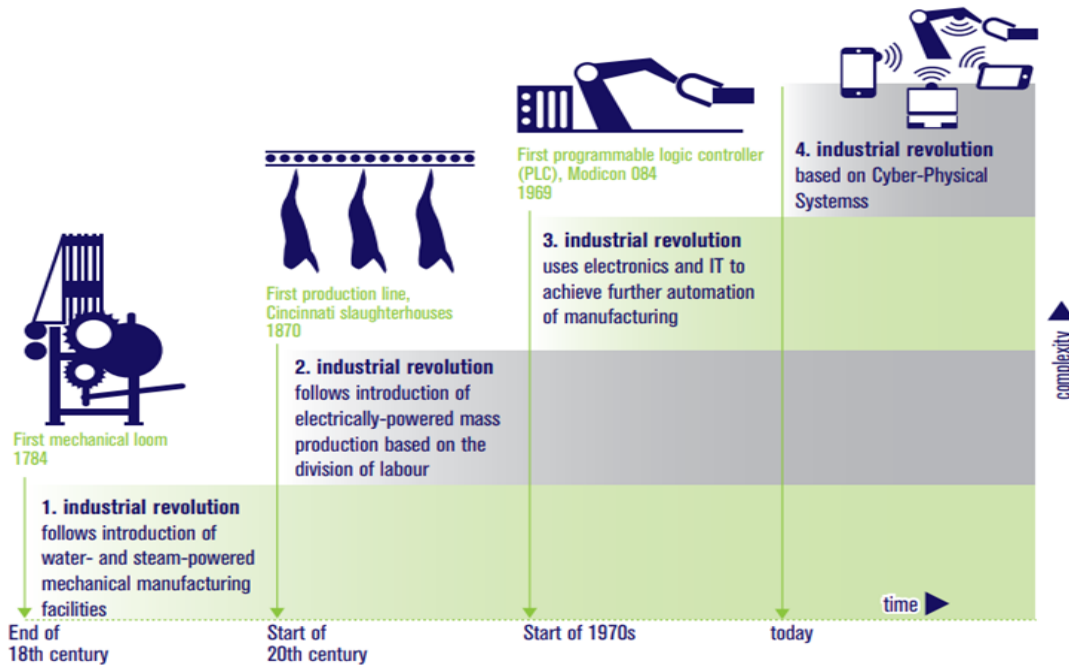


Figure 1. The four stages of the Industrial Revolution.

In 2013, with the support of the German Federal Ministry of Education and Research, Industry 4.0 study group (Prof. Dr. Henning Kagermann, Prof. Dr. Wolfgang Wahlster, Dr. Johannes Helbig) proposed strategic application for the Industry 4.0 initiative. Many academicians and sector representatives participated in the study. Under different titles, this initial study laid the foundations for the ongoing new industrial revolution all over the world. The industrial revolution has ensued from the need for a new and theoretical basis for the research, development of and design of complex industrial systems that can be adapted very dynamically to specific production needs.

In this context, the three standing out main definitions - the internet of services, the internet of things and cyber-physical systems (CPS) - have been adopted on a global basis. The internet of services represents the services and functions as sensitive software components and is offered via the cloud. Development platforms enable the development of web-compatible services. The internet of things is the technological vision for integrating all kinds of objects, devices and persons in a universal digital network. It is intended to help people in their business. The Internet of things and internet of services mainly target companies and public administrations. Cyber physical systems; communicated objects, devices, machines and logistics components that consist of integrated systems. They can communicate over the Internet and by using Internet services, form networks with each other and make autonomous decisions with humans in a decentralized manner. Cyber-physical systems help in completion or support of a series of tiring, slow, and insecure tasks for humans by machines. They are able to carry out tasks

autonomously by making decisions at a micro level as much as possible.

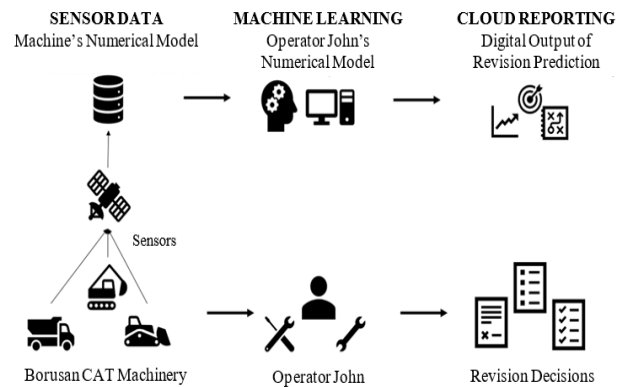


Figure 2. EWS Data Discovery Process

Contacting with most of the principles that Industry 4.0 is based upon, this study is one of the most comprehensive Industry 4.0 studies ever conducted in Turkey. It has been developed with the integration of industrial processes and information technologies. In general, the study involves the steps of remote detection, performance monitoring, and execution of tasks in real time. Immediate data flow from over 10,000 construction machinery is provided by GPS and GPRS by using internet of things technologies across Turkey. The machines are being monitored instantaneously with more than 1600 different sensor types, tens of them on each machine. The size of the generated data flow is more than 2 million data per minute. From this perspective, this study can be considered as a big data project in terms of data amount. However, this large amount of data alone does not mean much. For this reason, big data does not

focus on how much information is available, it focuses on what can be done with that information and provides the selection of the right approach for saving money time and sustaining the system in decision making processes. Accordingly, the streaming data is made ready for modelling via various data manipulation methods. The discovery process of the EWS proposed in this study consists of the following steps (Figure 2):

2. Literature Review

There are different studies available in this field in the literature. The most noteworthy of these are listed below:

Monitoring of the state of a motor vehicle using machine learning and data mining technology to generate component models that are then used to monitor components, predict failure, etc., such analysis being useful for repair, etc.

Component data analysis involves: provision of a version of a behavior model of the component; prediction of component behavior based on the first model version; collection of component performance data; comparison of the predicted behavior with the collected data; determination of any discrepancies between the two; determination of whether the discrepancy stems from a component failure; and if not modification of the first model version. Independent claims are also made for a component data analysis system, a method for monitoring component service lives and a diagnosis method for application to motor vehicle components so that their service life can be estimated (Bertsche and Engelhardt, 2001).

Method and system for condition monitoring of vehicles

A method and system for an improved vehicle monitoring system in order to provide a cost-effective and scalable system design for industrial application through the use of machine learning and data mining technologies on data acquired from a plurality of vehicles to create models. Frequent acquisition of vehicle sensor and diagnostic data enables comparison with the created models to provide continuing analysis of the vehicle with respect to repair, maintenance and diagnostics (Fiechter and Göker, 2001).

Service Life predicting method and system for machine and vulnerable component of generating set

The invention relates to a system for estimating the life of the mechanical and electrical wearing parts for the power set, which comprises a database, a server, a software,

a knowledge base, a man-machine interface, and a user terminal system, the life estimation method comprises: counting the history data of the life of the mechanical and electrical wearing parts, setting up a life database of the mechanical and electrical wearing parts, determining the distribution category and parameters of the mechanical and electrical wearing parts, counting the reliability $R(t)$ of the mechanical and electrical wearing parts, counting the reliable life t_{RO} of the mechanical and electrical wearing parts, setting up a knowledge base of the malfunction result of the mechanical and electrical wearing parts, determining the changeover period HRP of the mechanical and electrical wearing parts, estimating the residual life HRL , adjusting the planned maintenance interval to realize the optimized maintenance. The advantages of the invention are that the changeover period of the mechanical and electrical wearing parts can be calculated quantitatively in the using stage of the power set, and the residual life of the parts can be calculated quantitatively, thereby a computer online monitor of the changeover period and the residual life of the mechanical and electrical wearing parts can be realized (Jinyuan et al., 2007).

Systems and methods for analyzing equipment failures and maintenance schedules

A computer implemented method may be used for analyzing equipment failures and maintenance schedules. An equipment maintenance system generates a model of equipment and components of each piece of equipment. In one embodiment, the model is a tree representation. The equipment maintenance system may then determine estimated failure information for each component based on a selected statistical model. The equipment maintenance system may also generate a maintenance schedule based on the determined estimated failure information for each component of the equipment. In one embodiment, the equipment maintenance system displays the equipment maintenance information (Richards et al., 2011).

Failure diagnosis method, failure diagnosis apparatus, conveyance device, image forming apparatus, program, and storage medium

A failure diagnosis method diagnoses a failure occurring in a diagnosis target apparatus including a drive mechanism having a drive member that receives power supply to operate and a power transmission member that transmits drive force of the drive member to another member. The method includes automatically acquiring by a sensor an operation state signal indicating an operation state during the drive mechanism operating for a predetermined period; and analyzing the automatically acquired operation state signal based on a failure probability model, which is obtained by modeling a cause of failure occurring in the diagnosis target apparatus with

using probabilities, to execute failure diagnosis with respect to each of constituent members of the drive mechanism (Yasukawa et. al., 2004).

Intelligent fluid sensor for machinery diagnostics, prognostics, and control

A system that facilitates device and/or machinery diagnostics, prognostics and control by way of condition sensing, such as sensing the condition of the device and/or a fluid of the device (e.g., fluid health indicators). The system can employ a plurality of sensors to determine a current state and estimate a future state of the fluid and/or device, as well as providing control of the device, e.g., in order to increase the remaining useful life of the fluid and/or operation of the device. The sensors can communicate wirelessly with each other, with the device, and/or with a central control system that provides, e.g., sensor fusion, prognostics and control integration. In addition, the sensors can be powered locally based upon the physical or chemical properties of the environment (Frederick et al., 2006).

Industrial process data acquisition and analysis

A method for optimizing an industrial process data is disclosed. The method includes collecting data from a plurality of sensor elements, wherein each sensor element collects data from a portion of the industrial process and verifying the data collected. The method further includes analyzing the data collected for efficiency and generating at least one recommendation for optimizing the industrial process. The method further includes presenting the at least one recommendation generated to an administrator of the industrial process (Rumi et. al., 2005).

3. Data Abstraction and Comprehension of Data

Today, many vehicles that are being produced are equipped with electronic circuit elements and software. Thus, the operation and control of these vehicles have become easier and it has been made possible to detect any failure that may occur in the vehicle so that it does not stop the flow of the process. This has resulted in significant returns in terms of cost and time.

Borusan, one of the world's leading companies in the production and maintenance of construction machinery, aims to use electronic circuits in its vehicles in the most effective way. Construction machinery of Caterpillar Inc. sold by Borusan are equipped with a total of 1667 different sensors that are used to detect major failures that require replacement of parts. With the help of these sensors, the data related to the operation of the construction machinery are provided instantly via satellite over GPS and GPRS.

3.1. Data Acquisition

By storing the signals received from the satellite in the cloud, machine/system data are obtained using cloud computing. At the date of registration, the sensor codes that are obtained from the built-in sensors and microchips for the respective machine and that transmit signals are stored in the data warehouse in the format given in Table 1.

The descriptions for the variables in the data set are as follows:

MID: Signals acquired from integrated microchips on the machine to which sensors are connected.

CID/EID: Sensors used for diagnostic purposes.

FMI: Sensor signal strength.

Table 1. Sensor codes and an image of raw data.

SERIAL NUMBER	MODE L	RECEIVED TIME	DATE	MID	CID	FMI	EID	FAULT	Service Date	Service T-Sensor T
AXG01341	980H	15.07.2018 06:15	20180715	30	129	1	96	Engine	20180802	17
AXG01341	980H	5.07.2018 08:59	20180705	30	190	22	96	Transmission	20180802	27
SXG01341	980H	17.07.2018 09:58	20180717	30	155	2	96	0	-	
SXG01341	980H	22.06.2018 05:16	20180622	45	2	1	44	0	-	
RXG01432	315A	14.07.2018 06:15	20180714	20	100	1	22	0	20180713	29
RXG01432	315A	8.07.2018 08:59	20180708	20	33	2	222	Injector	20180713	35
RXG01432	315A	16/07/2018 09:58	20180716	20	20	2	960	Injector	20180713	27
NXG01432	315A	20/06/2018 05:16	20180620	20	10	3	13	0	-	

The variables listed above are considered as independent variables within the scope of the study and the variable containing machine failures gathered under the basic codes given in Table 2 is considered as the target variable.

Table 2. Equipment faults and codes

Component Code	Component Name
1000	Engine
3000	Transmission & Drive Line
3030	Transmission
5070	Piston Pump
5459	Swing Drive

3.2. Data Preprocessing (Data Manipulation)

Before the development of the EWS, the preprocessing procedure of the raw data from the sensors has to be carried out. The details of this procedure, which consists of two steps - data acquisition and recoding of the data - are as follows:

Step 1: Data acquisition

Since it has been indicated that the period for reporting the faults after machines had failed and the period for obtaining the service records and for the repair of the related components and their registry to the system, lasts a maximum of 30 days, the t-30-day signal data prior to failure notifications have not been taken into consideration. For non-defective machines, the 30-day data before the

model production date has not been taken into account. After generating a failure code the machine sensors generate a continuous signal for a certain period of time or until they are repaired, thus the last fault signal generated from each sensor within the range of t-30 and t-60 has been considered as data. In order for a machine to be considered faulty, it must generate a signal within the range of t-30 and t -60 starting from the date of the failure. Figure 3 shows the process of data acquisition.

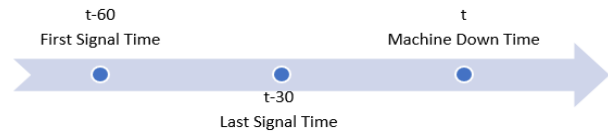


Figure 3. Data acquisition process

Step 2: Recoding of data

The signal data (CID, FMI, EID) incoming under the three main categories are rearranged so that each sensor is a dummy variable. If the sensor number is written to the data warehouse at the relevant date, the dummy variable of that sensor takes the value 0 if it is not written.

In Tables 3 (a) and (b), the conversion stage of the CID variable is examined. Accordingly, the codes 129, 190, 155 of the CID variables in (a) were converted each into a dummy variable in (b). The dummy variables have a value of 1 if the sensor number has been written in the data warehouse and a value of 0 if it has not been. The conversion procedures of FMI and EID variables were performed in a similar way.

Table 3. An image from the conversion stages of the CID variable



Serial Number	Component	CID	Date
RXG01432	Engine	129	1/8/2018
NXG01432	-	190	2/8/2018
NXG01432	-	190	3/8/2018
NXG01432	-	155	3/8/2018

(a)

Serial Number	Component	CID-129	CID-190	CID-155
RXG01432	Engine	1	0	0
NXG01432	-	0	1	1

(b)

4. Methodology

In the following sub-sections, models will be presented that are included within the content of the EWS that have been developed within the scope of the study. Then, details on preferred model comparison criteria in order to make comparisons of all developed and used models and the evaluation criterion that is used to measure the success of the model will be given. Finally, the parameter and variable selection will be explained in detail.

4.1. Models in Use

Supervised machine learning enables computers to learn from labelled training data without having been explicitly programmed. The tasks covered by this supervised learning are regression and classification. Accordingly, both non-linear decision tree techniques such as gradient boosting, C5.0 and linear machine learning algorithms such as logistic regression will be used in the modelling phase.

The application is based on the comparison of these techniques using several classification techniques on multiple dependent variables. For this reason, logistic regression analysis, gradient boosting and C5.0 decision rule derivation algorithm, which represent the three important components of data mining - statistics, machine learning and database technologies - were determined as classification techniques to be used in practice.

4.1.1. C5.0 Algorithm

Number of data and data quality play an important role in the success of the applications. In line with the need for up-to-date and fast decision making, the opinion gains importance which suggests that the most suitable model is going to be the C5.0 algorithm. The C5.0 algorithm uses the boosting algorithm to increase accuracy, particularly in large data sets. Therefore, the C5.0 algorithm is also known as boosting trees. Among its biggest advantages: the algorithm being fast and that it uses memory efficiently. The C5.0 algorithm uses the concept of information gain and entropy reduction to optimally separate nodes. For the variable X , k -fold probabilities are named p_1, p_2, \dots, p_k respectively. The entropy for the X variable or quality is given in the Equation 1 (Quinlan, 1996):

$$Entropy = H(X) = -\sum_{j=1}^k p_j \log_2(p_j) \quad (1)$$

It is assumed that the target quality is subdivided into T_1, T_2, \dots, T_k of T sub-clusters depending on the quality of the training set. The weighted average of the information required to determine the class of each T can be calculated as the weighted total number of entropies. The weighted average of the information is given in the Equation 2.

$$H_s(T) = \sum_{i=1}^k p_i H_s(T_i) \quad (2)$$

The information gain is then calculated to perform the separation. Thus, the C5.0 algorithm performs optimal separation by determining the separation criterion with the largest information gain at each decision node. Information gain is given in the Equation 3.

$$\text{Information Gain } (S) = H(T) - H_s(T) \quad (3)$$

4.1.2. Gradient Boosting

Gradient Boosting is a prediction algorithm that can be used for regression and classification models. It is generally based on the re-estimation of model errors with sequential models and weak estimators, and the reduction of variance as a result of the aggregation of the resulting models with a certain weight (Chen, 2014). This algorithm is preferred as sequential models are used in the system that is developed within the scope of the study and that the estimators are quite small. In the following formula flow, the gradient boosting algorithm that has been generalized in decision trees is being demonstrated (Friedman, 1999). It is a machine learning technique for gradient boosting, regression and classification problems. This, combined with weak prediction models, typically creates a model of decision trees. The purpose of any supervised learning algorithm is to identify and minimize a loss function. How mathematics works for Gradient Boosting algorithm is shown in Equation 4, Equation 5 and Equation 6 (Gradient Boosting.

<https://devhuntery.wordpress.com/2018/07/11/gradyan-artirma-gradient-boosting/> [11July 2018]). Assuming that we have a Mean Square Error (MSE) defined as in Equation 4:

$$Loss = MSE = \sum (y_i - y_i^p)^2 \quad (4)$$

Where y_i = i th target value, y_i^p = i th prediction, $L(y_i, y_i^p)$ is Loss function

Estimations are requested so that the loss function (MSE) is minimum. By using gradient descent and updating estimations according to the learning rate, values can be found where the MSE is minimum (Equation 6), Equation 7.

$$y_i^p = y_i^p + \alpha * \delta \sum_{i=1}^n (y_i - y_i^p)^2 / \delta y_i^p \quad (6)$$

$$\text{Which becomes, } y_i^p = y_i^p - \alpha * 2 * \sum_{i=1}^n (y_i - y_i^p) \quad (7)$$

Basically, the estimates need to be updated so that the sum of the residues is close to 0 (or minimum) and that the estimated values are close enough to the actual values.

4.1.3. Logistic Regression

Logistic regression is the analysis method used in linear classification problems in cases where the dependent variable is categorical and in the presence of independent variables determining the outcome. There are only two possible cases. The developed model is used to explain the relationship between dependent and independent variables. If the data is categorical (nominal/ordinal), non-parametric statistical methods are being used. Since logistic regression analysis aims to estimate the value of categorical dependent variable, a "membership" for two or more groups is estimated. Accordingly, it can be said that one of the aims of the analysis is to classify and the other is to analyze the relationships between dependent and independent variables (Mertler, et al., 2005). Logistic regression is a model-building technique in statistics and is used to determine the cause-effect relationship with independent variables when the dependent variable has binary (dichotomous), triple and multi-categories. The assumption that independent variables are suitable for normal distribution and that covariance matrixes are equal for all groups, is not sought after within the logistic regression analysis. Therefore, the estimations in logistic regression are focused on the probability of the dependent variable taking the value of 1, instead of focusing on the value of the dependent variable. Since the result obtained is a probability value, it can only take a value between 0 and 1. The logistic regression model formulation is shown in Equation 8, Equation 9, Equation 10 and Equation 11 (Yakut and Gemici, 2017).

X : The probability of occurrence of $Y=1$ is π when the data matrix for the independent variable is $X=x$. The specific form of the logistic regression model is shown in Equation 8, Equation 9 and Equation 10 as follows (Hosmer and Lemeshow, 2000):

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (8)$$

Logit conversion is applied to the non-linear logistic regression function given in Equation (9). When performing the logit operation, the natural logarithm of the Odds of an event is taken. Odds is the ratio of the probability that an event will occur to the probability that it will not. In other words For odds $Y=1$ $P(Y=1)/(1- P(Y=1))$ (Hosmer and Lemeshow, 2000; Menard, 2002).

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \quad (9)$$

The logit transformation function $g(x)$ covers most of the desired properties of a linear regression model. Logit or $g(x)$, is a linear equation that can take a value of $-\infty$ and $+\infty$, depending on the independent variable x . Logistic regression model can be written as following in Odds of the independent variable.

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0 + \beta_1 x} \quad (10)$$

When the natural logarithm of Odds is taken, the model becomes linear.

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln e^{\beta_0 + \beta_1 x} = \beta_0 + \beta_1 x \quad (11)$$

4.2. Model Comparison and Evaluation Criteria

4.2.1. ROC Curve and AUC Value

In the study, the **ROC (Receiver Operating Characteristic) curve** and corresponding **AUC (Area under the ROC Curve) value** have been used as model selection criteria. AUC value is used frequently in binominal dependent variables and it is essentially a statistic based on type 1 and type 2 error rates (Classification: ROC Curve and AUC, 2019. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc/> [1 December 2019]). This value, which cannot be calculated for a multinomial variable, has been made computable by following a different path within the scope of the study. Namely, each target category is considered separately and the AUC value could be calculated then by combining the other categories. This process was repeated for each category and the ROC curve and AUC values were calculated for different binominal target sets as many as the number of categories. Then, the average of these calculated AUC values has been used as model success criteria. The average AUC used in the study is given in Equation 12.

$$\begin{aligned} & \text{Avg AUC} \\ & = (AUC1 + AUC2 + AUC3 + \dots + AUCn) / n \quad (12) \end{aligned}$$

4.2.2. F Value

In this study, the **F-value** calculated by taking the harmonic average of the precision and recall values have been used as another comparison criterion. The F value is as shown in Equation 13.

$$F = 2(\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (13)$$

The precision and recall measurements obtained in the formula are shown in Equation 14 and Equation 15, respectively.

$$\text{Precision} = TP / TNF \quad (14)$$

$$\text{Recall} = TP / (TP + FN) \quad (15)$$

The descriptions of the abbreviations used in the formulas are as follows:

TP: Accurate Fault Classification Number

TNF: Total Number of Faults

FN: Incorrectly Estimated Number of Fail-Free Classes

FP: Incorrectly Estimated Number of Successful Classes

TN: True Fault Prediction

SC: Success Criterion

RECALL: The concept of precision is generally indicated by the letter p and is calculated as the ratio of the correct results in the incoming information to the whole incoming information.

PRECISION: The concept of recall is generally indicated by the letter r and is calculated as the ratio of the incoming correct results that are brought to those that have to be brought.

In order to calculate TP and TNF values, it is necessary to determine which category the probability values that are predicted by the model fall into. Since the data contains negative and positive values and the number of data is not balanced, the probability limit value was determined according to the Odds ratio. The odds ratio was calculated as in Equation 16.

$$Odds\ Ratio = \frac{TP / FP}{FN / TN} \quad (16)$$

4.2.3. Over-learning (Overfitting) and Recall

On the other hand, in order to determine the over-learning amount of the model and to measure the sampling recall, the percentage difference of the learning and test sample was taken. Over-learning is that the model begins to learn observations rather than patterns. In this case, the data set used for learning is learned. However, such models reduce the likelihood of making a successful prediction when faced with new and previously unseen observations. Over-learning was calculated as shown in Equation 17.

$$Overfitting = 1 - \frac{SC_{train}}{SC_{test}} \quad (17)$$

4.3. Selection of Variables and Parameters

In boosting algorithms, it is generally not advisable to pre-select variables based on predictive power. The algorithm selects them automatically. Nevertheless, there is an expectation that the variables may produce a minimum amount of loss value or a certain level of variance. For this reason, the following conditions are taken into consideration in determining the variable list to be used within the scope of the study:

1. Loss value analysis was performed but no loss value could be detected.
2. In the case of defective machines; sensor variables are expected to produce signals at least 4% of the number of faults.

Considering these limitations, the number of variables decreased from 1390 to 50.

4.3.1. Parameter Optimization with Cross-Validation

In the study, two different samplings were used to train

and test machine learning algorithms. In contrast, cross-validation was used for parameter optimization in the training cluster. This developed methodology is shown in Figure 4.

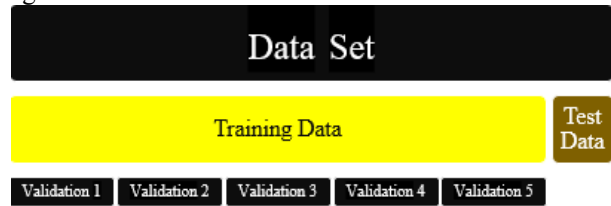


Figure 4. Preparing the data sets

In order to optimize the model parameters, the training set was divided into 5 equal parts with k-fold cross validation, the model was trained with the same parameters by keeping a different part out each time and the external part was estimated. In this way, when the parts that are excluded from the training set and that are estimated are combined, it covers all observations in the training set and a validation success statistic is calculated.

As a result of repeated iterations with different parameter values, the parameters giving the best validation success statistics were used in the model. Table 4 covers the details of the cross-validation method applied.

Table 4. Cross-validation

Training Data Set					
	Validation 1	Validation 2	Validation 3	Validation 4	Validation 5
1	Prediction 1	Train	Train	Train	Train
2	Train	Prediction 2	Train	Train	Train
3	Train	Train	Prediction 3	Train	Train
4	Train	Train	Train	Prediction 4	Train
5	Train	Train	Train	Train	Prediction 5

1	Prediction 1
2	Prediction 2
3	Prediction 3
4	Prediction 4
5	Prediction 5

In Table 4, it can be seen that one group of train data, which was divided into 5 equal parts, was excluded one by one and five different estimation results were achieved by the train data that entered the analysis. The combination of estimation results was used to measure success statistics. The formulations of the AUC values used for each cross validation are given in Equation 18-22 below:

AUC1=AUC[C_CODE1000], the code 1000 in the dependent variable is acknowledged as “one”, other codes are acknowledged as “null” (18)

AUC2=AUC[C_CODE3000], the code 3000 in the

dependent variable is acknowledged as “one”, other codes are acknowledged as “null” (19)

$AUC3=AUC[C_CODE3030]$, the code 3030 in the dependent variable is acknowledged as “one”, other codes are acknowledged as “null” (20)

$AUC4=AUC[C_CODE5070]$, the code 5070 in the dependent variable is acknowledged as “one”, other codes are acknowledged as “null” (21)

$AUC5=AUC[C_CODE5459]$, the code 5459 in the dependent variable is acknowledged as “one”, other codes are acknowledged as “null” (22)

The average of these AUC values have formed the average ROC value. The Equation 23 has the average ROC value involved.

$$AvgROC = (AUC1 + AUC2 + AUC3 + AUC4 + AUC5) / 5 \quad (23)$$

When calculating the average ROC, the AUC values of each category were aggregated and arithmetic means were obtained.

5. Application of the Models and Their Results

5.1. Gradient Boosting Model

The results of the model are given in Table 5. The best validation results obtained from 64 different trials with Gradient Boosting algorithm are as follows: Trees: 60, Interaction.depth: 2, Minobsinnode: 15, Shrinkage: 0.08. The average ROC value for these parameters is 82.0%.

5.2. C5.0 Model

The results of the model are given in Table 6. The best validation results obtained with the C5.0 algorithm are as follows: winnow:1, noGlobalPruning:1, earlyStopping :1, trails:1 The average ROC value is 65.2%.

Table 5. Result of Gradient Boosting Model

Iteration	Avg ROC	n.trees	interaction.depth	n.minobsinnode	shrinkage
73	82.0%	60	2	15	0.08
1	81.8%	80	3	20	0.05
85	81.7%	110	1	25	0.08
43	81.7%	40	2	15	0.08
4	81.5%	70	2	15	0.08
14	81.5%	100	1	20	0.05
72	81.2%	80	2	10	0.06
84	81.1%	150	1	25	0.1
66	81.0%	110	2	10	0.06
2	71.5%	160	2	45	0.09

Table 6. C5.0 model result

Iteration	avgROC	winnow	noGlobalPruning	earlyStopping	trails
1	65.2%	1	1	1	1
43	65.2%	1	1	0	4
4	65.2%	1	1	0	4
14	65.2%	1	1	1	1
42	65.2%	1	1	0	4
47	65.2%	1	1	0	4
16	65.2%	1	1	1	1
44	65.2%	1	1	0	4
41	65.2%	1	1	0	4
64	56.3%	0	0	0	11

Table 7. Result of Logistic Regression model

Full Validation Set	Validation 1	Validation 2	Validation 3	Validation 4	Validation 5
Intercept	Intercept	Intercept	Intercept	Intercept	Intercept
isold	isold	isold	isold	isold	isold
l2 level e	l2 level e	eid119 eid	l2 level e	l2 level e	l2 level e
mid36 mid e	mid36 mid e	l2 level e	mid113 mid	mid113 mid	mid113 mid
mid113 mid	mid113 mid	mid113 mid	mid36 mid e	mid36 mid e	mid36 mid e
c190 8 fmi	eid119 eid	mid36 mid e	eid284 eid	eid284 eid	c190 8 fmi
eid119 eid	c190 8 fmi	c190 8 fmi	c174 3 fmi	c174 3 fmi	l3 level e
l3 level e	l3 level e	l3 level e	l1 level	c190 8 fmi	c3547 3 fmi
eid284 eid	mid27 mid	l3 level	mid27 mid	mid27 mid	c174 3 fmi
l3 level	c91 8 fmi	eid284 eid	l3 level	l3 level	l3 level
mid27 mid	mid39 mid e	mid53 mid	c596 9 fmi	eid265 eid	eid284 eid
c596 9 fmi	mid39 mid	mid39 mid	mid82 mid	c590 9 fmi	mid27 mid
c91 8 fmi	l2 level	mid39 mid e	mid82 mid e		mid39 mid e
mid39 mid e	l3 level		mid53 mid e		mid39 mid
mid39 mid	eid284 eid		mid39 mid		
l2 level	mid53 mid		mid39 mid e		
mid82 mid			l2 level		
			mid263 mid		

6	Significant in all validation and train samplings
5	Significant in all validation samplings
4	Significant in Validation 4 samplings
3	Significant in Validation 3 samplings
2	Significant in Validation 2 samplings
1	Significant in 1 sampling

Table 8. Result of Logistic Regression model

Variable	Coefficient	Std. Dev	t value	Prob.
isOld	1.731662	0.147888	11.70931	0.0000
L2 LEVEL E	1.1510092	0.165821	9.106766	0.0000
MID36 MID E	-0.654071	0.149915	-4.362935	0.0000
MID113 MID	1.88977	0.312417	6.048875	0.0000
C190 8 FMI	1.051788	0.353036	2.979267	0.0029
MID27 MID	0.949149	0.256952	3.693878	0.0002
MID39 MID E	-1.028192	0.251712	-4.084792	0.0000
MID39 MID	1.038852	0.246243	4.218811	0.0000
EID284 EID	1.477105	0.400435	3.688748	0.0002
L3 LEVEL	-0.626918	0.216584	-2.894576	0.0038

	Limit Confidence			
LIMIT 2:C(11)	2.184592	0.138888	15.72922	0.0000
LIMIT 3:C(12)	2.592963	0.145968	17.76396	0.0000
LIMIT 4:C(13)	2.757336	0.148853	18.52386	0.0000
LIMIT 5:C(14)	2.819558	0.149993	18.79791	0.0000
LIMIT 6:C(15)	2.868769	0.150892	19.01211	0.0000

5.3. Logistic Regression Model

For logistic regression, backward variable selection method was used in each validation sampling. The aim of this analysis is to determine the variables that will provide the optimal cohesion. Variables used in the models have been selected by taking into consideration these criteria: in how many validation sets these variables are found to be significant and variables that have been tried in at least 5 models are considered to be stable. The results of the model are given in Table 7.

Table 7 shows the distribution of the significance order of the variables across all validation clusters. The order of significance was scaled within a range of 1 to 6, and the green color representing 6 predicates significance across all validation clusters whereas the color representing 1 represents significance in validation 1 samplings.

Table 8 lists the logistic regression results of the independent variables that are found to be significant in all validation and train clusters that provide the optimal cohesion.

Threshold-limit values determined according to this will help to categorize the variables examined in the model are given under "Limit Points". These limit values sampling.

Table 9. Model assessment tools

Model	Parameter	Categories	F-Value (train)	F-Value (Test)	F-Value Overfitting %	AUC (train)	AUC (Test)	AUC Overfitting%
Gradient Boosting	n trees: 150 interaction. depth: 1 minobsinnode:20 shrinkage: 0.08	C1000	37.1%	24.3%	53%	93.3%	80.2%	16%
		C3000	0.0%	0.0%	0%	98.0%	72.6%	35%
		C3030	0.0%	0.0%	0%	97.9%	26.9%	265%
		C5070	14.2%	10.3%	38%	99.1%	75.6%	31%
		C5459	12.5%	8.3%	51%	99.5%	65.5%	52%
C5.0	iteration:62 winnow:1 no global pruning early stopping no trails:10	C1000	27.4%	21.8%	26%	75.7%	72.9%	4%
		C3000	4.9%	1.7%	188%	71.9%	38.8%	85%
		C3030	6.1%	1.9%	221%	78.5%	64.2%	22%
		C5070	15.2%	8.7%	75%	94.9%	30.1%	215%
		C5459	11.7%	4.7%	149%	96.0%	66.1%	45%
Logit Regression	LIMIT_2:C(11) 2.184592	C1000 C3000 C3030 C5070 C5459	21.1% 4.3% 5.4% 6.7% 3.8%	19.6% 2.8% 1.5% 7.5% 3.1%	8% 54% 260% -11% 23%	78.2% 78.4% 82.8% 94.4% 90.3%	75.5% 47.7% 66.3% 81.0% 93.7%	4% 65% 25% 17% -4%
	LIMIT_3:C(12) 2.592963							
	LIMIT_4:C(13) 2.757336							
	LIMIT_5:C(14) 2.819558							
	LIMIT_6:C(15) 2.868769							

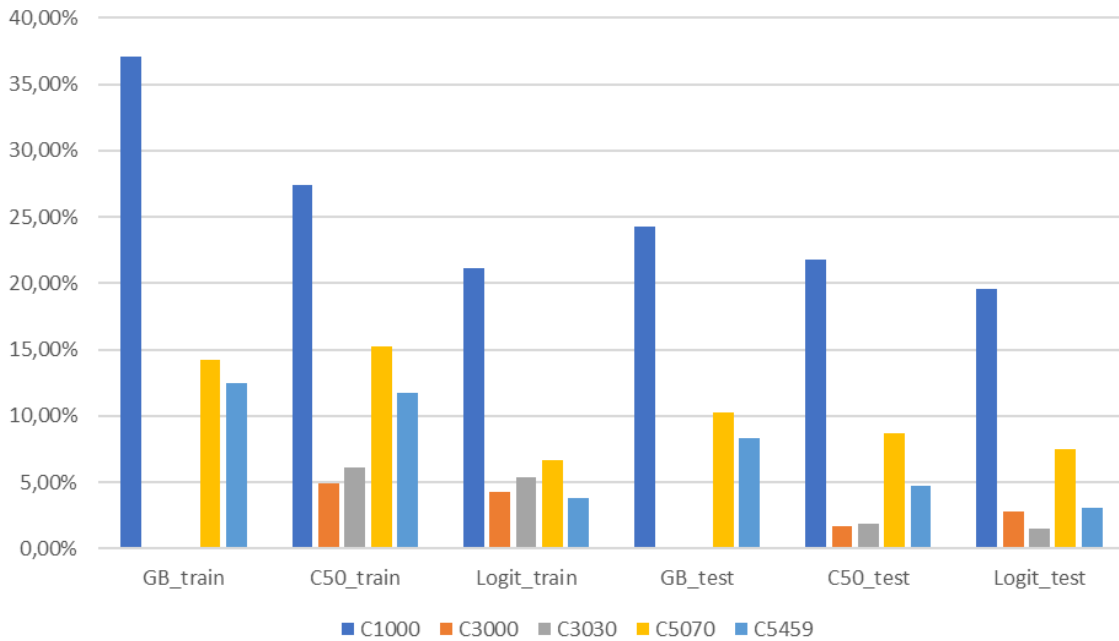


Figure 5. Comparison of methods according to the F-value

Continuing to review the results in Table 9, the model success status for categories of the train and test clusters according to the F-value is as follows:

For train data: It can be said that the C1000 category produces best train results via Gradient Boosting algorithm with a ratio of 37.1%. On the other hand, the best test results for the C3000, C3030, C5070 and C5459 categories were obtained with the C5.0 algorithm with ratios of 9%, 6.1%, 15.2% and 11.7% respectively.

Looking at the over-learning values of AUC; it can be said that the Gradient boosting algorithm has over-learned for the component C3030 and over-learned for the components C5070 and C3000 for the C5.0 algorithm.

As a result, if the EWS is processed for train data, it can be said that the best model success rates will be achieved when a hybrid structure will be built where according to F value for C1000 category the Gradient boosting algorithm is preferred and for the C3000, C3030, C5070 and C5459 categories the C5.0 algorithm is preferred. In other words, as a result of the application of the subject hybrid structure, creating warnings for the construction machinery to be taken to maintenance service before the occurrence of any failures is being performed in the most reliable way. Table 10 summarizes the subject situation.

Table 10. Structure of EWS according to the F-value for train data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	C5.0	C5.0	C5.0	C5.0

For test data: It can be said that the C1000 category produces best test results via Gradient Boosting algorithm with a ratio of 24.3%. On the other hand, the best test results for the C3030, C5070 and C5459 categories were obtained with the C5.0 algorithm. The ratios are 1.9%, 8.7% and 4.7% respectively. It can be said that the C3000 category yields the best test result on the logit regression model with a ratio of 2.8%. Figure 5 aims to explain this situation visually.

Table 11. Structure of EUS according to the F-Value for test data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	Logistic Regression	C5.0	C5.0	C5.0

Table 11 shows the architectural structure of the EWS according to the F-value that must be built for the test data.

Continuing to review the results in Table 9, the model success status for each category of the train and test clusters according to the AUC value is as follows:

For train data: It is clear that the Gradient Boosting algorithm produces the he best train results for all categories with ratios of 93.3% (C1000), 98.0% (C3000), 97.9% (C3030), 99.1% (C5070) and 99.5% (C5459) respectively. The architecture of the EWS is shown in Table 12.

For test data: It is clear that the Gradient Boosting algorithm produces the best test results in the C1000 and C3000 categories with ratios of 80.2% and 72.6% respectively. In the C3030, C5070 and C5459 categories, the Logit Regression model gives the best results with ratios of 66.3%, 81.0% and 93.7% respectively. Figure 6 aims to explain the results briefly. The architectural structure of the EWS is shown in Table 13.

Table 12. Structure of EUS according to the AUC-Value for train data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	Gradient Boosting	Gradient Boosting	Gradient Boosting	Gradient Boosting

Table 13. Structure of EWS according to the AUC-Value for test data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	Logistic Regression	Logistic Regression	Logistic Regression	Logistic Regression

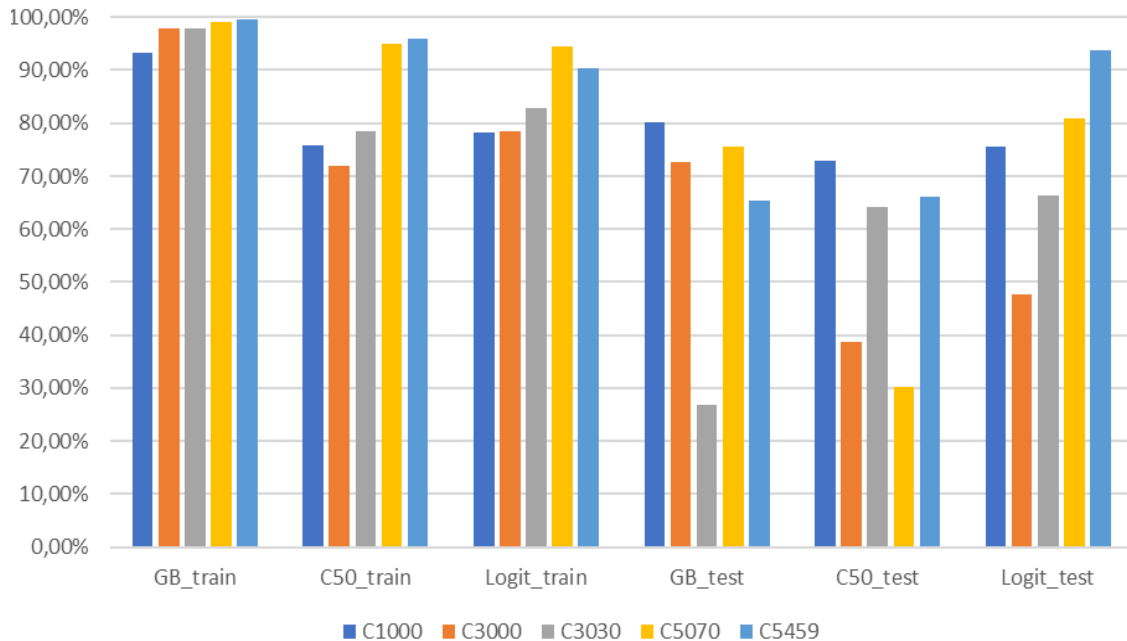


Figure 6. Comparison of methods according to the AUC value.

6. Conclusions

Data mining covers important points about the accurate, effective, goal-oriented use of the data and what to do with the information, rather than the size of the data. In this study model, where machine learning is used effectively, it can be seen that all 3 methods can yield better results for different categories when different success criteria are being taken into account. It is important here whether the criterion selected serves the performance logic that is wanted to be measured. For example, the AUC value attaches importance to the whole distribution as it deals with the sequencing of the wrong estimations, while the F value focuses only on false estimations above or below the specified limit value, as it is decision matrix-oriented.

In order to select the accurate model, it is necessary to set off from the train data set results. When the AUC value is considered as the criterion of success for the train data set results, it is clear that the best algorithm is Gradient Boosting for all categories. Gradient boosting yields the best result within the C1000 category for the F-Value while C5.0 yields the best result in other categories.

On the other hand, test results reveal, when the AUC criterion is taken into consideration, that the Logit and Gradient Boosting algorithms can be used together to produce the best results in certain categories. Taking F criteria into consideration, it can be seen that using all three model approaches together for different categories produces the best result.

In the model that has been used, the data from the sensors on the construction machinery are trained and learn the patterns obtained from machines under maintenance to predict the vehicles that may need possible maintenance in the future.

In brief, it is certain that the use of different algorithm approaches together will be beneficial in optimizing the results and reducing over-learning. A single combination for same categories is also possible just like the use of different algorithm results for different categories. In general, the main purpose of machine learning applications is learning patterns from data and using these patterns to create value. For this purpose, the decision support system has been evaluated by revealing valuable, usable information from the big data.

REFERENCES

- [1] Bertsche A, Engelhardt T. 2001. *Monitoring of The State of a Motor Vehicle Using Machine Learning*. US.
- [2] Chen T. 2014. *Introduction to Boosted Trees*. University of Washington. USA.
- [3] Classification: ROC Curve and AUC, 2019. <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc/> [1 December 2019].
- [4] Fiechter C-N, Göker M H. 2001. *Method and System for Condition Monitoring of Vehicles*. US.
- [5] Friedman J H. 1999. *Greedy Function Approximation a Gradient Boosting Machine*. USA.
- [6] Gradient Boosting. <https://devhunteryz.wordpress.com/2018/07/11/gradyan-arttirma-gradient-boosting/> [11 July 2018].
- [7] Hosmer D W, Lemeshow S. 2000, *Applied Logistic Regression*, 2nd Edition, Johnson Wiley & Sons Incorporation, New York, USA.
- [8] Jinyuan S, Yu Y, Zhicheng D, Xingping W. 2007. *Service Life Predicting Method and System for Machine and Vulnerable Component*. CN.
- [9] Menard S. 2002. *Applied Logistic Regression Analysis (Quantitative Applications in the Social Sciences)*. 2nd Edition, Volume 106, Beverly Hills, CA: Sage
- [10] Mertler C A, Vannatta R A. 2005. *Advanced and Multivariate Statistical Methods. Practical Application and Interpretation (third edition)*. United States: Pyczak Publishing.
- [11] Quinlan J R. 1996. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research* 4: 77-90.
- [12] Recommendations for implementing the strategic initiative industrie. <http://alvarestech.com/temp/tcn/CyberPhysicalSystems-Industrial4-0.pdf> [4 October 2019].
- [13] Richards W G, Martin K D, Lieurance W L, Kosler M-M. 2011. *Systems and Methods for Analyzing Equipment Failures and Maintenance Schedules*. USA.
- [14] Rumi E M, Zepf P J, Baird J, Lue C N, Rintjema T, Wu Y. 2005. *Industrial Process Data Acquisition and Analysis*. USA.
- [15] Yakut E, Gemici E. 2017. Predicting Stock Return Classification through LR, C5.0, CART and SVM methods, and Comparing the Methods Used: An Application at BIST in Turkey. *Ege Academic Review*, 17-4.
- [16] Yasukawa K, Adachi K, Uwatoko Kouki, Yamada N, Nakagawa E, Satonaga T. 2004. *Failure Diagnosis Method, Failure Diagnosis Apparatus, Conveyance Device, Image Forming Apparatus, Program and Storage Medium*. USA.
- [17] What is Industry. <http://www.armolis.com/akademi/endustr-i-40/> [1 December 2019].